

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

Predictive Analysis

Hinna Mary Steephen (301225437)

Table of Contents

Executive Summary.....	3
1.0. Introduction	4
2.0. Tools and Models used	4
3.0. Data Extraction.....	5
4.0. Modeling Steps	7
4.1. Create a Data Diagram	7
4.2. Data Source	7
4.3. Data Exploration	8
4.3.1 First level data Exporation	8
4.3.2. Second level data Exporation.....	10
4.4. Modifying and correcting Data	12
4.4.1. Setting zero as missing value	12
4.5. Data Partitioning	14
5.0. Modelling	15
5.1. Decision Tree.....	15
5.1.1. Maximal Tree	15
5.1.2. Misclassification Tree.....	17
5.1.3. Average square Error Tree	18
5.1.4. Decision Tree Summary	19
5.2. Regression	20
5.2.1. Imputation	20
5.2.2. Transformation or managing skewness and outlier variables	21
5.2.3. Regression Models.....	23
5.2.4. Regression Summary.....	29
6.0. Neural Network.....	31
6.1. Neural Network Models:.....	31
6.1.1. Impute neural Network.....	31
.....	34
7.0. Model Comparison.....	35
8.0. Results.....	36
8.1. Fit Statistics analysis of all the models.....	36

8.2. Ranking based ROC index and Gini Coefficient.....	36
8.3. Final Outcome:.....	37
8.5. Recommendation.....	37
9.0. Complete Diagram	38
10.0. Reference	39

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 dataset contains a column or feature called 'satisfaction,' which has two values: 'Dissatisfied' and 'satisfied,' describing the customer's overall satisfaction level.

Outcome

Based on this analysis, the neural network using variable selected by backward regression with 8 hidden units and 50 iteration is the best model.

1.0. 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 programmes 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 on-board 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.

2.0. 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

3.0. 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 129880 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 service	Interval	Satisfaction level of the inflight service	0-Not Applicable;1-5
8	Departure/Arrival time convenient	Interval	Satisfaction level of Departure/Arrival time 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, we need to create a process flow diagram

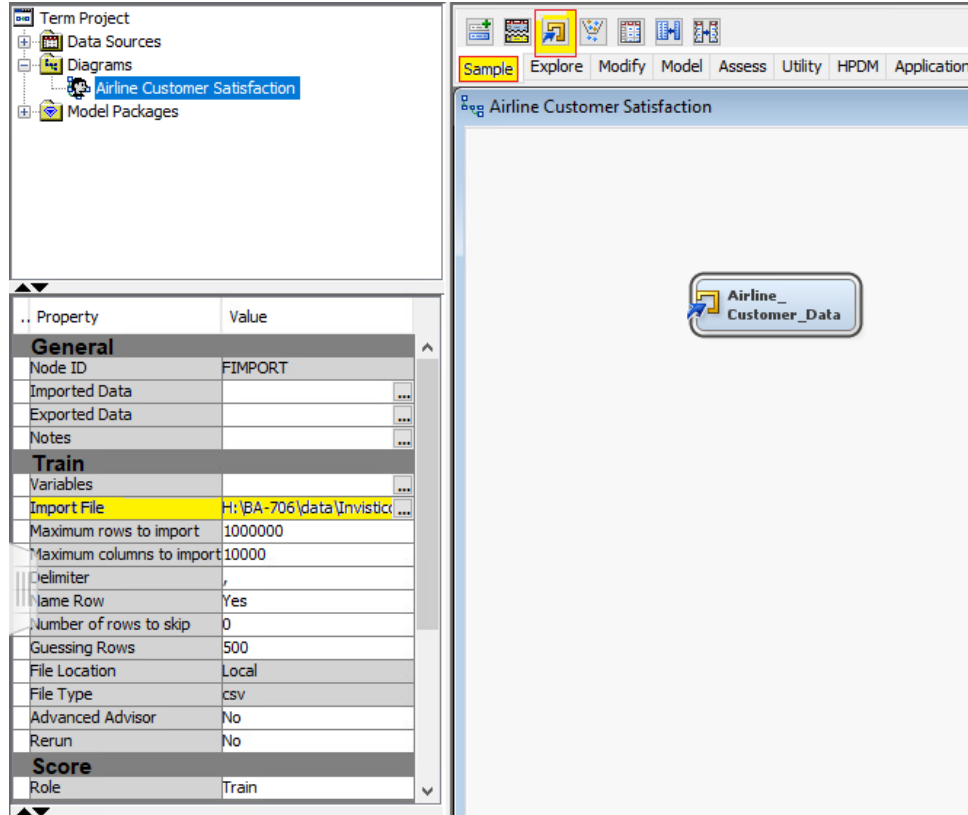
Steps:

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 utilise data visualisation and statistical tools to characterise dataset descriptions 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 categorised 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_convenience	Departure/Arrival time convenience	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 has 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 variable, as it's the dependent variable that relies on all the other input variable.

- Reject the following variable:
 - Gate_location : This variable gives **irrelevant** data as the gate location is predetermined for different airline and the type of journey and cannot be altered to improve customer satisfaction
 - flight distance: Flight distance is also an **irrelevant** data for this study as the flight distance is fixed which cannot be improved by any means
 - departure_arrival_time_convenience: Its **irrelevant** data as well since information about the flight time is available to the customer and the pricing is determined based on this factor as well.
 - Type of travel: its redundant data for this study as Class variable gives the same information as well.

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

Name	Role	Level	Report	Order	Drop	Lower Limit	Up
Inflight_entertainment	Input	Interval	No		No	.	
Gender	Input	Nominal	No		No	.	
Inflight_wifi_service	Input	Interval	No		No	.	
Ease_of_Online_booking	Input	Interval	No		No	.	
Food_and_drink	Input	Interval	No		No	.	
Online_support	Input	Interval	No		No	.	
Online_boarding	Input	Interval	No		No	.	
Seat_comfort	Input	Interval	No		No	.	
Leg_room_service	Input	Interval	No		No	.	
On_board_service	Input	Interval	No		No	.	
Class	Input	Nominal	No		No	.	
Checkin_service	Input	Interval	No		No	.	
Cleanliness	Input	Interval	No		No	.	
Arrival_Delay_in_Minutes	Input	Interval	No		No	.	
Age	Input	Interval	No		No	.	
Baggage_handling	Input	Interval	No		No	.	
Departure_Delay_in_Minutes	Input	Interval	No		No	.	
Customer_Type	Input	Nominal	No		No	.	
Type_of_Travel	Rejected	Nominal	No		No	.	
Flight_Distance	Rejected	Interval	No		No	.	
Gate_location	Rejected	Interval	No		No	.	
Departure_Arrival_time_convenience	Rejected	Interval	No		No	.	
satisfaction	Target	Nominal	No		No	.	

Variable summary:

Variable Summary		
Role	Measurement Level	Frequency Count
INPUT	INTERVAL	15
INPUT	NOMINAL	3
REJECTED	INTERVAL	3
REJECTED	NOMINAL	1
TARGET	NOMINAL	1

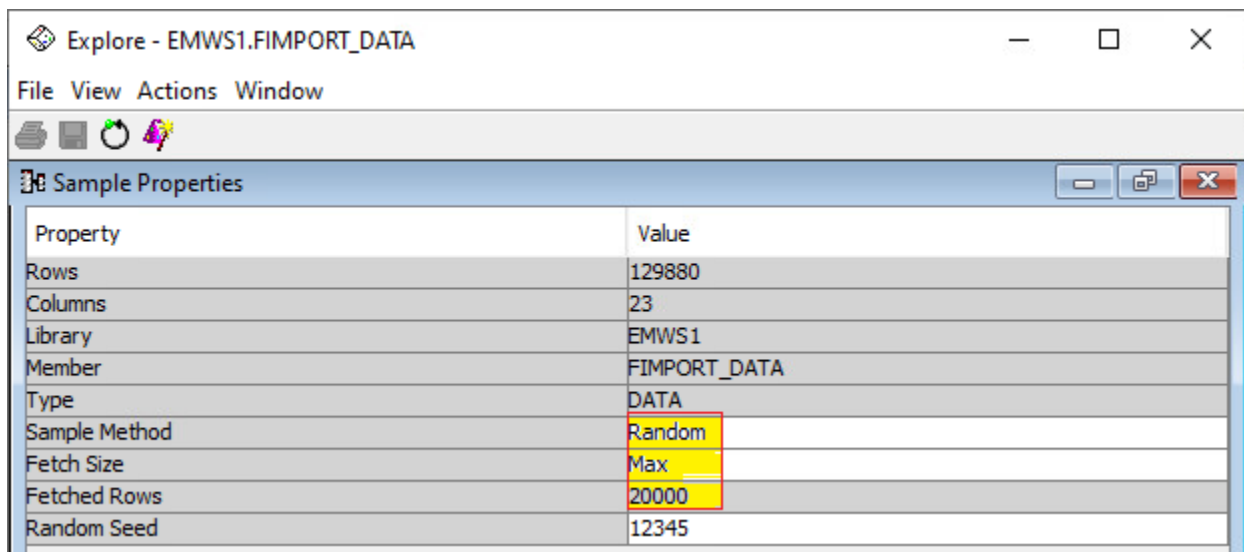
4.3.2. Second level data Exporation

To further analyse the data lets explore all the input variable.

Steps:

1. right click the Airline_Customer_Data node, press CTRL+A and select all the variable and 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 20000 results randomly for analysis.



