**Data Mining Project**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Prepared By** | **SID** | **TAID** |
| **Name** | Rajib Mandal | 118310 | TA17002 |
| **Submitted Date** | 20-JAN-2018 | | |

Table of Contents

[1. Data Set Information 3](#_Toc504223681)

[2. Attribute Information 3](#_Toc504223682)

[3. Project Objective 4](#_Toc504223683)

[4. Abstract 4](#_Toc504223684)

[5. Data Analysis 4](#_Toc504223685)

[5.1 Check the structure of the Data Set 4](#_Toc504223686)

[5.2 VISUALIZATION OF Data 5](#_Toc504223687)

[5.3 Data Cleaning and preparation 6](#_Toc504223688)

[5.4 Calculate mode for all categorical field and replace unknown value with mode 6](#_Toc504223689)

[5.5 TRANSFORM FOLLOWING categorical variables into dummy variables 7](#_Toc504223690)

[5.6 Remove unwanted fields from the dataset 8](#_Toc504223691)

[6. Modelling the Data 8](#_Toc504223692)

[7. Logistic Regression 8](#_Toc504223693)

[7.1 Check for multicollinearity 8](#_Toc504223694)

[7.2 Building best fitted model on training data 9](#_Toc504223695)

[7.3 Deviance Analysis 9](#_Toc504223696)

[7.4 Apply Logistic regression on test data 10](#_Toc504223697)

[7.5 ROC CURVE & performance of the model 10](#_Toc504223698)

[8. Decision Tree 11](#_Toc504223699)

[8.1 Build a model of recursive partitioning & regression tree 11](#_Toc504223700)

[8.2 ROC CURVE & performance of the model 12](#_Toc504223701)

[9. Random Forest 13](#_Toc504223702)

[9.1 Random Forest Initial Model 13](#_Toc504223703)

[9.2 Random Forest Error Rate 14](#_Toc504223704)

[9.3 Tune Random Forest Model 14](#_Toc504223705)

[9.4 Fit the Random Forest Model After Tuning 15](#_Toc504223706)

[9.5 ROC CURVE & performance of the model 16](#_Toc504223707)

[9.6 Calculate FPR & TPR 17](#_Toc504223708)

[1. Conclusion 17](#_Toc504223709)

[2. ANNEXTURE 17](#_Toc504223710)

# Data Set Information

The Consumer Complaint Database contains complaints CFPB (Consumer Financial Protection Bureau) have received about consumer financial products and services. The Consumer Complaint Database shows the consumer’s original product, sub-product, issue, and sub-issue selections consistent with the options available on the form at the time the consumer submitted the complaint.

The source of the data is(**https://catalog.data.gov/dataset/consumer-complaint-database**)

# Attribute Information

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Description | Type | Data Type |
| Date received | Complaint receive date by the agency(CFPB) | Categorical | Date Time |
| Product | The type of product the consumer identified in the complaint | Categorical | Text |
| Sub-product | The type of sub-product the consumer identified in the complaint | Categorical | Text |
| Issue | The issue the consumer identified in the complaint | Categorical | Text |
| Sub-issue | |  | | --- | | The sub-issue the consumer identified in the complaint | | Categorical | Text |
| Consumer complaint narrative | Consumer complaint detail about product | Categorical | Text |
| Company public response | The company's optional, public-facing response to a consumer's complaint | Categorical | Text |
| Company | |  | | --- | | The complaint is about the company | | Categorical | Text |
| State | The state of the mailing address provided by the consumer | Categorical | Text |
| ZIP code | The mailing ZIP code provided by the consumer | Categorical | Text |
| Tags | Categories of the consumer | Categorical   * Older American * Servicemember * Older American, Servicemember | Text |
| Consumer consent provided? | Consumer consent about the issue | Categorical   * Consent not provided * Consent provided * Consent withdrawn | Text |
| Submitted via | Complaintsubmittedthrough the channels | Categorical   * Email * Fax * Phone * Postal mail * Referral * Web | Text |
| Date sent to company | The date the agency sent the complaint to the company | Categorical | Date |
| Company response to consumer | Company response to the customer. | Categorical   * Closed * Closed with explanation * Closed with monetary relief * Closed with relief * Closed without relief * Closed with non-monetary relief | Text |
| Timely response? | If the   company gave a timely response | Categorical   * Yes * No | Binary |
| Consumer disputed? | If the bank disputes the allegations contained in the complaint | Categorical   * Yes * No | Binary |
| Complaint ID | The unique identification number for a complaint | Categorical | Numeric |

