**Contents**

[**1.1** **Introduction** 4](#_Toc531307719)

[**1.2** **Modelling Process** 4](#_Toc531307721)

[1.2.1. Business Understanding Stage – Project Objective 4](#_Toc531307722)

[1.2.2. Data Understanding Stage 4](#_Toc531307723)

[1.2.3. Data Preparation Stage 4](#_Toc531307724)

[**1.3** **Modelling** 5](#_Toc531307725)

[**1.3.1 Multiple Linear Regression (MLR)** 6](#_Toc531307726)

[1.3.1.1 MLR with All attributes considered 7](#_Toc531307727)

[1.3.1.2 MLR with Optimum attributes 7](#_Toc531307728)

[1.3.1.3MLR with Best Subsets 8](#_Toc531307729)

[1.3.1.4 Lift Chart Observation: 9](#_Toc531307730)

[1.3.1.5 RROC and AOC Observation: 9](#_Toc531307731)

[1.3.1.6 MLR Summary 10](#_Toc531307732)

[**1.3.2 Regression Tree**](#_Toc531307733) 9

[1.3.2.1 Regression Tree – Min Error 10](#_Toc531307734)

[1.3.2.2 Regression Tree – Best Pruned 12](#_Toc531307735)

[1.3.2.3 Regression Tree – Full Tree 13](#_Toc531307736)

[**1.3.3 Artificial Neural Network** 13](#_Toc531307737)

[**1.3.4 Ensemble Modelling** 14](#_Toc531307738)

[1.3.4.1 Regression Tree – Boosting 15](#_Toc531307739)

[1.3.4.2 Regression Tree – Bagging 16](#_Toc531307740)

[1.3.4.3 NNP – Boosting 16](#_Toc531307741)

[1.3.4.4 NNP – Bagging 16](#_Toc531307742)

[**1.4** **Summary** 17](#_Toc531307743)

[**1.5** **Comparing the Predictive Accuracy:** 20](#_Toc531307744)

# 

# **Introduction**

# Machine Learning is one of the most prominent research topics in computer science and engineering, which is applicable in many disciplines. It provides a collection of algorithms, methods and tools able to embody intelligence to machines. This process involves finding anomalies, patterns and correlations within large data sets to predict feasible outcomes. A unique range of techniques can be used to increase revenues, cut costs, improve customer relationships and furthermore reduce risks. It aids in understanding the organisational requirements and uses varying levels of human inputs and set of rules to arrive at a decision. The significance of ML is the modelling tools embedded within, which can be trained, via a learning procedure, with a set of data describing a certain problem and to respond to similar unseen data with a common way. Predictive modelling is one such tool which enables to classify the future events and helps in anticipating the unforeseen outcomes. The modelling techniques includes Regression, Neural networks and decisions trees. The abundance of data which are unstructured makes it essential to pre and post process the data’s such that the factors like internal and external variables and constraints which recommends various course of actions can be figured out.

# **Modelling Process**

## 1.2.1. Business Understanding Stage – Project Objective

The major cities in United States are exposed to new airports which would provide the passengers with varied routes to travel. Fares of different routes have not yet been confirmed by several major airlines which include the Southwest Airline that operates on a different model compared to others. The presence of discounted airlines has led to reduction in airfares significantly. The objective is to *build a model which best predicts the average fare on a new route for the airlines.* As the data that would be fed to the model is labelled, the learning method is ‘Supervised’ and since the objective is to predict a numerical value, the Modelling Algorithm selected is **Regression**.

## 1.2.2. Data Understanding Stage

The dataset consists of 18 attributes. An analysis has been done to understand the importance of each of these attributes. Observations were made that the dataset has the greatest number of flights starting from Chicago (90) and ending in New York (54).

* HI: On exploration it was figured that the market concentration is individually measured with this parameter. It is calculated by squaring the market share of each firm competing in a market and then summing the resulting numbers. The other parameters under consideration was evident from the description given.

