Predictive Modeling Assignment – Prediction of Customer Churn

Model Solution

GREAT LAKES
INSTITUTE OF MANAGEMENT

Table of Contents

1 Phase I – Discovery

1.1 The Business Domain

Customer attrition is an important issue for any industry. It is especially important in mature industries where the initial period of exponential growth has been left behind. Not surprisingly, attrition (or, to look on the bright side, retention) is a major application of data mining.

One of the first challenges in modeling attrition is deciding what it is and recognizing when it has occurred. This is harder in some industries than in others. At one extreme are businesses that deal in anonymous cash transactions.

When a once-loyal customer deserts his regular coffee bar for another down the block, the barista who knew the customer's order by heart may notice, but the fact will not be recorded in any corporate database. Even in cases where the customer is identified by name, telling the difference between a customer who has churned and one who just hasn't been around for a while may be hard. If a loyal Ford customer who buys a new F150 pickup every five years hasn't bought one for six years, has the customer defected to another brand?

Attrition is a bit easier to spot when a monthly billing relationship exists, as with credit cards. Even there, attrition might be silent. A customer may stop using the credit card, but not cancel it. Attrition is easiest to define in subscription-based businesses, and partly for that reason, attrition modeling is most popular in these businesses. Long-distance companies, mobile phone service providers, insurance companies, cable companies, financial services companies, Internet service providers, newspapers, magazines, and some retailers all share a subscription model where customers have a formal, contractual relationship that must be explicitly ended.

Lost customers must be replaced by new customers, and new customers are expensive to acquire. Often, new customers generate less revenue in the near term than established customers. This is especially true in mature industries where the market is fairly saturated — anyone likely to want the product or service probably already has it from somewhere, so the main source of new customers is people leaving a competitor.

Hence it is vital for such industries to understand the attrition.

This assignment focuses on one of such problems the Telecom Industry faces – Customer Churn.

1.2 The Problem in hand

Students are given a Cell Phone Data file and are requested to build a Logistic Regression Model which can tell the parameters contributing (and not contributing) for Customer Churn (attrition), along with the intensity of each attribute.

The input file needs to divide into Training Dataset, which should contain 70% of the data and Testing Dataset, which would contain remaining 30% of the data.

The cell phone data file contains following attributes:

1.2.1 Cell Phone Data Attributes

Sr. No.	Variable Name	Description	Model Name	Remark
1	Churn	1 if customer cancelled service, 0 if not	y	Dependent Variable
2	AccountWeeks	number of weeks customer has had active account	x1	Independent Variable
3	ContractRenewal	1 if customer recently renewed contract, 0 if not	x2	Independent Variable
4	DataPlan	1 if customer has data plan, 0 if not	x3	Independent Variable
5	DataUsage	gigabytes of monthly data usage	x4	Independent Variable
6	CustServCalls	number of calls into customer service	x5	Independent Variable
7	DayMins	average daytime minutes per month	x6	Independent Variable
8	DayCalls	average number of daytime calls	x7	Independent Variable
9	MonthlyCharge	average monthly bill	x8	Independent Variable
10	OverageFee	largest overage fee in last 12 months	x9	Independent Variable
11	RoamMins	average number of roaming minutes	x10	Independent Variable

1.3 The Initial Hypothesis

The Cellphone File contains one Dependent and 10 Predictor variables.

The assignment aim is to identify the predictor variables which are significant for Customer Churn.

Null Hypothesis (Ho) –No predictor is able to predict the Churn

Alternate Hypothesis (Ha)–At least one of the predictors is able to predict the churn.

2 Phase II – Data Preparation

2.1 Data Exploration and Visualization

2.1.1 Initial Data Exploration

1. Check for Variables names, Five Point Summary, and whether they contain any null value or not.

   This is the first opportunity to explore our data.

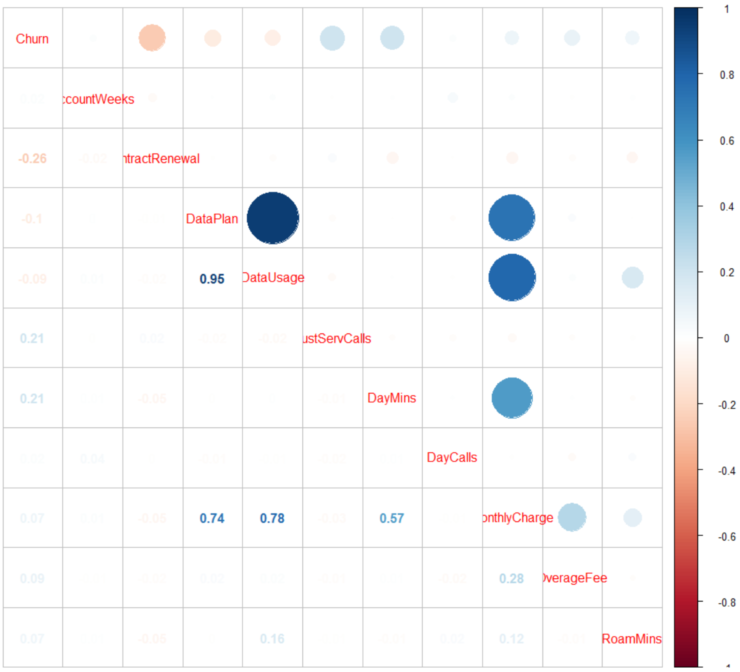
```
str(cellphone.df)
summary(cellphone.df)
head(cellphone.df)
tail(cellphone.df)
sum(is.na(cellphone.df)) # no null data
```

