IE 425

Spring 2022 - Homework 1

Ahmet Yiğit Doğan – 2018402105

1. a)

Overall	Training	Test
6,863%	6,863%	6,863%

Table 1: Percentage of companies in distress in the original and split data sets

The percentages look pretty much the same, which implies that the data set is suitable for a homogeneous split with high precision.

b) As the cp table generated by the R script suggests, the tree with 2 splits and 3 terminal nodes yields the lowest cross-validation error.

c)

Accuracy	93,682%
Sensitivity	33,333%
Specificity	98,129%
Precision	56,757%

Table 2: Prediction outcomes for Q1, part c

d) To obtain the tree with the smallest deviance, the *"minsize"* and *"mindev"* parameters should be set to the minimum values they can take, 2 and 0, in the *"tree"* function. This yields a tree with 85 leaf nodes.

e)

Accuracy	91,285%
Sensitivity	47,619%
Specificity	94,503%
Precision	38,910%

Table 3: Prediction outcomes for Q1, part e

As can be expected, the tree in part C has better accuracy, specificity, and precision compared to the one in part E which lost some of its general validity due to its perfect fit. The unexpected superiority of part E in terms of sensitivity might be related with the small number of positive financial distress values in the dataset.

2. b) Number of leaf nodes in the best tree in terms of cross validation error: 7

c) Root mean square error: 1309.853
Mean absolute error: 1005.777