## 1.2.3. Data Preparation Stage

Below are the steps executed to check if the data can be fed into the Modelling Tool. (Observations are denoted in italics)

* Checked if any values are missing – *No missing values*
* Checked if all attributes are Numerical – *There are a few Categorical values present in the data set. Upon analysis, the below attributes were converted into numeric data with the help of ‘Transform’ function.*

*-VACATION – 0 & 1 for NO & YES respectively*

*- SW – 0 & 1 for NO & YES respectively*

*- SLOT – 0 & 1 for CONTROLLED & FREE respectively*

*- GATE – 0 & 1 for CONSTRAINED & FREE respectively*

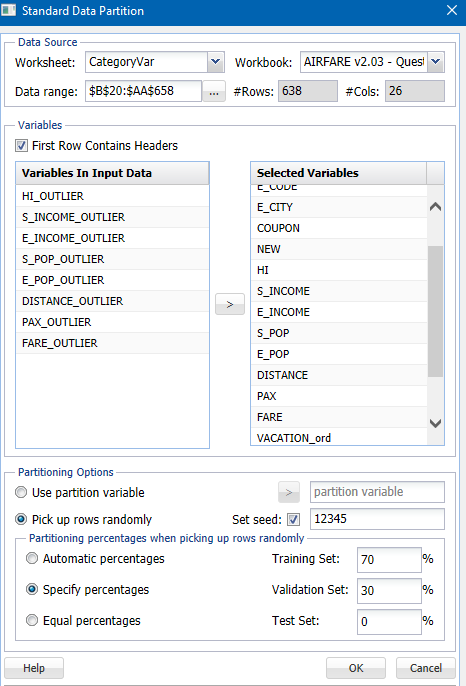
*The other Categorical attributes (S\_CODE, S\_CITY, E\_CODE, E\_CITY) were ignored as they are categorical (with many values) and hence irrelevant in determining the Dependent variable, FARE.*

* Checked for Outliers – *Outliers were identified outside the (µ ± 3σ) range. However, upon further analysis, decision was made not to eliminate these values, as the ‘potential outliers’ are all within the feasible range in this case. Below are the outliers identified upon analysis.*
* HI: The value of HI is expected to lie within the range [-730.65 : 9614.94]. Although there were few values in the given data which falls out of this range it had to be considered since the acceptable range of HI is up to 10000.

A similar observation was seen for three other parameters [S\_INCOME, PAX and FARE]. Hence these parameters are considered inevitable upon building the predictive models.

* Data Partitioning - *The original data under consideration has only* ***638 records*** *and hence to proceed with the modelling to obtain the best results, we need to train the model with some fair portion of the data set. Hence the partition was done with* ***70:30*** *(Training vs Validation) proportion. This has produced 447 samples for training the model and 191 samples for validating the same.*

Below provided is the Dialogue box of the Data Partitioning in XL Miner.



# **Modelling**

Modelling aims at processing the past data to predict the future efforts. It involves a group of process in which multiple sets of data are combined and analysed to reveal the relationships and patterns. The potency of data modelling is that it generalises multiple data sources and give independent judgments regarding what is relevant or not.

Several models were considered to determine which model fits into the training data set and produce a viable target value against the validation data set. The focus was mainly on the Root Mean Square Error values of the individual models, which may enable to come up with the decision of selecting the Best Model.

Below are the algorithms considered for the model preparing:

* Multiple Linear Regression (MLR)
* Regression Tree
* Minimum Error
* Best Pruned &
* Full Tree
* Neural Network
* Ensemble Models
* Regression Tree Boosting
* Regression Tree Bagging
* Neural Network Boosting
* Neural Network Bagging

## **1.3.1 Multiple Linear Regression (MLR)**

MLR attempts to model the relationship between two or more explanatory variables and the response variable by fitting a linear equation to observe data. All value of independent variable is associated with a value of dependent variable y.

In order to understand whether the attributes are correlated, a correlation table is prepared. From this, it is observed that the attributes DISTANCE & COUPON, FARE & DISTANCE are strongly correlated.

From the correlation plot, it can be assessed the attributes NEW, HI, and PAX are not correlated to the target, which is FARE. The only attribute which is strongly correlated to FARE is the DISTANCE.

Below provided is the Correlation plot of the attributes.



Inorder to understand the independent attributes are statistically significant to the target variable, there is need to analyse further on the basis of their *p* values.

The initial approach was to try and see whether the complete set of attributes fit into a linear model.

