**Project Name – Churn Reduction**

*Akhilesh Sayana*

*21 November 2018*

**Table Of Contents**

**1 Introduction 2**

* 1. Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
  2. Problem Category
  3. Tools Used
  4. Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1. **Methodology 4**
   1. Exploratory Data Analysis
   2. Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.1.2 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**Model Building**

**Model Evaluation**

**Conclusion**

**Important Features**

**Introduction**

* 1. **Problem Statement**

Understanding the problem statement is the first and foremost step. This would help you give an intuition of what you will face ahead of time. Let us see the problem 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.

Customer churn occurs when customers or subscribers stop doing business with a company or service. Also known as customer attrition, customer churn is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers – earning business from new customers means working leads all the way through the sales funnel, utilizing the marketing and sales resources throughout the process. Customer retention, on the other hand, is generally more cost-effective as the trust and loyalty of existing customers has already been earned.

Customer churn impedes growth, so companies should have a defined method for calculating customer churn in a given period of time. By being aware of and monitoring [churn rate](http://churn-rate.com/), organizations are equipped to determine their customer retention success rates and identify strategies for improvement.

Various organizations calculate customer churn rate in a variety of ways, as churn rate may represent the total number of customers lost, the percentage of customers lost compared to the company’s total customer count, the value of recurring business lost, or the percent of recurring value lost. Other organizations calculate churn rate for a certain period, such as quarterly periods or fiscal years. One of the most commonly used methods for calculating customer churn is to divide the total number of clients a company has at the beginning of a specified time period by the number of customers lost during the same period.

* 1. **Problem Category**

It is a classification problem where we need to predict if a customer with given details will leave the company.

In a classification problem, we have to predict discrete values based on a given set of independent variable(s). Classification can be of two types:

* **Binary Classification** : In this classification we have to predict either of the two given classes. For example: classifying the gender as male or female, predicting the result as win or loss, etc.
* **Multiclass Classification** : Here we have to classify the data into three or more classes. For example: classifying a movie's genre as comedy, action or romantic, classify fruits as oranges, apples, or pears, etc.

Customer Churn prediction is a very common real-life problem that each service-based company faces, and telecom sector is a very common example of the same. If done correctly, it can save the cost that will be accounted for the acquisition of new customers.

As given in the problem statement either the customer will leave or will be retained. There are only two discrete classes. Hence this is a “binary classification” or “binomial classification”.

* 1. **Objective**

The objective of this Case is to predict customer behavior. We are given with a

public dataset that has customer usage pattern and if the customer has moved or not.

We need to develop an algorithm to predict the churn score based on usage

pattern.

* 1. **TOOLS USED**

We will be using Python with Jupyter notebook and R with RStudio for EDA, Data Preprocessing, Model Building and Model Evaluation.

* 1. **Data**

For this practice problem, we have been given two CSV files: Train\_data.csv and Test\_data.csv. Both the files contain all the independent variables, and the target variable. The predictors provided are as follows:

● account length

● international plan

● voicemail plan

● number of voicemail messages

● total day minutes used

● day calls made

● total day charge

● total evening minutes

● total evening calls

● total evening charge

● total night minutes

● total night calls

● total night charge

● total international minutes used

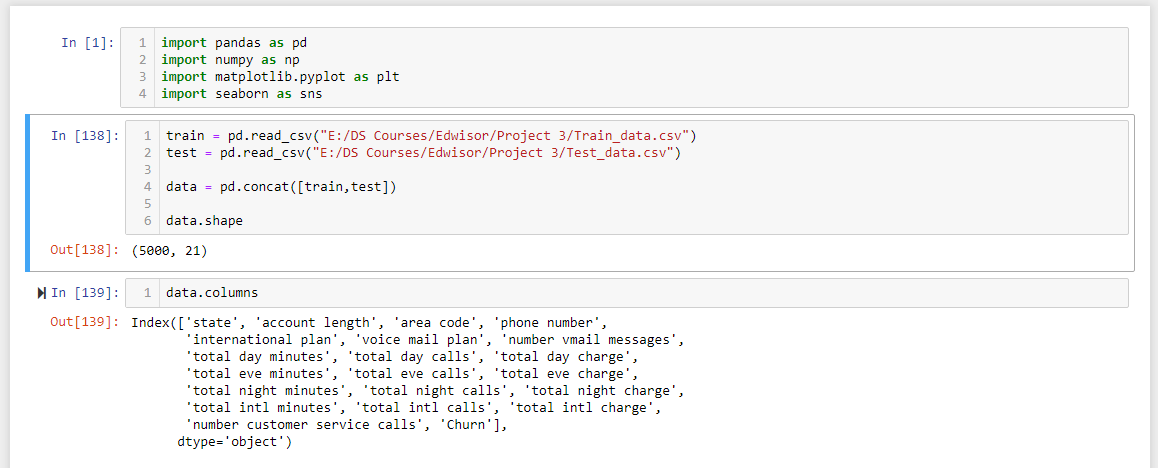
● total international calls made

● total international charge

● number of customer service calls made

**Target Variable**: move: if the customer has moved (1=yes; 0 = no)

**Reading Data:**

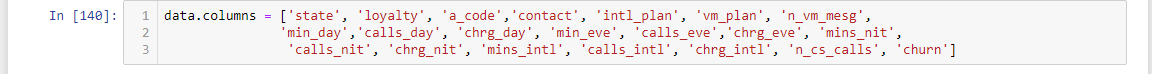


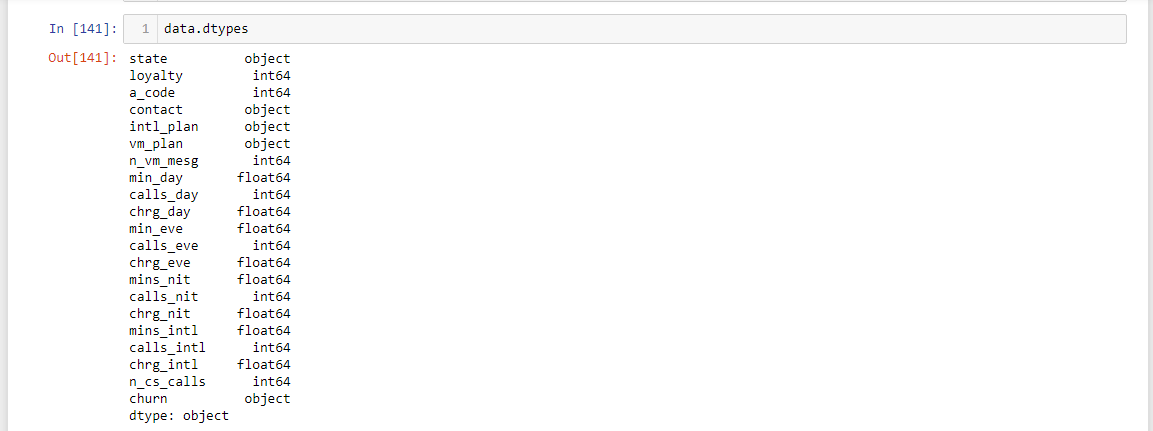
The data has been read from both the files into two different dataframes. Both the dataframes are then concatenated into a single dataframe- “data”. The complete dataset has 5000 rows and 21 columns.

I decided to concatenate both the dataframes because both the datasets(train and test) have all the values for the target variable- (Churn). So there is no pont of dealing with the dataframes separately. We will split the data later when required for modelling purpose.

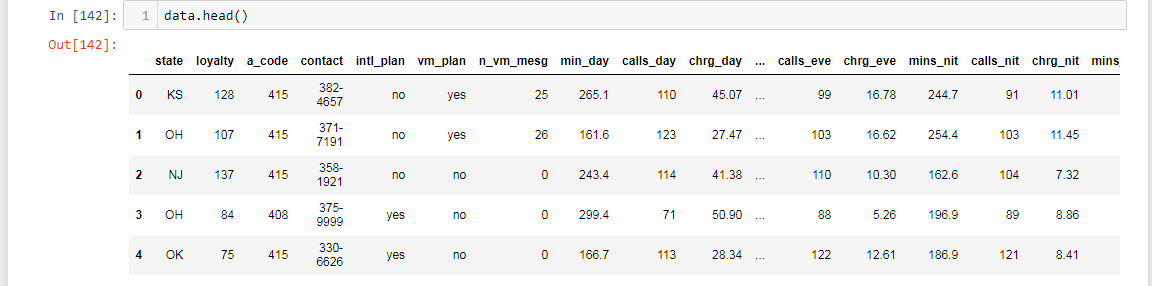


The columns have been renamed for the sake of convenience.



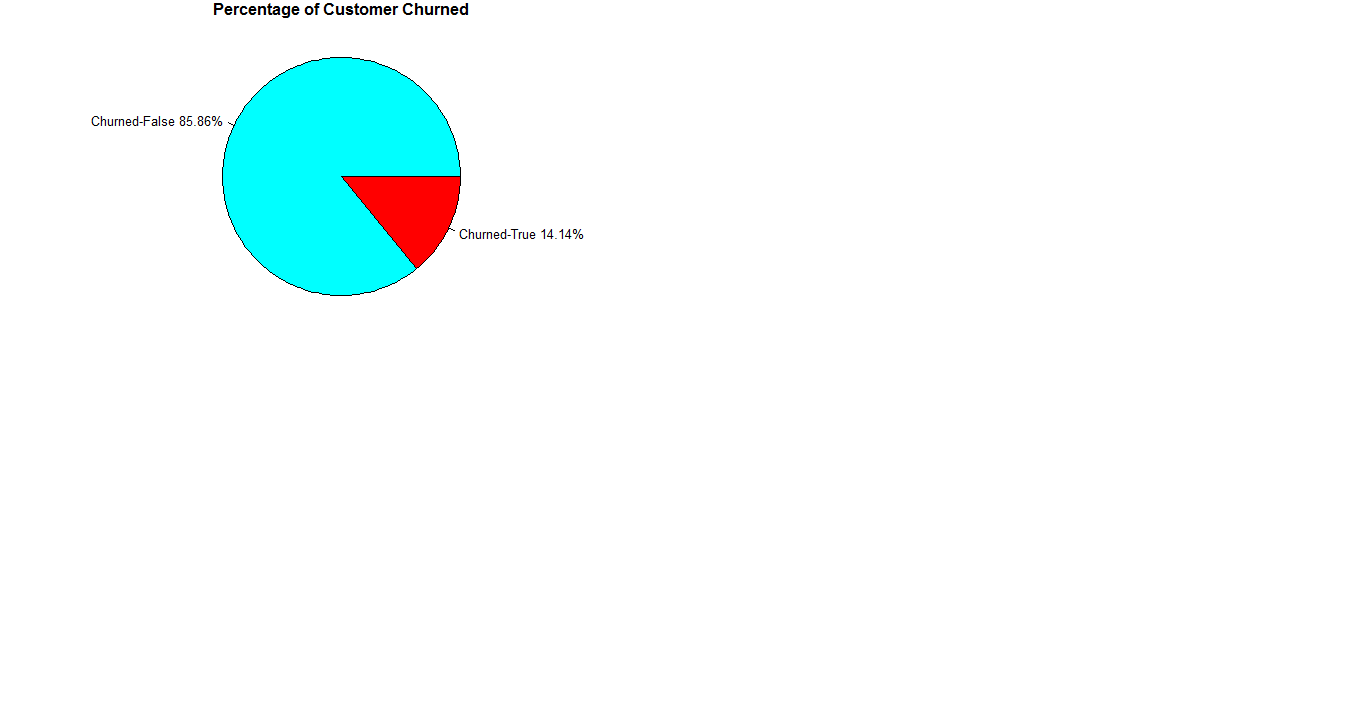


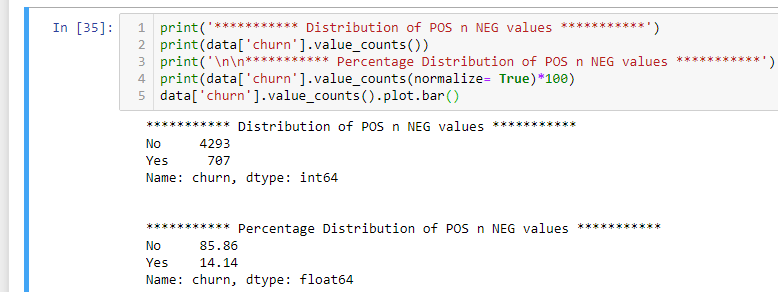
First look at the data below:

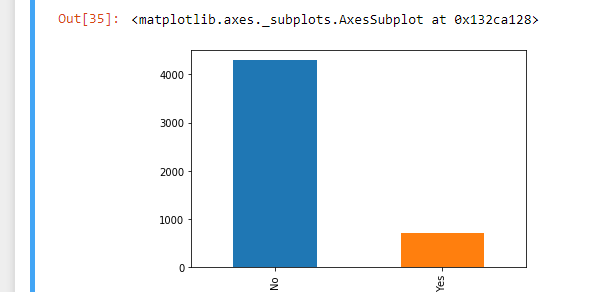


* 1. **Class Distribution in target variable**

707 (~14%) out of 5000 customers left the company.

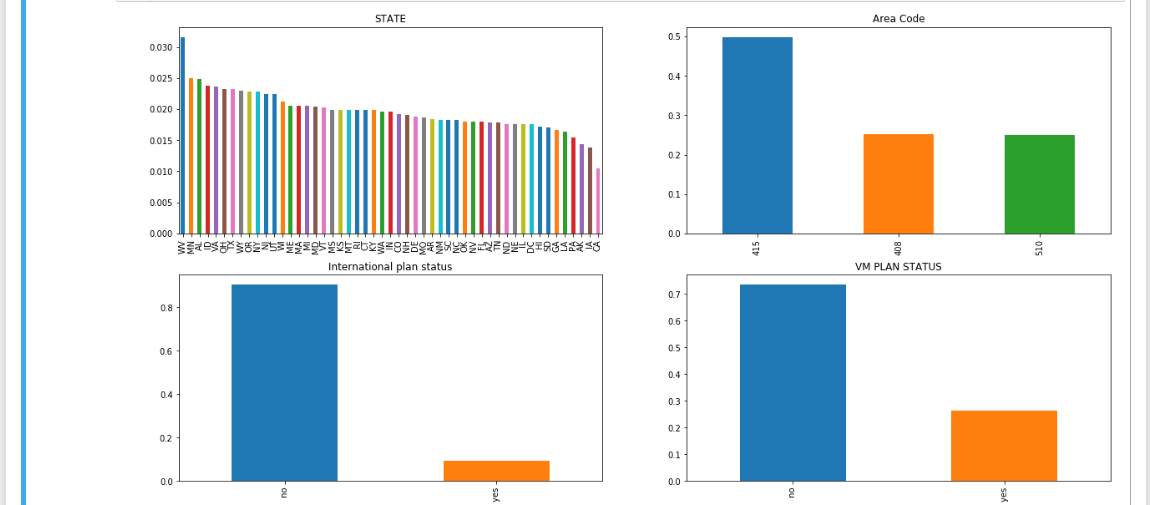




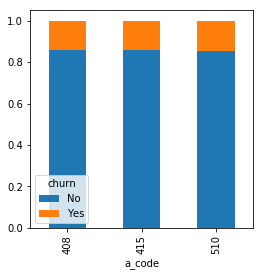
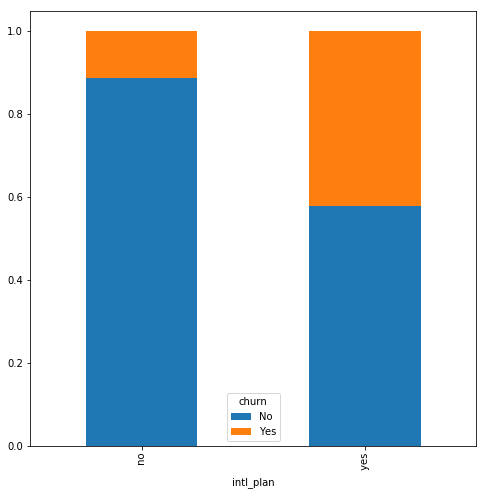
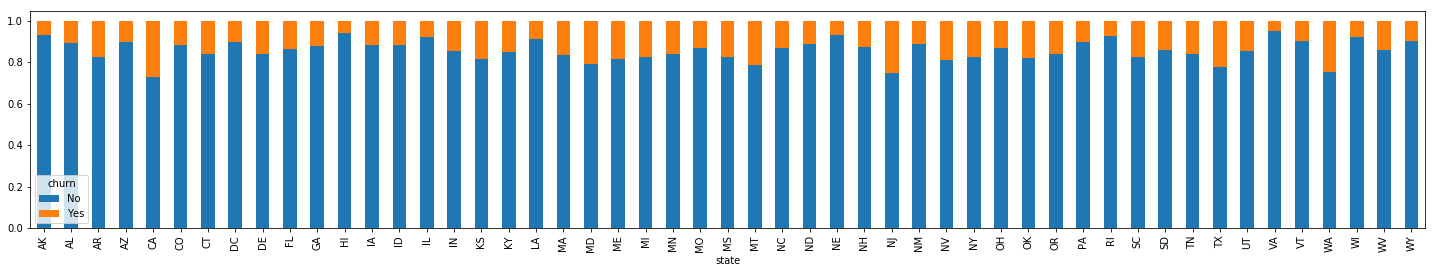
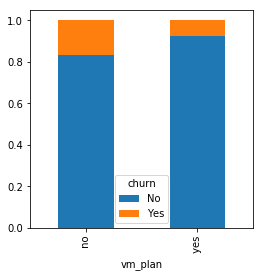


