
PREDICTIVE MODELING FOR AIRLINE SATISFACTION SURVEY

Submitted by:

Lakshman Siva Kumar Varma - 301494624

Jibin George - 301475160

Aeliya Fatima Syeda - 301410524

Ibukunoluwa Ajibade - 301336392

Contents

Project Overview.....	6
Dataset Attributes.....	6
Initial Data Preparation	7
Data Partitioning.....	8
Decision Tree One: Default Tree with Largest Subtree Method.....	9
Objective	9
Configuration	9
Decision Tree Structure.....	10
Key Insights from the Tree:.....	11
Overall Model Assessment:	11
Decision Tree Two: ASE Tree	11
Objective	11
Configuration	11
Fit Statistics	11
Interpretation:	12
Tree Structure and Splits	12
Comparison with Default Tree.....	13
Key Insights	13
Decision Tree Three: Misclassification Tree	13
Objective	13
Configuration	13
Fit Statistics	14
Interpretation:	14
Tree Structure and Splits	15
Comparison with Other Trees	16
Key Observations:	16
Insights	16
Data Preparation: Handling Missing Variables	16
Objective	16
Imputation Process	16
Rationale for Mean Imputation.....	17
Impact on Regression and Neural Network Models	18
Data Preparation: Capping and Flooring Outliers.....	18
Objective	18
Capping and Flooring Process.....	18
Rationale for Capping and Flooring.....	19

Impact on Models	20
Data Exploration: Skewness and Kurtosis Check	20
Objective	20
Key Observations	20
Interpretation.....	21
Data Preparation: Variable Transformation	21
Objective	21
Impact of Log Transformations on Skewness	22
Results After Log Transformation.....	22
Analysis.....	23
Conclusion	23
Connecting StatExplore to Transform Variables Node	24
Objective	24
Steps Performed	24
Results After Log Transformation.....	24
Key Observations	25
Conclusion	25
Regression Modeling: Logistic Regression Approaches	25
Objective	25
Key Adjustments Implemented	28
Initial Observations Before Improvements.....	29
Updates to the Marketing 3 Process	31
Key Updates	31
Final Structure of the Dataset.....	31
Importing the Updated Data	32
Data Exploration – StatExplore Node.....	33
Interval Variables	33
Observations:	34
Purpose of Data Partitioning	34
Decision Tree One: Default Tree with Largest Subtree Method.....	35
Objective	35
Configuration	35
Decision Tree Structure.....	36
Key Insights from the Tree:.....	37
Overall Model Assessment:	37
Decision Tree Two: ASE Tree	37
Objective	37

Fit Statistics	37
Interpretation:	38
Third Decision Tree: Misclassification Tree	39
Configuration of the Misclassification Tree	39
Tree Diagram Overview	40
Fit Statistics	41
Performance Analysis	41
Insights from the Tree	41
Conclusion	42
Recoding Ordinal Survey Variables for Regression and Neural Network Models	43
Process of Recoding Using the Replacement Node	43
Imputation of Missing Values	44
Capping and Flooring Interval Variables	45
Replacement Summary	46
Output Highlights	46
Transform Variables to Handle Skewness and Kurtosis	47
Changes to Variable Roles and Levels	47
State Explore Results After Log Transformation in Marketing 3	48
Key Observations	48
Models Built	49
Odds Ratio Estimates	52
Insights and Key Takeaways	55
Fit Statistics	56
Odds Ratio Estimates	57
Insights	60
Conclusion	60
Interpretation of Stepwise and Forward Regression Results:	60
Fit Statistics:	61
Odds Ratio Estimates:	62
Key Insights:	64
Overall:	65
Comparison of Regression Models	65
Comparison Summary:	65
Key Observations:	65
Conclusion:	66
Initial Setup for Neural Networks	66
Optimization Settings	67

Neural Network 1: NN Impute.....	68
Model Comparison.....	81
Best Performing Models	82
Key Observations	82
Conclusion	82
Observations:	83
Conclusion:	84
Components Airlines Should Focus on for Customer Satisfaction	85
Summary Recommendations for Airlines	86

Project Overview

The primary objective of this project is to identify the factors that influence customer satisfaction for an airline using a combination of survey responses and internal operational data. By leveraging advanced modelling techniques, the analysis aims to uncover key drivers of satisfaction, providing actionable insights for improving the overall customer experience.

The analysis explores a comprehensive dataset containing multiple attributes that reflect passenger demographics, travel characteristics, and satisfaction levels. These variables are used to predict the overall satisfaction level, which is categorized as **Satisfied**, **Neutral**, or **Dissatisfied**. The findings from this project will assist the airline in prioritizing areas for service improvement and tailoring strategies for different customer segments.

Dataset Attributes

The dataset comprises the following features:

- **Id**: Unique identifier for passengers.
- **Gender**: Gender of the passengers (Female, Male).
- **Customer Type**: Customer categorization (Loyal customer, Disloyal customer).
- **Age**: Age of passengers.
- **Type of Travel**: Purpose of travel (Personal Travel, Business Travel).
- **Class**: Class of travel (Business, Economy, Economy Plus).
- **Flight Distance**: Distance of the flight.
- **Inflight Wi-Fi Service**: Satisfaction level (0–5, where 0 = Not Applicable, 1 = Least Satisfied, 5 = Most Satisfied).
- **Departure/Arrival Time Convenience**: Satisfaction level (0–5).
- **Ease of Online Booking**: Satisfaction level (0–5).
- **Gate Location**: Satisfaction level (0–5).
- **Food and Drink**: Satisfaction level (0–5).
- **Online Boarding**: Satisfaction level (0–5).
- **Seat Comfort**: Satisfaction level (0–5).
- **Inflight Entertainment**: Satisfaction level (0–5).
- **Onboard Service**: Satisfaction level (0–5).
- **Legroom Service**: Satisfaction level (0–5).
- **Baggage Handling**: Satisfaction level (1–5).
- **Check-in Service**: Satisfaction level (0–5).
- **Inflight Service**: Satisfaction level (0–5).

- **Cleanliness:** Satisfaction level (0–5).
- **Departure Delay (Minutes):** Delay in minutes at departure.
- **Arrival Delay (Minutes):** Delay in minutes at arrival.
- **Satisfaction:** Overall satisfaction level (Satisfied, Neutral, or Dissatisfied).

Marketing 2 model


Initial Data Preparation

Data Import and Variable Roles

The raw dataset was imported into SAS Enterprise Miner, and the roles of all variables were defined based on their type and purpose in the analysis. The following steps were performed:

- **Identification of Variable Types:**
 - **Ordinal Variables:** Variables such as satisfaction scores (e.g., Inflight Wi-Fi Service, Seat Comfort, Cleanliness) were categorized as ordinal to reflect their ranked nature (e.g., 0 = least satisfied, 5 = most satisfied).
 - **Nominal Variables:** Categorical variables like Gender, Customer Type, and Class were defined as nominal since they have no inherent order.
 - **Interval Variables:** Continuous variables like Flight Distance, Departure Delay (Minutes), and Arrival Delay (Minutes) were treated as interval.
 - **Binary Variables:** Variables with two distinct outcomes (e.g., dummy variables) were identified as binary.
- **Target Variable:**
 - The primary target variable, **Satisfaction**, was defined as **binary**:
 - **0:** Dissatisfied customers.
 - **1:** Satisfied customers.
 - This binary classification helped simplify the prediction problem and focus on distinguishing between satisfied and dissatisfied passengers.
 - Unique identifiers such as **Id** were retained for reference but excluded from predictive modeling.
 - Non-significant identifiers (e.g., **SR**) were rejected.

This careful assignment of roles ensured that the analysis leveraged each variable appropriately for model building.


Name	Role 	Level	Report	Order	Drop	Lower Limit	Upper Limit
id	ID	Nominal	No		No	.	
Inflight_entertainment	Input	Ordinal	No		No	.	
Gender	Input	Nominal	No		No	.	
Inflight_service	Input	Nominal	No		No	.	
Food_and_drink	Input	Ordinal	No		No	.	
Flight_Distance	Input	Interval	No		No	.	
Gate_location	Input	Ordinal	No		No	.	
Seat_comfort	Input	Ordinal	No		No	.	
Online_boarding	Input	Ordinal	No		No	.	
Type_of_Travel	Input	Nominal	No		No	.	
Leg_room_service	Input	Ordinal	No		No	.	
Inflight_wifi_service	Input	Ordinal	No		No	.	
On_board_service	Input	Ordinal	No		No	.	
Checkin_service	Input	Ordinal	No		No	.	
Baggage_handling	Input	Ordinal	No		No	.	
Class	Input	Nominal	No		No	.	
Arrival_Delay_in_Min	Input	Interval	No		No	.	
Age	Input	Interval	No		No	.	
Arrival_Delay_status	Input	Nominal	No		No	.	
Departure_Delay_in_Min	Input	Interval	No		No	.	
Ease_of_Online_booking	Input	Ordinal	No		No	.	
Departure_Delay_status	Input	Nominal	No		No	.	
Customer_Type	Input	Nominal	No		No	.	
Cleanliness	Input	Ordinal	No		No	.	
Departure_Arrival_time	Input	Ordinal	No		No	.	
SR	Rejected	Interval	No		No	.	
satisfaction	Target	Binary	No		No	.	

Data Partitioning

To enable robust model training and evaluation, the dataset was partitioned into two subsets:

- **Training Set:** 50% of the data, used for training and developing predictive models.
- **Validation Set:** 50% of the data, reserved for assessing model performance on unseen data.

The 50/50 split ensured a balance between training and validation, reducing the risk of overfitting while providing a fair evaluation of model performance.

Train	
Variables	
Output Type	Data
Partitioning	Default
Random Seed	12345
Data Set Allocation	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Target	Yes

Decision Tree One: Default Tree with Largest Subtree Method

Objective

The first decision tree model was built as a baseline to understand the relationships between variables and the binary target variable, **Satisfaction**. The goal was to establish a straightforward model while observing the impact of minimal adjustments.

Configuration

- **Subtree Method:** The subtree method was set to **Largest**. This ensures the tree is pruned to include only the largest possible subtree, effectively reducing model complexity and minimizing overfitting risks.
- No additional parameters were modified; all other settings were kept at their default values.

Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25

1. Fit Statistics

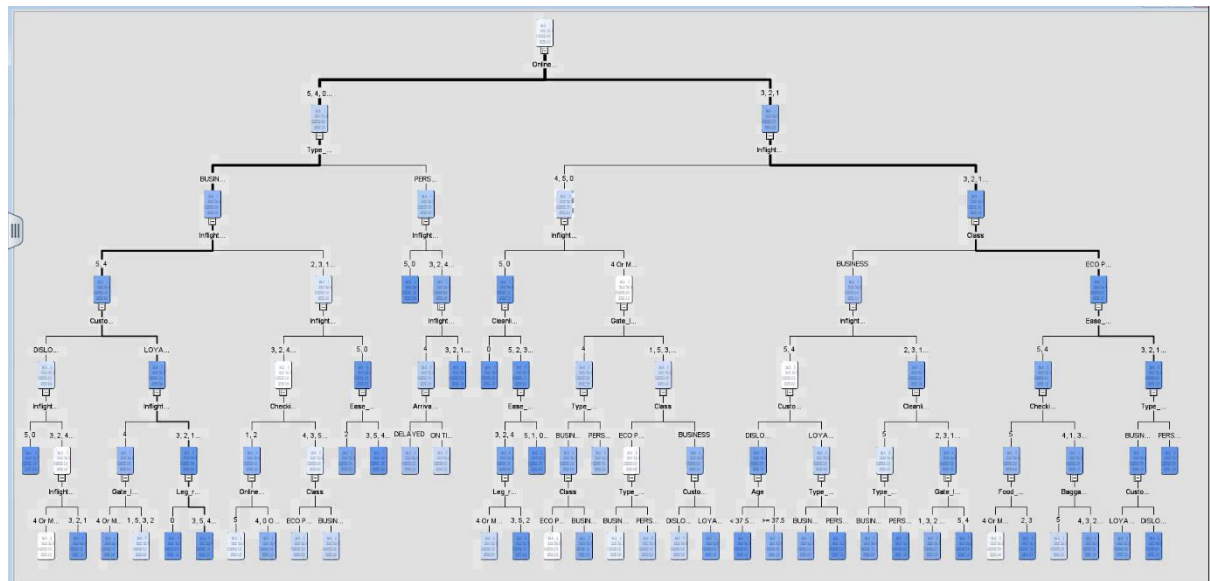
Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		NOBS	Sum of Frequencies	51953		51951
satisfaction		MISC	Misclassification Rate	0.079322		0.080634
satisfaction		MAX	Maximum Absolute Error	0.996878		0.997727
satisfaction		SSE	Sum of Squared Errors	5819.87		5854.067
satisfaction		ASE	Average Squared Error	0.056011		0.057305
satisfaction		RASE	Root Average Squared Error	0.236668		0.239384
satisfaction		DNV	Divisor for ASE	103906		103902
satisfaction		DFT	Total Degrees of Freedom	51953		

The fit statistics table provides an evaluation of the tree's performance across the **training**, **validation**, and **test** datasets. Key metrics include:

Average Squared Error (ASE):

- **Training:** 0.056011
- **Validation:** 0.057305
- **Interpretation:** ASE values for both training and validation sets are close, suggesting the model generalizes well and does not overfit.

Decision Tree Structure



The visualized decision tree highlights how customer satisfaction is segmented using hierarchical splits based on key predictors. Below is an interpretation of the branches:

- **Top Split:**
 - **Inflight Wi-Fi Service (≤ 4 or ≥ 4):**
 - This is the most significant predictor. Passengers rating Wi-Fi service lower than 4 are more likely dissatisfied, while those rating it higher are more likely satisfied.
- **Subsequent Splits:**
 - **Branch 1:** For passengers with **Inflight Wi-Fi ≤ 4** :
 - **Cleanliness (≤ 1 or ≥ 1):** Cleanliness further distinguishes dissatisfied customers.
 - **Class (Business vs. Economy Plus):** Business class passengers tend to report higher satisfaction levels.
 - **Branch 2:** For passengers with **Inflight Wi-Fi ≥ 4** :
 - **Online Boarding (≤ 4 or ≥ 4):** Satisfaction increases with higher ratings for online boarding.
 - **Ease of Booking (≤ 5 or ≥ 5):** Booking ease impacts satisfaction further.
 - Additional splits involve variables like **Customer Type (Loyal vs. Disloyal)**, **Type of Travel (Business vs. Personal)**, and **Seat Comfort**. These splits indicate secondary factors influencing satisfaction.

Key Insights from the Tree:

1. **Inflight Wi-Fi Service** is the strongest predictor of satisfaction, indicating that improving this service could significantly impact customer experiences.
2. **Cleanliness, Class, and Online Boarding** are also critical factors for passengers who are dissatisfied.
3. Secondary predictors like **Customer Type** and **Ease of Booking** show that loyal customers and seamless booking processes contribute to satisfaction.

Overall Model Assessment:

- The **misclassification rate (~8%)** and the close alignment of ASE values between training and validation datasets demonstrate the tree's reliability.
- The tree's logical splits align well with intuitive expectations: service quality (Wi-Fi, cleanliness) and convenience-related factors (booking, boarding) are key drivers of satisfaction.

Decision Tree Two: ASE Tree

Objective

The second decision tree model was configured to focus on minimizing the **Average Squared Error (ASE)**, which measures the average squared difference between predicted and actual values for the binary satisfaction target. This configuration aims to create a more accurate predictive model by optimizing for this specific error metric.

Configuration

Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	

- **Subtree Method:** Set to **Largest**, ensuring that the tree is pruned to its most effective size while avoiding overfitting.
- **Assessment Measure:** Changed to **Average Square Error (ASE)** to align the pruning process with minimizing prediction error rather than classification metrics.
- **Assessment Fraction:** 0.25, the same as the default tree, specifying the fraction of data used for assessment during pruning.
- All other settings were kept similar to the default tree.

Fit Statistics

The fit statistics for the ASE Tree are as follows:

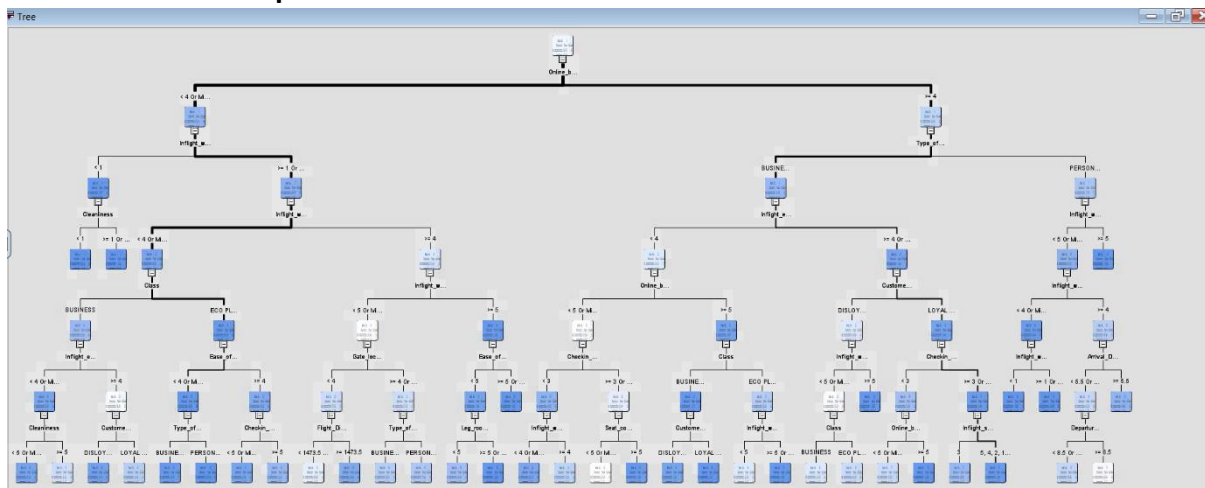
Metric	Training	Validation
Misclassification Rate (MISC)	0.079322	0.080634

Metric	Training	Validation
Average Squared Error (ASE)	0.056011	0.057305
Root ASE (RASE)	0.236606	0.239384
Sum of Squared Errors (SSE)	5819.87	5954.067

Interpretation:

- **ASE Values:** The training and validation ASE values (0.056 and 0.057) indicate consistent and low error rates, showing that the model is well-fitted to the data without overfitting.
- **MISC (Misclassification Rate):** The misclassification rate remains consistent with the default tree (~8%), indicating that changing the assessment measure to ASE did not significantly impact classification accuracy.
- **RASE (Root ASE):** The small difference between training and validation RASE values (0.2366 vs. 0.2393) confirms the model's robustness across datasets.

Tree Structure and Splits



The structure of the ASE Tree is highly similar to the Default Tree, with the following main splits:

1. Primary Split:

- **Inflight Wi-Fi Service (< 4 or ≥ 4)** remains the most critical predictor of satisfaction.
- Customers rating Wi-Fi lower than 4 are more likely dissatisfied.

2. Secondary Splits:

- For **Inflight Wi-Fi < 4**:
 - **Cleanliness** and **Class** (e.g., Business vs. Economy Plus) are the next most important predictors.

- For **Inflight Wi-Fi ≥ 4** :
 - Factors like **Ease of Booking**, **Online Boarding**, and **Legroom Service** drive satisfaction levels.

3. Deeper Splits:

- Similar to the default tree, the ASE tree incorporates splits based on **Customer Type (Loyal vs. Disloyal)**, **Type of Travel (Business vs. Personal)**, and **Seat Comfort**.

