Predicting Churn Rate

-Deepak Dogra

Contents

Sno Name page no

1 Introduction 3

1.1 Problem Statement 3

1.2 Data 3

2 Methodology 5

2.1 Pre Processing 5

2.1.1 Outlier Analysis 6

2.1.2 Feature Engineering 10

2.1.3 Normalization/standardization of datasets 10

2.1.4 Feature Selection 12

2.2 Modeling 17

2.2.1 Model Selection 17

2.2.2. Decision Tree 17

2.2.3 KNN Regression 19

2.2.4 Random Forest 21

2.2.5 Naïve Bayes 23

3 Conclusion 25

3.1 Visualisation 26

3.2 Model selection 30

4 Reference 31

Chapter 1

Introduction

* 1. Problem Statement

The objective of this project is to predict the churn score (loss of customers to competition).

* 1. Data

We are given a set of data containing details about various customer usage pattern such as total day calls,charges etc.Our aim is to develop a model which will help us to understand various predictors which are responsible for the churn and also predict the churn score .Given below is a sample of data set that we will be using to train our model.

Table 1.1 columns 1-7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| state | account length | area code | phone number | international plan | voice mail plan | number vmail messages |
| KS | 128 | 415 | 382-4657 | no | yes | 25 |
| OH | 107 | 415 | 371-7191 | no | yes | 26 |
| NJ | 137 | 415 | 358-1921 | no | no | 0 |
| OH | 84 | 408 | 375-9999 | yes | no | 0 |

Table 1.2 columns 8-14

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| total day minutes | total day calls | total day charge | total eve minutes | total eve calls | total eve charge | total night minutes |
| 265.1 | 110 | 45.07 | 197.4 | 99 | 16.78 | 244.7 |
| 161.6 | 123 | 27.47 | 195.5 | 103 | 16.62 | 254.4 |
| 243.4 | 114 | 41.38 | 121.2 | 110 | 10.3 | 162.6 |
| 299.4 | 71 | 50.9 | 61.9 | 88 | 5.26 | 196.9 |

Table 1.3 columns 15-21

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| total night calls | total night charge | total intl minutes | total intl calls | total intl charge | number customer service calls | Churn |
| 91 | 11.01 | 10 | 3 | 2.7 | 1 | False. |
| 103 | 11.45 | 13.7 | 3 | 3.7 | 1 | False. |
| 104 | 7.32 | 12.2 | 5 | 3.29 | 0 | False. |
| 89 | 8.86 | 6.6 | 7 | 1.78 | 2 | False. |

The observation regarding the dataset was as follows-

1)There are 20 independent variables out of which 5 are categorical predictors and 15 are continuous predictors.

2)There is one dependent variables namely- Churn.

Chapter 2

Methodology

2.1 Pre Processing

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

Various preprocessing techniques are applied to clean the data such as if there are missing values in the data then missing value analysis is done. Outlier analysis is done to remove the exceptional data sets from the data. Feature selection is done to remove the multicollinearity and to reduce the size of the dataset. Next step comes the feature scaling in which data is either normalized or standardized. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

2.1.1 Outlier Analysis

In statistics, an **outlier** is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. Thus it is very important to remove an outlier from our data set. Outliers can have many anomalous causes. A physical apparatus for taking measurements may have suffered a transient malfunction. There may have been an error in data transmission or transcription. Outliers arise due to changes in system behavior, fraudulent behavior, human error, instrument error or simply through natural deviations in populations. A sample may have been contaminated with elements from outside the population being examined. Alternatively, an outlier could be the result of a flaw in the assumed theory, calling for further investigation by the researcher. Additionally, the pathological appearance of outliers of a certain form appears in a variety of datasets, indicating that the causative mechanism for the data might differ at the extreme end (King effect).

The first analysis to be done was outlier analysis on the data set and following box plot was plotted -