* 1. **ANALYSIS of CATAGORICAL VARIABLES**

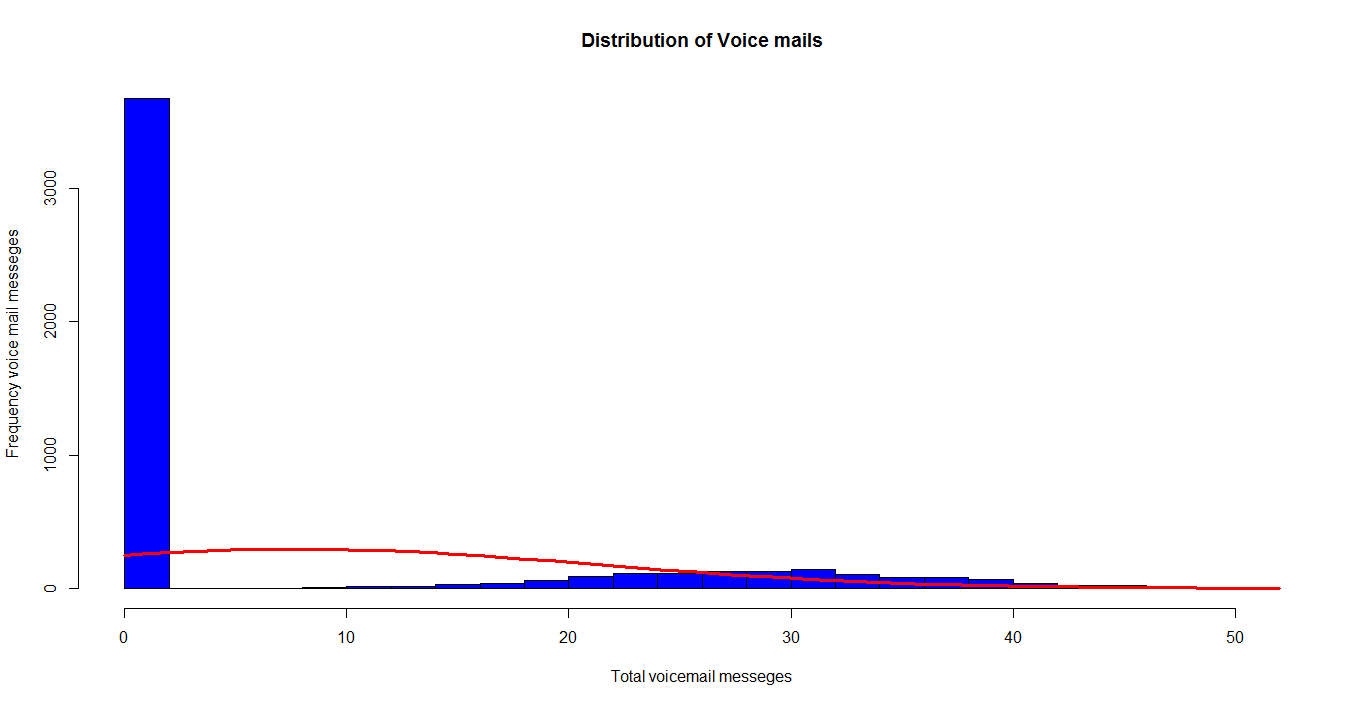
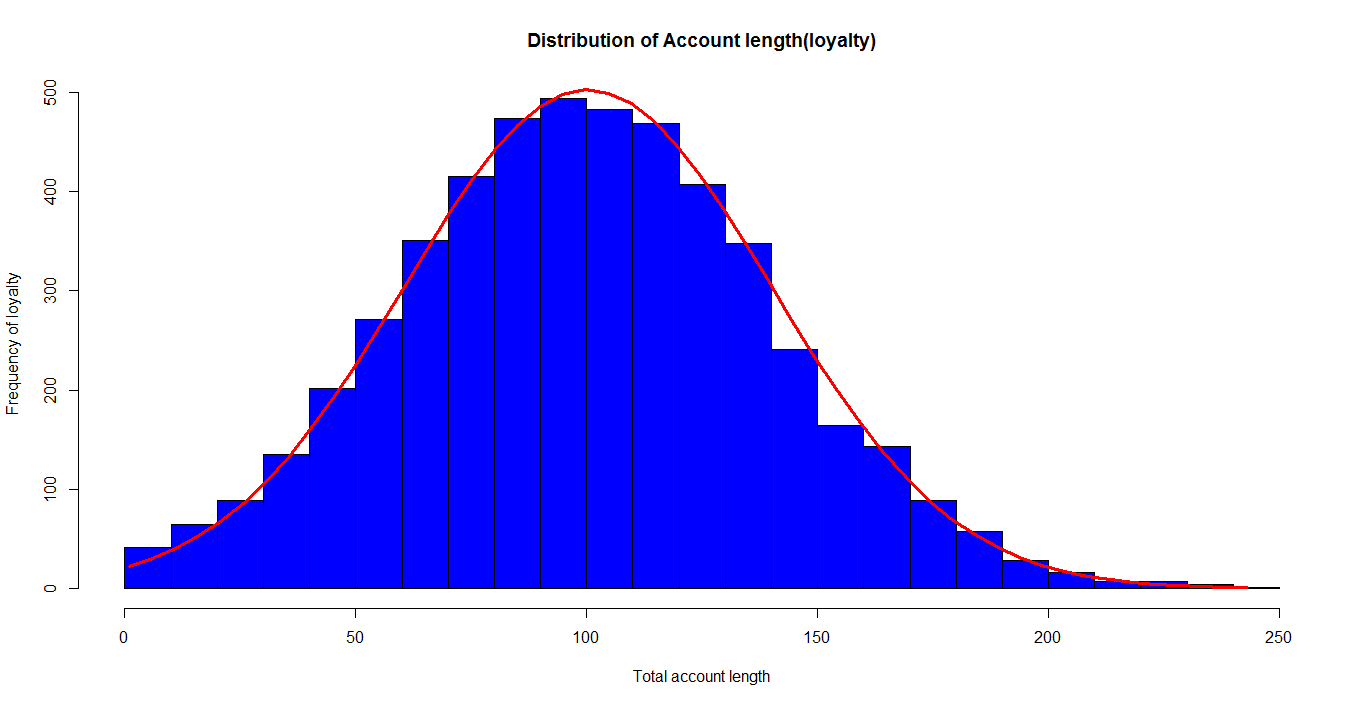
****

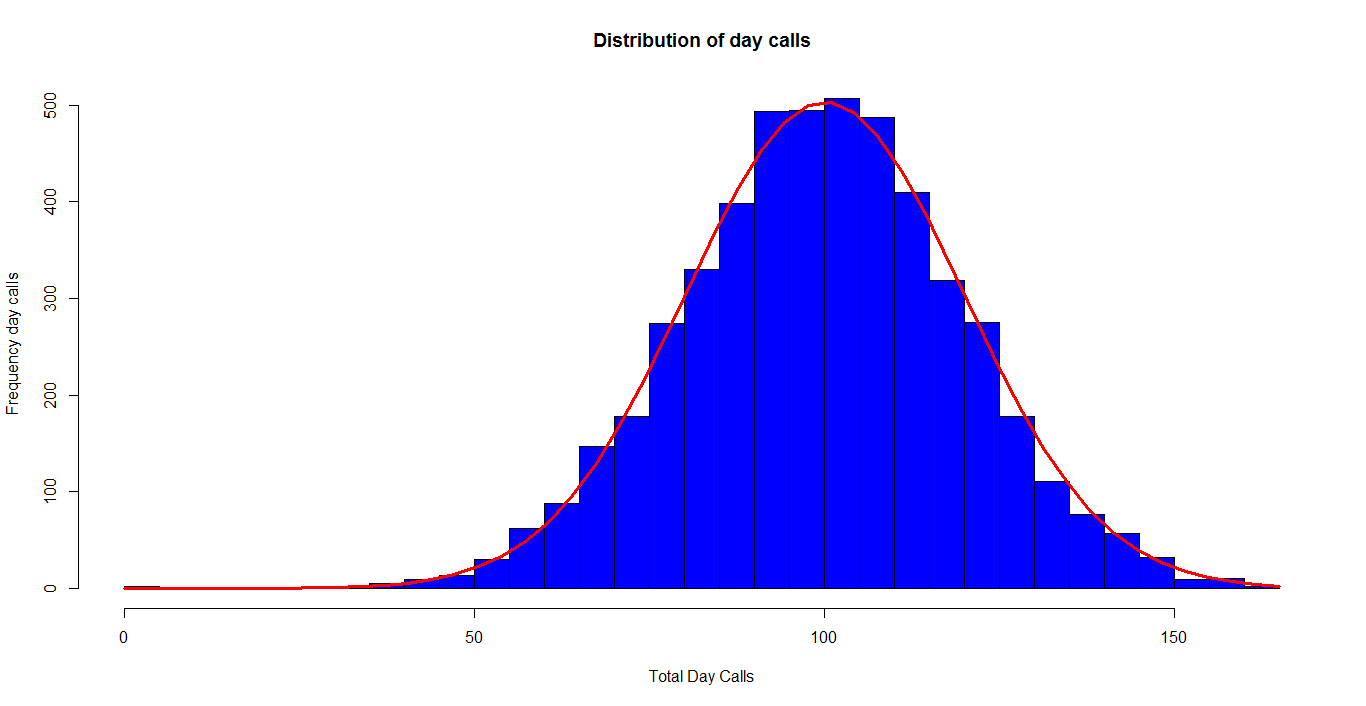
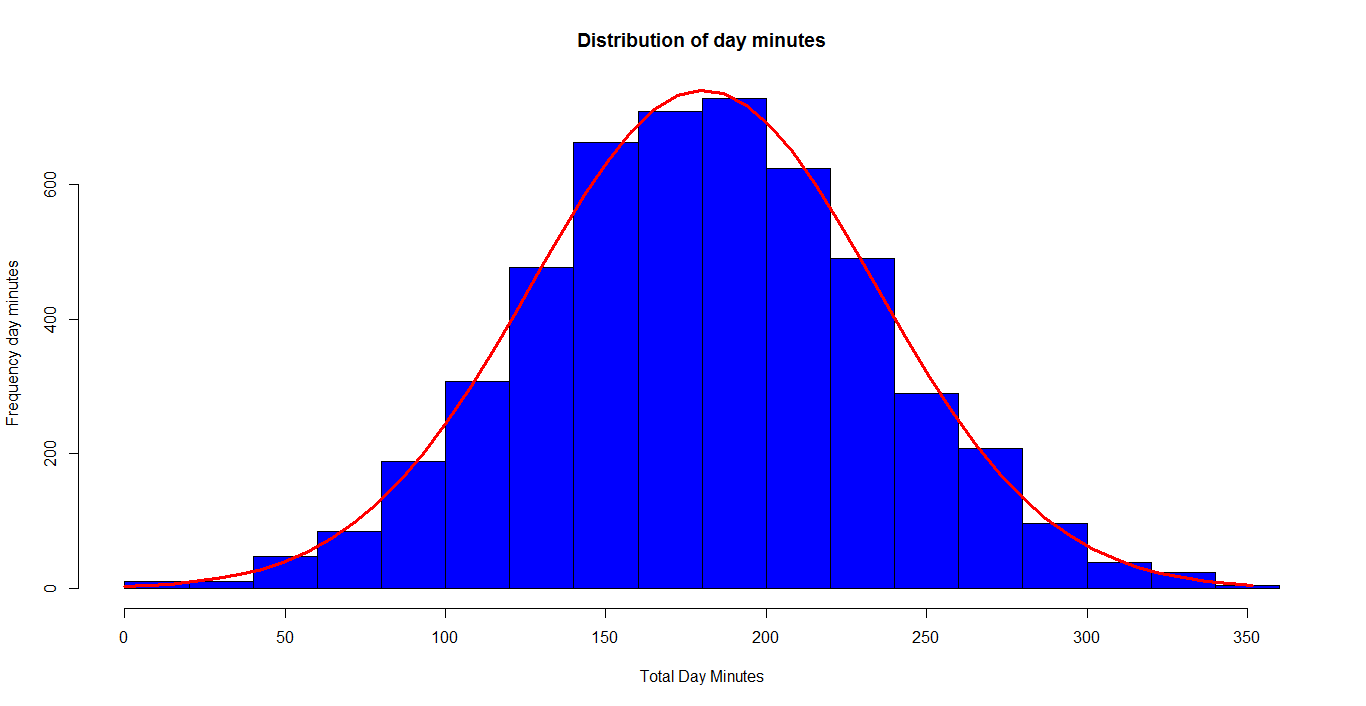


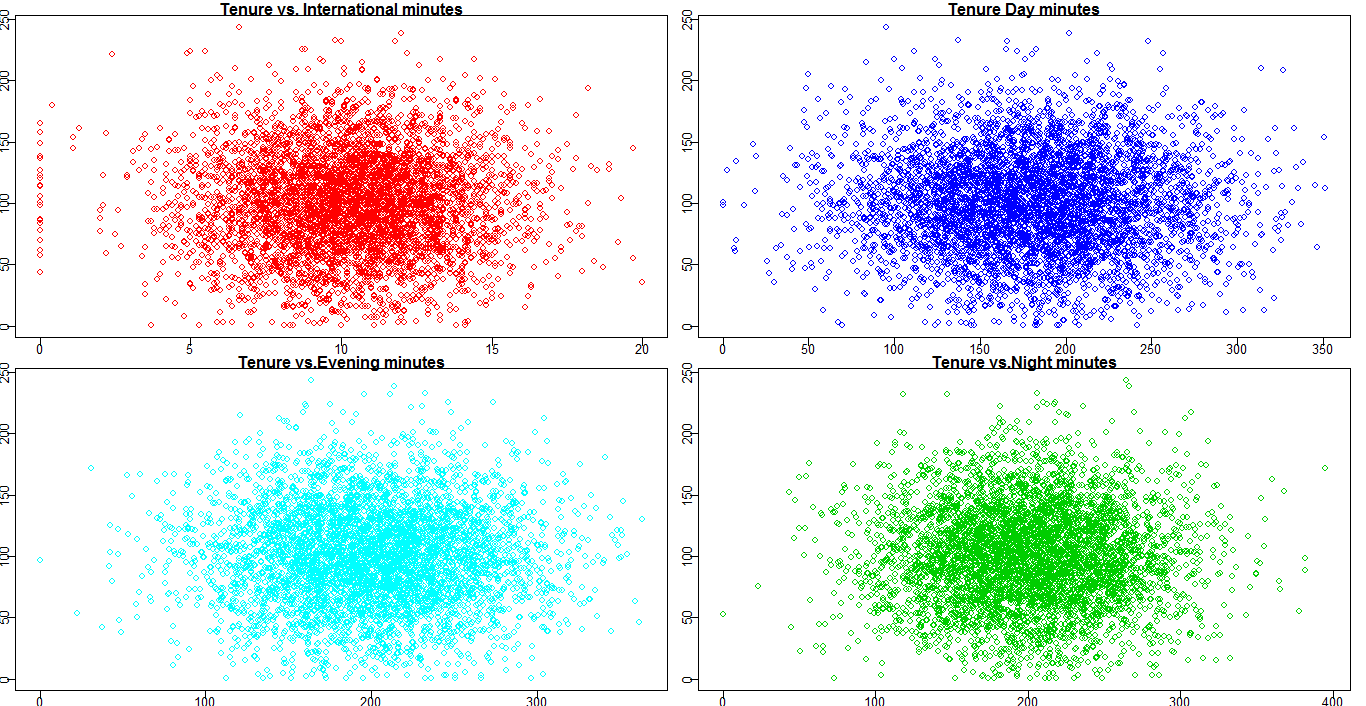
Class distribution with respect to the different categorical variable

