Churn Reduction

*Diksha P. Jagre*

**Contents**

**1** **Introduction**   **2**

1.1 Problem Statement …............................................................................................................. 3

1.2 Data ….................................................................................................................................... 3

**2** **Methodology**  **4**

2.1 Pre -Processing ….................................................................................................................. 4

2.1.1 Outlier Analysis -------------------------------------------------------------------------- 4

2.1.2 Feature Selection ------------------------------------------------------------------------- 12

2.2 Modeling ----------------------------------------------------------------------------------------------- 15

2.2.1 Model Selection -------------------------------------------------------------------------- 15

2.2.2 Logistic Regression----------------------------------------------------------------------- 15

2.2.3 Decision /Classification Tree------------------------------------------------------------ 19

2.2.4 Regression---------------------------------------------------------------------------------- 21

**3**  **Conclusion** 22

3.1 Model Evaluation -------------------------------------------------------------------------------------- 22

3.1.1 Confusion Matrix-------------------------------------------------------------------------- 22

3.1.2 False Negative Rate ---------------------------------------------------------------------- 24

3.2 Model Selection --------------------------------------------------------------------------------------- 25

**Appendix A - Extra Figures 26**

**Appendix B - R Code 37**

Boxplot for all predictor variables (Fig: 2.1) ----------------------------------------------------------- 37

Correlation Plot (Fig: 2.2) --------------------------------------------------------------------------------- 37

Random Forest (Fig: 2.4) ---------------------------------------------------------------------------------- 37

Logistic Regression Plot (Fig: 2.6)------------------------------------------------------------------------ 37

Decision Tree Using C5.0 and CART (Fig: 2.7 & Fig: 2.8) -------------------------------------------38

Complete R File .-------------------------------------------------------------------------------------------- 46

**Chapter 1**

**Introduction**

**1.1 Problem Statement**

The objective of this Case is to predict customer behaviour .We are providing you a public dataset that has customer usage pattern and if the customer has moved or not. We expect you to develop an algorithm to predict the churn score based on usage pattern.

**1.2 Data**

Our task is to build a classification models which will classify the customer behaviour depending on the churn score. Given below is a sample of data set that we are using to predict the customer behaviour:

Data Sets -

1) Test\_data.csv 2) Train\_data.csv

As you can see in the table below we have the following 17 variables, using which we have to correctly predict the customer behaviour:

|  |
| --- |
| *Account length* |
| *International plan* |
| *Voice mail plan* |
| *Number of voice mail messages* |
| *Total day minutes* |
| *Total day calls* |
| *Total day charge* |
| *Total evening minutes* |
| *Total evening calls* |
| *Total evening charge* |
| *Total night minutes* |
| *Total night calls* |
| *Total night charge* |
| *Total international minutes* |
| *Total international calls* |
| *Total international charge* |
| *Number of customer service calls* |

**Figure 1.1 Table : Predictor Variables**

**Chapter 2**

**Methodology**

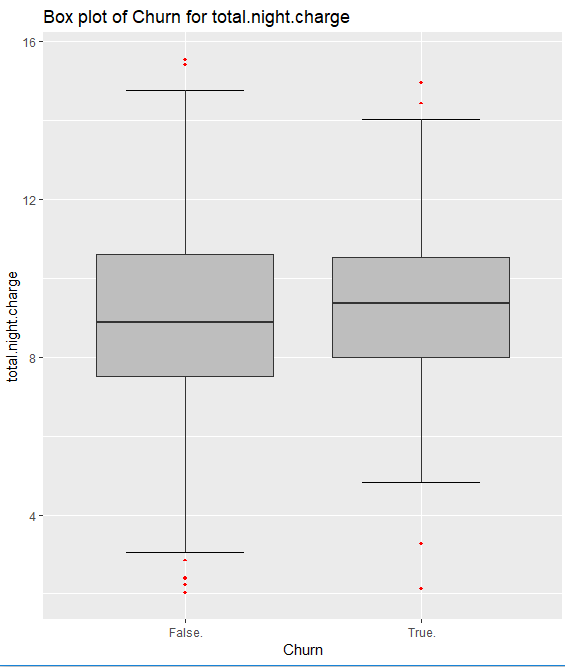
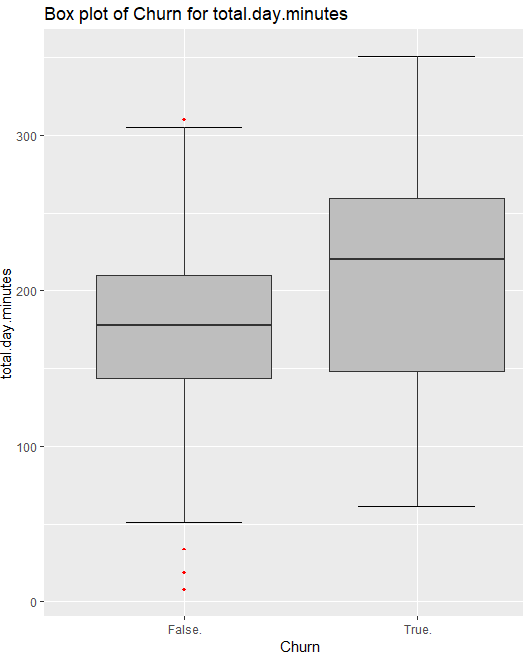
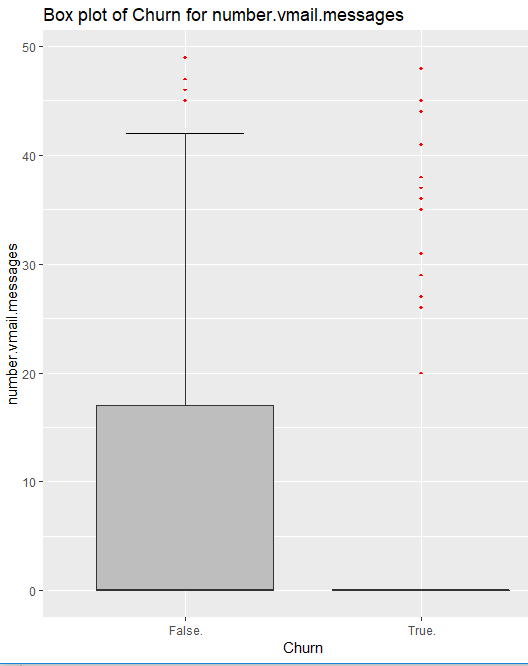
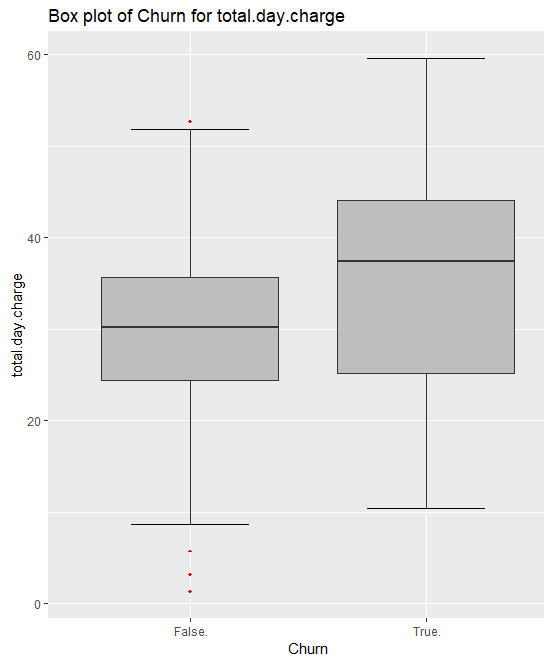
**2.1 Pre-Processing**

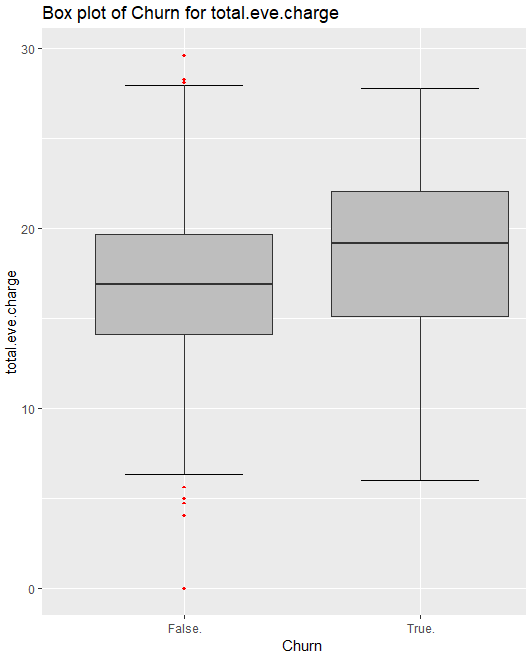
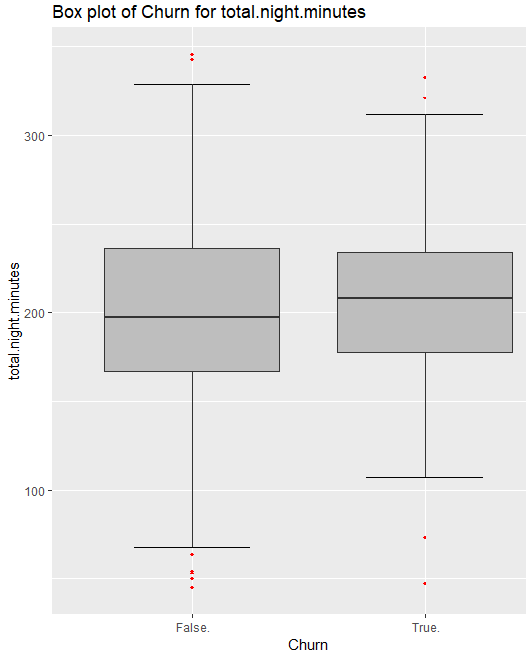
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is 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.

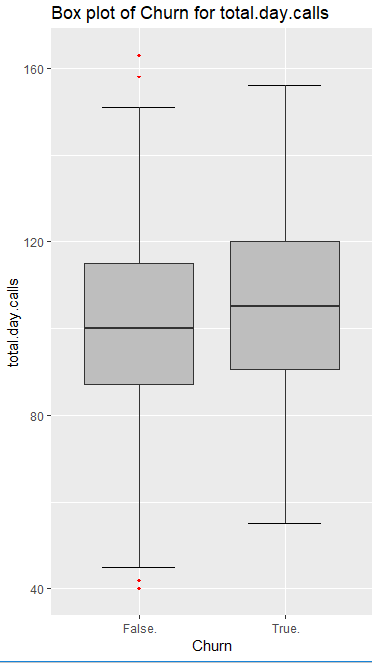
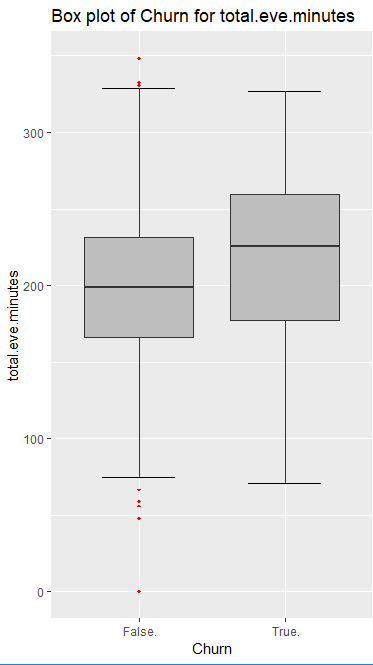
Here the dataset will be already given in probability like in test and train data. So, we will first try to look our train and test dataset for checking how many variables dataset contain and this dataset have balance or imbalance dataset. If the given dataset is balanced then used that dataset for further operations. but if the given dataset is imbalance then first you have to apply sampling technique. Sampling technique is used for equally distributed dataset that is in 70%-30% dataset or 60%-40% dataset. Once we apply sampling technique, dataset will be in proportion like in 70%-30% or 60%-40%.After that we will used that dataset for further operations.

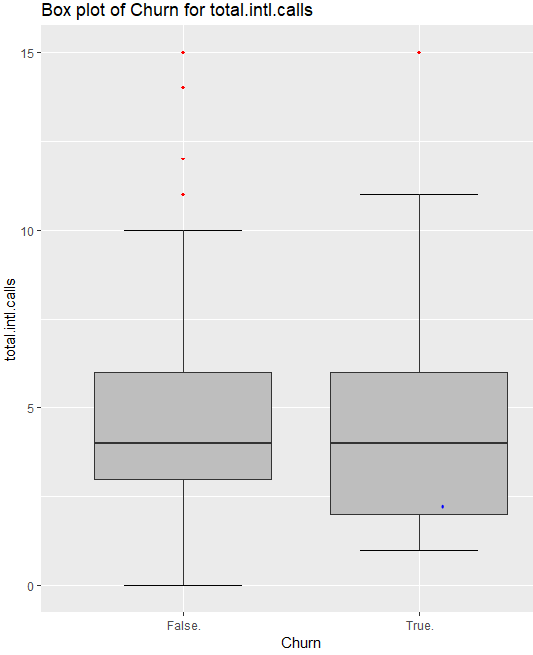
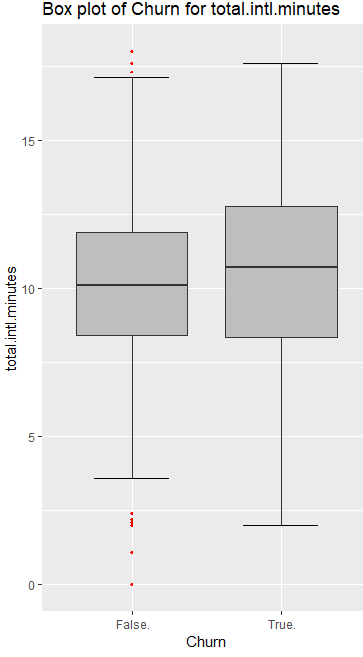
**2.1.1 Outlier Analysis**

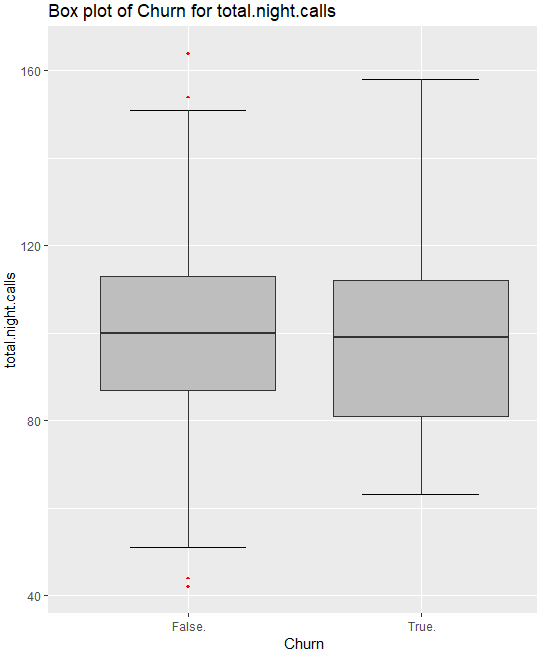
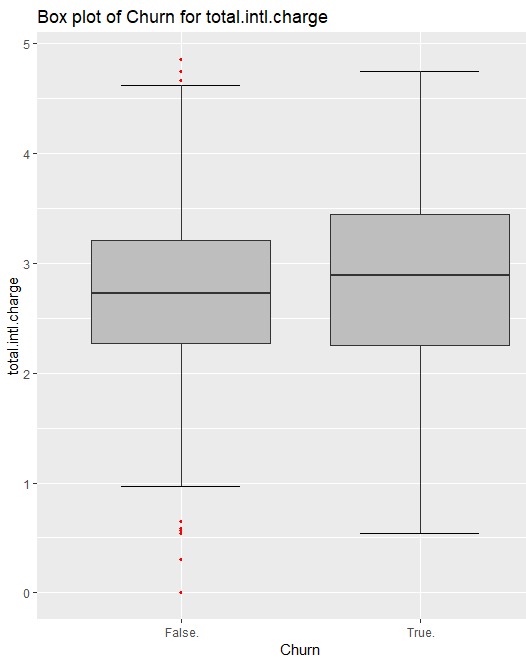
Outlier analysis is one of the technique to understand, clean and prepare data for building a predictive model. We can clearly observe from these probability distributions that most of the variables are skewed. The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. We can see the effect of the skew in below figure.

