By: Gohel Ravi VashramBhai

  2nd February 2019

Churn Reduction

Contents Page No.

1.Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2

1.1 Problem Statement\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2

1.2 Data\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2

2.Methodology\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6

2.1Pre Processing\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6

2.1.1Missing Value Analysis\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6

2.1.2Outlier Analysis\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6

2.1.3Feature Selection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_9

2.1.4Feature scaling\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11

2.2. Modeling\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14

2.2.1. Decision tree for classification\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14

2.2.2. Random Forest\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_15

2.2.3. KNN Implementation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_16

2.2.4. Naïve Bayes\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17

2.2.5 Logistic Regression\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_18

3. Conclusion \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_19

**1.Introduction**

1.1Preoblem Statement:

Churn (loss of customers to competition) is a problem for companies because it is more expensive to acquire a new customer than to keep your existing one from leaving. This problem statement is targeted at enabling churn reduction using analytics concepts.

The objective of this Project is to predict customer behavior. Using given data set which has customer’s usage patterns and if the customer has moved or not, developed an algorithm to predict the churn score based on usage patterns.

1.2DATA

Our task is to build classification models which will predict whether the customer will churn or not depending on multiple factors. Given below is a sample of the data set that we are using to predict the churn. Actual data has 5000 observations and 21, numeric and categorical variables:

Train Dataset.

* Train Dataset

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| state | account length | area code | phone number | international plan | voice mail plan | number vmail messages | total day minutes | total day calls | total day charge | total eve minutes |
| KS | 128 | 415 | 382-4657 | no | yes | 25 | 265.1 | 110 | 45.07 | 197.4 |
| OH | 107 | 415 | 371-7191 | no | yes | 26 | 161.6 | 123 | 27.47 | 195.5 |
| NJ | 137 | 415 | 358-1921 | no | no | 0 | 243.4 | 114 | 41.38 | 121.2 |
| OH | 84 | 408 | 375-9999 | yes | no | 0 | 299.4 | 71 | 50.9 | 61.9 |
| OK | 75 | 415 | 330-6626 | yes | no | 0 | 166.7 | 113 | 28.34 | 148.3 |

Column (1-11)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| total eve calls | total eve charge | total night minutes | total night calls | total night charge | total intl minutes | total intl calls | total intl charge | number customer service calls | Churn |
| 99 | 16.78 | 244.7 | 91 | 11.01 | 10 | 3 | 2.7 | 1 | False. |
| 103 | 16.62 | 254.4 | 103 | 11.45 | 13.7 | 3 | 3.7 | 1 | False. |
| 110 | 10.3 | 162.6 | 104 | 7.32 | 12.2 | 5 | 3.29 | 0 | False. |
| 88 | 5.26 | 196.9 | 89 | 8.86 | 6.6 | 7 | 1.78 | 2 | False. |
| 122 | 12.61 | 186.9 | 121 | 8.41 | 10.1 | 3 | 2.73 | 3 | False. |

Column (12-21)

* Test Dataset

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| state | account length | area code | phone number | international plan | voice mail plan | number vmail messages | total day minutes | total day calls | total day charge | total eve minutes |
| HI | 101 | 510 | 354-8815 | no | no | 0 | 70.9 | 123 | 12.05 | 211.9 |
| MT | 137 | 510 | 381-7211 | no | no | 0 | 223.6 | 86 | 38.01 | 244.8 |
| OH | 103 | 408 | 411-9481 | no | yes | 29 | 294.7 | 95 | 50.1 | 237.3 |
| NM | 99 | 415 | 418-9100 | no | no | 0 | 216.8 | 123 | 36.86 | 126.4 |
| SC | 108 | 415 | 413-3643 | no | no | 0 | 197.4 | 78 | 33.56 | 124 |

Column (1-11)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| total eve calls | total eve charge | total night minutes | total night calls | total night charge | total intl minutes | total intl calls | total intl charge | number customer service calls | Churn |
| 73 | 18.01 | 236 | 73 | 10.62 | 10.6 | 3 | 2.86 | 3 | False. |
| 139 | 20.81 | 94.2 | 81 | 4.24 | 9.5 | 7 | 2.57 | 0 | False. |
| 105 | 20.17 | 300.3 | 127 | 13.51 | 13.7 | 6 | 3.7 | 1 | False. |
| 88 | 10.74 | 220.6 | 82 | 9.93 | 15.7 | 2 | 4.24 | 1 | False. |
| 101 | 10.54 | 204.5 | 107 | 9.2 | 7.7 | 4 | 2.08 | 2 | False. |

