|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Predictive Modeling Project | /var/folders/2g/b4yp_nt14cb2s12zn8fq29br0000gn/T/com.microsoft.Word/Content.MSO/8C58AE03.tmp | | About Predictive Modeling News | Predictive Modeling: How Data Can Help You Predict the Future ... | |
| **Predicting Customer Churn in Telecom**  By – Shweta Gupta  PGPBABI |

INDEX

* BACKGROUND AND OBJECTIVES
* DATA DICTIONARY
* EXPLORATORY DATA ANALYSIS AND INTERPRETATION
* MODEL BUILDING – LOGISTIC REGRESSION
  + KEY STEPS, ASSUMPTIONS, MODEL RESULTS, EVALUATION MEASURES SUMMARY, INTERPRETATION
* MODEL BUILDING – KNN
  + KEY STEPS, MODEL RESULTS, EVALUATION MEASURES SUMMARY, INTERPRETATION
* MODEL BUILDING – NAÏVE BAYES
  + KEY STEPS, ASSUMPTIONS, MODEL RESULTS, EVALUATION MEASURES SUMMARY, INTERPRETATION
* CHOOSING THE BEST FIT MODEL & CONCLUSION

**Background and Objectives**

**Telecom Customer Churn Prediction Assessment**

* Customer Churn is a burning problem for Telecom companies.
* In this project, we simulate one such case of customer churn where we work on a data of postpaid customers with a contract.
* The data has information about the customer usage behavior, contract details and the payment details.
* The data also indicates which were the customers who canceled their service.
* Based on this past data, we need to build a model which can predict whether a customer will cancel their service in the future or not.

**Objectives**

1. **EDA** 
   * How does the data looks like, Univariate and bivariate analysis. Plots and charts which illustrate the relationships between variables
   * Look out for outliers and missing values
   * Check for multicollinearity & treat it
   * Summarize the insights you get from EDA
2. **Build Models and compare them to get to the best one**
   * Logistic Regression
   * KNN
   * Naive Bayes (is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?)
   * Model Comparison using Model Performance metrics & Interpretation
3. **Actionable Insights**
   * Interpretation & Recommendations from the best model

**Data Dictionary**

* The Dataset contains 3333 observations across 11 variables with description as below.
* Churn is our dependent variable which we will seek to predict

|  |  |
| --- | --- |
| Variables |  |
| Churn | 1 if customer cancelled service, 0 if not |
| AccountWeeks | number of weeks customer has had active account |
| ContractRenewal | 1 if customer recently renewed contract, 0 if not |
| DataPlan | 1 if customer has data plan, 0 if not |
| DataUsage | gigabytes of monthly data usage |
| CustServCalls | number of calls into customer service |
| DayMins | average daytime minutes per month |
| DayCalls | average number of daytime calls |
| MonthlyCharge | average monthly bill |
| OverageFee | largest overage fee in last 12 months |
| RoamMins | average number of roaming minutes |

**DATA STRUCTURE OF OUR DATASET**

tibble [3,333 × 11] (S3: tbl\_df/tbl/data.frame)

$ Churn : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

$ AccountWeeks : num [1:3333] 128 107 137 84 75 118 121 147 117 141 ...

$ ContractRenewal: Factor w/ 2 levels "0","1": 2 2 2 1 1 1 2 1 2 1 ...

$ DataPlan : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 2 1 1 2 ...

$ DataUsage : num [1:3333] 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...

$ CustServCalls : num [1:3333] 1 1 0 2 3 0 3 0 1 0 ...

$ DayMins : num [1:3333] 265 162 243 299 167 ...

$ DayCalls : num [1:3333] 110 123 114 71 113 98 88 79 97 84 ...

$ MonthlyCharge : num [1:3333] 89 82 52 57 41 57 87.3 36 63.9 93.2 ...

$ OverageFee : num [1:3333] 9.87 9.78 6.06 3.1 7.42 ...

$ RoamMins : num [1:3333] 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...

* All variables are continuous except for **Churn ,Contract Renewal And Data Plan** Which are factor variables
* Distribution of Churn in dataset:

0 1

0.855 0.144

* For our dependent variable Churn in our dataset , **14% have cancelled service** whereas 86% continue with the service

Exploratory Data Analysis

**FIVE POINT SUMMARY**

Churn AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls DayMins

0:2850 Min. : 1.0 0: 323 0:2411 Min. :0.0000 Min. :0.000 Min. : 0.0

1: 483 1st Qu.: 74.0 1:3010 1: 922 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:143.7

