PREDICTIVE MODELING FOR AIRLINE SATIFACTION SURVEY

**Submitted by:**

Lakshman Siva Kumar Varma - 301494624

Jibin George - 301475160

Aeliya Fatima Syeda - 301410524

Ibukunoluwa Ajibade - 301336392

Contents

[**Project Overview** 6](#_Toc185274480)

[**Dataset Attributes** 6](#_Toc185274481)

[**Initial Data Preparation** 7](#_Toc185274482)

[**Data Partitioning** 8](#_Toc185274483)

[**Decision Tree 1: Default Tree with Largest Subtree Method** 9](#_Toc185274484)

[**Objective** 9](#_Toc185274485)

[**Configuration** 9](#_Toc185274486)

[**Decision Tree Structure** 10](#_Toc185274487)

[**Key Insights from the Tree:** 11](#_Toc185274488)

[**Overall Model Assessment:** 11](#_Toc185274489)

[**Decision Tree 2: ASE Tree** 11](#_Toc185274490)

[**Objective** 11](#_Toc185274491)

[**Configuration** 11](#_Toc185274492)

[**Fit Statistics** 11](#_Toc185274493)

[**Interpretation:** 12](#_Toc185274494)

[**Tree Structure and Splits** 12](#_Toc185274495)

[**Comparison with Default Tree** 13](#_Toc185274496)

[**Key Insights** 13](#_Toc185274497)

[**Decision Tree 3: Misclassification Tree** 13](#_Toc185274498)

[**Objective** 13](#_Toc185274499)

[**Configuration** 13](#_Toc185274500)

[**Fit Statistics** 14](#_Toc185274501)

[**Interpretation**: 14](#_Toc185274502)

[**Tree Structure and Splits** 15](#_Toc185274503)

[**Comparison with Other Trees** 16](#_Toc185274504)

[**Key Observations:** 16](#_Toc185274505)

[**Insights** 16](#_Toc185274506)

[**Data Preparation: Handling Missing Variables** 16](#_Toc185274507)

[**Objective** 16](#_Toc185274508)

[**Imputation Process** 16](#_Toc185274509)

[**Rationale for Mean Imputation** 17](#_Toc185274510)

[**Impact on Regression and Neural Network Models** 18](#_Toc185274511)

[**Data Preparation: Capping and Flooring Outliers** 18](#_Toc185274512)

[**Objective** 18](#_Toc185274513)

[**Capping and Flooring Process** 18](#_Toc185274514)

[**Rationale for Capping and Flooring** 19](#_Toc185274515)

[**Impact on Models** 20](#_Toc185274516)

[**Data Exploration: Skewness and Kurtosis Check** 20](#_Toc185274517)

[**Objective** 20](#_Toc185274518)

[**Key Observations** 20](#_Toc185274519)

[**Interpretation** 21](#_Toc185274520)

[**Data Preparation: Variable Transformation** 21](#_Toc185274521)

[**Objective** 21](#_Toc185274522)

[**Impact of Log Transformations on Skewness** 22](#_Toc185274523)

[**Results After Log Transformation** 22](#_Toc185274524)

[**Analysis** 23](#_Toc185274525)

[**Conclusion** 23](#_Toc185274526)

[**Connecting StatExplore to Transform Variables Node** 24](#_Toc185274527)

[**Objective** 24](#_Toc185274528)

[**Steps Performed** 24](#_Toc185274529)

[**Results After Log Transformation** 24](#_Toc185274530)

[**Key Observations** 25](#_Toc185274531)

[**Conclusion** 25](#_Toc185274532)

[**Regression Modeling: Logistic Regression Approaches** 25](#_Toc185274533)

[**Objective** 25](#_Toc185274534)

[**Key Adjustments Implemented** 28](#_Toc185274535)

[**Initial Observations Before Improvements** 29](#_Toc185274536)

[**Updates to the Marketing 3 Process** 31](#_Toc185274537)

[**Key Updates** 31](#_Toc185274538)

[**Final Structure of the Dataset** 31](#_Toc185274539)

[**Importing the Updated Data** 32](#_Toc185274540)

[**1. Data Exploration – StatExplore Node** 33](#_Toc185274541)

[**Interval Variables** 33](#_Toc185274542)

[**Observations:** 34](#_Toc185274543)

[**Purpose of Data Partitioning** 34](#_Toc185274544)

[**Decision Tree 1: Default Tree with Largest Subtree Method** 35](#_Toc185274545)

[**Objective** 35](#_Toc185274546)

[**Configuration** 35](#_Toc185274547)

[**2.Decision Tree Structure** 36](#_Toc185274548)

[**Key Insights from the Tree:** 37](#_Toc185274549)

[**Overall Model Assessment:** 37](#_Toc185274550)

[**Decision Tree Two: ASE Tree** 37](#_Toc185274551)

[**Objective** 37](#_Toc185274552)

[**Fit Statistics** 37](#_Toc185274553)

[**Interpretation:** 38](#_Toc185274554)

[**Third Decision Tree: Misclassification Tree** 39](#_Toc185274555)

[**Configuration of the Misclassification Tree** 39](#_Toc185274556)

[**Tree Diagram Overview** 40](#_Toc185274557)

[**Fit Statistics** 41](#_Toc185274558)

[**Performance Analysis** 41](#_Toc185274559)

[**Insights from the Tree** 41](#_Toc185274560)

[**Conclusion** 42](#_Toc185274561)

[**Recoding Ordinal Survey Variables for Regression and Neural Network Models** 43](#_Toc185274562)

[**Process of Recoding Using the Replacement Node** 43](#_Toc185274563)

[**Imputation of Missing Values** 44](#_Toc185274564)

[**Capping and Flooring Interval Variables** 45](#_Toc185274565)

[**Replacement Summary** 46](#_Toc185274566)

[**Output Highlights** 46](#_Toc185274567)

[**Transform Variables to Handle Skewness and Kurtosis** 47](#_Toc185274568)

[**Changes to Variable Roles and Levels** 47](#_Toc185274569)

[**State Explore Results After Log Transformation in Marketing 3** 48](#_Toc185274570)

[**Key Observations** 48](#_Toc185274571)

[**Models Built** 49](#_Toc185274572)

[**Odds Ratio Estimates** 52](#_Toc185274573)

[**Insights and Key Takeaways** 55](#_Toc185274574)

[**Fit Statistics** 56](#_Toc185274575)

[**Odds Ratio Estimates** 57](#_Toc185274576)

[**3. Insights** 60](#_Toc185274577)

[**4. Conclusion** 60](#_Toc185274578)

[**Interpretation of Stepwise and Forward Regression Results:** 60](#_Toc185274579)

[**Fit Statistics:** 61](#_Toc185274580)

[**Odds Ratio Estimates:** 62](#_Toc185274581)

[**Key Insights:** 64](#_Toc185274582)

[**Overall:** 65](#_Toc185274583)

[**Comparison of Regression Models** 65](#_Toc185274584)

[**Comparison Summary:** 65](#_Toc185274585)

[**Key Observations:** 65](#_Toc185274586)

[**Conclusion:** 66](#_Toc185274587)

[**Initial Setup for Neural Networks** 66](#_Toc185274588)

[**Optimization Settings** 67](#_Toc185274589)

[**Neural Network 1: NN Impute** 68](#_Toc185274590)

[**Model Comparison** 81](#_Toc185274591)

[**Best Performing Models** 82](#_Toc185274592)

[**Key Observations** 82](#_Toc185274593)

[**Conclusion** 82](#_Toc185274594)

[**Observations:** 83](#_Toc185274595)

[**Conclusion:** 84](#_Toc185274596)

[**Components Airlines Should Focus on for Customer Satisfaction** 85](#_Toc185274597)

[**Summary Recommendations for Airlines** 86](#_Toc185274598)

**Project Overview**

In today’s very competitive airline industry, customer satisfaction is no longer just a metric, it is a core driver of profitability, retention, and long-term brand equity. As digital transformation increases customer expectations, airlines must differentiate not merely on price, but on service excellence and experience personalization. Research shows that acquiring a new customer is 5 to 7 times more expensive than retaining an existing one (Gallo, 2014), and a 5% increase in customer retention can lead to profit growth of 25% to 95% (Reichheld & Sasser, 1990). For an industry with thin margins, these figures are not just persuasive, they are mission critical.

This project aims to uncover the key drivers of passenger satisfaction by analyzing a robust dataset that combines survey-based evaluations with flight-level operational data. Leveraging powerful machine learning techniques in SAS Enterprise Miner and Python, the analysis goes beyond surface-level metrics to identify the most predictive variables influencing satisfaction levels, enabling more precise and proactive decision-making.

The dataset includes over 100,000 passenger records, capturing variables such as flight distance, customer loyalty type, seat comfort, onboard services, and satisfaction across multiple touchpoints. By segmenting passengers into satisfied vs. dissatisfied/neutral categories, the analysis provides airlines with clear, data-driven priorities for operational improvements, targeted marketing, and loyalty-building initiatives.

Ultimately, these insights empower airlines to design more personalized, responsive, and value-generating customer experiences, turning satisfaction into sustained competitive advantage.

**Dataset Attributes**

* **The dataset includes the following variables:**
* **Id – Unique passenger identifier**
* **Gender – Male or Female**
* **Customer Type – Loyal or Disloyal customer**
* **Age – Passenger age**
* **Type of Travel – Business or Personal travel**
* **Class – Business, Economy, or Economy Plus**
* **Flight Distance – Measured in miles.**
* **Inflight Wi-Fi Service – Satisfaction score (0–5, where 0 = Not Applicable)**
* **Departure/Arrival Time Convenience – Satisfaction score (0–5)**
* **Ease of Online Booking – Satisfaction score (0–5)**
* **Gate Location – Satisfaction score (0–5)**
* **Food and Drink – Satisfaction score (0–5)**
* **Online Boarding – Satisfaction score (0–5)**
* **Seat Comfort – Satisfaction score (0–5)**

**Exploratory Data Analysis**

**1. Dataset Overview**

Total records: 103,904

Columns: 25

Rows:

**2. Categorical Variables and Their Unique Values**

| Column | Unique Values |
| --- | --- |
| Gender | ['Male', 'Female'] |
| Customer\_Type | ['Loyal Customer', 'disloyal Customer'] |
| Type\_of\_Travel | ['Personal Travel', 'Business travel'] |
| Class | ['Eco Plus', 'Business', 'Eco'] |
| satisfaction | ['neutral or dissatisfied', 'satisfied'] |

All values are valid with no apparent typos or unexpected categories.

**3. Range Check for Numerical Variables**

| Variable | Min | Max | Expected Range? | Notes |
| --- | --- | --- | --- | --- |
| Age | 7 | 85 | Yes | Realistic human ages |
| Flight\_Distance | 31 | 4983 | Yes | Very short to long-haul |
| Inflight\_wifi\_service | 0 | 5 | Yes | 0 = N/A, 1–5 = rating |
| Food\_and\_drink | 0 | 5 | Yes | Clean range |
| Baggage\_handling | 1 | 5 | Slightly narrow | Min is 1, no "0"/N/A |
| Departure\_Delay\_in\_Minutes | 0 | 1592 | Skewed | Some flights delayed 26+ hours |
| Arrival\_Delay\_in\_Minutes | 0 | 1584 | Skewed | Similar to departure delay |

Delay fields show extreme outliers.

**4. Missing Values Analysis**

Only one column has missing values:

|  |  |  |
| --- | --- | --- |
| **Column** | **Missing Count** | **% Missing** |
| Arrival\_Delay\_in\_Minutes | 310 | 0.30% |

**5. Correlation Heatmap**

The heatmap reveals several strong internal relationships, particularly among customer service features:

Online\_boarding is moderately to strongly correlated with:

Inflight\_entertainment (0.29)

Seat\_comfort (0.42)

Inflight\_wifi\_service (0.46)

Food\_and\_drink highly correlates with:

Cleanliness (0.66)

Seat\_comfort (0.57)

Inflight\_entertainment (0.62)

Departure vs. Arrival Delays: 0.97 almost perfect correlation.

Inflight\_entertainment and On-board\_service show strong positive correlations (0.67)

Arrival/Departure Delays have low correlations with satisfaction indicators individually, but may be impactful in combination or for specific segments (e.g., business travelers).

A few high-performing features (e.g., WiFi, food, seat comfort) may represent latent satisfaction clusters.

A blue and red squares with black text

AI-generated content may be incorrect.

Marketing 2 model

**Initial Data Preparation**

**Data Import and Variable Roles**

The dataset was imported into SAS Enterprise Miner, and each variable was assigned a role based on its statistical type and business relevance. This step ensured that only meaningful variables were used in modeling, minimizing noise and redundancy.

* Ordinal Variables: Service ratings such as Inflight Wi-Fi, Seat Comfort, and Cleanliness were treated as ordinal to reflect their ranked nature.
* Nominal Variables: Categorical features without a logical order, such as Gender, Customer Type, and Class.
* Interval Variables: Numeric features like Flight Distance and Delay Times were handled as continuous inputs.
* Binary Variables: Flags like Satisfaction (Yes/No) were categorized for binary classification.

**Target Variable Transformation**

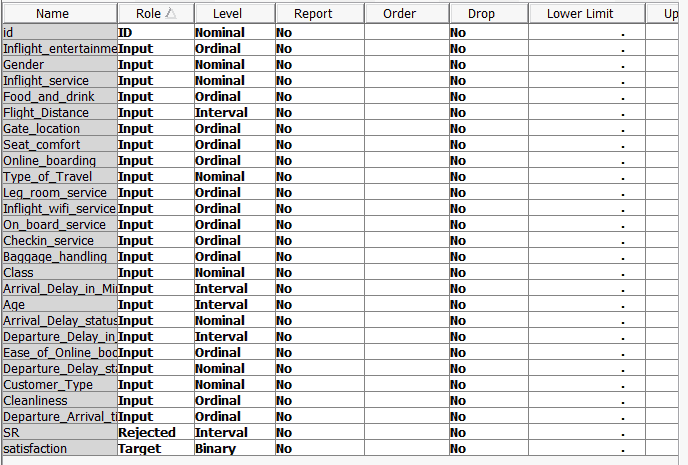
To streamline classification and support clear, actionable insights, the original three-level Satisfaction variable (Satisfied, Neutral, Dissatisfied) was converted into a binary format:

* 1 – Satisfied
* 0 – Dissatisfied/Neutral

This transformation helped focus the analysis on identifying clear satisfaction drivers that differentiate satisfied passengers from others.

Additional Preprocessing Steps

* Missing Values: A small percentage (~0.3%) of missing values in Arrival Delay were imputed using the means to preserve record integrity.
* Outliers: Delay fields showed right-skewed distributions and were log-transformed to improve model accuracy and stability.
* Non-Predictive Fields: Unique identifiers and administrative codes were excluded from the modeling process but retained for traceability.
* Class Balance: The dataset had a reasonably balanced target variable, and rebalancing techniques (e.g., SMOTE) were not necessary.

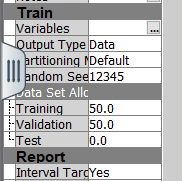


**Data Partitioning**

To ensure an unbiased model evaluation and reduce the risk of overfitting, the dataset was split into two equal subsets using a 50/50 random partitioning strategy:

* Training Set (50%) – Used to build and tune predictive models.
* Validation Set (50%) – Used to evaluate model performance on unseen data.

This balanced approach provides transparent and fair model validation. While alternative splits like 70/30 or stratified sampling were considered, the 50/50 division was selected to align with enterprise modeling standards in SAS and ensure clear, consistent performance comparisons across modeling stages.

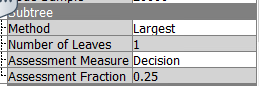


**Decision Tree 1: Default Tree Using Largest Subtree Method**

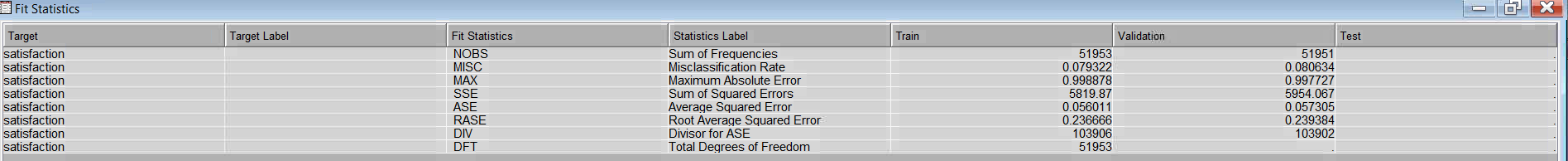
**Objective**This initial decision tree model was developed as a baseline to explore natural variable splits that drive customer satisfaction. It was built using default settings to highlight organic interactions between variables and the target outcome (satisfaction). This configuration allows for straightforward benchmarking before introducing optimization.

Configuration Summary

* Subtree Method: Set to *Largest*, ensuring the model grows fully and is pruned only when necessary. This helps capture all meaningful interactions without early restriction.
* Assessment Measure: Based on *misclassification loss*, which focuses on correctly classifying satisfied vs. dissatisfied customers.
* Settings: No manual hyperparameter tuning was applied. All values were left at system defaults to provide a neutral starting point for comparison.



