

**Group Assignment**

**Predictive Modeling- using R**

**Post-Graduate programme in Data Science & Business Analytics**

**Batch-October 2019.**

**By: Group 4**

**Ashish Nath Km**

**Athul Aravind PM**

**Indu Mohan**

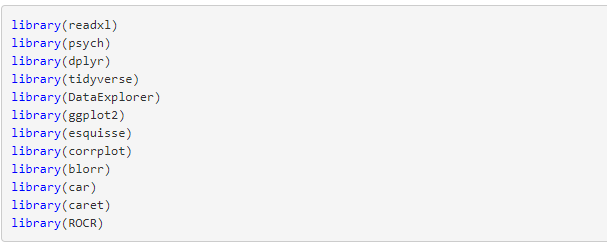
**Karthik R**

**Nishitha Ramesh**

**CELLPHONE-CUSTOMER CHURN PREDICTION CASE STUDY**

This case study attempts to build a model that will help the Telecom company predict the customers who have a higher probability of churning.

For the analysis of the problem of Cell phone dataset, we have utilised the below R-packages from the library.



We also need to understand the meaning of the variables from the given dataset.

**Details of the 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 |

Few important terminologies from the point of view of analysing this case study are as follows.

**Logistic Regression**

[Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic regression is named for the function used at the core of the method, the logistic function.

The [logistic function](https://en.wikipedia.org/wiki/Logistic_function), also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

Where e is the [base of the natural logarithms](https://en.wikipedia.org/wiki/E_(mathematical_constant)) (Euler’s number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.



Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b’s).

Logistic regression is a linear method, but the predictions are transformed using the logistic function. The impact of this is that we can no longer understand the predictions as a linear combination of the inputs as we can with linear regression, for example, continuing on from above, the model can be stated as:

p(X) = e^(b0 + b1\*X) / (1 + e^(b0 + b1\*X))

The coefficients (Beta values b) of the logistic regression algorithm must be estimated from your training data. This is done using maximum-likelihood estimation.

[Maximum-likelihood estimation](https://en.wikipedia.org/wiki/Maximum_likelihood) is a common learning algorithm used by a variety of machine learning algorithms, although it does make assumptions about the distribution of your data (more on this when we talk about preparing your data).

The best coefficients would result in a model that would predict a value very close to 1 (e.g. male) for the default class and a value very close to 0 (e.g. female) for the other class. The intuition for maximum-likelihood for logistic regression is that a search procedure seeks values for the coefficients (Beta values) that minimize the error in the probabilities predicted by the model to those in the data (e.g. probability of 1 if the data is the primary class).

**Model Performance Measures:**

Machine Learning fundamentally differentiates itself from classical statistics by not assuming that the data comes from a specific model. Hence ML, justifiably, can try different models on a given dataset to eventually pick the **best** one. Obviously, to do this one needs to define what it means to be **best.**

Different “model performance measures” exist and any of these can be used to compare models - largely depending on the context and the kind of output:

* + Regression outputs are continuous numbers
  + Classification outputs are either
    - Class output (from algorithms like SVM and KNN that usually give a classification) or
    - Probability output (from algorithms like Logistic Regression, Random Forest that can give probability outputs)

There are various model performance measures as stated below and we have used Confusion Matrix, ROC and AUC.

* Confusion Matrix
* ROC Curves, Gini Coefficient
* Gain and Lift Chart
* Kolmogorov-Smirnov (K-S) chart
* Concordance-Discordance ratio
* Root Mean Square Error, Mean Absolute Error

**Confusion Matrix:**

* For classification problem with a class output, the confusion matrix gives the counts of correct and erroneous predictions:

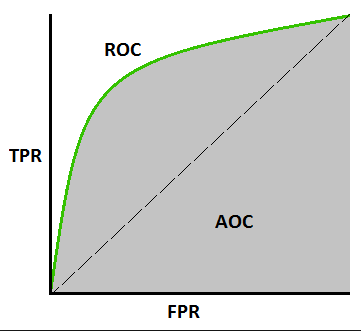


* + **Classification Error Rate:** sum of Type 1 (FP) and Type 2 (FN) Errors (in percentage). Accuracy is 1-(error rate)
  + **Sensitivity** (also called Recall or True Positive Rate): proportion of Total Positives that were correctly identified
  + **Specificity** (also called True Negative Rate): proportion of Total Negatives that were correctly identified

**ROC and AUC:**

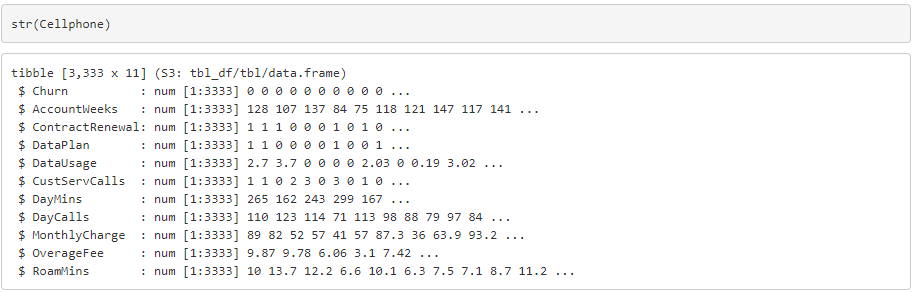
For classification problems with probability outputs, a **threshold** can convert probability outputs to classifications. This choice of threshold can change the confusion matrix. As the threshold changes, a plot of the false positive rate vs true positive rate is called the ROC curve - Receiver Operating Characteristic.

* Area under the curve (AUC) is a measure of how good the model is

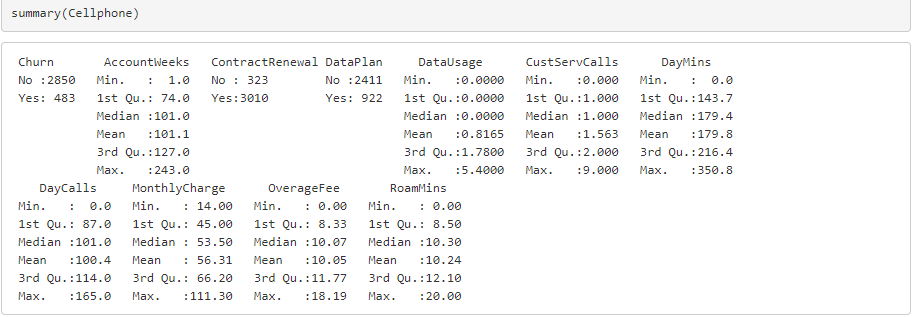


1. **Exploratory Data Analysis (Data Summary, Univariate and Bivariate Analysis)**

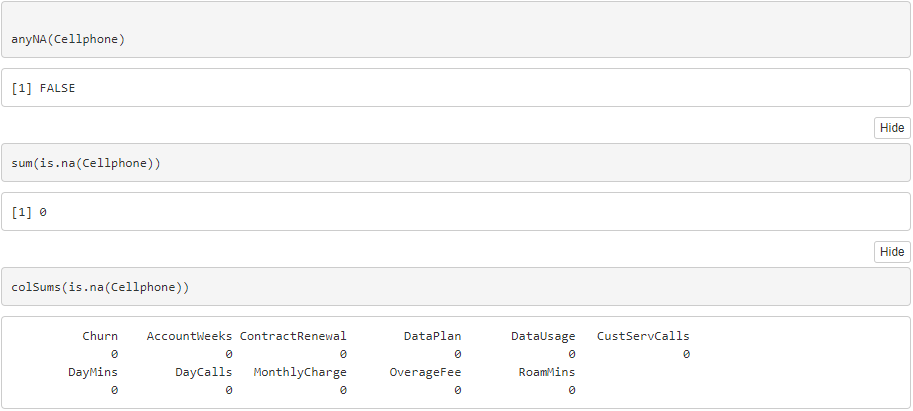
Structure of the data  
Observations: All variables are of num datatype. As variables like ‘Churn’, ‘Contract Renewal’ and ‘Data Plan’ are of Categorical type, this needs to be converted to factors.



