Project 3

Owen McGrattan

5/1/2017

Question 1

Read the train data into R

```r
library(readr)
train <- read_csv("~/stat28/projects/data/train_titanic.csv")
```

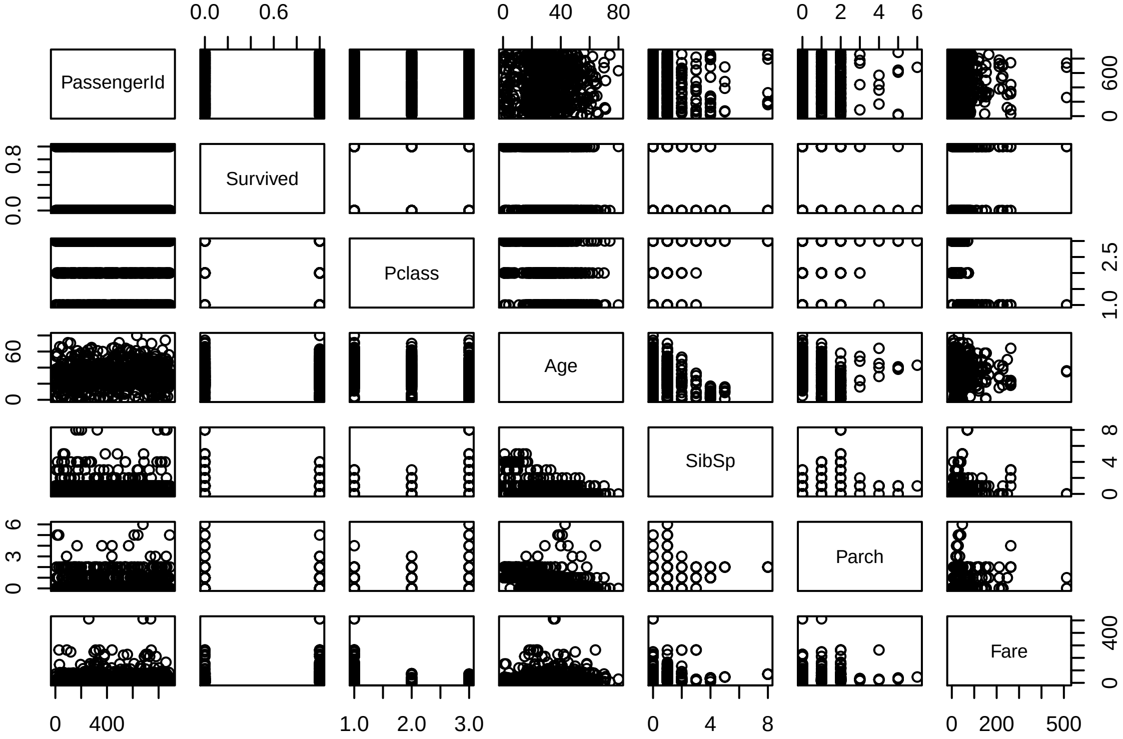
```
## Parsed with column specification:
## cols(
##   PassengerId = col_integer(),
##   Survived = col_integer(),
##   Pclass = col_integer(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_double(),
##   SibSp = col_integer(),
##   Parch = col_integer(),
##   Ticket = col_character(),
##   Fare = col_double(),
##   Cabin = col_character(),
##   Embarked = col_character()
## )
```

Question 2

```r
summary(train)
```

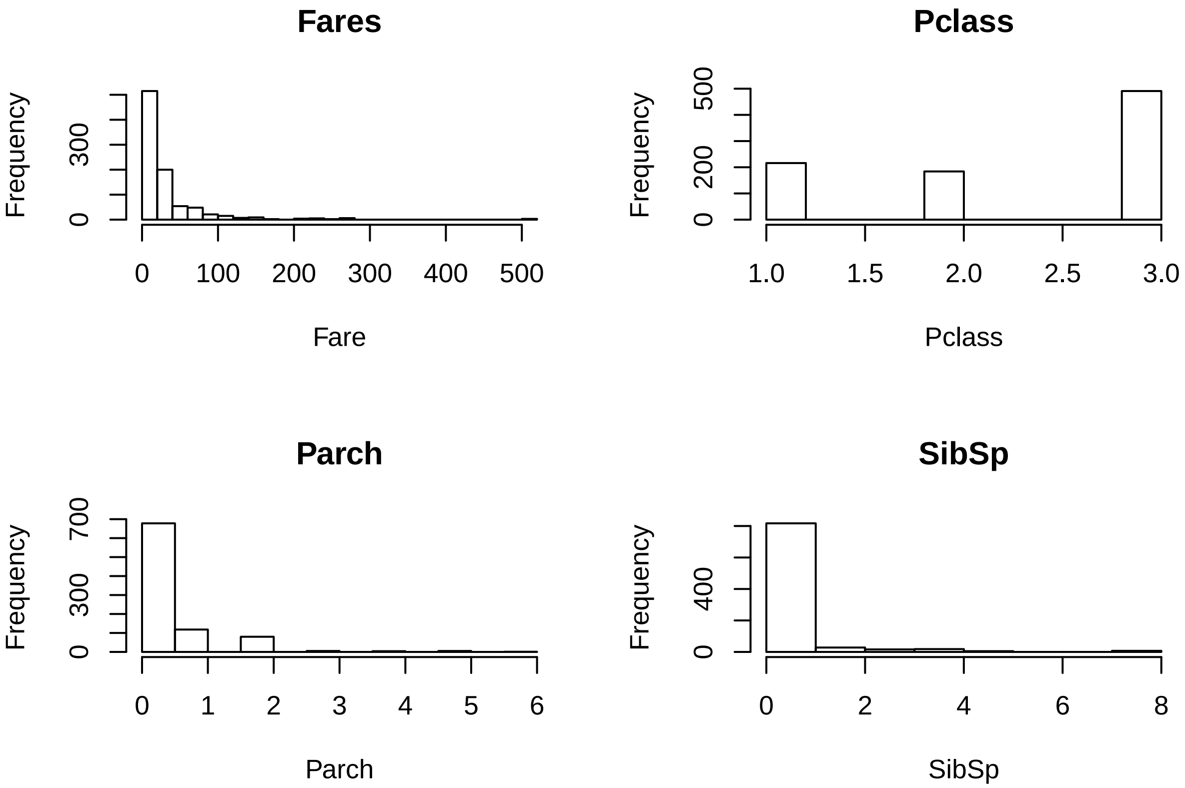
```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                          Max.   :512.33
##
```

```r
pairs(train[,-c(9,4,5,11,12)])
```