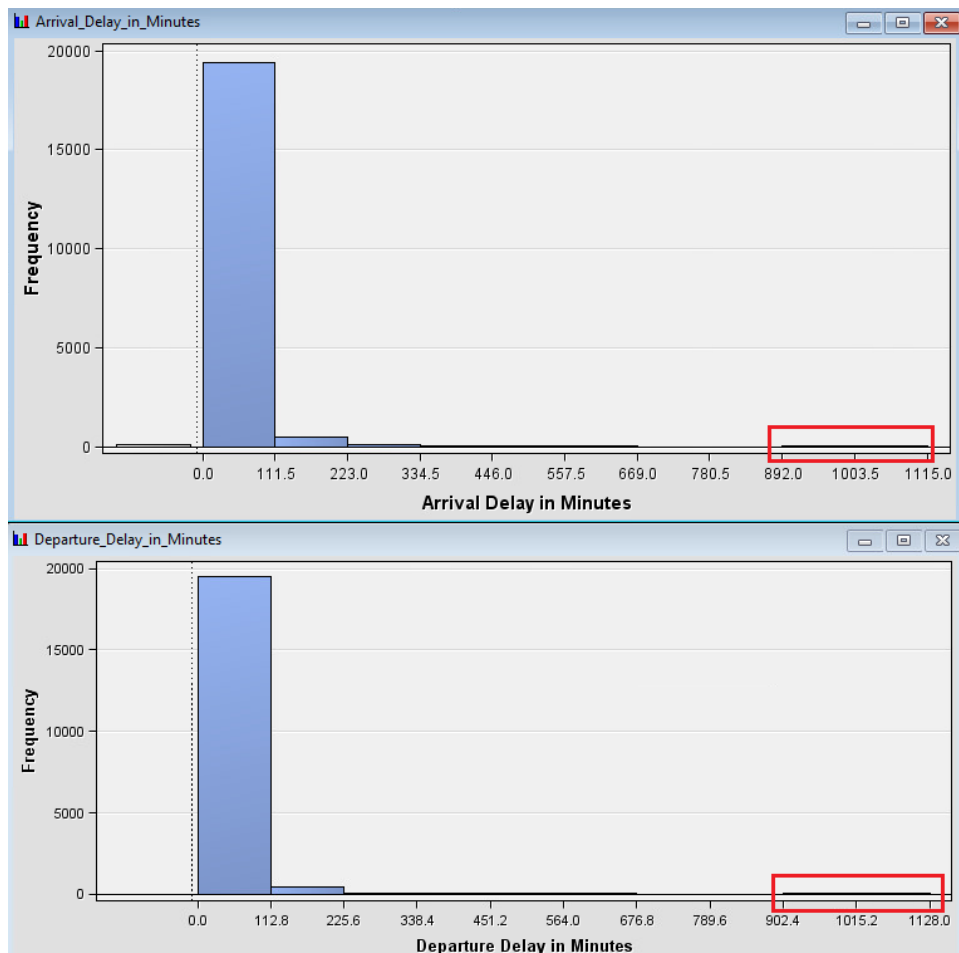
Sample statistics:

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_convenient	Departure/Arrival time convenient	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

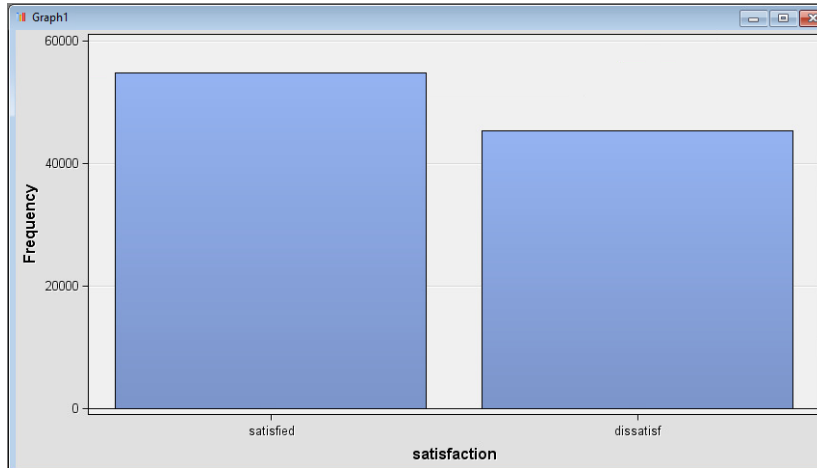
Let's study the sample statistic further:

Observation:

- There are 0.33 missing data in arrival delay in minutes variable
- For the following variable the minimum value is 0 and maximum is 5, but these variable gives us a scale of satisfaction rate of customer from 1 to 5 so the 0 value needs to be considered as missing value:
 - Cleanliness
 - Departure_arival_time_convinience(rejected)
 - Ease_of_online_booking
 - Food_and_drink
 - Gate_Location(rejected)
 - Inflight_entertainment
 - Inflight_wifi_service
 - Leg_room_service
 - On_board_service
 - Online_boarding
 - Seat_comfort
- Looking at the mean value of the Arrival_delay_in_minutes and Departure_delay_in_minutes, its evident that the data is skewed to the right (also analysed using the following histogram graphs)



Output Variable



This graph shows that our dataset has 54761 satisfied customers of 100000, i.e. 54.76%

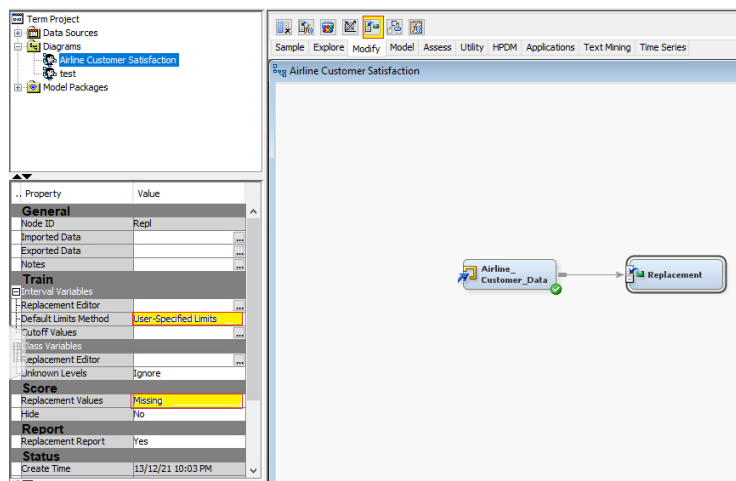
4.4. Modifying and correcting Data

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



- Inside replacement Editor, set the lower limits to 1 of the variables identified having missing value.

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

- Click ok and run the node

Sample statistics

From the sample statistics, its clear that Sas miner is now considering the variables for which we set lower limit has 1. If they have zero value those are considered to be missing value and the replaced variables are stored with a prefix REP as seen.

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_wif...	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 Dist...	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 practise in predictive modelling for evaluating model performance. We will divide the data into two parts for this project: training data (50 %) and validation data (50 %) . 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. In other words, 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 50
 - b. Set validation data set allocation to 50
 - c. Set test data to 0
4. Run the node

The screenshot displays the Orange3 software interface. On the left, the 'Properties' panel for the 'Data Partition' node is visible, showing the 'Train' tab with the following settings:

Property	Value
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	13/12/21 11:42 PM
Run ID	

On the right, the 'Diagram' workspace shows a workflow with three nodes: 'Airline_Customer_Data', 'Replacement', and 'Data Partition'. The 'Airline_Customer_Data' node is connected to the 'Replacement' node, which is then connected to the 'Data Partition' node. All nodes have a green checkmark indicating they are ready to run.

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Repl_TRAIN	129880
TRAIN	EMWS1.Part_TRAIN	64940
VALIDATE	EMWS1.Part_VALIDATE	64940

5.0. Modelling

5.1. Decision Tree

As the name implies, 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 logworth value determines the quality of the split in a decision tree. The threshold logworth 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.

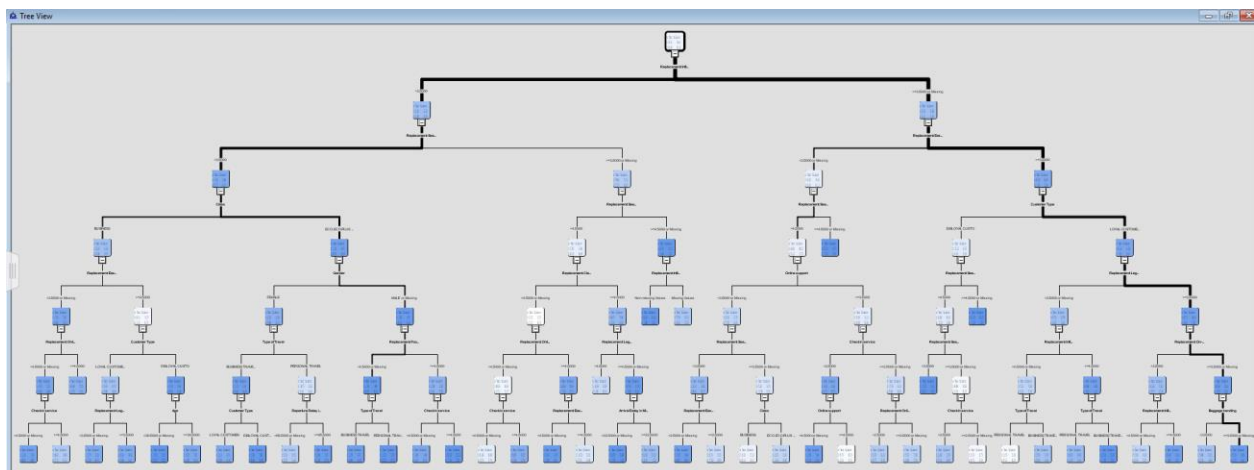
5.1.1. Maximal Tree

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 logworth drops below the threshold value.