# Project Objective

The data is related with customer complaints about financial products and services of a leading North American Bank. The classification goal is to predict if the bank disputes the allegations contained in the complaint.

# Abstract

To classify whether each complaint leads to dispute, I analysed customer compliant using 3 different models: **Logistic Regression, Decision Trees & Random Forest**. Overall, **Random Forest** proved to be the best in determining disputes with an **83.31 %** accuracy rate.

# Data Analysis

## Check the structure of the Data Set

consumercomplain<-read.csv("Consumer\_Complaints.csv",header=TRUE, na.strings=c("","NA"))

str(consumercomplain)

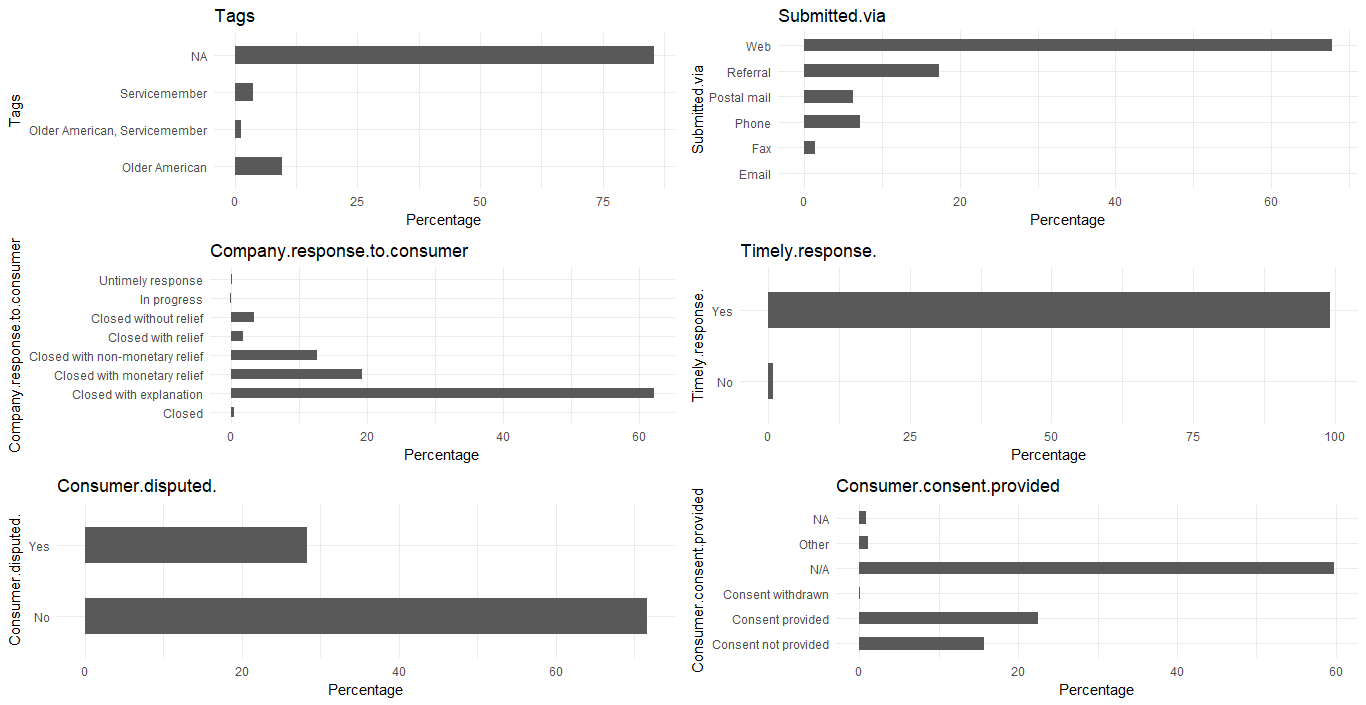
From the output, it is seen that there are 1 numerical variable and 16 categorical variables, excluding the variable that represents the dispute status, “Consumer disputed”, which will be used as the response variable the models are attempting to predict.