### 1.3.1.1 MLR with All attributes considered

Observations:

Below were the RMS Error obtained:

* Taining Sample – 35.53
* Validation Sample – 34.39





From the report,we could observe that three attributes – **COUPON**, **NEW** and **S\_INCOME** have their *p* values greater than 0.05, which means that these parameters are *statistically insignificant* for this model. Hence, it was considered to re-build the M LR model by eliminating these attributes.

### 1.3.1.2 MLR with Optimum attributes

Optimum output is determined by skimming at the *p* values of the attributes and eliminating the ones that have their pvalues greater than 0.05(they are COUPONS, NEW and S\_INCOME), obtained from the previous result.

Observations

The RMS errors for the model are provided below:-

* **Training data – 35.73**
* **Validation data – 34.51**

Adjusted **R2  - 0.7784**





### 1.3.1.3MLR with Best Subsets

Since there occurs many predictors and as it is better to limit the model to only the significant variables, ‘Variable Selection’ was performed to select the best subset of variables. For this, the algorithm was executed by selecting the ‘Best Subset’ option where searches for all combinations of variables are performed to find the best fit combination. Hence a number of best subsets were enabled and the below result was obtained.



For choosing the Best Subset below aspects were considered: -

* Minimum Number of coefficients – **at least 10**
* Probability value – **less than 0.05**
* Adjusted R2 – **Reasonably higher**

Observations:





### 1.3.1.4 Lift Chart Observation:

As the Lift Chart provides the effectiveness of the predictive model calculated as the ratio between the results obtained with and without the predictive model. From the Training and the Validating data set, it is thereby observed that, as the area between the lift curve and the baseline is considerably less, the performance of the predictive model is not satisfactory.

### 1.3.1.5 RROC and AOC Observation:

Unlike the Classification model, which could be assessed by the Receiver Operating Curve, the Regression models can be visually assessed using Regression ROC Curve (Over estimation vs Under estimation). The Area Over the Curve (AOC) obtained for the Training and Validation sets are 1.31 e+008 and 2.13 e+007.Hence, it was inferred that the model performed comparatively well for Validation dataset.





The model predicted using Multiple Linear Regression can be represented as below:-

**FARE = 73.2945 + 0.00945 \* HI + 3.8866E-06 \* S\_POP + 4.1017E-06 \* E\_POP + 0.0784 \* DISTANCE -0.0008 \* PAX - 37.9071 \* VACATION - 45.6111 \* SW - 20.4173 \* SLOT - 17.3639 \* GATE**

### 1.3.1.6 MLR Summary

Three models were build based on Multiple Linear Regression and below provided are the observations from each of the models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **# Attributes Selected** | **Attributes Eliminated** | **Adjusted R²** | **RMS Error** | |
| **Training Data** | **Validation Data** |
| Complete Attributes | 13 Attributes | - | 77.964% | 35.5346 | 34.3927 |
| Best Subset | 9 Attributes | COUPON  NEW  E\_INCOME  S\_INCOME | 77.281% | 36.2470 | 34.2726 |
| Statistically Significant Attributes | 10 Variables | COUPON NEW S\_INCOME | 77.846% | 35.6190 | 34.3347 |

It is observed that the value of R2 and RMS Error is almost same for all the three MLR Models; However, the decision was made to stick on with the Best Subset model which produced best result for the validation data set. The value of R2 suggests that proportion of variability of the dependant variable (FARE) is ~**78%** asexplained by the model.

## **1.3.2 Regression Tree**

### 1.3.2.1 Regression Tree – Min Error

A regression Tree is implemented through binary recursive partitioning, where the process is iterative that splits the data into branches and the process continues to split each of those partitions into smaller groups. The regression tree algorithm chooses the node that gives the minimum value sum of the squared deviations from the mean in the two separate partitions.

Next approach was to build a model based on the Regression Tree. As there are three options available, which are Full Tree, Minimum Error and Best Prune, the approach was to consider all the three models with Best Prune as this would prevent the Model from Overfitting to the training data. The partitioned data was fed into the Model without any normalisation of the attributes.

From, Minimum Error Regression Tree, it is observed that, the smallest Validation RMS error belongs to the tree with 13 Decision Nodes. Hence, this is considered as the Minimum Error Tree, as it holds the smallest misclassification error in the Validation set.