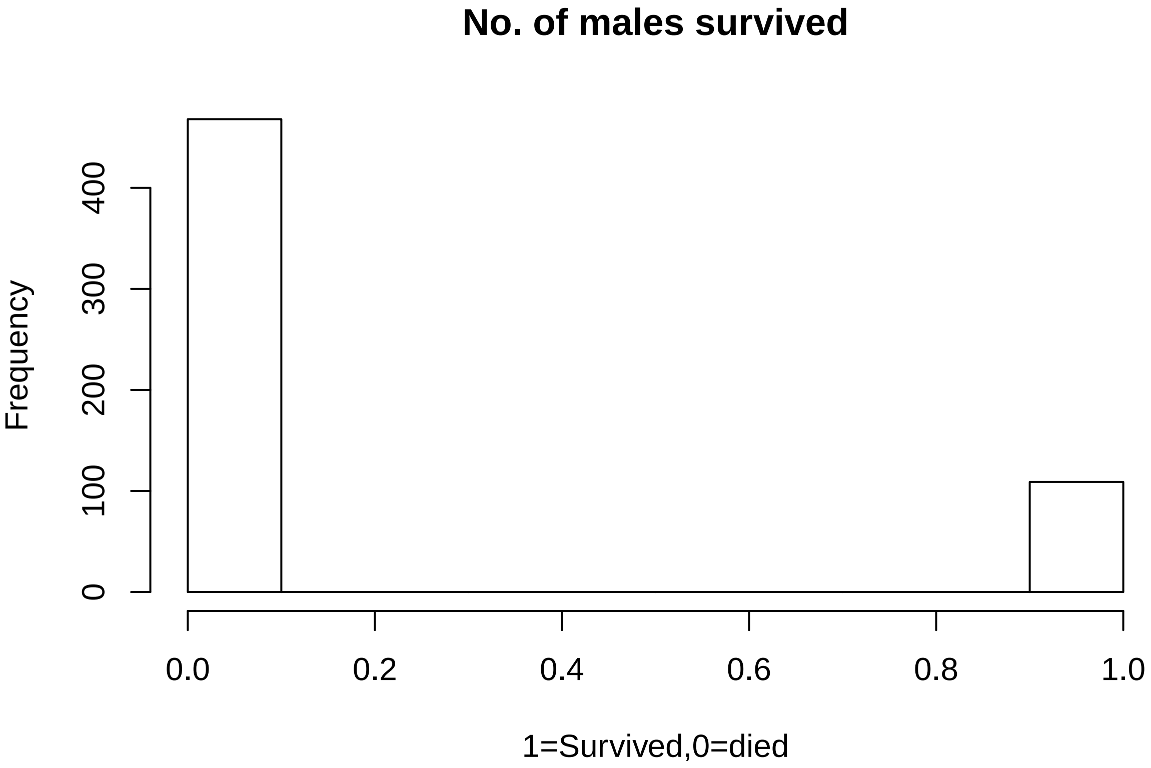
It's a little difficult to pull anything from this pairs plot but the differences between the male and female groups should be analyzed as well as the differences in whether or not the cost of the ticket had anything to do with survival as well.

```r
par(mfrow=c(2,2))
hist(train$Fare,xlab="Fare",main="Fares",breaks=25)
hist(train$Pclass,xlab="Pclass",main="Pclass")
hist(train$Parch,xlab="Parch",main="Parch")
hist(train$SibSp,xlab="SibSp",main="SibSp")
```

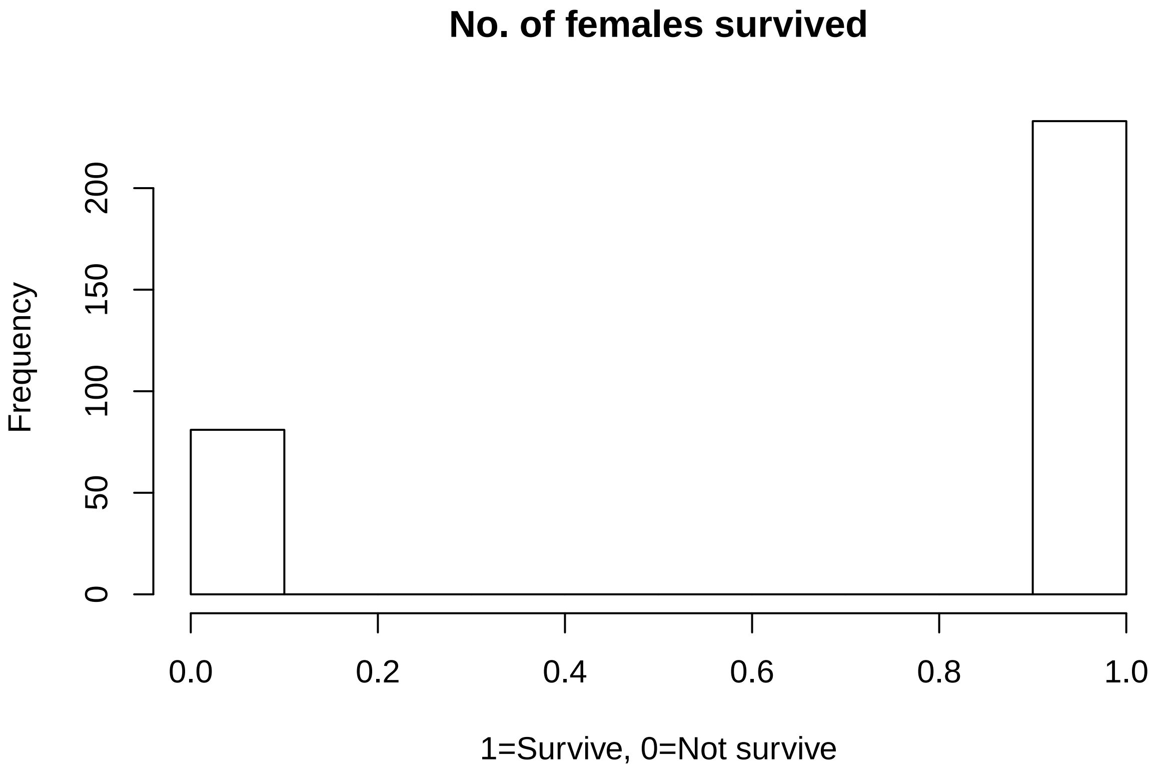


Based on the histograms of the explanatory variables, there appears to be skewness for Fares, Pclass, Parch, and SibSp. We should consider taking

```
males<-train[train$Sex=="male",]
hist(males$Survived,main="No. of males survived",xlab="1=Survived,0=died")
```

