Titanic Classification Models

Aaron Banlao

```
library(pacman)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
p_load(titanic, Amelia, naniar, DataExplorer, tidyverse, janitor, tidymodels, yardstick, randomForest, c
```

Data

```
library(titanic)

data(titanic_train)
data(titanic_test)

head(titanic_train)
head(titanic_test)
```

```
#create_report(titanic_train, y = "Survived")
```

Tidying the data

```
titanic_train <- titanic_train %>%
  mutate(Sex = as.factor(Sex),
         Survived = as.factor(Survived),
         Pclass = as.factor(Pclass),
         Embarked = as.factor(Embarked))

titanic_test <- titanic_test %>%
```

```
  mutate(Pclass = as.factor(Pclass),
         Sex = as.factor(Sex),
         Embarked = as.factor(Embarked))

head(titanic_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 Parch
## 1                             Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                              Heikkinen, Miss. Laina female  26     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                            Allen, Mr. William Henry   male  35     0     0
## 6                                    Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
head(titanic_test)
```

```
##   PassengerId Pclass                                          Name    Sex  Age
## 1         892      3                              Kelly, Mr. James   male 34.5
## 2         893      3              Wilkes, Mrs. James (Ellen Needs) female 47.0
## 3         894      2                     Myles, Mr. Thomas Francis   male 62.0
## 4         895      3                             Wirz, Mr. Albert   male 27.0
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
## 6         897      3                   Svensson, Mr. Johan Cervin   male 14.0
##   SibSp Parch  Ticket    Fare Cabin Embarked
## 1     0     0  330911  7.8292              Q
## 2     1     0  363272  7.0000              S
## 3     0     0  240276  9.6875              Q
## 4     0     0  315154  8.6625              S
## 5     1     1 3101298 12.2875              S
## 6     0     0    7538  9.2250              S
```

```
summary_null <- data.frame(missing = sapply(titanic_train, function(x) sum(is.na(x))))
print(summary_null)
```

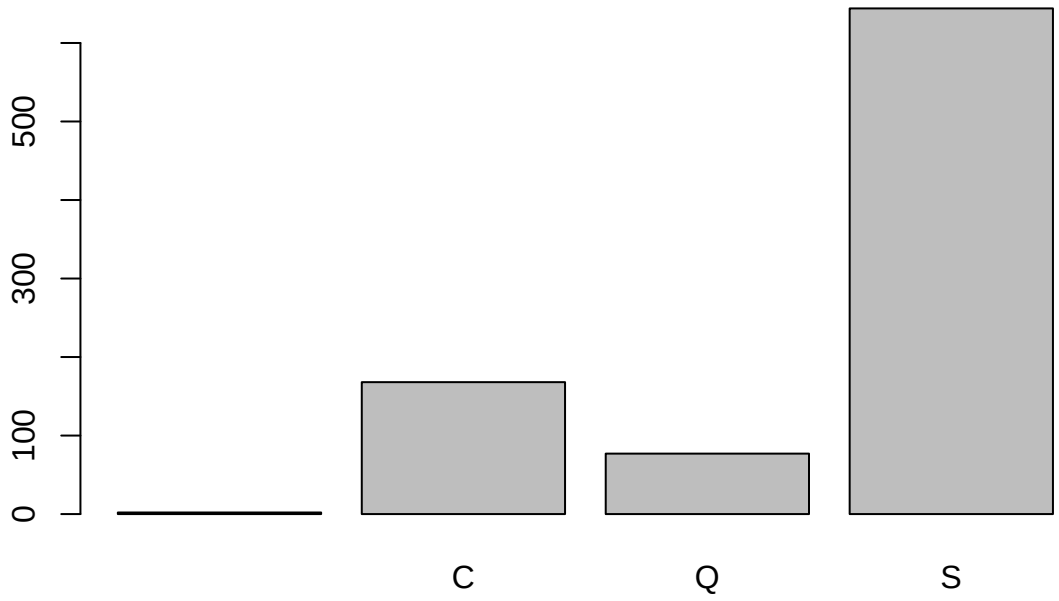
```
##             missing
## PassengerId       0
## Survived          0
## Pclass            0
```

```
## Name              0
## Sex               0
## Age             177
## SibSp             0
## Parch             0
## Ticket            0
## Fare              0
## Cabin             0
## Embarked          0
```

```
sum(nzchar(titanic_train$Cabin))
```

```
## [1] 204
```

```
barplot(table(titanic_train$Embarked))
```



