Mini Project – Telecom Customer Churn Prediction Assessment

Model Report

Table of Contents

1. [Project Objective 3](#_bookmark0)
2. [Exploratory Data Analysis – Step by step approach 3](#_bookmark2)
   1. [Environment Set up and Data Import 3](#_bookmark3)
   2. [Variable Identification](#_bookmark7) 3
      1. [Variable Identification – Inferences](#_bookmark8) 3
   3. [Univariate Analysis](#_bookmark9) 4-5
   4. Bivariate Analysis……………………………………………………………………………………………………………….6
   5. [Missing Value Identification](#_bookmark11) 6
   6. [Outlier Identification](#_bookmark12) 6-7

3 Multicollinearity……………………………………………………………………………………………………………………. 8

4 Logistic Regression…..……………………………………………………………………………………………………… 9-11

5 K-Nearest Neighbour……………………………………………………………………………………………………….12-13

6 Naïve Bayes ……………………….…………………………………………………………………………………………. 13-14

7 Model Perfromance ………….…………………………………………………………………………………………. 15-17

8 [Business Insights and Recommendations……………………………………………………………………..](#_bookmark14) 18

**Project Objectives**

The main objective of this project is to build a model to identify which group of customers will cancel the post-paid services in future. This is further divided into small objectives which are as follows:

1. To perform exploratory data analysis.
2. To check the presence of multicollinearity and deal with it.
3. To build logistic regression model, k – nearest neighbour model and Naïve Bayes model.
4. Selecting the best model by comparing different model performance measures like accuracy, KS stat, gini coefficient, concordance ratio, etc.

**Environment set up and Data Import**

Setting up of the working directory help in accessing the dataset easily. Different packages like “tidyverse”, “car”, “InformationValue”, “bloom”, “caTools”, “caret”, “dplyr”, “gini”, “ROCR”, “gridExtra” and “corrplot” are installed to make the analysis process slightly easier.

The data file is in “.xlsx” format. To import the dataset, read\_excel function(available in readxl package) is used. The dataset consists of 3333 rows and 11 columns.

**Variable Identification**

**“head”** function is used to show top six rows of the dataset.

**“tail”** function is used to show last six rows of the dataset.

**“str”** function is used to identify type of the variables. In this dataset, all the variables are numeric. Initially all the variables are numerical. Some of the variables like Churn, DataPlan and ContractRenewal are converted into factors using as.factor() function.

**“summary”** function is use to show the descriptive statistics of all the variables

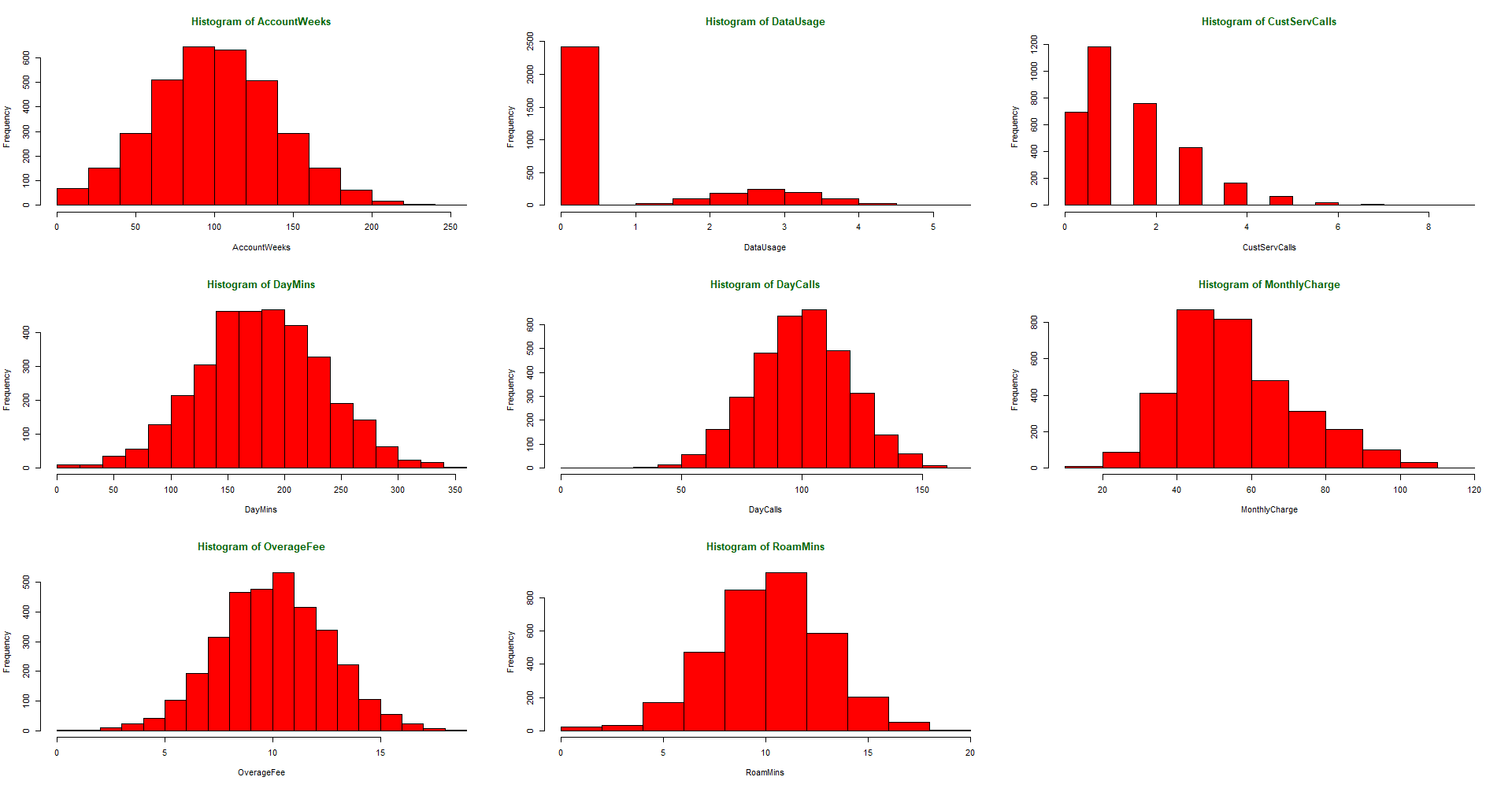
**Exploratory Data Analysis**

The historical data reflects that around 85.51% customers retained the service and only 14.49% cancelled the post-paid service.

