HW5

March 13, 2024

Problem 5.1 (Predicting Prices of Used Cars via Regression Trees)

The file `ToyotaCorolla.csv` contains the data on used cars (Toyota Corolla) on sale during late summer of 2004 in the Netherlands. It has 1436 records containing details on 38 attributes, including `Price`, `Age`, `Kilometers`, `HP`, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications.

Data Preprocessing. Split the data into training (60%), and validation (40%) datasets.

a. Run a full-grown regression tree (RT) with outcome variable `Price` and predictors `Age_08_04`, `KM`, `Fuel_Type` (first convert to dummies), `HP`, `Automatic`, `Doors`, `Quarterly_Tax`, `Mfr_Guarantee`, `Guarantee_Period`, `Airco`, `Automatic_ airco`, `CD_Player`, `Powered_Windows`, `Sport_Model`, and `Tow_Bar`. Set `random_state=1`.

i. Which appear to be the three or four most important car specifications for predicting the car's price?

ii. Compare the prediction errors of the training and validation sets by examining their mean squared error. How does the predictive performance of the validation set compare to the training set? Why does this occur?

iii. How might we achieve better validation predictive performance at the expense of training performance?

iv. Create a smaller tree by using `GridSearchCV()` with `cv = 5` to find a fine-tuned tree. Compared to the full-grown tree, what is the predictive performance on the validation set?

b. Let us see the effect of turning the price variable into a categorical variable. First, create a new variable that categorizes price into 20 bins. Now repartition the data keeping `Binned_Price` instead of `Price`. Run a classification tree with the same set of input variables as in the RT, and with `Binned_Price` as the output variable. As in the less deep regression tree, create a smaller tree by using `GridSearchCV()` with `cv = 5` to find a fine-tuned tree.

i. Compare the smaller tree generated by the CT with the smaller tree generated by RT. Are they different? (Look at structure, the top predictors, size of tree, etc.) Why?

ii. Predict the price, using the smaller RT and CT, of a used Toyota Corolla with the specifications listed in the following table.

[ ]:
```
from IPython.display import Image
Image(filename='images/car.png', width=800)
```

[ ]:

Variable	Value
Age_-08_-04	77
KM	117,000
Fuel_Type	Petrol
HP	110
Automatic	No
Doors	5
Quarterly_Tax	100
Mfg_Guarantee	No
Guarantee_Period	3
Airco	Yes
Automatic_airco	No
CD_Player	No
Powered_Windows	No
Sport_Model	No
Tow_Bar	Yes