# 타이타닉 생존 예측

이영석

2017-03-30

# 타이타닉호 데이터와 Kaggle

- Titanic 호 사건이란 ?
  - *1912년 4월 15일 대서양에서 침몰한 여객선 사건으로 2,224명 탑승자 중 1,514명 사망, 710명 생존함*
- Kaggle 사이트 <https://www.kaggle.com/c/titanic >
- Datacamp 사이트 <https://www.datacamp.com/community/open-courses/kaggle-tutorial-on-machine-learing-the-sinking-of-the-titanic >

# 공부할 것

- 예측문제를 어떻게 통계적 지식으로 해결하는가?
- R or Python 으로 decision tree, regression 사용하기
- Kaggle 사이트이용하기

# 데이터 읽어들이기

```r
# Import the training set: train
train_url <- "http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/train.csv"
train <- read.csv(train_url)

# Import the testing set: test
test_url <- "http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/test.csv"
test <- read.csv(test_url)

# Print train and test to the console
#train
#test
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
```

```r
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
## 1 34.5     0     0  330911  7.8292              Q
## 2 47.0     1     0  363272  7.0000              S
## 3 62.0     0     0  240276  9.6875              Q
## 4 27.0     0     0  315154  8.6625              S
## 5 22.0     1     1 3101298 12.2875              S
## 6 14.0     0     0    7538  9.2250              S
```

# 데이터 테이블로 살펴보기

```
# Your train and test set are still loaded
#str(train)
#str(test)

# Survival rates in absolute numbers
table(train$Survived)
```

```
## 
##   0   1 
## 549 342 
```

```
# Survival rates in proportions
prop.table(table(train$Survived))
```

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

```
# Two-way comparison: Sex and Survived
table(train$Sex, train$Survived)
```

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

```
# Two-way comparison: row-wise proportions

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

```
##         
##                  0         1
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

# 어린이 데이터 테이블

```r
# Your train and test set are still loaded in
#str(train)
#str(test)

# Create the column child, and indicate whether child or no child
train$Child <- NA
train$Child[train$Age < 18] <- 1
train$Child[train$Age >= 18] <- 0

# Two-way comparison
prop.table(table(train$Child, train$Survived), 1)
```

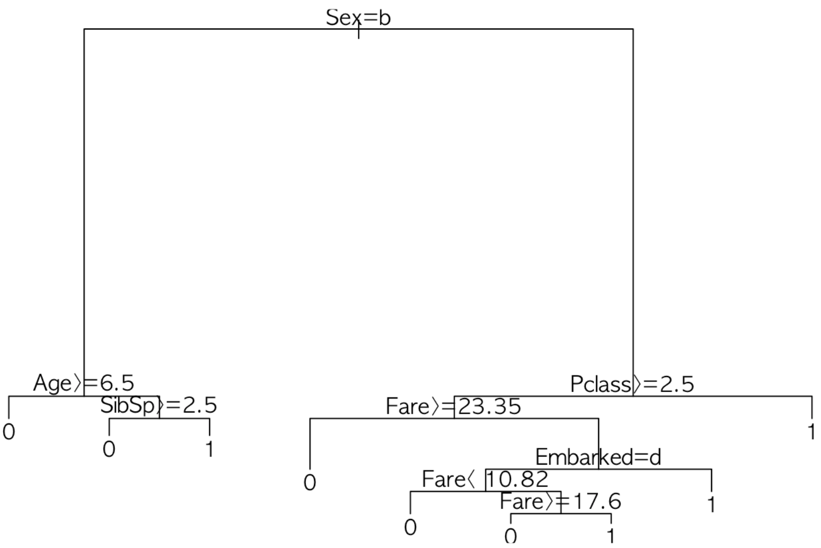
```
##
##              0         1
##   0 0.6189684 0.3810316
##   1 0.4601770 0.5398230
```

# 테스트 데이터 생성

```r
# Your train and test set are still loaded in
#str(train)
#str(test)

# Copy of test
test_one <- test

# Initialize a Survived column to 0
test_one$Survived <- 0

# Set Survived to 1 if Sex equals "female"
test_one$Survived[test$Sex == "female"] <- 1
```

# Decision Tree 만들기

```
library(rpart)
# Your train and test set are still loaded in
#str(train)
#str(test)

# Build the decision tree
my_tree_two <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked, data = train, method = "class")

# Visualize the decision tree using plot() and text()
plot(my_tree_two)
text(my_tree_two)
```



# 좀더 보기 좋은 Decision Tree

```
# Load in the packages to build a fancy plot
library(rattle)
```

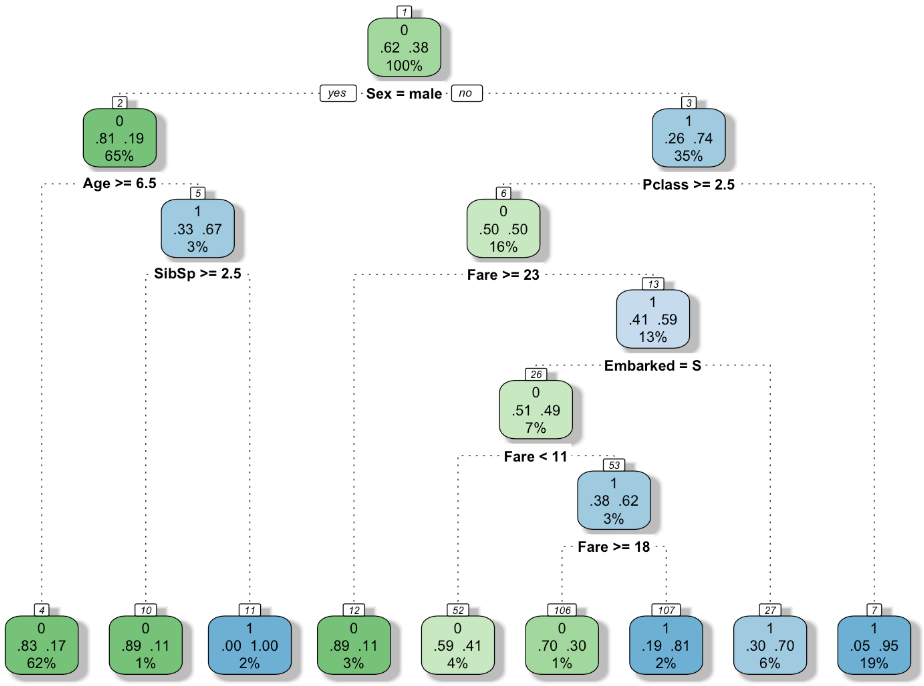
```
## Please install GTK+ from http://r.research.att.com/libs/GTK_2.24.17-X11.pkg
```

```
## If the package still does not load, please ensure that GTK+ is installed and that it is on yo
```

```
## IN ANY CASE, RESTART R BEFORE TRYING TO LOAD THE PACKAGE AGAIN
```

```
## Rattle: A free graphical interface for data mining with R.
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

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

# Time to plot your fancy tree
fancyRpartPlot(my_tree_two)
```



Rattle 2017-Mar-28 15:34:14 youngseoklee

# 예측결과 저장하기

```r
# my_tree_two and test are available in the workspace
# Make predictions on the test set
my_prediction <- predict(my_tree_two, newdata = test, type = "class")

# Finish the data.frame() call
my_solution <- data.frame(PassengerId = test$PassengerId, Survived = my_prediction)

# Use nrow() on my_solution
nrow(my_solution)
```

```
## [1] 418
```

```r
# Finish the write.csv() call
write.csv(my_solution, file = "my_solution.csv", row.names = FALSE)
```