A Project predicting the customer satisfaction of air travelers using SAS Enterprise Miner

Predictive Analysis

Hosain Ahmed (301209637)
Paola Yescas (301212529)

Table of Contents

Executive Summary 3

Introduction 4

Tools and Models used 4

Data Extraction 5

4.0. 7

4.1. 7

4.2. 7

4.3. 8

4.3.1 8

4.3.2. 10

4.4. 12

4.4.1. 12

4.5. 14

5.0. 15

5.1. 15

5.1.1. 15

5.1.2. 17

5.1.3. 18

5.2. 19

5.2.1. 19

5.2.2. 20

5.2.3. 21

5.2.4. 23

6.0. 26

6.1. 27

6.1.1. 27

7.0. 31

8.0. 32

8.1. 32

8.2. 32

8.3. 33

8.3.1. 33

8.4. 34

8.5. 34

9.0. Complete Diagram 35

10.0. References 36

Executive Summary

A detailed predictive analysis was undertaken as part of this project using SAS enterprise Miner to develop a machine learning model for forecasting the variables influencing the satisfaction rate of consumers travelling in an airline.

Kaggle provided the dataset for this project, which included data from an airline poll on passenger/customer contentment based on a variety of characteristics. Age, Gender, Travel Class, Travel Type, Customer Type, Flight Distance, Arrival and Departure Delays, as well as customer satisfaction variables such as On-board Service, Cleanliness, Seat Comfort, Baggage Handling, Inflight Entertainment, WIFI, Ease of Online Booking, Gate Location, Food and Drinks, Online Boarding, Leg Room Service, Check- in Service, and Inflight Service are all included in the dataset.

The outcome for this analysis will be the column feature called 'satisfaction,' which has two values: 'Dissatisfied' and 'Satisfied,' describing the customer's overall satisfaction level.

Outcome

Based on this analysis, the ASE 4B Tree model has proved to be the best model.

Introduction

The aviation industry has grown to be the most essential element of a country's economic development. It is critical in transferring people or goods from one location to another, whether domestically or internationally, especially when long distances are involved. The airline industry is fiercely competitive, and the most crucial aspect in the travel process is the client. In a highly competitive climate, providing high-quality services to passengers, in addition to improving flight safety and operation, is the key competitive advantage for an airline's profitability and long-term success.

A judgement made on the basis of a given service interaction is known as passenger satisfaction. Satisfaction and loyalty are not interchangeable terms. Customers can be loyal without being extremely satisfied, or they can be highly satisfied but not loyal. Airlines have begun customer engagement programs in order to improve customer relations and encourage them to travel with the same airline on a regular basis. People nowadays are extremely price sensitive, and they will switch airlines for a slight price difference. As a result, the airlines must now devise a strategy to retain customers while also satisfying them. Passengers' satisfaction varies from person to person; some desire more off-board amenities, while others prefer aboard amenities. Some like extra luggage, while others are content with good cuisine. Now the question arises as to how an airline can accommodate such a large number of passengers.

This study is undertaken to gain a better understanding of the clients. Determine what people expect from airlines and what they receive. The gap will reveal where airlines are falling short and how they may improve customer service.

Tools and Models used

SAS Enterprise Miner: It is a powerful tool that Streamlines data mining and use analytics to build predictive and descriptive models. SAS Enterprise Miner aids in the analysis of complicated data, the discovery of trends, and the development of models so that fraud may be detected more quickly, resource demands can be forecasted, and customer attrition can be reduced. In this project we use 3 predictive models:

- **Decision tree**: decision trees employ a tree structure to display the predictions that arise from a series of feature-based splits
    - Maximal Decision tree
    - Decision Tree with Misclassification
    - Decision Tree with Average square error (2 Branch)
    - Decision Tree with Average square error (3 Branch)
    - Decision Tree with Average square error (4 Branch)
- **Regression**: The statistical link between a dependent variable and one or more independent variables is determined using regression
    - Full Regression
    - Forward Regression
    - Backward regression
    - Stepwise Regression

- **Neural Network**: Without any task-specific rules, these systems learn to do tasks by being exposed to a variety of datasets and examples. Based on biological neural network
    - Neural network using impute with 50 iteration and 3 hidden units
    - Neural Network using impute with 100 iteration and 4,5,8 hidden units
    - With 100 iterations; 3 and 8 hidden units:
        - Neural network using log transformation
        - Neural Network using Full Regression

Data Extraction

The dataset for this project was received from Kaggle, which had data taken from an airline poll on passenger/customer happiness based on numerous parameters. The dataset contains 129,880 entries with 23 columns including Age, Gender, Travel Class, Arrival and Departure Delays, as well as variables that influence customer satisfaction such as On-board Service, Cleanliness, Seat Comfort, Baggage Handling, and so on.

The dataset includes a column or feature called 'satisfaction,' which has two values 'Dissatisfied' and 'Satisfied' that describes the customer's overall satisfaction level. This feature is referred to as the label feature since it represents the customer's overall experience based on the ratings given for other features.

Detailed data description:

	Name	Attribute	Description	Value
1	Gender	Nominal	Gender of the passengers	Female, Male
2	Customer Type	Nominal	The customer type	Loyal customer, disloyal customer
3	Age	Interval	The actual age of the passengers	Numeric
4	Type of Travel	Nominal	Purpose of the flight of the passengers	Personal Travel, Business Travel
5	Class	Nominal	Travel class in the plane of the passengers	Business, Eco, Eco Plus
6	Flight Distance	Interval	The flight distance of this journey	Numeric
7	Inflight Wi/Fi Service	Interval	Satisfaction level of the inflight Wi-Fi service	0: Not Applicable;1-5
8	Departure/Arrival Time Convenient	Interval	Satisfaction level of time Departure/Arrival convenient	0: Not Applicable;1-5

9	Ease of Online Booking	Interval	Satisfaction level of online booking	0: Not Applicable;1-5
10	Gate Location	Interval	Satisfaction level of Gate location	0: Not Applicable;1-5
11	Food and Drink	Interval	Satisfaction level of Food and drink	0: Not Applicable;1-5
12	Online Boarding	Interval	Satisfaction level of online boarding	0: Not Applicable;1-5
13	Seat Comfort	Interval	Satisfaction level of Seat comfort	0: Not Applicable;1-5
14	Inflight Entertainment	Interval	Satisfaction level of inflight entertainment	0: Not Applicable;1-5
15	On-board Service	Interval	Satisfaction level of On-board service	0: Not Applicable;1-5
16	Leg Room Service	Interval	Satisfaction level of Leg room service	0: Not Applicable;1-5
17	Baggage Handling	Interval	Satisfaction level of baggage handling	0: Not Applicable;1-5
18	Check-in Service	Interval	Satisfaction level of Check-in service	0: Not Applicable;1-5
19	Inflight Service	Interval	Satisfaction level of inflight service	0: Not Applicable;1-5
20	Cleanliness	Interval	Satisfaction level of Cleanliness	0: Not Applicable;1-5
21	Departure Delay in Minutes	Interval	Minutes delayed when departure	Numeric
22	Arrival Delay in Minutes	Interval	Minutes delayed when Arrival	Numeric
23	Satisfaction	Binary	Airline satisfaction level	Satisfaction, dissatisfaction

4.0. Modeling Steps

4.1. Create a Data Diagram

To begin building a process flow diagram we need to create a process flow diagram

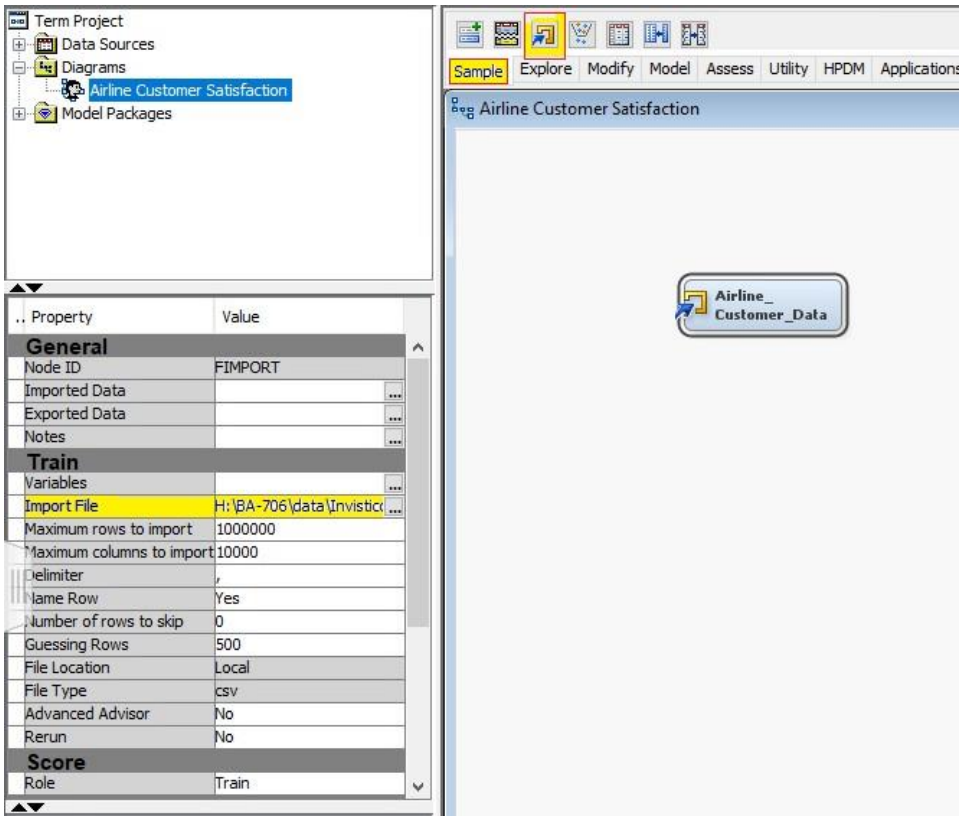
Steps:

1. On the File menu, select New -> Diagram.
2. Enter Airline Customer Satisfaction as the Diagram Name and click OK. An empty diagram opens in the Diagram Workspace.

4.2. Data Source

Since our raw data is in excel format, we cannot load the data directly from the data source in SAS Miner as it only allows SAS table through the library, instead we need to import it through the File Import node as shown in steps below:

1. Select the **Sample** tab on the Toolbar and drag the **File Import** node into the Diagram workspace.
2. Rename it to Airline_Customer_Data.
3. Go to the properties panel of Airline_Customer_Data node and click on import file under Property >Train
4. Locate and load the excel data.



4.3. Data Exploration

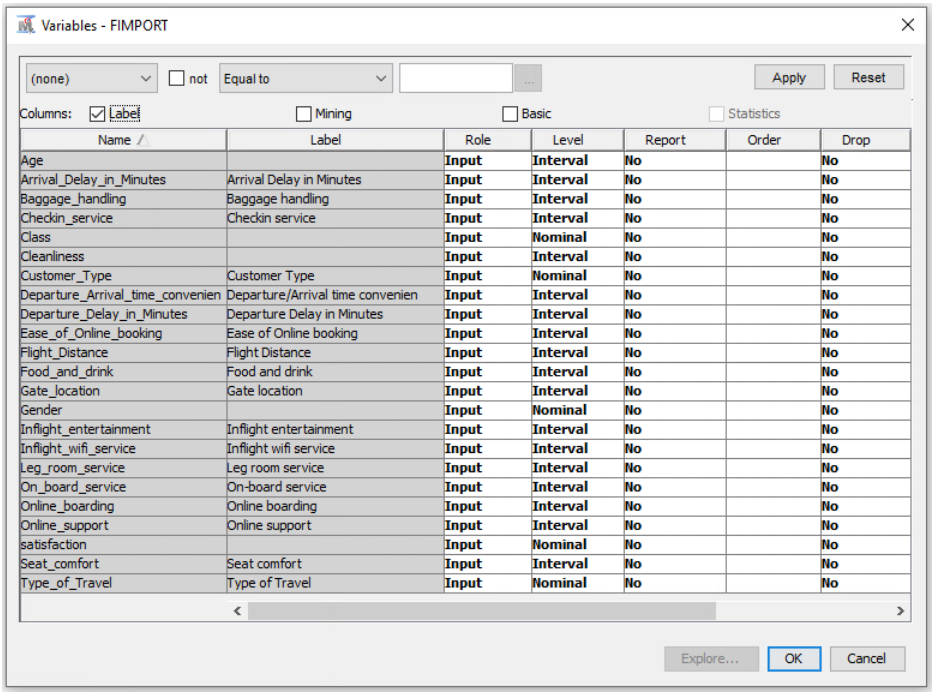
Data exploration is the first step in data analysis, during which data analysts utilize data visualization and statistical tools to characterize dataset characterizations like size, amount, and correctness in order to gain a better understanding of the data. Both manual analysis and automated data exploration software solutions are used to visually explore and identify relationships between different data variables.

4.3.1 First level data Exploration

To Explore the variable, first we need to determine if SAS Miner has categorized the variables to the correct measurement level.

