ASSIGNMENT 2

KSHITIJ R DANI
2015B4A70594G

INSIGHT QUESTIONS

Do you need data pre-processing? If Yes, mention all the pre-processing steps required and why?Else Why not?

Ans: Dataset contains no null or duplicate values. Normalisation also not required as the scale is not too large that it'll skew the graph. The categorical data columns are given dummy values using On-hot encoding.

How Optimal Depth Limit for Decision Tree was chosen and why do you think it is the best possible?

Ans: Plotted Elbow graph to see that the least distance between training and test error was around max-depth 5. This is appropriate because higher depth would lead to overfitting and lesser depth would lead to underfitting.

Which algorithm (mentioned in quest A (a)) will be chosen based on given metrics? Is Accuracy measure alone enough to decide which model is best? Why or Why not? .

Ans: Decision Tree is probably the best algorithm. It gives higher accuracy (89.8%) than the others (though i not much of a difference. KNN - 88.4% ; RF - 89.67%). Accuracy isn't not the only thing that helps us determine which algorithm is better. DT gives a higher Area Under the Curve as well. Another reason to choose DTs would be that the dataset is relatively small and running Random Forest on it seems to be like overkill.

What is the significance of AUC graph? Explain how AUC graph is useful to decide the admissibility of a model for the given problem.

Ans: The AUC Graph plots the values between False Positive Rate V/s True Positive Rate. Here False Positive Rate = FP/(FP+TN) and True Positive Rate = TP/(TP+FN) A higher area value implies that the model is fitting well as it's able to give higher TPR and lower FPR. In our case, the DT model gives highest area under the curve ((0.66) compared to KNN- 0.54 and RF-0.60.

Compare the two models (mentioned in Ques B) in terms of underfitting and overfitting. Explain.

Ans: The 2 models in the end give similar results on checking both accuracy and confusion matrix as well as classification report. Now it is possible to say that the model with max depth = 40 might be overfitting as this is a very high number of splits for a simple dataset.