**  **

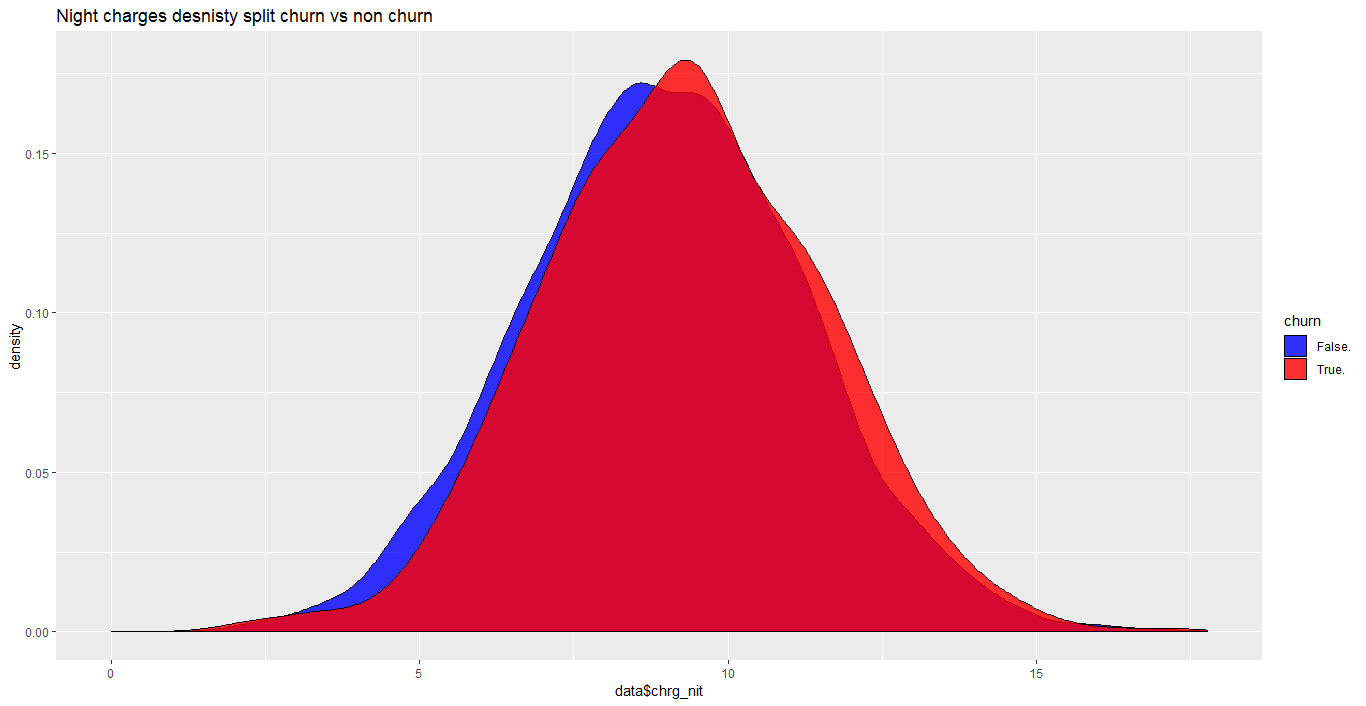
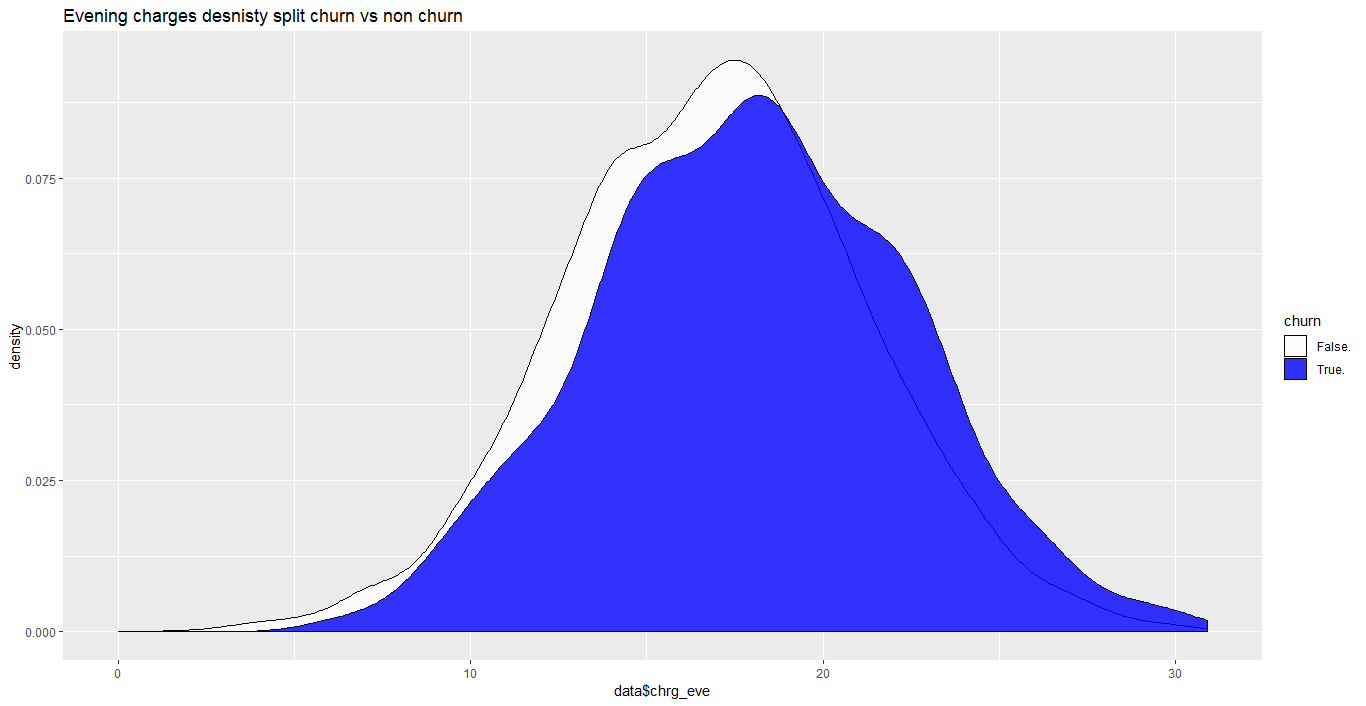
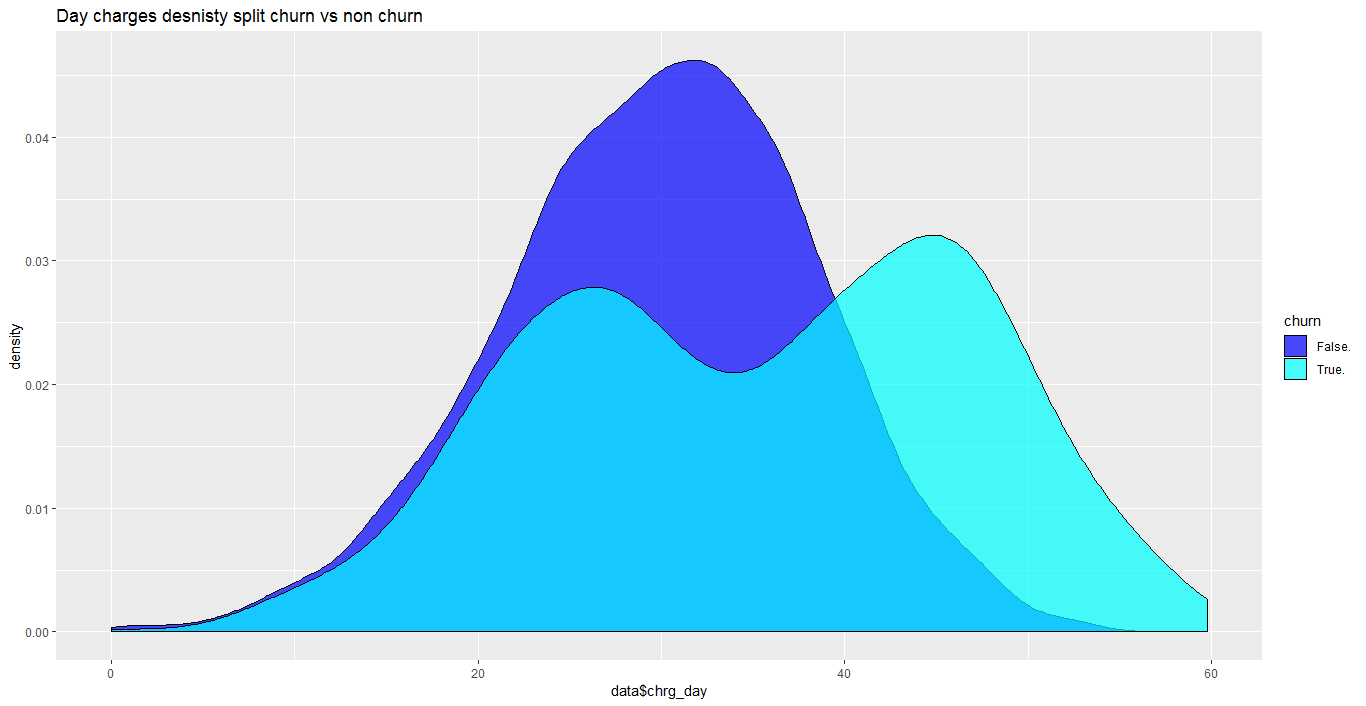
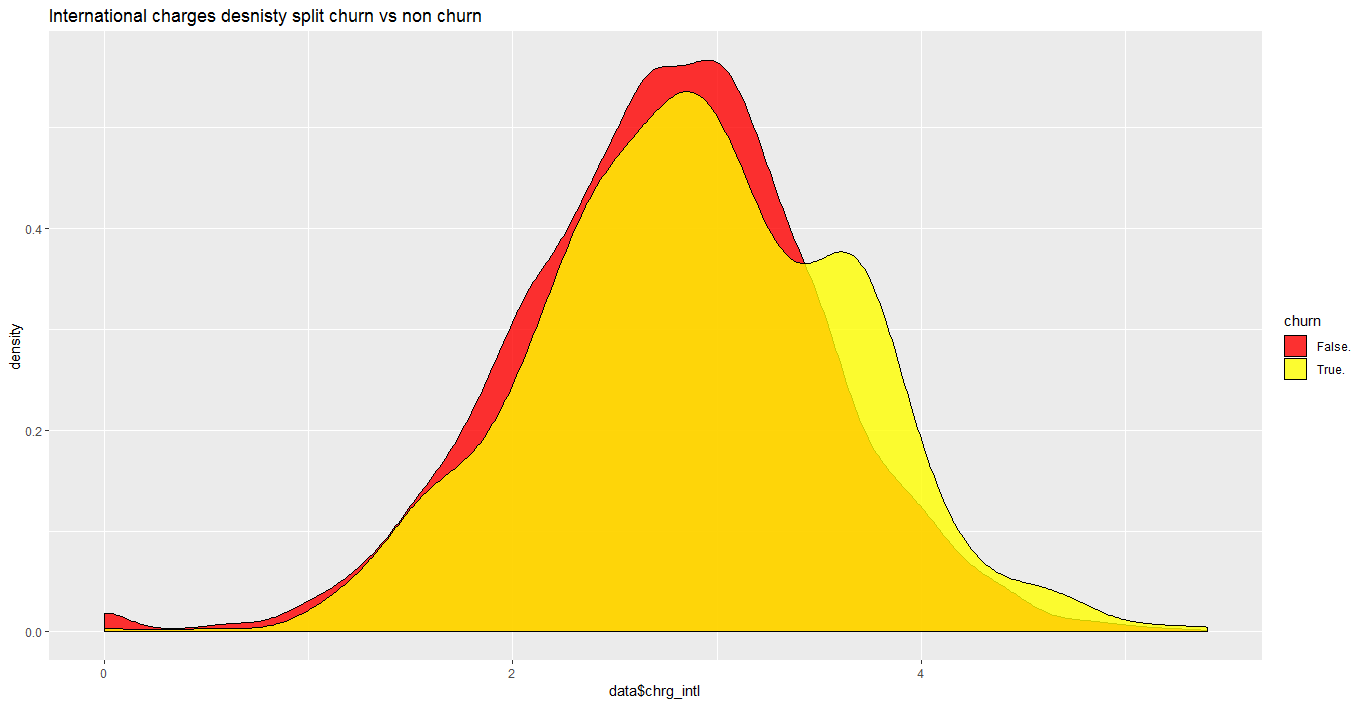
* 1. **Independent Variables(Numerical)**

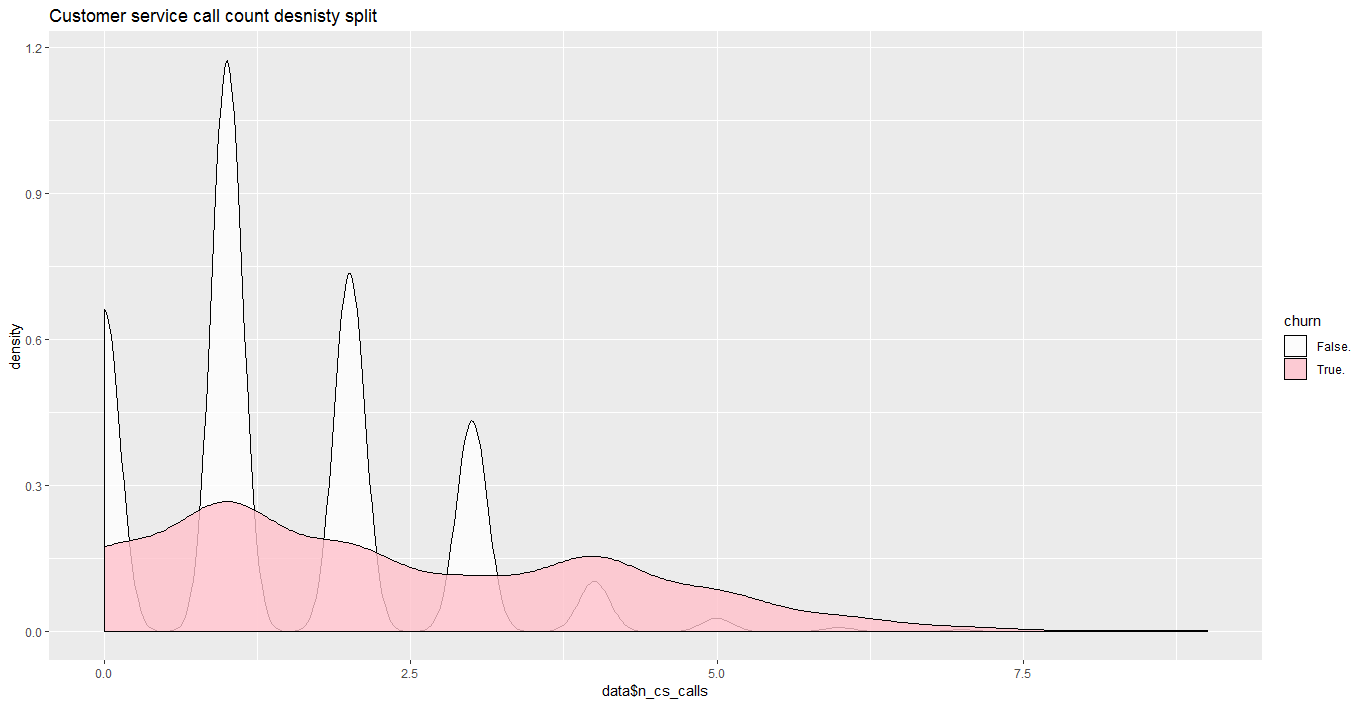




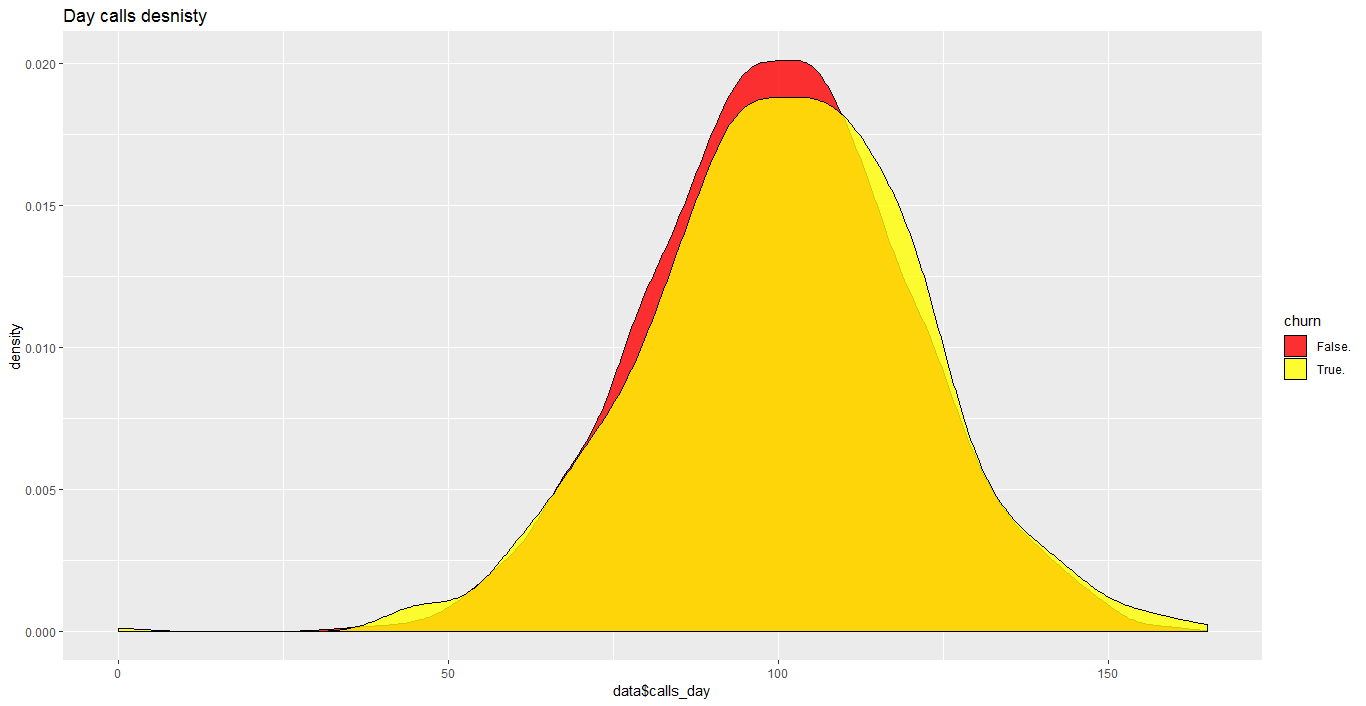
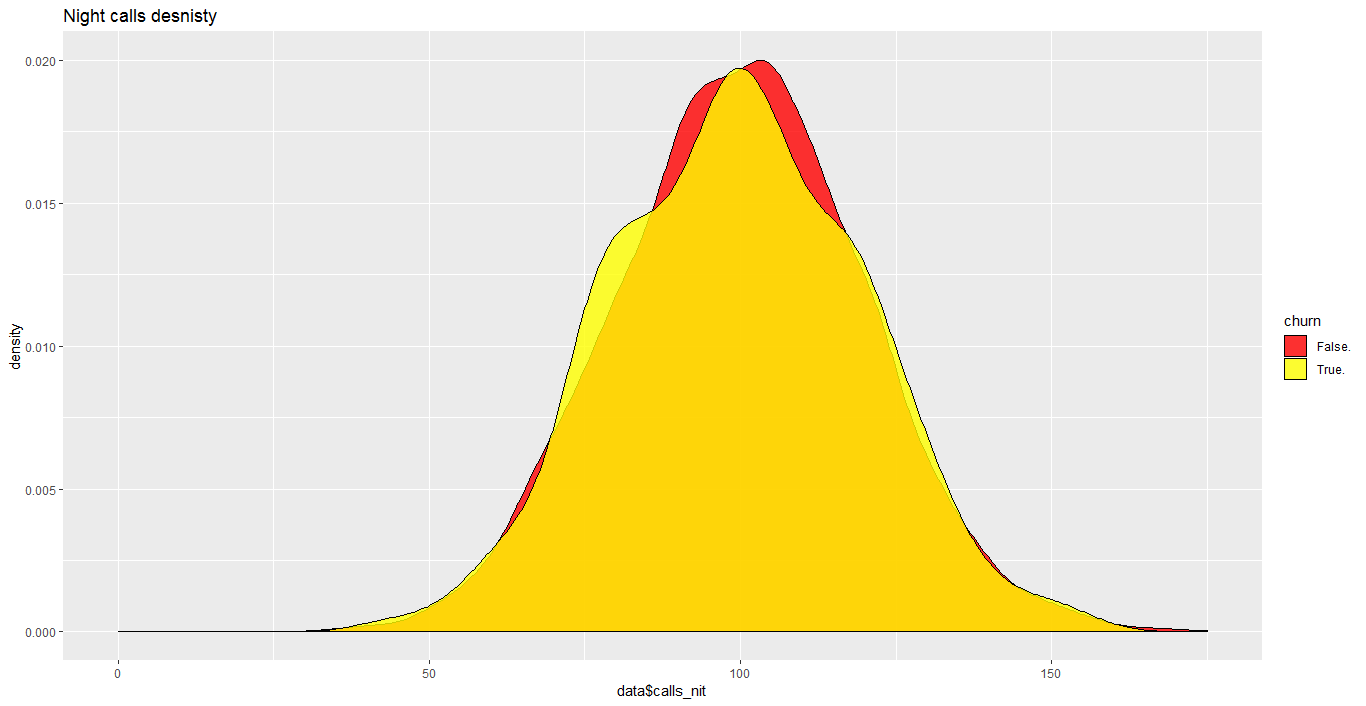
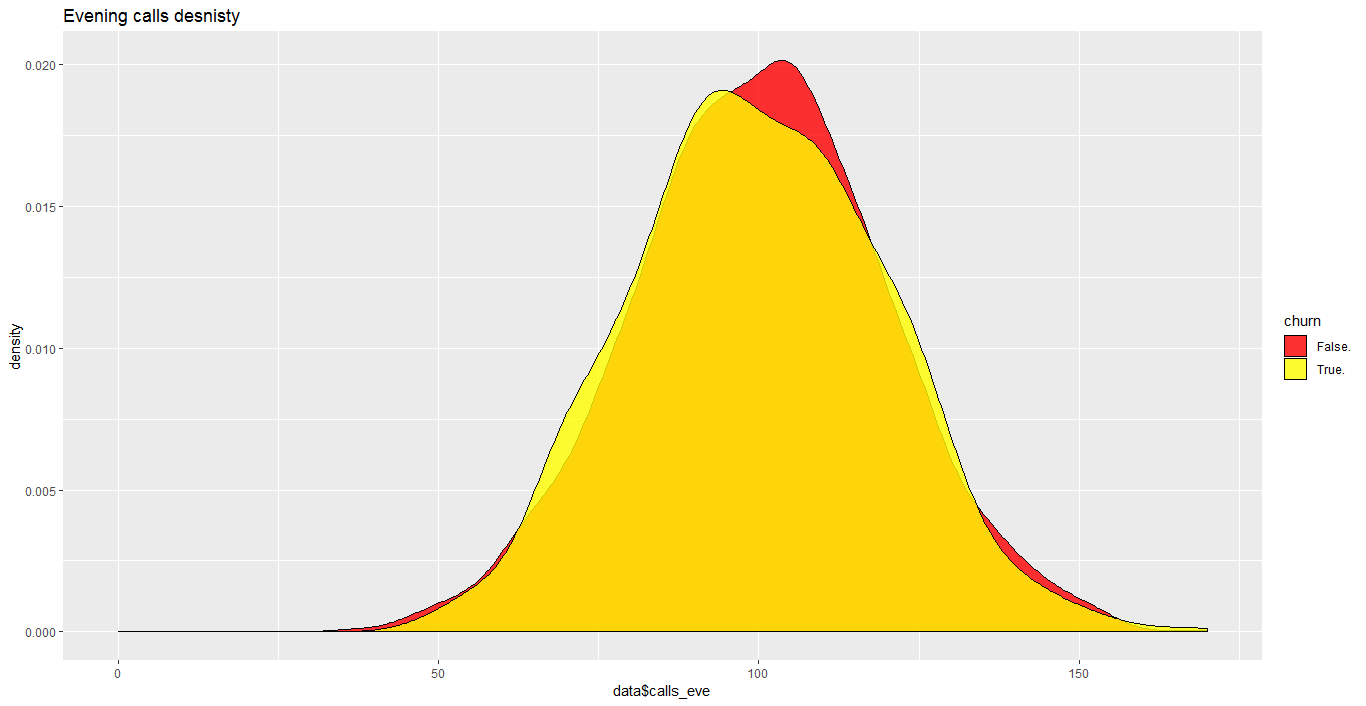
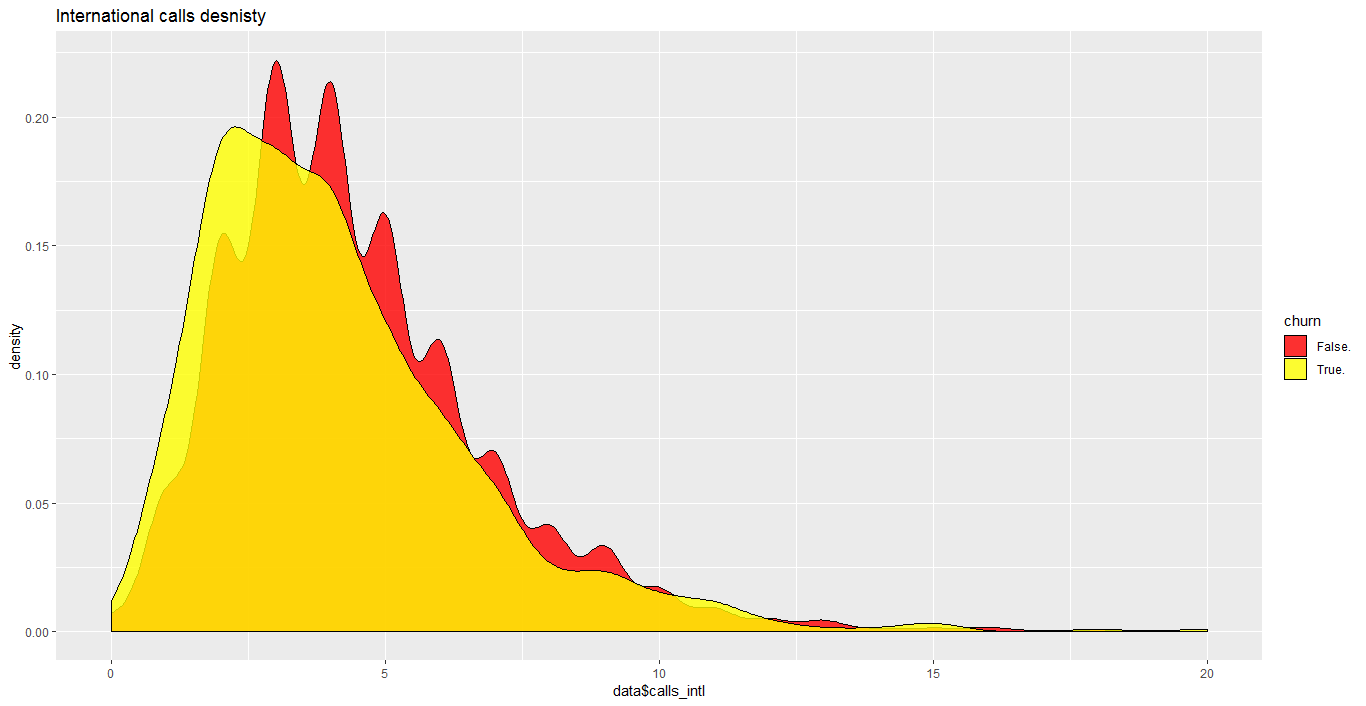


The above figures show that all the variables are normally distributed with number of voice mails as exception.

