**SUMMARY**

This report covers all the data analysis based on the excel sheet provided to us “Cellphone.xlsx”. Based on the same, we have exported the sheet to “R Studio” and further have analyzed the data set based on the question requirements with respect to the Predictive Modeling course.

The following points have been covered

* Detailed Exploratory Data Analysis report of the dataset along with the missing value treatment
* Multicollinearity check and summarization of problem statement for business stakeholders
* Logistic Regression Model: creation and interpretation of the results
* Comparing the model performances using confusion matrix
* Actionable Insights for the business stakeholders

|  |  |
| --- | --- |
| **Points to be covered** | **Page No** |
| 1) Exploratory Data Analysis a) Basic data summary, Univariate, Bivariate analysis, graphs | 3-15 |
| 1) Exploratory Data Analysis b) Check for Outliers and missing values and check the summary of the dataset | 5, 7-9 |
| 1) Exploratory Data Analysis c) Check for Multicollinearity - Plot the graph based on Multicollinearity | 15 |
| 1) Exploratory Data Analysis d) Interpreting the business problem and sharing the observations | 6, 15 |
| 2) Applying Supervised Machine Learning Techniques (Test & Train) a) Applying Logistic Regression and Interpret the Regression model output. | 16-19 |
| 2) Applying Supervised Machine Learning Techniques (Test & Train) b) Applying Logistic Regression with relevant variables and Interpret the Regression model output. | 20-29 |
| 3) Model Performance Measures (Test & Train) a) Confusion matrix interpretation for all models | 17, 20-21, 29 |
| 3) Model Performance Measures (Test & Train) b) Interpretation of other Model Performance Measures for logistic <KS, AUC, GINI> | 18,22-23,28-29 |
| 3) Model Performance Measures (Test & Train) c) Remarks on Model validation exercise <Which model performed the best> | 30 |
| 4) Actionable Insights and Recommendations | 30-31 |

**EXPLORATORY DATA ANALYSIS**

Dataset: Cellphone

Objective: A model to be built which can predict if the customer would churn ornot; that is cancel their service in future or not.

We would need to do EDA to understand the Data Types, Missing Values, Univariate and Bivariate analysis

Packages that need to be installed for this assignment

library(readxl)

> library(DataExplorer)

> library(corrplot)

> library(car)

> library(tidyverse)

> library(caret)

> library(lmtest)

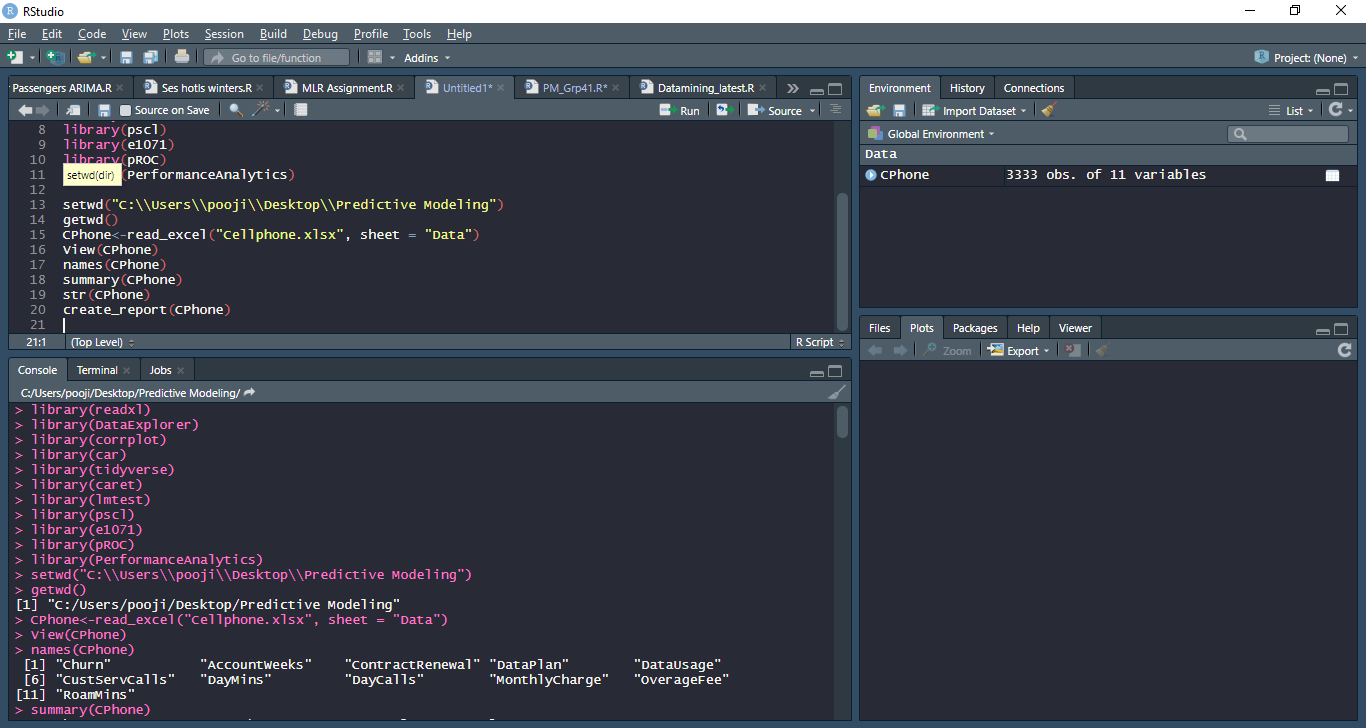
> library(pscl)

> library(e1071)

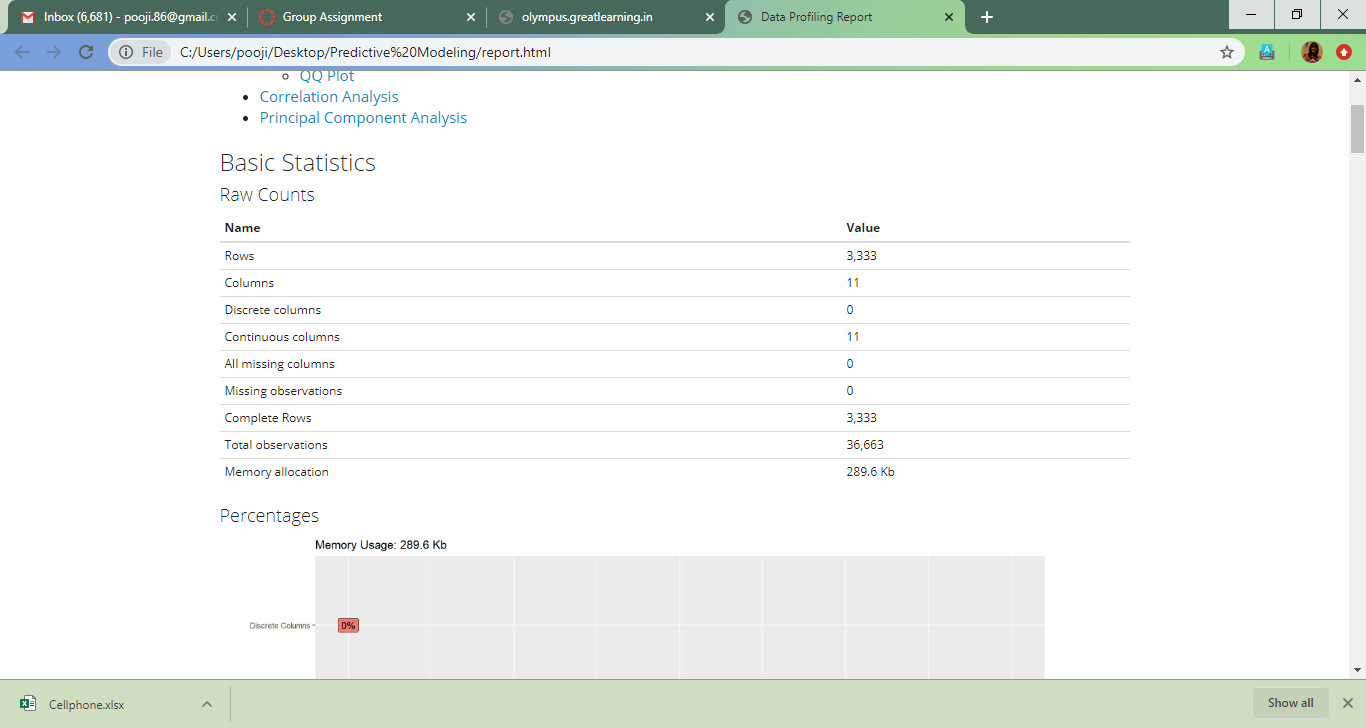
> library(pROC)

> library(PerformanceAnalytics)

The file is imported to R as the first step and renamed as “CPhone” for running the code



This report generation provides based out of the “Data Explorer” Package and automatically creates reports in a html browser



This report however says that all the columns are continuous in nature and there is an absence of discrete variables. Let’s first correct that try again

col<-c("Churn", "ContractRenewal", "DataPlan")

>CPhone[col]<-lapply(CPhone[col], factor)

>str(CPhone)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 3333 obs. of 11 variables:

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

$ AccountWeeks : num 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 2.7 3.7 0 0 0 0 2.03 0 0.19 3.02 ...

$ CustServCalls : num 1 1 0 2 3 0 3 0 1 0 ...

$ DayMins : num 265 162 243 299 167 ...

