KENT STATE UNIVERSITY, KENT, OHIO

Project Report On

CUSTOMER CHURN PREDICTION - ABC Wireless Inc

BUSINESS ANALYTICS MIS 64036, FINAL PROJECT REPORT

Submitted by:

AMRUTA DESHPANDE

MANOJ KUMAR CHALAMALA

SRUSHTI PADADE

ADITYA PUTTA

FALL 2019

COLLEGE OF BUSINESS ADMINISTRATION - MSBA

ABSTRACT

Churn is an issue for telecom companies since it is expensive to obtain new customer than to prevent existing current customer from leaving.

Customer Churn Modelling has lot of consideration of late, as there are signs that a huge portion of corporate benefit originates from the current customers.

Organizations are exceptionally keen on recognizing customers who are probably going to churn and regularly embrace Data Mining techniques to assist them.

In this project, we have identified the customers who are likely to churn and give an appropriate intercession to motivate them to stay based on available information.

INTRODUCTION

Customer retention and acquisition are significant elements that directly affect the benefit for any industry and to achieve a balance between the two isn't a simple task.

Retention of customer is likely to be an important factor since churn rate has effects on revenue of the organization.

Few of the reasons for customers switching service providers can be due to network issues, bad customer service and high monthly plan.

These issues can be tackled by providing discounts or better service etc. and prevents them from switching service providers.

In modern times where we have computational capability to read and analyze complex data and extract valuable information, we can utilize this information to extract patterns and predict the future outcome.

Based on this we are considering customer data provided by ABC wireless for modeling (C50 package).

The objective of this project is to analyze the data and find patterns using predictive model to determine if an existing customer will switch service provider. There are many predictive models which can be utilized to perform our analysis like Naïve bias, K nearest neighbor, regression. Here we will be building our model based on logistic regression.