Since there are missing values in the data set, we impute data and proceed to create a few visualizations to explore the data set and gain some initial insights.

## VISUALIZATION OF Data

We detect patterns and anomalies in the following categorical variables through explanatory data analysis using bar plot

* Consumer.consent.provided.
* Submitted.via
* Company.response.to.consumer
* Consumer.consent.provided
* Timely.response.
* Consumer.disputed.



From the above graph we can see that following variables are not evenly distributed

* Consumer.consent.provided.
* Submitted.via
* Company.response.to.consumer
* Consumer.consent.provided
* There seems to be minimal record for “Email” in Submitted.via field.
* There seems to be minimal record for “Consent withdrawn” in Consumer.consent.provided field
* There seems to be very few records for “Untimely response”,”In Progress”,”Closed”, “Closed with relief” in Company.response.to.consumer field

Here is the code snippet for bar plot.

p1 <- ggplot(consumercomplain, aes(x=consumercomplain$Tags)) + ggtitle("Tags") + xlab("Tags") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p2 <- ggplot(consumercomplain, aes(x=consumercomplain$Submitted.via)) + ggtitle("Submitted.via") + xlab("Submitted.via") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p3 <- ggplot(consumercomplain, aes(x=consumercomplain$ Company.response.to.consumer)) + ggtitle("Company.response.to.consumer") + xlab("Company.response.to.consumer") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p4 <- ggplot(consumercomplain, aes(x=consumercomplain$Timely.response.)) + ggtitle("Timely.response.") + xlab("Timely.response.") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p5 <- ggplot(consumercomplain, aes(x=consumercomplain$ Consumer.disputed.)) + ggtitle("Consumer.disputed.") + xlab("Consumer.disputed.") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p6 <- ggplot(consumercomplain, aes(x=consumercomplain$Consumer.consent.provided)) + ggtitle("Consumer.consent.provided") + xlab("Consumer.consent.provided") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

grid.arrange(p1, p2, p3, p4,p5,p6, ncol=2)

## Data Cleaning and preparation

The following tasks are performed during this phase

Replace “N/A" with “NA” in the data set

consumercomplain[consumercomplain=="N/A"] <- "NA"

## Calculate mode for all categorical field and replace unknown value with mode

We replace missing values with mode for the categorical field and with mean for numeric fields.

Here is the code.

Mode <- function (x, na.rm) {

xtab <- table(x)

xmode <- names(which(xtab == max(xtab)))

if (length(xmode) > 1) xmode <- ">1 mode"

return(xmode)

}

for (var in 1:ncol(consumercomplain))

{

if (class(consumercomplain[,var])=="numeric")

{

consumercomplain[is.na(consumercomplain[,var]),var] <- mean(consumercomplain[,var], na.rm = TRUE)

} else if (class(consumercomplain[,var]) %in% c("character", "factor"))

{

consumercomplain[is.na(consumercomplain[,var]),var] <- Mode(consumercomplain[,var], na.rm = TRUE)

}

}

## TRANSFORM FOLLOWING categorical variables into dummy variables

We convert following categorical variable into numeric variables.