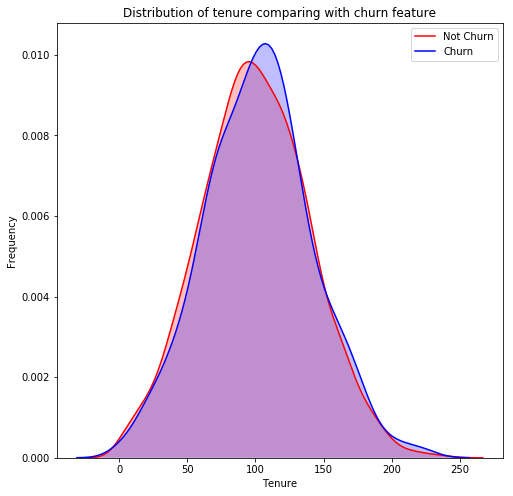
Above are the density plots for the charges w.r.t the customer churn: international, Day, Evening and Nights respectively.



Above are the density plots for the count of customer service calls w.r.t the customer churn.



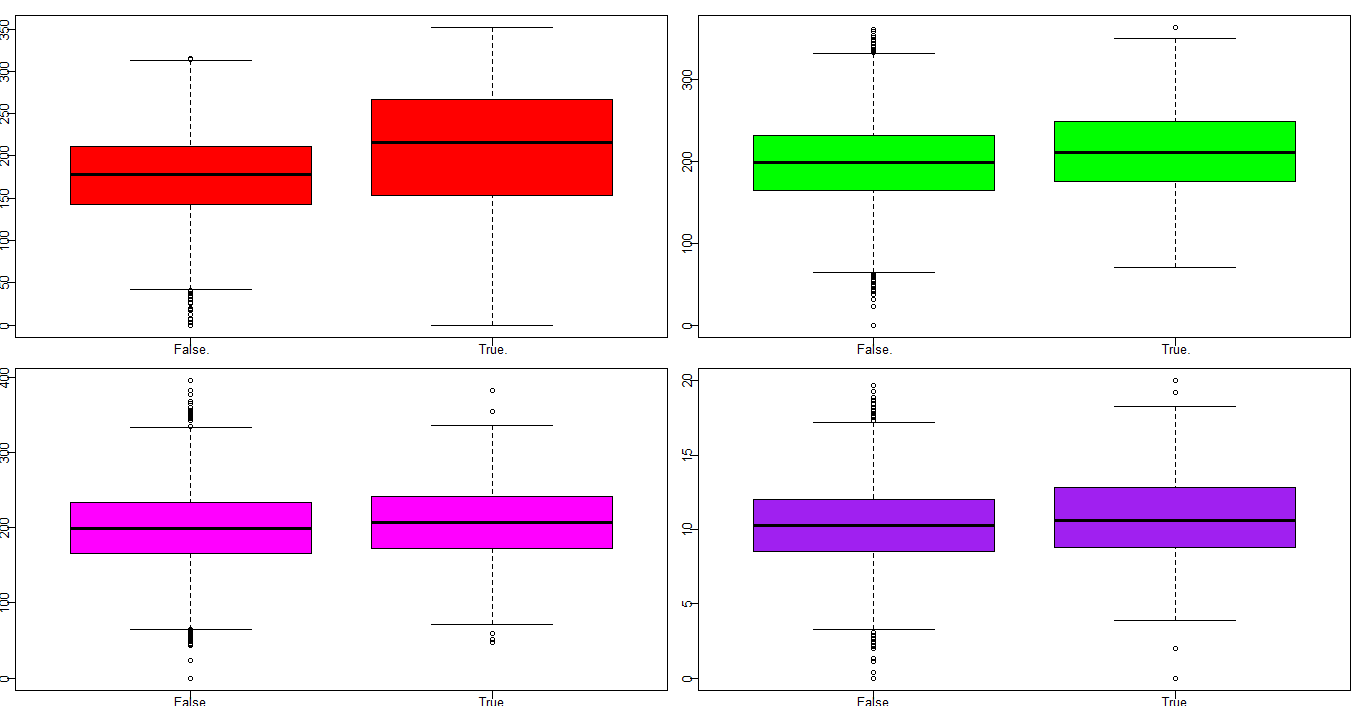
Above are the density plots for the calls w.r.t the customer churn: international, Evening, Night and Day respectively.



As the kdeplot is almost overlapping, we are unable to get any significant impact of tenure on the customer churn.

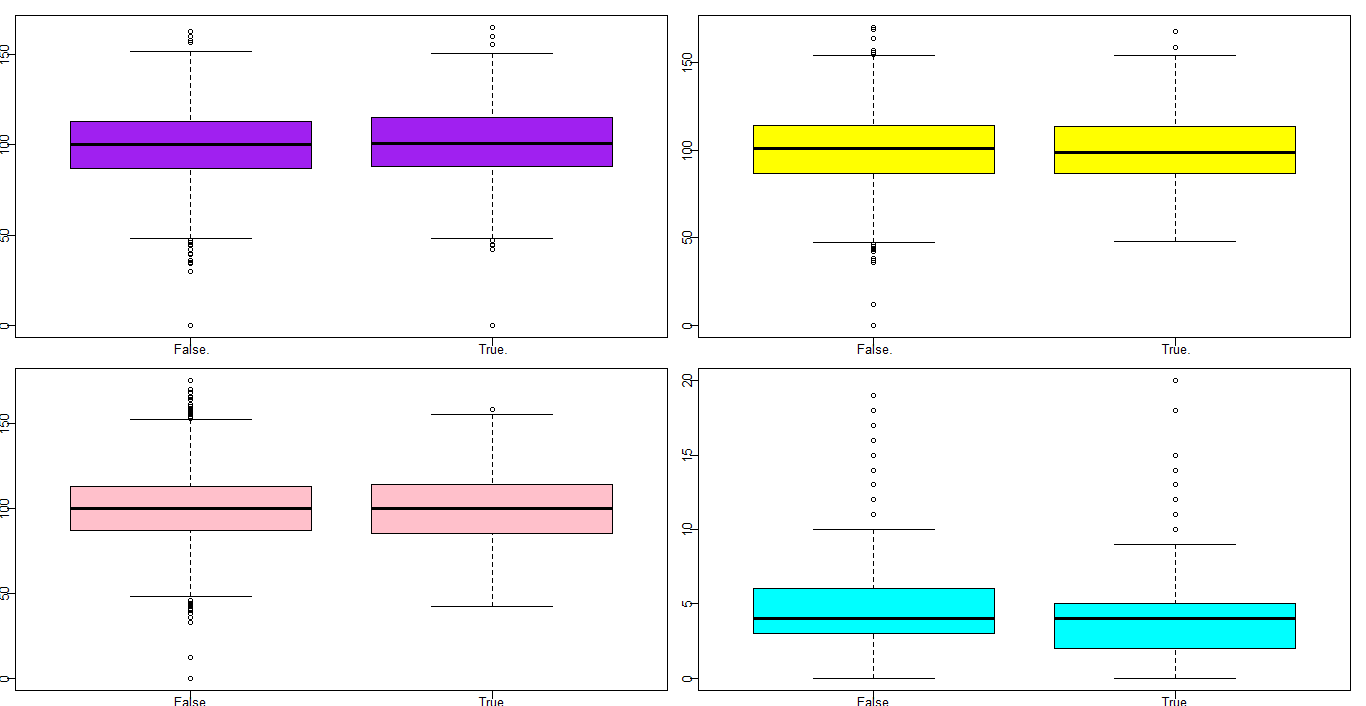
Below are the boxplots with individual classes in the target for different attributes.

1. # Boxplot to see the pattern of Total Minutes (Day/Evening/Night/International) for Churned / Not Churned Customers



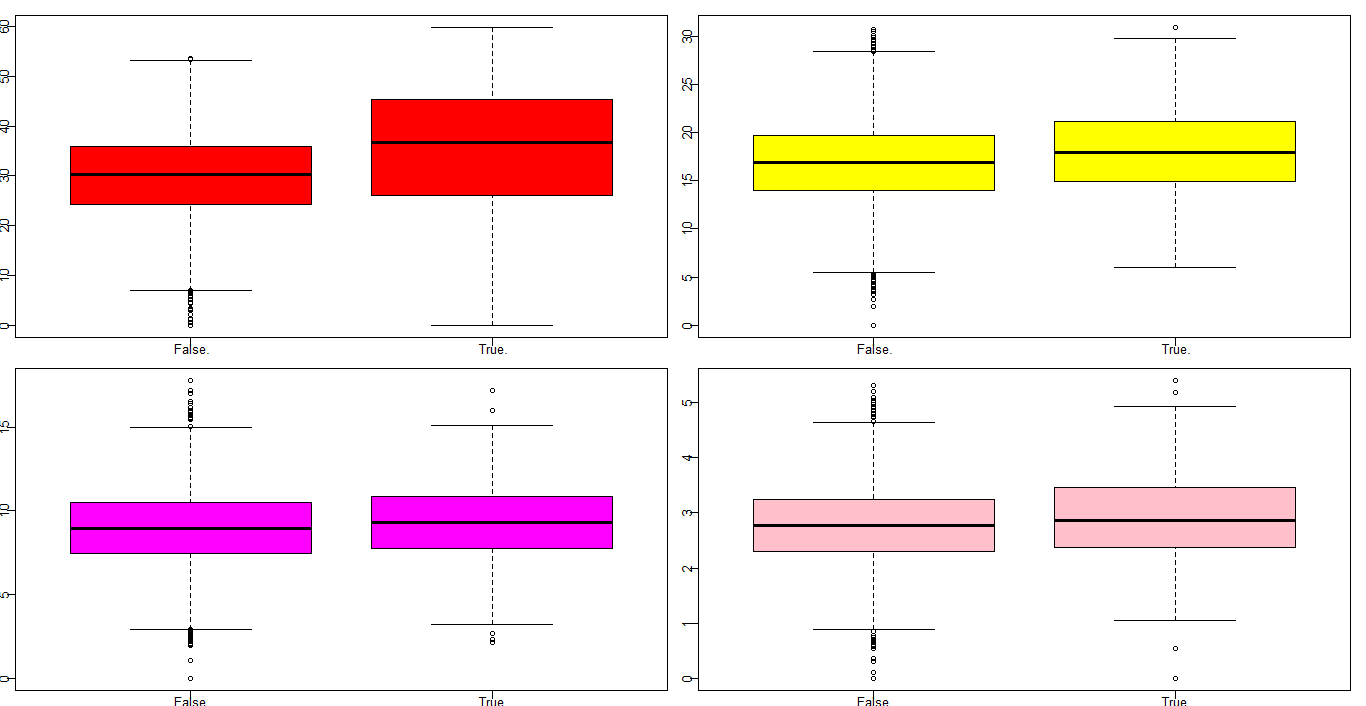
* The Q1,Q3, IQR & the area of the box for Total Day mins ,Total Evening mins & Total International mins for Churned Customer is higher than Not Churned Customers.
* Hence, usage pattern of churned customers is high as compared to Not Churned.
* There is no significant difference in the usage of Total Night Calls for churned and not churned Customers. Outliers are present in all the cases especially significant in Total Int mins, Total Evening Mins & Total Night mins.

1. Pattern of Total Cals (Day/Evening/Night/International) for Churned / Not Churned Customers



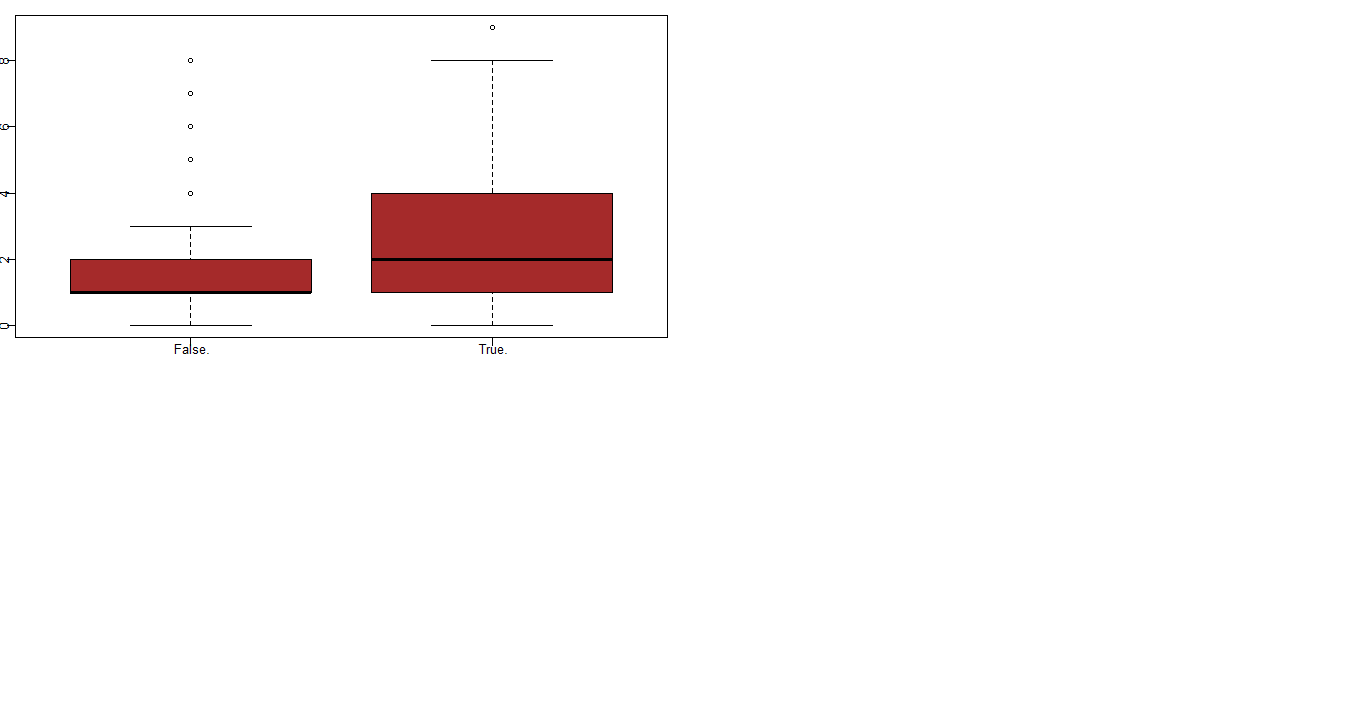
* There is no significant difference in the above plots for churned and non-churned segment of customers except Total Night Calls where spread of IQR for Chruned Customers is relatively higher.
* The Q1, Q3 and upper whiskers in Total International Calls for Not Churned Customers is relatively higher than the Churned Customers. So, can we say that Customers Churned, call less relatively but the duration of their calls are relatively high (mins used from the previous box plot)??

1. # Boxplot to see the pattern of Total Charges (Day/Evening/Night/International) for Churned / Not Churned Customers



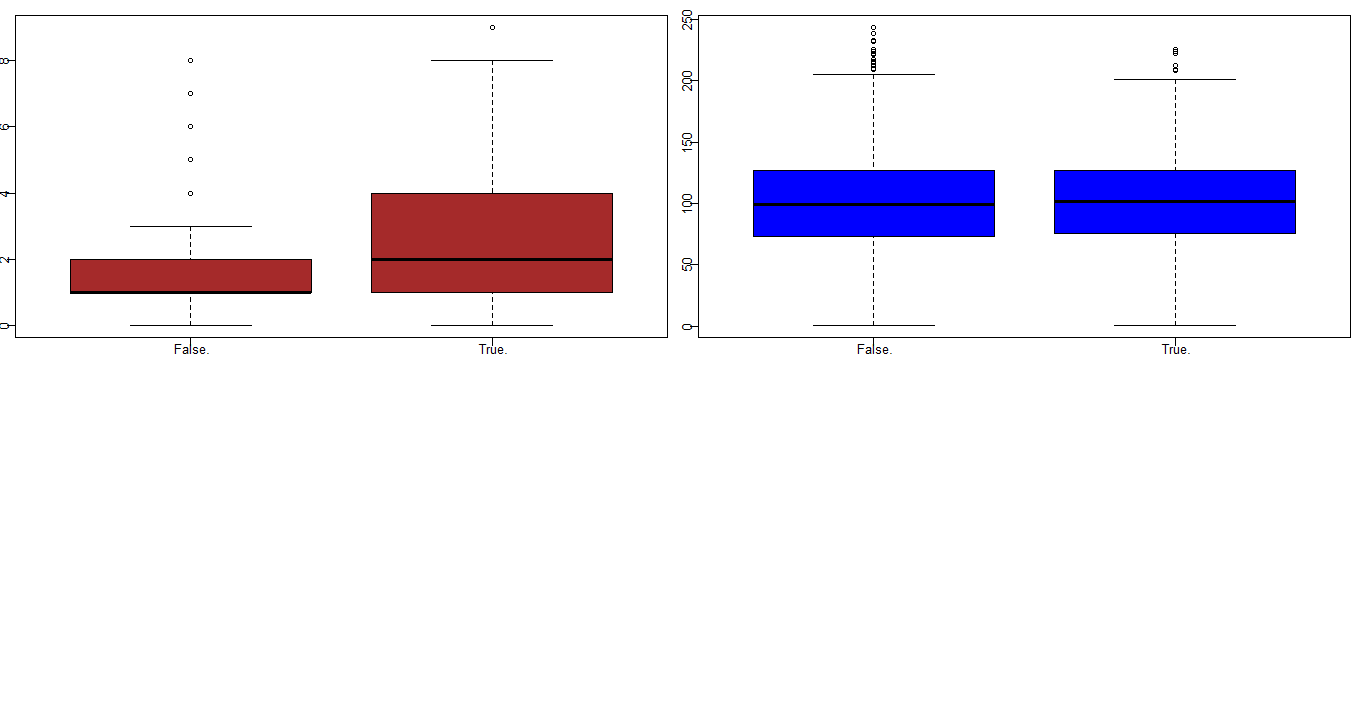
* The Q1,Q3, IQR & the area of the box for Total Day Charges for Churned Customer is significantly higher than Not Churned Customers. And relatively higher for Churned Customers for all other cases as well.

