

**MS5107 – BUSINESS MODELLING & ANALYTICS**

**ASSIGNMENT NO : 2**

**Group No : 32**

**1. Jayakarthi Boovendran - 19230487**

**2. SreeKrishnan Chelakara Hariharan - 19230828**

**3. Naveen Banagani - 19230199**

**4. Rajat Agarwal - 19230537**

**INTRODUCTION:**

The objective of this assignment is to build a business model that predicts the average air fare on a new flying route in the US. For this purpose, a major US airline collected information on 638 air routes in the United States. Some factors that this airline was able to identify about the new routes were the distance travelled, demographics of the city where the new city is located, and whether this city is a vacation destination. However, there are some important factors yet unknown such as the number of passengers that will travel this route, and whether Southwest Airlines or another discount airline are planning to travel on these new routes. The presence of these discount airlines is believed to reduce the fares greatly. Therefore, in order to come up with a strategy on pricing we analyzed both the situation where Southwest Airline will travel and a situation where it will not travel on these new routes. We performed various prediction models including Linear Regression, Regression tree, Neural Networks, and Ensemble (bagging, boosting and random Trees).

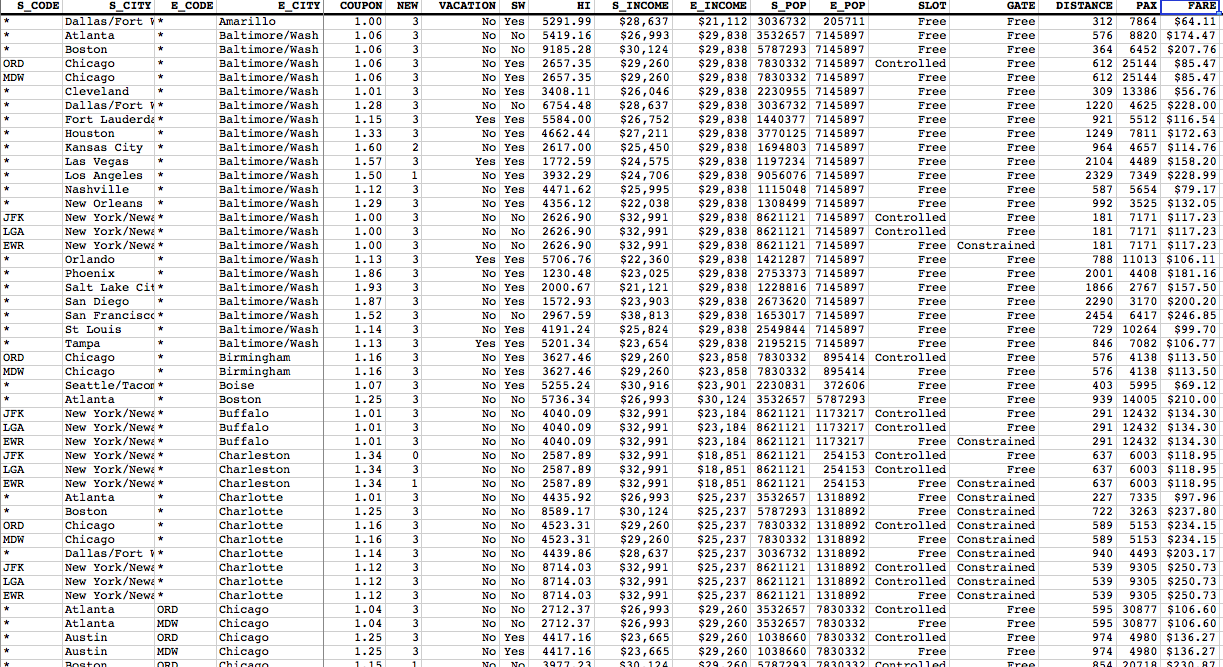
# **DATASET:**

The data set contains real-time data from 638 current U.S. air routes. The following are the 18 data points that we use in the prediction process.

Airfares on new routes.

|  |  |
| --- | --- |
| 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 |

**Table 1: Airfare Dataset variables**

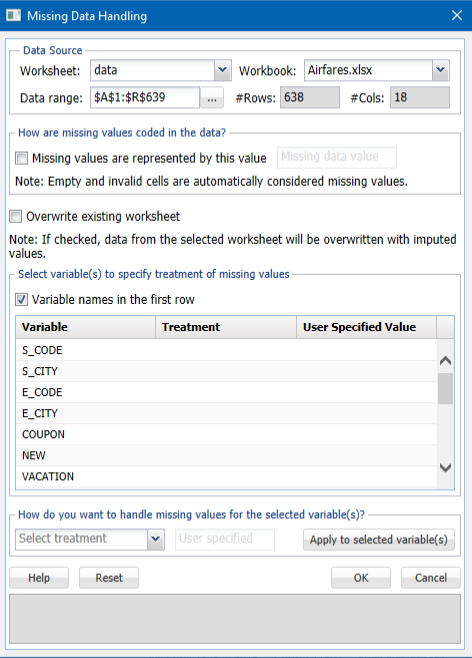


**Fig 1: Airfare dataset**

**DATA PREPARATION**

**Missing Data Handling**

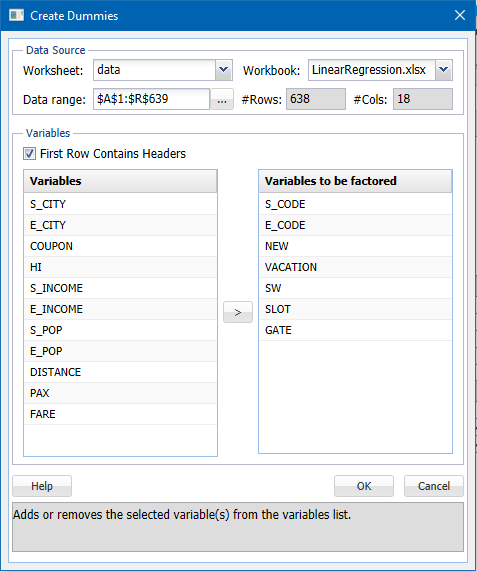
* Click on Data Mining Tab from the Airfare dataset
* Click on Transform -> Missing Data Handling under ‘Data Analysis’



**Creating Dummies For Categorical Data**

Here we see the data for Vacation and South West Airlines are in a Boolean format (Yes/No) which cannot be inputted into model as it is non-numeric. Therefore, we are converting the two columns into dummy variables where 0s would be No and the 1s would be Yes. This helps us to input these values into the models and for predicting the average airfare.

Click on Transform - Transform Categorical Data - Create Dummies from the ‘Data Analysis’ section



## **Removing Outliers**

In order to find the outliers, the following formula was used:

=IF(OR(E2<E$640-(3\*E$641),E2>E$640+(3\*E$641)),”OULIERS”,”OK”)

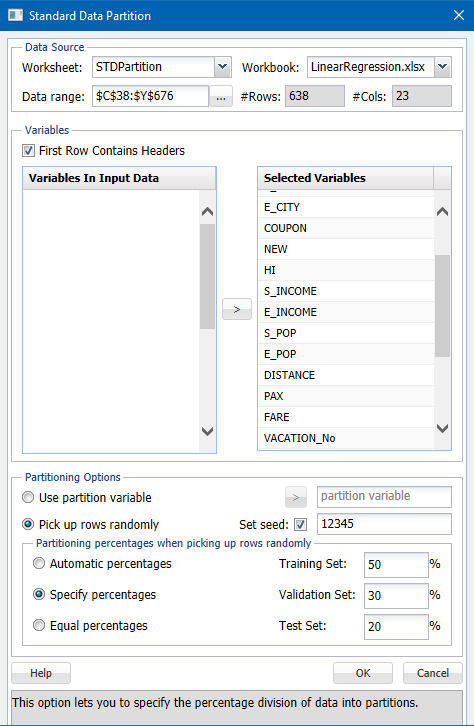
where, E$640 is the Average and E$641 is the Standard deviation of columns

In the given dataset, we found 82 outliers. It was agreed, however, not to eliminate the outliers because they are significant. In addition, some data in the data set could not be considered as outliers as they are major air routes with better income, length, etc. relative to other locations. Several outliers have been identified, therefore.

## **Data Partition**

Use the usual standard data partition with percentages of 50 percent of the data allocated randomly to the training set, 30 percent of the data allocated randomly to the validation set, and 20 percent of the data allocated randomly to the test set to divide the data into training and validation set.