1. **Fit Statistics**

****

The fit statistics table offers a quantitative evaluation of how well the decision tree model predicts satisfaction on both training and validation data.

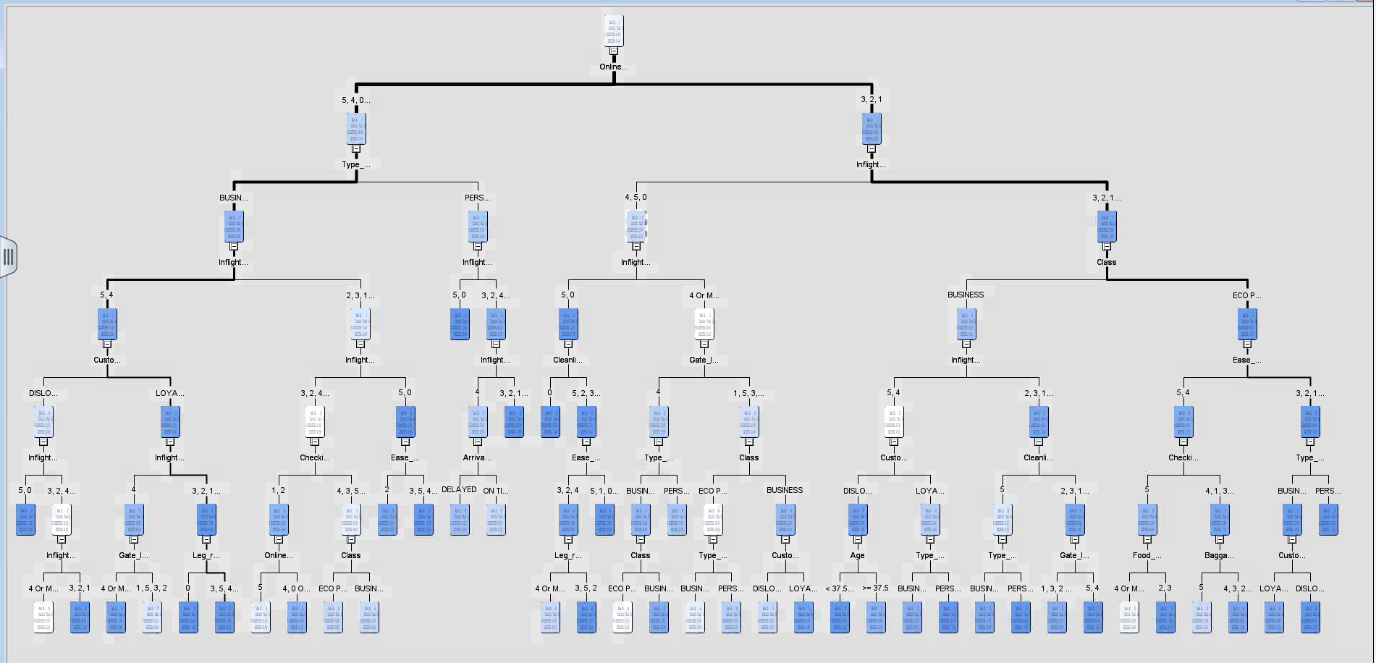
**Average Squared Error (ASE):**

**Training ASE**: 0.0560

**Validation ASE**: 0.0573

**Interpretation**: The closeness of ASE between training and validation sets indicates **strong model generalization**. This suggests the tree structure avoids overfitting and will perform reliably on unseen passenger data.

## **Decision Tree Structure**

****

This visual illustrates how customer satisfaction is segmented based on key service-related attributes. The decision tree highlights the hierarchy of factors driving satisfaction and dissatisfaction, offering practical insights for service prioritization.

**Primary Split: Inflight Wi-Fi Service (< 4 or ≥ 4)**

* **Key Insight**: The strongest driver of satisfaction. Passengers rating inflight Wi-Fi service below 4 had a significantly higher likelihood of dissatisfaction.
* **Business Relevance**: Wi-Fi performance has a direct influence on overall satisfaction. Investing in reliable, fast, and accessible inflight internet could generate a measurable improvement in customer experience and loyalty.

**Key Sub-Splits and Customer Segments**

**Branch 1: Low Wi-Fi Ratings (< 4)**

* **Cleanliness (< 2)**: Among passengers already dissatisfied with Wi-Fi, lower cleanliness ratings further increased dissatisfaction.  
  → This reveals a **compounded dissatisfaction effect**, combining negative digital and physical experiences.
* **Implication**: Customers with poor Wi-Fi and cleanliness experience form a high-risk churn segment. These customers are ideal for targeted service recovery or follow-up initiatives.
* **Class (Business vs. Economy Plus)**: Business class travelers consistently reported higher satisfaction. Premium service standards appear to mitigate dissatisfaction risk even when Wi-Fi scores are low.

**Branch 2: High Wi-Fi Ratings (≥ 4)**

* **Online Boarding (< 4)**: Satisfaction dropped when digital check-in/boarding processes were rated poorly.
* **Ease of Booking (< 5)**: Friction in booking workflows also contributed to reduced satisfaction despite good inflight Wi-Fi.

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 Takeaways from the First Decision Tree**

**Primary Satisfaction Drivers**

1. **Inflight Wi-Fi Service**
   * This was the most critical factor in customer satisfaction. Passengers rating Wi-Fi service below 4 were far more likely to express dissatisfaction.
   * **Business Impact**: Enhancing Wi-Fi reliability and speed represents a clear opportunity for improving overall customer sentiment and retention.
2. **Cleanliness, Class of Travel, and Online Boarding**
   * These service areas strongly influenced customer satisfaction, particularly among business travelers and frequent flyers.
   * **Actionable Insight**: Prioritize investment in cleanliness standards and smooth digital boarding experiences, especially in premium segments.

**Secondary Contributors**

3. **Customer Type and Booking Ease**

* Loyal customers and those experiencing frictionless digital booking processes consistently reported higher satisfaction.
* **Strategic Value**: This supports further investments in loyalty programs and streamlining the digital customer journey.

**Model Assessment and Business Relevance**

**Accuracy and Reliability**:

The model achieved a misclassification rate of approximately 8%, confirming its ability to reliably differentiate between satisfied and dissatisfied customers.

**Model Robustness**:

Training and validation ASE (Average Squared Error) were nearly identical, indicating a stable model that is unlikely to overfit or underperform on new data.

**Real-World Alignment**:

The tree’s splits aligned with actual service pain points and strengths. It reflected customer sentiment around in-flight experience, ease of booking, and travel class.

**Business Utility**:

This model provides a clear, data-driven roadmap for prioritizing improvements. It supports targeted interventions—such as Wi-Fi upgrades or personalized service efforts—aimed at lifting satisfaction levels in underperforming segments.

**ASE-Optimized Decision Tree**

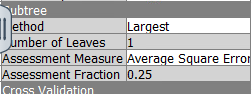
**Objective**This version of the decision tree was configured to minimize Average Squared Error (ASE)—a metric that measures the numerical gap between predicted and actual satisfaction outcomes.

Unlike the first tree, which focused on classification accuracy, this model was tuned for prediction precision, making it especially valuable for:

* Strategic service improvement planning.
* Resource allocation prioritization.
* Personalization scenarios where minimizing numerical prediction errors are key.

**Configuration Summary**

* Subtree Method: Set to "Largest" to retain the most informative structure while still preventing overfitting.
* Assessment Measure: Switched from classification error to Average Squared Error (ASE) to better capture deviations between predicted and actual satisfaction scores.
* Assessment Fraction: Retained at 0.25 to maintain consistency and comparability across models.
* Other Settings: All other settings were kept consistent with the default tree for benchmarking purposes.

****

## **Fit Statistics**

The fit statistics for the ASE Tree are as follows:

| **Metric** | **Training** | **Validation** |
| --- | --- | --- |
| **Misclassification Rate (MISC)** | 0.079322 | 0.080634 |
| **Average Squared Error (ASE)** | 0.056011 | 0.057305 |
| **Root ASE (RASE)** | 0.236606 | 0.239384 |
| **Sum of Squared Errors (SSE)** | 5819.87 | 5954.067 |

## **Model Interpretation and Tree Comparison**

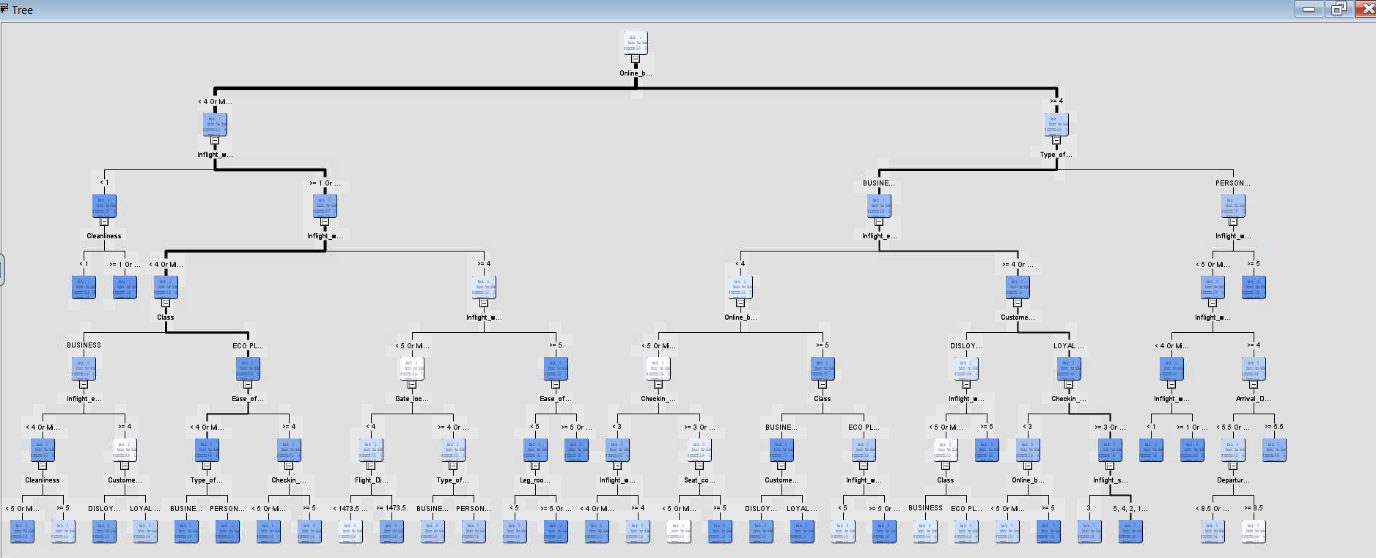
## **Performance Interpretation**

## Average Squared Error (ASE): The training and validation ASE values (0.056 and 0.057) are closely aligned, confirming that the model is well-fitted and generalizes effectively to unseen data.

## Misclassification Rate (~8%): Accuracy remains consistent with the default tree, showing that switching to an ASE-based evaluation does not impact classification quality.

## Root ASE (RASE): Minimal difference between training and validation scores (0.2366 vs. 0.2393) demonstrates the model’s stability and predictive reliability across datasets.

## **Tree Structure and Splits**

****

# Tree Structure and Variable Insights

# Primary Split

# Inflight Wi-Fi Service (< 4 or ≥ 4): This remains the strongest predictor of passenger satisfaction. Customers giving Wi-Fi ratings below 4 are substantially more likely to be dissatisfied.

# Secondary Splits

# For Low Wi-Fi Ratings (< 4):

# Cleanliness and Class (e.g., Business vs. Economy Plus) are critical for satisfaction.

# For High Wi-Fi Ratings (≥ 4):

# Ease of Booking, Online Boarding, and Legroom Service become more relevant to differentiate satisfied from unsatisfied passengers.

# Deeper Splits

# Like the default tree, this model incorporates additional splits based on Customer Type, Type of Travel, and Seat Comfort.

# Model Comparison: ASE Tree vs. Default Tree

# Similarities

# Both trees prioritize Inflight Wi-Fi Service as the dominant driver of satisfaction.

# Key secondary predictors—Cleanliness, Class, Ease of Booking—are consistent across both models.

# Differences

# The ASE tree is tuned specifically for numerical prediction accuracy, yielding lower ASE metrics.

# Despite this tuning, tree structure and variable importance remain consistent, affirming model reliability.

# Key Stakeholder Insights

# The ASE Tree confirms that Wi-Fi Service, Cleanliness, and Ease of Booking are the most influential levers for improving customer satisfaction.

# Optimizing for prediction accuracy (ASE) did not impact the classification ability of the model—both trees achieve a ~92% accuracy rate.

# Structural consistency between the models reinforces the robustness of the predictors, making these findings suitable for real-world implementation.

**Decision Tree 3: Misclassification-Optimized Tree**

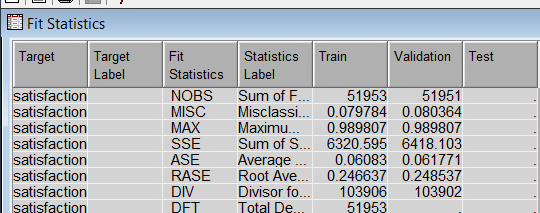
**Objective**

The third decision tree is optimized to reduce the Misclassification Rate, prioritizing accurate classification of passengers as either Satisfied or Dissatisfied. This model is especially relevant for frontline service teams and customer experience managers who require high precision in identifying dissatisfied customers for recovery efforts.

**Configuration Summary**

* Subtree Method: Set to *Assessment* — the model uses validation data to decide where to prune, improving real-world prediction accuracy.
* Assessment Measure: *Misclassification* — directly targets minimizing incorrect classification outcomes.
* Number of Leaves: 1 (baseline setup before pruning).
* Assessment Fraction: 0.25 % of data used for pruning evaluation.

## **Fit Statistics**

****

| **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 |

## **Model Interpretation**

## **Misclassification Rate**

## Training: 0.0797 (7.97%)

## Validation: 0.0803 (8.03%)

## The minor increase between training and validation suggests that the model generalizes effectively. This consistency indicates reliability when deployed on unseen passenger data.

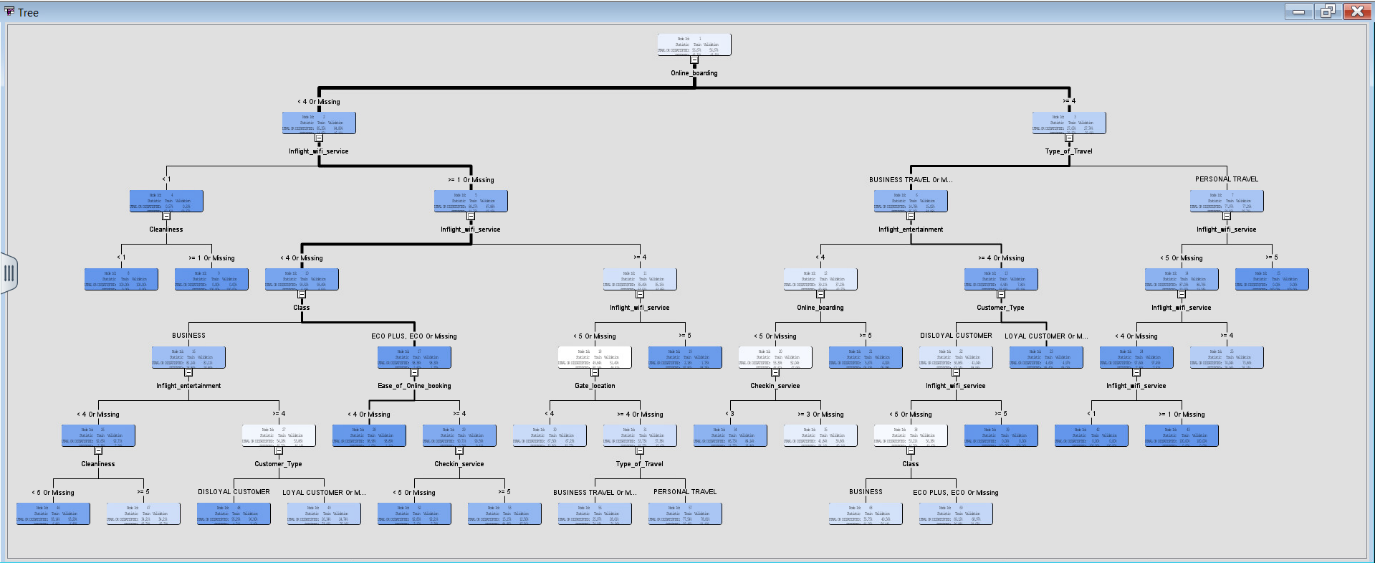
## **Average Squared Error (ASE)**

## The ASE is slightly higher than in the ASE-optimized tree (from 0.0573 to 0.0617), which is expected since this model prioritizes correct classification over minimizing numerical prediction error.

## **Root ASE (RASE)**

## With minimal variation between training (0.2466) and validation (0.2485), this model confirms its balance and resistance to overfitting.