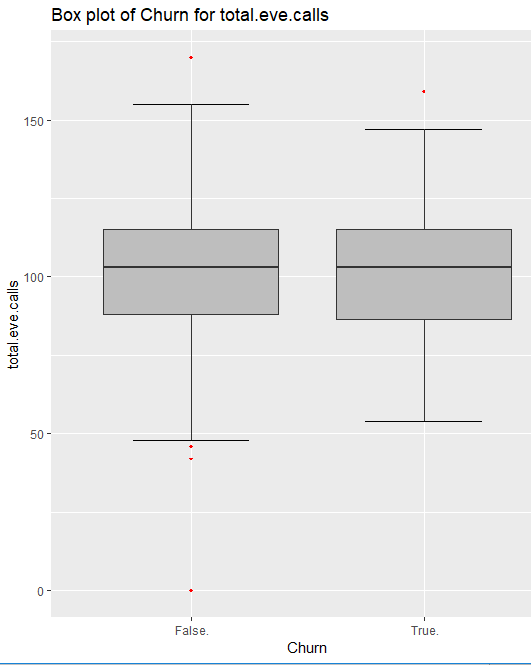
 







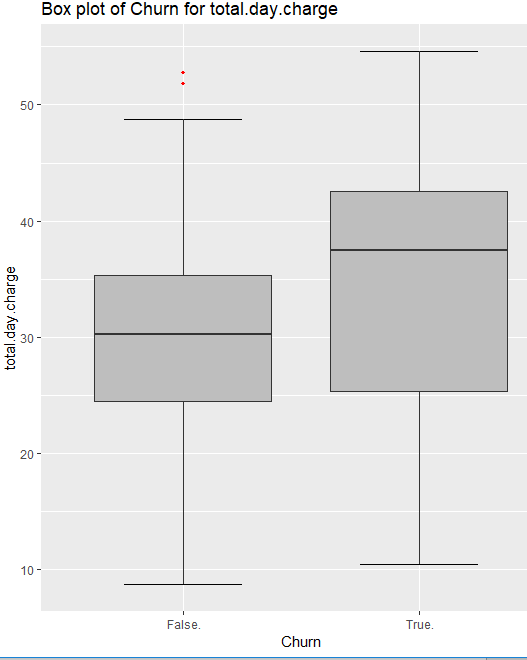
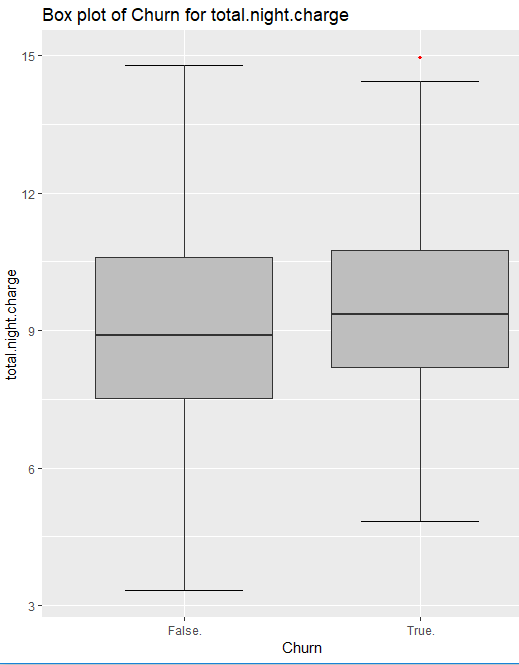


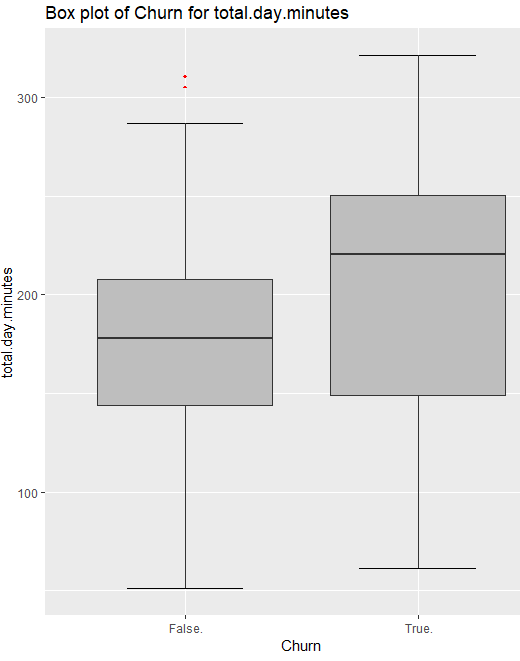
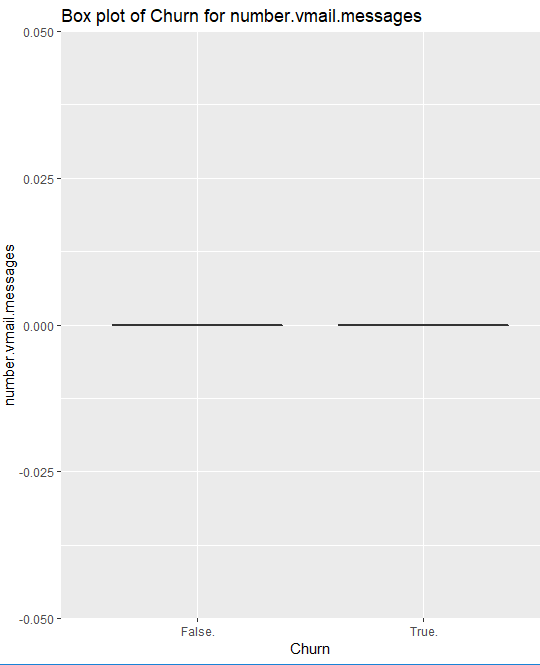


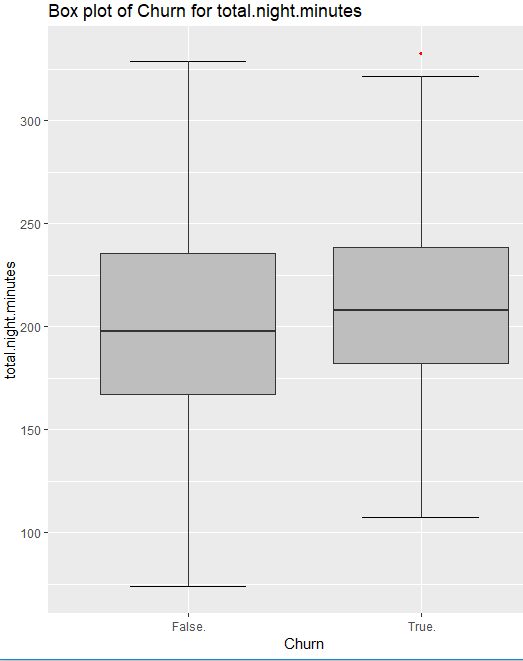
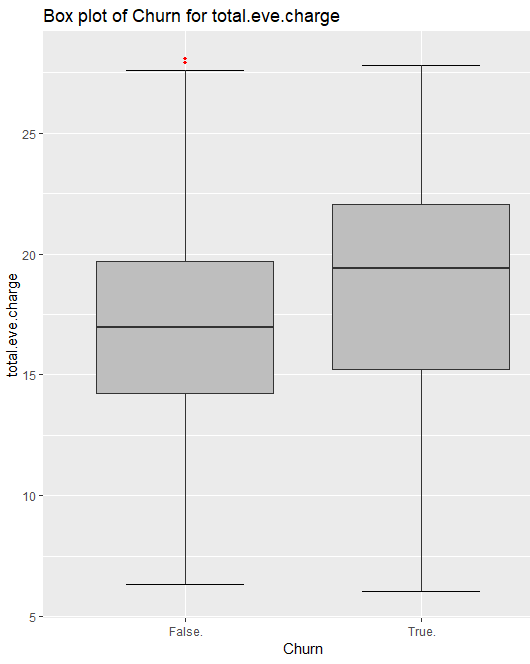
**Figure 2.1 Churn Vs Predictor Boxplots(**See R code in Appendix**)**

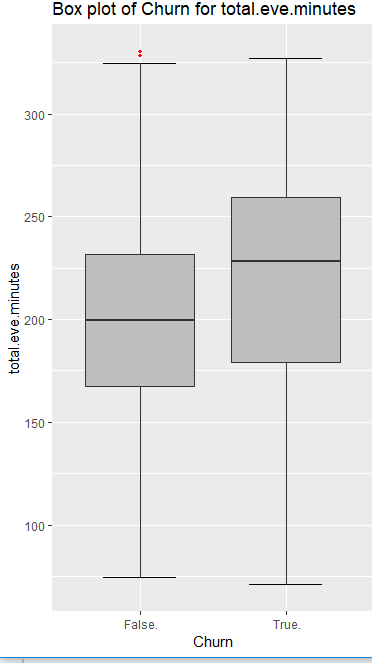
Here the red dots indicates the outliers or the values which is extreme on the given range. We visualize the outliers using boxplot. In figure 2.1 we have plotted the boxplots of the 13 predictor variables with respect to each Churn value. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set. For that we first detect the outliers and then remove it from the boxplot also there are multiple methods to impute those outliers. Using different techniques like KNN, Mean, Median & Mode we remove outliers.

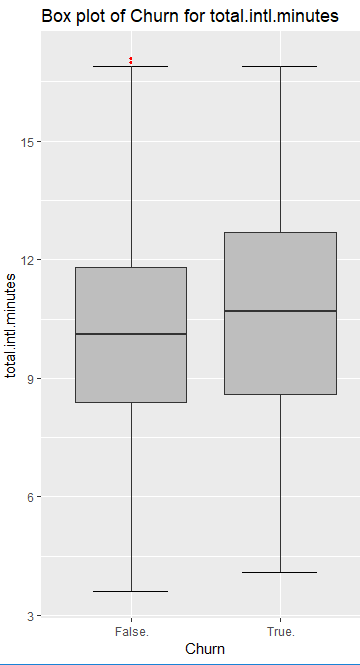
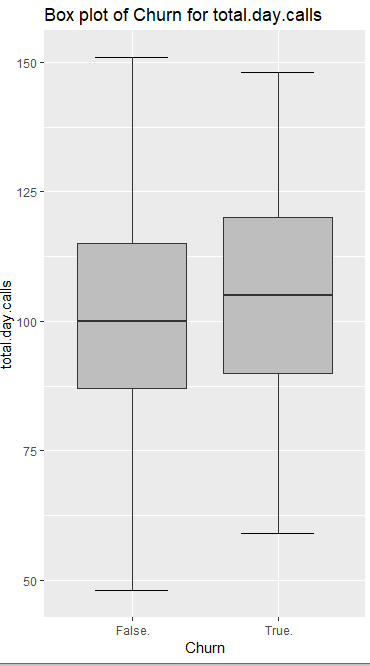
After removal of outliers the boxplot will be like this:

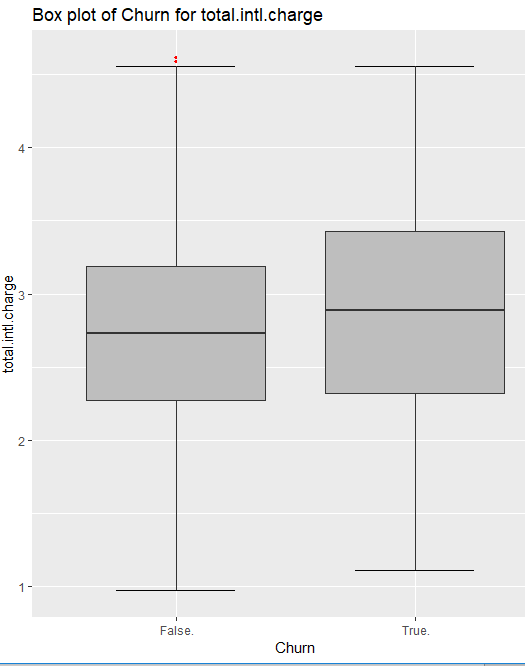
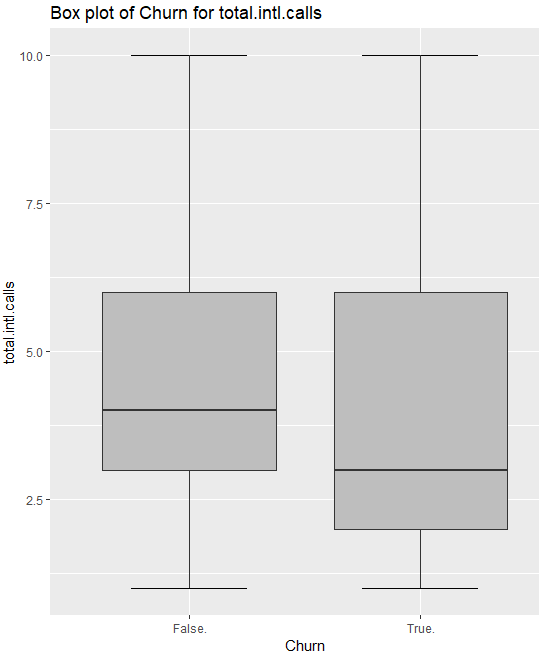


