Homework 2

Due on April 6, 11:59pm

We will use the cleaned Toyota corolla pricing data set in the attachment to do this homework.

a) Dichotomize the variable "Price" at its 75th percentile and call this new variable "HighPrice". Please note that you will not replace the "Price" variable but create a new variable "HighPrice".

b) Partition the data into training (60%) and validation (40%).

c) For the outcome variable "Price", fit a multiple linear regression with forward selection, regression-based kNN with k selected by the validation set, and regression tree. Use all variables in these three methods and PCA analysis is not needed.

d) Compare the above three models obtained and pick up the best one. Explain why you think it is the best one. Note that I already took out the "grey" as the reference group for color. You don't have to do any data cleaning in this homework.

e) For the outcome variable "HighPrice", fit a logistic regression with forward regression, classification-based kNN with k selected by the validation set, and classification tree.

f) Compare the models obtained and pick up the best one. Explain why you think it is the best one.