Krissy Wong
BSAN 6070
2/18/2022

CA03 Decision Tree Algorithm

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

It makes sense to discretize the columns there are many categorical values such as Occupation and Education. It would be more cumbersome to work with data that includes numerical values in the columns. It will be better to bin the values to minimizes the noise of the data.

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

If we did not discretize the columns then the data will be skewed if there are any outliers in the data. Binning the data will place the outliers with the nearest corresponding bin.

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)

CA03 - Deicison Tree

Name:	Krissy Wong

Decision Tree Hyperparameter Variations Vs. Tree Performance

========== Complete the following table ==========

Hyperparameter Variations				Model Perfromance			
Split Criteria (Entropy or Gini)	Minimum Sample Split	Minimum Sample Leaf	Maximum Depth	Accuracy	Recall	Precision	F1 Score
Gini Impurity	*10*	*15*	*10*	0.8423	0.84	0.83	0.84
	5	*10*	*5*	0.8423	0.83	0.82	0.82
	2	*5*	*2*	0.8165	0.82	0.8	0.8
	7	*10*	*7*	0.8361	0.84	0.83	0.83
Entropy	*2*	*15*	*10*	0.8424	0.84	0.83	0.84
	2	*15*	*7*	0.8373	0.84	0.83	0.83
	2	*20*	*7*	0.8368	0.84	0.83	0.83
	2	*20*	*10*	0.843	0.84	0.84	0.84

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

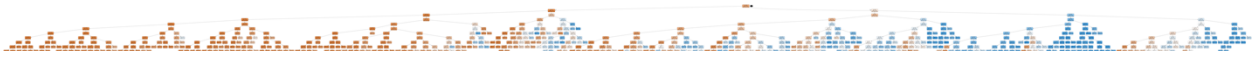
The total run time for the model is 0.01488.

Q.8.2 Did you find the BEST TREE?

The best tree had the highest accuracy. In this case it is Tree 8, which uses an Entropy Split Criteria, and has an accuracy of 0.843.

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

Refer to code to see result.



Q.8.4 What makes it the best tree?

Tree 8 is the best tree because it had the highest accuracy and the highest model performance scores, as seen in Q.7.1.

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?

My predicted decision based on my new trained model is “The income for this type of person is equal or below 50K”.

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

The probability that my outcome prediction is accurate is 0.714.