Titanic

Vaibhav Laturkar

13 March 2016

train <- read.csv(choose.files())  
  
head(train)

## PassengerId Survived Pclass  
## 1 1 0 3  
## 2 2 1 1  
## 3 3 1 3  
## 4 4 1 1  
## 5 5 0 3  
## 6 6 0 3  
## Name Sex Age SibSp  
## 1 Braund, Mr. Owen Harris male 22 1  
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38 1  
## 3 Heikkinen, Miss. Laina female 26 0  
## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35 1  
## 5 Allen, Mr. William Henry male 35 0  
## 6 Moran, Mr. James male NA 0  
## Parch Ticket Fare Cabin Embarked  
## 1 0 A/5 21171 7.2500 S  
## 2 0 PC 17599 71.2833 C85 C  
## 3 0 STON/O2. 3101282 7.9250 S  
## 4 0 113803 53.1000 C123 S  
## 5 0 373450 8.0500 S  
## 6 0 330877 8.4583 Q

## How many Male and Female   
summary(train$Sex)

## female male   
## 314 577

## How many Male and Female survived   
a <- table(train$Survived,train$Sex)  
a

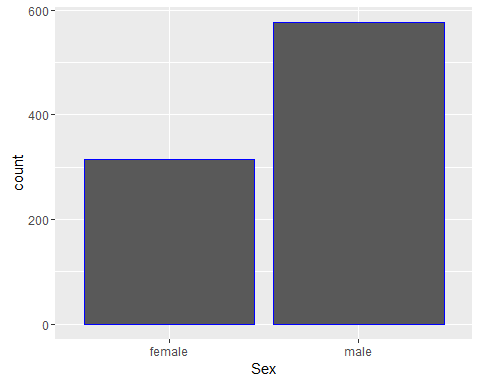
##   
## female male  
## 0 81 468  
## 1 233 109

## We want Proportion of Survived  
prop.table(a)

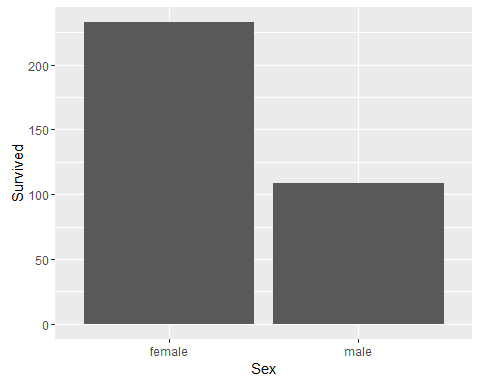
##   
## female male  
## 0 0.09090909 0.52525253  
## 1 0.26150393 0.12233446

After seeing the proportion table **9% female died and 26 % survived and male 52.52 % died and only 12 % survived**

library(ggplot2)  
  
ggplot(train,aes(x=Sex)) +geom\_bar(col="blue")



ggplot(train,aes(x=Sex,y=Survived)) +geom\_bar(stat = "identity")



## There were more Male compared to Female

**We will upload our score on Kaggle on gender base**

## read the test file   
test <- read.csv(choose.files())  
  
## We will add a column to test file name as Survived and assgin to 0  
  
test$Survived <- 0  
  
## Now we will now write our first Model  
  
test$Survived[test$Sex == "female"] <- 1  
  
head(test)

## PassengerId Pclass Name Sex  
## 1 892 3 Kelly, Mr. James male  
## 2 893 3 Wilkes, Mrs. James (Ellen Needs) female  
## 3 894 2 Myles, Mr. Thomas Francis male  
## 4 895 3 Wirz, Mr. Albert male  
## 5 896 3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female  
## 6 897 3 Svensson, Mr. Johan Cervin male  
## Age SibSp Parch Ticket Fare Cabin Embarked Survived  
## 1 34.5 0 0 330911 7.8292 Q 0  
## 2 47.0 1 0 363272 7.0000 S 1  
## 3 62.0 0 0 240276 9.6875 Q 0  
## 4 27.0 0 0 315154 8.6625 S 0  
## 5 22.0 1 1 3101298 12.2875 S 1  
## 6 14.0 0 0 7538 9.2250 S 0

Know we will upload our score

write.csv(test,file = "myfirstgendermodel.csv")