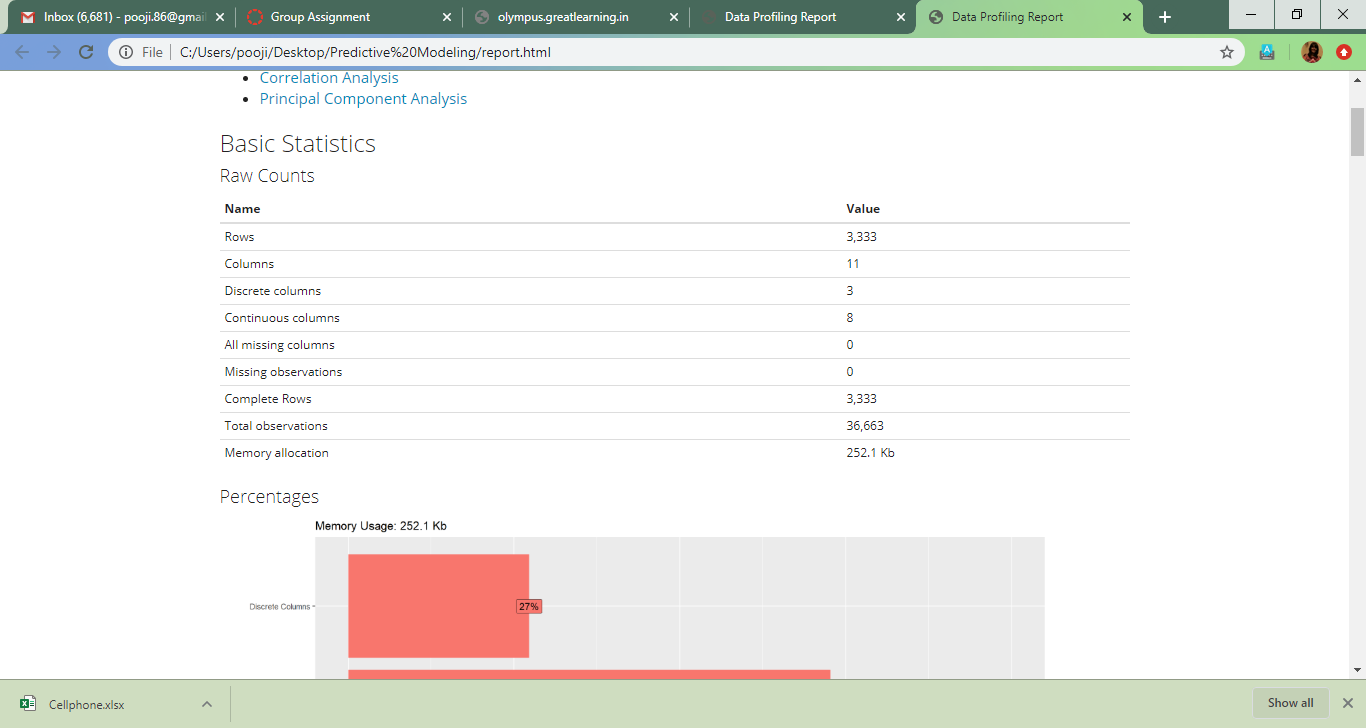
$ DayCalls : num 110 123 114 71 113 98 88 79 97 84 ...

$ MonthlyCharge : num 89 82 52 57 41 57 87.3 36 63.9 93.2 ...

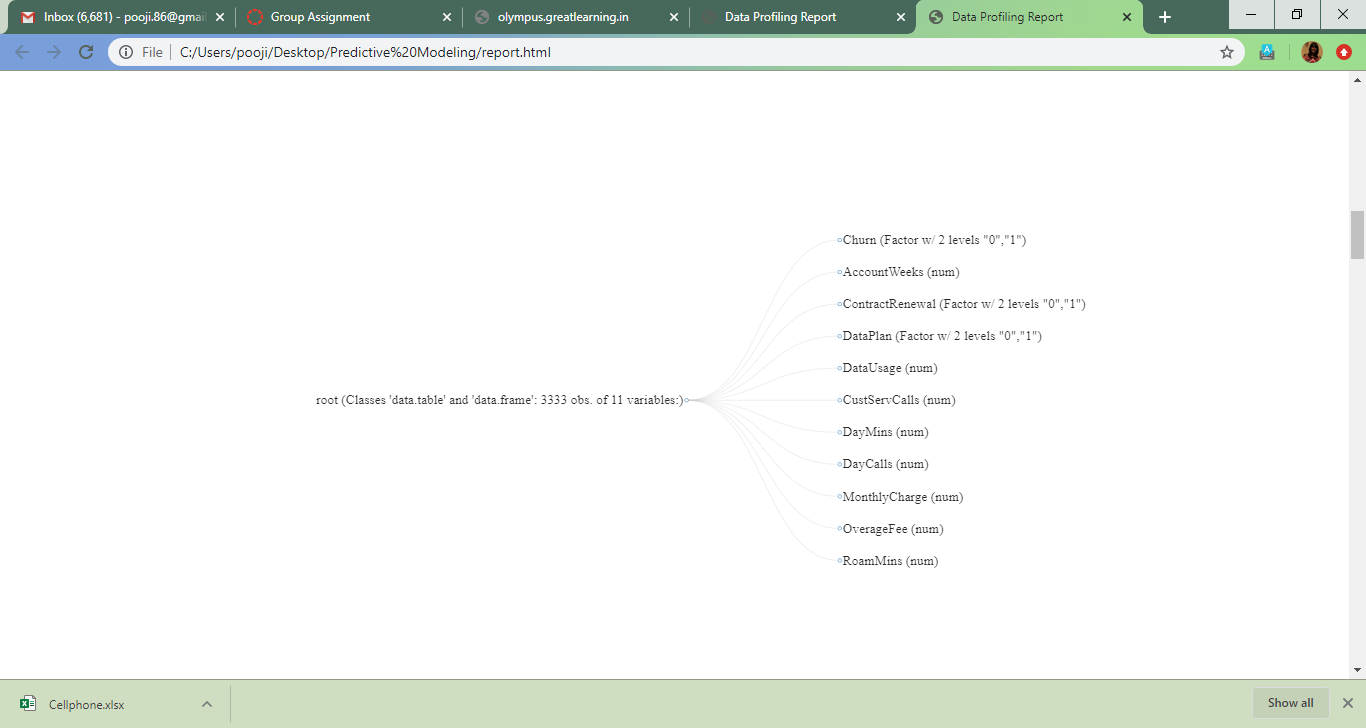
$ OverageFee : num 9.87 9.78 6.06 3.1 7.42 ...

$ RoamMins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...

Now that the required variables have converted to factors, lets try generating the report again



1. As we can see, there are in total 3,333 rows, 8 columns with continuous variables and 3 columns with discrete/ binary variables
2. Discrete variables are Churn, ContractRenewal and DataPlan
3. Continuous variables are AccountWeeks, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge, OverageFee and RoamMins



Churn AccountWeeksContractRenewalDataPlanDataUsageCustServCalls

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

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

Median :101.0 Median :0.0000 Median :1.000

Mean :101.1 Mean :0.8165 Mean :1.563

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

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

DayMinsDayCallsMonthlyChargeOverageFeeRoamMins

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

1st Qu.:143.7 1st Qu.: 87.0 1st Qu.: 45.00 1st Qu.: 8.33 1st Qu.: 8.50

Median :179.4 Median :101.0 Median : 53.50 Median :10.07 Median :10.30

Mean :179.8 Mean :100.4 Mean : 56.31 Mean :10.05 Mean :10.24

3rd Qu.:216.4 3rd Qu.:114.0 3rd Qu.: 66.20 3rd Qu.:11.77 3rd Qu.:12.10

Max. :350.8 Max. :165.0 Max. :111.30 Max. :18.19 Max. :20.00

Missing Variables Analysis

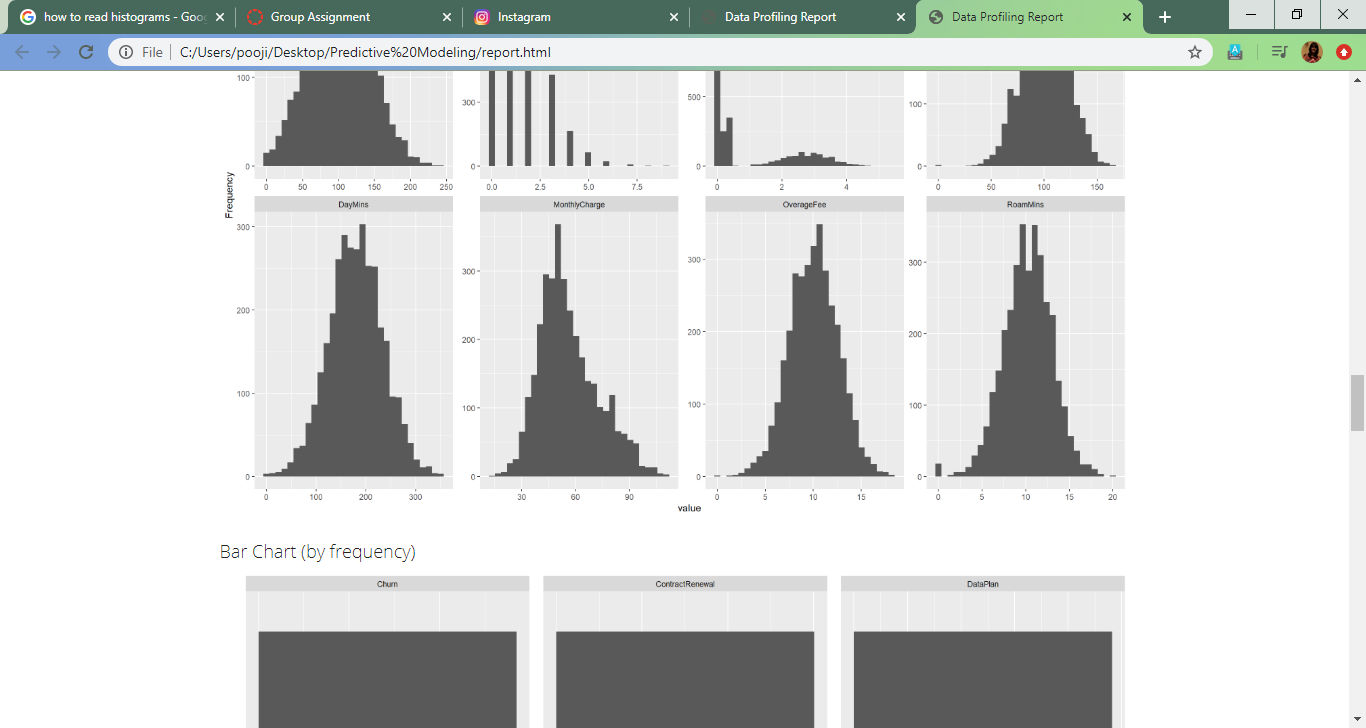


**Basic EDA Inference:**

* No Missing Values
* No Negative Values
* 3 variables converted from Numeric to Factor
* Churn rate is at 14% which is our Dependent Variable
* Monthly charge could have outliers, but we shall confirm that via Univariate analysis

