

ACADGILD

PROJECT-II

DATA ANALYTICS

Harshita Singh Parihar | Churn Prediction | 22-11-2017

Basic Data Cleaning and Exploring the Data Set

For splitting the data set into a Train and Test Dataset, Package I used is “catools”

Thus,

- Train_DF has 2592 Observations
- Test_DF has 741 Observations

Hypothesis testing

Using T-test,

Null Hypo: There is no significant difference between average number of customer churns and Account_length

Alternate Hypo: There is a significant difference between average number of customer churns and Account_length

Therefore as the p-value is less than 0.05, “we will reject null hypothesis and establish the fact that there is significant difference between average customer churns and customer service calls”.

LOGISTIC REGRESSION MODELING

```
datafile_glm<- glm(Churn~.,train_DF,family = "binomial")
```

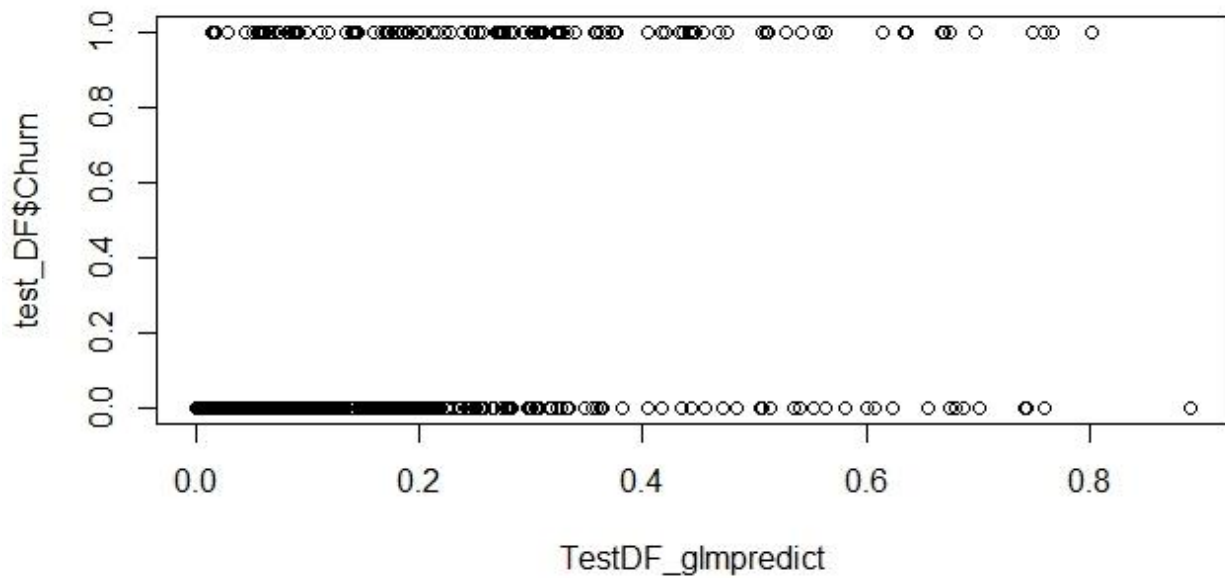
Independent variables with p-value <0.05 will be affecting the regression model relatively more than the ones with p-value>0.05

- Null Deviance is greater than Residual Deviance, which is better sign
- AIC value is also in control

Prediction of datafile_glm on Train_DF

```
TestDF_glmpredict<- predict(datafile_glm,test_DF,type = "response")
```

```
plot(test_DF$Churn~TestDF_glmpredict)
```



Checking with the threshold value of 0.50

```
table_testDF<- table(Actual=test_DF$Churn,Predicted=TestDF_glmpredict>0.5)
```

```
outcome1=floor(TestDF_glmpredict+0.50)
```

```
table(outcome1)
```

Accuracy of the Model

```
accuracy_testDF=(607+27)/(607+20+27+88)
```

Therefore,

Accuracy = **85.444%** (test_DF)

