QMB 6316: R for Business Analytics
College of Business
University of Central Florida
Fall 2024

Assignment 5

Due Friday, December 6, 2024 at 11:59 PM
in digital form in your GitHub repository.

Question 1:

Continue the in-class exercise using the script `data_mining_A5.R` in RStudio, which is a modified version of the script `data_mining_demo.R`. In this exercise, we generate simulated data for car prices, which depend on mileage, whether the car has been in an accident, and whether the car has sustained major structural damage, which can happen only as a result of an accident. Specifically, the regression model is

$$CAR\_PRICE_i = \beta_0 + \beta_1 \times MILEAGE_i + \beta_2 \times ACCIDENT_i + \beta_3 \times DAMAGE_i + \epsilon_i \qquad (1)$$

where:

CAR_PRICE_i	=	the value of car i
$MILEAGE_i$	=	the mileage of car i
$ACCIDENT_i$	=	whether or car i has been involved in an accident (i.e., $ACCIDENT_i = 1$ if car i has been in an accident, zero otherwise)
$DAMAGE_i$	=	whether or not car i car has sustained major structural damage (1 or 0)

The script also generates several additional variables, `rainfall_1` to `rainfall_20`, which measure whether it rained in other states, independent of where the car sale occurred. These variables are truly unrelated to the price a car sold somewhere else but we might find that they appear to have predictive value. To verify the results, we will calculate $\bar{R}^2$ from two samples: one on which we fit the model and one to evaluate the predictions out of sample.

Run the entire script and observe the output from the series of models.

a) Verify that damage occurs in both the samples. Observe the output under the code blocks in lines 123-139 to verify that the bottom right numbers (the number of observations with accidents with damages) are not zero. If one of these bottom-right numbers is zero, run the script again.

b) Copy and paste the table of selected models and R-squared values, under the command `print(best_models)` on line 315.

c) Compare the models according to the in-sample $\bar{R}^2$ under the column `R2_in_sample`. Which model is best under this criterion? The variables in the model are listed in the column `best_new_variable` up to the row of the chosen model.

d) Compare the models according to the out-of-sample $\bar{R}^2$ under the column `R2_out_sample`. Which model is best under this criterion? The variables in the model are listed in the column `best_new_variable` up to the row of the chosen model.

e) Compare the differences between the two models. Do any variables appear in one model and not the other? Which model do you recommend on the basis of this data? Is your recommended model the same as the true model?