**Univariate Analysis- Histogram**

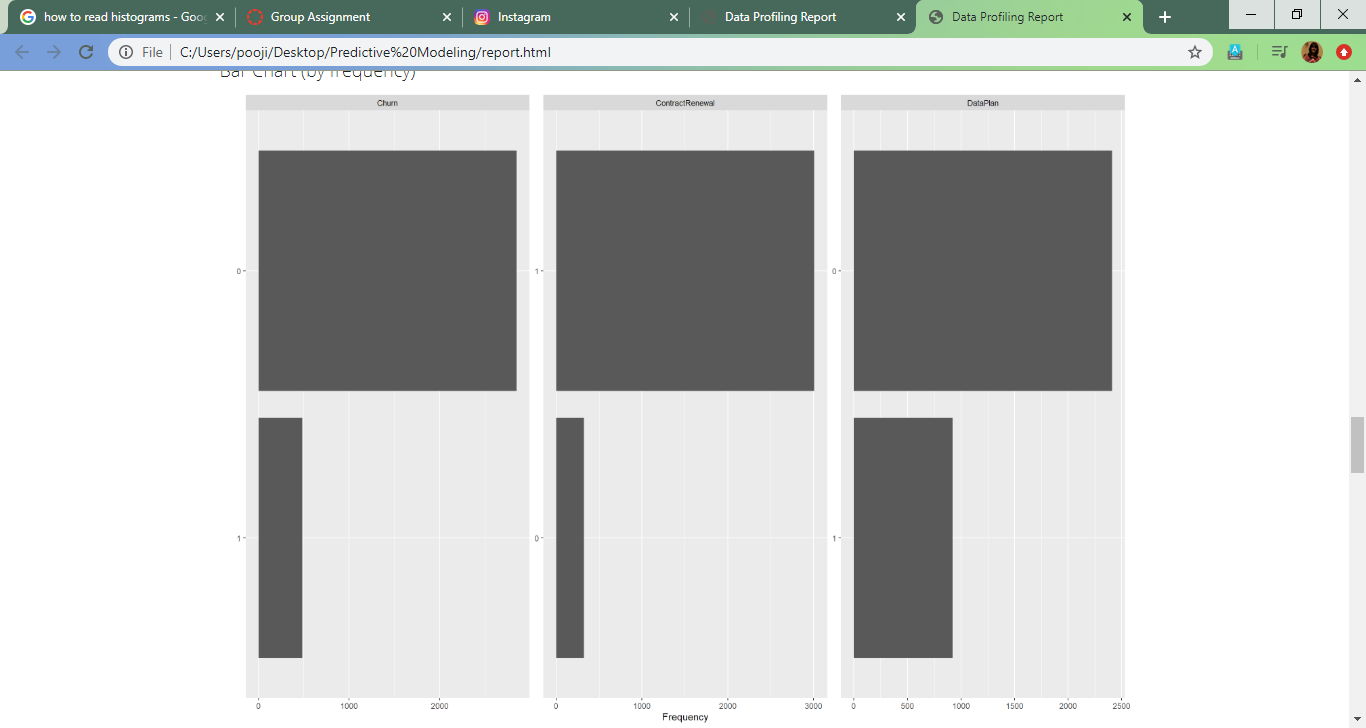




* Frequency of Active Account weeks per customer peaks at 100 weeks. Both the median and the mean stands at 101.1.
* The frequency of number of Customer service call are 1.5 as per the data
* But there is a high concentration of occurance of Data usage less that 1 GB
* Also, the repetition number of daytime calls peak at around 100

**Univariate Analysis- Barchart**

Barchart is created for Discrete variables and collates the frequency of the various factors



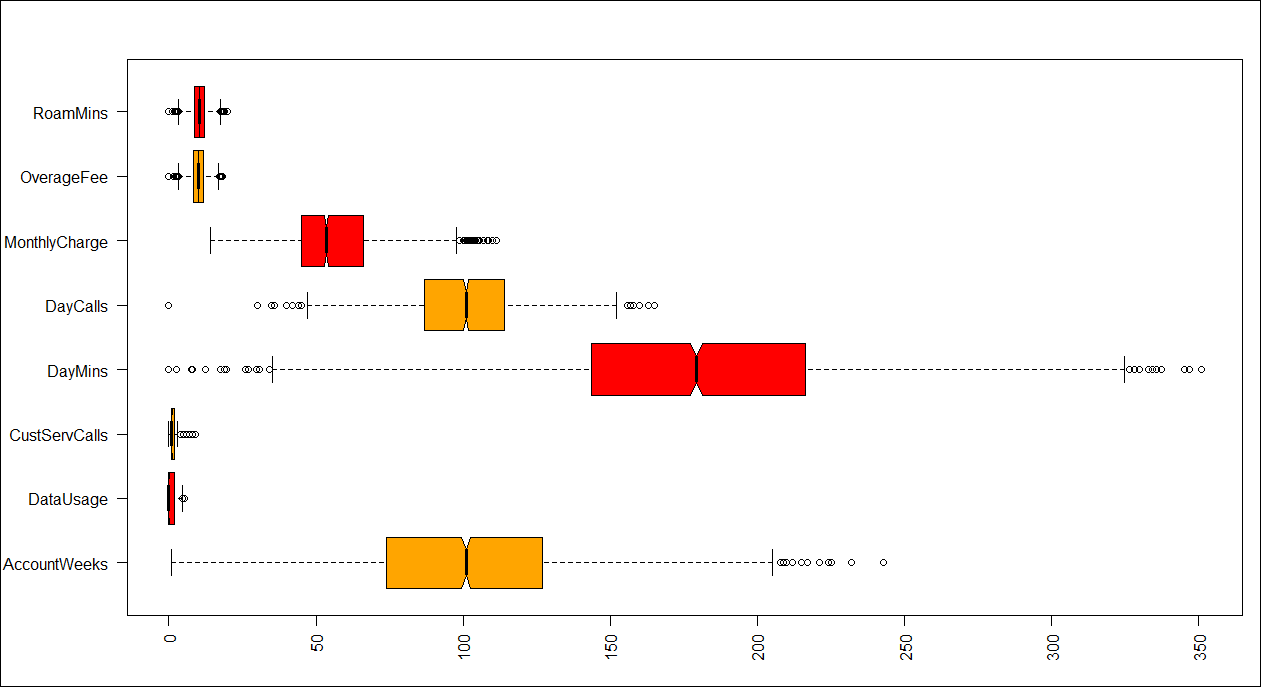
* As we can see Customer retention is much more that Churned customers
* Also, there is a higher rate of customers not having a data plan
* Customers who have renewed the contract recently also are at a higher rate and do not fall under churn

**Univariate Analysis- Box Plots**

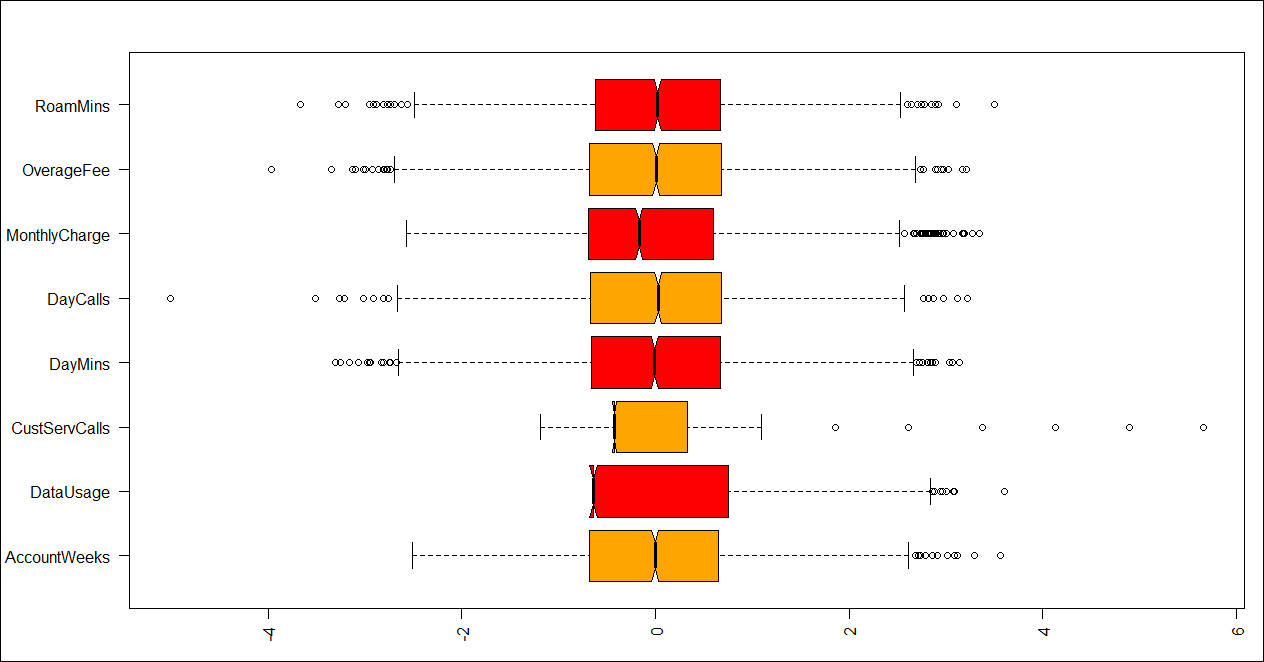
par(mar=c(5,6,4,1)+.8) #To fit the labels inside the chart

>boxplot(CPhone[c(2,5:11)],

+ names= c("AccountWeeks","DataUsage","CustServCalls","DayMins","DayCalls","MonthlyCharge","OverageFee","RoamMins"),las=2,col=c("orange","red"),border="black",horizontal = TRUE,notch = TRUE)



* Scaled Boxplot for the better presentation
* boxplot(scale(CPhone[c(2,5:11)]),
* + names= c("AccountWeeks","DataUsage","CustServCalls","DayMins","DayCalls","MonthlyCharge","OverageFee","RoamMins"),las=2,col=c("orange","red"),border="black",horizontal = TRUE,notch = TRUE)