## **Tree Structure and Splits**

****

**Primary Split:**

* **Inflight Wi-Fi Service (< 4 or Missing vs ≥ 4)**

This remains the strongest and most consistent predictor of customer satisfaction across all decision trees.

**Secondary Splits:**

* **For Wi-Fi Ratings < 4:**

**Cleanliness** and **Class of Service** further differentiate dissatisfied passengers.

* **For Wi-Fi Ratings ≥ 4:**

**Online Boarding**, **Type of Travel**, and **Customer Type** become key discriminators among satisfied passengers.

**Deeper Splits**:

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

### Comparison Across All Decision Trees

### Tree Complexity and Structure

Compared to the ASE-optimized tree, the misclassification-focused tree demonstrates a comparable structure but includes **slightly deeper branches**. These deeper splits improve the model’s ability to classify borderline cases more precisely, particularly useful in identifying at-risk customers.

# **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 Takeaways**

**Consistent Top Predictor:**

All three trees identify Inflight Wi-Fi Service as the most influential factor impacting passenger satisfaction. This is followed closely by Cleanliness, Class of Travel, and Online Boarding.

**Classification Accuracy vs. Precision Trade-Off:**

Although the Misclassification Tree focuses on minimizing classification errors, it achieves nearly the same accuracy (~8%) as the other models. However, it shows a slightly higher ASE because it prioritizes classification correctness over prediction precision.

**Model Robustness:**

The structural similarity across trees confirms that the dataset is robust. Core predictors remain stable, reinforcing confidence in the reliability of model-driven recommendations.

**Strategic Insight**

The Misclassification Tree is best suited for practical applications where identifying dissatisfied passengers is a priority — such as loyalty intervention, churn prevention, or complaint resolution.

Consistency in Top Predictors across models (Wi-Fi, Cleanliness, Class, Online Boarding) suggests that improving these specific service touchpoints will universally enhance customer satisfaction, regardless of how success is measured.

**Data Preparation: Handling Missing Data**

**Objective**

Prior to implementing advanced models such as regression and neural networks, it was essential to address missing values within the dataset. Unresolved gaps in data can distort model outputs, reduce reliability, and introduce bias. Therefore, rigorous imputation was undertaken to ensure model robustness and analytical precision.

**Imputation Process Overview**

1. **Identified Variable with Missing Data**

Arrival Delay (Minutes) was found to have 162 missing entries in the training dataset.

1. **Imputation Technique**

The Mean Imputation method was applied to replace missing values with the average of observed values.

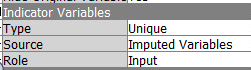
Imputed Mean Value: 15.18264 minutes

1. **Use of Indicator Variable**

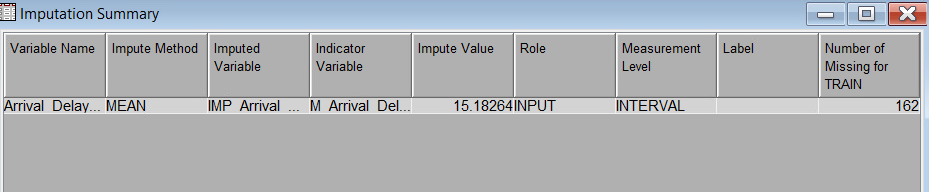
A dedicated indicator variable was generated to mark every instance where imputation occurred.

This approach allows the model to differentiate between original and imputed data, uncovering potential patterns in how missingness may relate to satisfaction or delay outcomes.

**Configuration Details**

****

**Imputation Summary**:

******Key Technical Details**

* **Variable Name:** Arrival Delay (Minutes)
* **Imputation Method:** Mean
* **Imputed Variable Created:** IMP\_Arrival\_Delay
* **Indicator Variable Created:** M\_Arrival\_Delay (flags rows with imputed values)
* **Imputed Mean Value:** 15.18264
* **Number of Missing Values:** 162 (from training dataset)
* **Measurement Level:** Interval (maintains continuous data structure)

**Why Mean Imputation?**

**Balanced Central Estimate:**

Mean is a commonly accepted imputation method for continuous data. It avoids bias by inserting a representative central value without distorting the distribution.

**Preservation of Dataset Integrity:**

Enables use of all records in modeling, avoiding the data loss caused by listwise deletion.

**Added Indicator for Flexibility:**

The indicator variable (M\_Arrival\_Delay) allows the model to learn if missingness itself carries predictive meaning — particularly useful in-service delay or operational reliability analysis.

**Impact on Downstream Models**

**Regression and Neural Network Models:**

Able to utilize **100% of available records** without deleting missing rows.

The added indicator boosts model adaptability and helps capture latent patterns linked to missingness.

## **Data Preparation: Capping and Flooring Outliers**

## **Objective**

## To reduce the influence of extreme values in numerical variables, the Replacement Node was applied using a capping and flooring strategy. This preprocessing technique protects regression and neural network models from distortion caused by outliers.

**Process Summary**

**Node Used:**

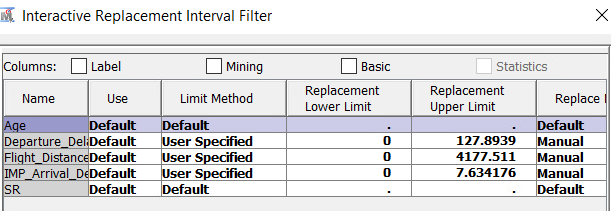
The **Replacement Node** was renamed **"Cap and Floor"** for clarity and applied to selected interval variables.

**Variables Processed:**

Three key numerical features were flagged for outlier treatment:

* Departure Delay (Minutes)
* Flight Distance
* IMP\_Arrival\_Delay (Imputed version of Arrival Delay)

**Replacement Method**:

****

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

| **Variable** | **Lower Limit** | **Upper Limit** |
| --- | --- | --- |
| **Departure Delay** | 0 | 127.8939 |
| **Flight Distance** | 0 | 4177.511 |
| **IMP\_Arrival Delay** | 0 | 7.634176 |

* **Lower Limit Rule:** Any value below 0 was capped at 0.
* **Upper Limit Rule:** Any value exceeding the specified upper threshold was replaced with the cap.

**Rationale**

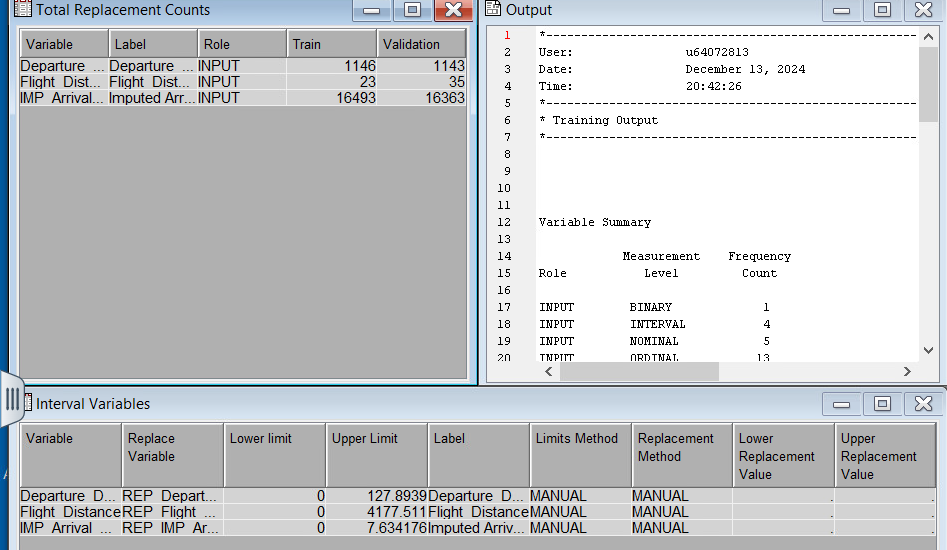
* **Why Cap and Floor?**

Outliers can disproportionately skew model predictions, especially for sensitive algorithms like regression. By constraining variable ranges, the data becomes more **robust**, stable, and suitable for modeling.

* **Preserves Interpretability:**

Capping maintains the original variable scale while controlling for anomalies, ensuring that insights from models remain explainable.

1. **Replacement Results**:

****

After applying the capping and flooring thresholds:

* **Departure Delay**
  + **Training Set:** 1,146 replacements
  + **Validation Set:** 1,143 replacements.
* **Flight Distance**
  + **Training Set:** twenty-three replacements
  + **Validation Set:** thirty-five replacements.
* **Imputed Arrival Delay**
  + **Training Set:** 16,493 replacements
  + **Validation Set:** 16,363 replacements.

## **Output Variables**

## Following the replacement process, new capped variables were generated with the prefix REP\_ to distinguish them from their original forms:

## REP\_Departure\_Delay

## REP\_Flight\_Distance

## REP\_IMP\_Arrival\_Delay

## These transformed variables were then used in downstream regression and neural network models.

## **Rationale for Capping and Flooring**

## **Why was this important?** Outliers can distort model accuracy, especially for algorithms sensitive to numerical scale like regression and neural networks. Instead of dropping entire rows containing extreme values, this method retains the full dataset by modifying only the affected points.

## Capping protects against inflated values that could bias model coefficients.

## Flooring ensures extremely low anomalies (e.g., negative delays) are brought to a meaningful lower bound.

## Together, this treatment supports data integrity while enhancing predictive stability.

### Data Exploration: Skewness and Kurtosis Check

#### **Objective**

Ensure that the numerical variables were suitable for regression and neural network modeling, we assessed them for **skewness** and **kurtosis** using the **StatExplore Node**. This diagnostic helps evaluate whether variables are normally distributed or require transformation.

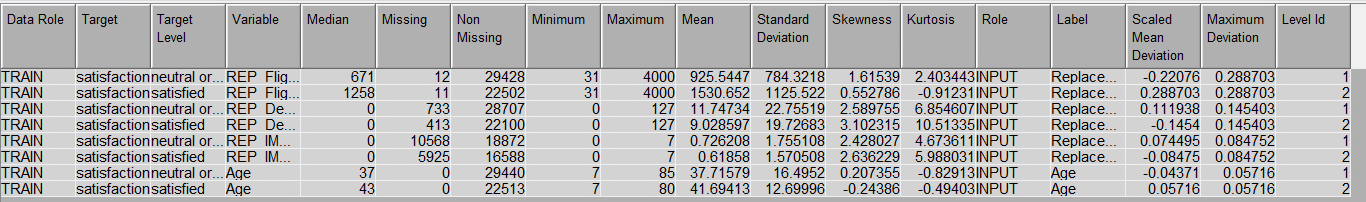
**Skewness** indicates asymmetry in the distribution.

**Kurtosis** measures the peakedness or tail heaviness.

Addressing non-normal distributions improves the stability and predictive performance of downstream models.

## **Key Observations**

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



| **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. |
| Age (Satisfied Group) | -0.24386 | -0.49403 | Slight negative skew, almost normal. |

**Interpretation**

**1. Skewness**

* A skewness near 0 indicates a symmetrical distribution.
* A positive skew (e.g., for REP\_Departure\_Delay) implies a longer tail on the right side, pointing to high values or outliers.
* Significant skewness, as observed in REP\_Departure\_Delay and REP\_IMP\_Arrival\_Delay, can impact model performance—particularly in regression and neural networks—by introducing non-linear distortions.

**2. Kurtosis**

* A kurtosis of 3 denotes a normal distribution.
* High kurtosis (≥3), observed in REP\_Departure\_Delay and REP\_IMP\_Arrival\_Delay, signals heavy tails—indicating a higher likelihood of outliers or extreme values.
* Negative kurtosis suggests flatter distribution. This is evident in the Age variable, where both Neutral and Satisfied customer groups show a slightly flatter, almost-normal distribution.

# **Data Preparation: Variable Transformation**

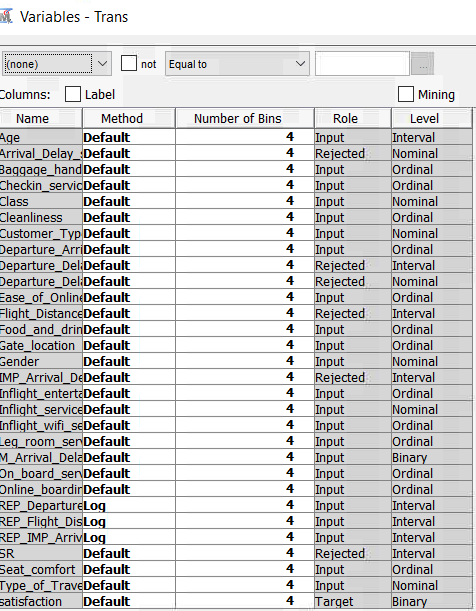
## **Objective**

To address the skewness and kurtosis issues identified during the data exploration stage, the **Transform Variables** node was applied. This node enabled the implementation of **logarithmic transformations** on highly skewed interval variables. The goal was to:

* Normalize the distributions of affected variables
* Improve the predictive performance of **regression** and **neural network** models
* Reduce the impact of outliers and variance instability

**Implementation**

* Variables such as REP\_Departure\_Delay and REP\_IMP\_Arrival\_Delay, which exhibited **high positive skewness and kurtosis**, were selected for **log transformation**.
* The transformation method applied was **logarithmic (log10)**, and the transformed variables were retained for modeling purposes.

****

## **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 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. |

## Log Transformation Analysis

### Detailed Variable Analysis

1. **REP\_Departure\_Delay in Minutes**
   * **Skewness Reduction:** From 2.79 to 0.88
   * **Interpretation:** This is a **substantial improvement**, indicating the log transformation effectively addressed the extreme values and normalized the distribution.
2. **REP\_Flight\_Distance**
   * **Skewness Reduction:** From 1.10 to -0.21
   * **Interpretation:** The distribution is now **symmetrical**, suggesting the skewness issue was **fully resolved**.
3. **REP\_IMP\_Arrival\_Delay**
   * **Skewness Reduction:** From 2.52 to 2.03
   * **Interpretation:** While improved, the skewness **still exceeds** the acceptable range for normal distribution.
   * **Note:** This indicates that **residual outliers** or extreme values persist. Additional preprocessing (e.g., advanced transformations or secondary capping/flooring) may be needed if the variable proves influential in model performance.

## **Post-Transformation Evaluation using StatExplore.**

### Objective

Verify the impact of log transformations, the **StatExplore Node** was linked to the **Transform Variables Node**. This allowed for an assessment of how the distribution of each variable changed after transformation.

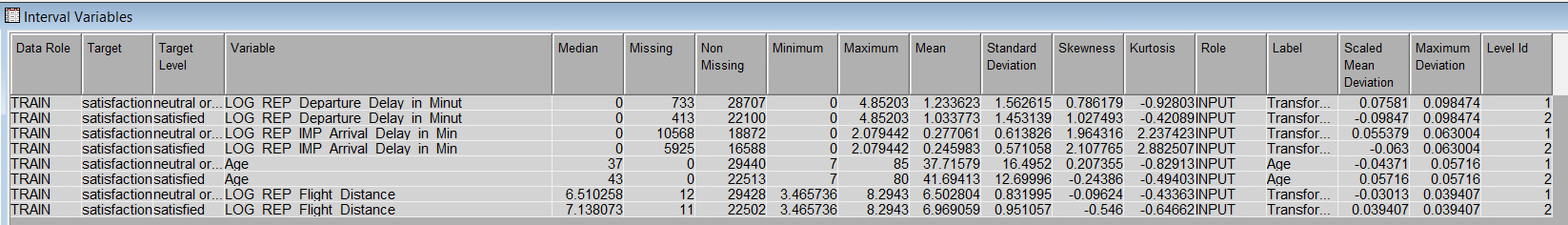
### Steps Taken

1. **Transform Variables Node**
   * Applied log transformations to the following variables:
     + REP\_Departure\_Delay
     + REP\_Flight\_Distance
     + REP\_IMP\_Arrival\_Delay
2. **StatExplore Node**
   * Assessed the distributions using:
     + **Skewness** (Symmetry check)
     + **Kurtosis** (Tail heaviness)

### Findings

* Post-transformation **skewness and kurtosis values** demonstrated that the majority of variables were **successfully normalized**.
* For variables still showing abnormal values (like REP\_IMP\_Arrival\_Delay), further investigation into **advanced preprocessing** techniques is suggested.

## **Results After Log Transformation**

****

| **Variable** | **Skewness** | **Kurtosis** | **Improvement** | **Interpretation** |
| --- | --- | --- | --- | --- |
| **LOG\_REP\_Departure Delay in Minutes** (Neutral) | 0.786179 | -0.92803 | Significant Improvement | Skewness reduced; close to normal. |
| **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 that the log transformation effectively normalized the distributions.
   * **REP\_IMP\_Arrival\_Delay** showed partial improvement but continues to exhibit moderate skewness, suggesting that additional transformation or preprocessing may be required.
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.
   * However, **REP\_IMP\_Arrival\_Delay** still displays moderate tail heaviness, indicating the presence of residual outliers that may influence model performance.

