**Group Assignment – PM**

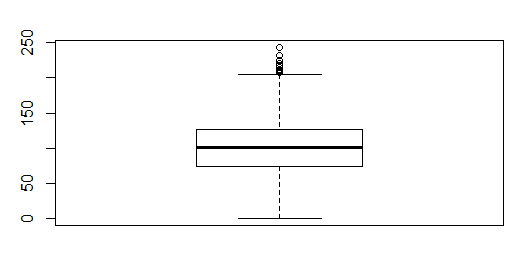
**Submitted by – Rohit Raj Singh, Sheetanshu Upadhyay, Jay Mehta, Vikas Jugran**

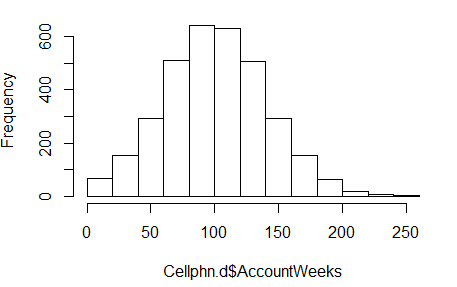
**Exploratory data analysis:**

* There are 3,333 objects and 11 variables in the dataset
* All 3,333 cases in the dataset are complete and there are no missing values

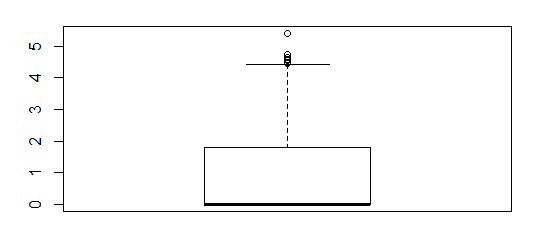
**Univariate analysis:**

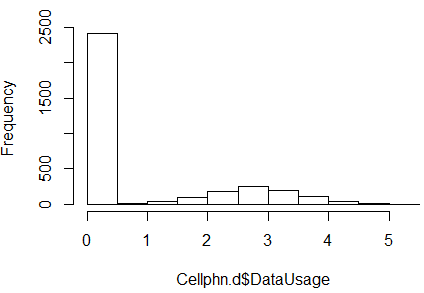
* Out of total, customers who have cancelled service (churn) = 483 and customers who have not = 2,850
* Number of contracts recently renewed = 3010, not renewed = 323
* AccountWeeks - There is a great variation in terms of number of weeks (duration) from which, accounts are active. The duration ranges from 1 week to 243 weeks with a mean of 101.1 and median of 101. The values are uniformly distributed as mean and median are equal. Few outliers are present on the higher side.



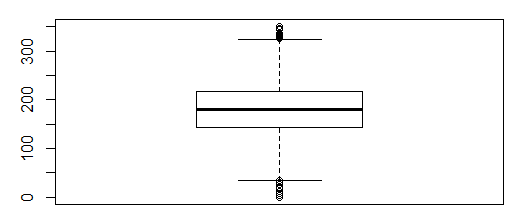


* DataUsage – Mean data usage is .82 gigabytes/month, however, there are significant number of customers who do not use data. The values are greatly skewed towards right which reflects relatively more data pointers on the lower end of the scale.

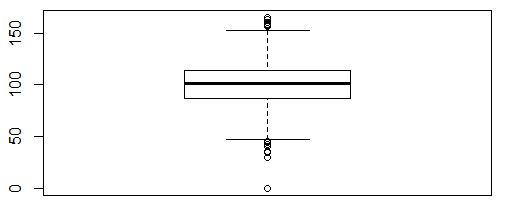




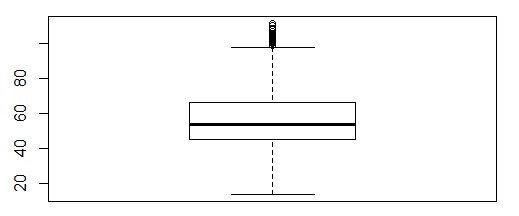
* DayMins – Average daytime minutes per month is uniformly distributed with significant variation (0 – 350). It has outliers on both upper as well as lower side.



* DayCalls – Average number of daytime calls is uniformly distributed with a mean and median of ~100. It has outliers on both upper as well as lower side.

