Problem Statement:

The project aims at predicting the churn likelihood in the telephone customers based on the past data on churn using a set of different attributes.

Steps involved:

Below is the brief description of the 12 detailed steps followed in R coding which is comprised into summarised high-level view. It depicts the sequential flow of how we build the prediction model and its subsequent testing on the test sample.

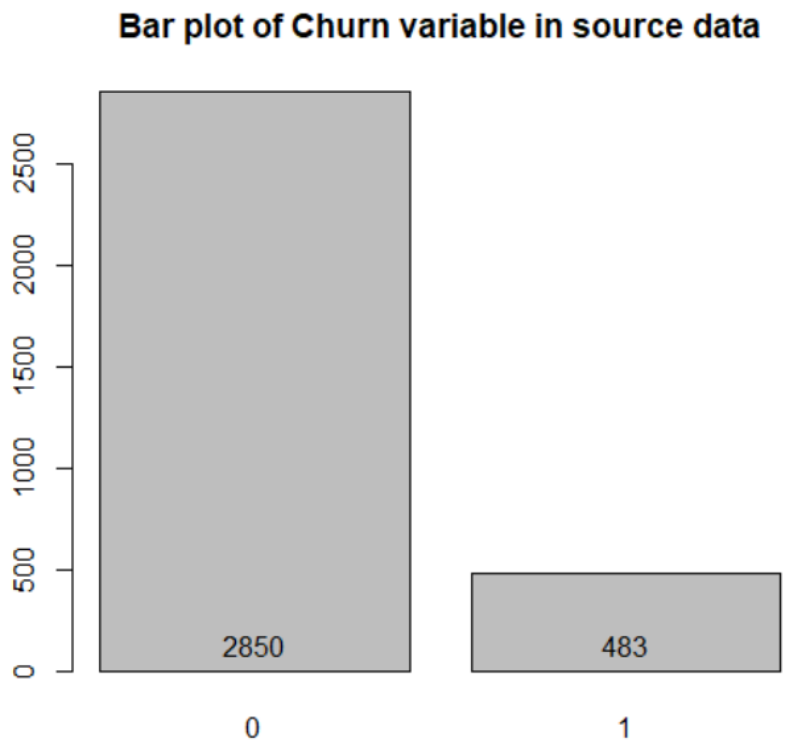
Step 1:

We pulled the given data file in the R environment and studied the structure of data.

Step 2:

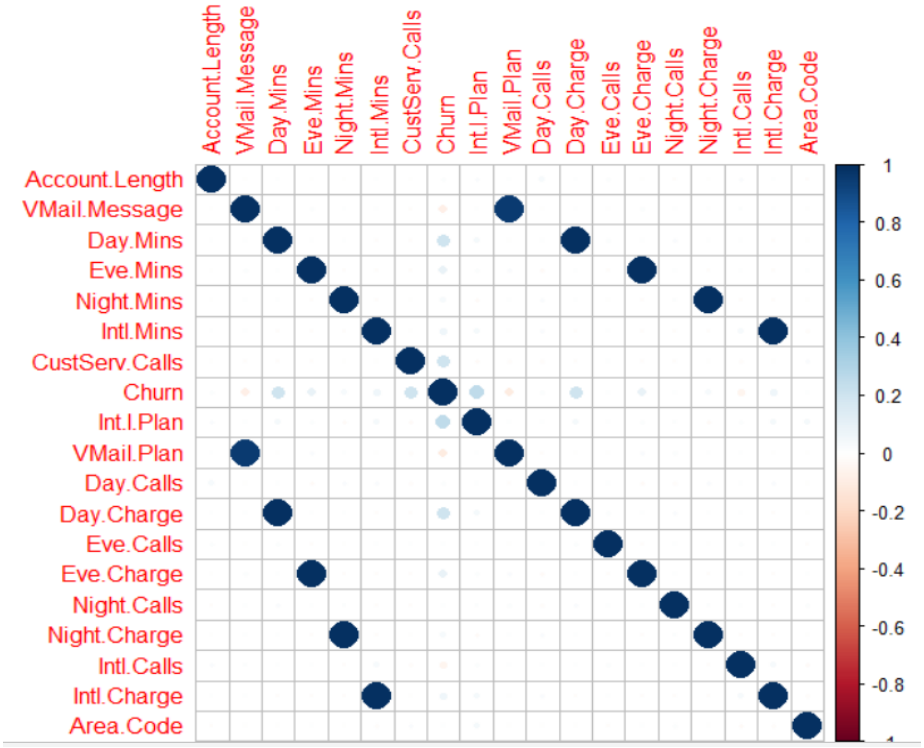
We inspect the distribution in current dataset using visualization and count.

The “0” here indicate the cases with no churn in the past and “1” are with churn.



Step 3:

We plot the correlation between the variables and avoid the ones with strong intra-correlation to avoid undue influence and erroneous models.



