# Stat 6620

# Project Work

# Asati Surabhi

# Topic: Quora Question Pairs

# Project Goal:To identify question pairs that have the same intent

library(ggplot2)  
library(readr)  
library(tm)

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':  
##   
## annotate

library(syuzhet)  
library(SnowballC)  
library(NLP)  
#install.packages("NLP")

# Step 1: Collecting data

# To develop the Quora Question Pairs model, I am using the training data with labels

# <https://www.kaggle.com/c/quora-question-pairs/data>

# Dataset includes over 400K Questions. I am working on the first 5000 samples.

# Data fields:

# Data includes 5 features and is\_duplicate outcome, as follows:

# id - the id of a training set question pair

# qid1, qid2 - unique ids of each question (only available in train.csv)

# question1, question2 - the full text of each question

# is\_duplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

mydata <-read.csv("train.csv", header = TRUE, nrows = 5000, stringsAsFactors = FALSE)  
str(mydata)

## 'data.frame': 5000 obs. of 6 variables:  
## $ id : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ qid1 : int 1 3 5 7 9 11 13 15 17 19 ...  
## $ qid2 : int 2 4 6 8 10 12 14 16 18 20 ...  
## $ question1 : chr "What is the step by step guide to invest in share market in india?" "What is the story of Kohinoor (Koh-i-Noor) Diamond?" "How can I increase the speed of my internet connection while using a VPN?" "Why am I mentally very lonely? How can I solve it?" ...  
## $ question2 : chr "What is the step by step guide to invest in share market?" "What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?" "How can Internet speed be increased by hacking through DNS?" "Find the remainder when [math]23^{24}[/math] is divided by 24,23?" ...  
## $ is\_duplicate: int 0 0 0 0 0 1 0 1 0 0 ...

# Step 2: Exploring and Preparing the data

count <- 1500  
train <- mydata[1:count,]   
test <- mydata[(count+1):(count\*2),]  
  
dim(train)

## [1] 1500 6

dim(test)

## [1] 1500 6

head(train)

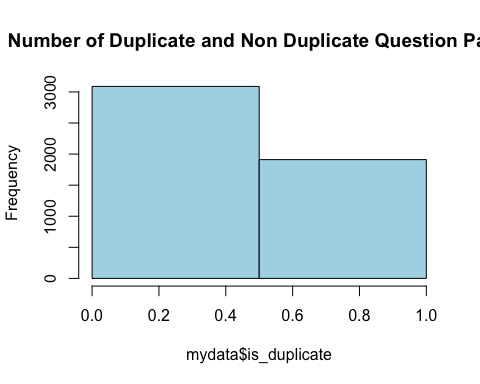
## id qid1 qid2  
## 1 0 1 2  
## 2 1 3 4  
## 3 2 5 6  
## 4 3 7 8  
## 5 4 9 10  
## 6 5 11 12  
## question1  
## 1 What is the step by step guide to invest in share market in india?  
## 2 What is the story of Kohinoor (Koh-i-Noor) Diamond?  
## 3 How can I increase the speed of my internet connection while using a VPN?  
## 4 Why am I mentally very lonely? How can I solve it?  
## 5 Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?  
## 6 Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?  
## question2  
## 1 What is the step by step guide to invest in share market?  
## 2 What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?  
## 3 How can Internet speed be increased by hacking through DNS?  
## 4 Find the remainder when [math]23^{24}[/math] is divided by 24,23?  
## 5 Which fish would survive in salt water?  
## 6 I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?  
## is\_duplicate  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 1

head(test)

## id qid1 qid2  
## 1501 1500 2987 2988  
## 1502 1501 2989 2990  
## 1503 1502 2991 2992  
## 1504 1503 2993 2994  
## 1505 1504 2995 2996  
## 1506 1505 2997 2998  
## question1  
## 1501 How do you write a beautiful introduction to an essay?  
## 1502 What do people from Myanmar think about Indians?  
## 1503 What are the best things to buy on Amazon?  
## 1504 What side dishes go well with hamburgers?  
## 1505 Where can I get best assistance in Sydney for any property purchasing?  
## 1506 Have you ever experienced anything supernatural or paranormal?  
## question2  
## 1501 How do I write the introduction of an essay?  
## 1502 What do people think about Indian women?  
## 1503 What is the best thing I can buy for 2\342\202\254 on Amazon?  
## 1504 What side dishes would go well with lamb chops?  
## 1505 Where can I get knowledgeable assistance for transfer of property in Sydney?  
## 1506 Is there a paranormal event that you have experienced, which science cannot explain?  
## is\_duplicate  
## 1501 1  
## 1502 0  
## 1503 0  
## 1504 0  
## 1505 1  
## 1506 0