To view the variable:

1. Right click on the Airline_Customer_Data node and click edit variable
2. We get the following window:



Name	Label	Role	Level	Report	Order	Drop
Age		Input	Interval	No		No
Arrival_Delay_in_Minutes	Arrival Delay in Minutes	Input	Interval	No		No
Baggage_handling	Baggage handling	Input	Interval	No		No
Checkin_service	Checkin service	Input	Interval	No		No
Class		Input	Nominal	No		No
Cleanliness		Input	Interval	No		No
Customer_Type	Customer Type	Input	Nominal	No		No
Departure_Arrival_time_convenien	Departure/Arrival time convenien	Input	Interval	No		No
Departure_Delay_in_Minutes	Departure Delay in Minutes	Input	Interval	No		No
Ease_of_Online_booking	Ease of Online booking	Input	Interval	No		No
Flight_Distance	Flight Distance	Input	Interval	No		No
Food_and_drink	Food and drink	Input	Interval	No		No
Gate_location	Gate location	Input	Interval	No		No
Gender		Input	Nominal	No		No
Inflight_entertainment	Inflight entertainment	Input	Interval	No		No
Inflight_wifi_service	Inflight wifi service	Input	Interval	No		No
Leg_room_service	Leg room service	Input	Interval	No		No
On_board_service	On-board service	Input	Interval	No		No
Online_boarding	Online boarding	Input	Interval	No		No
Online_support	Online support	Input	Interval	No		No
satisfaction		Input	Nominal	No		No
Seat_comfort	Seat comfort	Input	Interval	No		No
Type_of_Travel	Type of Travel	Input	Nominal	No		No

Observations:

1. SAS has assigned input(independent variable) role to all the variables.
2. SAS has determined that Class, Customer Type, Gender, Type of Travel and Satisfaction contains nominal data.
3. SAS has categories all the other variables as interval data.

After analysis the variable at a rudimentary level, following changes need to be made:

- Set satisfaction as target and binary variable, this will be the variable that depends on all the other input variable.

- Reject the following variables:
    - Considered as **irrelevant** data:
        - Gate_Location: the gate location is predetermined for each airline and the type of journey and cannot be altered to improve the customer satisfaction.
        - Flight Distance: for this study as the flight distance is fixed which cannot be improved by any means.
        - Departure_Arrival_Time_Convenience: as well as information about the flight time is available to the customer and the pricing is determined based of this factor as well.
    - Considered as **redundant** data
        - Type of Travel: Class variable gives the same information.

Variable summary:

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	16
INPUT	NOMINAL	3
REJECTED	INTERVAL	2
REJECTED	NOMINAL	1
TARGET	BINARY	1

After making the required changes as shown above, click okay and run the node.

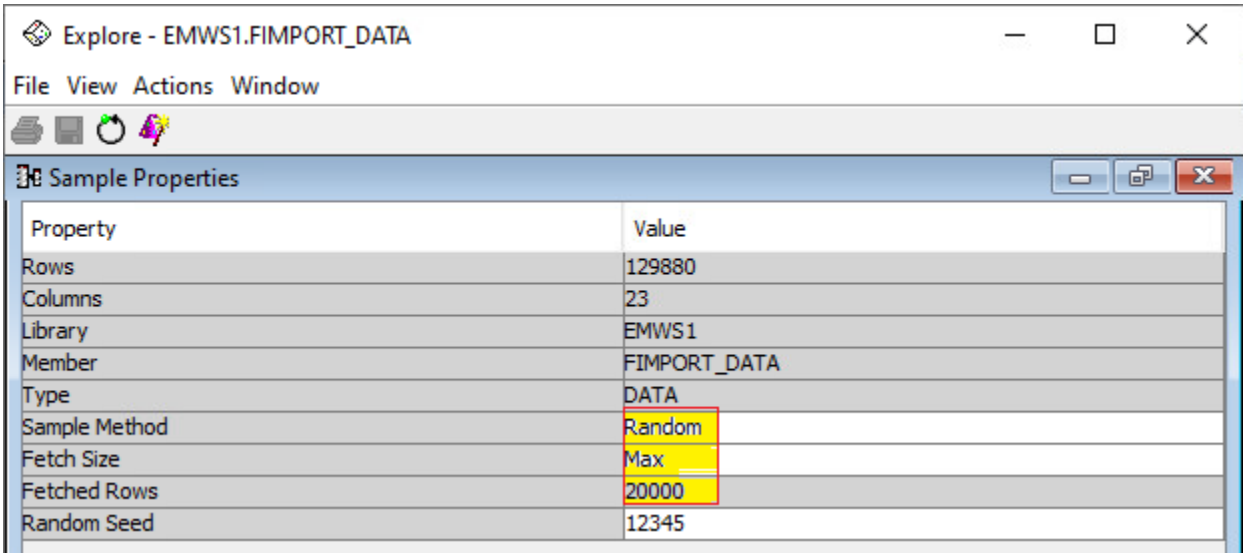
4.3.2. Second level data Exploration

Exploring all the input variables.

Steps:

1. Right click the Airline_Customer_Data node, press Ctrl and select all the variables, then click explore.
2. Go to Sample properties and set sample method to random, fetch size to max and click apply.
3. Then maximize the sample statistics.

It fetches 20,000 results randomly for analysis.



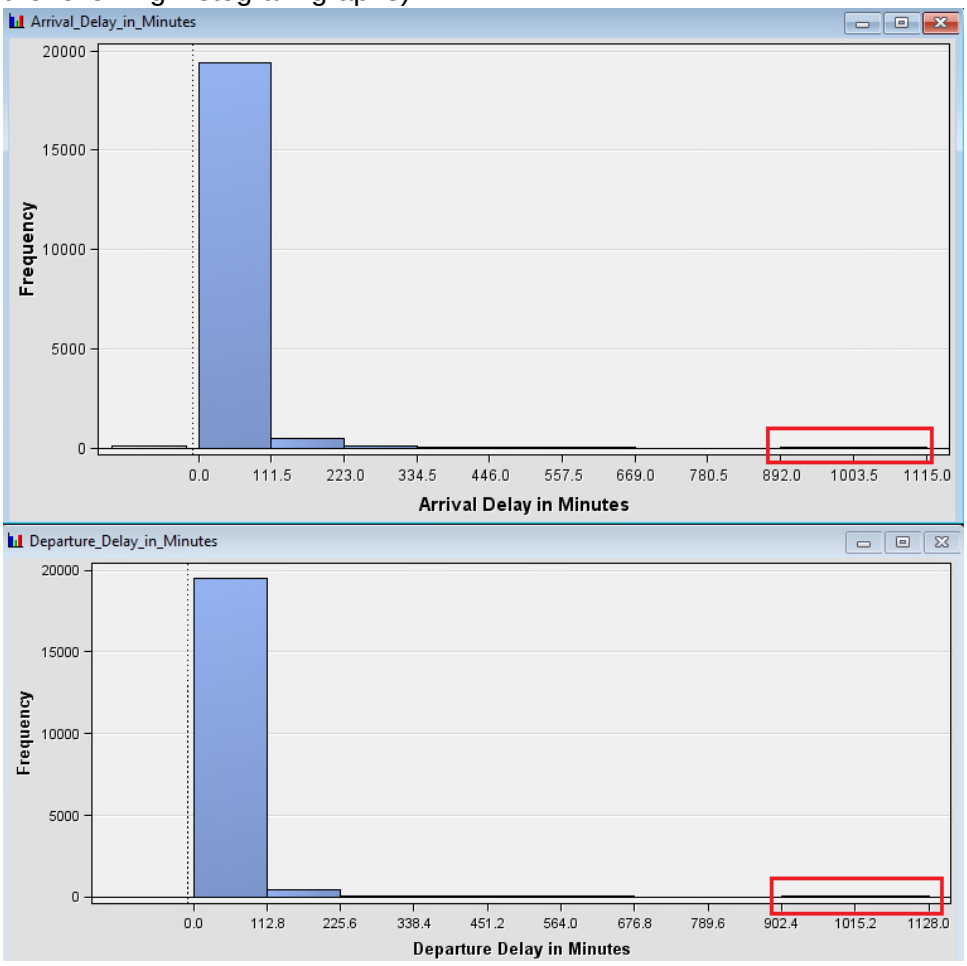
Property	Value
Rows	129880
Columns	23
Library	EMWS1
Member	FIMPORT_DATA
Type	DATA
Sample Method	Random
Fetch Size	Max
Fetched Rows	20000
Random Seed	12345

Sample statistics:

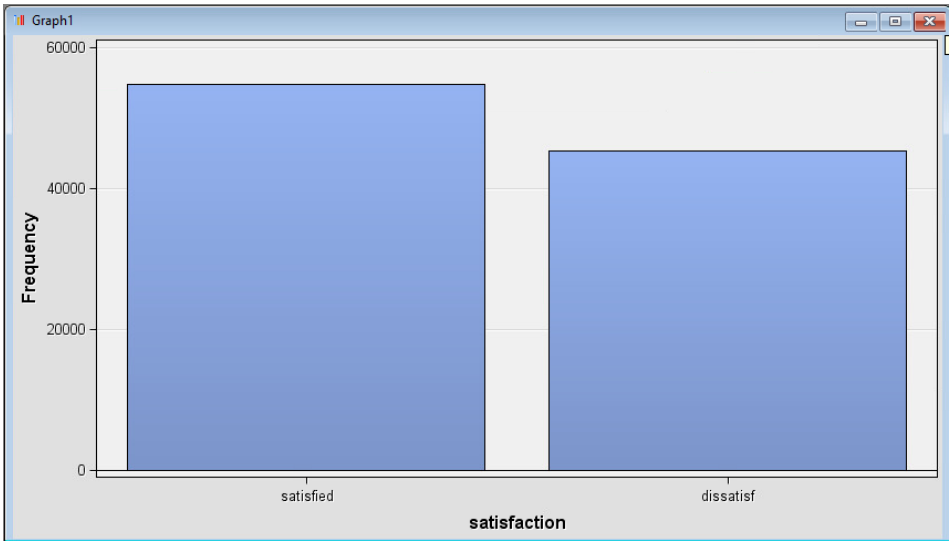
Obs #	Variable Name	Label	Type	Percent Missing	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
14	Flight_Distance	Flight Distance	VAR	0	50	6813	1986.594	.	.	
6	Age		VAR	0	7	85	39.3539	.	.	
8	Baggage_handling	Baggage handling	VAR	0	1	5	3.69575	.	.	
9	Checkin_service	Checkin service	VAR	0	1	5	3.33785	.	.	
22	Online_support	Online support	VAR	0	1	5	3.51225	.	.	
7	Arrival_Delay_in_Minutes	Arrival Delay in Minutes	VAR	0.33	0	1115	14.89877	.	.	
10	Cleanliness		VAR	0	0	5	3.70885	.	.	
11	Departure_Arrival_time_convenien	Departure/Arrival time convenien	VAR	0	0	5	2.98015	.	.	
12	Departure_Delay_in_Minutes	Departure Delay in Minutes	VAR	0	0	1128	14.54685	.	.	
13	Ease_of_Online_booking	Ease of Online booking	VAR	0	0	5	3.4738	.	.	
15	Food_and_drink	Food and drink	VAR	0	0	5	2.8643	.	.	
16	Gate_location	Gate location	VAR	0	0	5	2.9874	.	.	
17	Inflight_entertainment	Inflight entertainment	VAR	0	0	5	3.396	.	.	
18	Inflight_wifi_service	Inflight wifi service	VAR	0	0	5	3.25175	.	.	
19	Leg_room_service	Leg room service	VAR	0	0	5	3.48625	.	.	
20	On_board_service	On-board service	VAR	0	0	5	3.46525	.	.	
21	Online_boarding	Online boarding	VAR	0	0	5	3.34425	.	.	
23	Seat_comfort	Seat comfort	VAR	0	0	5	2.8506	.	.	
1	Class		CLASS	0	.	.	.	3	48.135	BUSINESS
2	Customer_Type	Customer Type	CLASS	0	.	.	.	2	81.34	LOYAL CU...
3	Gender		CLASS	0	.	.	.	2	50.565	FEMALE
4	Type_of_Travel	Type of Travel	CLASS	0	.	.	.	2	69.625	BUSINESS ...
5	satisfaction		CLASS	0	.	.	.	2	54.66	SATISFIED

Observations:

- 33% of missing data in Arrival Delay in Minutes variable.
- For the following variables is considered as minimum value “0” and a maximum of “5”, the scale of satisfaction rate to these variables must be from 1 to 5 so the 0 value is considered as missing value:
    - Cleanliness
    - Departure_Arival_Time_Convinience (rejected)
    - Ease_of_Online_Booking
    - Food_and_Drink
    - Gate_Location(rejected)
    - Inflght_Entertainment
    - Inflight_Wifi_Service
    - Leg_Room_Service
    - On_Board_Service
    - Online_Boarding
    - Seat_Comfort
- The mean value of the Arrival_Delay_In_Minutes (8 outliers) and Deparute_Delay_In_Minutes (6 outliers), the data is skewed to the right (analyzed using the following histogram graphs)



Output Variable



This graph shows that 54.66% are satisfied customers and the other 45.34% are dissatisfied customers.