Column (12-22)

As you can see in the table above we have the following 20 variables, using which we have to correctly predict the Churn:

1. state
2. account length
3. area code
4. phone number
5. international plan
6. voice mail plan
7. number vmail messages
8. total day minutes
9. total day calls
10. total day charge
11. total eve minutes
12. total eve calls
13. total eve charge
14. total night minutes
15. total night calls
16. total night charge
17. total intl minutes
18. total intl calls
19. total intl charge
20. number customer service calls

**2.Methodology**

2.1Pre Processing

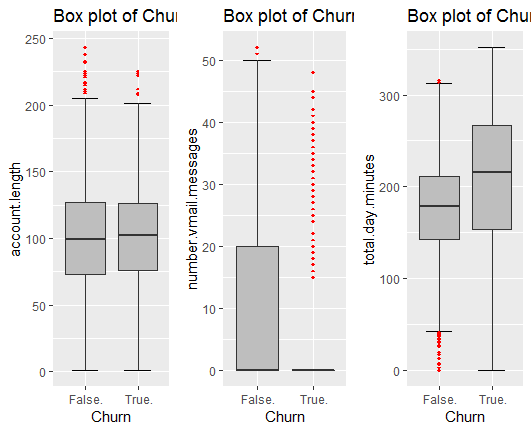
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process, we will first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

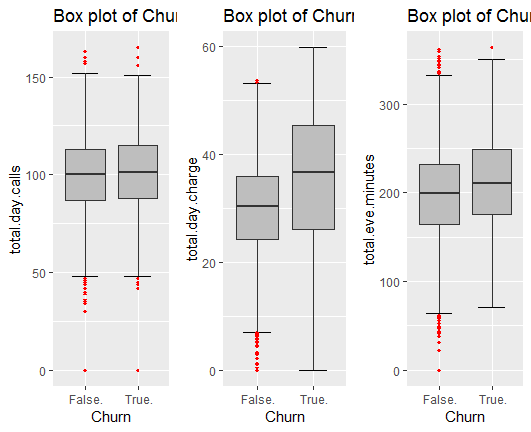
2.1.1 Missing Value Analysis:

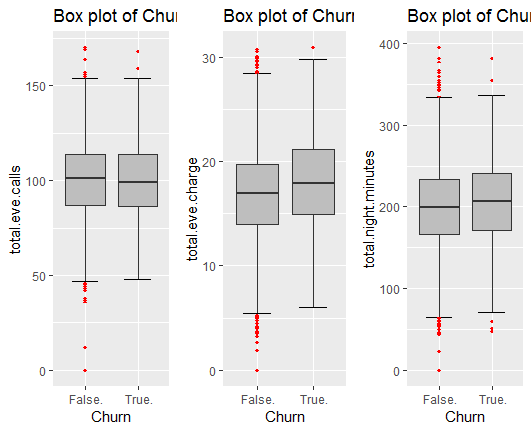
There is no any missing value in given datasets so there is no need for missing value analysis and imputation of missing values.

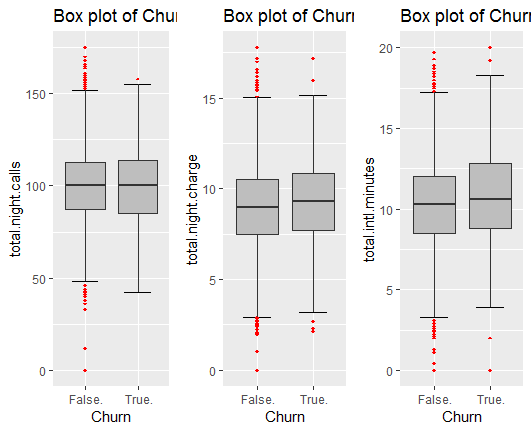
2.1.2. Outlier Analysis:

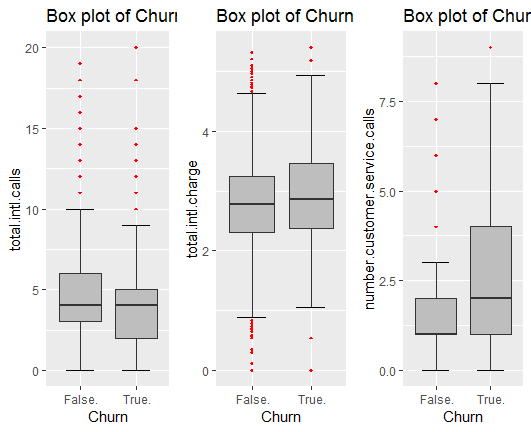
Here we are using a simple method to detect outliers which is box plot method. After detecting them we will replace them using appropriate method. Following are the box plot graphs of continuous variables.









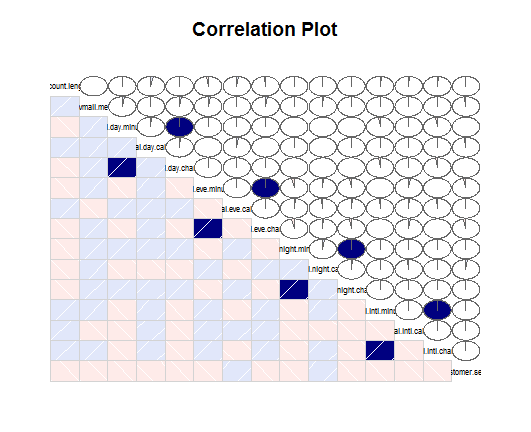


2.1.3 Feature Selection:

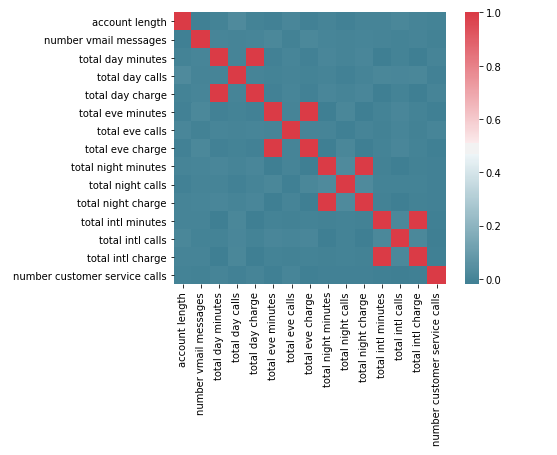
If the two or more features of the data set carry the same information and If we use both of them to develop a model, then it will consume the more time and memory so here feature selection must be needed. Thumb rule for feature selection is: there must be high co-relation between dependent and independent variable and there must be low co relation between two independent variables. Following are the co relation graph for the all continuous variables.

Using chi-square test we can check the co relation between categorical variable.

By both the test we will drop some variables for further analysis.

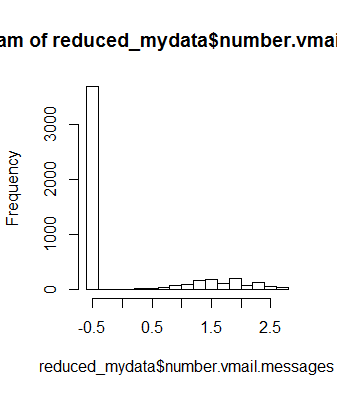
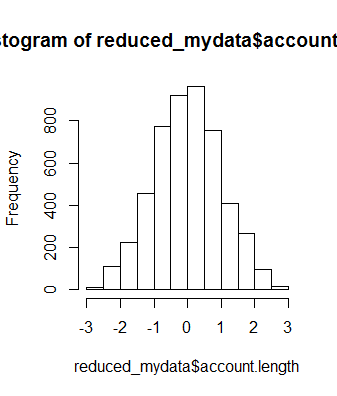