On the Data Mining ribbon, from the Data Mining tab, select Partition - Standard Partition to open the *Standard Data Partition* dialog. Data Partition is shown below:



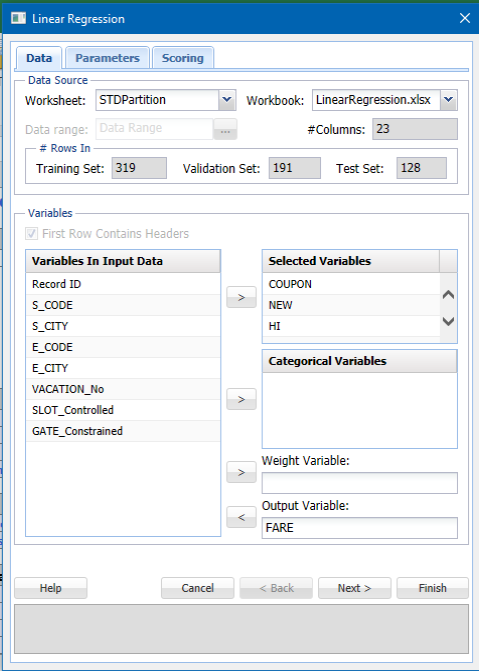
**Fig 2: Data Partition before MLR**

# **OBJECTIVE A**

# **Building Predictive Models**

# **LINEAR REGRESSION**:

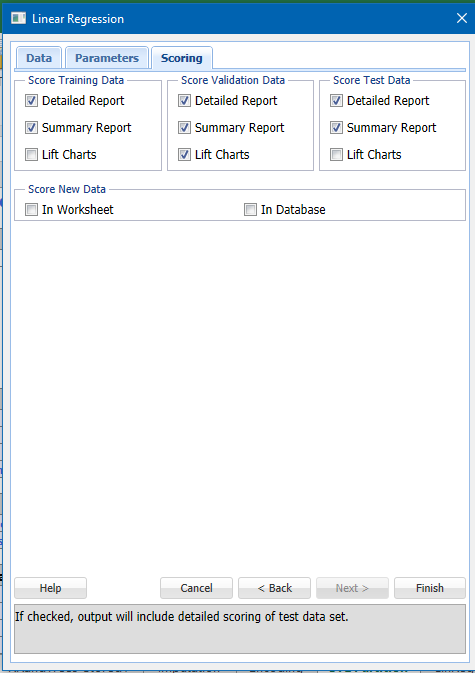
* Select a cell from the worksheet on the Data Partition. Select Predict-Multiple Linear Regression to open the Multiple Linear Regression dialog on the XLMINER ribbon from the Data Mining section.
* Select FARE at the Input Variable and select all remaining variables (except S CITY and E CITY) from the Selected output Variables list.
* S\_CITY and E\_CITY columns are not selected, since they do not contain any numeric data.



**Fig 3: LR Step 1**

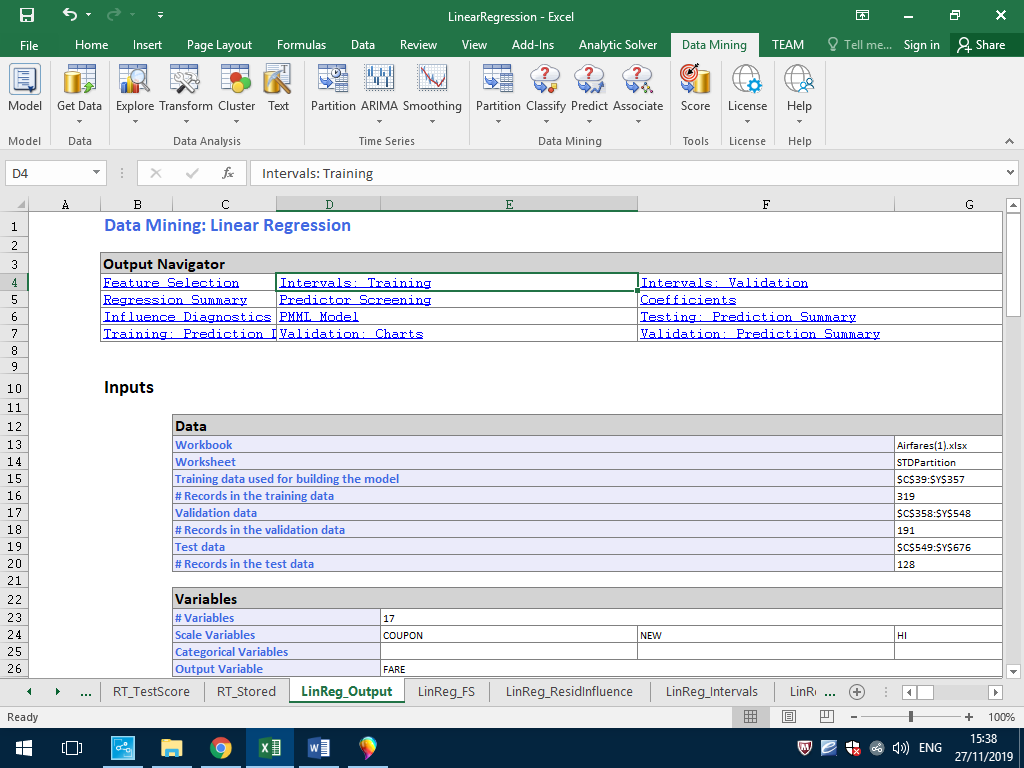
Click Next to advance to the Step 2 of 2 dialog.

* If there is a constant term in the formula, force is selected to zero.
* Select values that are fitted. The equipped values will be shown in the output when this choice is selected.
* Select ANOVA table. When this choice is chosen, the output displays the ANOVA table, which displays the effects of the variance.
* Under Residuals, select Standardized to display the Standardized Residuals in the output. Through dividing unstandardized residuals through the respective standard deviations, standardized residuals are obtained.
* Choose Detailed Report and Summary Report to generate all four documents in the performance under Score Training Data and Score Validation Information.



**Fig 4: MLR Step 2**

* Click finish. To find the output navigator, click the LR Output worksheet.
* To view the selected output or to show one of the selections made, press any reference on the output navigator.

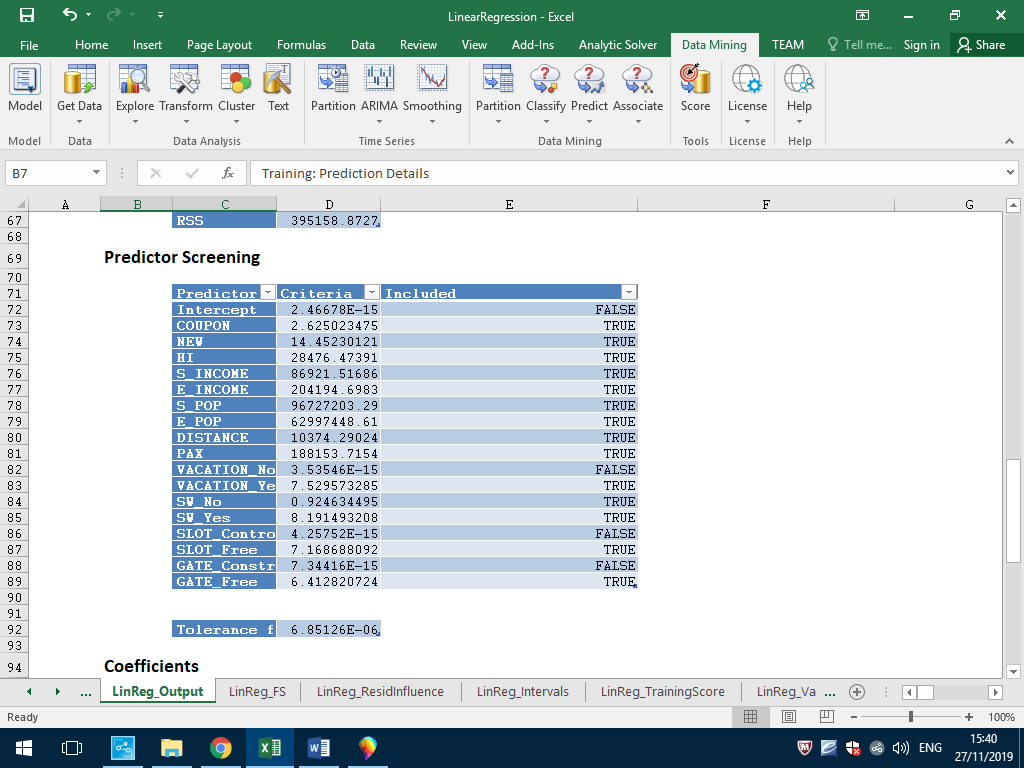


**Fig 5: LR – Output Navigator section**

* Inputs section shows a list of 17 variables
* Parameters/Options section shows the lists of options selected while selecting the LR model.