steps

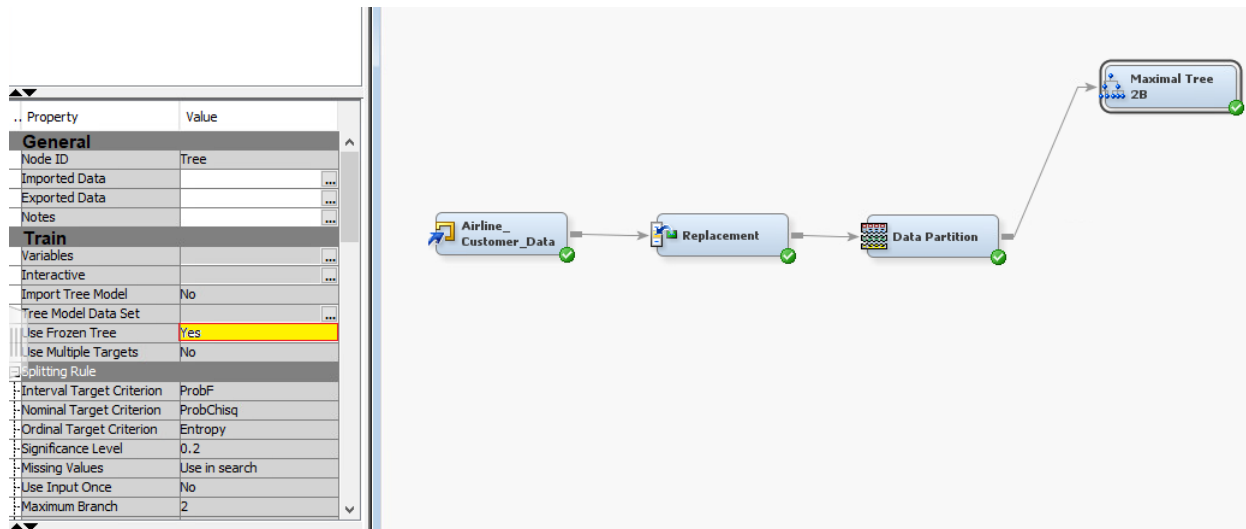
1. Select the **Model** tab on the Toolbar and drag the **Decision Tree** node into the Diagram workspace.
2. Rename the node to **Maximal Tree 2B**
3. Connect the **Data Partition** node to the **Decision Tree** node
4. On the properties panel, click on the interactive ellipse
5. Right click on the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

Maximal tree will get created



6. Close the interactive window

7. Change Use frozen tree from no to yes in properties panel, to restrict change to the maximal tree because of other changes
8. Run the maximal tree and record the average square error value



Observation:

No of leaves : 43

Average square error of validation data – 0.075896

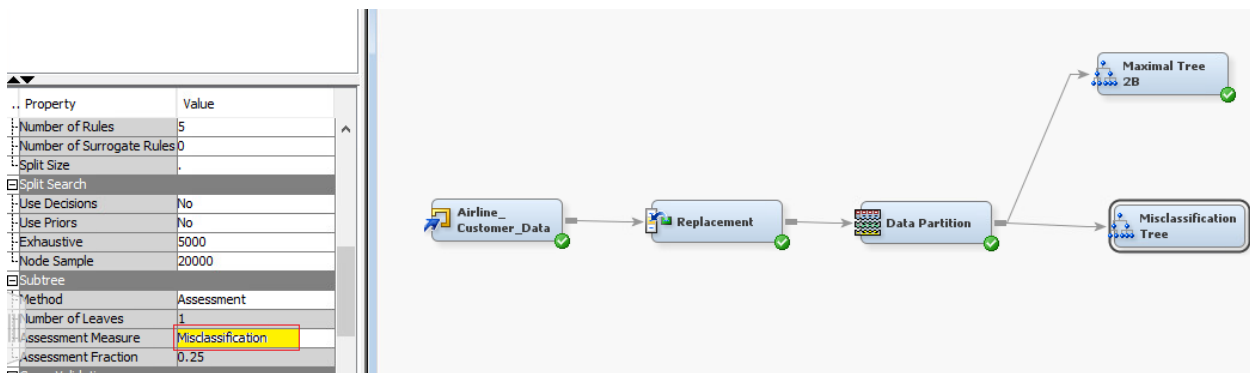
Fit statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		_NOBS_	Sum of Frequencies	64940	64940	.
satisfaction		_MISC_	Misclassification Rate	0.099307	0.101571	.
satisfaction		_MAX_	Maximum Absolute Err...	0.993389	1	.
satisfaction		_SSE_	Sum of Squared Errors	9582.507	9857.338	.
satisfaction		_ASE_	Average Squared Error	0.07378	0.075896	.
satisfaction		_RASE_	Root Average Squared...	0.271624	0.275492	.
satisfaction		_DIV_	Divisor for ASE	129880	129880	.
satisfaction		_DFT_	Total Degrees of Free...	64940	.	.

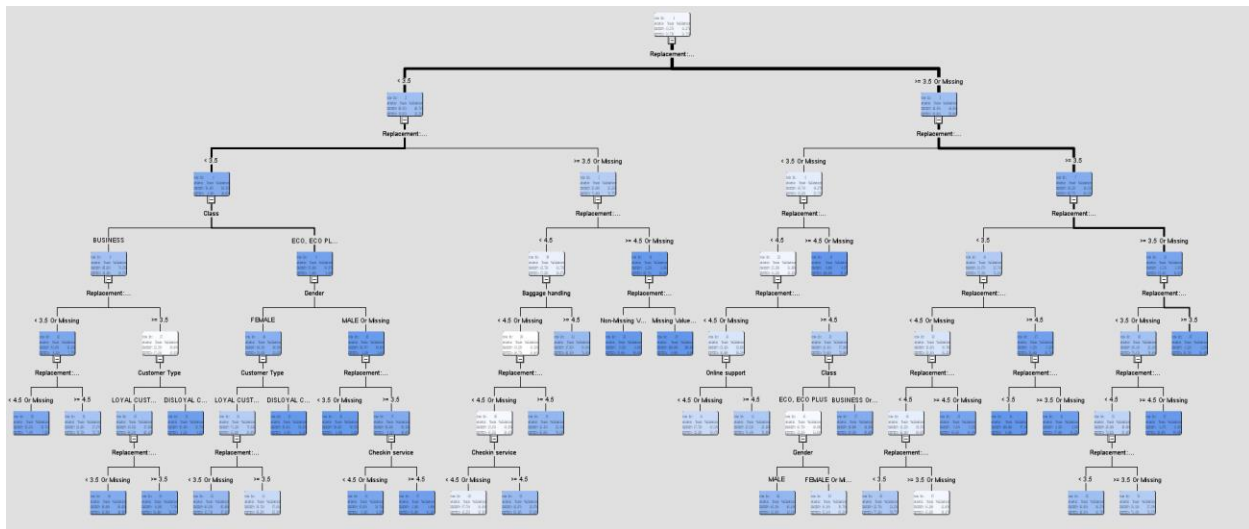
5.1.2. Misclassification Tree

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

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		_NOBS_	Sum of Frequencies	64940	64940	.
satisfaction		_MISC_	Misclassification Rate	0.103018	0.107977	.
satisfaction		_MAX_	Maximum Absolute Err...	0.993389	1	.
satisfaction		_SSE_	Sum of Squared Errors	9938.123	10237.58	.
satisfaction		_ASE_	Average Squared Error	0.076518	0.078823	.
satisfaction		_RASE_	Root Average Squared...	0.276618	0.280755	.
satisfaction		_DIV_	Divisor for ASE	129880	129880	.
satisfaction		_DFT_	Total Degrees of Free...	64940	.	.

Observation:

No of leaves - 32

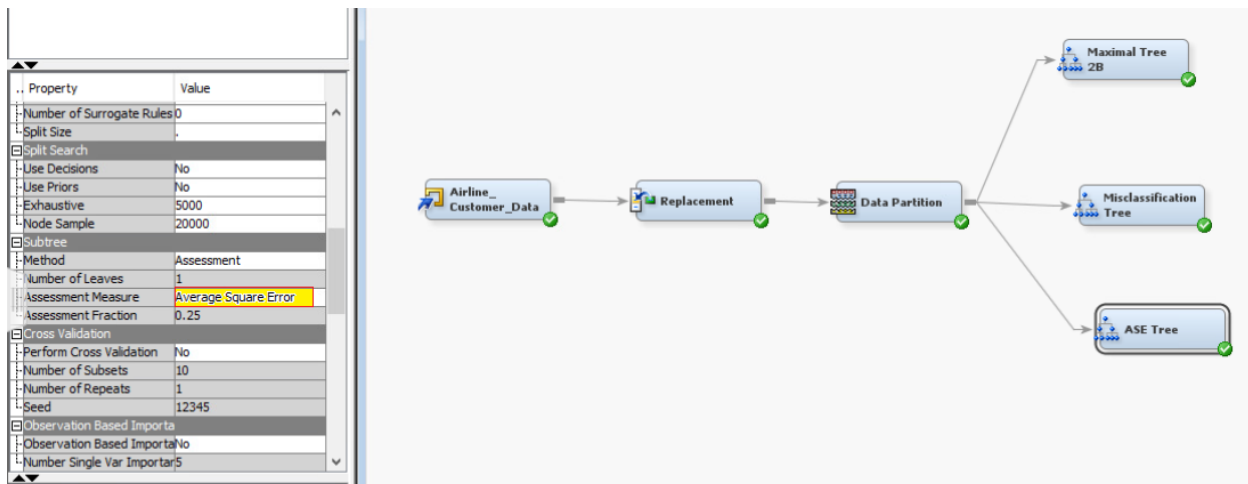
Average square error – 0.078823

5.1.3. Average square Error Tree

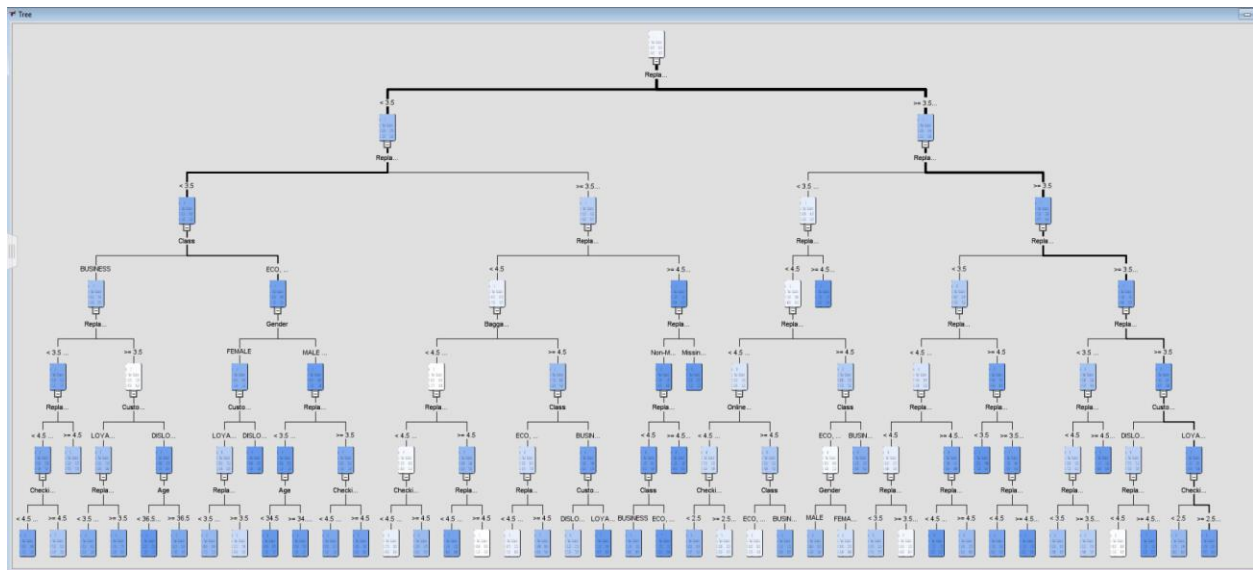
2 Branch

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
satisfaction		_NOBS_	Sum of Frequencies	64940	64940
satisfaction		_MISC_	Misclassification Rate	0.103018	0.107977
satisfaction		_MAX_	Maximum Absolute Err...	0.999072	1
satisfaction		_SSE_	Sum of Squared Errors	9528.374	9869.607
satisfaction		_ASE_	Average Squared Error	0.073363	0.07599
satisfaction		_RASE_	Root Average Squared...	0.270856	0.275663
satisfaction		_DIV_	Divisor for ASE	129880	129880
satisfaction		_DFT_	Total Degrees of Free...	64940	.

Observation:

No of leaves - 32

Average Square error – 0.073363

5.1.4. Decision Tree Summary

Maximal Tree has the least Average Square error at 0.075896 of the 3 decision tree models. Looking into the maximal tree <3 model we can see:

- Customers has the highest satisfaction rate if the following services are good:
 - Inflight entertainment
 - Ease of online booking
 - On board service

- On the other hand the satisfaction level decreases considerably because the following service were handled poorly:
 - Seat comfort
 - Onboard food and drinks
 - Baggage handling

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 utilize 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.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 \cdot X_1 + \hat{w}_2 \cdot X_2 \quad \text{logit scores}$$

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.

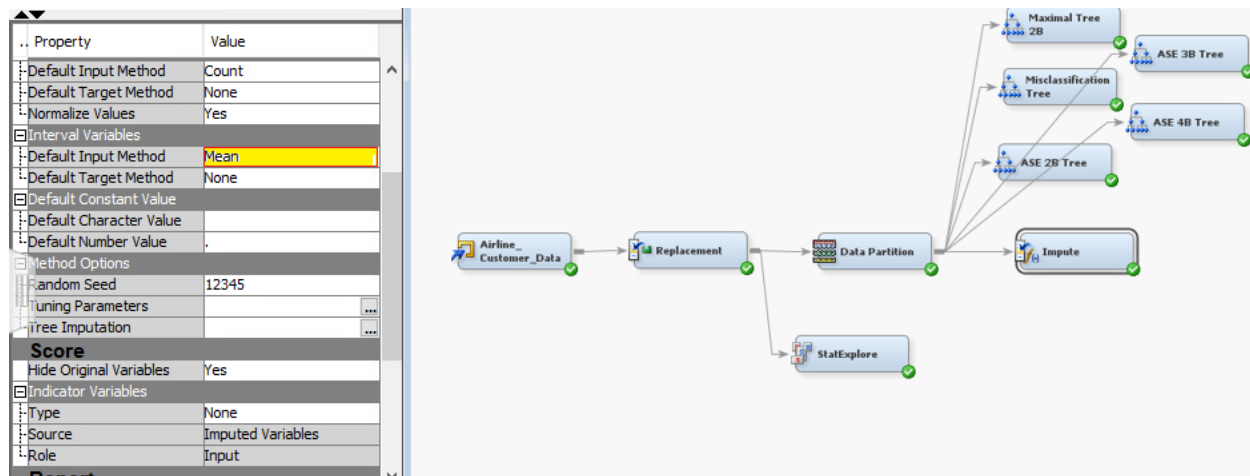
1. In regression, complete-case analysis is the default strategy for dealing with missing values. As a result, only full cases are considered in this method's analysis.
2. Cases with missing values are not scored by the prediction method.

Remedy for the missing values

1. Synthetic distribution: In this situation, the mean, mode, or median values are calculated and utilised 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

Variable Name	Impute Method	Imputed Variable	Impute Value ▼	Role	Number of Missing for TRAIN
Arrival_Delay_in_Minutes	MEAN	IMP_Arrival_Delay_in_Minutes	14.79266	INPUT	205
REP_Cleanliness	MEAN	IMP_REP_Cleanliness	3.704164	INPUT	2
REP_Leg_room_service	MEAN	IMP_REP_Leg_room_service	3.497666	INPUT	234
REP_Ease_of_Online_booking	MEAN	IMP_REP_Ease_of_Online_bo...	3.470884	INPUT	9
REP_Inflight_entertainment	MEAN	IMP_REP_Inflight_entertainment	3.46547	INPUT	1445
REP_On_board_service	MEAN	IMP_REP_On_board_service	3.463627	INPUT	2
REP_Online_boarding	MEAN	IMP_REP_Online_boarding	3.350808	INPUT	7
REP_Inflight_wifi_service	MEAN	IMP_REP_Inflight_wifi_service	3.253626	INPUT	61
REP_Food_and_drink	MEAN	IMP_REP_Food_and_drink	2.996855	INPUT	2938
REP_Seat_comfort	MEAN	IMP_REP_Seat_comfort	2.950284	INPUT	2364

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

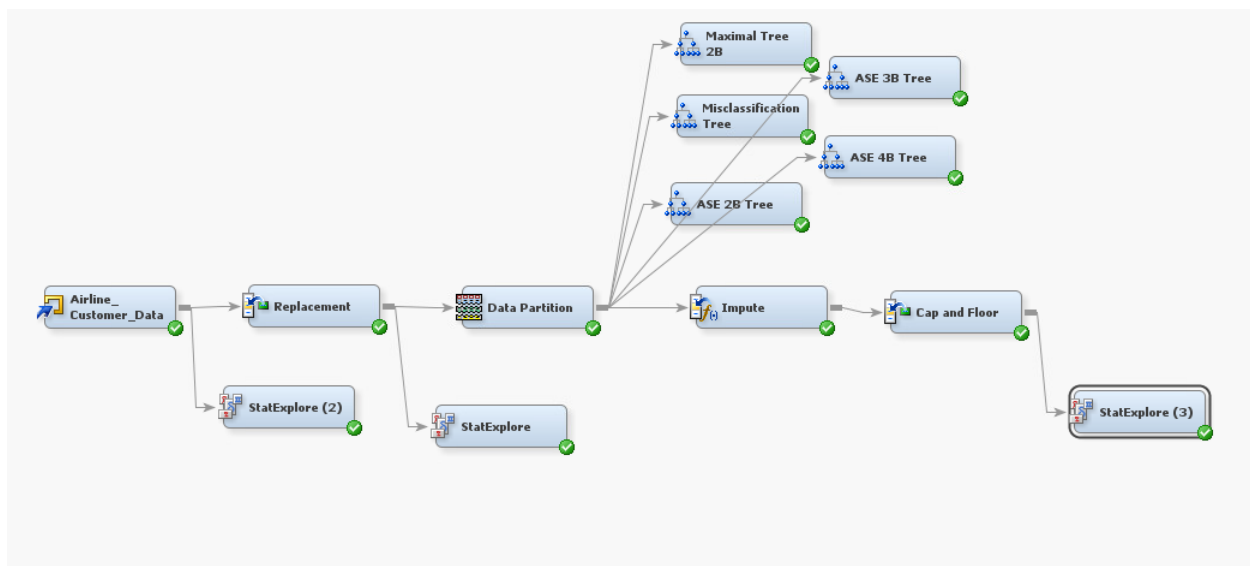
Data Role	Target	Target Level	Variable	Skewness ▼	Mean
TRAIN	satisfaction	satisfied	Departure_Delay_in_Minutes	7.184991	12.15477
TRAIN	satisfaction	satisfied	Arrival_Delay_in_Minutes	6.996799	12.26888
TRAIN	satisfaction	dissatisf	Departure_Delay_in_Minutes	6.607497	17.80775
TRAIN	satisfaction	dissatisf	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

5.2.2.1. 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 StatExplorer and check whether it improves the skewness



