PROJECT REPORT

ON

CHURN PREDICTION MODEL

By

Sheetal Nishad

Index

1. Problem Statement

Typical information that is available about customers' concerns demographics, behavioral data, and revenue information. At the time of renewing contracts, some customers do and some do not: they churn. It would be extremely useful to know in advance which customers are at risk of churning, as to prevent it – especially in the case of high revenue customers.

This is a prediction problem. Starting with a small training set, where we can see who has churned and who has not in the past, we want to predict which customer will churn (churn = 1) and which customer will not (churn = 0).

attr 1, attr 2, …, attr n => churn (0/1)

Building the Model in R

Building Model in Tableau

2. Data Set

The data pertains to Telecom Company. Past data set of 3333 customers is provided in the .csv file with different variables

3. Understanding the Data

Import the data set in R with read.csv

```
setwd("D:/Data Analytics/R AcGg/Projects/Projects/Project 2 - Churn Prediction")

library(readxl)
Churn <- read_excel("Churn.xls")
```

Further data set is understood with below codes

```
class(Churn)
View(Churn)
nrow(Churn)
str(Churn)
table(Churn$Churn)
summary(Churn)
```

It is observed that there are total 3333 rows that means data for past 3333 customers are available. Out of these 483 customers churn (didn't renew the contract) and balance continued.

Data is organized correct class as data frame.

Below are the variables in the data set:

"Account.Length"
"VMail.Message"
"Day.Mins"
"Eve.Mins"
"Night.Mins"
"Intl.Mins"
"CustServ.Calls"
"Churn"
"Int.l.Plan"
"VMail.Plan"
"Day.Calls"
"Day.Charge"
"Eve.Calls"
"Eve.Charge"
"Night.Calls"
"Night.Charge"
"Intl.Calls"
"Intl.Charge"
"State"
"Area.Code"
"Phone"

Variable 8 is Churn (Binomial Variable with '0' and '1')

Variable 21 should be omitted as it is just phone number and could not be predictor.

Variable 19 is 'string' which we will also omit.

Some data may be missing or having null values. Same is identified by

```
library(Amelia)
missmap(Churn,col=c("yellow","red"))
any(is.na(Churn))
# No missing values observed
```

Some of the variables may be dependent on others or having strong correlation with each other. This is identified by

```
cor(Churn[, c(-19,-21)])
library(corrplot)
corrplot(cor(Churn[,c(-19,-21)]), type = "upper")
library(DataExplorer)
plot_correlation(Churn, type = 'continuous')
```

there are some direct correlation viz., Day Mins~Day Charge, Eve Mins ~ Eve Charge, Int Min ~ Int Charge

4. Model Building Approach

We built following models mentioned below.

```
# Split the Data Set into 80 ~ 20 for Train and Test: Churn_train
# Split the data into 80~ 20 having equal proportion of '0' & '1' : Churn_train2
# undresampling with 483 'Zeros' and 483 'Ones': data_483
# undersampling with 1449 'zeros' and 483 'ones': data_1449
# undersampling with 2415 'zeros' and 483 'ones': data_2415
```

Testing the Models and finding the Accuracy, ROC and AUC
Choose the best model
Finally adjusting the threshold probability for prediction by building a funciton
Conclusion

5. Model Building

Total five Models were build based on the model building approach.

```
glm_model1 <- glm(formula = Churn ~ Int.l.Plan + CustServ.Calls + Day.Charge +
                    Eve.Mins + VMail.Plan + Intl.Calls + Night.Mins + Intl.Charge +
                    VMail.Message + Eve.Charge, family = binomial(link = "logit"),
                  data = Churn_train)

glm_model2 <- glm(formula = Churn ~ Int.l.Plan + Day.Mins + CustServ.Calls +
                    Eve.Mins + VMail.Plan + Intl.Charge + Intl.Calls + Night.Charge +
                    Intl.Mins + VMail.Message, family = binomial(link = "logit"),
                  data = Churn_train)

glm_model3<- glm(formula = Churn ~ `Day.Calls` + `VMail.Message` + `Intl.Calls` + `Night.Mins` +
`VMail.Plan` +  `Eve.Mins` + `Int.l.Plan` + `CustServ.Calls`, family = binomial(link = "logit"),
                 data = data_483)

glm_model4<- glm(formula = Churn ~ Int.l.Plan + CustServ.Calls + Day.Mins +
                   VMail.Plan + Eve.Mins + Intl.Charge + Night.Charge + Intl.Calls +
                   VMail.Message, family = binomial(link = "logit"),
                 data = data_1449)

glm_model5<- glm(formula = Churn ~ Int.l.Plan + CustServ.Calls + Day.Mins +
                   VMail.Plan + Eve.Mins + Intl.Charge + Intl.Calls + Night.Charge +
                   VMail.Message, family = binomial(link = "logit"),
                 data = data_2415)
```

6. Testing

We checked the dispersion of models with the ratio of deviance and residual deviance

ROC plot for all the Models and Area Under Curve was calculated.

