헬로 데이터 과학: 분류 실습

Jin Young Kim

이번에는 캐글에 입문하는 사람들을 위해 캐글에서 제공하는 예측 문제를 바탕으로 실제 캐글에서 제공하는 문제를 풀어 답을 제출하고, 퍼블릭 리더보드에서 순위를 확인하는 과정을 알아보자. 이 과정에서 앞서 공부한 데이터 문제 해결의 과정을 실습해보는 의미도 있다.

여기서 사용할 문제는 '타이타닉호의 생존자 예측'으로, 타이타닉 호에 탑승한 승객의 생존 여부를 예측하는 것이다. 최근에 있었던 세월호 사태와 같이 대형 인명사고 이후에 많은 사람들의 생존 여부가 밝혀지지 않은 경우 예측 모델을 사용하면 승객들의 생사를 확인하는데 도움을 받을 수 있을 것이다. (물론 이는 생사가 알려진 승객과 알려지지 않은 승객의 집단이 무작위로 추출되었을 때만 가능한 일이다.)

여기에 사용되는 데이터셋은 타이타닉호에 탑승한 승객들의 다양한 신상정보와 (나이, 객실, 성별 등) 생존여부를 기록한 것으로, 이 데이터셋은 캐글(Kaggle)의 홈페이지에서 구할 수 있다. 우선 이 데이터를 가지고 탐색적 데이터 분석을 수행해보고, 이를 바탕으로 승객들의 생존 여부를 예측하는 작업을 수행한다. 본 실습에는 R을 사용한다. (R은 캐글 참가자들이 가장 많이 사용하는 프로그래밍 언어이기도 하다.)

source("dbook.R")  
load.packages(c("ggplot2", "rpart", "rpart.plot"))

## [1] loading ggplot2  
## [1] loading rpart  
## [1] loading rpart.plot

train <- read.csv("titanic\_train.csv")  
test <- read.csv("titanic\_test.csv")

#### 개별 속성 분석

우선 데이터의 전체적인 분포를 알아보자. 아래는 데이터의 속성별 값의 분포를 보여준다.

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.:20.12   
## Abbott, Mrs. Stanton (Rosa Hunt) : 1 Median :28.00   
## Abelson, Mr. Samuel : 1 Mean :29.70   
## Abelson, Mrs. Samuel (Hannah Wizosky): 1 3rd Qu.:38.00   
## Adahl, Mr. Mauritz Nils Martin : 1 Max. :80.00   
## (Other) :885 NA's :177   
## 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  
## :687 : 2   
## B96 B98 : 4 C:168   
## C23 C25 C27: 4 Q: 77   
## G6 : 4 S:644   
## C22 C26 : 3   
## D : 3   
## (Other) :186

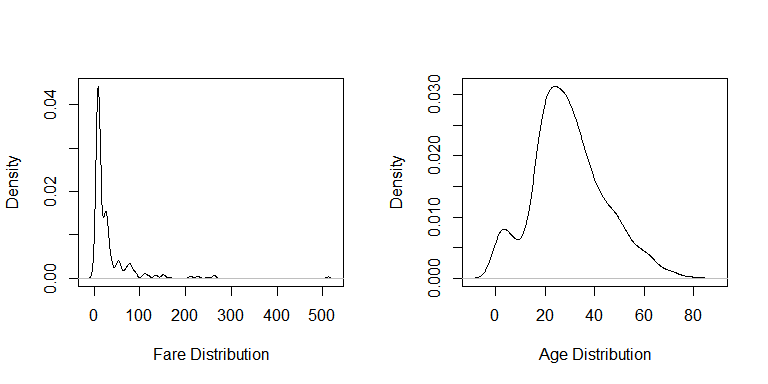
이제 예측 목표가 되는 생존 여부의 분포를 알아보자. 주어진 학습 데이터의 경우 약 38%의 승객이 생존한 것을 알 수 있다.

table(train$Survived)

##   
## 0 1   
## 549 342

이중 수치형 속성인 티켓 가격과 나이의 확률 분포를 알아보면 다음과 같다. 티켓 가격은 상대적으로 저렴한 50불 미만이 대부분이고, 나이는 20-40대의 청장년 층과 그들의 자녀일 것으로 추정되는 어린 아이가 많은 것을 알 수 있다.

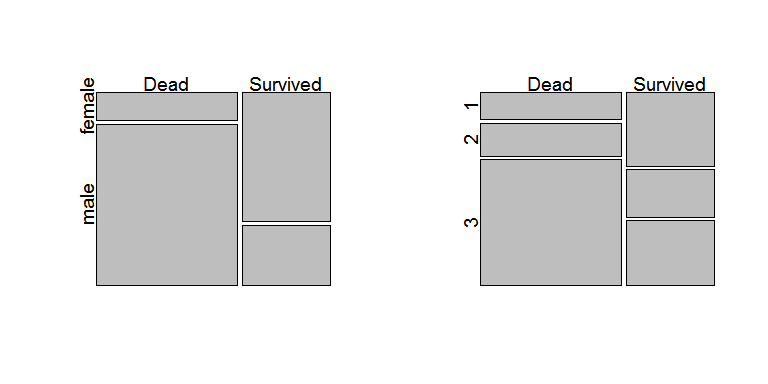
par(mfrow=c(1,2))  
plot(density(train$Fare), main="", xlab="Fare Distribution")  
plot(density(train[!is.na(train$Age),]$Age), main="", xlab="Age Distribution")