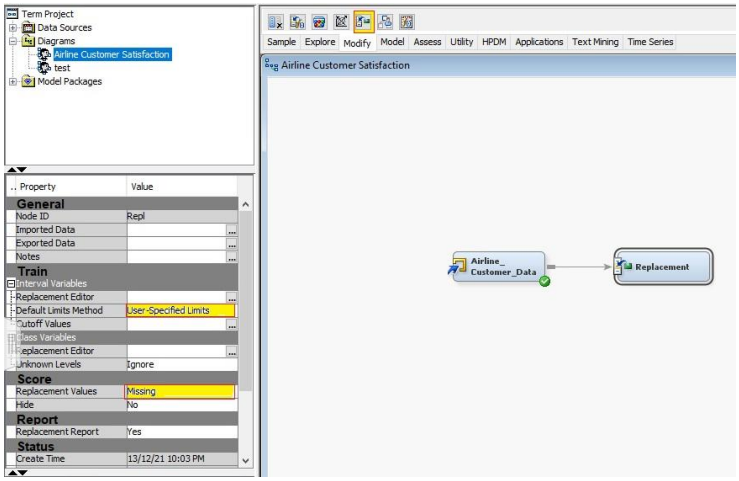
4.4. Modifying and correcting Data

After exploring the data, we need to set missing data criteria so that SAS can flag those data as missing.

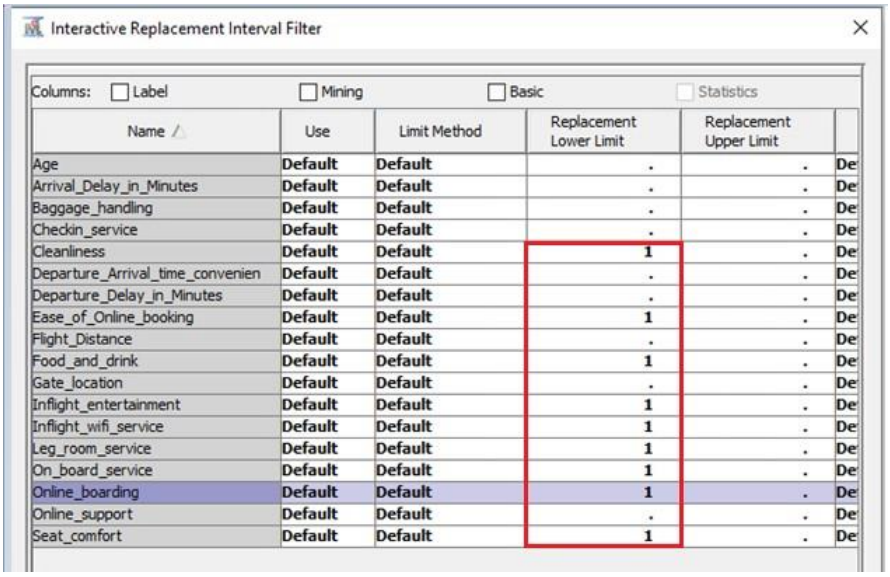
4.4.1. Setting zero as missing value

Steps:

1. Select the **modify** tab on the Toolbar and drag the **Replacement** node into the Diagram workspace.
2. Connect the data node to the replacement node.
3. Make the following changes in the properties panel:
    a. Set Default Limits method to User-specified Limits.
    b. Set Replacement Values to missing.



4. Inside replacement Editor, set the lower limits to 1 of the variables identified having missing value. Considering that the range of these variables is from 1 to 5.



Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	
Age	Default	Default	.	.	De
Arrival_Delay_in_Minutes	Default	Default	.	.	De
Baggage_handling	Default	Default	.	.	De
Checkin_service	Default	Default	.	.	De
Cleanliness	Default	Default	1	.	De
Departure_Arrival_time_convenien	Default	Default	.	.	De
Departure_Delay_in_Minutes	Default	Default	.	.	De
Ease_of_Online_booking	Default	Default	1	.	De
Flight_Distance	Default	Default	.	.	De
Food_and_drink	Default	Default	1	.	De
Gate_location	Default	Default	.	.	De
Inflight_entertainment	Default	Default	1	.	De
Inflight_wifi_service	Default	Default	1	.	De
Leg_room_service	Default	Default	1	.	De
On_board_service	Default	Default	1	.	De
Online_boarding	Default	Default	1	.	De
Online_support	Default	Default	.	.	De
Seat_comfort	Default	Default	1	.	De

5. Click ok and run the node

Sample statistics

From the sample statistics, SAS Miner is considering as “REP_”, those variables for which the lower limit set is 1 as follows:

Obs # ▼	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
32	Seat_comfort	Seat comfort	VAR	0	0	5	2.8506.		.	
31	REP_Seat_...	Replaceme...	VAR	3.5	1	5	2.95399.		.	
30	REP_Onlin...	Replaceme...	VAR	0.005	1	5	3.344417.		.	
29	REP_On_b...	Replaceme...	VAR	0.01	1	5	3.465597.		.	
28	REP_Leg_r...	Replaceme...	VAR	0.41	1	5	3.500602.		.	
27	REP_Infligh...	Replaceme...	VAR	0.08	1	5	3.254353.		.	
26	REP_Infligh...	Replaceme...	VAR	1.95	1	5	3.463539.		.	
25	REP_Food...	Replaceme...	VAR	4.345	1	5	2.994407.		.	
24	REP_Ease...	Replaceme...	VAR	0.015	1	5	3.474321.		.	
23	REP_Clean...	Replaceme...	VAR	0.01	1	5	3.709221.		.	
22	Online_sup...	Online sup...	VAR	0	1	5	3.51225.		.	
21	Online_boa...	Online boar...	VAR	0	0	5	3.34425.		.	
20	On_board_...	On-board s...	VAR	0	0	5	3.46525.		.	
19	Leg_room_...	Leg room s...	VAR	0	0	5	3.48625.		.	
18	Inflight_wifi...	Inflight wifi ...	VAR	0	0	5	3.25175.		.	
17	Inflight_ent...	Inflight ente...	VAR	0	0	5	3.396.		.	
16	Gate_locati...	Gate location	VAR	0	0	5	2.9874.		.	
15	Food_and_...	Food and d...	VAR	0	0	5	2.8643.		.	
14	Flight_Dist...	Flight Dista...	VAR	0	50	6813	1986.594.		.	
13	Ease_of_O...	Ease of Onl...	VAR	0	0	5	3.4738.		.	

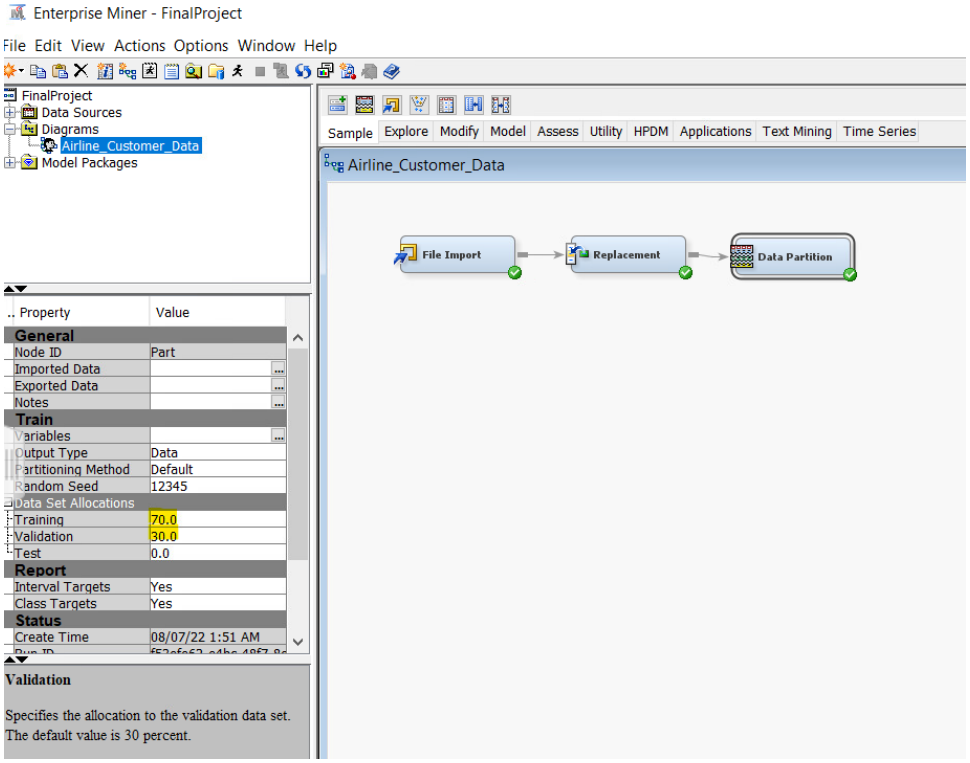
4.5. Data Partitioning

Data splitting is a common practice in predictive modelling for evaluating model performance. We will divide the data into two parts for this project: training data (80%) and validation data (20%) . The training data is used to fit the model, while the validation data is utilized to monitor and tune the model in order to improve its performance by optimizing the chosen model.

Validation data is used to choose the optimal model for various sorts of models and complexities, as well as to optimize the chosen model.

Steps:

1. Select the **Sample** tab on the Toolbar and drag the **Data Partition** node into the Diagram workspace.
2. Connect the replacement node to the **Data Partition** node
3. Make the following changes in the properties panel
    a. Set training data set allocations to 70
    b. Set validation data set allocation to 30
    c. Set test data to 0
4. Run the node



Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Repl_TRAIN	129880
TRAIN	EMWS1.Part_TRAIN	90915
VALIDATE	EMWS1.Part_VALIDATE	38965

5.0. Modelling

5.1. Decision Tree

Decision trees employ a tree structure to display the predictions that arise from a series of feature-based splits. It begins with a root node and finishes with a decision made by leaves. It covers all of the fundamentals of modelling essentials:

- Prediction rules are used to score cases.
- The input selection is aided by a split-search method.
- Pruning is used to deal with model complexity.

The log worth value determines the quality of the split in a decision tree. The threshold log worth value for the split is set to 0.7 by default and without any external imputation, decision trees can manage missing values on their own.

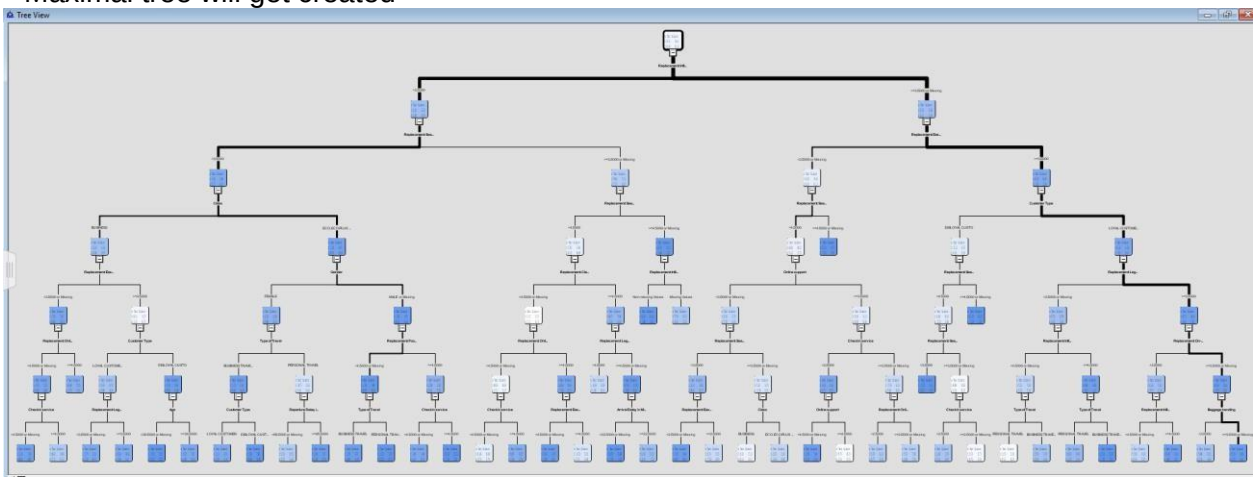
5.1.1. Maximal Tree

The maximal tree represents the most complicated model you are willing to construct from a set of training data. It has the maximum number of splits, and it splits till the log worth drops below the threshold value.

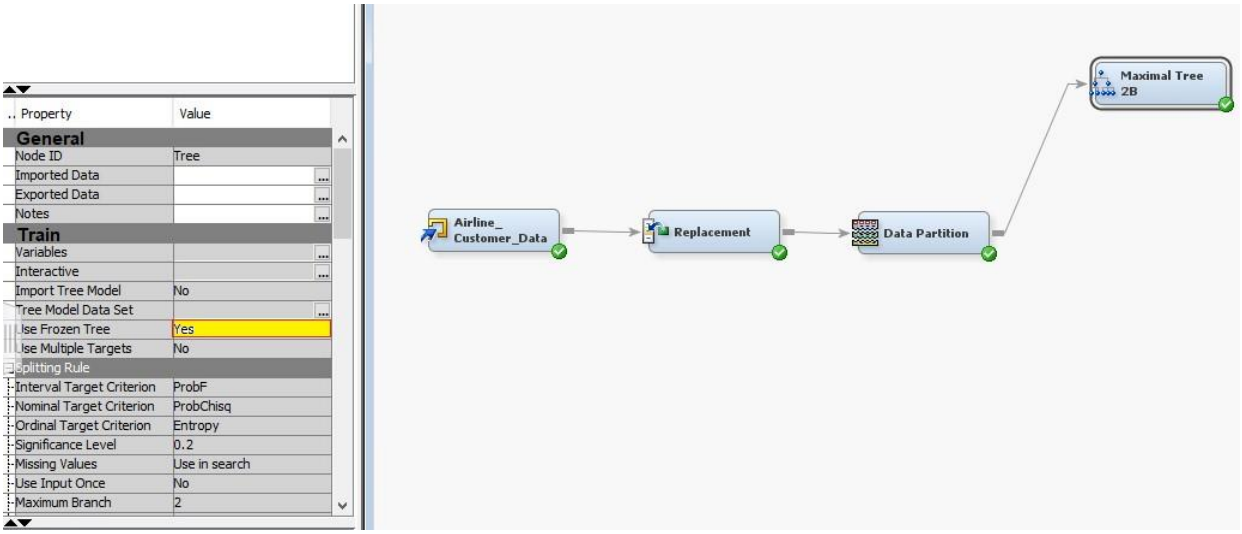
steps

1. Select the **Model** tab on the Toolbar and drag the **Decision Tree** node into the Diagram workspace.
2. Rename the node to **Maximal Tree 2B.**
3. Connect the **Data Partition** node to the **Decision Tree** node.
4. On the properties panel, click on the interactive ellipse.
5. Right click on the root node and select the train node.

