**HOMEWORK 4**

**BIG DATA ANALYTICS FOR COMPETITIVE ADVANTAGE**

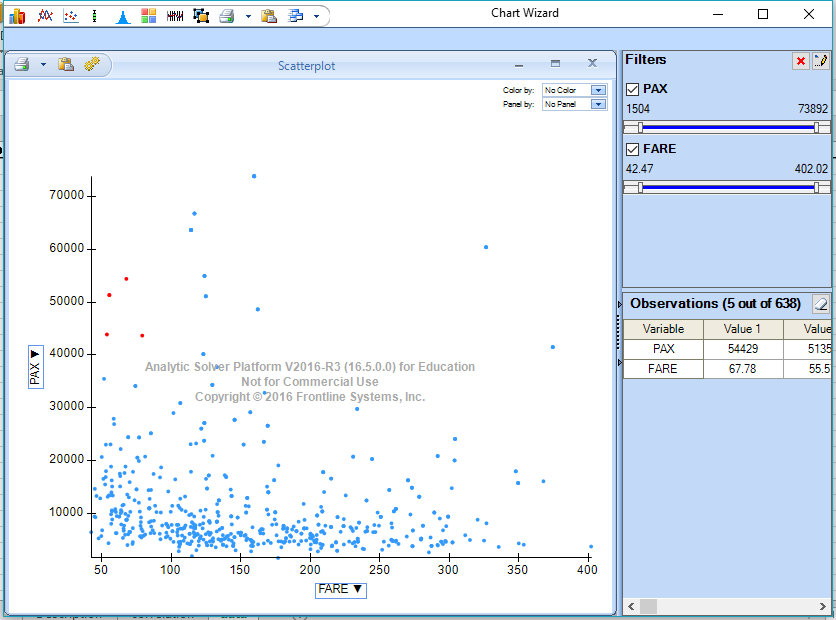
**NAME: SURABHI MAHESHWARI NINERNET:801028703**

1. The numerical predictors are as follows:

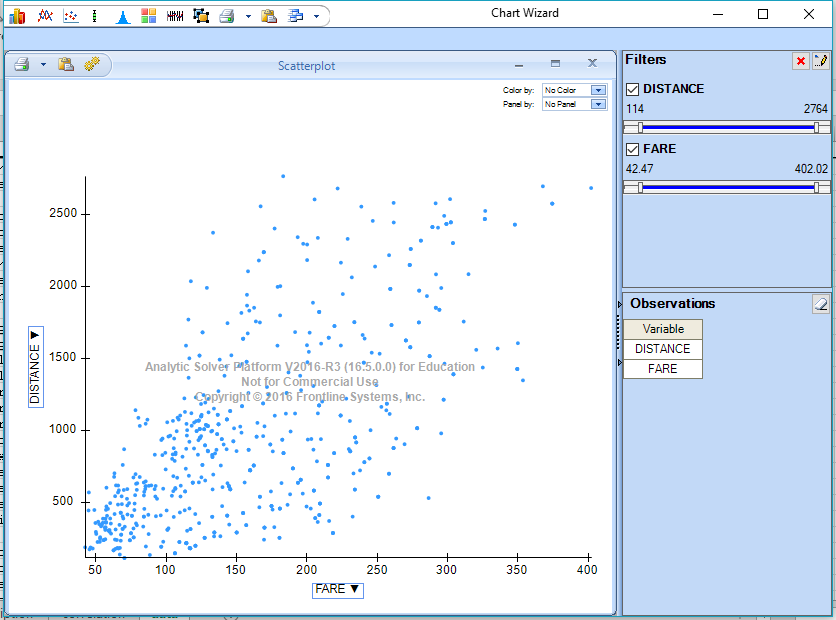
* COUPON
* NEW
* HI
* S\_INCOME
* E\_INCOME
* S\_POP
* E\_POP
* DISTANCE
* PAX