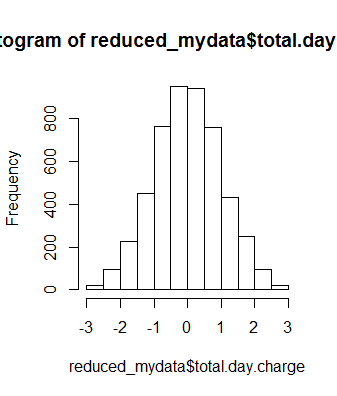
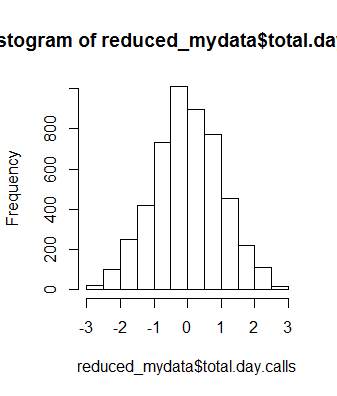
(Correlation graph in R and Python)

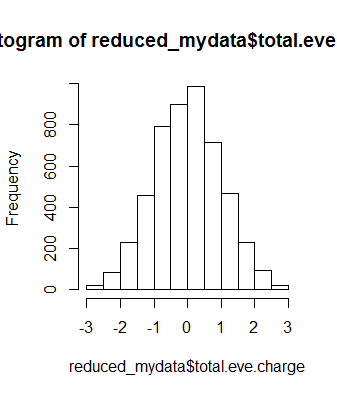
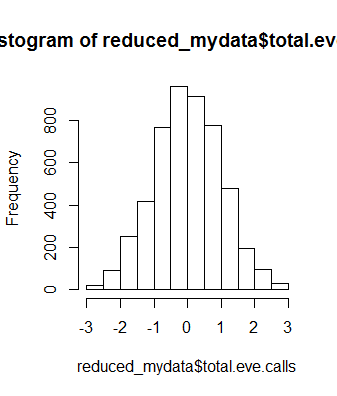


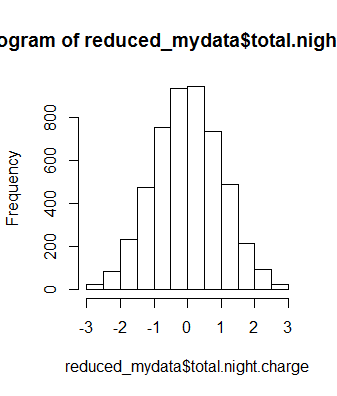
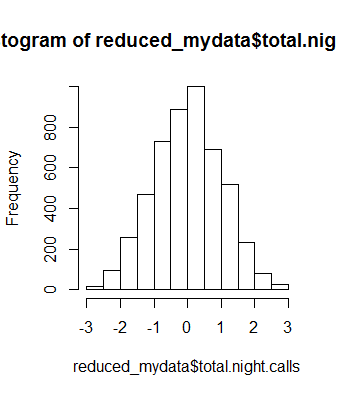
2.1.4Feature Scaling

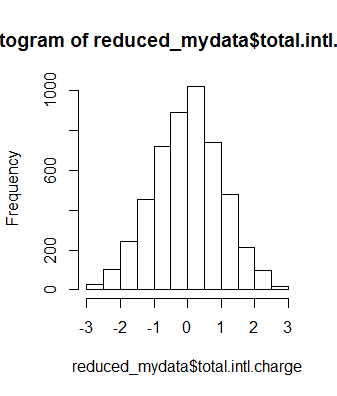
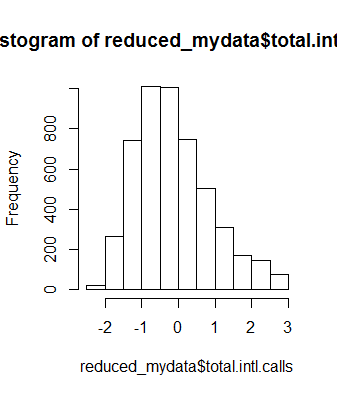
Data distribution also affects the performance the model. We can use normalization or standardization method to scale the data.by plotting histogram graph of the data we can see the data distribution. following are the histogram of the all continuous data.

