**Chapter 4 – Classification**

5) Examine the difference b/t LDA and QDA

1. Expect LDA to perform better on the training and test sets
   1. Blog says QDA is better on the training set b/c it’s higher flexibility will yield a closer fit.
   2. LDA performs better on test set b/c QDA could overfit the linearity of Bayes decision boundary.
2. Expect QDA to perform better on the training and test tests
3. As the size of the dataset increases, we expect QDA’s performance to improve relative to LDA b/c more samples decreases the variance risk
4. False. If the dataset’s distribution significantly drifts from a normal then QDA will perform poorly.

6) Logistic regression – Probability

1. 37.8%
2. 50 hours

7) Probability of distributing dividend = 78.9% to 80%

8) We cannot conclude which model is better. A model’s accuracy/error is not always the best indicator of how the algorithm performs. The KNN may have a high variance, signifying that it capture lots of noise. If so, it prevents the model from being applied to the population/external model

9) Odds

a) 32.5%

b) Odds are 0.19

**APPLIED**

10) Weekly data

1. Lag2 is statistically significant, coefficient p-value = 0.011. Volume could be significant given its coefficient’s p-value is 0.034.

h) KNN performed the best

12) The KNN model outperforms logistic, LDA and QDA based on accuracy, precision and recall.

The independent variables seem to suffer from multicollinearity.

When interpreting the coefficients of the logistic regression, it seems that crime is greatest in the welather areas. I suspect this dataset is reported crime; consequently, wealthier individuals are more likely to call the police.

The Black population seems to have a negative correlation with crime. One hypothesis may be that Black view the police as a threat, not a form of safety.