Course › Section 5: Classifica... › 5.3: Titanic Exercises › Titanic Exercises, p...

Titanic Exercises, part 2

Question 7: Survival by fare - LDA and QDA

2.0/2.0 points (graded)

Train a model using linear discriminant analysis (LDA) with the `caret` `lda` method using fare as the only predictor.

What is the accuracy on the test set for the LDA model?

0.693 ✓ **Answer:** 0.659 **or** 0.693

0.693

Explanation

The accuracy can be determined using the following code:

```
train_lda <- train(Survived ~ Fare, method = "lda", data = train_set)
lda_preds <- predict(train_lda, test_set)
mean(lda_preds == test_set$Survived)
```

Train a model using quadratic discriminant analysis (QDA) with the `caret` `qda` method using fare as the only predictor.

What is the accuracy on the test set for the QDA model?

0.693 ✓ **Answer:** 0.659 **or** 0.693

0.693

Explanation

The accuracy can be determined using the following code:

```
train_qda <- train(Survived ~ Fare, method = "qda", data = train_set)
qda_preds <- predict(train_qda, test_set)
mean(qda_preds == test_set$Survived)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 8: Logistic regression models

3.0/3.0 points (graded)

Train a logistic regression model with the `caret` `glm` method using age as the only predictor.

What is the accuracy on the test set using age as the only predictor?

0.615

✓ **Answer:** 0.615

0.615

Explanation

The accuracy can be determined using the following code:

```
train_glm_age <- train(Survived ~ Age, method = "glm", data = train_set)
glm_preds_age <- predict(train_glm_age, test_set)
mean(glm_preds_age == test_set$Survived)
```

Train a logistic regression model with the `caret` `glm` method using four predictors: sex, class, fare, and age.

What is the accuracy on the test set using these four predictors?

0.849

✓ **Answer:** 0.821 **or** 0.849

0.849

Explanation

The accuracy can be determined using the following code:

```
train_glm <- train(Survived ~ Sex + Pclass + Fare + Age, method = "glm", data =
train_set)
glm_preds <- predict(train_glm, test_set)
mean(glm_preds == test_set$Survived)
```

Train a logistic regression model with the `caret` `glm` method using all predictors. Ignore warnings about rank-deficient fit.

What is the accuracy on the test set using all predictors?

0.849

✓ **Answer:** 0.827 **or** 0.849

0.849

Explanation
The accuracy can be determined using the following code:

```
train_glm_all <- train(Survived ~ ., method = "glm", data = train_set)
glm_all_preds <- predict(train_glm_all, test_set)
mean(glm_all_preds == test_set$Survived)
```

Submit You have used 1 of 10 attempts

ℹ Answers are displayed within the problem

Question 9a: kNN model

1.0/1.0 point (graded)
Set the seed to 6. Train a kNN model on the training set using `caret`. Try tuning with `k = seq(3, 51, 2)`.

What is the optimal value of the number of neighbors `k`?

11 ✓ **Answer:** 15 **or** 11

11

Explanation
The optimal value can be calculated using the following code:

```
#set.seed(6)
set.seed(6, sample.kind = "Rounding")    # simulate R 3.5
train_knn <- train(Survived ~ .,
                   method = "knn",
                   data = train_set,
                   tuneGrid = data.frame(k = seq(3, 51, 2)))
train_knn$bestTune
```

Submit You have used 1 of 10 attempts

ℹ Answers are displayed within the problem

Question 9b: kNN model

1.0/1.0 point (graded)

Plot the kNN model to investigate the relationship between the number of neighbors and accuracy on the training set.

Of these values of k, which yields the highest accuracy?

○ 7

● 11

○ 17

○ 21

✓

Explanation

The plot can be generated using the following code:

```
ggplot(train_knn)
```

Submit

You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Question 9c: kNN model

1.0/1.0 point (graded)

What is the accuracy of the kNN model on the test set?

0.709 ✓ **Answer:** 0.732 **or** 0.709

0.709

Explanation

The accuracy can be calculated using the following code:

```
knn_preds <- predict(train_knn, test_set)
mean(knn_preds == test_set$Survived)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

Question 10: Cross-validation

2.0/2.0 points (graded)

Set the seed to 8 and train a new kNN model. Instead of the default training control, use 10-fold cross-validation where each partition consists of 10% of the total.

What is the optimal value of `k` using cross-validation?

5

✓ **Answer:** 23 **or** 5

5

Explanation

The optimal value of `k` can be found using the following code:

```
#set.seed(8)
set.seed(8, sample.kind = "Rounding")    # simulate R 3.5
train_knn_cv <- train(Survived ~ .,
                      method = "knn",
                      data = train_set,
                      tuneGrid = data.frame(k = seq(3, 51, 2)),
                      trControl = trainControl(method = "cv", number = 10, p = 0.9))
train_knn_cv$bestTune
```

What is the accuracy on the test set using the cross-validated kNN model?

