Lauren Dennis

BSAN 6070

Professor Brahma

February 20, 2022

CA03 – Decision Trees Questions

Q.1.1 Why does it makes sense to discretize columns for this prediction problem?

Because we are looking at specific attributes for thousands of people to be able to predict if their income is at, above, or below $50,000, it is important to bin the data into categories/levels. Some values of the independent variables can be continuous, but it would be best practice to bin the data for more encompassing data analysis. By using the binning technique, we can allow the decision tree model to parse the individuals based on levels within the independent variables and eventually come to a decision.

Q.1.2 What might be the issues (if any) if we DID NOT discretize the columns.

Some issues we may run into if we did not discretize the columns would be designing the model to create its own cutoffs for the independent variables. It would also make encoding the data very difficult because we would have to assign different labels to values that are potentially only one cent off from each other.

Table

Description automatically generated

Q.7.1 Decision Tree Hyper-parameter variation vs. performance (run your program   
manually for the following eight cases and enter the Model Performance values   
manually in the table)

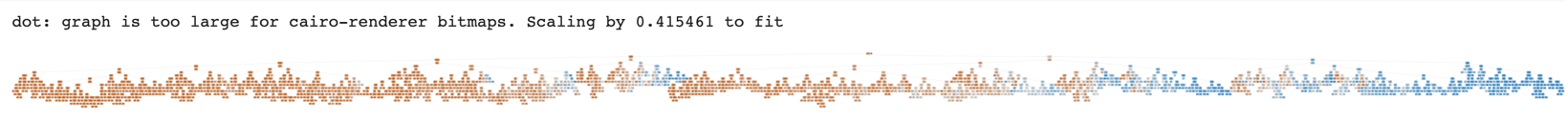
Q.8.1 How long was your total run time to train the model?

Total run time to train the model was 68.2 ms. This value was obtained by using the %%time function while training the model.

Q.8.2 Did you find the BEST TREE?

From the values that I chose for the table in Q7.1, the best tree appears to be the second option from the bottom. It must be taken into consideration that these values were arbitrarily chosen and I would like to try different variations to find an even better fitting model.

Q.8.3 Draw the Graph of the BEST TREE Using GraphViz



Q.8.4 What makes it the best tree?

This tree has an F1 score (considers the recall and precision) of .74 along with an accuracy of .84. This tree uses the Gini impurity split criterion. More research will have to be done to obtain all AUC scores, but this tree output an AUC value of .74, which is a fair score.

Q.10.1 What is the probability of the outcome of the prediction for this? What is   
your decision probability threshold and what is your predicted decision based on that?

The probability outcome is array([[0.27272727, 0.72727273]]), meaning that there is a 27% chance that the new record yields a [0], or the individual has an income level at or below $50k. On the other hand, there is a 72% chance that the new individual yields a [1], or above $50k income. I predict that the new record will be [1], or above $50k income, based on a decision probability threshold of 70%.

Q. 10.2 What is the probability that your outcome prediction is accurate?

The probability that my outcome is accurate is 72.7%.