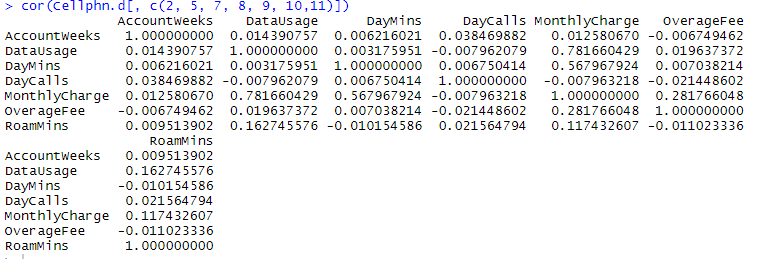


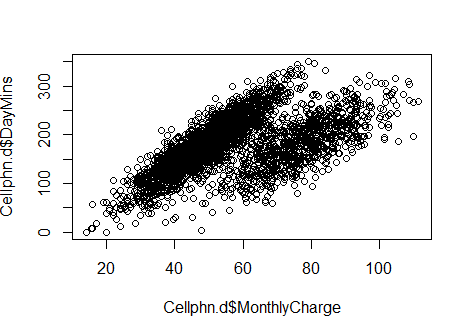
* Monthlycharge – Skewed towards right with a mean of 56 and median of 53. It has outliers on the upper side.

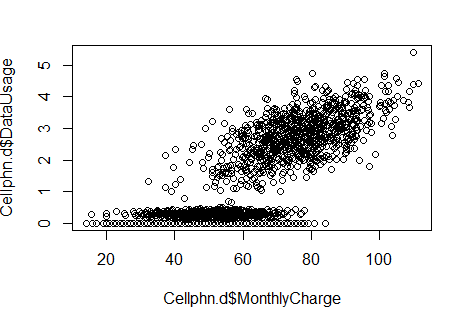


**Bivariate analysis:**

* There is not much of correlation between the continuous variables except between monthly charge and data usage(strong correlation), and Monthly charge and day minutes (moderate).







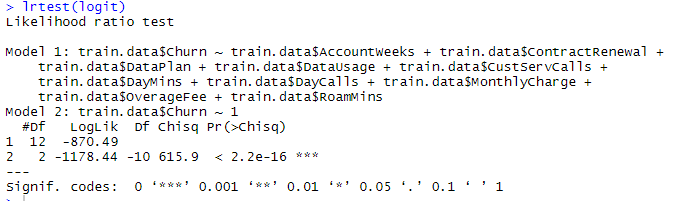
**Business problem:**

* Identification of customers who are likely to cancel services to plan a customer retention strategy targeting these customers
* Identification of variables/factors that are relevant for prediction of such customers and building a predictive model using these set of variables.

**Logistic regression model**

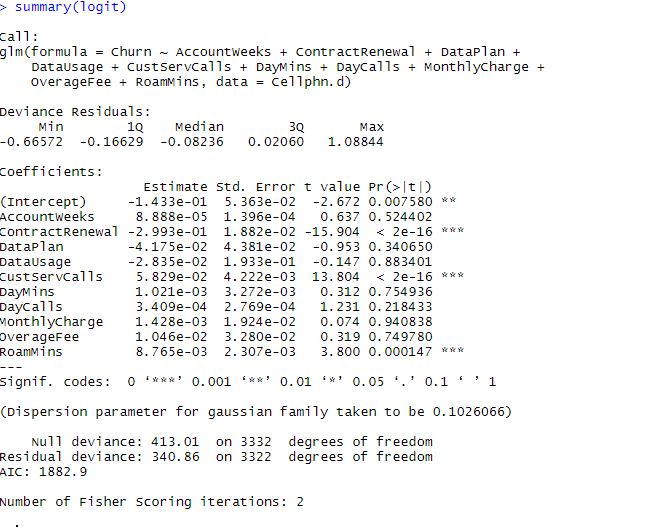
* Splitting the data into train and test (70 : 30)
* Checking the overall validity of the model using train data
* Running the model, taking into account all the variables.
* Checking the variables that are relevant and have p<.05
* Removing these variables one by one
* Running the model with final set of variables and checking how efficient is the prediction on test data

Checking overall validity of the model:

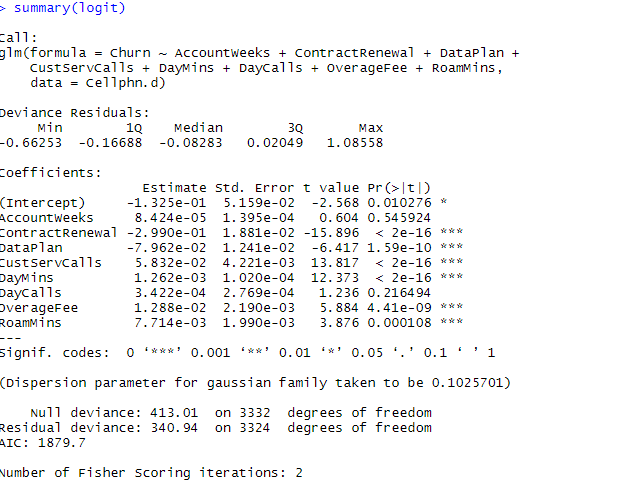


Since obtained p value <0.05, the model is valid and at least one of the coefficients of “X”is non-zero