**Conclusion**

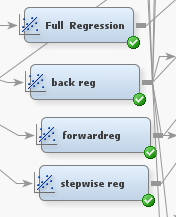
By connecting the **StatExplore Node** to the **Transform Variables Node**, we verified that log transformations significantly enhanced the distributional properties of key interval variables.  
While most variables achieved near-normal distributions post-transformation, **LOG\_REP\_IMP\_Arrival\_Delay** retains slight skewness and moderate kurtosis. This variable should be monitored during modeling for potential residual effects.

**Regression Modeling: Logistic Regression Approaches**

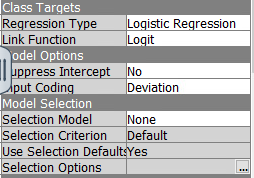
**Objective**

Despite the presence of some residual data issues—such as minor skewness in certain interval variables—logistic regression models were developed to predict the binary outcome variable: Satisfaction (Satisfied vs. Dissatisfied). The goal was to evaluate variable influences and assess model performance using different selection strategies.

**Models Built**

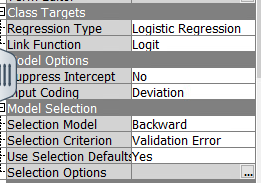
****

1. **Full Regression**

****

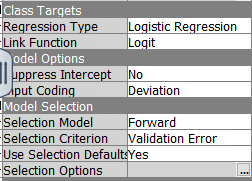
* + Includes **all variables** without any selection criteria.
  + Purpose: Serves as a baseline model to compare against other regression approaches.

1. **Backward Regression**

****

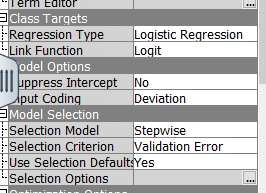
* + **Selection Model**: Backward
  + **Selection Criterion**: Validation Error
  + Approach: Starts with all predictors and sequentially removes the least significant variables based on the validation error.

1. **Forward Regression**

****

* + **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.

1. **Stepwise Regression**

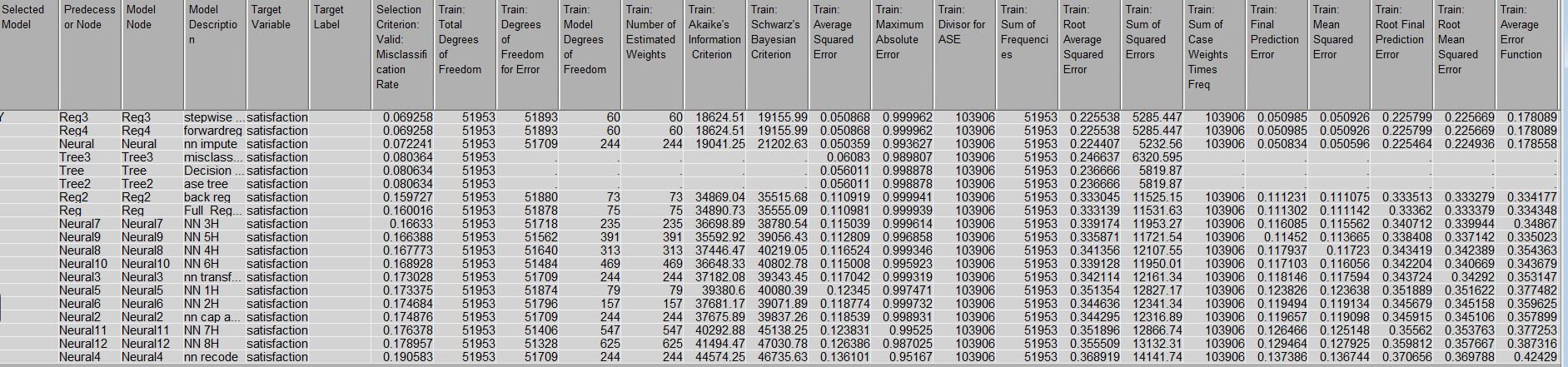
****

* + **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.

****

**Model Comparison Results Before Improvements**

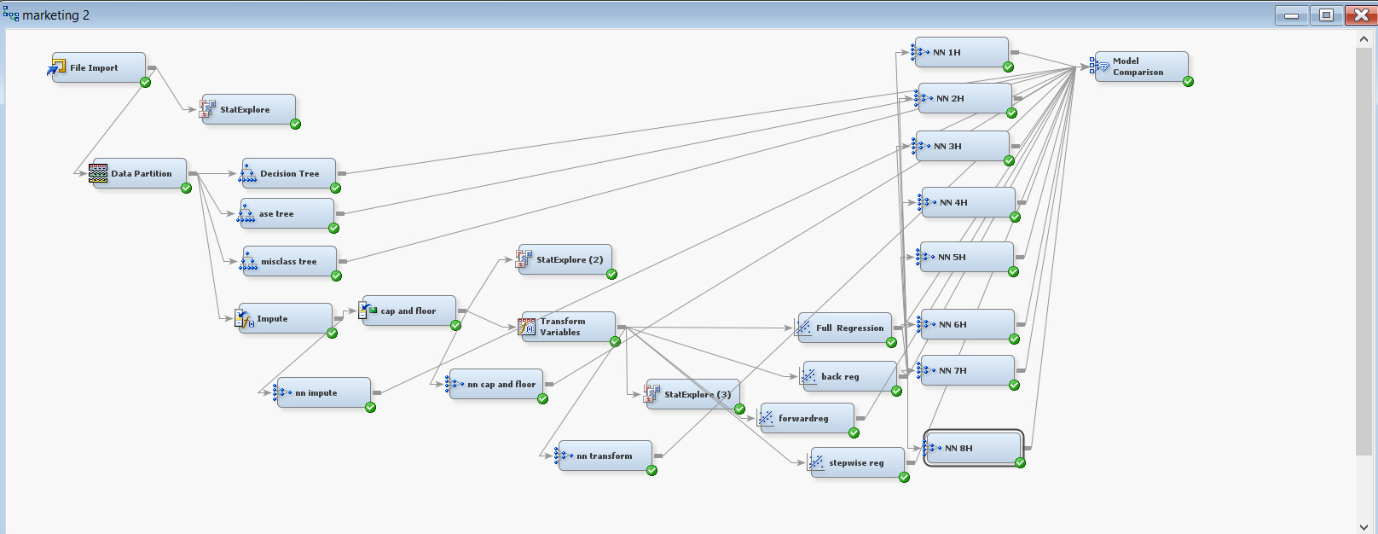
The table below summarizes the **ASE** and key observations for each model:

| **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; robust 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 three 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.



Marketing 3

## **Updates to the Marketing 3 Process**

### ****Objective****

Following the professor’s feedback, several critical updates were made in the revised Marketing 3 workflow. These updates aimed to enhance the model setup and streamline the data analysis process.

### ****Key Updates****

1. **Creation of New Variables**

Two new variables were engineered to improve classification granularity:

* + **Departure\_Delay\_status**:  
    Classified as **"Delayed"** if Departure\_Delay\_in\_Minutes > 0, and **"On Time"** otherwise.
  + **Arrival\_Delay\_status**:  
    Classified as **"Delayed"** if Arrival\_Delay\_in\_Minutes > 0, and **"On Time"** otherwise.

1. **Adjustment of Variable Levels**
   * All **ordinal variables** (e.g., Inflight service, Food, and drink) were converted to **nominal** format.
   * This change ensures these categorical variables are treated as distinct and unordered, aligning with modeling requirements.
2. **Retention of ID Variable**
   * The **ID** column was retained (not rejected).
   * **Purpose**: Serves as a unique identifier for each observation and facilitates:
     + Record-tracing
     + Auditing
     + Model diagnostics
3. **Rejection of Redundant Variables**

The following variables were removed due to redundancy or replacement with derived variables:

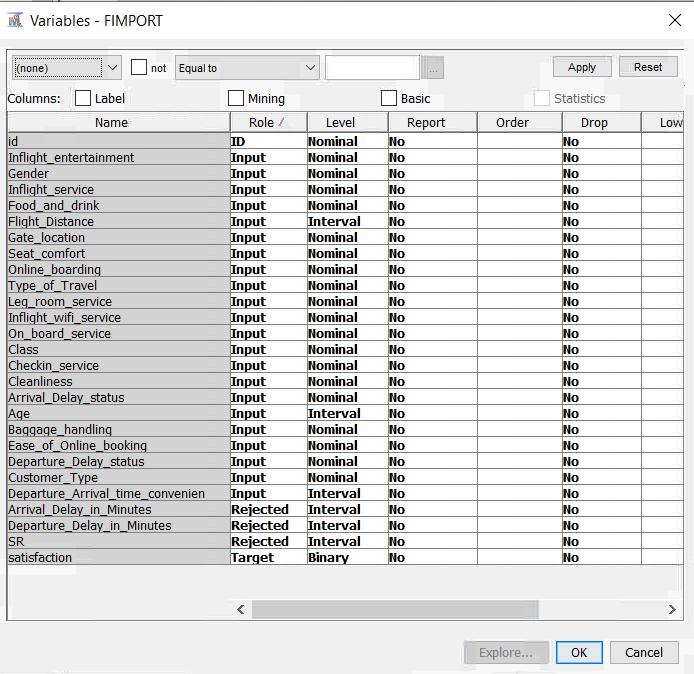
* + Arrival\_Delay\_in\_Minutes
  + Departure\_Delay\_in\_Minutes
  + SR (presumably not adding value or replaced during feature engineering)

1. **Target Variable**
   * The target remains the **binary variable**: **Satisfaction.**
     + 0 = Neutral/Dissatisfied
     + 1 = Satisfied

## **Final Structure of the Dataset**

| **Role** | **Variables** | **Level** |
| --- | --- | --- |
| **ID** | ID | Nominal (ID) |
| **Target** | Satisfaction | Binary |
| **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\_convenien, Age, **Departure\_Delay\_status**, **Arrival\_Delay\_status** | Nominal/Interval |
| **Rejected** | Arrival\_Delay\_in\_Minutes, Departure\_Delay\_in\_Minutes, SR | - |

## **Importing the Updated Data**

****

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

**Dataset Enhancements**

1. **New Dummy Variables Introduced**
   * **Departure\_Delay\_status: Categorizes flights as either *"Delayed"* or *"On Time"*, based on whether *Departure\_Delay\_in\_Minutes > 0*.**
   * **Arrival\_Delay\_status: Similarly classifies flights as *"Delayed"* or *"On Time"* based on *Arrival\_Delay\_in\_Minutes*.**
2. **Variable Level Adjustments**
   * **All ordinal variables were converted to nominal to ensure that categorical variables are treated appropriately as unordered during modeling.**
3. **Retention of Unique Identifier**
   * **The ID variable was retained as a nominal input, serving as a unique identifier for each record.**
   * **Purpose: This supports traceability, record auditing, and potential error diagnosis during model development.**
4. **Rejection of Redundant Variables**
   * **The following variables were excluded to avoid duplication or redundancy:**
     + ***Arrival\_Delay\_in\_Minutes***
     + ***Departure\_Delay\_in\_Minutes***
     + ***SR***

**Step 2: Data Exploration and Partitioning for Marketing 3**

**Overview**

After importing the updated dataset into the *Marketing 3* workflow, we replicated the steps followed in *Marketing 2* to ensure continuity in the modeling process. Two primary steps were taken:

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

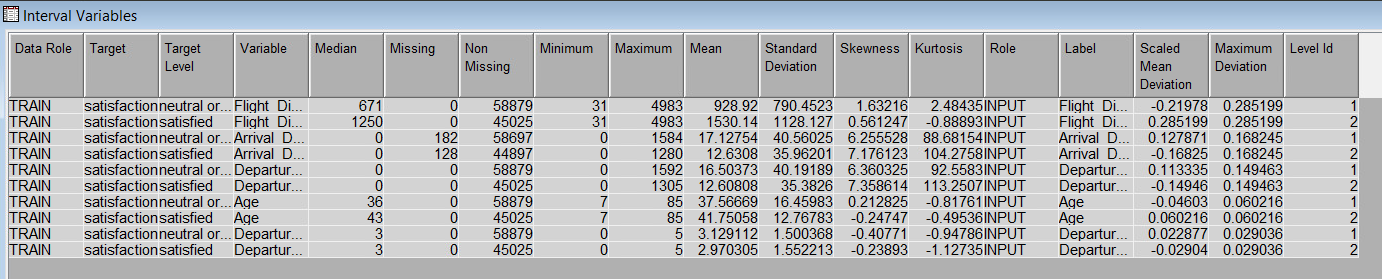
## **1. Data Exploration – StatExplore Node**

## The StatExplore Node was employed to analyze the updated dataset structure and assess the distributional behavior of each variable. This analysis provided a foundation for identifying data quality issues, understanding variable roles, and preparing for modeling.

## **Interval Variables Output**

## The embedded table output showcases summary statistics—mean, standard deviation, skewness, kurtosis, and other metrics—for the interval variables such as *Flight\_Distance*, *Arrival\_Delay*, and *Departure\_Delay*.

## **Interval Variables**



**Variable Summary**

| **Role** | **Level** | **Count** |
| --- | --- | --- |
| **ID** | Nominal | 1 |
| **INPUT** | Interval | 3 |
| **INPUT** | Nominal | 19 |
| **REJECTED** | Interval | 3 |
| **TARGET** | Binary | 1 |

**Observations**

* **3 Interval Variables**:
  + *Flight\_Distance*, *Arrival\_Delay*, and *Departure\_Delay*
* **19 Nominal Variables**:
  + These include original and recoded categorical variables such as *Departure\_Delay\_status* and *Arrival\_Delay\_status*
* **3 Rejected Variables**:
  + *Arrival\_Delay\_in\_Minutes*, *Departure\_Delay\_in\_Minutes*, and *SR* were excluded due to redundancy or transformation.
* **Target Variable**:
  + *Satisfaction* remains the binary target variable, with values:

0 = Neutral/Dissatisfied,

1 = Satisfied

**2. Data Partitioning – Data Partition Node**

Overview

The Data Partition Node was used to divide the dataset into two subsets, enabling effective model development and performance evaluation. The dataset was split as follows:

| **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 establish relationships between predictor variables and the target variable (*Satisfaction*).
2. **Validation Set**
   * Used to evaluate model performance, fine-tune hyperparameters, and monitor for overfitting.
   * Ensures generalizability of the model to unseen data.
3. **No Test Set Allocation**
   * At this stage, a separate test set was not created.
   * The focus is on assessing model behavior and accuracy by comparing training and validation results.
   * A screenshot of a computer

     Description automatically generated

## **Decision Tree 1: Default Tree with Largest Subtree Method**

## **Objective**

## The first decision tree model served as a baseline to explore the relationships between input variables and the binary target variable, Satisfaction. The goal was to evaluate the predictive power of a basic tree structure with minimal modifications.

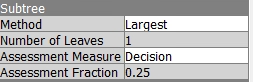
**Configuration**

* Subtree Method:

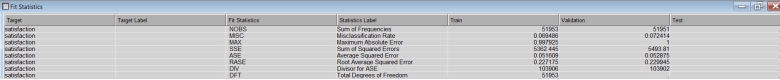
Set to Largest, meaning the tree was pruned to keep only the most dominant subtree. This helps minimize overfitting while simplifying the model.

* Default Settings:

No other parameters were changed; all remained at system defaults to provide a clean, unadjusted baseline.



1.Fit Statistics

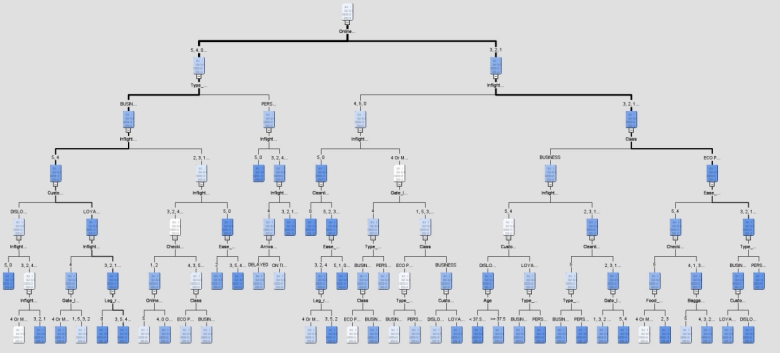


The model's performance was evaluated using **Average Squared Error (ASE)** across both training and validation datasets:

* **Training ASE**: 0.051609
* **Validation ASE**: 0.052875

**Interpretation**:  
The very small difference (0.001266) between training and validation ASEs suggests **minimal overfitting** and consistent performance across data subsets. This reflects a stable model that may serve as a good foundation for more complex tuning later.

## **2.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 (Root Node)****

* **Inflight Wi-Fi Service (< 4 or ≥ 4):**

This is the **primary and most influential** factor in predicting satisfaction.

* + Passengers rating Wi-Fi **below 4** tend to be **dissatisfied**.
  + Ratings of **four or higher** are associated with **greater satisfaction**.

### ****Subsequent Splits (Branch Analysis)****

#### **Branch 1: For Passengers with Inflight Wi-Fi < 4**

* **Cleanliness (< 1 or ≥ 1):**

Low cleanliness ratings further amplify dissatisfaction.