* Consumer.consent.provided.
* Timely.response
* Consumer.disputed.
* Tags
* Submitted.via

consumercomplain$Consumer.consent.provided.[consumercomplain$Consumer.consent.provided.=="NA"] <- "Consent provided"

consumercomplain$Timely.response<- ifelse(consumercomplain$Timely.response. == "Yes", 1, 0)

consumercomplain$Consumer.disputed\_Y<-ifelse(consumercomplain$Consumer.disputed.=='Yes',1,0)

consumercomplain$Tags\_Service <- ifelse(consumercomplain$Tags == "Older American", 1, 0)

consumercomplain$Consumer.consent.provided <- ifelse(consumercomplain$Consumer.consent.provided. == "Consent provided", 1, 0)

consumercomplain$Submitted.via[consumercomplain$Submitted.via=="Web"]<-"Email"

for(level in unique(consumercomplain$Submitted.via)){

consumercomplain[paste("Submitted.via", gsub("-","\_", gsub(" ","\_",level, fixed=TRUE), fixed=TRUE), sep = "\_")] <- ifelse(consumercomplain$Submitted.via == level, 1, 0)

}

## Remove unwanted fields from the dataset

We remove the following columns as we do not need them for the analysis.

* Timely.response
* Date.received
* Company.response.to.consumer
* Submitted.via
* Consumer.consent.provided.
* Tags
* Consumer.disputed.
* Sub.product
* Consumer.complaint.narrative
* Sub.issue

Here is the code snippet.

consumercomplain$Timely.response. <- NULL

consumercomplain$Date.received<-NULL

consumercomplain$Company.response.to.consumer<-NULL

consumercomplain$Submitted.via<-NULL

consumercomplain$Consumer.consent.provided.<-NULL

consumercomplain$Tags<-NULL

consumercomplain$Consumer.disputed.<-NULL

consumercomplain$Sub.product<-NULL

consumercomplain$Sub.issue<-NULL

consumercomplain$Consumer.complaint.narrative<-NULL

consumercomplain$ZIP.code<-NULL

consumercomplain$Date.sent.to.company<-NULL

# Modelling the Data

Create training and test data sets to run the model.   Split 70% of the observations into the training set and the remaining 30% into the test set.

set.seed(200)

splitdata<-sample.split(consumercomplain,SplitRatio=.7)

consumercomplaintrainig<-subset(consumercomplain,splitdata==TRUE)

consumercomplaintesting<-subset(consumercomplain,splitdata==FALSE)

# Logistic Regression

## Check for multicollinearity

res <-glm(Consumer.disputed\_Y~

Product

+Company.public.response

+Timely.response

+Tags\_Service

+Consumer.consent.provided

+Submitted.via\_Referral

+Submitted.via\_Postal\_mail

+Submitted.via\_Phone

+Submitted.via\_Fax

,data=consumercomplain,family = "binomial")

summary(res)

#Check for multicollinearity

vif(res)

Since the VIF of all the below variables are less than 10, we can conclude there is no relation between explanatory variables.

|  |  |
| --- | --- |
| Variable Name | VIF |
| Product | 1.102432 |
| Company.public.response | 1.067348 |
| Timely.response | 1.003773 |
| Tags\_Service | 1.024123 |
| Consumer.consent.provided | 1.129025 |
| Submitted.via\_Referral | 1.123441 |
| Submitted.via\_Postal\_mail | 1.044660 |
| Submitted.via\_Phone | 1.059998 |
| Submitted.via\_Fax | 1.014993 |

## Building best fitted model on training data

churn\_logistic<-glm(Consumer.disputed\_Y~

Product

+Company.public.response

+Timely.response

+Tags\_Service

+Consumer.consent.provided

+Submitted.via\_Referral

+Submitted.via\_Postal\_mail

+Submitted.via\_Phone

+Submitted.via\_Fax

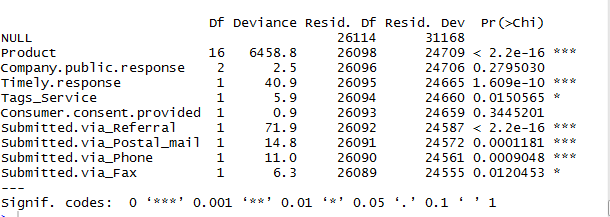
,data= consumercomplaintrainig,family = "binomial")

summary(churn\_logistic)

|  |  |  |
| --- | --- | --- |
| Summary | | Degree of Freedom |
| Null deviance | 31168 | 26114 |
| Residual deviance | 24 555 | 26089 |
| AIC | 245556 |  |

## Deviance Analysis

anova(churn\_logistic, test="Chisq")



Analysing the above deviance table we can see the drop in deviance when adding each variable one at a time. Adding **Product**, **Timely. Response** and **Submiited.via\_Referral** significantly reduces the residual deviance. The other variables such as **Submitted.via\_phone and Submitted.via\_Postal\_mail, Tag\_services & submitted\_via\_phone** seem to improve the model less even though they all have low p-values.