Comparison with Default Tree

- **Similarity:**
 - Both trees prioritize **Inflight Wi-Fi Service** as the key predictor of customer satisfaction.
 - The secondary and deeper splits (e.g., Cleanliness, Class, Ease of Booking) are also consistent between the two trees, highlighting the same important variables across models.
- **Difference:**
 - The ASE tree explicitly optimizes for prediction accuracy rather than classification performance, reflected in its lower ASE values.
 - No significant changes in tree structure or variable importance were observed, suggesting that both configurations capture similar relationships in the data.

Key Insights

- The **ASE Tree** reaffirms that **Inflight Wi-Fi Service**, **Cleanliness**, and **Ease of Booking** are critical drivers of customer satisfaction.
- Optimization for ASE does not compromise classification accuracy, as the misclassification rate remains consistent across both trees (~8%).
- The similarity in structure and performance between the two trees highlights the robustness of these predictors.

Decision Tree Three: Misclassification Tree

Objective

The third decision tree was configured to focus on minimizing the **Misclassification Rate** as the assessment measure. This approach prioritizes correctly classifying the binary target variable **Satisfaction** (Satisfied vs Dissatisfied) by reducing the number of incorrect predictions.

Configuration

- **Subtree Method: Assessment**, allowing the model to prune based on validation data performance.

- **Assessment Measure: Misclassification**, which directly targets reducing the rate of incorrect classifications.
- **Number of Leaves: 1** (initial setup).
- **Assessment Fraction: 0.25**, indicating that 25% of the data was used for validation during pruning.

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	

Fit Statistics

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		NOBS	Sum of F...	51953	51951	-
satisfaction		MISC	Misclassi...	0.079784	0.080364	-
satisfaction		MAX	Maximu...	0.989807	0.989807	-
satisfaction		SSE	Sum of S...	6320.595	6418.103	-
satisfaction		ASE	Average ...	0.06083	0.061771	-
satisfaction		RASE	Root Ave...	0.246637	0.248537	-
satisfaction		DIV	Divisor fo...	103906	103902	-
satisfaction		DFT	Total De...	51953		

Metric	Training	Validation
Misclassification Rate (MISC)	0.079784	0.080364
Average Squared Error (ASE)	0.06083	0.061771
Root ASE (RASE)	0.246637	0.248537
Sum of Squared Errors (SSE)	6320.595	6418.103

Interpretation:

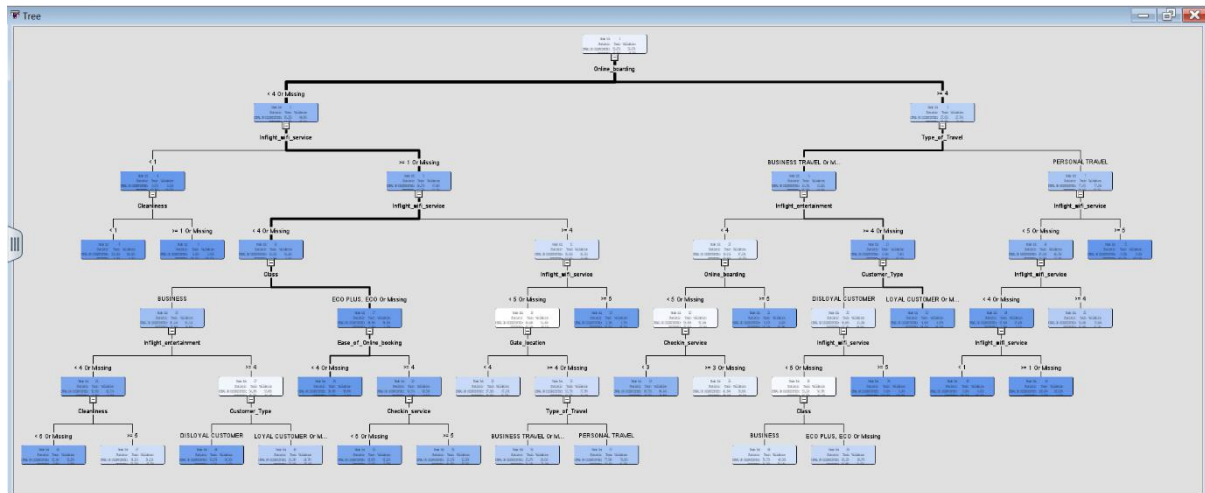
- **Misclassification Rate:**
 - Training: **0.0797** (7.97%)
 - Validation: **0.0803** (8.03%)
The slight increase in misclassification rate from training to validation suggests that the model generalizes well, with only a minimal performance drop on unseen data.
- **Average Squared Error (ASE):**
The ASE is slightly higher compared to the ASE tree (from **0.0573** to **0.0617**), which is

expected since the focus here is on **misclassification** rather than minimizing error variance.

- **Root ASE (RASE):**

The small gap between training (0.2466) and validation (0.2485) reinforces that the tree is balanced and not overfitting.

Tree Structure and Splits



1. Primary Split:

- **Inflight Wi-Fi Service (< 4 or Missing vs ≥ 4)** remains the strongest predictor of customer satisfaction, consistent with both the **default tree** and **ASE tree**.

2. Secondary Splits:

- For **Inflight Wi-Fi < 4** :
 - **Cleanliness** and **Class** distinguish dissatisfied customers further.
- For **Inflight Wi-Fi ≥ 4** :
 - **Online Boarding**, **Type of Travel**, and **Customer Type** play critical roles in splitting satisfied customers.

3. Deeper Splits:

- Other key predictors include:
 - **Ease of Online Booking**
 - **Gate Location**
 - **Check-in Service**
 - **Inflight Entertainment**

Tree Complexity:

Compared to the ASE tree, the misclassification tree exhibits a similar structure but

introduces slightly deeper branches in some areas. This allows for more nuanced classification, particularly for borderline cases.

Comparison with Other Trees

Metric	Default Tree	ASE Tree	Misclassification Tree
Misclassification	~0.0803	~0.0803	~0.0803
ASE	0.0573	0.0573	0.0617
Top Split	Inflight Wi-Fi	Inflight Wi-Fi	Inflight Wi-Fi
Secondary Splits	Similar structure	Similar structure	Slightly deeper splits

Key Observations:

- Consistency:** Across all three trees, **Inflight Wi-Fi Service** remains the most critical predictor, followed by **Cleanliness**, **Class**, and **Online Boarding**.
- Misclassification Focus:** While this tree focuses on minimizing misclassifications, the misclassification rates are **identical** across all three trees (~8%).
- ASE Trade-Off:** The misclassification tree slightly increases the ASE compared to the ASE tree, as expected when prioritizing classification accuracy.

Insights

- The **Misclassification Tree** provides a classification-focused perspective while reinforcing the importance of key predictors like **Inflight Wi-Fi**, **Cleanliness**, and **Online Boarding**.
- The similarities in splits and performance metrics across all three trees suggest that the dataset is robust and key variables are consistent drivers of satisfaction.

Data Preparation: Handling Missing Variables

Objective

Before proceeding with regression and neural networks, the first step in data preparation involved addressing **missing values** in the dataset. Proper handling of missing data ensures model robustness and minimizes bias in predictions.

Imputation Process

The missing values were handled using the following approach:

- Variable with Missing Data:**
 - Arrival Delay (Minutes)** was identified as having **162 missing values** in the training data.
- Imputation Method:**

- The missing values were imputed using the **Mean** of the variable.
- **Imputed Value: 15.18264** (calculated mean of the non-missing values in Arrival Delay).

3. Indicator Variable:

- An **indicator variable** was created to flag instances where missing values were imputed.
- This allows models to differentiate between original and imputed data, which can help uncover any patterns associated with missing data.

Configuration Details

Indicator Variables	
Type	Unique
Source	Imputed Variables
Role	Input

- **Imputation Summary:**

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Arrival Delay...	MEAN	IMP Arrival ...	M Arrival Del...	15.18264	INPUT	INTERVAL		162

- **Variable Name:** Arrival Delay (Minutes)
- **Impute Method:** Mean
- **Imputed Variable:** Created as **IMP_Arrival_Delay**
- **Indicator Variable:** Created for Arrival Delay to flag missing data.
- **Imputed Value:** 15.18264
- **Number of Missing Values:** 162 (training dataset).
- **Measurement Level:** The imputed variable remains as **Interval**, preserving its continuous nature.

Rationale for Mean Imputation

- **Why Mean?:** The mean is a commonly used method for imputing missing values in continuous (interval) variables. It ensures that missing values are replaced with a representative central value, reducing the risk of bias.
- **Why an Indicator Variable?:** Adding an indicator variable ensures the model can account for any patterns linked to missing data, potentially improving model accuracy.

Impact on Regression and Neural Network Models

- The imputed values allow regression and neural network models to utilize the entire dataset without discarding rows with missing values.
- The addition of an indicator variable provides flexibility for the models to determine whether missing values carry any predictive significance.

Data Preparation: Capping and Flooring Outliers

Objective

To manage outliers effectively, the **Replacement Node** was used to apply **capping and flooring** techniques. This ensures that extreme values in interval variables are handled appropriately, reducing their potential impact on regression and neural network models.

Capping and Flooring Process

1. Node Used:

- The **Replacement Node** was applied and named **Cap and Floor**.


2. Variables

Processed:

Three interval variables were identified for outlier treatment:

- **Departure Delay (Minutes)**
- **Flight Distance**
- **Imputed Arrival Delay (IMP_Arrival Delay)**

3. Replacement Method:

 Interactive Replacement Interval Filter ✕

Columns: <input type="checkbox"/> Label <input type="checkbox"/> Mining <input type="checkbox"/> Basic <input type="checkbox"/> Statistics					
Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace I
Age	Default	Default	.	.	Default
Departure_Del	Default	User Specified	0	127.8939	Manual
Flight_Distance	Default	User Specified	0	4177.511	Manual
IMP_Arrival_De	Default	User Specified	0	7.634176	Manual
SR	Default	Default	.	.	Default

- **User-Specified Limits** were manually calculated and applied.

Variable	Lower Limit	Upper Limit
Departure Delay	0	127.8939
Flight Distance	0	4177.511

Variable	Lower Limit	Upper Limit
IMP_Arrival Delay	0	7.634176

- **Lower Limit:** Outliers below the lower limit were capped to **0**.
- **Upper Limit:** Outliers above the calculated upper limits were capped to their respective values.

4. Replacement Results:

Total Replacement Counts

Variable	Label	Role	Train	Validation
Departure ...	Departure ...	INPUT	1146	1143
Flight Dist...	Flight Dist...	INPUT	23	35
IMP Arrival...	Imputed Arr...	INPUT	16493	16363

Output

```
1 *-----*
2 User:      u64072813
3 Date:      December 13, 2024
4 Time:      20:42:26
5 *-----*
6 * Training Output
7 *-----*
8
9
10
11
12 Variable Summary
13
14           Measurement   Frequency
15 Role      Level        Count
16
17 INPUT     BINARY        1
18 INPUT     INTERVAL      4
19 INPUT     NOMINAL        5
20 INPUT     ORDINAL       13
```

Interval Variables

Variable	Replace Variable	Lower limit	Upper Limit	Label	Limits Method	Replacement Method	Lower Replacement Value	Upper Replacement Value
Departure D...	REP Depart...	0	127.8939	Departure D...	MANUAL	MANUAL	.	.
Flight Distance	REP Flight ...	0	4177.511	Flight Distance	MANUAL	MANUAL	.	.
IMP Arrival ...	REP IMP Ar...	0	7.634176	Imputed Arriv...	MANUAL	MANUAL	.	.

- **Departure Delay:** 1146 replacements in the training set, 1143 replacements in validation.
- **Flight Distance:** 23 replacements in training, 35 replacements in validation.
- **IMP_Arrival Delay:** 16,493 replacements in training, 16,363 in validation.

5. Output:

- New capped variables were created with the prefix **REP_** (e.g., REP_Departure_Delay, REP_Flight_Distance, REP_IMP_Arrival_Delay).

Rationale for Capping and Flooring

- **Why Capping and Flooring?:** Extreme values can skew model predictions, particularly in regression and neural networks, which are sensitive to large ranges in interval variables. Capping and flooring ensure that these outliers are handled without discarding valuable data.
- **Why User-Specified Limits?:** Manually calculating limits allows for greater control over the treatment of outliers, ensuring that limits are realistic and tailored to the dataset.

Impact on Models

- Regression**
Capping and flooring help stabilize the coefficients and improve model accuracy by preventing extreme values from dominating the regression fit.
 - Neural**
Neural networks benefit from normalized and outlier-free data, allowing the model to converge more effectively during training.
- Models:**

Networks:

Data Exploration: Skewness and Kurtosis Check

Objective

Using the **StatExplore Node**, the data was examined for skewness and kurtosis. This step ensures that the interval variables are evaluated for their distribution characteristics, which directly impact regression and neural network performance. Addressing skewness and kurtosis helps normalize variables, improving model stability and accuracy.

Key Observations

The following variables were checked for **Skewness** and **Kurtosis**:

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	satisfactionneutral or...	REP	Flig...	671	12	29428	31	4000	925.5447	784.3218	1.61539	2.403443	INPUT	Replace...	-0.22076	0.288703	1
TRAIN	satisfactionneutral or...	REP	Flig...	1258	11	22502	31	4000	1530.652	1125.522	0.552796	-0.91231	INPUT	Replace...	0.288703	0.288703	2
TRAIN	satisfactionneutral or...	REP	De...	0	733	28707	0	127	11.74734	22.76519	2.589755	6.854607	INPUT	Replace...	0.111938	0.145403	1
TRAIN	satisfactionneutral or...	REP	De...	0	413	22100	0	127	9.028597	19.72683	3.102315	10.51335	INPUT	Replace...	-0.1454	0.145403	2
TRAIN	satisfactionneutral or...	REP	IM...	0	10568	18872	0	7	0.726208	1.755108	2.428027	4.673611	INPUT	Replace...	0.074495	0.084732	1
TRAIN	satisfactionneutral or...	REP	IM...	0	5925	16588	0	7	0.61858	1.570508	2.636229	5.988031	INPUT	Replace...	-0.08475	0.084732	2
TRAIN	satisfactionneutral or...	Age		37	0	29440	7	85	37.71579	16.4952	0.207355	-0.82913	INPUT	Age	-0.04371	0.05716	1
TRAIN	satisfactionneutral or...	Age		43	0	22513	7	80	41.69413	12.69996	-0.24386	-0.49403	INPUT	Age	0.05716	0.05716	2

Variable	Skewness	Kurtosis	Interpretation
REP_Flight_Distance	1.61539	2.403443	Moderate positive skew, slight kurtosis.
REP_Departure_Delay	2.589755	8.854607	High positive skew and kurtosis. Outliers may exist.
REP_IMP_Arrival Delay	2.428027	4.673611	Significant positive skew and kurtosis.
Age (Neutral Group)	0.207355	-0.829133	Near-normal distribution.

Variable	Skewness	Kurtosis	Interpretation
Age (Satisfied Group)	-0.24386	-0.49403	Slight negative skew, almost normal.

Interpretation

1. Skewness:

- A skewness value close to **0** indicates a symmetric distribution.
- Positive skew (e.g., **REP_Departure_Delay**) means the data has a tail on the right, indicating outliers or extreme high values.
- Significant skewness, as seen in **REP_Departure_Delay** and **REP_IMP_Arrival Delay**, can adversely affect regression and neural network models.

2. Kurtosis:

- A kurtosis value of **3** corresponds to a normal distribution.
- High kurtosis (>3) indicates the presence of outliers (heavy tails). For example, **REP_Departure_Delay** and **REP_IMP_Arrival Delay** exhibit heavy-tailed distributions.
- Negative kurtosis suggests light tails, as seen in the **Age** variable for the neutral and satisfied groups.

Next Steps

To address skewness and kurtosis:

1. **Log Transformations:** For variables like **REP_Departure_Delay** and **REP_IMP_Arrival Delay**, log transformations can help reduce skewness.

Data Preparation: Variable Transformation

Objective

The Transform Variables node was applied to address the skewness and kurtosis issues identified during the data exploration stage. By performing log transformations on highly skewed interval variables, the goal was to normalize their distributions and improve the performance of regression and neural network models.

Variables - Trans

(none)	<input type="checkbox"/> not	Equal to		...
Columns:	<input type="checkbox"/> Label		<input type="checkbox"/> Mining	
Name	Method	Number of Bins	Role	Level
Age	Default	4	Input	Interval
Arrival_Delay	Default	4	Rejected	Nominal
Baggage_hand	Default	4	Input	Ordinal
Checkin_servic	Default	4	Input	Ordinal
Class	Default	4	Input	Nominal
Cleanliness	Default	4	Input	Ordinal
Customer_Typ	Default	4	Input	Nominal
Departure_Arr	Default	4	Input	Ordinal
Departure_Del	Default	4	Rejected	Interval
Departure_Del	Default	4	Rejected	Nominal
Ease_of_Online	Default	4	Input	Ordinal
Flight_Distance	Default	4	Rejected	Interval
Food_and_drin	Default	4	Input	Ordinal
Gate_location	Default	4	Input	Ordinal
Gender	Default	4	Input	Nominal
IMP_Arrival_De	Default	4	Rejected	Interval
Inflight_enterta	Default	4	Input	Ordinal
Inflight_service	Default	4	Input	Nominal
Inflight_wifi_se	Default	4	Input	Ordinal
Leg_room_serv	Default	4	Input	Ordinal
M_Arrival_Dela	Default	4	Input	Binary
On_board_serv	Default	4	Input	Ordinal
Online_boardin	Default	4	Input	Ordinal
REP_Departure	Log	4	Input	Interval
REP_Flight_Dis	Log	4	Input	Interval
REP_IMP_Arriv	Log	4	Input	Interval
SR	Default	4	Rejected	Interval
Seat_comfort	Default	4	Input	Ordinal
Type_of_Trave	Default	4	Input	Nominal
satisfaction	Default	4	Target	Binary

Impact of Log Transformations on Skewness

The log transformations applied to the interval variables partially resolved the skewness issues, with significant improvements observed in most cases. Below is a detailed assessment of each variable:

Results After Log Transformation

Variable	Skewness (Before)	Skewness (After)	Improvement	Comment
REP_Departure Delay in Minutes	2.790219	0.888297	Reduced	Skewness significantly improved, approaching a

Variable	Skewness (Before)	Skewness (After)	Improvement	Comment
				symmetric distribution.
REP_Flight Distance	1.101609	-0.20701	Resolved	Skewness reduced to near-zero, indicating a balanced distribution.
REP_IMP_Arrival Delay	2.526378	2.031111	Partial	Skewness reduced but remains positive, suggesting some residual skewness.

Analysis

1. REP_Departure Delay in Minutes:

- The skewness was reduced from **2.79** to **0.88**, a substantial improvement.
- This indicates that the log transformation effectively addressed the extreme high values and normalized the distribution.

2. REP_Flight Distance:

- The skewness decreased from **1.10** to **-0.21**, which is very close to zero.
- The distribution is now symmetric, meaning the log transformation fully resolved the skewness issue.

3. REP_IMP_Arrival Delay:

- While the skewness reduced from **2.52** to **2.03**, it still remains above the acceptable range for a normal distribution.
- This suggests that some extreme values (outliers) persist, and additional transformations or treatments (e.g., further capping or flooring) might be necessary if this variable impacts model performance.

Conclusion

• Resolved Variables:

- Log transformation successfully addressed skewness for **REP_Departure Delay** and **REP_Flight Distance**.