With testing with actual data points, Confusion Matrix were obtained for all the Models and accuracy was calculated based on the formula:

Accuracy = (True Positive + True Negative) / (Total)

Accuracy for the Models is Given below:

Model	Accuracy
Model 1	0.866330390920555
Model 2	0.850609756097561
Model 3	0.721311475409836
Model 4	0.861286254728878
Model 5	0.867591424968474

Model 5 had the highest accuracy, but considering the overall factors we will select model 1 as the final model since the accuracies are almost same.

11. Deciding the Threshold Probability

Function was written to find the accuracy for different threshold probabilities. It was observed initially that accuracy reduces with threshold limits of 0.4 and 0.6. Hence it was tested for different probabilities between 0.4 and 0.6 wherein it was noticed that accuracy is same for 0.506.

Hence Threshold Probability is finalized as 0.506

The customers with probability greater than 0.506 will churn (Churn =1)

12. Conclusion

Thus, the final Model is as below with threshold probability = 0.506

```
glm_model_final <- glm(formula Churn ~ Int.l.Plan + CustServ.Calls + Day.Charge +
                Eve.Mins + VMail.Plan + Intl.Calls + Night.Mins + Intl.Charge +
                VMail.Message + Eve.Charge, family = binomial(link = "logit"), data = train_data)

summary(glm_model_final)
pred_final <- predict(glm_model_final,test_data,type="response")
pred_final <- ifelse(pred_final > 0.506 ,1,0)
```

Summary of the Model is given below

```
Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1989  -0.5207  -0.3427  -0.1912   2.8986
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.951285   0.583028 -13.638  < 2e-16 ***
## Int.l.Plan       2.067146   0.166065  12.448  < 2e-16 ***
## CustServ.Calls   0.507400   0.044288  11.457  < 2e-16 ***
## Day.Charge       0.074615   0.007176  10.398  < 2e-16 ***
## Eve.Mins         3.181745   1.864004   1.707 0.087833 .
## VMail.Plan      -2.363100   0.688112  -3.434 0.000594 ***
## Intl.Calls      -0.095363   0.028684  -3.325 0.000886 ***
## Night.Mins       0.003784   0.001276   2.967 0.003008 **
## Intl.Charge      0.248350   0.086004   2.888 0.003881 **
## VMail.Message    0.047322   0.021304   2.221 0.026332 *
## Eve.Charge     -37.336479  21.929129  -1.703 0.088643 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2133.4  on 2539  degrees of freedom
## Residual deviance: 1658.8  on 2529  degrees of freedom
## AIC: 1680.8
```

File Churn Prediction.R is attached

13. Churn Prediction in Tableau

In order to build Churn Prediction in Tableau, we need to use R Script

Hence Rserve Package may need to install depending upon the version of R

library(Rserve)

Rserve()

Following steps taken to build the Model in Tableau

1. Import the data set
2. Create Parameter 'Threshold Probability' to adjust between the values '0' to '1'
3. Show Parameter control in order to adjust the values through slider
4. Create calculated field 'Predictions'
    a. Use R Script with arguments with all variables
    b. GLM Model to predict the values
5. Create calculated field 'Threshold_ Predictions' to find whether the customer Churn OR do not
6. Create calculated field 'Model_Accuracy' using the R Script with calculation of accuracy from confusion matrix built using 'Churn' and 'Threshold_Predictions'
7. Add Measure Names to Filters and Show Filters
8. Add all variables one by one to Filter and Show Filters
    a. Numerical Variables set for adjusting the units between the range of the 8variable
    b. Categorical Variables set for '0' and '1'
9. Add measure names to Columns
10. Add Model Accuracy, Predictions and Measure Values to Rows
11. Add Model Accuracy, Predictions and Measure Values to Marks and Labels
12. Display Accuracy by adding Model_Accuracy to Labels
13. Selecting and adjusting the variables and values to see the change in predictions and Accuracy

File Churn Prediction Tableau.twbx attached.

Thus, the Model for Churn Prediction in Tableau.