Churn Analysis – Cellphone Company

Nipun Kansal

**Table of Contents**

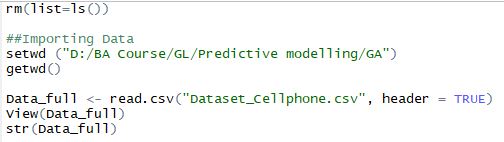
|  |  |
| --- | --- |
| S.No. | Description |
| 1 | Introduction |
| 2 | Importing data into R |
| 3 | Installing required libraries |
| 4 | Exploratory Data Analysis |
| 5 | Creating the Development and Holdout samples |
| 6 | Creating models using Logistic Regression Technique |
| 7 | Predicting churn of customers |
| 8 | ROCplots for Logistic regression models |
| 9 | Odds ratio – exp (Beta) |

**1 Introduction**

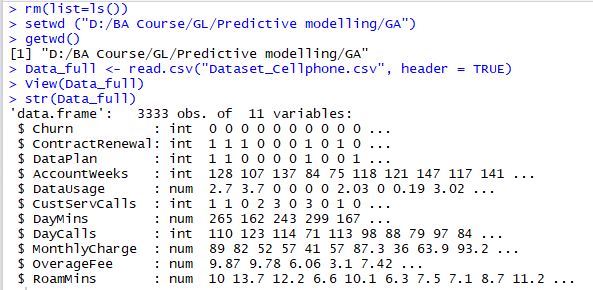
The dataset contains the data for 3333 customers of a cell phone company. There are 10 explanatory variables, with Churn being considered as the explained variable for the analysis. We need to predict whether the customer will churn or not by using logistic regression technique.

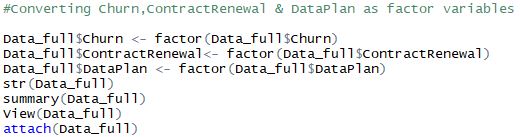
**2 Importing data into R**

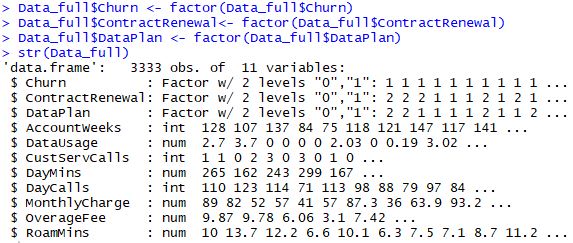
We begin by importing the data and summarizing & structuring data using R. Wherever required converting categorical explanatory variable into factors (if not).

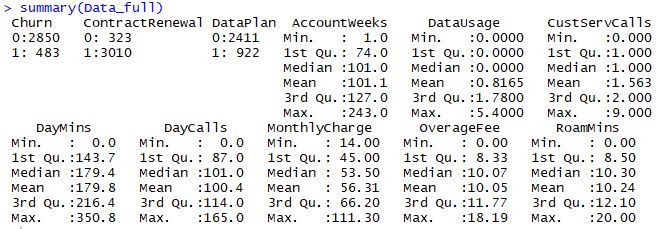


**Results:**





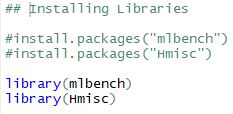




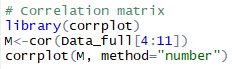
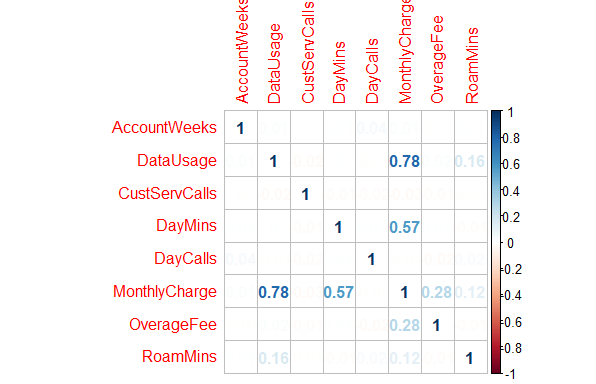
Observations:

* In the given data, out of 3333 customers 483 customers had cancelled the service while
* Out of 3333 customers 3010 customers had recently renewed the contract
* Out of 3333 customers only 922 have data plan
* For rest explanatory variables summary is mentioned above