Median :101.0 Median :0.0000 Median :1.000 Median :179.4

Mean :101.1 Mean :0.8165 Mean :1.563 Mean :179.8

3rd Qu.:127.0 3rd Qu.:1.7800 3rd Qu.:2.000 3rd Qu.:216.4

Max. :243.0 Max. :5.4000 Max. :9.000 Max. :350.8

DayCalls MonthlyCharge OverageFee RoamMins

Min. : 0.0 Min. : 14.00 Min. : 0.00 Min. : 0.00

1st Qu.: 87.0 1st Qu.: 45.00 1st Qu.: 8.33 1st Qu.: 8.50

Median :101.0 Median : 53.50 Median :10.07 Median :10.30

Mean :100.4 Mean : 56.31 Mean :10.05 Mean :10.24

3rd Qu.:114.0 3rd Qu.: 66.20 3rd Qu.:11.77 3rd Qu.:12.10

Max. :165.0 Max. :111.30 Max. :18.19 Max. :20.00

**PRESENCE OF MISSING VALUES**

Churn AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls DayMins

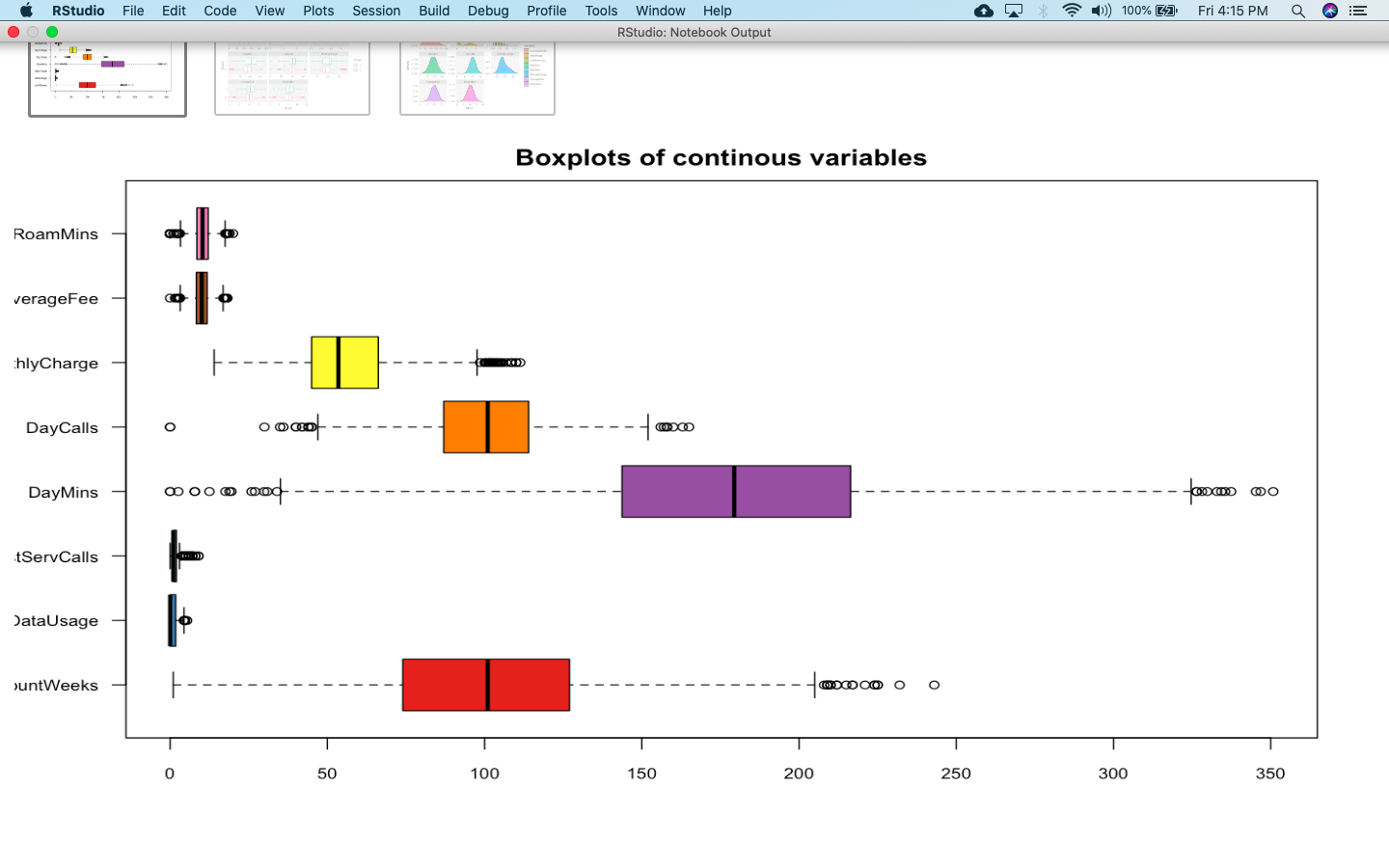
0 0 0 0 0 0 0

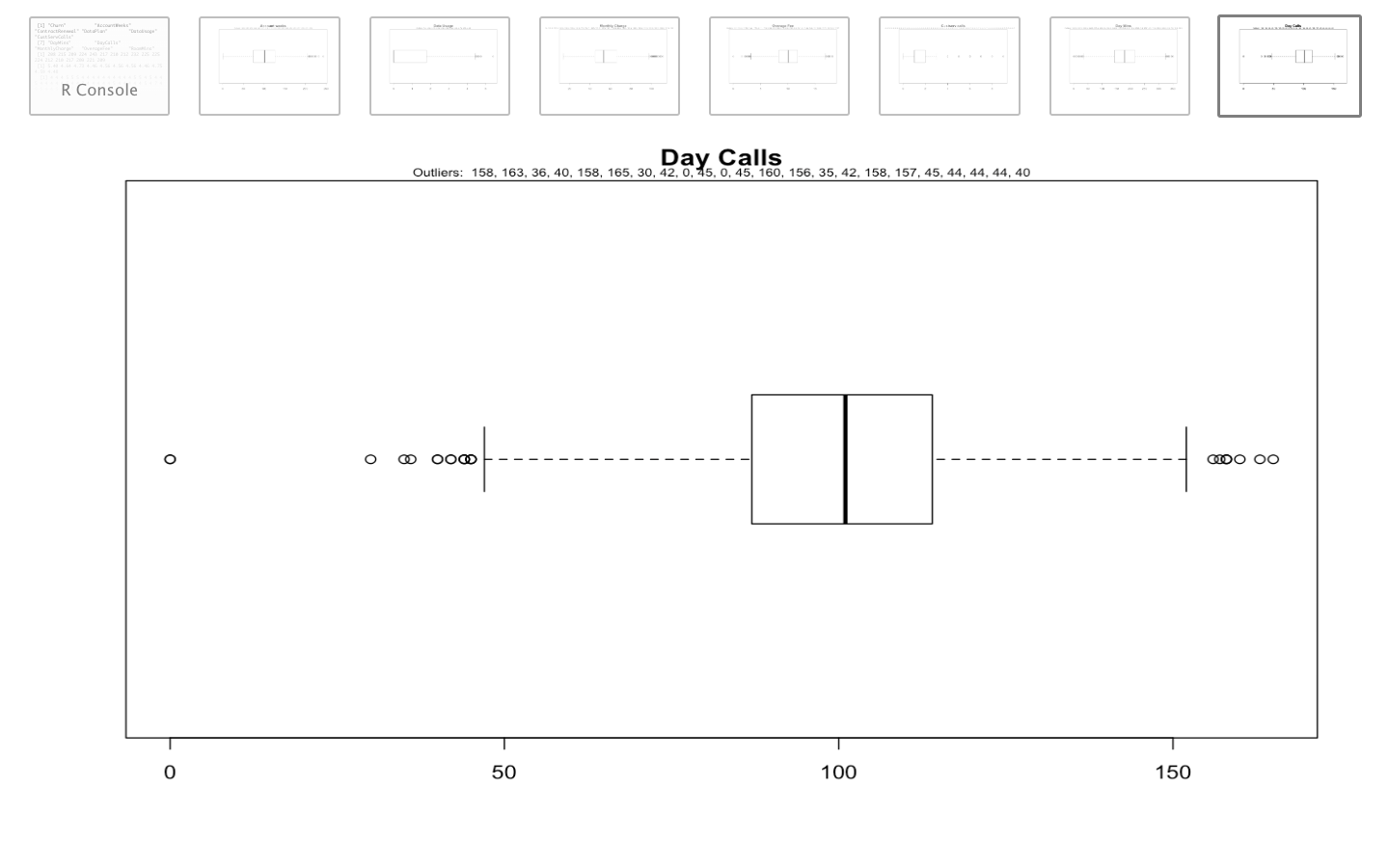
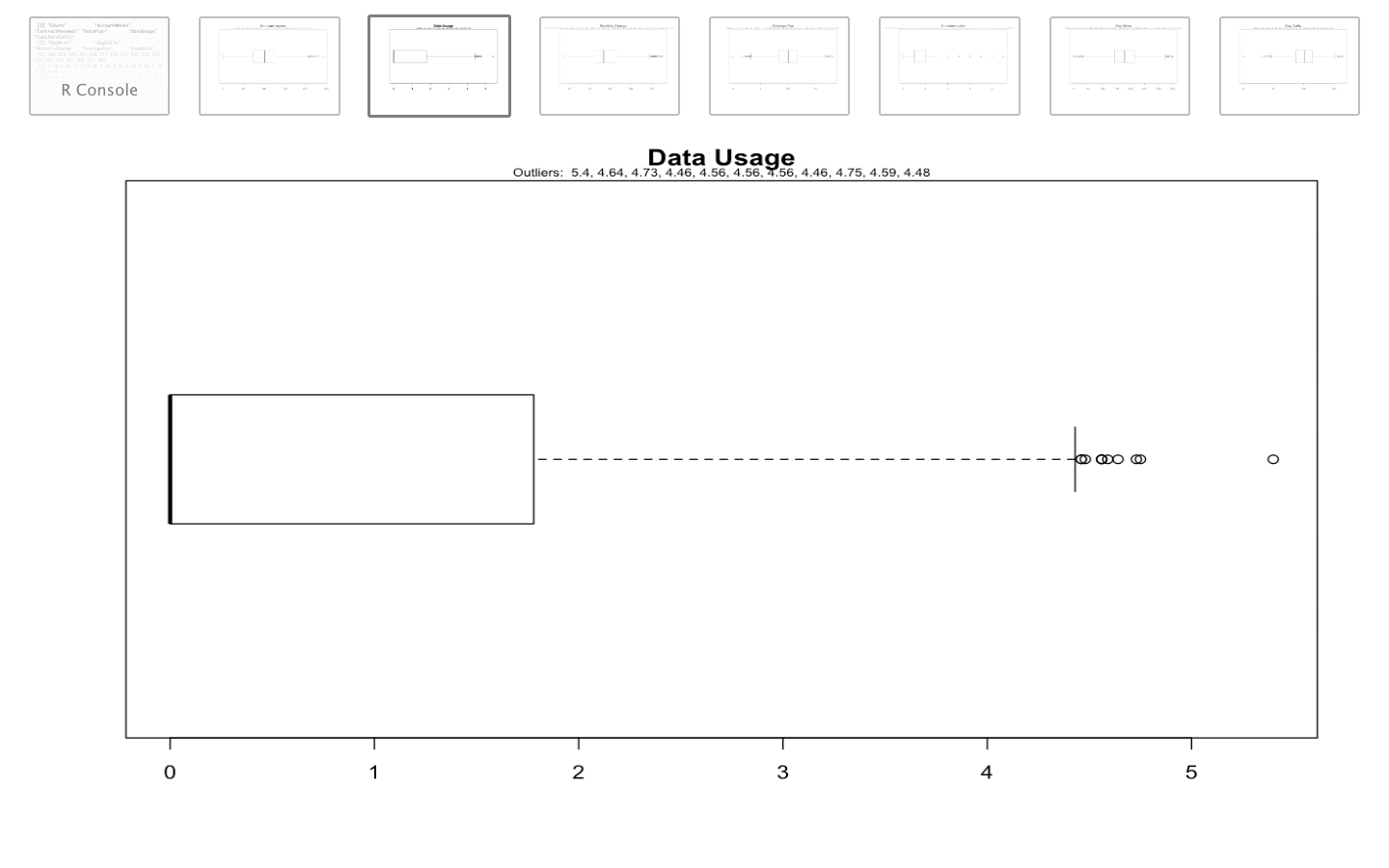
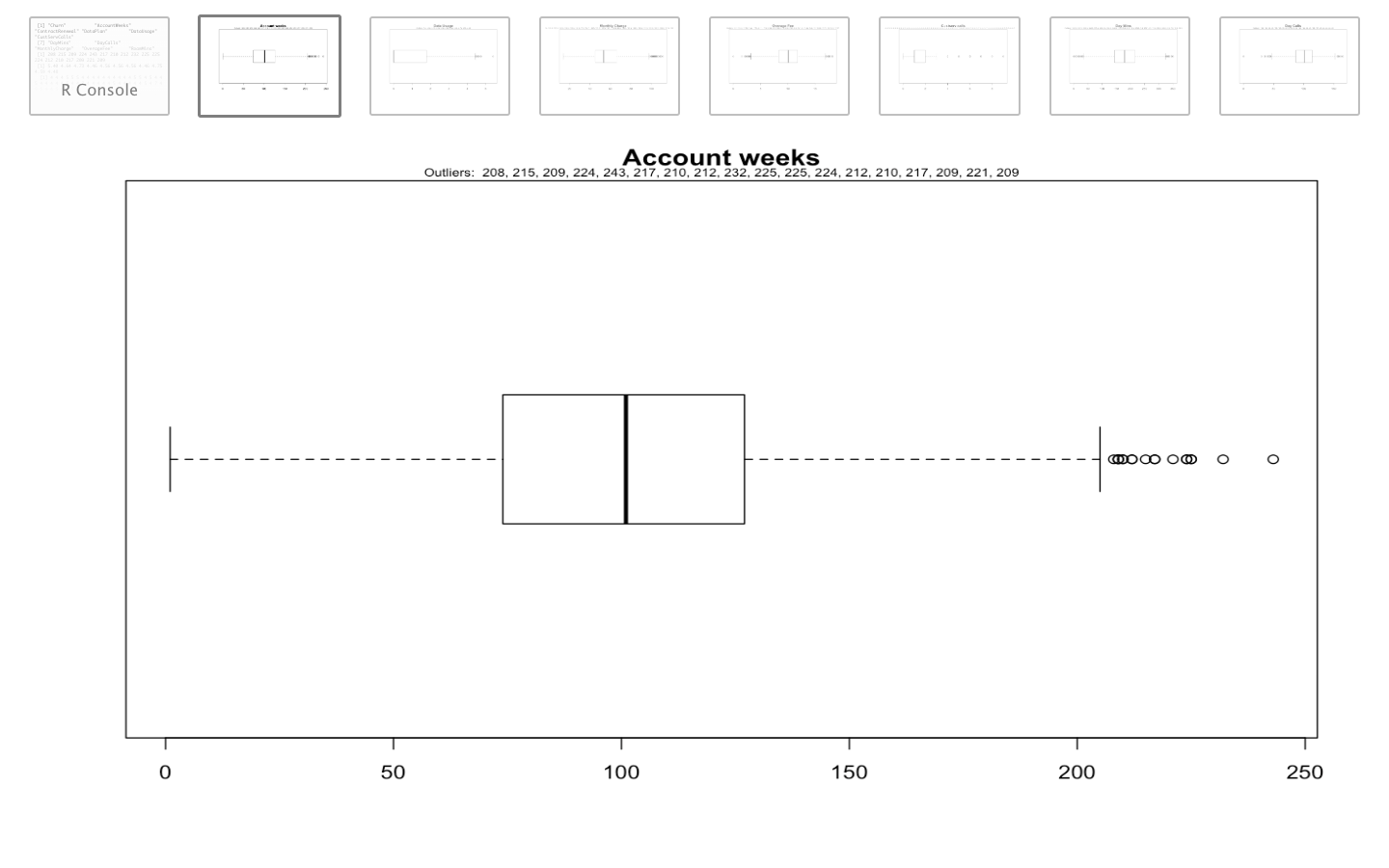
DayCalls MonthlyCharge OverageFee RoamMins

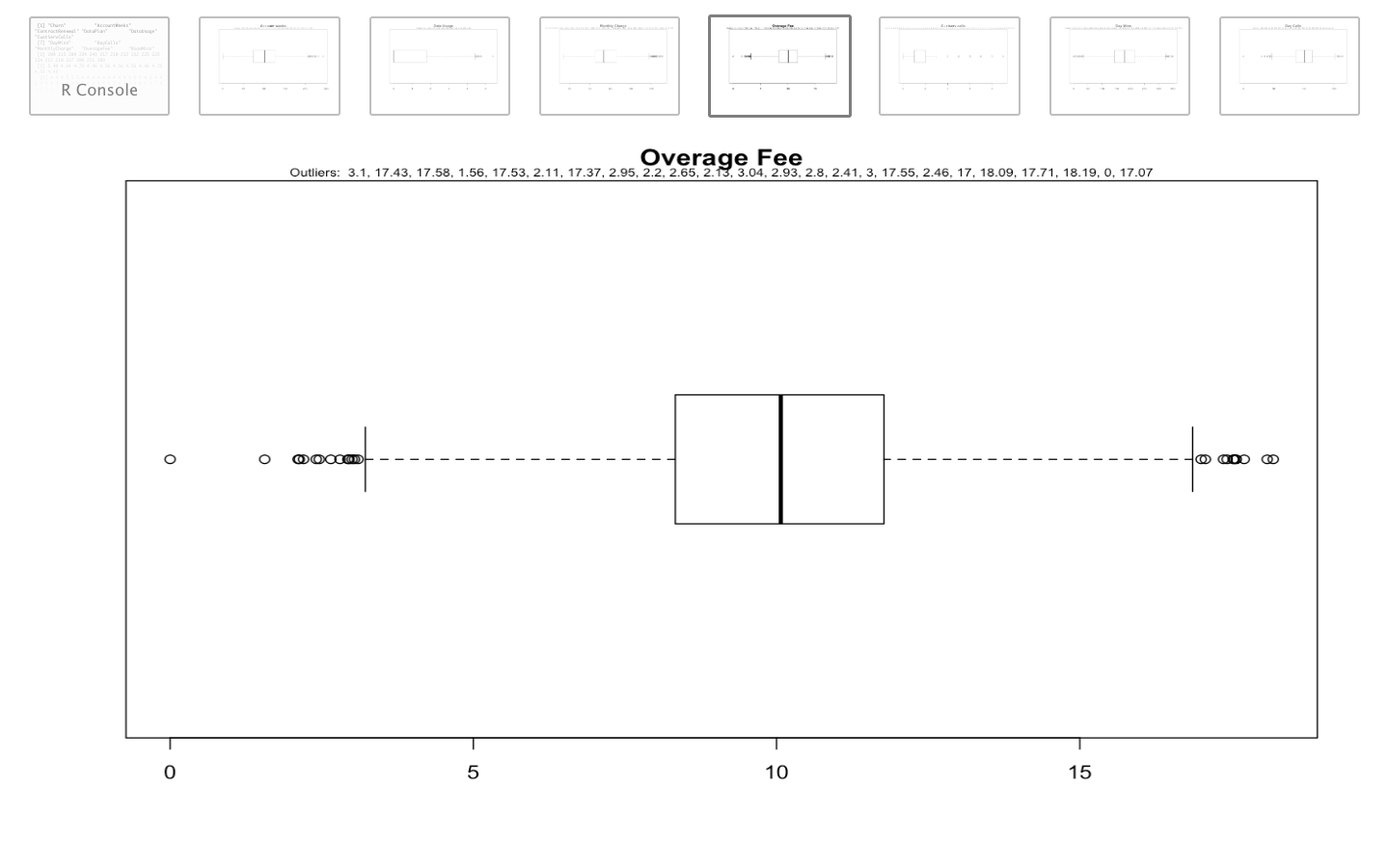
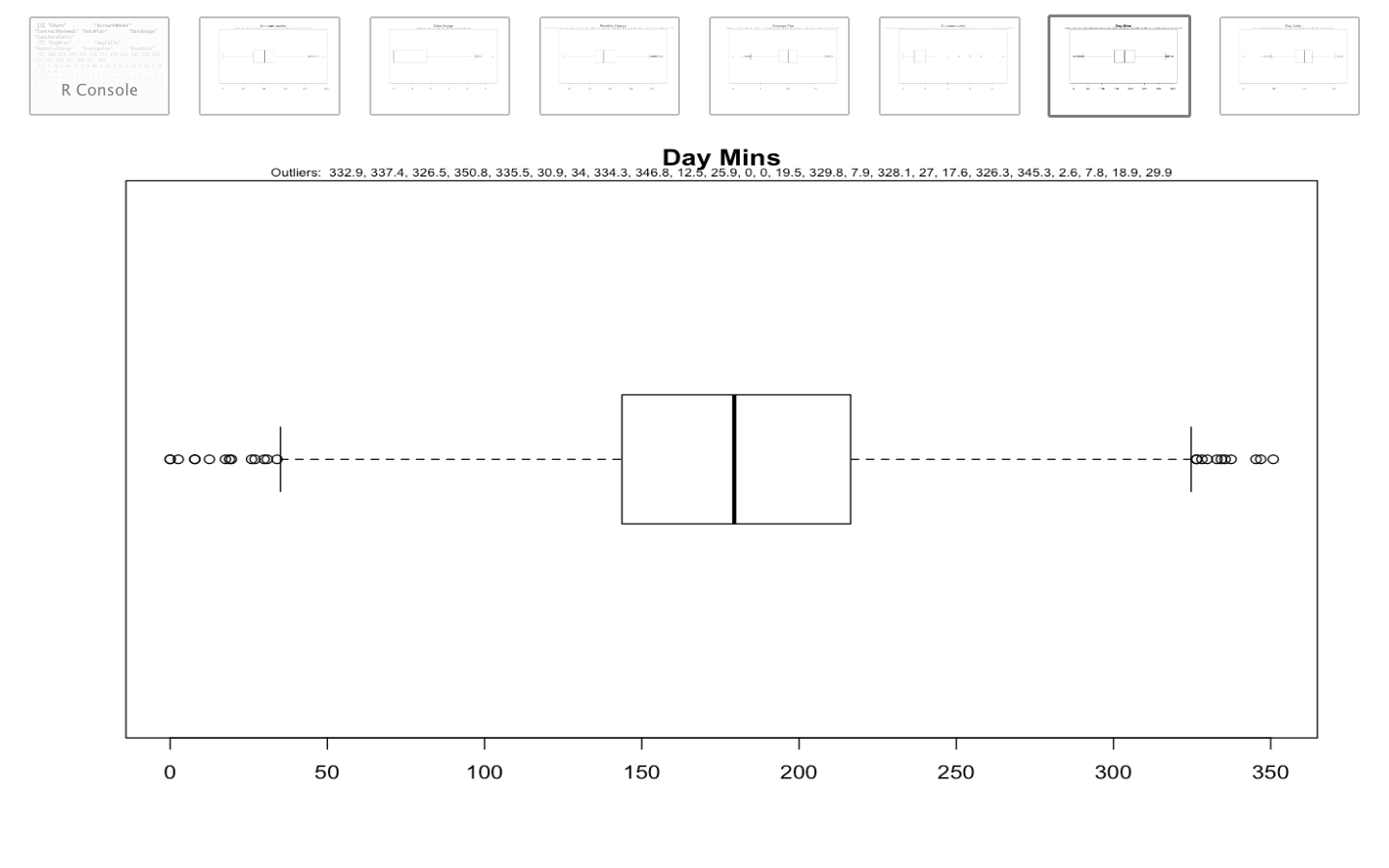
0 0 0 0