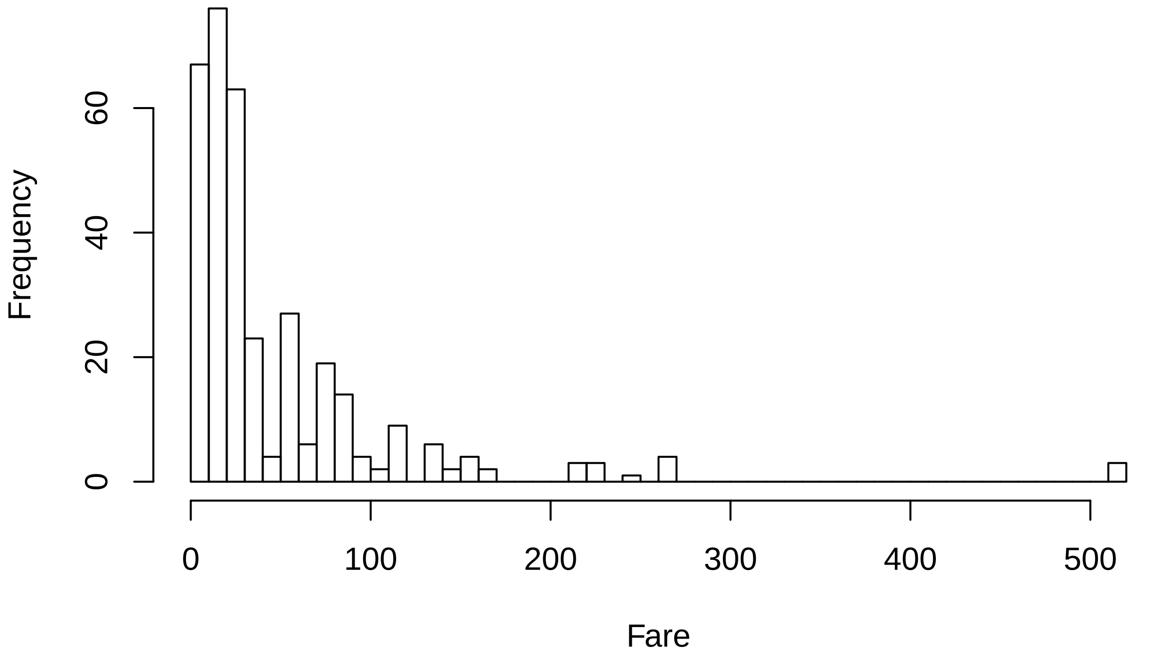


```r
females<-train[train$Sex=="female",]
hist(females$Survived,main="No. of females survived",xlab="1=Survive, 0=Not survive")
```

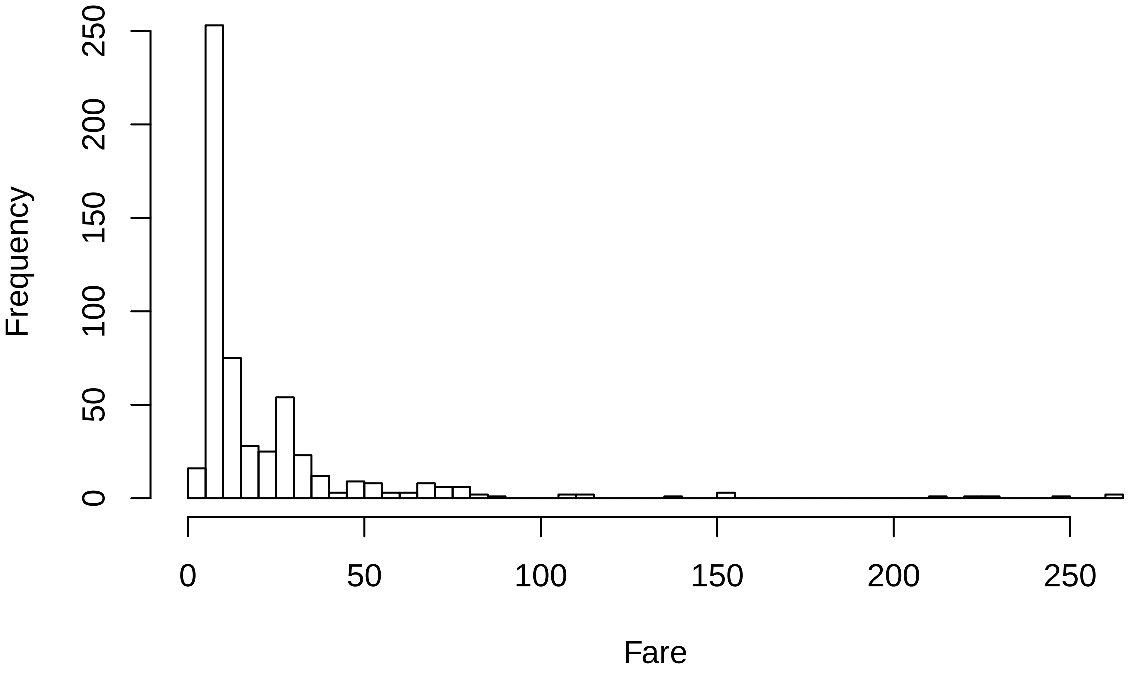


The people on the Titanic very much tried to save the women as evidenced by the few number of men who survived and the large number of women who survived.

```r
survived<-train[train$Survived=="1",]
hist(survived$Fare,breaks=50,xlab="Fare",main="")
```



```
not_survived<-train[train$Survived=="0",]
hist(not_survived$Fare,breaks=50,xlab="Fare",main="")
```



```
median(survived$Fare)
```

```
## [1] 26
```

```
median(not_survived$Fare)
```

```
## [1] 10.5
```

Looking at the distributions of both the fares for those who survived versus those who didn't shows that those who survived generally paid higher fares with the median fare ticket of the surviving group coming in at 26 with the median fare ticket of the non survivors at 10.5.