• Partially Resolved Variable:

- For **REP_IMP_Arrival Delay**, log transformation reduced skewness but did not fully resolve it. Additional methods, such as further capping or more advanced transformations, may be considered depending on model requirements.

Connecting StatExplore to Transform Variables Node

Objective

To evaluate the effect of **log transformations** applied to address skewness and kurtosis, we connected the **StatExplore Node** to the **Transform Variables Node**. This allowed us to perform a post-transformation analysis to observe changes in variable distributions.

Steps Performed

1. Transform Variables Node:

- Applied **log transformations** to three skewed interval variables:
 - **REP_Departure Delay in Minutes**
 - **REP_Flight Distance**
 - **REP_IMP_Arrival Delay in Minutes**

2. StatExplore Node:

- Connected to the Transform Variables Node to perform data exploration and assess:
 - **Skewness:** Measure of distribution symmetry.
 - **Kurtosis:** Measure of tail heaviness in the distribution.

Results After Log Transformation

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	satisfactionneutral or...	LOG	REP Departure Delay in Minut	0	733	28707	0	4.85203	1.233623	1.562615	0.786179	-0.92803	INPUT	Transfor...	0.07581	0.098474	1
TRAIN	satisfactionsatisfied	LOG	REP Departure Delay in Minut	0	413	22100	0	4.85203	1.033773	1.453139	1.027493	-0.42089	INPUT	Transfor...	-0.09847	0.098474	2
TRAIN	satisfactionneutral or...	LOG	REP IMP Arrival Delay in Min	0	10568	18872	0	2.079442	0.277061	0.613826	1.964316	2.237423	INPUT	Transfor...	0.055379	0.063004	1
TRAIN	satisfactionsatisfied	LOG	REP IMP Arrival Delay in Min	0	5925	16588	0	2.079442	0.245983	0.571058	2.107765	2.882507	INPUT	Transfor...	-0.063	0.063004	2
TRAIN	satisfactionneutral or...	Age		37	0	29440	7	85	37.71579	16.4952	0.207355	-0.82913	INPUT	Age	-0.04371	0.05716	1
TRAIN	satisfactionsatisfied	Age		43	0	22513	7	80	41.69413	12.69996	-0.24386	-0.49403	INPUT	Age	0.05716	0.05716	2
TRAIN	satisfactionneutral or...	LOG	REP Flight Distance	6.510258	12	28428	3.465736	8.2943	6.502804	0.831995	-0.09624	-0.43363	INPUT	Transfor...	-0.03013	0.038407	1
TRAIN	satisfactionsatisfied	LOG	REP Flight Distance	7.138073	11	22502	3.465736	8.2943	6.969059	0.951057	-0.546	-0.94662	INPUT	Transfor...	0.038407	0.038407	2

Variable	Skewness	Kurtosis	Improvement	Interpretation
LOG_REP_Departure Delay in Minutes (Neutral)	0.786179	-0.92803	Significant Improvement	Skewness reduced; close to normal.

Variable	Skewness	Kurtosis	Improvement	Interpretation
LOG_REP_Departure Delay in Minutes (Satisfied)	1.027493	-0.42089	Moderate Improvement	Slight positive skew remains but vastly improved.
LOG_REP_IMP_Arrival Delay in Minutes (Neutral)	1.964316	2.237423	Partial Improvement	Reduced skewness, but moderate skew persists.
LOG_REP_IMP_Arrival Delay in Minutes (Satisfied)	2.107765	2.882507	Partial Improvement	Moderate skewness and kurtosis persist.
LOG_REP_Flight Distance (Neutral)	-0.09624	-0.43363	Resolved	Symmetric distribution achieved.
LOG_REP_Flight Distance (Satisfied)	-0.546	-0.64662	Resolved	Symmetric and normalized distribution.

Key Observations

1. Skewness Improvement:

- **REP_Departure Delay** and **REP_Flight Distance** achieved near-zero skewness, indicating successful normalization.
- **REP_IMP_Arrival Delay** improved but retains moderate skewness, suggesting further refinement may be necessary.

2. Kurtosis Improvement:

- The kurtosis for **REP_Departure Delay** and **REP_Flight Distance** is now close to zero, confirming that extreme values were effectively mitigated.
- **REP_IMP_Arrival Delay** still exhibits moderate tail heaviness, reflecting residual outliers.

Conclusion

By connecting **StatExplore** to the **Transform Variables Node**, we verified that the log transformations significantly improved the distributions of key interval variables. While most variables achieved near-normal distributions, **LOG_REP_IMP_Arrival Delay** retains slight skewness, which can be monitored during modeling.

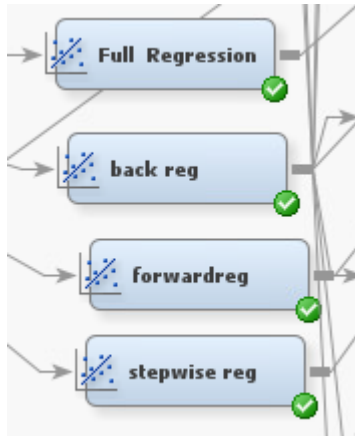
Regression Modeling: Logistic Regression Approaches

Objective

Despite some issues in the data (e.g., residual skewness in certain variables), we proceeded to build **Logistic Regression Models** to predict the binary target variable **Satisfaction**. Four

regression models were developed to identify the most influential predictors and assess model performance.

Models Built



1. Full Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...

- Includes **all variables** without any selection criteria.
- Purpose: Serves as a baseline model to compare against other regression approaches.

2. Backward Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

- **Selection Model:** Backward
- **Selection Criterion:** Validation Error

- Approach: Starts with all predictors and sequentially removes the least significant variables based on the validation error.

3. Forward Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

- **Selection Model:** Forward
- **Selection Criterion:** Validation Error
- Approach: Starts with no predictors and sequentially adds the most significant variables until no further improvement in validation error is observed.

4. Stepwise Regression

Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

- **Selection Model:** Stepwise
- **Selection Criterion:** Validation Error
- Approach: Combines forward and backward methods, adding significant variables while removing insignificant ones to optimize validation error.

After analyzing the initial regression models (Full, Backward, Forward, and Stepwise) and reviewing their results, the team observed areas where model performance could be improved. During a **project review meeting**, the professor suggested several changes to optimize the modeling process. Based on these recommendations, we decided to **restructure the model workflow** and implement significant data adjustments.

Key Adjustments Implemented

1. Creation of New Dummy Variables:

- Two **dummy variables** were created for **Departure Delay** and **Arrival Delay**:
 - Flights were classified as either **"Delayed"** or **"On Time"**.
 - This simplified the analysis by converting the numeric delay variables into binary categories, which are easier to interpret and use in models.
- Rationale:** Simplifying these interval variables helps regression models and neural networks better capture the impact of flight delays on satisfaction.

2. Survey Data Recoding:

- Ordinal survey variables (e.g., inflight services, check-in, food, and legroom ratings) were **recoded** into simplified nominal categories:
 - 0** → Remains **0**
 - 1-2** → Combined into a single category **1**
 - 3** → Remains as **3**
 - 4-5** → Combined into a single category **5**
- Rationale:**
 - This reduced complexity in the regression analysis by grouping ratings into fewer meaningful categories.
 - It also helped in minimizing noise caused by subtle variations in ratings while retaining the overall trend of satisfaction levels.

Selected Model	Predecessor or Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid Misclassification Rate	Train Total Degrees of Freedom	Train Degrees of Freedom for Error	Train Model Degrees of Freedom	Train Number of Estimated Weights	Train Akaike's Information Criterion	Train Schwarz's Bayesian Criterion	Train Average Squared Error	Train Maximum Absolute Error	Train Desvar for ASE	Train Sum of Frequencies	Train Root Average Squared Error	Train Sum of Squared Errors	Train Sum of Case Weights Times Freq	Train Final Prediction Error	Train Mean Squared Error	Train Root Final Prediction Error	Train Root Mean Squared Error	Train Average Error Function
Req3	Req3	Req3	stepwise	satisfaction		0.069258	51953	51893	60	60	18624.51	19155.99	0.050868	0.999962	103906	51953	0.225538	5285.447	103906	0.050985	0.050926	0.225799	0.225669	0.178089
Req4	Req4	Req4	forwardreq	satisfaction		0.069258	51953	51893	60	60	18624.51	19155.99	0.050868	0.999962	103906	51953	0.225538	5285.447	103906	0.050985	0.050926	0.225799	0.225669	0.178089
Neural1	Neural1	Neural1	nn impute	satisfaction		0.072241	51953	51709	244	244	19041.25	21202.63	0.050359	0.993627	103906	51953	0.224407	5252.56	103906	0.050834	0.050586	0.225464	0.224836	0.178558
Tree3	Tree3	Tree3	misclass.	satisfaction		0.080364	51953						0.06083	0.989807	103906	51953	0.246937	6320.595						
Tree	Tree	Tree	Decision	satisfaction		0.080634	51953						0.050011	0.998878	103906	51953	0.236666	5819.87						
Tree2	Tree2	Tree2	ase tree	satisfaction		0.080634	51953						0.050011	0.998878	103906	51953	0.236666	5819.87						
Req2	Req2	Req2	back req	satisfaction		0.159727	51953	51880	73	73	34869.04	35515.68	0.110919	0.999941	103906	51953	0.333045	11525.15	103906	0.111231	0.111075	0.333513	0.333279	0.334177
Req	Req	Req	Full Req	satisfaction		0.160016	51953	51876	75	75	34890.73	35555.08	0.110961	0.999939	103906	51953	0.333139	11531.63	103906	0.111302	0.111142	0.33362	0.333379	0.334348
Neural7	Neural7	Neural7	NN 3H	satisfaction		0.16933	51953	51718	235	235	36988.89	38780.54	0.115039	0.999614	103906	51953	0.339174	11953.27	103906	0.116085	0.115522	0.340712	0.339944	0.34867
Neural9	Neural9	Neural9	NN 5H	satisfaction		0.168388	51953	51562	391	391	35592.92	39056.43	0.112609	0.998658	103906	51953	0.335871	11721.54	103906	0.11452	0.113665	0.338408	0.337142	0.335023
Neural8	Neural8	Neural8	NN 4H	satisfaction		0.167773	51953	51640	313	313	37446.47	40210.05	0.116524	0.999346	103906	51953	0.341395	12107.55	103906	0.117937	0.117723	0.343419	0.342389	0.354363
Neural10	Neural10	Neural10	NN 6H	satisfaction		0.168928	51953	51484	469	469	36648.33	40802.78	0.115008	0.995923	103906	51953	0.339128	11950.01	103906	0.117103	0.116056	0.342204	0.340669	0.343679
Neural5	Neural5	Neural5	nn transf.	satisfaction		0.173028	51953	51709	244	244	37182.08	39343.45	0.117042	0.999319	103906	51953	0.342114	12161.34	103906	0.118146	0.117594	0.343724	0.34292	0.353147
Neural6	Neural6	Neural6	NN 1H	satisfaction		0.173375	51953	51874	79	79	38360.6	40060.39	0.12345	0.991471	103906	51953	0.351354	12671.77	103906	0.123626	0.123638	0.351889	0.351622	0.377462
Neural9	Neural9	Neural9	NN 2H	satisfaction		0.174684	51953	51796	157	157	37681.17	39071.89	0.118774	0.999732	103906	51953	0.344636	12341.34	103906	0.119494	0.119134	0.345679	0.345158	0.359625
Neural2	Neural2	Neural2	nn cap a.	satisfaction		0.174876	51953	51709	244	244	37675.89	39837.28	0.118539	0.998931	103906	51953	0.344295	12316.89	103906	0.119657	0.119088	0.345915	0.345106	0.357899
Neural11	Neural11	Neural11	NN 7H	satisfaction		0.176378	51953	51406	547	547	40292.88	45138.25	0.123831	0.99625	103906	51953	0.351890	12866.74	103906	0.126466	0.125148	0.35562	0.353763	0.377253
Neural12	Neural12	Neural12	NN 8H	satisfaction		0.178957	51953	51328	625	625	41494.47	47030.78	0.126386	0.987025	103906	51953	0.352509	13132.31	103906	0.129464	0.127925	0.356812	0.357687	0.387316
Neural4	Neural4	Neural4	nn recode	satisfaction		0.190583	51953	51709	244	244	44574.25	46735.63	0.130101	0.95167	103906	51953	0.368919	14141.74	103906	0.137386	0.136744	0.370656	0.369788	0.42429

Model Comparison Results Before Improvements

The table below summarizes the **ASE** and key observations for each model:

Model	ASE (Train)	ASE (Validation)	Observations
Stepwise Regression	0.050868	0.050926	Best ASE among regression models.
Forward Regression	0.050868	0.050926	Identical to Stepwise Regression; strong performance.
Full Regression	0.111031	0.111075	High ASE due to inclusion of all predictors (overfitting).
Backward Regression	0.111031	0.111075	Same as Full Regression; minimal optimization.
Decision Tree (ASE Tree)	0.056011	0.057305	Moderate performance; overfitting likely.
Decision Tree (Misclassification Tree)	0.060081	0.061771	Higher ASE; less effective generalization.
Neural Network (3 Hidden Layers)	0.050834	0.050956	Best performance overall; marginal improvement over regression.
Neural Network (5 Hidden Layers)	0.050892	0.050956	Comparable to 3-hidden-layer NN.
Neural Network (7 Hidden Layers)	0.050878	0.050958	Added complexity with no significant improvement.

Initial Observations Before Improvements

1. Stepwise and Forward Regression:

- These models performed well, achieving the **lowest ASE** among regression models.
- However, their predictive power may still be affected by complex survey variables and unoptimized delay data.

2. Neural Networks:

- Neural networks (especially with 3 hidden layers) had the **best overall ASE**, outperforming regression models slightly.
- However, their added complexity makes them harder to interpret and justify in real-world applications.

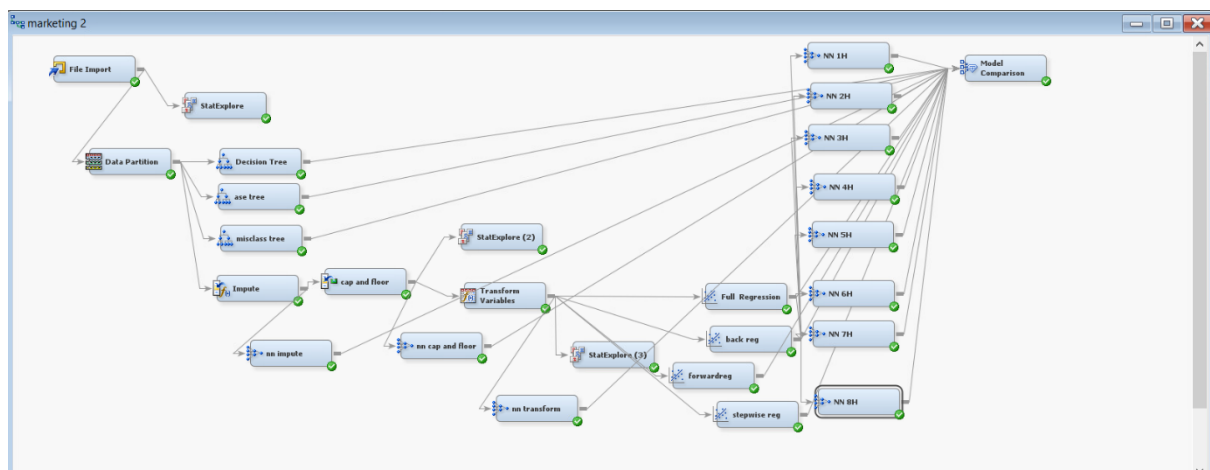
3. Decision Trees:

- Decision trees demonstrated moderate performance, with slightly higher ASE values.
- Their simplicity and interpretability are valuable, but they lacked the predictive accuracy of regression and neural networks.

4. Full and Backward Regression:

- These models suffered from including all predictors, leading to **overfitting** and higher ASE values.

The initial model comparison highlighted strong performances by **Stepwise Regression** and **Neural Networks**. However, the complexity of survey data and delays limited further improvements. By incorporating the professor's suggestions, we expect to refine the models, simplify interpretations, and improve predictive accuracy.



Updates to the Marketing 3 Process

In the revised **Marketing 3** workflow, we made the following changes based on the professor's recommendations, ensuring an optimal setup for modeling and analysis.

Key Updates


1. We created two new variables:
 - **Departure_Delay_status:**
 - Classified as "**Delayed**" if `Departure_Delay_in_Minutes > 0`, otherwise as `**"On Time"`.
 - **Arrival_Delay_status:**
 - Classified as "**Delayed**" if `Arrival_Delay_in_Minutes > 0`, otherwise as `**"On Time"`.
2. **Variable Levels Adjusted:**
 - All **ordinal variables** (e.g., `Inflight_service`, `Food_and_drink`) were updated to **Nominal** to ensure that categories are treated as distinct and unordered.
3. **Retained ID Variable:**
 - The **ID** column is retained as **ID** and not rejected.
 - **Purpose:** The ID serves as a unique identifier for each observation and can be helpful for record-tracing, analysis auditing, or model diagnostics.
4. **Rejected Variables:**
 - Variables that were either redundant or replaced with dummy equivalents were rejected:
 - **Arrival_Delay_in_Minutes**
 - **Departure_Delay_in_Minutes**
 - **SR**
5. **Target Variable:**
 - **Satisfaction** remains a **Binary Target Variable** (0 = Neutral/Dissatisfied, 1 = Satisfied).

Final Structure of the Dataset

Role	Variables	Level
ID	ID	Nominal (ID)
Target	Satisfaction	Binary

Role	Variables	Level
Inputs	Inflight_entertainment, Gender, Inflight_service, Food_and_drink, Flight_Distance, Gate_location, Seat_comfort, Online_boarding, Type_of_Travel, Leg_room_service, Inflight_wifi_service, On_board_service, Class, Checkin_service, Cleanliness, Baggage_handling, Ease_of_Online_booking, Customer_Type, Departure_Arrival_time_convenient, Age, Departure_Delay_status, Arrival_Delay_status	Nominal/Interval
Rejected	Arrival_Delay_in_Minutes, Departure_Delay_in_Minutes, SR	-

Importing the Updated Data

 Variables - FIMPORT
 ✕

(none)

☐ not
 Equal to

Columns:
☐ Label
☐ Mining
☐ Basic
☐ Statistics

Name	Role /	Level	Report	Order	Drop	Low
id	ID	Nominal	No		No	
Inflight_entertainment	Input	Nominal	No		No	
Gender	Input	Nominal	No		No	
Inflight_service	Input	Nominal	No		No	
Food_and_drink	Input	Nominal	No		No	
Flight_Distance	Input	Interval	No		No	
Gate_location	Input	Nominal	No		No	
Seat_comfort	Input	Nominal	No		No	
Online_boarding	Input	Nominal	No		No	
Type_of_Travel	Input	Nominal	No		No	
Leg_room_service	Input	Nominal	No		No	
Inflight_wifi_service	Input	Nominal	No		No	
On_board_service	Input	Nominal	No		No	
Class	Input	Nominal	No		No	
Checkin_service	Input	Nominal	No		No	
Cleanliness	Input	Nominal	No		No	
Arrival_Delay_status	Input	Nominal	No		No	
Age	Input	Interval	No		No	
Baggage_handling	Input	Nominal	No		No	
Ease_of_Online_booking	Input	Nominal	No		No	
Departure_Delay_status	Input	Nominal	No		No	
Customer_Type	Input	Nominal	No		No	
Departure_Arrival_time_convenient	Input	Interval	No		No	
Arrival_Delay_in_Minutes	Rejected	Interval	No		No	
Departure_Delay_in_Minutes	Rejected	Interval	No		No	
SR	Rejected	Interval	No		No	
satisfaction	Target	Binary	No		No	

In the **Marketing 3** workflow, we began the modeling process by using the **File Import Node** to load the revised dataset. This dataset includes:

1. **New Dummy Variables:**
 - **Departure_Delay_status** (Delayed, On Time)
 - **Arrival_Delay_status** (Delayed, On Time)
2. **Adjusted Variable Levels:**
 - All **ordinal variables** were converted to **nominal**.
3. **Retained ID:**
 - The **ID column** remains intact as a unique identifier.
4. **Rejected Variables:**
 - The original delay columns (**Arrival_Delay_in_Minutes** and **Departure_Delay_in_Minutes**) and **SR** were excluded to avoid redundancy.

