Titanic Survival - Kaggle

Tobias Ford

2017-05-17

Contents

Titanic Survival - Kaggle

Introduction

Provided initial training set of data (train.csv) containing passenger roster with some key details.

Load Data

```
train <- read.csv('../input/train.csv', stringsAsFactors = F)
test <- read.csv('../input/test.csv', stringsAsFactors = F)

str(train)
```

```
## 'data.frame':    891 obs. of  12 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       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Data Dictionary

```
head(train,1)
```

```
##   PassengerId Survived Pclass                     Name  Sex Age SibSp Parch    Ticket Fare Cabin Embar
## 1           1        0      3 Braund, Mr. Owen Harris male  22     1     0 A/5 21171 7.25
```

- PassengerId

- Survived
- Pclass
- Name
- Sex
- Age
- SibSp
- Parch
- Ticket
- Fare
- Cabin
- Embarked

First level of investigation

Distribution Based on Sex

```
table(train$Sex)
```

```
##
## female   male
##    314    577
```

```
summary(train$Sex)
```

```
##    Length     Class      Mode
##       891 character character
```

```
prop.table(table(train$Sex))
```

```
##
##   female     male
## 0.352413 0.647587
```

```
prop.table(table(train$Survived))
```

```
##
##         0         1
## 0.6161616 0.3838384
```

```
train$SurvivedBoolean <- as.logical(train$Survived)
train$SurvivedLabel[train$SurvivedBoolean == TRUE] <- 'Survived'
train$SurvivedLabel[train$SurvivedBoolean == FALSE] <- 'Died'

prop.table(table(train$Sex, train$SurvivedLabel))
```

```
##
##                Died   Survived
##   female 0.09090909 0.26150393
##   male   0.52525253 0.12233446
```

```
prop.table(table(train$Sex, train$SurvivedLabel),1)
```

```
##
##                Died  Survived
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

Diving into Age

```
summary(train$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

```
train$Child <- FALSE
train$Child[train$Age < 18] <- TRUE

table(train$Child)
```

```
##
## FALSE  TRUE
##   778   113
```

```
table(train$Child, train$Survived)
```

```
##
##           0   1
##   FALSE 497 281
##   TRUE   52  61
```

```
aggregate(Survived ~ Child + Sex, data=train, FUN=sum)
```

```
##   Child    Sex Survived
## 1 FALSE female      195
## 2  TRUE female       38
## 3 FALSE   male       86
## 4  TRUE   male       23
```

```
aggregate(Survived ~ Child + Sex, data=train, FUN=length)
```

```
##   Child    Sex Survived
## 1 FALSE female      259
## 2  TRUE female       55
## 3 FALSE   male      519
## 4  TRUE   male       58
```

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

```
##   Child    Sex  Survived
## 1 FALSE female 0.7528958
## 2  TRUE female 0.6909091
## 3 FALSE   male 0.1657033
## 4  TRUE   male 0.3965517
```

Diving into Fare

```
train$Fare2 <- '30+'
train$Fare2[train$Fare < 30 & train$Fare >= 20] <- '20-30'
train$Fare2[train$Fare < 20 & train$Fare >= 10] <- '10-20'
train$Fare2[train$Fare < 10] <- '<10'

aggregate(Survived ~ Fare2 + Pclass + Sex, data=train, FUN=function(x) {sum(x)/length(x)})
```