2. Convert the Categorical variables into Factor.
   In R, by default, if the Categorical variables contain numeric values, those are defined as Numeric (See the output of str(cellphone.df)). It’s better to convert those as Factor variables.

```
cellphone.df$Churn<- as.factor(cellphone.df$Churn)
cellphone.df$DataPlan<- as.factor(cellphone.df$DataPlan)
cellphone.df$ContractRenewal<-
as.factor(cellphone.df$ContractRenewal)
```

3. Check for Correlated Variables.

```
bc = cor(cellphone.df) #create an object of the features
bc
corrplot.mixed(bc)
```



As the plot shows, following variables are highly correlated:

1. dataplan and datausage highly correlated
2. dataplan and monthlycharge highly correlated
3. datausage and monthlycharge highly correlated
4. daymins and monthlycharge highly correlated

2.2 Data Preparation – Training and Testing Dataset

1. Split the Input dataset into Training (70%) and Testing(30%).

```
set.seed(100)
ind = sample(2, nrow(cellphone.df), replace = TRUE,
prob=c(0.7,0.3))
train.df = cellphone.df[ind == 1,]
test.df = cellphone.df[ind == 2,]
```

2. Verify the Training and Testing Dataset for:
    - If the records are distributed as expected and
    - If the Dependant variable is distributed in the same proportion.

```
dim(train.df)
dim(test.df)
prop.table(table(train.df$Churn))
prop.table(table(test.df$Churn))
```

We see almost equal representation in both training and testing set for the dependent or response variable.

3 Phase III – Model Planning and Building

3.1 Model Selection and initial build

3.1.1 Model 1 with all variables:

Let's build the initial Logistic Regression Model taking all independent variables into consideration.

```
#Model 1 with all the variables
Model1 <- glm(Churn ~ ., data = train.df,
      family = binomial(link="logit"))

> summary(Model1)

Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"),
data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0345  -0.5094  -0.3381  -0.1948   3.0376

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.267591   0.668315  -9.378  < 2e-16 ***
AccountWeeks      0.001969   0.001677   1.174  0.24029
ContractRenewal1 -2.065359   0.168834 -12.233  < 2e-16 ***
DataPlan1        -1.238548   0.637685  -1.942  0.05211 .
DataUsage         2.037504   2.324631   0.876  0.38077
CustServCalls     0.492662   0.045729  10.773  < 2e-16 ***
DayMins           0.046913   0.039194   1.197  0.23133
DayCalls          0.001900   0.003332   0.570  0.56861
MonthlyCharge    -0.194928   0.230311  -0.846  0.39735
OverageFee        0.496165   0.393093   1.262  0.20687
RoamMins          0.086529   0.026768   3.233  0.00123 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1956.4  on 2325  degrees of freedom
Residual deviance: 1511.7  on 2315  degrees of freedom
AIC: 1533.7

Number of Fisher Scoring iterations: 6
```

<u>Interpretation:</u>

The three significant variables highlighted in yellow are:

1. Contract Renewal: Please note, this has a negative impact on Customer Churn.
2. Customer Service Calls
3. Roaming Minutes

Also, the AIC[1] Score is 1533.7. This will be observed in subsequent stages when we refine the model. The model having **least AIC Score** would be the **most preferred and optimized one.**

3.2 Model 1 Significance Verification

Now since we have built our initial Regression Model considering all the Predictors, let’s check its significance.

3.2.1 Overall Significance of the Model – Log Likelihood Ratio Test[2]

The first check is on overall Significance, by verifying the Log Likelihood Ratio.

```
#Likelihood ratio test
lrtest(Model1)

> lrtest(Model1)
Likelihood ratio test

Model 1: Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage +
    CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee +
    RoamMins
Model 2: Churn ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  11 -755.87
2   1 -978.18 -10 444.61  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

H0: All betas are zero
H1: At least 1 beta is nonzero

From the log likelihood, we can see that, intercept only model -978.18 variance was unknown to us. When we take the full model, -755.87 variance was unknown to us. So we can say that, 1 – (-755.87 /-978.18)= **22.72%** of the uncertainty inherent in the intercept only model is calibrated by the full model.

[1] AIC: The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters.

[2] Log Likelihood Test: In statistics, a likelihood ratio test is a statistical test used for comparing the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model and hence accept the alternative model.

Chisq likelihood ratio is significant. Also the p value suggests that we can accept the Alternate Hypothesis that at least one of the beta is not zero. So Model is significant.

3.2.2 Model Robustness – McFadden's pseudo-R Squared Test

Now since we concluded that the model built is significant, let's find out how robust it is with the help of McFadden pseudo-R Squared Test[3].