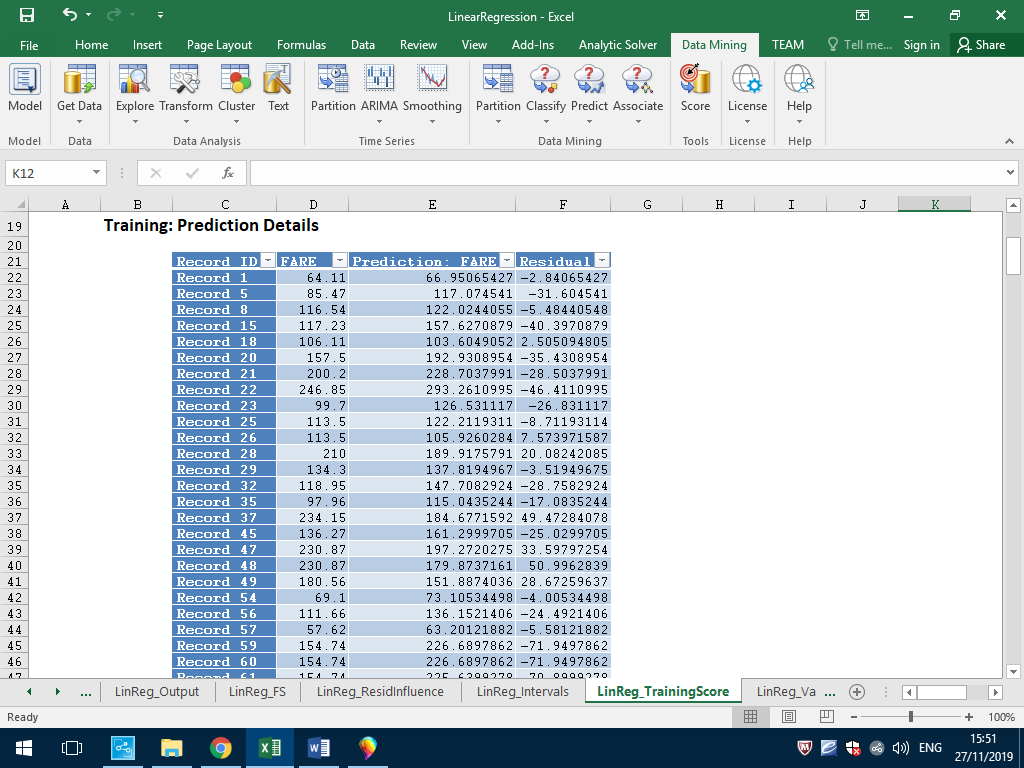
## **PREDICTORS SCREENING:**

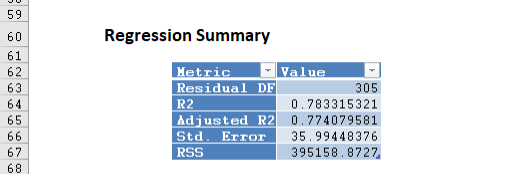
Clicking on the Predictors hyperlink will show the Model Predictors table from the output navigator. We can see from this model that no predictors are omitted. All predictors are eligible to enter the model that met the 6.85126E-06 sensitivity threshold. This denotes a threshold beyond which within machine precision a variance-covariance matrix is not exactly unique. Predictors that do not pass the test are omitted and we can see that the screenshot below does not include any excluded predictors



## **REGRESSION:**

On the Output Navigator, click the Regress.Model link to display the Regression Model Table.





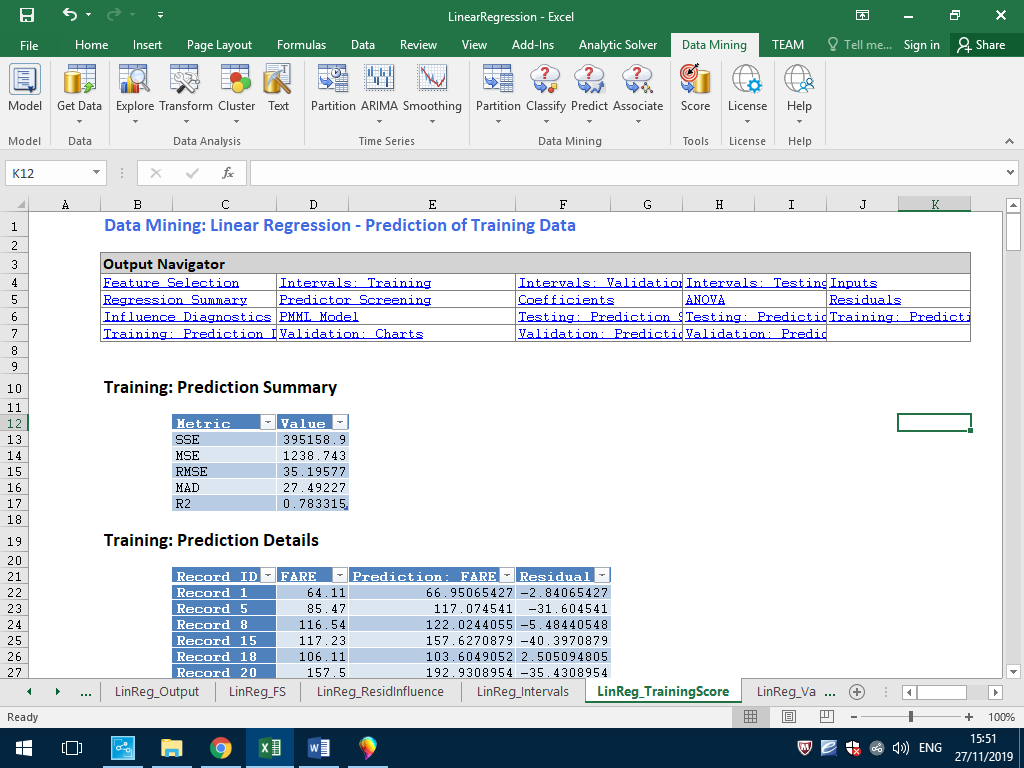
**Fig 7: MLR - Regression**

For each variable included in the model, the Regression table above shows coefficient, std.error, t-statistic, p-value, RSS reduction and squared error sum. The number of squared errors is determined by adding each variable in the system and proceeding with each variable as it appears in the dataset.

Residual degrees of freedom, R- Squared value, standard deviation type measure for the model (chi-square distribution) and the Residual Sum of square error (RSS) are shown in the regression model.

## **TRAINING DATA SCORING:**

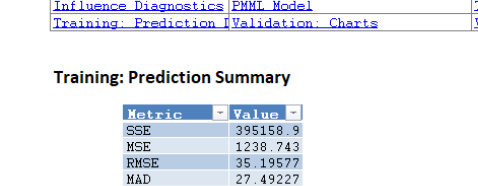
On the output navigator, click the Train.Score-Detailed Rep. to open the Linear Regression – Prediction of Training Data table.

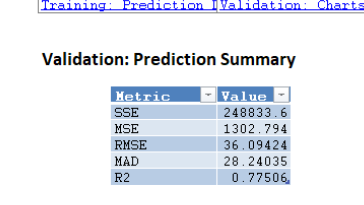


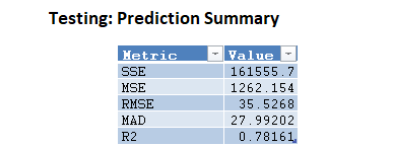
**Fig 8: LR - Training Score Detailed Report**

XLMiner provides 95% confidence intervals for the expected values. Prediction periods are usually used more commonly as they are a more reliable array for the expected value. The Interval of Prediction addresses possible future deviations from the average of the expected result. There is a 95% probability that the expected value falls within the period of estimation.

XLMiner displays Total sum of squared error summaries on the LR Output worksheet for both the training and validation sets. The total sum of square errors is the number of square errors (deviations from expected and real values) and the root mean square error (square root of average square error).



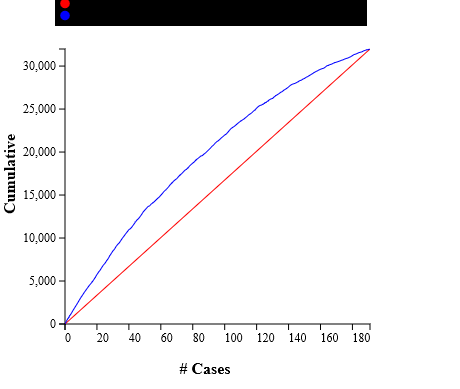


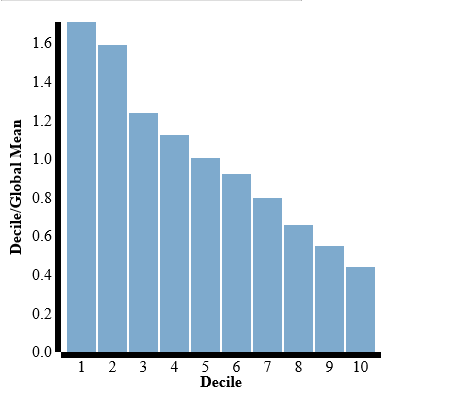


**Fig 9: MLR - Summary Reports**

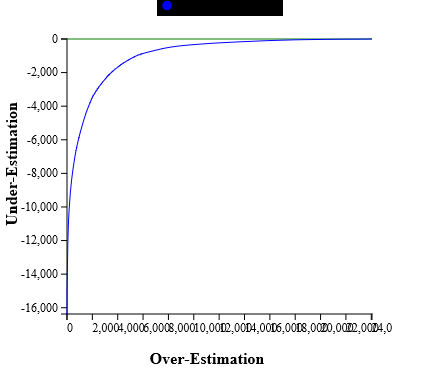
## **LIFT CHARTS AND RROC CURVES:**

* Lift charts and RROC curves (respectively on the MLR Training Lift Map and MLR Validation Lift Chart) are graphical aids for template performance assessment. Lift diagrams consist of a mean and a lift curve. The larger the area between the baseline and the lift curve, the better the model.
* RROC curves plot the output of retrogrades using graph overestimates (foreseen values too high) versus underestimations (foreseen values too low). The closer the curve to the upper left corner of the plot (the smaller the area above the curve) the more powerful the model is..