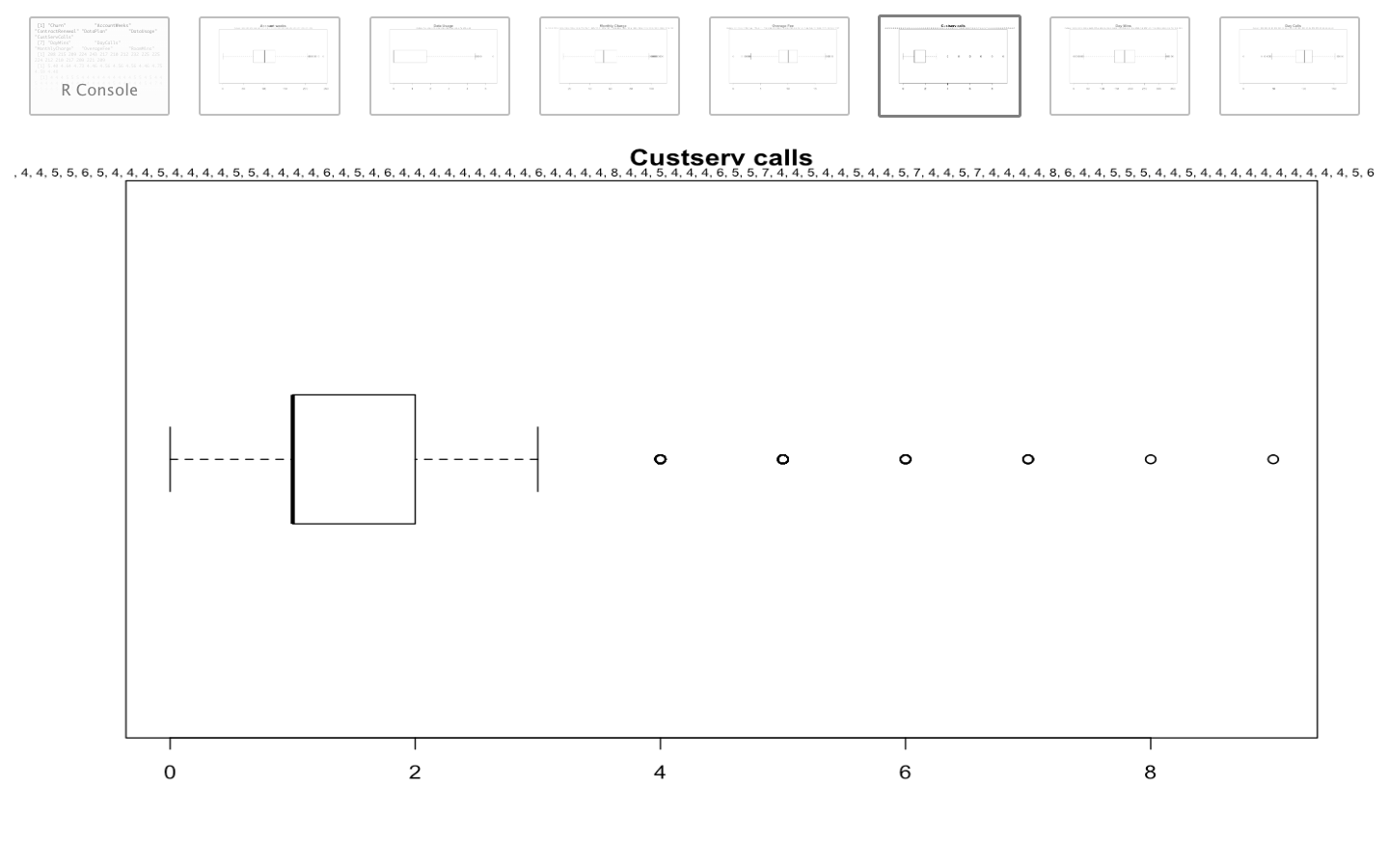
* Highly skewed data for most continuous variables indicating presence of outliers
* No missing values in the data

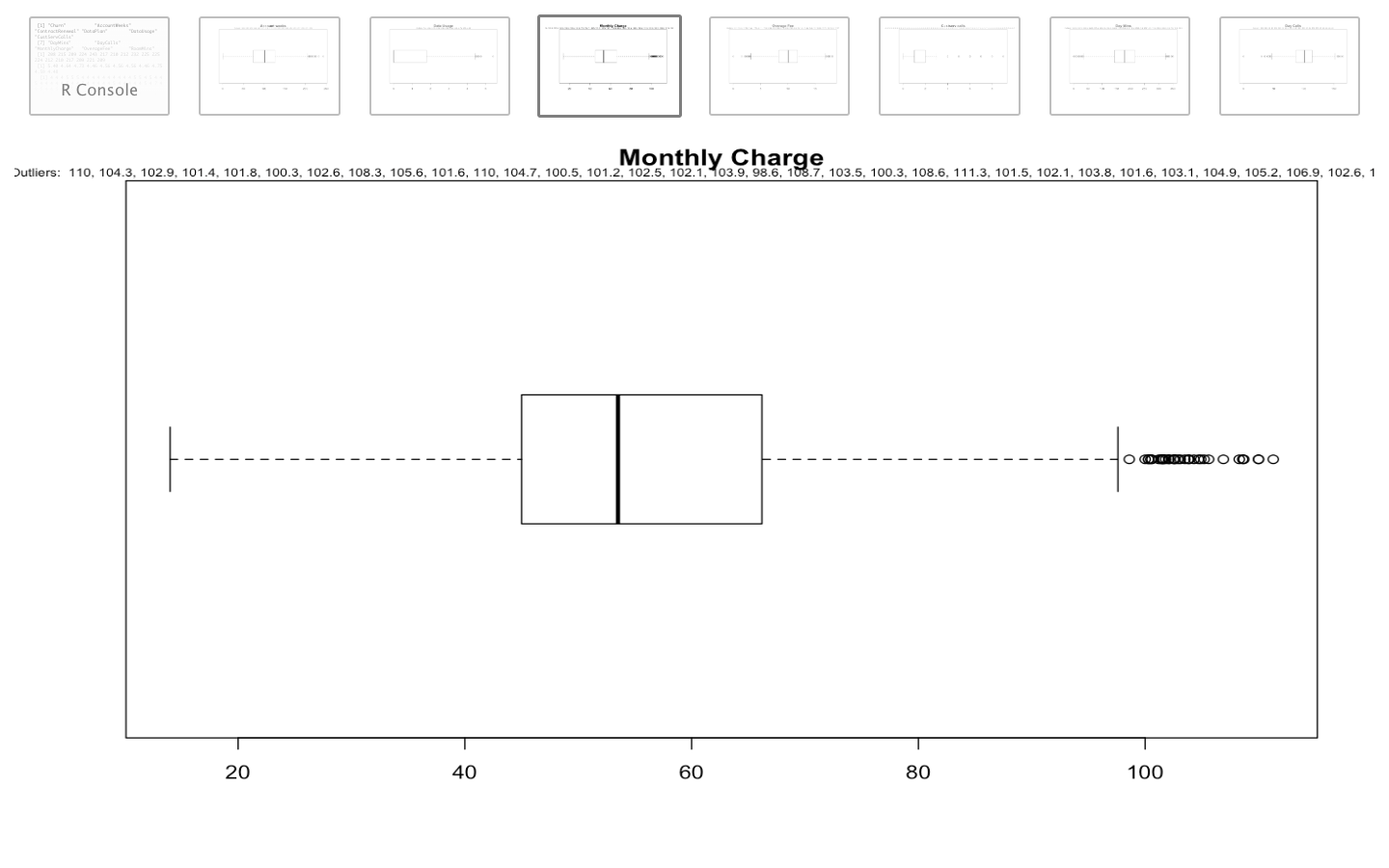
**IDENTIFYING THE OUTLIERS IN CONTINUOS VARIABLES**





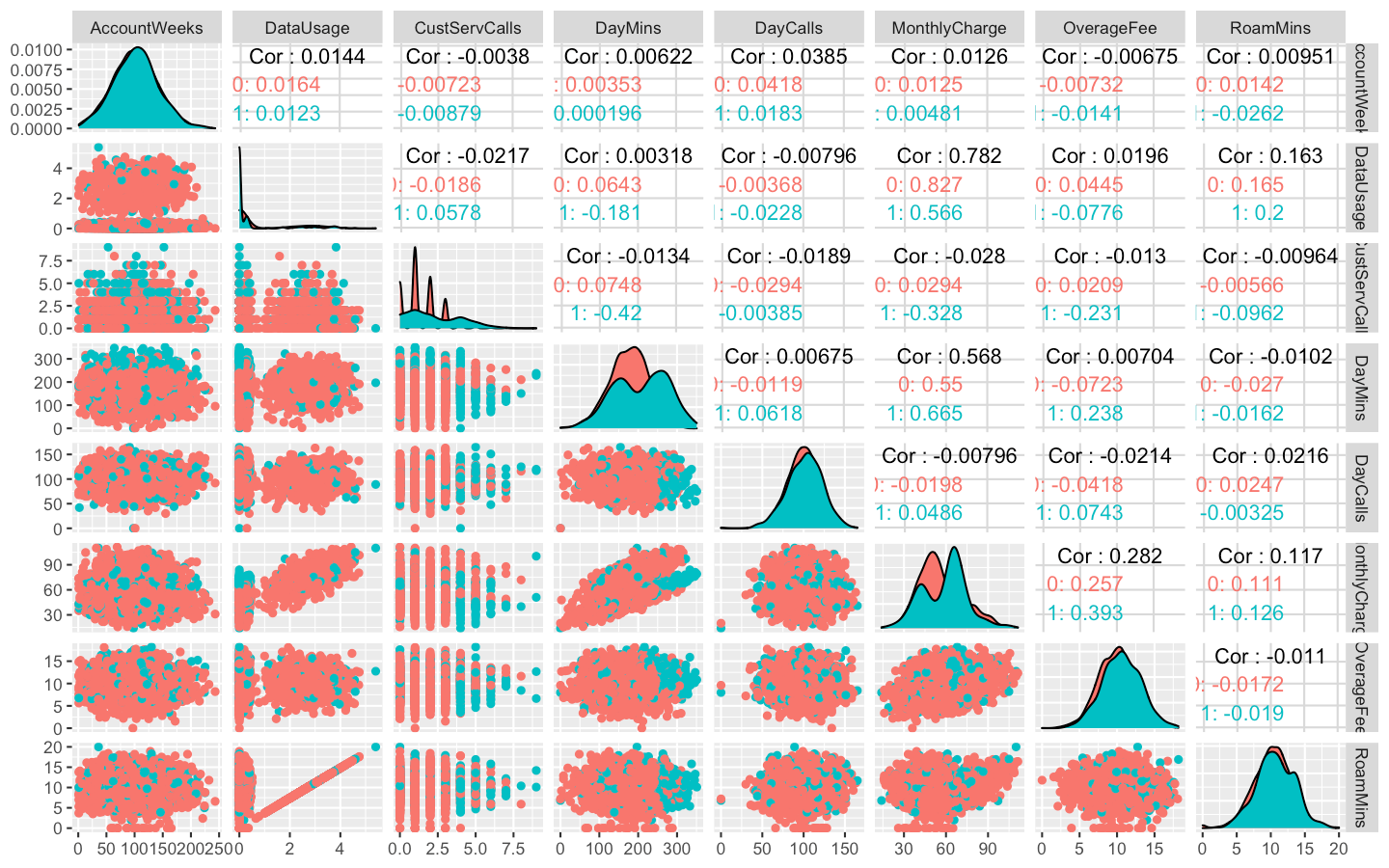






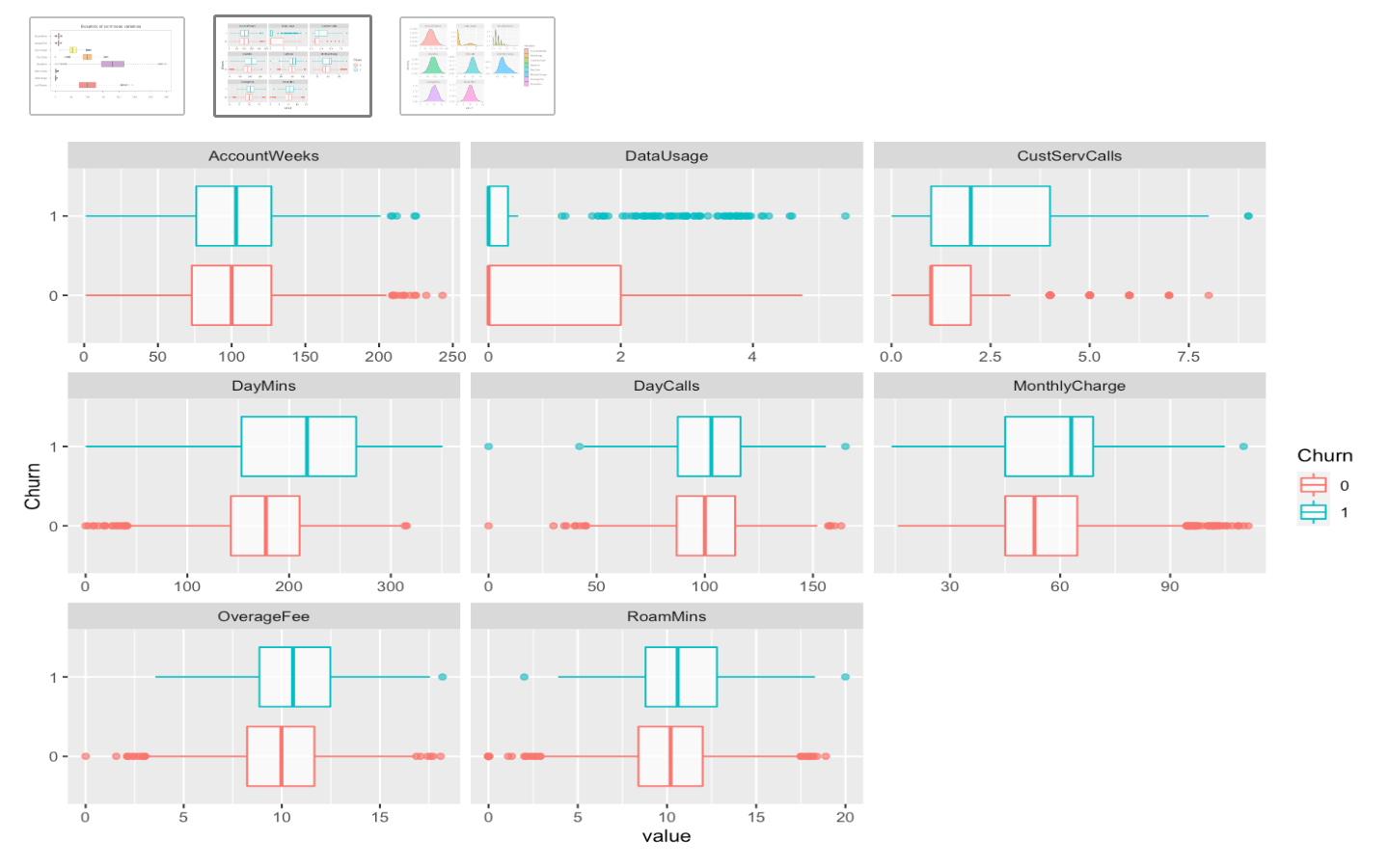
* All continuous variables in our data indicate presence of outliers
* We will however further look at how these outliers impact our model during model building stage and treat them accordingly