#### 속성간 관계 분석

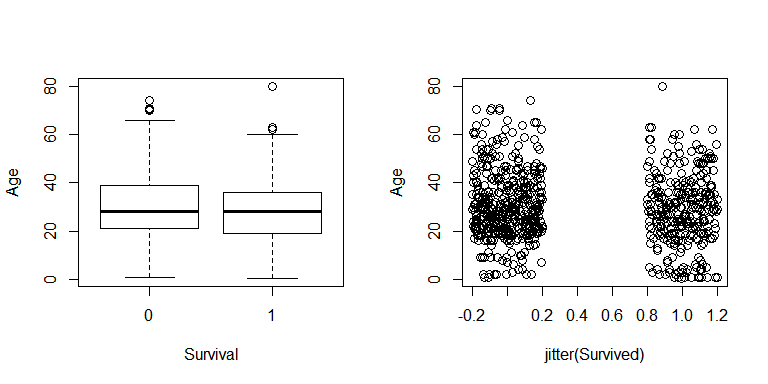
이제 속성간 관계를 살펴보자. 다음은 타이타닉호의 탑승자 가운데 남녀 및 각 객실의 생존 비율을 비교한 모자이크 플롯이다. 상대적으로 남성이, 그리고 3등석 손님의 사망률이 높았음을 알 수 있다.

par(mfrow=c(1,2))  
mosaicplot(table(ifelse(train$Survived==1, "Survived","Dead"), train$Sex), main="", cex=1.2)  
mosaicplot(table(ifelse(train$Survived==1, "Survived","Dead"), train$Pclass), main="", cex=1.2)



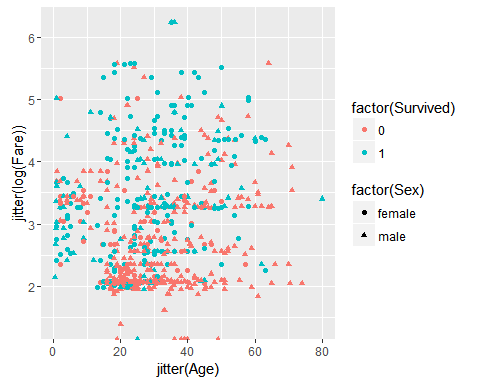
아래는 타이타닉호 탑승객의 나이와 생존 여부의 관계를 나타낸 그래프다. 왼쪽 박스플롯은 얼핏 생존자의 나이가 약간 어리지만 두 그룹간의 큰 차이가 없는 것으로 보인다. 하지만 오른쪽의 스케터플롯을 보면 사망자중 20대와 노인의 비율이 높고, 생존자중에는 아이의 비율이 높은 것을 알 수 있다. 이처럼 시각화의 종류에 따라 같은 데이터에서 다른 결론을 도출할 수 있다.

par(mfrow=c(1,2))  
boxplot(Age ~ Survived, train, xlab="Survival", ylab="Age", cex=1.2)  
plot(Age ~ jitter(Survived), train, cex=1.2)



마지막으로 두 수치 속성 간의 관계를 나타내는 데에는 스케터플롯이 적절하다. 스케터플롯의 각 점에 색상과 모양을 추가하면 더 많은 정보를 한눈에 볼 수 있다. 아래 플롯은 나이와 요금의 (로그 스케일) 스케터플롯에 색상과 (생존 여부) 모양을 (성별) 추가한 결과다. 젊은 남성으로 낮은 요금을 지불한 승객의 경우 사망 확률이 매우 높은 것을 알 수 있다.

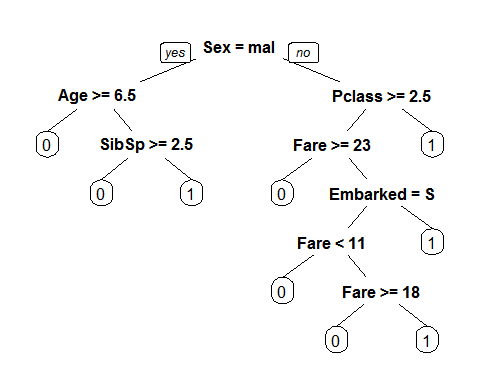
qplot(jitter(Age), jitter(log(Fare)), data=train, color=factor(Survived), shape=factor(Sex))



#### 모델 만들기

이제 기본적인 예측 모델을 만들어 보자. 기본 모델을 만드는 이유에는 데이터를 잘 이해하는 것도 있으니 해석이 쉬운 의사결정트리를 만들어 보자. 아래는 데이터에 주어진 속성을 가지고 만든 의사결정트리를 시각화한 결과다. 승객의 성과 나이, 객실 유형이 예측에 있어 첫번째 기준이 되는 것을 알 수 있다.

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



마지막으로 모델을 실제 테스트 데이터 적용하여 제출용 파일을 만들어 보자. R에서는 predict() 함수를 사용하여 만들어진 모델을 새로운 데이터에 적용할 수 있다.

Prediction <- predict(fit, test, type = "class")  
submit <- data.frame(PassengerId = test$PassengerId, Survived = Prediction)  
head(submit)

## PassengerId Survived  
## 1 892 0  
## 2 893 0  
## 3 894 0  
## 4 895 0  
## 5 896 1  
## 6 897 0

write.csv(submit, file = "titanic\_submission.tsv", row.names = FALSE)

### 맺음말

본 R 예제는 데이터 과학 입문서인 [헬로 데이터 과학](http://www.yes24.com/24/goods/24349374)의 부록으로 제공됩니다.