Observations:

Nodes for the Validation Data

|  |  |
| --- | --- |
| **#Decision Nodes** | 13 |
| **#Terminal Nodes** | 14 |



RMSE for Training Data – 32.71

RMSE for Validation Data – 38.02

Decision Tree – Min Error with Pruning



The Decision Tree Algorithm has formulated the Model by considering the DISTANCE variable as the Primary Node followed by SW (Southwest Airline’s operating decision) and VACATION (Vacation route or not). It is also observed that the Algorithm has not taken into consideration on the attributes like HI, S\_INCOME, E\_POP, GATE and PAX for this Model.

### 1.3.2.2 Regression Tree – Best Pruned

Form Best Pruned Regression Tree, the tree pruned using validation dataset was obtained. Here, the pruning creates a decision tree with cross validation error within 1standard error of the minimum error. And it is observed that, the smallest Validation RMS error belongs to the tree with 13 Decision Nodes.

Observations:

|  |  |  |
| --- | --- | --- |
| **#Decision Nodes** |  | 13 |
| **#Terminal Nodes** |  | 14 |

### 1.3.2.3 Regression Tree – Full Tree

From Full Tree, the full regression tree grown using the training data set is obtained. And it is observed that, the smallest RMS error belongs to the tree with 13 Decision Nodes.

Observations:

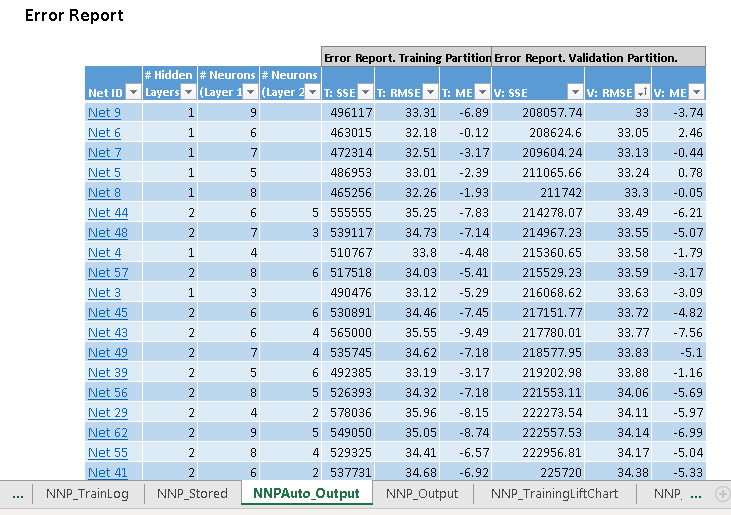
|  |  |
| --- | --- |
| **#Decision Nodes** | 13 |
| **#Terminal Nodes** | 14 |



## **1.3.3 Artificial Neural Network**

ANN’s are nonlinear data driven self-adaptive approach as opposed to traditional model-based methods. They can identify and learn correlated patterns between input data and corresponding target values. ANN’s are predominant tools for modelling especially when the relationship between the data is unknown. After training ANN’s can be used to predict the outcome of new independent input data.

The error report obtained below on running the model suggests that the Network 9 has the least RMSE error of 33 compared to all other networks.