DECISION TREE MODELING

Using the packages

- Rpart
- Rpart.plot

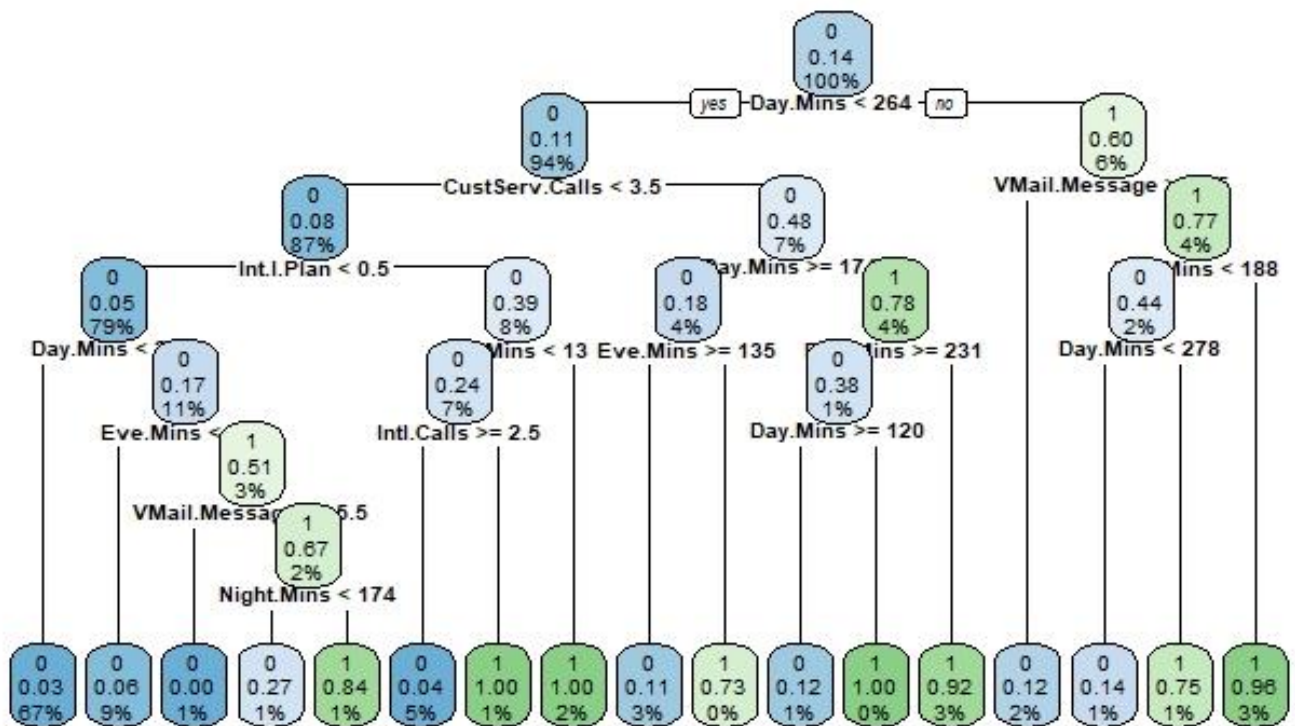
Running the model:

```
datafile_Decisiontree<- rpart(Churn~.,data = train_DF,method = "class")
```

```
plot(datafile_Decisiontree,cex=0.5)
```

```
text(datafile_Decisiontree,cex=0.5)
```

```
rpart.plot(datafile_Decisiontree,cex=0.6)
```



Predict Decision Tree Model on Test_DF

```
DecisionTree_Predict_TestDF<- predict(datafile_Decisiontree,test_DF,type = "class")
```

Confusion Matrix

```
confusionMatrix(DecisionTree_Predict_TestDF,test_DF$Churn)
```