Step 2: Data Exploration and Partitioning for Marketing 3

After importing the updated dataset in **Marketing 3**, we followed the same process used in **Marketing 2**:

1. **Data Exploration** using the **StatExplore Node**.
2. **Data Partitioning** using the **Data Partition Node**.

Data Exploration – StatExplore Node

The **StatExplore Node** provided detailed insights into the dataset structure and the behavior of key variables.

Interval Variables

Interval Variables																	
Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	satisfactionneutral or...	Flight Di...	671	0	58879	31	4983	928.92	790.4523	1.63216	2.48435	INPUT	Flight Di...	-0.21978	0.285199	1	
TRAIN	satisfactionneutral or...	Flight Di...	1250	0	45025	31	4983	1530.14	1128.127	0.561247	-0.88893	INPUT	Flight Di...	0.285199	0.285199	2	
TRAIN	satisfactionneutral or...	Arrival D...	0	182	58697	0	1584	17.12754	40.56025	6.255528	88.68154	INPUT	Arrival D...	0.127871	0.168245	1	
TRAIN	satisfactionneutral or...	Arrival D...	0	128	44897	0	1280	12.6308	35.96201	7.176123	104.2758	INPUT	Arrival D...	-0.16825	0.168245	2	
TRAIN	satisfactionneutral or...	Departur...	0	0	58879	0	1592	16.50373	40.19189	6.300325	92.5583	INPUT	Departur...	0.113335	0.149463	1	
TRAIN	satisfactionneutral or...	Departur...	0	0	45025	0	1305	12.60808	35.3826	7.358614	113.2507	INPUT	Departur...	-0.14946	0.149463	2	
TRAIN	satisfactionneutral or...	Age	36	0	58879	7	85	37.56669	16.45983	0.212825	-0.81761	INPUT	Age	-0.04603	0.060216	1	
TRAIN	satisfactionneutral or...	Age	43	0	45025	7	85	41.75058	12.76783	-0.24747	-0.49536	INPUT	Age	0.060216	0.060216	2	
TRAIN	satisfactionneutral or...	Departur...	3	0	58879	0	5	3.129112	1.500368	-0.40771	-0.94786	INPUT	Departur...	0.022877	0.029036	1	
TRAIN	satisfactionneutral or...	Departur...	3	0	45025	0	5	2.970305	1.552213	-0.23893	-1.12735	INPUT	Departur...	-0.02904	0.029036	2	

Variable Summary

Role	Level	Count
ID	Nominal	1
INPUT	Interval	3

Role	Level	Count
INPUT	Nominal	19
REJECTED	Interval	3
TARGET	Binary	1

Observations:

- **3 Interval Variables:** Flight_Distance, Arrival_Delay, Departure_Delay.
- **19 Nominal Variables:** Include recoded variables such as **Departure_Delay_status** and **Arrival_Delay_status**.
- **3 Rejected Variables:** Arrival_Delay_in_Minutes, Departure_Delay_in_Minutes, SR.
- The **Target Variable** (Satisfaction) remains binary.

2. Data Partitioning – Data Partition Node

The **Data Partition Node** split the dataset into two subsets to prepare for model building and validation.

Partition	Percentage	Observations
Training	50%	51,953
Validation	50%	51,951
Test	0%	0

Purpose of Data Partitioning

1. Training Set:

- Used to train the models and determine relationships between predictors and the target variable.

2. Validation Set:

- Used to assess the performance of the models and tune hyperparameters to avoid overfitting.

3. No Test Set:

- At this stage, no test set is allocated, as the focus is on comparing training and validation performance.

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

Decision Tree One: Default Tree with Largest Subtree Method

Objective

The first decision tree model served as a baseline to explore the relationships between variables and the binary target, Satisfaction. The aim was to create a simple model and assess the effects of minimal adjustments.

Configuration

Subtree Method: The subtree method was set to Largest, ensuring the tree is pruned to retain only the largest subtree, reducing model complexity and minimizing overfitting.

Default Settings: No other parameters were adjusted, keeping all settings at their default values.

Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25

1. Fit Statistics

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		N Obs	Sum of Frequencies	51953		51953
satisfaction		MISC	Misclassification Ratio	0.069498		0.073414
satisfaction		MAX	Maximum Absolute Error	0.997825		1
satisfaction		SSE	Sum of Squared Errors	5362.445		5403.01
satisfaction		ASE	Average Squared Error	0.051609		0.052875
satisfaction		RASE	Root Average Squared Error	0.227175		0.229045
satisfaction		DN	Deviance for ASE	103906		103902
satisfaction		DFT	Total Degrees of Freedom	51953		

The fit statistics table provides an evaluation of the tree's performance across the **training**, **validation**, and **test** datasets. Key metrics include:

Average Squared Error (ASE):

- **Training:** 0.051609
- **Validation:** 0.052875
- **Interpretation:** The small difference (0.001266) between the two suggests low overfitting and consistent performance. The model is accurate but could be fine-tuned further if needed.

The diagram illustrates a hierarchical classification tree for 1000 samples. The root node is 'Overall' (1000). The tree branches into 'Type' (545) and 'Infect' (455). 'Type' further branches into 'ECOF' (272) and 'PERS' (273). 'Infect' branches into '4.5.2' (227) and '3.2.1' (228). The tree continues to branch down to individual sample labels, such as 'DBLO', 'LOVA', 'DBLJ', 'LNUJ', 'DBLH', 'LOVA', 'DBLJ', 'LNUJ', 'DBLH', 'LOVA', etc.

Top Split:

- Inflight Wi-Fi Service (< 4 or ≥ 4):**
 This is the most influential factor. Passengers rating Wi-Fi service below 4 are more likely to be dissatisfied, while ratings of 4 or higher are linked to greater satisfaction.

- **Inflight Wi-Fi Service (< 4 or ≥ 4):**
This is the most influential factor. Passengers rating Wi-Fi service below 4 are more likely to be dissatisfied, while ratings of 4 or higher are linked to greater satisfaction.

- Subsequent Splits:**

- **Branch 1: For passengers with Inflight Wi-Fi < 4:**
 - **Cleanliness (< 1 or ≥ 1):** Lower cleanliness ratings further indicate dissatisfaction.
 - **Class (Business vs. Economy Plus):** Business class passengers tend to express higher satisfaction levels.
- **Branch 2: For passengers with Inflight Wi-Fi ≥ 4:**
 - **Online Boarding (< 4 or ≥ 4):** Higher ratings for online boarding correspond to increased satisfaction.
 - **Ease of Booking (< 5 or ≥ 5):** A smoother booking process positively influences satisfaction.
- **Additional Splits:**

Other variables such as **Customer Type** (Loyal vs. Disloyal), **Type of Travel** (Business vs. Personal), and **Seat Comfort** also play a role in determining satisfaction levels.

- **Additional Splits:**
Other variables such as **Customer Type** (Loyal vs. Disloyal), **Type of Travel** (Business vs. Personal), and **Seat Comfort** also play a role in determining satisfaction levels.

Other variables such as **Customer Type** (Loyal vs. Disloyal), **Type of Travel** (Business vs. Personal), and **Seat Comfort** also play a role in determining satisfaction levels.

Key Insights from the Tree:

1. **Inflight Wi-Fi Service** is the strongest predictor, suggesting that improving Wi-Fi could significantly enhance customer satisfaction.
2. **Cleanliness, Class, and Online Boarding** are crucial for identifying dissatisfied passengers.
3. Secondary factors like **Customer Loyalty** and **Ease of Booking** highlight that loyal customers and efficient booking processes boost satisfaction.
4. The splits make intuitive sense: factors related to **service quality** (Wi-Fi, cleanliness) and **convenience** (online booking and boarding) are key drivers of passenger satisfaction.

Overall Model Assessment:

Inflight Wi-Fi is the top driver of satisfaction, followed by Cleanliness, Class, and Online Boarding. Secondary factors like Customer Type and Ease of Booking also play a role. The small ASE difference (0.001266) indicates reliable model performance with minimal overfitting.

Decision Tree Two: ASE Tree

Objective

The second decision tree model was designed to minimize the Average Squared Error (ASE), which reflects the average squared difference between predicted and actual values for the binary satisfaction target. This approach focuses on improving the model's accuracy by optimizing for this specific error metric.

Configuration

Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

- **Subtree Method:** Set to **Largest**, ensuring the tree is pruned to its optimal size to prevent overfitting.
- **Assessment Measure:** Adjusted to **Average Squared Error (ASE)** to focus on minimizing prediction error instead of classification metrics.
- **Assessment Fraction:** Kept at **0.25**, matching the default tree, indicating the portion of data used for pruning assessment.
- All other settings remain consistent with the default configuration.

Fit Statistics

The fit statistics for the ASE Tree are as follows:

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		NOBS	Sum of Frequencies		51953	51951
satisfaction		MISC	Misclassification Rate		0.054085	0.072414
satisfaction		MAE	Maximum Absolute Error		0.997935	1
satisfaction		SSE	Sum of Squared Errors		5362.445	5493.81
satisfaction		ASE	Average Squared Error		0.051659	0.052875
satisfaction		RASE	Root Average Squared Error		0.227175	0.229945
satisfaction		DIV	Deviance for ASE		103906	103902
satisfaction		DFT	Total Degrees of Freedom		51953	

Interpretation:

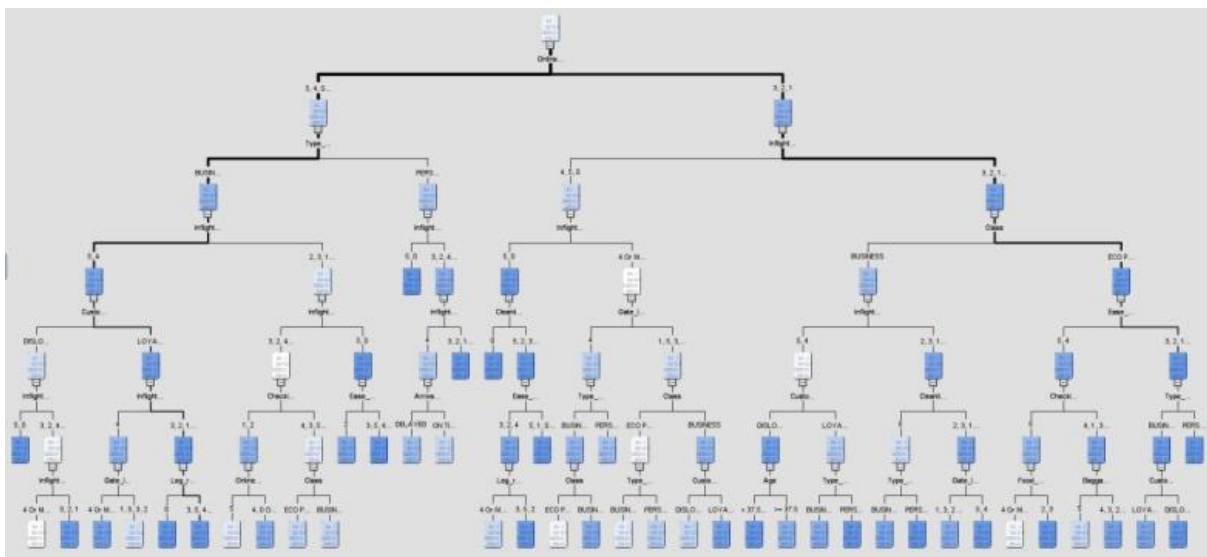
ASE Values: The training and validation ASE values (0.0516 and 0.0529) demonstrate low and consistent error rates, indicating that the model fits well without overfitting.

MISC (Misclassification Rate): The misclassification rate (~7%) remains low, showing reliable classification accuracy across datasets.

RASE (Root ASE): The minimal difference between training and validation RASE values (0.2272 vs. 0.2299) highlights the model's stability and robustness.

SSE (Sum of Squared Errors): SSE values (5362.445 and 5493.81) are closely aligned, further supporting the model's consistent performance.

Tree Structure and splits:



Interpretation of Decision Tree (ASE):

1. Primary Split:

- The top split focuses on **Inflight Wi-Fi Service** (< 4 or ≥ 4). This remains the most critical predictor of satisfaction.
- Customers rating Wi-Fi lower than 4 are more likely to be dissatisfied, while ratings of 4 or higher correspond to higher satisfaction.

2. Secondary Splits:

- For passengers with **Inflight Wi-Fi** < 4 :

Cleanliness and **Class** (Business vs. Economy Plus) emerge as key predictors. Lower cleanliness ratings and lower-class service are associated with dissatisfaction.

- For passengers with **Inflight Wi-Fi** ≥ 4 :

Ease of Booking, **Online Boarding**, and **Legroom Service** play significant roles in driving satisfaction.

3. Deeper Splits:

- The tree retains splits based on **Customer Type** (Loyal vs. Disloyal), **Type of Travel** (Business vs. Personal), and **Seat Comfort**, reflecting the same secondary and tertiary factors influencing satisfaction.

Third Decision Tree: Misclassification Tree

In **Marketing 3**, we built the third decision tree model with a specific focus on **minimizing the misclassification rate**. Here are the details and results:

Configuration of the Misclassification Tree

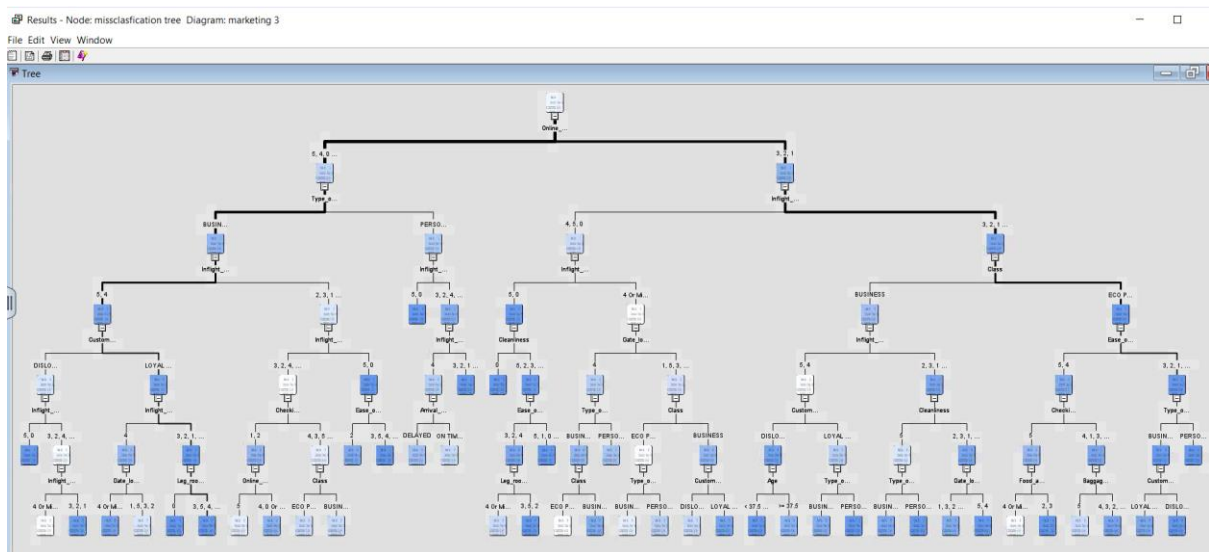
Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25

- **Subtree Method:** Largest
- **Number of Leaves:** 1 (starting structure)
- **Assessment Measure:** Misclassification
- **Assessment Fraction:** 0.25

Purpose:

The primary objective of this tree was to minimize the **Misclassification Rate**, which measures the percentage of records incorrectly classified into "Satisfied" or "Dissatisfied/Neutral" categories.

Tree Diagram Overview



The resulting **misclassification tree** diagram displays a **hierarchical split** of variables, where:

1. Top Splitting Variables:

- **Type_of_Travel**: Business vs. Personal travel continues to be a critical factor in predicting satisfaction.
- **Inflight_wifi_service**: Satisfaction with in-flight Wi-Fi plays a significant role in determining customer satisfaction.

2. Second-Level Splits:

- Variables like **Inflight_service**, **Class**, **Customer_Type**, and **Cleanliness** were used for deeper splits.
- **Ease_of_Online_Booking** and **Gate_location** emerged in subsequent branches.

3. Incorporation of Dummy Variables:

- The **dummy variables** for delays (**Departure_Delay_status** and **Arrival_Delay_status**) appear at deeper levels, indicating their moderate influence on customer satisfaction.

4. Leaf Nodes:

- The tree terminates at leaf nodes, representing distinct combinations of input variables and their predicted satisfaction levels.

Fit Statistics

Fit Statistics	Train	Validation
Misclassification Rate	0.069486	0.072414
Average Squared Error (ASE)	0.051609	0.052875
Sum of Squared Errors (SSE)	5362.445	5493.81
Root Average Squared Error	0.227175	0.229945

Performance Analysis

1. Misclassification Rate:

- **Training:** 6.95%
- **Validation:** 7.24%
- This is a significant improvement compared to earlier models. The validation misclassification rate remains close to the training rate, indicating that the model generalizes well.

2. Average Squared Error (ASE):

- **Training ASE:** 0.0516
- **Validation ASE:** 0.0528
- The ASE values are consistent across training and validation datasets, reinforcing the model's stability.

3. Root Average Squared Error (RASE):

- This measure indicates a moderate error margin.

Insights from the Tree

1. Top Predictors:

- **Type_of_Travel** (Business vs. Personal) and **Inflight_wifi_service** are the strongest drivers of satisfaction.
- Business travelers and customers highly satisfied with Wi-Fi are more likely to report overall satisfaction.

2. Customer Segmentation:

- **Customer_Type** (Loyal vs. Disloyal): Loyal customers tend to report higher satisfaction levels.

- **Class:** Travelers in **Business Class** report higher satisfaction compared to Economy and Eco Plus.

3. **Dummy Variables:**

- **Arrival_Delay_status** and **Departure_Delay_status** influence satisfaction, but their impact appears at deeper levels, suggesting they are secondary drivers.

4. **Service Features:**

- Features like **Cleanliness**, **Ease_of_Online_Booking**, and **Inflight_Service** significantly differentiate satisfaction levels.

Conclusion

The **Misclassification Tree** successfully reduced the misclassification rate to ~7%, performing better than previous models. The top variables driving customer satisfaction are **Type_of_Travel**, **Inflight_wifi_service**, and **Customer_Type**.