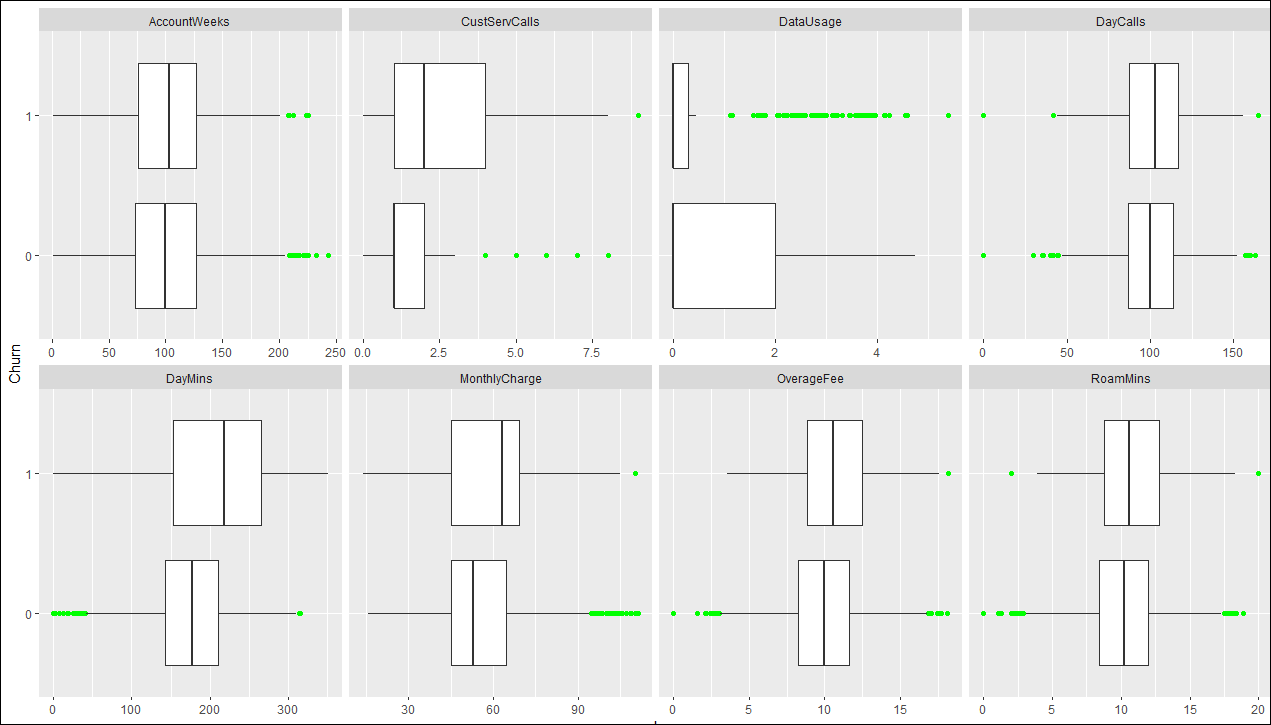


* We can see that the Data Usage is skewed towards the left, that us close to 0
* Inspite of the presence of outliers, we are not providing it any treatment as they are continuous variables and the presence of them might not disturb the data

**Bivariate Analysis- Box Plots**

We can also look the Boxplot against the Dependant variable, Churn to have better insight

plot\_boxplot(CPhone, by="Churn", geom\_boxplot\_args = list("outlier.color" = "green"))



Here we can see that the Data used is very minimally used by the customers who Churn. This could be an important insight when we do Logistic Regression. Also, the Average Daytime minutes in a month looks to be on a higher side for the Churn customers

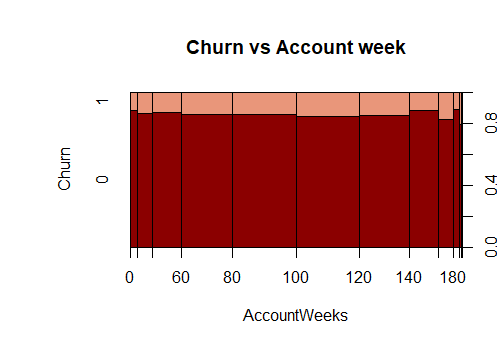
**Bivariate Analysis- Barplot**

Barplots can used to analyse two variables, when the thedependant variable is Categorical. Since, Churn is our dependant variables, we shall be using Bar plots to compare with the rest of variable

1. **Churn vs Account Weeks**

>plot(Churn~AccountWeeks,data = CPhone,col=colors()[100:102],

+ main="Churn vs Account week")



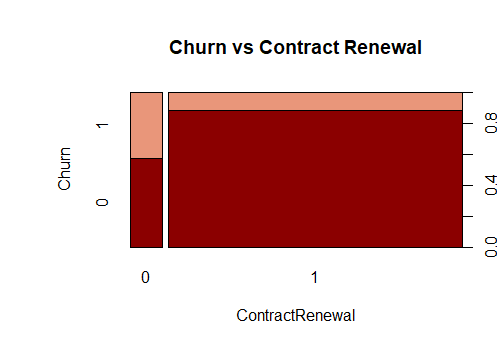
As we can see, Most customers fall under 80-120 weeks category, and not much insight on its effect on churn is see

1. **Churn Vs Contract Renewal**

>plot(Churn~ContractRenewal,data = CPhone,col=colors()[100:102],

+ main="Churn vs Contract Renewal")

>

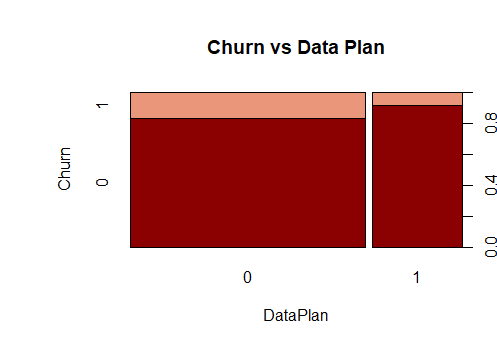


Customer who don’t renew contract are almost split between Churn and Not Churn. We have to see in further analysis is this variable has any effect on Churn.

1. **Churn Vs Data Plan**

plot(Churn~DataPlan,data = CPhone,col=colors()[100:102],

+ main="Churn vs Data Plan")



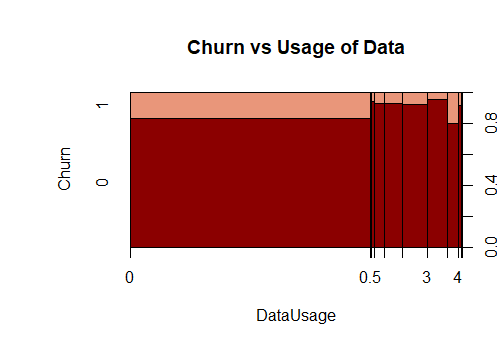
A large number of customers who have churned do not have a data plan.

This is clear in this chart.

1. **Churn Vs Data usage**

>plot(Churn~DataUsage,data = CPhone,col=colors()[100:102],

+ main="Churn vs Usage of Data")

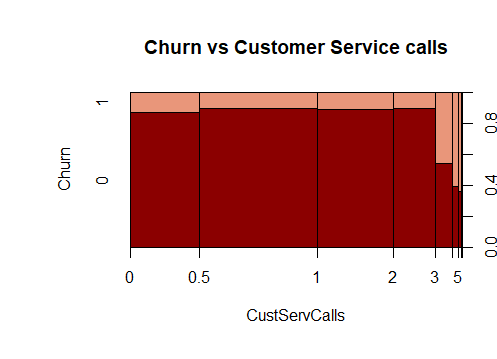


Again, here we can see that a large set of cutomers who do not use data also are retained customers. This is also be because of them not having any Data Plan in the first place

1. **Churn Vs Customer Service Calls**

>plot(Churn~CustServCalls,data = CPhone,col=colors()[100:102],

+ main="Churn vs Customer Service calls")

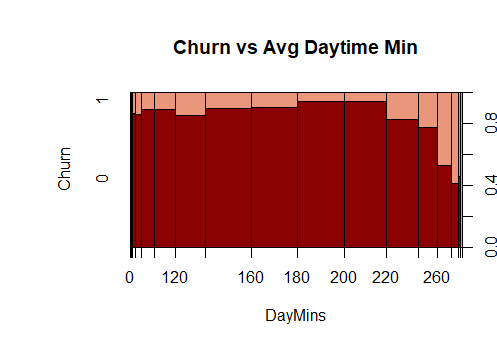


It is very evident here that the higher the number calls place, the higher the Churn rate. This could be a significant variable

1. **Churn Vs Average Daytime Min**

>plot(Churn~DayMins,data = CPhone,col=colors()[100:102],

+ main="Churn vs Avg Daytime Min")

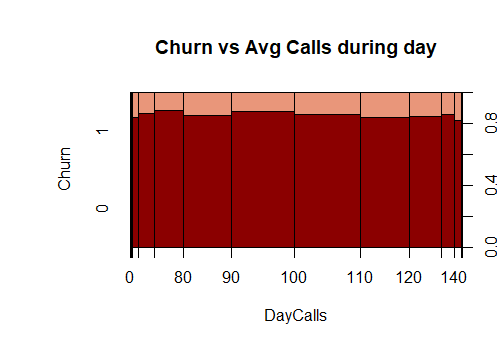


Interesting to note a pattern here, the higher time in Average Day calls, we can see the Churn rate going higher.