Summary of Data  
Once the required variables are converted to factors, the summary of data set looks like below.



Check for NA Values:  
We can observe that there are no NA values in the Cellphone dataset.



Univariate Analysis and Bivariate Analysis:

Univariate and Bivariate analysis on the variables to understand the distribution and empirical relationship between the variables respectively.

CHURN

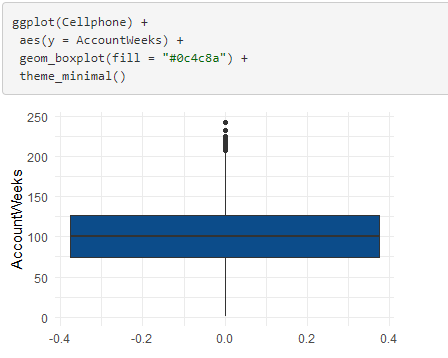
14.5% is the Churn Rate.

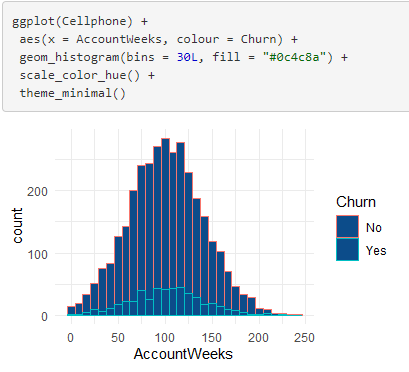


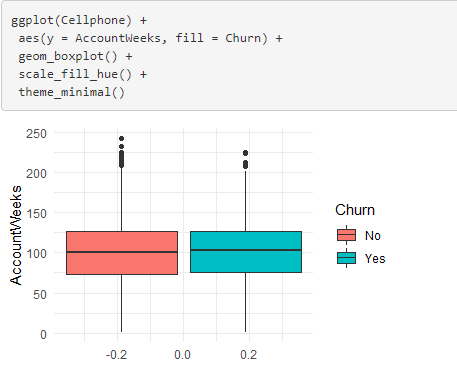
Accounts Week

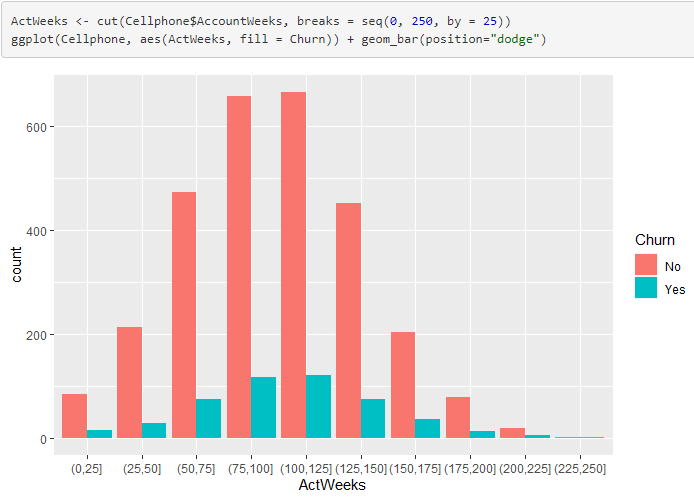
**Insight**

1. There are few outliers on the higher end of the data, i.e., there are few active customers who have been availing the services from a long time.
2. With the increase in the time (accounts week) of an account, one would expect a decreasing churn rate, but there is no clear trend visible





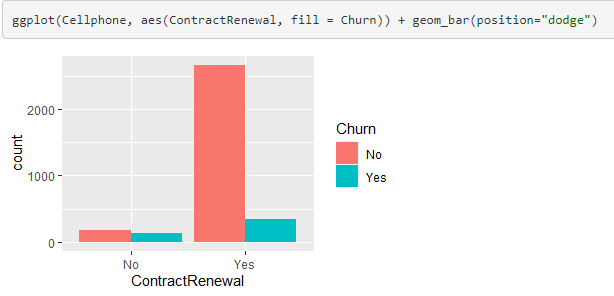




Contract Renewal

**Insight**

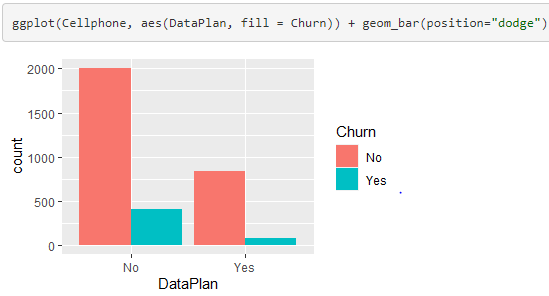
There is a good probability of an account churning if the account has not been renewed.



Data Plan

**Insight**

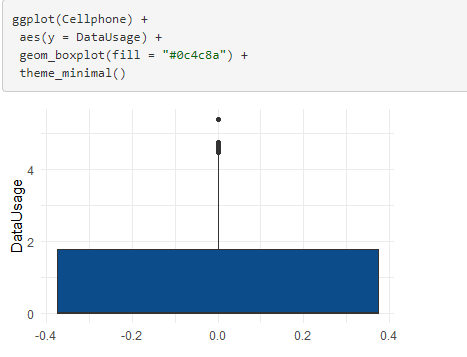
The probability of an account churning is higher if the account has not subscribed for Data Plan.

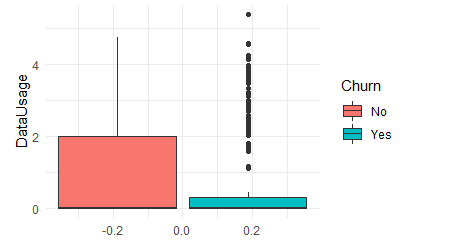


Data Usage

**Insight**

1. 50% of the customers do not use Data and there are few customers who use more than 4GB data
2. The probability of customers churning is higher when they aren’t using Data

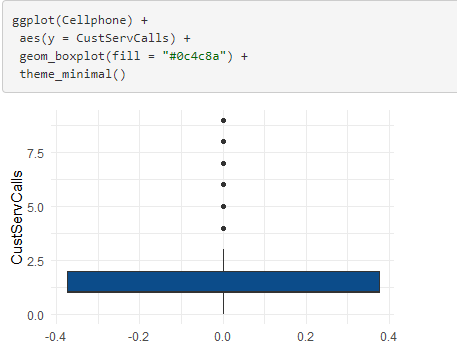


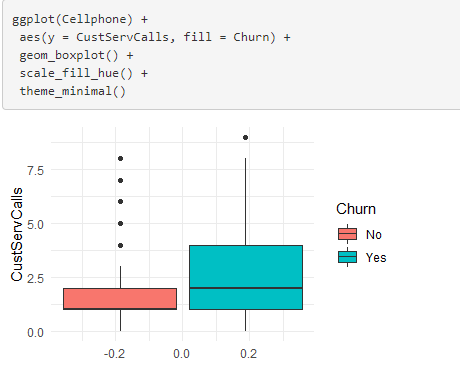


Customer Service Calls

**Insight**

1. There are few outliers on the higher end of the data, which means few customers have called the Customer service for more times when compared to others
2. Churn Rate increases if a customer has called for more than 4 times