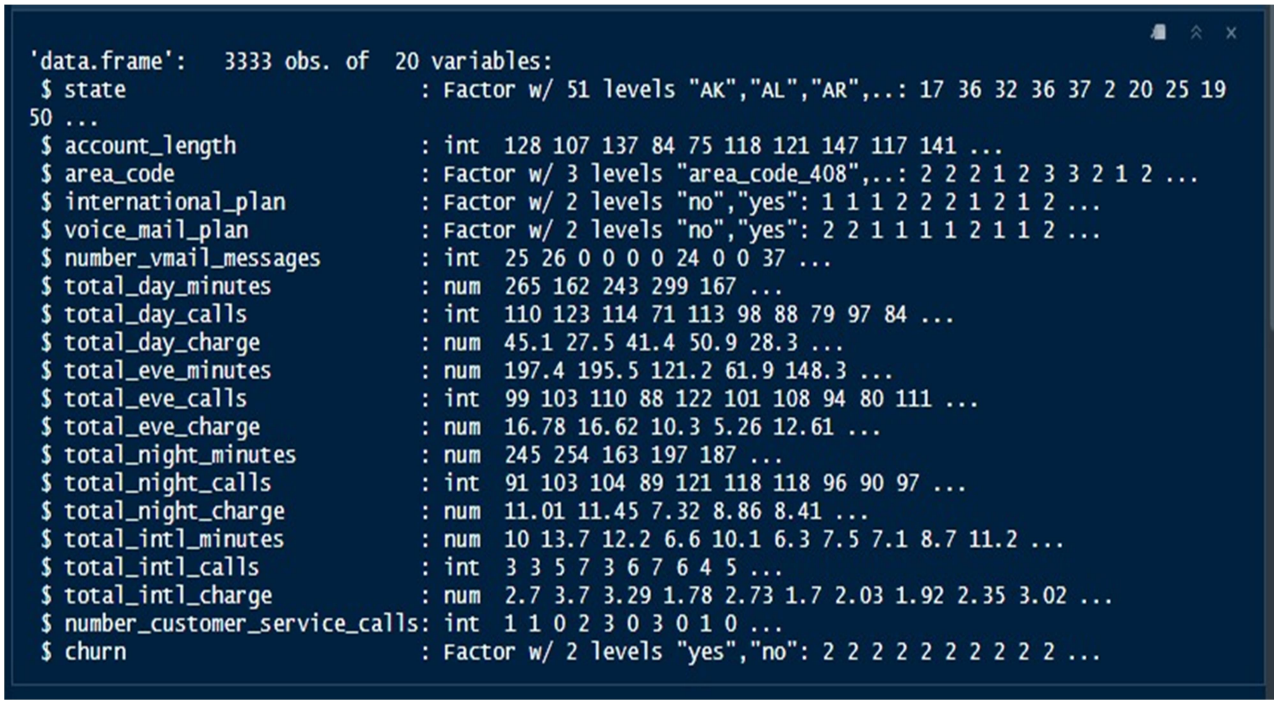
Overview of Data

Our data consist of 2 sets **churnTrain** and **churnTest.**

For building our model we have used Churntrain as training dataset which has **3333 observation with 20 variables** and **churnTest which has 1667 observation with 20 variable**s as a dataset for testing our model.

Data Preprocessing

Data have different variables which are numerical, categorical and binary (used for prediction). The categorical variables are state, area code, international plan and voice mail plan. Numerical variables include total minutes, total call, and total charges. This can be seen in the Figure 1.



```
'data.frame':   3333 obs. of  20 variables:
 $ state                         : Factor w/ 51 levels "AK","AL","AR",..: 17 36 32 36 37 2 20 25 19
50 ...
 $ account_length                : int  128 107 137 84 75 118 121 147 117 141 ...
 $ area_code                     : Factor w/ 3 levels "area_code_408",..: 2 2 2 1 2 3 3 2 1 2 ...
 $ international_plan            : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 1 2 ...
 $ voice_mail_plan               : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 1 2 ...
 $ number_vmail_messages         : int  25 26 0 0 0 0 24 0 0 37 ...
 $ total_day_minutes             : num  265 162 243 299 167 ...
 $ total_day_calls               : int  110 123 114 71 113 98 88 79 97 84 ...
 $ total_day_charge              : num  45.1 27.5 41.4 50.9 28.3 ...
 $ total_eve_minutes             : num  197.4 195.5 121.2 61.9 148.3 ...
 $ total_eve_calls               : int  99 103 110 88 122 101 108 94 80 111 ...
 $ total_eve_charge              : num  16.78 16.62 10.3 5.26 12.61 ...
 $ total_night_minutes           : num  245 254 163 197 187 ...
 $ total_night_calls             : int  91 103 104 89 121 118 118 96 90 97 ...
 $ total_night_charge            : num  11.01 11.45 7.32 8.86 8.41 ...
 $ total_intl_minutes            : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
 $ total_intl_calls              : int  3 3 5 7 3 6 7 6 4 5 ...
 $ total_intl_charge             : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
 $ number_customer_service_calls: int  1 1 0 2 3 0 3 0 1 0 ...
 $ churn                         : Factor w/ 2 levels "yes","no": 2 2 2 2 2 2 2 2 2 2 ...
```

