Shruti Mehta
MS in Information Systems
Northeastern University

INFO 7390: Advances in Data Sciences and Architecture

Report

Load Titanic dataset along with Test data

```
train_data <- read.csv("./datasets/train.csv")
test_data <- read.csv("./datasets/test.csv")
```

Exploring the data

```
str(train_data)

## '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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109
191 358 277 16 559 520 629 417 581 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2
1 1 ...
##  $ 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     : Factor w/ 681 levels "110152","110413",..: 524 597
670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1
1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4
4 2 ...

head(train_data)

##   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 Si
bSp
## 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
```

```
tail(train_data)
```

```
##     PassengerId Survived Pclass
Name
## 886         886        0      3     Rice, Mrs. William (Margaret No
rton)
## 887         887        0      2                    Montvila, Rev. J
uozas
## 888         888        1      1             Graham, Miss. Margaret 
Edith
## 889         889        0      3 Johnston, Miss. Catherine Helen "Ca
rrie"
## 890         890        1      1                    Behr, Mr. Karl H
owell
## 891         891        0      3                      Dooley, Mr. Pa
trick
##        Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886 female  39     0     5     382652 29.125              Q
## 887   male  27     0     0     211536 13.000              S
## 888 female  19     0     0     112053 30.000   B42        S
## 889 female  NA      1     2 W./C. 6607 23.450              S
## 890   male  26      0     0     111369 30.000  C148        C
## 891   male  32      0     0     370376  7.750              Q
```

Age column have some missing values

```
summary(train_data$Age)

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

Imputing the missing values from Age columns as replace them with mean

```
train_data$Age[is.na(train_data$Age)] <- mean(train_data$Age, na.rm = TRUE)
summary(train_data$Age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   22.00   29.70   29.70   35.00   80.00
```

Age and Fare columns in test data is also missing, so we fix them by replacing with mean

```
summary(test_data$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_data$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_data$Age[is.na(test_data$Age)] <- mean(test_data$Age, na.rm = TRU
E)
test_data$Fare[is.na(test_data$Fare)] <- mean(test_data$Fare, na.rm =
TRUE)

summary(test_data$Age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.17   23.00   30.27   30.27   35.75   76.00

summary(test_data$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   7.896  14.450  35.630  31.500 512.300
```

Survived column is integer class type

```
class(train_data$Survived)

## [1] "integer"

levels(as.factor(train_data$Survived))

## [1] "0" "1"
```

Converting it to factor with yes and no level

```
head(train_data$Survived)

## [1] 0 1 1 1 0 0

train_data$Survived <- ifelse(train_data$Survived == 1, "yes", "no")
train_data$Survived <- as.factor(train_data$Survived)
head(train_data$Survived)

## [1] no  yes yes yes no  no 
## Levels: no yes

class(train_data$Survived)

## [1] "factor"

library(rpart)


table(as.factor(train_data$Survived))

## 
##  no yes 
## 549 342

train_data$Survived <- as.factor(train_data$Survived)
str(train_data$Survived)

##  Factor w/ 2 levels "no","yes": 1 2 2 2 1 1 1 1 2 2 ...

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

```
##
##        no       yes
## 0.6161616 0.3838384
```

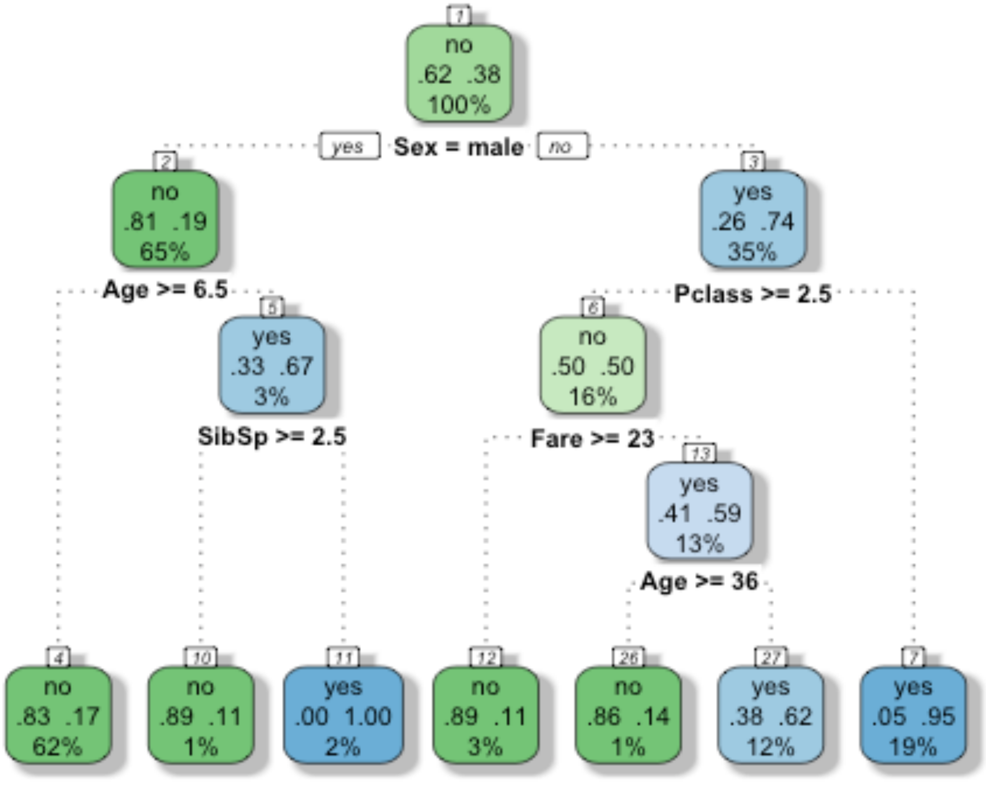
Identity columns like passenger id, name, cabin ignored for predictor variables

```
tree <- rpart(formula = Survived ~ Sex+Age+SibSp+Parch+Fare+Embarked+P
class,
              data = train_data,
              method = "class")

library(rattle)

## 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)
fancyRpartPlot(tree)
```



Now predicting the Survival status for test data

```
test_data$Survived <- as.factor(c("yes","no"))
test_data$Survived <- predict(tree, test_data, type="class")

table(test_data$Survived)

##
##  no yes
## 272 146

prop.table(table(test_data$Survived))

##
##        no       yes
## 0.6507177 0.3492823
```

Conclusion

- After loading the data, summary shows that Age columns have some missing value, so I replaced them with the mean of Age.
- Survived column was integer type so for classification I converted it to the Factor also set the labeled it with "Yes" and "No" values for 1 and 0 respectively.
- The identity variables like Passenger Id and Name are not considered in the predictor variables.
- The generated Decision Tree shows that Survival Rate. At the top node, 62% passengers have died, and 38% have survived. 100% of the sample is used here as shown in the top node.
- The first Split is based on Sex, if person is male then check left.
- For males, 81% of them died as compare to 19% who survived.
- For females, on right side, "yes" is voted for survival, 74% are survived and 26% died. We can conclude, more females are survived as compare to males.
- Same process will follow for other branches in the tree.
- From prediction we say that the our model did Good for Test data because number of people died is 65% and 35% survived which is close to the Trained data numbers.