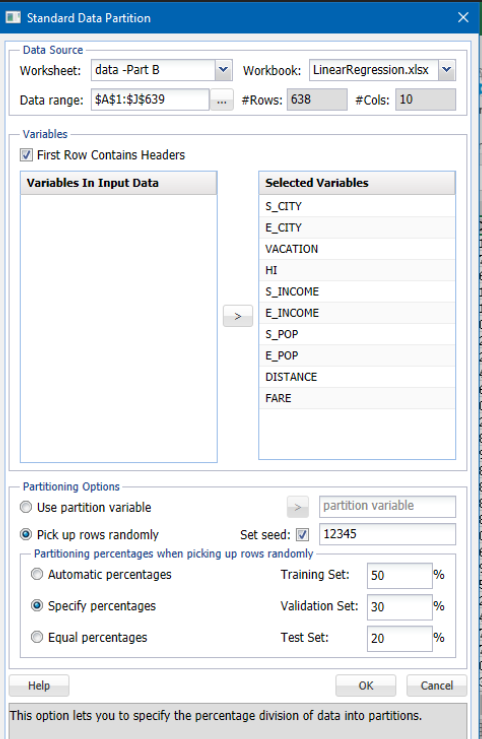
**Fig 10: LR - Lift chart & Decile wise lift chart**



**Fig 11: MLR - RROC Curve**

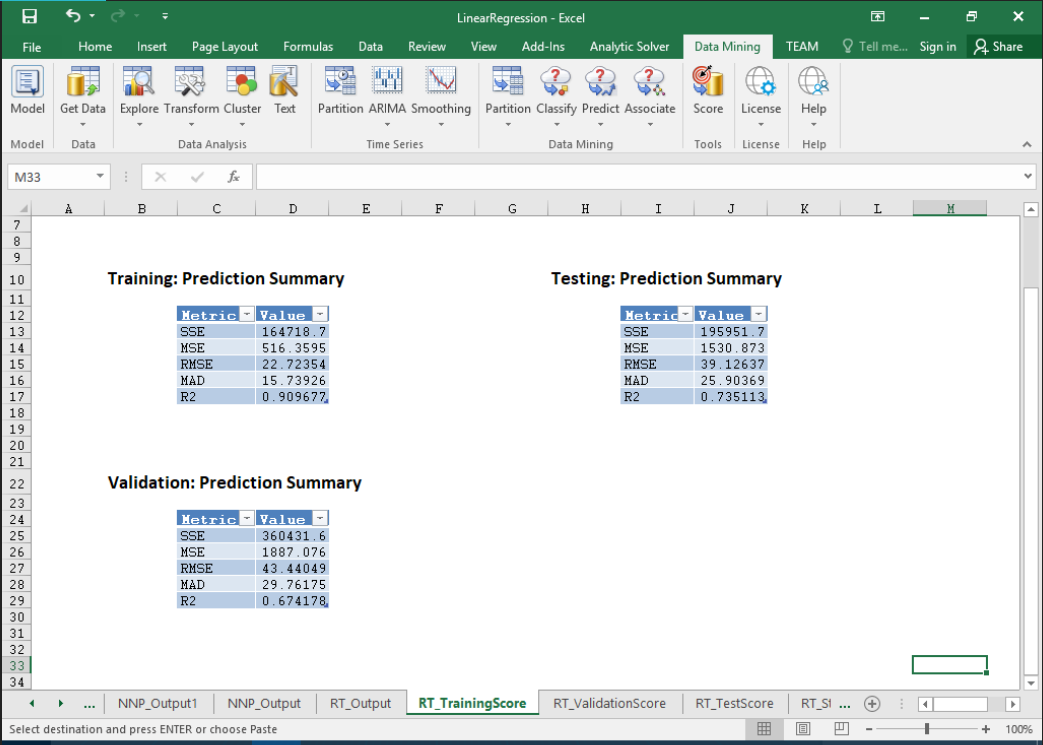
# **USING REGRESSION TREE MODEL:**

* The Regression Tree Model has been constructed the same way as that of the LR model.
* Dataset is to be partitioned into 3 categories: Validation, Training and Test Data, 50%,30% and 20% each.



**Fig 12: Regression Model - Data Partition**

* This partition the data as we have specified and now we can proceed by choosing the Prediction Model for the Regression Tree and selecting the required reports.
* In this case the Summary Reports are selected for all the partitioned data as seen below:

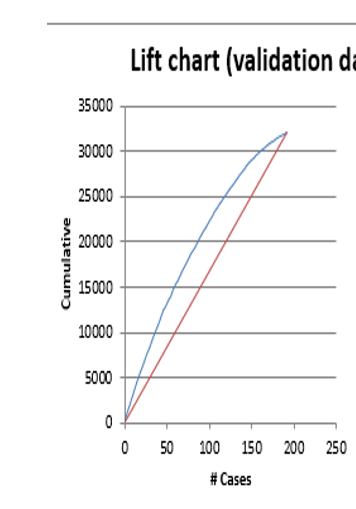


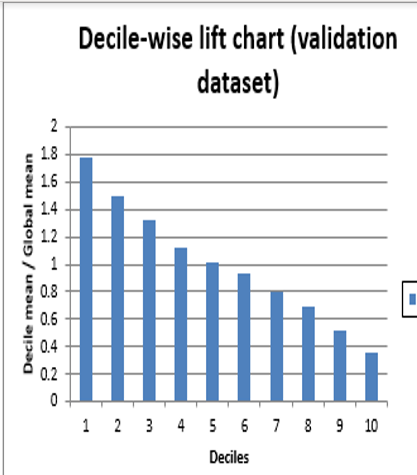
**Fig 13: Regression Model - Summary Report**

* RMS error for the Validation Summary Report was 35 when the first report was produced for the learning, validation and test data, respectively, with a partition of 50% 30% 20%.
* We modified the partition to the value s shown above and reduced the RMS to 43.44.

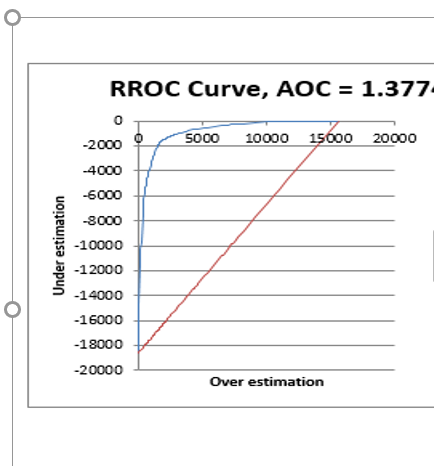
## **LIFT CHARTS AND RROC CURVES:**

* Lift Charts and RROC Curves (on the MLR\_Training Lift Chart and MLR\_Validation Lift Chart, respectively) are visual aids for measuring model performance. Lift Charts consist of a lift curve and a baseline. The greater the area between the lift curve and the baseline, the better the model.
* RROC curves plot the output of retrogrades using graph overestimates (foreseen values too high) versus underestimations (foreseen values too low). The closer the curve to the upper left corner of the plot (the smaller the area above the curve) the more powerful the model is.





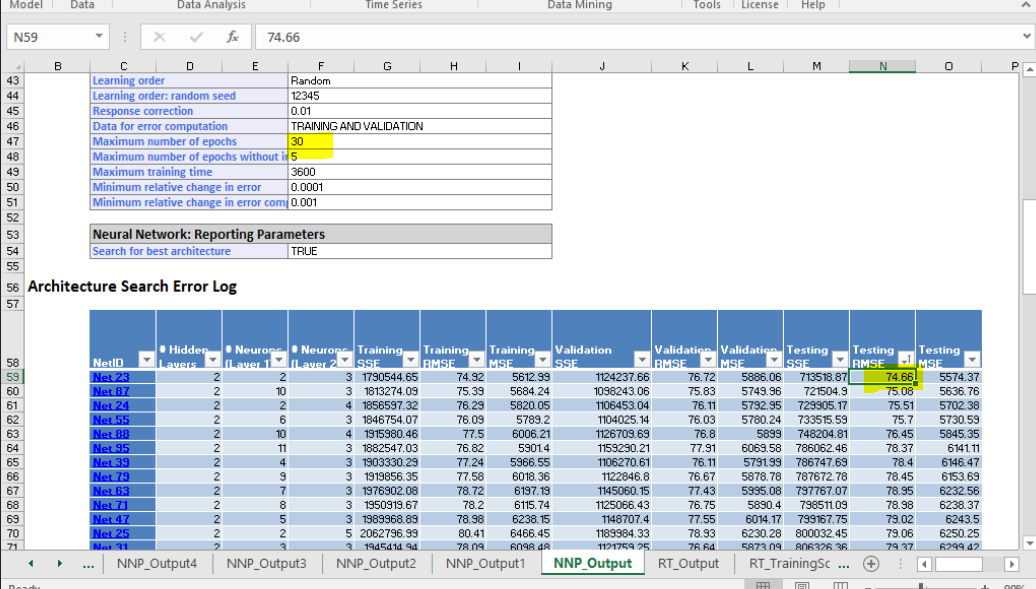
**Fig 14: Regression Model – Lift and Decile-wise Lift chart**