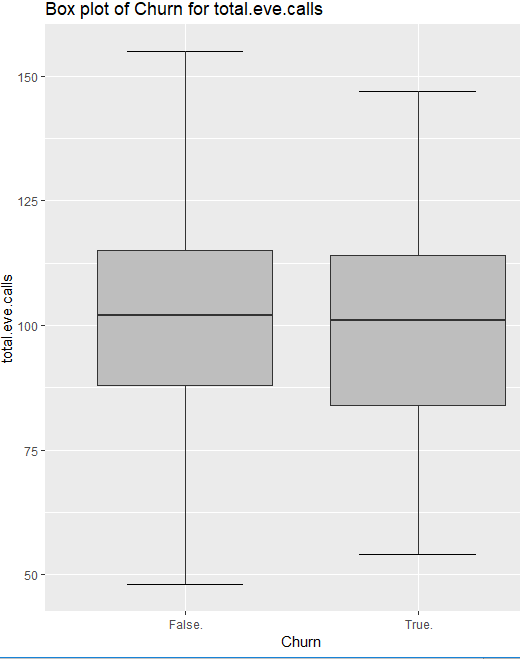
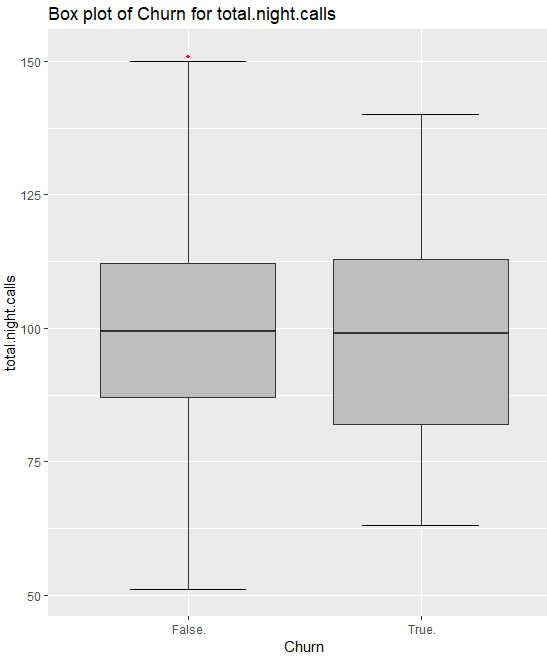












**Figure 2.2 Churn Vs Predictor Boxplot After Outlier Removal.(**See R code inAppendix**)**

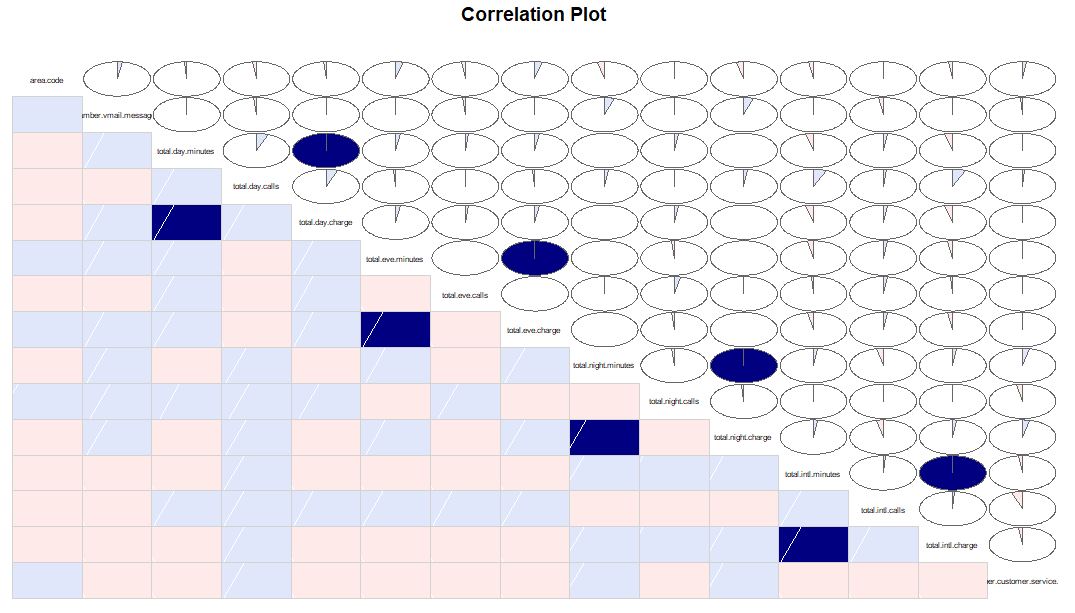
**2.1.2 Feature Selection**

Before performing any type of modeling, we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of churn prediction. There are several methods of doing that.

But before going for feature selection using different method, we first need to know domain knowledge. The dataset contains numerical as well as categorical variables. So, we apply the different analysis on different dataset. The correlation analysis is applied on only numerical data and chi-square test is apply on categorical data. In correlation plot analysis we check the variables which is positively correlated, negatively correlated and zero. The visualization tell about the variables are positively correlated or negatively correlated and according to negatively correlated variables we can drop.

Likewise, in chi-square test, if the p-value > 0.05 then reject the alternate hypothesis and variables are independent of each other, and if p-value < 0.05 then reject the null hypothesis and variables are dependent on each other.

Collect the variable whose p-value less than 0.05 and drop them in given dataset.



**Figure 2.3 Visualization of Correlation Analysis(**See R code in Appendix**)**

And Using the chi-square test on categorical variable where none of the variables are < than p-value means we cannot drop the variable through chi-square test.

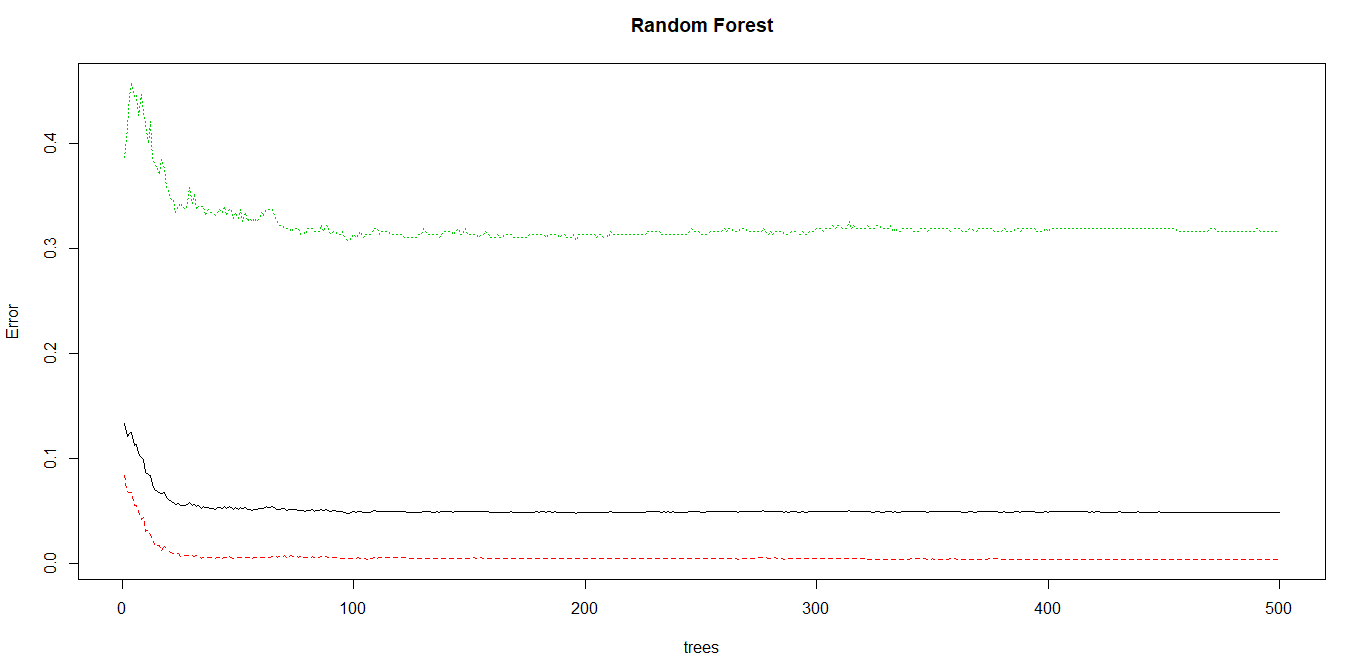
Below we have used Random Forests to perform features selection.

Random forest is a tree-based algorithm which involves building several trees (decision trees), then combining their output to improve generalization ability of the model. The method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner.

Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

So, In our given dataset the dependent (target ) variable is categorical so we solve this problem using classification. There is no chance to solve the problem using regression because the dependent(target) variable is not categorical, it is continuous.

The Visualization of Random Forest is given below:

***Random Forest :***

**Figure 2.4 Visualization of Random Forest (**See R code in Appendix**)**

The above visualization tell us about the number of trees with respective the error.

The first plot indicates the error for your different classes (colored) and out-of-bag samples (black) over the amount of trees. Classes are in the same order as the results you get from print(model), so will be red=False and green=True. You essentially see that the error seems to be lowest around 100 trees is the given example.

For the variable importance as **MeanDecreaseGini** :

MeanDecreaseGini

|  |  |
| --- | --- |
| number.vmail.messages | 28.83 |
| total.day.minutes | 84.27 |
| total.day.calls | 40.70 |
| total.day.charge | 86.41 |
| total.eve.minutes | 47.51 |
| total.eve.calls | 36.20 |
| total.eve.charge | 46.83 |
| total.night.minutes | 39.03 |
| total.night.calls | 33.01 |
| total.night.charge | 37.02 |
| total.intl.minutes | 33.58 |
| total.intl.calls | 26.83 |
| total.intl.charge | 33.21 |

**Figure 2.5 Table for MeanDecreaseGini**

The MeanDecreaseGini measures the Gini importance = how important the features are *over all splits* done in the tree/forest - whereas for each individual split the Gini importance indicates how much the Gini criterion = "unequality/heterogeneity" was reduced using this split. Because a classification tree essentially tries to built homogeneous groups of samples, so that one (homogeneous) class label can be predicted per group. So it makes sense to check how much features contributed to obtaining such homogeneous groups - which is the end is the MeanDecreaseGini = "variable importance" you see. So, as you can clearly see, **total.day.charge** and **total.day.minutes** contributed most to obtaining such splits, so they are considered more important.

**2.2 Modeling**

**2.2.1 Model Selection**

In our early stages of analysis during pre-processing we have come to understand that ,using correlation plot and chi-square test the result will be different. Therefore, we can neither combine the data sets nor use a single model for predicting variables. Hence, we need to analyse the data sets separately and generate separate models for data sets.

The dependent variable can fall in either of the four categories:

1. Nominal

2. Ordinal

3. Interval

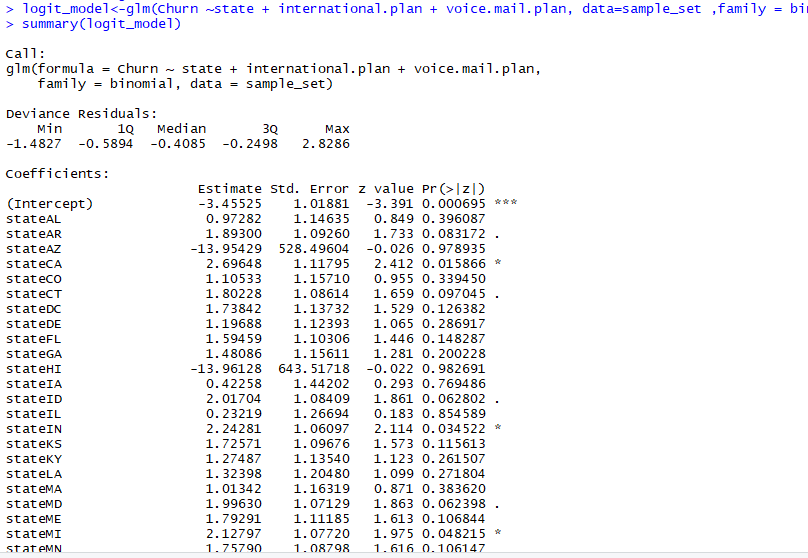
4. Ratio

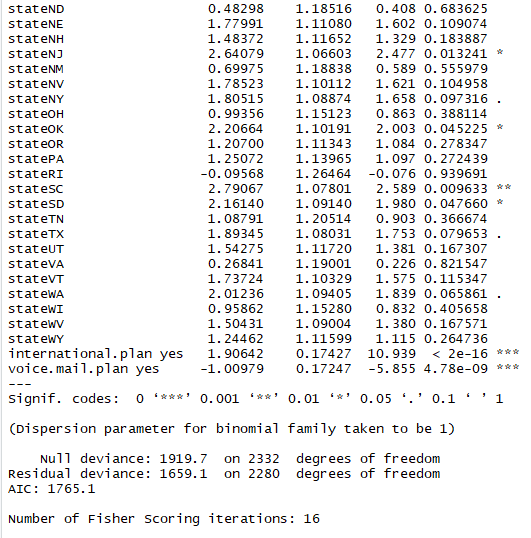
If the dependent variable, in our case Churn, is Nominal the only predictive analysis that we can perform is Classification, and if the dependent variable is Interval or Ratio the normal method is to do a Regression analysis, or classification after binning. But the dependent variable we are dealing with is Nominal, for which only classification can be done, because the Churn variable has categories.

You always start your model building from the simplest to more complex. Therefore, we use Logistic Regression.

**2.2.2 Logistic Regression**

Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. To represent binary / categorical outcome, we use dummy variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.