Figure 1

Data Summary

Figure 2 show how our training data Churntrain has been distributed across all variable. It shows minimum, maximum, median, 1st quartile and 3rd quartile for numerical data and either count for categorical data

```
     state          account_length          area_code     international_plan voice_mail_plan
WV      : 106   Min.    :  1.0   area_code_408: 838   no :3010           no :2411
MN      :  84   1st Qu.: 74.0   area_code_415:1655   yes: 323           yes: 922
NY      :  83   Median :101.0   area_code_510: 840
AL      :  80   Mean   :101.1
OH      :  78   3rd Qu.:127.0
OR      :  78   Max.   :243.0
(Other):2824
number_vmail_messages total_day_minutes total_day_calls total_day_charge total_eve_minutes
Min.   : 0.000        Min.   :  0.0     Min.   :  0.0   Min.   : 0.00    Min.   :  0.0
1st Qu.: 0.000        1st Qu.:143.7     1st Qu.: 87.0   1st Qu.:24.43    1st Qu.:166.6
Median : 0.000        Median :179.4     Median :101.0   Median :30.50    Median :201.4
Mean   : 8.099        Mean   :179.8     Mean   :100.4   Mean   :30.56    Mean   :201.0
3rd Qu.:20.000        3rd Qu.:216.4     3rd Qu.:114.0   3rd Qu.:36.79    3rd Qu.:235.3
Max.   :51.000        Max.   :350.8     Max.   :165.0   Max.   :59.64    Max.   :363.7

total_eve_calls total_eve_charge total_night_minutes total_night_calls total_night_charge
Min.   :  0.0   Min.   : 0.00    Min.   : 23.2       Min.   : 33.0     Min.   : 1.040
1st Qu.: 87.0   1st Qu.:14.16    1st Qu.:167.0       1st Qu.: 87.0     1st Qu.: 7.520
Median :100.0   Median :17.12    Median :201.2       Median :100.0     Median : 9.050
Mean   :100.1   Mean   :17.08    Mean   :200.9       Mean   :100.1     Mean   : 9.039
3rd Qu.:114.0   3rd Qu.:20.00    3rd Qu.:235.3       3rd Qu.:113.0     3rd Qu.:10.590
Max.   :170.0   Max.   :30.91    Max.   :395.0       Max.   :175.0     Max.   :17.770

total_intl_minutes total_intl_calls total_intl_charge number_customer_service_calls churn
Min.   : 0.00      Min.   : 0.000   Min.   :0.000     Min.   :0.000                  yes: 483
1st Qu.: 8.50      1st Qu.: 3.000   1st Qu.:2.300     1st Qu.:1.000                  no :2850
Median :10.30      Median : 4.000   Median :2.780     Median :1.000
Mean   :10.24      Mean   : 4.479   Mean   :2.765     Mean   :1.563
3rd Qu.:12.10      3rd Qu.: 6.000   3rd Qu.:3.270     3rd Qu.:2.000
Max.   :20.00      Max.   :20.000   Max.   :5.400     Max.   :9.000
```

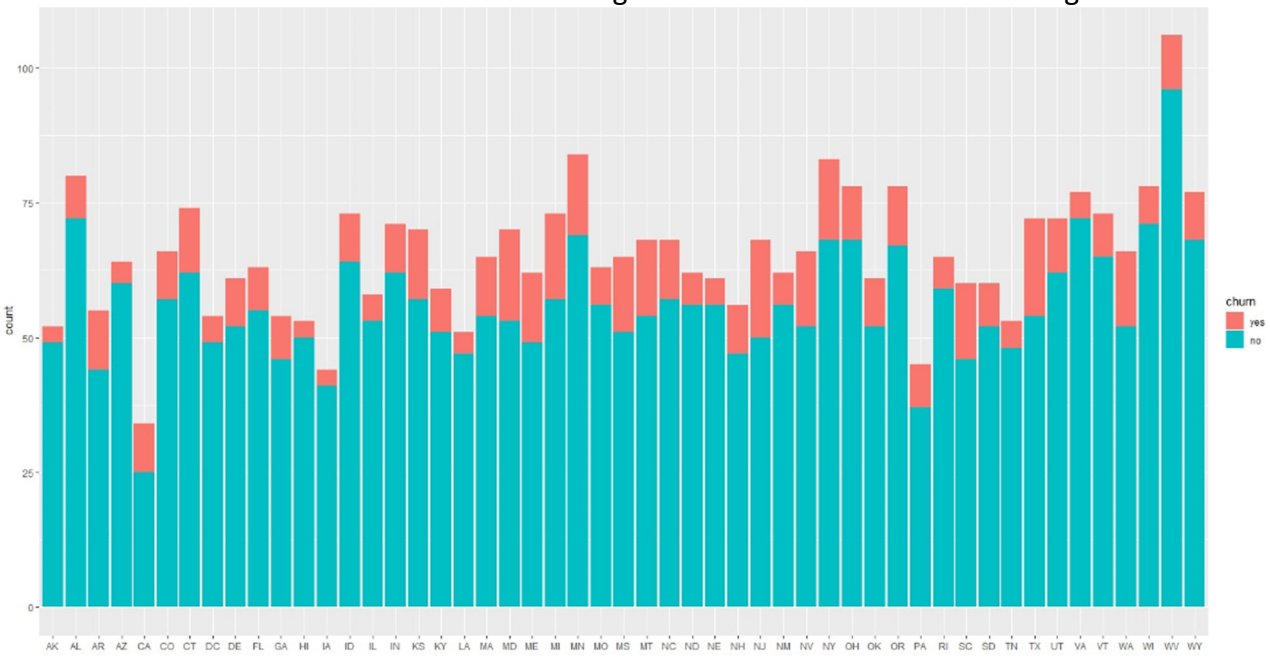
Figure 2

The 'churn' variable is 'yes' and 'no'. This implies expected probabilities would compare the likelihood of 'yes'.