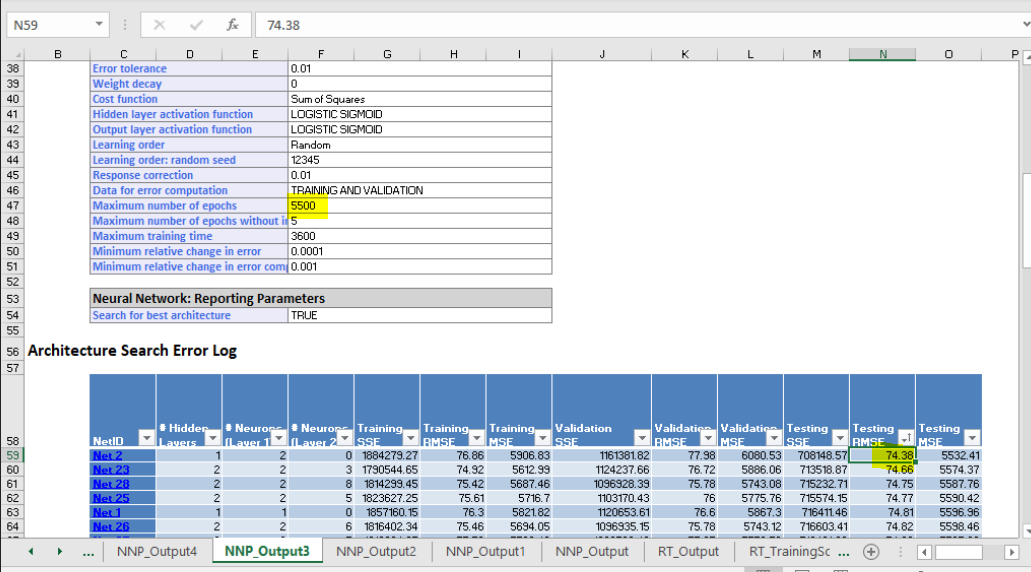


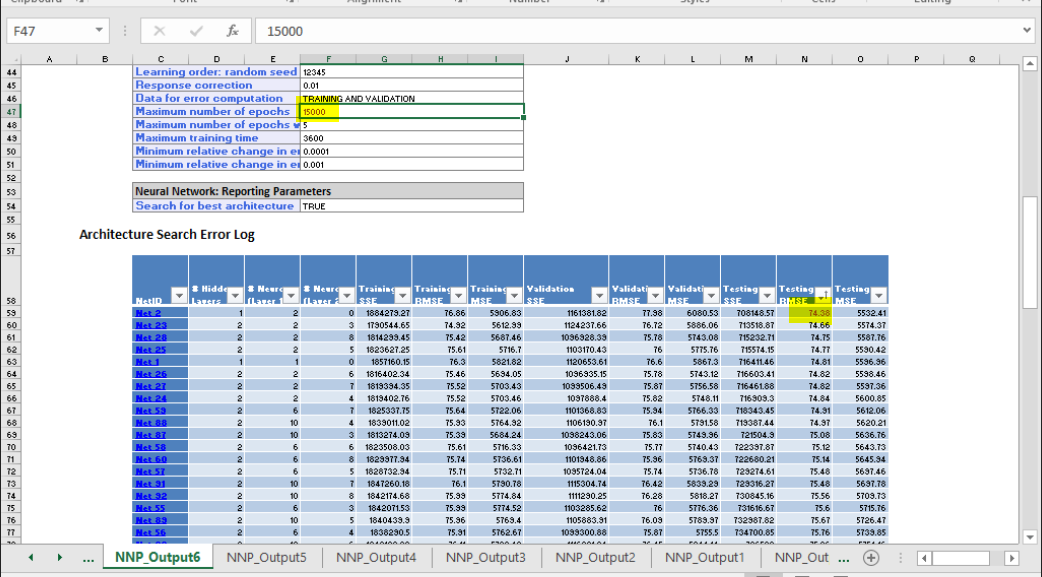
**Fig 15: Regression Model – RROC Curve**

* The area between the lift curve and the baseline is higher and from the RROC curve can be seen from the lift graph. We can see that the curve is nearest to the graph's top-left corner (the smaller the area above the curve).
* Based on our observations from the construction of the models and the graphs above, we concluded that the Regression Tree model is much better because it has a much better RMS value.

## **Neural Network:**



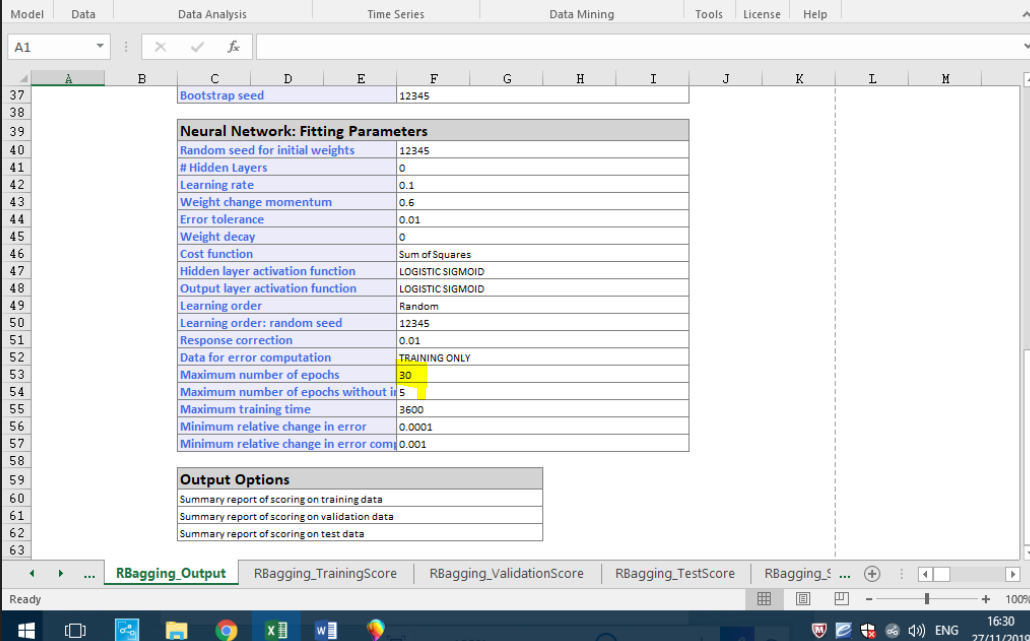


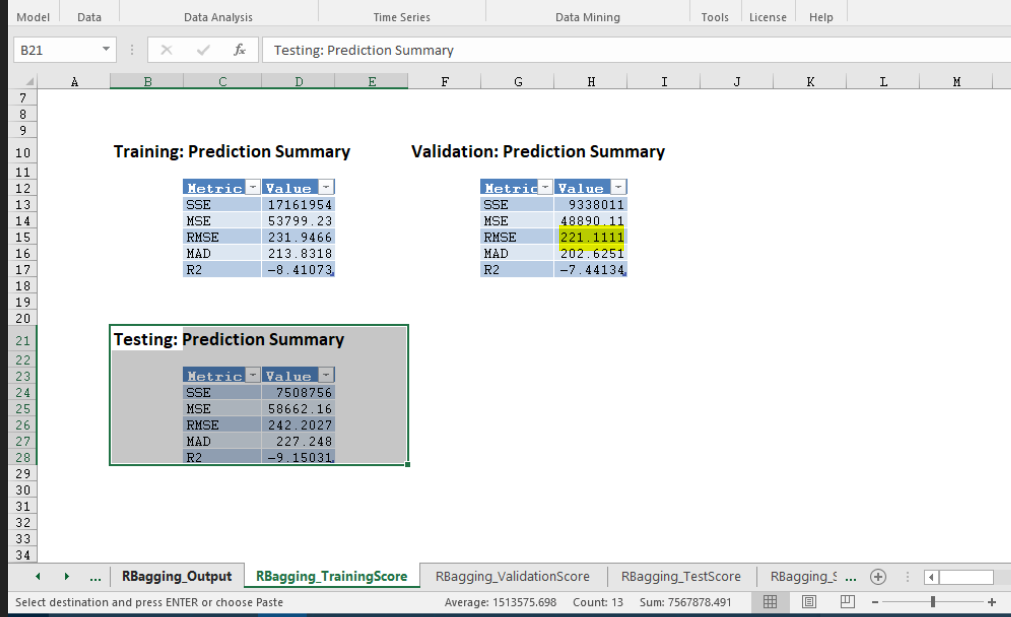


RMSE value remained constant which is 74.38.