Hence,

- Accuracy of Decision Tree Model on Test DF = **93.13%**
- Sensitivity= **0.9777**
- specificity= **0.6783**

ROC - AUC

Packages

- InformationValue
- pRoc

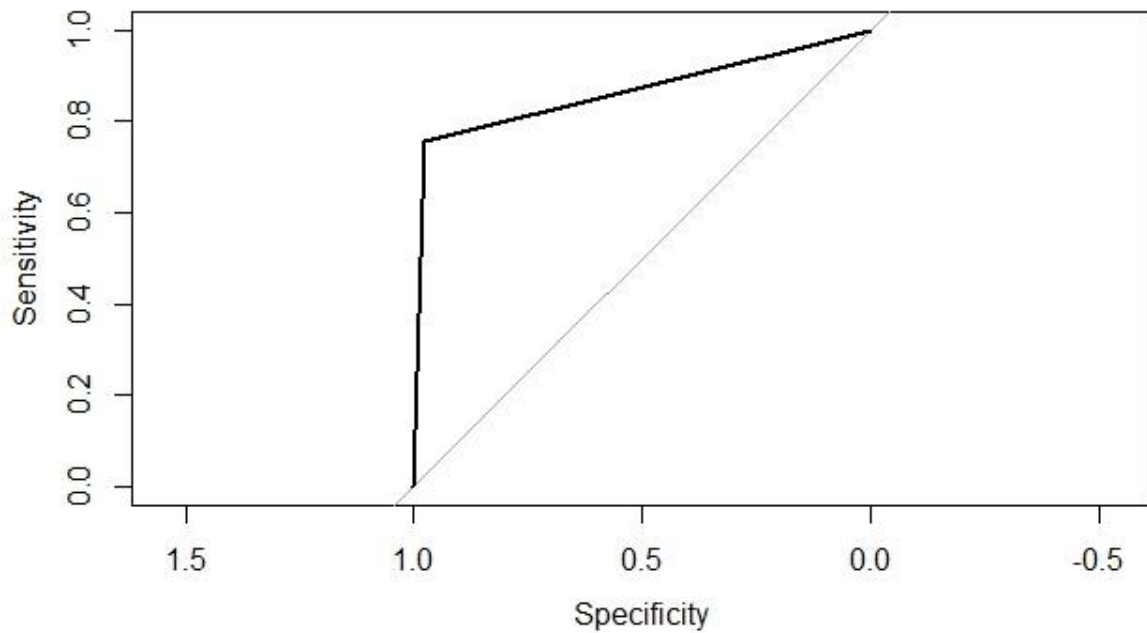
```
rpartpredict_TestDF<- predict(datafile_Decisiontree,test_DF,type = "vector")
```

```
AUC_TestDF<-auc(test_DF$Churn,rpartpredict_TestDF)
```

Thus,

- AUC is **82.8%**

```
plot(roc(test_DF$Churn,rpartpredict_TestDF))
```



RANDOM FOREST MODELING

```
Datafile_RF_TrainDF<-randomForest(Churn~.,data = train_DF,ntrees=500,do.trace=100)
```

```
Datafile_RF_TrainDF$predicted
```

```
Datafile_RF_TrainDF$importance
```

Prediction

```
RFpredict_TestDF<- predict(Datafile_RF_TrainDF,test_DF,type = "class")
```

```
table(Actual=test_DF$Churn,Predicted= RFpredict_TestDF>0.5)
```

```
RFpredict_TestDF$predicted
```

Accuracy

```
accuracy_RF_TestDF<- (620+78)/(620+7+78+37)
```

Accuracy of Random Forest Model on Test DF = **94.07008%**

Confusion Matrix

```
confusionMatrix(RFpredict_TestDF>0.50,test_DF$Churn)
```

```
getTree(Datafile_RF_TrainDF,k=40,labelVar = TRUE)
```