**Univariate Analysis**

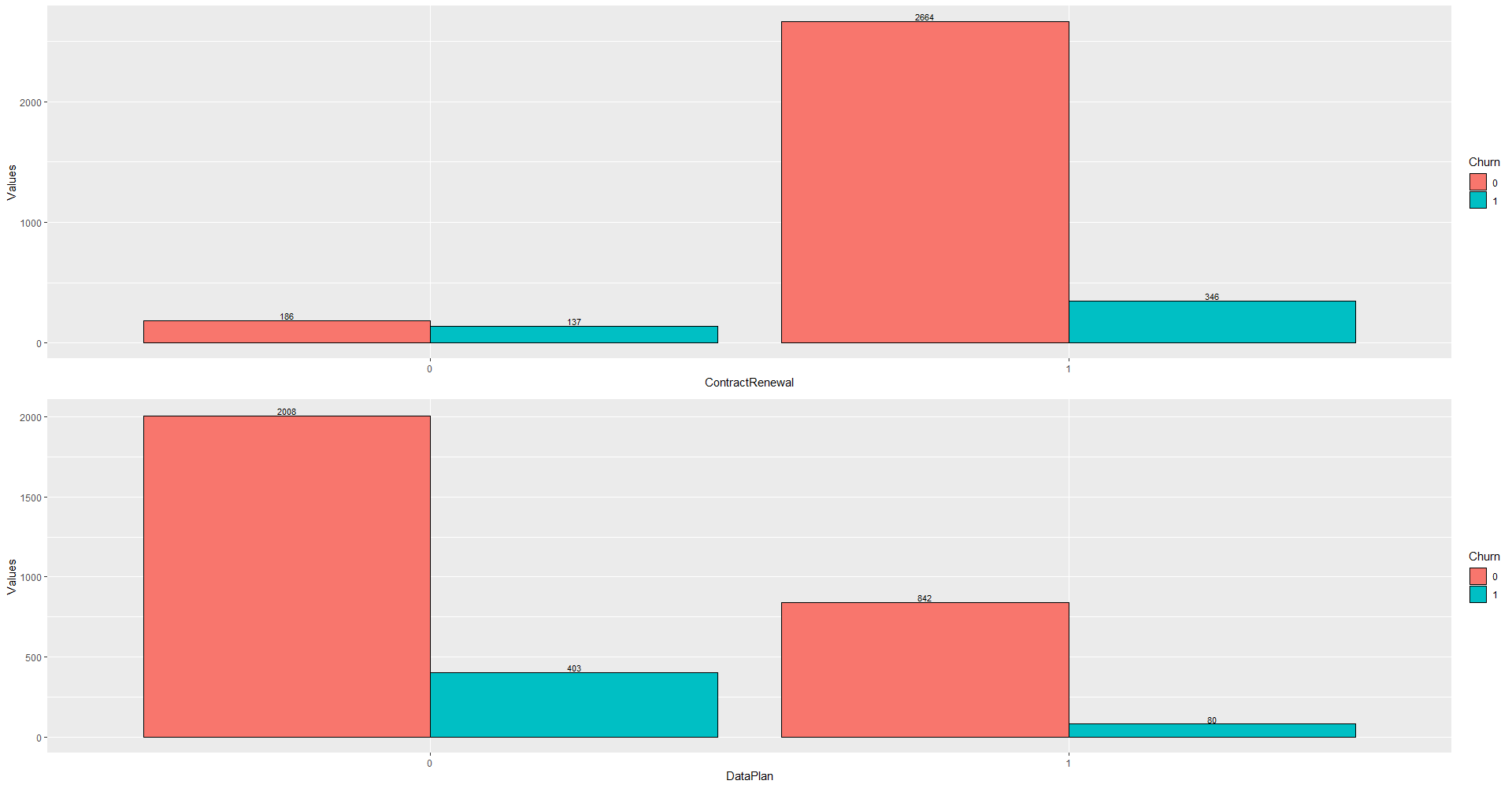
1. **Histograms**

The histogram is used to signify the normality and skewness of the data. It also signifies central location, shape and size of the data.



From above figure we can infer that:

1. Variables like OverageFee, RoamMins, DayCalls, DayMins and AccoutnWeeks are normally distributed.
2. Variables like CustServCalls and MonthlyCharge shows sign have right skewness.
3. **Barplots**

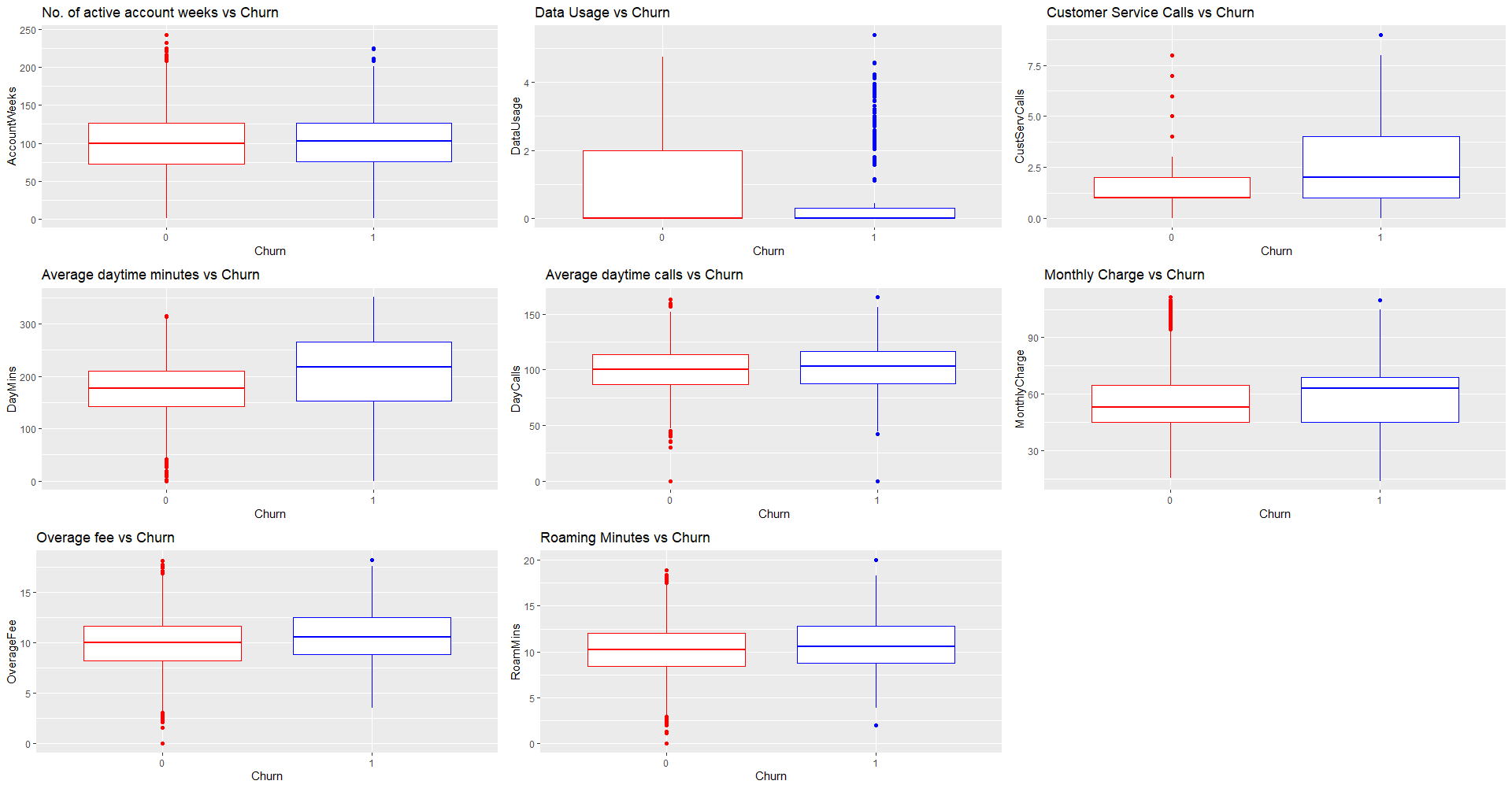


From above figure we can infer that:

1. Out of total customers who do not renewed the post-paid services contract, 137 customers cancelled the service and 186 customers retained the service.
2. Out of total customers who have renewed the post-paid services contract, only 346 customers cancelled the service and 2664 customers retained the service.
3. Out of total customers who have taken DataPan, only 80 customers cancelled the service and 642 customers retained the service.
4. Out of total customers who do not have DataPan, only 403 customers cancelled the service and 2008 customers retained the service.

**Bivariate Analysis**

1. **Boxplot**

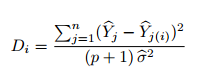


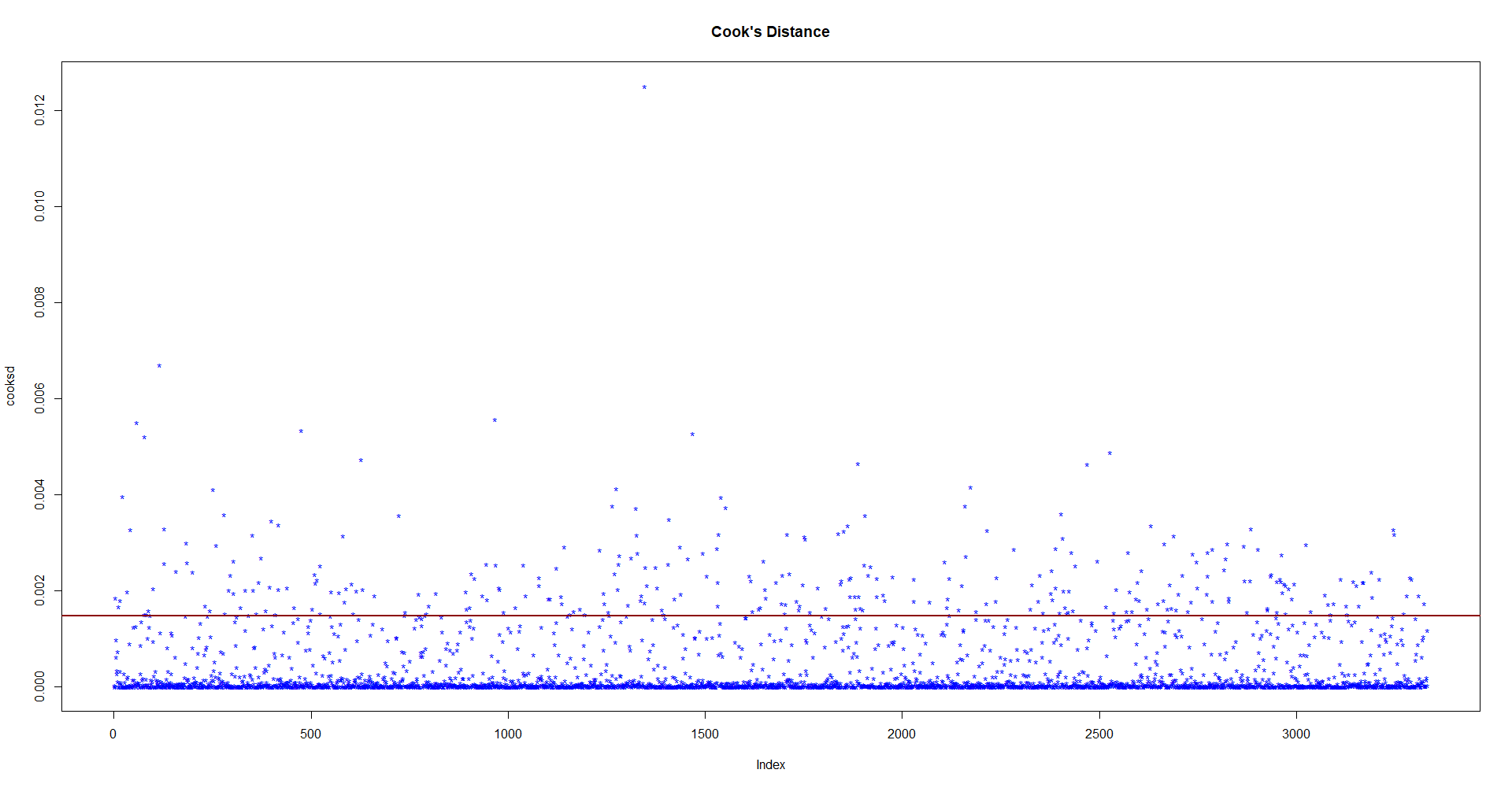
It can be easily seen that RoamMins have highest outliers followed by MonthlyCharge, DayMins and so on.

**Outlier detection and missing values**

There are no missing values in the given dataset.

To check the presence of the outliers in the dataset, cook’s distance is used. Cook’s distance shows how removing a particular observation from the data can affect the predicted values.

The general rule of thumb suggests than any observation above the threshold i.e. 4\* mean of cook’s coefficient D  is considered as an influential value or outlier.



It is clearly seen that all the values above red line are outliers and need to be treated accordingly.