1. Boxplot for total number of customer service calls.

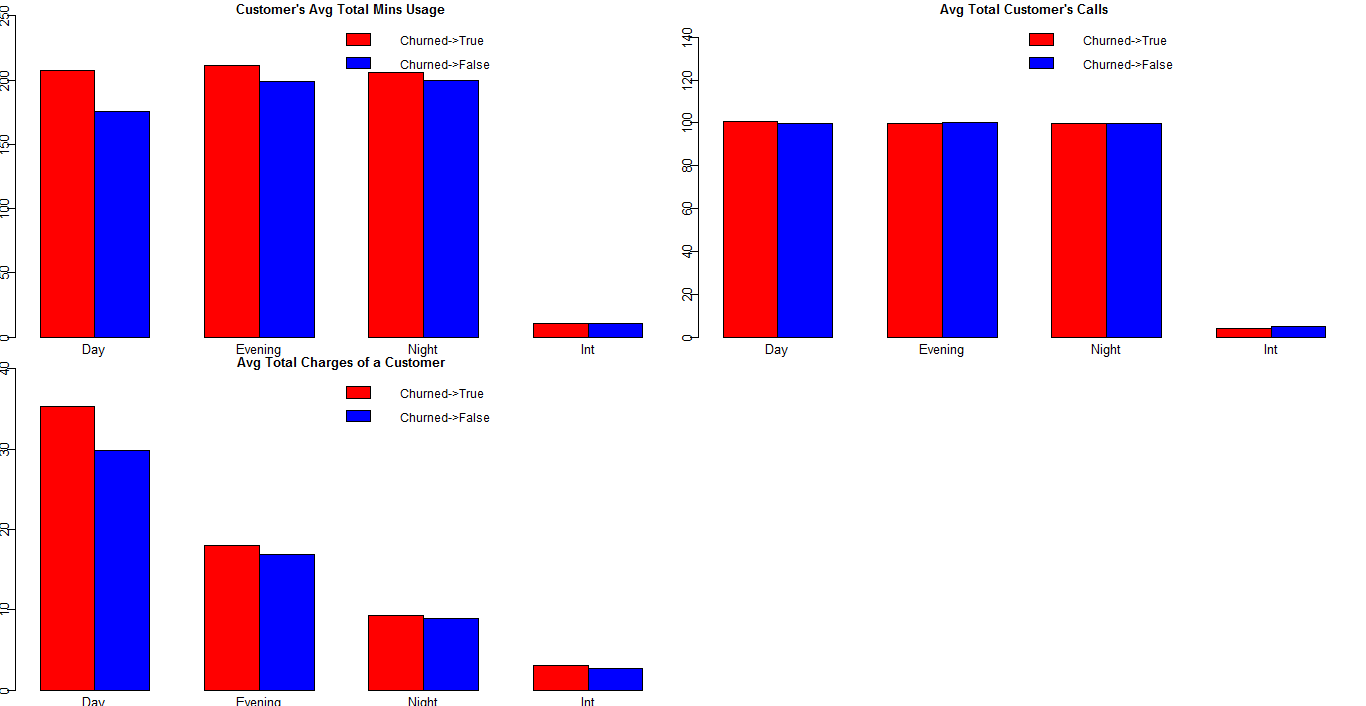


* As observed, the spread for Customer Calls made by Churned Customer is significantly more than Not churned ones. It seems Customers going to be churned call Customer Service a lot with their issues regarding service provided.

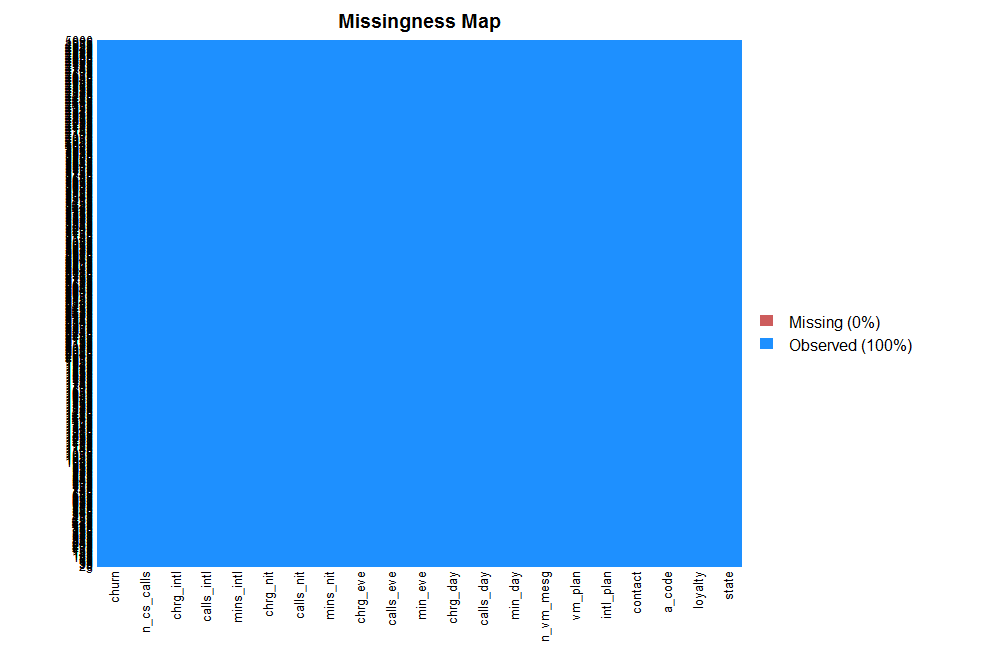
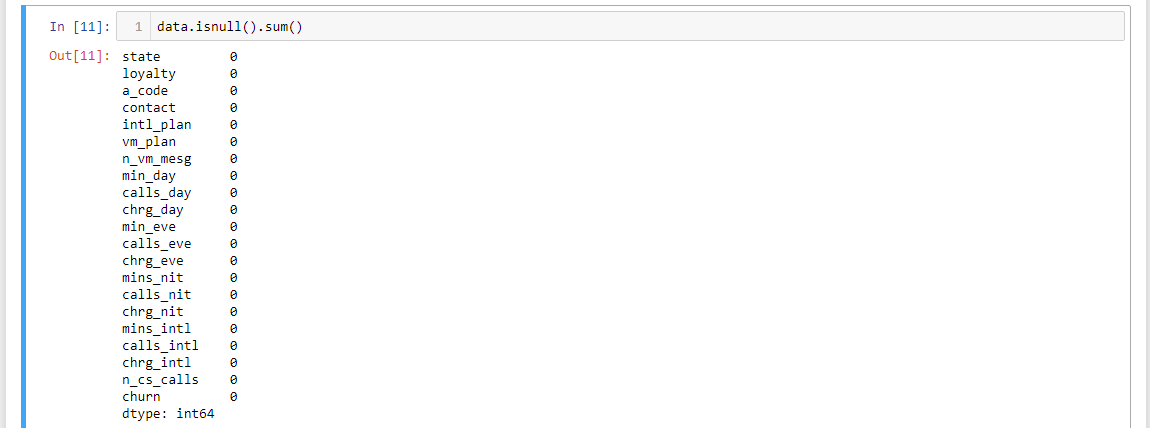
1. # Boxplot to see the pattern of tenure/loyalty for Churned / Not Churned Customers



1. Barplot for different attributes grouped with respect to the churned and retained(non-churned) customers.

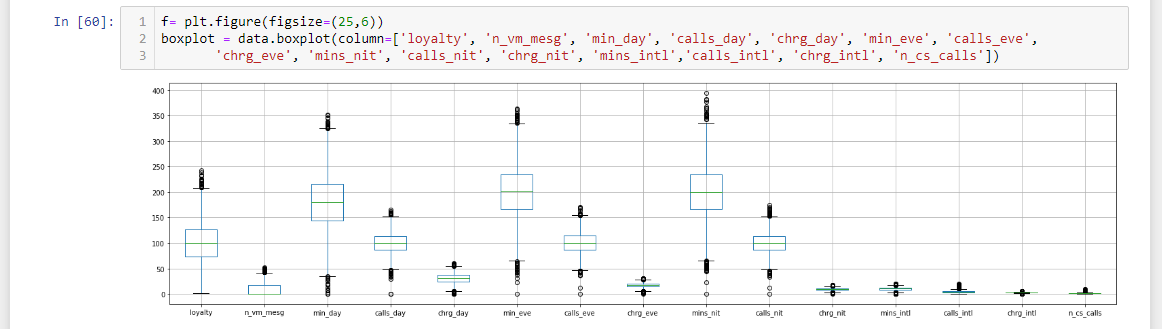


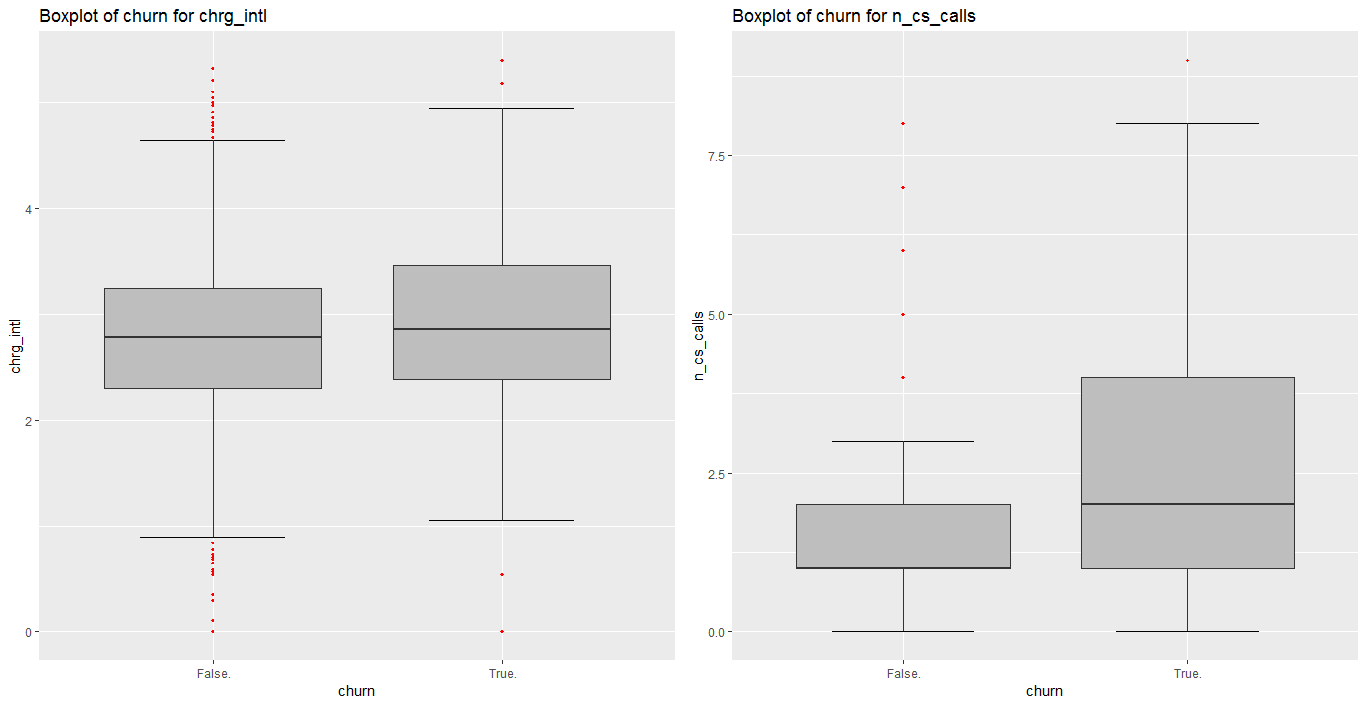
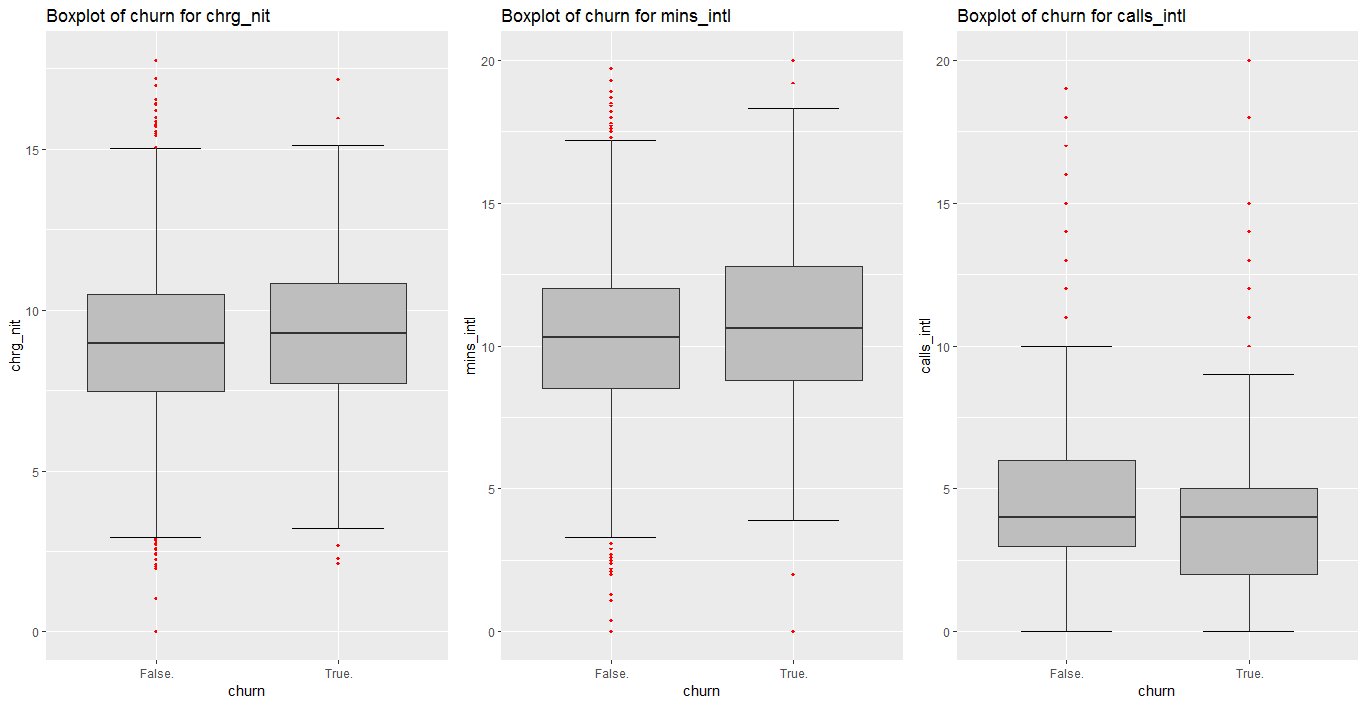
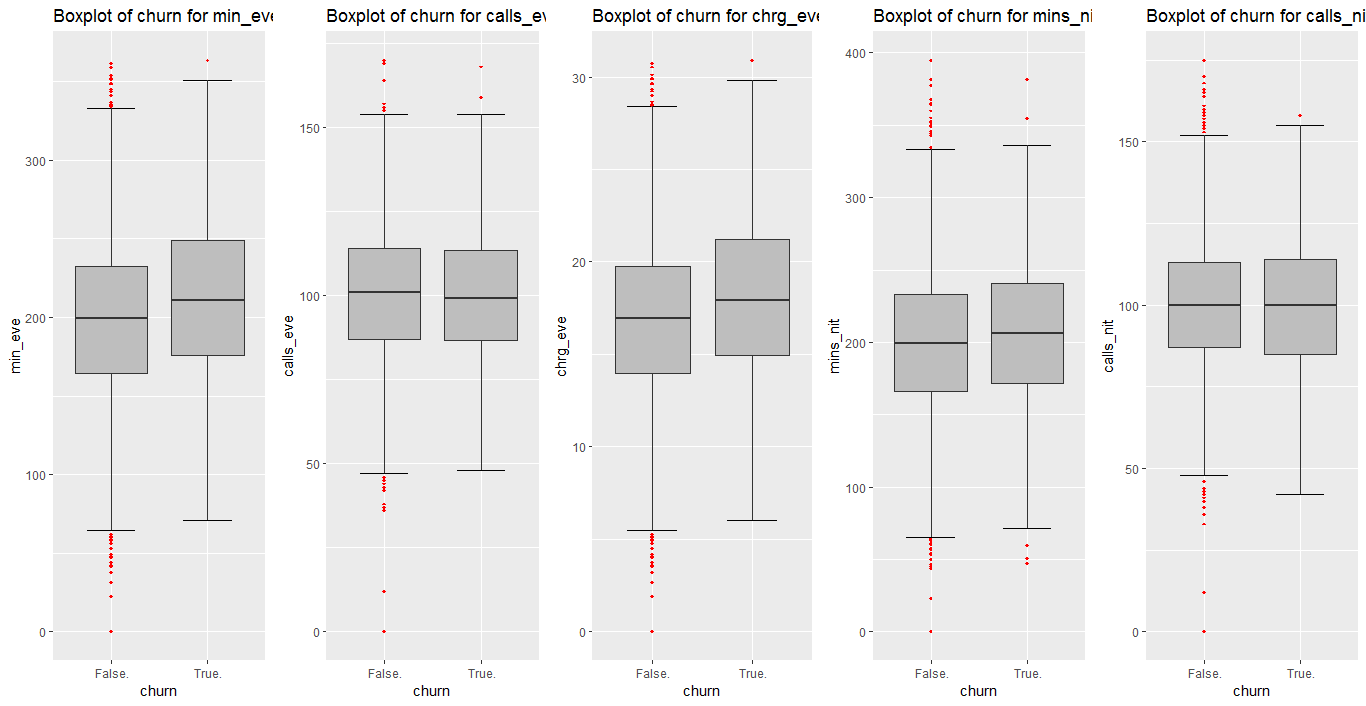
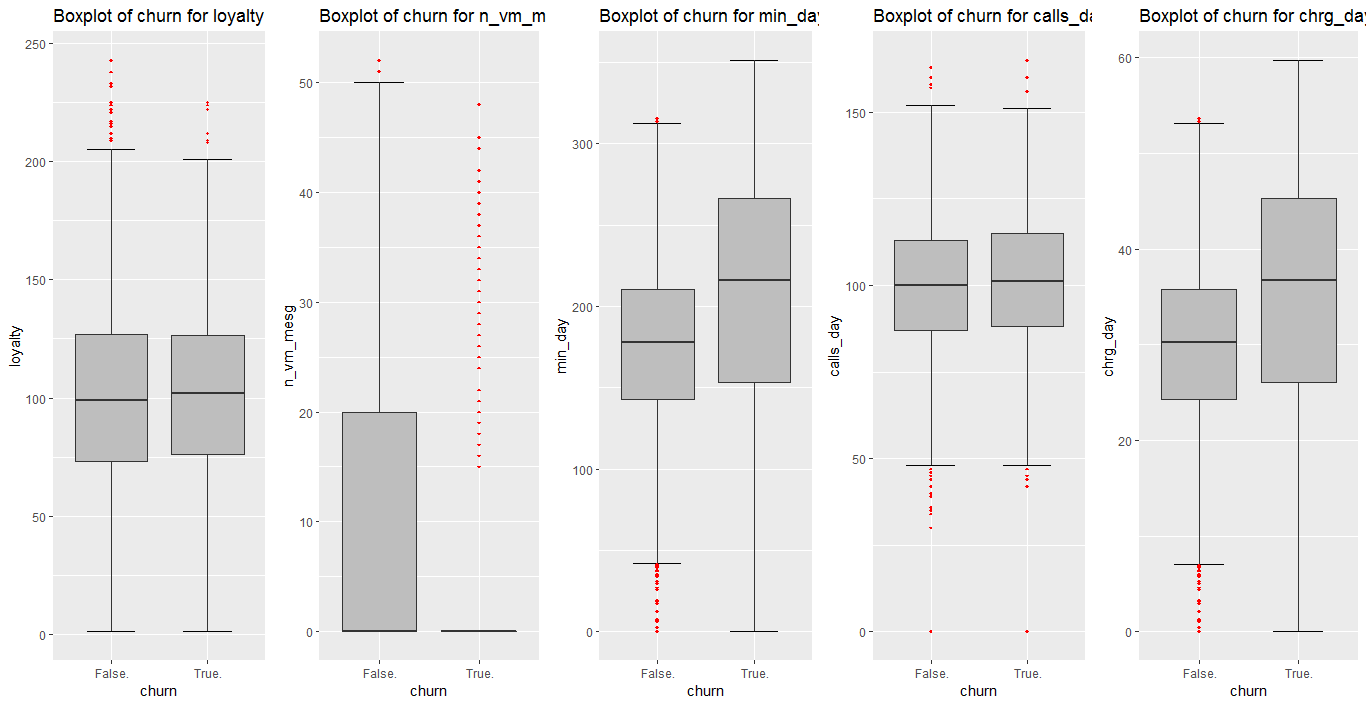
* 1. **Missing Value Analysis**