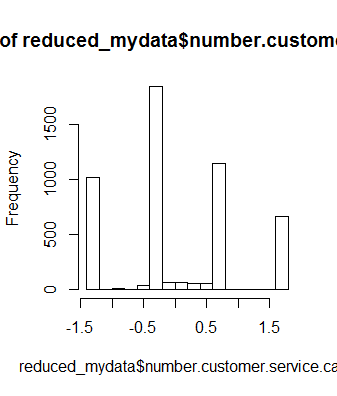










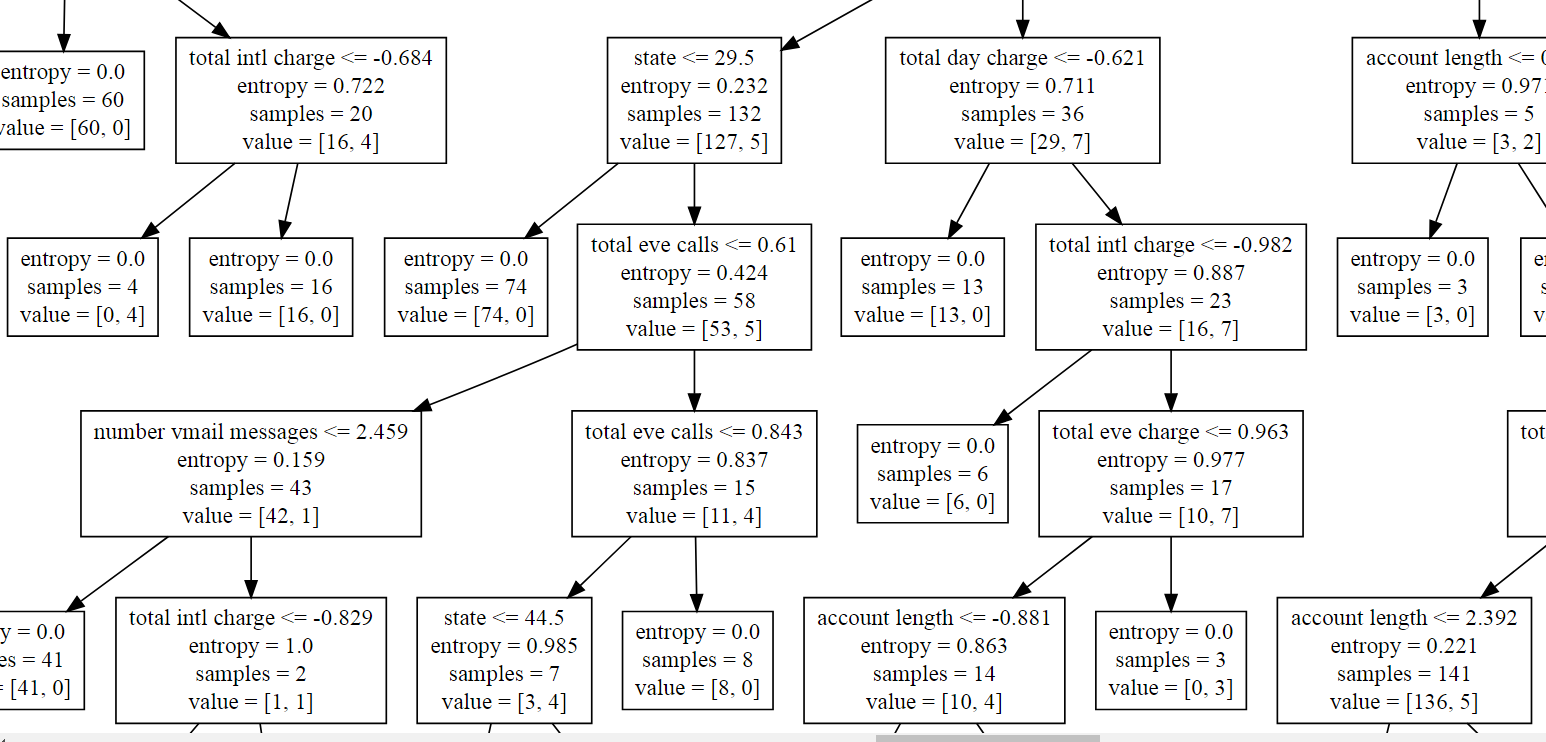


As we can see most of the data are distributed normally, so here we are using standardization method to scale the data.

