**Homework 5**

Name: \_\_\_\_\_\_\_\_\_\_\_Tyler DiNapoli-Chiappelli\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

There is 20 regular points in this assignment. The minimum increment is 0.5 point. Solve them and fill the answers in the blank space.

1. Briefly explain in what situations AUC is a better predictive performance measure than accuracy and why it is better in those situations. (2 points)

AUC is a better predictive performance measure than accuracy when the dependent variable in the historical data contains a relatively small amount of events coded as “1”. In these situations, relying solely on measures of accuracy will result in naive models appearing as though they are the optimal models. However, while naive models might be statistically optimal, they will likely not be financially optimal, as a business following a naive model in this situation may miss out on profit opportunities. On the other hand, relying on AUC as the measure of predictive performance will guide business decisions more precisely due the explicit incorporation of the FPR and TPR into the overall model performance.

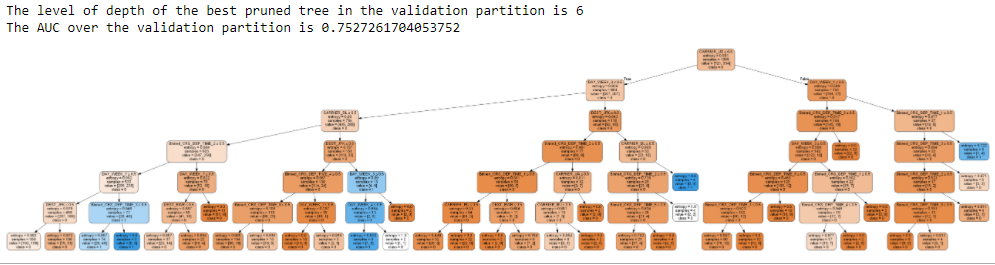
2. **Predicting Delayed Flights (cont.)**

Continue the case of predicting delayed flights from homework assignment 4. Use the same historical dataset. Allocate 20% of the data to the test partition. Set the number of folds for cross-validation to 5. If we need to drop redundant dummies during categorical variable coding, drop the dummy that corresponds to the mode value of each categorical predictor.

Classification tree:

1. Set the maximum level of depth for the best pruned tree to 8. Use cross-validation to select the classification tree based on AUC. Attach a screenshot of the best pruned tree. What is its level of depth? (2 point)

Below is a screenshot of the best pruned tree. It has level of depth = 6

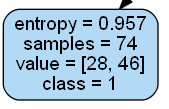


b. Does the tree in (a) max out all the levels of depth we allow it to have? If yes, please explain why it does. If no, please also explain why it does not. (2 point)

No, we allowed the tree to have up to 8 levels of depth and this tree maxed out at 6. This tree did not grow to its maximum allowed depth simply because the AUC in the validation partition was the smallest when the tree was pruned at level 6

c. Choose a leaf node on the tree in (a). Attach a screenshot of that leaf node. What is the predicted probability of being delayed if an observation reaches that particular leaf node? Show the calculation. (2 points)

If an observation reaches this particular node, the probability of the flight being delayed is 46 / 74 = 62.16%



d. Write down the top three effective English rules of the tree in (a). (2 points)

Please note: the first two rules are 100% effective and the third rule is 8 / 9 = 89% effective.

**1 (Leaf node ID 4, see Python output):**

- The carrier is US Airways or Delta Airlines

- The day of the week is Saturday

- The flight leaves between 8:00am and 10:00am

**2 (Leaf node ID 25, see Python output):**

- The carrier is US Airways

- The day of the week is not Saturday

- The destination is not JFK

**3 (Leaf node ID 12, see Python output):**

- The carrier is US Airways

- The destination is not JFK

- The day of the week is Wednesday, Thursday, or Saturday

1. Rank the English rules in (d) based on their significance. (1 points)

1: Leaf # 12 Rule (sample = 9)

2:Leaf # 25 Rule (sample = 4)

3: Leaf # 4 Rule (sample = 3)

1. If you want to choose one rule among all the rules in (d) to report, which one do you want to choose? Justify your choice (2 point)

I would elect to report Leaf Rule #12. Though it does not have the 100% effectiveness carried by the other two leaves, it does have the highest significance, in terms of sample size. The increased sample size, combined with the fact that the rule does not predict a 100% probability, likely makes it the most realistic rule to report.

1. Report the AUC of the tree in (a) over the test partition (1 point)

The AUC over the test partition is 0.7604818523153942

k-NN:

1. Use cross-validation to select the k-NN model based on AUC. What is the optimal k? (1 point)

The optimal k in the validation partition is 28

1. What is the potential risk that the model can run into when k = 1? (1 point)

If k ever equals 1, the model is vulnerable to using outliers to make predictions. If we are unlucky enough to have said outlier be the point of shortest distance away from the point we are trying to predict, the KNN model will end up using that outlier value as the sole basis of prediction.

1. Report the AUC of the k-NN model in (h) over the test partition. (1 point)

The AUC in the test partition is 0.6995306633291615

Logistic regression:

1. Use cross-validation to select the logistic regression model based on AUC. Report the optimal penalty level of the final selected model. (1 point)

The optimal alpha over the validation partition is 0.4013963963963964

1. Report the AUC of the final selected logistic regression model over the test partition. (1 point)

The AUC in the test partition is 0.7265957446808511

1. Cross compare the AUCs in (g), (j) and (l). Which predictive technique generates the model with the best performance over the test partition? (1 point)

The model which generates the highest performance over the test partition is the Classification Tree, with an AUC over the test partition of 0.7604818523153942.

Submit your Python code

* for the classification tree model with filename [DM2020] HW5\_CT\_YOURFULLNAME.ipynb
* for the k-NN model with filename [DM2020] HW5\_kNN\_YOURFULLNAME.ipynb
* for the logistic regression with filename [DM2020] HW5\_LOGIT\_YOURFULLNAME.ipynb