Assignment 4: EE522

1/2022

Instructions There are 1 question on this assignment. This question involves coding, so start early. Please submit your completed homework to Google Classroom by **Midnight, Thursday, September 22, 2022** with subject: "EE522–your name –" Submit your homework as the Jupiter Notebook code and figures. (so I can easily split it up for grading). Include your name and email address on each set.

Predicting Airfare on New Routes.

The following problem takes place in the United States in the late 1990s, when many major US cities were facing issues with airport congestion, partly as a result of the 1978 deregulation of airlines. Both fares and routes were freed from regulation, and low-fare carriers such as Southwest (SW) began competing on existing routes and starting nonstop service on routes that previously lacked it. Building completely new airports is generally not feasible, but sometimes decommissioned military bases or smaller municipal airports can be reconfigured as regional or larger commercial airports. There are numerous players and interests involved in the issue (airlines, city, state and federal authorities, civic groups, the military, airport operators), and an aviation consulting firm is seeking advisory contracts with these players. The firm needs predictive models to support its consulting service. One thing the firm might want to be able to predict is fares, in the event a new airport is brought into service. The firm starts with the file Airfares.csv, which contains real data that were collected between Q3—1996 and Q2—1997. The variables in these data are believed to be important in predicting FARE. Some airport-to-airport data are available, but most data are at the city-to-city level. One question is that will be of interest in the analysis is the effect that the presence or absence of Southwest has on FARE.

Description of Variables for Airfare Example

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 (a one-coupon flight is a nonstop flight, a two-coupon flight is a one-stop flight, etc.) for that route
NEW	Number of new carriers entering that route between Q3—1996 and Q2—1997
VACATION	Whether (Yes) or not (No) a vacation route
SW	Whether (Yes) or not (No) Southwest Airlines serves that route
HI	Herfindahl index: measure of market concentration
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

Task:

(1) Explore the numerical predictors and outcome (FARE) by creating a correlation table, heat map, and examining some scatterplots between FARE and those predictors. What seems to be the best single predictor of FARE?

(2) Explore the categorical predictors (Excluding the first four) by computing the percentage of flights in each category. Create the pivot table with the average fare in each category. Which categorical predictors seems best for predicting FARE?

(3) Find a model with the regression model for predicting the average fare on a new route:

- (3.1) Convert categorical variables into dummy variables. Then, partition the data into training and test sets. The model will be fit to the training data and evaluated on the test set.
- (3.2) Use regression to predict the following models.
  - (a) Using COUPON,DISTANCE,VACATION to be the features.
  - (b) Using COUPON,DISTANCE,VACATION, GATE to be the features.
  - (c) Using All features. You can ignore the first four predictors (S_CODE,S_CITY,E_CODE,E_CITY).
  - (d) Report the estimated results by creating the table to compare the estimated results as you have learned in class.
  - (e) Then, compute the RMSE and MAE of train and test data.
- (3.3) Compare the predictive performance of the models above. Which one you select? And Why?.