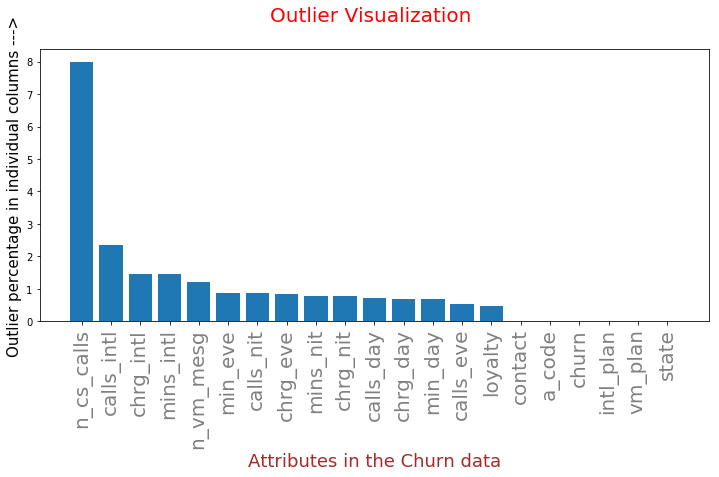
The dataset is clean and doesn’t have any missing value. The missing value map is plain blue. missmap() function in from Amelia package in R is used to plot the same.

* 1. **Outlier Analysis**



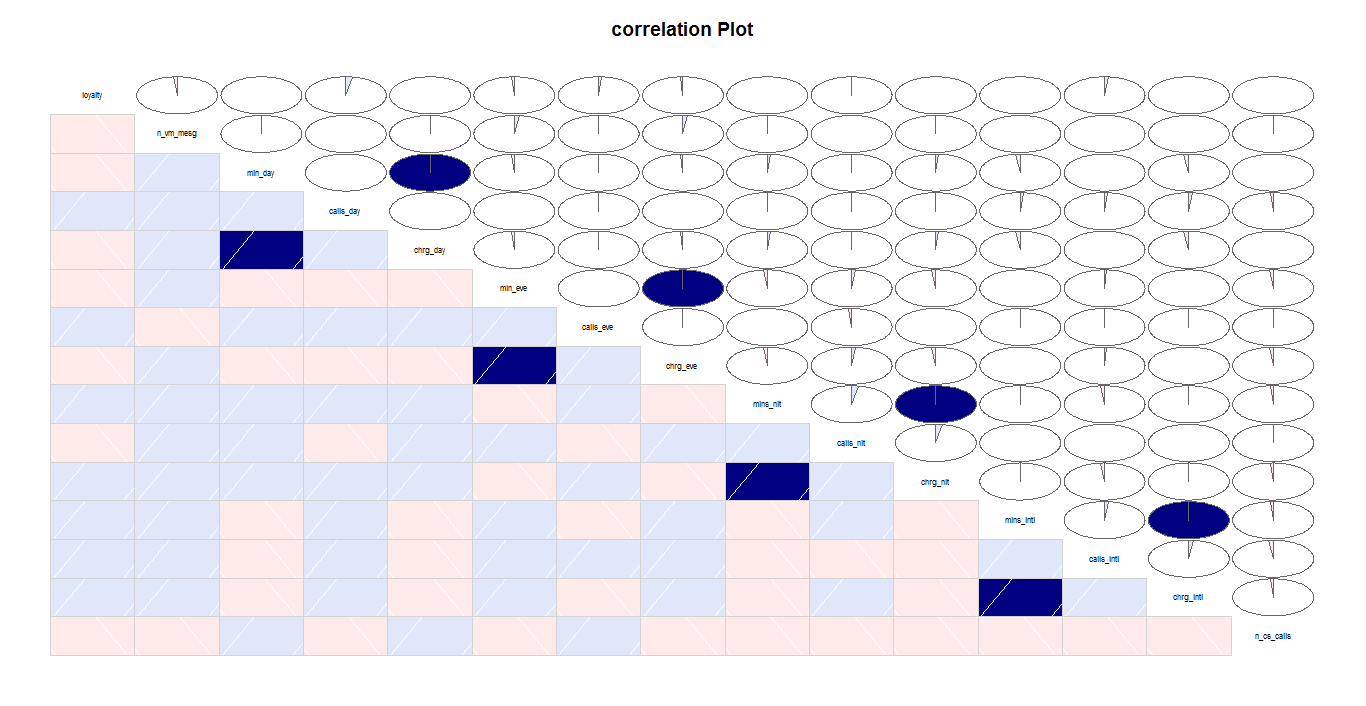


The outliers in the data are significant in number.



The outliers have been deleted and replaced with column means.

* 1. **Correlation**



The correlated variables are mins\_day~chrge\_day, mins\_eve~chrge\_eve, mins\_nit~chrge\_nit and mins\_intl~chrge\_intl

* 1. **Feature Selection**

Two different methodologies have been used for feature selection.

In R we have use PCA as the dimensionality reduction technique to select the features. PCA extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible.

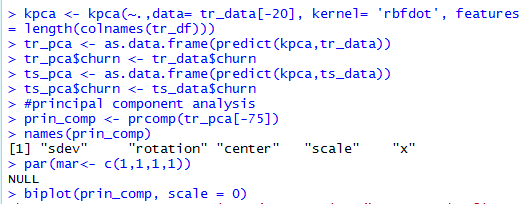
A principal component is a normalized linear combination of the original predictors in a data set. These principal components are orthogonal to the original dimensions and aim to capture as much information as possible with high explained variance.

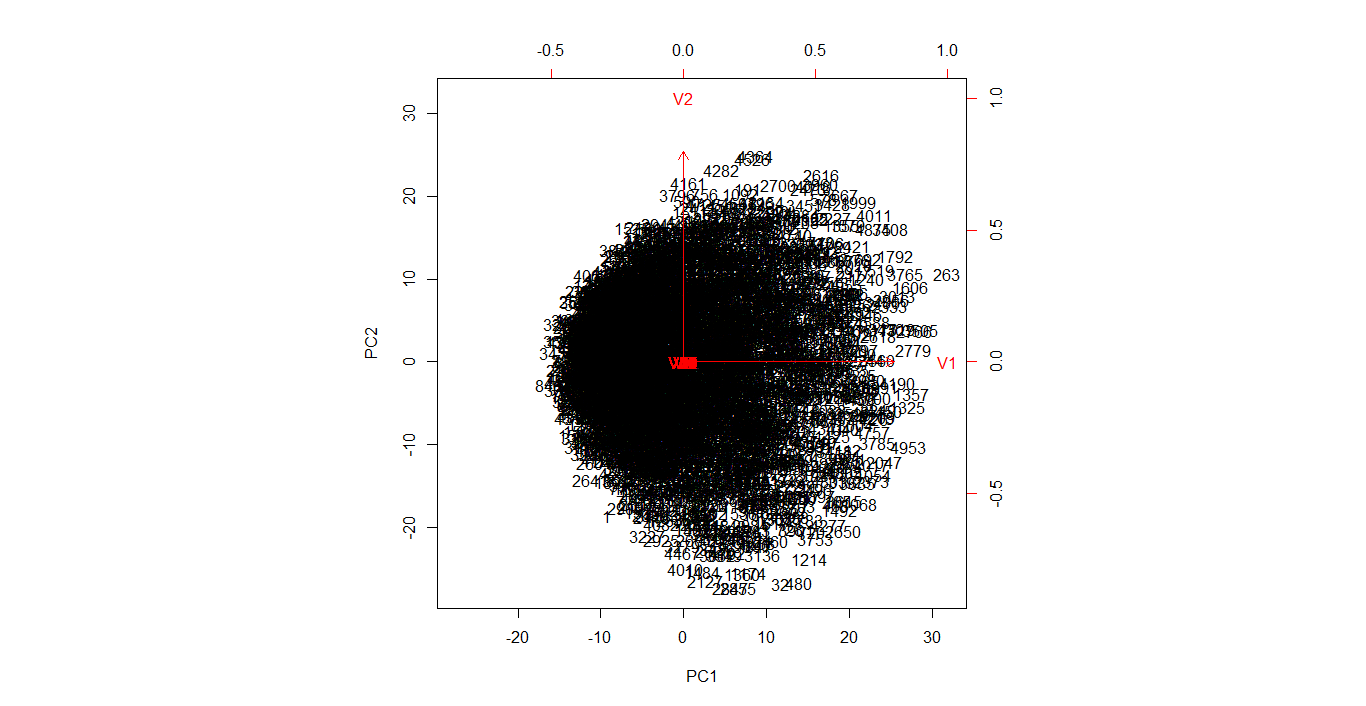
These components are resultant of normalized linear combination of original predictor variables. The first component has the highest variance followed by second, third and so on. Normalizing data becomes extremely important when the predictors are measured in different units.

R is used for PCA.

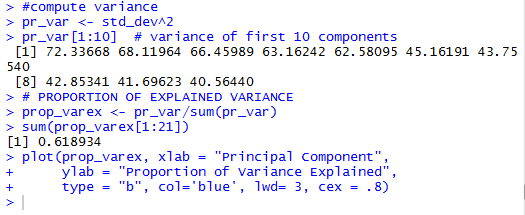
In the dataset we have 21 variables. After one hot encoding and standardizing we have 74 dimensions in total.

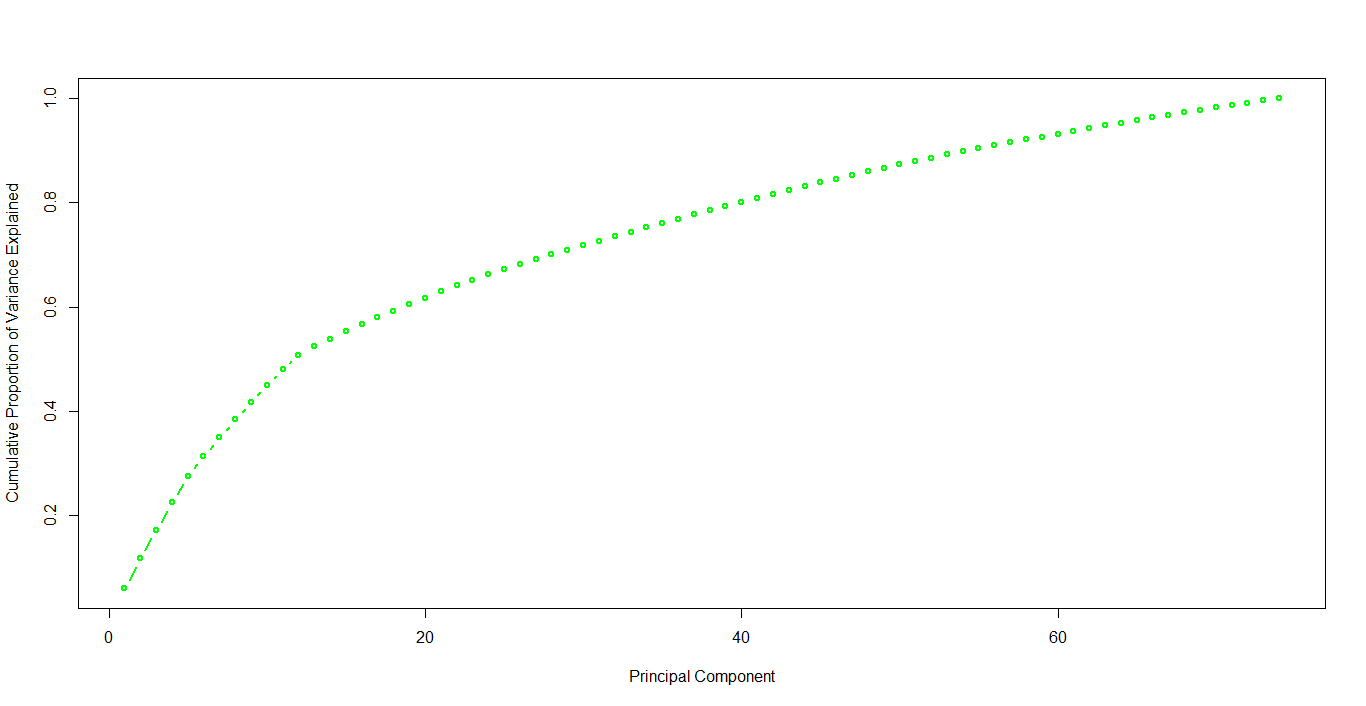
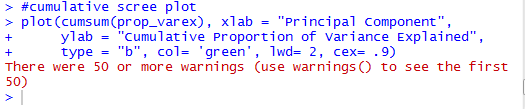
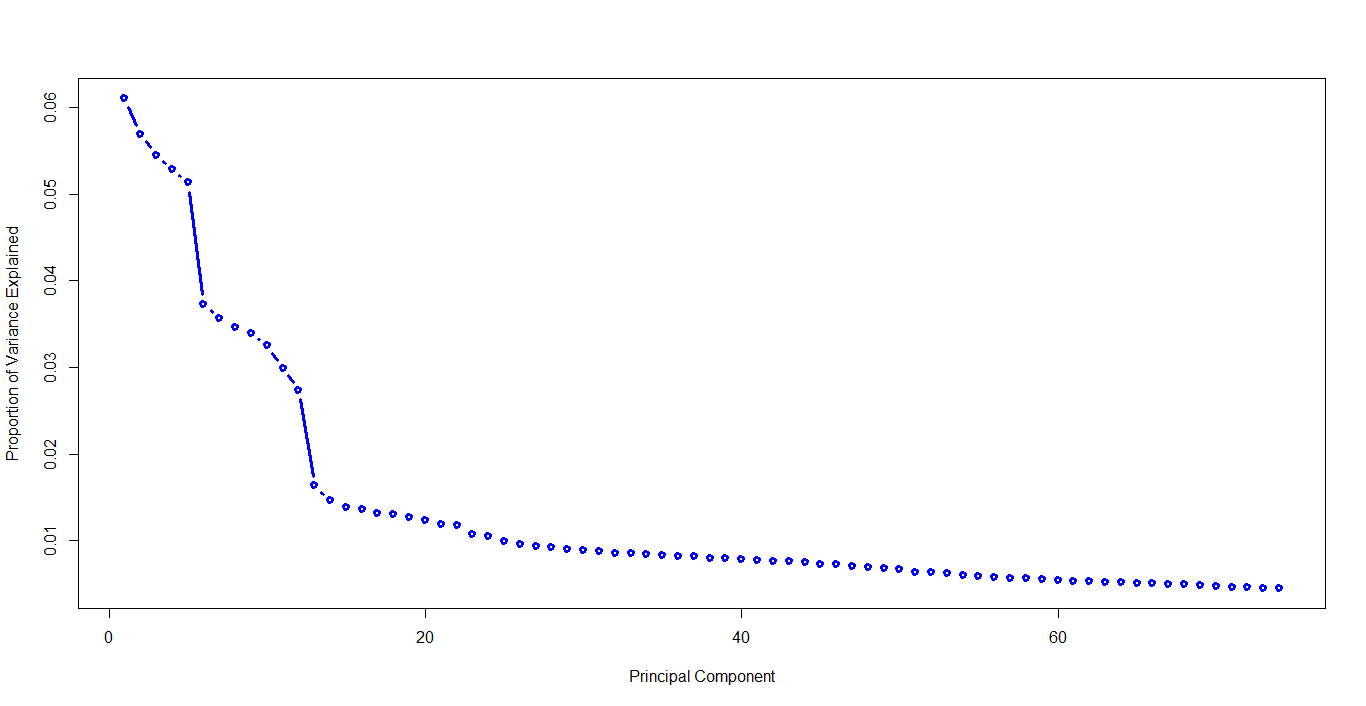
After applying PCA we have got the plot below showing the components.





Now the question is how to select the number of principal components to cover the maximum variance in the data. We can find the answer in the plot below.