1. **Churn Vs Avg Calls during the day**

>plot(Churn~DayCalls,data = CPhone,col=colors()[100:102],

+ main="Churn vs Avg Calls during day")

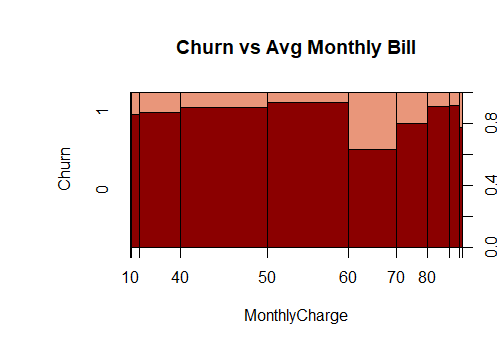


Not much of an insight from this

1. **Churn Vs Monthly Charge**

>plot(Churn~MonthlyCharge,data = CPhone,col=colors()[100:102],

+ main="Churn vs Avg Monthly Bill")

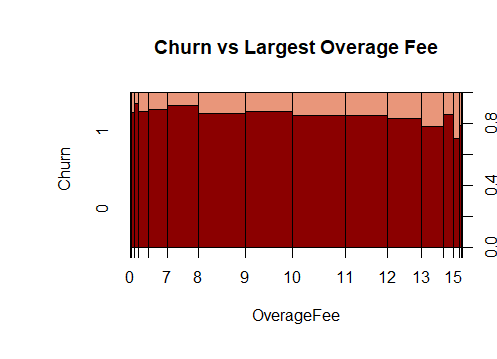


The highest Churn rate is in between 60-70 range of Average Monthly Bill

1. **Churn Vs Overage Fee**

>plot(Churn~OverageFee,data = CPhone,col=colors()[100:102],

+ main="Churn vs Largest Overage Fee")

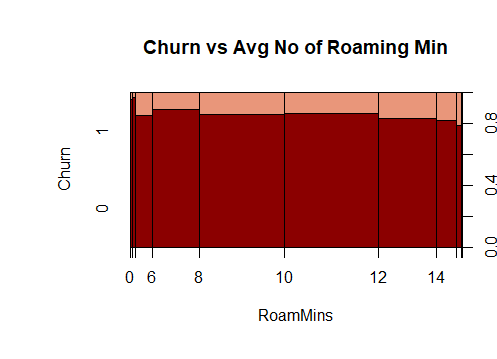


Overage fee doesn’t seem to affect the Churn rate

1. **Churn Vs Avg Roaming Min**

>plot(Churn~RoamMins,data = CPhone,col=colors()[100:102],

+ main="Churn vs Avg No of Roaming Min")



Though not a exponential decrease, but the Churn rate seems to slide down as the Average Roaming Min increases

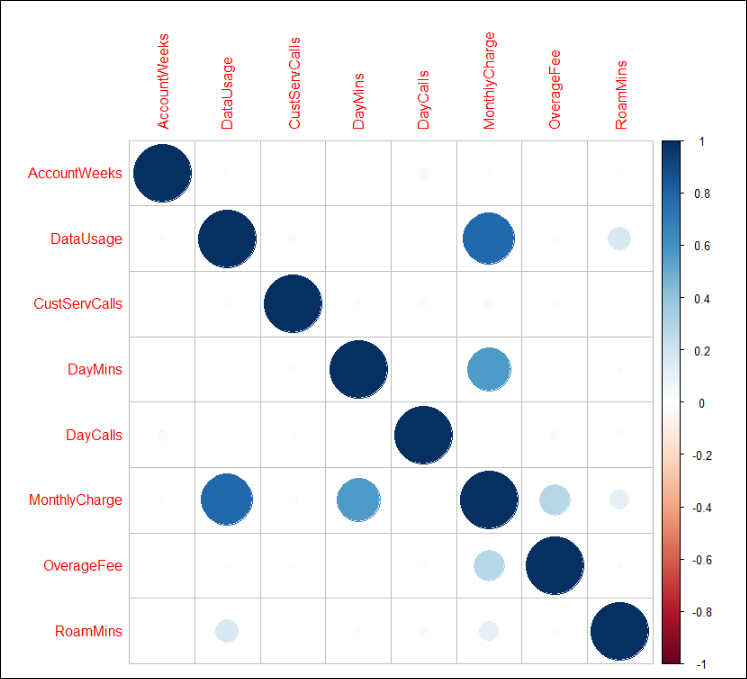
**EDA Inference& Demand for Predictive Modelling in this Data set**

* Frequency of Active Account weeks per customer peaks at 100 weeks
* Customers who have renewed their contract do not seem to churn
* Churned customers have used very minimal data
* Most Customers do not have a data plan
* Higher the Customer Service calls and Daytime calls, higher is the Churn rate. This is needs to be looked in detail
* There is no clarity on which variable is affecting the Churn rate more.
* Thus, we need to build a model to understand the Churn rate on the available dataset
* As the Defendant variable is Binary in nature, we can go ahead and try out Logistic Regression of model building

**MULTICOLLINEARITY CHECKING**

When the independent variables exist with multicollinearity, then it would be a challenging task for building the logistic regression model which will give us a precise prediction of customer churn out or not. Below is the correlation plot for the “Cellphone” dataset

**corrplot(cor(CPhone[,-c(1,3,4)]))**

****

* It is very clear from the correlation plot that the independent variable of MonthlyCharge is highly correlated with DataUsage, DayMins and RoamMins.
* However, we still have decided to build the model with all the independent variables.
* We would need to further check with vif at a later stage of model building to see the effect of correlated variables

**MODEL 1- Model building with all independent variables**

Initial we have taken is splitting the entire dataset for train and test

**set.seed(1234)**

**CPhone.index<-sample(1:nrow(CPhone), nrow(CPhone)\*0.70)**

**train<-CPhone[CPhone.index,]**

**test<-CPhone[-CPhone.index,]**

**dim(train)**

**dim(test)**

**MODEL 1**

**model1<-glm(train$Churn~., data = train, family = binomial)**

**summary(model1)**

Call:

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

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0073 -0.5200 -0.3645 -0.2259 2.8712

Coefficients:

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

(Intercept) -5.4302463 0.6401307 -8.483 < 2e-16 \*\*\*

AccountWeeks0.0013389 0.0016559 0.809 0.41876

ContractRenewal1 -2.0167863 0.1683616 -11.979 < 2e-16 \*\*\*

DataPlan1 -0.8229771 0.6103364 -1.348 0.17753

DataUsage -2.7903565 2.2654057 -1.232 0.21805

CustServCalls0.5091948 0.0452992 11.241 < 2e-16 \*\*\*

DayMins -0.0360141 0.0382466 -0.942 0.34638

DayCalls0.0009007 0.0031937 0.282 0.77792

MonthlyCharge0.2774728 0.2246527 1.235 0.21679

OverageFee -0.3462631 0.3829918 -0.904 0.36594

RoamMins0.0854158 0.0262351 3.256 0.00113 \*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1969.1 on 2332 degrees of freedom

Residual deviance: 1584.0 on 2322 degrees of freedom

AIC: 1606

Number of Fisher Scoring iterations: 5

From the summary of the model we have found that most of the independent variables are insignificant. Hence we decided that, due to multicollinearity exists between the independent variables this model1 have some insignificant independent variables and proceeded to run vif() for the model1.

**vif(model1)**

>vif(model1)

AccountWeeksContractRenewalDataPlanDataUsageCustServCalls

1.006176 1.064696 13.516598 1625.265693 1.093255

DayMinsDayCallsMonthlyChargeOverageFeeRoamMins

968.656528 1.003295 2904.037236 216.228842 1.216822

**FINDING ODD ‘s**

**exp(coef(model1))**

>exp(coef(model1))

(Intercept) AccountWeeks ContractRenewal1 DataPlan1

0.004382017 1.001339802 0.133082460 0.439122403

DataUsageCustServCallsDayMinsDayCalls

0.061399321 1.663950815 0.964626696 1.000901127

MonthlyChargeOverageFeeRoamMins

1.319790240 0.707326360 1.089169831

**FINDING PROBAILITIES**

**exp(coef(model1))/(1+exp(coef(model1)))**

>exp(coef(model1))/(1+exp(coef(model1)))

(Intercept) AccountWeeks ContractRenewal1 DataPlan1

0.004362898 0.500334726 0.117451699 0.305132074

DataUsageCustServCallsDayMinsDayCalls

0.057847523 0.624617694 0.490997449 0.500225180

MonthlyChargeOverageFeeRoamMins

0.568926542 0.414288900 0.521340972

**LOG LIKELIHOOD RATIO TEST**

Likelihood ratio test

Model 1: train$Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +

CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee +

RoamMins

Model 2: train$Churn ~ 1

#Df LogLik Df ChisqPr(>Chisq)

1 11 -792.00

2 1 -984.53 -10 385.06 < 2.2e-16 \*\*\*

---

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

* Contract Renewal and CustomerServCalls are the most significant variables. With one unit increase in CustServCalls there is a 62% increase in chance of customer churning. However with Contract Renewal it increases only by 13%.
* The Likelihood ratio test indicates that the model is significant. However we find the residual deviance value to be lesser than the null deviance.

**MODEL EVALUATION**

**pR2(model1)**

>pR2(model1)

llhllhNull G2 McFadden r2ML r2CU

-791.9997775 -984.5309127 385.0622704 0.1955562 0.1521489 0.2669210

**model1.pred<-predict(model1, type = 'response')**

**cutoff<-floor(model1.pred+0.8)**

**confmat<-table(Predicted=cutoff, Actual=train$Churn)**

**model1.eval<-confusionMatrix(confmat, positive = "1", mode = "everything")**

**model1.eval**

>model1.eval

Confusion Matrix and Statistics

Actual

Predicted 0 1

0 1655 126

1 329 223

Accuracy : 0.805

95% CI : (0.7883, 0.8209)

No Information Rate : 0.8504

P-Value [Acc> NIR] : 1

Kappa : 0.3817

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

Sensitivity : 0.63897

Specificity : 0.83417

PosPredValue : 0.40399

Neg PredValue : 0.92925

Precision : 0.40399

Recall : 0.63897

F1 : 0.49501

Prevalence : 0.14959

Detection Rate : 0.09559

Detection Prevalence : 0.23661

Balanced Accuracy : 0.73657

'Positive' Class : 1

**pred1<-prediction(model1.pred, train$Churn)**

**model1.perf<-performance(pred1, "tpr", "fpr")**

**model1.KS<-max(attr(model1.perf, 'y.values')[[1]]-attr(model1.perf, "x.values")[[1]])**

**model1.auc<-performance(pred1,"auc")**

**model1.auc<-as.numeric(model1.auc@y.values)**

**model1.gini<-ineq(model1.pred, type="Gini")**

>model1.KS

[1] 0.5144292

>model1.auc

[1] 0.811037

>model1.gini

[1] 0.5067753

* McFadden R squared of 0.19 indicated that the model is reasonable
* Confusion matrix the cutoff has been changed to 0.8 in order to improve the sensitivity of the model and also the F1 score. The cost of acquiring a new customer is much higher than cost of retaining an existing one.This is the reason for improving the sensitivity is because, in telecom industry, the cost of acquiring new customers 3-5 time high. F score enables us to get the reasonable balance between type 1 and type 2 error. We also cannot just focus on sensitivity because then the specificity and overall model accuracy deteriorates. If the model accuracy becomes low then business users would not even accept the model.
* AUC value is good at 0.81 and KS statistic is also reasonably good at 0.51

**VIF CHECK ON INDEPENDENT VARIABLES**

**vif(glm(train$Churn~. -MonthlyCharge, data = train, family = binomial))**

>vif(glm(train$Churn~. -MonthlyCharge, data = train, family = binomial))

AccountWeeksContractRenewalDataPlanDataUsageCustServCalls

1.006280 1.062349 13.498418 13.779695 1.092629

DayMinsDayCallsOverageFeeRoamMins

1.031771 1.003229 1.024802 1.215407

Even after removing the MonthlyCharge independent variable we still see DataPlan and DataUsage have high correlation values and we know that, only variables which has correlation value less than 5 are acceptable to include in the model.