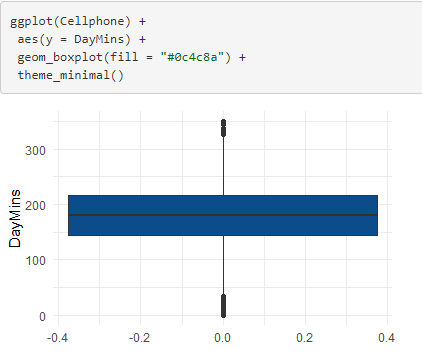


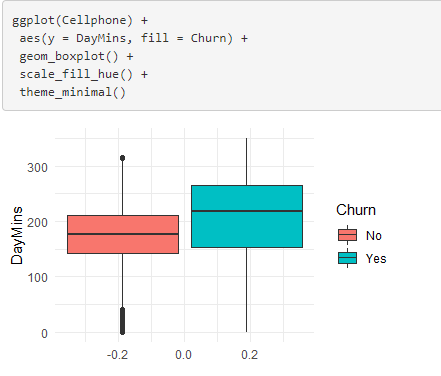


Day Minutes

**Insight**

1. There are outliers on both lower and higher end of the data, which means there are customers who uses the service for very less minutes and customers who tend to use it for much longer minutes.
2. Customers who tend to churn have greater monthly average daytime minutes.

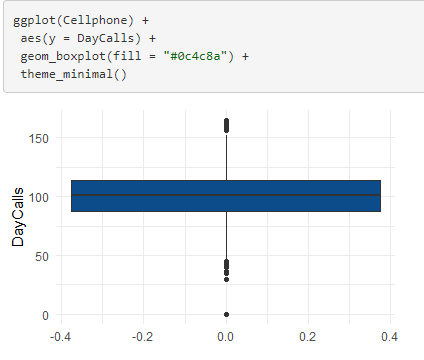


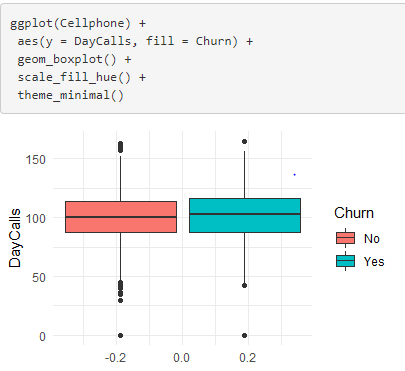


Day Calls

**Insight**

1. There are outliers on both lower and higher end of the data, which means there are customers who makes very few calls and customers who tend to make more calls.
2. There is no clear churn pattern visible from Day Calls.

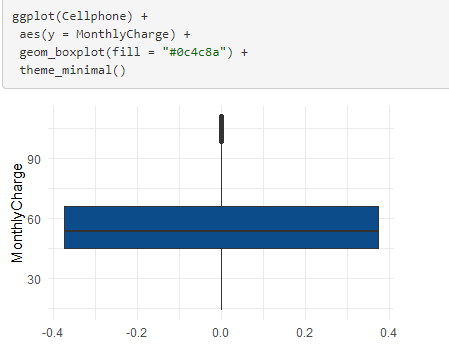


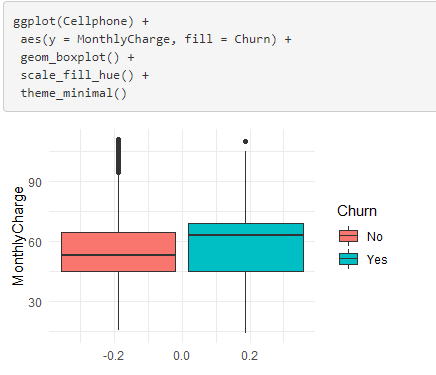


Monthly Charge

**Insight**

1. There are outliers on the higher end of the data, i.e., there are customers who have an average monthly bill that is higher than other customers.
2. Customers who tend to churn, comparatively have a greater average monthly bill.

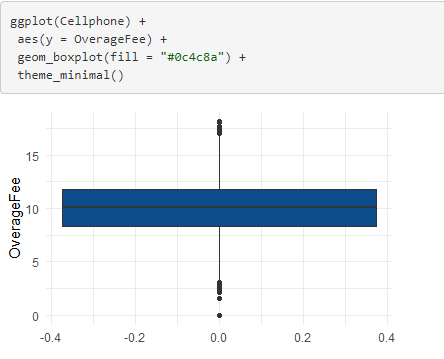


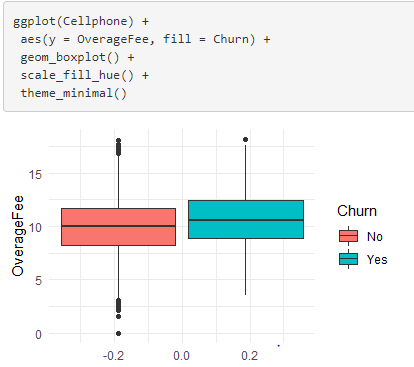


**Overage Fee**

**Insight**

1. There are outliers on both the lower and higher end of the data, i.e., there are few customers who have zero or very less overage fee and there are few customers to have a higher overage fee.
2. There is no clear churn pattern visible, but comparatively customers with higher overage fee tend to churn.

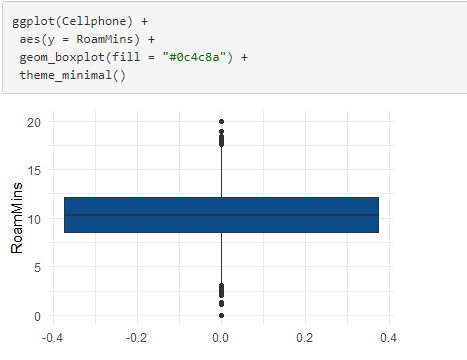


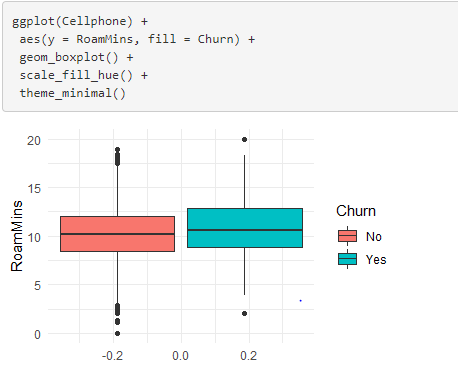


**Roam Minutes**

**Insight**

1. There are outliers on both the lower and higher end of the data, i.e., there are few customers who have zero or very less Roaming Minutes and there are few customers to have a higher Roaming Minutes.
2. There is no clear churn pattern visible wrt overage fee.