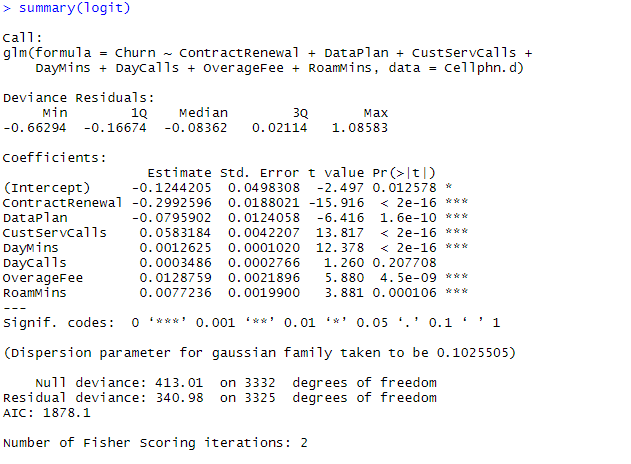
Running the model with all variables:



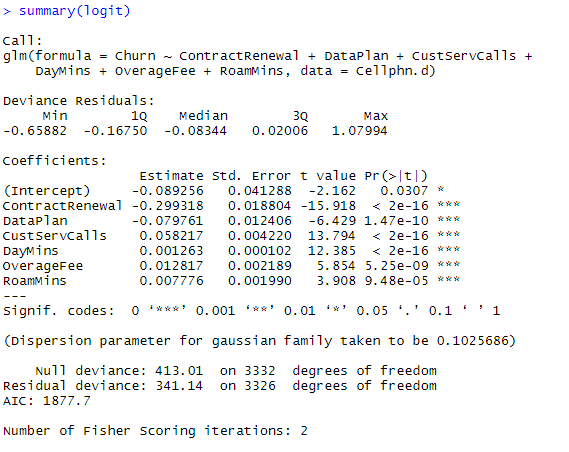
Removing data usage and monthly charge variables and checking the p values of remaining variables



Removing account weeks variable and checking the p values of remaining variables



Removing daily calls variable and checking the p values of remaining variables:



**Checking pseudo R square**



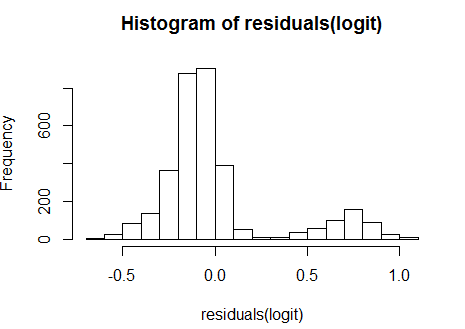
**Multicollinearity check of final set of variables**



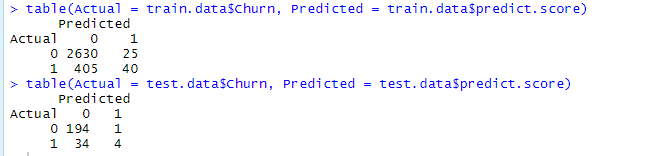
No multicollinearity detected as vif values for the values are not above 5.

**Plotting residuals**

Plot of residuals show that they are not exactly normally distributed but, a somewhat normal distribution. Eliminating the outliers from the model may normalize this graph better.



**Prediction using confusion matrix:**



**Confusion matrix interpretation for test and train dataset:**

Train dataset:

Sensitivity (Recall) = 40/(40+405) = 9%

Specificity = 2630/(2630+25) = 99.1%

Accuracy = (TP+TN)/Total = (2630+40)/(2630+40+405+25) = 86%

Misclassification rate = 14%

Precision = 40/(40+25) = 62.5%

Train dataset:

Sensitivity (Recall) = 4/(38) = 10.5%

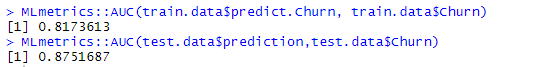
Specificity = 194/(194+1) = 99.5%

Accuracy = (TP+TN)/Total = 85%

Misclassification rate = 15%

Precision = 4/(4+1) = 80%

Area under curve:



**Final interpretation on the model and recommendation**

* The model is not suitable for predicting customers who would cancel the services (as evident from sensitivity). Hence, the model is not appropriate for predicting customer churn.
* However, the model can be used to understand which are the factors (variables) that are impacting customer churn and which are not significant at all. For instance, variables that are not significant are - data usage, monthly charge, account weeks, daily calls.
* To improve the model, next step could be removing the outliers and checking whether this leads to better prediction.
* Other models such as random forest and neural networks need to be tried out to check if they give a better prediction.