Maximal tree will get created



1. Close the interactive window.
2. Change Use frozen tree from no to yes in properties panel, to restrict change to the maximal tree because of other changes.
3. Run the maximal tree and record the average square error value



Observation:

Number of leaves : 45
Average square error of validation data – 0.077241

Fit statistics

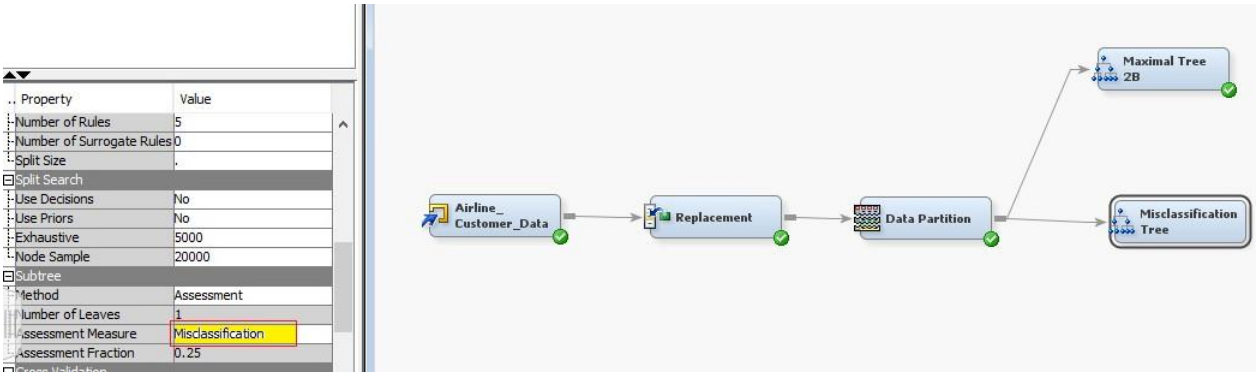
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		NOBS	Sum of Frequencies	90915	38965	.
satisfaction		MISC	Misclassification Rate	0.106264	0.109072	.
satisfaction		MAX	Maximum Absolute Error	0.999487	1	.
satisfaction		SSE	Sum of Squared Errors	13628.73	6019.388	.
satisfaction		ASE	Average Squared Error	0.074953	0.077241	.
satisfaction		RASE	Root Average Squared Error	0.273776	0.277923	.
satisfaction		DIV	Divisor for ASE	181830	77930	.
satisfaction		DFT	Total Degrees of Freedom	90915	.	.

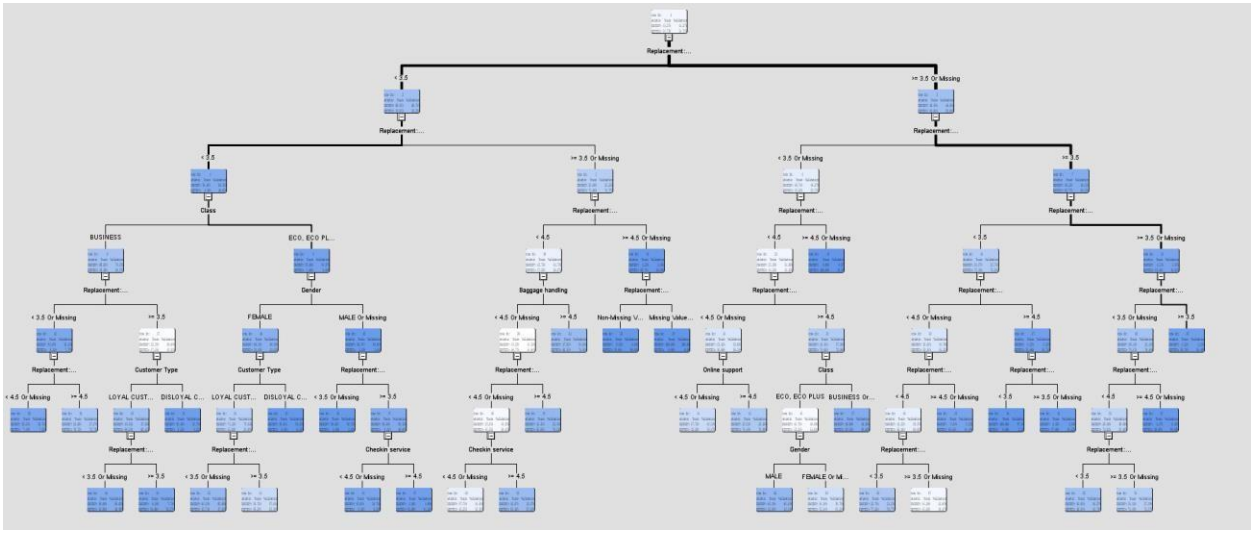
5.1.2. Misclassification Tree

Steps:

1. Select the **Model** tab on the Toolbar and drag the **Decision Tree** node into the Diagram workspace.
2. Rename the node to **Misclassification Tree.**
3. Connect the **Data Partition** node to the **Misclassification Tree** node.
4. On the properties panel, change the Assessment Measure to **Misclassification.**
5. Run the node.



Tree



Fit statistics

Fit Statistics

Target	Target Label	Fit Statistics ▲	Statistics Label	Train	Validation	Test
satisfaction		ASE	Average Squared Error	0.076787	0.07834	
satisfaction		DFT	Total Degrees of Freedom	90915		
satisfaction		DIV	Divisor for ASE	181830	77930	
satisfaction		MAX	Maximum Absolute Error	0.999749	1	
satisfaction		MISC	Misclassification Rate	0.10358	0.104812	
satisfaction		NOBS	Sum of Frequencies	90915	38965	
satisfaction		RASE	Root Average Squared Error	0.277104	0.279892	
satisfaction		SSE	Sum of Squared Errors	13962.11	6105.014	

Observation:

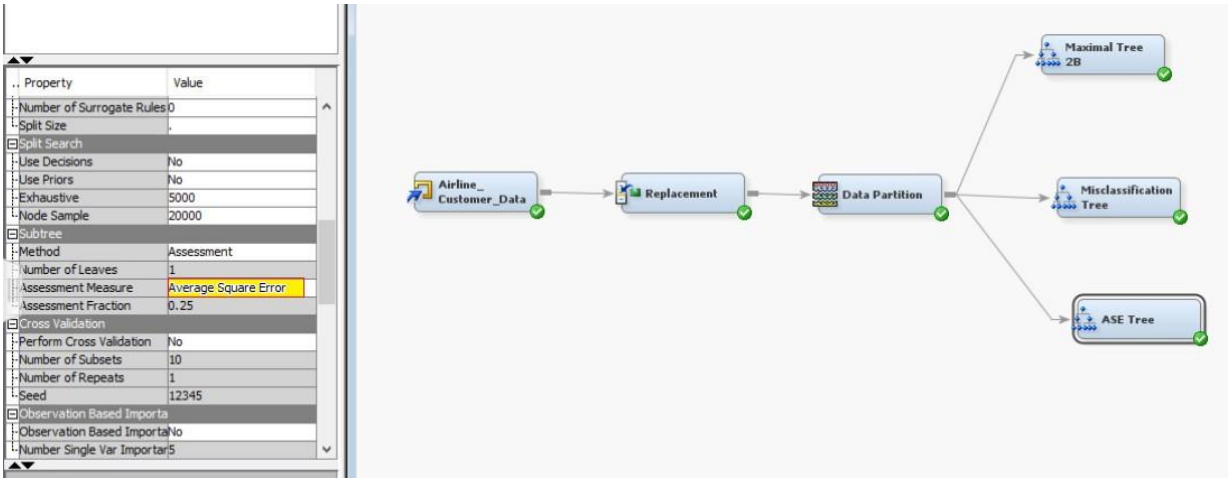
Average square error – 0.07834

5.1.3. Average square Error Tree

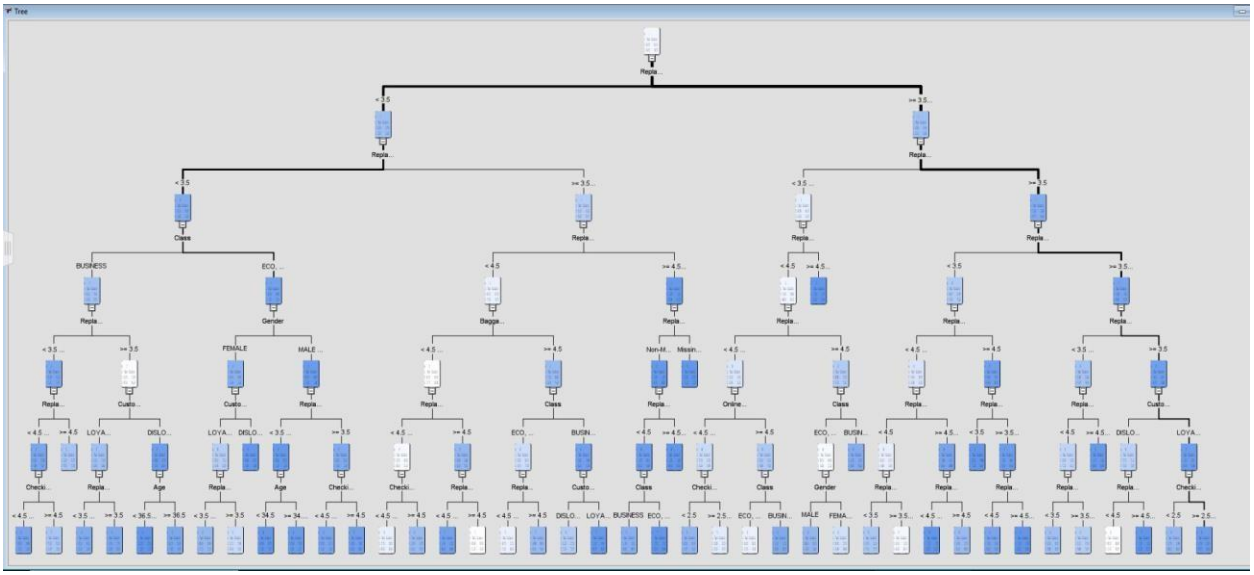
2 Branches

Steps

1. Select the **Model** tab on the Toolbar and drag the **Decision Tree** node into the Diagram workspace.
2. Rename the node to **ASE 2B Tree**
3. Connect the **Data Partition** node to the **ASE 2B Tree** node
4. On the properties panel, change the Assessment Measure to **Average Square Error**



Tree



Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		NOBS	Sum of Frequencies	90915	38965	
satisfaction		MISC	Misclassification Rate	0.10358	0.104812	
satisfaction		MAX	Maximum Absolute Error	0.999749	1	
satisfaction		SSE	Sum of Squared Errors	13439.65	5891.89	
satisfaction		ASE	Average Squared Error	0.073913	0.075605	
satisfaction		RASE	Root Average Squared Error	0.27187	0.274963	
satisfaction		DIV	Divisor for ASE	181830	77930	
satisfaction		DFT	Total Degrees of Freedom	90915		

Observation:

Average square error – 0.075605

5.2. Regression

The statistical link between a dependent variable and one or more independent variables is determined using regression. If our target has an interval variable, we will use a linear regression model. If our target has a binary value, logistic regression will be our model of choice. The following prediction formula is used in the logistic regression model.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2$$ logit scores

5.2.1. Imputation

Finding a replacement for a missing value is known as imputation. After agreeing on the formula, we must impute the missing value, as logistic regression, unlike decision trees, cannot work with missing data. Treating the missing variable as zero in regression will result in a skewed prediction outcome.

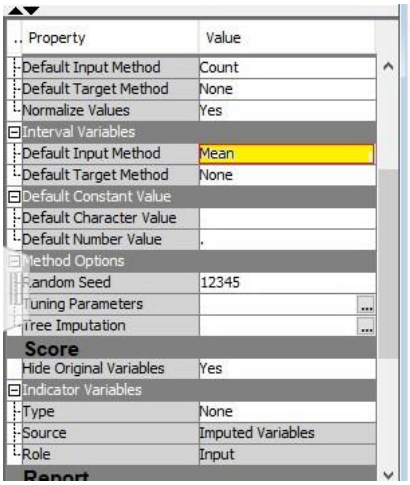
1. In regression, complete-case analysis is the default strategy for dealing with missing values. As a result, only full cases are considered in this method's analysis.
2. Cases with missing values are not scored by the prediction method.