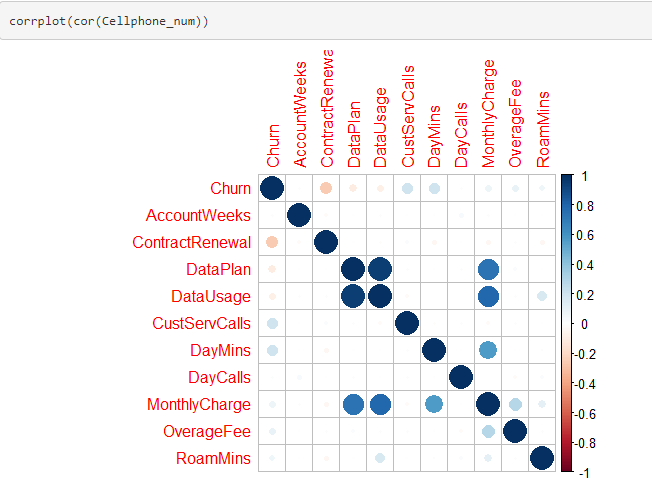


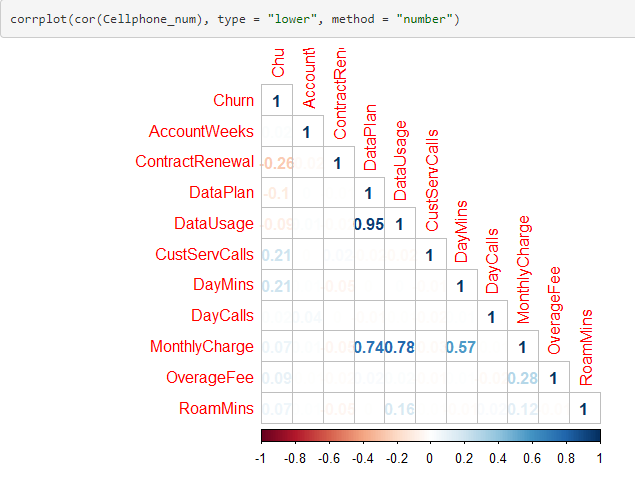


Correlation and Multicollinearity:

Churn does not seem to be highly correlated with any of the variables. Churn has maximum correlation with Contract Renewal, CustServCalls and DayMins.

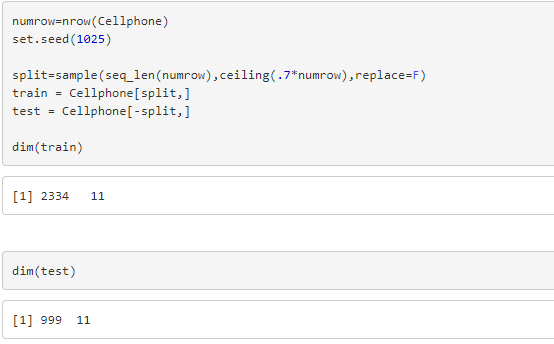
There is a high multicollinearity between Data Usage and Data Plan. Monthly Charge also has high multicollinearity with Data Plan, Data Usage and Day Mins.



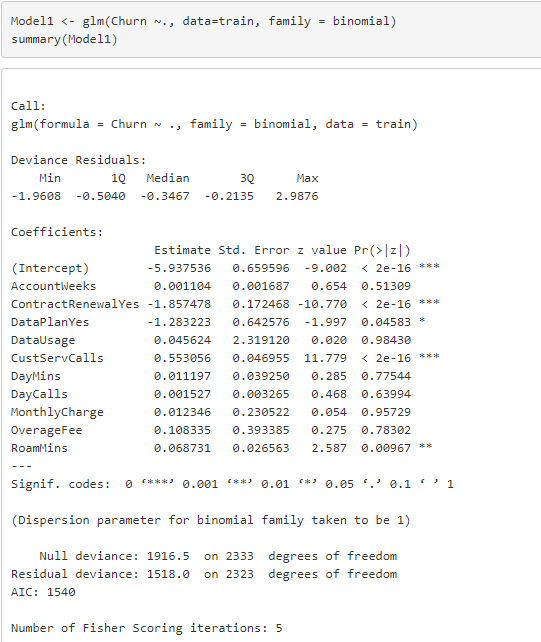


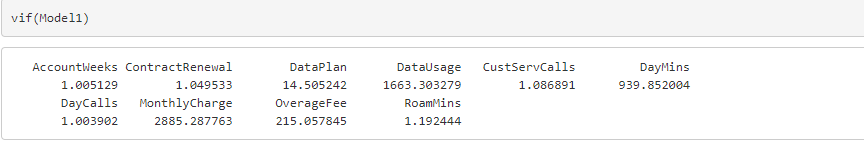
MODEL

**Splitting the data into Train & Test set**



**Applying Logistic Regression on the Full Model**

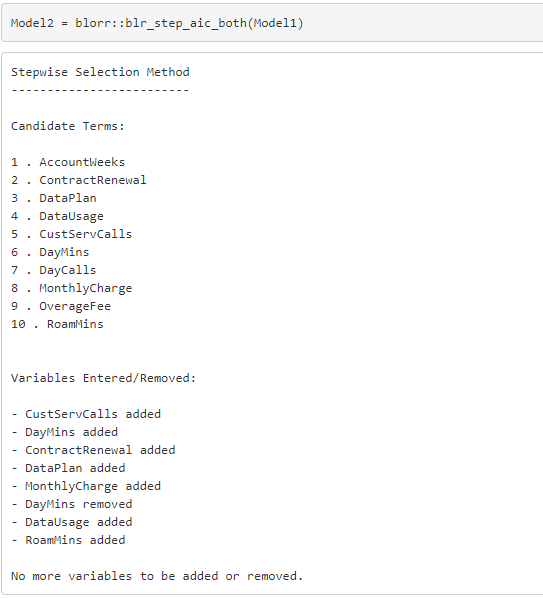


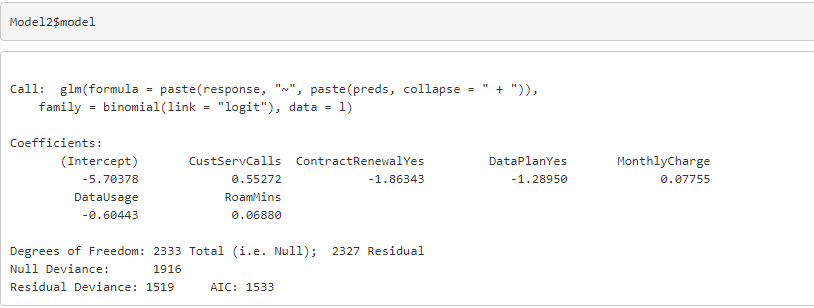


The individual p-values in the model suggests the significance of the variables. AccountWeeks, DataUsage, DayMins, DayCalls, MonthlyCharge, OverageFee has higher p-values.

The high multicollinearity has resulted in inflated vif values for correlated variables which makes the model highly unreliable.