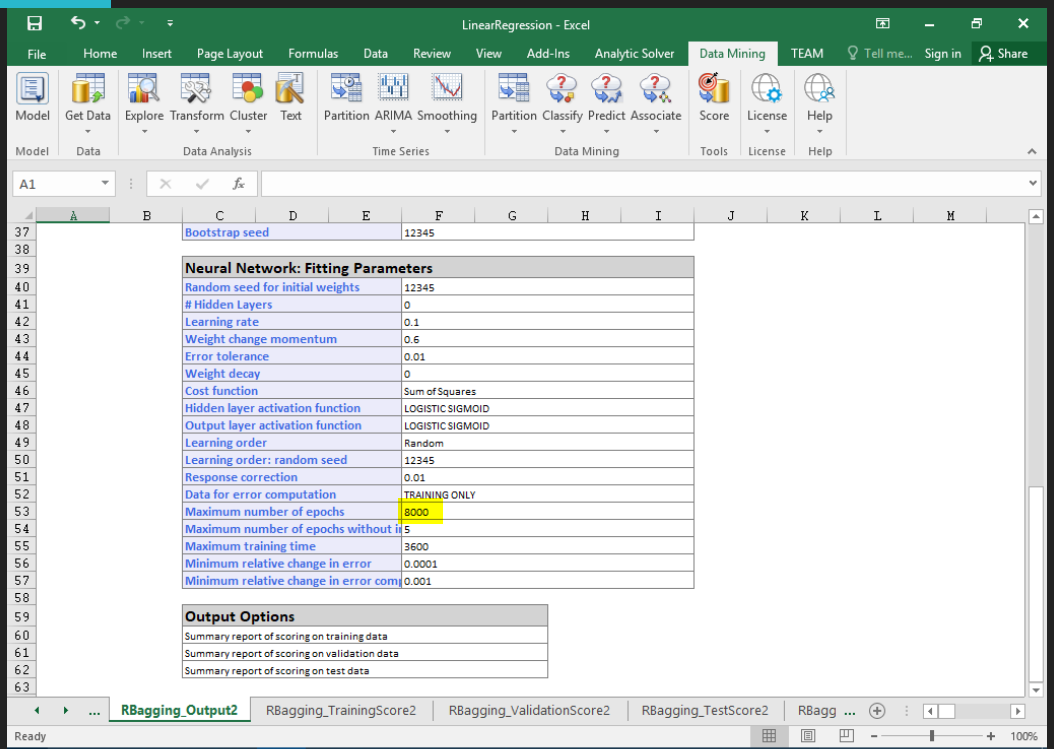
**Bagging:**

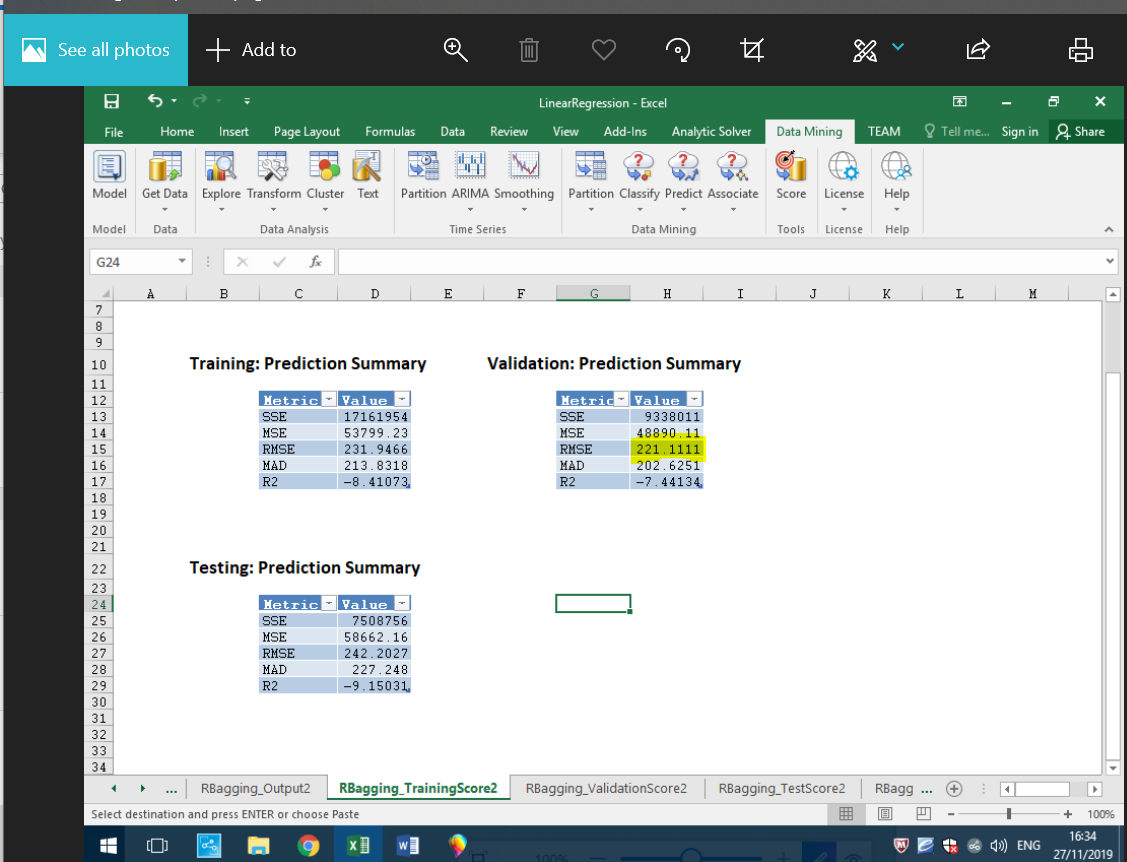
Epoch value – 30:





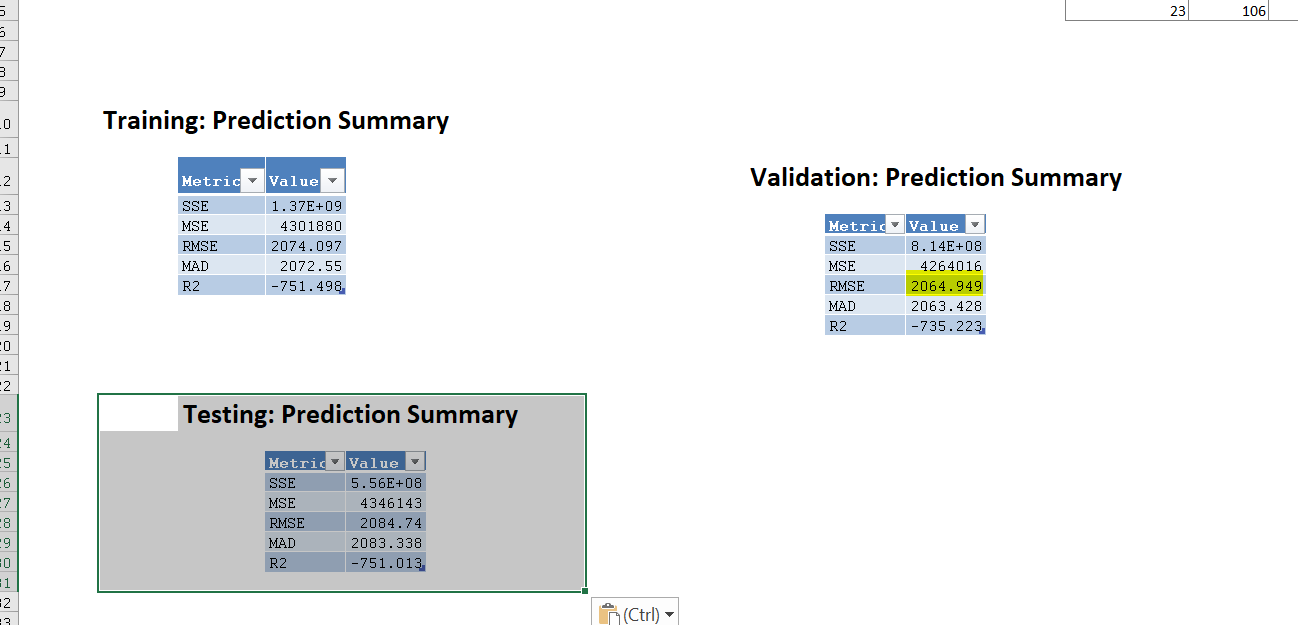
Epoch value 8000:





RMSE value remained constant which is 221.1111

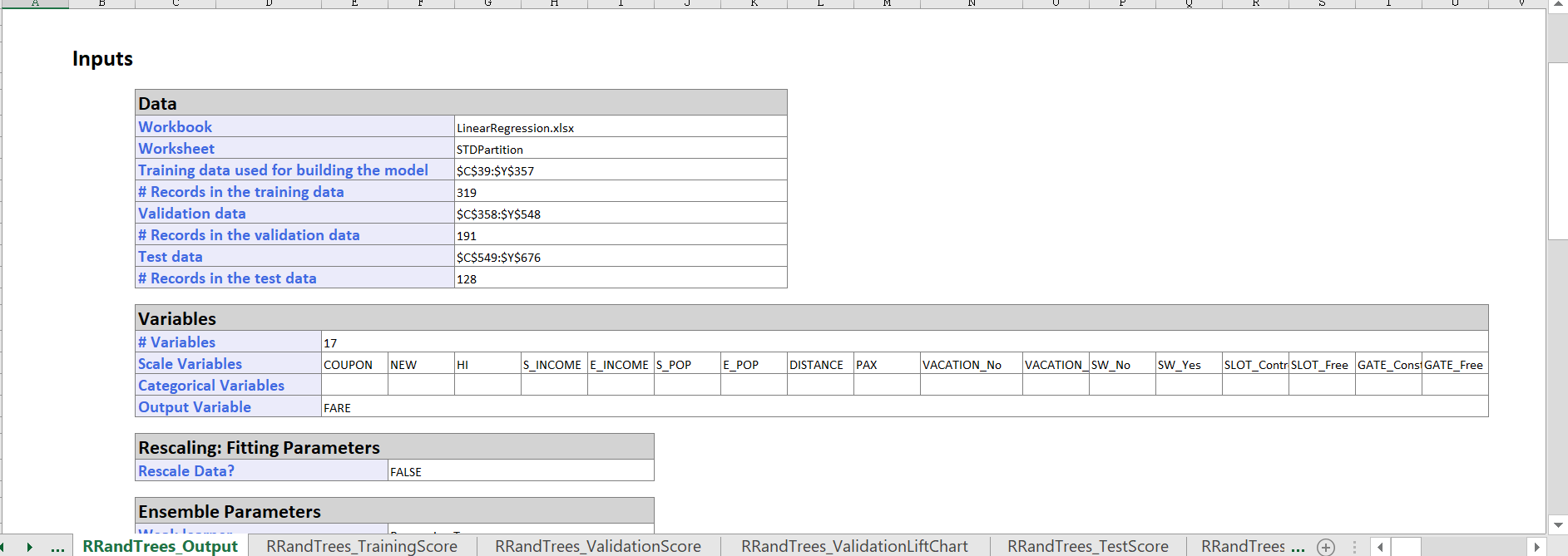
**Boosting Technique:**

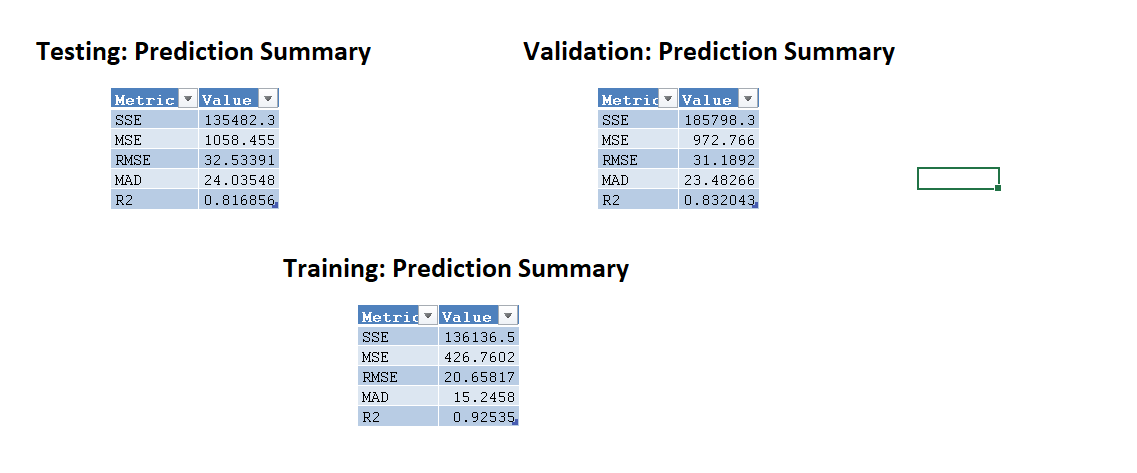


RMSE Value is 2064.949

**Random Tree:**

Random Tree Output:



