**PROJECT REPORT**

ON

**CHURN PREDICTION MODEL**

**By**

**Jaydeep Potdar**

**Index**

|  |  |
| --- | --- |
| 1 | Problem Statement |
| 2 | Data Set |
| 3 | Understanding Data Set |
| 4 | Model Building Approach |
| 5 | Working with Data |
| 6 | Sampling |
| 7 | Model Building |
| 8 | Testing Models |
| 9 | ROC and AUC |
| 10 | Accuracy of derived models |
| 11 | Deciding Threshold Probability |
| 12 | Conclusion |
| 13 | Preparing for Tableau |
| 14 | Prediction Model in Tableau |

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

1. **Data Set**

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

File reference Churn.csv uploaded in the git hub

Associated data file can be made available from the web

<https://drive.google.com/file/d/0Bxr27gVaXO5sQUlPTmZfMHdIRGM/view?usp=sharing>

1. **Understanding the Data**

Import the data set in R with read.csv

data<-read.csv("D:/Data Analytics with Excel R and Tabaleu/Churn Prediction/Churn.csv")

Further data set is understood with below codes

View(data)

nrow(data)

head(data)

str(data)

names(data)

table(data$Churn)

class(data)

summary(data)

It is observed that there are total 3333 rows that means data for past 3333 customers is 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 needs to be converted to ‘factor’

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

missmap(data, main = "Missing values vs observed") # No missing values observed

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

pairs(~Churn+Account.Length+VMail.Message+Day.Mins+Eve.Mins+Night.Mins

+Intl.Mins+CustServ.Calls+Int.l.Plan+VMail.Plan+Day.Calls+Day.Charge

+Eve.Calls+Eve.Charge+Night.Calls+Intl.Calls+Intl.Charge, data=data)

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

1. **Model Building Approach**

I did the below models for 'original data set' and 'data set after removing outliers'

# Split the Data Set into 80 ~ 20 for Train and Test

# Split the data into 80~ 20 having equal proportion of '0' & '1'

# undresampling with 483 'Zeros' and 483 'Ones'

# undersampling with 1449 'zeros' and 483 'ones'

# undersampling with 2415 'zeros' and 483 'ones'

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

1. **Working with Data**

The data set was considered in two ways:

1. Original data set ( named as data in the R file)
2. After removing Outliers ( named as data\_9 in the R file)
3. **Sampling**

Sampling was done in below ways for original data set

1. Split the Data Set into 80 ~ 20 : train\_data
2. Split the data into 80~ 20 having equal proportion of '0' & '1 : train\_data1
3. undresampling with 483 'Zeros' and 483 'Ones' : train\_data483
4. undersampling with 1449 'zeros' and 483 'ones' : train\_data1449
5. undersampling with 2415 'zeros' and 483 'ones': train\_data2415

Sampling was done in below ways for ‘data\_9’

1. Split the Data Set into 80 ~ 20 : train\_data9
2. Split the data into 80~ 20 having equal proportion of '0' & '1 : train\_data91
3. undresampling with 483 'Zeros' and 483 'Ones' : train\_9\_data483
4. undersampling with 1449 'zeros' and 483 'ones' : train\_9\_data1449
5. undersampling with 2415 'zeros' and 483 'ones': train\_9\_data2415
6. **Model Building**

Total ten Models were build based on data given in point no. 6 earlier

glm\_model1 <- glm(formula = Churn ~ Int.l.Plan + Day.Mins + CustServ.Calls +

VMail.Plan + Eve.Charge + Intl.Charge + Intl.Calls + Night.Mins +

Intl.Mins + VMail.Message, family = binomial(link = "logit"),

data = train\_data)

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 = train\_data1)

glm\_model3<- glm(formula = Churn ~ Int.l.Plan + CustServ.Calls + Day.Mins +

VMail.Plan + Eve.Mins + Intl.Charge + Night.Mins + Intl.Calls +

VMail.Message + Day.Calls + Intl.Mins + Day.Charge, 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)

glm\_model91 <- glm(formula = Churn ~ Int.l.Plan + Day.Mins + CustServ.Calls +

VMail.Plan + Eve.Mins + Intl.Charge + Intl.Calls + Night.Charge,

family = binomial(link = "logit"), data = train\_data9)

glm\_model92 <- glm(formula = Churn ~ Int.l.Plan + Day.Mins + CustServ.Calls +

VMail.Plan + Eve.Mins + Intl.Charge + Intl.Calls + Night.Charge +

Intl.Mins + VMail.Message + Night.Calls, family = binomial(link = "logit"),

data = train\_data91)

glm\_model93<- glm(formula = Churn ~ Int.l.Plan + CustServ.Calls + Day.Mins +

Eve.Mins + VMail.Plan + Intl.Calls + Intl.Charge + Day.Calls,

family = binomial(link = "logit"), data = data\_9\_483)

glm\_model94<- glm(formula = Churn ~ Int.l.Plan + CustServ.Calls + Day.Charge +

Eve.Mins + VMail.Plan + Intl.Charge + Night.Charge + Intl.Calls +

Day.Calls + Intl.Mins, family = binomial(link = "logit"),

data = data\_9\_1449)

1. **Testing**

Check the dispersion of models with the ration of deviance and residual deviance

ROC were 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 | 0.865067466266867 |
| Model 2 | 0.850609756097561 |
| Model 3 | 0.848575712143928 |
| Model 4 | 0.845833333333333 |
| Model 5 | 0.857571214392804 |
| Model 91 | 0.862745098039216 |
| Model 92 | 0.833673469387755 |
| Model 93 | 0.761689291101056 |
| Model 94 | 0.856711915535445 |
| Model 95 | 0.856711915535445 |

**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 + Day.Mins + CustServ.Calls + VMail.Plan +**

**Eve.Charge + Intl.Charge + Intl.Calls + Night.Mins + Intl.Mins + VMail.Message,**

**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

-2.0781 -0.5172 -0.3393 -0.1915 3.1852

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.064335 0.575364 -14.016 < 2e-16 \*\*\*

Int.l.Plan 2.038804 0.163306 12.485 < 2e-16 \*\*\*

Day.Mins 0.013776 0.001228 11.216 < 2e-16 \*\*\*

CustServ.Calls 0.477234 0.043270 11.029 < 2e-16 \*\*\*

VMail.Plan -1.977392 0.654829 -3.020 0.00253 \*\*

Eve.Charge 0.077843 0.015160 5.135 2.82e-07 \*\*\*

Intl.Charge 39.757754 22.009236 1.806 0.07085 .

Intl.Calls -0.091727 0.027968 -3.280 0.00104 \*\*

Night.Mins 0.003768 0.001249 3.017 0.00255 \*\*

Intl.Mins -10.647690 5.942958 -1.792 0.07319 .

VMail.Message 0.032049 0.020541 1.560 0.11869

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2198.0 on 2665 degrees of freedom

Residual deviance: 1720.1 on 2655 degrees of freedom

AIC: 1742.1

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’
   1. Use R Script with arguments with all variables
   2. 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
   1. Numerical Variables set for adjusting the units between the range of the variable
   2. 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.**