R Lab #5 – Solutions

1. Work through the lab sections 4.6.1 and 4.6.2 (logistic regression) on pages 154-161. For the logistic regression model fit to just Lag1 and Lag2 on page 160, which, if any, of the two predictors are significant at the 5% level?

a) Lag1 only
b) Lag2 only
c) Neither
d) Both

Solution: C

```
glm.fit <- glm(Direction ~ Lag1 + Lag2, data = Smarket, family = binomial, subset = train)
summary(glm.fit)
coef(glm.fit) # get the coefficients of the fitted model
summary(glm.fit)$coef[,4] # get the p-values for the coefficients
```

2. What is the test set sensitivity (in %) for the logistic regression model fit to just Lag1 and Lag2 on page 160? (Round your answer to the nearest whole number.)

Solution: 75

```
106/(35+106)  # 0.752 = 75%
```

3. Work through the lab section 4.6.3 (LDA) on pages 161-162. What is the test set specificity (in %) for the LDA model fit to just Lag1 and Lag2 on page 161? (Round your answer to the nearest whole number.)

Solution: 32

```
35/(35+76)  # 0.315 = 32%
```

4. Confirm that the greatest posterior probability of decrease in 2005 for the LDA model fit to just Lag1 and Lag2 on page 161-2 is 52%. What is the lowest posterior probability of decrease in 2005 (in %) for this model? (Round your answer to the nearest whole number.)

Solution: 46

```
max(lda.pred$posterior[,1]) # = 0.5202 = 52%
min(lda.pred$posterior[,1]) # = 0.4578 = 46%
```

5. Work through the lab section 4.6.4 (QDA) on pages 162-163 and review Table 4.7. What is the test set precision (in %) for the QDA model fit to just Lag1 and Lag2 on page 163? (Round your answer to the nearest whole number.)

Solution: 60

```
121 / (121 + 81)  # 0.599 = 60%
```

6. Work through the lab section 4.6.5 (KNN) on pages 163-164. In the lab, the command set.seed(1) is entered immediately before fitting the KNN model with K=1. However, this command is not entered immediately before fitting the KNN model with K=3. Fit the KNN model with K=3 again, but this time enter the command set.seed(1) immediately before fitting the model. True or False? The confusion matrix that results is exactly the same as the one reported in the book for K=3.

Solution: False

```
set.seed(1)
knn.pred <- knn(train.X, test.X, train.Direction, k = 3)
table(knn.pred, Direction.2005) # determines the confusion matrix
#         Direction.2005
# knn.pred Down Up
#     Down   48 55
#     Up     63 86
```

7. Work through the lab section 4.6.6 (Caravan Insurance Data) on pages 164-167. Then enter the following code to fit an LDA model to the data:

```
lda.fit=lda(Purchase~.,data=Caravan,subset=-test)
lda.probs=predict(lda.fit, Caravan[test,])$posterior[,2]
lda.pred=rep("No",1000)
lda.pred[lda.probs>.25]="Yes"
table(lda.pred,test.Y)
```

Match the resulting entries in the confusion matrix.

a) Actual purchasers predicted to purchase. – 13
b) Actual purchasers not predicted to purchase. – 46
c) Actual non-purchasers predicted to purchase. – 27
d) Actual non-purchasers not predicted to purchase. – 914

Solution:

```
lda.fit=lda(Purchase~.,data=Caravan,subset=-test)
lda.probs=predict(lda.fit, Caravan[test,])$posterior[,2]
lda.pred=rep("No",1000)
lda.pred[lda.probs>.25]="Yes"
table(lda.pred,test.Y)
#         test.Y
# lda.pred  No Yes
#      No  914  46
#      Yes  27  13
```

8. True or False? The LDA model fit to the Caravan Insurance data for Question 7 outperforms the logistic regression model in terms of the percentage of predicted purchasers who actual purchase insurance. (Use a 0.25 probability cut-off for both models.)

Solution: False

```
13/(27+13)  # 32.5% are correctly predicted "Yes" to purchasing insurance
```