Author: Ludovic Henry

7.1/

a.

Dataset1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Is Stat | Is IT | Last contact | Took course |
| Customer 1 | 1 | 0 | 1 | 0 |
| Customer 2 | 0 | 0 | 1.1 | 1 |
| Prospect 1 | 0 | 1 | 1 |  |

Dataset2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Is Stat | Is IT | Is Other | Last contact | Took course |
| Customer 1 | 1 | 0 | 0 | 1 | 0 |
| Customer 2 | 0 | 0 | 1 | 1.1 | 1 |
| Prospect 1 | 0 | 1 | 0 | 1 |  |

b.

Dataset1:

D(Customer1, Prospect1) = = 1.4142

D(Customer2, Prospect1) = = 1.0049

Dataset2:

D(Customer1, Prospect1) = = 1.4142

D(Customer2, Prospect1) = = 1.4177

c.

The closest neighbor in Dataset1 is Customer2, while the closest neighbor in Dataset2 is Customer1. Using 2 or 3 dummies thus have an effect on the model.

7.2/

a.

This customer is classified as Personal.Loan = 0

b.

The *k* that balance between overfitting and underfitting is *k* = 61.

c.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1786 | 166 |
| 1 | 25 | 23 |

d.

The customer is classified as Personal.Loan = 0

e.

9.3/

a.

i.

The three most important factor are: AGE\_08\_04, HP and KM.

ii.

The training set predictions are not perfect, but they are more accurate than on the validation set. This occurs because the Decision Tree model tends to overfit easily to the training set if appropriate measures are not taken.

iii.

We can do that by limiting the complexity of the decision tree.

iv.

The predictive performance of the two models on the validation set are:

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | MAPE | MAE |
| Full tree | 1529.568 | 0.1126398 | 1122.904 |
| Pruned tree | 1315.646 | 0.09561973 | 960.1732 |

The pruned tree has a better performance on the validation set than the full tree.

b.

i.

The difference between the trees relies mostly in the values in the terminal nodes, and the depth of the tree. This can be explained by the reduced number of outcome which, in the case of the CT (with the Binned\_Price), is lower.

ii.

iii.

The output for the RT is 7425.556, and the output for the CT is 7428.472.