Question 3

```
#Fit a logistic regression equation
fitted<-glm(Survived~as.factor(Pclass)+as.factor(Sex)+Age+SibSp+Parch+Fare+as.factor(Embarked)+as.facto
summary(fitted)
```

```
##
## Call:
## glm(formula = Survived ~ as.factor(Pclass) + as.factor(Sex) +
##     Age + SibSp + Parch + Fare + as.factor(Embarked) + as.factor(Embarked):as.factor(Pclass) +
##     Age:as.factor(Pclass) + as.factor(Sex):as.factor(Pclass) +
##     as.factor(Pclass):Fare + Age:Fare + Parch:SibSp + as.factor(Pclass):SibSp +
##     as.factor(Pclass):Parch + Age:as.factor(Embarked), family = "binomial",
##     data = na.omit(train[, -11]))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7014  -0.6433  -0.3660   0.3662   3.0590
##
## Coefficients:
```

```
##                                           Estimate Std. Error z value
## (Intercept)                              6.3098052  1.4119224   4.469
## as.factor(Pclass)2                      -1.0027249  1.8782476  -0.534
## as.factor(Pclass)3                      -4.0353571  1.3276173  -3.040
## as.factor(Sex)male                      -3.6000994  0.6452697  -5.579
## Age                                     -0.0777492  0.0293930  -2.645
## SibSp                                    0.3153884  0.3819422   0.826
## Parch                                   -0.4125748  0.3421451  -1.206
## Fare                                    -0.0108607  0.0087497  -1.241
## as.factor(Embarked)Q                    -1.1649455  3.4286838  -0.340
## as.factor(Embarked)S                    -1.2860611  1.0406613  -1.236
## as.factor(Pclass)2:as.factor(Embarked)Q  4.3555511  6.2389663   0.698
## as.factor(Pclass)3:as.factor(Embarked)Q  0.9799635  2.7001097   0.363
## as.factor(Pclass)2:as.factor(Embarked)S  1.1660711  1.2075173   0.966
## as.factor(Pclass)3:as.factor(Embarked)S  0.0251254  0.7534836   0.033
## as.factor(Pclass)2:Age                  -0.0298523  0.0321328  -0.929
## as.factor(Pclass)3:Age                  -0.0013565  0.0269978  -0.050
## as.factor(Pclass)2:as.factor(Sex)male   -1.1155413  0.9156382  -1.218
## as.factor(Pclass)3:as.factor(Sex)male    2.0477832  0.7153558   2.863
## as.factor(Pclass)2:Fare                 -0.0243752  0.0312895  -0.779
## as.factor(Pclass)3:Fare                  0.0258043  0.0197148   1.309
## Age:Fare                                 0.0003543  0.0002337   1.516
## SibSp:Parch                             -0.0684397  0.1671300  -0.409
## as.factor(Pclass)2:SibSp                -0.2447923  0.6509883  -0.376
## as.factor(Pclass)3:SibSp                -0.8080794  0.4210054  -1.919
## as.factor(Pclass)2:Parch                 1.6197739  0.6163762   2.628
## as.factor(Pclass)3:Parch                 0.3585276  0.3696060   0.970
## Age:as.factor(Embarked)Q                -0.0279169  0.0653288  -0.427
## Age:as.factor(Embarked)S                 0.0269324  0.0238074   1.131
##                                         Pr(>|z|)
## (Intercept)                             7.86e-06 ***
## as.factor(Pclass)2                       0.59344
## as.factor(Pclass)3                       0.00237 **
## as.factor(Sex)male                      2.42e-08 ***
## Age                                      0.00817 **
## SibSp                                    0.40895
## Parch                                    0.22788
## Fare                                     0.21451
## as.factor(Embarked)Q                     0.73403
## as.factor(Embarked)S                     0.21653
## as.factor(Pclass)2:as.factor(Embarked)Q  0.48510
## as.factor(Pclass)3:as.factor(Embarked)Q  0.71665
## as.factor(Pclass)2:as.factor(Embarked)S  0.33421
## as.factor(Pclass)3:as.factor(Embarked)S  0.97340
## as.factor(Pclass)2:Age                   0.35287
## as.factor(Pclass)3:Age                   0.95993
## as.factor(Pclass)2:as.factor(Sex)male    0.22310
## as.factor(Pclass)3:as.factor(Sex)male    0.00420 **
## as.factor(Pclass)2:Fare                  0.43597
## as.factor(Pclass)3:Fare                  0.19057
## Age:Fare                                 0.12948
## SibSp:Parch                              0.68217
## as.factor(Pclass)2:SibSp                 0.70689
## as.factor(Pclass)3:SibSp                 0.05493 .
## as.factor(Pclass)2:Parch                 0.00859 **
## as.factor(Pclass)3:Parch                 0.33203
## Age:as.factor(Embarked)Q                 0.66914
## Age:as.factor(Embarked)S                 0.25795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 573.19  on 684  degrees of freedom
## AIC: 629.19
##
## Number of Fisher Scoring iterations: 6
```

```r
#Variable selection for fitted
step(fitted,direction="both")
```

```
## Start:  AIC=629.19
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Embarked) + as.factor(Embarked):as.factor(Pclass) +
##     Age:as.factor(Pclass) + as.factor(Sex):as.factor(Pclass) +
##     as.factor(Pclass):Fare + Age:Fare + Parch:SibSp + as.factor(Pclass):SibSp +
##     as.factor(Pclass):Parch + Age:as.factor(Embarked)
##
##                                          Df Deviance    AIC
## - as.factor(Pclass):as.factor(Embarked)  4   574.69 622.69
## - as.factor(Pclass):Age                  2   574.44 626.44
## - Age:as.factor(Embarked)                2   575.24 627.24
## - SibSp:Parch                            1   573.36 627.36
## - as.factor(Pclass):Fare                 2   575.52 627.52
## <none>                                       573.19 629.19
## - as.factor(Pclass):SibSp                2   577.30 629.30
## - Age:Fare                               1   575.40 629.40
## - as.factor(Pclass):Parch                2   581.20 633.20
## - as.factor(Pclass):as.factor(Sex)       2   602.14 654.14
##
## Step:  AIC=622.69
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Embarked) + as.factor(Pclass):Age +
##     as.factor(Pclass):as.factor(Sex) + as.factor(Pclass):Fare +
##     Age:Fare + SibSp:Parch + as.factor(Pclass):SibSp + as.factor(Pclass):Parch +
##     Age:as.factor(Embarked)
##
##                                          Df Deviance    AIC
## - as.factor(Pclass):Age                  2   575.86 619.86
## - SibSp:Parch                            1   574.87 620.87
## - as.factor(Pclass):Fare                 2   577.22 621.22
## - Age:as.factor(Embarked)                2   577.53 621.53
## - as.factor(Pclass):SibSp                2   578.62 622.62
## <none>                                       574.69 622.69
## - Age:Fare                               1   576.90 622.90
## - as.factor(Pclass):Parch                2   582.24 626.24
## + as.factor(Pclass):as.factor(Embarked)  4   573.19 629.19
## - as.factor(Pclass):as.factor(Sex)       2   604.36 648.36
```

```
##
## Step:  AIC=619.86
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Embarked) + as.factor(Pclass):as.factor(Sex) +