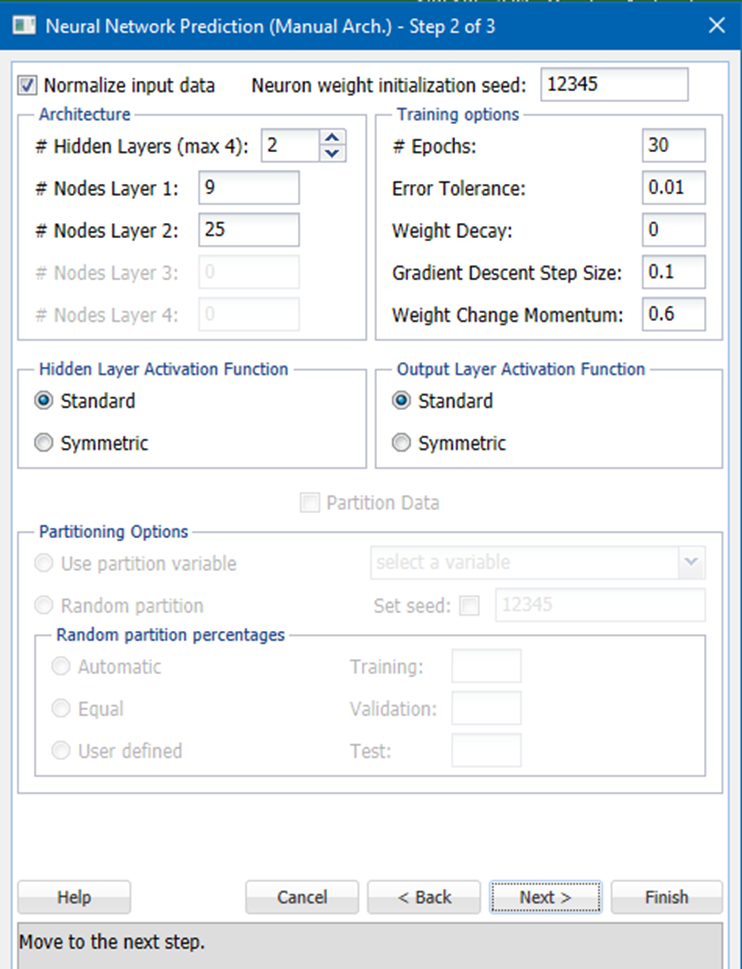
Observations:

The ANN produced a quite satisfactory RMSE scores for both the training and validation data.



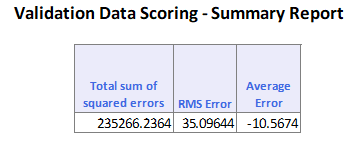
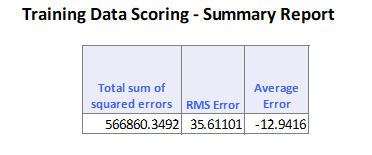
The next approach made was to identify if the increase of Hidden layers and the nodes have any impact on the Model’s performance.

Step 1 – Increased the Hidden Layers to Two and kept the Nodes count as default.



Observation was made that the increase of Hidden Layers and Nodes did not make any significant improvement in the Model’s performance.

Below provided is the RMS values obtained for the Model with 2 Hidden Layers.



## **1.3.4 Ensemble Modelling**

In order to improve the accuracy of predictive analytics, it is also essential to consider the ensemble models. This process includes various analytical models which are executed and eventually the results are synthesized into a single core. It is often observed that using specific modelling techniques on one data set can have biases, variability or even inaccuracies which would inadvertently affect the reliability of the analytical findings. All these drawbacks are overawed, and better and reliable information is suggested to the business with the help of ensemble models. Bagging and boosting are the two most widely used techniques to ensemble different models which are briefed below.

Bootstrap aggregating (Bagging)

Bagging process subsamples the training data set and produces models on each of the bootstrapped subsamples. Sampling is done with replacement. A mean of the predictions or the majority voted prediction would be selected as the Final Prediction. All these predictions are done parallelly.

Boosting

Boosting technique employs sequential prediction. The model learns from the mistakes of the previous predictors. Therefore, the observations have an unequal probability of appearing in subsequent models and ones with the highest error appear most.

### 1.3.4.1 Regression Tree – Boosting

The algorithm was able to generate exceptionally well Model, which almost fit into the Training dataset (3.75 RMSE). For Validation dataset, there is a significant difference in the RMSE score which is 22.33. It is obvious that the model is *Overfitting* to the Training data.

It is also observed that the algorithm has eliminated two attributes – COUPON & NEW

Observations:



### 1.3.4.2 Regression Tree – Bagging

The results obtained on running this model brings out the below observations which indicates that the RMSE for validation data (32.375) is way above the RMSE for the training data (27.765). Hence this model do not best suited to obtain the predicted fare.



### 1.3.4.3 NNP – Boosting

The algorithm on execution generates a model which has a higher RMSE value for Validation data (35.35) with respect to the Training data (32.74). Thereby, rejecting this model with the above justification.



### 1.3.4.4 NNP – Bagging

The model gives a summary report as indicated below with an overwhelming RMSE value for the Validation data (97.09) which is not suitable to predict the airline fare. Hence rejecting this model.