**Outliers Treatment**

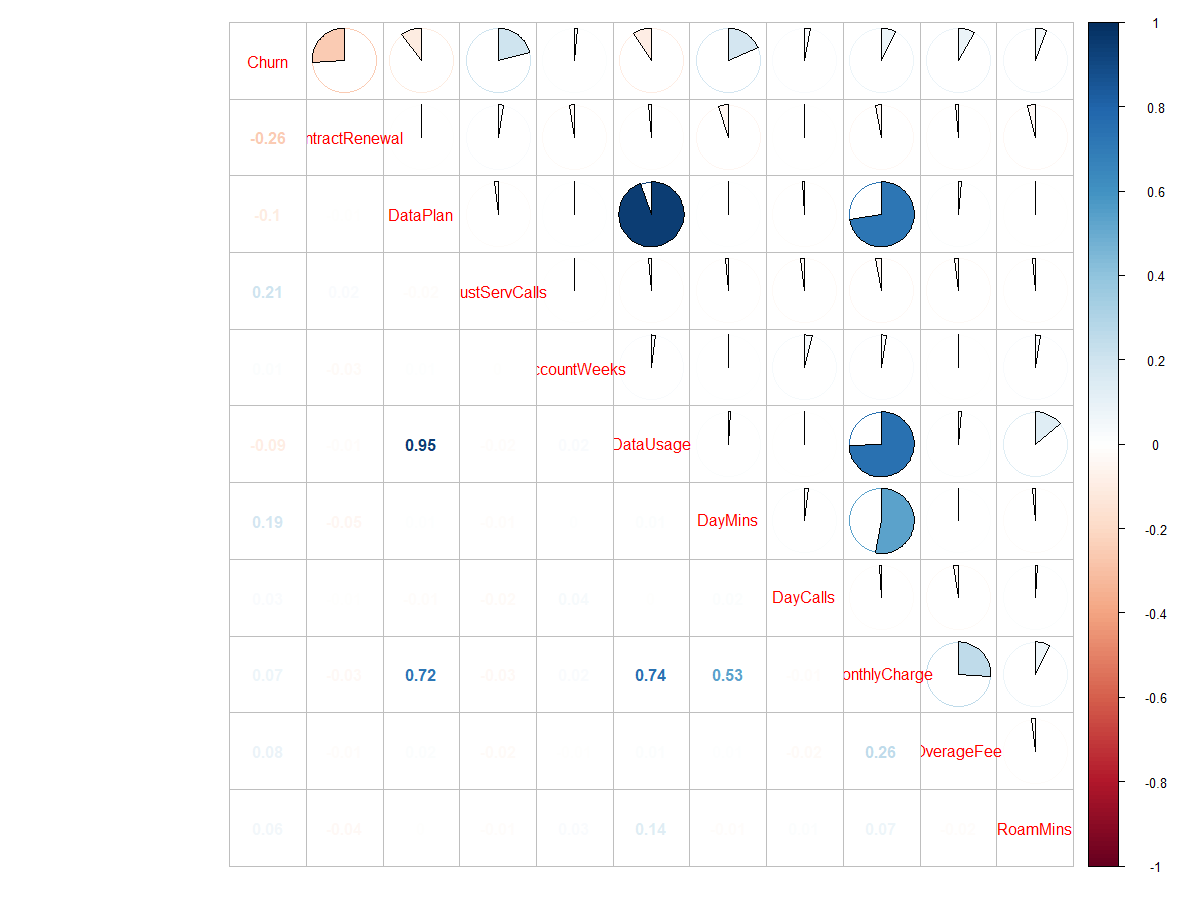
Mean imputation method is used to treat the outliers. The table below shows variables having outliers, how many outliers are present and mean value used to replace the outliers.

|  |  |  |
| --- | --- | --- |
| **Variables** | **Number of outliers** | **Mean Imputation Value** |
| AccountWeeks | 18 | 100 |
| DayMins | 25 | 180 |
| DayCalls | 23 | 101 |
| MonthlyCharge | 34 | 56 |
| OverageFee | 24 | 10 |
| RoamMins | 46 | 10 |
| DataUsage | 11 | 1 |

**Multicollinearity**

The problem of multicollinearity exists when the independent variables are highly correlated with each other. Variance Inflation Factor (VIF) and Tolerance level are the key factors to analyse the level of the multicollinearity among independent variables. In VIF, any value closer to 1 signifies low level of correlation and any value above 10 signifies high level of correlation.

As per the given dataset, Data Usage and Monthly Charge has VIF value greater than 5. Thus, they depict problem of multicollinearity and need to be treated.



It can be seen in the above figure that:

1. There exists high correlation (0.95) between Data usage and Data Plan.
2. Data plan and Monthly Charges also shows signs of high correlation (0.72).
3. Monthly Charges and DayMins shows signs of high correlation (0.53).
4. There exists high correlation (0.74) between Data Usage and Monthly Charges.