Stat Explorer Result after Cap and floor

Data Role	Target	Target Level	Variable	Skewness	Mean	Median	Missing
	Data Role	Level					
TRAIN	satisfaction	satisfied	REP_Departure_Delay_in_Minutes	3.022354	10.9921	0	0
TRAIN	satisfaction	satisfied	REP_IMP_Arrival_Delay_in_Minutes	3.014425	11.09059	0	0
TRAIN	satisfaction	dissatisf	REP_Departure_Delay_in_Minutes	2.748417	15.10523	0	0
TRAIN	satisfaction	dissatisf	REP_IMP_Arrival_Delay_in_Minutes	2.734282	15.77006	0	0
TRAIN	satisfaction	dissatisf	REP_IMP_REP_Seat_comfort	0.024848	2.470176	2	0
TRAIN	satisfaction	dissatisf	REP_IMP_REP_Food_and_drink	-0.00515	2.727625	3	0
TRAIN	satisfaction	dissatisf	REP_Age	-0.00652	37.59017	36	0
TRAIN	satisfaction	dissatisf	REP_IMP_REP_Inflight_wifi_servic	-0.18634	2.922105	3	0
TRAIN	satisfaction	satisfied	REP_IMP_REP_Food_and_drink	-0.21499	3.21953	3	0
TRAIN	satisfaction	satisfied	REP_Age	-0.23726	41.02321	42	0

Observation

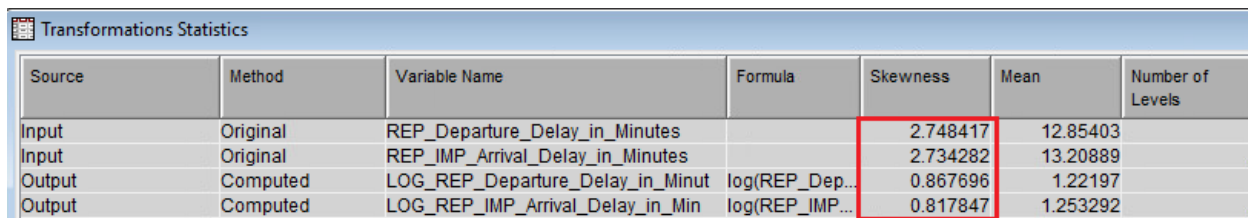
1. We can see that cap and floor has improved the skewness when we compare it to the original data.
2. Its still positively skewed, so we need to further transform the data to handle the skewness, this can be done by log transformation

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



Source	Method	Variable Name	Formula	Skewness	Mean	Number of Levels
Input	Original	REP_Departure_Delay_in_Minutes		2.748417	12.85403	
Input	Original	REP_IMP_Arrival_Delay_in_Minutes		2.734282	13.20889	
Output	Computed	LOG_REP_Departure_Delay_in_Minut	log(REP_Dep...	0.867696	1.22197	
Output	Computed	LOG_REP_IMP_Arrival_Delay_in_Min	log(REP_IMP...	0.817847	1.253292	

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

5.2.3. Regression Models

Four types of regression models are executed in this project

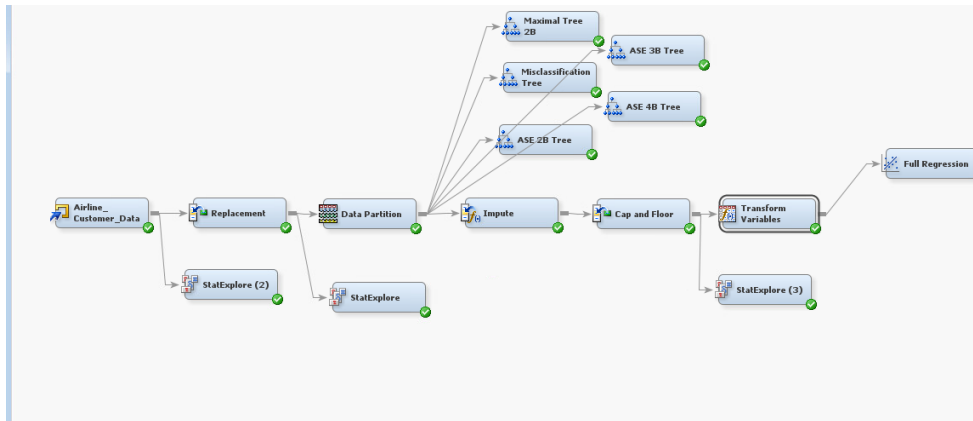
1. Full Regression
2. Forward Regression
3. Backward Regression
4. Stepwise Regression

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

Target	Target Label	Fit Statistics	Statistics Label	Train ▲	Validation	Test
satisfaction		_ASE_	Average Squared Error	0.108645	0.110523	
satisfaction		_MSE_	Mean Square Error	0.108679	0.110523	
satisfaction		_FPE_	Final Prediction Error	0.108712		
satisfaction		_MISC_	Misclassification Rate	0.152402	0.156036	
satisfaction		_RASE_	Root Average Sum of ...	0.329614	0.33245	
satisfaction		_RMSE_	Root Mean Squared E...	0.329664	0.33245	
satisfaction		_RFPE_	Root Final Prediction ...	0.329715		

Odds Ratio

Odds Ratio Estimates				
Effect			satisfaction	Point Estimate
Class	Business vs Eco Plus	satisfied		3.685
Class	Eco vs Eco Plus	satisfied		0.976
Customer_Type	Loyal Customer vs disloyal Custo	satisfied		5.165
Gender	Female vs Male	satisfied		2.776
LOG_REP_Departure_Delay_in_Minut		satisfied		0.997
LOG_REP_IMP_Arrival_Delay_in_Min		satisfied		0.863
REP_Age		satisfied		0.997
REP_Baggage_handling		satisfied		1.062
REP_Checkin_service		satisfied		1.279
REP_IMP_REP_Cleanliness		satisfied		1.015
REP_IMP_REP_Ease_of_Online_booki		satisfied		1.406
REP_IMP_REP_Food_and_drink		satisfied		0.837
REP_IMP_REP_Inflight_entertainme		satisfied		2.546
REP_IMP_REP_Inflight_wifi_servic		satisfied		0.905
REP_IMP_REP_Leg_room_service		satisfied		1.273
REP_IMP_REP_On_board_service		satisfied		1.339
REP_IMP_REP_Online_boarding		satisfied		1.145
REP_IMP_REP_Seat_comfort		satisfied		1.775
REP_Online_support		satisfied		1.034

Observation:

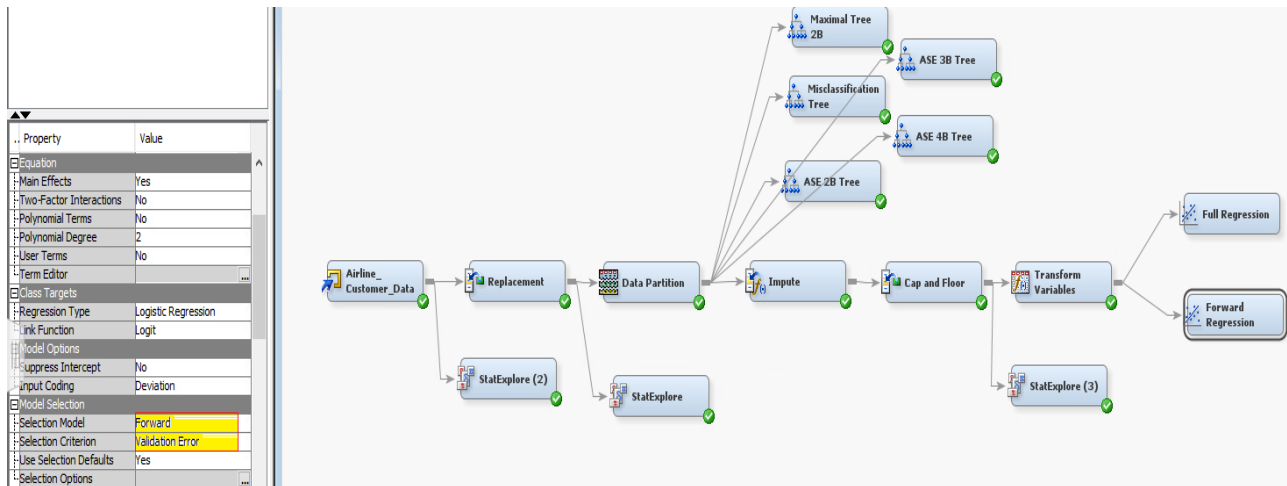
1. Average Square Error – 0.110523
2. Misclassification Error - 0.156036

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

Target	Target Label	Fit Statistics	Statistics Label	Train ▲	Validation	Test
satisfaction		_ASE_	Average Squared Error	0.108652	0.11053	.
satisfaction		_MSE_	Mean Square Error	0.108682	0.11053	.
satisfaction		_FPE_	Final Prediction Error	0.108712	.	.
satisfaction		_MISC_	Misclassification Rate	0.152572	0.15579	.
satisfaction		_RASE_	Root Average Sum of ...	0.329623	0.332461	.
satisfaction		_RMSE_	Root Mean Squared E...	0.329669	0.332461	.
satisfaction		_RFPE_	Root Final Prediction ...	0.329715	.	.