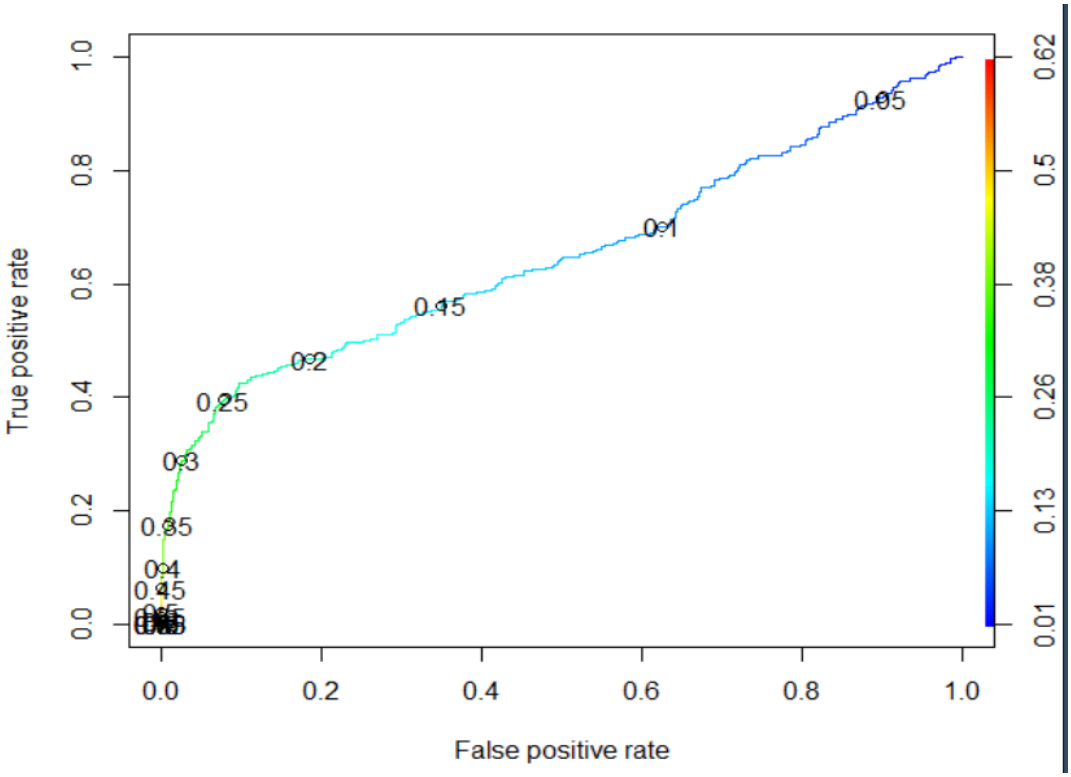
Step 4:

With selected set of influential variables we build the Logistic Regression model

```
> log_model <- glm(source_data$Churn ~.,train,family = "binomial")
> log_model

Call:  glm(formula = source_data$Churn ~ ., family = "binomial", data = train)

Coefficients:
(Intercept)   VMail.Plan   Day.Charge  Intl.Charge
   -4.61122     -0.77768      0.06744      0.29157

Degrees of Freedom: 3332 Total (i.e. Null);  3329 Residual
Null Deviance:      2758
Residual Deviance: 2558         AIC: 2566
> summary(log_model)

Call:
glm(formula = source_data$Churn ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0936  -0.5974  -0.4746  -0.3326   2.8606

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.611222   0.290718 -15.861  < 2e-16 ***
VMail.Plan  -0.777683   0.131750  -5.903 3.58e-09 ***
Day.Charge   0.067440   0.005792  11.644  < 2e-16 ***
Intl.Charge  0.291567   0.068928   4.230 2.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2758.3  on 3332  degrees of freedom
Residual deviance: 2557.5  on 3329  degrees of freedom
AIC: 2565.5
```

Step 5:

We plot the ROC curve to get the cut-offs for validating our confusion matrix result of the actual vs predicted results. We can see the graph hinting **at 0.25** as the threshold for True positive and False Positive classification. This helps us arrive at the Performance of the test.



Step 6:

We create a confusion matrix to gauge the accuracy of our model by deploying it over the test sample of our data and arrive at decent level of about 84%.

```
         Predicted
Observed FALSE TRUE
       0   857   75
       1   106   73
> accuracy <- (857+73)/(857+75+73+106)
> accuracy
[1] 0.8370837
```

Step 7: Odds ratio

We compute the confidence interval here to gauge the odds of how a per unit change in an independent variable can possibly influence the predicted outcome.

So for example, a unit change in Day. Charge is likely to increase the chances of churning by about 6 %.

```
> exp(confint(log_model))
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) 0.0055795 0.01744493
VMail.Plan  0.3528173 0.59170921
Day.Charge  1.0577909 1.08209019
Intl.Charge 1.1701198 1.53323155
>
```

So, the logistic regression model can help us build the model and validate it on the split data. The model strength and performance can be inferred using the ROC curve and the accuracy measures of confusion matrix.

----------**END**----------