## Apply Logistic regression on test data

prediction\_data<-predict(churn\_logistic,consumercomplaintesting,type="response",se.fit=FALSE)

predicted\_outcome<-ifelse(prediction\_data>0.5,1,0)

churn<-data.frame(consumercomplaintesting,prediction\_data,predicted\_outcome)

tb<-table(predicted\_outcome,consumercomplaintesting$Consumer.disputed\_Y)

acc<-sum(diag(tb))/sum(tb)

acc

miss\_class<-1-acc

Here we find the accuracy of the logistic model is **80.75%**

## ROC CURVE & performance of the model

roc\_pred<-predict(churn\_logistic,consumercomplaintesting,type="response")

roc\_obj<-prediction(roc\_pred,consumercomplaintesting$Consumer.disputed\_Y)

roc\_performance<-performance(roc\_obj,"tpr","fpr")

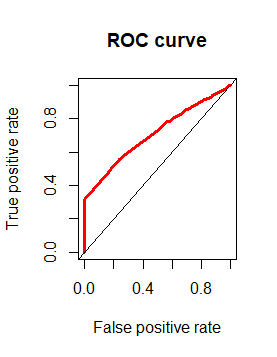
plot(roc\_performance,col=2, lwd=3, main="ROC curve")

abline(a=0,b=1)

accuracy<-performance(roc\_obj,"auc")

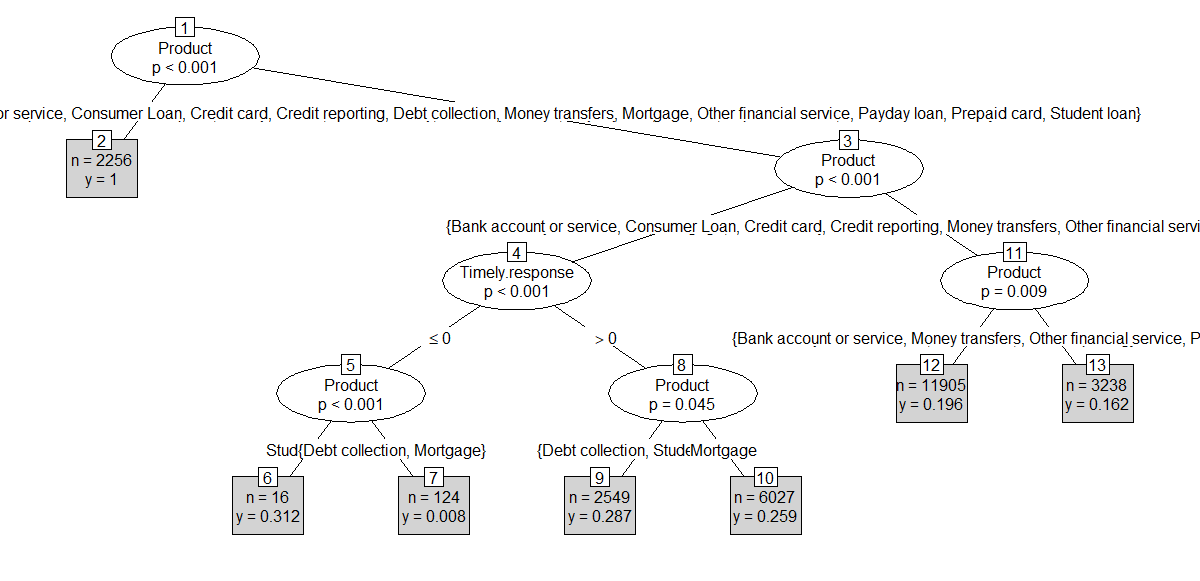
accuracy

The area under the curve (AUC) is used to summarize the performance of the model, and it is currently about **0.709031**. This is a pretty high value, which is an indicator of a good performance.