##     as.factor(Pclass):Fare + Age:Fare + SibSp:Parch + as.factor(Pclass):SibSp +
##     as.factor(Pclass):Parch + Age:as.factor(Embarked)
##
##                                       Df Deviance    AIC
## - SibSp:Parch                          1   576.04 618.04
## - Age:as.factor(Embarked)              2   578.22 618.22
## - as.factor(Pclass):Fare               2   578.61 618.61
## <none>                                     575.86 619.86
## - as.factor(Pclass):SibSp              2   580.53 620.53
## - Age:Fare                             1   580.22 622.22
## + as.factor(Pclass):Age                2   574.69 622.69
## - as.factor(Pclass):Parch              2   585.44 625.44
## + as.factor(Pclass):as.factor(Embarked)  4   574.44 626.44
## - as.factor(Pclass):as.factor(Sex)     2   604.39 644.39
##
## Step:  AIC=618.04
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Embarked) + as.factor(Pclass):as.factor(Sex) +
##     as.factor(Pclass):Fare + Age:Fare + as.factor(Pclass):SibSp +
##     as.factor(Pclass):Parch + Age:as.factor(Embarked)
##
##                                       Df Deviance    AIC
## - Age:as.factor(Embarked)              2   578.45 616.45
## - as.factor(Pclass):Fare               2   578.74 616.74
## <none>                                     576.04 618.04
## - as.factor(Pclass):SibSp              2   581.79 619.79
## + SibSp:Parch                          1   575.86 619.86
## - Age:Fare                             1   580.67 620.67
## + as.factor(Pclass):Age                2   574.87 620.87
## - as.factor(Pclass):Parch              2   585.62 623.62
## + as.factor(Pclass):as.factor(Embarked)  4   574.61 624.61
## - as.factor(Pclass):as.factor(Sex)     2   604.40 642.40
##
## Step:  AIC=616.45
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Embarked) + as.factor(Pclass):as.factor(Sex) +
##     as.factor(Pclass):Fare + Age:Fare + as.factor(Pclass):SibSp +
##     as.factor(Pclass):Parch
##
##                                       Df Deviance    AIC
## - as.factor(Embarked)                  2   580.93 614.93
## - as.factor(Pclass):Fare               2   581.07 615.07
## <none>                                     578.45 616.45
## - as.factor(Pclass):SibSp              2   583.69 617.69
## + Age:as.factor(Embarked)              2   576.04 618.04
## + SibSp:Parch                          1   578.22 618.22
## - Age:Fare                             1   582.70 618.70
## + as.factor(Pclass):Age                2   577.77 619.77
## - as.factor(Pclass):Parch              2   588.09 622.09
## + as.factor(Pclass):as.factor(Embarked)  4   576.31 622.31
## - as.factor(Pclass):as.factor(Sex)       2   606.78 640.78
##
## Step:  AIC=614.93
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Pclass):as.factor(Sex) + as.factor(Pclass):Fare +
##     Age:Fare + as.factor(Pclass):SibSp + as.factor(Pclass):Parch
##
##                                    Df Deviance    AIC
## - as.factor(Pclass):Fare            2   583.32 613.32
## <none>                                  580.93 614.93
## - as.factor(Pclass):SibSp           2   585.58 615.58
## + as.factor(Embarked)               2   578.45 616.45
## + SibSp:Parch                       1   580.51 616.51
## - Age:Fare                          1   585.51 617.51
## + as.factor(Pclass):Age             2   580.35 618.35
## - as.factor(Pclass):Parch           2   590.98 620.98
## - as.factor(Pclass):as.factor(Sex)  2   609.73 639.73
##
## Step:  AIC=613.32
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Pclass):as.factor(Sex) + Age:Fare +
##     as.factor(Pclass):SibSp + as.factor(Pclass):Parch
##
##                                    Df Deviance    AIC
## - as.factor(Pclass):SibSp           2   586.46 612.46
## <none>                                  583.32 613.32
## + as.factor(Pclass):Fare            2   580.93 614.93
## + SibSp:Parch                       1   582.99 614.99
## + as.factor(Embarked)               2   581.07 615.07
## - Age:Fare                          1   587.83 615.83
## + as.factor(Pclass):Age             2   582.62 616.62
## - as.factor(Pclass):Parch           2   592.49 618.49
## - as.factor(Pclass):as.factor(Sex)  2   612.97 638.97
##
## Step:  AIC=612.46
## Survived ~ as.factor(Pclass) + as.factor(Sex) + Age + SibSp +
##     Parch + Fare + as.factor(Pclass):as.factor(Sex) + Age:Fare +
##     as.factor(Pclass):Parch
##
##                                    Df Deviance    AIC
## <none>                                  586.46 612.46
## + SibSp:Parch                       1   585.20 613.20
## + as.factor(Pclass):SibSp           2   583.32 613.32
## + as.factor(Embarked)               2   584.65 614.65
## - Age:Fare                          1   590.69 614.69
## + as.factor(Pclass):Age             2   585.28 615.28
## + as.factor(Pclass):Fare            2   585.58 615.58
## - as.factor(Pclass):Parch           2   597.34 619.34
## - SibSp                             1   596.07 620.07
## - as.factor(Pclass):as.factor(Sex)  2   617.79 639.79

##
## Call:  glm(formula = Survived ~ as.factor(Pclass) + as.factor(Sex) +
##     Age + SibSp + Parch + Fare + as.factor(Pclass):as.factor(Sex) +
##     Age:Fare + as.factor(Pclass):Parch, family = "binomial",
##     data = na.omit(train[, -11]))
##
## Coefficients:
##                            (Intercept)
##                              5.9065744
##                     as.factor(Pclass)2
##                             -1.6213696
##                     as.factor(Pclass)3
##                             -4.3031338
##                     as.factor(Sex)male
##                             -3.7123918
##                                    Age
##                             -0.0647196
##                                  SibSp
##                             -0.3901625
##                                  Parch
##                             -0.4590655
##                                   Fare
##                             -0.0101914
## as.factor(Pclass)2:as.factor(Sex)male
##                             -0.8568955
## as.factor(Pclass)3:as.factor(Sex)male
##                              2.1674560
##                               Age:Fare
##                              0.0003845
##               as.factor(Pclass)2:Parch
##                              1.6272539
##               as.factor(Pclass)3:Parch
##                              0.4258068
##
## Degrees of Freedom: 711 Total (i.e. Null);  699 Residual
## Null Deviance:       960.9
## Residual Deviance: 586.5     AIC: 612.5
```

```
fitted<-glm(formula = Survived ~ as.factor(Pclass) + as.factor(Sex) +
    Age + SibSp + Parch + Fare + as.factor(Pclass):as.factor(Sex) +
    Age:Fare + as.factor(Pclass):Parch, family = "binomial",
    data = na.omit(train[, -11]))
```