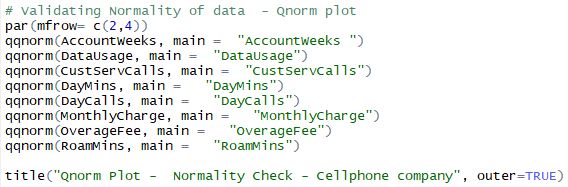
**3** **Installing required libraries**

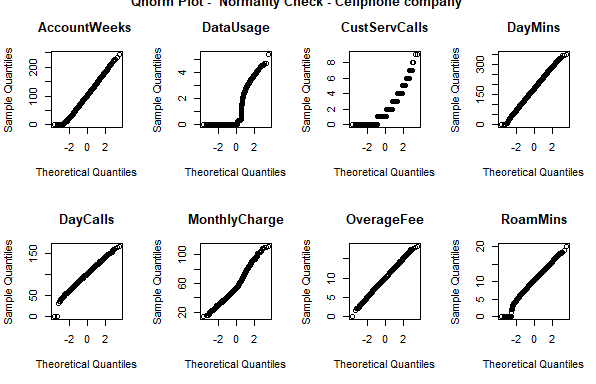
****

**4 Exploratory data analysis**

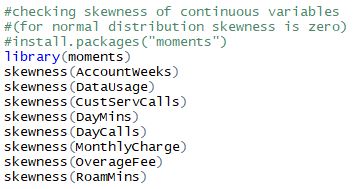
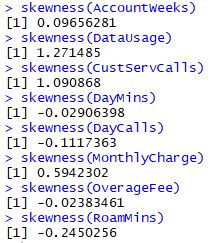
****

Correlation can be observed between the variables DataUsage, DayaMins & MonthlyCharge.

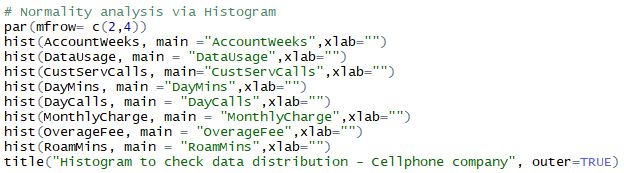


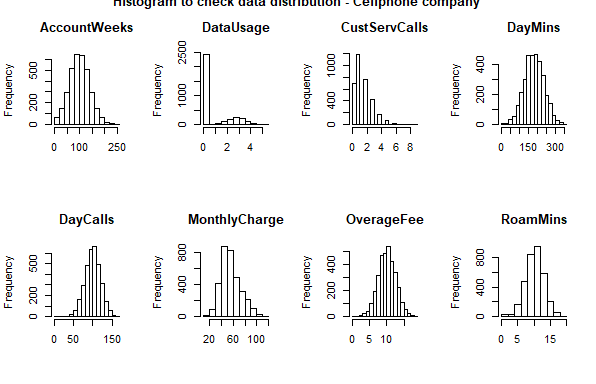


**Almost all the continuous explanatory variables are normally distributed (maybe variables like DataUsage, CustServCalls may have some discripancies)**

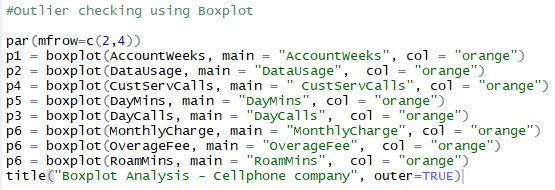
 

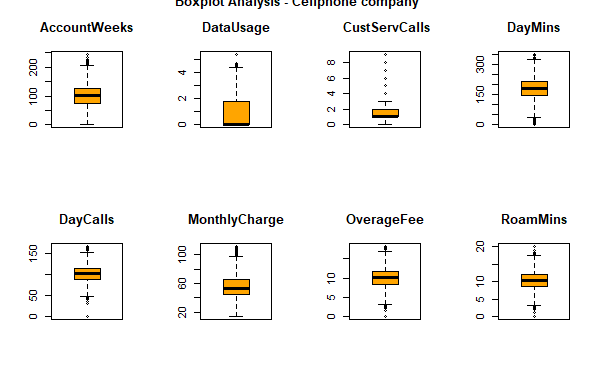
Explanatory continuous variables DataUsage, CustServCalls & MonthlyCharge are skewed positively and also the variables RoamMins is a bit negatively skewed. Rest are close to zero.