Data Visualization

To get more sense of data, we break down the relation between Churn Cases and different Variables.

Churn Rate based on states can be viewed in Figure 3. To see which states have high churn rate



Train$state	Train$churn	count	prop
CA	yes	9	26.470588
NJ	yes	18	26.470588
TX	yes	18	25.000000
MD	yes	17	24.285714
SC	yes	14	23.333333
MI	yes	16	21.917808
MS	yes	14	21.538462
NV	yes	14	21.212121
WA	yes	14	21.212121
ME	yes	13	20.967742
MT	yes	14	20.588235
AR	yes	11	20.000000

Model Strategy

Predictive model has been build based on regression where it can be used to showcase the influence of different variable and their significance to predict outcome of target variable. Regression can be performed either linearly or logistically. In this data, target variable is categorical hence choosing logistic regression model is an appropriate choice. It's tempting to use the linear regression as a model, although probability of the output can be negative or greater than 1 which is not informative while predicting a binomial attribute. Logistic regression provides likelihood or probability of odds between 0 and 1 which is desired outcome for this model.

Logistic Regression

Model 1: Assuming all the available attributes are significant for the Churn prediction we have built a logistic regression model.

```
##
## Call:
## glm(formula = churn ~ ., family = binomial(link = "logit"), data = Train_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9487  -0.4995  -0.3123  -0.1661   3.0431
...
## number_customer_service_calls  5.366e-01  4.100e-02  13.089  < 2e-16 ***
## ---
## 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: 2070.8  on 3263  degrees of freedom
## AIC: 2210.8
##
```

There are certain variables having **very less z value** and thus can be excluded while model building.

The **States and Area Cod**e do not provide any Statistical Significance as per the above model.

Based on the information gathered from above model we see accuracy as **87.34%**, with **high sensitivity and low level of specificity. ROC is 83.31%.**

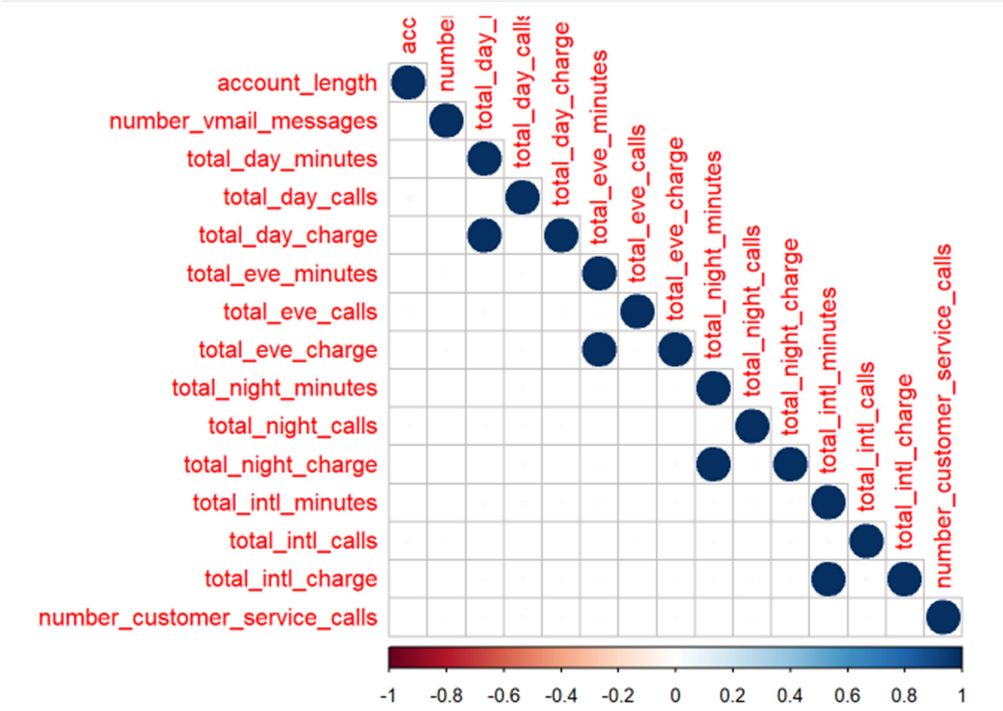
```
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  1418  186
       yes   25   38

               Accuracy : 0.8734
                 95% CI : (0.8565, 0.889)
    No Information Rate : 0.8656
    P-Value [Acc > NIR] : 0.1851

                  Kappa : 0.2187

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

            Sensitivity : 0.9827
            Specificity : 0.1696
         Pos Pred Value : 0.8840
         Neg Pred Value : 0.6032
             Prevalence : 0.8656
         Detection Rate : 0.8506
   Detection Prevalence : 0.9622
      Balanced Accuracy : 0.5762

       'Positive' Class : no
```