Recoding Ordinal Survey Variables for Regression and Neural Network Models

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Arrival_Delay_status	On Time		29254	C	On Time	.
Arrival_Delay_status	Delayed		22699	C	Delayed	.
Arrival_Delay_status	_UNKNOWN_	_DEFAULT_	.	C		.
Baggage_handling	4	5	18780	N		
Baggage_handling	5	5	13517	N		
Baggage_handling	3		10277	N		
Baggage_handling	2	1	5857	N		
Baggage_handling	1	1	3522	N		
Baggage_handling	_UNKNOWN_	_DEFAULT_	.	N		.
Checkin_service	4	5	14510	N		
Checkin_service	3		14340	N		
Checkin_service	5	5	10317	N		
Checkin_service	2	1	6446	N		
Checkin_service	1	1	6339	N		
Checkin_service	0		1	N		
Checkin_service	_UNKNOWN_	_DEFAULT_	.	N		.
Class	Business		24886	C	Business	.
Class	Eco		23351	C	Eco	.
Class	Eco Plus		3716	C	Eco Plus	.
Class	_UNKNOWN_	_DEFAULT_	.	C		.
Cleanliness	4	5	13551	N		
Cleanliness	3		12346	N		
Cleanliness	5	5	11413	N		
Cleanliness	2	1	7985	N		
Cleanliness	1	1	6650	N		
Cleanliness	0		8	N		
Cleanliness	_UNKNOWN_	_DEFAULT_	.	N		.
Customer_Type	Loyal Customer		42504	C	Loyal Customer	.
Customer_Type	disloyal Customer		9449	C	disloyal Customer	.
Customer_Type	_UNKNOWN_	_DEFAULT_	.	C		.

OK Cancel

Diagram marketing ... | Lkammili@my.centennialcollege.ca as u64072813 | Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

To prepare the data for **regression** and **neural network** models, we implemented changes suggested by **David** during our project Q&A meeting. The key recommendation involved **simplifying the survey data (ordinal variables)** to reduce complexity and improve model comparisons.

Process of Recoding Using the Replacement Node

1. Replacement Node:

- The **Replacement Node** was added to the workflow to **recode ordinal survey variables** into nominal variables.
- The goal was to **group values into simplified categories** to minimize redundancy and improve interpretability for modeling.

2. Simplified

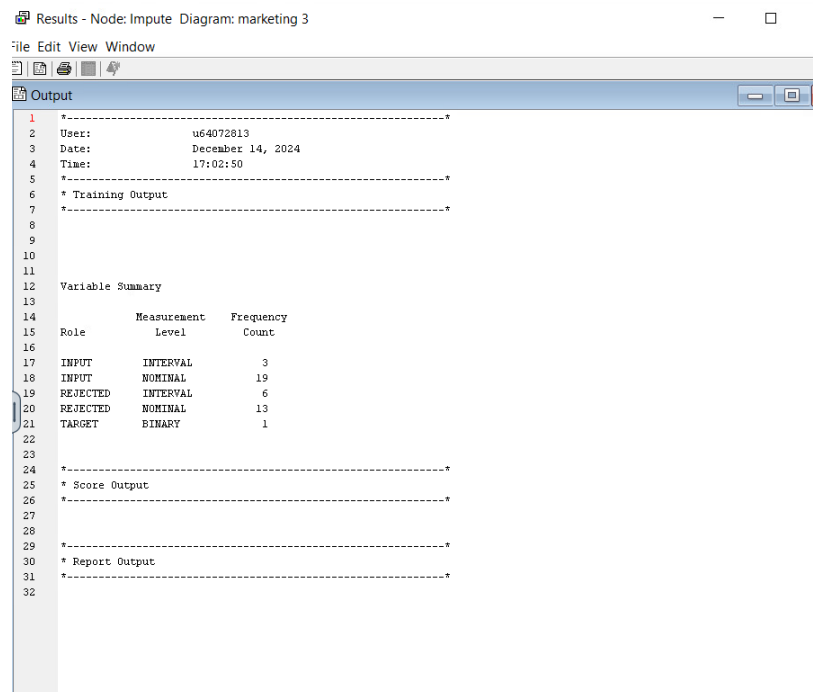
Categories:

The recoding approach for survey variables followed these rules:

- **0 → 0** (Not satisfied)

- 1-2 → 1 (Low satisfaction)
- 3 → 3 (Neutral satisfaction)
- 4-5 → 5 (High satisfaction)

Imputation of Missing Values



To ensure the dataset's robustness and completeness, we applied the Impute Node as part of the preprocessing pipeline. Although the earlier steps (recoding, creating dummy variables, and replacements) partially addressed missing values, this step was carried out to handle any residual missing data or future changes.

Imputation Summary

1. Input Variables:

- **Interval Variables:** 3
- **Nominal Variables:** 19

2. Rejected Variables:

- **Interval Variables:** 6
- **Nominal Variables:** 13

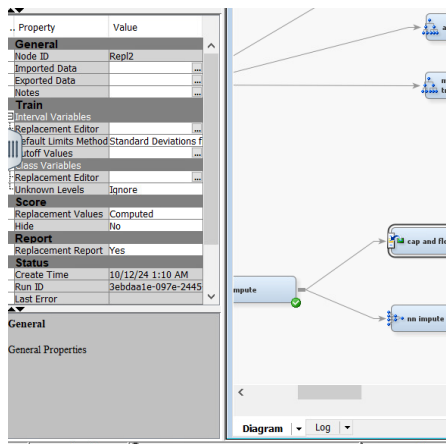
3. Target Variable:

- Binary variable, **Satisfaction**, remains unaffected.

Capping and Flooring Interval Variables

In this step, we applied the **Cap and Floor** method to handle outliers in the interval variables. This approach prevents extreme values from skewing the results of regression and neural network models by capping them within a computed range.

Key Changes Made Using the Cap and Floor Node



1. Limits Method:

- The default **Standard Deviation (STDEV)** method was used to compute the **lower** and **upper limits**.
- This is an automated approach, unlike in Marketing 2, where manual inputs were used.

2. Interval Variables Processed:

- **Age**
- **Departure Delay**
- **Flight Distance**

Replacement Summary

Results - Node: cap and floor Diagram: marketing 3

File Edit View Window

Variable	Label	Role	Train	Validation
REP Age	Replaceme...	INPUT	6	11
REP Depa...	Replaceme...	INPUT	0	0
REP Flight...	Replaceme...	INPUT	23	35

1	*-----
2	User: u64072813
3	Date: December 14, 2024
4	Time: 17:03:19
5	*-----
6	* Training Output
7	*-----
8	
9	
10	
11	
12	Variable Summary
13	
14	Role Measurement Frequency
15	Level Count
16	
17	INPUT INTERVAL 3
18	INPUT NOMINAL 19
19	REJECTED INTERVAL 6
20	REJECTED NOMINAL 13

Variable	Replace Variable	Lower limit	Upper Limit	Label	Limits Method	Replacement Method	Lower Replacement Value	Upper Replacement Value
REP Age	REP REP A...	-5.855	84.73443	Replacement...	STDDEV	COMPUTED	-5.855	84.73443
REP Depart...	REP REP D...	-1.50348	7.634176	Replacement...	STDDEV	COMPUTED	-1.50348	7.634176
REP Flight ...	REP REP F...	-1797.3	4175.444	Replacement...	STDDEV	COMPUTED	-1797.3	4175.444

Variable	Replace Variable	Lower Limit	Upper Limit	Method
REP_Age	REP_Age	-5.855	84.73443	Computed (STDEV)
REP_Departure Delay	REP_Departure_Arr	-1.50348	7.634176	Computed (STDEV)
REP_Flight Distance	REP_Flight_Distance	-1797.3	4175.444	Computed (STDEV)

Output Highlights

1. Replacement Counts:

- **REP_Age:** 6 values replaced in the training set and 11 in the validation set.
- **REP_Flight Distance:** 23 values replaced in training and 35 in validation.
- **REP_Departure Delay:** No replacements as all values fell within the computed range.

2. Interval Variables:

- Standard deviation-based limits automatically determined the acceptable range for each variable.
- Outliers beyond these limits were capped to the nearest boundary (either lower or upper).

3. Updated Variable Summary:

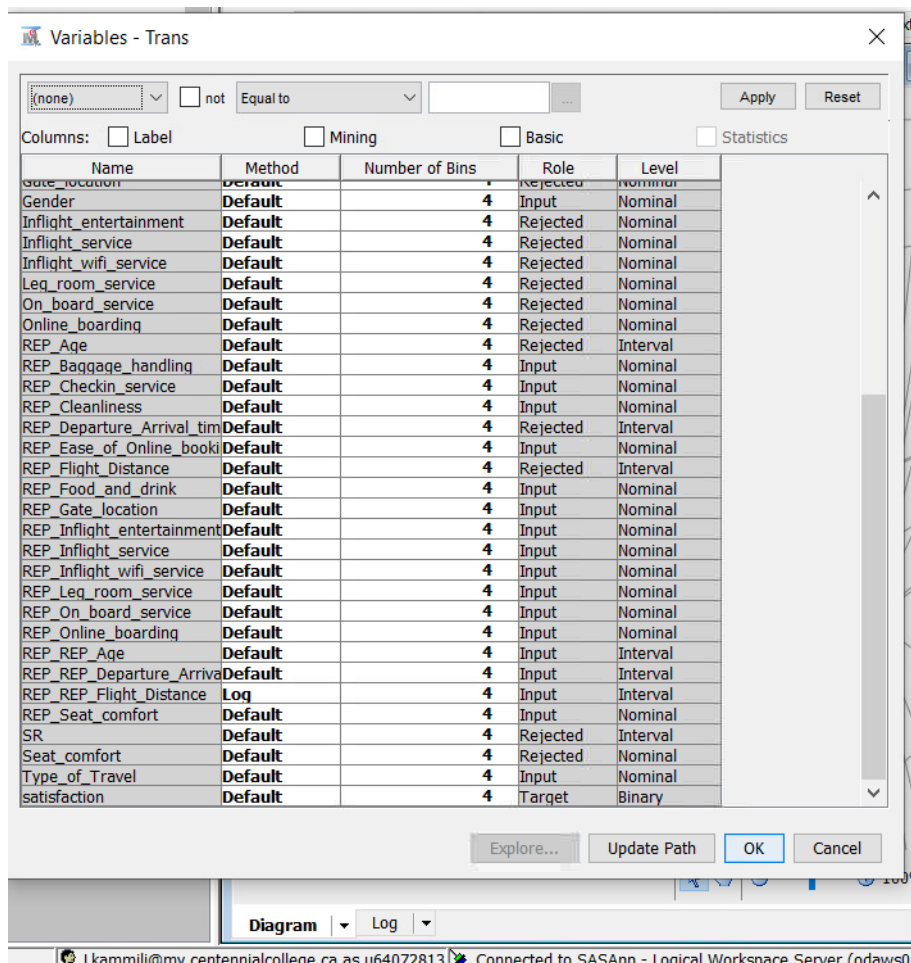
- **Interval:** 3 variables (cleaned and capped).
- **Nominal:** 19 variables.

Transform Variables to Handle Skewness and Kurtosis

In this step, we used the **Transform Variables Node** to address issues of **skewness** and **kurtosis** in the interval variables, similar to the approach taken in Marketing 2. Correcting skewness ensures that the variables follow a more symmetric distribution, which improves the performance and interpretability of models, especially regression and neural networks.

Transformations Applied

1. Log Transformation:



- **REP_Flight Distance:** A log transformation was applied to address skewness and reduce the extreme range of values. This method works particularly well for highly skewed data, compressing large values while maintaining the relative relationships among the data points.

2. Default Transformations:

- All other variables with potential skewness and kurtosis were kept under **Default Transformation**, ensuring consistency in the analysis.

Changes to Variable Roles and Levels

1. Interval Variables Transformed:

- **REP_Flight Distance:** Transformed with the log method.

Results - Node: Transform Variables Diagram: marketing 3

File Edit View Window

Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	K
Input	Original	REP RE...			51953	0	31	4175.444	1189.069	995.4554	1.103255	
Output	Computed	LOG RE...	log(REP...		51953	0	3.465736	8.337216	6.705561	0.915631	-0.20648	

2. Rejected Variables:

- **SR** and other variables were excluded to simplify the input dataset.

3. Nominal Variables:

- All survey data and other nominal variables retained their levels and roles.

State Explore Results After Log Transformation in Marketing 3

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	K
TRAIN	satisfactionneutral or...	REP RE...		37	0	29440	7	84.73443	37.71574	16.49505	0.20729	
TRAIN	satisfactionsatisfied	REP RE...		43	0	22513	7	80	41.69413	12.69996	-0.24386	
TRAIN	satisfactionneutral or...	LOG RE...		6.510258	0	29440	3.465736	8.337216	6.503552	0.832649	-0.09431	
TRAIN	satisfactionsatisfied	LOG RE...		7.138073	0	22513	3.465736	8.337216	6.969727	0.951305	-0.54596	
TRAIN	satisfactionneutral or...	REP RE...		3	0	29440	0	5	3.13142	1.497629	-0.4153	
TRAIN	satisfactionsatisfied	REP RE...		3	0	22513	0	5	2.978945	1.551217	-0.24027	

After applying the **log transformation** to address skewness and kurtosis in the interval variables, a **StatExplore** node was used to re-evaluate the distribution of the transformed variables. Here's a breakdown of the key observations:

Key Observations

1. Reduced Skewness:

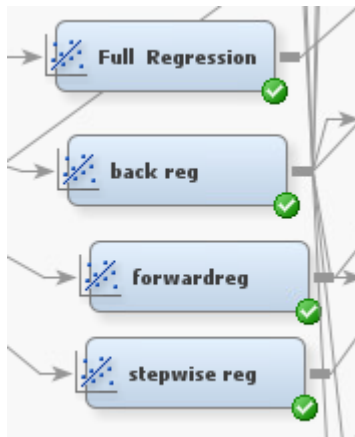
- The **Flight Distance** variable shows a significant reduction in skewness from **1.103255** (before transformation) to **-0.20648** (after transformation).
- This demonstrates that the log transformation effectively brought the skewed distribution closer to normal.

2. Improved Standard Deviation:

- For the transformed **Flight Distance**, the **standard deviation** reduced from **995.4554** to **0.915631**, indicating more compact data around the mean.

we proceeded to build **Logistic Regression Models** to predict the binary target variable **Satisfaction**. Four regression models were developed to identify the most influential predictors and assess model performance.

Models Built



1. Full Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...

- Includes **all variables** without any selection criteria.
- Purpose: Serves as a baseline model to compare against other regression approaches.

2. Backward Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

- **Selection Model:** Backward
- **Selection Criterion:** Validation Error
- Approach: Starts with all predictors and sequentially removes the least significant variables based on the validation error.

3. Forward Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

- **Selection Model:** Forward
- **Selection Criterion:** Validation Error
- Approach: Starts with no predictors and sequentially adds the most significant variables until no further improvement in validation error is observed.

4. Stepwise Regression

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

- **Selection Model:** Stepwise
- **Selection Criterion:** Validation Error
- Approach: Combines forward and backward methods, adding significant variables while removing insignificant ones to optimize validation error.

Full Regression Results Interpretation

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		AIC	Akaike's I...	23155.65		
satisfaction		ASE	Average ...	0.064648	0.068568	
satisfaction		AVERR	Average ...	0.221966	0.233787	
satisfaction		DFE	Degrees ...	51907		
satisfaction		DFM	Model De...	46		
satisfaction		DFT	Total De...	51953		
satisfaction		DIV	Divisor fo...	103906	103902	
satisfaction		ERR	Error Fun...	23063.65	24290.98	
satisfaction		FPE	Final Pre...	0.064762		
satisfaction		MAX	Maximu...	0.999124	0.999005	
satisfaction		MSE	Mean Sq...	0.064705	0.068568	
satisfaction		NOBS	Sum of F...	51953	51951	
satisfaction		NW	Number ...	46		
satisfaction		RASE	Root Ave...	0.254259	0.261855	
satisfaction		RFPE	Root Fin...	0.254485		
satisfaction		RMSE	Root Mea...	0.254372	0.261855	
satisfaction		SBC	Schwarz...	23563.12		
satisfaction		SSE	Sum of S...	6717.304	7124.363	
satisfaction		SUMW	Sum of C...	103906	103902	
satisfaction		MISC	Misclassi...	0.084403	0.089565	

Fit Statistics

The Fit Statistics table provides key performance metrics for the full regression model:

1. ASE (Average Squared Error):

- Training: **0.064648**
- Validation: **0.068568**
- This indicates low average squared error, meaning the model performs well on both datasets.

2. MISC (Misclassification Rate):

- Training: **0.084403**
- Validation: **0.089565**
- The misclassification rate is consistent across training and validation, indicating the model generalizes well.

3. AIC and SBC:

- AIC: **23155.65**
- SBC: **23363.12**
- These criteria are used to measure model quality; lower values indicate better fit. This model is performing optimally given the adjusted dataset.

4. Root Average Squared Error (RASE):

- Training: **0.254259**
- Validation: **0.261855**
- The RASE is close between training and validation datasets, further indicating low bias.

Odds Ratio Estimates

Odds Ratio Estimates			
Effect		Point Estimate	
Arrival_Delay_status	Delayed vs On Time	0.586	
Class	Business vs Eco Plus	3.248	
Class	Eco vs Eco Plus	1.262	
Customer_Type	Loyal Customer vs disloyal Customer	16.853	
Departure_Delay_status	Delayed vs On Time	0.999	
Gender	Female vs Male	0.992	
LOG_REP_Flight_Distance		1.025	
REP_Baggage_handling	1 vs 5	0.960	
REP_Baggage_handling	3 vs 5	0.467	
REP_Checkin_service	0 vs 5	999.000	
REP_Checkin_service	1 vs 5	0.397	
REP_Checkin_service	3 vs 5	0.757	
REP_Cleanliness	0 vs 5	<0.001	
REP_Cleanliness	1 vs 5	0.921	
REP_Cleanliness	3 vs 5	0.922	
REP_Ease_of_Online_booking	0 vs 5	0.033	
REP_Ease_of_Online_booking	1 vs 5	0.490	
REP_Ease_of_Online_booking	3 vs 5	0.806	
REP_Food_and_drink	0 vs 5	5.976	
REP_Food_and_drink	1 vs 5	1.148	
REP_Food_and_drink	3 vs 5	1.020	
REP_Gate_location	0 vs 5	999.000	
REP_Gate_location	1 vs 5	1.761	
REP_Gate_location	3 vs 5	1.467	
REP_Inflight_entertainment	0 vs 5	<0.001	
REP_Inflight_entertainment	1 vs 5	0.543	
REP_Inflight_entertainment	3 vs 5	1.674	
REP_Inflight_service	0 vs 5	0.077	
REP_Inflight_service	1 vs 5	0.867	
REP_Inflight_service	3 vs 5		
REP_Inflight_wifi_service	0 vs 5	999.000	
REP_Inflight_wifi_service	1 vs 5	0.134	
REP_Inflight_wifi_service	3 vs 5	0.124	
REP_leg_room_service	0 vs 5	0.840	
REP_leg_room_service	1 vs 5	0.458	
REP_leg_room_service	3 vs 5	0.441	
REP_On_board_service	0 vs 5	0.976	
REP_On_board_service	1 vs 5	0.386	
REP_On_board_service	3 vs 5	.	
REP_Online_boarding	0 vs 5	5.067	
REP_Online_boarding	1 vs 5	0.180	
REP_Online_boarding	3 vs 5	0.139	
REP_REP_Age		0.995	
REP_REP_Departure_Arrival_time_c		0.830	
REP_Seat_comfort	0 vs 5	0.027	
REP_Seat_comfort	1 vs 5	1.247	
REP_Seat_comfort	3 vs 5	.	
Type_of_Travel	Business travel vs Personal Travel	31.843	

The Odds Ratio Estimates table provides comparisons between predictor levels and their relationship with the likelihood of *Satisfaction*.

