Homework 4

Instructions:

You should work individual for this assignment. The data file for this homework is Airfares.xls, which is to be downloaded from Canvas. Create a new Word document and save it as HW4Answers_xx (where xx is your last name). Where required, write your answers or paste screenshots into this Word document. Your response should not exceed 100 words for each below question. You need to submit this Word document and XLMiner solution.

Problem Description for 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 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.xls, which contains real data that were collected between Q3-1996 and Q2-97. The variables in these data are listed in the following table, and 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 that will be of interest in the analysis is the effect that the presence or absence of Southwest (SW) has 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 (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route
NEW	number of new carriers entering that route between Q3-96 and Q2-97
VACATION	whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes
SW	whether Southwest Airlines serves that route (Yes) or not (No)
HI	Herfindel Index – measure of market concentration (refer to BMGT 681)
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 either endpoint airport is slot controlled or not; this is a measure of

	airport congestion
GATE	whether either endpoint airport has gate constraints or not; 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

a) Explore the numerical predictors and response (FARE) by creating a correlation table and examining some scatterplots between FARE and those predictors. What seems to be the best single predictor of FARE?
b) Explore the categorical predictors (excluding the first four) 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?
c) Find a model for predicting the average fare on a new route:
    i. Convert categorical variables (e.g., SW) into dummy variables. Then partition the data into training and validation sets. The model will be fit to the training data and evaluated on the validation set.
    ii. Why should the data be partitioned into training, and validation? What will the training set be used for? What will the validation set be used for?
    iii. Use stepwise regression to reduce the number of predictors. You can ignore the first four predictors (S CODE, S CITY, E CODE, E CITY). Report the estimated model selected.
    iv. Repeat (iii) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (iii) in terms of the predictors that are in the model.
    v. Compare the predictive accuracy of both models (iii) and (iv) using measures such as RMSE and average error and lift charts.
    vi. Using model (iv), 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.
    vii. Using model (iv), predict the reduction in average fare on the route in (vi) if Southwest decides to cover this route.
    viii. In reality, which of the factors will not be available for predicting the average fare from a new airport (i.e., before flights start operating on those routes)? Which ones can be estimated? How?
    ix. Select a model that includes only factors that are available before flights begin to operate on the new route. Use an exhaustive search to find such a model.
    x. Use the model in (ix) to predict the average fare on a route with 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.

xi. Compare the predictive accuracy of this model with model (iv). Is this model good enough, or is it worthwhile re-evaluating the model once flights begin on the new route?

Important submission instructions

Save your Word file and XLMiner solution. Use the link "Homework 4" to upload these files. **Due by 11.59 P.M. Mar. 12, 2018.**