Homework 5 – Due 5/4 at 9 AM Eastern Time – Prepared by Michael Wacey

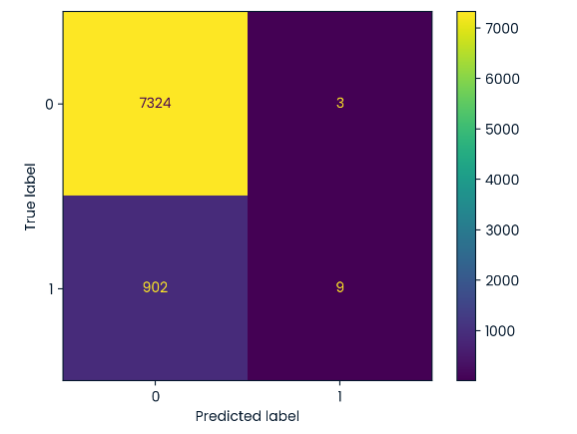
1. Complete the tutorial at https://www.datacamp.com/tutorial/random-forests-classifier-python and answer the following questions in paragraph form:
   1. How is the sample data split and what are some reasons why the test data and training data should not overlap?  
        
      This code is used to split the data into training and test data sets. The test data set will be about 20% of the data and the training data set will be the remaining data, about 80%. We do not want any overlap so that there is no information leakage from the training data. The goal is that we train the model on the training data and then test it on the test data. If there is overlap, the test data will not be a good test of our model.  
        
      # Split the data into training and test sets  
      X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2)
   2. What were your first three decision trees and what do they show?  
        
      Here are the first three decision trees:  
      Diagram

      Description automatically generated  
        
      Diagram

      Description automatically generated  
        
      Diagram

      Description automatically generated  
        
      The first tree selected cons.price.idx as the top variable while the other two selected cons.conf.idx. For the second two trees they used the same value to split on for the first two levels. Of the four variables available, none if the trees used ‘default’.  
        
      I changed the max depth to 6 and got the results below. There are many levels to these trees.  
        
      Chart, timeline

      Description automatically generated
   3. Compare the results of using large max\_depth decision trees to a using small max\_depth decision trees.   
        
      For this question, I retained the unconstrained max depth and added the depths of 3 for small and 30 for large. I then calculated the accuracy of each of them. Just looking at the accuracy numbers they did not seem to be very different. So, I would conclude that depth has li  
        
      I ran the hyper parameter tuning twice. The first time it had a depth of 3 but the second time it had a depth of 7. So, it seems that depth is not critical.  
      Graphical user interface, text

      Description automatically generated
   4. Show your confusion matrix and explain the results, including why you believe you got them.  
        
        
        
      The confusion matrix shows that of the 911 people who actually did subscribe, the model only predicts that 9 of them will. In addition, of the 911 people who will subscribe, it only predicts that11 will. This is not very useful. Based on this, I assume that the attributes age, default, consumer price index, and consumer confidence index are not good predictors of whether someone will subscribe. Maybe income or assets would be a better predictor.
   5. What does recall tell us in this problem, specifically with regards to bank’s direct marketing campaign?  
        
      From <https://builtin.com/data-science/precision-and-recall> the definition of recall:  
        
      Recall: The ability of a model to find all the relevant cases within a data set. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives.  
        
      In this case, that is 9 / (9 + 902) or 0.0099. This is a low value and indicates that the model is not good at finding relevant cases. In this case, it would not be good for the bank to use this model to select people to call.
2. Generate a random variable in excel with 100 independent experiments from the binomial distribution and plot the probability mass function. Use Excel’s Analysis ToolPak to generate the numbers. <https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel6a63e598-cd6d-42e3-9317-6b40ba1a66b4>  
     
   Steps:
   1. Create a random variable with 100 values.
   2. Fin the maximum and minimum values
   3. Create a table of values from the minimum to the maximum
   4. Count the number of occurrences of each value
   5. Calculate the percentage
   6. Plot the results.

The first 10 values of the random variable:

|  |
| --- |
| A |
| 51 |
| 52 |
| 53 |
| 48 |
| 56 |
| 48 |
| 56 |
| 54 |
| 52 |
| 52 |

The Minimum and maximum:

|  |  |
| --- | --- |
| Min A | 38 |
| Max A | 64 |

The distribution table:

|  |  |  |
| --- | --- | --- |
| A | Count | Percent |
| 38 | 1 | 1% |
| 39 | 1 | 1% |
| 40 | 1 | 1% |
| 41 | 1 | 1% |
| 42 | 2 | 2% |
| 43 | 2 | 2% |
| 44 | 1 | 1% |
| 45 | 4 | 4% |
| 46 | 9 | 9% |
| 47 | 2 | 2% |
| 48 | 12 | 12% |
| 49 | 10 | 10% |
| 50 | 11 | 11% |
| 51 | 8 | 8% |
| 52 | 4 | 4% |
| 53 | 4 | 4% |
| 54 | 4 | 4% |
| 55 | 4 | 4% |
| 56 | 5 | 5% |
| 57 | 2 | 2% |
| 58 | 2 | 2% |
| 59 | 4 | 4% |
| 60 | 0 | 0% |
| 61 | 3 | 3% |
| 62 | 2 | 2% |
| 63 | 0 | 0% |
| 64 | 1 | 1% |
| Total | 100 | 100% |