# **Summary**

After the analysis of various models, a Summary of all the models considered has been prepared and displayed in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl No** | **Training Model** | **RMS Error** | | **Decision** | **Time Taken(mS)** |
| **Training Data** | **Validation Data** |
| 1 | MLR | 36.2470 | 34.2726 |  | 4 |
| 2 | RT - Min Error | 32.7187 | 38.0190 |  | 37 |
| 3 | RT - Best Pruned | 33.2077 | 38.3280 |  | 37 |
| 4 | RT - Full Tree | 31.6604 | 38.6898 |  | 37 |
| 5 | NNP | 33.3149 | 33.0047 |  | 18 |
| 6 | RT-Boost | 3.7579 | 22.3326 |  | 2103 |
| 7 | RT-Bagging | 27.7657 | 32.3759 |  | 1269 |
| 8 | NNP-Boost | 32.7369 | 35.3512 |  | 1758 |
| 9 | NNP-Bagging | 102.8285 | 97.0907 |  | 1340 |

MLR model showed decent performance against the Training and Validation data sets. Regression models produced quite reasonable performance for the Training data, however their performance against the Validation samples were not appreciable when compared to that of MLR. Neural Network Model produced significant performance during the training and validation stages. While the Ensemble models (except NNP-Bagging which produced 102.8285 which unfits the training data) were able to generate substantial performance during the training phase (3.76 for RT-Boost which overfits the training data), they showed a significant difference with respect to the validation data score value, which made them not to consider as the best model for this.

The RMSE for NNP Bagging which was 102.8285 suggests that the model produced is *underfitting* to the dataset.

Hence, the decision was made to consider the Neural Network as the best model among these as it showed consistent performance and is not suffering from any high bias or high variance with both the training and validation data samples.

A.2.

A separate sheet was prepared to predict the Average Fare on a route with the built model. Below mentioned are the steps performed:-

- The Categorical values were converted into Numerical in line with the data set.

- The Model was made to predict the Fare by *Matching the Names*.

The result obtained is provided below:-



Average Fare - **$281.57.**

A.3

The Attribute corresponding to the Southwest Airlines’ decision on whether the route will be served or not is SW. So the value SW has changed to 1 from 0 in the sample and the following results were obtained:



Since Southwest Airlines operates on low fare strategy, we expect the fare to reduce. As expected, the Average Fare has come down to **$207.23.**



B.

For a new Airport, most of the attributes considered for the above model may not be available. For instance, Attributes like COUPON (Average number of Coupons), HI, NEW, SLOT, GATE (Airport Congestion), PAX (Number of Passengers) & SW may not be relevant or will not be available while setting up a new Airport. The Market Concentration is found to be unique and it is Route specific. So this parameter is available only after the Airport is fully functional.

Hence, the following relevant/available attributes were considered to predict the model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S\_INCOME** | **E\_INCOME** | **S\_POP** | **E\_POP** | **DISTANCE** | **VACATION** |

S\_INCOME: Data can be collected for the Starting City’s average personal income.

E\_INCOME: Similarly, data will be available for the average personal income for the Ending City.

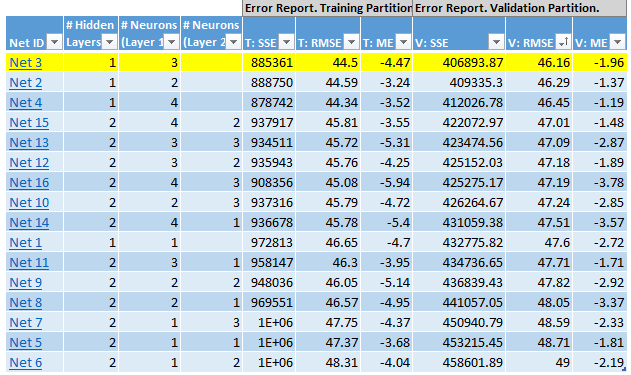
S\_POP and E\_POP: Population for both the Starting and Ending city can be collected before opening a new airport.

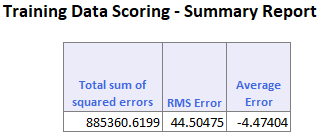
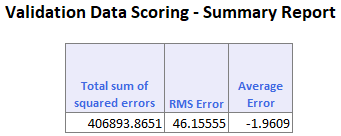
DISTANCE: Distance between the new and destination airports can be calculated.

