Homework4

Yingshan Li (7937790)

April 26, 2022

Load the data

```
titanic <- read.csv(file = "titanic.csv" )
titanic1 <- titanic %>% mutate(survived = factor(survived, levels = c("Yes", "No"))) %>%
  mutate(pclass = factor(pclass))
```

```
set.seed(3435)
```

Question 1

```
titanic_split <- initial_split(titanic1, prop = 0.70, strata = survived)
titanic_train <- training(titanic_split)
titanic_test <- testing(titanic_split)
```

Verify correct number of observations in each data set

```
dim(titanic_train)
```

```
## [1] 623  12
```

```
dim(titanic_test)
```

```
## [1] 268  12
```

Create Recipe same as HW3

```
titanic_recipe <- recipe(survived ~ pclass + sex + age + sib_sp + parch + fare, data = titanic_train) %
  step_impute_linear(age) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(~ starts_with("sex"):fare + age:fare)
```

Question 2

Fold the training data

```
titanic_folds <- vfold_cv(titanic_train, k = 10)
titanic_folds
```

```
## #  10-fold cross-validation
## # A tibble: 10 x 2
##    splits           id
##    <list>           <chr>
##  1 <split [560/63]> Fold01
##  2 <split [560/63]> Fold02
##  3 <split [560/63]> Fold03
##  4 <split [561/62]> Fold04
##  5 <split [561/62]> Fold05
##  6 <split [561/62]> Fold06
```

```
## 7 <split [561/62]> Fold07
## 8 <split [561/62]> Fold08
## 9 <split [561/62]> Fold09
## 10 <split [561/62]> Fold10
```

Question 3

k-fold cross-validation is a resampling method. The training data are randomly partitioned into specified sets of roughly equal size for which we called each set the folds. For example, for 10-fold cross validation, for each iterations of resampling, one fold is held out as assessment set to evaluate the model , and all the 9 remaining folds are used as analysis set to fit the model. The final resampling estimate of model performance is the averages of each of the iteration. It is a better model evaluation method because simply fitting and testing models on the training set will result in an artificially optimistic estimate of the performance since the model is built based on the training data set. If we use the entire training set, the resampling method would be the validation set approach.

Question 4

Logistic Regression

```
log_reg <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")

log_wkflow <- workflow() %>%
  add_model(log_reg) %>%
  add_recipe(titanic_recipe)
```

Linear discriminant analysis

```
lda_mod <- discrim_linear() %>%
  set_mode("classification") %>%
  set_engine("MASS")

lda_wkflow <- workflow() %>%
  add_model(lda_mod) %>%
  add_recipe(titanic_recipe)
```

Quadratic discriminant analysis

```
qda_mod <- discrim_quad() %>%
  set_mode("classification") %>%
  set_engine("MASS")

qda_wkflow <- workflow() %>%
  add_model(qda_mod) %>%
  add_recipe(titanic_recipe)
```

In total, I will be fitting 30 models, 10 for each type of model because fitting 1 times for each of 10 folds.

Question 5

```
log_res <- log_wkflow %>%
  fit_resamples(resamples = titanic_folds)
```

```
lda_res <- lda_wkflow %>%
  fit_resamples(resamples = titanic_folds)
```

```r
qda_res <- qda_wkflow %>%
  fit_resamples(resamples = titanic_folds)
```

```r
log_acc <- collect_metrics(log_res)
log_acc
```

```
## # A tibble: 2 x 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.809    10  0.0156 Preprocessor1_Model1
## 2 roc_auc  binary     0.836    10  0.0161 Preprocessor1_Model1
```

```r
lda_acc <- collect_metrics(lda_res)
lda_acc
```

```
## # A tibble: 2 x 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.787    10  0.0172 Preprocessor1_Model1
## 2 roc_auc  binary     0.837    10  0.0157 Preprocessor1_Model1
```

```r
qda_acc <- collect_metrics(qda_res)
qda_acc
```

```
## # A tibble: 2 x 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.775    10  0.0171 Preprocessor1_Model1
## 2 roc_auc  binary     0.838    10  0.0137 Preprocessor1_Model1
```

```r
mean_accuracy <- c(log_acc$mean[1],lda_acc$mean[1], qda_acc$mean[1])
Standard_error <- c(log_acc$std_err[1],lda_acc$std_err[1], qda_acc$std_err[1])
models <- c("Logistic Regression", "LDA", "QDA")
results <- tibble(accuracies = mean_accuracy, Standard_error = Standard_error, models = models)
results %>%
  arrange(-accuracies)
```

```
## # A tibble: 3 x 3
##   accuracies Standard_error models
##        <dbl>          <dbl> <chr>
## 1      0.809         0.0156 Logistic Regression
## 2      0.787         0.0172 LDA
## 3      0.775         0.0171 QDA
```

The logistic regression model performs the best because it has the highest mean accuracy, and its standard error is only a little bit larger than the QDA model that has the smallest standard error.

Question 7

```r
log_fit <- fit(log_wkflow, titanic_train)
log_fit %>%
  tidy()
```

```
## # A tibble: 10 x 5
##    term             estimate std.error statistic  p.value
##    <chr>               <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)      -4.40      0.683       -6.45  1.13e-10
##  2 age               0.0629    0.0136       4.62  3.77e- 6
```

```
##  3 sib_sp           0.437      0.132        3.30  9.52e- 4
##  4 parch            0.151      0.153        0.989 3.23e- 1
##  5 fare            -0.00116    0.0107      -0.108 9.14e- 1
##  6 pclass_X2        1.25       0.363        3.46  5.48e- 4
##  7 pclass_X3        2.44       0.382        6.39  1.62e-10
##  8 sex_male         2.15       0.303        7.09  1.32e-12
##  9 sex_male_x_fare  0.0139     0.00836      1.66  9.65e- 2
## 10 fare_x_age      -0.000360   0.000206    -1.75  8.03e- 2
```

Question 8

```r
predict(log_fit, new_data = titanic_test, type = "prob")
```

```
## # A tibble: 268 x 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.933    0.0671
##  2    0.924    0.0763
##  3    0.119    0.881
##  4    0.183    0.817
##  5    0.230    0.770
##  6    0.239    0.761
##  7    0.120    0.880
##  8    0.119    0.881
##  9    0.0456   0.954
## 10    0.174    0.826
## # ... with 258 more rows
```

```r
augment(log_fit, new_data = titanic_test) %>%
  accuracy(truth = survived, estimate = .pred_class)
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.832
```

Model's testing accuracy is slightly higher than the average accuracy across folds.