```
> pR2(Model1)
          llh       llhNull            G2      McFadden          r2ML        r2CU
-755.8740291 -978.1767837   444.6055090     0.2272623     0.1739880 0.3059099
```

Interpretation:

The McFadden's pseudo-R Squared test suggests that atleast 22.72% variance of the data is captured by our Model, which suggests **it's a robust model.**

3.2.3 Test for Individual Coefficients

When looked closely at the output of the Model in section 3.1.1, we get following information about the Individual coefficients:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.267591   0.668315  -9.378  < 2e-16 ***
AccountWeeks      0.001969   0.001677   1.174  0.24029
ContractRenewal1 -2.065359   0.168834 -12.233  < 2e-16 ***Significant
DataPlan1        -1.238548   0.637685  -1.942  0.05211 .
DataUsage         2.037504   2.324631   0.876  0.38077
CustServCalls     0.492662   0.045729  10.773  < 2e-16 *** Significant
DayMins           0.046913   0.039194   1.197  0.23133
DayCalls          0.001900   0.003332   0.570  0.56861
MonthlyCharge    -0.194928   0.230311  -0.846  0.39735
OverageFee        0.496165   0.393093   1.262  0.20687
RoamMins          0.086529   0.026768   3.233  0.00123 ** Significant
```

[3] McFadden pseudo-R Squared: Logistic regression models are fitted using the method of maximum likelihood - i.e. the parameter estimates are those values which maximize the likelihood of the data which have been observed. McFadden's R squared measure is defined as $\boxed{R^2_{McFadden} = 1 - \frac{LogLc}{LogLnull}}$ where Lc denotes the (maximized) likelihood value from the current fitted model, and Lnull denotes the corresponding value but for the null model - the model with only an intercept and no covariates.

Interpretation:

As can be seen, following **three Variables are statistically significant**

- Contract Renewal,
- Customer Service Calls and
- RoamMins

Also, Contract Renewal, which is a categorical variable has negative impact on the Customer Churn.

However, **are these only significant variables we need to consider?**

Let's find out the power of Odds and Probability of the variables impacting on Customer Churn.

3.2.4 Odds Explanatory Power

```
> exp(coef(Model1)) # Odds Ratio
     (Intercept) AccountWeeks ContractRenewal1    DataPlan1      DataUsage
     0.001896793  1.001970708      0.126772710  0.289804648    7.671441061
   CustServCalls      DayMins     DayCalls  MonthlyCharge   OverageFee
     1.636667072  1.048031090      1.001901319  0.822894107    1.642410330
        RoamMins
     1.090383062

> exp(coef(Model1))/(1+exp(coef(Model1)))  # Probability
     (Intercept) AccountWeeks ContractRenewal1    DataPlan1     DataUsage
     0.001893202  0.500492192      0.112509567  0.224688792  0.884678914
   CustServCalls      DayMins        DayCalls MonthlyCharge   OverageFee
     0.620733307  0.511726162     0.500474878  0.451421782  0.621557640
        RoamMins
     0.521618780
```

Interpretation:

If a particular Variable as shown in following table is increased by 'One Unit', the odds of customer churn (Vs. not churning) and the probability of Customer Churn is shown in the following table.

Variable	The odds of Customer will Churn	Probability of Customer Churn increases by	Remark
AccountWeeks	1.00	50%	*These fields have negative impact on Customer Churn, as indicated in Section 3.2.3. The odds for these fields too need to interpret accordingly.
ContractRenewal1	0.12	-11%*	
DataPlan1	0.29	-22%*	
DataUsage	7.67	88%	
CustSrvCalls	1.64	62%	
DayMins	1.04	51%	
DayCalls	1.00	50%	
MonthlyCharge	0.82	-45%*	
Overage Fee	1.64	62%	
RoamMin	1.09	52%	

For Categorical Variables, e.g. when the Customer renews Contract (1) the odds customer will churn is 0.12 compared to when the customer doesn’t renew. Similarly, when the customer opts for Data Plan (Value 1) the odds the customer will churn is 0.29 compared to when the customer doesn’t opt.

3.2.5 Classification Table

Since we have confirmed the importance of additional significant variables, let’s check performance of our Model using a Classification Table / Confusion Matrix.

Classification Table on Training Dataset:

```
> pred<-predict(Model1,newdata=train.df,type="response")
> y_pred_num <- ifelse(pred>0.5,1,0)
> y_pred <- factor(y_pred_num, levels=c(0,1))
> y_act <- train.df$Churn
> confusionMatrix(y_pred,y_act,positive="1")

Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1922  275
         1   58   71

               Accuracy : 0.8568
                 95% CI : (0.8419, 0.8708)
    No Information Rate : 0.8512
    P-Value [Acc > NIR] : 0.2342

                  Kappa : 0.2373
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.20520
            Specificity : 0.97071
         Pos Pred Value : 0.55039
         Neg Pred Value : 0.87483
             Prevalence : 0.14875
         Detection Rate : 0.03052
   Detection Prevalence : 0.05546
      Balanced Accuracy : 0.58795

       'Positive' Class : 1
```

Interpretation:

1. 71 out of (58+71) Customers identified **correctly** which have been churned out. This translates to **55% of Positive Predictive Value.**
2. 1922 out of (1922+275) Customers identified **correctly** which have not been churned out. This translates to **87.5% of Negative Predictive Values.**
3. At **85.68 %,** the Model provides good accuracy measures.

