Gender_Model.R

JingbinXu

Thu Dec 22 16:16:03 2016

```
# Jinbin Xu - 25 Dec 2016
# Titanic-R

# Set working directory and import datafiles
train <- read.csv("train.csv")
test <- read.csv("test.csv")

# Look at gender patterns
summary(train$Sex)
```

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

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

```
##
##                    0          1
##   female 0.09090909 0.26150393
##   male   0.52525253 0.12233446
```

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

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

```
# Create new column in test set with our prediction that everyone dies
test$Survived <- 0
# Update the prediction to say that all females will survive
test$Survived[test$Sex == 'female'] <- 1

# Create submission dataframe and output to file
submit <- data.frame(PassengerId = test$PassengerId, Survived = test$Survived)
write.csv(submit, file = "gendermodel.csv", row.names = FALSE)

# Look at age patterns
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$Child <- 0
train$Child[train$Age < 18] <- 1
aggregate(Survived ~ Child + Sex, data=train, FUN=sum)
```

```
##   Child    Sex Survived
## 1     0 female      195
## 2     1 female       38
## 3     0   male       86
## 4     1   male       23
```

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

```
##   Child    Sex Survived
## 1     0 female      259
## 2     1 female       55
## 3     0   male      519
## 4     1   male       58
```

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

```
##   Child    Sex  Survived
## 1     0 female 0.7528958
## 2     1 female 0.6909091
## 3     0   male 0.1657033
## 4     1   male 0.3965517
```

```
# Look at class and fare patterns
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  10-20      3 female 0.5813953
## 7  20-30      3 female 0.3333333
## 8    30+      3 female 0.1250000
## 9    <10      3 female 0.5937500
## 10 20-30      1   male 0.4000000
## 11   30+      1   male 0.3837209
## 12   <10      1   male 0.0000000
## 13 10-20      2   male 0.1587302
## 14 20-30      2   male 0.1600000
## 15   30+      2   male 0.2142857
## 16   <10      2   male 0.0000000
## 17 10-20      3   male 0.2368421
## 18 20-30      3   male 0.1250000
## 19   30+      3   male 0.2400000
## 20   <10      3   male 0.1115385
```

```
# Create new column in test set with our prediction that everyone dies
test$Survived <- 0
# Update the prediction to say that all females will survive
test$Survived[test$Sex == 'female'] <- 1
# Update once more to say that females who pay more for a third class fare also peris
h
test$Survived[test$Sex == 'female' & test$Pclass == 3 & test$Fare >= 20] <- 0

# Create submission dataframe and output to file
submit <- data.frame(PassengerId = test$PassengerId, Survived = test$Survived)
write.csv(submit, file = "genderclassmodel.csv", row.names = FALSE)
```