Data Mining - CSC550

Final Project: Airfares Analysis

Surav Shrestha

James Shostak

Mohammad Rahbar

Sullivan University

Spring 2017

Professor Tran

**Introduction**

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 Q2-1997. The variables in these data are listed in Table 1, and most data are at the city-to-city level. One question that will be of interest in the analysis is the effect that the presences of absence of Southwest (SW) has on FARE.

In this project we try to answer to following criteria:

* Finding the single best predictor of fare by exploring the numerical predictors and response (FARE) with the help of a correlation table and examining some scatter plots between FARE and those predictors.
* Exploring the categorical predictors by computing the percentage of flights in each category with the help of pivot tables.
* Comparing the best model in terms of the predictors between the stepwise regression to reduce the number of predictors and exhaustive search instead of stepwise regression.
* Predicting the average fare on a route using exhaustive search.
* Comparing the predictive accuracy of the model.

The following table shows the code description for the attributes in the file “Airfare.xls”.

Table 1 - Data Codes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Variable | Description | Type | Unit |
| 1 | S\_CODE | starting airport’s code |  |  |
| 2 | S\_CITY | starting city |  |  |
| 3 | E\_CODE | ending airport’s code |  |  |
| 4 | E\_CITY | ending city |  |  |
| 5 | 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 | Numerical |  |
| 6 | NEW | number of new carriers entering that route between Q3-96 and Q2-97 | Numerical |  |
| 7 | VACATION | whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes | Categorical |  |
| 8 | SW | whether Southwest Airlines serves that route (Yes) or not (No) | Categorical |  |
| 9 | HI | Herfindel Index – measure of market concentration (refer to BMGT 681) | Numerical |  |
| 10 | S\_INCOME | starting city’s average personal income | Numerical | $ |
| 11 | E\_INCOME | ending city’s average personal income | Numerical | $ |
| 12 | S\_POP | starting city’s population | Numerical |  |
| 13 | E\_POP | ending city’s population | Numerical |  |
| 14 | SLOT | whether either endpoint airport is slot controlled or not; this is a measure of airport congestion | Categorical |  |
| 15 | GATE | whether either endpoint airport has gate constraints or not; this is another measure of airport congestion | Categorical |  |
| 16 | DISTANCE | distance between two endpoint airports in miles | Numerical | mile |
| 17 | PAX | number of passengers on that route during period of data collection | Numerical |  |
| 18 | FARE | average fare on that route | Numerical | $ |

**Exploring Numerical Predictors**

To find the best numerical predictor we use scatter plot to derive the correlation between two different characteristics in order to detect possible relationships. As Scatter plot shows a cause and effective relationship between the two characteristics, in case of any correlation if one variable changes then the other variable will also change.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *COUPON* | *NEW* | *HI* | *S\_INCOME* | *E\_INCOME* | *S\_POP* | *E\_POP* | *DISTANCE* | *PAX* | *FARE* |
| COUPON | 1 |  |  |  |  |  |  |  |  |  |
| NEW | 0.020223069 | 1 |  |  |  |  |  |  |  |  |
| HI | -0.347251894 | 0.054146918 | 1 |  |  |  |  |  |  |  |
| S\_INCOME | -0.088402648 | 0.026596726 | -0.027382247 | 1 |  |  |  |  |  |  |
| E\_INCOME | 0.046889195 | 0.113376642 | 0.082392574 | -0.138864196 | 1 |  |  |  |  |  |
| S\_POP | -0.107763359 | -0.01667212 | -0.172495472 | 0.517187183 | -0.144058568 | 1 |  |  |  |  |
| E\_POP | 0.094969942 | 0.058568182 | -0.062456209 | -0.27228027 | 0.458418059 | -0.28014283 | 1 |  |  |  |
| DISTANCE | 0.746805214 | 0.080965198 | -0.312374454 | 0.028153344 | 0.176530736 | 0.018436672 | 0.115639695 | 1 |  |  |
| PAX | -0.336973577 | 0.010495275 | -0.1689611 | 0.1381971 | 0.259961055 | 0.28461056 | 0.314697502 | -0.1024816 | 1 |  |
| FARE | 0.49653696 | 0.091729687 | 0.025195009 | 0.209134853 | 0.326092288 | 0.145097079 | 0.285042985 | 0.670015995 | -0.090705409 | 1 |

**Scatter Plot of all Numerical Predictor**

Based on the results of the data seen on the correlation table and the scatter plot, the best single predictor of FARE seems to be the PAX (number of passengers on that route during period of data collection).