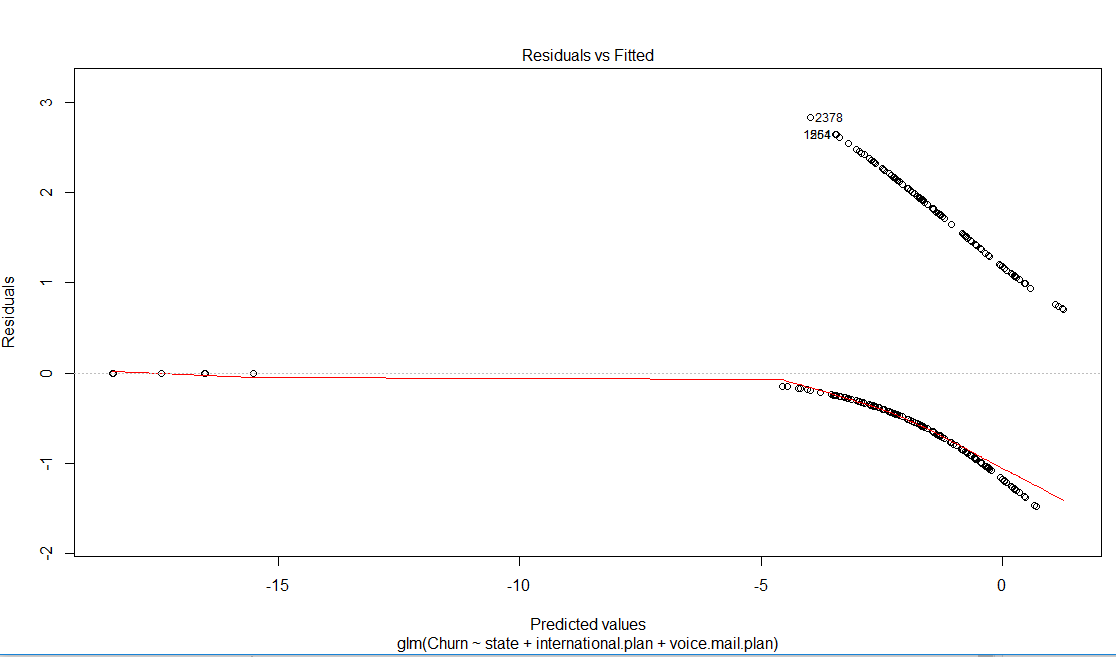
As you can see the *probability value,* if p-value > 0.05 then we reject the alternate hypothesis that variables are independent of each other. That variables are StateNY StateTX, StateWA.And the \*\*\* indicates the if p-value < 0.05 then we reject null hypothesis that target variable does not depend on **international.plan** and **voice.mail.plan** of the predictor variable.

The Accuracy of the model will be less but the FNR will be greater. We can also display the prediction using plots.

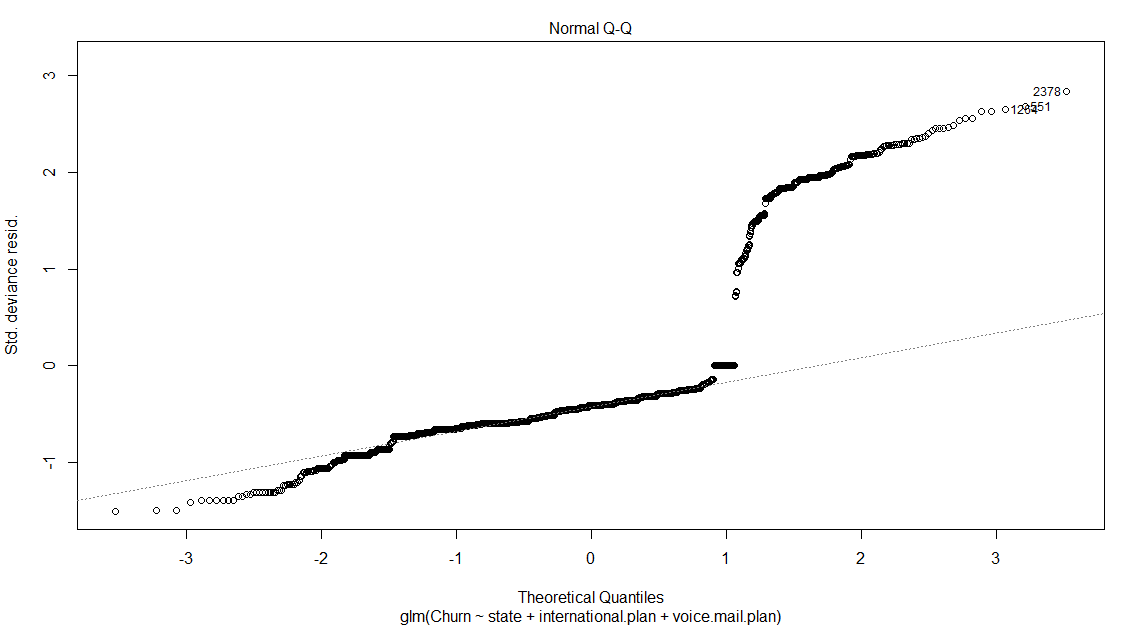
Logistic Regression is part of a larger class of algorithm known as Generalized Linear Model(glm).

Using glm() we can predict the output. The following are the different different plots which identify the logistic regression model.

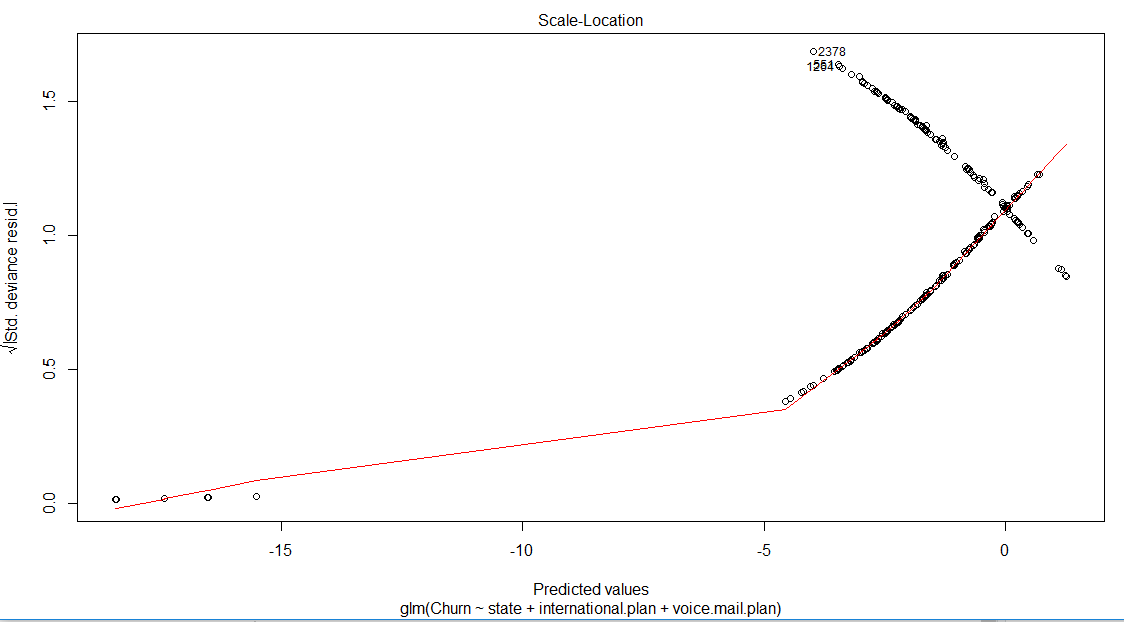
* + **The Residuals vs Fitted plot** can help you see, for example, if there are curvilinear trends that you missed. But the fit of a logistic regression is curvilinear by nature, so you can have odd looking trends in the residuals with nothing amiss.
  + **The Normal Q-Q plot** helps you detect if your residuals are normally distributed. But the deviance residuals don't have to be normally distributed for the model to be valid, so the normality / non-normality of the residuals doesn't necessarily tell you anything.
  + **The Scale-Location plot** can help you identify heteroscedasticity. But logistic regression models are pretty much heteroscedastic by nature.
  + **The Residuals vs Leverage** can help you identify possible outliers. But outliers in logistic regression don't necessarily manifest in the same way as in linear regression, so this plot may or may not be helpful in identifying them.



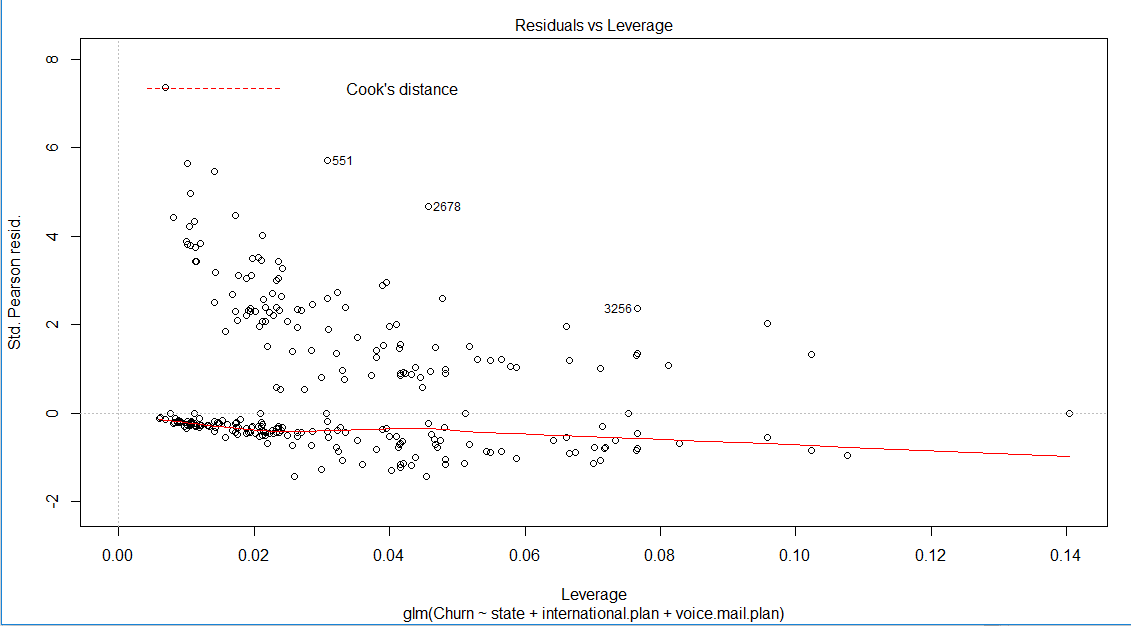
**Figure : Residuals Vs Fitted Plot**



**Figure :Normal Q-Q Plot**



**Figure : Scale – Location**



**Figure : Residuals Vs Leverage**

**Figure 2.6 Logistic Regression Plots (**See R code in Appendix**)**

**2.2.3 Decision Tree / Classification Tree**

Decision Tree is a supervised machine learning algorithm. Supervised means have a target variable. Decision tree used in a variety of ways to solve regression as well as classification model.

**Classification** **trees**, as the name implies are used to separate the dataset into classes belonging to the response variable. Usually the response variable has two classes:Yes or No (1 or 0). If the target variable has ***more*** than 2 categories, then a variant of the algorithm, called C4.5, is used. For binary splits however, the standard CART procedure is used. Thus classification trees are used when the response or target variable is categorical in nature.

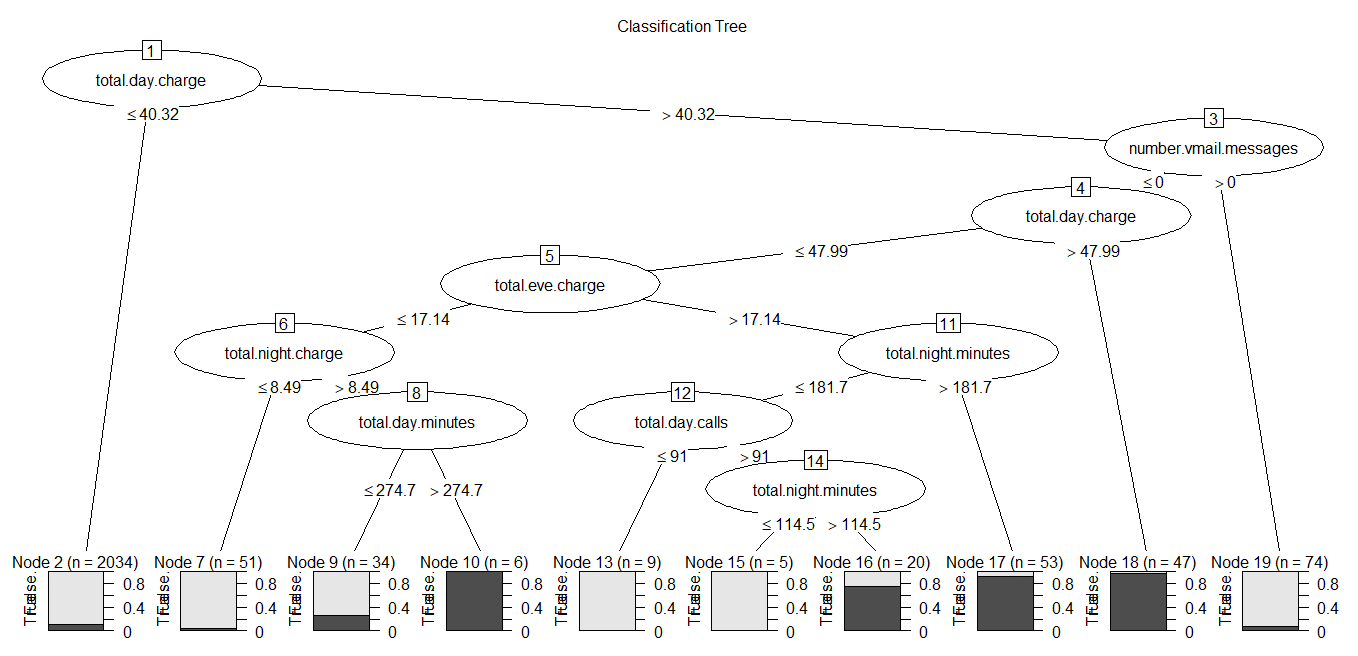
**Regression** **trees** are needed when the response variable is numeric or continuous. For example, the predicted price of a consumer good. Thus, regression trees are applicable for *prediction* type of problems as opposed to *classification*.

So, here in given dataset the target variable (Churn) is present which contain two category that is TRUE and FALSE. So, we solve this problem using C5.0 and CART as well.

Here we not used Regression Tree because the target or response variable is not numeric or continuous. The difference between these two algorithms is not so far, accuracy is also near. The Visualization of both algorithms is different.C5.0 Model is multi split tree and CART has binary split tree.