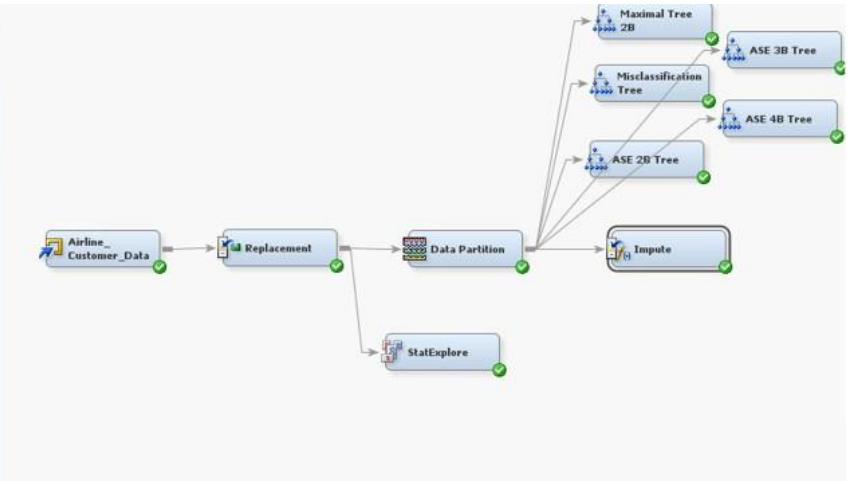
Remedy for the missing values

1. Synthetic distribution: In this situation, the mean, mode, or median values are calculated and utilized to fill in the database's missing value.
2. Estimation approach: In this method, missing data are handled as a prediction problem, and the analyst develops a model to locate them.

Steps:

1. Select the **Modify** tab on the Toolbar and drag the **Impute** node into the Diagram workspace.
2. Connect the **Data Partition** node to the **Impute** node.
3. On the properties panel make the following changes:
    a. Change the Default input Method to Mean.
4. Run the nodes.





Imputation Summary

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Number of Missing for TRAIN
Arrival Delay in Minutes	MEAN	IMP_Arrival Delay in Minutes	14.94606	INPUT	280
REP_Cleanliness	MEAN	IMP_REP_Cleanliness	3.704142	INPUT	3
REP_Ease of Online booking	MEAN	IMP_REP_Ease of Online booking	3.471811	INPUT	12
REP_Food and drink	MEAN	IMP_REP_Food and drink	2.993017	INPUT	4127
REP_Inflight entertainment	MEAN	IMP_REP_Inflight entertainment	3.464392	INPUT	2072
REP_Inflight wifi service	MEAN	IMP_REP_Inflight wifi service	3.254429	INPUT	88
REP_Leg room service	MEAN	IMP_REP_Leg room service	3.496507	INPUT	316
REP_On board service	MEAN	IMP_REP_On board service	3.461292	INPUT	3
REP_Online boarding	MEAN	IMP_REP_Online boarding	3.353094	INPUT	10
REP_Seat comfort	MEAN	IMP_REP_Seat comfort	2.950526	INPUT	3333

5.2.2. Transformation or managing skewness and outlier variables

Since regression is sensitive to very skewed and outlier values, these skewed and outlier values could get selected over significant variables diminishing model quality.

From the Stat Explorer result of Original Data shown below its clear that the Departure_Delay_In_Minutes and Arrival_Delay_In_Minutes is positively skewed. We can handle the skewness by performing Cap and Floor and/or LOG Transformation

Stat Explorer Result of Original Data

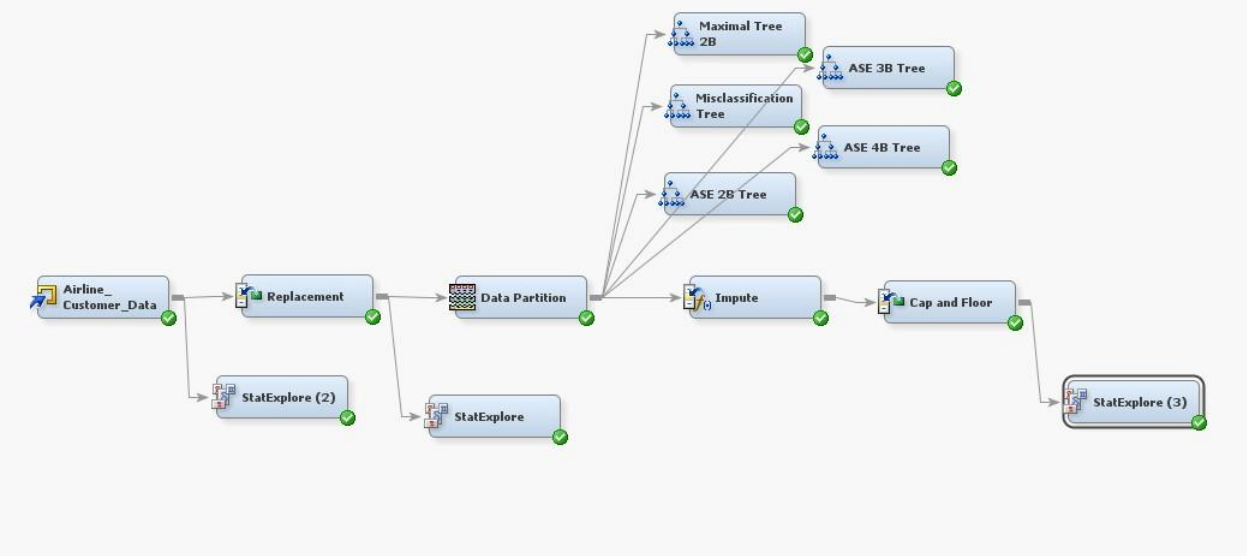
Data Role	Target	Target Level	Variable	Skewness ▼	Mean
TRAIN	satisfaction	satisfied	Departure_Delay_in_Minutes	7.184991	12.15477
TRAIN	satisfaction	satisfied	Arrival_Delay_in_Minutes	6.996799	12.26888
TRAIN	satisfaction	dissatisf	Departure_Delay_in_Minutes	6.607497	17.80775
TRAIN	satisfaction	dissatisf	Arrival_Delay_in_Minutes	6.477211	18.5046
TRAIN	satisfaction	dissatisf	Age	0.134164	37.46667
TRAIN	satisfaction	dissatisf	Seat_comfort	-0.05666	2.467335
TRAIN	satisfaction	dissatisf	Inflight_wifi_service	-0.07982	2.919854
TRAIN	satisfaction	dissatisf	Food_and_drink	-0.12869	2.660419
TRAIN	satisfaction	dissatisf	Online_boarding	-0.19416	2.869695

5.2.3. Cap and floor

We use a replacement node to lessen the skewness. The standard deviation from the mean is used as the default limit approach for decreasing skewness in the replacement node. We call it Cap and Floor because we are bringing the data within the standard deviation.

Steps:

1. Select the **Modify** tab on the Toolbar and drag the **replacement** node into the Diagram workspace.
2. Rename it to **Cap and Floor**
3. Connect the **Cap and Floor** node to the **Impute** node
4. Run the nodes
5. Connect a StatExplore and check whether it improves the skewness



Stat Explore Result after Cap and floor

Interval Variables

Data Role	Target	Target Level	Variable	Skewness ▼
TRAIN	satisfaction	satisfied	REP Departu...	3.040366
TRAIN	satisfaction	satisfied	REP IMP Arr...	3.031125
TRAIN	satisfaction	dissatisf	REP Departu...	2.490265
TRAIN	satisfaction	dissatisf	REP IMP Arr...	2.472444
TRAIN	satisfaction	dissatisf	REP Age	0.26767
TRAIN	satisfaction	dissatisf	REP IMP R...	0.166417
TRAIN	satisfaction	dissatisf	REP IMP R...	0.145385
TRAIN	satisfaction	dissatisf	REP IMP R...	0.100161
TRAIN	satisfaction	dissatisf	REP IMP R...	0.090283
TRAIN	satisfaction	dissatisf	REP IMP R...	0.041403
TRAIN	satisfaction	dissatisf	REP Online ...	-0.00655

Observation

1. We can see that cap and floor has improved the skewness when we compare it to the original data.
2. It still positively skewed, so we need to further transform the data to handle the skewness, this can be done by log transformation

5.2.2.2. LOG Transformation

A Transform variable node can be used to solve this problem. This node manages the input distributions by applying log to these skewed and outlier values

Steps

1. Select the **Modify** tab on the Toolbar and drag the **Transform Variable** node into the Diagram workspace.
2. Connect the **Transform Variable** node to the **Cap and Floor** node
3. Right click on the transform variable node and select edit variable
4. Apply log method to REP_Departure_Delay_in_Minutes and REP_IMP_Arrival_Delay_in_Minutes
5. Run the nodes

Transformations Statistics

Source	Method	Variable Name	Skewness
Input	Original	REP Departure ...	2.77065
Input	Original	REP IMP Arrival ...	2.754659
Output	Computed	LOG REP Depar...	0.869052
Output	Computed	LOG REP IMP A...	0.818077

Observation:

1. LOG Transformation has further improved the skewness as seen in the transformation statistics
2. Now that the missing values and skewness is handled, we can run the regression model

5.2.4. Regression Models

Four types of regression models are executed in this project

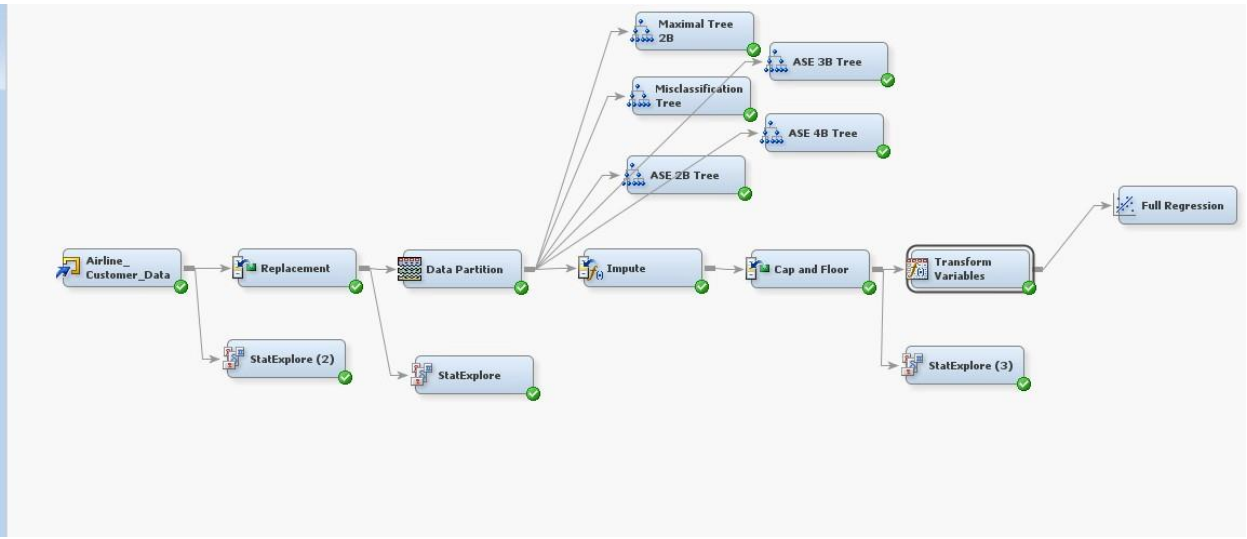
1. Full Regression
2. Forward Regression
3. Backward Regression
4. Stepwise Regression

5.2.3.1. Full Regression

Full regression is executed without setting any selection model.

Steps:

1. Select the **Model** tab on the Toolbar and drag the **Regression** node into the Diagram workspace.
2. Rename it to **Full Regression.**
3. Connect the **Full Regression** node to the **Transform Variable** node.
4. Run the nodes.