**CHECK VISUAL ASSOCIATION PATTERN FOR CONTINOUS PREDICTOR VARIABLES**

  
***Legend: Color – Churn variable***

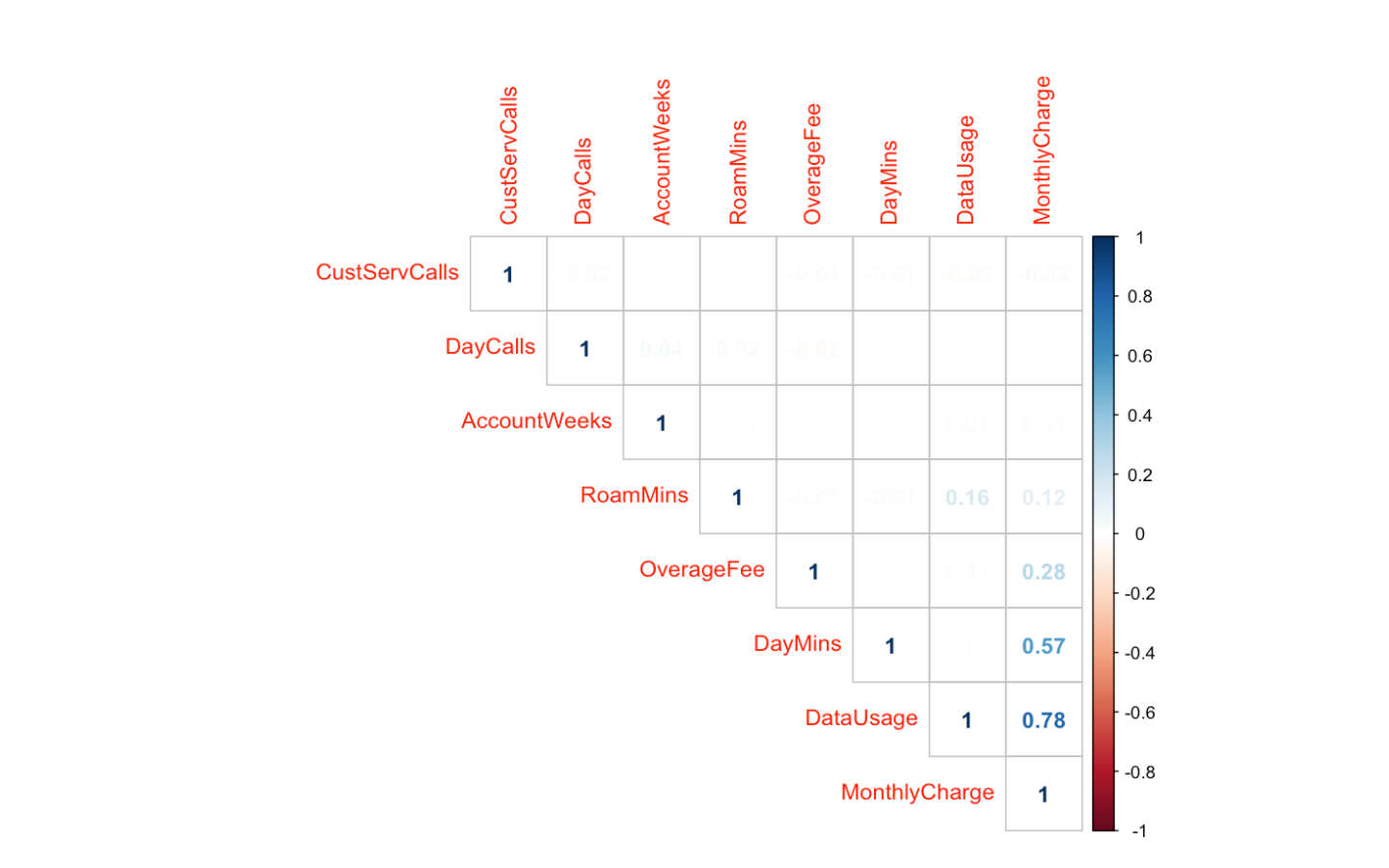
* From the above chart we can deduce that Roam Mins, Overage Fee, Monthly Charges, Day Calls, Day mins and Account weeks follow normal distribution.
* Other continuous variables - Data Usage and Custservcalls are left skewed.
* Data Usage and Roam Mins seem to have linear relationship.
* Data Usage and Monthly charges also seem to have a linear relationship.
* Similarly Day mins and Monthly charges also seem to have a linear relationship.

**BOXPLOTS FOR CONTINUOUS PREDICTOR VARIABLES**



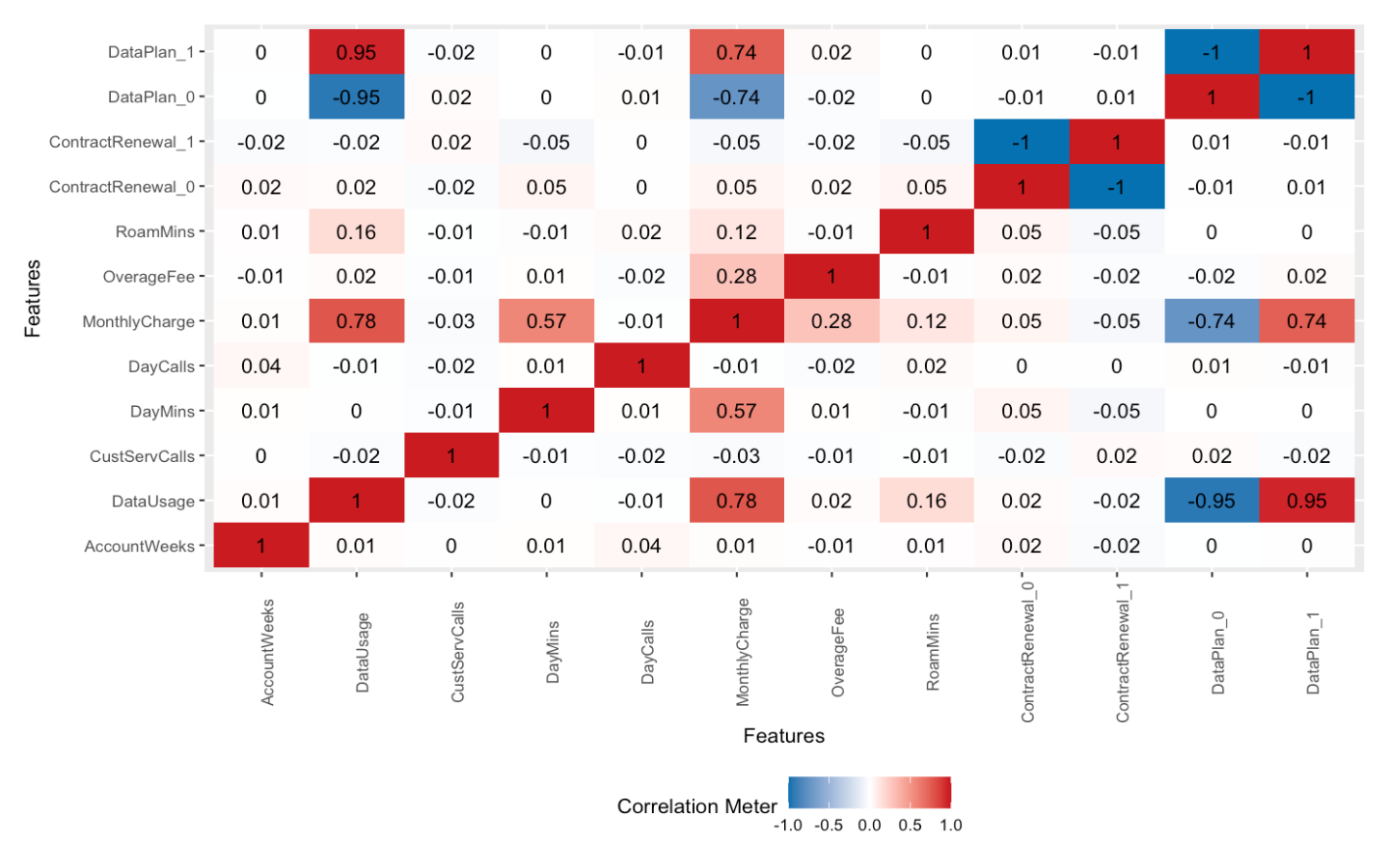
* The Boxplots for continuous variables by Churn indicate **higher Custservcalls, Daymins, Monthly Charges ,Overage Fee ad Roam mins for those who cancelled their service** compared to those who continue the service
* Data usage is apparently on the lower side for those who cancelled their service compared to those who continue the service

**CORRELATION MATRIX FOR CONTINUOS PREDICTOR VARIABLES**

****

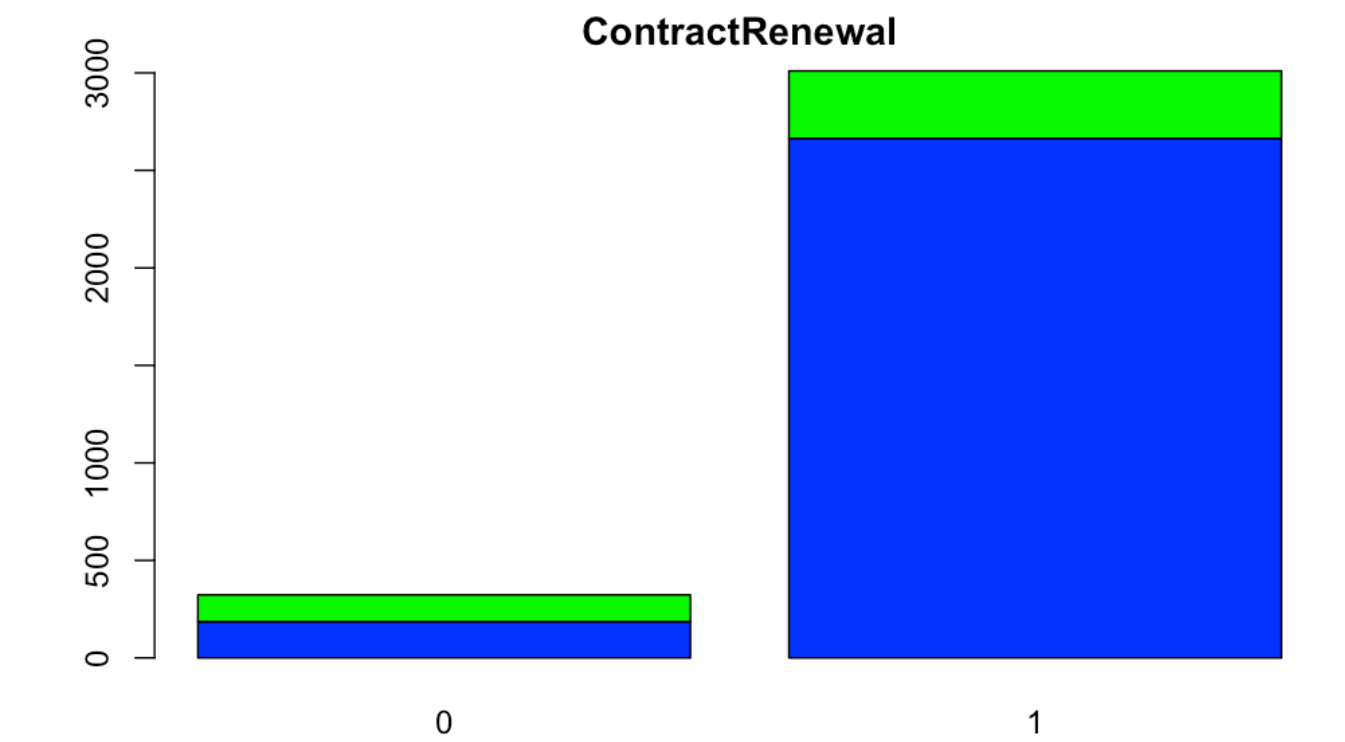
* Data usage and monthly charge seem to be correlated with r = 0.78
* Day mins and monthly charge seem to be mildly correlated with r = 0.57