Since we have built model based on all attributes of data, we further analyse to see if there are possibilities to reduce certain variable and check to interpret the predictive information from the model.

Hence, we are reassembling the model with extracting informative variables.

CORRELATION

The below image can help us determine correlation among the variables.



We can infer that total minutes and total charge for day, evening, night and international are highly correlated. This implies that higher the minute, higher is the charges- thus we can exclude either one of them.

Model 2: Building a simplified model with reduced number of variables.

While building this model we have excluded the categorical variables like States and Area Code since they do not hold much predictive power.

Also, a set of highly correlated variables are ignored to gain more productivity. Here we removed Variables which had the minutes in it and stored the ones with Charges.

```
Call:
glm(formula = churn ~ ., family = binomial(link = "logit"), data = Train_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1341  -0.5120  -0.3398  -0.1954   3.2626

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -8.6558865  0.7242648 -11.951  < 2e-16 ***
account_length               0.0008334  0.0013913   0.599 0.549142
international_planyes        2.0373143  0.1453553  14.016  < 2e-16 ***
voice_mail_planyes          -2.0075786  0.5732018  -3.502 0.000461 ***
number_vmail_messages        0.0353455  0.0179863   1.965 0.049399 *
total_day_calls              0.0032139  0.0027575   1.166 0.243805
total_day_charge             0.0763955  0.0063738  11.986  < 2e-16 ***
total_eve_calls              0.0010730  0.0027805   0.386 0.699580
total_eve_charge             0.0852189  0.0134450   6.338 2.32e-10 ***
total_night_calls            0.0006893  0.0028398   0.243 0.808206
total_night_charge           0.0822072  0.0246791   3.331 0.000865 ***
total_intl_calls            -0.0920574  0.0250065  -3.681 0.000232 ***
total_intl_charge            0.3253095  0.0755019   4.309 1.64e-05 ***
number_customer_service_calls 0.5137234  0.0392394  13.092  < 2e-16 ***
---
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: 2159.7  on 3319  degrees of freedom
AIC: 2187.7
```

By building a model with lesser number of variables we can see that the predictive capabilities of the model have increased.

This can be proven by the AIC values of the model.

AIC:

Model 1 = 2210

Model 2 = 2187

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  1426  189
##        yes   17   35
##
##                Accuracy : 0.8764
##                  95% CI : (0.8597, 0.8918)
##     No Information Rate : 0.8656
##     P-Value [Acc > NIR] : 0.1034
##
##                   Kappa : 0.2138
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9882
##             Specificity : 0.1562
##          Pos Pred Value : 0.8830
##          Neg Pred Value : 0.6731
##              Prevalence : 0.8656
##          Detection Rate : 0.8554
##    Detection Prevalence : 0.9688
##       Balanced Accuracy : 0.5722
##
##        'Positive' Class : no
##
```

Also the confusion matrix values have improved.

Area under the curve: 0.8408

We can infer, by reducing variables we not only increase the AUC of the model we are reducing complexity of the model

Model 3: Excluding the attributes with negligible significance.

Building model to reduce variable which are least significant and see we have any change in AUC of the model