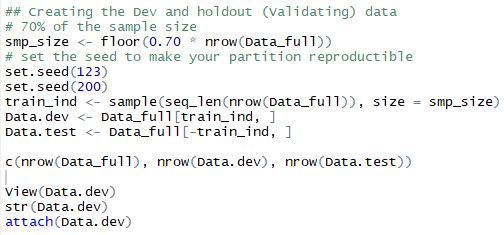
**DataUsage and CustServCalls are highly skewed variables and have normality issue.**





**From the boxplot analysis also it is observed that DataUsage and CustServCalls are not normally distributed.**

**5** **Creating Development and Holdout (Validating) Sample**

****

**Note:**

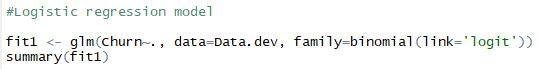
There are 2333 records in development sample and 1000 records in holdout sample

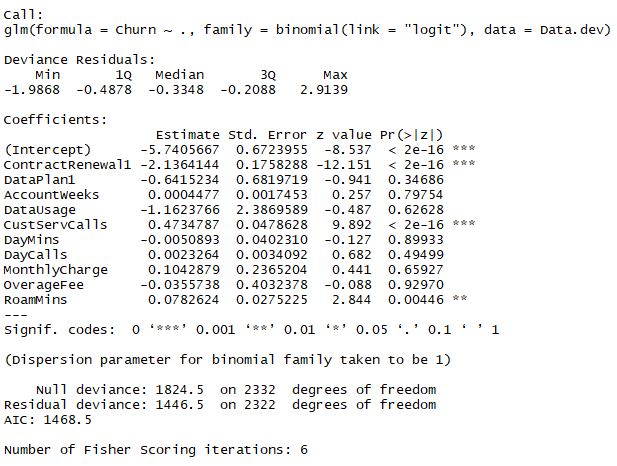
Results:

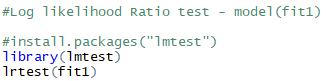


**6** **Creating models using Logistic Regression Technique**

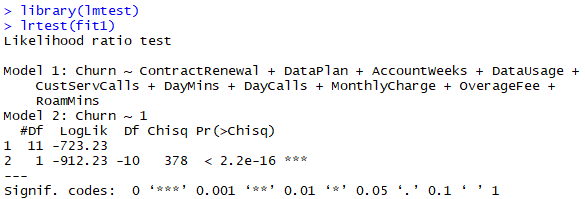
**6.1 First model**

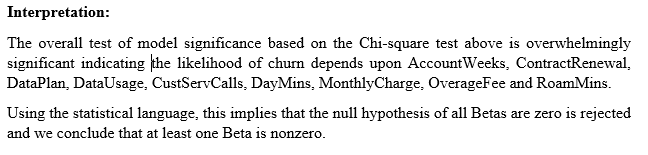


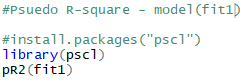




**Results:**

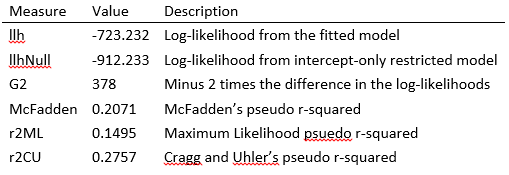


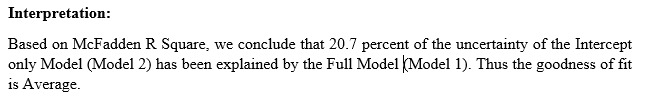




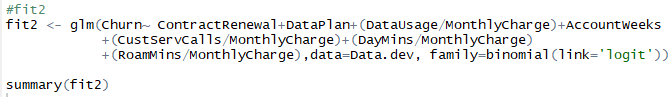
**Results:**

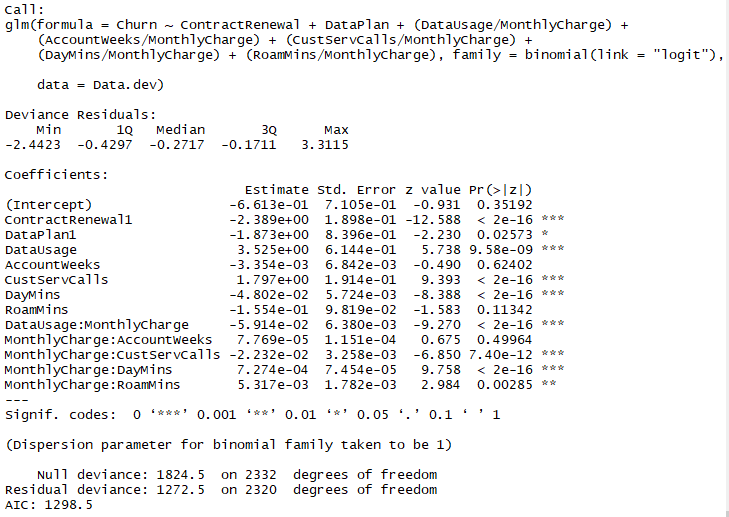


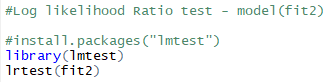




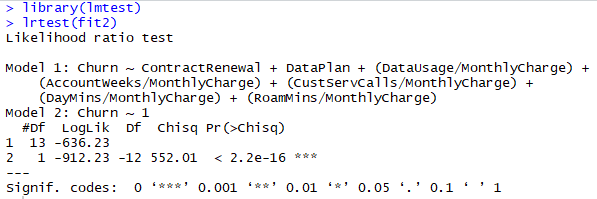
**6.2 Second model**

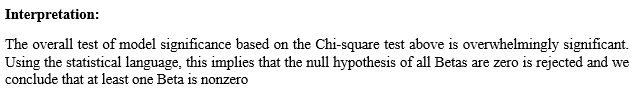


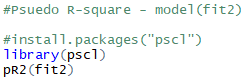




**Results:**

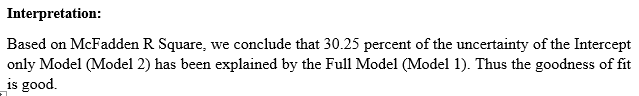


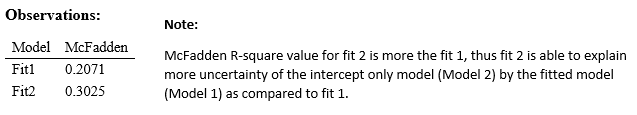




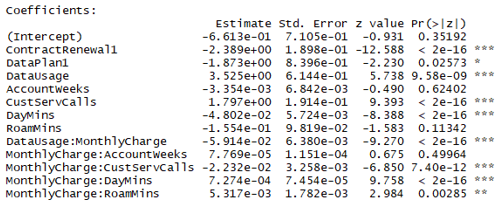
**Results:**

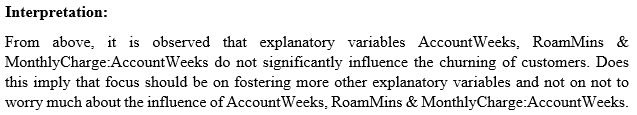






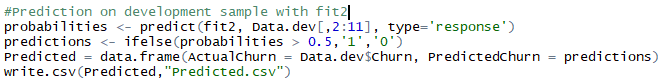
**Individual Coefficients**





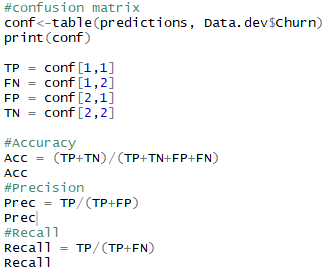
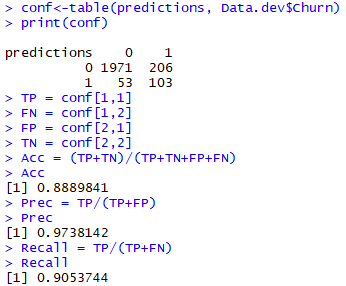
**7** **Predicting churn of customers**