**SCATTERPLOTS**

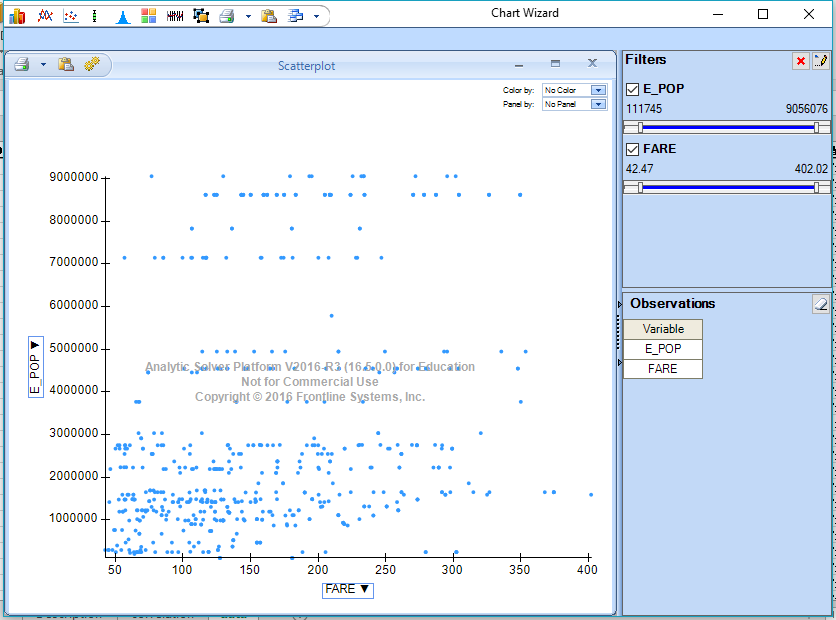
1. PAX AND FARE



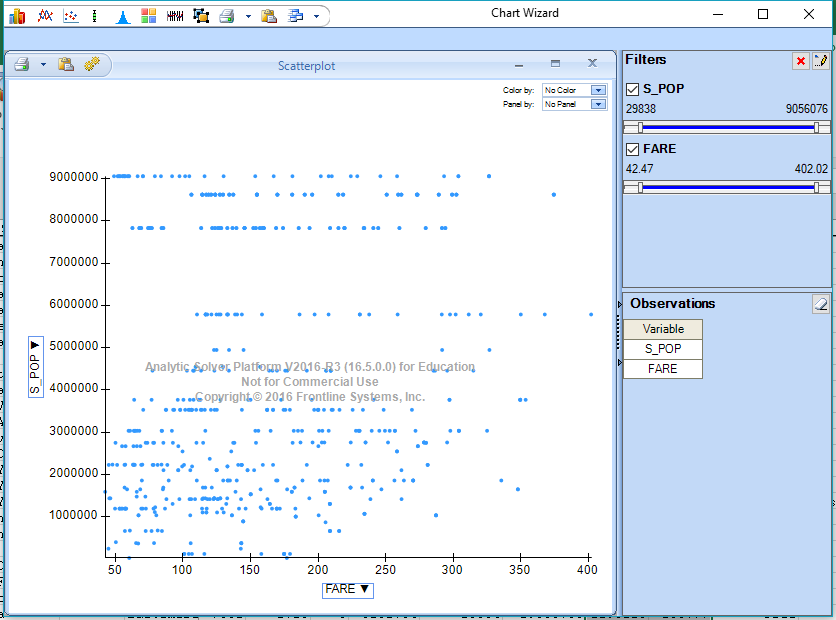
1. DISTANCE AND FARE



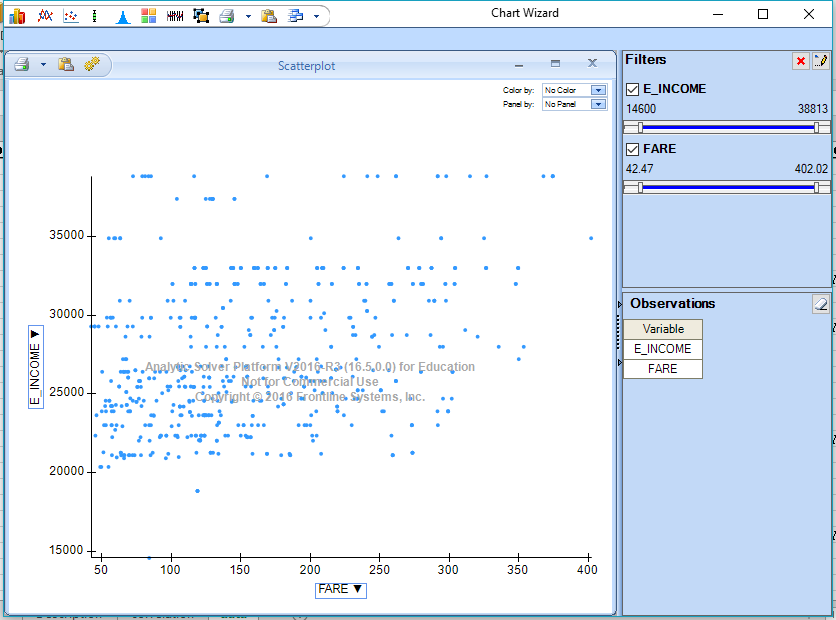
1. E\_POP AND FARE



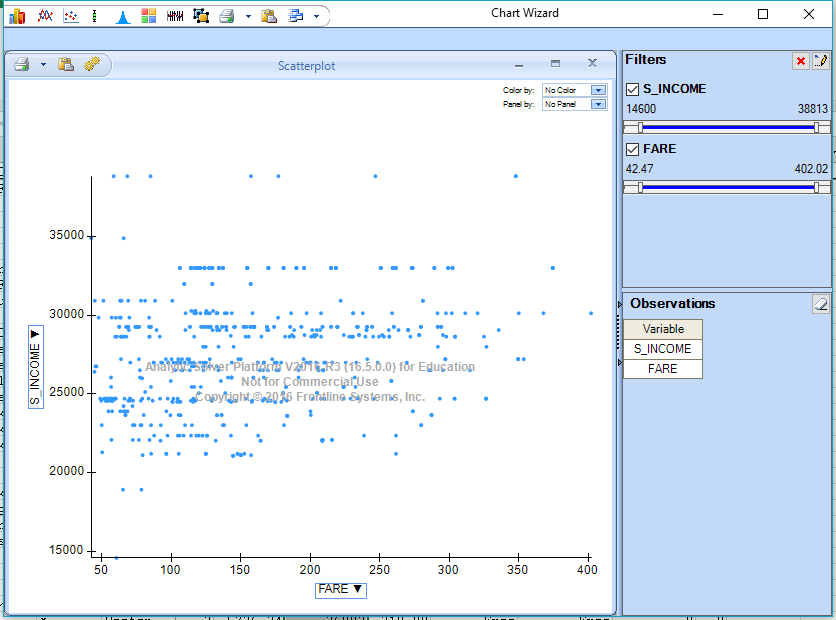
1. S\_POP AND FARE



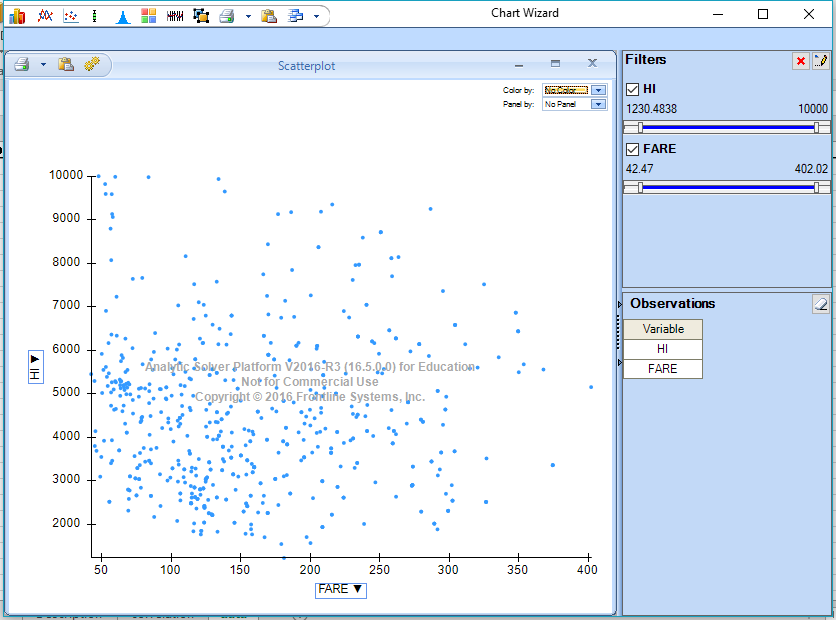
1. E\_INCOME AND FARE



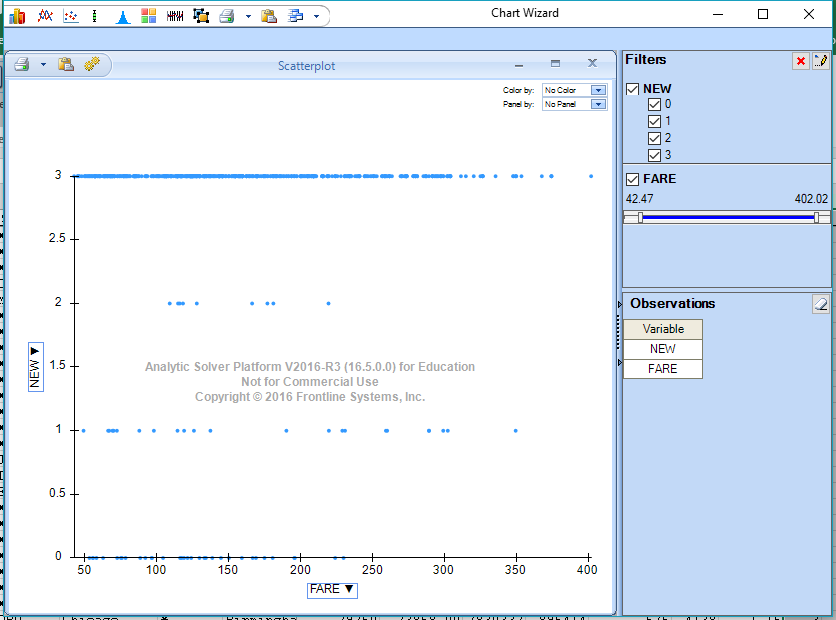
1. S\_INCOME AND FARE



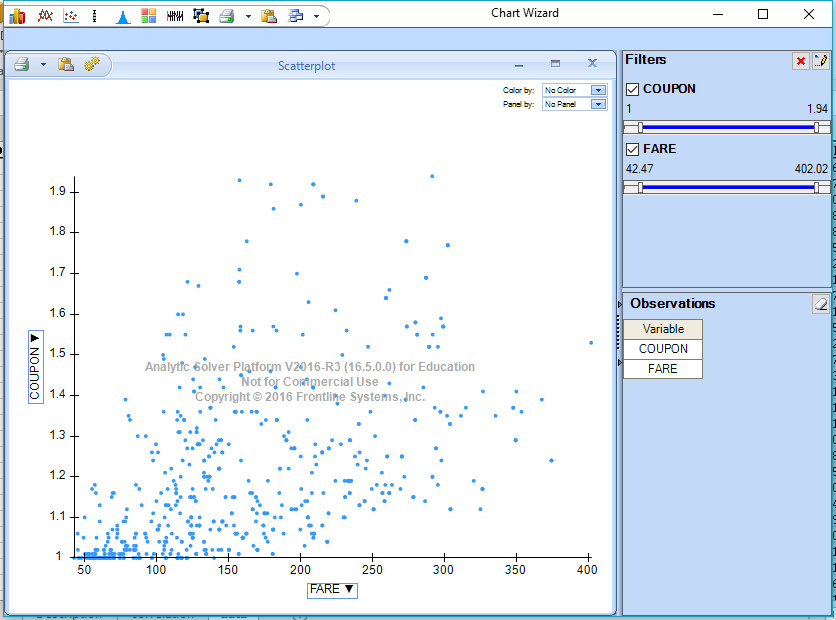
1. HI AND FARE



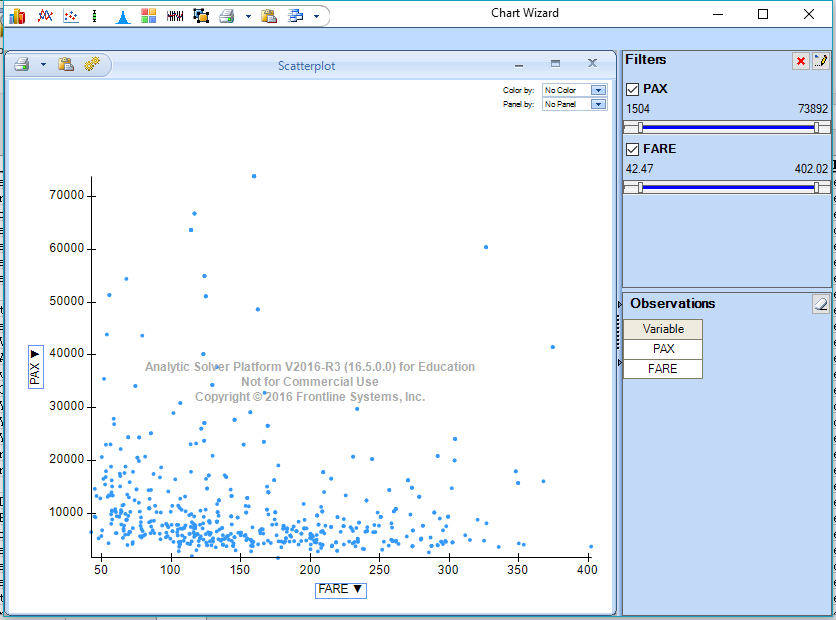
1. NEW AND FARE



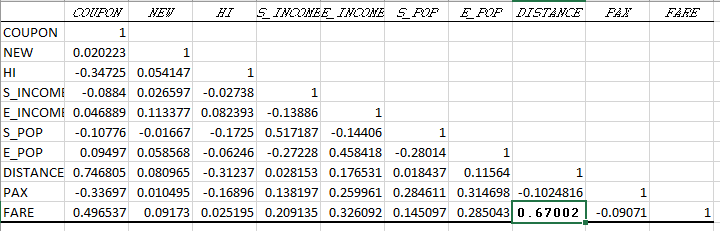
1. COUPON AND FARE



1. PAX AND FARE



CORRELATION TABLE BETWEEN NUMERICAL PREDICTORS AND FARE

Distance is the best single predictor for FARE.

1. The categorical predictors are as follows:
2. GATE
3. SLOT
4. SW
5. VACATION

The percentage of flights in each category are as follows:

|  |  |
| --- | --- |
| **Category** | **Percentage** |
|  |  |
| **GATE** |  |
| Constrained | 19.44 |
| Free | 80.56 |
|  |  |
| **SLOT** |  |
| Controlled | 28.52 |
| Free | 71.48 |
|  |  |
| **SW** |  |
| Yes | 30.41 |
| No | 69.59 |
|  |  |