```
## Call:
## glm(formula = churn ~ ., family = binomial(link = "logit"), data = Train_3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1203  -0.5133  -0.3375  -0.1969   3.2420
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -8.067137   0.515875 -15.638  < 2e-16 ***
## international_planyes         2.040325   0.145243  14.048  < 2e-16 ***
## voice_mail_planyes           -2.003143   0.572347  -3.500 0.000465 ***
## number_vmail_messages         0.035258   0.017964   1.963 0.049674 *
## total_day_charge              0.076590   0.006371  12.022  < 2e-16 ***
## total_eve_charge              0.084495   0.013433   6.290 3.17e-10 ***
## total_night_charge            0.082549   0.024653   3.348 0.000813 ***
## total_intl_calls             -0.092175   0.024988  -3.689 0.000225 ***
## total_intl_charge             0.326138   0.075453   4.322 1.54e-05 ***
## number_customer_service_calls 0.512257   0.039141  13.087  < 2e-16 ***
## ---
## 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: 2161.6  on 3323  degrees of freedom
## AIC: 2181.6
##
## Number of Fisher Scoring iterations: 6
```

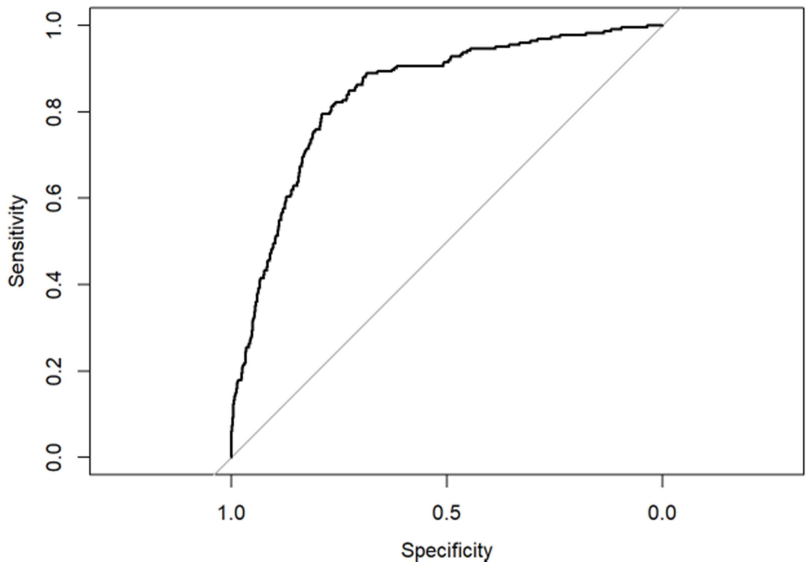
AUC of the model is same with less no. of variable with **threshold as: 0.6**

By looking at the significance of the variable after the model is built all the considered attributes are significant hence we cannot reduce any other variable.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  1428  191
##        yes   15   33
##
##               Accuracy : 0.8764
##                 95% CI : (0.8597, 0.8918)
##    No Information Rate : 0.8656
##    P-Value [Acc > NIR] : 0.1034
##
##                  Kappa : 0.2049
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9896
##             Specificity : 0.1473
##          Pos Pred Value : 0.8820
##          Neg Pred Value : 0.6875
##              Prevalence : 0.8656
##          Detection Rate : 0.8566
##    Detection Prevalence : 0.9712
##       Balanced Accuracy : 0.5685
##
##        'Positive' Class : no
##
```

The AIC value does not show much deviation between the model 2 and 3 thus we have considered this Model 3 as our final model with 9 variables to predict the churn.

The **ROC we computed is 84.13%** with good confusion matrix calculations.



Predicting Model based on data provided for test

We are using data set containing the list of consumers that we need to predict their future churn.

On basis of the model built we were able to predict **3.3%** of the customer switching from ABC wireless to other network

```
## Model_Pre_lables_Test
##   no  yes
## 96.7  3.3
```

Insights and conclusions

- Improve Customer Service across all channels
- Improve Network Coverage
- Right Pricing Strategies
- Offer Friends and Family Plan

Contributions

TASK	AMRUTHA	ADITHYA	SRUSHTI	MANOJ
Classification	✓	✓	✓	✓
Modelling	✓	✓	✓	✓
Development in R	✓	✓	✓	✓
Project Report	✓	✓	-	-
Presentation	-	-	✓	✓