**2.2 Modeling**

2.2.1 Decision Tree for classification:

Following is the partial image of the rule based tree generated by the decision tree model:



|  |  |  |
| --- | --- | --- |
| Decision Tree (Classification) | R | Python |
| Accuracy | 94.06 % | 91.96 % |
| False Negative Rate | 43.30 % | 33.03 % |

Results:

2.2.2 Random Forest:

|  |
| --- |
| [1] "account.length<=-2.51524009823046 & international.plan %in% c('1') & number.vmail.messages<=-0.110360215414717 & total.day.charge<=1.60503235232378 & total.eve.charge<=0.730176149691144 & total.eve.charge<=-1.18076808450286"  [2] "account.length<=-2.51524009823046 & international.plan %in% c('1') & number.vmail.messages<=-0.110360215414717 & total.day.charge<=1.60503235232378 & total.eve.charge<=0.730176149691144 & total.eve.charge>-1.18076808450286"  Results: |
|  |
| |  | | --- | |  | |

|  |  |  |
| --- | --- | --- |
| Random Forest | R | Python |
| Accuracy | 92.08 % | 92.92 % |
| False Negative Rate | 52.67 % | 50.89 % |

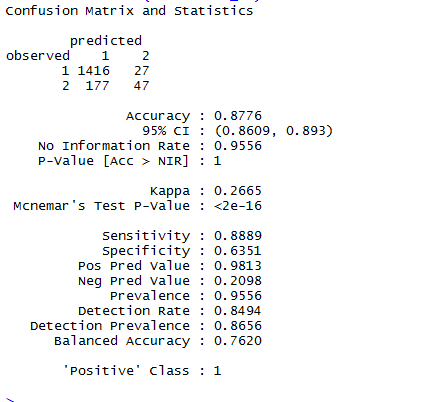
Followings are some rules generated by the model:

2.2.3 KNN Implementation:

Following are the results of the model

|  |  |  |
| --- | --- | --- |
| KNN Implementation | R | Python |
| Accuracy | 86.98 % | 86.74 % |
| False Negative Rate | 33.33 % | 96.42 % |

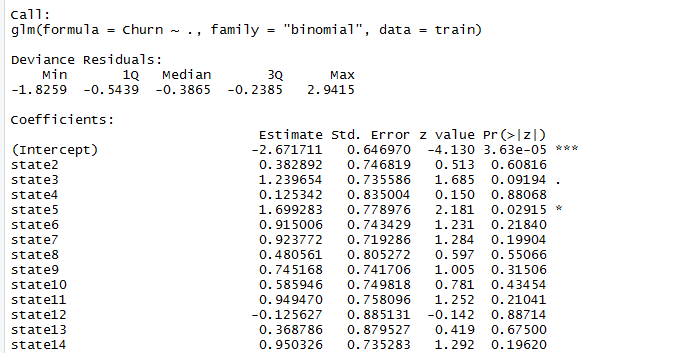
2.2.4 Naïve Bayes:



Following are the results of the model:

|  |  |  |
| --- | --- | --- |
| Naïve Bayes | R | Python |
| Accuracy | 87.76 % | 85.84 % |
| False Negative Rate | 79.01 % | 60.26 % |

2.2.5 Logistic Regression:



Followings are the results of the model

|  |  |  |
| --- | --- | --- |
| Logistic Regression | R | Python |
| Accuracy | 87.43 % | 87.06 % |
| False Negative Rate | 80.80 % | 81.15 % |

**3. Conclusion**

Following are the results of the different model in ‘R’ and ‘Python’.

|  |  |  |
| --- | --- | --- |
| R | | |
|  | Accuracy | False Negative Rate |
| 1.Decision Tree(Classification) | 94.06 % | 43.03 % |
| 2.Random Forest | 92.08 % | 52.67 % |
| 3.KNN Implementation | 86.98 % | 80.80 % |
| 4.Naive Bayes | 87.76 % | 33.33 % |
| 5.Logistic Regression | 87.43 % | 79.01 % |

|  |  |  |
| --- | --- | --- |
| Python | | |
|  | Accuracy | False Negative Rate |
| 1.Decision Tree(Classification) | 91.96 % | 33.03 % |
| 2.Random Forest | 92.92 % | 50.89 % |
| 3.KNN Implementation | 86.74 % | 96.42 % |
| 4.Naive Bayes | 85.84 % | 60.26 % |
| 5.Logistic Regression | 87.06 % | 81.15 % |

Accuracy of the model is important but we can’t ignore the false negative rate of the model so from the above summary we can choose any of the model. whose accuracy is greatest and having minimum false negative rate.