**Exploring Categorical Predictor**

To find out which variables have the most effect on target variables, we use Pivot table. Pivot tables mixes information from multiple variables and calculate a range of summary statistics. By using the pivot table, we could determine what sizes the different categories have and that how the outcome will be at each category. In particular, it helps to detect and summarize selected columns and rows of data in a spreadsheet or database to achieve a desired report.

By plotting pivot table, it could be realized that the best categorical variable is the one the has the highest total between the categories. Therefore, in this case PAX and DISTANCE seem to be best categorical predictors.

**Exploring The Best Suited Model**

Multiple linear regression is the most popular model for predictionsand helps to find the linear relationship between a quantitative dependent variables and a set of predictors. Multiple linear regression is derived by using least squares in a way that the sum-of-squares of differences of observed and predicted values is minimized.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **XLMiner : Multiple Linear Regression** | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **Output Navigator** | | |  |  |  |  |  |  |
| [Inputs](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B8) | [Train. Score - Summary](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B71) | | Valid. Score - Summary | | Test Score - Summary | | Database Score | |
| [Elapsed Time](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B77) | [Train. Score - Detailed Rep.](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#'MLR_TrainScore1'!A1) | | Valid. Score - Detailed Rep. | | Test Score - Detailed Rep. | | New Score - Detailed Rep. | |
| [ANOVA](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B48) | [Training Lift Charts](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#'MLR_TrainLiftChart1'!A1) | | Validation Lift Charts | | Test Lift Charts | | Subset selection | |
| [Reg. Model](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B33) | [Residuals-Fitted Values](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#'MLR_Resi-FitVal1'!A1) | | [Var. Covar. Matrix](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B56) | | Collinearity Diagnostics | |  | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Inputs** |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Data** | | | | | | |  |  |  |  |  |
|  | Training data used for building the model | | | Airfares.xlsx | | | |  |  |  |  |  |
|  | # Records in the training data | | | 200 | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Variables** | | | | | | | | | | | |
|  | # Input Variables | | | 9 | | | | | | | | |
|  | Input variables | | | COUPON | NEW | HI | S\_INCOME | E\_INCOME | S\_POP | E\_POP | DISTANCE | PAX |
|  | Output variable | | | FARE | | | | | | | | |
|  | Constant term present | | | Yes | | | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameters/Options** | | | | | | |
| Show ANOVA table | | | Yes | | | |
| Show fitted values on training data | | | Yes | | | |
| Show variance covariance matrix | | | Yes | | | |
| Show standardized residuals | | | Yes | | | |
| Show unstandardized residuals | | | Yes | | | |
|  |  |  |  |  |  |  |
| **Output options chosen** | | | |  |  |  |
| Summary report of scoring on training data | | | |  |  |  |
| Detailed report of scoring on training data | | | |  |  |  |
| Lift charts on training data | | | |  |  |  |

Above result from XLMiner indicates that by using predictors that are uncorrelated with dependent variable, the variance of predictors would be increased. Hence, we eliminate correlated predictors with the dependent variable to increase the average error of predictors.

**Exploring The Result Set of Exhaustive Search**

To evaluate subsets for the moderate values of predictors, we use exhaustive search. To do so, we leverage regression models with all possible combinations of predictors. Exhaustive search prevents from increasing that could be achieved by increasing the number of predictors, without increasing the amount of information.

We also used Mallow’s *Cp* to assume that the full model includes all predictors that is unbiased and causes reducing prediction variability.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **XLMiner : Logistic Regression** | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **Output Navigator** | | |  |  |  |  |  |  |
| [Inputs](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B8) | [Train. Score - Summary](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B130) | | Valid. Score - Summary | | Test Score - Summary | | Database Score | |
| [Elapsed Time](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B147) | [Train. Score - Detailed Rep.](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#'LR_TrainScore1'!A1) | | Valid. Score - Detailed Rep. | | Test Score - Detailed Rep. | | New Score - Detailed Rep. | |
| [Prior Class Pr](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B40) | [Training Lift Charts](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#'LR_TrainLiftChart1'!A1) | | Validation Lift Charts | | Test Lift Charts | | [Subset selection](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B96) | |
| [Reg. Model](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B49) | [Residuals](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#'LR_Residuals1'!A1) | | [Var. Covar. Matrix](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B64) | | [Collinearity Diagnostics](file:///C:\Users\Shyama\Desktop\Airfare_project.xlsx#RANGE!B79) | |  | |