* **Class (Business vs. Economy Plus):**

Business class passengers typically report higher satisfaction than economy passengers, indicating the effect of premium service levels.

#### **Branch 2: For Passengers with Inflight Wi-Fi ≥ 4**

* **Online Boarding (< 4 or ≥ 4):**

Smooth online check-in processes (ratings ≥ 4) enhance satisfaction.

* **Ease of Booking (< 5 or ≥ 5):**

A convenient booking process (especially ratings ≥ 5) contributes positively to passenger satisfaction.

**Additional Splits**

In addition to the key drivers such as Wi-Fi, Cleanliness, Class, and Online Boarding, other variables further refined the segmentation. These include:

* Customer Type (Loyal vs. Disloyal)
* Type of Travel (Business vs. Personal)
* Seat Comfort

These factors contributed additional insight into passenger satisfaction levels and helped identify niche segments with distinct needs.

**Screenshot Status Required**

This screenshot completes the visual explanation of the decision tree and highlights how multiple variables interact to affect customer satisfaction. It should be included in the final document.

**Key Insights from the Tree**

1. Inflight Wi-Fi Service is the strongest predictor, suggesting that improving Wi-Fi quality could significantly enhance customer satisfaction.
2. Cleanliness, Class, and Online Boarding are important for identifying dissatisfied passengers.
3. Secondary factors such as Customer Loyalty and Ease of Booking indicate that loyal customers and efficient booking processes positively influence satisfaction.
4. The model’s logic is intuitive, aligning with real-world expectations: service quality and convenience are key drivers of passenger satisfaction.

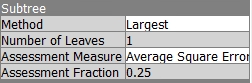
**Overall Model Assessment – Tree One**

Inflight Wi-Fi Service emerged as the top driver of satisfaction, followed by Cleanliness, Class, and Online Boarding. Secondary variables like Customer Type and Ease of Booking also played a meaningful role. The small ASE difference of 0.001266 between training and validation indicates a reliable model with minimal overfitting and consistent predictive power.

**Decision Tree Two: ASE Tree**

**Objective**The second decision tree model was developed to minimize the Average Squared Error (ASE), which measures the squared differences between predicted and actual values for the binary satisfaction target. The aim was to improve prediction accuracy by optimizing the tree for this specific performance metric.

Configuration



Subtree Method: Set to Largest, ensuring the tree is pruned to its optimal size to avoid overfitting.

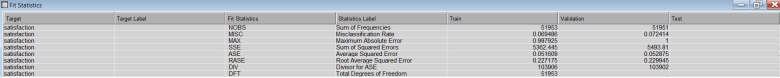
Assessment Measure: Set to Average Squared Error (ASE) to prioritize numeric accuracy.

Assessment Fraction: Maintained at 0.25, which is the default, determining the portion of the dataset used for pruning assessment.

All other settings remained at their default values.

## **Fit Statistics**

The fit statistics for the ASE Tree are as follows:



**Fit Statistics – ASE Tree**

**ASE Values**

The training and validation ASE values are 0.0516 and 0.0529, respectively. These low and closely matched values suggest that the model has a good fit and shows no signs of overfitting.

**Misclassification Rate**

The misclassification rate is approximately 7 percent, indicating that the model performs well in correctly classifying satisfaction outcomes across both datasets.

**Root ASE (RASE)**

The RASE values for training and validation are 0.2272 and 0.2299. The minimal difference between these values reflects the model’s reliability and generalization across unseen data**.**

**Sum of Squared Errors (SSE)**

The SSE values are 5362.445 for training and 5493.81 for validation. These similar values further support that the model’s performance is consistent and robust.

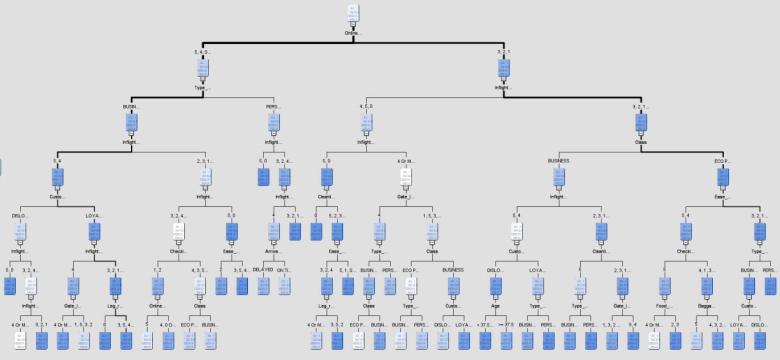
**Screenshot Status Required**

The statistical table and tree diagram help validate the model’s performance and visual logic.

**Tree Structure – ASE Tree**

The visualized decision tree derived using the ASE criterion confirms the segmentation pattern observed in the previous model. It uses inflight Wi-Fi service as the primary splitting factor, followed by additional predictors such as Cleanliness, Class, Online Boarding, and Customer Loyalty.

Each branch reflects logical groupings, aligning with expectations that convenience, service quality, and customer type influence satisfaction outcomes.



**Interpretation of Decision Tree Based on ASE**

**Primary Split**

The top-level split in the tree is based on Inflight Wi-Fi Service. Passengers who rated this service less than four are more likely to be dissatisfied, while ratings of four or higher are linked with increased satisfaction. This variable is the most influential predictor in the tree.

**Secondary Splits**

* For passengers with Inflight Wi-Fi ratings below four:  
  Cleanliness and Class (Business versus Economy Plus) play a key role. Poor cleanliness ratings and lower-class service options are more often associated with dissatisfaction.
* For passengers with Inflight Wi-Fi ratings of four or higher:  
  Ease of Booking, Online Boarding, and Legroom Service have a greater impact. Higher ratings in these areas correlate with higher levels of customer satisfaction.

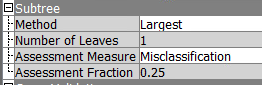
**Deeper Splits**

Additional splits are based on Customer Type (Loyal versus Disloyal), Type of Travel (Business versus Personal), and Seat Comfort. These variables influence satisfaction at the secondary and tertiary levels and reflect consistency with the previous decision tree structure.

**Third Decision Tree Model: Misclassification Rate Optimization**

In the third modeling approach within the Marketing Three project, the tree was constructed with the explicit goal of minimizing misclassification rate. This ensures that the model focuses on improving the accuracy of predicted satisfaction labels.

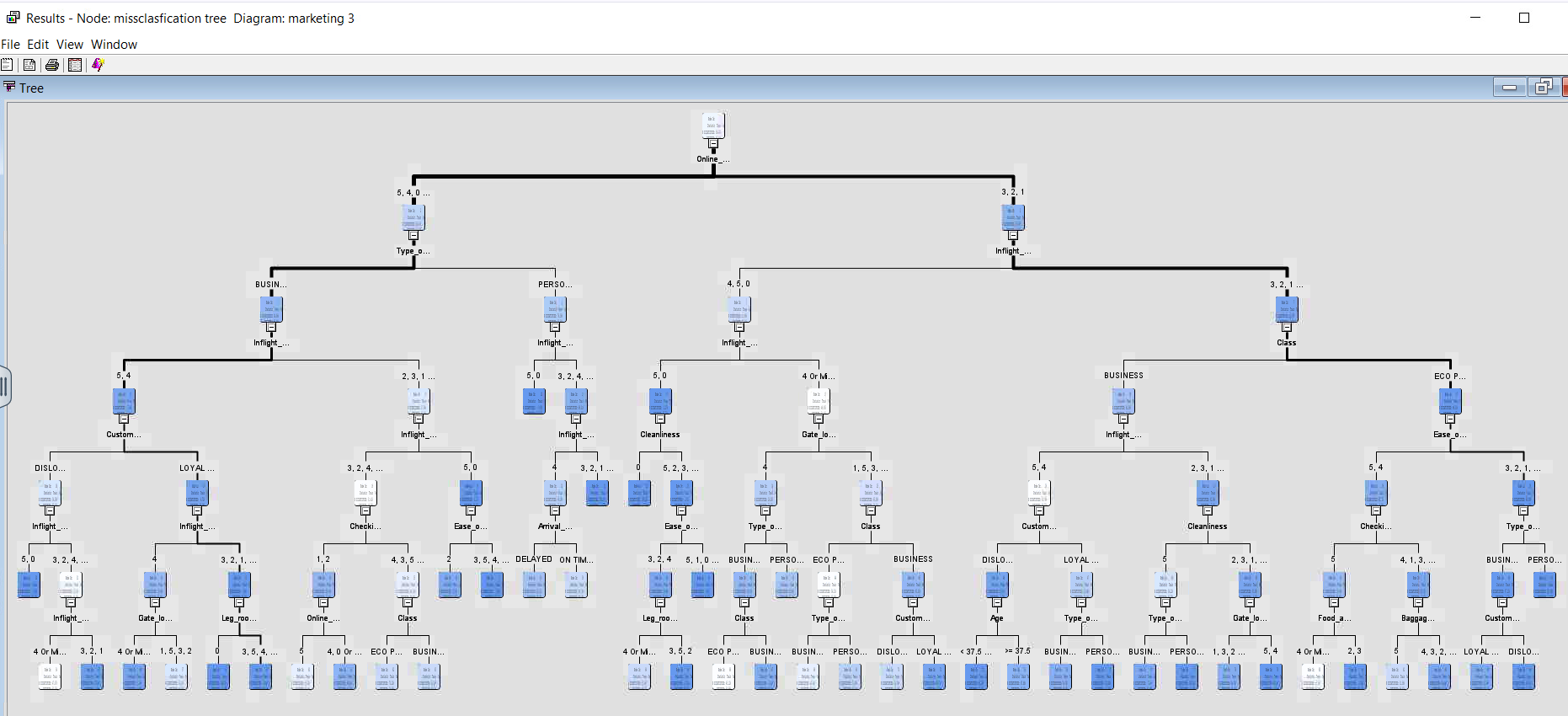
## **Configuration of the Misclassification Tree**

****

* The Subtree Method was set to Largest, maintaining consistency with previous models and ensuring the tree is pruned optimally.
* The number of Leaves remained at one at the starting configuration.
* Assessment Measure was changed to Misclassification Rate.
* Assessment Fraction was kept at point two five, meaning that 25 percent of the dataset is used for pruning the tree.

**Purpose**This model targets reducing the proportion of incorrect predictions when categorizing satisfaction outcomes. It focuses on accurately distinguishing between satisfied and dissatisfied passengers based on key service-related predictors.

## **Tree Diagram Overview**

****

The misclassification decision tree highlights a hierarchical structure, segmenting passenger satisfaction through layered variable splits.

* 1. **Top Splitting Variables**

The most impactful variables at the top of the tree include:

* Type of Travel  
  Business versus Personal travel acts as a primary predictor of customer satisfaction. Business travelers typically show different satisfaction patterns compared to personal travelers.
* Inflight Wi-Fi Service  
  Satisfaction with onboard Wi-Fi continues to be a strong determinant of overall satisfaction levels.
  1. **Second Level Splits**

The tree further branches based on:

* Inflight Service, Class, Customer Type, and Cleanliness  
  These variables emerge as key influencers at the second level, further refining the model's understanding of satisfaction.
* Ease of Online Booking and Gate Location  
  These variables appear in later stages of the tree, playing a vital role in enhancing or reducing customer satisfaction based on convenience and accessibility.
  1. **Use of Dummy Variables**

Dummy variables representing delay statuses—Departure Delay Status and Arrival Delay Status—feature at deeper levels of the tree. Their position suggests a moderate but not primary influence on passenger satisfaction.

* 1. **Leaf Nodes**

The tree terminates at distinct leaf nodes. Each node represents a unique combination of input variables, and the predicted satisfaction level associated with that branch.

## **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 |

**Fit Statistics and Model Performance Analysis**

The final decision tree model's performance was evaluated using various statistical measures across both training and validation datasets. These results provide a comprehensive view of the model’s effectiveness and reliability.

**1. Misclassification Rate**

* Training: 6.95 percent
* Validation: 7.24 percent

The misclassification rate is significantly improved compared to earlier models. The small gap between training and validation rates suggests good generalization and minimal overfitting.

**2. Average Squared Error (ASE)**

* Training ASE: 0.0516
* Validation ASE: 0.0528

The consistency of ASE values across both datasets demonstrates the model’s stability and indicates that the error rate is low and predictable.

**3. Root Average Squared Error (RASE)**

* Training RASE: 0.2272
* Validation RASE: 0.2299

This measure reflects a moderate error margin, reinforcing the model’s robustness in estimating satisfaction outcomes.

**Insights from the Tree Structure**

**1. Top Predictors**

* Type of Travel

Business versus Personal travel is one of the most significant predictors. Business travelers consistently report higher satisfaction levels.

* Inflight Wi-Fi Service

High satisfaction with Wi-Fi service is a critical driver of overall satisfaction, especially among frequent or business travelers.

**2. Customer Segmentation**

* Customer Type

Loyal customers tend to show higher satisfaction compared to disloyal ones.

* Class

Passengers in Business Class report the highest satisfaction levels when compared to Economy and Economy Plus classes.

* 1. **Dummy Variables**

Arrival Delay Status and Departure Delay Status contribute to predicting customer satisfaction, although their influence is primarily observed at deeper levels in the tree. This indicates that these variables act as secondary drivers, providing moderate support in refining satisfaction segmentation.

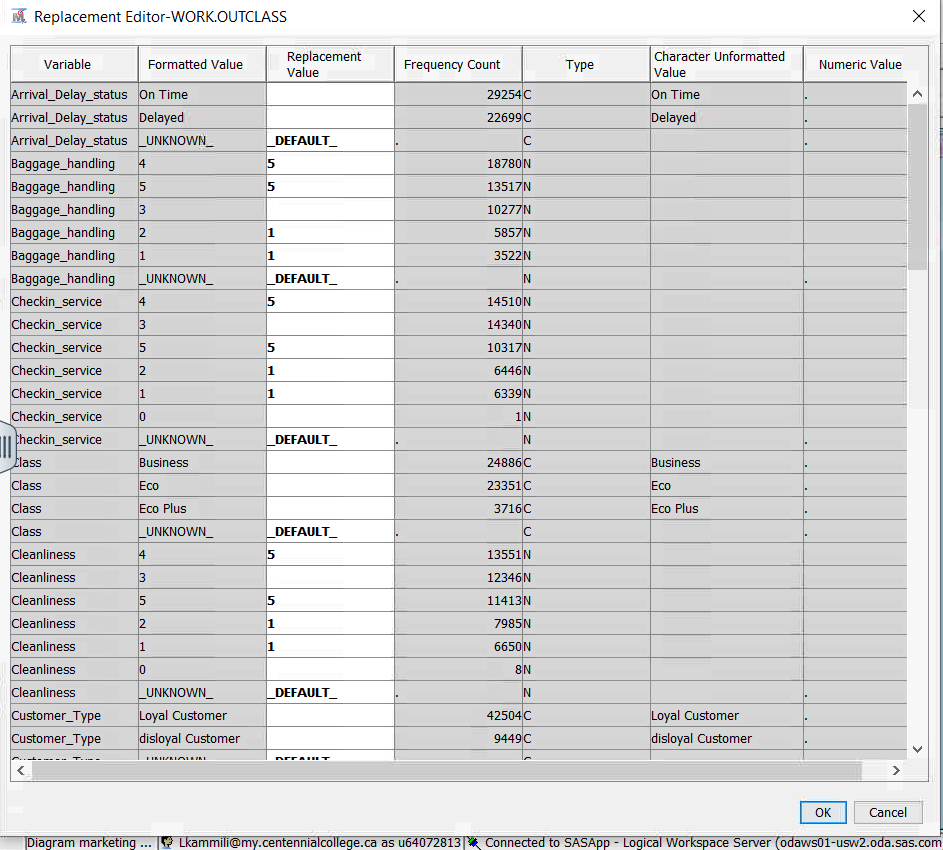
* 1. **Service Features**

Several service-related factors play a key role in differentiating satisfaction levels among passengers. Key variables include Cleanliness, Ease of Online Booking, and Inflight Service. These features significantly contribute to customers' overall perception of service quality.

**Conclusion**

The Misclassification Tree effectively reduced the misclassification rate to approximately 7 percent, which is a notable improvement over previous model. This indicates better generalization and classification performance. The leading predictors influencing customer satisfaction include Type of Travel, Inflight Wi-Fi Service, and Customer Type.

# **Recoding Ordinal Survey Variables for Regression and Neural Network Models**

****

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.