0.648

✓ **Answer:** 0.737 **or** 0.648

0.648

Explanation

The accuracy can be calculated using the following code:

```
knn_cv_preds <- predict(train_knn_cv, test_set)
mean(knn_cv_preds == test_set$Survived)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

Question 11a: Classification tree model

2.0/2.0 points (graded)

Set the seed to 10. Use `caret` to train a decision tree with the `rpart` method. Tune the complexity parameter with `cp = seq(0, 0.05, 0.002)`.

What is the optimal value of the complexity parameter (`cp`)?

0.016

✓ **Answer:** 0.02 **or** 0.016

0.016

Explanation

The optimal value of `cp` can be found using the following code:

```
#set.seed(10)
set.seed(10, sample.kind = "Rounding")    # simulate R 3.5
train_rpart <- train(Survived ~ .,
                     method = "rpart",
                     tuneGrid = data.frame(cp = seq(0, 0.05, 0.002)),
                     data = train_set)
train_rpart$bestTune
```

What is the accuracy of the decision tree model on the test set?

0.838

✓ **Answer:** 0.849 **or** 0.838

0.838

Explanation

The accuracy can be calculated using the following code:

```
rpart_preds <- predict(train_rpart, test_set)
mean(rpart_preds == test_set$Survived)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

Question 11b: Classification tree model

1.0/1.0 point (graded)

Inspect the final model and plot the decision tree.

Which variables are used in the decision tree?

Select ALL that apply.

☐ Survived

☑ Sex

☑ Pclass

☑ Age

☑ Fare

☐ Parch

☐ Embarked

✓

Explanation
The decision tree can be made using the following code:

```
train_rpart$finalModel # inspect final model

# make plot of decision tree
plot(train_rpart$finalModel, margin = 0.1)
text(train_rpart$finalModel)
```

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Question 11c: Classification tree model

2.0/2.0 points (graded)
Using the decision rules generated by the final model, predict whether the following individuals would survive.

A 28-year-old male

would NOT survive ✓ **Answer:** would NOT survive

A female in the second passenger class

would survive ✓ **Answer:** would survive

A third-class female who paid a fare of $8

would survive ✓ **Answer:** would survive

A 5-year-old male with 4 siblings

would NOT survive ✓ **Answer:** would NOT survive

A third-class female who paid a fare of $25

would NOT survive ✓ **Answer:** would NOT survive

A first-class 17-year-old female with 2 siblings

would survive ✓ **Answer:** would survive

A first-class 17-year-old male with 2 siblings

would NOT survive ✓ **Answer:** would NOT survive

Explanation

For each case, follow the decision tree to determine whether it results in survived=0 (didn't survive) or survived=1 (did survive).

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 12: Random forest model

3.0/3.0 points (graded)

Set the seed to 14. Use the `caret` `train` function with the `rf` method to train a random forest. Test values of mtry ranging from 1 to 7. Set `ntree` to 100.

What mtry value maximizes accuracy?

2 ✓ **Answer:** 3 **or** 2

2

Explanation

The mtry value can be calculated using the following code:

```
#set.seed(14)
set.seed(14, sample.kind = "Rounding")    # simulate R 3.5
train_rf <- train(Survived ~ .,
                  data = train_set,
                  method = "rf",
                  ntree = 100,
                  tuneGrid = data.frame(mtry = seq(1:7)))
train_rf$bestTune
```

What is the accuracy of the random forest model on the test set?

0.844

✓ **Answer:** 0.877 **or** 0.844

0.844

Explanation
The accuracy can be calculated using the following code:

```
rf_preds <- predict(train_rf, test_set)
mean(rf_preds == test_set$Survived)
```

Use `varImp` on the random forest model object to determine the importance of various predictors to the random forest model.

What is the most important variable?
Be sure to report the variable name exactly as it appears in the code.

Sexmale

✓ **Answer:** Sexmale

Explanation
The most important variable can be found using the following code:

```
varImp(train_rf)    # first row
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

Ask your questions or make your comments about Titanic Exercises, part 2 here! **Remember, one of the best ways to reinforce your own learning is by explaining something to someone else, so we encourage you to answer each other's questions (without giving away the answers, of course).**

Some reminders:

- Search the discussion board before posting to see if someone else has asked the same thing before asking a new question.
- Please be specific in the title and body of your post regarding which question you're asking about to facilitate answering your question.
- Posting snippets of code is okay, but posting full code solutions is not.
- If you do post snippets of code, please format it as code for readability. If you're not sure how to do this, there are instructions in a pinned post in the "general" discussion forum.

Discussion: Titanic Exercises, part 2

Show Discussion

Topic: Section 5: Classification with more than two classes and the caret package / 5.3.2: Titanic Exercises, part 2

© All Rights Reserved