| **VACATION** |  |
| Yes | 26.64 |
| No | 73.36 |

Pivot table can be created in excel by selecting the row for the pivot table and INSERT 🡪 PivotTable and marking the selection of columns. The category which has the maximum difference in their values can be treated as the best categorical predictor for predicting the fare.

The pivot table is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **SW** | **VACATION** | **GATE** | **SLOT** | **Average of FARE** |
| **No** | **No** | Constrained | Controlled | 206.7977778 |
|  |  |  | Free | 210.5816667 |
|  |  | Constrained Total |  | 209.8721875 |
|  |  | Free | Controlled | 208.3298246 |
|  |  |  | Free | 196.1837815 |
|  |  | Free Total |  | 202.1264807 |
|  | **No Total** |  |  | **204.3866261** |
|  | **Yes** | Constrained | Free | 137.3135 |
|  |  | Constrained Total |  | 137.3135 |
|  |  | Free | Controlled | 139.9017857 |
|  |  |  | Free | 143.9767164 |
|  |  | Free Total |  | 142.7756842 |
|  | **Yes Total** |  |  | **141.8257391** |
| **No Total** |  |  |  | **188.1827928** |
| **Yes** | **No** | Constrained | Controlled | 74.28 |
|  |  |  | Free | 139.96 |
|  |  | Constrained Total |  | 131.75 |
|  |  | Free | Controlled | 110.1705882 |
|  |  |  | Free | 96.95149123 |
|  |  | Free Total |  | 98.66694656 |
|  | **No Total** |  |  | **100.5710072** |
|  | **Yes** | Free | Controlled | 131.605 |
|  |  |  | Free | 89.81117647 |
|  |  | Free Total |  | 92.85072727 |
|  | **Yes Total** |  |  | **92.85072727** |
| **Yes Total** |  |  |  | **98.38226804** |
| **Grand Total** |  |  |  | **160.8766771** |

Since, SW has the maximum difference between categories, hence SW can be treated as the best predictor for predicting FARE.

1. **i.** The categorical attributes are:
2. SW
3. VACATION
4. GATE
5. SLOT

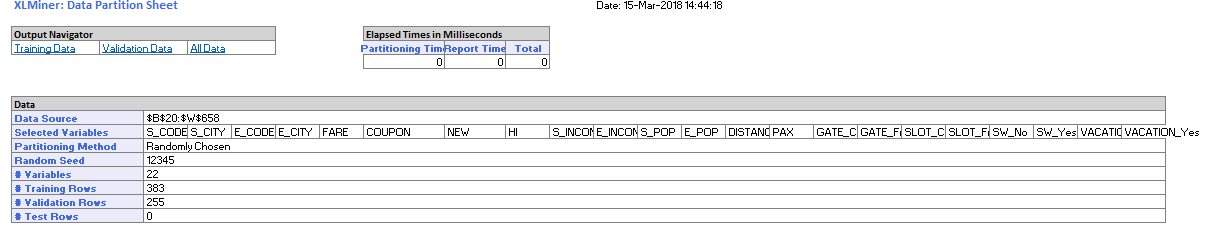
For creating dummies for categorical attributes, we can follow the following steps:

Go to XLMiner Platform🡪 Transform 🡪 Transform Categorical Data 🡪 Create Dummies. When you click Create Dummies, select all the categorical variables for which dummies must be created. As soon as we click OK, a new sheet will be formed with dummy values of all the categorical attributes.

