**Churn Analysis Report**

**Objective**:

To obtain a Logistic Regression Model of the Telecom data which includes various attributes of customers. Attributes such as their active calling plans & different tariffs. This model will allow us to predict customers who will opt for contract renewal (Churn=0) and customers who will not (Churn=1) so that it will help us to make better business decisions to retain the Churned customers.

**Approach**:

Building a model with all variables isn’t a feasible way because the variables comes with multi-collinearity meaning, they are dependent on each other. To get a good accurate model, we need to build the model based on certain variables which are significant. Also, we need to partition our data into two parts for training & testing purposes.

*Steps taken for selecting significant variables for our model:*

1. The data is divided into 2 parts. 80% of the data is used for training and 20% is used for testing.
2. Fit1 is created in relation to Churn and ‘1’ variable to check the number of significant variables.
3. Fit2 is created in relation to Churn and all the variables.
4. For further reduction of Fit 2 to obtain the most significant variables, Step function is used.
5. After the Step executed, we have obtained the variables with most significance. We can now proceed building the model Fit3 with variables obtained.
6. Summary of the Fit3 indicates all the variables used are significant.
7. Now we need to build a confusion matrix which will indicate the accuracy of the model Fit3 but before that we need to obtain the Predicted values in terms of 0 & 1.

*Steps taken for predicting the Churn:*

1. Apparently, Churn is a dichotomous (binary) variable. Its value can either be 1 or 0. So prediction has to be made based on these values.
2. Next step is to store the predicted values which are greater than 0.5 (test threshold) as 1 and less than 0.5 as 0.
3. A Confusion Matrix is created which will allow us determine the Accuracy of the model based on the Predicted values of the model.
4. I ran three iterations with 0.4, 0.5, 0.6 as probability thresholds to check the Accuracy variations.
5. The best probability threshold obtained was 0.5 with 86% Accuracy.
6. Plotted ROC curve to determine the area under the curve.

*Steps taken for Tableau visualisations:*

1. Created a Final\_model with all the significant variable observed from the R code above.
2. Final\_model Accuracy is created with the help of Confusion Matrix function in R.
3. Plotted Actual Churn, Final\_model (predicted churn) & Accuracy together against Phone numbers to know which Phone numbers might churn.
4. For the Odds sheet, created a filter with all the measure values & threshold parameter.
5. Created a new Model with R code including all the significant variables.

**Observations:**

Significant variables:

1. International Plans
2. Day Minutes
3. Customer Service Calls
4. Voice Mail Plans
5. Evening Charge
6. International Charge
7. International Calls
8. Night Minutes
9. International Minutes
10. Voice Mail Messages

Predicted

0 1

Actual

0 2254 28

1 334 50

**Results:**

We can say that the model Fit3 predicted that there are 334 customers who will not Churn and there are 28 Customers who might Churn with an Accuracy of 86.42%.

~End of the report~

Submitted by:

Nibhir Ghosh