# Decision Tree

## Build a model of recursive partitioning & regression tree

For illustration purpose, we will use only two variables for plotting Decision Trees, they are “Timely.response” and “Product” .****

We are using following the variables for confusion matrix table and make predictions.

* Product
* Issue
* Company.public.response
* Company
* Complaint.ID
* Timely.response
* Tags\_Service
* Consumer.consent.provided
* Submitted.via\_Email
* Submitted.via\_Referral
* Submitted.via\_Postal\_mail
* Submitted.via\_Phone
* Submitted.via\_Fax

fit <-rpart(Consumer.disputed\_Y~

Product

+Issue

+Company.public.response

+Company

+Complaint.ID

+Timely.response

+Tags\_Service

+Consumer.consent.provided

+Submitted.via\_Email

+Submitted.via\_Referral

+Submitted.via\_Postal\_mail

+Submitted.via\_Phone

+Submitted.via\_Fax,

data=consumercomplaintrainig

,method="class"

)

Here we find the accuracy of the decision tree model is 83.17%

Here is the code snippet

aabpredict<-predict(fit,consumercomplaintesting,type="class")

head(aabpredict)

tb<-table(consumercomplaintesting$Consumer.disputed\_Y,aabpredict)

sum(diag(tb))/sum(tb)

## ROC CURVE & performance of the model

aoc<-predict(fit,consumercomplaintesting,type="prob")

aad<-aoc[,2]

head(aad)

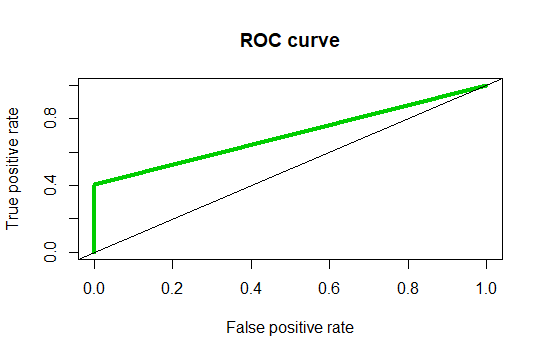
aoe<-prediction(aad,consumercomplaintesting$Consumer.disputed\_Y)

rocc<-performance(aoe,"tpr","fpr")

plot(rocc,col=3, lwd=4, main="ROC curve")

abline(a=0,b=1)

performance(aoe,"auc")

The area under the curve (AUC) is used to summarize the performance of the model, and it is currently about **0.7018**. This is a pretty high value, which is an indicator of a good performance. The accuracy for Decision Tree has hardly improved.

# Random Forest

## Random Forest Initial Model

rfmodel<-randomForest(as.factor(Consumer.disputed\_Y)~

Product

+Company.public.response

+Complaint.ID

+Timely.response

+Tags\_Service

+Consumer.consent.provided

+Submitted.via\_Email

+Submitted.via\_Referral

+Submitted.via\_Postal\_mail

+Submitted.via\_Phone

, data=consumercomplaintrainig

, method="class"

,ntree=500

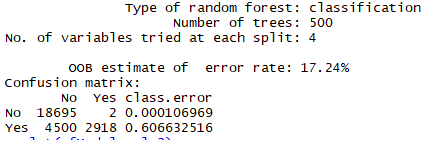
,mtry=4

,keep.forest=TRUE

)

print(rfmodel)