For partitioning data, select the categories and go to XLMiner Platform🡪 Partition 🡪 Select the dataset to be partitioned 🡪 Pick up rows randomly 🡪 select the percentage as 60% for training set and 40% for validation set 🡪 click OK.

The generated file will show the data partition.

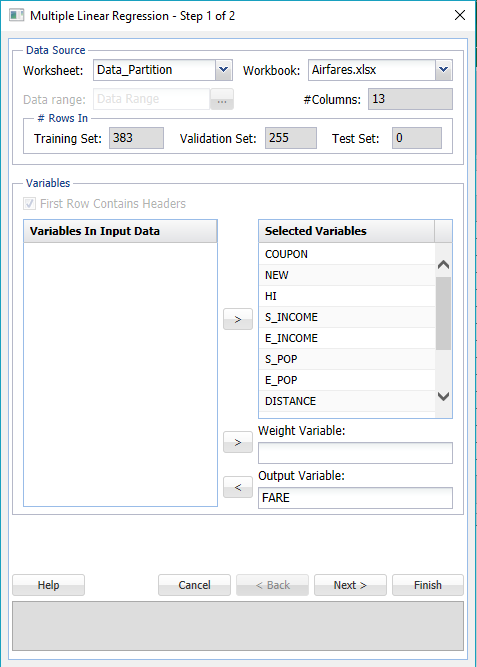
In this case, the output is

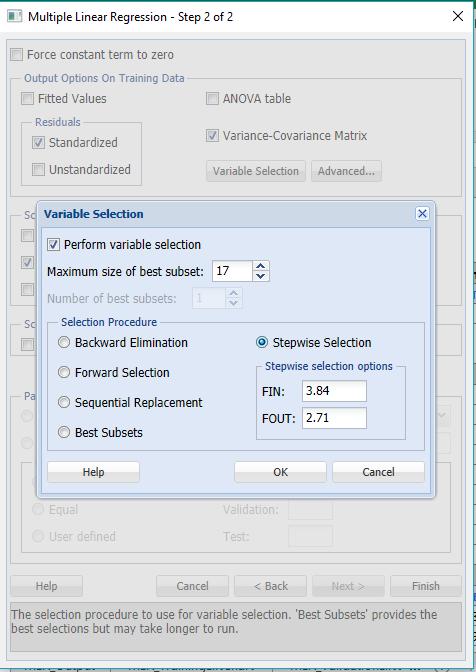


**ii.** The dataset must be divided in 2 parts namely training set and validation set as training set can be used to train the model according to the data and what output it should generate. Further, validation set will validate the accuracy and performance of the model. To generate an efficient application will good accuracy, training and validation must be applied to the dataset. The training set is used to train the model for estimating what the output should be and validation set is used to test the data if it is performing efficiently and the accuracy is reasonable or not.

**iii.**  For generating stepwise regression model, follow the following steps:

Go to XLMiner Platform 🡪 Predict 🡪 Multiple Linear Regression 🡪 select the datasource, input variables and FARE as the output variable 🡪 Next 🡪 Variable Selection 🡪 Check Perform Variable Selection 🡪 Click Stepwise Selection 🡪 Ok 🡪 Finish.





The generated file will have the following output:

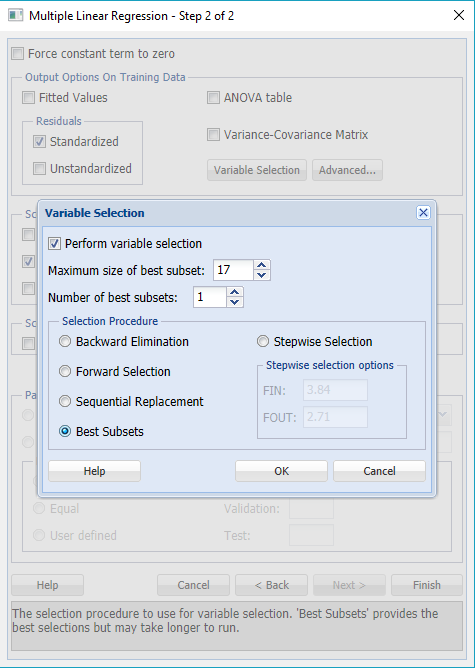
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Predictors** | |  |  |  |
|  |  |  |  |  |
|  | **Tolerance for Entering the Model** | | | 9.07264E-06 |
|  |  |  |  |  |
|  | **Included** | | **Excluded** | |
|  | **Predictor** | **Criteria** | **Predictor** | **Criteria** |
|  | **Intercept** | 1.01429635 | **GATE\_Constrained** | 3.23185E-15 |
|  | **COUPON** | 2.877009329 | **SLOT\_Controlled** | 7.77959E-15 |
|  | **NEW** | 15.83731177 | **SW\_No** | 8.34823E-15 |
|  | **HI** | 32251.39024 | **VACATION\_No** | 6.21439E-15 |
|  | **S\_INCOME** | 99174.25422 |  |  |
|  | **E\_INCOME** | 219391.7811 |  |  |
|  | **S\_POP** | 106685127.2 |  |  |
|  | **E\_POP** | 69411958.24 |  |  |
|  | **DISTANCE** | 11519.3443 |  |  |
|  | **PAX** | 215642.951 |  |  |
|  | **GATE\_Free** | 6.943145248 |  |  |
|  | **SLOT\_Free** | 8.020632568 |  |  |
|  | **SW\_Yes** | 8.779115831 |  |  |
|  | **VACATION\_Yes** | 8.149752609 |  |  |

