Homework 2

BUAN 6356

Read the instructions below before you start your analysis.

1. Create an R Markdown document to prepare your answers. Your group should upload **two (2)** files on eLearning: (i) an **.RMD** file; and (ii) a **.PDF** file that is generated using “knit” in the .RMD file. Both of these files should contain the required R code, R tables and charts, and all the required explanations and answers to the questions in the homework.

2. **DO NOT** use an absolute directory path. I should be able to “knit” your R Markdown document to an .html/.pdf document without trying to find the input data in another directory. Test the “knit” process before uploading files on eLearning. Assume that I have the .csv file(s) mentioned below.

3. **DO NOT** change the dataset name before importing it into R. If you rename the dataset or any variable(s), use your R script to do that.

4. Label the charts and/or tables appropriately. Your reader should be able to figure out what information a chart is providing by looking at the chart title and its labels.

5. Any assignment submitted after the deadline will be considered late and will not be graded.

Homework 2

A consulting firm working for Southwest Airlines would like to predict airfares using ***Airfares.csv***, which contains real data that were collected between Q3-1996 and Q2-1997. The variables in these data are listed below. Some airport-to-airport data (*e.g.,* JFK-BWI) are available, but most data are at the city-to-city level (*e.g.,* Atlanta-Boston). A key question is whether the presence or absence of Southwest Airlines (a low-cost entrant) would have any effect on *fare*.

Variable Description:

S_CODE: Starting airport's code

S_CITY: Starting city

E_CODE: Ending airport's code

E_CITY: Ending city

COUPON: Average number of coupons for that route *(a one-coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, etc.)*

NEW: Number of new carriers entering that route between Q3-96 and Q2-97

VACATION: Whether (Yes) or not (No) a vacation route

SW: Whether (Yes) or not (No) Southwest Airlines serves that route

HI: Herfindahl index, a measure of market concentration *(higher number means smaller number of available carriers on that route)*

S_INCOME: Starting city's average personal income

E_INCOME: Ending city's average personal income

S_POP: Starting city's population

E_POP: Ending city's population

SLOT: Whether or not either endpoint airport is slot-controlled *(this is a measure of airport congestion)*

GATE: Whether or not either endpoint airport has gate constraints *(this is another measure of airport congestion)*

DISTANCE: Distance between two endpoint airports in miles

PAX: Number of passengers on that route during period of data collection

FARE: Average fare on that route

Remove the first four predictors (S_CODE, S_CITY, E_CODE, E_CITY) from all your analysis below.

1. Create a correlation table and scatterplots between FARE and the predictors. What seems to be the best single predictor of FARE? Explain your answer.

2. Explore the categorical predictors by computing the percentage of flights in each category. Create a pivot table with the average fare in each category. Which categorical predictor seems best for predicting FARE? Explain your answer.

3. Create data partition by assigning 80% of the records to the training dataset. Use rounding if 80% of the index generates a fraction. Also, set the seed at *42*.

4. Using *leaps* package, run stepwise regression to reduce the number of predictors. Discuss the results from this model.

5. Repeat the process in (4) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (4) in terms of the predictors included in the final model.

6. Compare the predictive accuracy of both models—stepwise regression and exhaustive search—using measures such as RMSE.

7. Using the exhaustive search model, predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28,760, E_INCOME = $27,664, S_POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.

8. Predict the reduction in average fare on the route in question (7.), if Southwest decides to cover this route [using the exhaustive search model above].

9. Using *leaps* package, run backward selection regression to reduce the number of predictors. Discuss the results from this model.

10. Now run a backward selection model using *stepAIC()* function. Discuss the results from this model, including the role of AIC in this model.