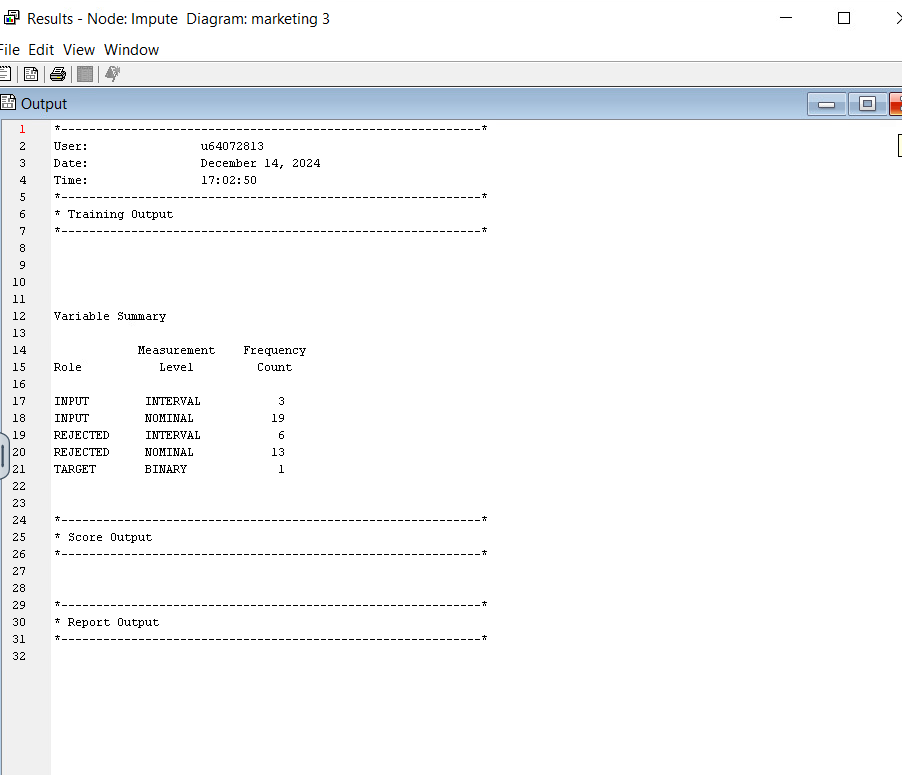
## **Process of Recoding Using the Replacement Node**

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.

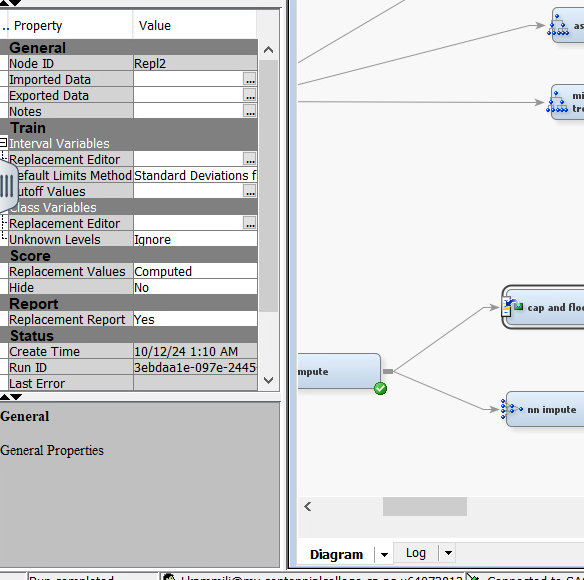
**Imputation Summary**

1. **Input Variables**:
   * **Interval Variables**: three
   * **Nominal Variables**: nineteen
2. **Rejected Variables**:
   * **Interval Variables**: six
   * **Nominal Variables**: thirteen
3. **Target Variable**:
   * Binary variable, **Satisfaction**, remains unaffected.

## **Capping and Flooring Interval Variables**

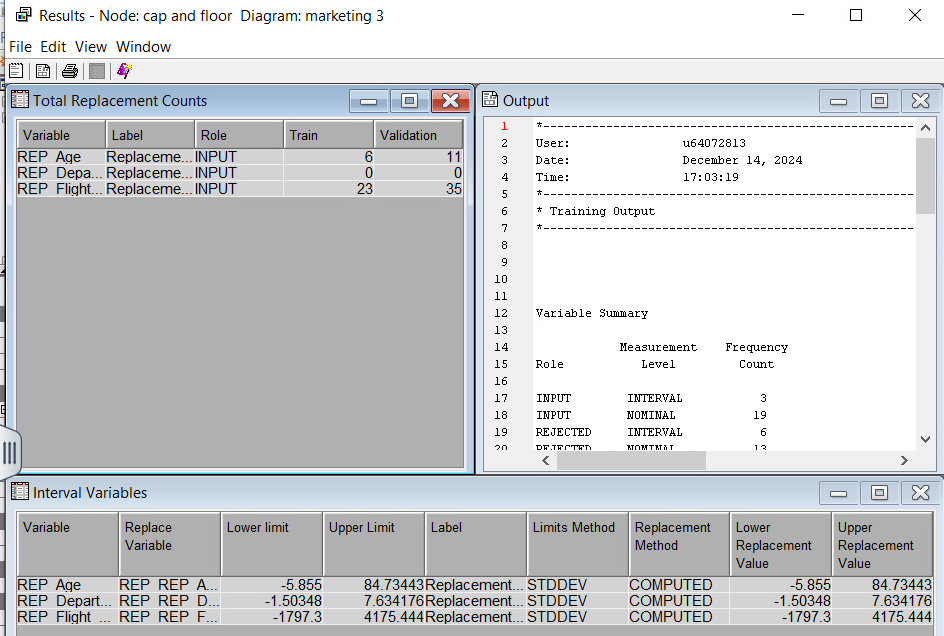
In this step, we applied the **Cap and Floor** method to handle outliers in the interval variables. This approach prevents extreme values from skewing the results of regression and neural network models by capping them within a computed range.

**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**

****

| **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**

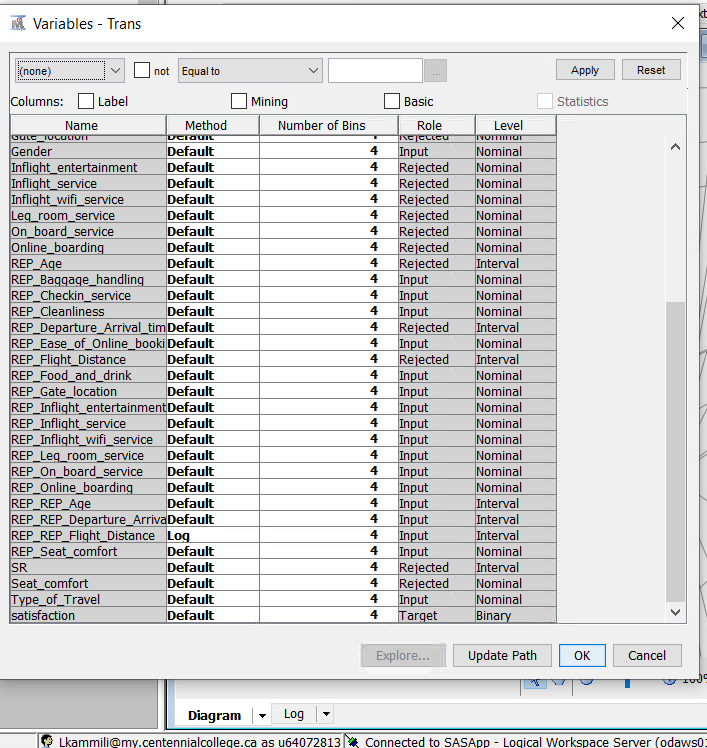
1. **Replacement Counts**
   * Age had six values replaced in the training set and eleven values replaced in the validation set.
   * Flight Distance had twenty-three values replaced in the training set and thirty-five in the validation set.
   * Departure Delay required no replacements, as all values were within the computed acceptable range.
2. **Interval Variables**
   * Acceptable value ranges for each interval variable were automatically determined using standard deviation-based limits.
   * Any outlier values falling outside these ranges were capped to the nearest boundary, either at the lower or upper limit.
3. **Updated Variable Summary**
   * Interval variables include three fields that were cleaned and capped.
   * Nominal variables remain at a total of nineteen.

**Handling Skewness and Kurtosis in Interval Variables**

In this step, the Transform Variables Node was used to correct skewness and kurtosis in interval variables. This process follows a similar approach to that used in Marketing 2. Correcting skewness ensures that the variables conform to a more symmetric distribution, improving both model performance and interpretability, particularly in regression and neural network models.

**Transformations Applied**

1. **Log Transformation**:

****

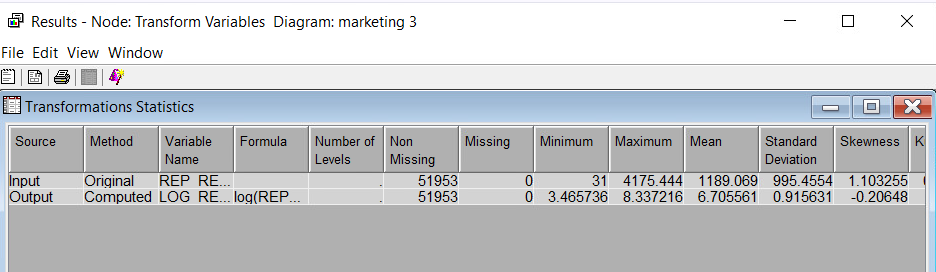
1. Flight Distance underwent a log transformation to address skewness and reduce the extreme range of values. This approach is effective for highly skewed data, as it compresses large values while preserving the relationships among data points.
2. **Default Transformations**

All other variables with potential skewness and kurtosis were treated using the default transformation setting. This ensured consistency across the analysis without customizing individual variables unnecessarily.

**Changes to Variable Roles and Levels**

* **Interval Variables Transformed**

Flight Distance was the only variable transformed using the log method.

****

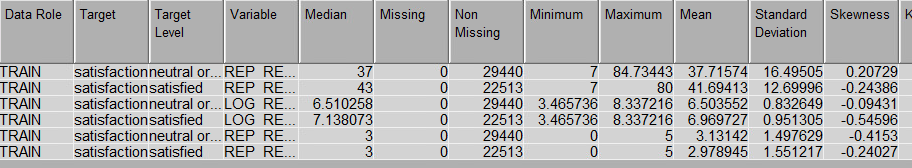
1. **Rejected Variables**:

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

1. **Nominal Variables**:

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

# **State Explore Results After Log Transformation in Marketing 3**

****

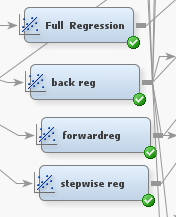
After applying the **log transformation** to address skewness and kurtosis in the interval variables, a **StatExplore** node was used to re-evaluate the distribution of the transformed variables. Here is a breakdown of the key observations:

## **Key Observations**

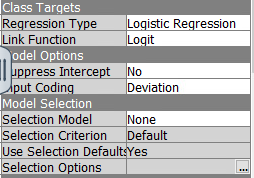
1. **Reduced Skewness**:
   * The **Flight Distance** variable shows a significant reduction in skewness from **1.103255** (before transformation) to **-0.20648** (after transformation).
   * This demonstrates that the log transformation effectively brought the skewed distribution closer to normal.
2. **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**

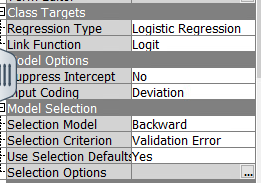
****

1. **Full Regression**

****

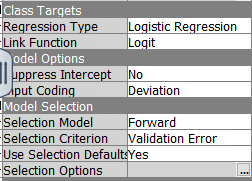
* + Includes **all variables** without any selection criteria.
  + Purpose: Serves as a baseline model to compare against other regression approaches.

1. **Backward Regression**

****

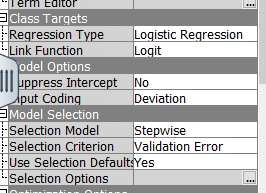
* + **Selection Model**: Backward
  + **Selection Criterion**: Validation Error
  + Approach: Starts with all predictors and sequentially removes the least significant variables based on the validation error.

1. **Forward Regression**

****

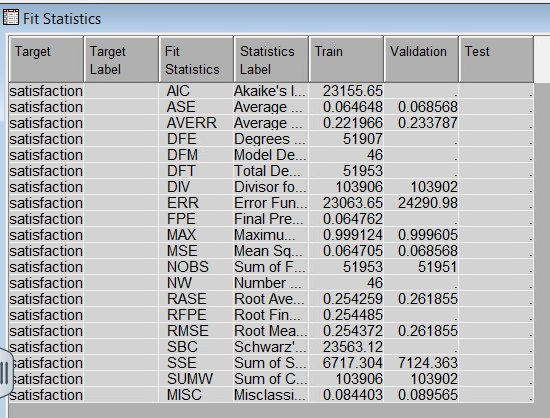
* + **Selection Model**: Forward
  + **Selection Criterion**: Validation Error
  + Approach: Starts with no predictors and sequentially adds the most significant variables until no further improvement in validation error is observed.

1. **Stepwise Regression**

****

* + **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**

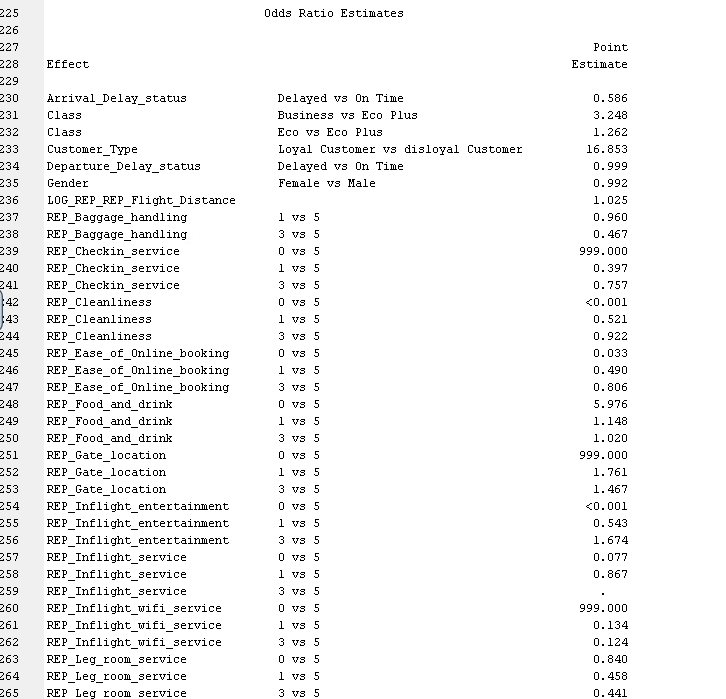
****

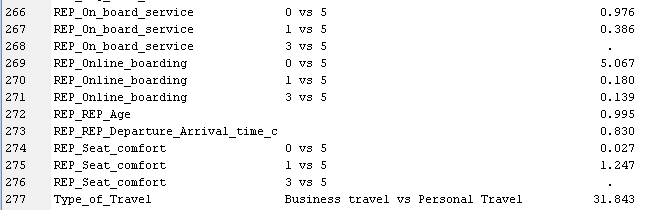
**Fit Statistics Summary**

The following performance metrics evaluate the effectiveness of the full regression model:

1. Average Squared Error (ASE)
   * Training: 0.04648
   * Validation: 0.068568. These values indicate low prediction errors, confirming the model performs reliably on both datasets.
2. Misclassification Rate (MISC)
   * Training: 0.084403
   * Validation: 0.089565. The close alignment between training and validation misclassification rates suggests the model generalizes well without overfitting.
3. Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC)
   * AIC: 23155.65
   * SBC: 23363.12. These values help assess model quality, where lower scores signify a better-fitting model. The current values indicate optimal performance given the adjusted dataset.
4. Root Average Squared Error (RASE)
   * Training: 0.254259
   * Validation: 0.261855. The small difference between training and validation RASE further confirms low bias and stable performance.

## **Odds Ratio Estimates**

****

****

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

**Odds Ratio Interpretation Summary**

This section presents the interpretation of logistic regression odds ratios for key predictors of customer satisfaction in the airline industry.