NEURAL NETWORKS MODELING

Packages used

- nnet
- neuralnet
- caret

Running the model:

```
datafile_nnet= nnet(Churn~.,data=train_DF,size=5,maxit=1000)
```

```
summary(datafile_nnet)
```

Predict Neural Networks Model on Test_DF

```
TestDF_nnetpredict=predict(datafile_nnet, type = "raw")
```

Confusion Matrix

```
confusionMatrix(TestDF_nnetpredict,test_DF$Churn)
```

Accuracy from Confusion Matrix = **85.26%**

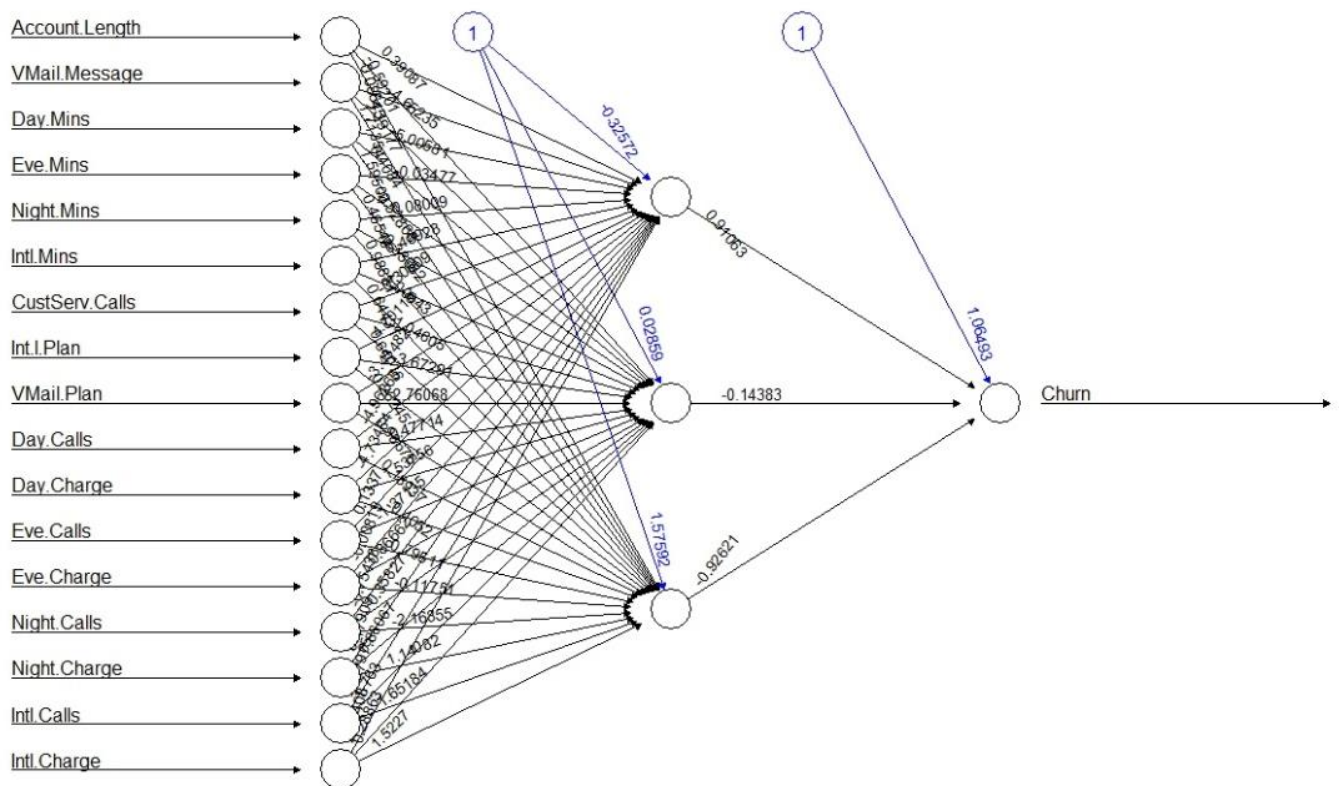
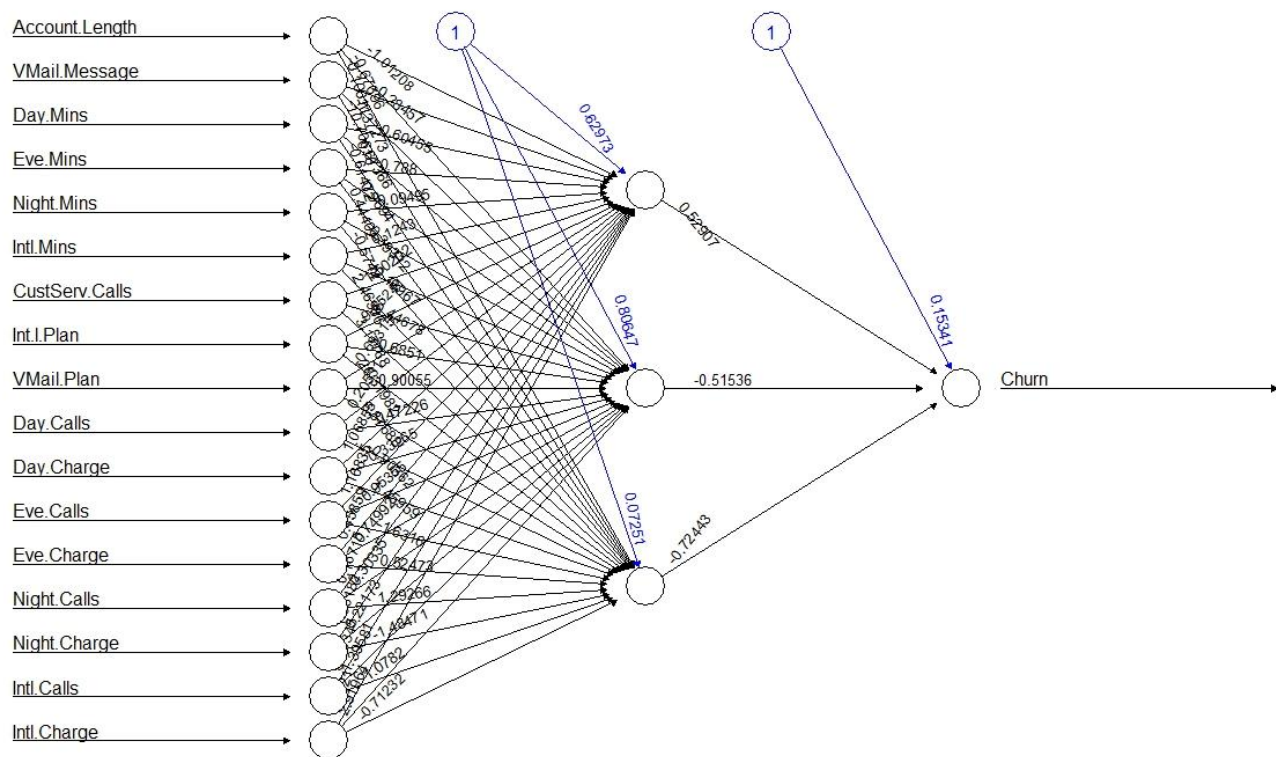
Plotting Neural Network

```
Datafile_neuralnet=neuralnet(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+Night.Charge+Intl.Calls+Intl.Charge, data=train_DF, hidden=3)
```

```
plot(Datafile_neuralnet)
```

```
Datafile_TestDF_neuralnet=neuralnet(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+Night.Charge+Intl.Calls+Intl.Charge, data=test_DF, hidden=3)
```

```
plot(Datafile_TestDF_neuralnet)
```

STARTING THE SERVER

For integrating R with Tableau, Package used is

- Rserve