*Mini Project – Customer Churn Rate - Cellphone*

*By Sakshi Jain*

*Table of Contents*

1. *Project Objective* 
   1. *Problem in hand*

*b. Initial Hypothesis*

1. *Exploratory Data Analysis – Step by step approach*
   1. *Descriptive Statistics*
   2. *Missing Values*
   3. *Visual interpretation*
   4. *Data Preparation – Training and Testing Dataset*
2. *Logistic Regression*
   1. [*Model*](https://rstudio-pubs-static.s3.amazonaws.com/514452_3124107a4f444a81bb3fe549b58207b5.html#evidence-of-multicollinearity) *1* 
      1. *Model Significance*
      2. *Odds Explanatory Power*
      3. *Confusion Matrix with ROC Plot*

*b. Model 2*

1. *Model Significance*
2. *Odds Explanatory Power*
3. *Confusion Matrix with ROC Plot*
4. *Conclusion*

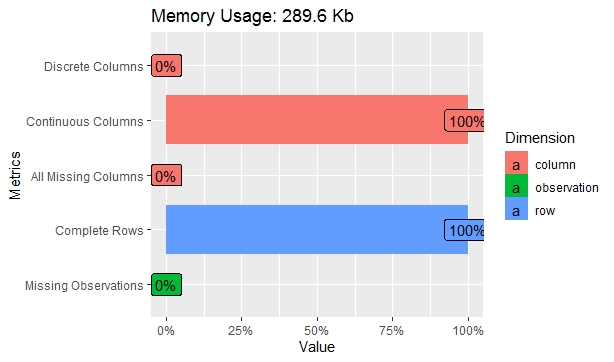
*Project Objective*

* *The objective of the report is to explore the data set “Cellphone.xlsx” in R and generate insights about the data set.*
* *Problem: Students are given a Cell Phone Data file and are requested to build a Logistic Regression Model which can tell the parameters contributing (and not contributing) for Customer Churn (attrition), along with the intensity of each attribute.*
* *The input file needs to divide into Training Dataset, which should contain 70% of the data and Testing Dataset, which would contain remaining 30% of the data.*
* *The Initial Hypothesis*
* *The Cell phone File contains one Dependent & Predictor variables. The assignment aim is to identify the predictor variables which are significant for Customer Churn.*
* *Null Hypothesis (Ho) –No predictor is able to predict the Churn*
* *Alternate Hypothesis (Ha)–At least one of the predictors is able to predict the churn.*

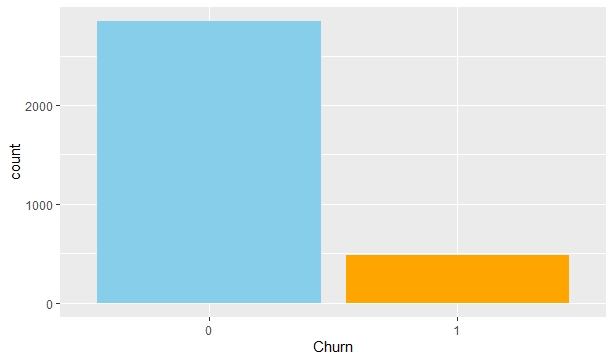
*Exploratory Data Analysis*

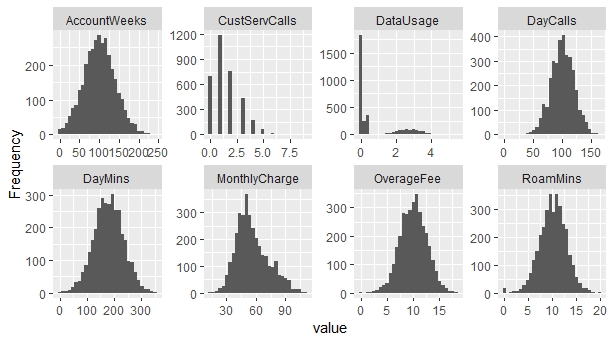
*Cellphone ==================================================================================*

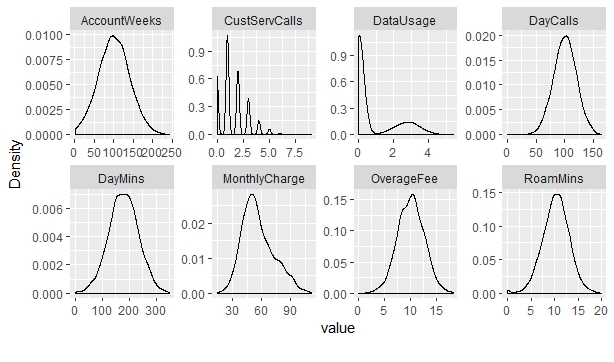
* + *Before building logistic regression model, we looked at the summary information to understand the data that we are dealing with. We will check for the variable names, five point summary & Missing values*
* *The data set used has 2333 rows & 11 columns. There were no missing values in the dataset. I have run the following graph to make it evident that there is no missing values present*