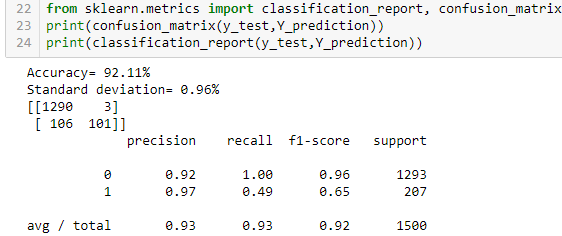


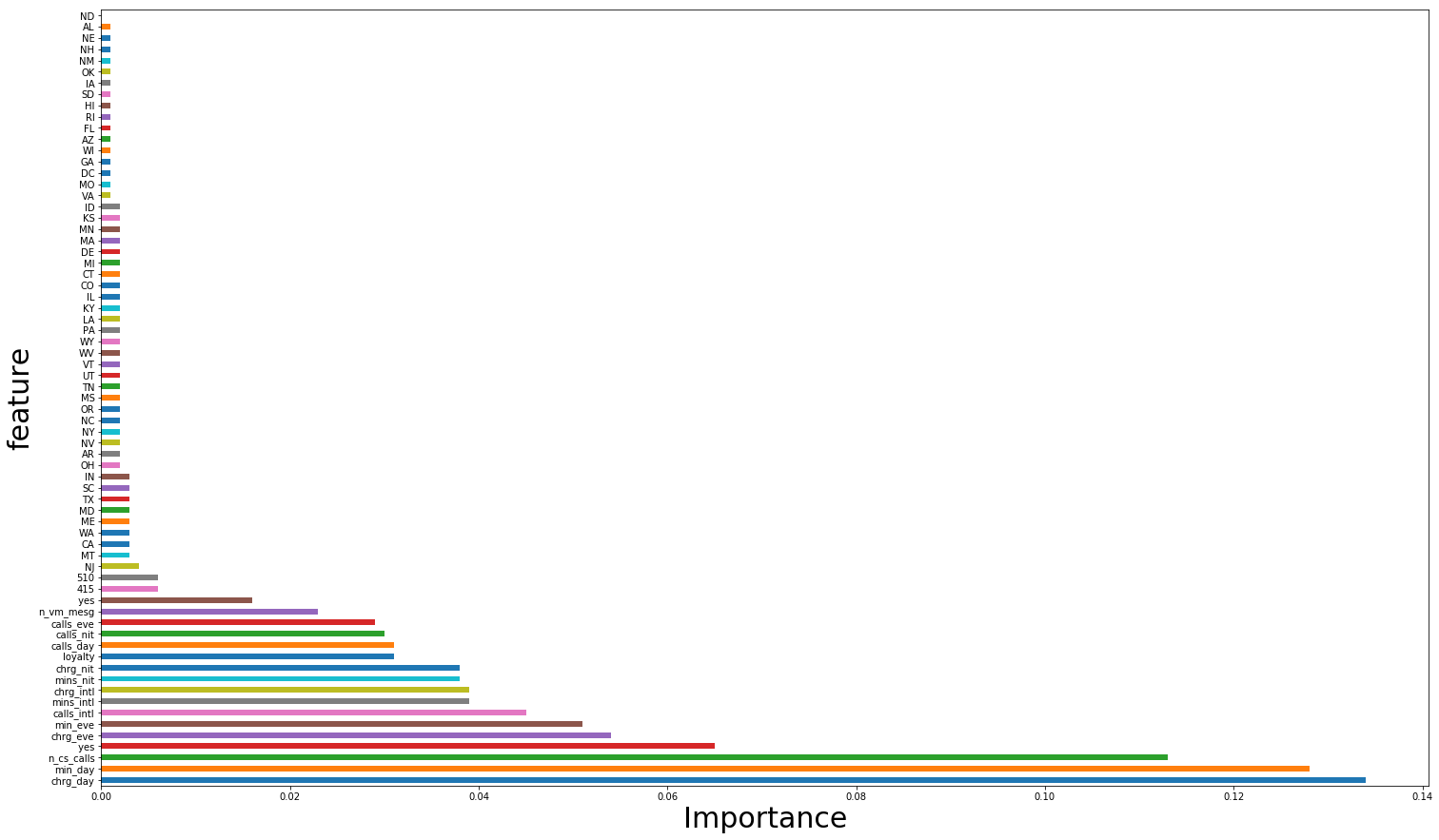


The above images show that the first 20 components contain a greater proportion of the explained variance. Hence, we will be using these components for model building.

**Method2**: Using python we have used Random forest as the feature selection technique with the feature importance.

A default random-forest has been fitted to the preprocessed data with all the features.

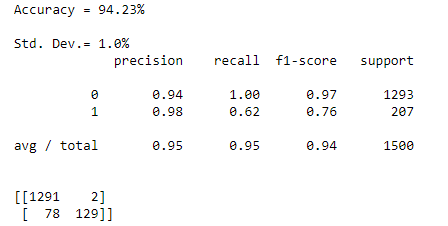




Above shown is the feature importance plot from the Random Forest. We found that the encoded features from the state column have very low importance. So, we will not be using these features in the further modelling. Or we can also drop the state variable for modelling purpose.

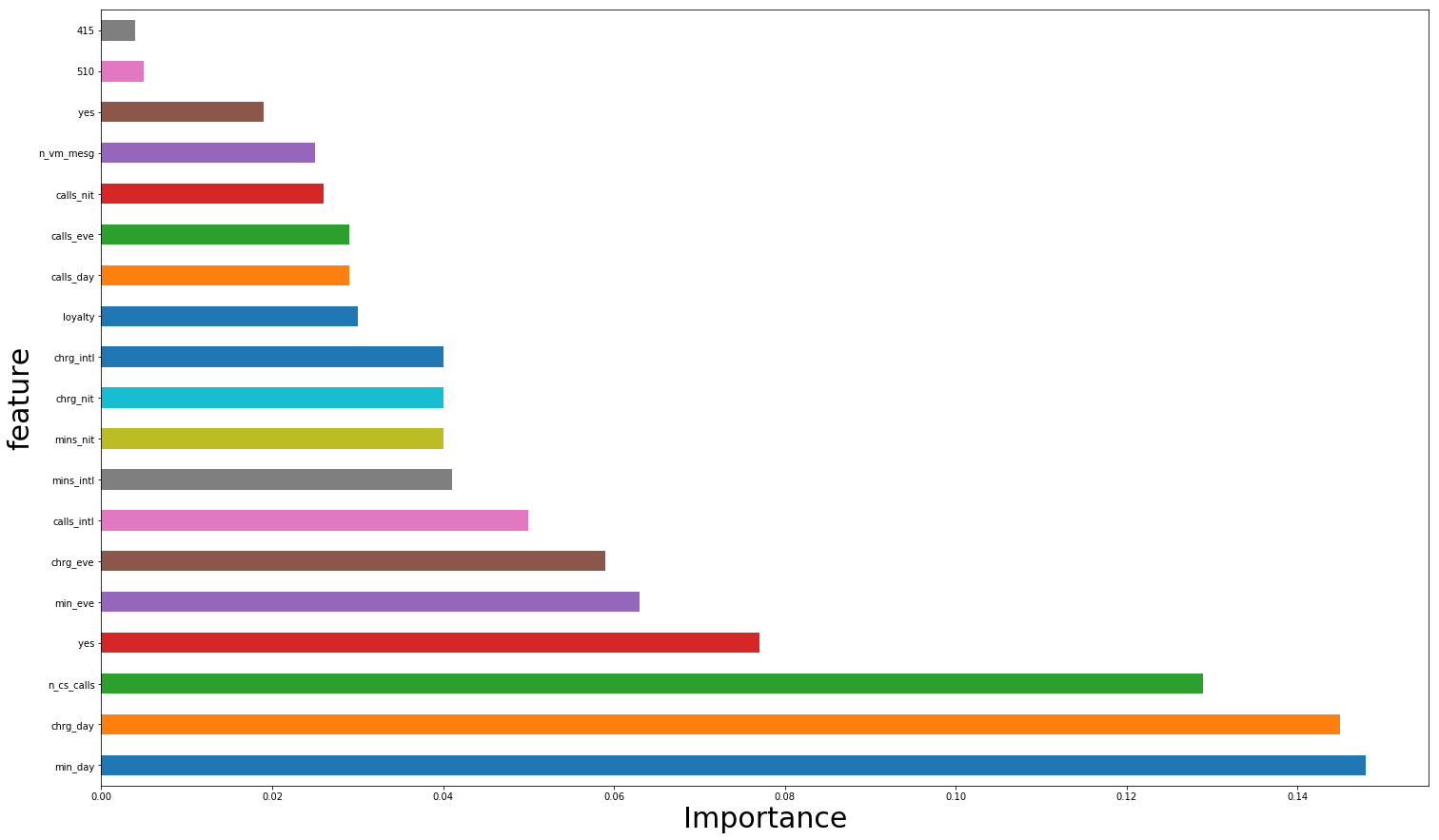
* 1. **Model Building**

After selecting the required features, Random classifier is fitted again.

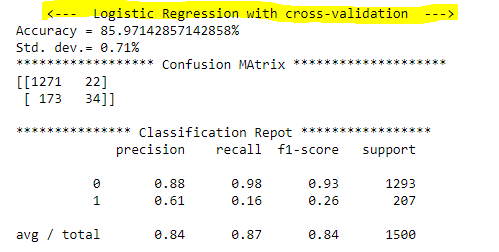
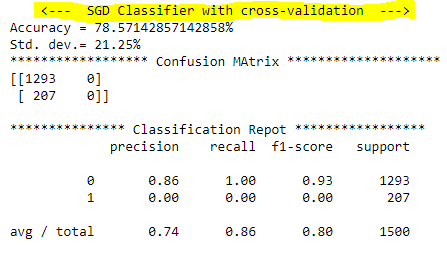


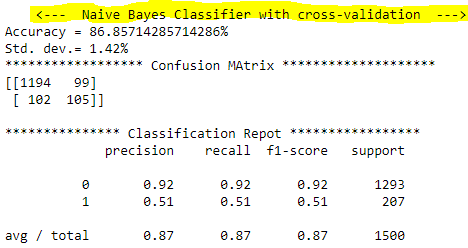
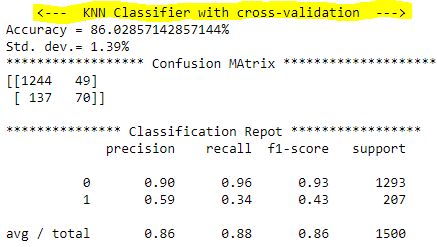
The accuracy of the model improved by more than 1%(92.11% to 94.23%). If we check the f1-score, it has also improved(jumped from 92% to 95%). There is a slight increase in the standard deviation fom 0.96% to 1.0%.

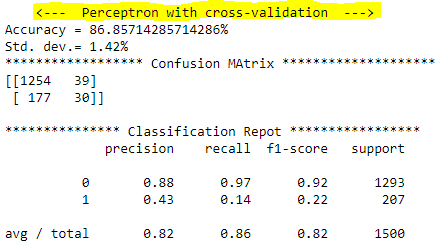
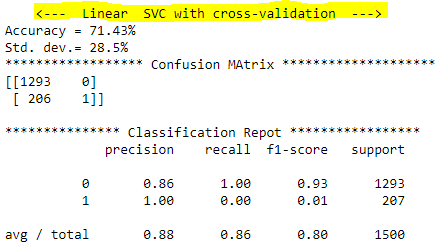
We have got the model improved on almost all the required parameters. So I am overall happy so far with this model.

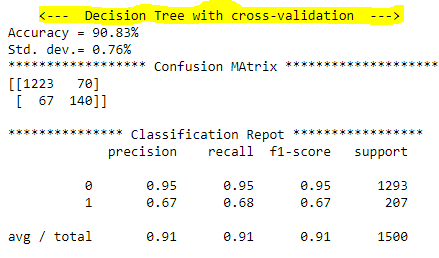


Now we will fit and check the other models for the best fit. We have fitted: Logistic Regression, Stochastic Gradient Descent, KNN, Naïve Baye’s, Linear SVC,

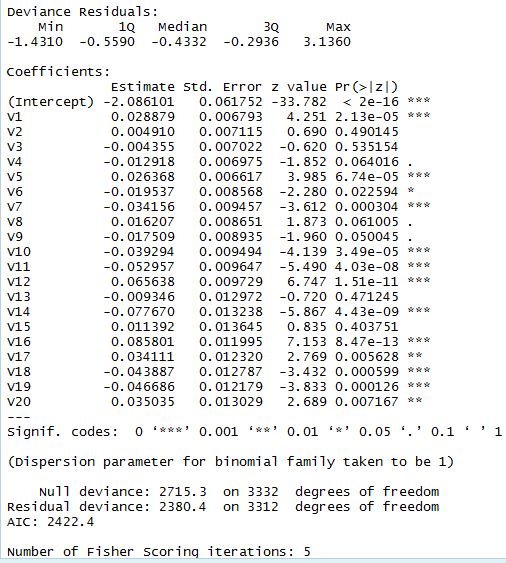
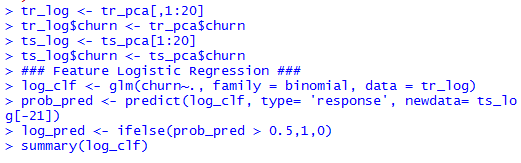


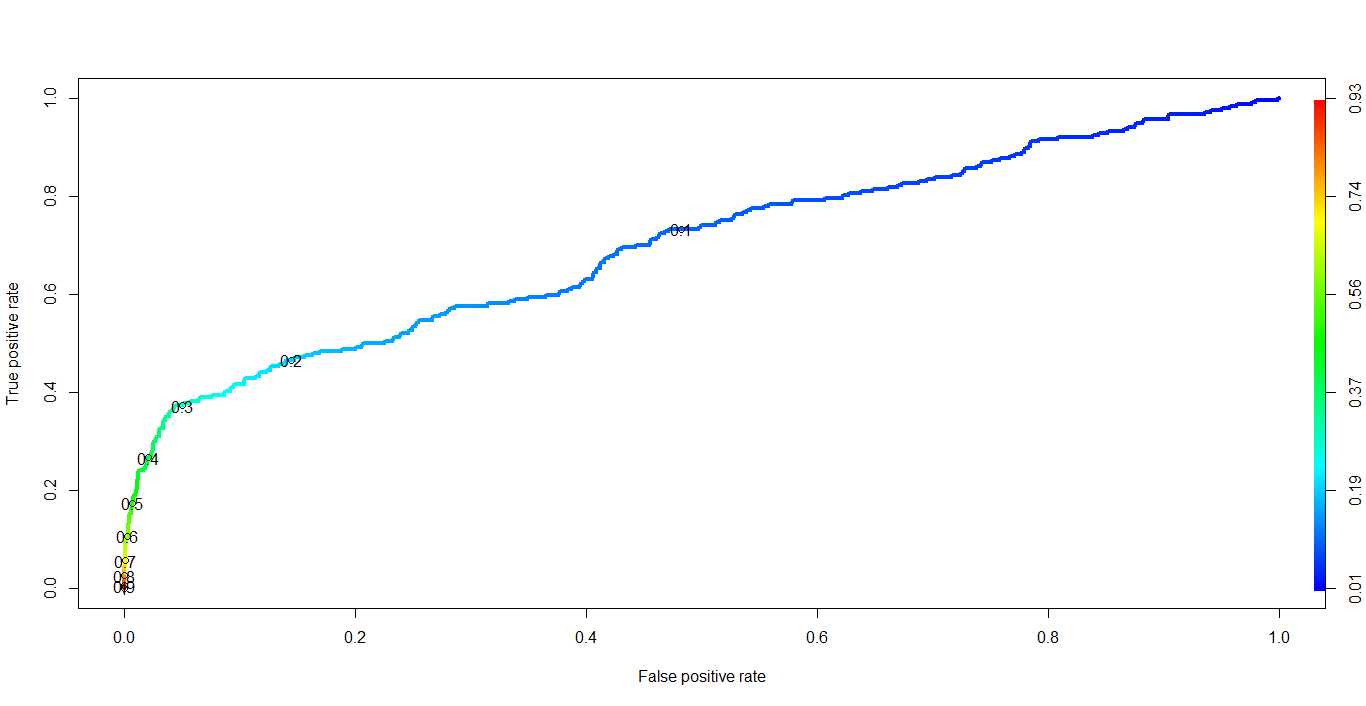
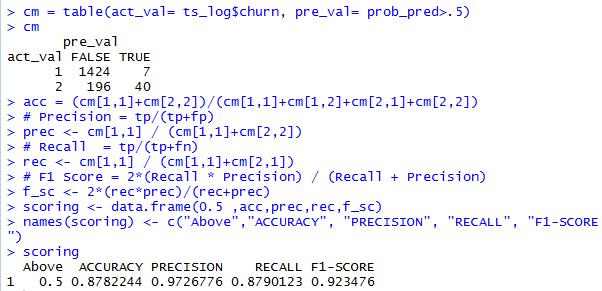
 



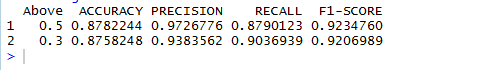
|  |  |
| --- | --- |
|  | **We can see that random forest has delivered the best accuracy with highest f1-score.**  **Also, in terms of standard deviation Random Forest has outperformed most of the models.**  **Now, we can say that RF best serves the purpose in this use-case.** |

**\*\*\*\*\*\*\*Further interpretation of Logistic Regression using R\*\*\*\*\*\*\***



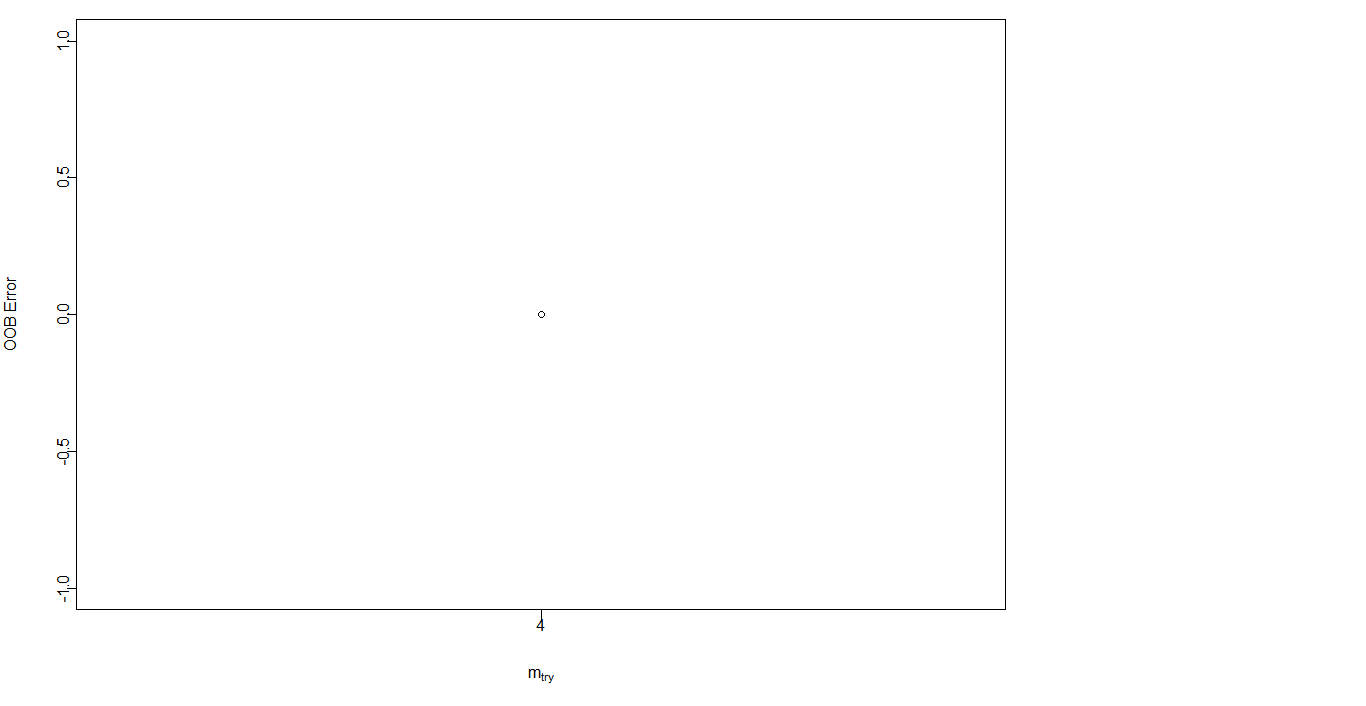
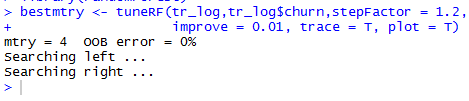