Obtain a suitable threshold by minimizing misclassification error

```
conf <- matrix(0, nrow = 21, ncol = 5)
colnames(conf) <- c("thr", "a", "b", "c", "d")
conf[, 1] <- seq(0, 1, by = 0.05)
omitted<-na.omit(train[,-11])
y <- omitted$Survived
yhat <- fitted$fitted.values
for (i in 1:21) {
    a <- sum((!y) & (yhat <= conf[i, 1]))
    b <- sum((!y) & (yhat > conf[i, 1]))
    c <- sum((y) & (yhat <= conf[i, 1]))
    d <- sum((y) & (yhat > conf[i, 1]))
    conf[i, 2:5] <- c(a, b, c, d)
}
```

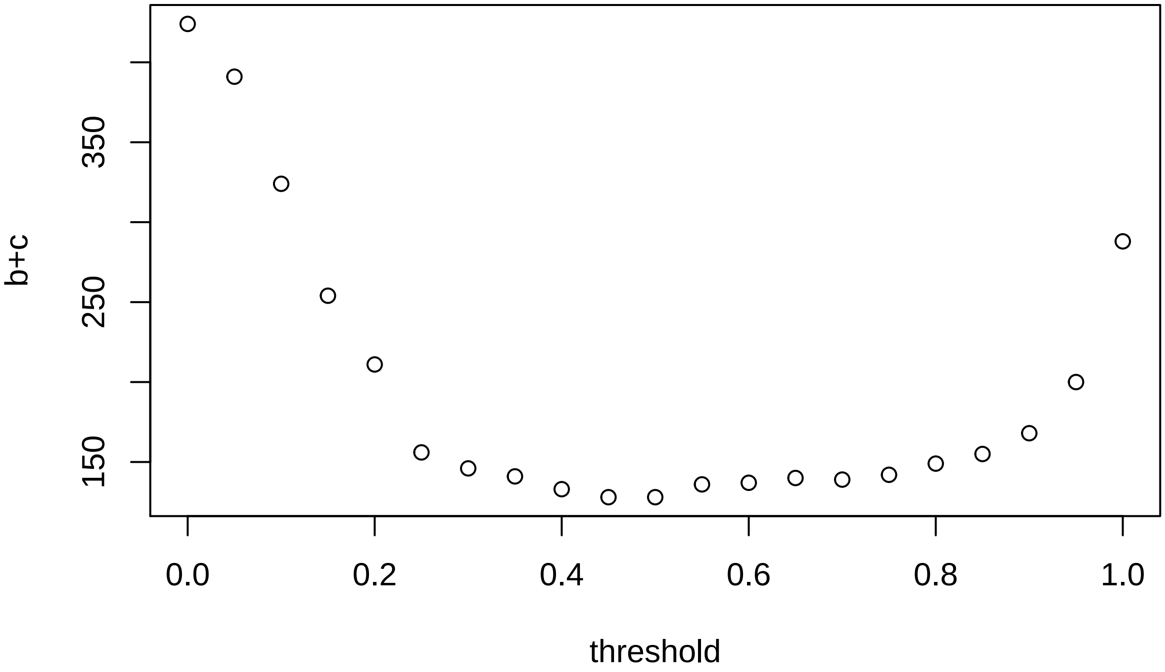
```
conf
```

```
##        thr   a   b   c   d
##  [1,] 0.00   0 424   0 288
##  [2,] 0.05  34 390   1 287
##  [3,] 0.10 110 314  10 278
##  [4,] 0.15 192 232  22 266
##  [5,] 0.20 247 177  34 254
##  [6,] 0.25 310 114  42 246
##  [7,] 0.30 330  94  52 236
##  [8,] 0.35 344  80  61 227
##  [9,] 0.40 359  65  68 220
## [10,] 0.45 375  49  79 209
## [11,] 0.50 386  38  90 198
## [12,] 0.55 395  29 107 181
## [13,] 0.60 404  20 117 171
## [14,] 0.65 412  12 128 160
## [15,] 0.70 415   9 130 158
## [16,] 0.75 415   9 133 155
## [17,] 0.80 417   7 142 146
## [18,] 0.85 417   7 148 140
## [19,] 0.90 419   5 163 125
## [20,] 0.95 422   2 198  90
## [21,] 1.00 424   0 288   0
```

```
conf[,"b"]+conf[,"c"]
```

```
##  [1] 424 391 324 254 211 156 146 141 133 128 128 136 137 140 139 142 149
## [18] 155 168 200 288
```

```
plot(conf[, 1], conf[, 3] + conf[, 4], xlab = "threshold",
    ylab = "b+c")
```



Based on the plot above and looking at the b+c values, the optimal threshold is at 0.5 where the b+c is lowest.