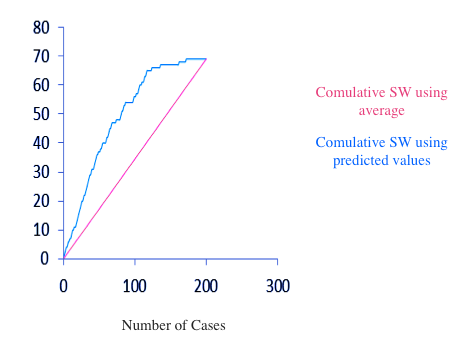
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Inputs** |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Data** | | | | | | |  |  |  |  |  |
|  | Training data used for building the model | | | Airfares.xlsx | | | |  |  |  |  |  |
|  | # Records in the training data | | | 200 | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **Variables** | | | | | | | | | | | |
|  | # Input Variables | | | 9 | | | | | | | | |
|  | Input variables | | | COUPON | NEW | HI | S\_INCOME | E\_INCOME | S\_POP | E\_POP | DISTANCE | PAX |
|  | Output variable | | | SW | | | | | | | | |
|  | Constant term present | | | Yes | | | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameters/Options** | | | | | | |
| # Iterations | | | 50 | | | |
| Marquardt overshoot factor | | | 1 | | | |
| Initial cutoff probability value | | | 0.5 | | | |
| Confidence Level % | | | 95 | | | |
| Peform collinearity diagnostics | | | Yes | | | |
| # Collinearity components | | | 2 | | | |
| Peform subset selection | | | Yes | | | |
| Subset selection procedure | | | Backward elimination | | | |
| Maximum size of best subsets | | | 9 | | | |
| # Best subsets | | | 10 | | | |
| Show coviance matrix | | | Yes | | | |
| Show residuals | | | Yes | | | |
|  |  |  |  |  |  |  |
| **Output options chosen** | | | |  |  |  |
| Summary report of scoring on training data | | | |  |  |  |
| Detailed report of scoring on training data | | | |  |  |  |
| Lift charts on training data | | | |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| According to relative occurrences in training data | | | |  |  |  |
|  |  |  |  |  |  |  |
| **Class** | | **Prob.** |  |  |  |  |
| Yes | | 0.345 | <-- Success Class | |  |  |
| No | | 0.655 |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

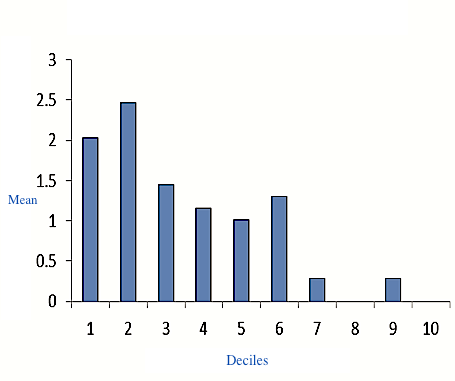
**Deciding The Best Suited Model for Airfare Dataset**

To determine the best model, we need to consider the fact that when is higher than the number of predictions increases, it influences the model’s performance. Hence, we can conclude that the best model is the one with minimum number of predictors.

***Logistic Regression***

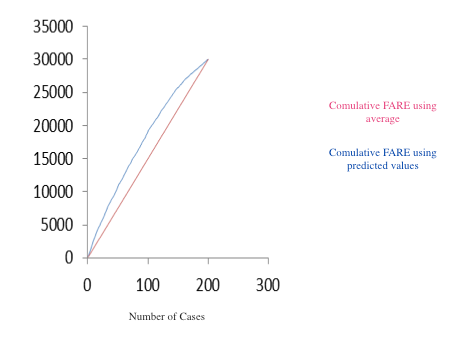


Lift Chart (training dataset)

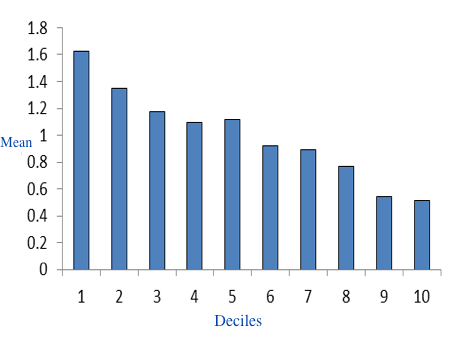


Decile-Wise Lift Chart (training dataset)

***Multiple linear regression***



Lift Chart (training dataset)



Decile-Wise Lift Chart (training dataset)

**Conclusion**

Figures above indicate MLR and LR analysis, by comparing root-mean-square-division, average errors, lift charts, and deciles charts of both models. As expected, the result proves that the model with minimum number of predictors has a very good performance. Also from the lift charts it is observed that LR can preserve a linear structure.

In conclusion, considering the comparison above, we could choose the Logistics Regression Model as the best model in order to analyze and compute airfares for Southwest airline in newly opened airports.

.

# Reference

Shmueli, G., Bruce, P. C., & Patel, N. R. (2007). *Data Mining for Business Analytics: Concepts, Techniques, and Applications With XLMINER.* Wiley.