REGRESSION MODEL

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| **Input Variables** | **Coefficient** | **Std. Error** | **t-Statistic** | **P-Value** | **CI Lower** | **CI Upper** | **RSS Reduction** |
| **Intercept** | -23.21351 | 35.41757 | -0.655423 | 0.512604 | -92.859101 | 46.432089 | 9869485.5 |
| **COUPON** | 10.24575 | 16.168454 | 0.6336878 | 0.526677 | -21.548118 | 42.039621 | 597050.18 |
| **NEW** | -1.759538 | 2.3379897 | -0.752586 | 0.452179 | -6.3569933 | 2.8379166 | 7462.8744 |
| **HI** | 0.009123 | 0.0012978 | 7.0292571 | 1.01E-11 | 0.0065708 | 0.011675 | 116785.51 |
| **S\_INCOME** | 0.001255 | 0.0006629 | 1.8927249 | 0.059177 | -4.885E-05 | 0.0025583 | 165817.2 |
| **E\_INCOME** | 0.001846 | 0.0004849 | 3.8071014 | 0.000165 | 0.00089259 | 0.0027997 | 229129.62 |
| **S\_POP** | 4.42E-06 | 8.332E-07 | 5.299898 | 2E-07 | 2.7776E-06 | 6.055E-06 | 62539.538 |
| **E\_POP** | 4.31E-06 | 9.833E-07 | 4.3846666 | 1.52E-05 | 2.3778E-06 | 6.245E-06 | 84148.763 |
| **DISTANCE** | 0.074126 | 0.0046118 | 16.072962 | 1.93E-44 | 0.06505705 | 0.0831946 | 266386.4 |
| **PAX** | -0.000894 | 0.0001971 | -4.5348 | 7.81E-06 | -0.0012816 | -0.000506 | 57550.662 |
| **GATE\_Constrained** | 0 | 0 | N/A | N/A | 0 | 0 | 0 |
| **GATE\_Free** | -17.39758 | 5.3407146 | -3.257538 | 0.001228 | -27.899637 | -6.895528 | 20061.387 |
| **SLOT\_Controlled** | 0 | 0 | N/A | N/A | 0 | 0 | 0 |
| **SLOT\_Free** | -17.45964 | 5.0223942 | -3.476357 | 0.000569 | -27.335741 | -7.583531 | 25940.037 |
| **SW\_No** | 0 | 0 | N/A | N/A | 0 | 0 | 0 |
| **SW\_Yes** | -35.10769 | 4.9025704 | -7.161078 | 4.37E-12 | -44.748172 | -25.46720 | 46856.119 |
| **VACATION\_No** | 0 | 0 | N/A | N/A | 0 | 0 | 0 |
| **VACATION\_Yes** | -33.74915 | 4.7615284 | -7.087881 | 6.97E-12 | -43.112281 | -24.38601 | 64833.593 |

**iv.** The exhaustive regression model can also be generated in the same way as stepwise regression.

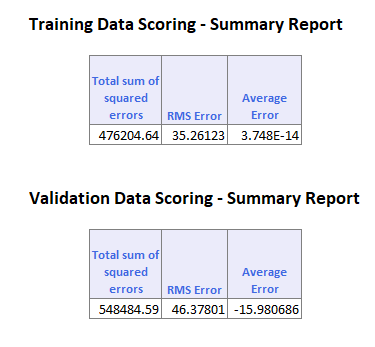
The steps are: go to XLMiner Platform 🡪 Predict 🡪 Multiple Linear Regression 🡪 select the datasource, input variables and FARE as the output variable 🡪 Next 🡪 Variable Selection 🡪 Perform Variable Selection 🡪 Best Subsets 🡪 OK.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| The output is as follows:  **Model Predictors** | |  |  |  |
|  |  |  |  |  |
|  | **Tolerance for Entering the Model** | | | 9.07264E-06 |
|  |  |  |  |  |
|  | **Included** | | **Excluded** | |
|  | **Predictor** | **Criteria** | **Predictor** | **Criteria** |
|  | **COUPON** | 2.877009329 | **Intercept** | 2.47302E-15 |
|  | **NEW** | 15.83731177 | **GATE\_Constrained** | 7.15644E-15 |
|  | **HI** | 32251.39024 | **SLOT\_Controlled** | 1.09931E-14 |
|  | **S\_INCOME** | 99174.25422 | **VACATION\_No** | 4.17366E-15 |
|  | **E\_INCOME** | 219391.7811 |  |  |
|  | **S\_POP** | 106685127.2 |  |  |
|  | **E\_POP** | 69411958.24 |  |  |
|  | **DISTANCE** | 11519.3443 |  |  |
|  | **PAX** | 215642.951 |  |  |
|  | **GATE\_Free** | 6.943145248 |  |  |
|  | **SLOT\_Free** | 8.020632568 |  |  |
|  | **SW\_No** | 1.01429635 |  |  |
|  | **SW\_Yes** | 8.779115831 |  |  |
|  | **VACATION\_Yes** | 8.149752609 |  |  |

There is a change in the attributes when we compare model (iii) and model (iv).