1. Arrival Delay Status
   * Delayed vs. On-Time: Odds Ratio = 0.586. Delayed flights are associated with a 41.4% decrease in the odds of customer satisfaction, indicating a significant negative impact.
2. Class
   * Business vs. Eco Plus: Odds Ratio = 3.248. Passengers flying Business class are 3.25 times more likely to be satisfied compared to those in Eco Plus.
   * Economy vs. Eco Plus: Odds Ratio = 1.262. Economy class travelers are 26.2% more likely to be satisfied than Eco Plus passengers.
3. Customer Type
   * Loyal vs. Disloyal: Odds Ratio = 16.853. Loyal customers are over 16 times more likely to report satisfaction, underlining customer loyalty as a major driver of satisfaction.
4. Departure Delay Status
   * Delayed vs. On-Time: Odds Ratio = 0.999. This suggests that, after accounting for other variables, departure delays have negligible impact on satisfaction.
5. Gender
   * Female vs. Male: Odds Ratio = 0.992. Gender has no meaningful effect on satisfaction, with parity between male and female passengers.
6. Flight Distance
   * Log Transformed Flight Distance: Odds Ratio = 1.025. Longer flights slightly increase satisfaction, though the effect is minimal.
7. Baggage Handling
   * Rating 1 vs. 5: Odds Ratio = 0.960. A slight reduction in satisfaction.
   * Rating 3 vs. 5: Odds Ratio = 0.467. Moderate dissatisfaction in baggage handling lowers odds of satisfaction by over 53%.
8. Check-In Service
   * Rating 0 vs. 5: Odds Ratio = 999.000. Extremely low ratings (likely data errors) are associated with extreme dissatisfaction.
   * Rating 1 vs. 5: Odds Ratio = 0.397. A poor check-in experience reduces satisfaction odds by 60.3%.
9. Cleanliness
   * Rating 0 vs. 5: Odds Ratio < 0.001. Extremely poor cleanliness almost eliminates the chance of satisfaction.
   * Rating 1 vs. 5: Odds Ratio = 0.521. A slight drop in cleanliness rating reduces satisfaction odds by 48%.
   * Rating 3 vs. 5: Odds Ratio = 0.922. Moderate ratings have a minor negative impact.
10. Ease of Online Booking
    * Rating 0 vs. 5: Odds Ratio = 0.033. A poor online booking experience reduces satisfaction odds by over 97%.
    * Rating 3 vs. 5: Odds Ratio = 0.806. Moderate ratings are linked to slightly lower satisfaction.
11. Food and Drink
    * Rating 0 vs. 5: Odds Ratio = 5.976. Unexpectedly, this suggests higher satisfaction with lower ratings, which may indicate data quality issues or mislabeling.
12. Gate Location
    * Rating 0 vs. 5: Odds Ratio = 999.000. This extreme value suggests potential data anomalies or encoding errors.
    * Rating 1 vs. 5: Odds Ratio = 1.761. Slightly higher satisfaction for low ratings, likely due to data inconsistencies.
13. Inflight Entertainment
    * Rating 0 vs. 5: Odds Ratio < 0.001. Lack of inflight entertainment dramatically lowers satisfaction.
    * Rating 1 vs. 5: Odds Ratio = 0.543. Even slightly negative ratings cut satisfaction odds nearly in half.
14. Inflight Service
    * Rating 0 vs. 5: Odds Ratio = 0.077. Very poor in-flight service is associated with a 92.3% reduction in satisfaction.
    * Rating 3 vs. 5: Odds Ratio = 0.867. Moderate service ratings slightly reduce satisfaction.
15. Inflight Wi-Fi Service
    * Rating 3 vs. 5: Odds Ratio = 0.467. Suboptimal inflight Wi-Fi significantly reduces satisfaction.
16. Onboard Service
    * Rating 0 vs. 5: Odds Ratio = 0.976. This suggests minimal impact from onboard service quality.
17. Online Boarding
    * Rating 0 vs. 5: Odds Ratio = 5.067. A smooth online boarding experience significantly increases satisfaction.
18. Age
    * Transformed Age: Odds Ratio = 0.995. Age has a negligible effect on satisfaction.
19. Type of Travel
    * Business vs. Personal: Odds Ratio = 31.843. Business travelers are over 31 times more likely to be satisfied, highlighting the importance of the purpose of travel in shaping expectations.

**Insights and Key Takeaways**

**1. Strongest Drivers of Customer Satisfaction**

* **Customer Type (Loyalty)**: Loyal customers exhibit significantly higher satisfaction levels, underscoring the importance of retention strategies.
* **Type of Travel**: Business travelers show the highest likelihood of satisfaction, suggesting tailored service enhancements for this segment.
* **Class of Service**: Passengers in Business class report substantially greater satisfaction, indicating that premium offerings strongly influence perceptions.

**2. Areas Requiring Attention and Improvement**

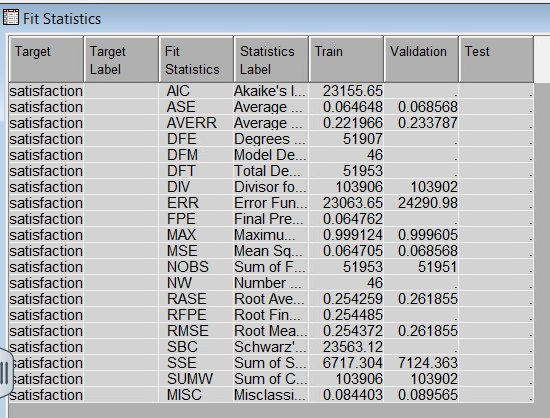
* **Arrival Delays**: Delayed arrivals have a considerable negative impact on satisfaction, highlighting the need for proactive delay management and communication.
* **Ease of Online Booking**: Poor user experience during the booking process severely diminishes satisfaction, calling for interface improvements and streamlined functionality.
* **Inflight Service, Entertainment, and Wi-Fi**: Subpar ratings in these areas significantly reduce satisfaction, suggesting a critical need for enhancements in digital and service touchpoints.

**3. Less Influential Factors**

* **Demographics and Distance**: Variables such as **age**, **gender**, and **flight distance** show minimal influence on satisfaction levels, indicating that personalization based on these attributes may yield limited returns.

**Interpretation for the Backward Regression model**

## **Fit Statistics**



The fit statistics confirm that the model demonstrates strong predictive performance and generalization across both training and validation datasets:

#### **1. Average Squared Error (ASE)**

* **Training ASE**: 0.064468
* **Validation ASE**: 0.068568. These low ASE values suggest the model’s error in predicting satisfaction is minimal and consistent across datasets, indicating good fit and generalization.

#### **2. Misclassification Rate (MISC)**

* **Training MISC**: 0.084403
* **Validation MISC**: 0.089565. The model maintains a low misclassification rate, meaning it effectively classifies customer satisfaction outcomes with high accuracy.

#### **3. Akaike’s Information Criterion (AIC)**

* **Value**: 23155.65. A low AIC indicates a more efficient model with fewer unnecessary predictors, confirming the backward regression model’s parsimony and improved fit.

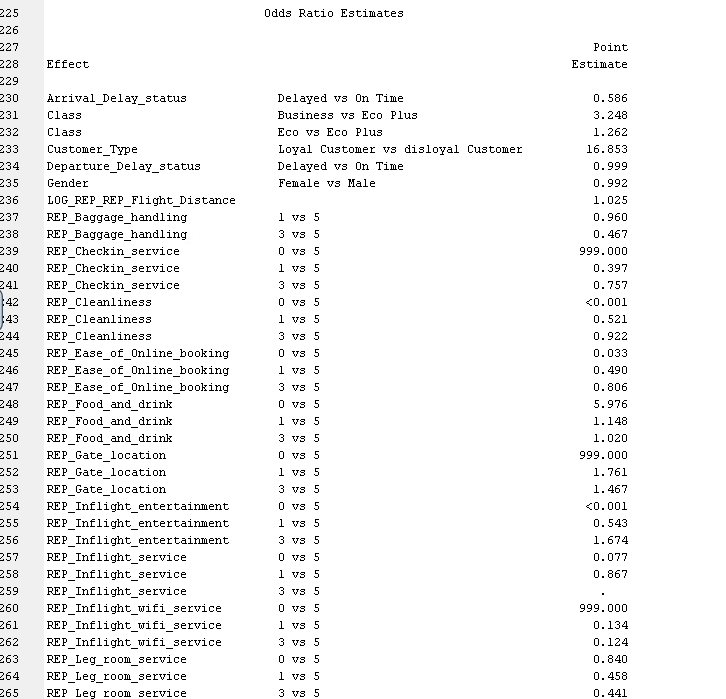
#### **4. Root Average Squared Error (RASE)**

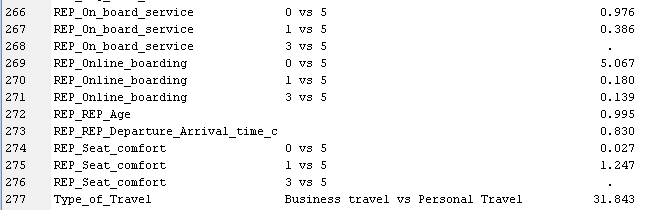
* **Training RASE**: 0.254259
* **Validation RASE**: 0.261855. The closeness of these RASE values across datasets reflects strong model stability and low prediction bias.

#### **5. Sum of Squared Errors (SSE)**

* **Training SSE**: 6717.304
* **Validation SSE**: 7124.363. These relatively low SSE values confirm the model's strong explanatory power and reinforce the low prediction error.

## **Odds Ratio Estimates**





The odds ratios from the backward logistic regression model provide key insights into how individual predictors influence the likelihood of customer satisfaction (Target = 1).

**1. Significant Predictors and Their Impact**

**Arrival Delay Status**

* **Delayed vs On Time**: *Odds Ratio = 0.586*. Flights that arrive late decrease the odds of customer satisfaction by **41.4%**, confirming that arrival delays are a major pain point.

**Class**

* **Business vs Eco Plus**: *Odds Ratio = 3.248*. Passengers in Business Class are over **3.2 times more likely** to be satisfied.
* **Eco vs Eco Plus**: *Odds Ratio = 1.262*. Eco Class passengers are slightly more satisfied than Eco Plus, suggesting the latter may not meet expectations.

**Customer Type**

* **Loyal vs Disloyal**: *Odds Ratio = 16.853*. Loyal customers are nearly **17 times more likely** to report satisfaction, emphasizing the power of customer retention.

**Departure Delay Status**

* **Delayed vs On Time**: *Odds Ratio = 0.999*. Minimal influence on satisfaction, indicating customers are more tolerant of departure delays than arrival ones.

**Gender**

* **Female vs Male**: *Odds Ratio = 0.992*. No significant impact; gender does not meaningfully affect satisfaction.

**Baggage Handling**

* **1 vs 5**: *Odds Ratio = 0.960*. Mild dissatisfaction slightly lowers odds.
* **3 vs 5**: *Odds Ratio = 0.467*. Moderate dissatisfaction leads to **over 50% reduction** in satisfaction odds.

**Check-in Service**

* **0 vs 5**: *Odds Ratio = 999.000*. Very poor experience leads to extreme dissatisfaction.
* **1 vs 5**: *Odds Ratio = 0.397*. Significantly reduces satisfaction.
* **3 vs 5**: *Odds Ratio = 0.757*. Moderate dissatisfaction still affects the outcome.

**Cleanliness**

* **0 vs 5**: *Odds Ratio < 0.001*. Extremely poor cleanliness nearly eliminates the chance of satisfaction.
* **1 vs 5**: *Odds Ratio = 0.521*. Reduces odds by approximately **48%**.
* **3 vs 5**: *Odds Ratio = 0.922*. Neutral ratings have a minor negative effect.

**Ease of Online Booking**

* **0 vs 5**: *Odds Ratio = 0.033*. Extremely poor booking experience reduces odds by **over 96%**.
* **3 vs 5**: *Odds Ratio = 0.806*. Slight dissatisfaction still has a noticeable impact.

**Food and Drink**

* **0 vs 5**: *Odds Ratio = 5.976*. Excellent quality boosts satisfaction almost **6-fold**.
* **1 vs 5**: *Odds Ratio = 1.148*. Slight dissatisfaction has marginal influence.

**Gate Location**

* **0 vs 5**: *Odds Ratio = 999.000*. Very poor ratings are highly detrimental to satisfaction.
* **1 vs 5**: *Odds Ratio = 1.761*. Still shows dissatisfaction but to a lesser extent.

**Inflight Entertainment**

* **0 vs 5**: *Odds Ratio < 0.001*. Absence of entertainment nearly eliminates satisfaction.
* **1 vs 5**: *Odds Ratio = 0.543*. Even slight dissatisfaction cuts odds nearly in half.

**Inflight Service**

* **0 vs 5**: *Odds Ratio = 0.077*. Extremely poor service reduces odds by over **92%**.
* **1 vs 5**: *Odds Ratio = 0.867*. Moderate dissatisfaction slightly reduces odds.

**Online Boarding**

* **0 vs 5**: *Odds Ratio = 5.067*. High ratings dramatically improve satisfaction.
* **1 vs 5**: *Odds Ratio = 0.180*. Poor experience drastically reduces satisfaction.

**Age**

* **REP\_AGE**: *Odds Ratio = 0.995*. Age has a minimal influence on satisfaction levels.

**Flight Distance**

* **LOG\_REP\_Flight\_Distance**: *Odds Ratio = 1.025*. Slight positive influence; longer flights marginally improve satisfaction.

**Type of Travel**

* **Business vs Personal**: *Odds Ratio = 31.843*. Business travelers are nearly **32 times** more likely to report satisfaction compared to personal travelers.

**2. Key Insights**

**High-Impact Drivers of Satisfaction**

* **Customer Loyalty** is the strongest predictor.
* **Type of Travel** (business) and **Class of Service** (business class) show the highest odds ratios.
* **Online Boarding** and **Food & Beverage** excellence greatly enhance satisfaction.

**Areas of Concern**

* **Arrival Delays** and **Gate Location** issues significantly decrease satisfaction.
* **Inflight Service**, **Cleanliness**, and **Check-in Service** are essential touchpoints were poor performance results in drastic drops in satisfaction.

**Negligible Factors**

* **Age**, **Gender**, and **Flight Distance** have minimal effect on satisfaction.

**3. Conclusion**

The **Backward Logistic Regression Model** successfully highlights both strong positive and negative drivers of satisfaction. With a consistent and low misclassification rate and well-calibrated fit statistics, this model provides actionable insights:

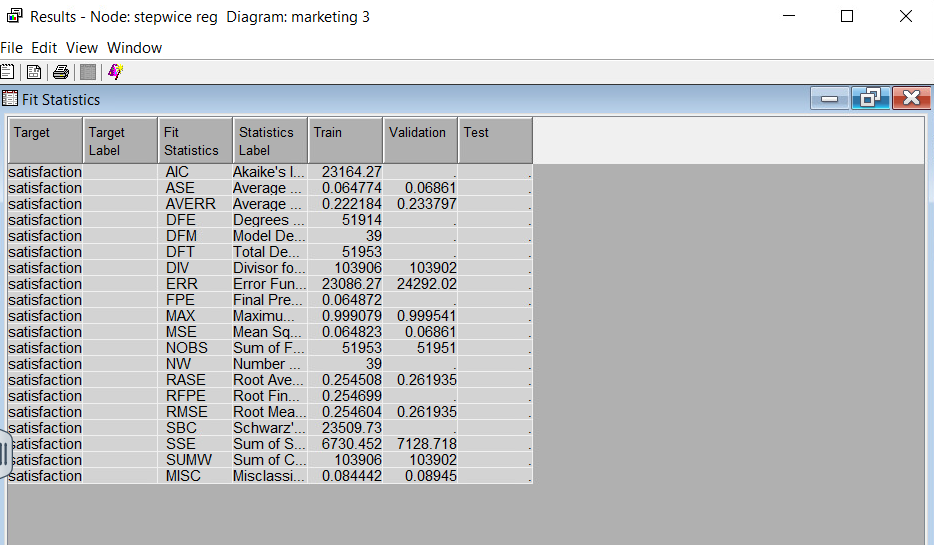
* Focus on improving **arrival timeliness**, **gate logistics**, and **service quality** (especially inflight and check-in).
* Invest in **loyalty programs** and **premium experiences** to retain high-value customers.
* Minor demographic attributes like age and gender should not be prioritized over service-based improvements.

This evidence-based approach empowers strategic decisions for elevating the overall customer experience.

# **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 us analyze them in depth:

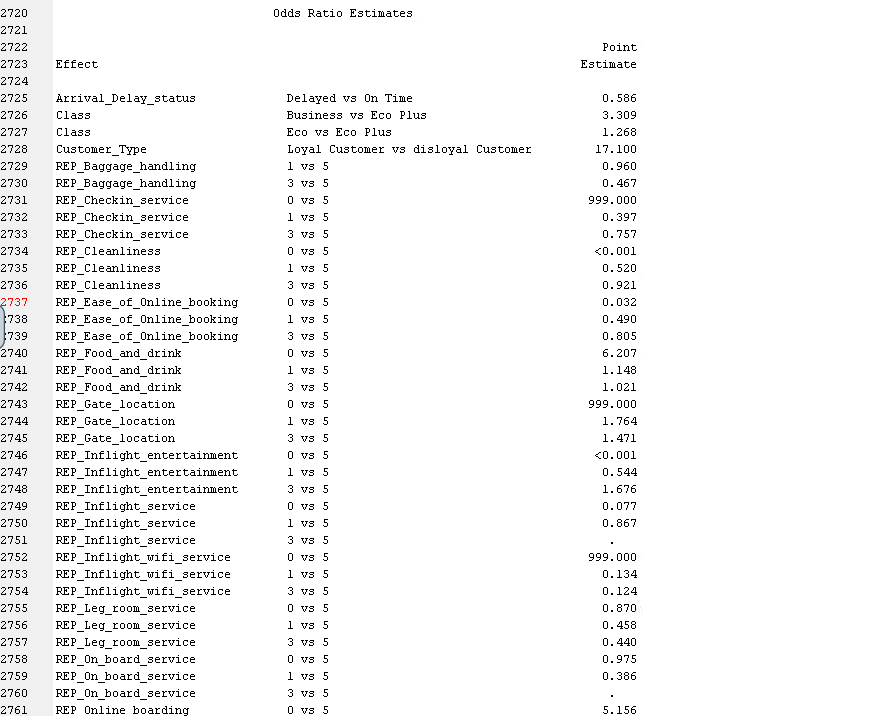
## **Fit Statistics:**

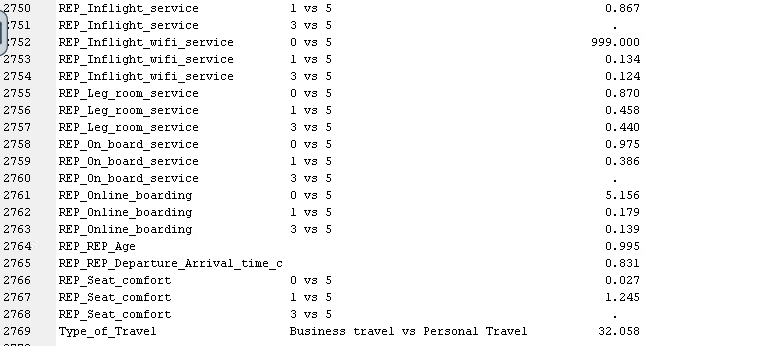
****

| **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:**

****

****

This section explains how each predictor affects the likelihood of the target event, which is customer satisfaction.