**vif(glm(train$Churn~. -MonthlyCharge -DataPlan, data = train, family = binomial))**

>vif(glm(train$Churn~. -MonthlyCharge -DataPlan, data = train, family = binomial))

AccountWeeksContractRenewalDataUsageCustServCallsDayMins

1.005960 1.062151 1.029088 1.088493 1.031757

DayCallsOverageFeeRoamMins

1.003285 1.023580 1.023757

We tried removing only the DataPlan and the result of vif have suggesting to go with the rest of variables since the values are acceptable.

**CHECKING GLM SUMMARY WITHOUT MULTICOLLINEARITY IDV’s**

Performing pre check to find if there is any insignificant independent variables even though after removing the high multicollinearity variables.

**summary(glm(train$Churn~. -MonthlyCharge -DataPlan, data = train, family = binomial))**

>summary(glm(train$Churn~. -MonthlyCharge -DataPlan, data = train, family = binomial))

Call:

glm(formula = train$Churn ~ . - MonthlyCharge - DataPlan, family = binomial,

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0124 -0.5236 -0.3641 -0.2317 2.9156

Coefficients:

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

(Intercept) -5.4314915 0.6252992 -8.686 < 2e-16 \*\*\*

AccountWeeks0.0013791 0.0016562 0.833 0.405

ContractRenewal1 -2.0124144 0.1679594 -11.982 < 2e-16 \*\*\*

DataUsage -0.2879068 0.0574603 -5.011 5.43e-07 \*\*\*

CustServCalls0.5051101 0.0452039 11.174 < 2e-16 \*\*\*

DayMins0.0112277 0.0012468 9.005 < 2e-16 \*\*\*

DayCalls0.0008613 0.0031922 0.270 0.787

OverageFee0.1245042 0.0263012 4.734 2.20e-06 \*\*\*

RoamMins0.0992388 0.0240972 4.118 3.82e-05 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1969.1 on 2332 degrees of freedom

Residual deviance: 1587.5 on 2324 degrees of freedom

AIC: 1605.5

Number of Fisher Scoring iterations: 5

1. Summary reveals us that, AccountWeeks and DayCalls are being insignificant but we have decided to remove only AccountWeeks since it is the most insignificant variable (40%) when comparing with DayCalls of (78%)
2. Obviously rest of the independent variables are being significant

**MODEL 2- Model building without insignificant variables**

Building the second logistic regression model without insignificance variables and high multicollinearity variables.

* Insignificant Variables- AccountWeeks
* Multicollinearity Variables- Monthly Charge, DataPlan

**model2<-glm(train$Churn~., data = train[,-c(2,4,9)], family = binomial)**

**summary(model2)**

>summary(model2)

Call:

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

-c(2, 4, 9)])

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0135 -0.5226 -0.3660 -0.2319 2.9151

Coefficients:

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

(Intercept) -5.2866084 0.5990357 -8.825 < 2e-16 \*\*\*

ContractRenewal1 -2.0145893 0.1679226 -11.997 < 2e-16 \*\*\*

DataUsage -0.2854577 0.0573308 -4.979 6.39e-07 \*\*\*

CustServCalls0.5054924 0.0452042 11.182 < 2e-16 \*\*\*

DayMins0.0112610 0.0012467 9.033 < 2e-16 \*\*\*

DayCalls0.0008858 0.0031903 0.278 0.781

OverageFee0.1237987 0.0262586 4.715 2.42e-06 \*\*\*

RoamMins0.0985773 0.0240701 4.095 4.21e-05 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1969.1 on 2332 degrees of freedom

Residual deviance: 1588.2 on 2325 degrees of freedom

AIC: 1604.2

Number of Fisher Scoring iterations: 5

1. We can understand that model2 have done its work by looking at the difference between the Null deviance and Residual deviance values
2. It is clear that model2 have performed better then model since the AIC values of model2 is 1604.2 which is lesser than the model1 AIC value of 1606
3. We will proceed to check Odds, Probabilities of each predictor variables and perform log likelihood ratio test to do deep check on the model2 performance
4. Excluding Data calls, all the variables have very high significance. Even 1 unit increase in other variables shall increase the chances of Churn significantly

**FINDING ODD ‘s**

**exp(coef(model2))**

>exp(coef(model2))

(Intercept) ContractRenewal1 DataUsageCustServCalls

0.005058889 0.133375175 0.751670132 1.657801608

DayMinsDayCallsOverageFeeRoamMins

1.011324688 1.000886195 1.131787992 1.103599703

**FINDING PROBAILITIES**

**exp(coef(model2))/(1+exp(coef(model2)))**

>exp(coef(model2))/(1+exp(coef(model2)))

(Intercept) ContractRenewal1 DataUsageCustServCalls

0.005033425 0.117679633 0.429116258 0.623749193

DayMinsDayCallsOverageFeeRoamMins

0.502815231 0.500221451 0.530910201 0.524624386

* From probabilities we can understand that if there is 1 unit increase in the CustServCalls then it will increases the chances of Churn by 62%
* If there is 1 unit increase in DayMins, DayCalls, OverageFee and RoamMins then they will increases the chances of Churn by 50%

**LOG LIKELIHOOD RATIO TEST**

**lrtest(model2)**

>lrtest(model2)

Likelihood ratio test

Model 1: train$Churn ~ ContractRenewal + DataUsage + CustServCalls + DayMins +

DayCalls + OverageFee + RoamMins

Model 2: train$Churn ~ 1

#Df LogLik Df ChisqPr(>Chisq)

1 8 -794.12

2 1 -984.53 -7 380.83 < 2.2e-16 \*\*\*

---

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

* The log likelihood ratio test confirms that the model is good for further process

**MODEL2 PREDICTION AND EVALUATION**

First we will check the McFadden R Square test result to evaluate the model2

>pR2(model2)

llhllhNull G2 McFadden r2ML r2CU

-794.1178994 -984.5309127 380.8260265 0.1934048 0.1506080 0.2642177

* The round off value of McFadden value is 0.2 and hence model can be considered

#Model2 prediction and evaluation

**model2.pred<-predict(model2, type = 'response')**

**cutoff1<-floor(model2.pred+0.8)**

**confmat1<-table(Predicted=cutoff1, Actual=train$Churn)**

**model2.eval<-confusionMatrix(confmat1, positive = "1", mode = "everything")**

**model2.eval**

>model2.eval

Confusion Matrix and Statistics

Actual

Predicted 0 1

0 1657 123

1 327 226

Accuracy : 0.8071

95% CI : (0.7905, 0.8229)

No Information Rate : 0.8504

P-Value [Acc> NIR] : 1

Kappa : 0.389

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

Sensitivity : 0.64756

Specificity : 0.83518

PosPredValue : 0.40868

Neg PredValue : 0.93090

Precision : 0.40868

Recall : 0.64756

F1 : 0.50111

Prevalence : 0.14959

Detection Rate : 0.09687

Detection Prevalence : 0.23703

Balanced Accuracy : 0.74137

'Positive' Class : 1

* We have set the cutoff value to 0.8 in order to improve the sensitivity of the model and its performance metrics

**MODEL2 PERFORMANCE**

Model performance is identified using the metrics of KS statistics, AUC and Gini values

**pred2<-prediction(model2.pred, train$Churn)**

**model2.perf<-performance(pred2, "tpr", "fpr")**

**model2.KS<-max(attr(model2.perf, 'y.values')[[1]]-attr(model2.perf, "x.values")[[1]])**

**model2.auc<-performance(pred2,"auc")**

**model2.auc<-as.numeric(model2.auc@y.values)**

**model2.gini<-ineq(model2.pred, type="Gini")**

>model2.KS

[1] 0.504776

>model2.auc

[1] 0.8100492

>model2.gini

[1] 0.5037524

From the above result, the model performance is good since the AUC is good at 0.81 and KS statistics is good at 0.50

**MODEL 2 PERFORMANCE ON TEST DATASET**

**model2.pred.test<-predict(model2, test, type = 'response')**

**cutoff2<-floor(model2.pred.test+0.8)**

**confmat2<-table(Predicted=cutoff2, Actual=test$Churn)**

**model2.eval.test<-confusionMatrix(confmat2, positive = "1", mode = "everything")**

**model2.eval.test**

>model2.eval.test

Confusion Matrix and Statistics

Actual

Predicted 0 1

0 726 37

1 140 97

Accuracy : 0.823

95% CI : (0.7979, 0.8462)