```
##    Fare2 Pclass    Sex  Survived
```

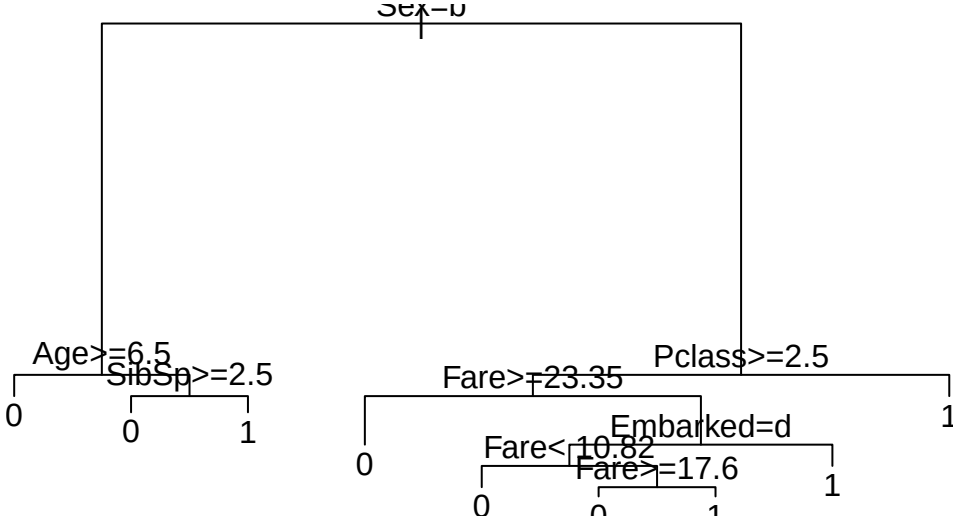
```
## 1  20-30      1 female 0.8333333
## 2    30+      1 female 0.9772727
## 3  10-20      2 female 0.9142857
## 4  20-30      2 female 0.9000000
## 5    30+      2 female 1.0000000
## 6    <10      3 female 0.5937500
## 7  10-20      3 female 0.5813953
## 8  20-30      3 female 0.3333333
## 9    30+      3 female 0.1250000
## 10   <10      1   male 0.0000000
## 11 20-30      1   male 0.4000000
## 12   30+      1   male 0.3837209
## 13   <10      2   male 0.0000000
## 14 10-20      2   male 0.1587302
## 15 20-30      2   male 0.1600000
## 16   30+      2   male 0.2142857
## 17   <10      3   male 0.1115385
## 18 10-20      3   male 0.2368421
## 19 20-30      3   male 0.1250000
## 20   30+      3   male 0.2400000
```

Determined something meaningful here and added it to the submission. Women in class 3 who paid $20 or more for a fare didn't fair well (pun intended).

Using Decision Trees

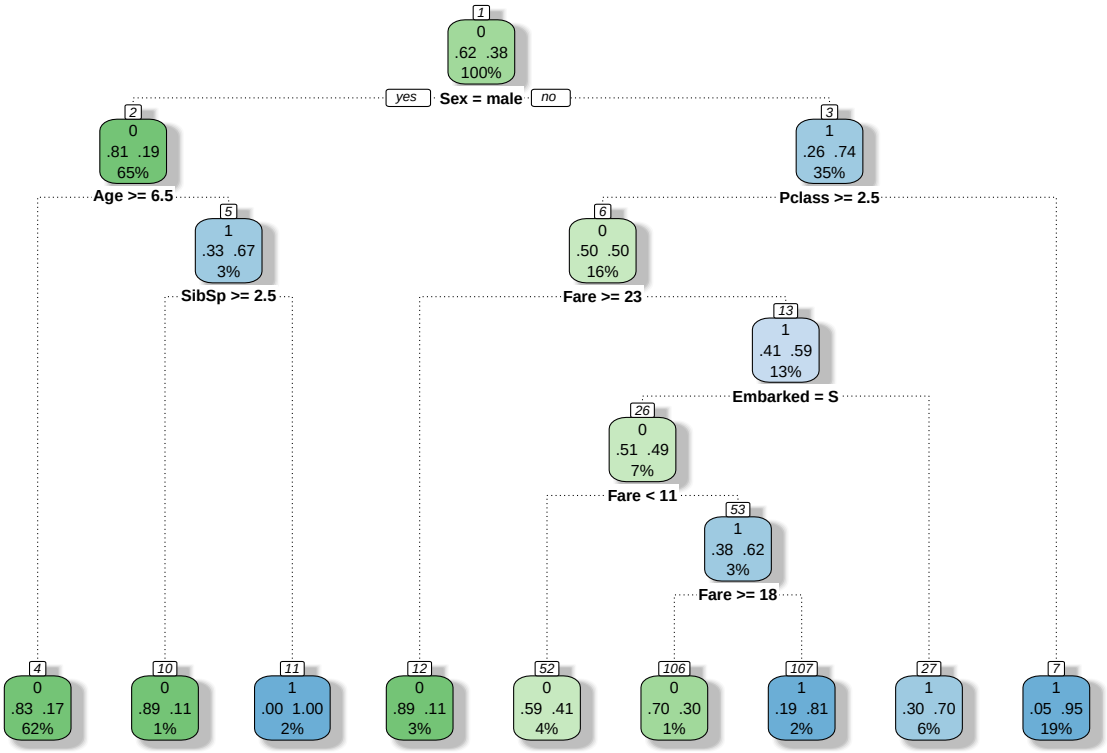
```
library(rpart)
fit <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,
               data=train,
               method="class")

plot(fit)
text(fit)
```




```
library(rattle)
library(rpart.plot)
library(RColorBrewer)
```

```
fancyRpartPlot(fit)
```



Rattle 2017−May−17 10:25:38 tobybot11

```
Prediction <- predict(fit, test, type = "class")
```

Submission

first attempt

```
test$Survived <- rep(0, 418)
test$Survived <- 0
test$Survived[test$Sex == 'female'] <- 1
test$Survived[test$Sex == 'female' & test$Pclass == 3 & test$Fare >= 20] <- 0

head(test,10)
```

```
##    PassengerId Pclass                                          Name    Sex  Age SibSp Parch    Ticket
## 1          892      3                              Kelly, Mr. James   male 34.5     0     0    330911
## 2          893      3              Wilkes, Mrs. James (Ellen Needs) female 47.0     1     0    363272
## 3          894      2                     Myles, Mr. Thomas Francis   male 62.0     0     0    240276
## 4          895      3                              Wirz, Mr. Albert   male 27.0     0     0    315154
## 5          896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0     1     1   3101298
## 6          897      3                    Svensson, Mr. Johan Cervin   male 14.0     0     0      7538
## 7          898      3                          Connolly, Miss. Kate female 30.0     0     0    330972
## 8          899      2                 Caldwell, Mr. Albert Francis   male 26.0     1     1    248738
## 9          900      3    Abrahim, Mrs. Joseph (Sophie Halaut Easu) female 18.0     0     0      2657
## 10         901      3                      Davies, Mr. John Samuel   male 21.0     2     0 A/4 48871
```

```
##    Embarked Survived
## 1         Q        0
## 2         S        1
## 3         Q        0
## 4         S        0
## 5         S        1
## 6         S        0
## 7         Q        1
## 8         S        0
## 9         C        1
## 10        S        0
```

```r
submit <- data.frame(PassengerId = test$PassengerId, Survived = test$Survived)
submit <- data.frame(PassengerId = test$PassengerId, Survived = Prediction)
write.csv(submit, file = "theyallperish.csv", row.names=FALSE)
```