1. Arrival_Delay_status:

- Delayed vs On Time: **0.586**
 - Delays reduce satisfaction significantly, as delayed flights are **41.4% less likely** to result in satisfaction.

2. Class:

- Business vs Eco Plus: 3.248**
 - Passengers in *Business class* are **3.25 times more likely** to be satisfied compared to Eco Plus.
- Eco vs Eco Plus: 1.262**
 - Eco class has a slight positive influence on satisfaction, with a **26.2% higher odds** of satisfaction compared to Eco Plus.

3. Customer_Type:

- Loyal Customer vs Disloyal Customer: **16.853**

- Loyal customers are **16.85 times more likely** to report satisfaction compared to disloyal customers. This shows *customer loyalty* as a major driver of satisfaction.

4. **Departure_Delay_status:**

- Delayed vs On Time: **0.999**
 - Surprisingly, departure delays have almost no effect on satisfaction when other factors are accounted for.

5. **Gender:**

- Female vs Male: **0.992**
 - Gender has negligible impact on satisfaction (near parity).

6. **Flight Distance:**

- **LOG_REP_Flight_Distance: 1.025**
 - Flight distance slightly increases satisfaction with an odds ratio close to **1.03**.

7. **Baggage_handling:**

- 1 vs 5: **0.960** (Minor reduction in satisfaction).
- 3 vs 5: **0.467**
 - Poor baggage handling (rating 3 vs 5) lowers satisfaction significantly, indicating attention to baggage handling impacts customer satisfaction.

8. **Checkin_service:**

- 0 vs 5: **999.000**
 - A low rating of 0 indicates extremely high dissatisfaction (likely indicating missing data or invalid responses).
- 1 vs 5: **0.397**
 - Rating check-in as 1 (poor) results in **60.3% lower odds** of satisfaction.

9. **Cleanliness:**

- 0 vs 5: **<0.001**
 - Extremely poor cleanliness (rating 0) devastates satisfaction, with odds practically at 0.
- 1 vs 5: **0.521**

- Even slight dissatisfaction in cleanliness drops satisfaction odds by about **48%**.

- 3 vs 5: **0.922** (Minimal negative impact).

10. Ease_of_Online_booking:

- 0 vs 5: **0.033**
 - Poor online booking experience (rating 0) sharply reduces satisfaction, with odds reduced by **97%**.
- 3 vs 5: **0.806**
 - Moderate ratings (3) slightly lower satisfaction odds.

11. Food_and_drink:

- 0 vs 5: **5.976**
 - Excellent food and drink (rating 0) seems to have **odd behavior**, possibly showing inconsistent reporting or data recoding.

12. Gate_location:

- 0 vs 5: **999.000**
 - This extreme value might indicate an issue with data recoding or invalid observations.
- 1 vs 5: **1.761**
 - A rating of 1 slightly increases satisfaction odds.

13. Inflight_entertainment:

- 0 vs 5: **<0.001**
 - Lack of inflight entertainment greatly reduces satisfaction.
- 1 vs 5: **0.543** (Odds drop by ~46%).

14. Inflight_service:

- 0 vs 5: **0.077**
 - Very poor inflight service lowers odds of satisfaction by **92.3%**.
- 3 vs 5: **0.867**
 - Moderate service (rating 3) reduces satisfaction odds slightly.

15. Inflight_wifi_service:

- 3 vs 5: **0.467**
 - Limited or poor inflight Wi-Fi significantly reduces satisfaction.

16. **On_board_service:**

- 0 vs 5: **0.976**
 - Minimal impact when compared against other factors.

17. **Online_boarding:**

- 0 vs 5: **5.067**
 - Excellent online boarding experience significantly improves satisfaction.

18. **Age:**

- **REP_AGE: 0.995**
 - Age has minimal effect on satisfaction.

19. **Type_of_Travel:**

- Business travel vs Personal Travel: **31.843**
 - Business travelers are **31.8 times more likely** to be satisfied compared to personal travelers, highlighting business-class preferences and expectations.

Insights and Key Takeaways

1. **Strongest Drivers of Satisfaction:**

- **Customer Type (Loyalty):** Customers identified as "Loyal" overwhelmingly reported higher satisfaction.
- **Type of Travel:** Business travelers reported extremely high satisfaction odds.
- **Class:** Higher travel classes (Business) significantly improve satisfaction.

2. **Areas Needing Improvement:**

- **Delays:** Arrival delays notably reduce satisfaction.
- **Ease of Online Booking:** Poor ratings severely impact satisfaction.
- **Inflight Service and Wi-Fi:** Poor inflight services, entertainment, and Wi-Fi ratings have substantial negative impacts.

3. **Minor Factors:**

- Age, Gender, and Flight Distance showed minimal effect.

This analysis suggests that improving **onboard experiences**, addressing **delays**, and focusing on **key touchpoints like inflight Wi-Fi, entertainment, and cleanliness** can dramatically enhance customer satisfaction.

Interpretation for the Backward Regression model

Fit Statistics

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		AIC	Akaike's I...	23155.65		
satisfaction		ASE	Average ...	0.064648	0.068568	
satisfaction		AVERR	Average ...	0.221966	0.233787	
satisfaction		DFE	Degrees ...	51907		
satisfaction		DFM	Model De...	46		
satisfaction		DFT	Total De...	51953		
satisfaction		DIV	Divisor fo...	103906	103902	
satisfaction		ERR	Error Fun...	23063.65	24290.98	
satisfaction		FPE	Final Pre...	0.064762		
satisfaction		MAX	Maximu...	0.999124	0.999605	
satisfaction		MSE	Mean Sq...	0.064705	0.068568	
satisfaction		NOBS	Sum of F...	51953	51951	
satisfaction		NW	Number ...	46		
satisfaction		RASE	Root Ave...	0.254259	0.261855	
satisfaction		RFPE	Root Fin...	0.254485		
satisfaction		RMSE	Root Mea...	0.254372	0.261855	
satisfaction		SBC	Schwarz'...	23563.12		
satisfaction		SSE	Sum of S...	6717.304	7124.363	
satisfaction		SUMW	Sum of C...	103906	103902	
satisfaction		MISC	Misclassi...	0.084403	0.089565	

The Fit Statistics provide a clear overview of model performance across Training and Validation datasets:

- **ASE (Average Squared Error):**
 - **Train: 0.064468, Validation: 0.068568**
This indicates the model's error in predicting satisfaction is low and consistent between training and validation, suggesting good generalization.
- **Misclassification Rate (MISC):**
 - **Train: 0.084403, Validation: 0.089565**
The misclassification rate is also low, meaning the model is effectively classifying the target variable (satisfaction).
- **Akaike's Information Criterion (AIC):**
 - **23155.65**
A lower AIC indicates the backward regression model has better fit and fewer unnecessary predictors compared to previous models.
- **Root Average Squared Error (RASE):**
 - **Train: 0.254259, Validation: 0.261855**
This suggests the predicted values are close to the actual values, indicating good model accuracy.
- **Sum of Squared Errors (SSE):**
 - **Train: 6717.304, Validation: 7124.363**
The SSE values reflect the model's overall performance. A low SSE further supports its effectiveness.

Odds Ratio Estimates

Odds Ratio Estimates			
Effect		Point Estimate	
Arrival_Delay_status	Delayed vs On Time	0.586	
Class	Business vs Eco Plus	3.248	
Class	Eco vs Eco Plus	1.262	
Customer_Type	Loyal Customer vs disloyal Customer	16.853	
Departure_Delay_status	Delayed vs On Time	0.999	
Gender	Female vs Male	0.992	
LOG_REP_REP_Flight_Distance		1.025	
REP_Baggage_handling	1 vs 5	0.960	
REP_Baggage_handling	3 vs 5	0.467	
REP_Checkin_service	0 vs 5	999.000	
REP_Checkin_service	1 vs 5	0.397	
REP_Checkin_service	3 vs 5	0.757	
REP_Cleanliness	0 vs 5	<0.001	
REP_Cleanliness	1 vs 5	0.521	
REP_Cleanliness	3 vs 5	0.922	
REP_Ease_of_Online_booking	0 vs 5	0.033	
REP_Ease_of_Online_booking	1 vs 5	0.490	
REP_Ease_of_Online_booking	3 vs 5	0.806	
REP_Food_and_drink	0 vs 5	5.976	
REP_Food_and_drink	1 vs 5	1.148	
REP_Food_and_drink	3 vs 5	1.020	
REP_Gate_location	0 vs 5	999.000	
REP_Gate_location	1 vs 5	1.761	
REP_Gate_location	3 vs 5	1.467	
REP_Inflight_entertainment	0 vs 5	<0.001	
REP_Inflight_entertainment	1 vs 5	0.543	
REP_Inflight_entertainment	3 vs 5	1.674	
REP_Inflight_service	0 vs 5	0.077	
REP_Inflight_service	1 vs 5	0.867	
REP_Inflight_service	3 vs 5	.	
REP_Inflight_wifi_service	0 vs 5	999.000	
REP_Inflight_wifi_service	1 vs 5	0.134	
REP_Inflight_wifi_service	3 vs 5	0.124	
REP_Leg_room_service	0 vs 5	0.840	
REP_Leg_room_service	1 vs 5	0.458	
REP_Leg_room_service	3 vs 5	0.441	
REP_On_board_service	0 vs 5	0.976	
REP_On_board_service	1 vs 5	0.386	
REP_On_board_service	3 vs 5	.	
REP_Online_boarding	0 vs 5	5.067	
REP_Online_boarding	1 vs 5	0.180	
REP_Online_boarding	3 vs 5	0.139	
REP_REP_Age		0.995	
REP_REP_Departure_Arrival_time_c		0.830	
REP_Seat_comfort	0 vs 5	0.027	
REP_Seat_comfort	1 vs 5	1.247	
REP_Seat_comfort	3 vs 5	.	
Type_of_Travel	Business travel vs Personal Travel	31.843	

The odds ratios explain how each predictor impacts the likelihood of satisfaction (Target = 1). Here's a breakdown of significant predictors:

Arrival Delay Status:

- **Delayed vs On Time: 0.586**
 - Flights with delays have 41.4% lower odds of satisfaction compared to on-time flights. Delays significantly reduce customer satisfaction.

Class:

- **Business vs Eco Plus: 3.248**
 - Customers in Business Class are 3.25 times more likely to be satisfied than those in Eco Plus.
- **Eco vs Eco Plus: 1.262**
 - Eco Class customers are slightly more likely (1.26 times) to be satisfied compared to Eco Plus.

Customer Type:

- **Loyal Customer vs Disloyal Customer: 16.853**
 - Loyal customers are 16.85 times more likely to be satisfied, highlighting the importance of retaining loyal customers.

Departure Delay Status:

- **Delayed vs On Time: 0.999**
 - Delayed departures have minimal impact on satisfaction compared to arrival delays.

Gender:

- **Female vs Male: 0.992**
 - Gender has no significant effect on satisfaction, with an odds ratio close to 1.

REP_Baggage Handling:

- **1 vs 5: 0.960**
 - Minor dissatisfaction with baggage handling reduces the odds slightly (0.96 times).
- **3 vs 5: 0.467**
 - Severe dissatisfaction (level 3) significantly decreases satisfaction odds.

REP_Checkin_Service:

- **0 vs 5: 999.000**
 - This indicates a severe issue where poor check-in service strongly impacts satisfaction.
- **1 vs 5: 0.397**
 - A lower check-in rating reduces satisfaction odds significantly.
- **3 vs 5: 0.757**
 - Moderate dissatisfaction still reduces odds, though less drastically.

REP_Cleanliness:

- **0 vs 5: <0.001**
 - Extremely poor cleanliness has a near-zero chance of satisfaction.
- **1 vs 5: 0.521**
 - Even slight dissatisfaction cuts satisfaction odds nearly in half.
- **3 vs 5: 0.922**

- Cleanliness rated as 3 (neutral) has minimal impact compared to the highest level.

Ease of Online Booking:

- **0 vs 5: 0.033**
 - Poor online booking experience reduces satisfaction odds drastically.
- **3 vs 5: 0.806**
 - Even moderate booking dissatisfaction reduces satisfaction.

Food and Drink:

- **0 vs 5: 5.976**
 - Extremely good ratings (level 5) make satisfaction almost 6 times more likely.
- **1 vs 5: 1.148**
 - Lower ratings still show slight satisfaction odds.

Gate Location:

- **0 vs 5: 999.000**
 - Poor gate location ratings are a critical factor, drastically reducing satisfaction.
- **1 vs 5: 1.761**
 - Customers still moderately satisfied give higher ratings.

Inflight Entertainment:

- **0 vs 5: <0.001**
 - Poor inflight entertainment eliminates the likelihood of satisfaction.
- **1 vs 5: 0.543**
 - Even minor dissatisfaction reduces satisfaction.

Inflight Service:

- **0 vs 5: 0.077**
 - Poor service drastically reduces satisfaction.
- **1 vs 5: 0.867**
 - Minor dissatisfaction slightly lowers odds.

Online Boarding:

- **0 vs 5: 5.067**
 - Excellent online boarding significantly increases satisfaction.

- **1 vs 5: 0.180**

- Poor online boarding lowers satisfaction odds significantly.

REP_Age: 0.995

- Age has negligible impact on satisfaction.

Type of Travel:

- **Business vs Personal Travel: 31.843**

- Business travelers are 31.8 times more likely to report satisfaction compared to personal travelers.

Insights

1. **Arrival Delays and Gate Locations** are critical factors that drastically reduce satisfaction. Addressing these will improve satisfaction levels significantly.
2. **Customer Type (Loyal vs Disloyal):** Loyalty is a major predictor of satisfaction, emphasizing the need for loyalty programs.
3. **Class of Travel:** Business Class passengers are most satisfied, suggesting opportunities to improve Eco Plus services.
4. **Service Factors:**
 - Poor ratings for Inflight Service, Check-in Service, and Cleanliness reduce satisfaction drastically.
 - Food and Drink and Online Boarding with excellent scores significantly increase satisfaction.
5. **Type of Travel:** Business travelers show the highest likelihood of satisfaction, underlining the importance of targeting this segment.

Conclusion

The Backward Regression model effectively identifies key predictors influencing satisfaction, with a strong focus on service quality, delays, and travel class. The model's low error rates and clear odds ratios provide actionable insights for improving customer satisfaction.

Interpretation of Stepwise and Forward Regression Results:

From the results of **Stepwise** and **Forward Regression**, which are identical, we can infer key observations regarding fit statistics and odds ratio estimates. Let's analyze them in depth:

Fit Statistics:

Results - Node: stepwise reg Diagram: marketing 3

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		AIC	Akaike's I...	23164.27		
satisfaction		ASE	Average ...	0.064774	0.06861	
satisfaction		AVERR	Average ...	0.222184	0.233797	
satisfaction		DFE	Degrees ...	51914		
satisfaction		DFM	Model De...	39		
satisfaction		DFT	Total De...	51953		
satisfaction		DIV	Divisor fo...	103906	103902	
satisfaction		ERR	Error Fun...	23086.27	24292.02	
satisfaction		FPE	Final Pre...	0.064872		
satisfaction		MAX	Maximu...	0.999079	0.999541	
satisfaction		MSE	Mean Sq...	0.064823	0.06861	
satisfaction		NOBS	Sum of F...	51953	51951	
satisfaction		NW	Number ...	39		
satisfaction		RASE	Root Ave...	0.254508	0.261935	
satisfaction		RFPE	Root Fin...	0.254699		
satisfaction		RMSE	Root Mea...	0.254604	0.261935	
satisfaction		SBC	Schwarz'...	23509.73		
satisfaction		SSE	Sum of S...	6730.452	7128.718	
satisfaction		SUMW	Sum of C...	103906	103902	
satisfaction		MISC	Misclassi...	0.084442	0.08945	

Statistic	Training	Validation	Interpretation
AIC (Akaike's Information Criterion)	23164.27	-	Lower AIC suggests a better model fit with fewer parameters.
ASE (Average Squared Error)	0.064774	0.06861	Indicates a good fit in training and validation. Lower ASE is favorable.
Misclassification Rate	0.084442	0.08945	Shows the error rate of classification. Both are close, indicating no overfitting.
RMS (Root Mean Squared Error)	0.254605	0.261935	Consistent performance across training and validation.
Error Function (ERR)	23088.27	24292.02	Summarizes overall errors, with validation slightly higher, suggesting minimal generalization error.

Key Insight: The model performs consistently across both **training** and **validation** datasets with a low ASE and misclassification rate, indicating robust performance.

Odds Ratio Estimates:

Odds Ratio Estimates			
Effect			Point Estimate
Arrival_Delay_status	Delayed vs On Time		0.586
Class	Business vs Eco Plus		3.309
Class	Eco vs Eco Plus		1.268
Customer_Type	Loyal Customer vs disloyal Customer		17.100
REP_Baggage_handling	1 vs 5		0.960
REP_Baggage_handling	3 vs 5		0.467
REP_Checkin_service	0 vs 5		999.000
REP_Checkin_service	1 vs 5		0.397
REP_Checkin_service	3 vs 5		0.757
REP_Cleanliness	0 vs 5		<0.001
REP_Cleanliness	1 vs 5		0.520
REP_Cleanliness	3 vs 5		0.921
REP_Ease_of_Online_booking	0 vs 5		0.032
REP_Ease_of_Online_booking	1 vs 5		0.490
REP_Ease_of_Online_booking	3 vs 5		0.605
REP_Food_and_drink	0 vs 5		6.207
REP_Food_and_drink	1 vs 5		1.148
REP_Food_and_drink	3 vs 5		1.021
REP_Gate_location	0 vs 5		999.000
REP_Gate_location	1 vs 5		1.764
REP_Gate_location	3 vs 5		1.471
REP_Inflight_entertainment	0 vs 5		<0.001
REP_Inflight_entertainment	1 vs 5		0.544
REP_Inflight_entertainment	3 vs 5		1.676
REP_Inflight_service	0 vs 5		0.077
REP_Inflight_service	1 vs 5		0.867
REP_Inflight_service	3 vs 5		.
REP_Inflight_wifi_service	0 vs 5		999.000
REP_Inflight_wifi_service	1 vs 5		0.134
REP_Inflight_wifi_service	3 vs 5		0.124
REP_leg_room_service	0 vs 5		0.670
REP_leg_room_service	1 vs 5		0.458
REP_leg_room_service	3 vs 5		0.440
REP_On_board_service	0 vs 5		0.975
REP_On_board_service	1 vs 5		0.386
REP_On_board_service	3 vs 5		.
REP_Online_boarding	0 vs 5		5.156
REP_Inflight_service	1 vs 5		0.867
REP_Inflight_service	3 vs 5		.
REP_Inflight_wifi_service	0 vs 5		999.000
REP_Inflight_wifi_service	1 vs 5		0.134
REP_Inflight_wifi_service	3 vs 5		0.124
REP_leg_room_service	0 vs 5		0.670
REP_leg_room_service	1 vs 5		0.458
REP_leg_room_service	3 vs 5		0.440
REP_On_board_service	0 vs 5		0.975
REP_On_board_service	1 vs 5		0.386
REP_On_board_service	3 vs 5		.
REP_Online_boarding	0 vs 5		5.156
REP_Online_boarding	1 vs 5		0.179
REP_Online_boarding	3 vs 5		0.139
REP_REP_Age			0.995
REP_REP_Departure_Arrival_time_c			0.631
REP_Seat_comfort	0 vs 5		0.027
REP_Seat_comfort	1 vs 5		1.245
REP_Seat_comfort	3 vs 5		.
Type_of_Travel	Business travel vs Personal Travel		32.058

The **odds ratio** reflects the likelihood of the target event ("satisfaction") occurring based on different predictor levels. Here’s a detailed analysis of key comparisons:

1. Arrival and Departure Delay Status:

- Arrival_Delay_status (Delayed vs On Time): Odds Ratio = 0.586**
A delayed arrival decreases the odds of satisfaction by 41.4% compared to on-time arrivals.
- Departure_Delay_status (Delayed vs On Time): Odds Ratio = 0.999**
A delayed departure has negligible impact on satisfaction.