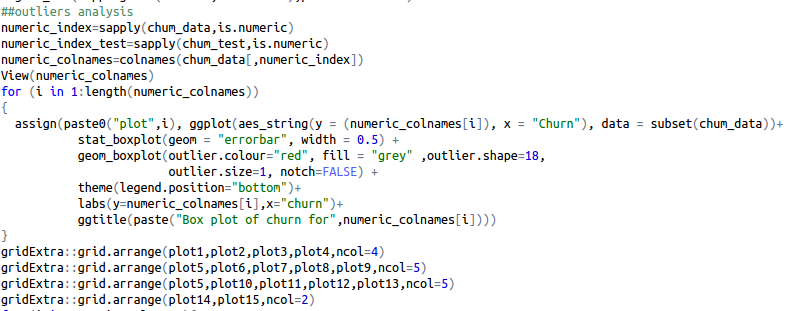
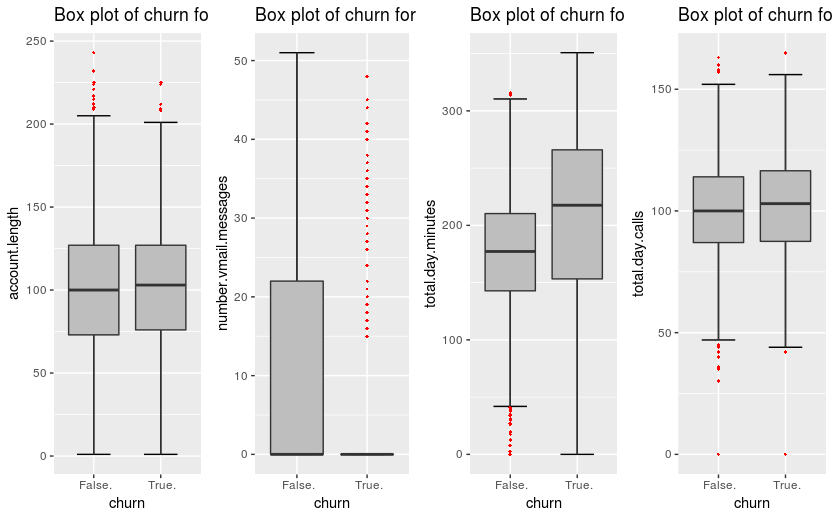
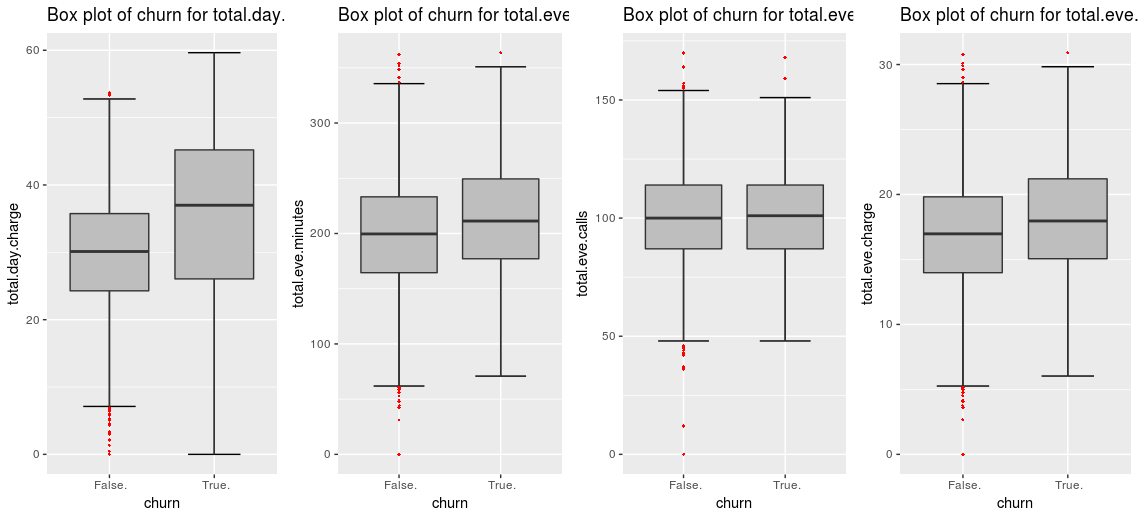
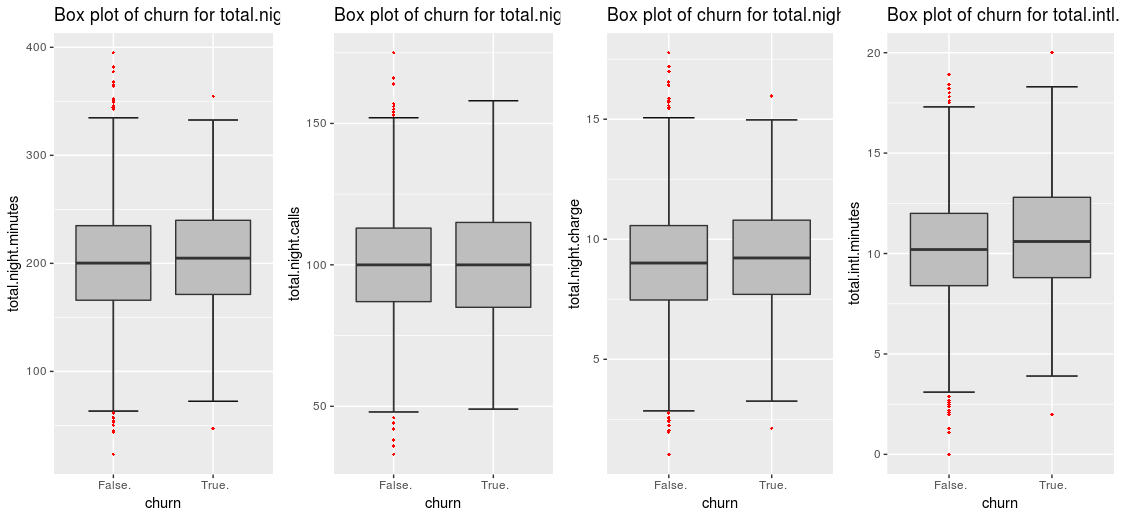
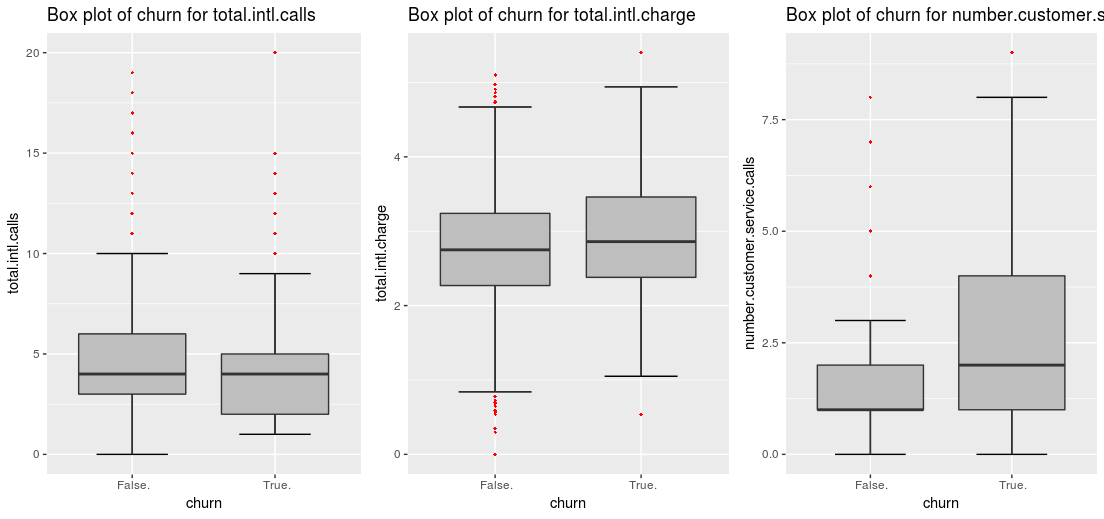