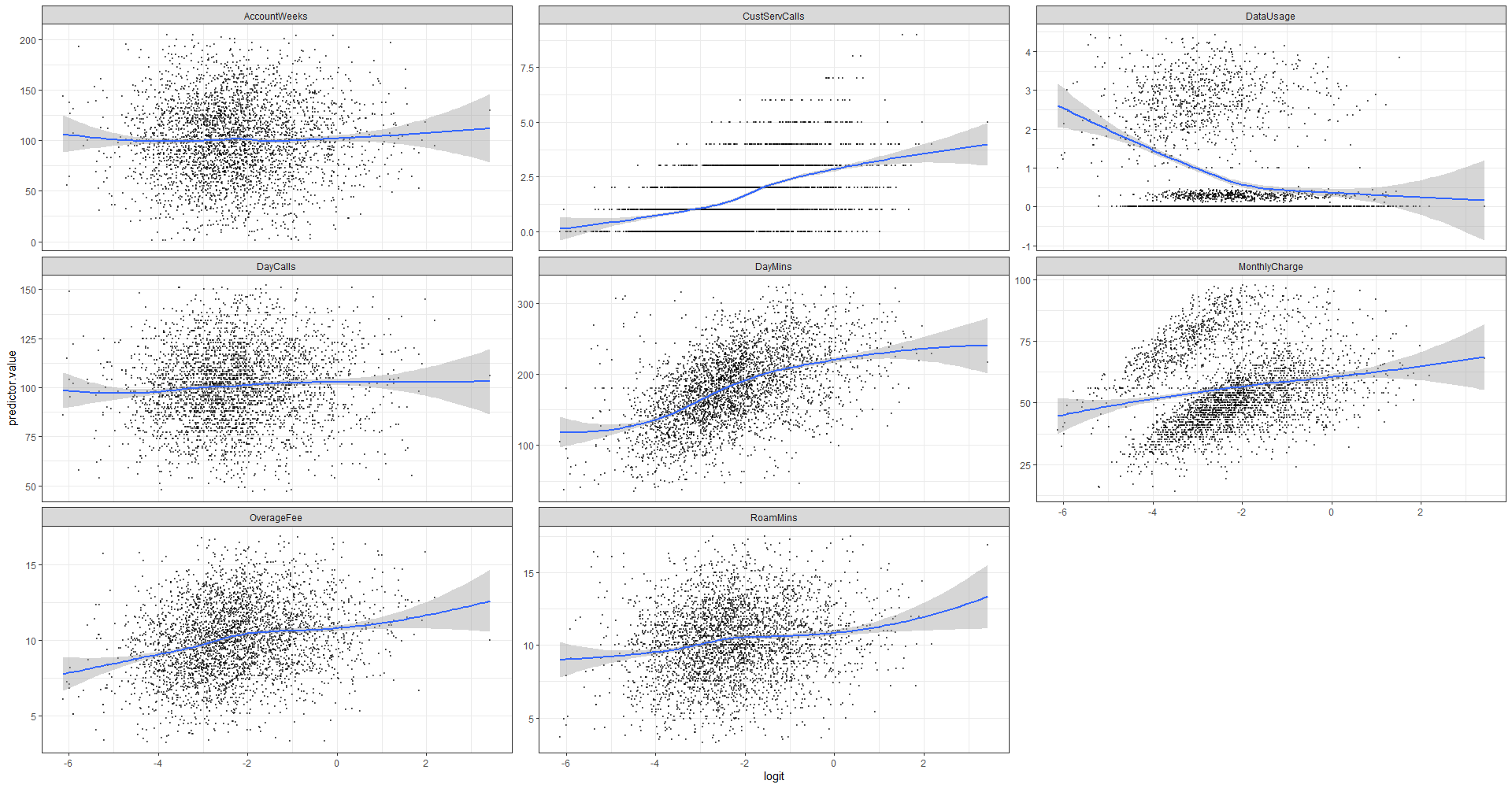
**Logistic Regression**

Logistic Regression is a type of generalized linear model which is used to solve binary classification problem. Logistic regression uses maximum likelihood method to obtain a best fit line. It uses logit function to estimate the coefficients. The function is given by

log(p/1-p) = beta0 + beta1\*X1 +beta2\*X2 + ……… + beta(n)\* Xn

**Assumptions:**

1. **Linear Relationship**: There should exist a linear relationship between log(odds) and regressors. It can be seen in the figure that most of the variables depicts this linear relationship.



1. **Outliers Treatment:** The dataset should be free from outliers. Using cook’s distance, it was found that there is presence of outliers in the model and mean imputation method is used to remove these outliers.
2. **No Multicollinearity:** The independent variables should not be highly correlated with each other. Using VIF, it was found that variables like DataUsage and MonthlyCharge are highly correlated variables. Thus, they were dropped from the model.

**Model Building**

The data has been split randomly into training set and test set with a split ratio of 2/3. K-fold validation technique was used to build the model on training data set and the model which best fits the training set (i.e. model with highest accuracy) was selected. Then the best fit model is used to predict the test dataset.

**Model 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Estimate | z-value | p-value | Status |
| Intercept | -2.25910 | -26.666 | < 2e-16 | Significant |
| ContractRenewal | -0.57744 | -11.092 | < 2e-16 | Significant |
| DataPlan | -0.43683 | -5.457 | 4.83e-08 | Significant |
| CustServCalls | 0.67521 | 10.866 | < 2e-16 | Significant |
| AccountWeeks | -0.03394 | -0.500 | 0.61692 | Not -Significant |
| DayMins | 0.63410 | 9.023 | < 2e-16 | Significant |
| DayCalls | 0.15874 | 2.383 | 0.01716 | Significant |
| OverageFee | 0.34363 | 4.957 | 7.15e-07 | Significant |
| RoamMins | 0.18964 | 2.763 | 0.00573 | Significant |

Churn = -2.26 – 0.58(ContractRenewal) – 0.44(DataPlan) + 0.68(CustServCalls) – 0.034(AccountWeeks) + 0.63(DayMins) + 0.16(DayCalls) + 0.34(OverageFee) + 0.19(RoamMins)

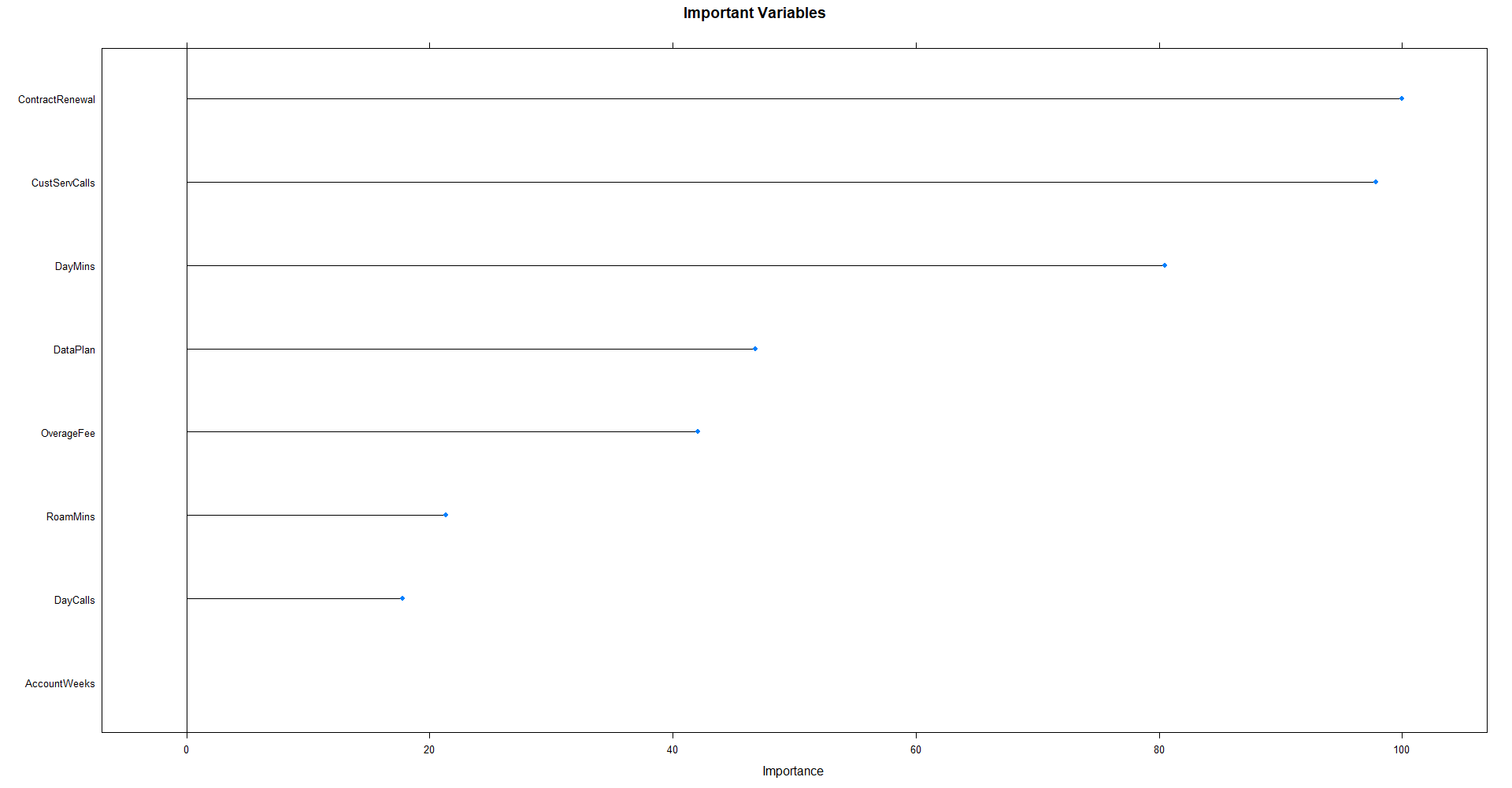
From the analysis, we can infer that variables like AccountWeeks doesn’t has any significant effect on the Churn. ContractRenewal has the highest effect on the Churn variable followed by CustServCalls, DayMins, DataPlan, OverageFee, RoamMins and DayCalls.

