Titanic Data Project

Nathan Kim

2023-12-29

```
library(readr)
```

```
## Warning: package ’readr’ was built under R version 4.2.2
```

```
train <- read_csv("train.csv")
```

```
## Rows: 891 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Sex, Ticket, Cabin, Embarked
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
test <- read_csv("test.csv")
```

```
## Rows: 418 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Sex, Ticket, Cabin, Embarked
## dbl (6): PassengerId, Pclass, Age, SibSp, Parch, Fare
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
## Create factors for our datasets

train$Survived <- factor(train$Survived)
train$Pclass <- factor(train$Pclass)
train$Name <- factor(train$Name)
train$Sex <- factor(train$Sex)
train$Ticket <- factor(train$Ticket)
train$Embarked <- factor(train$Embarked)
train$Pclass<-factor(train$Pclass)

test$Pclass <- factor(test$Pclass)
test$Name <- factor(test$Name)
test$Sex <- factor(test$Sex)
```

```r
test$Ticket <- factor(test$Ticket)
test$Embarked <- factor(test$Embarked)
test$Pclass<-factor(test$Pclass)
```

```r
set.seed(1000)

# Impute the NA values with mice function
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.2.3
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.2.3
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
# focus on the numerical columns in our coding
train_temp <- train[,c(-1,-4,-5,-11,-12)]
test_temp <-test[,c(1,3,4,5,7)]

# impute with mice variable
embarked <- mice(train_temp,m=5,maxit=5,meth='pmm',seed=10) ## mice is only numerical variables
```

```
##
##  iter imp variable
##   1   1  Age
##   1   2  Age
##   1   3  Age
##   1   4  Age
##   1   5  Age
##   2   1  Age
##   2   2  Age
##   2   3  Age
##   2   4  Age
##   2   5  Age
##   3   1  Age
##   3   2  Age
##   3   3  Age
##   3   4  Age
##   3   5  Age
##   4   1  Age
##   4   2  Age
##   4   3  Age
##   4   4  Age
##   4   5  Age
##   5   1  Age
##   5   2  Age
##   5   3  Age
##   5   4  Age
##   5   5  Age
```

```
## Warning: Number of logged events: 25
```

```
embarked2 <- mice(test_temp,m=5,maxit=5,meth='pmm',seed=10)
```

```
## 
##  iter imp variable
##   1   1  Age
##   1   2  Age
##   1   3  Age
##   1   4  Age
##   1   5  Age
##   2   1  Age
##   2   2  Age
##   2   3  Age
##   2   4  Age
##   2   5  Age
##   3   1  Age
##   3   2  Age
##   3   3  Age
##   3   4  Age
##   3   5  Age
##   4   1  Age
##   4   2  Age
##   4   3  Age
##   4   4  Age
##   4   5  Age
##   5   1  Age
##   5   2  Age
##   5   3  Age
##   5   4  Age
##   5   5  Age
```

```
## Warning: Number of logged events: 30
```

```
train_temp <- complete(embarked,5)
test_temp<-complete(embarked2,5)
```

```r
train$Age <- train_temp$Age
test$Age<-test_temp$Age

train <- train[,c(-1,-4,-9,-11,-12)]
test <- test[,c(-1,-3,-8,-10,-11)]

# determine important variables from our mode
model <- randomForest(Survived~.,data=train)
importance(model) #sex, fare, and age are important predictors
```

```
##        MeanDecreaseGini
## Pclass         35.50017
## Sex           106.96703
## Age            57.72841
## SibSp          16.91558
## Parch          13.33767
## Fare           69.09435
```

```r
library(MASS)
```

```
## Warning: package ’MASS’ was built under R version 4.2.2
```

```r
library(ISLR)
```

```
## Warning: package ’ISLR’ was built under R version 4.2.3
```

```r
set.seed(10001)

## utilize llogistic regression for function
glm_model <- glm(factor(Survived)~.,data=train,family=binomial())
glm_predict=predict(glm_model,newdata=test,type="response")

pred_test <- rep(0,418)
pred_test[glm_predict>0.5]=1

write.csv(data.frame(PassengerID=892:1309,Survived=pred_test),
  "C:\\Users\\natha\\OneDrive\\Documents\\survivalprediction.csv",row.names = FALSE)
```