**CORRELATION MATRIX FOR CONTINUOS AND CATAGORICAL PREDICTOR VARIABLES**

****

* Data Plan and Data Usage are highly correlated
* Data Plan and Monthly charges are also moderately correlated
* We will treat multicollinearity based on our results from base Logistic regression model

**CONTINGENCY TABLE OF DICHOTOMOUS VARIABLES WITH TARGET VARIABLE**

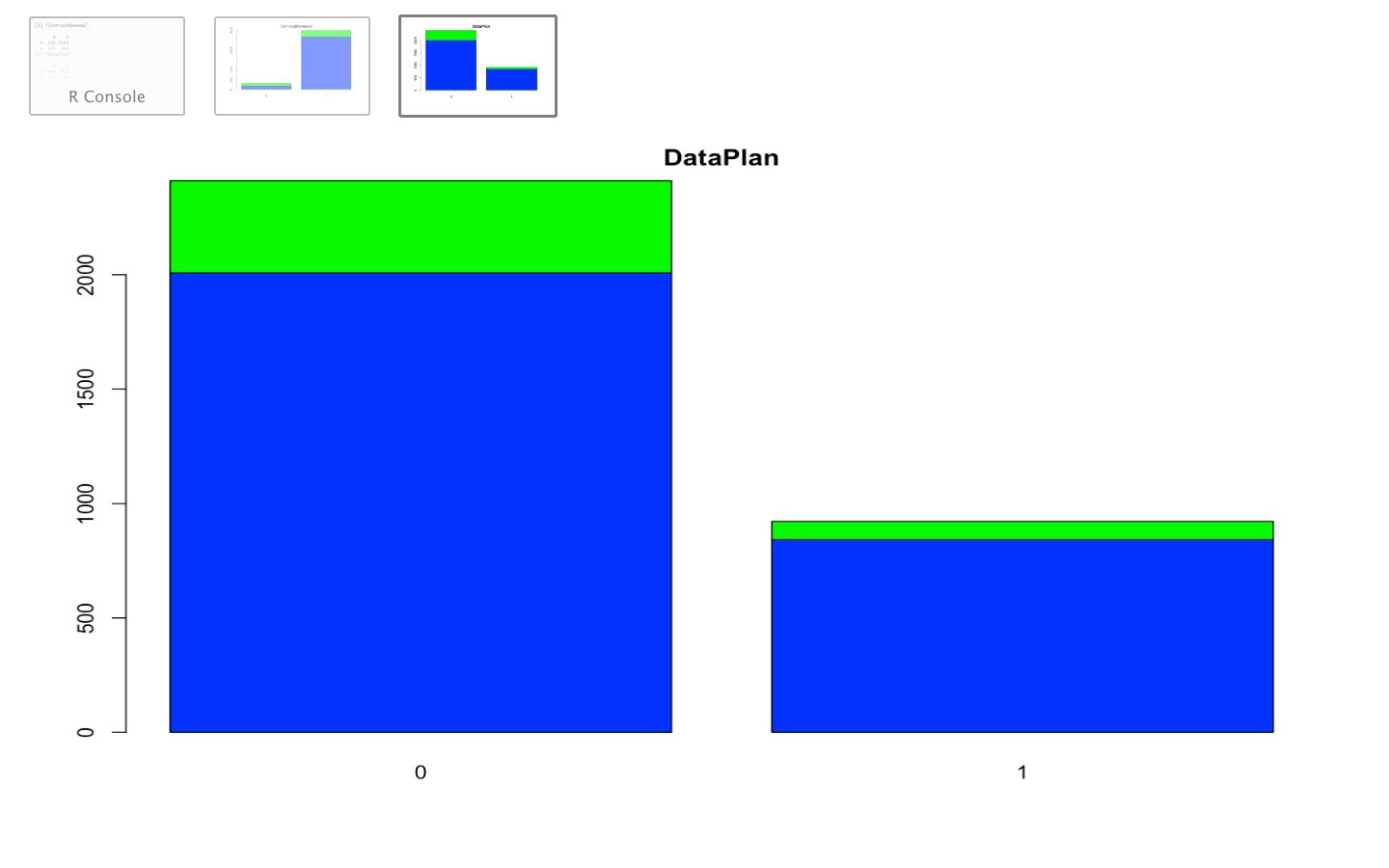
(table(data$Churn,ct.data[[i]]))

[1] "ContractRenewal"

0 1

0 186 2664

1 137 346



[1] "DataPlan"

0 1

0 2008 842

1 403 80

**CHI SQUARE TEST WITH TARGET VARIABLE**

| **Row**  <fctr> | **Column**  <fctr> | **Chi.SQuare**  <dbl> | **df**  <int> | **p.value**  <dbl> |
| --- | --- | --- | --- | --- |
| Churn | ContractRenewal | 222.56576 | 1 | 2.493108e-50 |
| Churn | DataPlan | 34.13166 | 1 | 5.150640e-09 |

* Contract renewal seem to have an inverse relationship with Churn
* Those who do not have a data plan are more like to leave, compared to those who have a data plan
* Both categorical variables **Contract renewal and Data plan are significant towards explaining our target variable** since their pvalue in chi sqare test is below .05 confidence level

**SUMMARISING EDA**

* The Dataset contains 3333 observations across 11 variables.
* All Variables Are Continuous Except For Churn ,Contract Renewal And Data Plan Which Are Factor Variables
* For our dependent variable Churn, in our dataset 14% have cancelled service whereas 86% continue with the service
* All continuous variables in our data indicate presence of outliers
* We will however further look at how these outliers impact our model during model building stage and treat them accordingly
* Roam Mins, Overage Fee, Monthly Charges, Day Calls, Day mins and Account weeks follow normal distribution.
* Other continuous variables - Data Usage and Custservcalls are left skewed.
* Data Usage and Roam Mins seem to have linear relationship.
* Data Usage and Monthly charges also seem to have a linear relationship.
* Similarly Day mins and Monthly charges also seem to have a linear relationship.
* Data usage and monthly charge seem to be correlated with r = 0.78
* Day mins and monthly charge seem to be mildly correlated with r = 0.57
* Data Plan and Data Usage are highly correlated
* Data Plan and Monthly charges are also moderately correlated
* We will treat multicollinearity based on our results from base Logistic regression model
* The Boxplots for continuous variables by Churn indicate higher Custservcalls, Daymins, Monthly Charges ,Overage Fee ad Roam mins for those who cancelled their service compared to those who continue the service
* Data usage is on the lower side for those who cancelled their service compared to those who continue the service
* Contract renewal seem to have an inverse relationship with Churn
* Those who do not have a data plan are more like to leave, compared to those who have a data plan
* Both categorical variables contract renewal and data plan are significant towards explaining our target variables since our pvalue in chi sqare test is below .05 confidence level

Data Preparation

**SPLIT THE DATA INTO TEST AND TRAIN**

Split ratio – Train : Test = 70:30

* TRAIN - 2333 observations for 11 variables
* TEST - 1000 observations for 11 variables

**CHECK SPLIT CONSISTENCY FOR CHURN VARIABLE ACROSS ALL DATA SETS**

**CHURN = = 1 / TOTAL OBSERVATIONS**

* Base Data = 0.1449145
* Test Data = 0.145
* Train Data= 0.1448778
* Hence the split data is consistent across all datasets

Model Building – Logistic Regression

**Logistic regression** uses a **logistic** function to model a binary dependent variable.

**MODEL BUILDING – LOGISTIC REGRESSION**

**ASSUMPTIONS**

* The **outcome is a binary** or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.
* There is a **linear relationship between the logit of the outcome and each predictor** variables.
* There is **no influential values (extreme values or outliers)** in the continuous predictors
* There is **no high intercorrelations (i.e. multicollinearity) among the predictors.**

**Key Steps**

* Split data in train and test in 70:30 ratio
* Run the model on train dataset using all variables , we call it base model
* Check for assumptions as above mentioned
* Take corrective steps on data and rerun the model to fine tune the model accuracy
* Check model evaluation measures on train and test dataset

**BASE MODEL WITH ALL VARIABLES**

Call:

glm(formula = Churn ~ ., family = binomial, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8569 -0.5001 -0.3388 -0.2002 3.0667

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.518167 0.663878 -9.818 < 2e-16 \*\*\*

AccountWeeks -0.000860 0.001696 -0.507 0.61202

ContractRenewal -1.955876 0.173443 -11.277 < 2e-16 \*\*\*

DataPlan -1.034234 0.666421 -1.552 0.12068

DataUsage -0.245846 2.305724 -0.107 0.91509

CustServCalls 0.530913 0.046981 11.301 < 2e-16 \*\*\*

DayMins 0.008667 0.039015 0.222 0.82420

DayCalls 0.007532 0.003294 2.286 0.02223 \*

MonthlyCharge 0.027078 0.229248 0.118 0.90597

OverageFee 0.109975 0.390682 0.281 0.77833

RoamMins 0.073374 0.026355 2.784 0.00537 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1930.4 on 2332 degrees of freedom

Residual deviance: 1511.0 on 2322 degrees of freedom

AIC: 1533

Number of Fisher Scoring iterations: 6

**CHECK FOR ASSUMPTIONS**

**VIF – CHECK MULTICOLLINEARITY AMONG PREDICTORS**

AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls

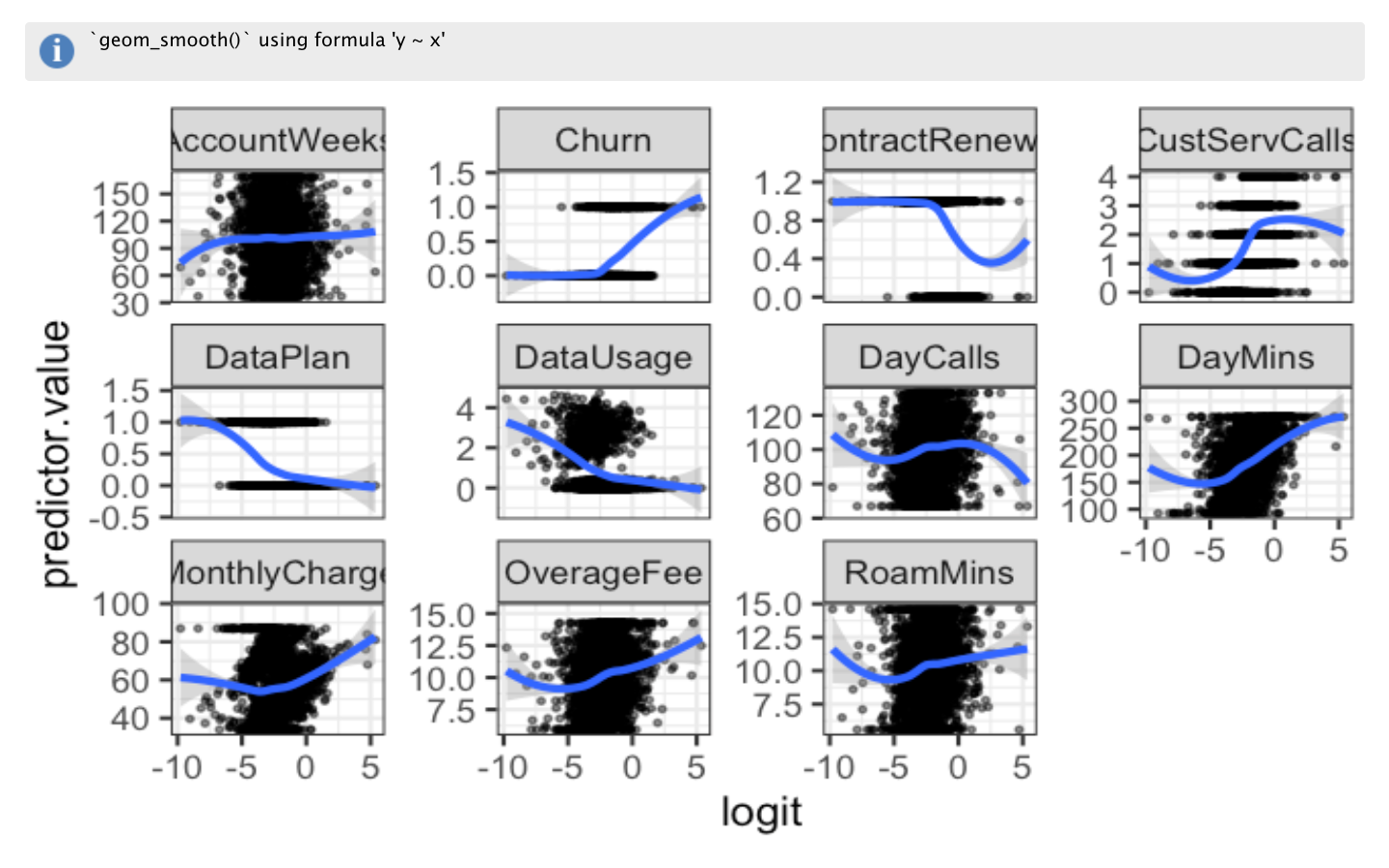
1.010827 1.065929 14.669441 1503.808964 1.081597