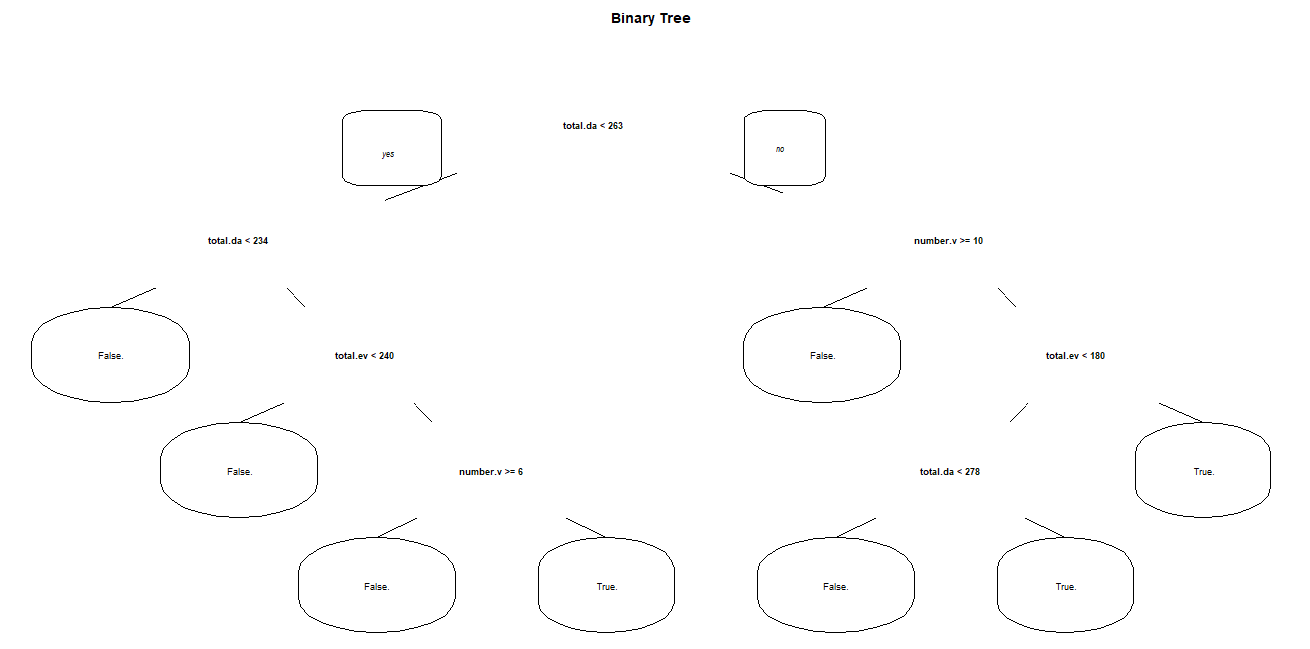
So here we conclude that both algorithms give the high accuracy. No one is better than other, both give the results.

The Classification/Decision Tree will be shown:



**Figure 2.7 Decision tree using C5.0 Model (**See R code in Appendix**)**

The Classification/Decision Tree will be shown:



**Figure 2.8 Decision Tree using CART Model(**See R code in Appendix**)**

**2.2.3 Regression**

Using Regression for prediction analysis in this case is not interval or ratio ,though it can be done. The reason is, the values of the target variable, Churn is categorical that is the form of TRUE,FALSE or YES,NO. Though Regression predictions have been done on interval or ratio variables it is not a recommended approach, because the information stored in terms of categorical variables.

**Chapter 3**

**Conclusion**

**3.1 Model Evaluation**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Confusion Matrix

2. False Negative Rate

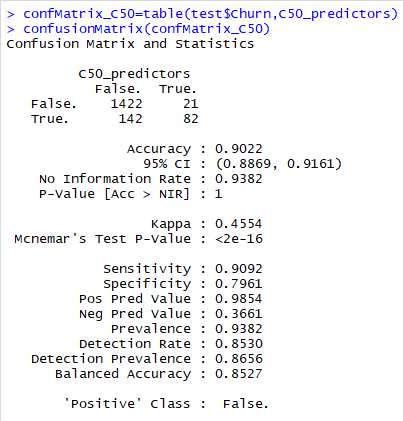
Predictive performance can be measured by comparing Predictions of the models with real values of the target variables, and calculating some average error measure.

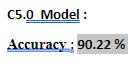
**3.1.1 Confusion Matrix**

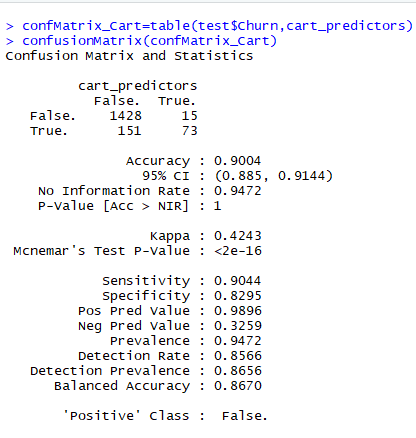
Confusion Matrix is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our models that we have generated in the previous section.

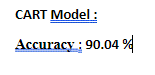
One by one we can display Confusion Matrix of models.

**A) Decision Tree:**

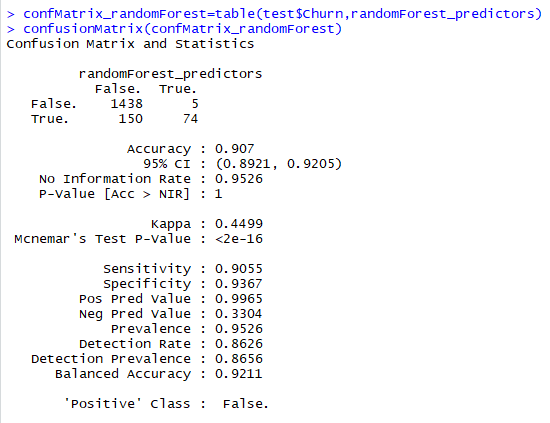






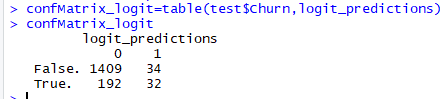


**B) Random Forest:**



**Accuracy = 90.7 %**

**C) Logistic Regression:**



**Accuracy**  = (TP + TN) / (TP + TN +FP +FN)

=(1409 + 32 )/ 1667

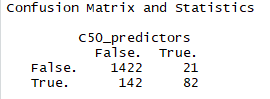
=1441 / 1667

**Accuracy = 86.44 %**

**3.1.1 False Negative Rate :**

**A) Decision Tree:**

**The Confusion Matrix for C5.0 Model is:**



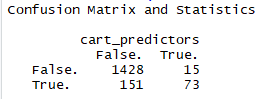
**The False Positive Rate for C5.0 Model is:**

**FNR =** FN/(FN+TP)

=142 / (142 + 82)

**FNR** = 63.39%

**The Confusion Matrix for CART Model is:**



**The False Positive Rate for CART Model is:**

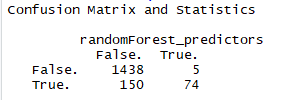
**FNR =** FN/(FN+TP)

=151 / (151 + 73)

**FNR** = 67.41%

**B) Random Forest:**

**The Confusion Matrix is :**



**The False Negative Rate is :**

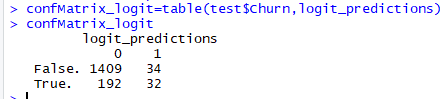
**FNR =** FN/(FN+TP)

=150 / (150 + 74)

**FNR** = 66.96%

**C) Logistic Regression:**

**The Confusion Matrix is :**



**The False Negative Rate is :**

**FNR =** FN/(FN+TP)

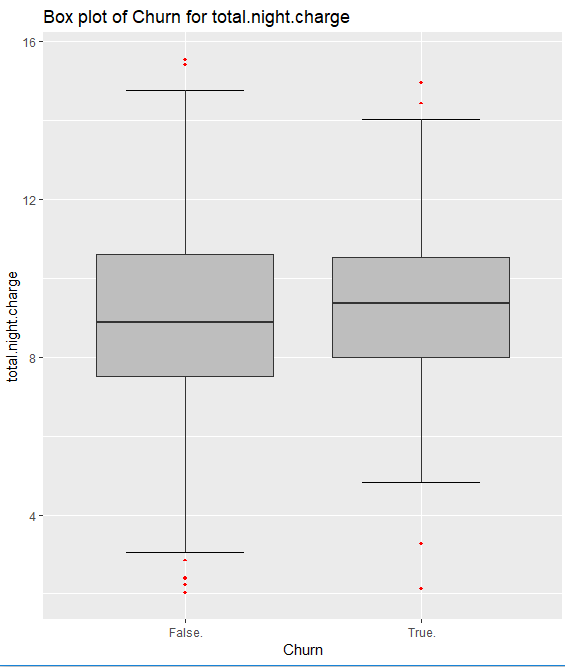
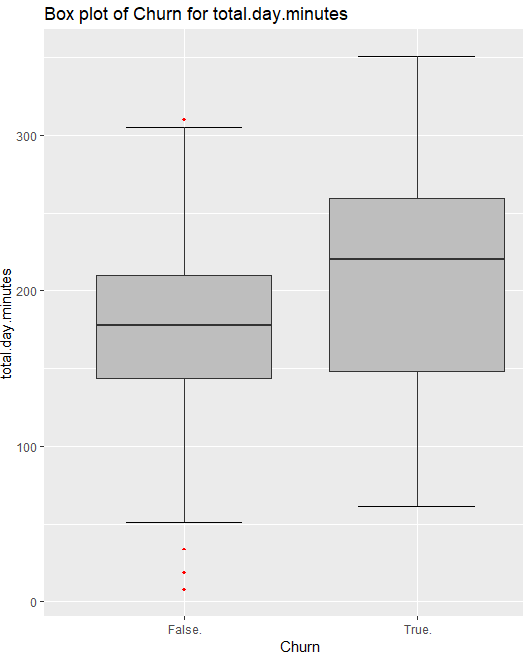
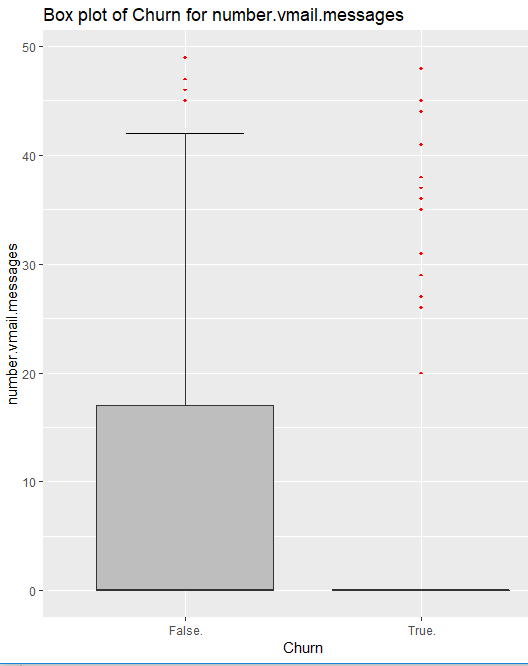
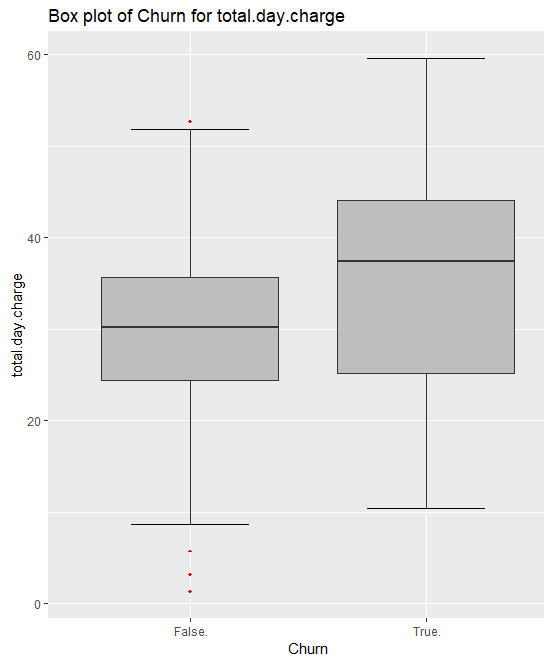
=192 / (192 + 32)

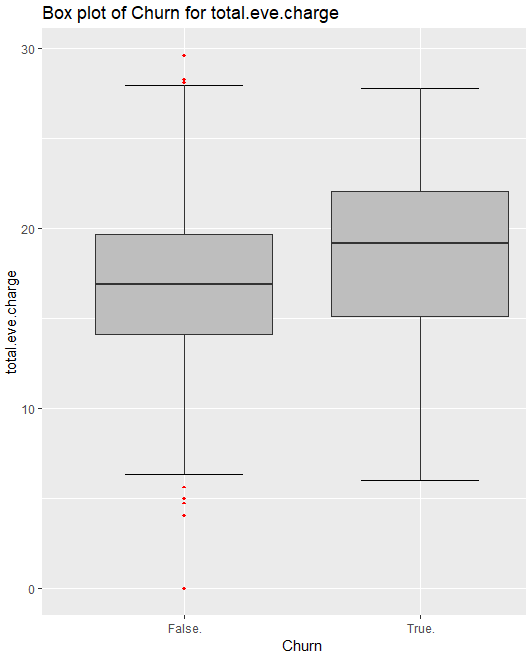
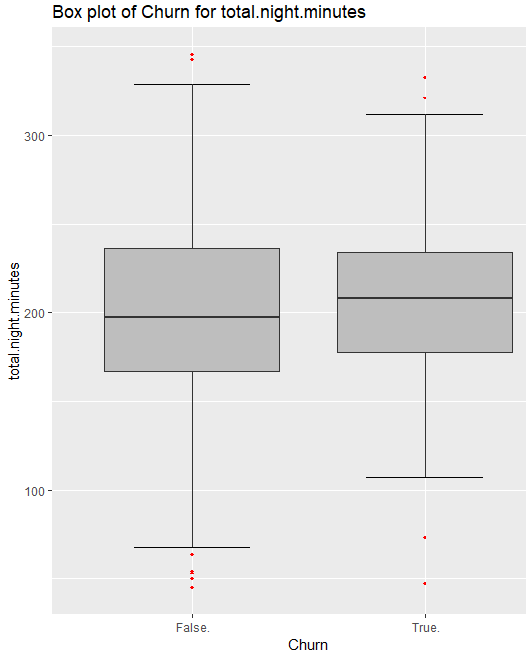
**FNR** = 85.71%