Predictions

```
#Read in test data
test = read_csv("~/stat28/projects/data/test_titanic.csv")
```

```
## Parsed with column specification:
## cols(
##   PassengerId = col_integer(),
##   Pclass = col_integer(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_double(),
##   SibSp = col_integer(),
##   Parch = col_integer(),
##   Ticket = col_character(),
##   Fare = col_double(),
##   Cabin = col_character(),
##   Embarked = col_character()
## )
```

```
test<-test[,-11]
#Fill in missing Age and Fare values using Amelia
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## Warning: package 'Rcpp' was built under R version 3.3.2
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2017c.1.0/
## zoneinfo/America/Los_Angeles'
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.4, built: 2015-12-05)
## ## Copyright (C) 2005-2018 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
test$Sex<-as.factor(test$Sex)
a.out<- amelia(test[,-c(3,8,10)],ts="PassengerId",noms = c("Sex"))
```

```
## -- Imputation 1 --
##
##   1  2  3  4  5  6  7
```

```
## Warning in is.na(value): is.na() applied to non-(list or vector) of type
## 'NULL'
```

```
##
## -- Imputation 2 --
##
##   1  2  3  4  5
```

```
## Warning in is.na(value): is.na() applied to non-(list or vector) of type
## 'NULL'
```

```
##
## -- Imputation 3 --
```

```
##
##   1  2  3  4  5  6  7  8
```

```
## Warning in is.na(value): is.na() applied to non-(list or vector) of type
## 'NULL'
```

```
##
## -- Imputation 4 --
##
##   1  2  3  4  5  6  7  8  9 10 11 12 13
```

```
## Warning in is.na(value): is.na() applied to non-(list or vector) of type
## 'NULL'
```

```
##
## -- Imputation 5 --
##
##   1  2  3
```

```
## Warning in is.na(value): is.na() applied to non-(list or vector) of type
## 'NULL'
```

```
test$Age<-a.out$imputations[[1]]$Age
test$Fare<-a.out$imputations[[1]]$Fare

pred.val = predict(fitted, test)
pred = as.numeric(pred.val > 0.50)
preds<-data.frame("PassengerId"=test$PassengerId,
                  "Survived"=pred)
pred.file = cbind(test$PassengerId, pred)
colnames(pred.file) = c("PassengerId", "Survived")
write.csv(preds, "Predictions.csv",row.names = F)
```

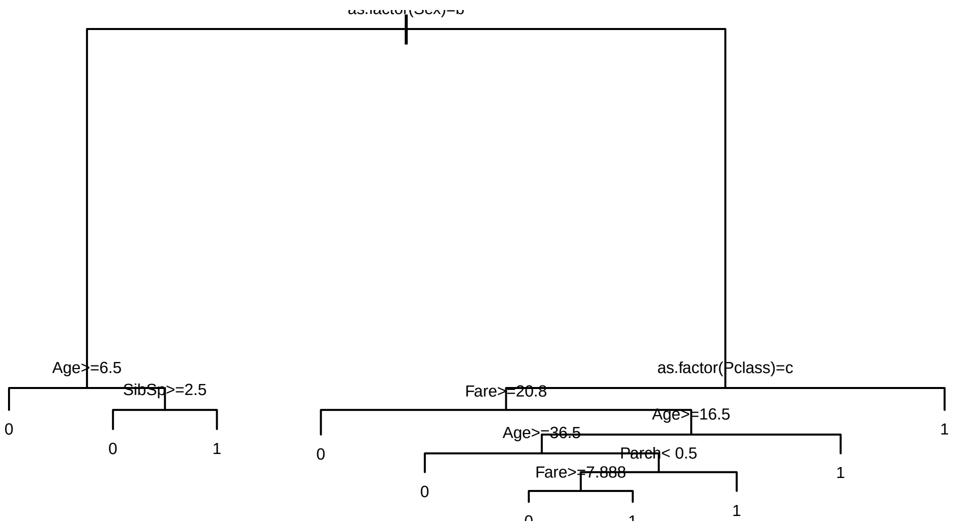
Kaggle Score=0.7512

The Kaggle score is computed by dividing the number of correct predictions by the total number of predictions.

Question 4

Fit a classification tree

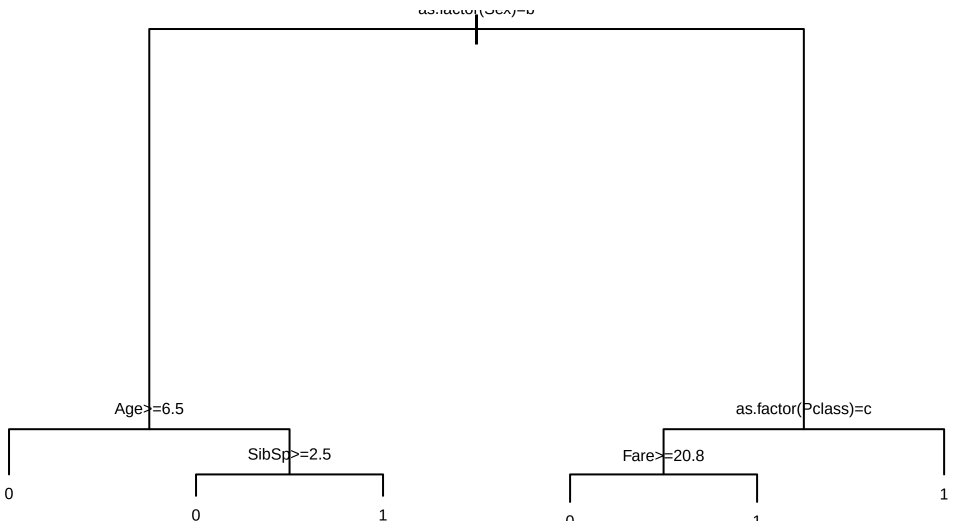
```
library(rpart)
tr<-rpart(Survived ~ as.factor(Pclass) + as.factor(Sex) +
    Age + SibSp + Parch + Fare,method="class" ,
    data = na.omit(train[, -11]))
plot(tr)
text(tr,cex=0.5)
```



```
#Pick the proper cp for tr
printcp(tr)
```

```
##
## Classification tree:
## rpart(formula = Survived ~ as.factor(Pclass) + as.factor(Sex) +
##     Age + SibSp + Parch + Fare, data = na.omit(train[, -11]),
##     method = "class")
##
## Variables actually used in tree construction:
## [1] Age               as.factor(Pclass) as.factor(Sex)    Fare
## [5] Parch             SibSp
##
## Root node error: 288/712 = 0.40449
##
## n= 712
##
##         CP nsplit rel error  xerror     xstd
## 1 0.454861      0   1.00000 1.00000 0.045472
## 2 0.029514      1   0.54514 0.54514 0.038412
## 3 0.027778      3   0.48611 0.56944 0.039010
## 4 0.024306      4   0.45833 0.57986 0.039258
## 5 0.010417      5   0.43403 0.53472 0.038146
## 6 0.010000      9   0.39236 0.54514 0.038412
```