Observation:

1. Average Square Error – 0.11053
2. Misclassification Error - 0.15579

Odds Ratio

Odds Ratio Estimates

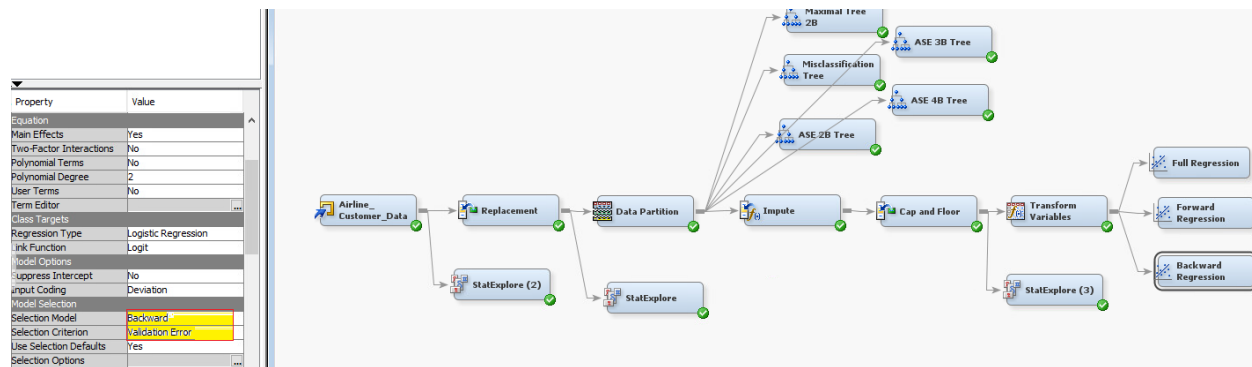
Effect		satisfaction	Point Estimate
Class	Business vs Eco Plus	satisfied	3.687
Class	Eco vs Eco Plus	satisfied	0.976
Customer_Type	Loyal Customer vs disloyal Custo	satisfied	5.156
Gender	Female vs Male	satisfied	2.774
LOG_REP_IMP_Arrival_Delay_in_Min		satisfied	0.861
REP_Age		satisfied	0.997
REP_Baggage_handling		satisfied	1.067
REP_Checkin_service		satisfied	1.281
REP_IMP_REP_Ease_of_Online_booki		satisfied	1.412
REP_IMP_REP_Food_and_drink		satisfied	0.837
REP_IMP_REP_Inflight_entertainme		satisfied	2.546
REP_IMP_REP_Inflight_wifi_servic		satisfied	0.904
REP_IMP_REP_Leg_room_service		satisfied	1.274
REP_IMP_REP_On_board_service		satisfied	1.343
REP_IMP_REP_Online_boarding		satisfied	1.144
REP_IMP_REP_Seat_comfort		satisfied	1.775
REP_Online_support		satisfied	1.033

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

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		_ASE_	Average Squared Error	0.108645	0.110521	
satisfaction		_MSE_	Mean Square Error	0.108677	0.110521	
satisfaction		_FPE_	Final Prediction Error	0.108709		
satisfaction		_MISC_	Misclassification Rate	0.152387	0.156052	
satisfaction		_RASE_	Root Average Sum of ...	0.329614	0.332447	
satisfaction		_RMSE_	Root Mean Squared E...	0.329662	0.332447	
satisfaction		_RFPE_	Root Final Prediction ...	0.329711		

Odds Ratio

Odds Ratio Estimates

Effect		satisfaction	Point Estimate
Class	Business vs Eco Plus	satisfied	3.684
Class	Eco vs Eco Plus	satisfied	0.976
Customer_Type	Loyal Customer vs disloyal Custo	satisfied	5.165
Gender	Female vs Male	satisfied	2.776
LOG_REP_IMP_Arrival_Delay_in_Min		satisfied	0.861
REP_Age		satisfied	0.997
REP_Baggage_handling		satisfied	1.062
REP_Checkin_service		satisfied	1.279
REP_IMP_REP_Cleanliness		satisfied	1.015
REP_IMP_REP_Ease_of_Online_booki		satisfied	1.406
REP_IMP_REP_Food_and_drink		satisfied	0.837
REP_IMP_REP_Inflight_entertainme		satisfied	2.546
REP_IMP_REP_Inflight_wifi_servic		satisfied	0.905
REP_IMP_REP_Leg_room_service		satisfied	1.273
REP_IMP_REP_On_board_service		satisfied	1.339
REP_IMP_REP_Online_boarding		satisfied	1.145
REP_IMP_REP_Seat_comfort		satisfied	1.775
REP_Online_support		satisfied	1.034

Observation:

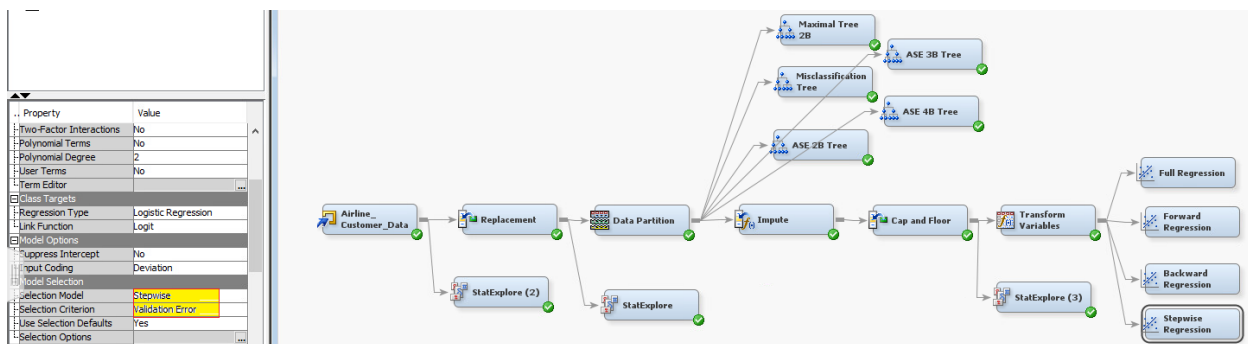
1. Average Square Error – 0.110521
2. Misclassification Error - 0.156052

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

Target	Target Label	Fit Statistics	Statistics Label	Train ▲	Validation	Test
satisfaction		_ASE_	Average Squared Error	0.108652	0.11053	.
satisfaction		_MSE_	Mean Square Error	0.108682	0.11053	.
satisfaction		_FPE_	Final Prediction Error	0.108712	.	.
satisfaction		_MISC_	Misclassification Rate	0.152572	0.15579	.
satisfaction		_RASE_	Root Average Sum of ...	0.329623	0.332461	.
satisfaction		_RMSE_	Root Mean Squared E...	0.329669	0.332461	.
satisfaction		_RFPE_	Root Final Prediction ...	0.329715	.	.

Observation:

1. Average Square Error – 0.11053
2. Misclassification Error - 0.15579

Odds Ratio

Odds Ratio Estimates			
Effect		satisfaction	Point Estimate
Class	Business vs Eco Plus	satisfied	3.687
Class	Eco vs Eco Plus	satisfied	0.976
Customer_Type	Loyal Customer vs disloyal Custo	satisfied	5.156
Gender	Female vs Male	satisfied	2.774
LOG_REP_IMP_Arrival_Delay_in_Min		satisfied	0.861
REP_Age		satisfied	0.997
REP_Baggage_handling		satisfied	1.067
REP_Checkin_service		satisfied	1.281
REP_IMP_REP_Ease_of_Online_booki		satisfied	1.412
REP_IMP_REP_Food_and_drink		satisfied	0.837
REP_IMP_REP_Inflight_entertainme		satisfied	2.546
REP_IMP_REP_Inflight_wifi_servic		satisfied	0.904
REP_IMP_REP_Leg_room_service		satisfied	1.274
REP_IMP_REP_On_board_service		satisfied	1.343
REP_IMP_REP_Online_boarding		satisfied	1.144
REP_IMP_REP_Seat_comfort		satisfied	1.775
REP_Online_support		satisfied	1.033

5.2.4. Regression Summary

Comparing the 4-regression model, Backward regression has the least Average square error at 0.110521. Inference drawn by interpreting the Odds ration of backward regression model as follows:

Odds ratio Interpretation

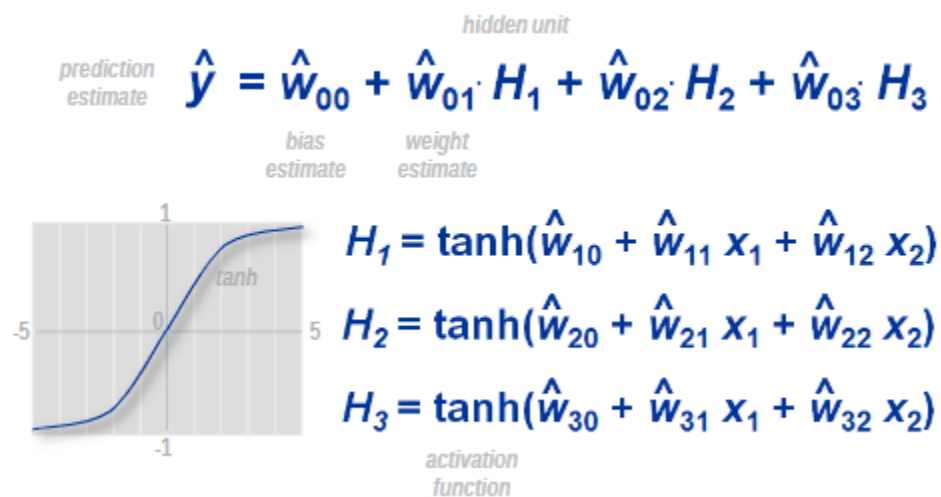
	Variable	Odds Ratio	Interpretation
1	Class(Business vs Eco plus)	3.687	The dependent variable is 3 times more likely to be predicted by Business class than cases with Eco plus
2	Class (Eco vs Eco plus)	0.976	The dependent variable is 2.4% less likely to be predicted by Economy class than cases with Eco plus
3	Customer_Type(loyal vs disloyal)	5.156	The dependent variable is 5 times more likely to be predicted by loyal customer
4	Gender(F vs M)	2.774	The dependent variable is 2 times more likely to be predicted Female than male
5	Log_REP_Departure_Delay_in_Minutes	0.861	The dependent variable is 13.9% less likely to be predicted by departure delay