**

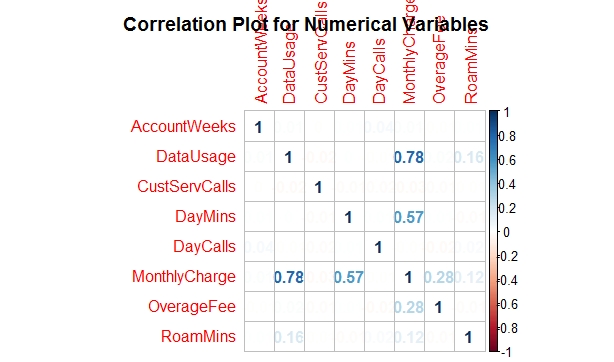
* *As this graph suggests missing values is 0%*
* *Let’s check the % of proportion of the customers who churned vs. who didn’t churned*

**

* *Let’s check for the normality of all the variables*

**

* *Except CustServCalls all the Variables seems to be normally distributed*
* *Check for Correlated Variables*

**

* *As the plot shows, following variables are highly correlated:*
* *datausage and monthlycharge highly correlated*
* *daymins and monthlycharge highly correlated*
* *We will Split the Input dataset into Training (70%) and Testing (30%). After splitting both the datasets. We see almost equal representation in both training and testing set for the dependent or response variable.*

*Logistic Regression – Model 1*

* *Model 1 with all variables: Building the initial Logistic Regression Model taking all independent variables into consideration*
* *We are using the GLM function*
* *When we check the summary of the combined model, below mentioned insights was generated:*
  + - *The three significant variables*:
      * *Contract Renewal: Please note, this has a negative impact on Customer Churn.*
      * *Customer Service Calls*
      * *Roaming Minutes*
* *Goodness of fit (Pseudo R²): The McFadden’s pseudo-R Squared test suggests that atleast 19% variance of the data is captured by our Model, which suggests it’s a robust model*
* *Akaike information criterion (AIC): AIC rewards the goodness of fit with an AIC of 1586.5. Contract Renewal (Yes), Customer Service Call and Roam Mins are significant whereas Data Plan (Yes) is marginally significant*
* *Odds Ratio: In Logistic Regression, the odds Ratio represent the constant effect of a predictor on the likelihood that one unit will occur. If a particular Variable as shown in following table is increased by ‘One Unit’, the odds of customer churn (Vs. not churning ) and the probability of Customer Churn is shown in the following table*

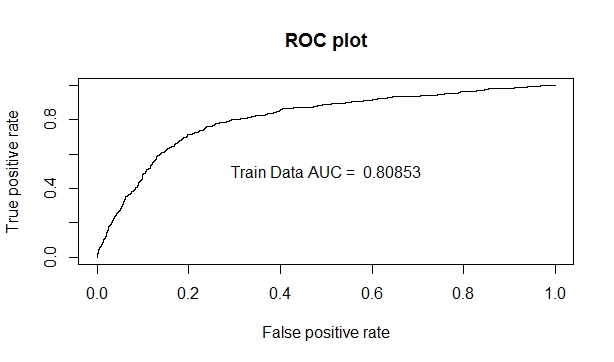
|  |  |  |
| --- | --- | --- |
| *Variable* | *The odds of*  *Customer will*  *Churn* | *Probability of*  *Customer Churn*  *increases by* |
| *Account Weeks* | *1.00* | *50%* |
| *ContractRenewal1* | *0.15* | *13%* |
| *DataPlan1* | *0.41* | *29%* |
| *DataUsage* | *1.02* | *51%* |
| *CustSrvCalls* | *1.66* | *62%* |
| *DayMins* | *1.01* | *50%* |
| *DayCalls* | *1.00* | *50%* |

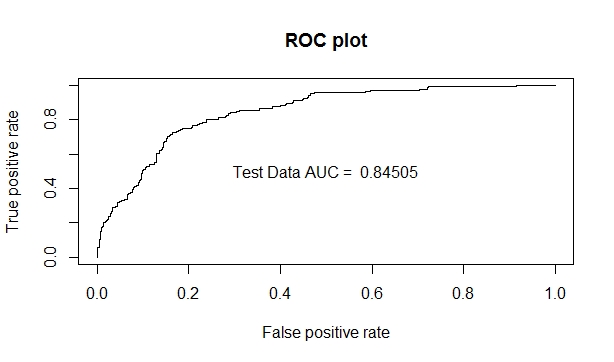
*MonthlyCharge 1.00 50%*

|  |  |  |
| --- | --- | --- |
| *Overage Fee* | *1.13* | *53%* |
| *RoamMin* | *1.07* | *52%* |