No Information Rate : 0.866

P-Value [Acc> NIR] : 0.9999

Kappa : 0.4244

Mcnemar's Test P-Value : 1.764e-14

Sensitivity : 0.7239

Specificity : 0.8383

PosPredValue : 0.4093

Neg PredValue : 0.9515

Precision : 0.4093

Recall : 0.7239

F1 : 0.5229

Prevalence : 0.1340

Detection Rate : 0.0970

Detection Prevalence : 0.2370

Balanced Accuracy : 0.7811

'Positive' Class : 1

**pred2t<-prediction(model2.pred.test, test$Churn)**

**model2.perf.t<-performance(pred2t, "tpr", "fpr")**

**model2.KS.t<-max(attr(model2.perf.t, 'y.values')[[1]]-attr(model2.perf.t, "x.values")[[1]])**

**model2.auc.t<-performance(pred2t,"auc")**

**model2.auc.t<-as.numeric(model2.auc.t@y.values)**

**model2.gini.t<-ineq(model2.pred.test, type="Gini")**

>model2.KS.t

[1] 0.5878115

>model2.auc.t

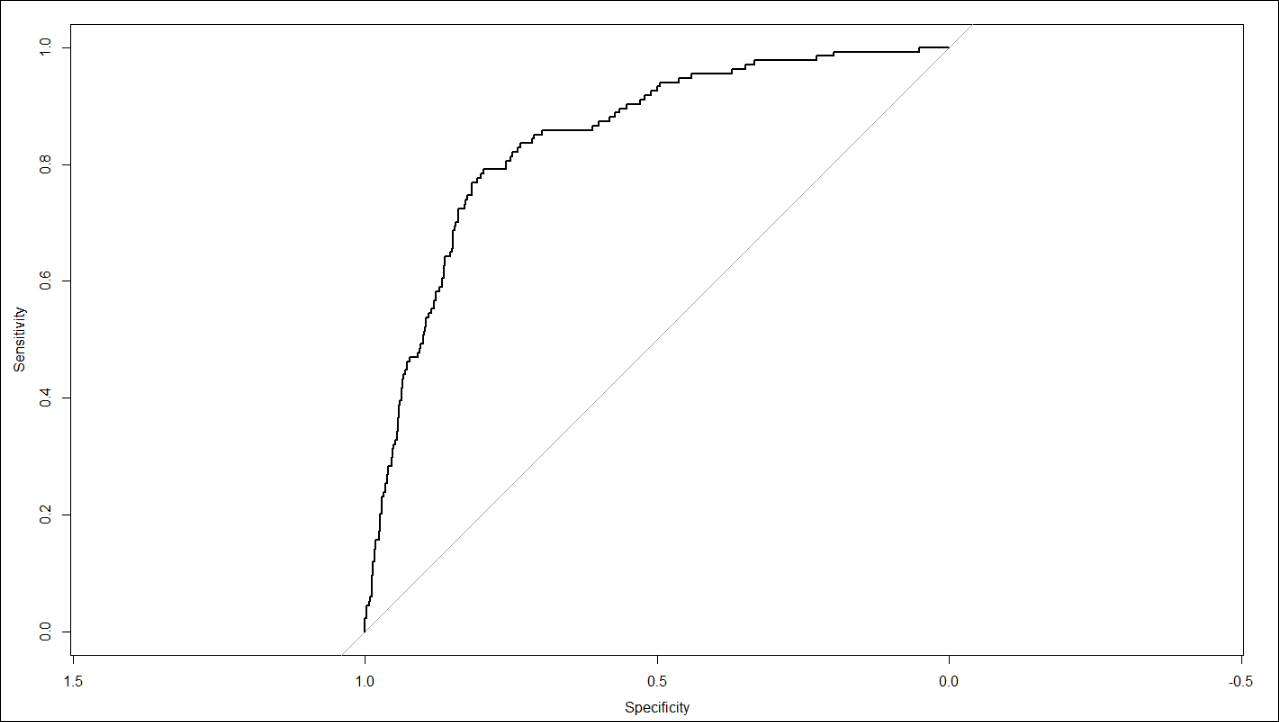
[1] 0.8421202

>model2.gini.t

[1] 0.5078639

* From the performance result, the AUC of 0.84 and KS statistics of 0.58 which are good when comparing to the model1 AUC and KS statistics

**ROC CURVE FOR MODEL2 - TEST DATA**



**BALANCING THE DATASET**

We have used the oversampling technique to balance the data under SMOTE method

**output<-train$Churn**

**input<-train[,-1]**

**data<-ubBalance(X= input, Y=output, type="ubSMOTE", percOver=300, percUnder=150, verbose=TRUE)**

**train.bal<-cbind(data$X,data$Y)**

**names(train.bal)[11]<-"Churn"**

>dim(train.bal)

[1] 2966 11

>table(train.bal$Churn)

0 1

1570 1396

>prop.table(table(train.bal$Churn))

0 1

0.5293324 0.4706676

* The balanced dataset has 2966 observations
* We see that the number of retained customers are at 52% and churned customers are at 47%

**CHECK FOR MULTICOLLINEARITY AND REMOVE**

**vif(glm(train.bal$Churn~., data = train.bal, family = binomial))**

>vif(glm(train.bal$Churn~., data = train.bal, family = binomial))

AccountWeeksContractRenewalDataPlanDataUsageCustServCalls

1.006053 1.039499 2.024587 1911.835667 1.113244

DayMinsDayCallsMonthlyChargeOverageFeeRoamMins

1157.672520 1.011080 3274.371999 230.127964 1.067814

**vif(glm(train.bal$Churn~. -MonthlyCharge, data = train.bal, family = binomial))**

>vif(glm(train.bal$Churn~. -MonthlyCharge, data = train.bal, family = binomial))

AccountWeeksContractRenewalDataPlanDataUsageCustServCalls

1.005206 1.036706 2.030271 2.105226 1.112667

DayMinsDayCallsOverageFeeRoamMins

1.077124 1.011156 1.036211 1.069237

**vif(glm(train.bal$Churn~. -MonthlyCharge -DataPlan, data = train.bal, family = binomial))**

>vif(glm(train.bal$Churn~. -MonthlyCharge -DataPlan, data = train.bal, family = binomial))

AccountWeeksContractRenewalDataUsageCustServCallsDayMins

1.005157 1.035973 1.041189 1.112689 1.075935

DayCallsOverageFeeRoamMins

1.011316 1.035474 1.036674

* We will remove MonthlCharge and DataPlan variables
* Now the multicollinearity values of rest of the variables are less than 2 which can considered to use in the model formula

**summary(glm(train.bal$Churn~. -MonthlyCharge -DataPlan, data = train.bal, family = binomial))**

>summary(glm(train.bal$Churn~. -MonthlyCharge -DataPlan, data = train.bal, family = binomial))

Call:

glm(formula = train.bal$Churn ~ . - MonthlyCharge - DataPlan,

family = binomial, data = train.bal)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.7868 -0.8229 -0.4565 0.9143 2.3016

Coefficients:

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

(Intercept) -2.5801633 0.4151415 -6.215 5.13e-10 \*\*\*

AccountWeeks0.0035173 0.0011975 2.937 0.00331 \*\*

ContractRenewal1 -2.3793694 0.1211164 -19.645 < 2e-16 \*\*\*

DataUsage0.1227471 0.0381744 3.215 0.00130 \*\*

CustServCalls0.4872556 0.0330100 14.761 < 2e-16 \*\*\*

DayMins0.0102591 0.0008474 12.106 < 2e-16 \*\*\*

DayCalls0.0010684 0.0022887 0.467 0.64064

OverageFee0.1207806 0.0186602 6.473 9.63e-11 \*\*\*

RoamMins -0.0243246 0.0168222 -1.446 0.14818

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4101.5 on 2965 degrees of freedom

Residual deviance: 3141.9 on 2957 degrees of freedom

AIC: 3159.9

Number of Fisher Scoring iterations: 4

**summary(glm(train.bal$Churn~. -MonthlyCharge -DataPlan -DayCalls -RoamMins, data = train.bal, family = binomial))**

>summary(glm(train.bal$Churn~. -MonthlyCharge -DataPlan -DayCalls -RoamMins, data = train.bal, family = binomial))

Call:

glm(formula = train.bal$Churn ~ . - MonthlyCharge - DataPlan -

DayCalls - RoamMins, family = binomial, data = train.bal)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.7724 -0.8285 -0.4603 0.9106 2.3154

Coefficients:

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

(Intercept) -2.7201095 0.3080723 -8.829 < 2e-16 \*\*\*

AccountWeeks0.0036003 0.0011956 3.011 0.00260 \*\*

ContractRenewal1 -2.3847371 0.1209320 -19.720 < 2e-16 \*\*\*

DataUsage0.1126024 0.0374809 3.004 0.00266 \*\*

CustServCalls0.4864129 0.0329660 14.755 < 2e-16 \*\*\*

DayMins0.0102529 0.0008469 12.107 < 2e-16 \*\*\*

OverageFee0.1217032 0.0185879 6.547 5.85e-11 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4101.5 on 2965 degrees of freedom

Residual deviance: 3144.2 on 2959 degrees of freedom

AIC: 3158.2

Number of Fisher Scoring iterations: 4

After checking multicollinearity, we have identified that there are two insignificant variables which are DayCalls and RoamMins will be removed in the model3

**MODEL 3: MODEL BUILDING FOR BALANCED DATA**

**model3<-glm(train.bal$Churn~. -MonthlyCharge -DataPlan -DayCalls -RoamMins, data = train.bal, family = binomial)**

**summary(model3)**

>model3<-glm(train.bal$Churn~. -MonthlyCharge -DataPlan -DayCalls -RoamMins, data = train.bal, family = binomial)

>summary(model3)

Call:

glm(formula = train.bal$Churn ~ . - MonthlyCharge - DataPlan -

DayCalls - RoamMins, family = binomial, data = train.bal)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.7724 -0.8285 -0.4603 0.9106 2.3154

Coefficients:

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

(Intercept) -2.7201095 0.3080723 -8.829 < 2e-16 \*\*\*

AccountWeeks0.0036003 0.0011956 3.011 0.00260 \*\*

ContractRenewal1 -2.3847371 0.1209320 -19.720 < 2e-16 \*\*\*

DataUsage0.1126024 0.0374809 3.004 0.00266 \*\*

CustServCalls0.4864129 0.0329660 14.755 < 2e-16 \*\*\*

DayMins0.0102529 0.0008469 12.107 < 2e-16 \*\*\*

OverageFee0.1217032 0.0185879 6.547 5.85e-11 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4101.5 on 2965 degrees of freedom

Residual deviance: 3144.2 on 2959 degrees of freedom

AIC: 3158.2

Number of Fisher Scoring iterations: 4

**FINDING ODD’S**

**exp(coef(model3))**

>exp(coef(model3))

(Intercept) AccountWeeks ContractRenewal1 DataUsage

0.06586754 1.00360683 0.09211320 1.11918686

CustServCallsDayMinsOverageFee

1.62647150 1.01030564 1.12941879

**FINDING PROBABILITIES**

**exp(coef(model3))/(1+exp(coef(model3)))**

>exp(coef(model3))/(1+exp(coef(model3)))

(Intercept) AccountWeeks ContractRenewal1 DataUsage

0.06179712 0.50090008 0.08434400 0.52812089

CustServCallsDayMinsOverageFee

0.61926105 0.50256320 0.53038829

**LOG LIKELIHOOD RATIO TEST**

**lrtest(model3)**

>lrtest(model3)