Fig 2.1 boxplots for the all 15 predictors









As we can clearly see that there are outliers present in many predictors.

The outliers were removed from the dataset as outliers are exceptional datapoints which can contribute to errors in the models and also reduce the accuracy of the model.

2.1.2 Feature Engineering

**Feature engineering** is the process of using [domain knowledge](https://en.wikipedia.org/wiki/Domain_knowledge) of the data to create [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) that make [machine learning](https://en.wikipedia.org/wiki/Machine_learning)algorithms work. Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive.

Here we created 8 new variables on the basis of domain knowledge that are-

a)total day charge per min

b)total eve charge per min

c)total night charge per min

d)total intl charge per min

e)total day avg min percall

f)total eve avg min percall

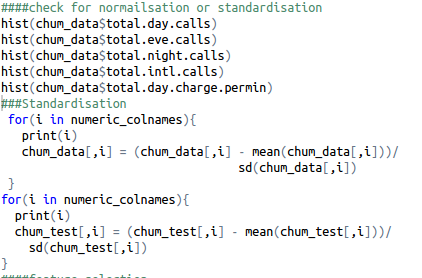
g)total night avg mincall

h)total intl avg min percall

2.1.3 Normalization/standardization of datasets

In the overall knowledge discovery process, before data mining itself, data preprocessing plays a crucial role. One of the first steps concerns the normalization of the data. This step is very important when dealing with parameters of different units and scales. For example, some data mining techniques use the Euclidean distance. Therefore, all parameters should have the same scale for a fair comparison between them.

The criteria for choosing between normalization and standardization is that if data is normally distributed then standardization should be done or else normalization is better option.



2.1.4 Feature Selection

The central premise when using a feature selection technique is that the data contains some features that are either *redundant* or *irrelevant*, and can thus be removed without incurring much loss of information. *Redundant* and *irrelevant* are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated.