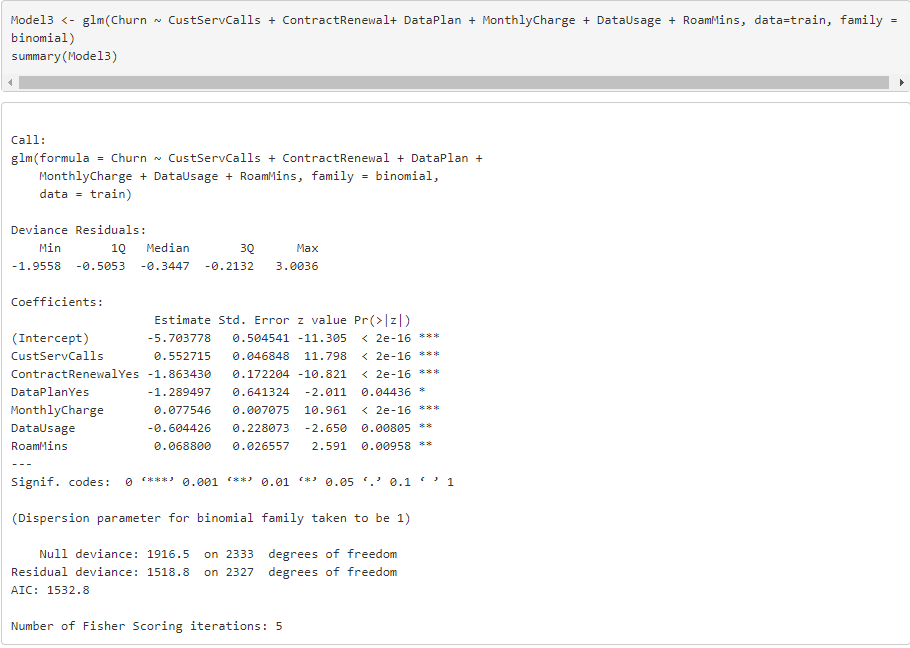
To solve this, we will use a step wise regression containing both forward and backward selection, which will calculate the variable importance of each variable in both forward and backwards using blorr package.





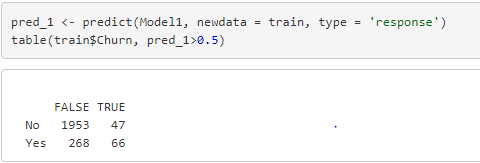
This also suggests the significant variables that needs to be taken into account for Model building.

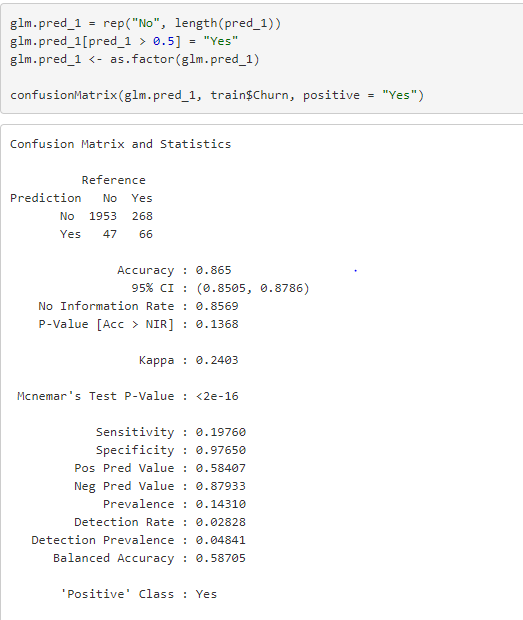
Now, let us build the Model with the significant variables.



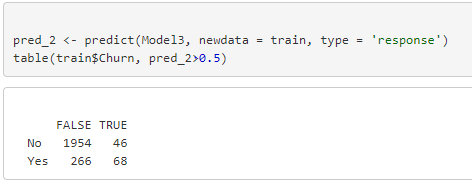
**Prediction on Train Data**

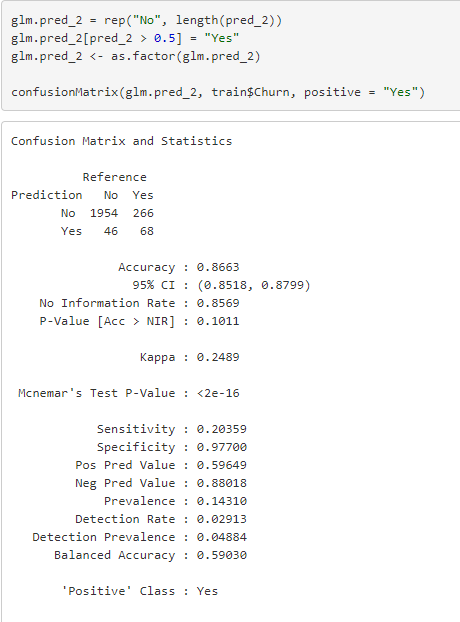
**For Model1 – Full Model**





**For Model3 – Significant Variables**





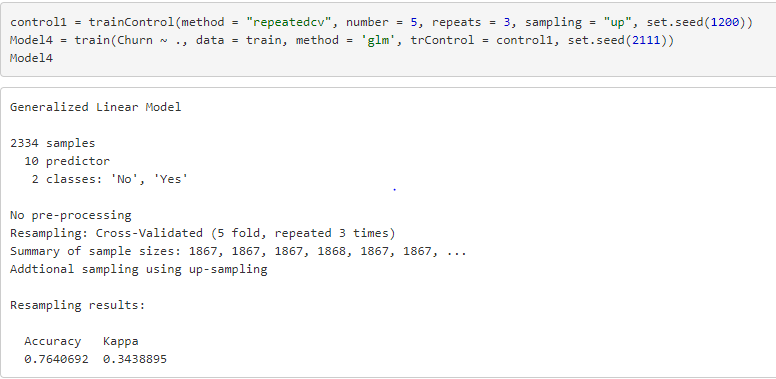
Though the accuracy of the Model is 86.6%, the sensitivity of the model is very less 20%.

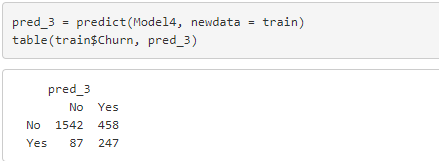
This could be because of the huge class imbalance, i.e., there is more data available for Customers not Churning and the Model is able to predict Customers who do not churn in a better way. Hence the Specificity of the Model is very high 97%. But there is less data available for Customers who churn, therefore the model isn’t able to predict this well.

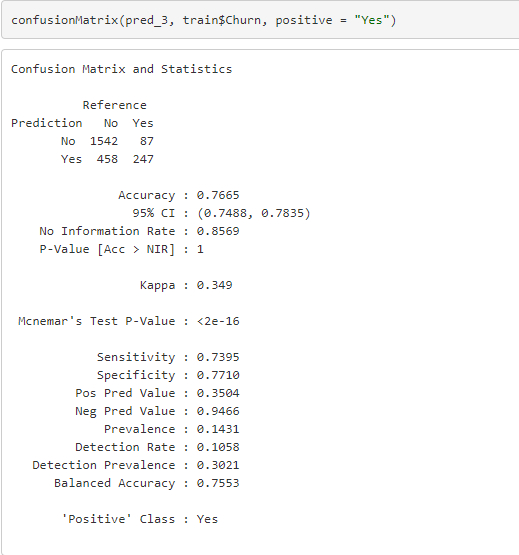
Let us try some sampling techniques to deal with the class imbalnce using the caret package.