\*\*\* Score = 0.7655 \*\*\*

\*\* Now we will do my Pclass \*\*

## We will check with Class   
b <- table(train$Survived,train$Pclass)  
b

##   
## 1 2 3  
## 0 80 97 372  
## 1 136 87 119

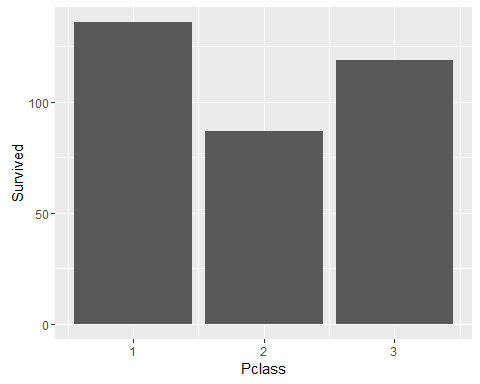
## we want proportion  
prop.table(b)

##   
## 1 2 3  
## 0 0.08978676 0.10886644 0.41750842  
## 1 0.15263749 0.09764310 0.13355780

from this we came to know that more person died from 3rd class

Now we will do **Visualization**  of the data

ggplot(train,aes(x=Pclass,y=Survived)) + geom\_bar(stat = "identity")

 We will write our model for Pclass

View(test)  
  
test$Survived <- 1  
  
test$Survived <- ifelse (test$Sex == "male" & test$Pclass==3,0,1)

Know we will upload our score

write.csv(test,file = "myfirstgenderclassmodel.csv")

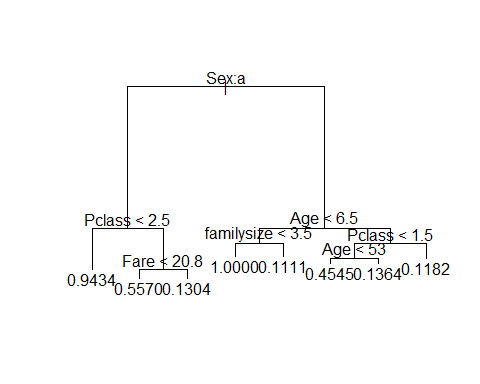
**Score = 0.5980 score reduced**

Now we will try **Decision tree Model**

train$familysize <- train$SibSp+train$Parch  
  
library(tree)  
model <- tree(Survived ~ Sex+Pclass+Fare+Age+Embarked+familysize,train)  
  
  
summary(model)

##   
## Regression tree:  
## tree(formula = Survived ~ Sex + Pclass + Fare + Age + Embarked +   
## familysize, data = train)  
## Variables actually used in tree construction:  
## [1] "Sex" "Pclass" "Fare" "Age" "familysize"  
## Number of terminal nodes: 8   
## Residual mean deviance: 0.124 = 87.55 / 706   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.9434 -0.1182 -0.1182 0.0000 0.0566 0.8889

plot(model)  
text(model)



test$familysize <- test$SibSp+test$Parch  
  
test$Survived <- predict(model,test)  
test$Survived1 <- 1  
test$Survived1[test$Survived < 0.5] <- 0  
  
View(test)  
  
write.csv(test,file = "treemodel.csv")

**Score is 0.7799**

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.2.4

## randomForest 4.6-12

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

##   
## Attaching package: 'randomForest'

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

##we wil impute missing values by decision tree  
  
sum(is.na(train$Age))

## [1] 177

Agefit <- tree(Age ~ Pclass + Sex + familysize + Fare + Embarked + SibSp + Parch,  
 data=train[!is.na(train$Age),], method="anova")  
train$Age[is.na(train$Age)] <- predict(Agefit, train[is.na(train$Age),])

## we have imputed age by decision tree

## we will check the NA in AGE for test

sum(is.na(test$Age))

## [1] 86

Agefit <- tree(Age ~ Pclass + Sex + SibSp + Parch + Fare + Embarked + familysize,  
 data=test[!is.na(test$Age),], method="anova")  
test$Age[is.na(test$Age)] <- predict(Agefit, test[is.na(test$Age),])

we will check missing value in train data and test data

summary(train)