4. **Sensitivity**[4] **is 0.20** and **Specificity is 0.97**

Classification Table on Testing Dataset:

```
> pred<-predict(Model1,newdata=test.df,type="response")
> y_pred_num <- ifelse(pred>0.5,1,0)
> y_pred <- factor(y_pred_num, levels=c(0,1))
> y_act <- test.df$Churn
> confusionMatrix(y_pred,y_act,positive="1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 852 114
         1  18  23

               Accuracy : 0.8689
                 95% CI : (0.8465, 0.8892)
    No Information Rate : 0.864
    P-Value [Acc > NIR] : 0.343

                  Kappa : 0.2088
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.16788
            Specificity : 0.97931
         Pos Pred Value : 0.56098
         Neg Pred Value : 0.88199
             Prevalence : 0.13605
         Detection Rate : 0.02284
   Detection Prevalence : 0.04071
      Balanced Accuracy : 0.57360

       'Positive' Class : 1
```

Interpretation:

1. 23 out of (18+23) Customers identified **correctly** which have been churned out. This translates to **56% of Positive Predictive Value.**
2. 852 out of (852+114) Customers identified **correctly** which have not been churned out. This translates to **88% of Negative Predictive Values.**
3. At **86.7 %,** the Model provides good accuracy measures.
4. **Sensitivity is 0.17** and **Specificity is 0.98**

Thus, the model shows pretty much similar performance on both Training as well as Testing datasets.

[4] Sensitivity is the probability that a model will indicate 'Churn' among those 'Churned Records'.
Specificity is the fraction of those 'not churned' who actually are not churned.

3.2.6 ROC Plot

Finally, let’s draw the Receiver Operating Characteristic (ROC) plot.

It is a plot of the True Positive Rate against the False Positive Rate for the different possible cut-points of a diagnostic test.

An ROC curve demonstrates several things:

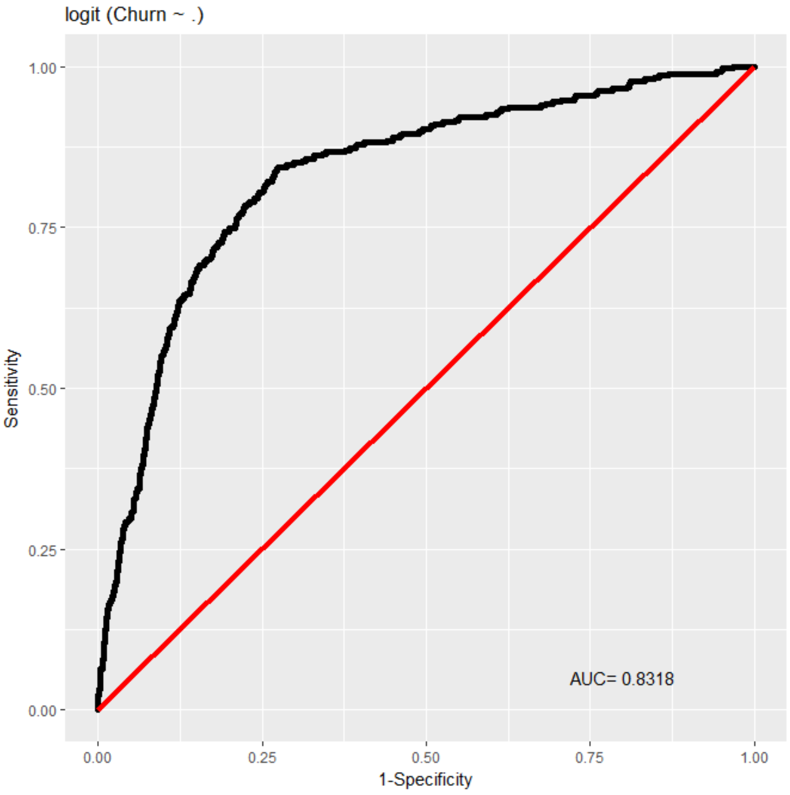
1. It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, **the more accurate the test**.
3. The closer the curve comes to the 45-degree diagonal of the ROC space, **the less accurate the test**.
4. The slope of the tangent line at a cut-point gives the likelihood ratio (LR) for that value of the test.
5. The area under the curve (AUC) is a measure of text accuracy.

Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test; an area of 0.5 represents a worthless test. A rough guide for classifying the accuracy of a model verification test is the traditional academic point system, as follows:

- 0.90-1 = excellent (A)
- 0.80-0.90 = good (B)
- 0.70-0.80 = fair (C)
- 0.60-0.70 = poor (D)
- 0.50-0.60 = fail (F)

```
> rocplot(Model1)
```

AUC At 0.8318, the ROC Curve of our model demonstrates fairly good results.