**Model 2: Without AccountWeeks**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Estimate | z-value | p-value | Status |
| Intercept | -2.25882 | -26.664 | < 2e-16 | Significant |
| ContractRenewal | -0.57532 | -11.093 | < 2e-16 | Significant |
| DataPlan | -0.43793 | -5.471 | 4.46e-08 | Significant |
| CustServCalls | 0.67404 | 10.858 | < 2e-16 | Significant |
| DayMins | 0.63445 | 9.027 | < 2e-16 | Significant |
| DayCalls | 0.15829 | 2.377 | 0.01746 | Significant |
| OverageFee | 0.34411 | 4.964 | 6.89e-07 | Significant |
| RoamMins | 0.19041 | 2.774 | 0.00553 | Significant |

After removing the insignificant variable (AccountWeeks), we get our final model which consists of ContractRenewal, DataPlan, CustSevCalls, DayMins, DayCalls, OverageFee and RoamMins as independent variable and Churn as dependent variable.

**Churn = -2.26 – 0.58(ContractRenewal) – 0.44(DataPlan) + 0.68(CustServCalls) + 0.63(DayMins) + 0.16(DayCalls) + 0.34(OverageFee) + 0.19(RoamMins)**



From the above figure we can see that ContractRenewal is the most important variable followed by CustServCalls, DayMins and so on. ContractRenewal and Dataplan have a negative relationship with dependent variable.

**K- Nearest Neighbour**

KNN algorithm is a supervised machine learning model which is used for both regression and classification problems. K- Nearest Neighbour is estimated using the concept of the shortest distance between the observation. There are majorly 3 ways to calculate the distance: Euclidian Distance, Manhattan Distance and Minkowski Distance.

Feature Scaling

Since we are calculating the Euclidian distance between different variables, thus standardization or scaling becomes a necessary part in data pre-processing. The formula used to standardize the variables is given by

**Standardized value = (Old value – Mean(n)) / Standard Deviation(n)**

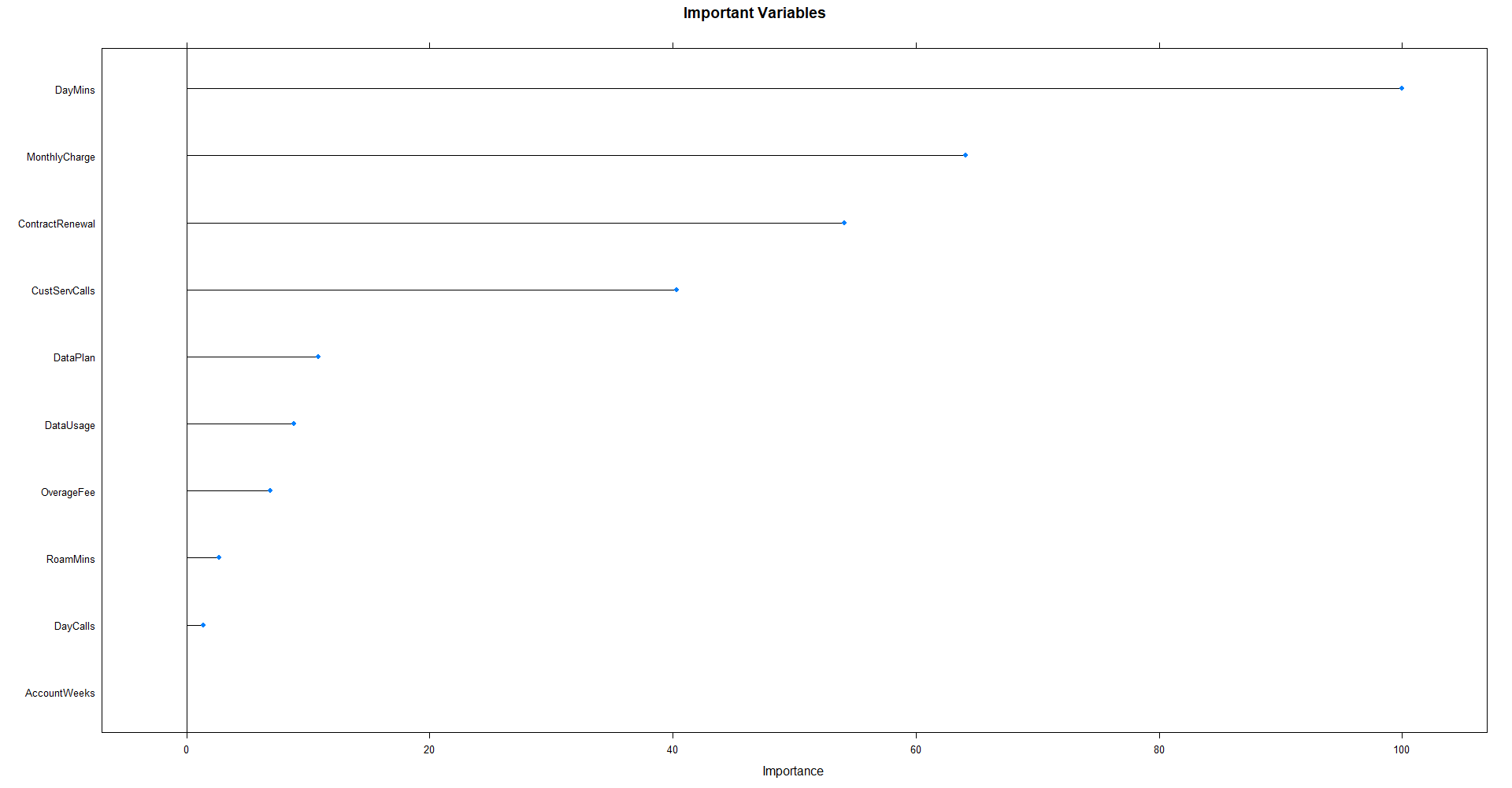
Where n is total number of observations

**Model Building**

The data has been split randomly into training set and test set with a split ratio of 2/3. K-fold validation technique was used to build the model on training data set and the model which best fits the training set (i.e. model with highest accuracy) was selected. Then the best fit model is used to predict the test dataset.

**Selection of K-Value**

The optimal number of K nearest neighbour is decided using a hyperparameter tuning method called grid search. Initially, a grid of 20 K-values is selected and “accuracy” metric is used to assess the performance of each model. It is found that accuracy is highest when the optimal value for K- nearest neighbour is 7.



It can be seen in the above figure that the most important variable influencing whether a customer will cancel the post-paid service or not is DayMins, followed by MonthlyCharge, ContractRenewal and so on. AccountWeeks is the least important variable in the model.