Fit statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		AIC	Akaike's Information Criterion	63491.91		
satisfaction		ASE	Average Squared Error	0.10924	0.110263	
satisfaction		AVERR	Average Error Function	0.348963	0.351995	
satisfaction		DFE	Degrees of Freedom for Error	90895		
satisfaction		DFM	Model Degrees of Freedom	20		
satisfaction		DFT	Total Degrees of Freedom	90915		
satisfaction		DIV	Divisor for ASE	181830	77930	
satisfaction		ERR	Error Function	63451.91	27430.98	
satisfaction		FPE	Final Prediction Error	0.109288		
satisfaction		MAX	Maximum Absolute Error	0.99753	0.998243	
satisfaction		MSE	Mean Square Error	0.109264	0.110263	
satisfaction		NOBS	Sum of Frequencies	90915	38965	
satisfaction		NW	Number of Estimate Weights	20		
satisfaction		RASE	Root Average Sum of Squares	0.330515	0.332058	
satisfaction		RFPE	Root Final Prediction Error	0.330588		
satisfaction		RMSE	Root Mean Squared Error	0.330552	0.332058	
satisfaction		SBC	Schwarz's Bayesian Criterion	63680.27		
satisfaction		SSE	Sum of Squared Errors	19863.17	8592.77	
satisfaction		SUMW	Sum of Case Weights Times Freq	181830	77930	
satisfaction		MISC	Misclassification Rate	0.153231	0.155909	

Observation:

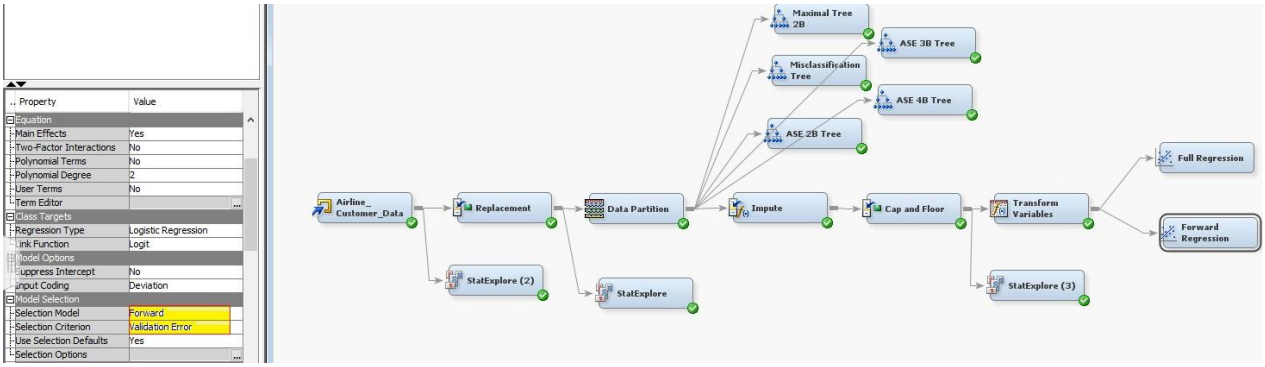
Average square error – 0.110263

5.2.3.2. Forward Regression

Forward Regression is a regression approach that begins with an empty model and at each step gradually adds variables to the regression model to find a model that best explains the data.

Steps:

1. Select the **Model** tab on the Toolbar and drag the **Regression** node into the Diagram workspace.
2. Rename it to **Forward Regression**
3. Connect the **Forward Regression** node to the **Transform Variable** node
4. On the properties panel make the following changes:
    a. Change the Selection model to Forward
    b. Change Selection criterion to Validation error
5. Run the nodes



Fit statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		AIC	Akaike's Information Criterion	63489		
satisfaction		ASE	Average Squared Error	0.109245	0.110277	
satisfaction		AVERR	Average Error Function	0.348969	0.35203	
satisfaction		DFE	Degrees of Freedom for Error	90897		
satisfaction		DFM	Model Degrees of Freedom	18		
satisfaction		DFT	Total Degrees of Freedom	90915		
satisfaction		DIV	Divisor for ASE	181830	77930	
satisfaction		ERR	Error Function	63453	27433.7	
satisfaction		FPE	Final Prediction Error	0.109288		
satisfaction		MAX	Maximum Absolute Error	0.997575	0.998279	
satisfaction		MSE	Mean Square Error	0.109266	0.110277	
satisfaction		NOBS	Sum of Frequencies	90915	38965	
satisfaction		NW	Number of Estimate Weights	18		
satisfaction		RASE	Root Average Sum of Squares	0.330522	0.33208	
satisfaction		RFPE	Root Final Prediction Error	0.330587		
satisfaction		RMSE	Root Mean Squared Error	0.330555	0.33208	
satisfaction		SBC	Schwarz's Bayesian Criterion	63658.52		
satisfaction		SSE	Sum of Squared Errors	19863.97	8593.901	
satisfaction		SUMW	Sum of Case Weights Times Freq	181830	77930	
satisfaction		MISC	Misclassification Rate	0.153055	0.155858	

Observation:

Average square error – 0.110277

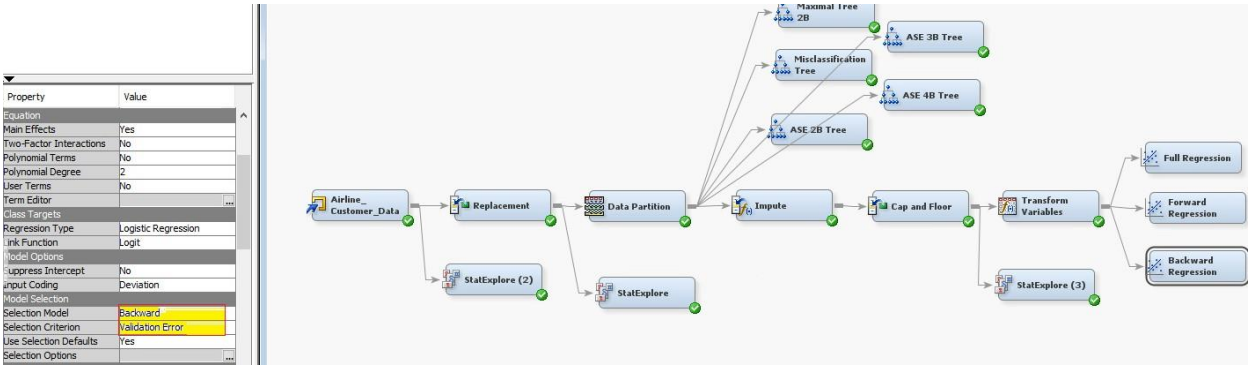
5.2.3.3. Backward Regression

Backward is a regression approach that begins with a full model and at each step gradually eliminates variables from the regression model to find a reduced model that best explains the data. Also known as Backward Elimination Regression.

Steps:

1. Select the **Model** tab on the Toolbar and drag the **Regression** node into the Diagram workspace.
2. Rename it to Backward **Regression**
3. Connect the Backward **Regression** node to the **Transform Variable** node

4. On the properties panel make the following changes:
    a. Change the Selection model to Backward
    b. Change Selection criterion to Validation error
5. Run the nodes



Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		AIC	Akaike's Information Criterion	63491.91		
satisfaction		ASE	Average Squared Error	0.10924	0.110263	
satisfaction		AVERR	Average Error Function	0.348963	0.351995	
satisfaction		DFE	Degrees of Freedom for Error	90895		
satisfaction		DFM	Model Degrees of Freedom	20		
satisfaction		DFT	Total Degrees of Freedom	90915		
satisfaction		DIV	Divisor for ASE	181830	77930	
satisfaction		ERR	Error Function	63451.91	27430.98	
satisfaction		FPE	Final Prediction Error	0.109288		
satisfaction		MAX	Maximum Absolute Error	0.99753	0.998243	
satisfaction		MSE	Mean Square Error	0.109264	0.110263	
satisfaction		NOBS	Sum of Frequencies	90915	38965	
satisfaction		NW	Number of Estimate Weights	20		
satisfaction		RASE	Root Average Sum of Squares	0.330515	0.332058	
satisfaction		RFPE	Root Final Prediction Error	0.330588		
satisfaction		RMSE	Root Mean Squared Error	0.330552	0.332058	
satisfaction		SBC	Schwarz's Bayesian Criterion	63680.27		
satisfaction		SSE	Sum of Squared Errors	19863.17	8592.77	
satisfaction		SUMW	Sum of Case Weights Times Freq	181830	77930	
satisfaction		MISC	Misclassification Rate	0.153231	0.155909	

Observation:

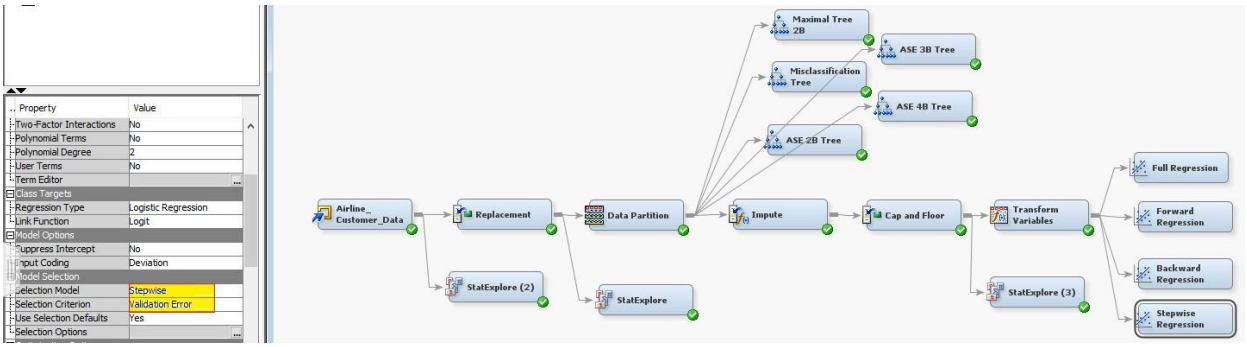
Average square error – 0.110263

5.2.3.4. Stepwise Regression

Stepwise Regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables

Steps:

1. Select the **Model** tab on the Toolbar and drag the **Regression** node into the Diagram workspace.
2. Rename it to Stepwise **Regression**
3. Connect the Stepwise **Regression** node to the **Transform Variable** node
4. On the properties panel make the following changes:
    a. Change the Selection model to Stepwise
    b. Change Selection criterion to Validation error
5. Run the nodes



Fit Statistics

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		AIC	Akaike's Information Criterion	63489		
satisfaction		ASE	Average Squared Error	0.109245	0.110277	
satisfaction		AVERR	Average Error Function	0.348969	0.35203	
satisfaction		DFE	Degrees of Freedom for Error	90897	.	
satisfaction		DFM	Model Degrees of Freedom	18	.	
satisfaction		DFT	Total Degrees of Freedom	90915		
satisfaction		DIV	Divisor for ASE	181830	77930	
satisfaction		ERR	Error Function	63453	27433.7	
satisfaction		FPE	Final Prediction Error	0.109288	.	
satisfaction		MAX	Maximum Absolute Error	0.997575	0.998279	
satisfaction		MSE	Mean Square Error	0.109266	0.110277	
satisfaction		NOBS	Sum of Frequencies	90915	38965	
satisfaction		NW	Number of Estimate Weights	18	.	
satisfaction		RASE	Root Average Sum of Squares	0.330522	0.33208	
satisfaction		RFPE	Root Final Prediction Error	0.330587	.	
satisfaction		RMSE	Root Mean Squared Error	0.330555	0.33208	
satisfaction		SBC	Schwarz's Bayesian Criterion	63658.52	.	
satisfaction		SSE	Sum of Squared Errors	19863.97	8593.901	
satisfaction		SUMW	Sum of Case Weights Times Freq	181830	77930	
satisfaction		MISC	Misclassification Rate	0.153055	0.155858	

Observation:

Average square error – 0.110277

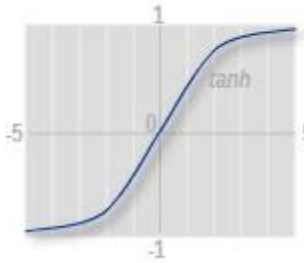
6.0. Neural Network

Neural networks are systems that are based on biological neural networks. Without any task-specific rules, these systems learn to do tasks by being exposed to a variety of datasets and examples. It is a regression model extension because it employs the same formula as regression but with some modifications. This allows the neural network's variables to interact with one another. In other words, A neural network is a regression model based on a set of hidden units.

Prediction formula:

$$\hat{y} = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$

prediction estimate; bias estimate; weight estimate; hidden unit



$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$

$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$

$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

activation function

6.1. Neural Network Models:

X types of neural networks are executed under this project.