# Number of Duplicate and Non Duplicate Question Pairs

#Examining the distribution of the outcome (is\_duplicate) variable  
  
hist(mydata$is\_duplicate, breaks = 3, col = "lightblue", border = "black", main = "Number of Duplicate and Non Duplicate Question Pairs")



# Text Preprocessing

# Building features(5):

# 1. Matching word counts (1 feature)

# 2. Sentiment Analysis (4 features, 2 per question)

# Process:

# a) Text cleaning: Tokenizatiom, convert all tokens to lower case,

# remove punctuations, special tokens, etc.

# b) Text normalization

prepdata <- function(inputdata){  
 df = data.frame()  
 df.new = data.frame()  
 for (i in 1:nrow(inputdata))  
 {  
 q1 <- Corpus(VectorSource(as.character(inputdata$question1[i])))  
   
 #Text cleaning  
 #Removing punctuations  
 q1 <- tm\_map(q1, removePunctuation)   
   
 #Removing Numbers  
 q1 <- tm\_map(q1, removeNumbers)   
 q1 <- tm\_map(q1, tolower)   
 q1 <- tm\_map(q1, stemDocument)   
 q1 <- tm\_map(q1, stripWhitespace)   
   
   
 #Removing Stop Words   
   
 q1 <- tm\_map(q1, removeWords, stopwords("english"))   
   
 #Converting documents into text documents  
 q1 <- tm\_map(q1, PlainTextDocument)   
 doc = TermDocumentMatrix(q1)   
 a11 = doc$dimnames$Terms  
   
 q2 <- Corpus(VectorSource(as.character(inputdata$question2[i])))  
 q2 <- tm\_map(q2, removePunctuation)   
 q2 <- tm\_map(q2, removeNumbers)   
 q2 <- tm\_map(q2, tolower)   
 q2 <- tm\_map(q2, stemDocument)   
 q2 <- tm\_map(q2, stripWhitespace)   
 q2 <- tm\_map(q2, removeWords, stopwords("english"))   
 q2 <- tm\_map(q2, PlainTextDocument)   
   
 doc = TermDocumentMatrix(q2)   
   
 b11 = doc$dimnames$Terms  
 c11 = a11 %in% b11  
   
 same\_items = sum(c11)  
 distinct\_items = length(a11) + length(b11)  
   
 match\_count = (2\*same\_items)/(distinct\_items)  
   
 #Sentiment Analysis   
 sentiment1 <- get\_nrc\_sentiment(as.character(inputdata$question1[i]))  
 sentiment2 <- get\_nrc\_sentiment(as.character(inputdata$question2[i]))  
   
 p1 = sum(sentiment1$positive)  
 p2 = sum(sentiment2$positive)  
 n1 = sum(sentiment1$negative)  
 n2 = sum(sentiment2$negative)  
   
 df.new = cbind(match\_count,p1,p2,n1,n2)  
 df = rbind(df,df.new)  
   
 }  
 return(df)  
}

# Construct train and Test data sets with additional features that express relationship between measured characterstics

#Calulating features  
mydata\_df <- prepdata(mydata[1:(count\*2),])  
mydata\_new = cbind(mydata[1:(count\*2),],mydata\_df[1:(count\*2),])  
  
#Creating a data frame with features: is\_duplicate, match\_count, p1,p2,n1,n2  
mydata\_new = mydata\_new[,6:11]  
mydata\_new[,c(1,3:6)] = lapply(mydata\_new[,c(1,3:6)],as.factor)  
  
mydata\_new = na.omit(mydata\_new)  
  
#Seggragate data into two datasets: Train and Test  
train\_new = mydata\_new[1:count,]  
test\_new = mydata\_new[(count+1):(count\*2),]  
  
#Remove labels (is\_duplicate) from Test dataset  
test\_new <- test\_new[-c(1)]

# Step 3: Training Logistic Regression model on data

# fit the logistic regression model, with all predictor variables  
  
logit\_model <- glm(is\_duplicate ~.,family=binomial(link='logit'),data=train\_new)  
logit\_model

##   
## Call: glm(formula = is\_duplicate ~ ., family = binomial(link = "logit"),   
## data = train\_new)  
##   
## Coefficients:  
## (Intercept) match\_count p11 p12 p13   
## -2.718501 3.625967 0.019715 0.011898 0.358236   
## p14 p15 p21 p22 p23   
## -14.153379 -13.341115 0.126256 0.086699 -0.180648   
## p24 p25 n11 n12 n13   
## -13.331292 -13.786351 0.006426 -0.049550 -0.429531   
## n14 n21 n22 n23 n24   
## -14.713803 0.193525 0.289593 -0.444689 -14.200802   
##   
## Degrees of Freedom: 1499 Total (i.e. Null); 1480 Residual  
## Null Deviance: 1986   
## Residual Deviance: 1691 AIC: 1731

summary(logit\_model)

