Daniel Watabe

Titanic Survival Kaggle Competition

Given a training set, predict whether passengers survived or not from another set of data

Questions/Prior Assumptions/Knowledge

-Women/Children were prioritized to get on lifeboats

-If men had a chance to get on a boat they were typically of a higher social status

-Higher passenger classes also prioritized for lifeboats (1 being 1st class and 3 being 3rd class)

-People in the lower decks were more likely to not make it to boats in time

-People closer to iceberg hull breach were more likely to not make it to boats in time

-Were families most likely to split or stay together? (Die together or live together)

-Maybe just fathers were left behind.

-By looking at the deck plans majority of first class was towards the middle of the ship (potentially closer to lifeboats)

-Generally the 2nd and 3rd class were below the 1st class

https://www.encyclopedia-titanica.org/titanic-deckplans/

What we observe from the train set

This is the number that survived and number that did not

```
table(train$Survived)
0(did not survive)   1(survived)
549                  342
```

About 60%

This is the proportion that survived based on sex

```
prop.table(table(train$Sex, train$Survived),1)
                0              1
 female  0.2579618  0.7420382
 male     0.8110919  0.1889081
```

This is the proportion that survived based on Pclass and Sex

```
aggregate(Survived ~ Pclass + Sex, data = train, FUN=function(x){sum(x)/length(x)})
  Pclass  Sex        Survived
1  1    female       0.9680851
2  2      female     0.9210526
3  3      female     0.5000000
4  1      male       0.3688525
5  2      male       0.1574074
6  3      male       0.1354467
```

Summary of Fare Prices

```
summary(train$Fare)
Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
0.00   7.91   14.45   32.20   31.00  512.30
```

The following makes a new variable “Fare Categories” and organizes them fares by price

```
train$FareCategories[train$Fare >= 30] = ’30+’
train$FareCategories[train$Fare < 30 & train$Fare >=20] = ’20-30’
train$FareCategories[train$Fare < 20 & train$Fare >=10] = ’10-20’
train$FareCategories[train$Fare < 10] = ’<10’
```

The following makes a new variable “isChild” and sets it to 1(true) if age is less than 18

```
train$isChild = ’0’
train$isChild[train$Age < 18] = ’1’
```

This outputs the proportion that survived when FareCategories, Pclass, Sex, and isChild is used as a “key”.

```
aggregate(Survived ~ FareCategories + Pclass + Sex + isChild, data = train, FUN=function(x) {sum(x)/length(x)})

      FareCategories Pclass   Sex isChild   Survived
1        20-30     1 female      0 0.83333333
2          30+     1 female      0 0.98750000
3        10-20     2 female      0 0.90625000
```

```
4   20-30 2 female   0 0.88000000
5     30+ 2 female   0 1.00000000
6     <10 3 female   0 0.56140351
7   10-20 3 female   0 0.50000000
8   20-30 3 female   0 0.40000000
9     30+ 3 female   0 0.11111111
10    <10 1   male   0 0.00000000
11  20-30 1   male   0 0.40000000
12    30+ 1   male   0 0.35365854
13    <10 2   male   0 0.00000000
14  10-20 2   male   0 0.11864407
15  20-30 2   male   0 0.04761905
16    30+ 2   male   0 0.00000000
17    <10 3   male   0 0.10931174
18  10-20 3   male   0 0.12903226
19  20-30 3   male   0 0.07142857
20    30+ 3   male   0 0.41666667
21    30+ 1 female   1 0.87500000
22  10-20 2 female   1 1.00000000
23  20-30 2 female   1 1.00000000
24    30+ 2 female   1 1.00000000
25    <10 3 female   1 0.85714286
26  10-20 3 female   1 0.73333333
27  20-30 3 female   1 0.16666667
28    30+ 3 female   1 0.14285714
29    30+ 1   male   1 1.00000000
30  10-20 2   male   1 0.75000000
31  20-30 2   male   1 0.75000000
32    30+ 2   male   1 1.00000000
33    <10 3   male   1 0.15384615
34  10-20 3   male   1 0.71428571
35  20-30 3   male   1 0.20000000
36    30+ 3   male   1 0.07692308
```

Just for simplicity I just filled all empty ages and fares in the test file to be the median age because some rich elderly passengers were pulling up the mean.

```
> summary(test$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s 
   0.17   21.00   27.00   30.27   39.00   76.00      86 
> summary(test$Fare)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s 
  0.000   7.896  14.450  35.630  31.500 512.300       1 
```

```
test$Age[which(is.na(test$Age))] = 27
test$Fare[which(is.na(test$Fare))] = 14.45
test$isChild = 0
> test$isChild[test$Age < 18] =    ‘1’
> test$FareCategories[test$Fare >= 30] = ’30+’
> test$FareCategories[test$Fare < 30 & test$Fare >=20] = ’20-30’
> test$FareCategories[test$Fare < 20 & test$Fare >=10] = ’10-20’
> test$FareCategories[test$Fare < 10] = ’<10’
> 
> fit <- rpart(Survived ~ Pclass + Sex + Age + FareCategories + isChild, data=train, method="class")
> Prediction <- predict(fit, test, type = "class")
submit <- data.frame(PassengerId = test$PassengerId, Survived = Prediction)
write.csv(submit, file = "titanic_test.csv", row.names = FALSE)
```