6.1.1. Impute Neural Network

6.1.1.1. Neural network with 3 hidden unit and 50 iterations

Steps:

1. Select the **Model** tab on the Toolbar and drag the **Neural Network** node into the Diagram workspace.
2. Rename it to NN 3H 100I
3. Connect the **Neural Network** node to the **Impute** node
4. On the properties panel make the following changes:
    a. Change the Model Selection criteria to Average Error
    b. Click the optimization ellipse and set maximum iteration to 100 and set preliminary training Enable to No.
5. Run the nodes

6.1.1.2. Neural Network with 50 iteration and Different hidden units

We are running the Neural Network model at 50 iterations because the data converges at 80

iterations Steps:

1. Select the **Model** tab on the Toolbar and drag the **Neural Network** node into the Diagram workspace.
2. Connect the **Neural Network** node to the **Impute** node.
3. On the properties panel make the following changes:
    a. Change the Model Selection criteria to Average Error.
    b. Click the optimization ellipse and set preliminary training Enable to No.
4. On the Properties panel click on Network ellipse and set the Hidden Unit for the following renamed neural networks.
5. Run the nodes

Neural Network	Hidden Units
NN 3H 50I	3
NN 4H 50I	4
NN 5H 50I	5
NN 8H 50I	8

6.1.1.3. Neural Network using Transform Variable

Steps:

1. Select the **Model** tab on the Toolbar and drag the **Neural Network** node into the Diagram workspace.
2. Connect the **Neural Network** node to the **Transform Variable** node
3. On the properties panel make the following changes:
    a. Change the Model Selection criteria to Average Error.
    b. Click the optimization ellipse and set preliminary training Enable to No.
4. On the Properties panel click on Network ellipse and set the Hidden Unit for the following renamed neural networks.
5. Run the nodes

Neural Network	Hidden Units
NN Transform 3H 50I	3
NN Transform 8H 50I	8

6.1.1.4. Neural Network using Backward Regression

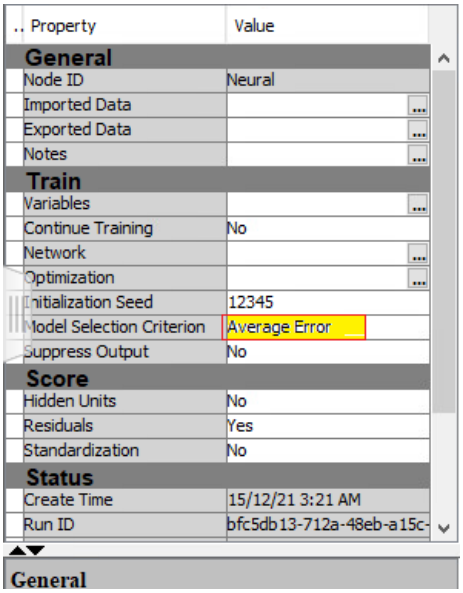
We are considering running neural network from Backward regression, as Backward regression has the least ASE value of all the regression model

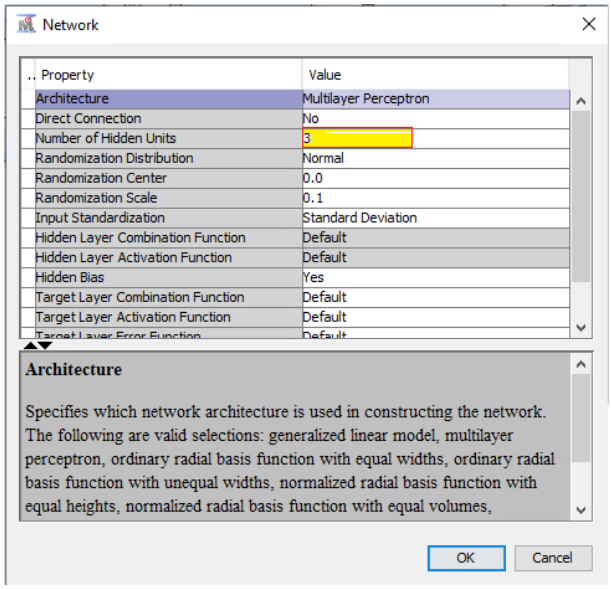
Steps:

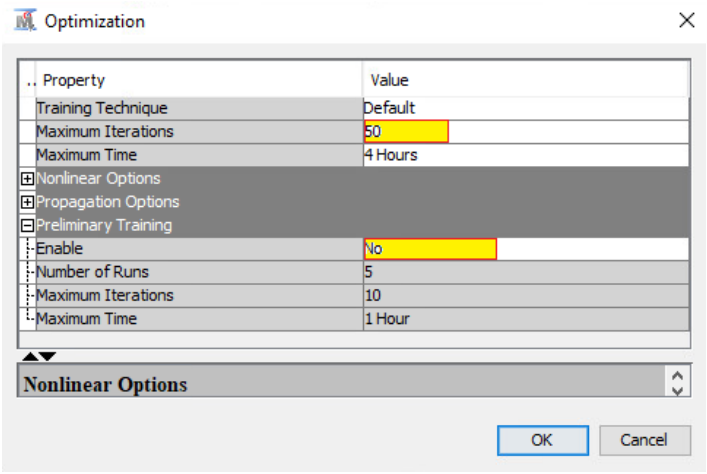
1. Select the **Model** tab on the Toolbar and drag the **Neural Network** node into the Diagram workspace.
2. Connect the **Neural Network** node to the **Backward Regression** node
3. On the properties panel make the following changes:
    a. Change the Model Selection criteria to Average Error.
    b. Click the optimization ellipse and set preliminary training Enable to No.
4. On the Properties panel click on Network ellipse and set the Hidden Unit for the following renamed neural networks
5. Run the nodes

Neural Network	Hidden Units
NN BR 3H 50I	3
NN BR 8H 50I	8

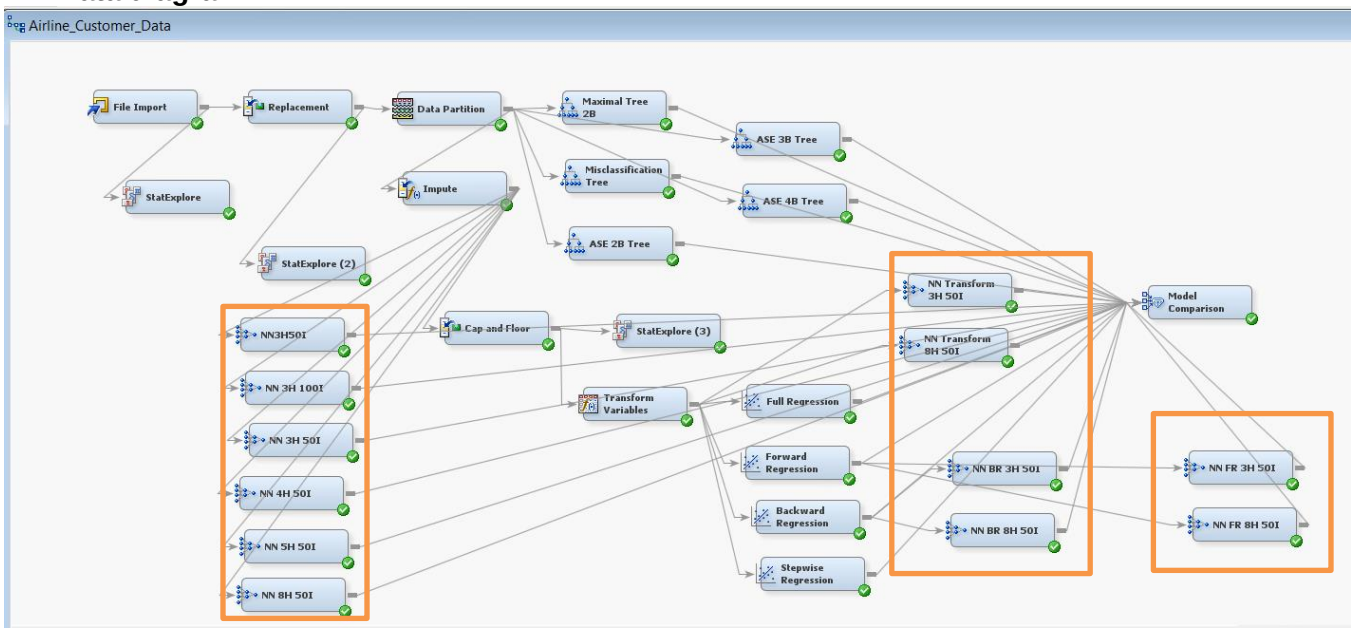
Panel screen shots







Data diagram

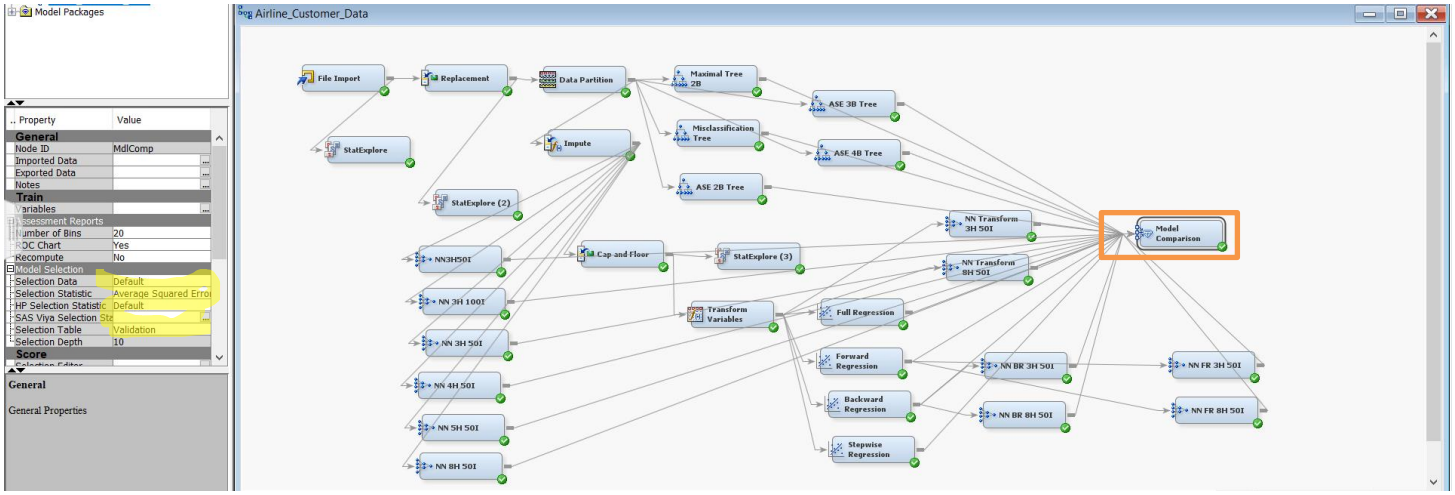


7.0. Model Comparison

The model is rated using the average square error or misclassification rate, profit or loss, and KS statistics in the model comparison tool. We will use the average square error rate to grade the models in this assignment.

Steps:

1. Select the **Assess** tab on the Toolbar and drag the **Model Comparison** node into the Diagram workspace.
2. Connect the **Model Comparison** node to the all the models **(Decision Trees, Regression Models and Neural Network Models)**
3. On the properties panel make the following changes:
    a. Change the Selection statistic to Average Square Error
    b. Click the Selection Table to Validation
4. Run the nodes



8.0. Results

8.1. Fit Statistics analysis of all the models

We are going to analyze the data based on the Average square error of validation data as we have set it as the selection criteria for model comparison

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Average Squared Error ▲	Valid: Roc Index	Valid: Gini Coefficient
Y	Tree5	Tree5	ASE 4B Tree	satisfaction	0.058006	0.975	0.95
	Tree4	Tree4	ASE 3B Tree	satisfaction	0.061058	0.974	0.949
	Neural10	Neural10	NN BR 8H 50I	satisfaction	0.069939	0.968	0.936
	Neural8	Neural8	NN Transform 8H 50I	satisfaction	0.069939	0.968	0.936
	Neural6	Neural6	NN 8H 50I	satisfaction	0.070389	0.967	0.935
	Neural12	Neural12	NN FR 8H 50I	satisfaction	0.072917	0.965	0.93
	Tree3	Tree3	ASE 2B Tree	satisfaction	0.075605	0.959	0.918
	Neural4	Neural4	NN 5H 50I	satisfaction	0.077043	0.961	0.922
	Tree	Tree	Maximal Tree 2B	satisfaction	0.077241	0.959	0.918
	Tree2	Tree2	Misclassification Tree	satisfaction	0.07834	0.953	0.906
	Neural3	Neural3	NN 4H 50I	satisfaction	0.07928	0.958	0.917
	Neural11	Neural11	NN FR 3H 50I	satisfaction	0.0851	0.953	0.907
	Neural5	Neural5	NN 3H 100I	satisfaction	0.085683	0.953	0.907
	Neural7	Neural7	NN Transform 3H 50I	satisfaction	0.085795	0.953	0.905
	Neural9	Neural9	NN BR 3H 50I	satisfaction	0.085795	0.953	0.905
	Neural	Neural	NN3H50I	satisfaction	0.08616	0.952	0.905
	Neural2	Neural2	NN 3H 50I	satisfaction	0.08616	0.952	0.905
	Reg	Reg	Full Regression	satisfaction	0.110263	0.923	0.846
	Reg3	Reg3	Backward Regression	satisfaction	0.110263	0.923	0.846
	Reg2	Reg2	Forward Regression	satisfaction	0.110277	0.923	0.846
	Reg4	Reg4	Stepwise Regression	satisfaction	0.110277	0.923	0.846

Based on the Fit statistics, Decision Tree with 4 branches and with Average Squared Error (ASE 4B Tree) is the best model as it has the least Average square error at 0.058006.

8.2. Ranking based ROC index and Gini Coefficient

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Average Squared Error ▲	Valid: Roc Index	Valid: Gini Coefficient
Y	Tree5	Tree5	ASE 4B Tree	satisfaction	0.058006	0.975	0.95
	Tree4	Tree4	ASE 3B Tree	satisfaction	0.061058	0.974	0.949
	Neural10	Neural10	NN BR 8H 50I	satisfaction	0.069939	0.968	0.936
	Neural8	Neural8	NN Transform 8H 50I	satisfaction	0.069939	0.968	0.936
	Neural6	Neural6	NN 8H 50I	satisfaction	0.070389	0.967	0.935
	Neural12	Neural12	NN FR 8H 50I	satisfaction	0.072917	0.965	0.93
	Tree3	Tree3	ASE 2B Tree	satisfaction	0.075605	0.959	0.918
	Neural4	Neural4	NN 5H 50I	satisfaction	0.077043	0.961	0.922
	Tree	Tree	Maximal Tree 2B	satisfaction	0.077241	0.959	0.918
	Tree2	Tree2	Misclassification Tree	satisfaction	0.07834	0.953	0.906
	Neural3	Neural3	NN 4H 50I	satisfaction	0.07928	0.958	0.917
	Neural11	Neural11	NN FR 3H 50I	satisfaction	0.0851	0.953	0.907
	Neural5	Neural5	NN 3H 100I	satisfaction	0.085683	0.953	0.907
	Neural7	Neural7	NN Transform 3H 50I	satisfaction	0.085795	0.953	0.905
	Neural9	Neural9	NN BR 3H 50I	satisfaction	0.085795	0.953	0.905
	Neural	Neural	NN3H50I	satisfaction	0.08616	0.952	0.905
	Neural2	Neural2	NN 3H 50I	satisfaction	0.08616	0.952	0.905
	Reg	Reg	Full Regression	satisfaction	0.110263	0.923	0.846
	Reg3	Reg3	Backward Regression	satisfaction	0.110263	0.923	0.846
	Reg2	Reg2	Forward Regression	satisfaction	0.110277	0.923	0.846
	Reg4	Reg4	Stepwise Regression	satisfaction	0.110277	0.923	0.846

Decision Tree with 4 branches and with Average Squared Error (ASE 4B Tree) has the Highest ROC index i.e., area under the curve and highest Gini coefficient.