6	REP_Age	0.997	The dependent variable is 0.3% less likely to be predicted by per unit change in age
7	REP_Baggage_Handling	1.067	The dependent variable is 6.7% more likely to be predicted by per unit change in satisfaction level for baggage handling
8	RESP_Checkin_Service	1.281	The dependent variable is 28.1% more likely to be predicted by per unit change in satisfaction level for checkin services
9	REP_IMP_REP_Ease_of_Online_Booking	1.412	The dependent variable is 41.2% more likely to be predicted by per unit change in satisfaction level for ease of booking
10	REP_IMP_REP_Food_and_drink	0.837	The dependent variable is 16.3% less likely to be predicted by per unit change in satisfaction level for food and drinks
11	REP_IMP_REP_Inflight_entertainment	2.546	The dependent variable is 2 times more likely to be predicted by per unit change in satisfaction level for inflight entertainment
12	REP_IMP_REP_Inflight_wifi_service	0.904	The dependent variable is 9.6% less likely to be predicted by per unit change in satisfaction level for inflight wifi service
13	REP_IMP_REP_Leg_room_service	1.274	The dependent variable is 27.4% more likely to be predicted by per unit change in satisfaction level for leg room service
14	REP_IMP_REP_On_board_service	1.343	The dependent variable is 34.3% more likely to be predicted by per unit change in satisfaction level for on board service
15	REP_IMP_REP_Online_booking	1.144	The dependent variable is 14.4% more likely to be predicted by per unit change in satisfaction level for online booking
16	REP_IMP_REP_Seat_comfort	1.745	The dependent variable is 74.5% more likely to be predicted by per unit change in satisfaction level for seat comfort
17	Rep_Online_support	1.044	The dependent variable is 4.4% more likely to be predicted by per unit change in satisfaction level for online support

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



The diagram illustrates the prediction formula for a neural network with three hidden units. At the top, the prediction estimate \hat{y} is calculated as the sum of a bias estimate \hat{w}_{00} and the weighted sum of three hidden unit outputs H_1, H_2, H_3 . Each hidden unit output H_i is the result of a tanh activation function applied to a weighted sum of input features x_1 and x_2 plus a bias term \hat{w}_{i0} . A graph of the tanh activation function is shown on the left, ranging from -1 to 1.

$$\hat{y} = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$
$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$
$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$
$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

6.1. Neural Network Models:

9 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

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

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

5. Run the nodes

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

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

5. Run the nodes

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

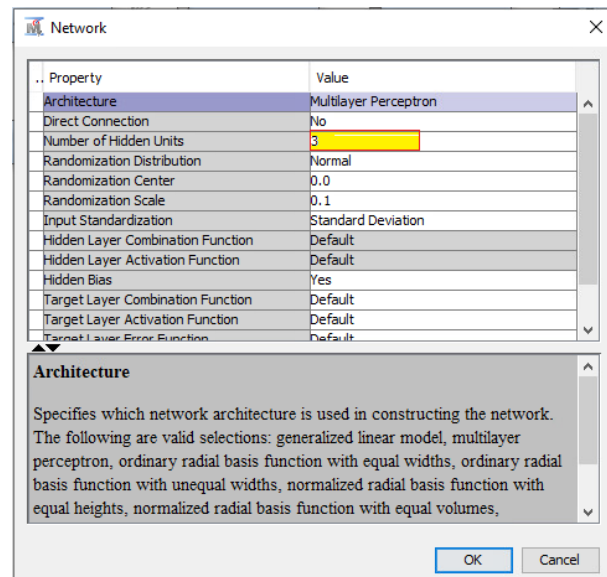
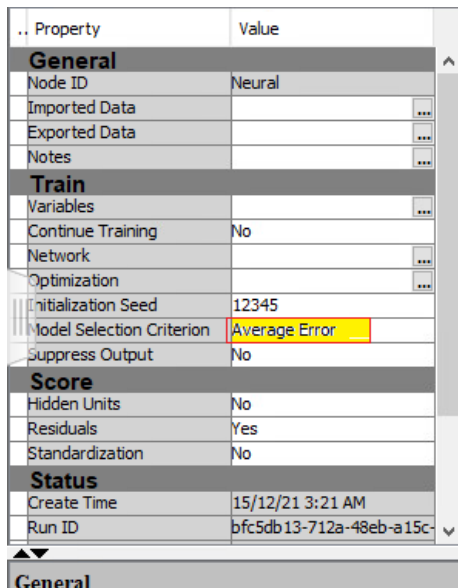
Steps:

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

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

5. Run the nodes

Panel screen shots



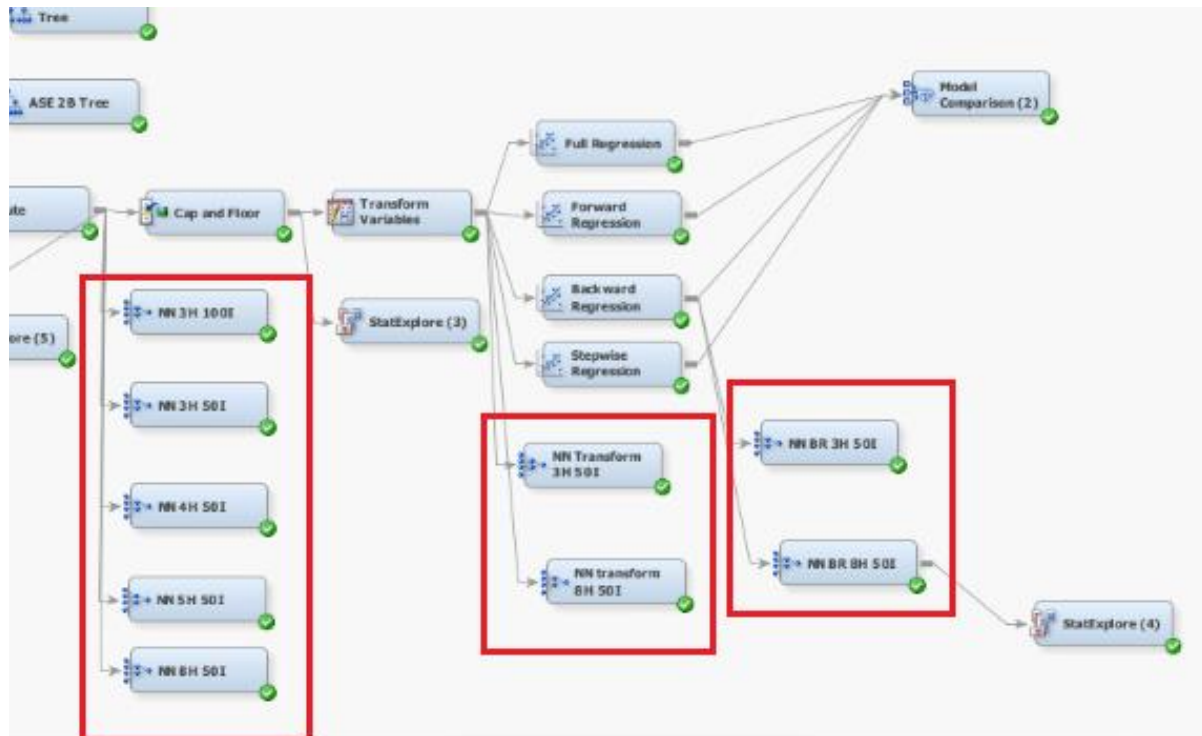
Optimization

Property	Value
Training Technique	Default
Maximum Iterations	50
Maximum Time	4 Hours
Nonlinear Options	
Propagation Options	
Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

Nonlinear Options

OK Cancel

Data diagram



Observation

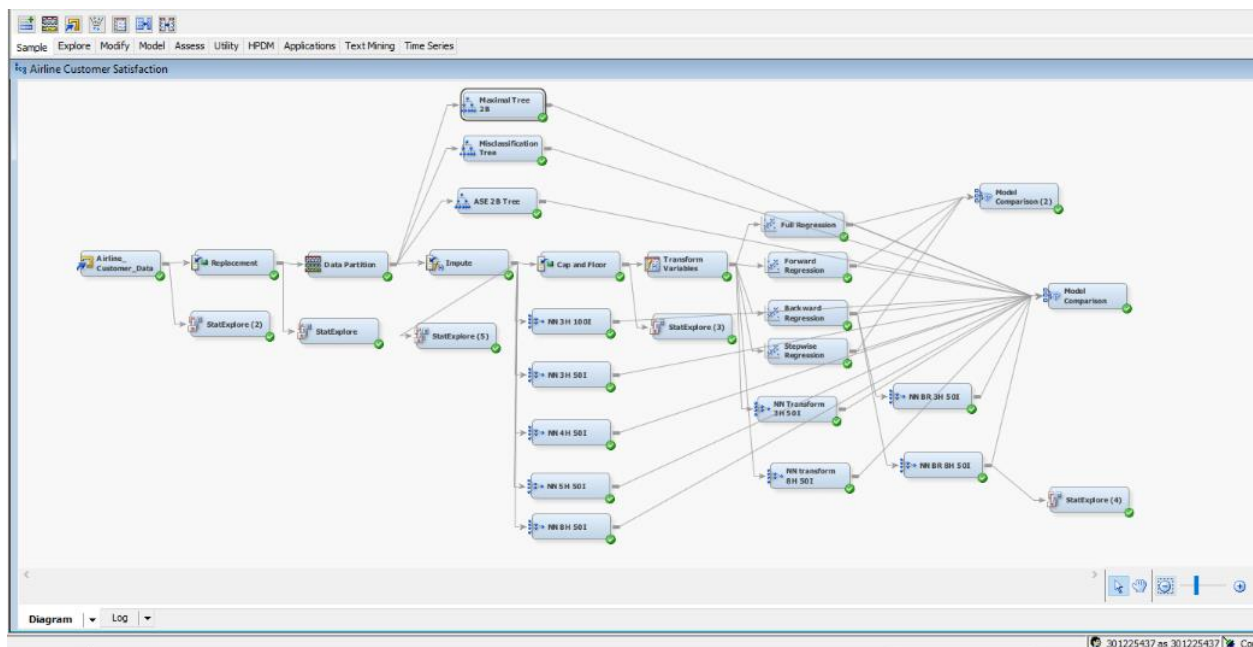
- We are running the neural network model at 50 iterations because the data converges at 80 iterations
- We need to determine the right number of hidden layers to run the neural network as a smaller number of hidden layers can lead to underfitting and more layers can lead to overfitting
- We are also running the model at 3 hidden layers, 4 hidden layers and 8 hidden layers to see if the accuracy of the model increase with depth. Beyond 8 hidden layers the accuracy does not improve.
- After completing the analysis, it was seen that The neural network drawn from the backward regression model with 8 hidden layers and 50 Iteration was the best model.