The output is given below



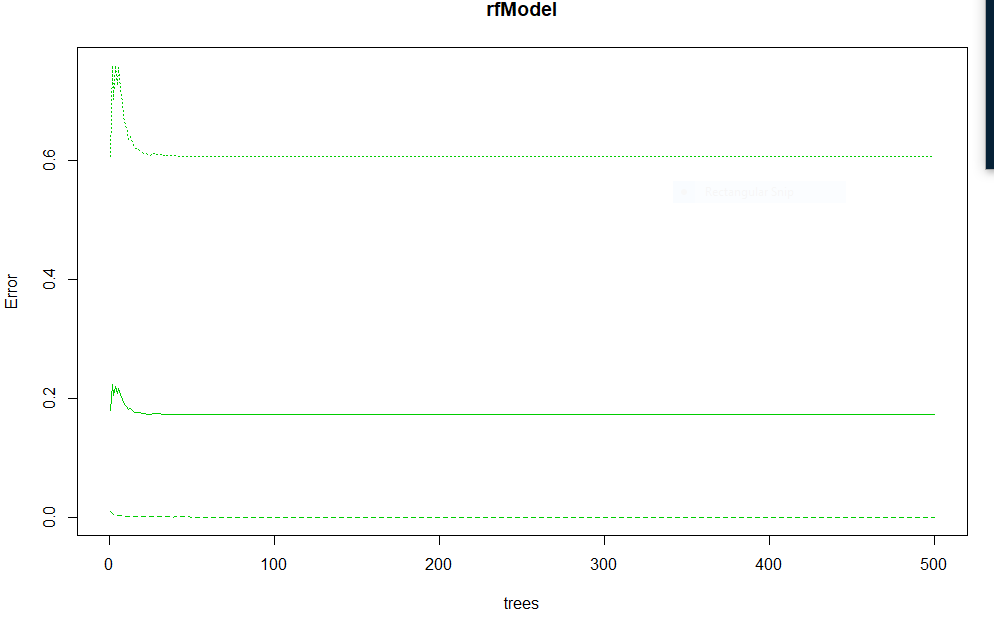
Here OOB error rate is **17.24%**

The error rate is relatively low when predicting “No”, and the error rate is much higher when predicting “Yes”

## Random Forest Error Rate

plot(rfmodel)

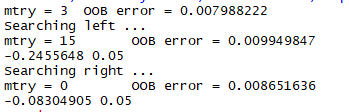
This plot to help us determine the number of trees. As the number of trees increases, the OOB error rate decreases, and then becomes almost constant. We are not able to decrease the OOB error rate after about 50 trees.



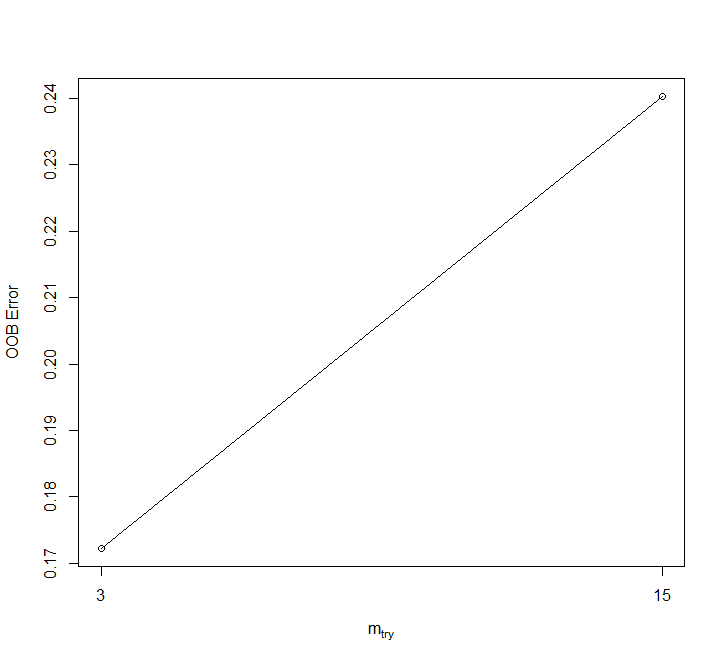
## Tune Random Forest Model

tuneRandomForest <- tuneRF(consumercomplaintrainig[, -5], consumercomplaintrainig[, 5], stepFactor = 0.2, plot = TRUE, ntreeTry = 200, trace = TRUE, improve = 0.05)