**Up-sampling for Full Model**



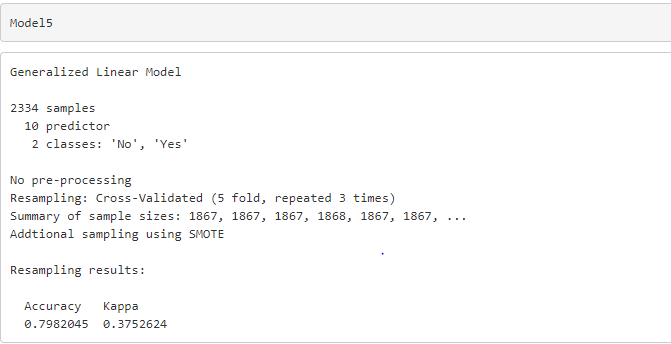


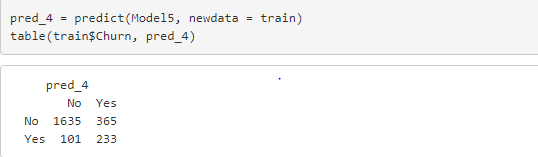


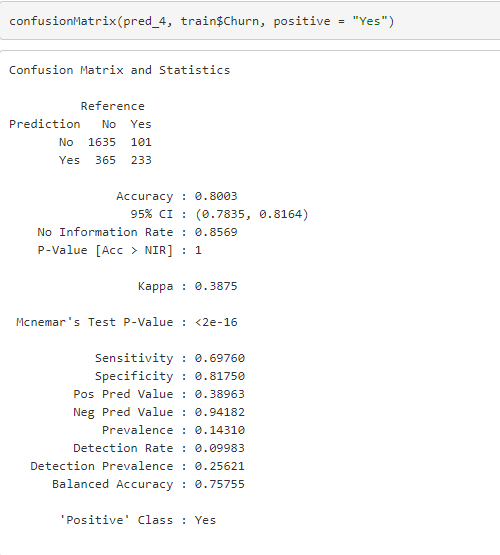
With Up-sampling using the Full Model for the Train data, we have got an Accuracy of 76.65%, Specificity of 77.1% and the Sensitivity has improved from 20% to 73.95%.

**Smote-Sampling for Full Model**



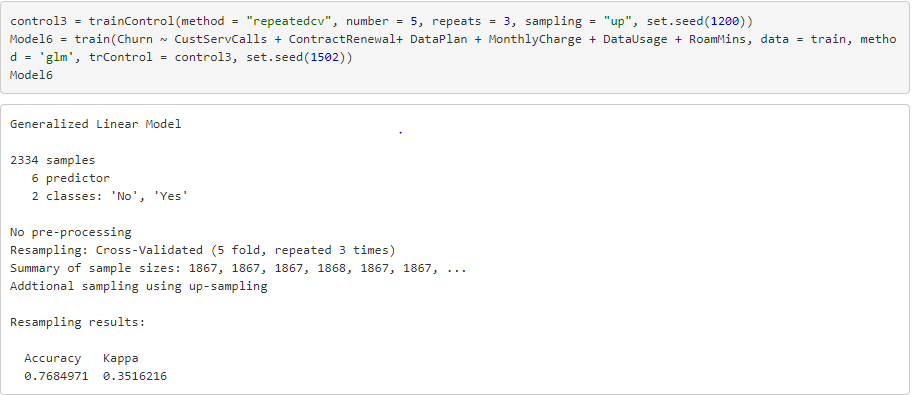


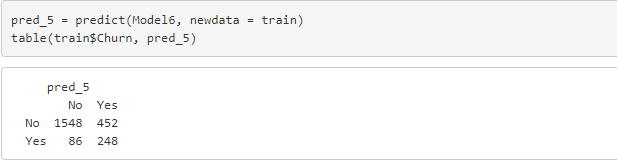


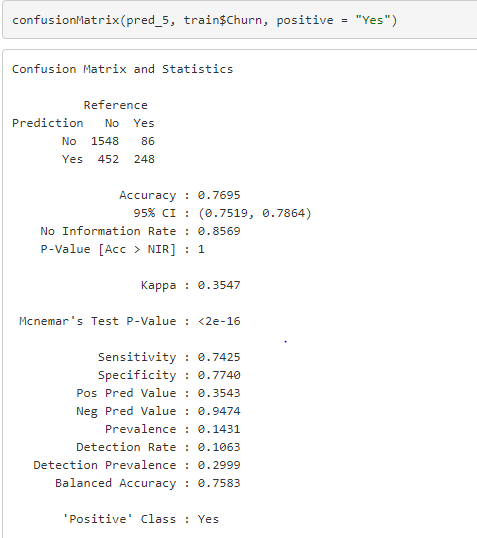


With Smote-sampling using the Full Model for the Train data, we have got an Accuracy of 80%, Specificity of 81.7% and the Sensitivity of 69.76%.

**Up-Sampling for Significant Variables Model**

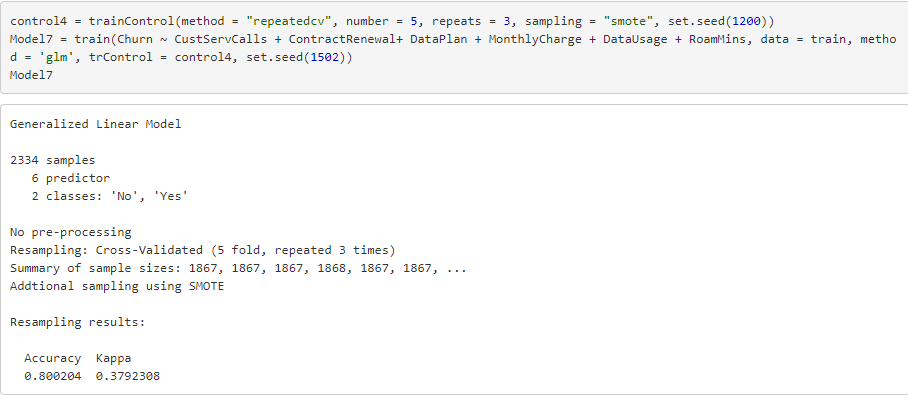


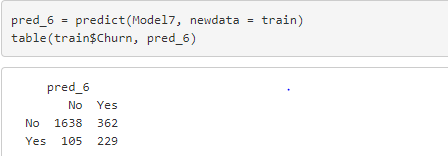


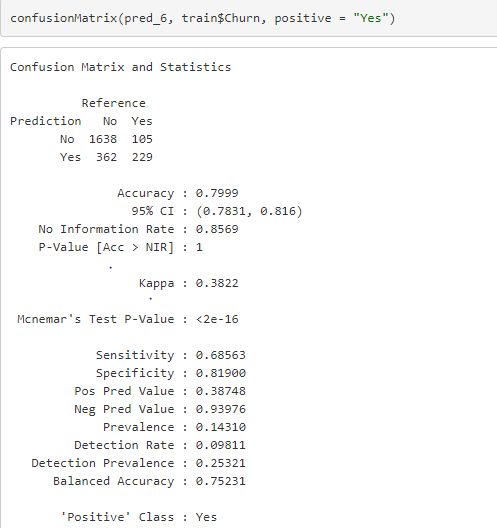


With Up-sampling using the Significant variables Model for the Train data, we have got an Accuracy of 76.95%, Specificity of 77.4% and the Sensitivity of 74.25%.

**Smote-Sampling** **for Significant Variables Model**





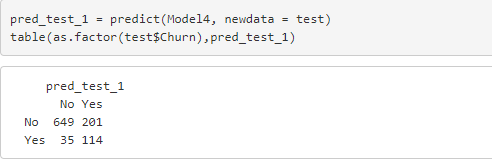


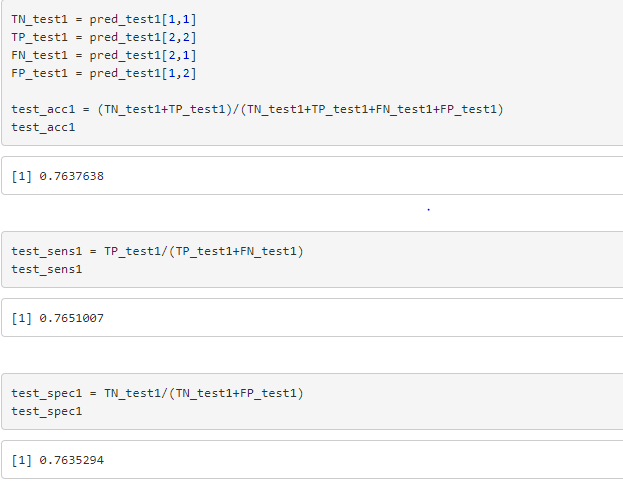
With Smote-sampling using the Significant variables Model for the Train data, we have got an Accuracy of 80%, Specificity of 81.9% and the Sensitivity of 68.56%.