## PassengerId Survived Pclass   
## Min. : 1.0 Min. :0.0000 Min. :1.000   
## 1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000   
## Median :446.0 Median :0.0000 Median :3.000   
## Mean :446.0 Mean :0.3838 Mean :2.309   
## 3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :891.0 Max. :1.0000 Max. :3.000   
##   
## Name Sex Age   
## Abbing, Mr. Anthony : 1 female:314 Min. : 0.42   
## Abbott, Mr. Rossmore Edward : 1 male :577 1st Qu.:21.54   
## Abbott, Mrs. Stanton (Rosa Hunt) : 1 Median :27.84   
## Abelson, Mr. Samuel : 1 Mean :29.53   
## Abelson, Mrs. Samuel (Hannah Wizosky): 1 3rd Qu.:36.19   
## Adahl, Mr. Mauritz Nils Martin : 1 Max. :80.00   
## (Other) :885   
## SibSp Parch Ticket Fare   
## Min. :0.000 Min. :0.0000 1601 : 7 Min. : 0.00   
## 1st Qu.:0.000 1st Qu.:0.0000 347082 : 7 1st Qu.: 7.91   
## Median :0.000 Median :0.0000 CA. 2343: 7 Median : 14.45   
## Mean :0.523 Mean :0.3816 3101295 : 6 Mean : 32.20   
## 3rd Qu.:1.000 3rd Qu.:0.0000 347088 : 6 3rd Qu.: 31.00   
## Max. :8.000 Max. :6.0000 CA 2144 : 6 Max. :512.33   
## (Other) :852   
## Cabin Embarked familysize   
## :687 : 2 Min. : 0.0000   
## B96 B98 : 4 C:168 1st Qu.: 0.0000   
## C23 C25 C27: 4 Q: 77 Median : 0.0000   
## G6 : 4 S:644 Mean : 0.9046   
## C22 C26 : 3 3rd Qu.: 1.0000   
## D : 3 Max. :10.0000   
## (Other) :186

summary(test)

## PassengerId Pclass   
## Min. : 892.0 Min. :1.000   
## 1st Qu.: 996.2 1st Qu.:1.000   
## Median :1100.5 Median :3.000   
## Mean :1100.5 Mean :2.266   
## 3rd Qu.:1204.8 3rd Qu.:3.000   
## Max. :1309.0 Max. :3.000   
##   
## Name Sex   
## Abbott, Master. Eugene Joseph : 1 female:152   
## Abelseth, Miss. Karen Marie : 1 male :266   
## Abelseth, Mr. Olaus Jorgensen : 1   
## Abrahamsson, Mr. Abraham August Johannes : 1   
## Abrahim, Mrs. Joseph (Sophie Halaut Easu): 1   
## Aks, Master. Philip Frank : 1   
## (Other) :412   
## Age SibSp Parch Ticket   
## Min. : 0.17 Min. :0.0000 Min. :0.0000 PC 17608: 5   
## 1st Qu.:22.00 1st Qu.:0.0000 1st Qu.:0.0000 113503 : 4   
## Median :26.00 Median :0.0000 Median :0.0000 CA. 2343: 4   
## Mean :29.75 Mean :0.4474 Mean :0.3923 16966 : 3   
## 3rd Qu.:36.38 3rd Qu.:1.0000 3rd Qu.:0.0000 220845 : 3   
## Max. :76.00 Max. :8.0000 Max. :9.0000 347077 : 3   
## (Other) :396   
## Fare Cabin Embarked Survived   
## Min. : 0.000 :327 C:102 Min. :0.1111   
## 1st Qu.: 7.896 B57 B59 B63 B66: 3 Q: 46 1st Qu.:0.1182   
## Median : 14.454 A34 : 2 S:270 Median :0.2053   
## Mean : 35.627 B45 : 2 Mean :0.3999   
## 3rd Qu.: 31.500 C101 : 2 3rd Qu.:0.5570   
## Max. :512.329 C116 : 2 Max. :1.0000   
## NA's :1 (Other) : 80   
## familysize Survived1   
## Min. : 0.0000 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.0000   
## Median : 0.0000 Median :0.0000   
## Mean : 0.8397 Mean :0.3589   
## 3rd Qu.: 1.0000 3rd Qu.:1.0000   
## Max. :10.0000 Max. :1.0000   
##

#there is data missing fare will impute it with median  
  
which(is.na(test$Fare))

## [1] 153

test$Fare[153] <- median(test$Fare, na.rm=TRUE)  
  
#there is missing value in embarked   
  
which(train$Embarked == '')

## [1] 62 830

train$Embarked[c(62,830)] <- "S"  
  
#check the stcture of data  
  
str(train)

## 'data.frame': 891 obs. of 13 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...  
## $ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...  
## $ familysize : int 1 1 0 1 0 0 0 4 2 1 ...