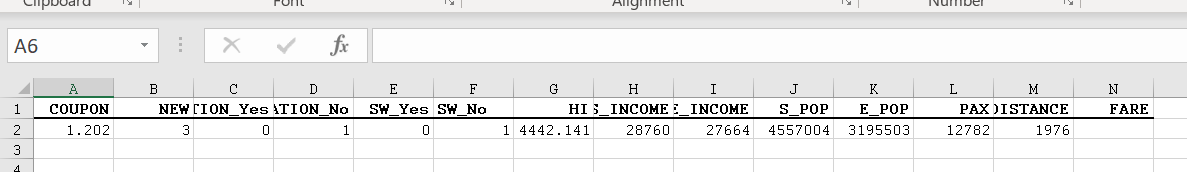


RMSE Value is 31.1892

After running all Regression Techniques Random Tree give optimal RMSE Value which is 31.1892.

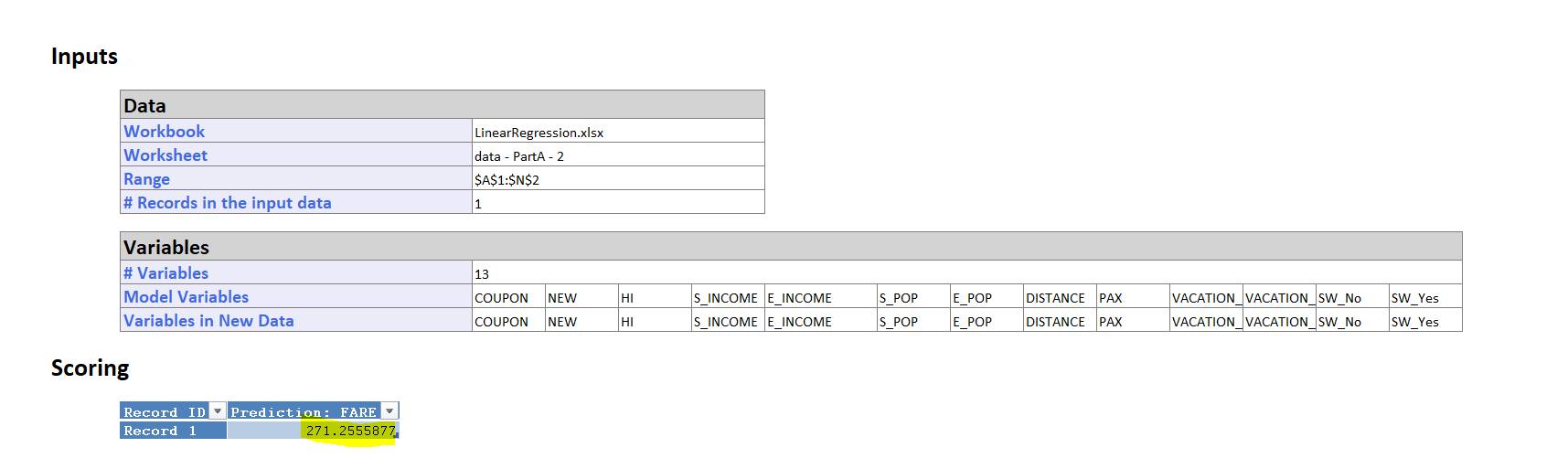
## **2. FARE PREDICTION:**

* Now, the average fare that needs to be predicted is fed to the model:



**Fig 16: Score input in Regression Model (If SW=No)**

* The predicted fare is determined by selecting Score a Match by Name in which all variables except fare are taken as the model input to predict the remaining variable(FARE) as shown below:

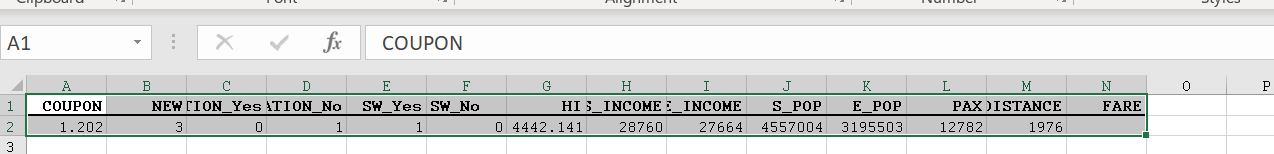


**Fig 17: Regression Model - Score Report**

* We see that the Predicted Average Fare is 271.25$.

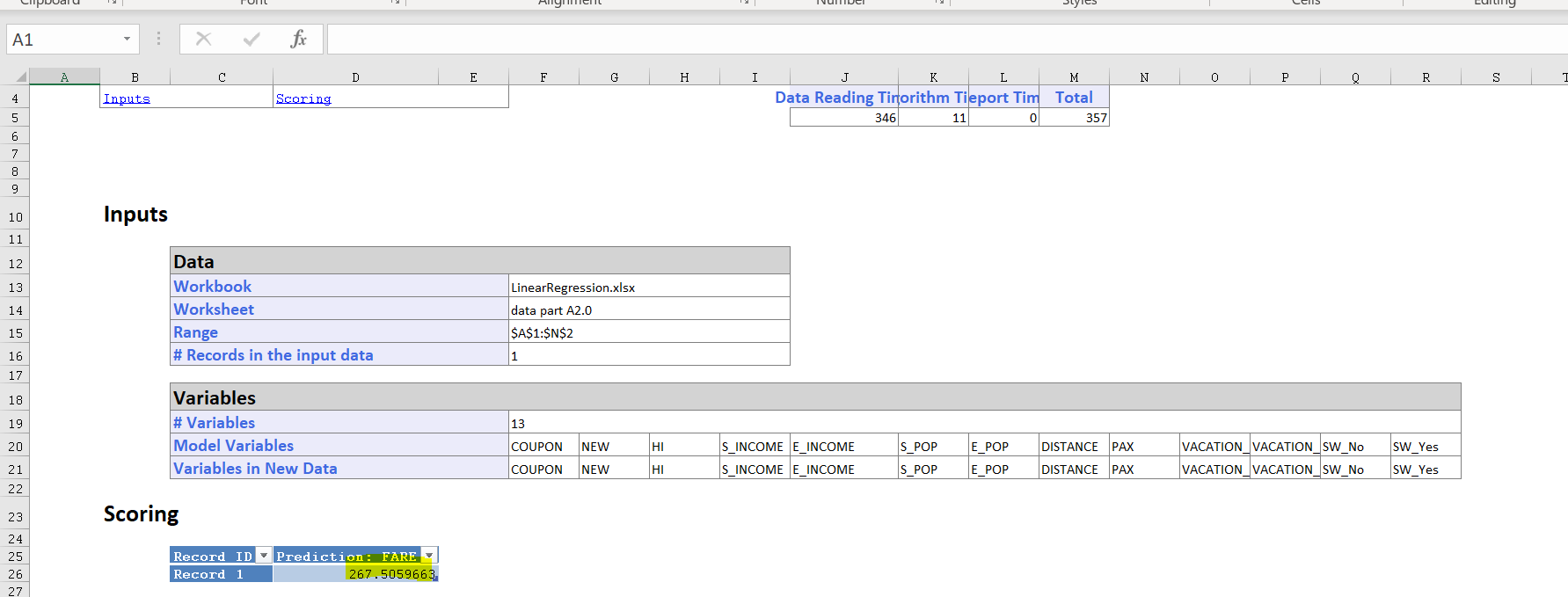
**3. Predict the reduction in average fare on the above route if South West Airlines decides to cover this route.**

* Bearing in mind that South West Airlines decides to cover this route, we change the SW column to 1 and run the model with the same data as below:



**Fig 18: Score input in Regression Model (If SW=Yes)**

* We predict the fare by selecting Score a Match by Name and we run it to get the FARE, as shown :



**Fig 19: Regression Model - Score Report**

* The estimated fare for this is $267.50 when this route is served by SW airlines.
* We can see that it is $4.75 (271.25-267.50) to minimize the air fare between the two.

**Part B**

**In reality, which of the factors (predictor variables) will not be available for predicting the average fare from a new airport? (i.e. before flights start operating on those new routes)**

1. **Briefly comment on your assumptions.**

We are excluding the below variables for predicting the average fare from a new airport,

S\_CODE: Since we already have the title S CITY in the dataset we should ignore / remove the code S CODE.

