Predictive Modeling in R

**Predicting Airfare on New Routes**. The question provides the Airfares.csv data. 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 non- stop 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 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 has on FARE.

**Table. Description of Variables for Airfare Example**

|  |  |
| --- | --- |
| 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 nonstop flight, a two-coupon flight is a one-stop flight, etc.) for that route |
| NEW | Number of new carriers entering that route between Q3-96 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: measure of market concentration of the airline industry |
| 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 |

1. 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?
2. Explore the categorical predictors (excluding the first four) by computing the percentage of flights in each category. Create a graphical chart to visualize the average fare in each category (Hint: using the barplot() function). Which categorical predictor seems best for predicting FARE?
3. Find a model for predicting the average fare on a new route:
   1. Convert categorical variables (e.g., SW) into dummy variables. Then, partition the data into training (60%) and validation (40%) sets. The model will be fit to the training data and evaluated on the validation set.
   2. 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.
   3. Repeat (ii) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (ii) in terms of the predictors that are in the model.
   4. Compare the predictive accuracy of both models (ii) and (iii) using measures such as RMSE and MAPE.
   5. Using model (iii), 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.
   6. Predict the reduction in average fare on the route in (v) if Southwest decides to cover this route [using model (iii)].
   7. 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?
   8. 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.
   9. Use the model in (viii) 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 = 12782, DISTANCE = 1976 miles.
   10. Compare the predictive accuracy of this model with model (iii). Is this model good enough, or is it worthwhile reevaluating the model once flights begin on the new route?
4. In competitive industries, a new entrant with a novel business plan can have a disruptive effect on existing firms. If a new entrant’s business model is sustainable, other players are forced to respond by changing their business practices. If the goal of the analysis was to evaluate the effect of Southwest Airlines’ presence on the airline industry rather than predicting fares on new routes, how would the analysis be different? Describe technical and conceptual aspects.

It would be important to see if Southwest’s presence on a route causes a change in the fares of other airlines along that same route. A hypothesis test could be performed to determine if a significant difference exists, attempting to control for other factors, between fares charged by other airlines for Southwest and non-Southwest routes. Additionally, the HI and PAX variables would become more important to examine if the purpose of this analysis was changed. Does the airline market concentration change on routes with Southwest? Does it favor Southwest specifically, or does it create a more even concentration between multiple airlines? For PAX, are more passengers traveling on routes with Southwest? Is this caused by Southwest having cheaper flights, or other airlines adjusting their rates along those routes as well? Does the coupon rate for a route change as other airlines try to change their customers’ experience, perhaps by offering more direct flights than Southwest, for a higher fare? Aiming to analyze Southwest’s effect on the industry as a whole would allow an analyst to gain more insight from the other variables in this dataset, beyond fare.