The output is given below



The graph is given below



We use this plot to give us some ideas on the number of mtry to choose. OOB error(0.00798)

rate is at the lowest when mtry is 3. Therefore, we choose mtry=3

## Fit the Random Forest Model After Tuning

# Fitted Random Forest

rfModel<-randomForest(as.factor(Consumer.disputed\_Y)~

Product

+Company.public.response

+Complaint.ID

+Timely.response

+Tags\_Service

+Consumer.consent.provided

+Submitted.via\_Email

+Submitted.via\_Referral

+Submitted.via\_Postal\_mail

+Submitted.via\_Phone

+Submitted.via\_Fax

,data=consumercomplaintrainig

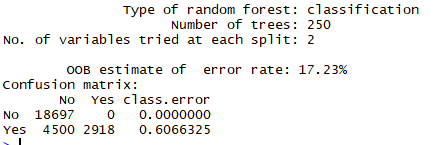
,ntree=250

,mtry=3

,importance = TRUE

, method="class")

print(rfModel)



OOB error rate decreased to 17.23% from 17.24% earlier.

Here we find the accuracy of the Random Forest model is **83.1765%**

The code snippet is given below.

aapr1<-predict(rfmodel,consumercomplaintesting,type="class")

aapr2<-ifelse(aapr1>=.5,1,0)

tb1<-table(consumercomplaintesting$Consumer.disputed\_Y,aapr2)

sum(diag(tb1))/sum(tb1)

## ROC CURVE & performance of the model

aapr3<-prediction(aapr1,consumercomplaintesting$Consumer.disputed\_Y)

aapr4<-performance(aapr3,"tpr","fpr")

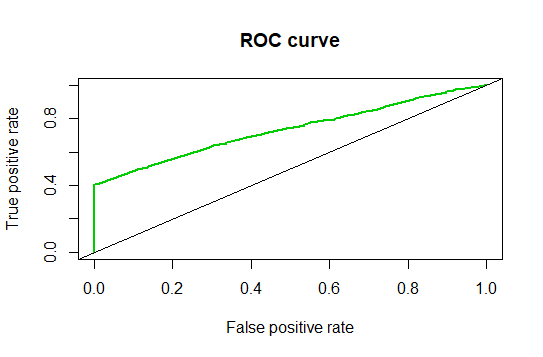
plot(aapr4, col=3,lwd=2, main="ROC curve")

abline(a=0,b=1)

aapr5<-performance(aapr3,"auc")

aapr5

The area under the curve (AUC) is used to summarize the performance of the model, and it is currently about **0.7467** This is a pretty high value, which is an indicator of a good performance.

****

## Calculate FPR & TPR

The following code provides FPR & TPR

alpha<-round(as.numeric(unlist(aapr4@alpha.values)),4)

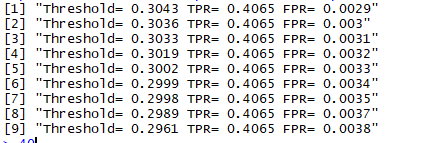
fpr<-round(as.numeric(unlist(aapr4@x.values)),4)

tpr<-round(as.numeric(unlist(aapr4@y.values)),4)

i<-which(round(alpha,2)==0.35)

paste("Threshold=",(alpha[i]),"TPR=", tpr[i], "FPR=", fpr[i])

The output is



# Conclusion

Comparison between different Model

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy(%)** | **Area under the curve(%)** |
| Logistic Regression | 80.75 | 70.90 |
| Decision Tree | 83.17 | 70.18 |
| Random Forest | 83.31 | 74.67 |

From the above example, we can see that **Random Forest** performed better than Decision Tree and Logistic regression customer complaint analysis for this dataset.

Throughout the analysis, we have learned several important things:

* 40% consumer complaint which is disputed are properly classified as disputed and .3% consumer complaint are improperly classified as disputed(Refer section 9.5)
* Features such as Product, Timely.response,Submiited.via appear to play a role in consumer dispute.
* There does not seem to be a relationship between Company public response and consumer dispute.
* There does not seem to be a relationship between Tags and consumer dispute.

# ANNEXTURE

1. Dataminig\_Project\_TA17002.R **-Source code**
2. Dataminig\_Project\_CONSOLE\_OUTPUT\_TA17002.txt -**Console output**
3. Consumer\_Complaints.csv – **Customer complaint data**