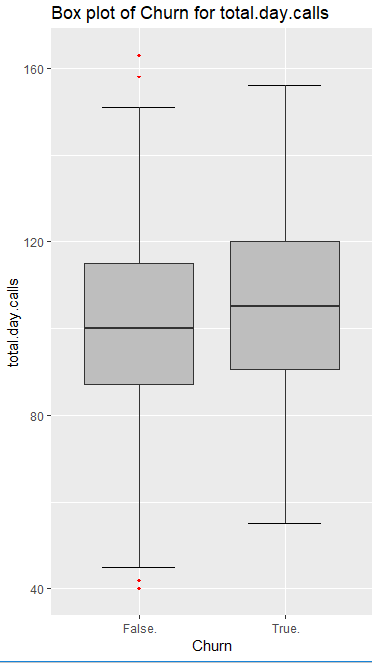
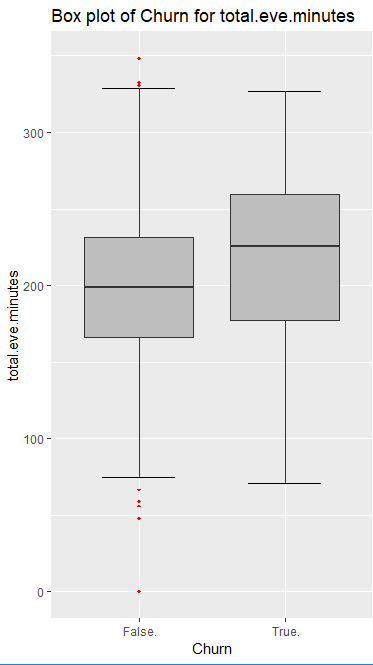
**3.2 Model Selection**

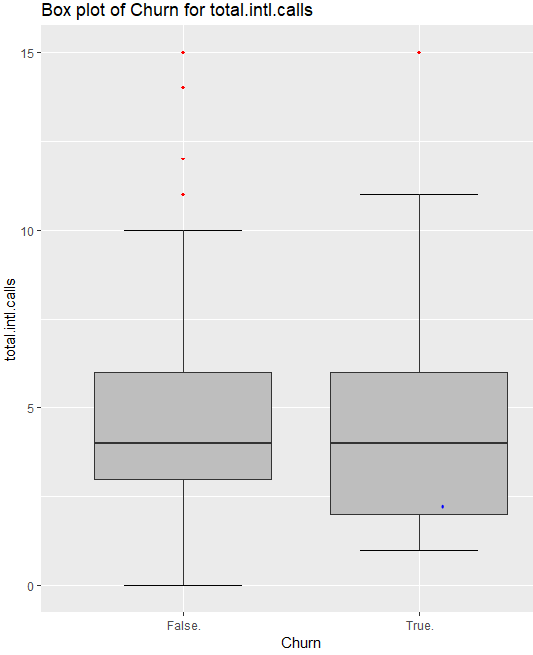
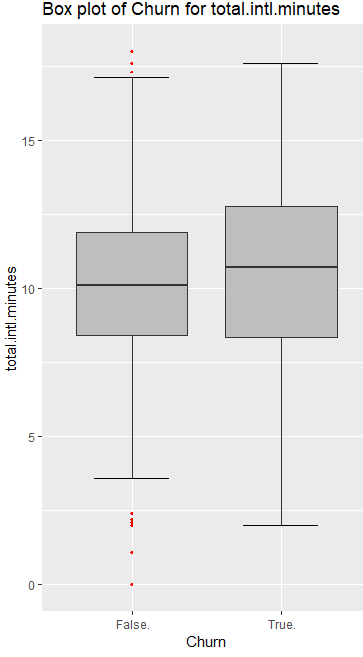
We can see that out of three models two models perform comparatively on average and therefore we can select either of the two models without any loss of information.

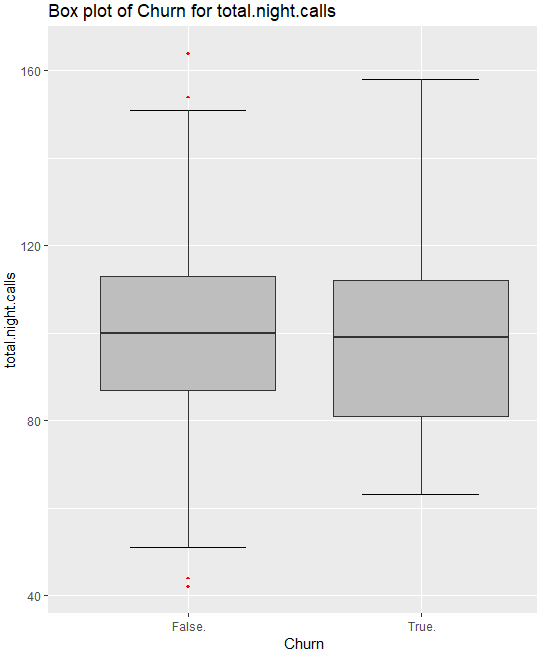
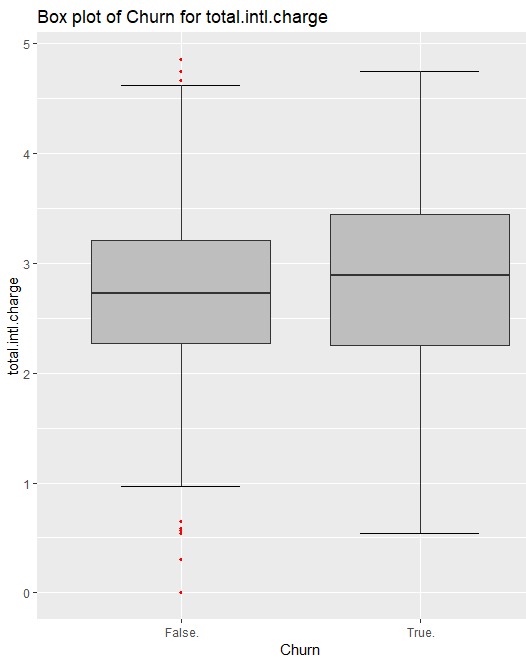
**Appendix A - Extra Figures**

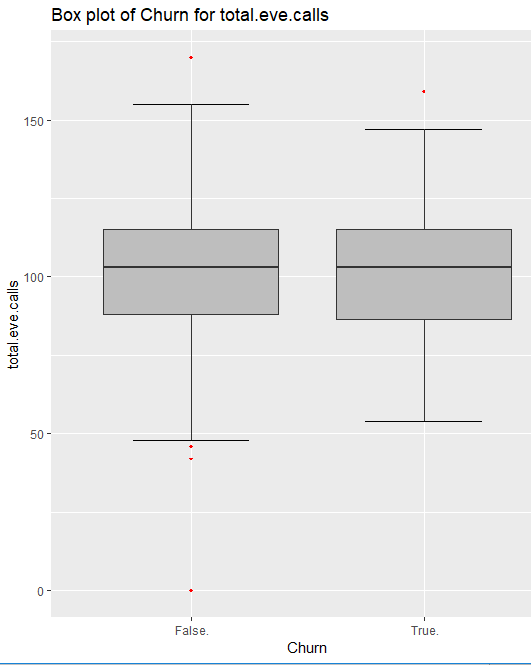
 





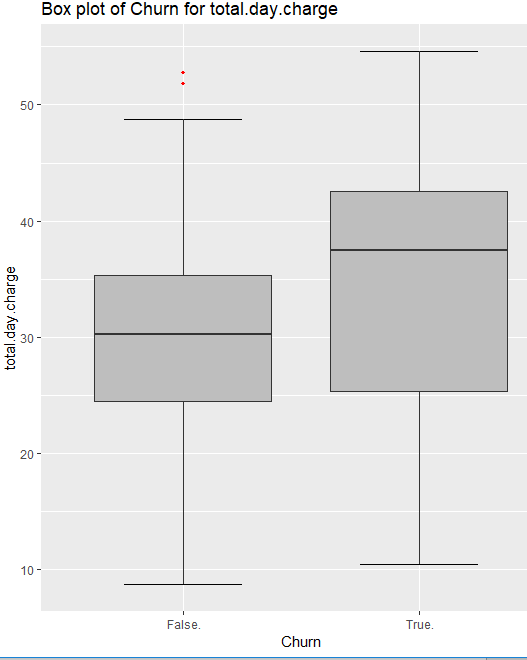
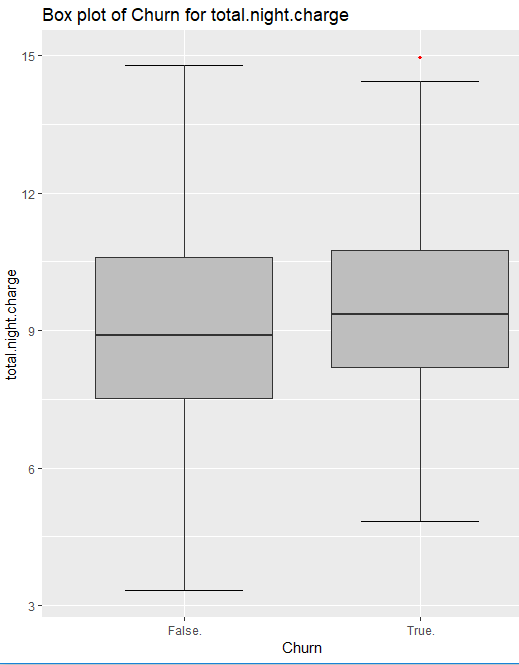


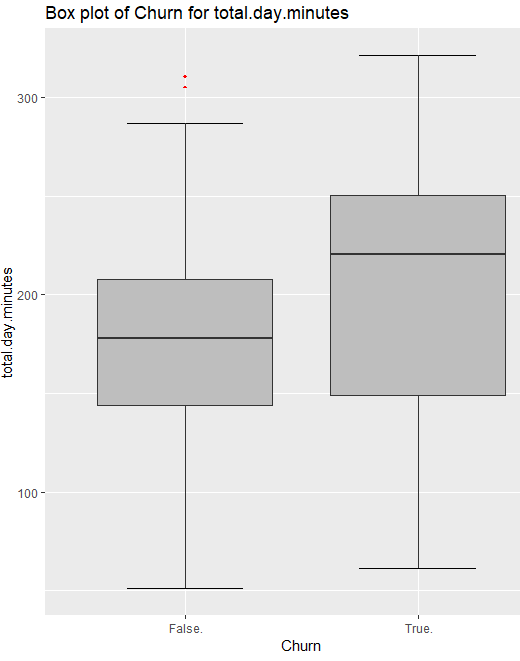
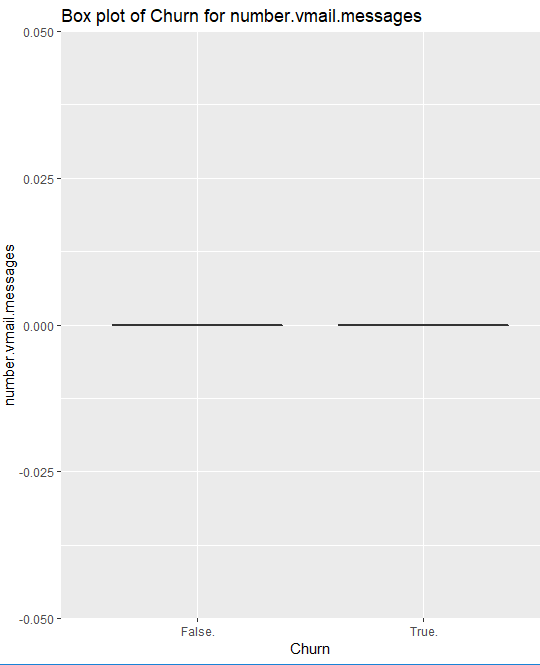


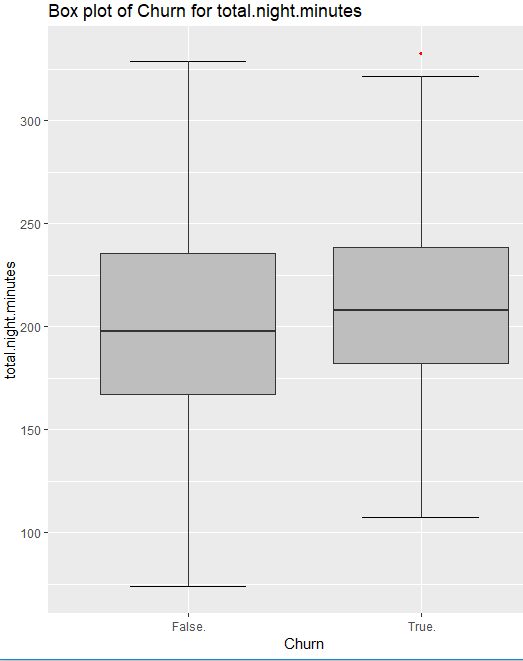
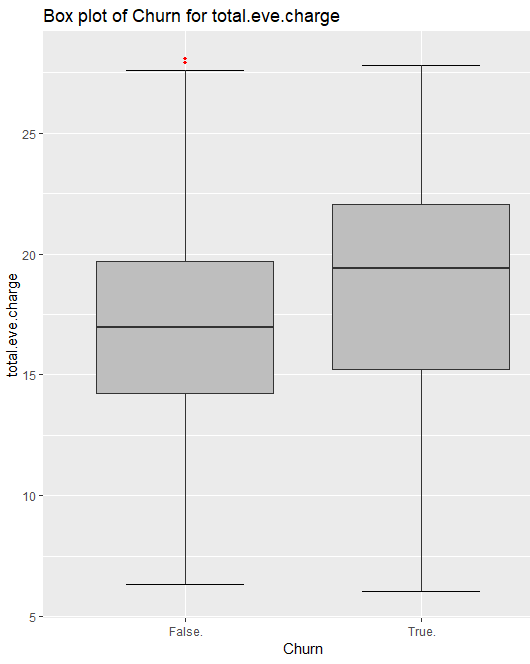


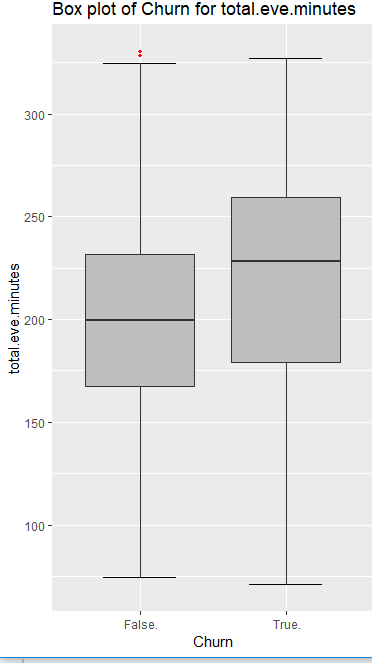
**Figure 2.1 Churn Vs Predictor Boxplots** (See R code in Appendix)