3.3 Model Refining

So far we have checked the Model's overall significance, and we are pretty happy with its results.

Hence shall we conclude and finalize the Model? Or we can do some more refining?

Let's revisit the predictors which are less significant and also see if there is any interaction between variables.

As seen in section 3.2.4, variable "MonthlyCharge" is insignificant (odds less than 1) and also as per section 2.1.1, it is correlated with the following three variables:

- Data Plan
- Data Usage
- DayMins

So we may decide to exclude it in our refinement step.

3.3.1 Variable selection using Step Function

So far in our Logistic Regression journey, we have included all the explanatory variables in our model. However, selecting the one's which really matters for the model becomes really important.

There are two main approaches towards selecting variables: the all possible regression approach and automatic methods.

The all possible regressions approach considers all possible subsets of the pool of explanatory variables and finds the model that best fits the data according to some criteria (e.g. Adjusted R2, AIC and BIC). These criteria assign scores to each model and allow us to choose the model with the best score.

Automatic methods are useful when the number of explanatory variables is large and it is not feasible to fit all possible models. In this case, it is more efficient to use a search algorithm (e.g., Forward selection, Backward elimination and Stepwise regression) to find the best model.

Let's use the R function step() to perform the Variable Selection.

The Model giving minimum value of the AIC would be the best one to choose from.

```
> Model2 <- step(glm(Churn ~ ., data = train.df, family = binomial(link="l
ogit")))
Start:  AIC=1533.75
Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage + 
    CustServCalls + DayMins + DayCalls + MonthlyCharge + OverageFee + 
    RoamMins
```

```
                  Df Deviance    AIC
- DayCalls         1   1512.1 1532.1
- MonthlyCharge    1   1512.5 1532.5
- DataUsage        1   1512.5 1532.5
- AccountWeeks     1   1513.1 1533.1
- DayMins          1   1513.2 1533.2
- OverageFee       1   1513.3 1533.3
<none>                 1511.8 1533.8
- DataPlan         1   1515.7 1535.7
- RoamMins         1   1522.5 1542.5
- CustServCalls    1   1630.3 1650.3
- ContractRenewal  1   1657.2 1677.2

Step:  AIC=1532.07
Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage + 
    CustServCalls + DayMins + MonthlyCharge + OverageFee + RoamMins

                  Df Deviance    AIC
- MonthlyCharge    1   1512.8 1530.8
- DataUsage        1   1512.8 1530.8
- AccountWeeks     1   1513.5 1531.5
- DayMins          1   1513.5 1531.5
- OverageFee       1   1513.7 1531.7
<none>                 1512.1 1532.1
- DataPlan         1   1515.9 1533.9
- RoamMins         1   1522.9 1540.9
- CustServCalls    1   1630.4 1648.4
- ContractRenewal  1   1657.5 1675.5

Step:  AIC=1530.78
Churn ~ AccountWeeks + ContractRenewal + DataPlan + DataUsage + 
    CustServCalls + DayMins + OverageFee + RoamMins

                  Df Deviance    AIC
- DataUsage        1   1512.9 1528.9
- AccountWeeks     1   1514.3 1530.3
<none>                 1512.8 1530.8
- DataPlan         1   1516.5 1532.5
- RoamMins         1   1523.9 1539.9
- OverageFee       1   1549.8 1565.8
- CustServCalls    1   1631.3 1647.3
- DayMins          1   1637.0 1653.0
- ContractRenewal  1   1658.7 1674.7

Step:  AIC=1528.9
Churn ~ AccountWeeks + ContractRenewal + DataPlan + CustServCalls + 
    DayMins + OverageFee + RoamMins

                  Df Deviance    AIC
- AccountWeeks     1   1514.4 1528.4
<none>                 1512.9 1528.9
- RoamMins         1   1527.3 1541.3
- OverageFee       1   1549.9 1563.9
- DataPlan         1   1549.9 1563.9
- CustServCalls    1   1631.3 1645.3
- DayMins          1   1637.2 1651.2
- ContractRenewal  1   1659.2 1673.2

Step:  AIC=1528.4
```

```
Churn ~ ContractRenewal + DataPlan + CustServCalls + DayMins +
    OverageFee + RoamMins
                  Df Deviance    AIC
<none>                 1514.4 1528.4
- RoamMins         1   1528.8 1540.8
- OverageFee       1   1551.1 1563.1
- DataPlan         1   1551.2 1563.2
- CustServCalls    1   1633.4 1645.4
- DayMins          1   1638.3 1650.3
- ContractRenewal  1   1661.2 1673.2
```

Interpretation:

Recall in Section 3.1.1 where we built our all-inclusive model wherein we had got the AIC value as 1533.7.

Using the step() function, now we got the best AIC of 1528.4, with only 6 significant predictor variables.