**Prediction on Test Data**

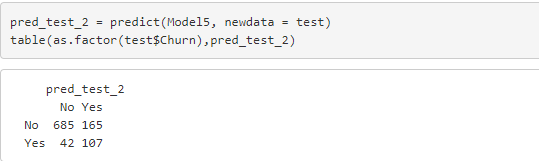
**Prediction on Test data for 4 Models obtained from Up & Smote sampling for Full Model and Significant Models.**

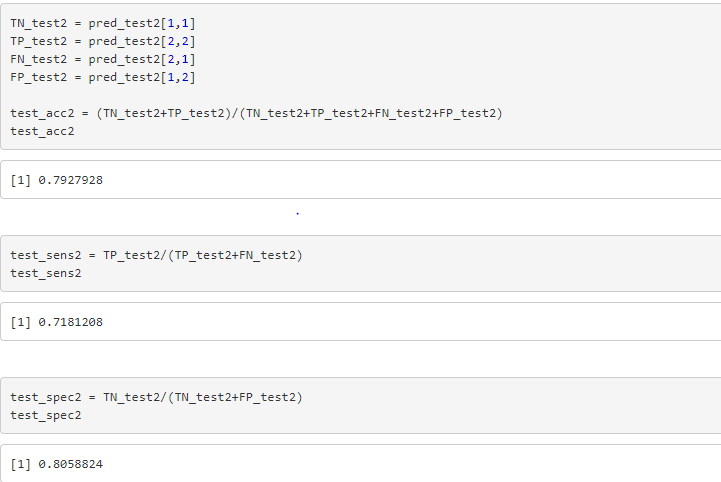
Prediction on the Up-sampling -Full Model for the Test data, we have got an Accuracy of 76.37%, Sensitivity of 76.5% and the Specificity of 76.35%



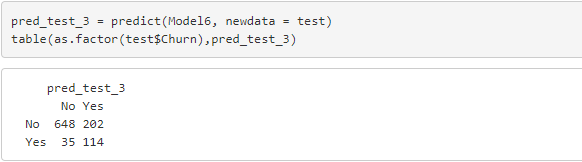


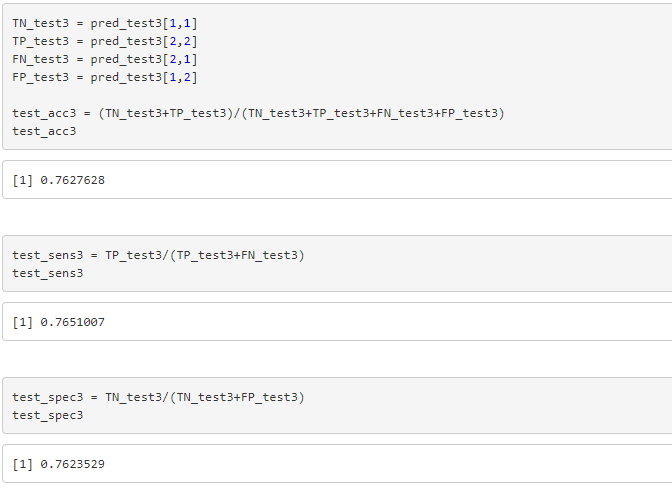
Prediction on the Smote-sampling -Full Model for the Test data, we have got an Accuracy of 79.27%, Sensitivity of 71.81% and the Specificity of 80.5%



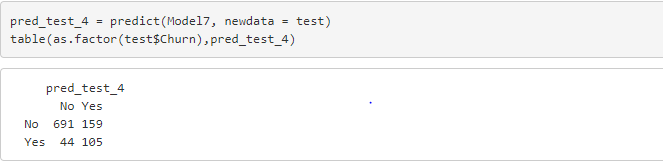


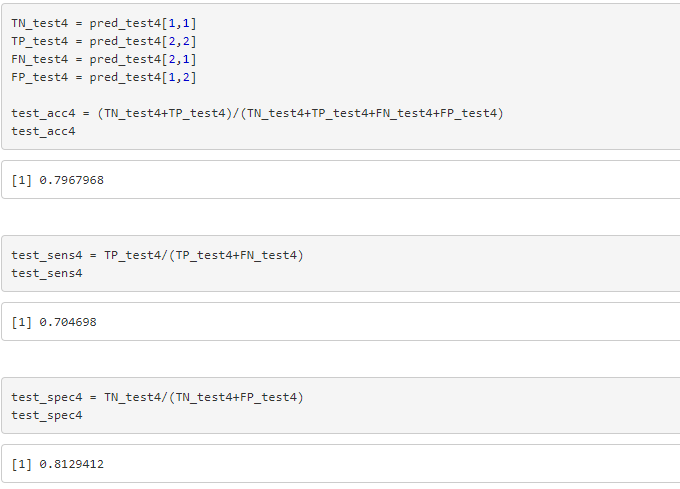
Prediction on the Up-sampling -Significant Model for the Test data, we have got an Accuracy of 76.27%, Sensitivity of 76.51% and the Specificity of 76.23%





Prediction on the Smote-sampling -Significant Model for the Test data, we have got an Accuracy of 79.67%, Sensitivity of 70.46% and the Specificity of 81.29%

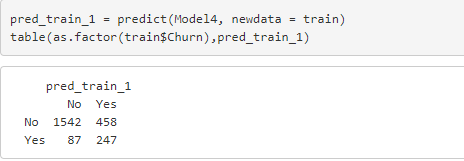




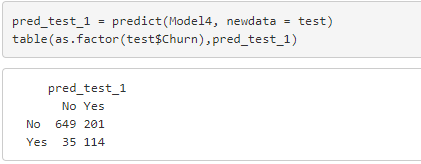
**Model Performance**

**Confusion Matrix**

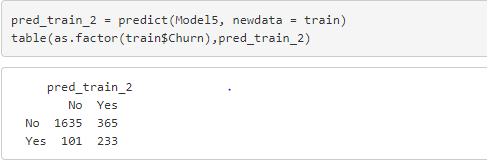
**Up-sampling Full Model Train Data**



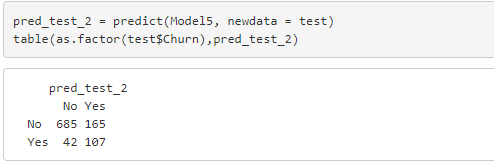
**Up-sampling Full Model Test Data**



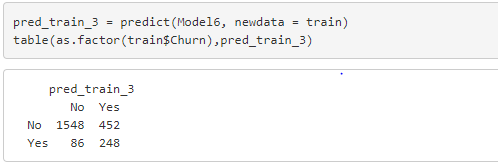
**Smote-sampling Full Model Train Data**