Feature Selection is a very critical component in a Data Scientist’s workflow. When presented data with very high dimensionality, models usually choke because

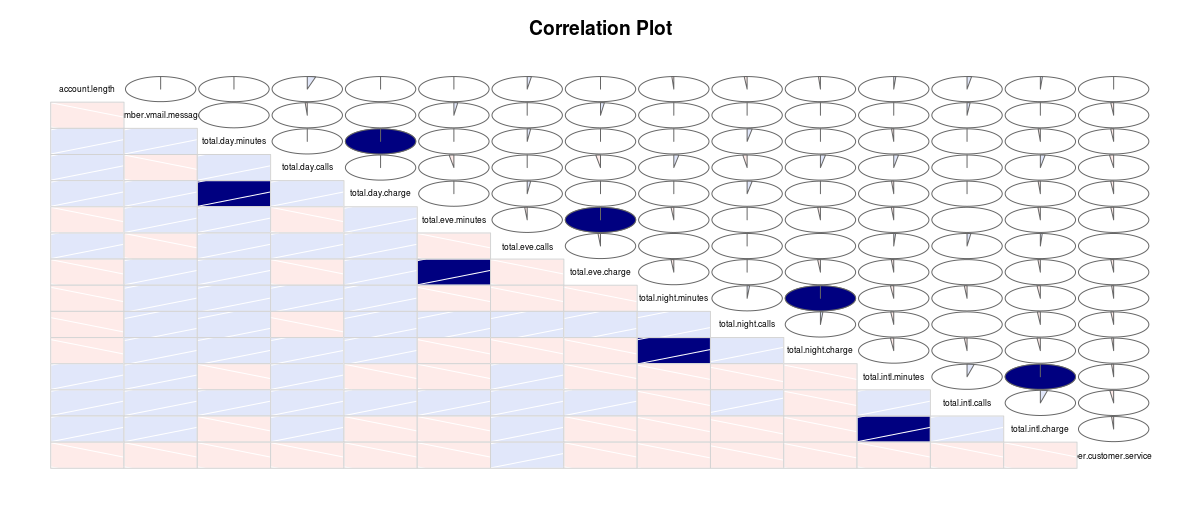
1. **Training time** increases exponentially with number of features.
2. Models have increasing risk of **over fitting**with increasing number of features.

Feature Selection methods helps with these problems by reducing the dimensions without much loss of the total information. It also helps to make sense of the features and its importance.

In our data set we first carried out correlation analysis on categorical predictors so as to deal with the problem of multicollinearity as well as remove predictors which are supplying redundant information.

The correlation plot plotted is as given below-

Fig 2.2 correlation plot



The intensity of blue color indicates how positively correlated the predictor is with the other variable while the intensity of pink color indicates how negatively correlated the predictor is with the other variable.

Studying the correlation plot we get the understanding that the total call charges of day,night,evening and international ,all are highly correlated with their respective total call predictors Thus we can drop all of these variable as they are carrying same information .

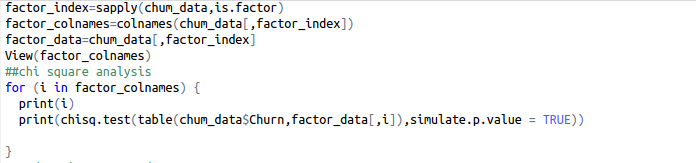
Chi –square test analysis

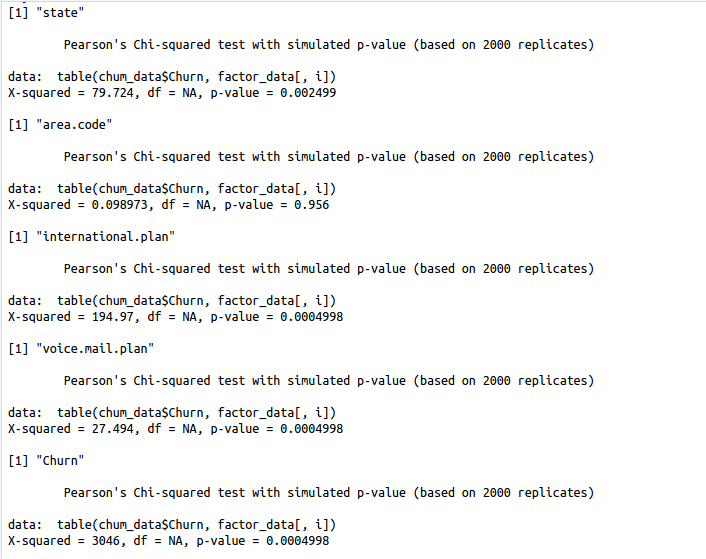
The next step was to carry out chi square test analysis. This analysis is carried out on categorical predictors. In this analysis two assumptions are made first, one is that there should be high dependency between dependent and independent variable .the second assumption is that there should be no dependency between two independent variables. After this two hypothesis are stated-

Null hypothesis- the two predictors are highly dependent

Alternate hypothesis- the two predictors are highly independent.

In this analysis we get p value ,if p value is less then 0.05(significant value) then we reject the null hypothesis and say that the two predictors are highly dependent.if p value is greater than 0.05, then we accept the null hypothesis and say that the two predictors are independent.