#set embarked as factor of missing value  
  
train$Embarked <- factor(train$Embarked)

## Now we will fit a model

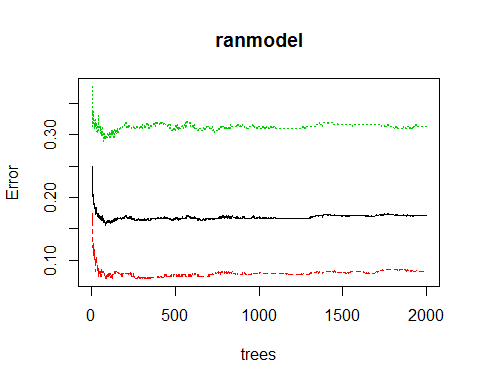
set.seed(1)  
ranmodel <- randomForest(as.factor(Survived) ~ Pclass+Sex+Age+SibSp+Parch+Fare+Embarked+ familysize,data=train, importance=TRUE,ntree=2000)  
  
summary(ranmodel)

## Length Class Mode   
## call 5 -none- call   
## type 1 -none- character  
## predicted 891 factor numeric   
## err.rate 6000 -none- numeric   
## confusion 6 -none- numeric   
## votes 1782 matrix numeric   
## oob.times 891 -none- numeric   
## classes 2 -none- character  
## importance 32 -none- numeric   
## importanceSD 24 -none- numeric   
## localImportance 0 -none- NULL   
## proximity 0 -none- NULL   
## ntree 1 -none- numeric   
## mtry 1 -none- numeric   
## forest 14 -none- list   
## y 891 factor numeric   
## test 0 -none- NULL   
## inbag 0 -none- NULL   
## terms 3 terms call

ranmodel

##   
## Call:  
## randomForest(formula = as.factor(Survived) ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked + familysize, data = train, importance = TRUE, ntree = 2000)   
## Type of random forest: classification  
## Number of trees: 2000  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 17.06%  
## Confusion matrix:  
## 0 1 class.error  
## 0 504 45 0.08196721  
## 1 107 235 0.31286550

plot(ranmodel)



test$Survived6 <- predict(ranmodel,test)  
View(test)  
  
write.csv(test,file = "Ranmodel.csv")

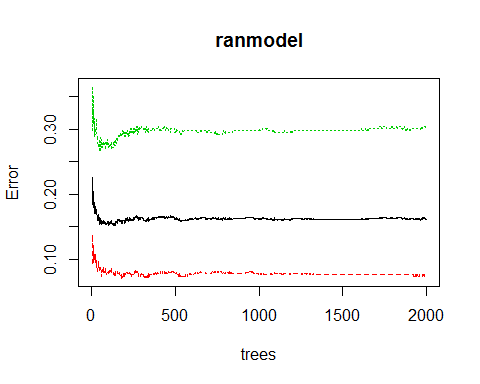
# Score is 0.78469

# We will remove Sibsp and Parch because we have sum them into one column

set.seed(1)  
ranmodel <- randomForest(as.factor(Survived) ~ Sex+Age+familysize+Pclass+Fare+Embarked,data=train, importance=TRUE,ntree=2000)  
  
ranmodel

##   
## Call:  
## randomForest(formula = as.factor(Survived) ~ Sex + Age + familysize + Pclass + Fare + Embarked, data = train, importance = TRUE, ntree = 2000)   
## Type of random forest: classification  
## Number of trees: 2000  
## No. of variables tried at each split: 2  
##   
## OOB estimate of error rate: 16.16%  
## Confusion matrix:  
## 0 1 class.error  
## 0 508 41 0.07468124  
## 1 103 239 0.30116959

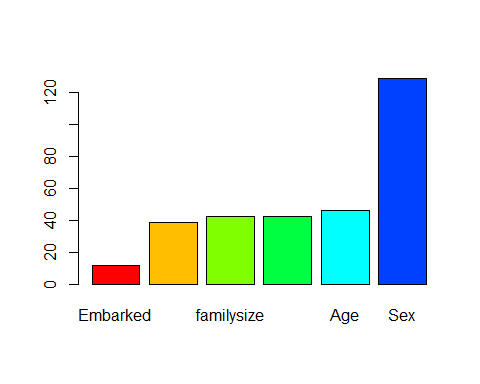
plot(ranmodel)



imp <- importance(ranmodel)[,1]  
imp

## Sex Age familysize Pclass Fare Embarked   
## 128.36570 46.01225 42.11995 42.25434 38.56442 11.54769

barplot(sort(imp),col=rainbow(8))



test$Survived7 <- predict(ranmodel,test)  
View(test)  
  
write.csv(test,file = "Ranmodel1.csv")

# Score= 0.78947

* We have not taken other variables like name and cabin variable we will check if there is any improvement or not
* We can do Feature Engineering