8.3. Outcome:

8.3.1. Variable importance of best regression model

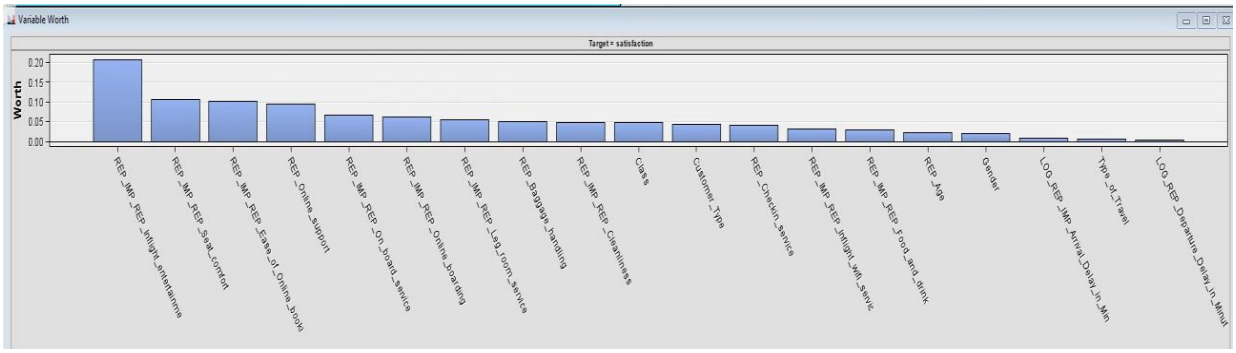
The following is the variable importance of ASE 4B Tree:

Fit Statistics

Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Tree5	ASE 4B Tree	0.05801	0.05544	0.07533	0.07812
	Tree4	ASE 3B Tree	0.06106	0.05941	0.08108	0.08343
	Neural10	NN BR 8H 50I	0.06994	0.06906	0.09795	0.09757
	Neural8	NN Transform 8H 50I	0.06994	0.06906	0.09795	0.09757
	Neural6	NN 8H 50I	0.07039	0.06962	0.09784	0.09829
	Neural12	NN FR 8H 50I	0.07292	0.07187	0.10079	0.10148
	Tree3	ASE 2B Tree	0.07560	0.07391	0.10358	0.10481
	Neural4	NN 5H 50I	0.07704	0.07620	0.10821	0.10894
	Tree	Maximal Tree 2B	0.07724	0.07495	0.10626	0.10907
	Tree2	Misclassification Tree	0.07834	0.07679	0.10358	0.10481
	Neural3	NN 4H 50I	0.07928	0.07894	0.11018	0.11107
	Neural11	NN FR 3H 50I	0.08510	0.08487	0.12130	0.12052
	Neural5	NN 3H 100I	0.08568	0.08594	0.12288	0.12308
	Neural7	NN Transform 3H 50I	0.08579	0.08601	0.12248	0.12306
	Neural9	NN BR 3H 50I	0.08579	0.08601	0.12248	0.12306
	Neural	NN3H50I	0.08616	0.08631	0.12274	0.12339
	Neural2	NN 3H 50I	0.08616	0.08631	0.12274	0.12339
	Reg	Full Regression	0.11026	0.10924	0.15323	0.15591
	Reg3	Backward Regression	0.11026	0.10924	0.15323	0.15591
	Reg2	Forward Regression	0.11028	0.10924	0.15306	0.15586
	Reg4	Stepwise Regression	0.11028	0.10924	0.15306	0.15586

Variable worth



Exported Validation data

EMWS1.Neural10_VALIDATE

	Observation Number	satisfaction	Gender	Predicted: satisfaction=satisfied	Replacement: Imputed: Replacement: Inflight entertainment	Replacement: Imputed: Replacement: Seat comfort	Replacement: Imputed: Replacement: Ease of Online booking	Replacement: Online support	Replacement: Imputed: Replacement: On-board service	Replacemer
1	114263.0	satisfied	Male	0.9999996828994572	5.0	5.0	5.0	5.0	5.0	5.0
2	116976.0	satisfied	Male	0.9999996405907589	5.0	5.0	5.0	4.0	5.0	5.0
3	122845.0	satisfied	Male	0.9999995875528442	5.0	5.0	5.0	5.0	5.0	5.0
4	121711.0	satisfied	Female	0.9999994389261391	5.0	5.0	5.0	5.0	5.0	5.0
5	122325.0	satisfied	Female	0.9999994339130395	5.0	5.0	5.0	4.0	5.0	5.0
6	119697.0	satisfied	Male	0.9999994290846882	5.0	5.0	5.0	5.0	5.0	5.0
7	125573.0	satisfied	Female	0.9999993923002902	5.0	5.0	5.0	5.0	5.0	5.0
8	117408.0	satisfied	Female	0.9999993514414849	5.0	5.0	5.0	5.0	5.0	5.0
9	121715.0	satisfied	Female	0.9999993470038117	5.0	5.0	5.0	4.0	5.0	5.0
10	119333.0	satisfied	Male	0.9999993422750989	5.0	5.0	5.0	4.0	5.0	5.0
11	119995.0	satisfied	Male	0.9999993356718396	5.0	5.0	5.0	4.0	5.0	5.0
12	126794.0	satisfied	Female	0.9999992824779013	5.0	5.0	5.0	4.0	5.0	5.0
13	119416.0	satisfied	Female	0.9999992819621671	5.0	5.0	5.0	5.0	5.0	5.0
14	123908.0	satisfied	Female	0.9999992665642853	5.0	4.0	3.0	5.0	5.0	5.0
15	120484.0	satisfied	Female	0.9999991800627741	5.0	5.0	5.0	4.0	5.0	5.0
16	124121.0	satisfied	Female	0.9999991577973075	5.0	5.0	5.0	5.0	5.0	5.0
17	126862.0	satisfied	Female	0.9999991475059402	5.0	5.0	5.0	5.0	5.0	5.0
18	119887.0	satisfied	Male	0.9999991258067407	5.0	5.0	5.0	5.0	5.0	5.0
19	114617.0	satisfied	Male	0.9999991021360736	5.0	5.0	5.0	4.0	5.0	5.0
20	113972.0	satisfied	Female	0.9999990973922206	5.0	5.0	5.0	4.0	5.0	5.0
21	128988.0	satisfied	Male	0.9999990644463155	5.0	5.0	5.0	4.0	5.0	5.0
22	124511.0	satisfied	Male	0.9999990590120524	5.0	5.0	5.0	5.0	5.0	4.0
23	124708.0	satisfied	Male	0.9999989960349497	5.0	5.0	5.0	5.0	5.0	5.0
24	113862.0	satisfied	Male	0.9999989794773779	5.0	4.0	4.0	5.0	5.0	5.0
25	129486.0	satisfied	Male	0.9999989266370328	5.0	5.0	5.0	4.0	5.0	5.0
26	114083.0	satisfied	Male	0.9999988623922624	5.0	5.0	5.0	5.0	5.0	5.0
27	128852.0	satisfied	Female	0.9999988282663019	5.0	5.0	5.0	5.0	5.0	5.0
28	128970.0	satisfied	Female	0.9999988101551638	5.0	5.0	5.0	5.0	5.0	5.0
29	123809.0	satisfied	Male	0.9999987722524947	5.0	5.0	5.0	4.0	5.0	5.0
30	110373.0	satisfied	Female	0.9999987569802675	5.0	5.0	5.0	5.0	5.0	5.0
31	127617.0	satisfied	Male	0.9999987545173048	5.0	5.0	5.0	4.0	5.0	5.0
32	116463.0	satisfied	Female	0.9999987434133076	5.0	5.0	5.0	4.0	5.0	5.0
33	124649.0	satisfied	Female	0.9999987201483655	5.0	5.0	5.0	5.0	5.0	5.0

8.4. Limitation

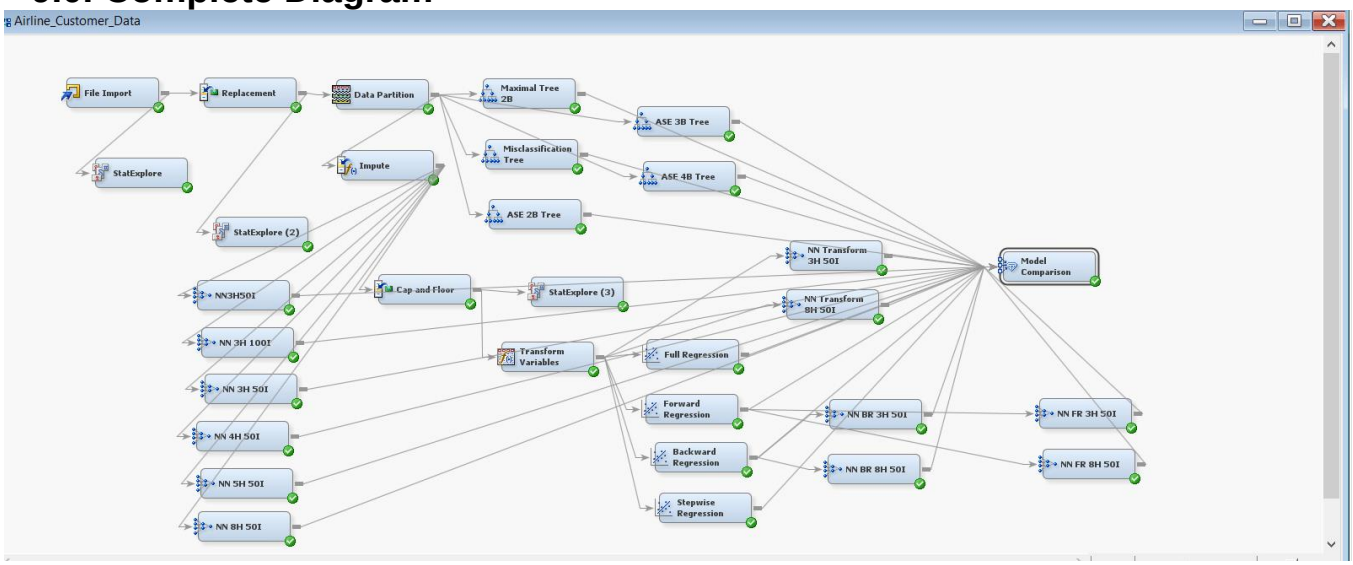
We can see that 12 selected input variable (Inflight Wi-Fi Service, Ease of Online booking, Food and Drink, Online Boarding, Seat Comfort, Inflight Entertainment, On-board Service, Leg Room Service, Baggage Handling, Check-in Service, Inflight Service, Cleanliness) are correlated with the target variable satisfaction. Since the target variable is measured on the concept satisfaction and certain input variable are also measured on the concept satisfaction it becomes difficult for us to compartmentalize the impact on each input on the target. Such leaks in the model are one of the major limitations of the study.

8.5. Recommendation

Based on the analysis, the ASE 4B Tree is the model which best fits the dataset. We recommend the following:

1. The passengers are most satisfied if they have good Inflight Entertainment Service. Hence most focus should be given to keeping the customers entertained and this service should never deteriorate to maintain the loyalty of customer.
2. Business class customers are more satisfied than the Eco Plus customers, new promotions, deals, and service needs to be given to ensure the satisfaction of eco plus customer.
3. Loyal Customers remain loyal to the airline even with decline in service.
4. Female passengers are seen to be more satisfied than male customer. New offers to accommodate female customers can help us turn the majority female customers into loyal customers.
5. Improving on the Seat Comfort, Ease of Online Booking, Online Support and Leg Room can also help the airline tremendously in converting their disloyal customers to loyal customers.

9.0. Complete Diagram



10.0. References

- *Center. Forward Regression | Center Based Statistics. (n.d.). Retrieved December 16, 2021, from https://center-based-statistics.com/html/forwardReg.html*
- *Data Exploration - a complete introduction. HEAVY.AI. (n.d.). Retrieved July 15, 2022, from https://www.heavy.ai/learn/data-exploration*
- *Jana, S. (2020, March 19). Airlines customer satisfaction. Kaggle. Retrieved July 15, 2022, from https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction*
- *Multicollinearity. Corporate Finance Institute. (2022, January 20). Retrieved July 15, 2022, from https://corporatefinanceinstitute.com/resources/knowledge/other/multicollinearity/*
- *SAS enterprise miner. SAS Enterprise Miner | SAS Support. (n.d.). Retrieved July 15, 2022, from https://support.sas.com/en/software/enterprise-miner-support.html#get-started*