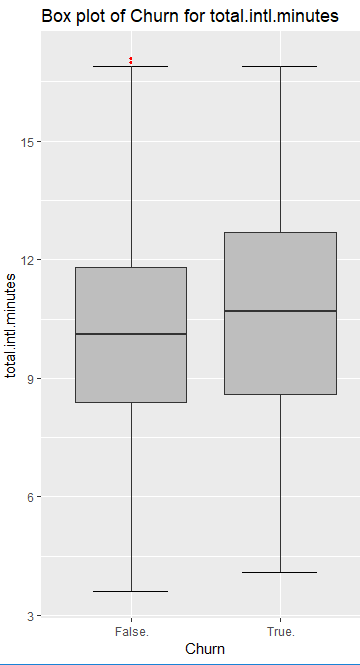
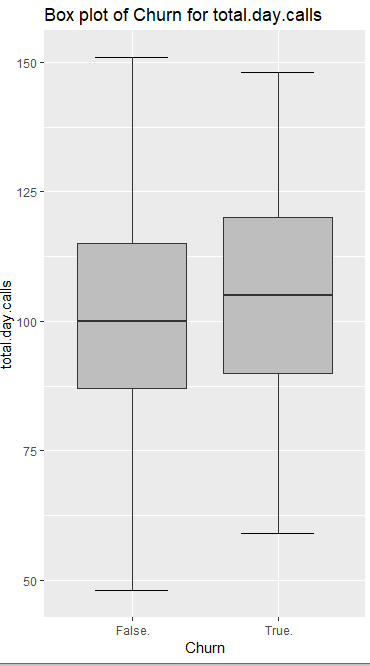
***After removal of outliers the boxplot will be like this:***

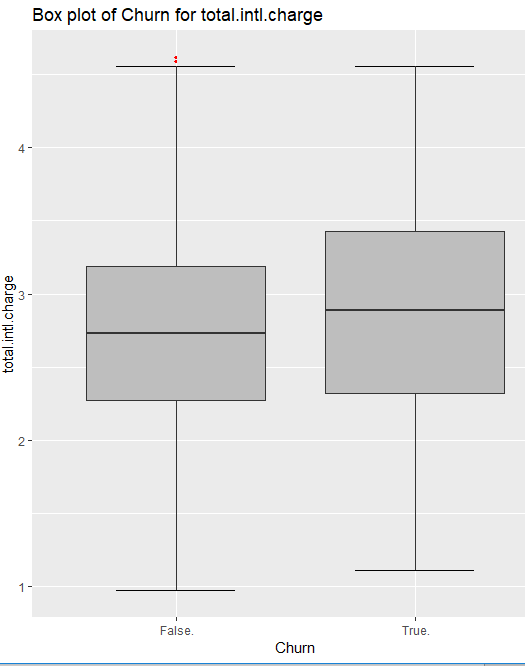
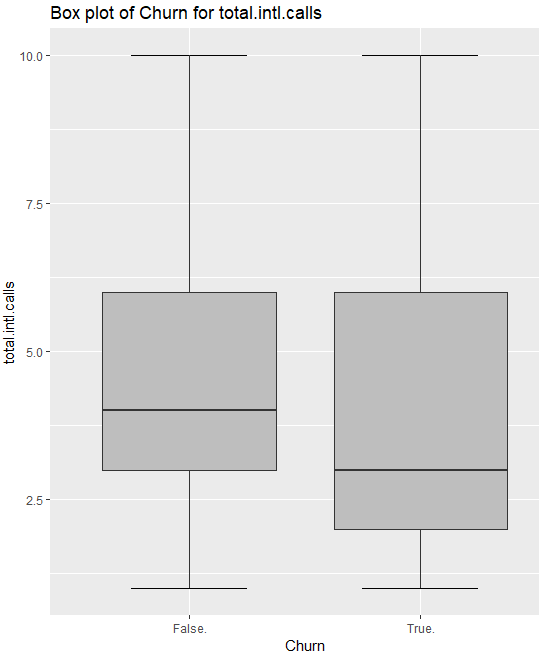


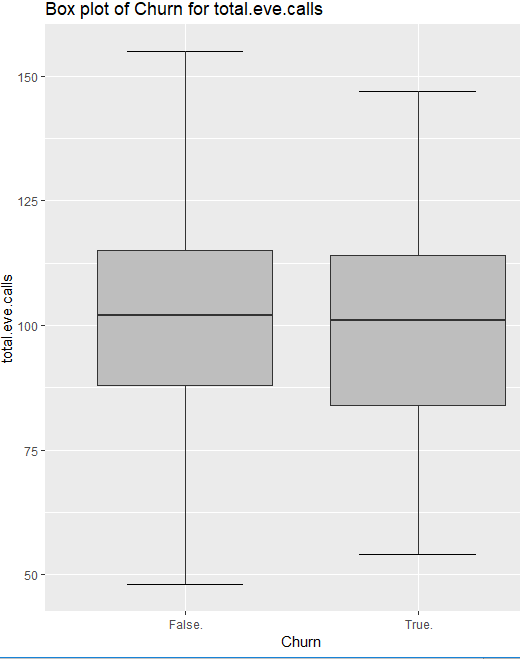
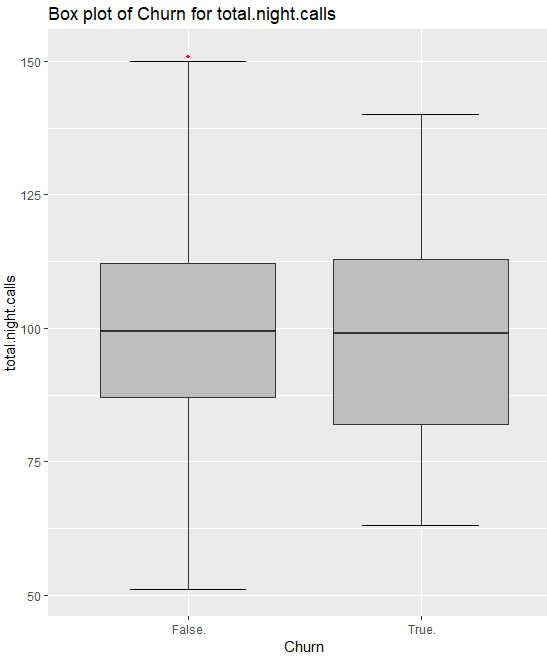




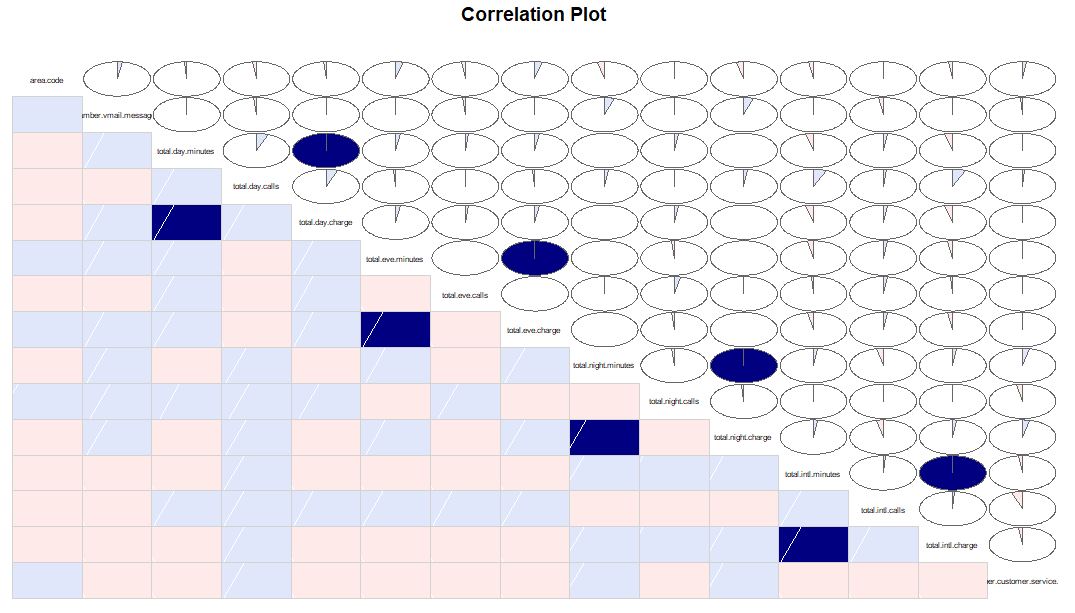




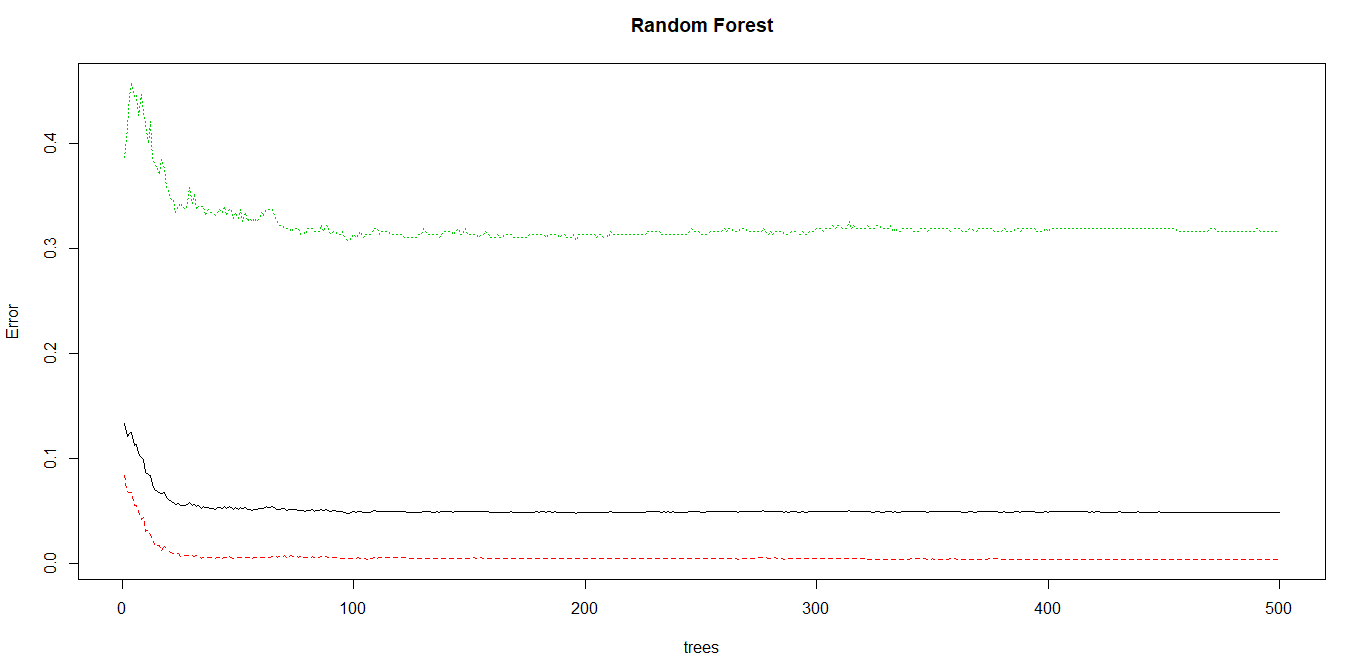




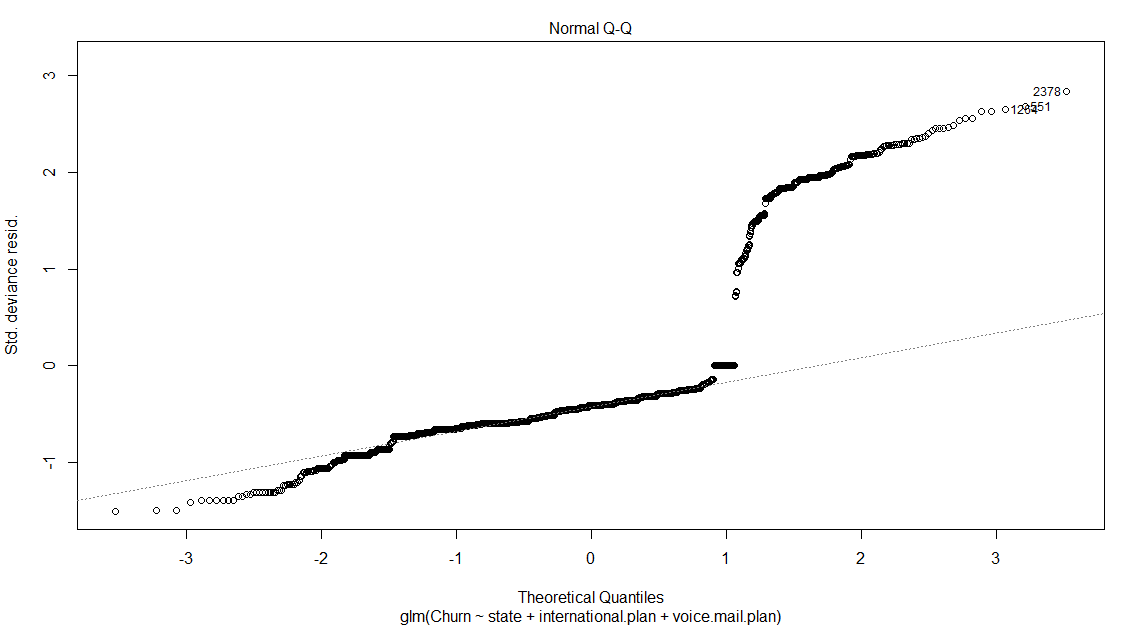
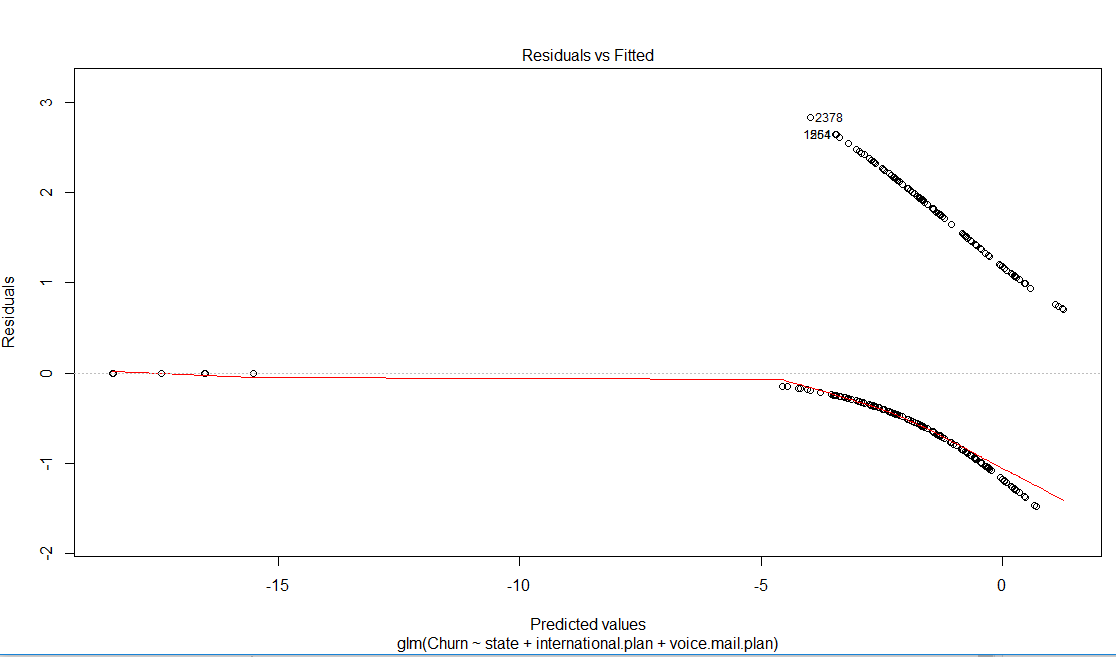
**Figure 2.2 Churn Vs Predictor Boxplot After Outlier Removal.(**See R code inAppendix**)**