In our test we found out that the area code predictor has a p value greater than 0.05,showing it is highly independent . Thus the final decision was to drop area code predictor from our dataset.

Thus following predictors were dropped from the data set before using the dataset to train the models.

1) phone-number- not significant for the analysis.

2) area code= shows high independency with target as indicated in correlation plot.

3) total day charge,total night charge,total eve charge and total intl charge-

2.2 Modeling

2.2.1 Model Selection

We have 20 independent variable and 1 dependent variable. As our dependent variable is categorical .we have to choose a classification model. The first model we chose for our study is decision tree.

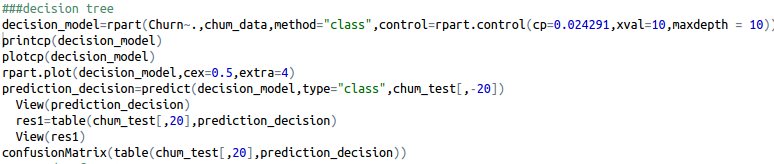
2.2.2 Decision Tree

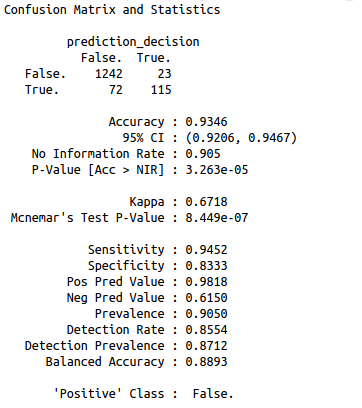
Decision trees, as the name suggest, is a tree shaped visual representation of one can reach to a particular decision by laying down all options and their probability of occurrances.Decision trees are extremely easy to understand and interpret. At each node of the tree, one can interpret what would be the consequence of selecting that node or option.

#decision tree model

desicion\_model = rpart(cnt ~ ., data = trainingdata, method = "class"

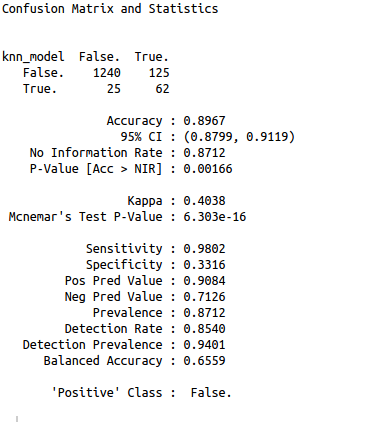
The important observation was made when we looked at the summary of this model –



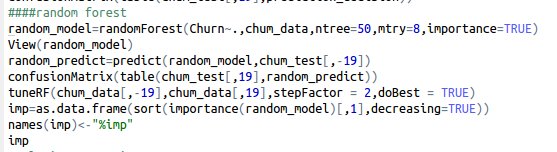


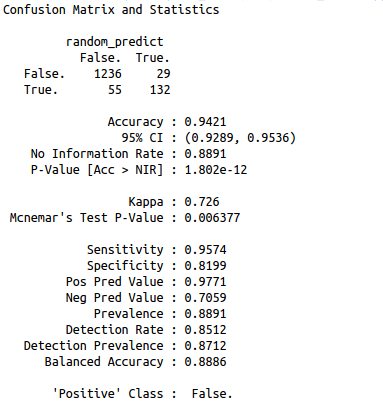
2.2.3 KNN Regression



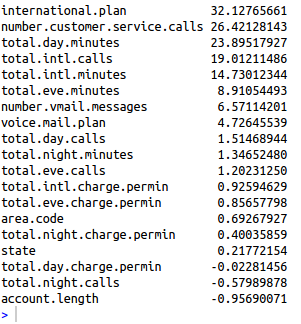


2.2.4 Random Forest



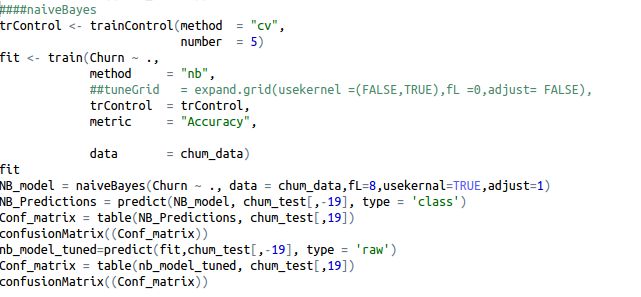


Importance of predictors as shown by the model-



As shown international plan ,number customers service calls ,total day minutes,total intl calls and total intl minutes are important predictors.