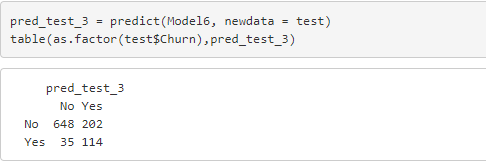
**Smote-sampling Full Model Test Data**



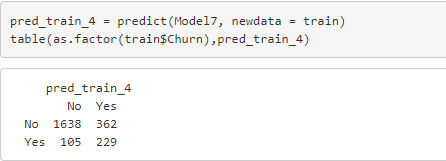
**Up-sampling Significant Model Train Data**



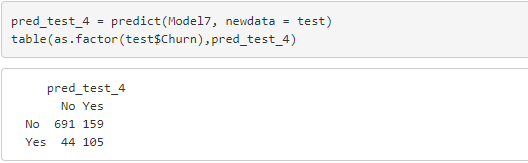
**Up-sampling Significant Model Test Data**



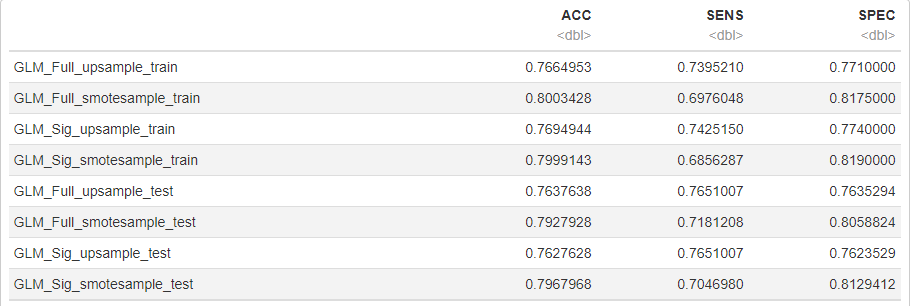
**Smote-sampling Significant Model Train Data**



**Smote-sampling Significant Model Test Data**

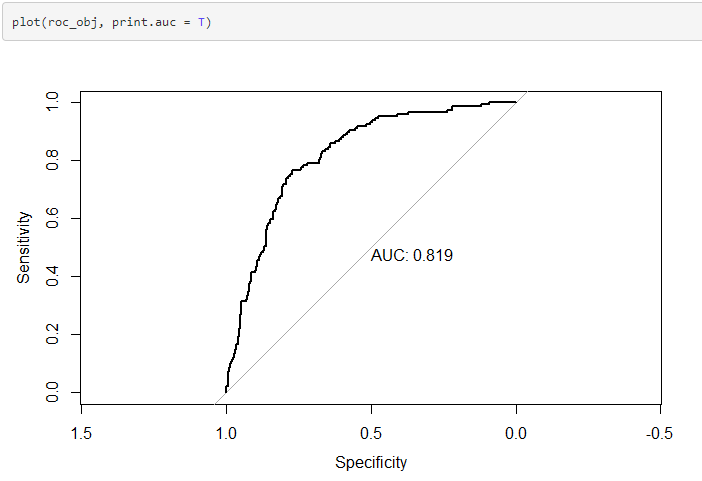


**Calculated Accuracy, Sensitivity and Specificity for Logistic Regression Full Models and Significant Models**

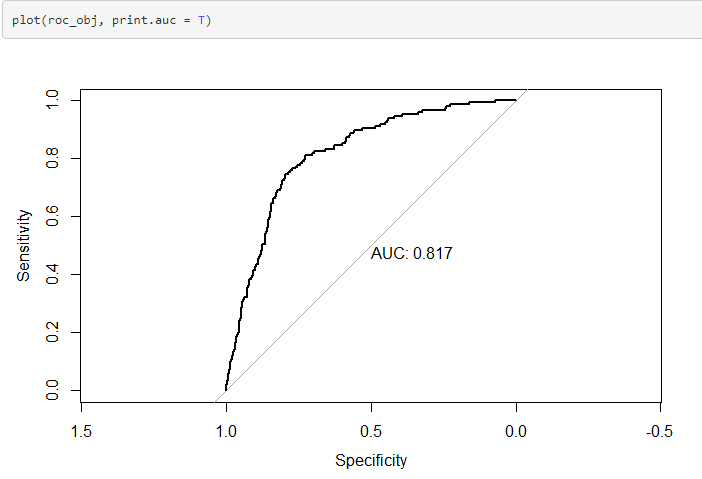


**ROC and AUC for Full Models and Significant Models (Test Data)**

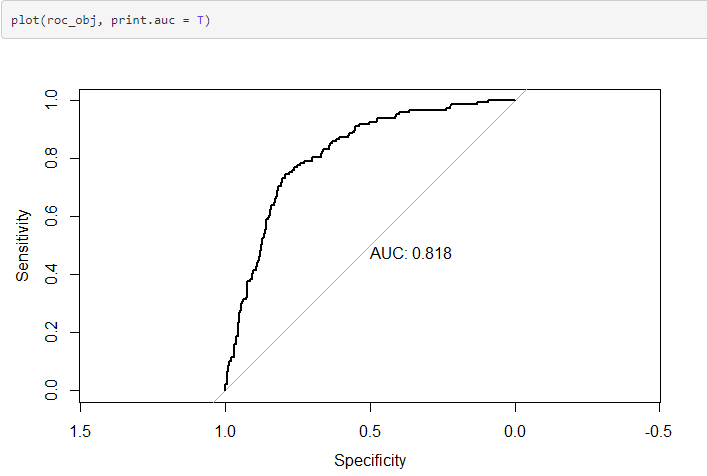
**Model 4 – Up-sampling Full Model**



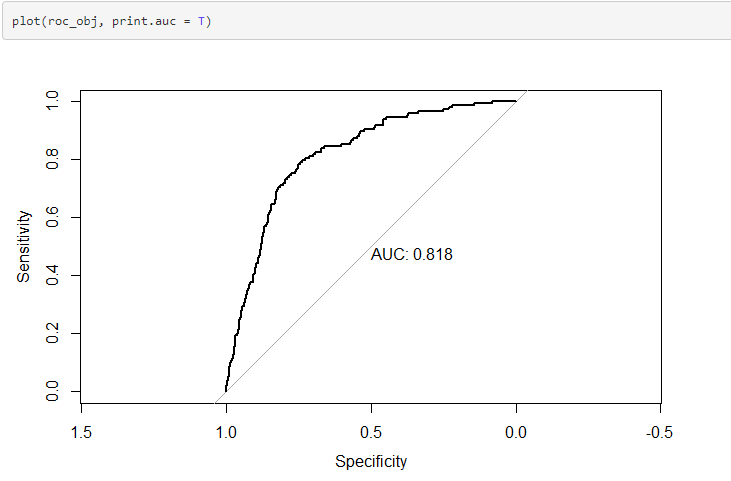
**Model 5 – Smote-sampling Full Model**



**Model 6 – Up-sampling Significant Model**



**Model 7 – Smote-sampling Significant Model**



**From the above Confusion Matrix, Calculated Accuracy, Sensitivity and Specificity for the Test model, we can observe that GLM Up-sampling Significant variables Model has performed the best**

**Also, from the ROC curve the AUC is around 81.8 for all the models.**

**The initial sensitivity of 20% has increased to 76.5% after the sampling.**

**As, this model predicts the customers who churn, we need a greater sensitivity and thus the accuracy of initial 80% is compromised.**

**Telecom company can concentrate on promoting the customers to renew the contracts and to opt for Data Plans because there is a lower Churn rate for Customers who renew contracts and use Data Plans**