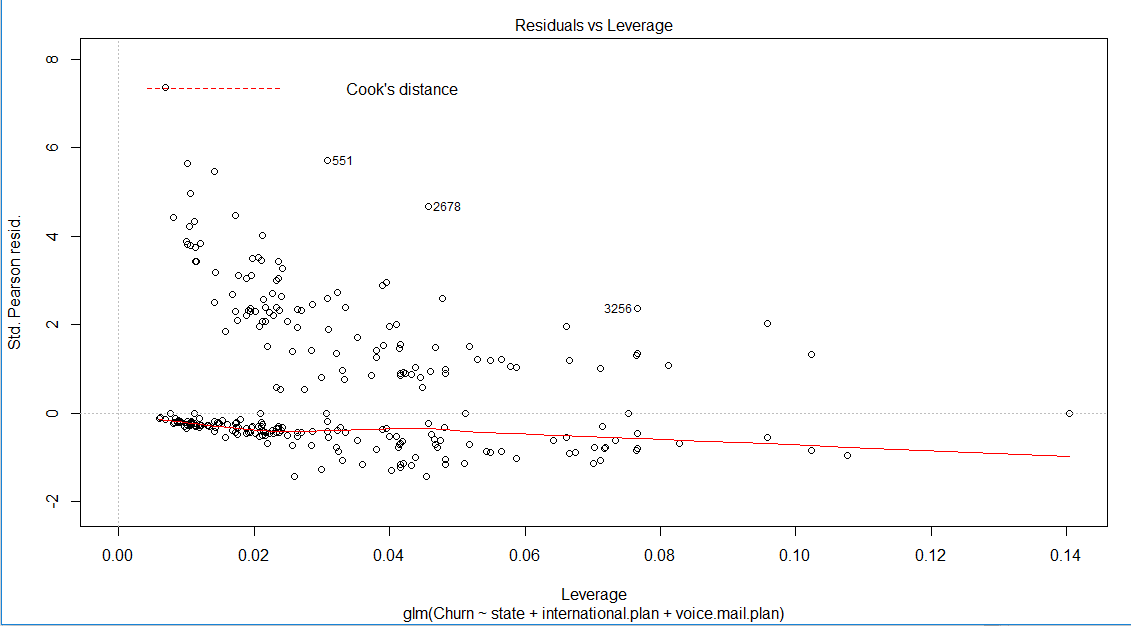
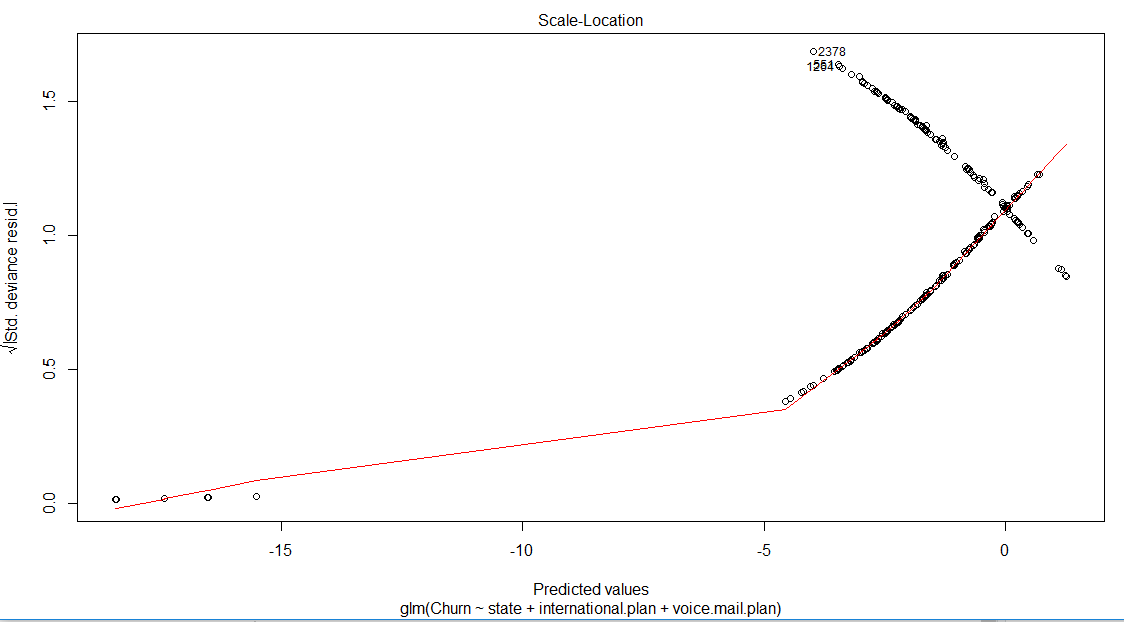


**Figure 2.3 Visualization of Correlation Analysis(**See R code in Appendix**)**

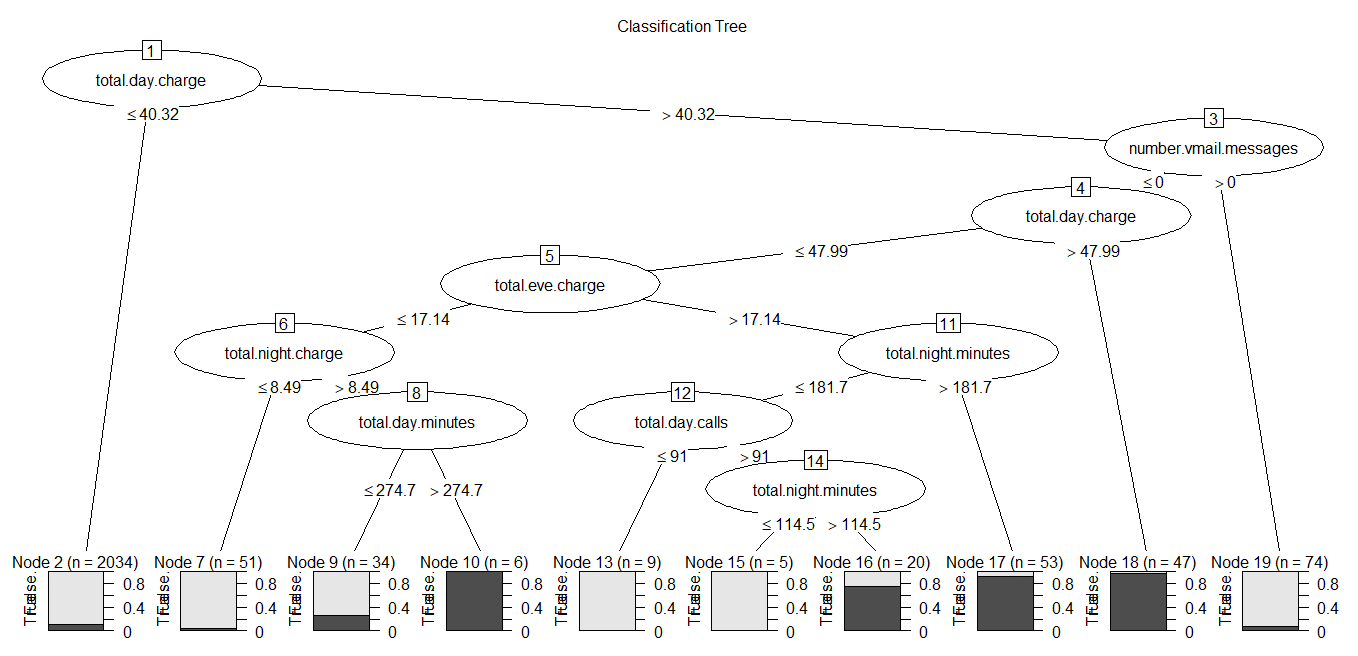


**Figure 2.4 Visualization of Random Forest(**See R code in Appendix**)**

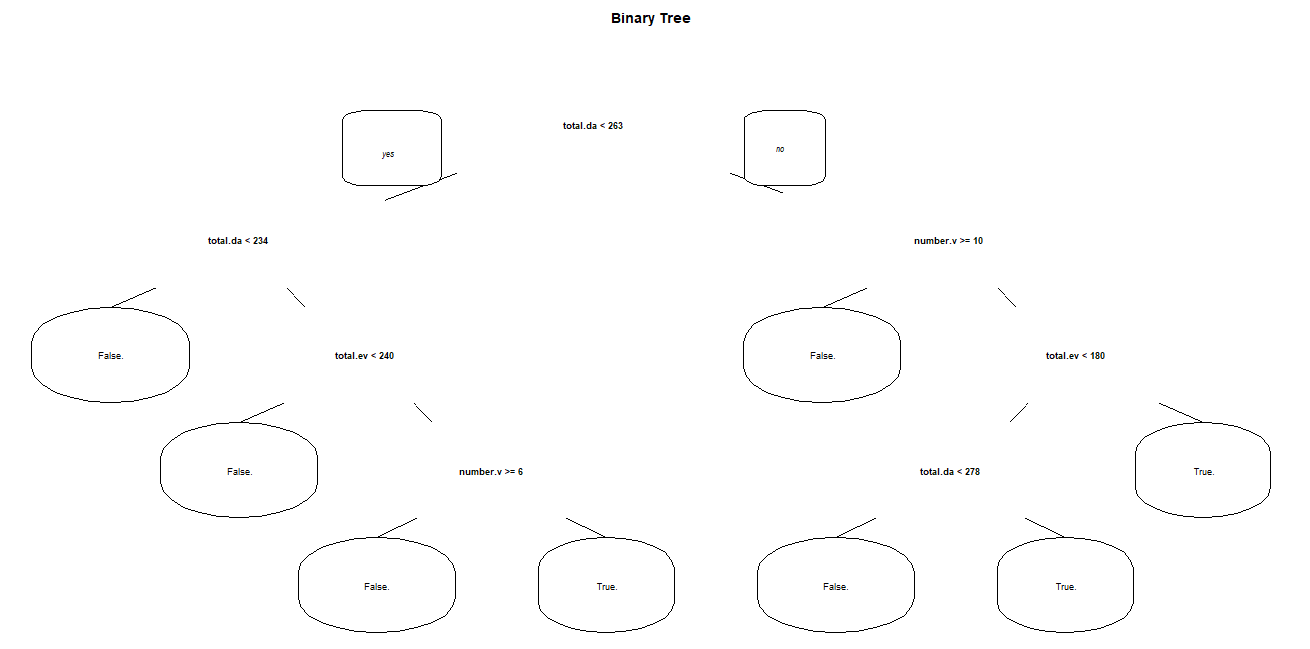




**Figure 2.6 Logistic Regression Plots(**See R code in Appendix**)**



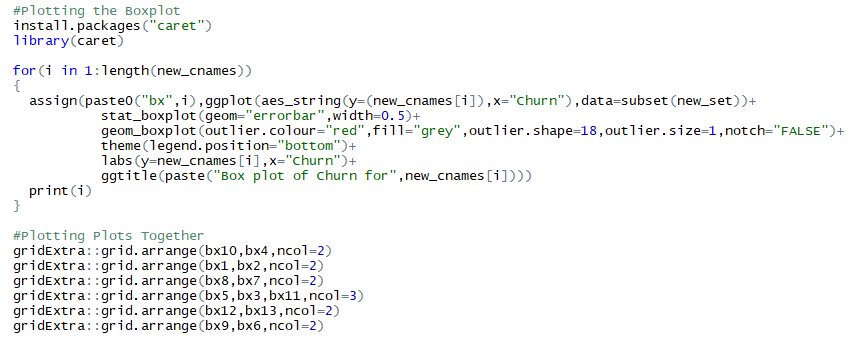
**Figure 2.7 Decision tree using C5.0 Model (**See R code in Appendix**)**



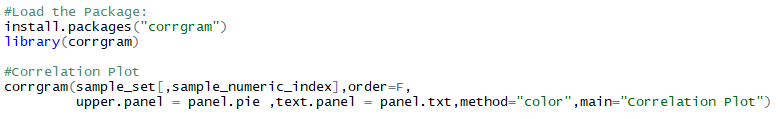
**Figure 2.8 Decision Tree using CART Model (**See R code in Appendix**)**

**Appendix B - R Code**

***Boxplots for all predictors variables (Fig. 2.1)***



***Correlation Plot (Fig. 2.3)***



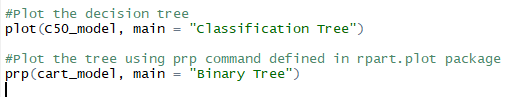
***Random Forest Plot (Fig. 2.4)***



***Logistic Regression Plot (Fig. 2.6)***



***Decision Tree Using C5.0 and CART (Fig. 2.7 & Fig. 2.8)***



***Complete R File***