##   
## Call:  
## glm(formula = is\_duplicate ~ ., family = binomial(link = "logit"),   
## data = train\_new)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7347 -0.9179 -0.5115 1.0500 2.3590   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.719e+00 1.858e-01 -14.633 <2e-16 \*\*\*  
## match\_count 3.626e+00 2.504e-01 14.480 <2e-16 \*\*\*  
## p11 1.971e-02 1.527e-01 0.129 0.897   
## p12 1.190e-02 2.482e-01 0.048 0.962   
## p13 3.582e-01 5.607e-01 0.639 0.523   
## p14 -1.415e+01 4.912e+02 -0.029 0.977   
## p15 -1.334e+01 9.308e+02 -0.014 0.989   
## p21 1.263e-01 1.502e-01 0.840 0.401   
## p22 8.670e-02 2.566e-01 0.338 0.735   
## p23 -1.806e-01 4.855e-01 -0.372 0.710   
## p24 -1.333e+01 4.348e+02 -0.031 0.976   
## p25 -1.379e+01 9.308e+02 -0.015 0.988   
## n11 6.426e-03 1.973e-01 0.033 0.974   
## n12 -4.955e-02 4.113e-01 -0.120 0.904   
## n13 -4.295e-01 1.455e+00 -0.295 0.768   
## n14 -1.471e+01 1.455e+03 -0.010 0.992   
## n21 1.935e-01 1.979e-01 0.978 0.328   
## n22 2.896e-01 4.002e-01 0.724 0.469   
## n23 -4.447e-01 8.850e-01 -0.502 0.615   
## n24 -1.420e+01 1.455e+03 -0.010 0.992   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1986.2 on 1499 degrees of freedom  
## Residual deviance: 1691.4 on 1480 degrees of freedom  
## AIC: 1731.4  
##   
## Number of Fisher Scoring iterations: 14

anova(logit\_model, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: is\_duplicate  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 1499 1986.2   
## match\_count 1 279.538 1498 1706.7 <2e-16 \*\*\*  
## p1 5 7.892 1493 1698.8 0.1623   
## p2 5 3.841 1488 1694.9 0.5725   
## n1 4 1.322 1484 1693.6 0.8576   
## n2 4 2.200 1480 1691.4 0.6989   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Step 4: Evaluating model performance

# Checking Accuracy  
  
fitted.results <- predict(logit\_model,newdata=train\_new,type='response')  
fitted.results <- ifelse(fitted.results > 0.5,1,0)  
  
misClasificError <- mean(fitted.results != train$is\_duplicate)  
print(paste('Accuracy',1-misClasificError))

## [1] "Accuracy 0.674"

# Accuracy = 0.674

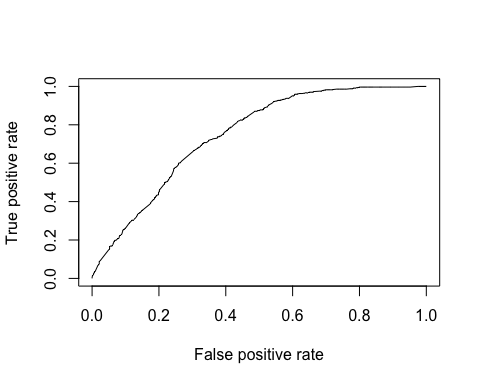
# ROC  
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

p <- predict(logit\_model, newdata=train\_new, type="response")  
pr <- prediction(p, train\_new$is\_duplicate)  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf)



auc <- performance(pr, measure = "auc")  
auc <- auc@y.values[[1]]  
auc

## [1] 0.7470497

# Accuracy = 0.747

# Step 5: Improving the Model: Training Random Forest Model on data

#Building the model  
  
library(randomForest)

## Warning: package 'randomForest' was built under R version 3.4.4

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

model <- randomForest(is\_duplicate ~ ., data = train\_new)  
model

##   
## Call:  
## randomForest(formula = is\_duplicate ~ ., data = train\_new)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 33.47%  
## Confusion matrix:  
## 0 1 class.error  
## 0 702 234 0.2500000  
## 1 268 296 0.4751773

library(party)