**Naïve Bayes**

It is based on the concept of Bayes Theorem but it follows a naïve assumption that all the variables are independent of each other. This method is used for both classification and regression problems. It is relatively faster as compared to other models.

**Model Building**

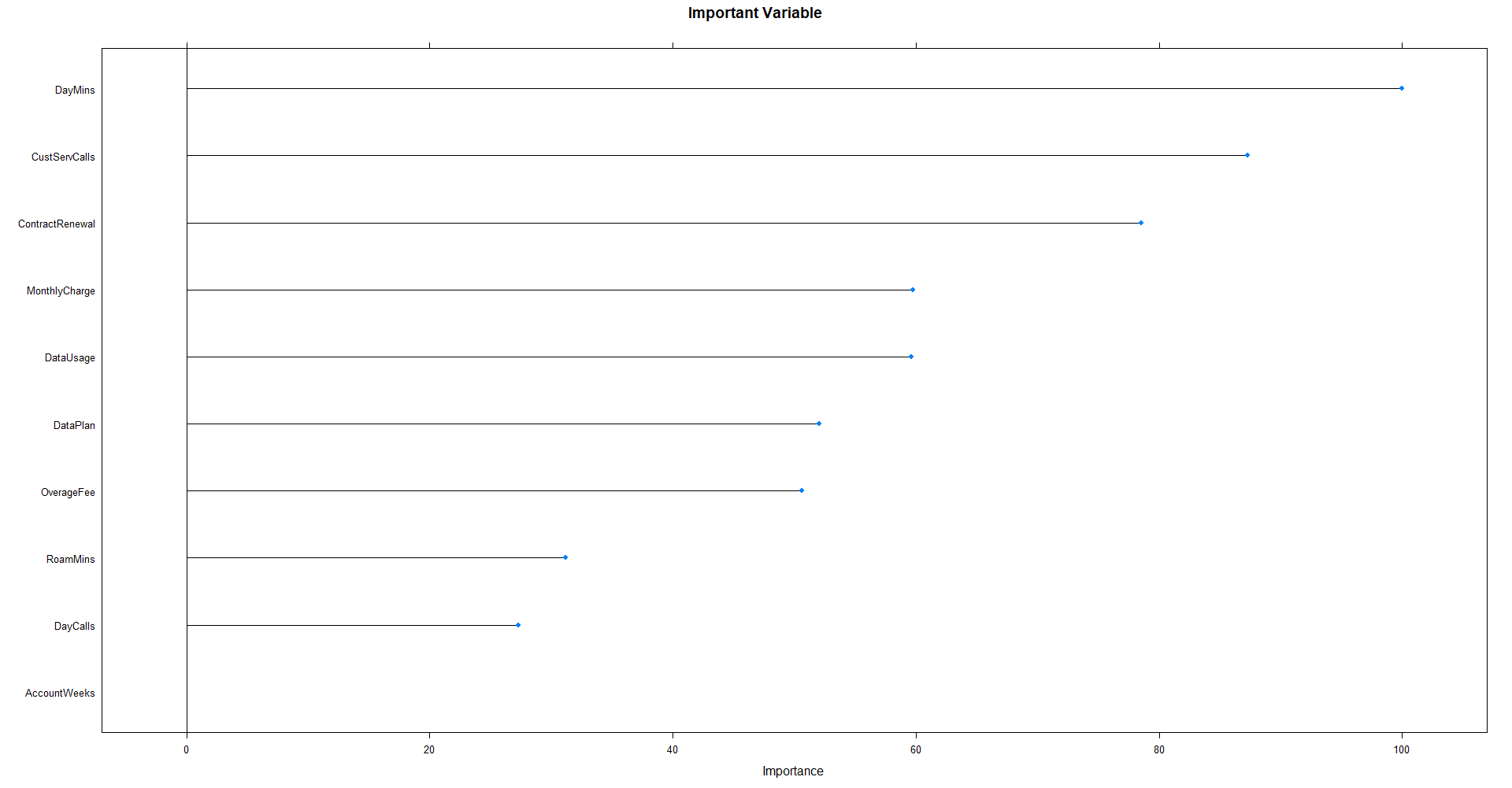
The data has been split randomly into training set and test set with a split ratio of 2/3. K-fold validation technique was used to build the model on training data set and the model which best fits the training set (i.e. model with highest accuracy) was selected. Then the best fit model is used to predict the test dataset.

**Hyperparameter Tuning**

**The grid search method is used to tune the hyperparameter which are given under:**

1. Laplace Smoothing – This method is generally used to smooth the categorical variable. It is used when conditional probability for a case is zero. Values taken for tuning are from 0 to 5.
2. Use of kernel: - This parameter allows us to use a kernel density estimate for a gaussian density estimate vs a continuous variable. It takes only two values – “True” and “False”
3. Bandwidth of kernel Density: - “adjust” parameter allows us to adjust the bandwidth of the kernel density. Values taken for tuning are 1 to 5

The final value used for each parameter is 0 for Laplace smoothing, 1 for adjust and TRUE for usekernel.



It can be seen in the above figure that the most important variable influencing whether a customer will cancel the post-paid service or not is DayMins, followed by CustServCalls ContractRenewal , MonthlyCharges and so on. AccountWeeks is the least important variable in the model.

**Model Performance**

Once the model is prepared on the training dataset, next step to is to measure the performance of the model on test dataset. Since the models predicts the test values in the form of probability, a threshold is selected to convert it to either 0 or 1. In these model, 0.5 is selected as threshold. Any probability less than 0.5 will be shifted to 0 and any probability above 0.5 will be shifted to 1.

Different key performance measures are used to check the efficiency and effectiveness of the model.

1. Accuracy: - It is the ratio of total number of correct predictions to total number of samples.

**Accuracy = (True Positive + False Negative)/ (True Positive + False Negative + True Negative + False Positive)**

1. Classification Error: - It is the ratio of total number of incorrect predictions to total number of samples.

**Classification Error = (False Positive + True Negative)/ (True Positive + False Negative + True Negative + False Positive)**

1. Sensitivity: - It is the proportion of customers who didn’t cancel the post-paid services got predicted correctly to total number of customers who didn’t cancel the services.

**Sensitivity = True Positive/ (True Positive + False Negative)**

1. Specificity: - It is the proportion of customers who cancelled the post-paid services got correctly predicted to total number of customers who cancelled the post-paid services.