Likelihood ratio test

Model 1: train.bal$Churn ~ (AccountWeeks + ContractRenewal + DataPlan +

DataUsage + CustServCalls + DayMins + DayCalls + MonthlyCharge +

OverageFee + RoamMins) - MonthlyCharge - DataPlan - DayCalls -

RoamMins

Model 2: train.bal$Churn ~ 1

#Df LogLik Df ChisqPr(>Chisq)

1 7 -1572.1

2 1 -2050.8 -6 957.34 < 2.2e-16 \*\*\*

---

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

**McFADDEN R SQUARED TEST**

**pR2(model3)**

>pR2(model3)

llhllhNull G2 McFadden r2ML

-1572.0977619 -2050.7677624 957.3400010 0.2334101 0.2758606

r2CU

0.3682376

From the McFadden R Squared of 0.23, we can confirm that model3 can be considered

**MODEL3 PREDICTION AND EVALUATION**

**model3.pred<-predict(model3, type = 'response')**

**cutoff2<-floor(model3.pred+0.5)**

**confmat2<-table(Predicted=cutoff2, Actual=train.bal$Churn)**

**model3.eval<-confusionMatrix(confmat2, positive = "1", mode = "everything")**

**model3.eval**

>model3.eval

Confusion Matrix and Statistics

Actual

Predicted 0 1

0 1284 466

1 286 930

Accuracy : 0.7465

95% CI : (0.7304, 0.762)

No Information Rate : 0.5293

P-Value [Acc> NIR] :< 2.2e-16

Kappa : 0.4875

Mcnemar's Test P-Value : 6.69e-11

Sensitivity : 0.6662

Specificity : 0.8178

PosPredValue : 0.7648

Neg PredValue : 0.7337

Precision : 0.7648

Recall : 0.6662

F1 : 0.7121

Prevalence : 0.4707

Detection Rate : 0.3136

Detection Prevalence : 0.4100

Balanced Accuracy : 0.7420

'Positive' Class : 1

**MODEL3 PERFORMANCE**

**pred3<-prediction(model3.pred, train.bal$Churn)**

**model3.perf<-performance(pred3, "tpr", "fpr")**

**model3.KS<-max(attr(model3.perf, 'y.values')[[1]]-attr(model3.perf, "x.values")[[1]])**

**model3.auc<-performance(pred3,"auc")**

**model3.auc<-as.numeric(model3.auc@y.values)**

**model3.gini<-ineq(model3.pred, type="Gini")**

>model3.KS

[1] 0.5211825

>model3.auc

[1] 0.8153829

>model3.gini

[1] 0.3249562

KS statistics is at 0.52 and AUC is at 0.81 are good from the model3 for the balanced data

**MODEL3 ON TEST DATASET**

**model3.pred.test<-predict(model3, test, type = 'response')**

**cutoff4<-floor(model3.pred.test+0.8)**

**confmat4<-table(Predicted=cutoff4, Actual=test$Churn)**

**model3.eval.test<-confusionMatrix(confmat2, positive = "1", mode = "everything")**

**model3.eval.test**

>model3.eval.test

Confusion Matrix and Statistics

Actual

Predicted 0 1

0 1284 466

1 286 930

Accuracy : 0.7465

95% CI : (0.7304, 0.762)

No Information Rate : 0.5293

P-Value [Acc> NIR] :< 2.2e-16

Kappa : 0.4875

Mcnemar's Test P-Value : 6.69e-11

Sensitivity : 0.6662

Specificity : 0.8178

PosPredValue : 0.7648

Neg PredValue : 0.7337

Precision : 0.7648

Recall : 0.6662

F1 : 0.7121

Prevalence : 0.4707

Detection Rate : 0.3136

Detection Prevalence : 0.4100

Balanced Accuracy : 0.7420

'Positive' Class : 1

**TEST MODEL 3 PERFORMANCE**

**pred3t<-prediction(model3.pred.test, test$Churn)**

**model3.perf.t<-performance(pred3t, "tpr", "fpr")**

**model3.KS.t<-max(attr(model3.perf.t, 'y.values')[[1]]-attr(model3.perf.t, "x.values")[[1]])**

**model3.auc.t<-performance(pred3t,"auc")**

**model3.auc.t<-as.numeric(model3.auc.t@y.values)**

**model3.gini.t<-ineq(model3.pred.test, type="Gini")**

>model3.KS.t

[1] 0.5164937

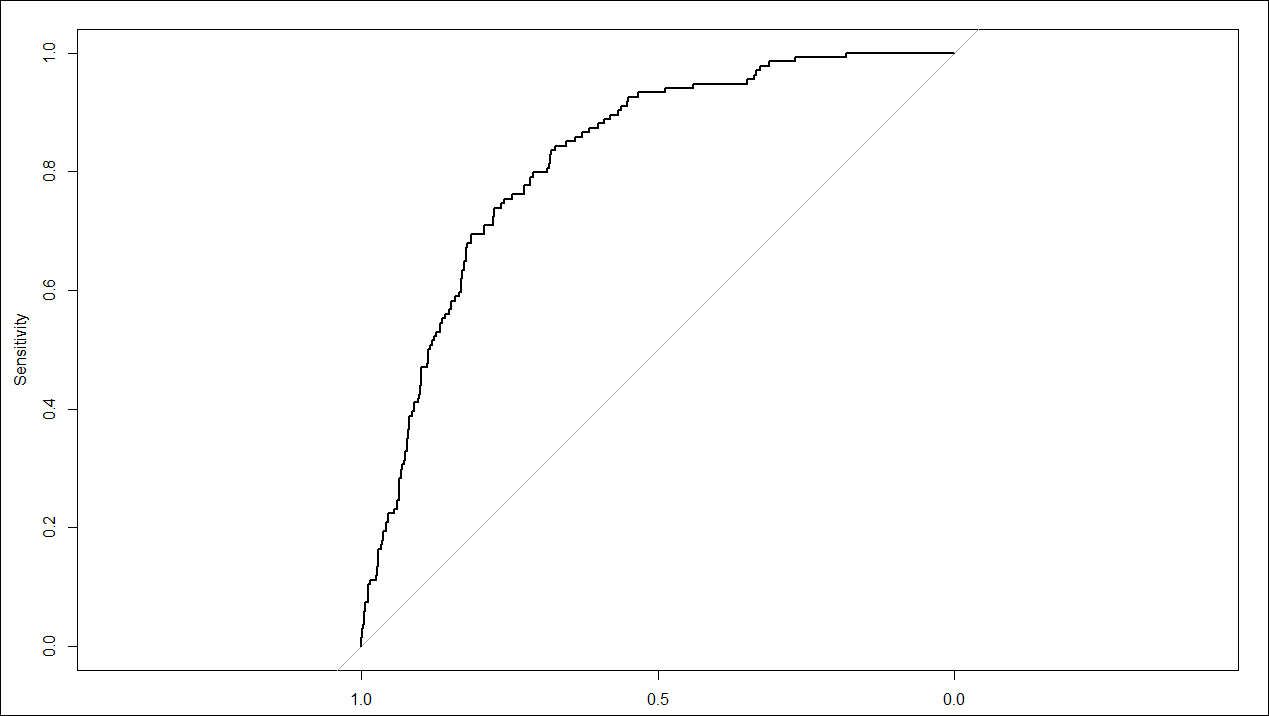
>model3.auc.t

[1] 0.8207835

>model3.gini.t

[1] 0.3340034

**ROC CURVE FOR MODEL3 - TEST DATA**

****

**MODEL COMPARISONS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models Output Comparison** | | | | | | |
| **Parameters** | **Model1** | | **Model2** | | **Model3** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| KS Statistics | 0.51 | **0.59** | 0.50 | **0.58** | 0.52 | **0.51** |
| AUC | 0.81 | **0.84** | 0.81 | **0.84** | 0.81 | **0.82** |
| GINI | 0.50 | **0.51** | 0.50 | **0.50** | 0.32 | **0.33** |
| Accuracy | 80% | **82%** | 80% | **82%** | 74% | **74%** |
| Precision | 40% | **40%** | 40% | **40%** | 76% | **76%** |
| Recall | 64% | **72%** | 64% | **72%** | 66% | **66%** |
| F1 Score | 50% | **52%** | 50% | **52%** | 71% | **71%** |
| Sensitivity | 64% | **72%** | 64% | **72%** | 66% | **66%** |
| Specificity | 83% | **83%** | 83% | **83%** | 81% | **81%** |

**MODEL OUTPUT INSIGHTS:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **Null Deviance** | **Residual Deviance** | **Deviance Difference** | **AIC** | **McFadden R^2** |
| **Model1** | **1969.1** | **1584** | **385.1** | **1606** | **0.1955562** |
| **Model2** | **1969.1** | **1588.2** | **380.9** | **1604.2** | **0.1934048** |
| **Model3** | **4101.5** | **3144.2** | **957.3** | **3158.2** | **0.2334101** |

* **Model1**: Built using all the independent variables and without multicollinearity treatment
* **Model2**: Built after removing the high multicollinearity variables and insignificance variables
* **Model3**: Built after doing the over sampling using SMOTE technique, removed high multicollinearity and insignificance variables
* Model 1 cannot be considered due to the issue of Multicollinearity between the Independent variables. Hence, we cannot take any actionable insights with respect to business perspective
* When we compare Model 2 and 3, we can see that the accuracy has decreased by 8 % in Model 3. Also, considering this is a balanced data, and the low performance output of the the model, we can very well go ahead with the interpretations of Model 2
* **Model2 is better than any other models and can be used for predicting customer Churn.**

**BUSINESS RECOMMENDATIONS**

* Customer Service Support Calls to be addressed &improved. Customer centric measures to be taken to provide FCR(First Call Resolution)
* Improved Training & Transaction calls monitoring to be improved to understand why customer are calling repeatedly
* Self-help Options can be introduced for customer to limit wait time in service operation calls & addressed through IVR or Customer Portal to reduce Churn
* Overage Fees to be rechecked by Business teams along with market trend & initiate the survey along with Customers of expectations and revision
* Proactive Customer Offers to be included to customers with the tenurity and usage pattern of the services
* Non-Data plans customers should be clustered and attractive data plans to be introduced with combo-plans for them to start off
* Customer using reduced data plans again to be targeted by rationalising the Data and call plans
* To Summarize attracting calling plans & improved calling(reception) Improved Customer centric measures like Quick resolution, Improved CSAT would make the churn rate under control