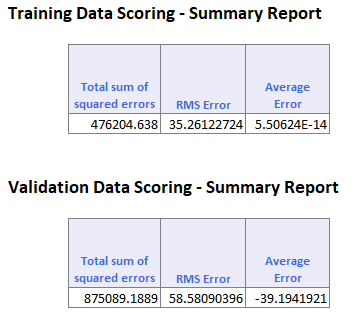
1. The measure for model (iii) are:



**LIFT CHARTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decile** | **Mean** | **Std.Dev.** | **Min.** | **Max.** |
| **1** | 310.9672 | 43.47528805 | 224.17 | 402.02 |
| **2** | 254.0036 | 41.26156059 | 125.9 | 325.02 |
| **3** | 210.4856 | 63.05994417 | 83.74 | 297.2 |
| **4** | 170.6156 | 60.30814083 | 59.8 | 297.2 |
| **5** | 135.8732 | 47.32854017 | 47.85 | 252.97 |
| **6** | 147.1364 | 45.66369281 | 57.29 | 219.38 |
| **7** | 123.4676 | 46.11051193 | 57.57 | 207.17 |
| **8** | 98.5292 | 34.47457323 | 55.16 | 188.11 |
| **9** | 91.5732 | 34.96007636 | 42.47 | 199.8 |
| **10** | 76.1772 | 22.22266308 | 49.02 | 114.35 |

The measures for model (iv) are:



**LIFT CHART**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decile** | **Mean** | **Std.Dev.** | **Min.** | **Max.** |
| **1** | 310.9672 | 43.47528805 | 224.17 | 402.02 |
| **2** | 254.0036 | 41.26156059 | 125.9 | 325.02 |
| **3** | 210.4856 | 63.05994417 | 83.74 | 297.2 |
| **4** | 170.6156 | 60.30814083 | 59.8 | 297.2 |
| **5** | 135.8732 | 47.32854017 | 47.85 | 252.97 |
| **6** | 147.1364 | 45.66369281 | 57.29 | 219.38 |
| **7** | 123.4676 | 46.11051193 | 57.57 | 207.17 |
| **8** | 98.5292 | 34.47457323 | 55.16 | 188.11 |
| **9** | 91.5732 | 34.96007636 | 42.47 | 199.8 |
| **10** | 76.1772 | 22.22266308 | 49.02 | 114.35 |

1. The formula used for calculating the average fare is:

Y = http://www.stat.yale.edu/Courses/1997-98/101/beta.gif0 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif1*x*1 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif2*x*2 + ... + http://www.stat.yale.edu/Courses/1997-98/101/beta.gifp*x*p

The values of beta can be found out from the regression model. The values of beta are coefficient of all the categories in regression model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Input variable** | **Coefficient** | **Given value** | **Product** |
| **Intercept** | 0 |  |  |
| **COUPON** | 10.24575 | 1.202 | 12.31539 |
| **NEW** | -1.75954 | 3 | -5.27862 |
| **HI** | 0.009123 | 4442.141 | 40.52565 |
| **S\_INCOME** | 0.001255 | 28760 | 36.0938 |
| **E\_INCOME** | 0.001846 | 27664 | 51.06774 |
| **S\_POP** | 4.42E-06 | 4557004 | 20.14196 |
| **E\_POP** | 4.31E-06 | 3195503 | 13.77262 |
| **DISTANCE** | 0.074126 | 1976 | 146.47 |
| **PAX** | -0.00089 | 12782 | -11.3 |
| **GATE\_Free** | -17.3976 | 1 | -17.3976 |
| **SLOT\_Free** | -17.4596 | 1 | -17.4596 |
| **SW\_No** | -23.2135 | 1 | -23.2135 |
| **SUM** |  |  | **245.7378** |

Thus, the average fare is **$245.73.**

1. If southwest decides to cover this route, the value of SW\_YES will be one.

The average fare can be calculated as:

|  |  |  |  |
| --- | --- | --- | --- |
| **Input variable** | **Coefficient** | **Given value** | **Product** |
| **Intercept** | 0 |  |  |
| **COUPON** | 10.24575 | 1.202 | 12.31539 |
| **NEW** | -1.75954 | 3 | -5.27862 |
| **HI** | 0.009123 | 4442.141 | 40.52565 |
| **S\_INCOME** | 0.001255 | 28760 | 36.0938 |
| **E\_INCOME** | 0.001846 | 27664 | 51.06774 |
| **S\_POP** | 4.42E-06 | 4557004 | 20.14196 |
| **E\_POP** | 4.31E-06 | 3195503 | 13.77262 |
| **DISTANCE** | 0.074126 | 1976 | 146.47 |
| **PAX** | -0.00089 | 12782 | -11.3 |
| **GATE\_Free** | -17.3976 | 1 | -17.3976 |
| **SLOT\_Free** | -17.4596 | 1 | -17.4596 |
| **SW\_Yes** | -58.3212 | 1 | -58.3212 |
|  |  |  |  |
| **SUM** |  |  | **210.630** |

Thus, the predicted average fare is **$210.630**.

The reduction in average fare would be 245.7378- 210.630= $**35.1077**

We cannot predict PAX which is the count of number of passengers on that route for a period as the number of passengers may vary. Also, SW which is whether the flight is travelling in the SouthWest direction or not since a flight may travel in SouthWest direct which cannot be predicted earlier.

From a new airport, we can predict the values of following attributes:

S\_CODE: starting airport’s code a

E\_CODE: ending airport’s code

S\_CITY: starting city

E\_CODE: ending city

COUPON: average number of coupons applied on a flight

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

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

FARE: average fare on that route

Since, the values of S\_C0DE, E\_CODE, S\_CITY AND E\_CITY will always be fixed, it is easy to predict these attributes. The COUPON can be predicted as it may not vary with new airport.

The number of NEW carriers can be predicted. It can be predicted whether a flight goes on VACATION route or not. The S\_INCOME, E\_INCOME, S\_POP AND E\_POP are fixed measures as well which do not change with new airport. SLOT, GATE, DISTANCE and FARE also do not change with new airport.