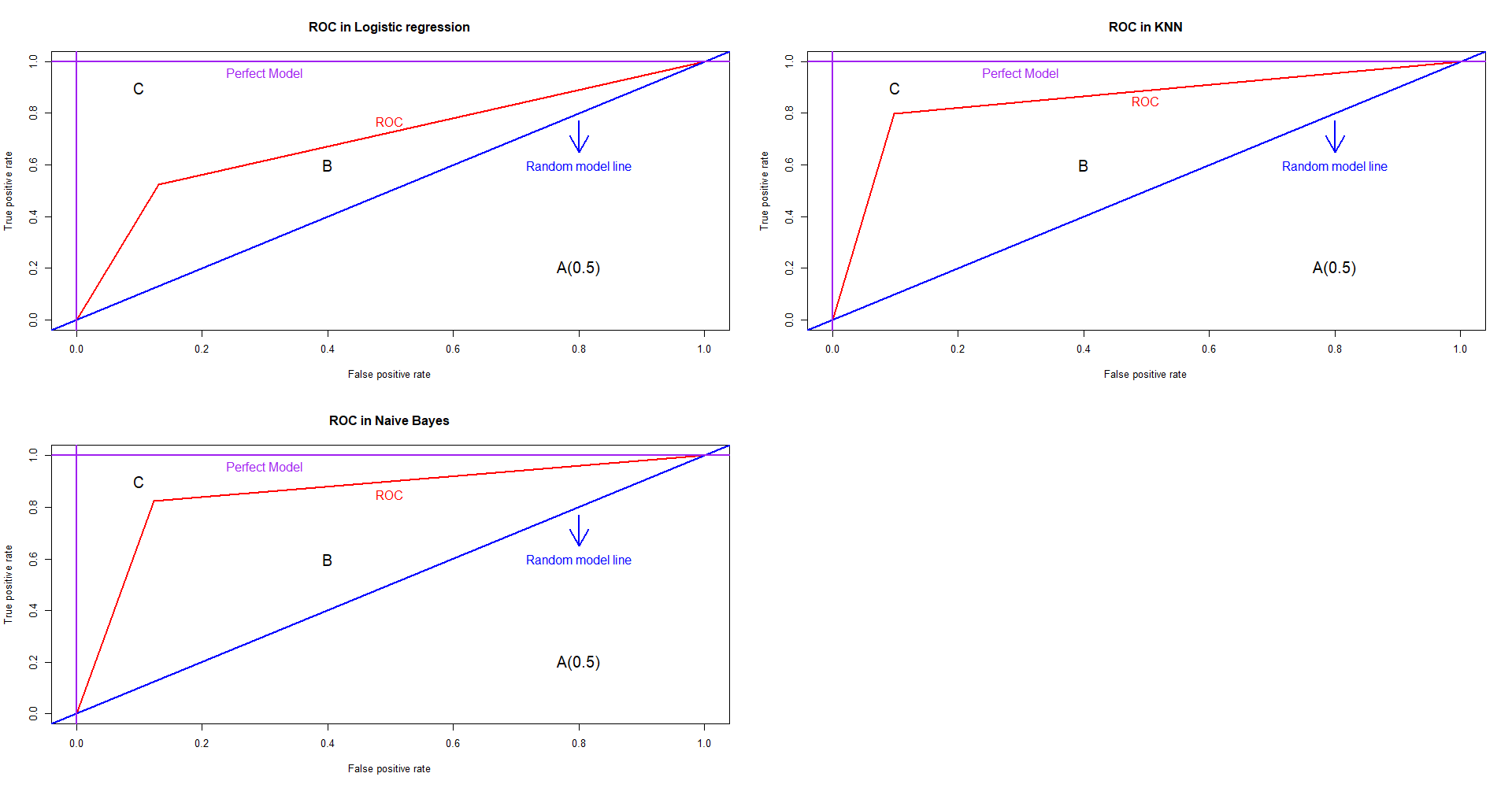
**Sensitivity = True Negative/ (True Negative + False Positive)**

1. Concordance Ratio: - It is the ratio of concordant pairs to the total number of pairs. After making all the pairs of alternative classes, the pairs in which the predicted probability for class 1 is higher than the predicted probability for class 0 is considered as concordant pairs. Higher the concordance ratio, better the model.

**Concordance Ratio = Number of concordant pairs / Total number of pairs**

1. KS stat: - The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.
2. Area Under Curve/ Receiver Operating Characteristics (AUC/ROC): - It signifies the degree of correct predictions made by the model. Higher the AUC, better the model.
3. Gini Coefficient: - It is the ratio of area between ROC curve and random line to the area between the perfect ROC model and random model line. Higher the Gini coefficient, better the model.

**Gini = 2AUC - 1**



The above figure shows Receiver Operating Characteristics of all the models. We can use ROC curve to derive the Gini coefficient and Area under curve (AUC).

Gini coefficient = B/B+C = B/0.5

AUC = A + B = 0.5 + B

Thus, using both equation we get, **Gini = 2AUC – 1**

The results drawn from performance measure are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Performance Measures** | **KNN** | **Logistic** | **Naïve Bayes** |
| **Accuracy** | 0.8946895 | 0.8568857 | 0.8748875 |
| **Classification error** | 0.1053105 | 0.1431143 | 0.1251125 |
| **Sensitivity** | 0.9016393 | 0.8692810 | 0.8765088 |
| **Specificity** | 0.7972973 | 0.5250000 | 0.8235294 |
| **KS stat** | 0.6989366 | 0.3942810 | 0.7000382 |
| **AUC** | 0.8494683 | 0.6971405 | 0.8500191 |
| **Gini** | 0.7485327 | 0.5122880 | 0.7449367 |
| **Concordance** | 0.7894148 | 0.7952860 | 0.8393266 |

We can draw following conclusion from the above table:

1. Logistic regression is the worst model as it performed poorly on all the measure as compared to other models.
2. Performance measures like Gini, AUC and KS stat are same for KNN model and Naïve Bayes model.
3. Naïve Bayes model outperform KNN model in respect of Specificity and Concordance Ratio.
4. But KNN model outperforms Naïve Bayes model in respect of Sensitivity and Accuracy. Thus, we can say that KNN model is the best model out of three models with accuracy of 90%.

**Business Insights**

1. Total number of calls made to customer service has a significant positive effect on whether a customer will cancel post-paid service or not.
2. Total number minutes customer talks in a daytime and total number of calls made in a day also influence the decision of the customer.
3. If a customer has renewed the post-paid service contract, it is highly likely that he will not cancel the service in future.
4. The total spending on the monthly bills also defines whether a person will cancel the service or not.
5. The decision can also be affected by whether a customer has availed data plan or not and even if he/she has availed it then how many gigabytes of data he/she is using in a month.
6. The decision is also influenced by the average number of roaming minutes and the largest overhead fee, if any, charged from the customer.

**Recommendations**

To reduce the proportion of customers who cancelled the post-paid service, following actions should be taken:

1. Strengthening the compliant resolution mechanism by addressing the customer’s problems on time and training the customer support staff.
2. Some segment of group can be persuaded by providing special services like higher talktime, monetary concessions on data plans or by increasing amount of data in a plan.
3. A fee waiver scheme can be introduced for those customers whose overage fee and roaming charges are relatively low.
4. Customers whose monthly bill is above a particular threshold can be provided special benefits like discount coupons on next payment, shopping vouchers, free 1-month subscription for Netflix, extra discount on making payment via a particular debit/credit card or online wallets, etc.