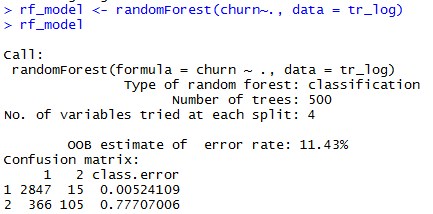
**Checked with .2, .3 and .5**

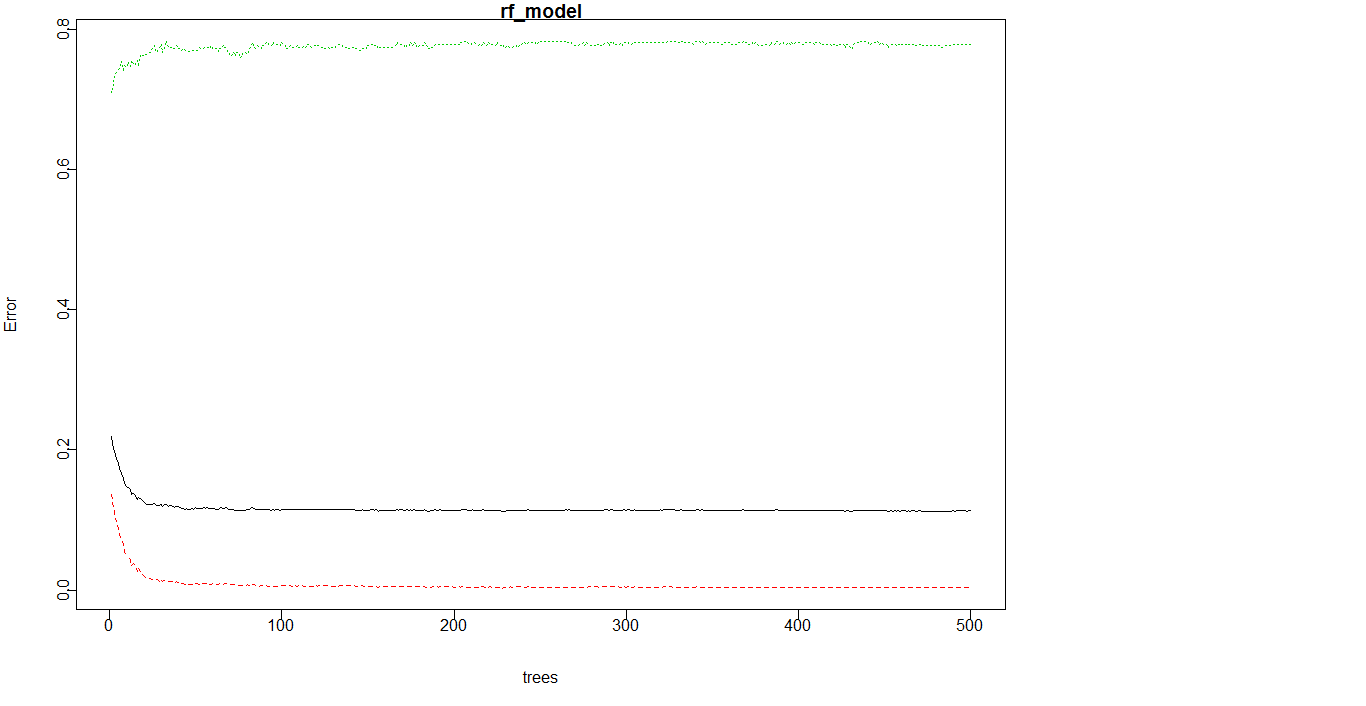


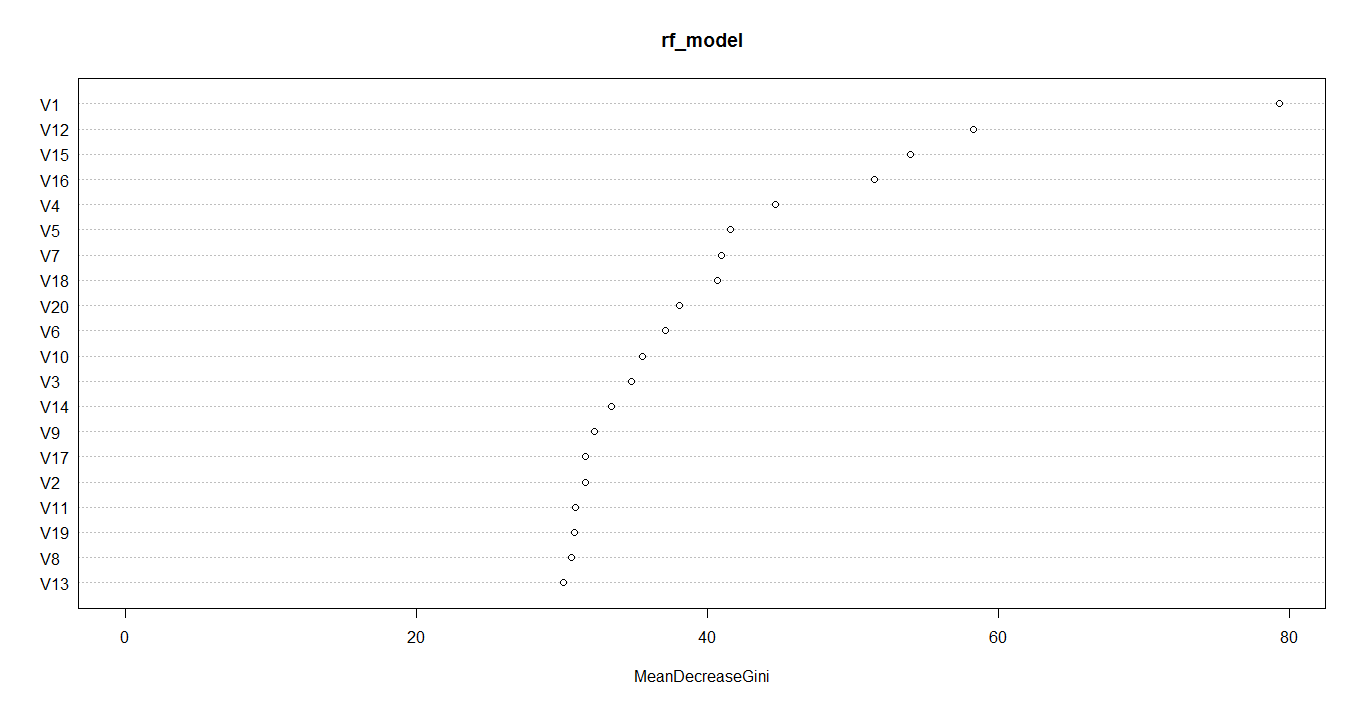
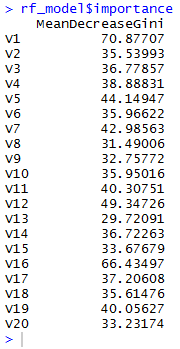
**\*\*** Getting to know the better performance, the random-forest model has also been fitted in R on top of PCA.

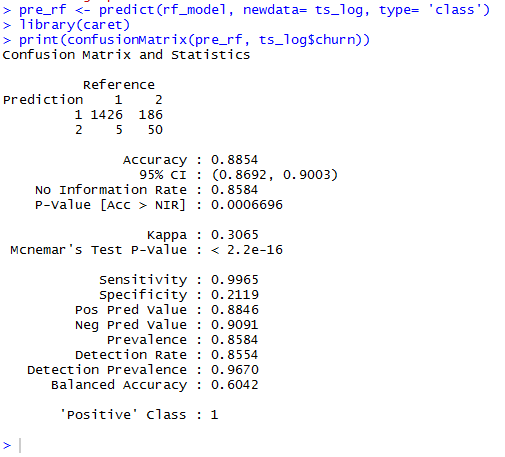
Got the optimal value of mtry with tuneRF function.

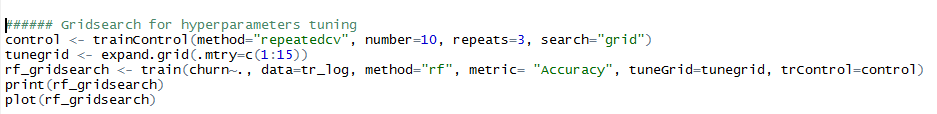












* 1. **Neural Network (MLP Classification) ~python**

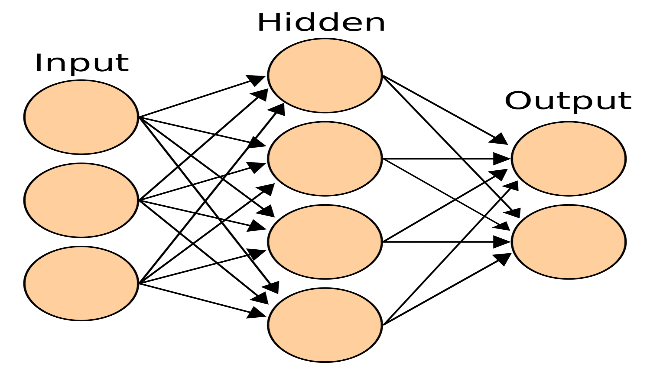
 Artificial neural networks are computation systems that intend to imitate human learning capabilities via a complex architecture that resembles the human nervous system.

The simplest neural network consists of only one neuron and is called a [perceptron](https://en.wikipedia.org/wiki/Perceptron), as shown in the figure below:



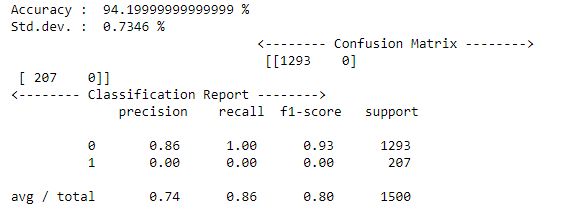
A single layer perceptron can solve simple problems where data is linearly separable in to 'n' dimensions, where 'n' is the number of features in the dataset. However, in case of non-linearly separable data, the accuracy of single layer perceptron decreases significantly. Multilayer perceptrons, on the other hand, can work efficiently with non-linearly separable data.

An artificial neural network (MLP) has an input layer, one or more hidden layers, and an output layer. This is shown in the image below:

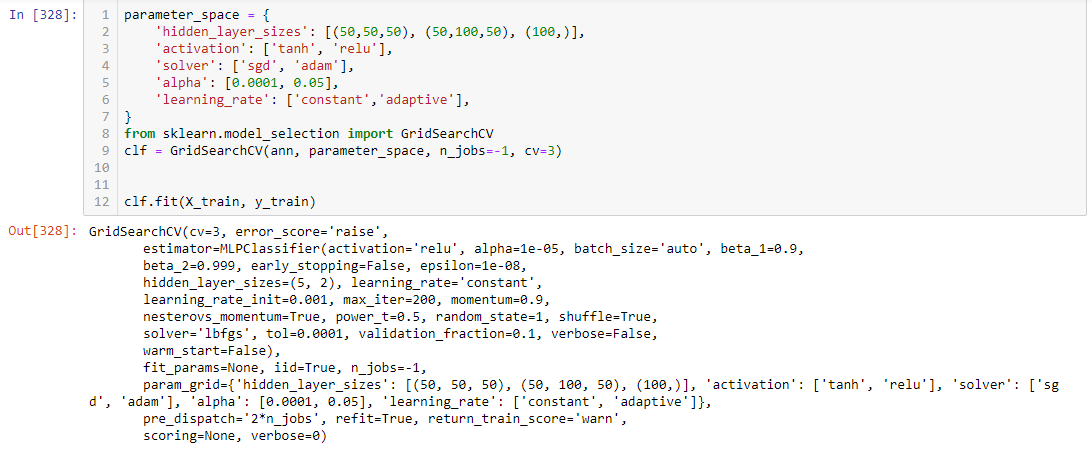


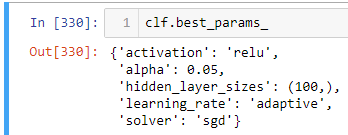
A neural network executes in two phases: Feed-Forward and Back Propagation.

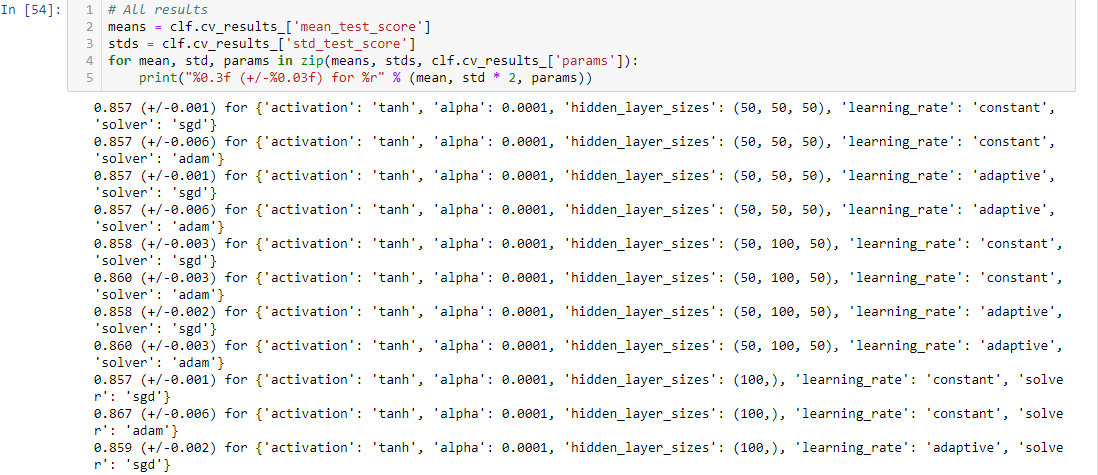
Below is the result of the neural network for classification(MLP). However we can see that for this case the random forest has outperformed the neural network.

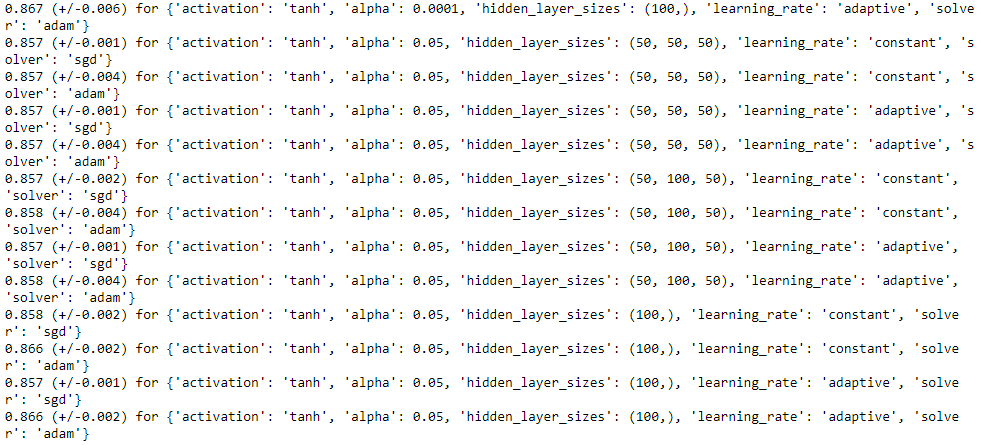


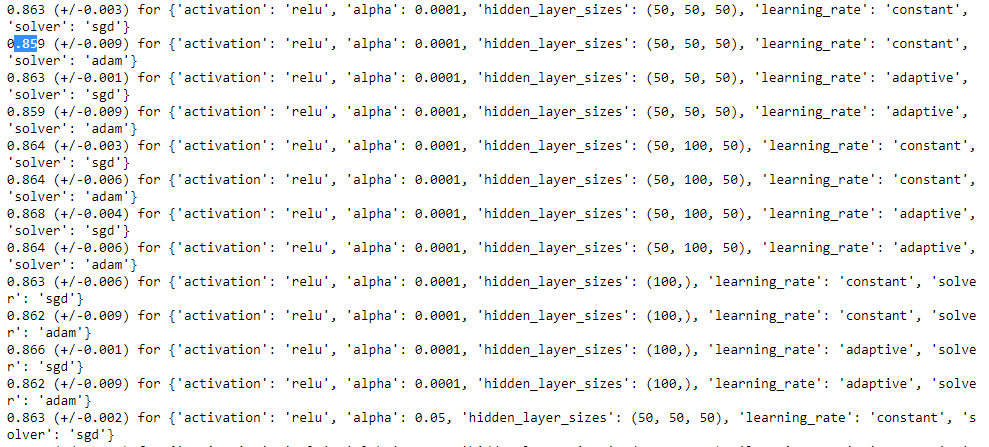
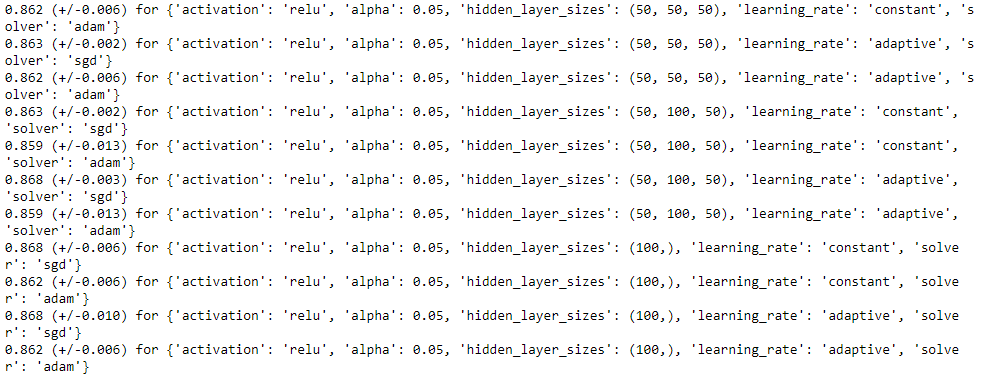
* 1. **Hyperparameter Tuning (MLP with Gridsearch)**

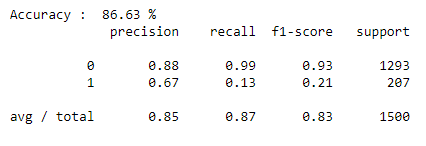
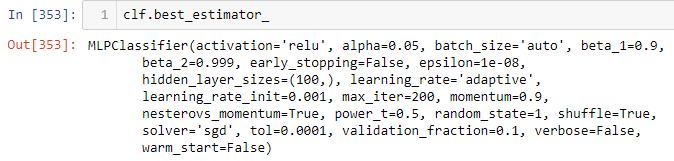


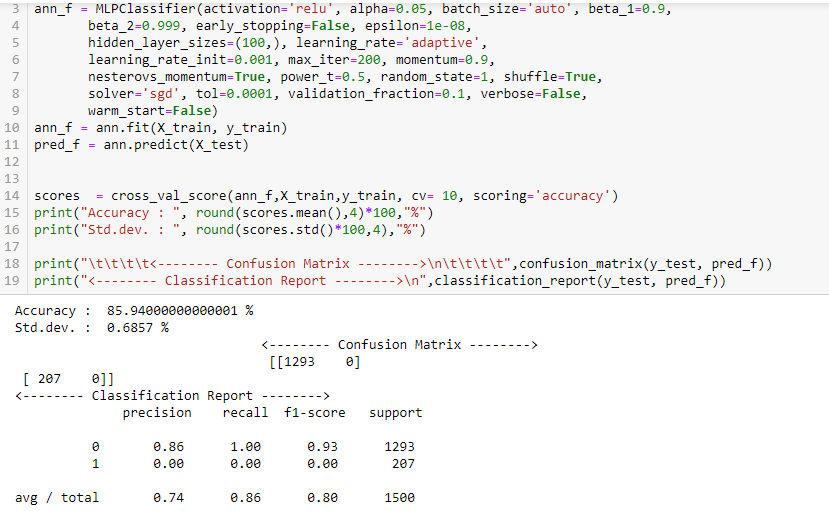








* 1. **Model Selection**

For this use case Random forest has outperformed all other models, even the neural network as well. So we will be selecting RANDOM FOREST model.

* 1. **CONCLUSION**



Out of various models developed for Prediction of Customers going to be churned, Random Forest suits best for our business case. RF is able to correctly predict 62%(sensitivity: 61.8357%) of customers to be churned as true at the same keeping the false out ratio of 0.15%. That means apart from predicting 62% correctly it is also predicting incorrectly for 0.15% customers as churned (which is not true).

But it is providing the best results of all the models tested with almost amazing results.