2.2.4 Naïve Bayes

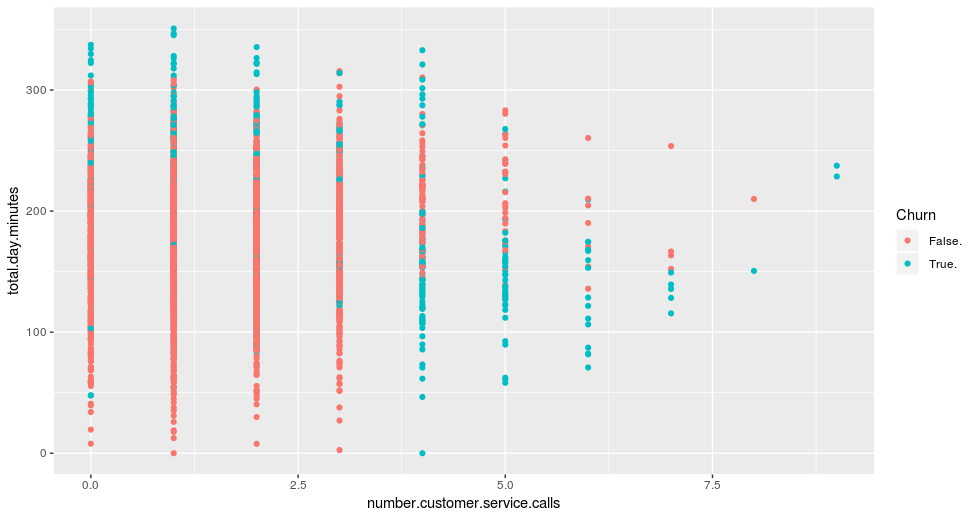


Chapter 3

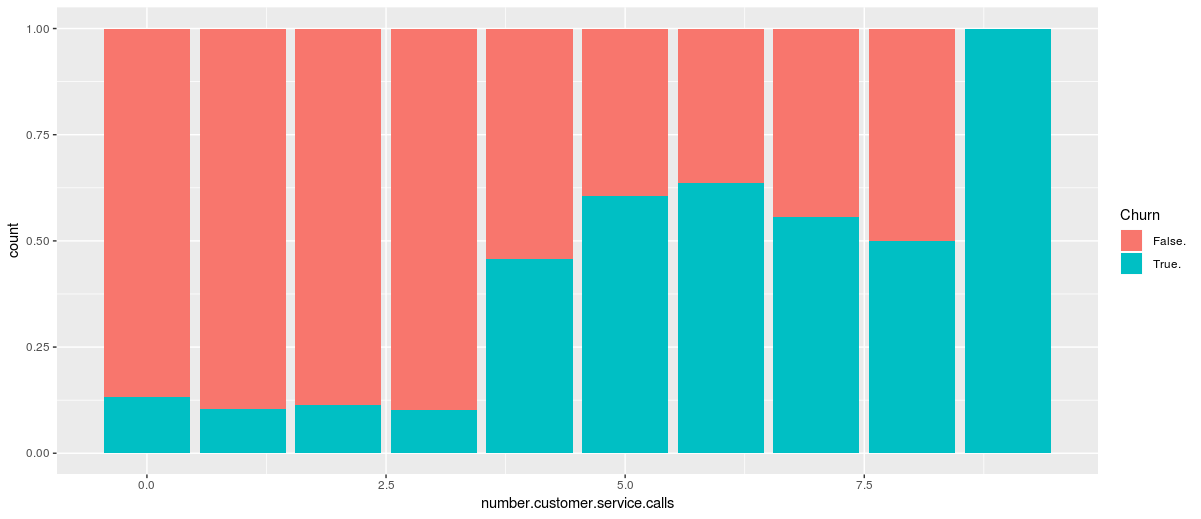
Conclusion

|  |  |  |
| --- | --- | --- |
| Name | Accuracy | FNR |
| Decision tree | 93.46 | 38% |
| Knn | 89.67 | 28% |
| Random Forest | 94.21 | 29% |
| Naïve Bayes | 88.71 | 37% |