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 analyse 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	Target Label
Y	Neural10	Neural10	NN BR 8H 50I	satisfaction	0.069651	
	Neural8	Neural8	NN transform 8H 50I	satisfaction	0.071284	
	Tree	Tree	Maximal Tree 2B	satisfaction	0.075896	
	Tree3	Tree3	ASE 2B Tree	satisfaction	0.07599	
	Neural5	Neural5	NN 8H 50I	satisfaction	0.076115	
	Neural4	Neural4	NN 5H 50I	satisfaction	0.077497	
	Tree6	Tree6	Misclassification Tree	satisfaction	0.078823	
	Neural3	Neural3	NN 4H 50I	satisfaction	0.080075	
	Neural9	Neural9	NN BR 3H 50I	satisfaction	0.085795	
	Neural7	Neural7	NN Transform 3H 50I	satisfaction	0.086684	
	Neural	Neural	NN 3H 50I	satisfaction	0.086903	
	Neural2	Neural2	NN 3H 100I	satisfaction	0.086903	
	Reg3	Reg3	Backward Regression	satisfaction	0.110521	
	Reg	Reg	Full Regression	satisfaction	0.110523	
	Reg2	Reg2	Forward Regression	satisfaction	0.11053	
	Reg4	Reg4	Stepwise Regression	satisfaction	0.11053	

Based on the Fit statistics, Neural network using backward regression (NN BR 8H 50I) is the best model as it has the least Average square error at 0.069651.

8.2. Ranking based ROC index and Gini Coefficient

Fit Statistics						
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Roc Index	Valid: Gini Coefficient
Y	Neural10	Neural10	NN BR 8H 50I	satisfaction	0.968	0.937
	Neural8	Neural8	NN transform 8H 50I	satisfaction	0.967	0.935
	Tree	Tree	Maximal Tree 2B	satisfaction	0.957	0.914
	Tree3	Tree3	ASE 2B Tree	satisfaction	0.96	0.919
	Neural5	Neural5	NN 8H 50I	satisfaction	0.962	0.923
	Neural4	Neural4	NN 5H 50I	satisfaction	0.961	0.921
	Tree6	Tree6	Misclassification Tree	satisfaction	0.954	0.907
	Neural3	Neural3	NN 4H 50I	satisfaction	0.958	0.915
	Neural9	Neural9	NN BR 3H 50I	satisfaction	0.952	0.904
	Neural7	Neural7	NN Transform 3H 50I	satisfaction	0.952	0.904
	Neural	Neural	NN 3H 50I	satisfaction	0.952	0.903
	Neural2	Neural2	NN 3H 100I	satisfaction	0.952	0.903
	Reg3	Reg3	Backward Regression	satisfaction	0.923	0.845
	Reg	Reg	Full Regression	satisfaction	0.923	0.845
	Reg2	Reg2	Forward Regression	satisfaction	0.923	0.845
	Reg4	Reg4	Stepwise Regression	satisfaction	0.923	0.845

Neural network using backward Regression with 8 hidden layers and 50 Iteration has the Highest ROC index i.e area under the curve and highest Gini coefficient as well.

8.3. Final Outcome:

The best interpretable model build in this analysis was the maximal tree. Drawing some insights from this model we have make the following observation:

- The data analysed contained 45.27% of satisfied customers with the current facilities provided by the airline
- When Inflight entertainment rating is 3 or above the satisfaction rate has increased to 81.04%,
- The satisfaction rate further increased to 89.55% when the ease of booking was 3 and above
- From the data we were able to determine that the satisfaction rate reached 94.36% when the legroom service was 3 and above.

The above statistics shows gives us an insight to what the airline company is doing right as these insights are drawn from the satisfied customers.

Looking into the dissatisfied customers we can draw the following inference:

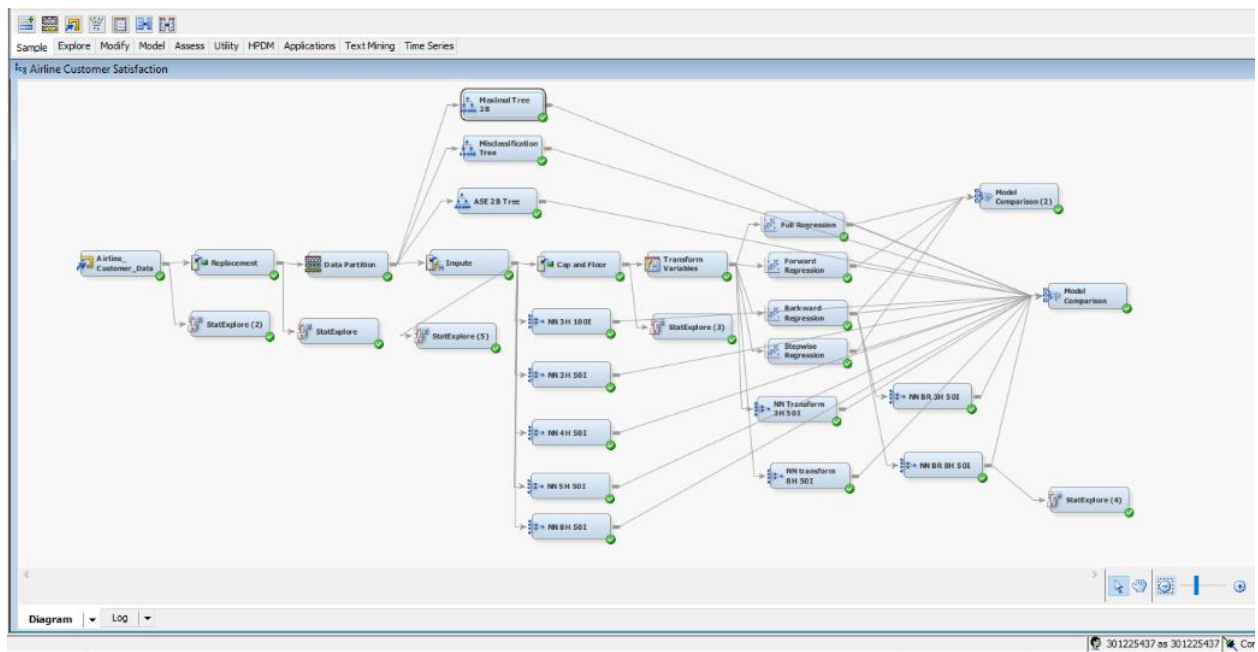
- The data analysed contained 54.73% of dissatisfied customers with the current facilities provided by the airline
- When Inflight entertainment rating is below 3 the satisfaction rate has decreased to 80.74%,
- The satisfaction rate further decreases to 90.10% when the seat comfort was 3 and below
- From the data we were able to determine that the dissatisfaction rate reached 94.36% when the ease of online booking was 3 and below.

8.5. Recommendation

Based on the analysis, The neural network using Backward regression with 8 hidden units and 50 iteration is the model which best fits the dataset. Since using only the result of neural network we cannot determine the variable contributing to satisfied Air customer, we are going to draw our recommendation by comparing the results of the third best interpretable model in this project which is the maximal tree.

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. The satisfaction rate is greatly impacted by the ease of online booking. The Airline should ensure that the online booking website is up to date and working efficiently without any bugs or buffer issues.
3. 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.
4. Loyal customers remain loyal to the airline even with decline in service.
5. Improving on the seat comfort and leg room can also help the airline tremendously in converting their disloyal customers to loyal customer
6. 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.

9.0. Complete Diagram



10.0. Reference

Center. Forward Regression | Center Based Statistics. (n.d.). Retrieved December 16, 2021, from <https://center-based-statistics.com/html/forwardReg.html>

Multicollinearity. Corporate Finance Institute. (2021, July 10). Retrieved December 16, 2021, from <https://corporatefinanceinstitute.com/resources/knowledge/other/multicollinearity/>

SAS enterprise miner. SAS Enterprise Miner | SAS Support. (n.d.). Retrieved December 16, 2021, from <https://support.sas.com/en/software/enterprise-miner-support.html#get-started>

Data Exploration - a complete introduction. OmniSci. (n.d.). Retrieved December 16, 2021, from <https://www.omnisci.com/learn/data-exploration#:~:text=What%20is%20Data%20Exploration%3F%20Data%20exploration%20definition%3A%20Data,to%20better%20understand%20the%20nature%20of%20the%20data.>

Jana, S. (2020, March 19). *Airlines customer satisfaction*. Kaggle. Retrieved December 16, 2021, from <https://www.kaggle.com/sjleshtrac/airlines-customer-satisfaction>