* + - *Confusion Matrix on Training Dataset:* 
      * *52 out of (52+48) Customers identified correctly which have been churned out. This translates to 52% of Positive Predictive Value.*
      * *1947 out of (1947+286) Customers identified correctly which have not been churned out. This translates to 87.19% of Negative Predictive Values.*
      * *At 85.68 %, the Model provides good accuracy measures*
    - *Confusion Matrix on Test Dataset:* 
      * *30 out of (30+18) Customers identified correctly which have been churned out. This translates to 62.5% of Positive Predictive Value.*
      * *837 out of (837+115) Customers identified correctly which have not been churned out. This translates to 87.9% of Negative Predictive Values.*
      * *At 86.7 %, the Model provides good accuracy measures*
    - *ROC & AUC on Training Dataset & Test Dataset:*

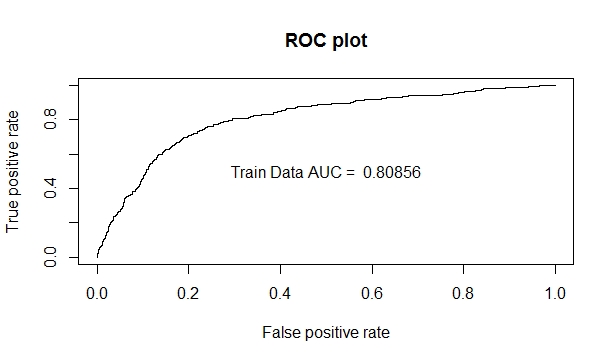
* *Thus, the model shows pretty much similar performance on both Training as well as Testing datasets*

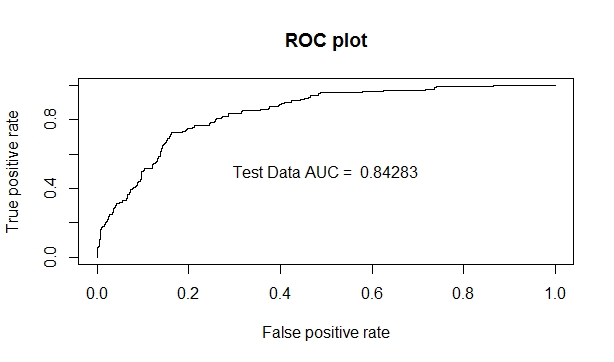
**

**

*Model 2 – With Step Function*

* + *We have included all the explanatory variables in our model. However, selecting the one’s which really matters for the model becomes really important. I have used Step() function in R to check my model using independent variables The Model with the minimum AIC will be the best model*
  + *From above step, we select our second model as:*
    - *Churn ~ ContractRenewal +*
    - *DataPlan +*
    - *CustServCalls +*
    - *DayMins +*
    - *OverageFee +*
    - *RoamMins*
  + *When we check the summary of the new model, below mentioned insights was generated:*
    - *Goodness of fit (Pseudo R²): The McFadden’s pseudo-R Squared test suggests that atleast 19% variance of the data is captured by our Model, which suggests it’s a robust model*
    - *Akaike information criterion (AIC): AIC rewards the goodness of fit with an AIC of 1579.1.*
    - *Confusion Matrix on Training Dataset:* 
      * *52 out of (52+49) Customers identified correctly which have been churned out. This translates to 51.4% of Positive Predictive Value.*
      * *1946 out of (1946+286) Customers identified correctly which have not been churned out. This translates to 87.18% of Negative Predictive Values.*
      * *At 85.64%, the Model provides good accuracy measures*
    - *Confusion Matrix on Test Dataset:* 
      * *31 out of (31+19) Customers identified correctly which have been churned out. This translates to 62% of Positive Predictive Value.*
      * *836 out of (836+114) Customers identified correctly which have not been churned out. This translates to 88% of Negative Predictive Values.*
      * *At 86.70%, the Model provides good accuracy measures*
  + *ROC & AUC on Training Dataset & Test Dataset*

**

**

*Conclusion*

* + *Based on the data set provided and the detailed analysis we have done, the organization should focus on targeted marketing targeting the customers predicted to churn by our best predictive model Model 3*
  + *Key observations: The following kind of customers has higher probability to churn* 
    - *Customer calling service center more than once who has not renewed his contract recently*
    - *Customer calling service center more five time who has a data plan*
    - *Customer having data plan who haven’t renewed his contract*
    - *Customer who haven’t renewed his contract*
    - *Customer having higher roaming mins*
    - *Customer having higher overage fee*
    - *Customer having higher monthly charge*