Homework 4

Kabir Snell

2022-11-07

Homework 4

```
# Loading Libraries
library(tidymodels)
library(discrim)
library(poissonreg)
library(corrr)
library(klaR)
library(resample)
library(ISLR)
library(ISLR2)
tidymodels_prefer()

set.seed(727)
```

```
titanic <- read.csv("data/titanic.csv")

titanic$survived <- factor(titanic$survived, levels= c("Yes", "No"))
titanic$pclass <- factor(titanic$pclass)
```

Question 1

```
# splitting the data
titanic_split <- initial_split(titanic, prop = .8, strata = survived)
titanic_train <- training(titanic_split)
titanic_test <- testing(titanic_split)

titanic_split
```

```
## <Training/Testing/Total>
## <712/179/891>
```

Question 2

```
# Folding the training data
# I had to run this line in a different file for some reason
titanic_folds <- vfold_cv(titanic_train, v = 10)
titanic_folds
```

```
## #  10-fold cross-validation
## # A tibble: 10 x 2
##    splits           id
##    <list>           <chr>
##  1 <split [640/72]> Fold01
##  2 <split [640/72]> Fold02
##  3 <split [641/71]> Fold03
##  4 <split [641/71]> Fold04
##  5 <split [641/71]> Fold05
##  6 <split [641/71]> Fold06
##  7 <split [641/71]> Fold07
##  8 <split [641/71]> Fold08
##  9 <split [641/71]> Fold09
## 10 <split [641/71]> Fold10
```

Question 3

We split the training data into 10 different groups. We do this in order to create subgroups of testing and training data. We then fit models to each of the subgroups. K-fold cross-validation takes out the first group (fold) to find the MSE for the rest of the groups/fold. This is repeated so that each fold is used as a validation set. The cross validation becomes the average of each of the MSEs found. This allows us to find a much more consistent model. If we were to use the entire training set instead, we would use the validation set approach. This would allow us to get better results if using the entire training set.

Question 4

```
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(terms = ~ starts_with("sex"):fare) %>%
  step_interact(~age:fare)
```

```
# Logistic regression model
log_mod <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")

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

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

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

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

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

Since there are 10 folds and 3 models. We will be fitting a total of 30 models to the data

Question 5

```
titanic_log_fit <- fit_resamples(log_wkflow, titanic_folds, metrics = metric_set(accuracy))

titanic_lda_fit <- fit_resamples(lda_wkflow, titanic_folds, metrics = metric_set(accuracy))

titanic_qda_fit <- fit_resamples(qda_wkflow, titanic_folds, metrics = metric_set(accuracy))
```

Question 6

```
log_metrics <- collect_metrics(titanic_log_fit)
log_average = tibble(sum(log_metrics$mean), sum(log_metrics$std_err))
log_average
```

```
## # A tibble: 1 x 2
##   `sum(log_metrics$mean)` `sum(log_metrics$std_err)`
##                     <dbl>                      <dbl>
## 1                   0.802                     0.0122
```

```
lda_metrics <- collect_metrics(titanic_lda_fit)
lda_average = tibble(sum(lda_metrics$mean), sum(lda_metrics$std_err))
lda_average
```

```
## # A tibble: 1 x 2
##   `sum(lda_metrics$mean)` `sum(lda_metrics$std_err)`
##                     <dbl>                      <dbl>
## 1                   0.788                     0.0148
```

```
qda_metrics <- collect_metrics(titanic_qda_fit)
qda_average = tibble(sum(qda_metrics$mean), sum(qda_metrics$std_err))
qda_average
```

```
## # A tibble: 1 x 2
##   `sum(qda_metrics$mean)` `sum(qda_metrics$std_err)`
##                     <dbl>                      <dbl>
## 1                   0.770                     0.0122
```

The logistic regression model has performed the best. We can prove this because:

```
mean(Logistic) - std_err > mean(LDA) + std_err
```

```
mean(Logistic) - std_err > mean(QDA) + std_err
```

These equations show us that the lower bound range value for the logistic model is still bigger than the upper bound range value for both LDA and QDA models

Question 7

```
final_fit <- fit(log_wkflow, titanic_train)
```

Question 8

```
results <- bind_cols(titanic_test$survived, c(predict(final_fit, titanic_test)))
```

```
## New names:
## * `` -> `...1`
```

```
log_acc <- accuracy(results, truth = ...1, estimate = .pred_class)
log_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.838
```