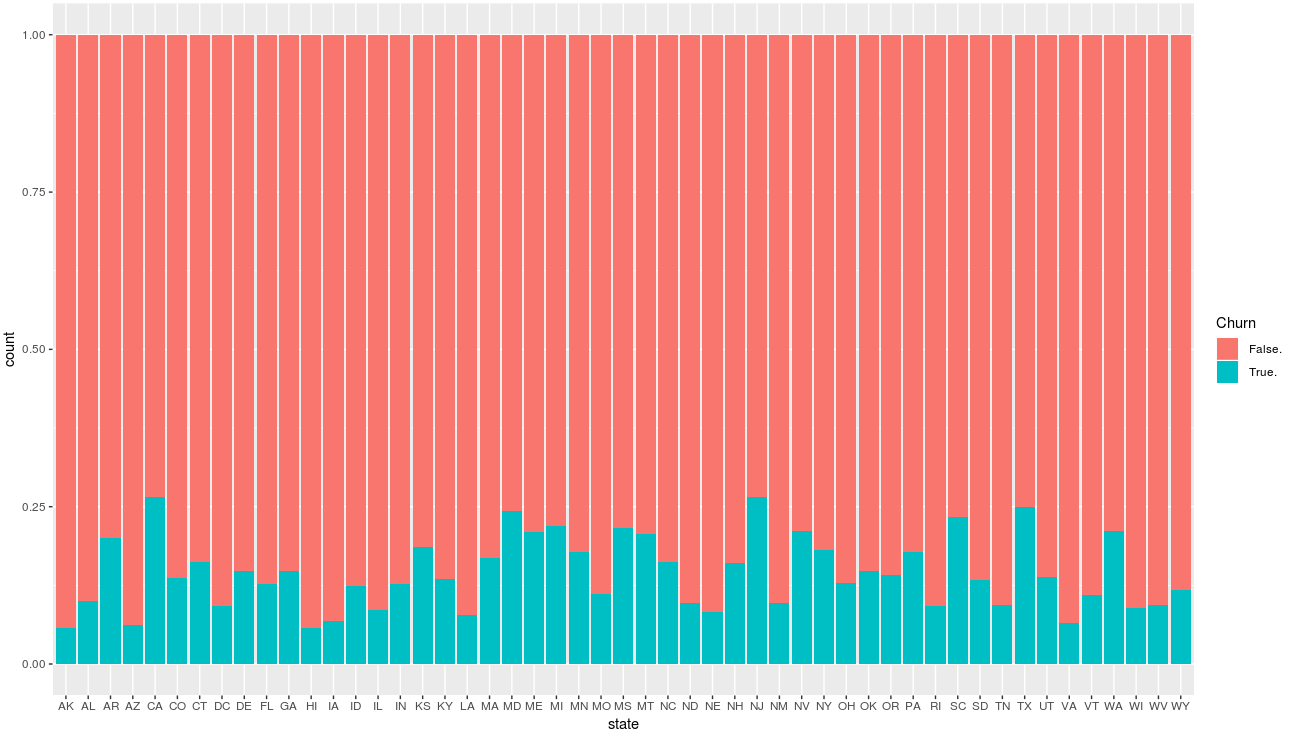
Visualizations



As seen from the scatter plot more customers are tend to churn out if number of customer calls are made more than 4 . So making this customers a high priority and solving their problems can help in reducing the churn rate.