VACATION: It can be identified whether the new airport location is a vacation spot or not.

Based on considering these attributes, another Neural Network Model has been made and the below observations were obtained:

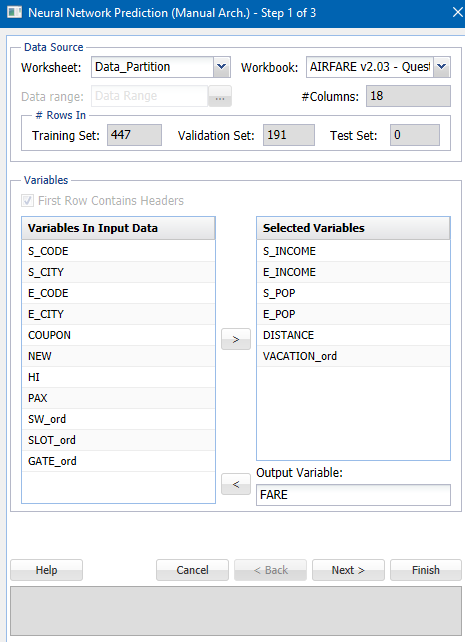
Network 3 produced the lowest RMSE of 46.16 and the corresponding model is generated with this network.



The RMS errors for the model are provided below:-

* **Training data – 44.50**
* **Validation data – 46.16**



By creating the predictive model from Network 3, and the Model was made to predict the Fare by *Matching the Names* and produced the below result with a Fare of **$273.43.**



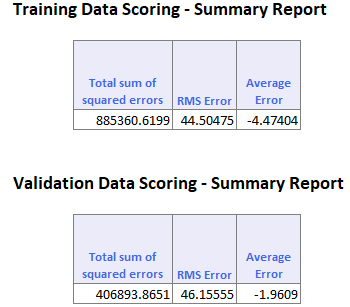
# **Comparing the Predictive Accuracy:**

The Predictive Accuracy of a Machine Learning Algorithm can be interpreted using several methods, like Classification Accuracy, Logarithmic Loss, Confusion Matrix, Area Under Curve in ROC Curve etc. for the Classification Models), Mean Absolute Error, Root Mean Square Error, Area Over Curve etc. for Regression Models.

Since the target value is a continuous or Regression variable, the focus was given to the RMSE, Lift Chart and the RROC Curve of each model to evaluate their predictive accuracy.

For the first case (existing route), the Model had taken into consideration all of the attributes and produced a RMSE of 33.31 and 33.00 for the Training and Validation sets respectively. However, for the new Airport, when the Model is formulated with the set of available attributes, the Score predicted found to be unfit as the RMSE was 44.50 and 46.15 respectively for the Training and Validation Data. Below provided are the Individual Model’s performance scores.

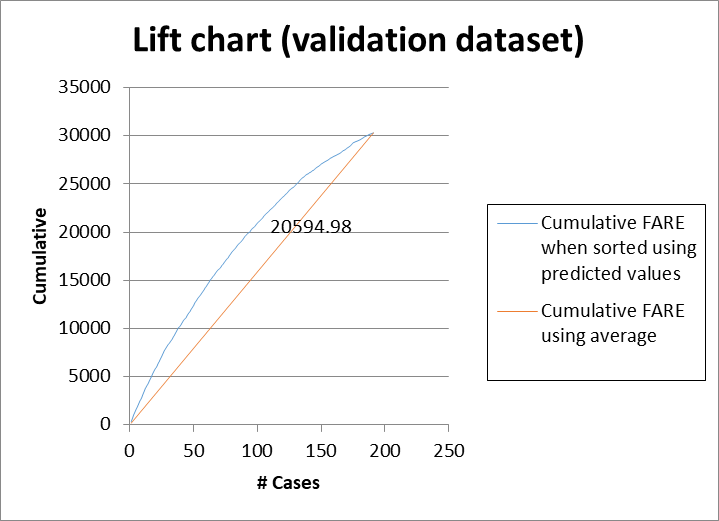
* Comparing the RMSE for the two cases; for the existing route and for the new route