3.3.2 Model 2: Using step() function

From above step, we select our second model as:

```
Churn ~ ContractRenewal +
        DataPlan +
        CustServCalls +
        DayMins +
        OverageFee +
        RoamMins
```

We may proceed with this model, to check its overall significance now, by performing the following tests:

1. Log Likelihood Test
2. McFadden’s pseudo-R Squared Test
3. Test for Individual coefficient
4. Odds Explanatory Power
5. Classification Table and
6. ROC Plot.

3.4 Model 2: Significance Verification

3.4.1 Overall Significance of the Model – The Log Likelihood Test

```
> lrtest(Model2)
Likelihood ratio test

Model 1: Churn ~ ContractRenewal + DataPlan + CustServCalls + DayMins + 
    OverageFee + RoamMins
Model 2: Churn ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)    
1   7 -757.20                         
2   1 -978.18 -6 441.96  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

From the log likelihood, we can see that, intercept only model -978.18 variance was unknown to us. When we take the full model, -757.20 variance was unknown to us. So we can say that, 1 – (-757.20 /-978.18)= **22.59%** of the uncertainty inherent in the intercept only model is calibrated by the full model.

Chisq likelihood ratio is significant. Also the p value suggests that we can accept the Alternate Hypothesis that at least one of the beta is not zero. So Model is significant.

3.4.2 Model Robustness – McFadden’s pseudo-R Squared Test

```
> pR2(Model2)
         llh      llhNull           G2     McFadden         r2ML         r2CU 
-757.1984213 -978.1767837  441.9567247    0.2259084    0.1730468    0.3042551 
```

Interpretation:

The McFadden’s pseudo-R Squared test suggests that atleast **22.59**% variance of the data is captured by our Model, which suggests **it’s a robust model.**

3.4.3 Test for Individual Coefficients

```
> summary(Model2)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -5.983593   0.525910 -11.378  < 2e-16 ***
ContractRenewal1 -2.072655   0.168627 -12.291  < 2e-16 ***
DataPlan1        -0.984993   0.173519  -5.677 1.37e-08 ***
CustServCalls     0.491729   0.045538  10.798  < 2e-16 ***
DayMins           0.013722   0.001302  10.539  < 2e-16 ***
OverageFee        0.162964   0.027446   5.938 2.89e-09 ***
RoamMins          0.091768   0.024459   3.752 0.000176 ***
```

Interpretation:

For Model 1, we had seen that statistically, three variables were significant. As can be seen above, for Model2, all the six variables are significant, with Contract Renewal and Data Plan having negative impact on Customer churn, whereas the rest of the variables contribute to Customer Churn.

3.4.4 Odds Explanatory Power

```
> exp(coef(Model2)) # Odds Ratio
   (Intercept) ContractRenewal1     DataPlan1  CustServCalls        DayMins
   0.002519756      0.125851166  0.373441725    1.635140434    1.013816928
      OverageFee         RoamMins
     1.176993755      1.096110106

> exp(coef(Model2))/(1+exp(coef(Model2))) # Probability
   (Intercept) ContractRenewal1    DataPlan1  CustServCalls        DayMins
   0.002513423      0.111783129 0.271902126    0.620513584    0.503430532
      OverageFee         RoamMins
     0.540650956      0.522925825
```

Interpretation:

If a particular Variable as shown in following table is increased by 'One Unit', the odds of customer churn (Vs. not churning) and the probability of Customer Churn is shown in the following table.

Variable	The odds of Customer will Churn	Probability of Customer Churn increases by
ContractRenewal1	0.13	-11%*
DataPlan1	0.37	-27%*
CustSrvCalls	1.63	62%
DayMins	1.01	50%
Overage Fee	1.17	54%
RoamMin	1.09	52%

* These fields have negative impact on Customer Churn, as indicated in Section 3.3.2.3.
The odds for these fields too need to interpret accordingly.

For Categorical Variables, e.g. when the Customer renews Contract (Value 1) the odds customer will churn is 0.13 compared to when the customer doesn't renew the contract. Similarly, when the customer opts for Data Plan (Value 1) the odds the customer will churn is 0.39 compared to when the customer doesn't opt for a data plan.

3.4.5 Classification Table

Classification Table on Training Dataset:

```
> pred<-predict(Model2,newdata=train.df,type="response")
> y_pred_num <- ifelse(pred>0.5,1,0)
> y_pred <- factor(y_pred_num, levels=c(0,1))
> y_act <- train.df$Churn
> confusionMatrix(y_pred,y_act,positive="1")
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1926  273
         1   54   73

               Accuracy : 0.8594
                 95% CI : (0.8446, 0.8733)
    No Information Rate : 0.8512
    P-Value [Acc > NIR] : 0.1403

                  Kappa : 0.2487
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.21098
            Specificity : 0.97273
         Pos Pred Value : 0.57480
         Neg Pred Value : 0.87585
             Prevalence : 0.14875
         Detection Rate : 0.03138
   Detection Prevalence : 0.05460
      Balanced Accuracy : 0.59185