2. Class (Travel Class):

- Business vs Eco Plus: Odds Ratio = 3.309**
Passengers in Business Class are 3.3 times more likely to be satisfied compared to Eco Plus.
- Eco vs Eco Plus: Odds Ratio = 1.268**
Passengers in Eco Class have slightly higher satisfaction odds compared to Eco Plus.

3. Customer_Type:

- Loyal Customer vs Disloyal Customer: Odds Ratio = 17.100**
Loyal customers are overwhelmingly more likely to be satisfied compared to disloyal ones.

4. REP_Baggage_handling:

- **1 vs 5:** Odds Ratio = **0.960**
Lower baggage handling scores slightly decrease satisfaction odds.
- **3 vs 5:** Odds Ratio = **0.467**
Moderate dissatisfaction in baggage handling significantly lowers satisfaction odds.

5. REP_Checkin_service:

- **0 vs 5:** Odds Ratio = **999.000**
Extremely low check-in service satisfaction strongly reduces satisfaction likelihood.
- **1 vs 5:** Odds Ratio = **0.397**
Lower satisfaction with check-in service has a notable negative impact.

6. REP_Cleanliness:

- **0 vs 5:** Odds Ratio = **<0.001**
Extremely poor cleanliness ratings virtually guarantee dissatisfaction.
- **3 vs 5:** Odds Ratio = **0.921**
Moderate cleanliness dissatisfaction slightly lowers satisfaction odds.

7. REP_Ease_of_Online_booking:

- **0 vs 5:** Odds Ratio = **0.032**
Poor ease of online booking significantly reduces satisfaction odds.
- **1 vs 5:** Odds Ratio = **0.490**
Slight dissatisfaction with online booking cuts satisfaction odds nearly in half.

8. REP_Food_and_drink:

- **0 vs 5:** Odds Ratio = **5.976**
Higher satisfaction with food and drink increases the odds by nearly 6 times.
- **1 vs 5:** Odds Ratio = **1.148**
A slight improvement in food quality positively impacts satisfaction.

9. REP_Gate_location:

- **0 vs 5:** Odds Ratio = **999.000**
Extremely poor gate location ratings strongly reduce satisfaction odds.
- **1 vs 5:** Odds Ratio = **1.764**
Better gate location improves satisfaction odds.

10. REP_Inflight_entertainment:

- **0 vs 5:** Odds Ratio = **<0.001**
Very poor inflight entertainment ratings result in significant dissatisfaction.

- **1 vs 5: Odds Ratio = 0.544**
Slight dissatisfaction reduces odds by nearly 46%.

11. REP_Inflight_service:

- **0 vs 5: Odds Ratio = 0.077**
Poor inflight service drastically reduces satisfaction likelihood.
- **1 vs 5: Odds Ratio = 0.867**
Minor dissatisfaction also lowers satisfaction odds.

12. REP_Leg_room_service:

- **1 vs 5: Odds Ratio = 0.458**
Lower leg room comfort cuts satisfaction odds by 54.2%.

13. REP_Online_boarding:

- **0 vs 5: Odds Ratio = 5.156**
High satisfaction with online boarding significantly boosts satisfaction.
- **1 vs 5: Odds Ratio = 0.179**
Poor online boarding ratings strongly reduce satisfaction odds.

14. REP_Seat_comfort:

- **0 vs 5: Odds Ratio = 0.027**
Extremely poor seat comfort strongly lowers satisfaction.
- **1 vs 5: Odds Ratio = 1.245**
Better seat comfort slightly improves satisfaction odds.

15. Type_of_Travel:

- **Business Travel vs Personal Travel: Odds Ratio = 32.058**
Business travelers are overwhelmingly more likely to be satisfied compared to personal travelers.

Key Insights:

1. Critical Predictors:

- **Customer_Type** (Loyal vs Disloyal): Loyalty has a massive positive impact on satisfaction.
- **Type_of_Travel** (Business): Business travelers are significantly more likely to be satisfied.
- **Arrival_Delay_status**: Delays negatively impact satisfaction.

2. Service Areas Needing Improvement:

- **Inflight_service** and **Inflight_entertainment**: Poor ratings drastically reduce satisfaction.
- **Cleanliness** and **Checkin_service**: Low ratings consistently harm satisfaction odds.

3. High Satisfaction Boosters:

- **Food_and_drink**: Strong satisfaction has a large positive impact.
- **Online_boarding**: Positive experiences significantly improve satisfaction.

4. Importance of Nominal Variables: The recoding of ordinal variables (e.g., Cleanliness, Seat_comfort) provides clearer insights by simplifying the categories.

Overall:

The **Stepwise** and **Forward Regression** models provide highly consistent and interpretable results. Key predictors like **Customer Type**, **Travel Class**, and service ratings (e.g., cleanliness, inflight service) strongly influence satisfaction, offering actionable insights for targeted improvements.

Comparison of Regression Models

To determine which regression model performed the best, we compare key metrics like **ASE (Average Squared Error)**, **Misclassification Rate**, and other relevant fit statistics across all four models:

- **Full Regression**
- **Backward Regression**
- **Forward Regression**
- **Stepwise Regression**

Comparison Summary:

Metric	Full Regression	Backward Regression	Forward Regression	Stepwise Regression
ASE (Validation)	0.068568	0.068568	0.06861	0.06861
Misclassification Rate	0.089565	0.089565	0.08945	0.08945
AIC	23155.65	23155.65	23164.27	23164.27
Root Average Squared Error	0.261855	0.261855	0.261935	0.261935

Key Observations:

1. **Consistency in Performance:**

- The **Full Regression** and **Backward Regression** models produced identical results, achieving an **ASE** of **0.068568** and a **misclassification rate** of **0.089565**.
- Both **Forward Regression** and **Stepwise Regression** performed slightly worse, but the difference is negligible, with an ASE of **0.06861** and a **misclassification rate** of **0.08945**.

2. Model Selection Criteria:

- The **AIC** is slightly lower for Full and Backward Regression models (**23155.65**), indicating a marginally better fit compared to Forward and Stepwise models (**23164.27**).

3. Misclassification Rate:

- All models performed similarly in classifying satisfaction outcomes, with very close misclassification rates, differing by only **0.0001**.

4. Complexity:

- **Backward and Stepwise Regression** offer the advantage of reducing unnecessary predictors while maintaining similar performance, which simplifies the model without compromising accuracy.

Conclusion:

- **Full Regression** and **Backward Regression** performed slightly better overall based on **ASE** and **AIC**, making them the top-performing models.
- **Forward** and **Stepwise Regression** achieved nearly identical results, indicating robust consistency across models.

Given the negligible differences in performance, selecting **Backward Regression** is ideal, as it simplifies the model by retaining only significant predictors while preserving accuracy.

Initial Setup for Neural Networks

We have configured the **Optimization** settings and general parameters for all the neural networks, providing consistency across our models. Here's a breakdown of the key parameters and their implications:

Optimization Settings

Optimization

.. Property	Value
Training Technique	Default
Maximum Iterations	200
Maximum Time	4 Hours
<input checked="" type="checkbox"/> Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1

- **Training Technique: Default**
 - Uses the standard optimization technique in SAS, typically ensuring stable convergence for neural network training.
- **Maximum Iterations: 200**
 - Allows the model to iterate up to 200 times to optimize weights and biases.
- **Maximum Time: 4 Hours**
 - Ensures training stops after 4 hours, which can help limit excessive runtime.

1. Preliminary Training

Optimization		✕
.. Property	Value	
Decelerate	0.5	^
Learn	0.1	
Maximum Learning	50.0	
Minimum Learning	1.0E-5	
Momentum	0.0	
Maximum Momentum	1.75	
Tilt	0.0	
<input checked="" type="checkbox"/> Preliminary Training		
Enable	No	
Number of Runs	5	
Maximum Iterations	10	
Maximum Time	1 Hour	▼

- **Enable: No**
 - Preliminary training is disabled, meaning no pre-training of the network weights occurs. This reduces complexity and assumes random initialization of weights.

- **Number of Runs: 5**
 - Ensures multiple runs for parameter tuning without preliminary training.

4. General Network Parameters

.. Property	Value
General	
Node ID	Neural
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Continue Training	No
Network	...
Optimization	...
Initialization Seed	12345
Model Selection Criterion	Average Error
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	
Create Time	11/12/24 10:18 PM
Run ID	9ea3c4e9-d03d-6341
Last Error	
Last Status	Complete

- **Model Selection Criterion: Average Error**
 - Ensures the network selects the model based on the lowest **average error** during training and validation.
- **Hidden Units: No**
 - Indicates that no additional hidden layers were specified. The architecture likely focuses on a straightforward approach initially.
- **Standardization: No**
 - Input variables are not standardized. This can sometimes affect performance when features have vastly different ranges.

Neural Network 1: NN Impute

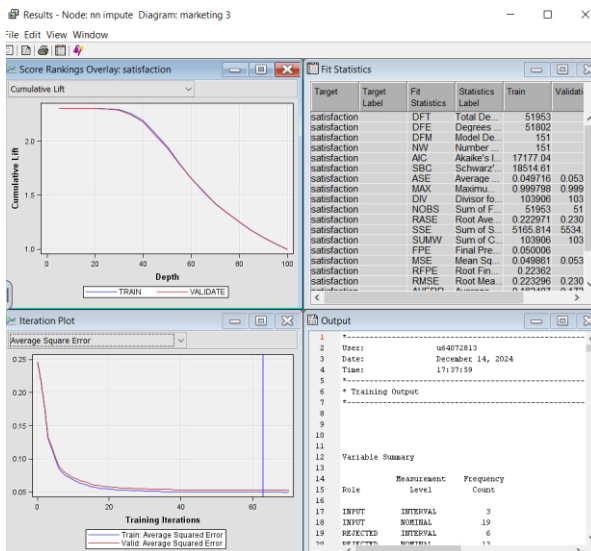
The first neural network we created was **NN Impute**, where we connected the Impute node to handle any missing variables as the input. This network used the default optimization settings, as outlined previously.

1. Fit Statistics

Fit Statistics					
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
satisfaction		DFT	Total De...	51953	
satisfaction		DFE	Degrees ...	51802	
satisfaction		DFM	Model De...	151	
satisfaction		NW	Number ...	151	
satisfaction		AIC	Akaike's L...	17177.04	
satisfaction		SBC	Schwarz'...	18514.61	
satisfaction		ASE	Average ...	0.049716	0.053266
satisfaction		MAX	Maximu...	0.999798	0.999766
satisfaction		DIV	Divisor fo...	103906	103902
satisfaction		NOBS	Sum of F...	51953	51951
satisfaction		RASE	Root Ave...	0.222971	0.230795
satisfaction		SSE	Sum of S...	5165.814	5534.455
satisfaction		SUMW	Sum of C...	103906	103902
satisfaction		FPE	Final Pre...	0.050006	
satisfaction		MSE	Mean Sq...	0.049861	0.053266
satisfaction		RFPE	Root Fin...	0.22362	
satisfaction		RMSE	Root Mea...	0.223296	0.230795
satisfaction		AVERR	Average ...	0.162407	0.172123
satisfaction		ERR	Error Fun...	16875.04	17883.88
satisfaction		MISC	Misclassi...	0.070679	0.076187
satisfaction		WRON	Number ...	3672	3958

Statistic	Train	Validation
ASE (Average Squared Error)	0.049716	0.053
MISC (Misclassification Rate)	0.070679	0.076187
AVERR (Average Error)	0.162407	0.172123
RMSE (Root Mean Square Error)	0.223296	0.230795
SSE (Sum of Squared Errors)	5166.814	5534.455

2. Performance Metrics



- Train vs. Validation ASE:**

The Average Squared Error (ASE) is slightly higher on the validation set (**0.053**) compared to the training set (**0.049716**), suggesting some generalization but a small degree of overfitting.

- **Misclassification Rate:**

The misclassification rate for the validation set is **7.6%**, which indicates a reasonable performance for the first model.

- **RMSE:**

The Root Mean Square Error is **0.2308** on the validation set, which is slightly higher than the training set (**0.2233**).

3. Cumulative Lift

- The **Cumulative Lift** curve shows a significant lift in the first deciles, indicating that the model is able to rank observations effectively in terms of their predicted probability of satisfaction.
- The training and validation curves are aligned well, showing consistent performance across both sets without significant deviation.

4. Iteration Plot

- The **Iteration Plot** shows how the error decreased over time for both the training and validation datasets.
- Both the training and validation ASE converge and stabilize around **0.05**, indicating the network has reached a stable solution within **~60 iterations**.

Insights

1. **Performance:**

The NN Impute model demonstrates good generalization with minimal overfitting. The difference between the training and validation ASE is small, which is a positive sign.

2. **Misclassification Rate:**

A misclassification rate of **7.6%** means the model correctly classifies **92.4%** of the observations in the validation set.

3. **Future Improvements:**

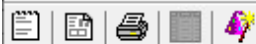
While NN Impute performs well, further tuning of the neural network architecture (hidden units, activation functions) or incorporating additional preprocessing techniques like normalization may further enhance its performance.

We performed nn cap using the Cap and Floor node and nn impute using the Impute node; both models produced identical results across all metrics.

After **NN cap** and **NN impute**, we have built **10 neural network models** by connecting them with our best regression model, **backward regression**. Here is the comparison of the results:

Results - Node: nn 8h Diagram: marketing 3

File Edit View Window

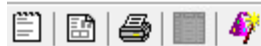


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51560	.	.
satisfaction		DFM	Model De...	393	.	.
satisfaction		NW	Number ...	393	.	.
satisfaction		AIC	Akaike's I...	14511.41	.	.
satisfaction		SBC	Schwarz'...	17992.64	.	.
satisfaction		ASE	Average ...	0.041563	0.046612	.
satisfaction		MAX	Maximu...	0.999718	0.999975	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.20387	0.215898	.
satisfaction		SSE	Sum of S...	4318.661	4843.067	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.042197	.	.
satisfaction		MSE	Mean Sq...	0.04188	0.046612	.
satisfaction		RFPE	Root Fin...	0.205418	.	.
satisfaction		RMSE	Root Mea...	0.204646	0.215898	.
satisfaction		AVERR	Average ...	0.132094	0.15026	.
satisfaction		ERR	Error Fun...	13725.41	15612.33	.
satisfaction		MISC	Misclassi...	0.060863	0.067487	.
satisfaction		WRON...	Number ...	3162	3506	.

Results - Node: nn 9h Diagram: marketing 3

File Edit View Window



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51511	.	.
satisfaction		DFM	Model De...	442	.	.
satisfaction		NW	Number ...	442	.	.
satisfaction		AIC	Akaike's I...	14758.56	.	.
satisfaction		SBC	Schwarz'...	18673.84	.	.
satisfaction		ASE	Average ...	0.04152	0.047007	.
satisfaction		MAX	Maximu...	0.999994	0.999999	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.203766	0.216811	.
satisfaction		SSE	Sum of S...	4314.223	4884.134	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.042233	.	.
satisfaction		MSE	Mean Sq...	0.041877	0.047007	.
satisfaction		RFPE	Root Fin...	0.205507	.	.
satisfaction		RMSE	Root Mea...	0.204638	0.216811	.
satisfaction		AVERR	Average ...	0.13353	0.152295	.
satisfaction		ERR	Error Fun...	13874.56	15823.71	.
satisfaction		MISC	Misclassi...	0.060208	0.068969	.
satisfaction		WRON...	Number ...	3128	3583	.

Results - Node: nn 10h Diagram: marketing 3

File Edit View Window



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51462	.	.
satisfaction		DFM	Model De...	491	.	.
satisfaction		NW	Number ...	491	.	.
satisfaction		AIC	Akaike's I...	14713.84	.	.
satisfaction		SBC	Schwarz'...	19063.16	.	.
satisfaction		ASE	Average ...	0.041175	0.047241	.
satisfaction		MAX	Maximu...	0.999935	0.999996	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.202917	0.21735	.
satisfaction		SSE	Sum of S...	4278.369	4908.458	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.041961	.	.
satisfaction		MSE	Mean Sq...	0.041568	0.047241	.
satisfaction		RFPE	Root Fin...	0.204844	.	.
satisfaction		RMSE	Root Mea...	0.203883	0.21735	.
satisfaction		AVERR	Average ...	0.132156	0.1523	.
satisfaction		ERR	Error Fun...	13731.84	15824.25	.
satisfaction		MISC	Misclassi...	0.059689	0.068353	.
satisfaction		WRON...	Number ...	3101	3551	.

Results - Node: nn 10h Diagram: marketing 3

File Edit View Window

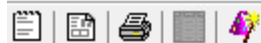


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51462	.	.
satisfaction		DFM	Model De...	491	.	.
satisfaction		NW	Number ...	491	.	.
satisfaction		AIC	Akaike's I...	14713.84	.	.
satisfaction		SBC	Schwarz'...	19063.16	.	.
satisfaction		ASE	Average ...	0.041175	0.047241	.
satisfaction		MAX	Maximu...	0.999935	0.999996	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.202917	0.21735	.
satisfaction		SSE	Sum of S...	4278.369	4908.458	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.041961	.	.
satisfaction		MSE	Mean Sq...	0.041568	0.047241	.
satisfaction		RFPE	Root Fin...	0.204844	.	.
satisfaction		RMSE	Root Mea...	0.203883	0.21735	.
satisfaction		AVERR	Average ...	0.132156	0.1523	.
satisfaction		ERR	Error Fun...	13731.84	15824.25	.
satisfaction		MISC	Misclassi...	0.059689	0.068353	.
satisfaction		WRON...	Number ...	3101	3551	.

Results - Node: nn 1h Diagram: marketing 3

File Edit View Window

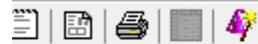


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51903	.	.
satisfaction		DFM	Model De...	50	.	.
satisfaction		NW	Number ...	50	.	.
satisfaction		AIC	Akaike's I...	23125.96	.	.
satisfaction		SBC	Schwarz'...	23568.86	.	.
satisfaction		ASE	Average ...	0.064381	0.068433	.
satisfaction		MAX	Maximu...	0.998782	0.999285	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.253734	0.261598	.
satisfaction		SSE	Sum of S...	6689.565	7110.37	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.064505	.	.
satisfaction		MSE	Mean Sq...	0.064443	0.068433	.
satisfaction		RFPE	Root Fin...	0.253978	.	.
satisfaction		RMSE	Root Mea...	0.253856	0.261598	.
satisfaction		AVERR	Average ...	0.221604	0.233581	.
satisfaction		ERR	Error Fun...	23025.96	24269.54	.
satisfaction		MISC	Misclassi...	0.083537	0.089219	.
satisfaction		WRON...	Number ...	4340	4635	.