DayMins DayCalls MonthlyCharge OverageFee RoamMins

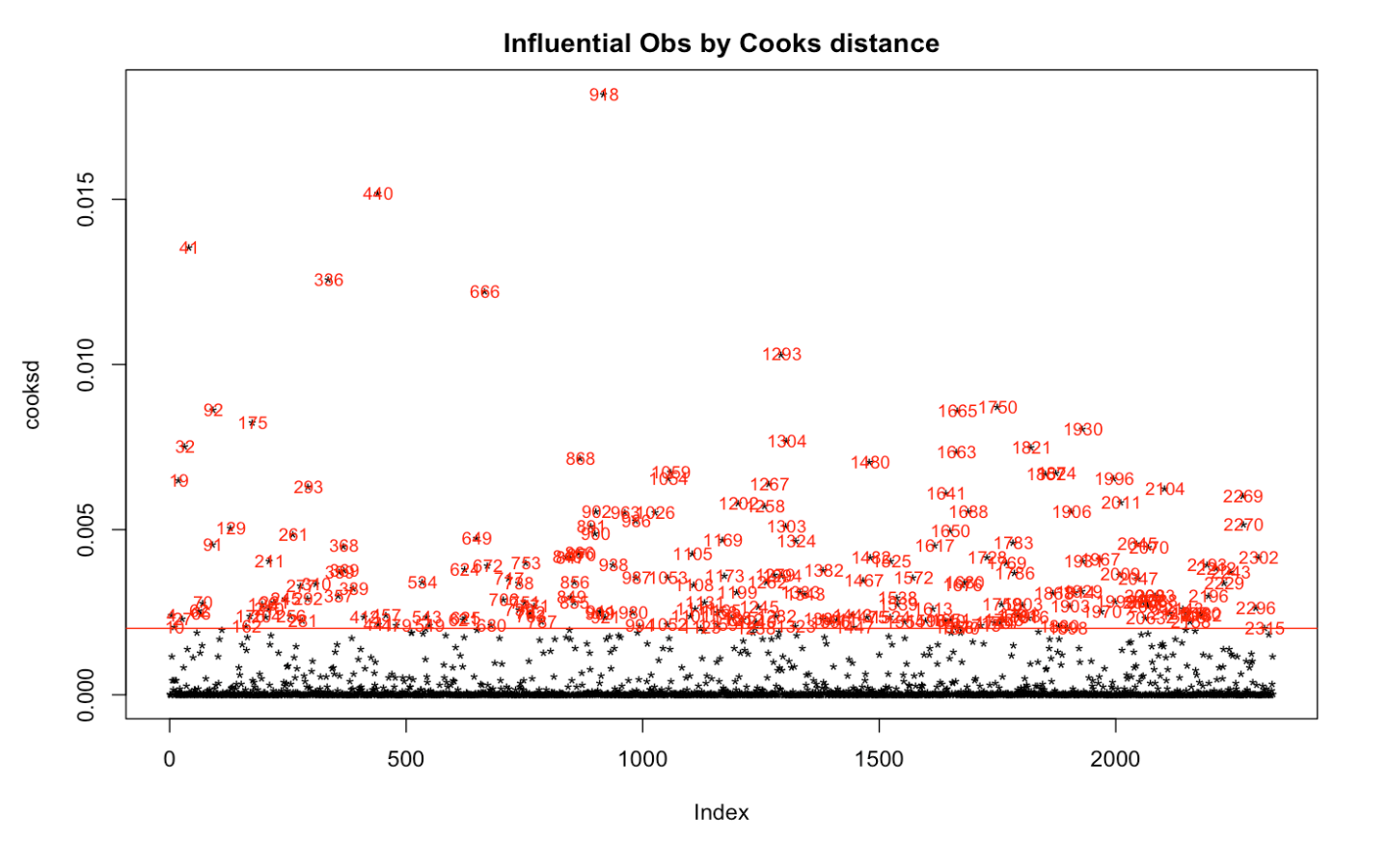
950.788675 1.011359 2653.526161 213.752869 1.165149

* Dataplan, Datausage, Day mins, Monthly charges and Overage fee have a higher VIF value >4
* Many insignificant variables in the model – account weeks, data plan, data usage, day mins, monthly charge and overage fee.

**CHECK NORMALITY ASSUMPTION OF PREDICTORS WITH LOGIT OF OUTCOME**



**CHECK FOR INFLUENTIAL POINTS – COOKS DISTANCE**

****

**ASSUMPTIONS CHECK FINDINGS**

* Many insignificant variables present in the model – Account Weeks, Data Plan, Data Usage, Day Mins, Monthly Charge And Overage Fee
* Data Plan, Data Usage, Day Mins, Monthly Charges And Overage Fee have a higher VIF value >4, indicating impact of multicollinearity in the model from these variables
* Cooks distance graph shows presence of influential points/outliers
* The normality graph shows somewhat linear relationship of predictors with logit

**TREATING THE OUTLIERS**

Identify the variables with influential points

**Dataset with influential points using cooks distance**

Churn AccountWeeks ContractRenewal DataPlan DataUsage

0: 29 Min. : 12.00 Min. :0.0000 Min. :0.000 Min. :0.0000

1:177 1st Qu.: 77.25 1st Qu.:1.0000 1st Qu.:0.000 1st Qu.:0.0000

Median :103.50 Median :1.0000 Median :0.000 Median :0.0000

Mean :105.25 Mean :0.7524 Mean :0.267 Mean :0.8284

3rd Qu.:131.75 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.6650

Max. :224.00 Max. :1.0000 Max. :1.000 Max. :4.5900

CustServCalls DayMins DayCalls MonthlyCharge

Min. :0.000 Min. : 0.0 Min. : 0.0 Min. : 14.00

1st Qu.:1.000 1st Qu.:132.2 1st Qu.: 87.0 1st Qu.: 42.00

Median :2.000 Median :187.4 Median :104.0 Median : 62.00

Mean :2.432 Mean :190.5 Mean :101.6 Mean : 58.41

3rd Qu.:4.000 3rd Qu.:246.7 3rd Qu.:117.8 3rd Qu.: 70.00

Max. :7.000 Max. :350.8 Max. :165.0 Max. :104.90

OverageFee RoamMins

Min. : 3.100 Min. : 3.90

1st Qu.: 8.275 1st Qu.: 8.30

Median :10.180 Median :10.10

Mean :10.146 Mean :10.33

3rd Qu.:11.893 3rd Qu.:12.60

Max. :17.530 Max. :17.80

**Dataset without influential points using cooks distance**

Churn AccountWeeks ContractRenewal DataPlan DataUsage

0:1966 Min. : 1.0 Min. :0.0000 Min. :0.000 Min. :0.0000

1: 161 1st Qu.: 74.0 1st Qu.:1.0000 1st Qu.:0.000 1st Qu.:0.0000

Median :101.0 Median :1.0000 Median :0.000 Median :0.0000

Mean :101.1 Mean :0.9158 Mean :0.276 Mean :0.8103

3rd Qu.:127.0 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.7700

Max. :232.0 Max. :1.0000 Max. :1.000 Max. :4.7300

CustServCalls DayMins DayCalls MonthlyCharge OverageFee

Min. :0.000 Min. : 0.0 Min. : 0 Min. : 15.70 Min. : 1.560

1st Qu.:1.000 1st Qu.:144.2 1st Qu.: 87 1st Qu.: 45.00 1st Qu.: 8.245

Median :1.000 Median :181.1 Median :100 Median : 53.20 Median :10.030

Mean :1.498 Mean :180.0 Mean :100 Mean : 56.19 Mean : 9.999

3rd Qu.:2.000 3rd Qu.:216.2 3rd Qu.:113 3rd Qu.: 65.90 3rd Qu.:11.690

Max. :9.000 Max. :346.8 Max. :158 Max. :111.30 Max. :18.190

RoamMins

Min. : 0.00

1st Qu.: 8.50

Median :10.30

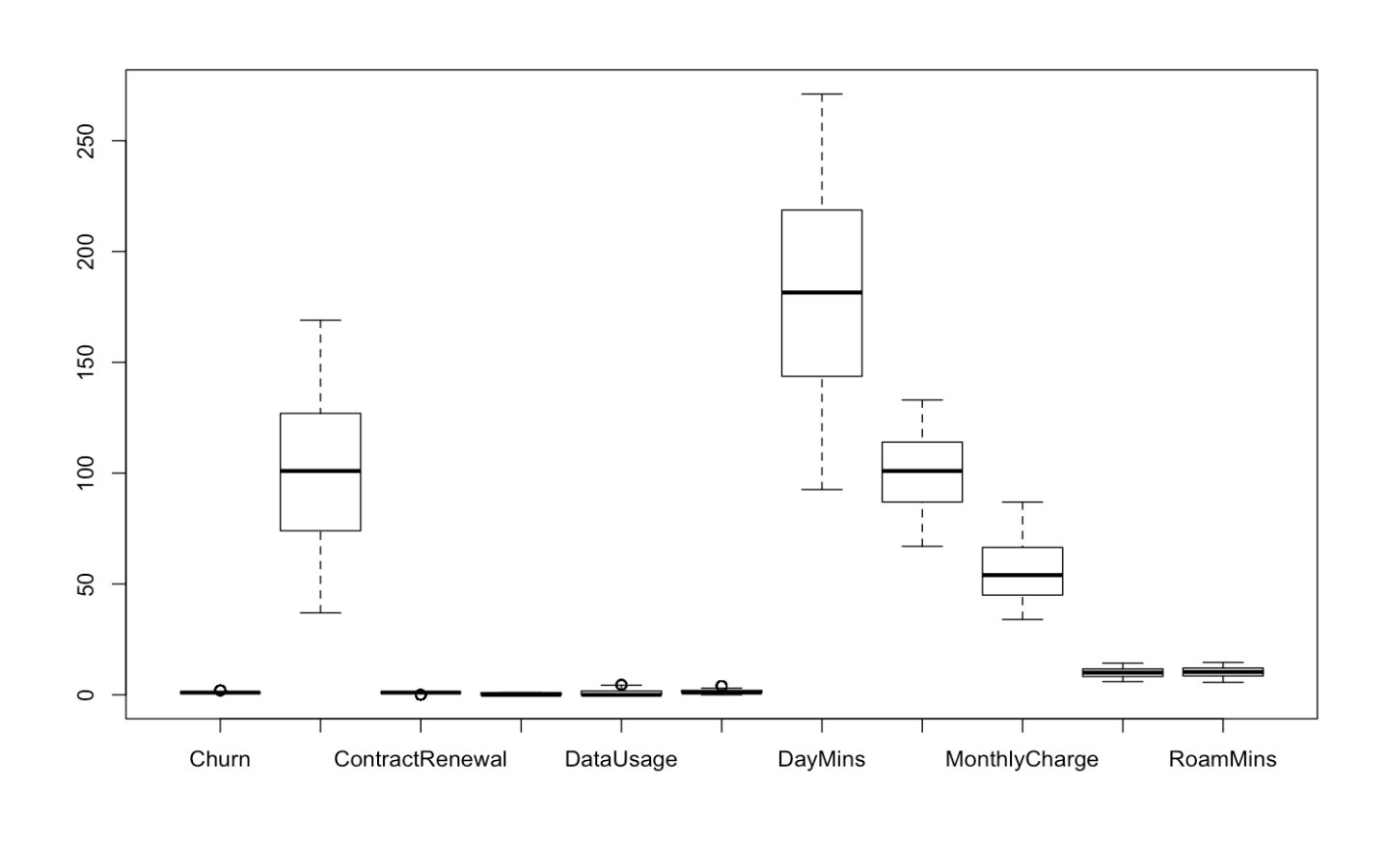
Mean :10.22

3rd Qu.:12.10

Max. :18.90

* Comparing the two datasets i.e. with and without the influential points we realise that **Account weeks, Custservcalls, Daymins, Daycalls, Monthly charge have a higher median value** compared to the dataset without influential points.
* Therefore we will treat these variables in both train and test data using flooring and ceiling method for outliers to be capped at 5 percentile and 95 percentile.

**BOXPLOT OF TRAIN DATA AFTER TREATING THE OUTLIERS**



**CREATING BEST FIT LOGISTIC REGRESSION MODEL**

* We will use the **step aic both method to arrive at best fit** logistic regression model

**blr\_step\_aic\_both(LRModel, details = TRUE)**

Candidate Terms:

1 . AccountWeeks

2 . ContractRenewal

3 . DataPlan

4 . DataUsage

5 . CustServCalls

6 . DayMins

7 . DayCalls

8 . MonthlyCharge

9 . OverageFee

10 . RoamMins

Step 0: AIC = 1932.419

Stepwise Summary

---------------------------------------------------------------

Variable Method AIC BIC Deviance

---------------------------------------------------------------

DayMins addition 1823.935 1835.445 1819.935

CustServCalls addition 1713.926 1731.190 1707.926

ContractRenewal addition 1597.954 1620.974 1589.954

DataPlan addition 1567.763 1596.537 1557.763

OverageFee addition 1538.383 1572.913 1526.383

RoamMins addition 1530.539 1570.823 1516.539

DayCalls addition 1527.269 1573.308 1511.269

**FINAL MODEL**

* As indicated above our final model will constitute only 7 of the 10 independent variables and will **exclude Data Usage, Monthly Charges and Account weeks** from our model

**final.model <- glm(Churn~DayMins+CustServCalls+ContractRenewal+DataPlan+OverageFee+RoamMins+DayCalls, data = train, family = binomial)**

Call:

glm(formula = Churn ~ DayMins + CustServCalls + ContractRenewal +

DataPlan + OverageFee + RoamMins + DayCalls, family = binomial,

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8516 -0.5182 -0.3580 -0.2175 3.0378

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.659324 0.661647 -10.065 < 2e-16 \*\*\*

DayMins 0.012699 0.001369 9.274 < 2e-16 \*\*\*

CustServCalls 0.529998 0.053666 9.876 < 2e-16 \*\*\*

ContractRenewal -1.888204 0.169805 -11.120 < 2e-16 \*\*\*

DataPlan -0.933698 0.172304 -5.419 6.00e-08 \*\*\*

OverageFee 0.159987 0.029016 5.514 3.51e-08 \*\*\*

RoamMins 0.073816 0.026669 2.768 0.00564 \*\*

DayCalls 0.009008 0.003570 2.523 0.01162 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