       'Positive' Class : 1
```

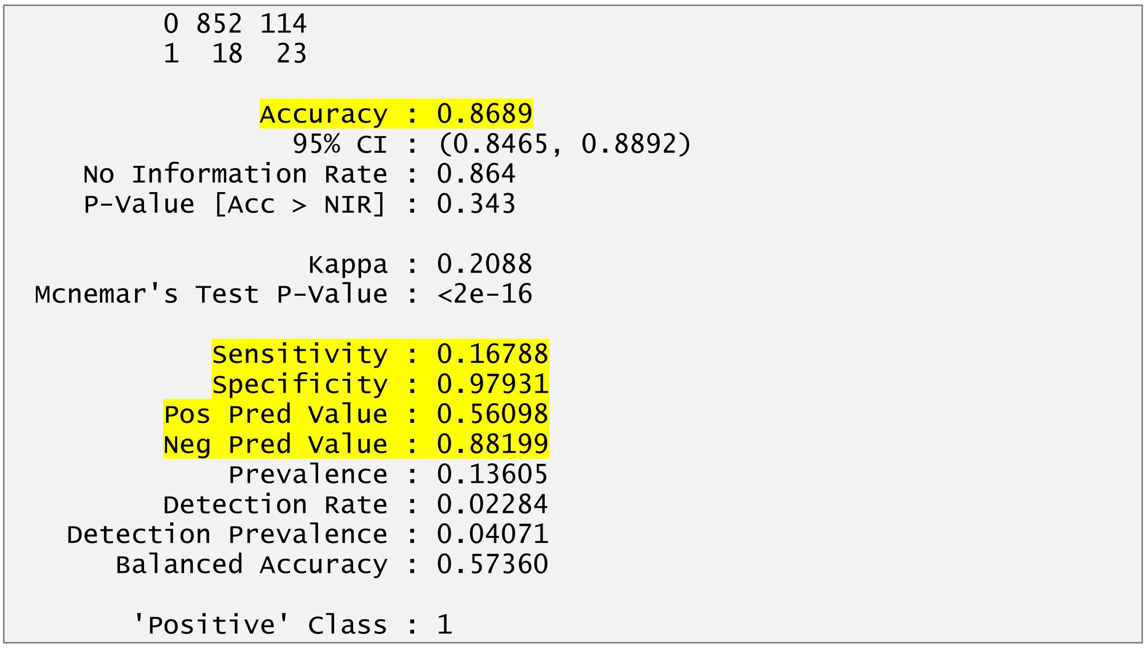
Interpretation:

1. 73 out of (54+73) Customers identified **correctly** which have been churned out. This translates to **57% of Positive Predictive Value.**
2. 1926 out of (1926+273) Customers identified **correctly** which have not been churned out. This translates to **87.5% of Negative Predictive Values.**
3. At **85.94 %,** the Model provides good accuracy measures.
4. **Sensitivity is 0.21** and **Specificity is 0.97**

Classification Table on Testing Dataset:

```
> pred<-predict(Model2,newdata=test.df,type="response")
> y_pred_num <- ifelse(pred>0.5,1,0)
> y_pred <- factor(y_pred_num, levels=c(0,1))
> y_act <- test.df$Churn
> confusionMatrix(y_pred,y_act,positive="1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
```



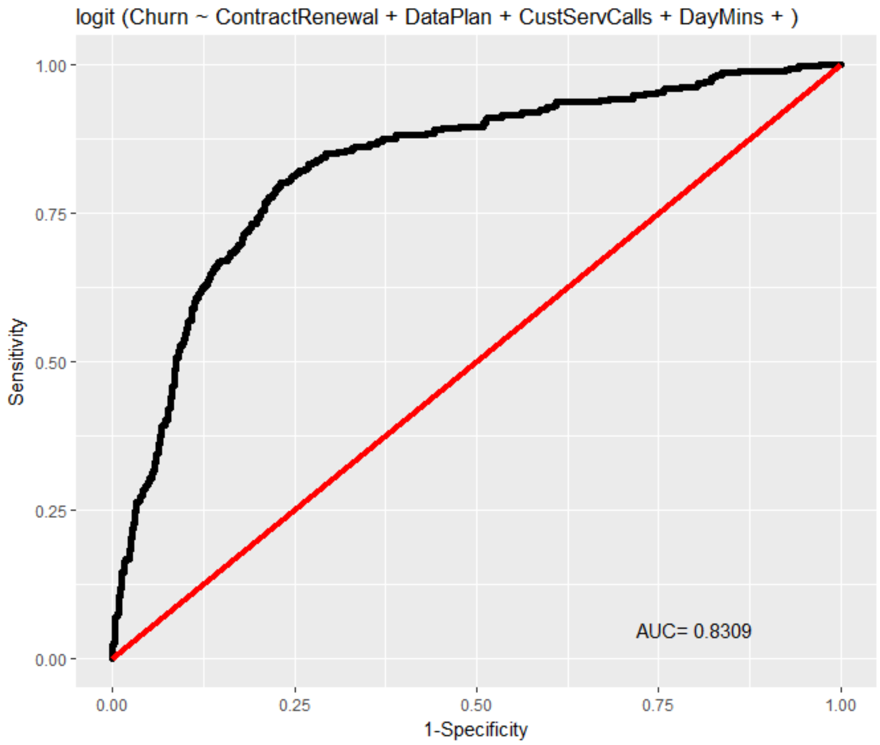
```
        0 852 114
        1  18  23
                                          