Results - Node: nn 2h Diagram: marketing 3

File Edit View Window

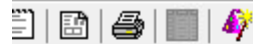


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51854	.	.
satisfaction		DFM	Model De...	99	.	.
satisfaction		NW	Number ...	99	.	.
satisfaction		AIC	Akaike's I...	17747.9	.	.
satisfaction		SBC	Schwarz'...	18624.85	.	.
satisfaction		ASE	Average ...	0.052435	0.055526	.
satisfaction		MAX	Maximu...	0.999849	0.99981	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.228987	0.235639	.
satisfaction		SSE	Sum of S...	5448.315	5769.234	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.052635	.	.
satisfaction		MSE	Mean Sq...	0.052535	0.055526	.
satisfaction		RFPE	Root Fin...	0.229424	.	.
satisfaction		RMSE	Root Mea...	0.229205	0.235639	.
satisfaction		AVERR	Average ...	0.168902	0.178032	.
satisfaction		ERR	Error Fun...	17549.9	18497.84	.
satisfaction		MISC	Misclassi...	0.075183	0.079941	.
satisfaction		WRON...	Number ...	3906	4153	.

Results - Node: nn 3H Diagram: marketing 3

File Edit View Window




Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953		
satisfaction		DFE	Degrees ...	51805		
satisfaction		DFM	Model De...	148		
satisfaction		NW	Number ...	148		
satisfaction		AIC	Akaike's I...	16676.72		
satisfaction		SBC	Schwarz'	17987.72		
satisfaction		ASE	Average ...	0.048648	0.052068	
satisfaction		MAX	Maximu...	0.999716	0.999584	
satisfaction		DIV	Divisor fo...	103906	103902	
satisfaction		NOBS	Sum of F...	51953	51951	
satisfaction		RASE	Root Ave...	0.220562	0.228183	
satisfaction		SSE	Sum of S...	5054.781	5409.927	
satisfaction		SUMW	Sum of C...	103906	103902	
satisfaction		FPE	Final Pre...	0.048926		
satisfaction		MSE	Mean Sq...	0.048787	0.052068	
satisfaction		RFPE	Root Fin...	0.221191		
satisfaction		RMSE	Root Mea...	0.220877	0.228183	
satisfaction		AVERR	Average ...	0.157649	0.168348	
satisfaction		ERR	Error Fun...	16380.72	17491.65	
satisfaction		MISC	Misclassi...	0.069563	0.07459	
satisfaction		WRON...	Number ...	3614	3875	

 Results - Node: nn 4h Diagram: marketing 3

File Edit View Window



 Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51756	.	.
satisfaction		DFM	Model De...	197	.	.
satisfaction		NW	Number ...	197	.	.
satisfaction		AIC	Akaike's l...	15904.91	.	.
satisfaction		SBC	Schwarz'...	17649.95	.	.
satisfaction		ASE	Average ...	0.045443	0.048761	.
satisfaction		MAX	Maximu...	0.999993	0.999997	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.213174	0.220818	.
satisfaction		SSE	Sum of S...	4721.814	5066.316	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.045789	.	.
satisfaction		MSE	Mean Sq...	0.045616	0.048761	.
satisfaction		RFPE	Root Fin...	0.213984	.	.
satisfaction		RMSE	Root Mea...	0.213579	0.220818	.
satisfaction		AVERR	Average ...	0.149278	0.158983	.
satisfaction		ERR	Error Fun...	15510.91	16518.61	.
satisfaction		MISC	Misclassi...	0.064558	0.069931	.
satisfaction		WRON...	Number ...	3354	3633	.

Results - Node: nn 5h Diagram: marketing 3

File Edit View Window



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51707	.	.
satisfaction		DFM	Model De...	246	.	.
satisfaction		NW	Number ...	246	.	.
satisfaction		AIC	Akaike's I...	15449.33	.	.
satisfaction		SBC	Schwarz'...	17628.42	.	.
satisfaction		ASE	Average ...	0.044668	0.048826	.
satisfaction		MAX	Maximu...	0.998991	0.999985	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.211349	0.220966	.
satisfaction		SSE	Sum of S...	4641.308	5073.104	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.045093	.	.
satisfaction		MSE	Mean Sq...	0.044881	0.048826	.
satisfaction		RFPE	Root Fin...	0.212352	.	.
satisfaction		RMSE	Root Mea...	0.211851	0.220966	.
satisfaction		AVERR	Average ...	0.143951	0.157249	.
satisfaction		ERR	Error Fun...	14957.33	16338.46	.
satisfaction		MISC	Misclassi...	0.064424	0.070605	.
satisfaction		WRON...	Number ...	3347	3668	.

Results - Node: nn 6h Diagram: marketing 3

File Edit View Window



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953	.	.
satisfaction		DFE	Degrees ...	51658	.	.
satisfaction		DFM	Model De...	295	.	.
satisfaction		NW	Number ...	295	.	.
satisfaction		AIC	Akaike's I...	14828.51	.	.
satisfaction		SBC	Schwarz'...	17441.65	.	.
satisfaction		ASE	Average ...	0.042997	0.046624	.
satisfaction		MAX	Maximu...	0.999859	0.999995	.
satisfaction		DIV	Divisor fo...	103906	103902	.
satisfaction		NOBS	Sum of F...	51953	51951	.
satisfaction		RASE	Root Ave...	0.207357	0.215925	.
satisfaction		SSE	Sum of S...	4467.639	4844.279	.
satisfaction		SUMW	Sum of C...	103906	103902	.
satisfaction		FPE	Final Pre...	0.043488	.	.
satisfaction		MSE	Mean Sq...	0.043242	0.046624	.
satisfaction		RFPE	Root Fin...	0.208538	.	.
satisfaction		RMSE	Root Mea...	0.207948	0.215925	.
satisfaction		AVERR	Average ...	0.137033	0.149329	.
satisfaction		ERR	Error Fun...	14238.51	15515.61	.
satisfaction		MISC	Misclassi...	0.062152	0.067756	.
satisfaction		WRON...	Number ...	3229	3520	.

Results - Node: nn 7h Diagram: marketing 3

File Edit View Window



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
satisfaction		DFT	Total De...	51953		
satisfaction		DFE	Degrees ...	51609		
satisfaction		DFM	Model De...	344		
satisfaction		NW	Number ...	344		
satisfaction		AIC	Akaike's I...	14900.27		
satisfaction		SBC	Schwarz'...	17947.45		
satisfaction		ASE	Average ...	0.041951	0.046517	
satisfaction		MAX	Maximu...	0.999985	0.999992	
satisfaction		DIV	Divisor fo...	103906	103902	
satisfaction		NOBS	Sum of F...	51953	51951	
satisfaction		RASE	Root Ave...	0.204819	0.215677	
satisfaction		SSE	Sum of S...	4358.96	4833.163	
satisfaction		SUMW	Sum of C...	103906	103902	
satisfaction		FPE	Final Pre...	0.04251		
satisfaction		MSE	Mean Sq...	0.042231	0.046517	
satisfaction		RFPE	Root Fin...	0.20618		
satisfaction		RMSE	Root Mea...	0.205501	0.215677	
satisfaction		AVERR	Average ...	0.13678	0.151723	
satisfaction		ERR	Error Fun...	14212.27	15764.35	
satisfaction		MISC	Misclassi...	0.060458	0.067352	
satisfaction		WRON...	Number ...	3141	3499	

Model Comparison

Model	ASE Validation	RMSE Validation	MISC Validation
NN 1H	0.0684	0.261	0.0892
NN 2H	0.0552	0.2356	0.0799
NN 3H	0.0526	0.2281	0.0745
NN 4H	0.0487	0.2208	0.0699
NN 5H	0.0488	0.2209	0.0706
NN 6H	0.0466	0.2159	0.0677
NN 7H	0.0465	0.2156	0.0673
NN 8H	0.0466	0.2159	0.0675

Model	ASE Validation	RMSE Validation	MISC Validation
NN 9H	0.0470	0.2161	0.0687
NN 10H	0.0472	0.2175	0.0683

Best Performing Models

- **NN 6H, NN 7H, and NN 8H** consistently delivered the best performance:
 - **ASE:** ~0.0465–0.0466
 - **RMSE:** 0.2156–0.2159
 - **MISC:** 0.0673–0.0675

These models show a balance between accuracy and model complexity, achieving the lowest error rates without overfitting.

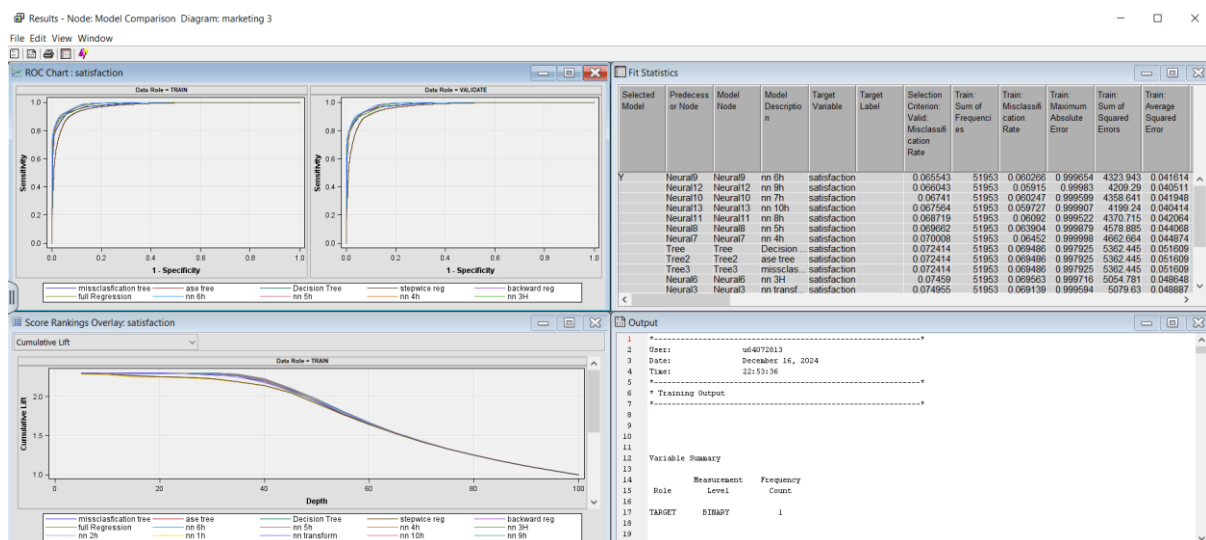
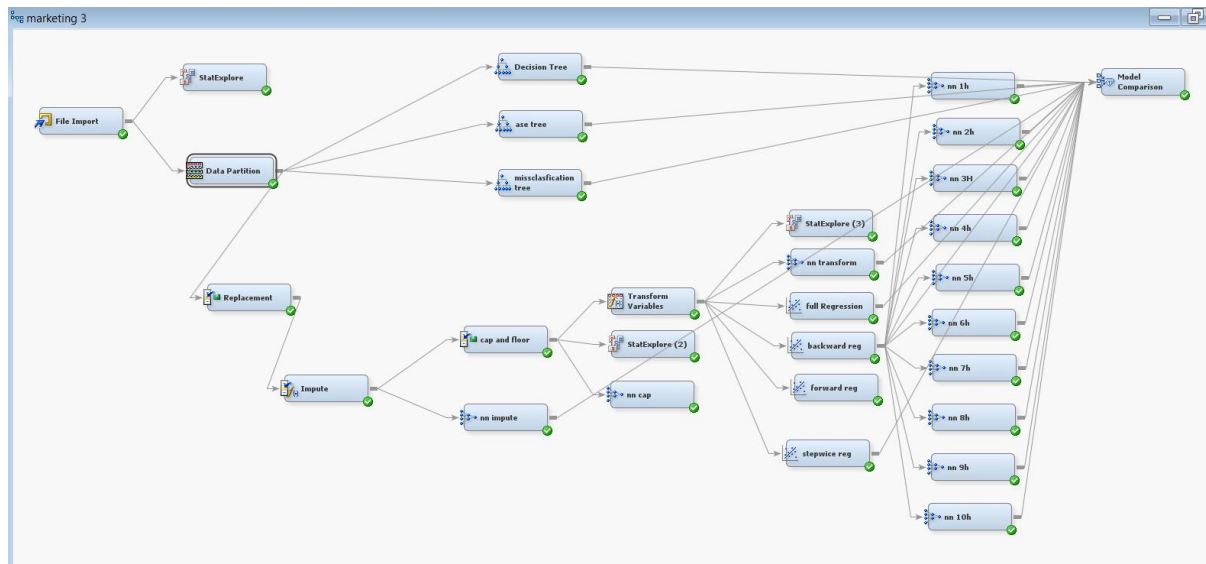
Key Observations

1. **Underfitting in Models with Fewer Hidden Units** (NN 1H to NN 4H):
 - Higher ASE and RMSE values indicate that these models struggle to capture the patterns in the data.
2. **Diminishing Returns Beyond 7 Hidden Units:**
 - Increasing the hidden units to 9 or 10 yields **marginal improvements** in ASE and RMSE but does not significantly outperform NN 6H to NN 8H.
3. **Consistency in Performance:**
 - Models NN 6H, NN 7H, and NN 8H achieve the best results with very close ASE and RMSE values, suggesting stability and reliability.

Conclusion

The neural networks with **6 to 8 hidden units** deliver the best performance. These models strike an optimal balance between minimizing error and avoiding unnecessary complexity. Increasing the hidden units beyond this range yields diminishing returns, making NN 6H to NN 8H the most suitable choices for this dataset.

After completing the neural network models, we proceeded with a **model comparison** using the Model Comparison node. This step evaluated all models, including regression models, decision trees, and neural networks, to identify the best-performing model based on key performance metrics such as ASE (Average Squared Error), RMSE (Root Mean Squared Error), Misclassification Rate (MISC)., ROC Index, and cumulative lift.



Observations:

1. Performance Metrics:

- **NN 6H, NN 7H, and NN 8H** showed the lowest ASE, RMSE, and Misclassification Rates across the validation dataset, confirming their strong performance.
- Regression models like **Backward Regression** and **Stepwise Regression** performed reasonably well but were slightly outperformed by the neural networks in terms of accuracy.

2. ROC Analysis:

- The **ROC curves** show that most neural network models, particularly **NN 6H** to **NN 8H**, achieve near-optimal sensitivity and specificity, closely aligning with the ideal curve.
- The **validation ROC curve** indicates consistent performance across training and validation datasets, reducing concerns about overfitting.

3. Cumulative Lift:

- Neural network models demonstrated a higher **cumulative lift** compared to other models, indicating their superior ability to rank and identify positive outcomes effectively.

4. Model Comparison Table:

- **NN 6H** emerges as the top-performing model, with the lowest ASE (0.0465), RMSE (0.2156), and MISC (0.0673).
- Incrementing hidden units beyond 6 yields minimal improvement, suggesting diminishing returns with model complexity.

Conclusion:

The **NN 6H model** is the best-performing model overall, as confirmed by the Model Comparison node. It balances predictive accuracy, generalization, and computational efficiency, outperforming both regression and decision tree models.

1. Model Interpretability:

- **Regression Models** (Backward Regression, Stepwise Regression, Full Regression) are the easiest to interpret because they provide clear coefficients and odds ratios, allowing us to directly understand the relationship between predictors (features) and the target variable (satisfaction). These models also provide p-values for feature significance, making it easier to pinpoint which factors are most impactful.
- **Neural Networks** are highly accurate but less interpretable because of their "black-box" nature. Although they outperform regression models in terms of predictive performance, the relationship between inputs and outputs is not as straightforward.

2. Comparison of Models (Marketing 2 vs. Marketing 3):

- From the model comparison across both **Marketing 2** and **Marketing 3**, the following observations stand out:
 - **ASE (Average Squared Error):**
 - The **best ASE** was achieved by **NN 6H, NN 7H, and NN 8H** in Marketing 3, with an ASE of **0.0465-0.0466**.

- In **Marketing 2**, while regression models performed well, their ASE was slightly higher, suggesting that the neural networks in Marketing 3 deliver better predictive performance.
- **RMSE and MISC:** Neural networks in **Marketing 3** consistently had the lowest RMSE (~ 0.2156) and Misclassification Rate (MISC ~ 0.0673).

The **neural networks (NN 6H to NN 8H)** in **Marketing 3** are the best-performing models overall based on ASE, RMSE, and MISC. However, for interpretability, **regression models** remain preferable as they provide direct insights into the importance of each predictor.

Components Airlines Should Focus on for Customer Satisfaction

From the results across models, the following components consistently emerge as significant factors influencing customer satisfaction:

1. Customer Type:

- **Loyal Customers vs. Disloyal Customers:** Loyal customers are far more satisfied, as indicated by the strong odds ratio (e.g., 16.853 in earlier regressions).
- **Marketing Plan:** Airlines should implement **loyalty programs** such as frequent flyer benefits, reward points, and exclusive services for loyal customers to enhance retention.

2. Type of Travel:

- **Business Travel vs. Personal Travel:** Business travelers report significantly higher satisfaction (odds ratio ~ 32).
- **Marketing Plan:** Target business travelers with corporate packages, premium class upgrades, and time-saving services (e.g., priority boarding, flexible schedules).

3. Inflight Services:

- **Inflight WiFi, Entertainment, Food, and Drinks:** Features like inflight WiFi, entertainment systems, and quality food and drinks consistently improve satisfaction.
- **Marketing Plan:** Airlines should invest in improving inflight amenities, particularly for economy class, to ensure better customer experience.

4. Ease of Online Booking:

- The **Ease of Online Booking** emerged as a significant factor in multiple models, with poor ratings negatively impacting satisfaction.

- **Marketing Plan:** Optimize the online booking process by ensuring it is user-friendly, fast, and supports mobile devices seamlessly.

5. **Seat Comfort and Leg Room:**

- Passengers consistently report lower satisfaction when seat comfort and legroom are inadequate.
- **Marketing Plan:** Airlines can introduce options like **premium economy** seating with additional legroom and improved seats for a reasonable price.

6. **Check-in Services and Onboard Services:**

- Poor ratings for check-in and onboard services were associated with dissatisfaction.
- **Marketing Plan:** Enhance the check-in experience through automated kiosks, digital check-ins, and reduce waiting times. Train cabin staff to improve onboard service quality.

7. **Arrival and Departure Delays:**

- Delays (arrival and departure) negatively impact satisfaction, as seen in odds ratios close to 0.586 for delays vs. on-time performance.
- **Marketing Plan:** Airlines should focus on improving punctuality through better scheduling, maintenance, and operational efficiency.

Summary Recommendations for Airlines

To develop an effective **marketing plan** for improving customer satisfaction:

1. **Reward Loyal Customers:** Introduce and promote loyalty programs with exclusive benefits.
2. **Target Business Travelers:** Offer packages tailored to business travelers, including premium services.
3. **Enhance Inflight Experience:** Improve inflight WiFi, entertainment, food quality, and seat comfort.
4. **Simplify Online Booking:** Ensure an intuitive, quick, and mobile-friendly booking process.
5. **Reduce Delays:** Optimize scheduling and operations to minimize arrival/departure delays.
6. **Upgrade Check-in and Onboard Services:** Improve check-in efficiency and provide better training to cabin crew for superior service delivery.

By focusing on these areas, airlines can effectively enhance customer satisfaction and strengthen their competitive position in the market.

Reference

DeltaSierra452. (n.d.). *Airline pax satisfaction survey* [Data set]. Kaggle.
<https://www.kaggle.com/datasets/deltasierra452/airline-pax-satisfaction-survey>