John Lehne

BSAN 6070

CA03 – Decision Tree Algorithms

20 February 2022

Q1.1: Why does it makes sense to discretize columns for this prediction problem?

It makes sense to discretize columns for this prediction problem because having a large number of numeric values would require unique records for each distinct value in order to measure frequency. For decision tree analysis, these unique records would all have to be accounted for creating a very difficult and large analysis. Through discretizing, the volume of the analysis can be limited to a manageable number that makes reasonable sense.

Q1.2: What might be the issues (if any) if we DID NOT discretize the columns?

As mentioned above, if discretizing did not happen then there would be an excessive number of leaves. This would make the decision tree part of the analysis very difficult to conduct. With each unique numeric value needing to be a leaf, there could be an enormous number of leaves for some columns. Discretizing allows for the number of leaves to be limited by grouping numbers into bins in order to make the analysis simpler and make more sense. This process of discretizing the columns can help with limiting overfitting of the model.

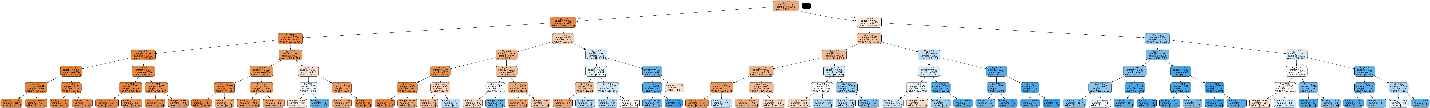
Q8.1: How long was your total run time to train the model?

To manually run the model and input all the results manually into an Excel spreadsheet took roughly 30 minutes. I would imagine automating this process would allow for faster times and the analysis of a larger volume of hyperparameters and their impact on the decision tree. Manually running the model only allowed me to test 8 different hyperparameter combinations.

Q8.2: Did you find the BEST TREE?

After manually running my various combinations of hyperparameters, I was able to find a tree with improved accuracy, precision, and f1 score. The recall was marginally smaller, but not significantly. With these results from the model, I would assess this tree to be the best of the hyperparameter combinations I tested. The criterion was also set to entropy for this best tree.

Q8.3: Draw the Graph of the BEST TREE Using GraphViz



Q8.4: What makes it the best tree?

As mentioned above in the previous question, I am assessing the impact of the hyperparameter combinations on the accuracy, recall, precision, and f1 score. This best tree had the highest values for accuracy, precision, and f1 score. Only a very small decrease was seen in the recall value. In addition to these, the AUC also was the highest with this tree. With all of these scores seeing an increase with this combination of hyperparameters, I have judged that this is the best tree compared to the other models created from the hyperparameters.

Q10.1: 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 of the outcome of the prediction is .64. The decision probability threshold is .50, meaning that the probability of the outcome of the prediction being at .64 would return a predicted value of 1. This value equates to labeling the income category of the new individual as >50k.

Q10.2: What is the probability that your outcome prediction is accurate?

The probability that the outcome prediction of 1 (income category >50k) is accurate is .64 and the probability of the outcome actually being 0 (income category <=50k) is .36.