1. The factors which are known before the flights begin to operate on new route are as follows:

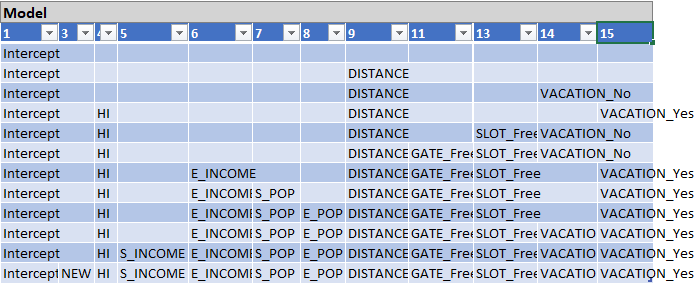
* COUPON
* NEW
* HI
* S\_INCOME
* E\_INCOME
* S\_POP
* E\_INCOME
* DISTANCE
* VACATION\_YES
* VACATION \_NO
* SLOT\_CONTROLLED
* SLOT\_FREE
* GATE\_CONTROLLED
* GATE\_FREE.

Use the same steps as in (iv) for selecting a model.

The output for the Exhaustive regression is

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Predictors** | |  |  |  |
|  |  |  |  |  |
|  | **Tolerance for Entering the Model** | | | 9.07264E-06 |
|  |  |  |  |  |
|  | **Included** | | **Excluded** | |
|  | **Predictor** | **Criteria** | **Predictor** | **Criteria** |
|  | **COUPON** | 3.231043246 | **Intercept** | 1.71763E-15 |
|  | **NEW** | 15.84554467 | **GATE\_Constrained** | 1.04348E-14 |
|  | **HI** | 32444.09091 | **SLOT\_Controlled** | 2.74231E-15 |
|  | **S\_INCOME** | 99432.72324 |  |  |
|  | **E\_INCOME** | 219391.7811 |  |  |
|  | **S\_POP** | 106685127.2 |  |  |
|  | **E\_POP** | 69411958.24 |  |  |
|  | **DISTANCE** | 11844.66036 |  |  |
|  | **GATE\_Free** | 7.367263209 |  |  |
|  | **SLOT\_Free** | 8.448469283 |  |  |
|  | **VACATION\_No** | 1.089264248 |  |  |
|  | **VACATION\_Yes** | 8.233710356 |  |  |

Model



1. The formula used for calculating the average fare is:

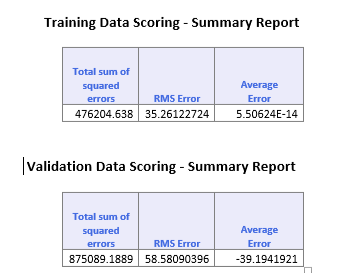
Y = http://www.stat.yale.edu/Courses/1997-98/101/beta.gif0 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif1*x*1 + http://www.stat.yale.edu/Courses/1997-98/101/beta.gif2*x*2 + ... + http://www.stat.yale.edu/Courses/1997-98/101/beta.gifp*x*p

The values of beta can be found out from the regression model. The values of beta are coefficient of all the categories in regression model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Variables** | **Coefficient** | **Values** | **Product** |
| **Intercept** | 0 |  |  |
| **COUPON** | 56.12835 | 1.202 | 67.46628 |
| **NEW** | -2.14355 | 3 | -6.43065 |
| **HI** | 0.012048 | 4442.141 | 53.51891 |
| **S\_INCOME** | 0.002521 | 28760 | 72.50396 |
| **E\_INCOME** | 0.002023 | 27664 | 55.96427 |
| **S\_POP** | 3.23E-06 | 4557004 | 14.71912 |
| **E\_POP** | 3.65E-06 | 3195503 | 11.66359 |
| **DISTANCE** | 0.073642 | 1976 | 145.51 |
| **GATE\_Free** | -27.2937 | 1 | -27.2937 |
| **SLOT\_Free** | -23.4959 |  | -23.4959 |
| **VACATION\_No** | -131.961 | 1 | -131.961 |
| **SW\_No** | 0 | 0 |  |
| **SUM** |  |  | **232.1653** |

Thus, the average predicted fare is **$232.16**

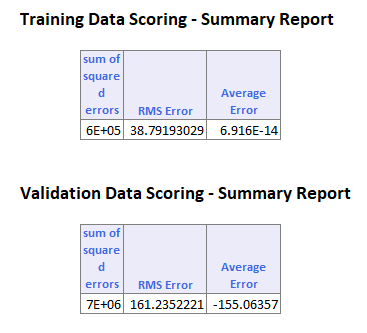
1. The measures for model (iv) are:



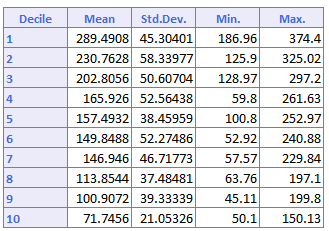
**LIFT CHART**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Decile** | **Mean** | **Std.Dev.** | **Min.** | **Max.** |
| **1** | 291.4016 | 43.32816 | 169.41 | 374.4 |
| **2** | 234.832 | 55.51679 | 125.9 | 325.02 |
| **3** | 190.9232 | 50.12512 | 97.46 | 297.2 |
| **4** | 178.1868 | 54.48251 | 59.8 | 286.54 |
| **5** | 154.5276 | 35.3632 | 107.51 | 244.5 |
| **6** | 156.0028 | 49.37171 | 56.43 | 240.88 |
| **7** | 143.17 | 43.8678 | 52.92 | 229.84 |
| **8** | 117.3976 | 39.17206 | 57.57 | 202.77 |
| **9** | 84.1832 | 32.75395 | 45.11 | 199.8 |
| **10** | 80.894 | 24.29225 | 50.1 | 116.78 |

The measures for model (ix) is:



**LIFT CHART**



We can observe from the above measures that the RMSE error and average error for model (ix) is more as compared to that of model (iv). Thus, model (ix) is less accurate as compared to model (iv).