**Null deviance: 1930.4 on 2332 degrees of freedom**

**Residual deviance: 1560.2 on 2325 degrees of freedom**

**AIC: 1576.2**

Number of Fisher Scoring iterations: 5

**RSQUARE MC FADDEN : 0.1917628**

**VIF**

DayMins CustServCalls ContractRenewal DataPlan OverageFee

1.018585 1.053418 1.050347 1.012737 1.024754

RoamMins DayCalls

1.004691 1.011002

**COEFFICIENTS EXPONENTS**

(Intercept) DayMins CustServCalls ContractRenewal

0.001282013 1.012779806 1.698928319 0.151343355

DataPlan OverageFee RoamMins DayCalls

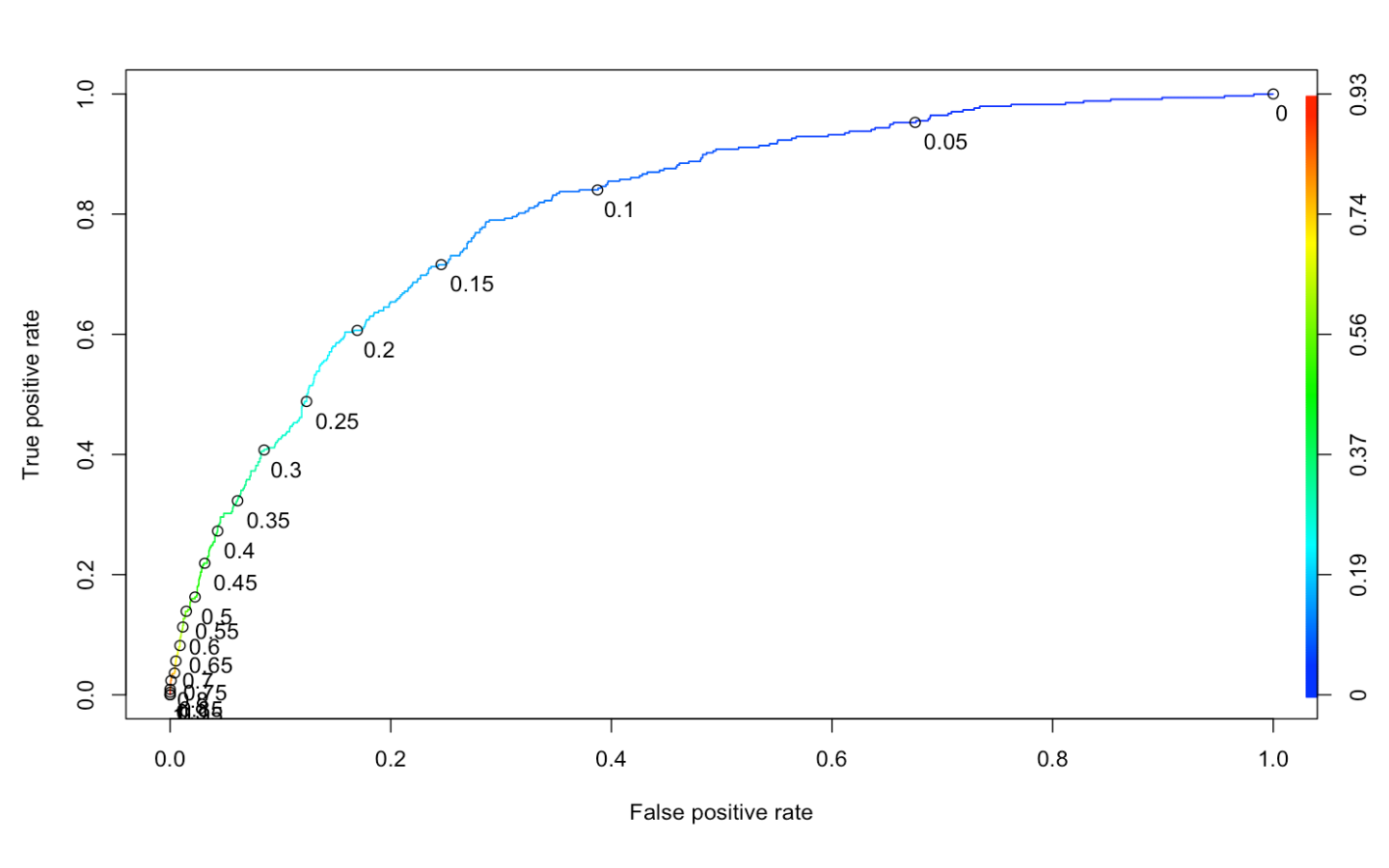
0.393097228 1.173495656 1.076609010 1.009049139

**INTERPRETING THE MODEL**

* The p values in coefficients table indicate **all variables are significant** in determining the Churn
* The VIF values are lower than threshold value of 4 indicating the **model is free from impact of multicollinearity**
* Each one-unit change in **Daymins** will increase the log odds of Churn =1. OR Customer leaving by 0.012699 OR increase the probability of Churn = 1 , by (odds / (1 + odds) = 1.012779806 / (1.012779806 + 1) = 0.50) = 50%
* The difference between Null deviance and Residual deviance tells us that the model is a good fit. Greater the difference better the model. W**e have a difference of 370 points in our model**
  + Null deviance is the value when we only have intercept in the equation with no variables and Residual deviance is the value when we are taking all the variables into account.
* The  AIC an estimator of out-of-sample prediction error and given n number of models for the data, AIC estimates the quality of each model, relative to other models. Lower the AIC its better. For our model AIC is 1576, which is a decent decline from step 0 AIC of 1932 in step both modelling to identify best fit model.
* Rsquare Mcfadden - should have values between 0.2 and 0.4 is valid model, Below 0.2 is an Underfit model and Over 0.4 is an Overfit model – **Our value of 0.19 is decent and passes as a valid model**

**IDENTIFYING THE RIGHT CUT OFF FOR CONFUSION MATRIX IN OUR BEST FIT MODEL**

**ROC CURVE**

****

* From the above graph, if we take a benchmark of 30% False Positive rate, then the **ideal cut of rate to balance and maximize sensitivity and specificity is 0.13**
* Therefore we will modify the cut off rate at 13% to report our model evaluation measures

**MODEL EVALUATION MEASURES FOR LOGISTIC REGRESSION MODEL**

**TRAIN**  **TEST**

Evaluation measures for Test dataset

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 623 36

1 232 109

Accuracy : 0.732

95% CI : (0.7034, 0.7592)

No Information Rate : 0.855

P-Value [Acc > NIR] : 1

Kappa : 0.3077

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.7517

Specificity : 0.7287

Pos Pred Value : 0.3196

Neg Pred Value : 0.9454

Prevalence : 0.1450

Detection Rate : 0.1090

Detection Prevalence : 0.3410

Balanced Accuracy : 0.7402

**AUC : 0.780**

Gini Index : 0.506

KS : 0.46

'Positive' Class : 1

Evaluation measures for Train dataset

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 1432 76

1 563 262

Accuracy : 0.7261

95% CI : (0.7075, 0.7441)

No Information Rate : 0.8551

P-Value [Acc > NIR] : 1

Kappa : 0.3084

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.7751

Specificity : 0.7178

Pos Pred Value : 0.3176

Neg Pred Value : 0.9496

Prevalence : 0.1449

Detection Rate : 0.1123

Detection Prevalence : 0.3536

Balanced Accuracy : 0.7465

AUC : 0.808

Gini Index : 0.508

KS : 0.48

'Positive' Class : 1

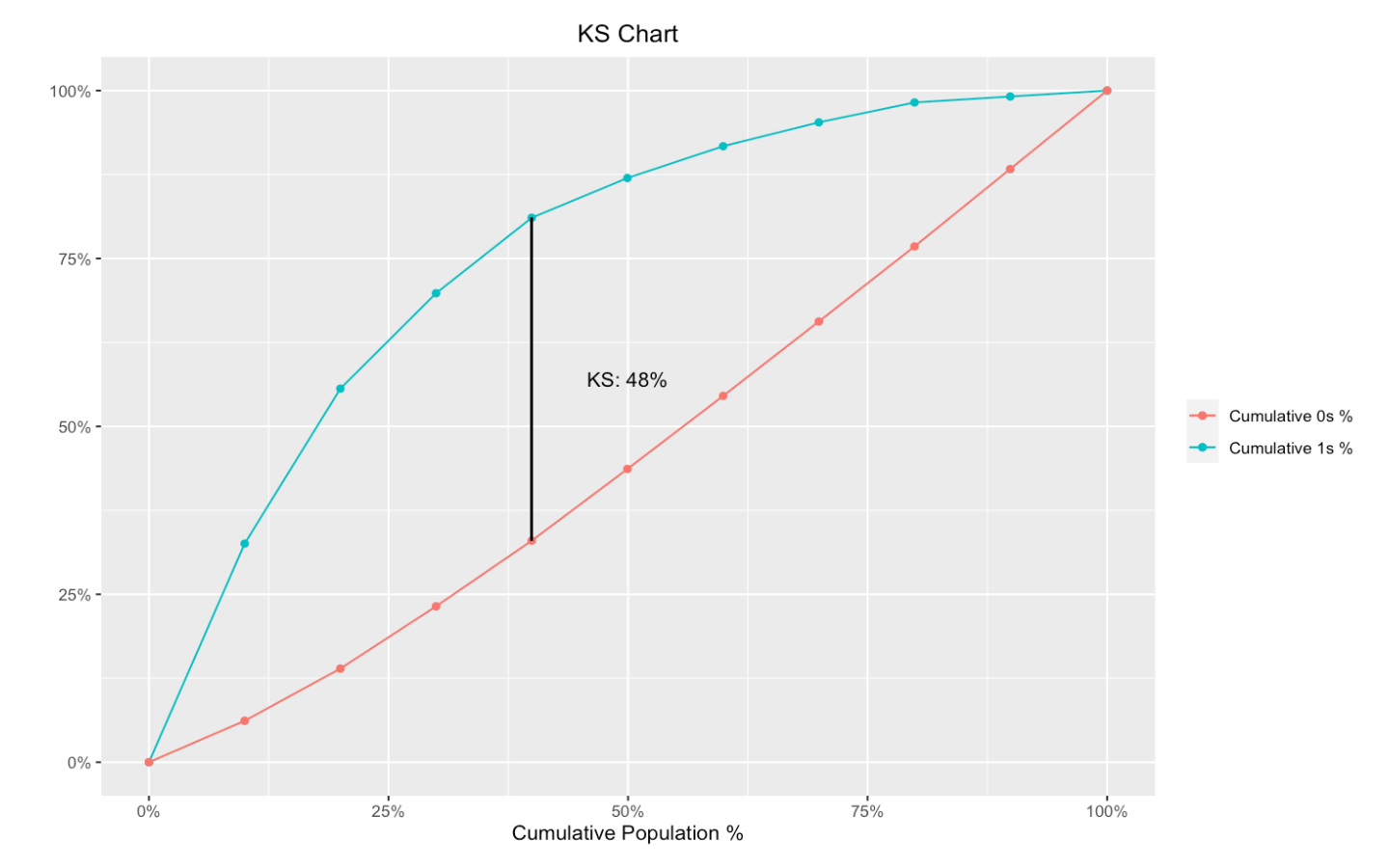
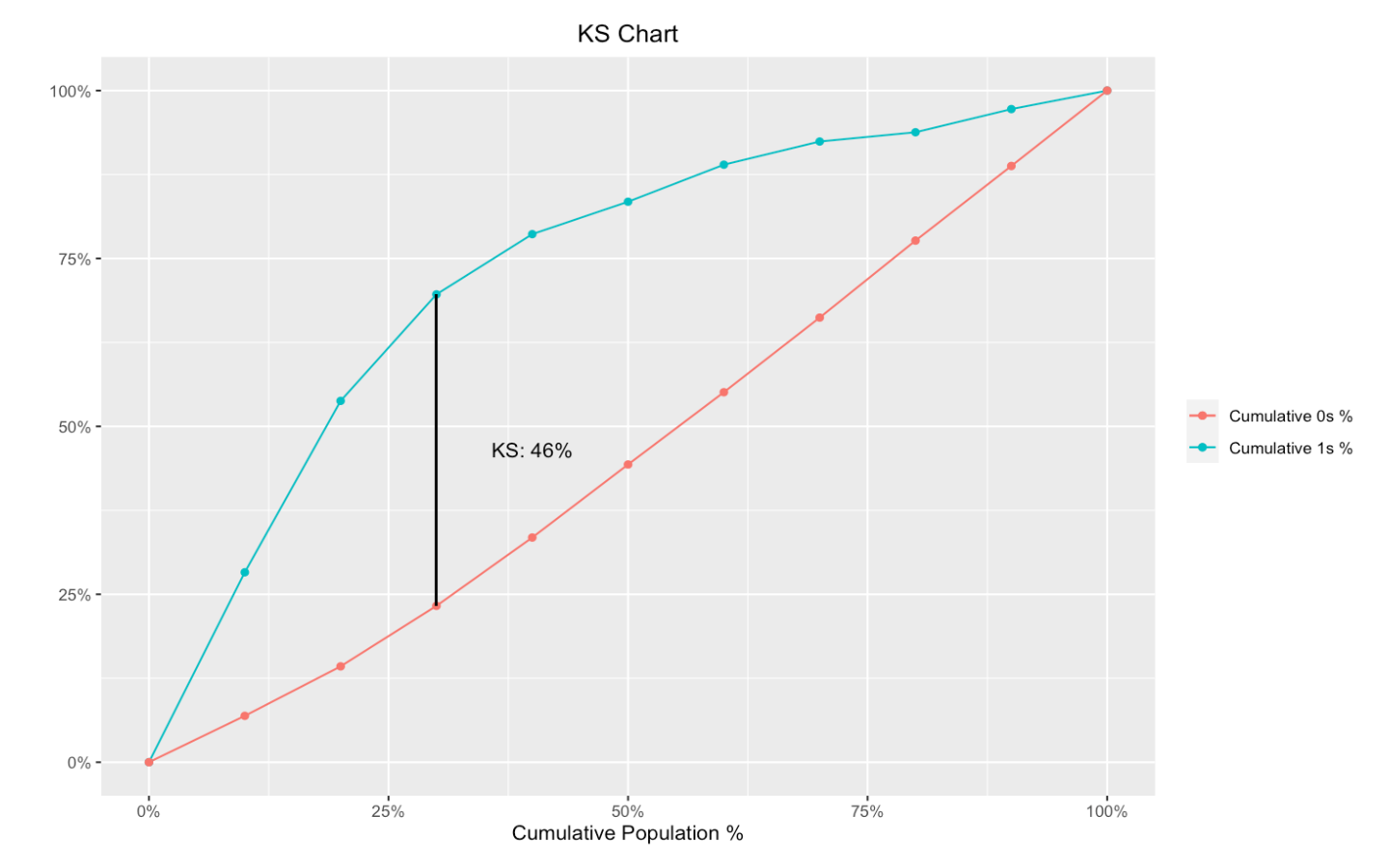
**DEFINING THE ABOVE EVALUATION MEASURES**

