ML Assignment 2 Report

Aryan GD Singh
2019459

Q1.

We first read the dataset and then start preprocessing

Preprocessing:

1. We examine the data and notice that some samples contain NULL values, so those samples are removed from the data.
2. No column is removed.
3. Next we check which features are highly correlated
4. cbwd is encoded as integer
5. Data is split into training and test sets, and also into input and labels

A

Entropy provides better accuracy

B

The accuracy increases in both cases with increase in depth, the best results are given by depth 30

C

Ensembling with depth 3 trees does not lead to very good results, with an accuracy of around 34% for both training and test sets. This is similar to the performance of a single tree of depth 3 from the graph in part B.

A single tree leads to more than 80% accuracy using the default parameters in part A, and also with more depth in part B, which means our ensemble is limited due to the small depth of the individual trees.

D

All 3 graphs follow similar patterns, initial accuracy is heavily dependent on depth but not on the number of trees.
As the max depth increases, the number of trees has more influence on the accuracy
For training set, increasing the value of the parameters increases the accuracy, this makes sense because as the model becomes more complex it is able to better fit to the training data. There are highly diminishing returns after about 20 depth and 20 trees as the model is close to perfect
For validation set, increasing the number of trees does increase performance, but depth 20 gives better results than depth 30. This is because as the depth increases the model becomes more complex and starts overfitting on the training data

Depth 20 is chosen as it gives us the best results and avoid overfitting

E

ADAboost is stochastic process so there are variations in the result, but generally 8, 10, 15 estimators give us consistently high accuracy of around 89%

Performance is similar to Random forest with 20 depth, with both giving around 90%. But Random Forest is not susceptible to variance, so it may be the better choice here

Q2.

Theory