 Model Score for Old Routes (with all attributes) Model Score for New Route (with six attributes)

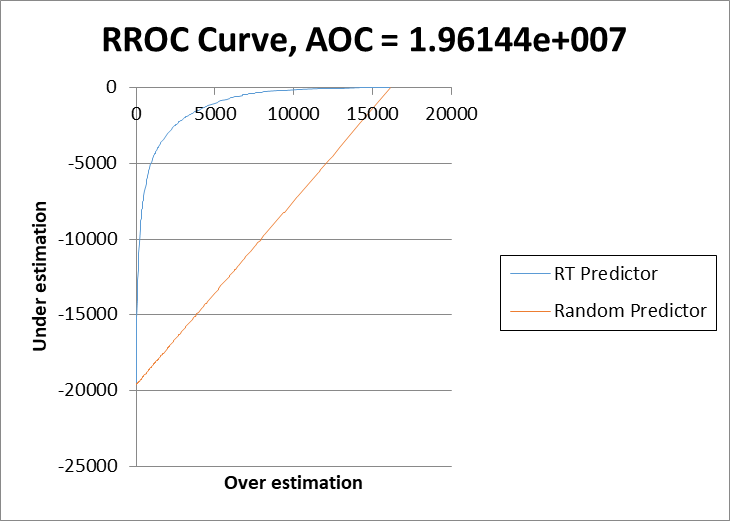
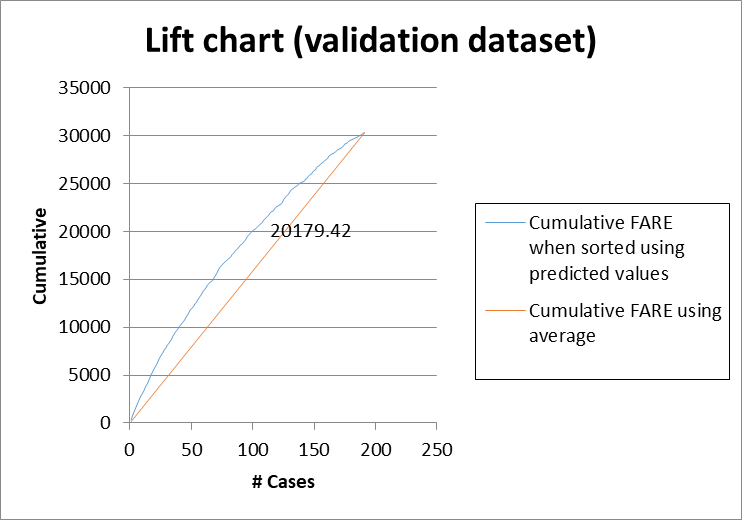
Observation:

From the above obtained Training and Validation results, it can be observed that, RMS Error for Validation data is considerably less (with value **33.00)** in the model for old routes while that for new route is comparatively more, i.e., with RMS error – **46.16**.

* Comparing Lift Charts and Regression ROC Curve, AOC between models for old routes and the new route:



**Lift Chart and RROC Curve, AOC For the Model for Old Routes**



**Lift Chart and RROC Curve, AOC For the Model for New Route**

Observation:

From the above Charts, it is interpreted that when measured against 100 routes (cases) the cumulative airfare is 20594.98 when considering the model for existing routes. For the new routes, the cumulative airfare obtained is 20179.42. Also it is observed that the Area Over the Curve when considering the model for new route is **3.87882e+007** while that for the old routes is **1.96144e+007.** That is the first Model has produced superior results when compared to the second one.

Inference:

From the above comparison and its observations, it can be inferred that, since the RMS Error and Area Over the Curve for model for old routes is less than that of the new routes, the predictive accuracy is more for the Model considering the old routes. Moreover, the observation from the Lift Charts suggests that, the cumulative airfare is higher with the model for existing routes. That is even though the Model makes fair predictions for the existing routes, there is a significant difference in the performance while predicting the Fare for the new routes.

Thus it makes the former model to be considered more robust and accurate. Hence, it is worthwhile to re-evaluate the model once the flights commence on the new route.

*References made:*

* Data-mining-for-business-analytics-galit-shmueli
* [*https://en.wikipedia.org/wiki*](https://en.wikipedia.org/wiki)
* *https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9*