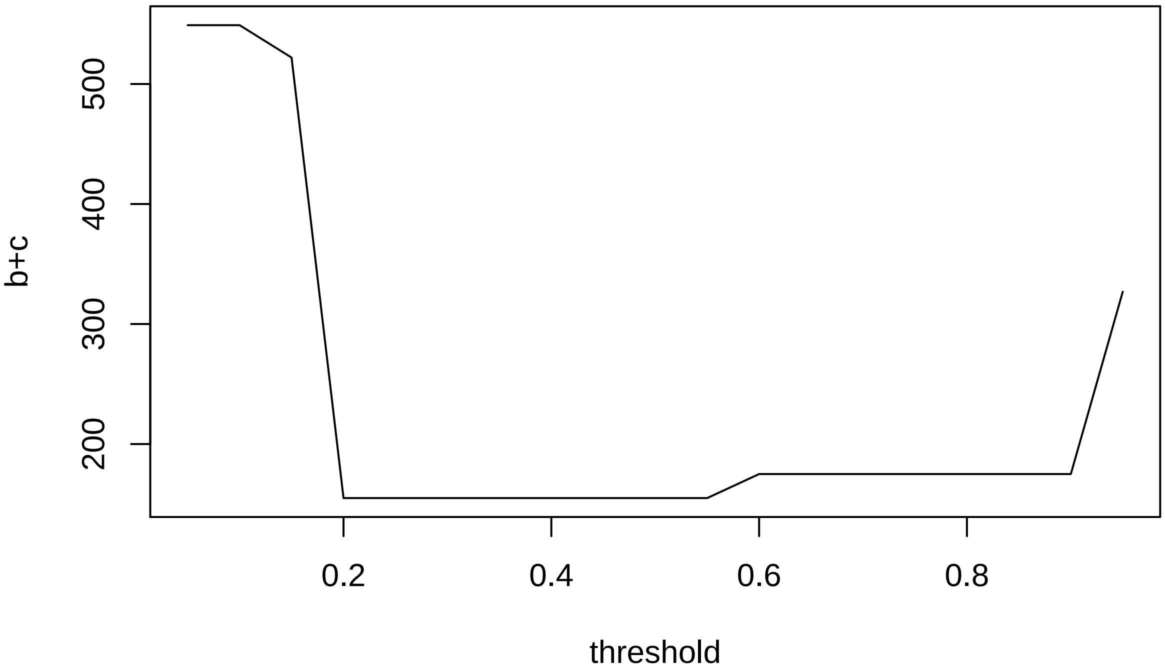
```
c<-0.010417
tr<-rpart(Survived ~ as.factor(Pclass) + as.factor(Sex) +
    Age + SibSp + Parch + Fare,method="class",cp=c,
    data = na.omit(train[, -11]))
plot(tr)
text(tr,cex=0.5)
```



Predict the values for the test set

```
y.tr = predict(tr, train)[, 2]
confusion <- function(y, yhat, thres) {
n <- length(thres)
conf <- matrix(0, length(thres), ncol = 4)
colnames(conf) <- c("a", "b", "c", "d")
for (i in 1:n) {
        a <- sum((!y) & (yhat <= thres[i]))
        b <- sum((!y) & (yhat > thres[i]))
        c <- sum((y) & (yhat <= thres[i]))
        d <- sum((y) & (yhat > thres[i]))
        conf[i, ] <- c(a, b, c, d)

}
return(conf) }
v = seq(0.05, 0.95, by = 0.05)
y = as.numeric(train$Survived == 1)
tree.conf = confusion(y, y.tr, v)
plot(v, tree.conf[, 2] + tree.conf[, 3], xlab = "threshold",
    ylab = "b+c", type = "l")
```



Based on the table of thresholds, the threshold with the lowest b+c value is 0.2

```
#Get predicted values using threshold
pred.val = predict(tr, test)
pred = as.numeric(pred.val > 0.2)

#Create data frame and write csv
preds<-data.frame("PassengerId"=test$PassengerId,
                  "Survived"=pred)
#Weird error where rows repeat themselves, eliminate duplicates
preds<-preds[419:836,]
write.csv(preds,"Classification.csv",row.names = F)
```

Kaggle score: 0.76555

Question 5:

Random Forests

Create a random forest

```
library(randomForest)
```

```
## randomForest 4.6-12
```

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

```
train$Pclass<-as.factor(train$Pclass)
train$Sex<-as.factor(train$Sex)
#Create a random forest
ft<-randomForest(as.factor(Survived) ~ Pclass + Sex +
    Age + SibSp + Parch + Fare + Pclass:Sex +
    Age:Fare + Pclass:Parch,importance=TRUE,
    data = na.omit(train[, -11]))
```

```
#Create a dataframe of predictions
test$Pclass<-as.factor(test$Pclass)
```

```r
test$Sex<-as.factor(test$Sex)
res<-predict(ft,test)
results<-data.frame("PassengerId"=test$PassengerId,
                    "Survived"=res)
write.csv(results,"Results.csv",row.names=F)
```

Kaggle Score: 0.75598

My Kaggle scores were not horrible, but certainly nowhere near the top of the leaderboard. In the future I would look to improve my predictions for the missing Age values to improve my final predictions. I would also try different methods of variable selection to see if I could find a model better than the one I chose.