## Warning: package 'party' was built under R version 3.4.4

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

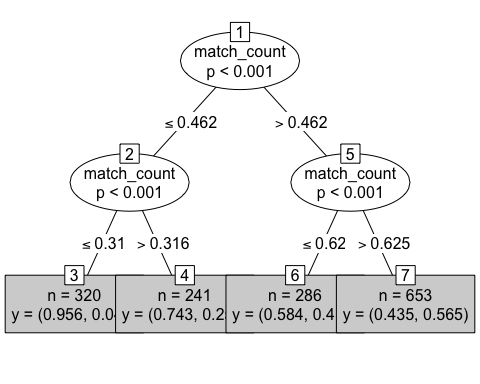
## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

x <- ctree(is\_duplicate ~ ., data = train\_new)  
plot(x, type="simple")



#Making predictions on training data  
rf <- predict(model, newdata = train\_new, type = "response")  
  
table(rf,train\_new$is\_duplicate)

##   
## rf 0 1  
## 0 765 185  
## 1 171 379

#Accuracy  
  
sum(rf==train\_new$is\_duplicate) / nrow(train\_new)

## [1] 0.7626667

# Accuracy = 0.796

# Making predictions on testing data

rf <- predict(model, newdata = test\_new, type = "response")  
  
table(rf,test$is\_duplicate)

##   
## rf 0 1  
## 0 698 279  
## 1 230 293

# Evaluating model performance

#Accuracy  
  
sum(rf==test$is\_duplicate) / nrow(test\_new)

## [1] 0.6606667

# Accuracy = 0.661

# Evaluating Random forest model performance

# Comparing an auto-tuned random forest to the best auto-tuned boosted C5.0 model

library(caret)

## Warning: package 'caret' was built under R version 3.4.4

## Loading required package: lattice

library(lattice)  
library(ggplot2)  
ctrl <- trainControl(method = "repeatedcv",number = 10, repeats = 10)  
grid\_rf <- expand.grid(.mtry = c(1, 2, 3, 4))  
  
set.seed(300)  
m\_rf <- train(is\_duplicate ~ ., data = train\_new, method = "rf",   
 metric = "Kappa", trControl = ctrl, tuneGrid = grid\_rf)  
  
grid\_c50 <- expand.grid(.model = "tree",  
 .trials = c(10, 20, 30, 40),  
 .winnow = "FALSE")  
set.seed(300)  
m\_c50 <- train(is\_duplicate ~ ., data = train\_new, method = "C5.0",  
 metric = "Kappa", trControl = ctrl,  
 tuneGrid = grid\_c50)

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials  
  
## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials  
  
## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials  
  
## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials  
  
## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials  
  
## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials  
  
## Warning: 'trials' should be <= 5 for this object. Predictions generated  
## using 5 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials  
  
## Warning: 'trials' should be <= 3 for this object. Predictions generated  
## using 3 trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials  
  
## Warning: 'trials' should be <= 1 for this object. Predictions generated  
## using 1 trials

## Warning in Ops.factor(x$winnow): '!' not meaningful for factors

m\_rf

## Random Forest   
##   
## 1500 samples  
## 5 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 1349, 1351, 1349, 1350, 1351, 1351, ...   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa   
## 1 0.6240025 0.0000000000  
## 2 0.6242012 0.0007336575  
## 3 0.6565988 0.2035944282  
## 4 0.6684745 0.2743775213  
##   
## Kappa was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 4.

m\_c50

## C5.0   
##   
## 1500 samples  
## 5 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 1349, 1351, 1349, 1350, 1351, 1351, ...   
## Resampling results across tuning parameters:  
##   
## trials Accuracy Kappa   
## 10 0.6746262 0.3263416  
## 20 0.6746262 0.3263416  
## 30 0.6746262 0.3263416  
## 40 0.6746262 0.3263416  
##   
## Tuning parameter 'model' was held constant at a value of tree  
##   
## Tuning parameter 'winnow' was held constant at a value of FALSE  
## Kappa was used to select the optimal model using the largest value.  
## The final values used for the model were trials = 10, model = tree  
## and winnow = FALSE.

# Plot: Random Forest model performance (Accuracy) across various sample sizes for both - Train and Test data

sample\_sizes <- c(1000, 2000, 3000, 4000, 5000)  
train\_accuracy <- c(.794, .769, .796, .759, .746)  
test\_accuracy <- c(.634, .554, .657, .668, .668)  
  
resampling\_accuracy <- data.frame(samplesize=rep(sample\_sizes, 2),  
 results=c(train\_accuracy, test\_accuracy), group = rep(c("train","test"),each = 5))  
ggplot(data=resampling\_accuracy, aes(x=samplesize, y=results, group = group)) +  
 geom\_line(aes(color=group))+  
 geom\_point(aes(color=group))+  
 labs(title="Resampling Accuracy results for train and test data",x="Sample Size", y = "Accuracy")