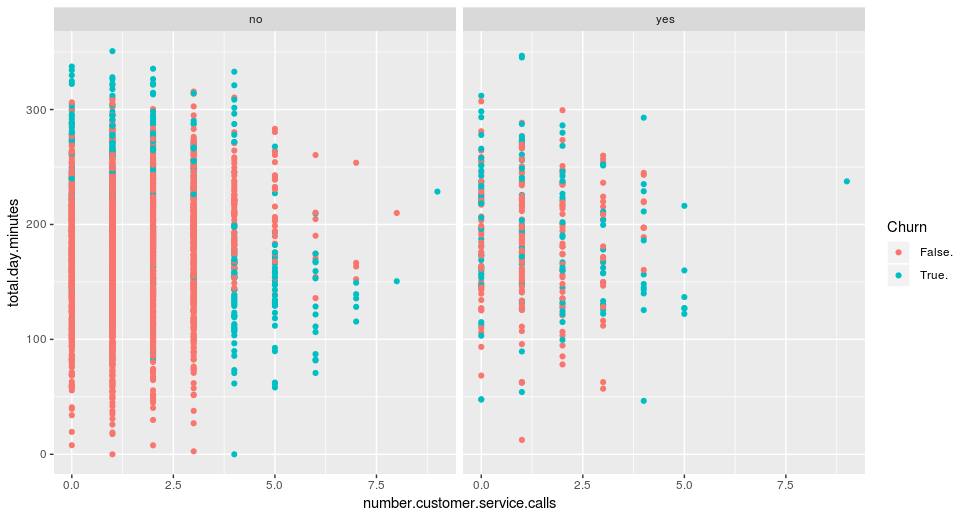


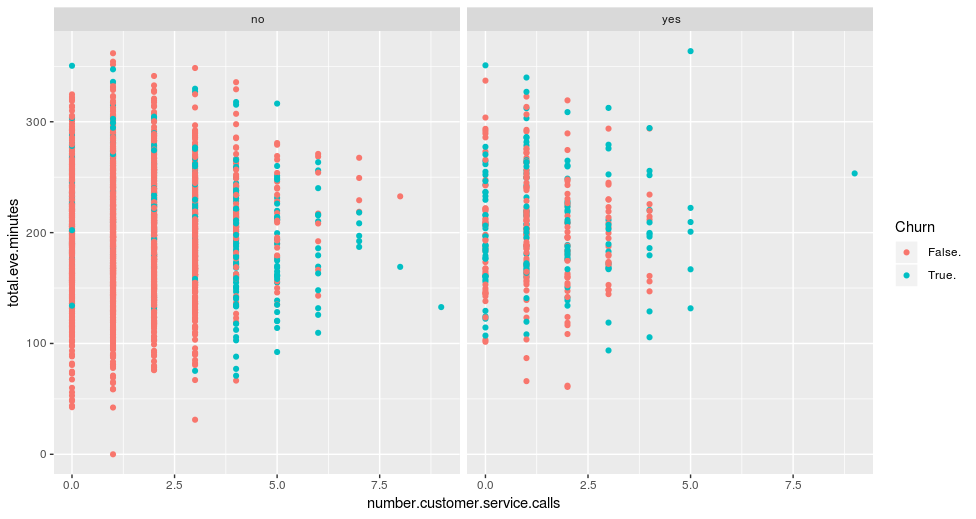
Bar plot of statewise churn rates{\displaystyle {\mbox{M}}={\frac {1}{n}}\sum \_{t=1}^{n}\left|{\frac {A\_{t}-F\_{t}}{A\_{t}}}\right|,}

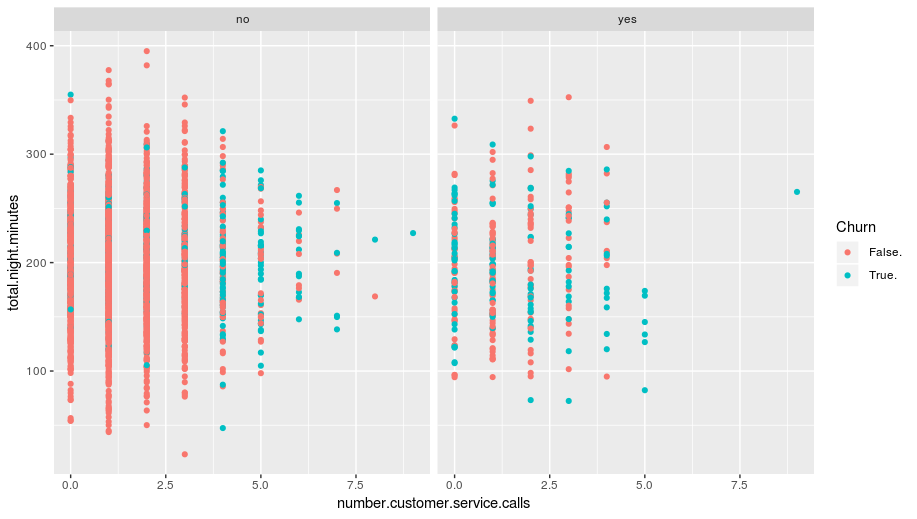


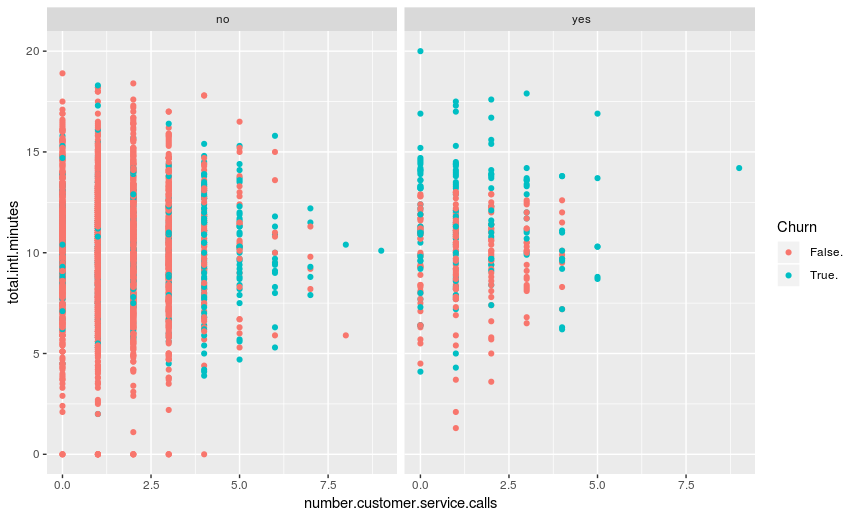
Scatter plot of total day,evening,night and international minutes of calling wrt to number of service calls made

.





.



3.2 Model Selection

We can see that the random forest gives the best compromise between accuracy and FNR,Thus the preferred model for this problem statement is random forest.

References

1. Wikipedia

2. Edwisor

3. YouTube