Selecting relevant variables to model

```
titanic_train <- titanic_train %>%
  dplyr::select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked)

head(titanic_train)
```

```
##   Survived Pclass    Sex Age SibSp Parch    Fare Embarked
## 1        0      3   male  22     1     0  7.2500        S
## 2        1      1 female  38     1     0 71.2833        C
## 3        1      3 female  26     0     0  7.9250        S
## 4        1      1 female  35     1     0 53.1000        S
## 5        0      3   male  35     0     0  8.0500        S
## 6        0      3   male  NA     0     0  8.4583        Q
```

```
n <- nrow(titanic_train)
titanic_train_split <- titanic_train %>%
  initial_split(prop = 0.8)

titanic_train_split %>%
  training() %>%
  head()
```

```
##     Survived Pclass    Sex Age SibSp Parch     Fare Embarked
## 888        1      1 female  19     0     0  30.0000        S
## 885        0      3   male  25     0     0   7.0500        S
## 714        0      3   male  29     0     0   9.4833        S
## 660        0      1   male  58     0     2 113.2750        C
## 321        0      3   male  22     0     0   7.2500        S
## 835        0      3   male  18     0     0   8.3000        S
```

```
titanic_train_recipe <- training(titanic_train_split) %>%
  recipe(Survived ~ .) %>%
  step_rm(Pclass, Sex, Embarked) %>%
  step_nzv(all_predictors()) %>%
  step_impute_mean(Age) %>%
  prep()
```

```
titanic_train_test <- titanic_train_recipe %>%
  bake(testing(titanic_train_split))

titanic_train_test
```

```
## # A tibble: 179 x 5
##      Age SibSp Parch  Fare Survived
##    <dbl> <int> <int> <dbl> <fct>
##  1  26       0     0  7.92 1
##  2  35       0     0  8.05 0
##  3   2       3     1 21.1  0
##  4   2       4     1 29.1  0
##  5  29.8     0     0 13    1
##  6  35       0     0 26    0
##  7  29.8     0     0  7.88 1
##  8  40       0     0 27.7  0
##  9  14       1     0 11.2  1
## 10  40       1     0  9.48 0
## # ... with 169 more rows
```

```
titanic_train_training <- juice(titanic_train_recipe)

titanic_train_training
```

```
## # A tibble: 712 x 5
##      Age SibSp Parch   Fare Survived
##    <dbl> <int> <int>  <dbl> <fct>
##  1    19     0     0  30    1
##  2    25     0     0   7.05 0
##  3    29     0     0   9.48 0
##  4    58     0     2 113.   0
##  5    22     0     0   7.25 0
##  6    18     0     0   8.3  0
##  7    28     0     0  13.5  0
##  8    14     5     2  46.9  0
##  9    37     1     0  26    0
## 10    28     0     0  10.5  0
## # ... with 702 more rows
```

Creating the Models

Null Model

```
titanic_train_training %>%
  count(Survived) %>%
  mutate(pct = n / sum(n))
```

```
## # A tibble: 2 x 3
##   Survived     n   pct
##   <fct>    <int> <dbl>
## 1 0          448 0.629
## 2 1          264 0.371
```

```
titanic_mod_null <- logistic_reg(mode = "classification") %>%
  set_engine("glm") %>%
  fit(Survived ~ 1, data = titanic_train_training)
```

```
pred <- titanic_train_training %>%
  dplyr::select(Survived, Age, SibSp, Parch, Fare) %>%
  bind_cols(
    predict(titanic_mod_null, new_data = titanic_train_training, type = "class")
  ) %>%
  rename(survived_null = .pred_class)

accuracy(pred, Survived, survived_null)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.629
```

```
confusion_null <- pred %>%
  conf_mat(truth = Survived, estimate = survived_null)

confusion_null
```

```
##           Truth
## Prediction   0   1
##          0 448 264
##          1   0   0
```

KNN

```
titanic_mod_knn <- nearest_neighbor(mode = "classification", neighbors = 11) %>%
  set_engine("kknn") %>%
  fit(Survived ~ ., data = titanic_train_training)
```

```
titanic_mod_knn %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  accuracy(truth = Survived, estimate = .pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.654
```

```
titanic_mod_knn %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  conf_mat(truth = Survived, estimate = .pred_class)
```

```
##           Truth
## Prediction  0  1
##          0 80 41
##          1 21 37
```

Random Forest

```
titanic_train_forest <- rand_forest(
  mode = "classification",
  mtry = 4,
  trees = 300
) %>%
  set_engine("randomForest") %>%
  fit(Survived ~ ., data = titanic_train_training)
```

```
titanic_train_forest %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  accuracy(truth = Survived, estimate = .pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.609
```

```
titanic_train_forest %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  conf_mat(truth = Survived, estimate = .pred_class)
```

```
##            Truth
## Prediction  0  1
##          0 75 44
##          1 26 34
```

Naive Bayes

```
titanic_train_nb <- naive_Bayes(mode = "classification") %>%
  set_engine("klaR") %>%
  fit(Survived ~ ., data = titanic_train_training)
```

```
titanic_train_nb %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  accuracy(truth = Survived, estimate = .pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.631
```

```
titanic_train_nb %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  conf_mat(truth = Survived, estimate = .pred_class)
```

```
##            Truth
## Prediction  0  1
##          0 84 49
##          1 17 29
```

Logistic Regression using Regularlization

```
titanic_train_glm <- logistic_reg(mode = "classification", penalty = 0.001, mixture = 0.5) %>%
  set_engine("glmnet") %>%
  fit(Survived ~ ., data = titanic_train_training)
```

```
titanic_train_glm %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  accuracy(truth = Survived, estimate = .pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.682
```

```
titanic_train_glm %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  conf_mat(truth = Survived, estimate = .pred_class)
```

```
##           Truth
## Prediction   0   1
##          0 100  56
##          1   1  22
```

XGBoost

```
titanic_train_xgb <- boost_tree(mode = "classification",trees = 20) %>%
  set_engine("xgboost") %>%
  fit(Survived ~ ., data = titanic_train_training)
```

```
titanic_train_xgb %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  accuracy(truth = Survived, estimate = .pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.682
```

```
titanic_train_xgb %>%
  predict(titanic_train_test) %>%
  bind_cols(titanic_train_test) %>%
  conf_mat(truth = Survived, estimate = .pred_class)
```

```
##           Truth
## Prediction  0  1
##          0 88 44
##          1 13 34
```

Out of all the models presented, it seems like the XGboost has the best performance. So this is what model we will run on the full titanic_train dataset.

```
titanic_train_xgb2 <- boost_tree(mode = "classification",trees = 20) %>%
  set_engine("xgboost") %>%
  fit(Survived ~ ., data = titanic_train)
```

```
prediction <- predict(titanic_train_xgb2, titanic_test)

solution <- data.frame(PassengerID = titanic_test$PassengerId, Survived = prediction)

write.csv(solution, file = "titanic_prediction.csv", row.names = F)
```