COUPON: There are no airlines flying on this path so it is not possible to calculate an average number of vouchers.

PAX: Since this is a new airport, the number of passengers arriving at this stage cannot be determined.

NEW: As the airport has not yet been established, we cannot calculate the number of new carriers joining the road.

E\_CODE: We should deny E CODE like S CODE because both have similar characteristics.

SW: This being a new airport, we cannot confirm if SW airlines would be covering this route.

GATE: People traveling through a new route wouldn't be more worried with GATE traffic.

SLOT: Like GATE, SLOT is refused here because citizens would be less worried about the airport traffic.

With these assumptions, we can start predicting the average fare using the model we have developed.

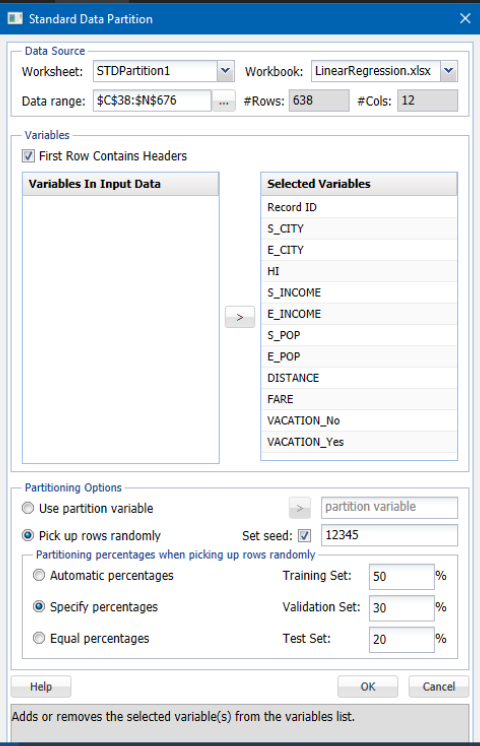
1. **Based on the settings and findings of the model from item A, build another model using the available (in your opinion) variables only. Comment on your solution.**

* Having tried various prediction models such as MLR (Multiple Linear Regression), NN (Neural Network) and Regression Trees, we decided to build our model using the Random Tree model as it proved to be the most effective model predictor during our analysis in part A.
* Now, based on our assumptions, we are going to build the model using Random Trees.

1. **Use this model to predict the average fare using only the available (in your opinion) data from the record in item A.2.**

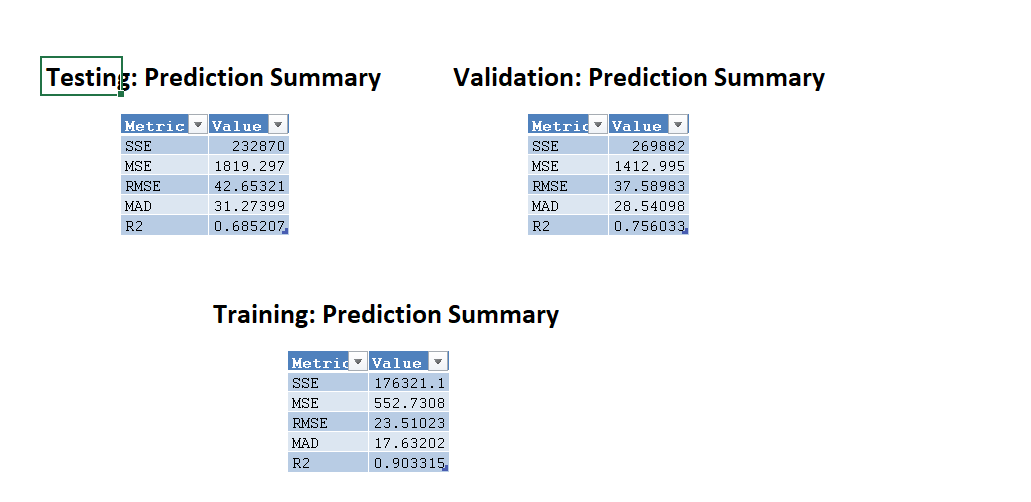
# **USING RANDOM TREE TECHNIQUE:**

* We have constructed the Random Tree Model; the same way we have constructed the LR model.
* Dataset is going to be partitioned into 3 categories: Validation, Training and Test Data, 50%,30% and 20% each.



**Fig. 20: Random Model - Data Partition**

* This partitions the data as we have specified and now we can go ahead by selecting the Random Tree Prediction Model and select the required reports which need to be generated.
* In this case we have selected the Summary Reports for all the partitioned data as we see below:



* Validation Summary report for Random Model shows a RMS error value of 37.58.

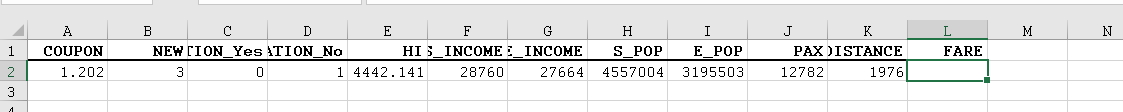
## **FARE PREDICTION:**

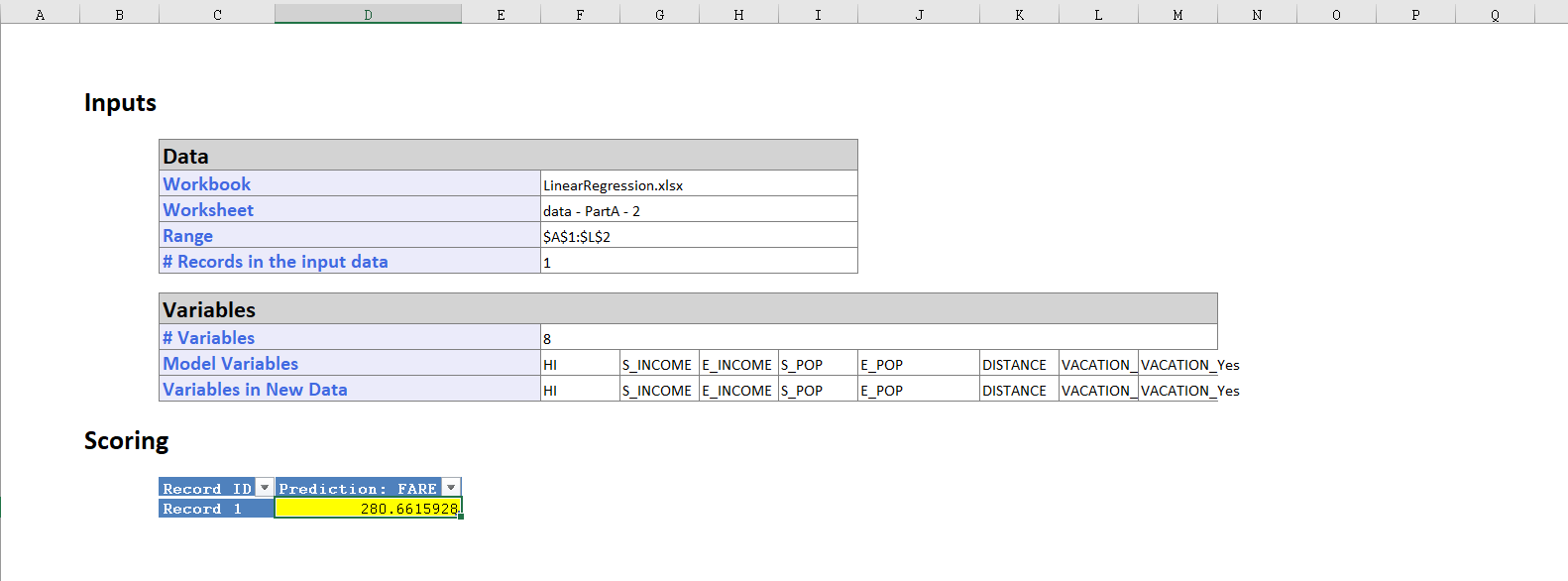
* Now, the average fare that needs to be predicted is fed to the model:



**Fig 24: Regression Model – Score Input**

* The predicted fare is calculated by selecting Score  Match by Name where all the variables except fare would be taken as the input for the model to predict the remaining variable(FARE) as shown below:





**Fig 25: Regression Model – Score Report**

* We see that the Predicted Average Fare is 280.69$ for the data provided into the model.

1. **Compare the predictive accuracy of this model with model from item A. Is this model good enough, or is it worthwhile re-evaluating the model once flights commence on the new route?**

In the first model only four variables were excluded, these variables are S\_CITY, E\_CITY, SLOT and GATE. The second model has only 10 variables since COUPON, PAX, NEW, SW along with the pervious variables were excluded.

In order to tell the accuracy of each model we read the RMS error value from Training Data, Validation Data and Test Data. The initial model from A predicted scores of 36.094, 43.40,74.38,221.11,206494 and 30.86 respectively. In the second model scores of Random Tree Validation data is 37.58 predicted. The fact that the scores in the first model are lower than the second suggests to us that the first model has better predictive accuracy.

Firstly, considering that we are building a new airport we had the compulsion to exclude certain variables, which would not influence the model for making the correct prediction. Hence, we had to exclude 8 variables for the second model as shown above. Due to having more robust variables, the first model can be used for used for training and predicting purposes, compared to the second one.

# **CONCLUSION:**

To sum up, in this exercise we tested on different prediction models to predict the airfares for new routes and airports. Out of all the prediction models that we implemented; Random Trees provided the most accurate average fares because of its lower RMS error value. However, it is to be noted that the RMS error is an indicator of the performance of the predictive model and not the ultimate deciding factor.