              Accuracy : 0.8689          
                95% CI : (0.8465, 0.8892)
   No Information Rate : 0.864           
   P-Value [Acc > NIR] : 0.343           
                                          
                 Kappa : 0.2088          
Mcnemar's Test P-Value : <2e-16          
                                          
           Sensitivity : 0.16788         
           Specificity : 0.97931         
        Pos Pred Value : 0.56098         
        Neg Pred Value : 0.88199         
            Prevalence : 0.13605         
        Detection Rate : 0.02284         
  Detection Prevalence : 0.04071         
     Balanced Accuracy : 0.57360         
                                          
      'Positive' Class : 1               
```

Interpretation:

1. 23 out of (18+23) Customers identified **correctly** which have been churned out. This translates to **56% of Positive Predictive Value.**
2. 852 out of (852+114) Customers identified **correctly** which have not been churned out. This translates to **88% of Negative Predictive Values.**
3. At **86.7 %,** the Model provides good accuracy measures.
4. **Sensitivity is 0.17** and **Specificity is 0.98**

Thus, the model shows pretty much similar performance on both Training as well as Testing datasets.

3.4.6 ROC Plot



```
> rocplot(Model2)
```

Interpretation:

AUC At 0.8309, the ROC Curve of our model demonstrates fairly good results.

3.5 Model Refining – Uncovering interactions

Having more than one input variable in a regression model brings up several issues that do not come up when there is only a single input.

- Ideally, all inputs should be linearly independent with respect to each other.
- There may be interactions between inputs that should be explicitly included in the model.
- Adding a new input changes the coefficient values for any inputs added previously.

So, what exactly interaction is?

Even when two variables are completely independent, their effect on the target may not be. The attractiveness of an ice-cream cone may depend on both its price and the weather — especially how hot the day is. These variables may safely be assumed to be independent. (Certainly, the price of ice cream does not determine the temperature; temperature could conceivably affect the price of ice cream, but let us assume it does not.) Despite the independence of these variables, the effect of price may still be affected by temperature. On a very hot day, people may buy ice cream at almost any price, whereas only a really good deal might tempt them when the weather is cold, damp, and drizzly.

In Model2, we considered 6 input variables, and there are chances of interactions between any of those.

When interactions are considered important, they are often included in the model by adding a new variable that is the product of the standardized values of the variables involved in the interaction.

A step() function and product of the input variables can be used to add interactions and build a further refined Model. Step() function will help us in coming up with the appropriate combination of input variables and the interaction terms.

Refining the Model by understanding interactions between variables is not in scope for this exercise; hence we are not going ahead with it. However, using interactions, we can further optimize the model and achieve accuracy beyond 90%, as well as Sensitivity beyond 50%.

4 Phase IV – Communicating Results

4.1 Conclusion on final Model

Model 1 and Model 2 demonstrates fairly good and matching results. However, following the parsimonious principle, we choose **Model 2 as our final mode.**

The Model 2 equation is as follows:

$$\text{Log(Odds)} = -5.98 - 2.07(\text{ContractRenewal}) - 0.98(\text{DataPlan}) + 0.49(\text{CustServCalls}) + 0.013(\text{DayMins}) + 0.16(\text{OverAgeFee}) + 0.09(\text{RoamMins})$$

4.2 Final interpretation

An organization loses its customers to its competition for various reasons. Churn can affect the company's overall growth. The reputation of the company also goes down in the market if the percentage churn increases year on year. For a company to expand its clientele, its growth rate, as measured by the number of new customers, must exceed its churn rate.

For the data provided to our assignment, the Customer Churn is significantly affected by following variables:

1. **Contract Renewal:** Our Model suggests that if a Customer renews contract, then there is 11% probability that the customer will not churn compared to the one who has not renewed his contract.
2. **Data Plan:** Data plan also has negative impact on Customer Churn. The customer opting for Data Plan has 27% probability that the customer will not churn compared to the one who has not opted for Data Plan.
3. **Customer Service Calls:** The odds of Customer churning out are 1.63 when he makes one unit of Service Calls, who is not making Service Calls. This translates to 62% probability of the customer churning out. The Telecom companies need to ensure that their customers are happy with their services, so that the Customer Churn is reduced.
4. **Day Mins:** The odds of Customer churning out are 1.01 when there is an increase of one unit of average daytime minutes per month. This translates to 50% probability of that customer churning out. The Telecom companies need to keep an eye on this parameter.
5. **Overage Fee:** The odds of Customer churning out are 1.01 when there is an increase of one unit in overage fee. This translates to 54% probability in the customer churn.
6. **Roaming Minutes:** The odds of Customer churning out are 1.09 when there is an increase of one unit in Roaming Minutes of a Customer usage, compared to the one who is not availing roaming minutes. This translates to 52% probability of that customer churning out. The Telecom companies need to keep an eye on this parameter.