* **Sensitivity** - Total no correct predictions of 1 out of total predictions of 1
* **Specificity** - Total no correct predictions of 0 out of total predictions of o
* **Accuracy** - Rati0 of correct predictions to total observations
* **Error Rate** - 1- Accuracy
* **AUC**  - AUC gives the rate of successful classification by the logistic model.
* **Gini Index** - This measures how much better the model is performing compared to random selection.
* **KS**  - Degree of separation among positive and negative distributions in the decile table. Higher the better

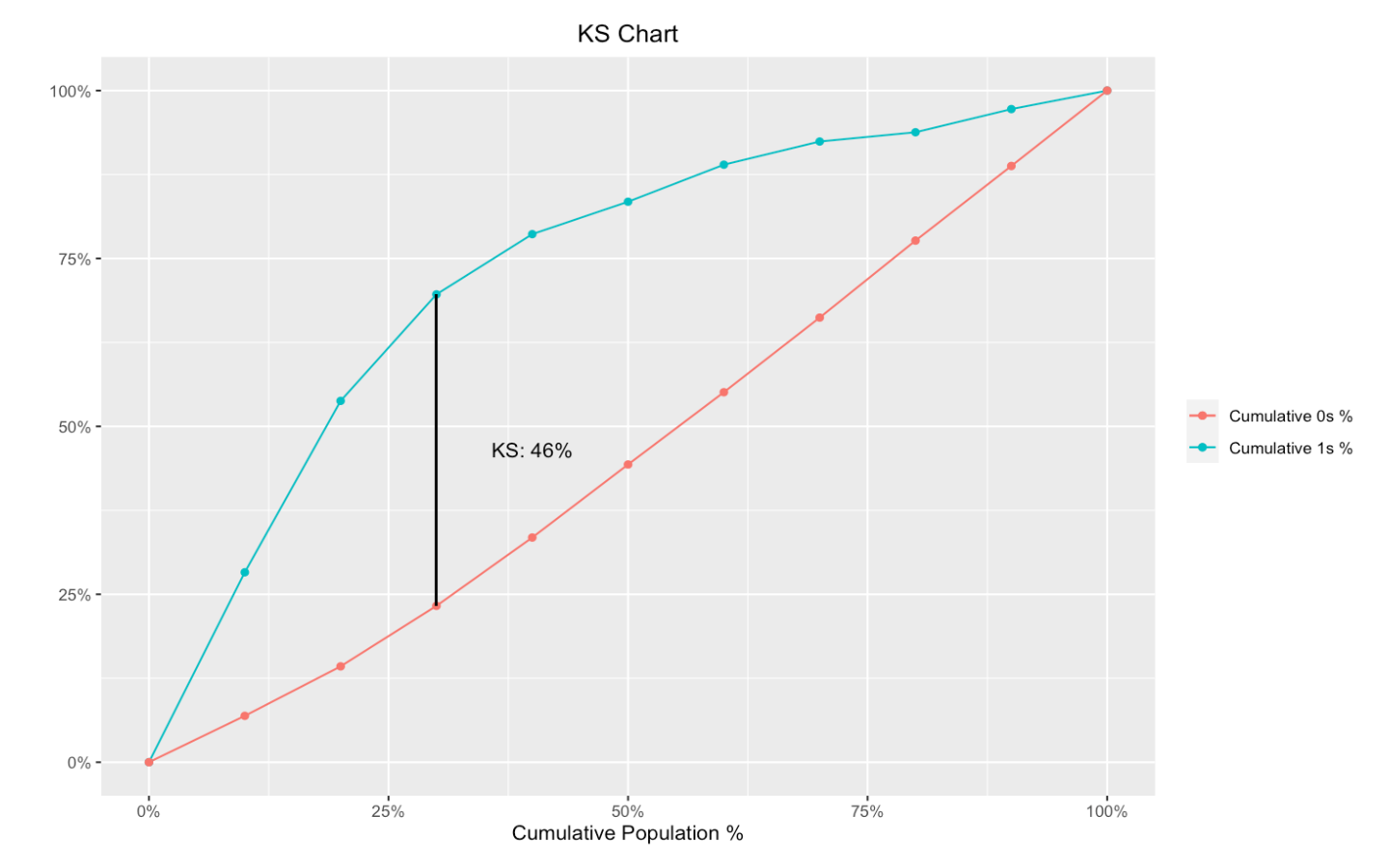
**INTERPRETING THE ABOVE RESULTS**

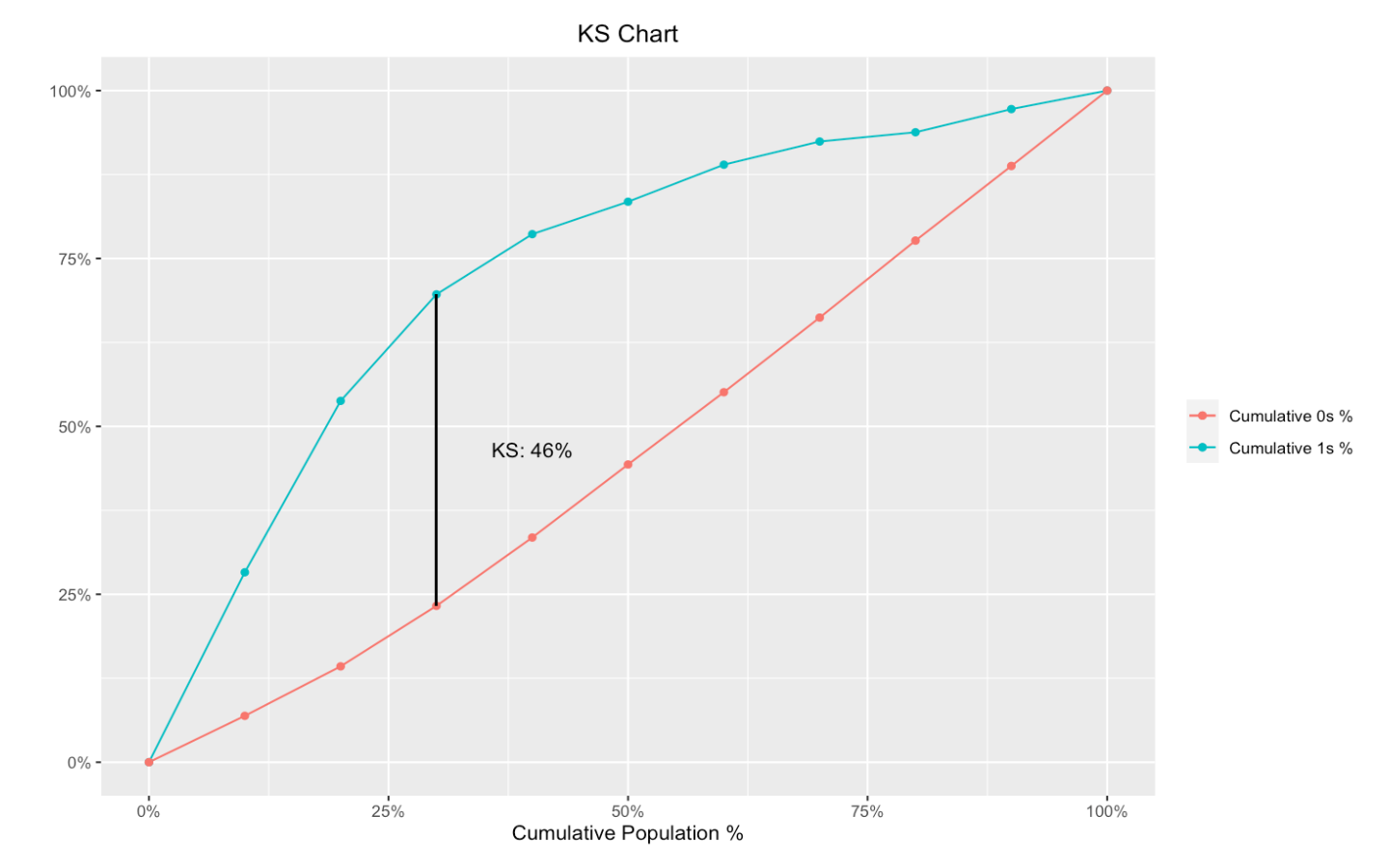
* AUC of 80% for Train and 78% for Test Model indicates a decent fit model
* Sensitivity, Specificity, Accuracy , AUC ,Gini Index , KS are all with in 10% difference for Test and Train data indicating a valid model
* Our test Gini index of 50% indicates a good model as typical values range from 40% to 60% for a good model
* Our test KS of 46% is higher then the benchmark value and hence it indicates a decent model
* Our test AUC of 78% indicates that model is giving successful classification 78% of times.

**KS chart for Train Data set**



**KS chart for Test Data set**





Model Building – KNN

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions like eucilidean distance)

**BUILDING KNN MODEL**

**KEY STEPS**

* Normalise all continuous variables, so that they are measured on similar scale to avoid scale bias in the model
* Split the data in train and test – 70:30 ratio
* Run knn model
* Identify the k value (no of nearest neighbour cases/observations to target variable) which maximises accuracy or any other evaluation measure(for eg sensitivity, specificity )of interest, we chose accuracy in our case
* In our case k =13 gives maximum accuracy
* Check evaluation measures on Train and Test

**SUMMARISING THE EVALUATION MEASURES**

**Evaluation Measures on Test set**

Confusion Matrix

pred.knn 0 1

0 846 106

1 9 39

Accuracy : 0.885

Sensitivity : 0.2689655

Specificity : 0.9894737

Error. : 0.115

**Evaluation Measures on Train set**

Confusion Matrix

pred.knn.train 0 1

0 1986 198

1 9 140

Accuracy : 0.911273

Sensitivity : 0.4142012

Specificity : 0.9954887

Error. : 0.08872696

* The accuracy on Test set is 88.5% and difference between Test and Train is under 10%, indicating a valid model
* However Sensitivity for the model is quite low at 26% for the Test Dataset, hence while the model is good at explaining who will not cancel, it is not so good at explaining who will cancel the service

Model Building – Naïve Bayes

Naïve Bayes - It is a classification technique based on Bayes’ Theorem

**BUILDING THE NAÏVE BAYES MODEL**

**ASSUMPTIONS**

* It considers Presence of a particular feature /independent variable in a class to be unrelated to the presence of any other feature/independent variable
* Considers categorical Independent variables, however also includes continuous Independent variables
* For continuous Independent variables it assumes they are normally distributed

**Whether Naïve Bayes is applicable in our model**

1. We have 10 independent variables out of which only two are categorical, while Naïve Bayes works best with only categorical independent variables, it also accommodates the continuous /numerical independent variables however numerical variable must be binned and converted to categorical
2. We have many independent variables which are correlated as seen during EDA phase - hence it voids the assumption of independence of predictor variables , however the model will take care of this assumption
3. Some of our continuous independent variables do not follow normal distribution, but we will take care of this assumption by converting all variables to categorical

**KEY STEPS**

* Convert numerical variables to categorical
* Split data in train and test in 70:30 ratio
* Run the model on train dataset
* Check model evaluation measures on train and test dataset

**SUMMARISING THE EVALUATION MEASURES**

**Evaluation Measures on Train set**

Confusion Matrix

**predNB.train 0 1**

**0 1942 267**

**1 53 71**

Accuracy : 0.8628375

Sensitivity : 0.2100592

Specificity : 0.9734336

Error. : 0.1371625

**Evaluation Measures on Test set**

Confusion Matrix

**predNB 0 1**

**0 845 136**

**1 10 9**

Accuracy : 0.854

Sensitivity : 0.06206897

Specificity : 0.9883041

Error. : 0.146

* The **accuracy** on Test set is 85.4% and difference between Test and Train is very low i.e. under 10%, hence indicating a valid model
* However Sensitivity for the model is quite low at only 6% for the Test Dataset, hence while the model is good at explaining who will not cancel, it is not so good at explaining who will cancel the service.

CHOOSING THE BEST FIT MODEL

**EVALUATION MEASURES ON TEST SET - KNN**

Confusion Matrix

pred.knn 0 1

0 846 106

1 9 39

Accuracy : 0.885

Sensitivity : 0.2689655

Specificity : 0.9894737

Error. : 0.115

**EVALUATION MEASURES ON TEST SET – NAÏVE BAYES**

Confusion Matrix

**predNB 0 1**

**0 845 136**

**1 10 9**

Accuracy : 0.854

Sensitivity : 0.06206897

Specificity : 0.9883041

Error. : 0.146

**Choosing the best fit model**

* Accuracy is highest for KNN followed by Naïve Bayes and Logistic Regression
* However if we observe sensitivity, which is a key measure that indicates how well is the model predicting customers who are cancelling among total customers who actually cancelled , it is quite low for KNN and Naïve Bayes.
* Sensitivity is highest for our logistic regression model
* Its AUC also stands at 78% indicating a decent model
* Since **logistic regression model is balanced in terms of accuracy,sensitivity and specificity, we will recommend to go ahead with the logistic regression Model**

**EVALUATION MEASURES FOR TEST DATASET – LOGISTIC REGRESSION**

Confusion Matrix

Reference

Prediction 0 1

0 623 36

1 232 109

Accuracy : 0.732

Sensitivity : 0.7517

Specificity : 0.7287

Balanced Accuracy : 0.7402

**AUC : 0.780**

Gini Index : 0.506

KS : 0.46