**7.1 On Development sample (Data.dev)**

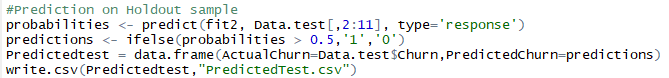


[****](Predicted.csv)

**Computing the prediction Accuracy, Precision and Recall**

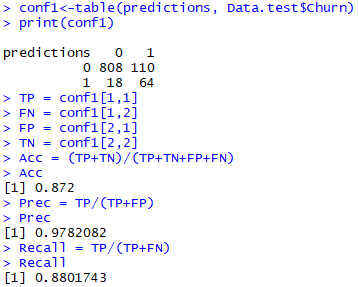
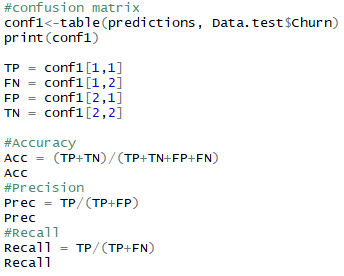
 

**7.2 On Holdout (Validating) sample (Data.test)**



[](PredictedTest.csv)

**Computing the prediction Accuracy, Precision and Recall**



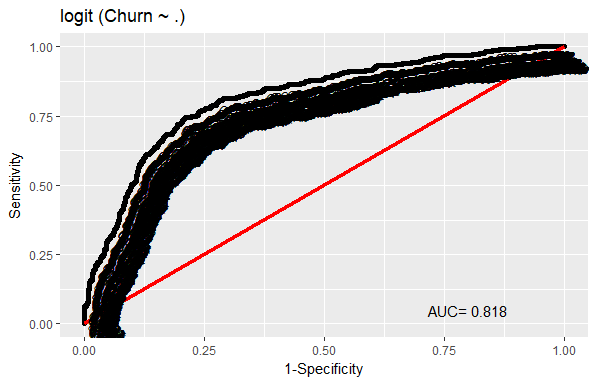
**Table 1:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Measurement** | **Data.dev** | **Data.test** |
| **1** | Accuracy | 0.8898 | 0.872 |
| **2** | Precision | 0.9738 | 0.9782 |
| **3** | Recall | 0.9053 | 0.8801 |

From above table it can be concluded that Fit2 is a good fit on holdout sample also.

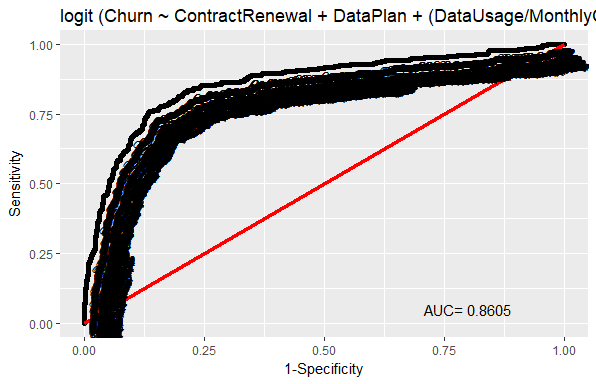
**8** **ROCplots for Logistic regression models**

**8.1 ROCplot for fit1**

****

AUC for fit1 model is 81.8%, which is good.

**8.1 ROCplot for fit2**

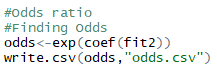
****

AUC for fit2 model is 86.05%, which is good.

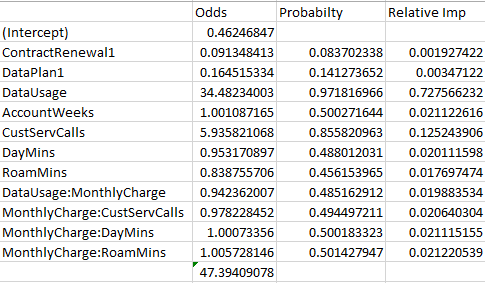
**Observation:**

The measure, AUC value of 87.54% for fit2 model is greater than AUC value of 81.8% for fir1 model. Thus fit2 model is better model for prediction.

**9** **Odds ratio – exp (Beta)**

[****](odds.csv)

**Results:**



**Interpretation:**

If DataUsage increases by one unit, the odds for not