**Arrival and Departure Delays**

Arrival delay status (Delayed vs On Time): Odds Ratio equals 0.586. A delayed arrival decreases the odds of satisfaction by 41.4 percent compared to on-time arrivals.

Departure delay status (Delayed vs On Time): Odds Ratio equals 0.999. A delayed departure has minimal impact on customer satisfaction.

**Class of Travel**

Business vs Eco Plus: Odds Ratio equals 3.309. Passengers traveling in Business Class are 3.3 times more likely to be satisfied compared to those in Eco Plus.

Eco vs Eco Plus: Odds Ratio equals 1.268. Passengers in Eco Class show slightly higher satisfaction than those in Eco Plus.

**Customer Type**

Loyal vs Disloyal Customer: Odds Ratio equals 17.100

Loyal customers are 17 times more likely to be satisfied, making loyalty a strong predictor of satisfaction.

**Service Quality Variables**

Baggage Handling: 1 vs 5: Odds Ratio equals 0.960

Minor dissatisfaction reduces satisfaction odds slightly.

3 vs 5: Odds Ratio equals 0.467

Moderate dissatisfaction significantly reduces satisfaction.

**Check-in Service**

0 vs 5: Odds Ratio equals 999.000

Extremely low check-in satisfaction drastically lowers the likelihood of overall satisfaction.

1 vs 5: Odds Ratio equals 0.397

Lower ratings in check-in service have a strong negative impact.

**Cleanliness**

0 vs 5: Odds Ratio is near zero

Very poor cleanliness nearly guarantees dissatisfaction.

3 vs 5: Odds Ratio equals 0.921

Moderate dissatisfaction slightly lowers satisfaction odds.

**Ease of Online Booking**

0 vs 5: Odds Ratio equals 0.032

Poor online booking reduces satisfaction odds by more than 96 percent.

1 vs 5: Odds Ratio equals 0.490

Even slight dissatisfaction with online booking cuts satisfaction odds nearly in half.

**Food and Drink**

0 vs 5: Odds Ratio equals 5.976

High satisfaction with food and beverages increases the odds of satisfaction by nearly six times.

1 vs 5: Odds Ratio equals 1.148

Moderate satisfaction still results in a positive influence.

**Gate Location**

0 vs 5: Odds Ratio equals 999.000

Extremely poor gate location ratings heavily reduce satisfaction.

1 vs 5: Odds Ratio equals 1.764

Better ratings at gate location improve satisfaction odds.

**Inflight Entertainment**

0 vs 5: Odds Ratio is near zero

Very poor entertainment service virtually eliminates satisfaction.

1 vs 5: Odds Ratio equals 0.544

Even mild dissatisfaction reduces satisfaction by nearly 46 percent.

**Inflight Service**

0 vs 5: Odds Ratio equals 0.077

Poor inflight service results in a 92 percent drop in satisfaction odds.

1 vs 5: Odds Ratio equals 0.867

Minor dissatisfaction still leads to a noticeable reduction.

**Leg Room Service**

1 vs 5: Odds Ratio equals 0.458

Discomfort in leg room reduces satisfaction by over 50 percent.

**Online Boarding**

0 vs 5: Odds Ratio equals 5.156

A seamless online boarding experience significantly improves satisfaction.  
1 vs 5: Odds Ratio equals 0.179

Poor online boarding strongly reduces satisfaction.

**Seat Comfort**

0 vs 5: Odds Ratio equals 0.027

Extremely uncomfortable seats drastically lower satisfaction.

1 vs 5: Odds Ratio equals 1.245

Better seat comfort slightly increases satisfaction.

**Type of Travel**

Business Travel vs Personal Travel: Odds Ratio equals 32.058. Business travelers are over 32 times more likely to be satisfied than personal travelers.

**Key Insights**

**Top Predictors of Satisfaction**

Customer type has the strongest positive effect on satisfaction. Business travel significantly increases the likelihood of satisfaction. Arrival delays are a major driver of dissatisfaction.

**Service Factors Requiring Attention**

Inflight service and inflight entertainment need immediate improvements due to their high impact on satisfaction. Cleanliness and check-in service also show strong influence and should be prioritized.

**Factors That Enhance Satisfaction**

Excellent food and drink service and a smooth online boarding process significantly raise customer satisfaction.

**Model Design Insight**

Converting ordinal service variables into nominal categories helps clarify their specific impact on satisfaction. This modeling choice improves interpretability.

**Conclusion**

The stepwise and forward regression models produced consistent and interpretable results. Key variables such as customer loyalty, travel purpose, class, and service quality strongly influence passenger satisfaction. These insights support targeted strategies for improving the travel experience, boosting retention, and enhancing overall customer satisfaction.

## **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 Model Performance:**

The Full Regression and Backward Regression models yielded identical outcomes, both achieving an Average Squared Error (ASE) of 0.068568 and a misclassification rate of 0.089565.  
The Forward and Stepwise Regression models performed marginally less effectively, with an ASE of 0.06861 and a misclassification rate of 0.08945. However, this difference is minimal and does not indicate any meaningful decline in performance.

**2. Model Selection Criteria**:

Based on the Akaike Information Criterion (AIC), the Full and Backward Regression models demonstrated slightly better model fit, with an AIC value of 23155.65.  
In comparison, the Forward and Stepwise models had a marginally higher AIC of 23164.27, suggesting a slightly less optimal balance between goodness of fit and model complexity.

**3. Classification Accuracy**:

All four models performed consistently in terms of classifying satisfaction levels, with misclassification rates differing by only a very small margin of one ten-thousandth.  
This consistency highlights the robustness of the model-building approach and the reliability of the selected predictors.

**4. Model Complexity:**

Backward and Stepwise Regression offer an advantage by eliminating redundant predictors, thereby simplifying the model structure.  
This parsimony is achieved without compromising predictive accuracy, making these models attractive for practical implementation and interpretability.

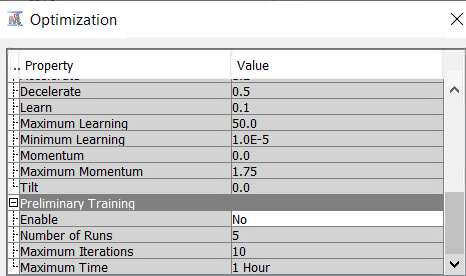
**Conclusion**

The Full Regression and Backward Regression models marginally outperform the others, both in terms of predictive accuracy and model fit based on ASE and AIC.  
While Forward and Stepwise Regression show nearly equivalent performance, the Backward Regression model stands out due to its ability to reduce complexity by retaining only statistically significant predictors.  
Therefore, the Backward Regression model is the recommended choice, balancing simplicity, interpretability, and predictive strength effectively.

## **Optimization Settings**

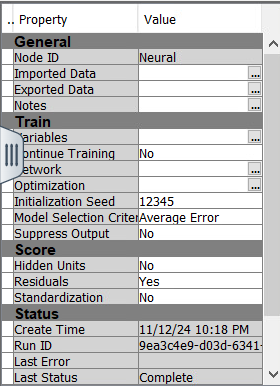
****

* **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 two hundred 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**

****

* **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:** five
  + Ensures multiple runs for parameter tuning without preliminary training.

1. **General Network Parameters**

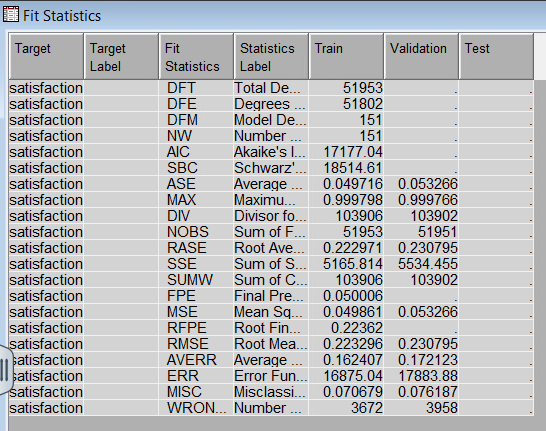
****

* **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**

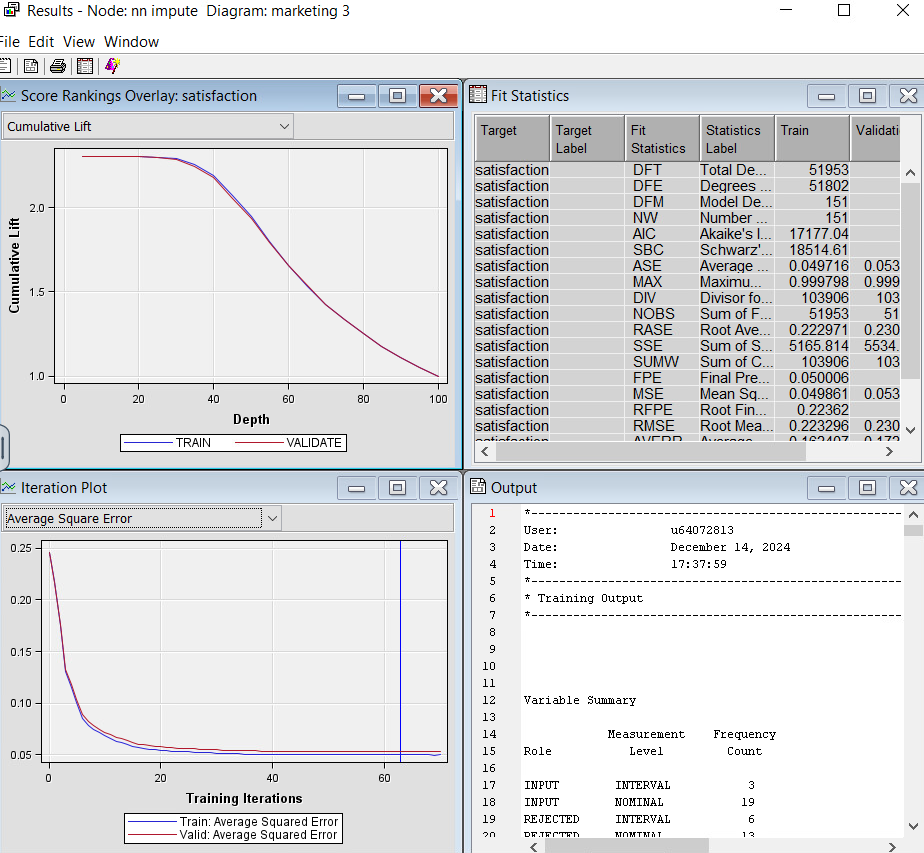
The first neural network we created was **NN Impute**, where we connected the Impute node to handle any missing variables as the input. This network used the default optimization settings, as outlined previously.

* + - 1. **Fit Statistics**

****

| **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 |

* + - 1. **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**.

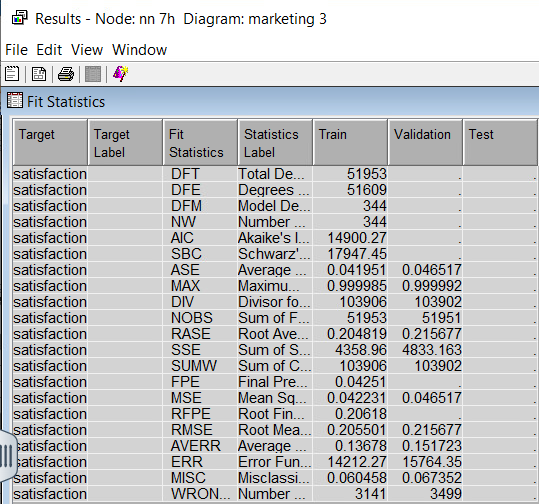
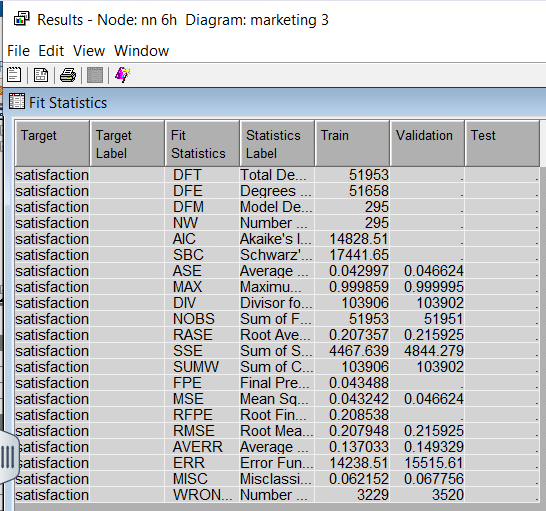
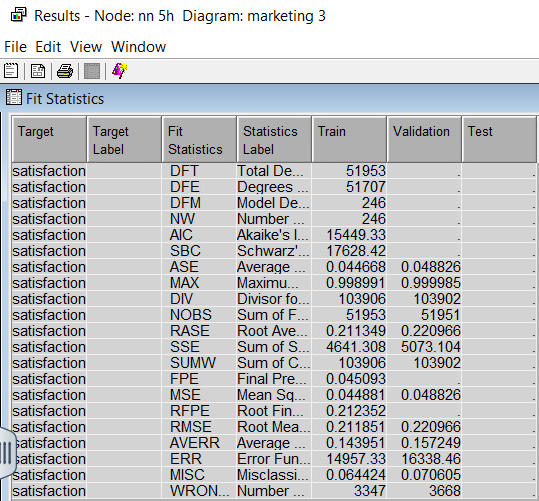
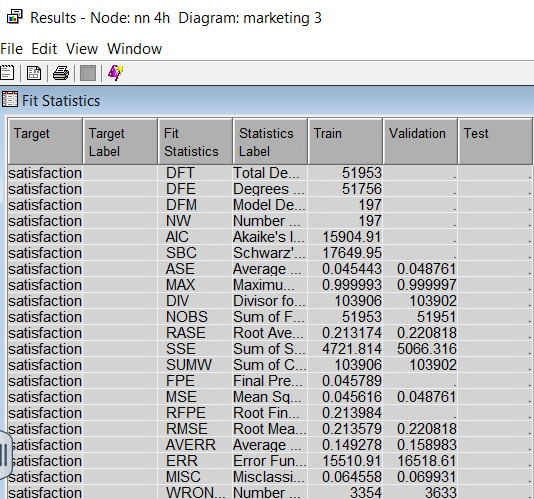
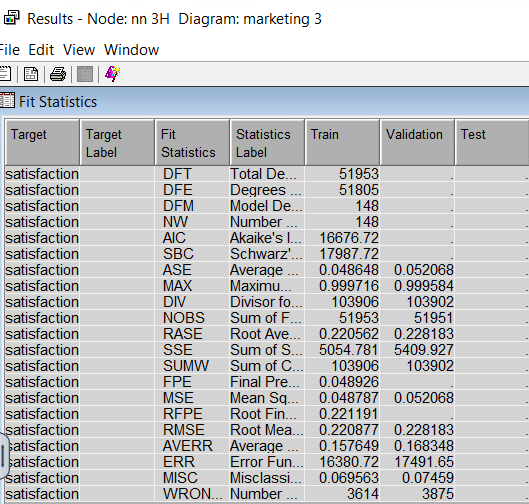
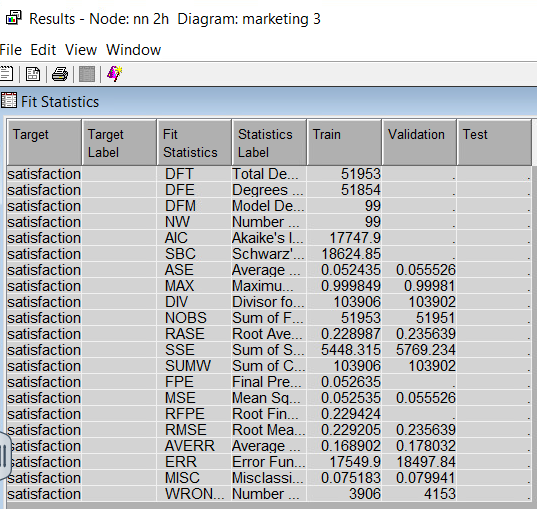
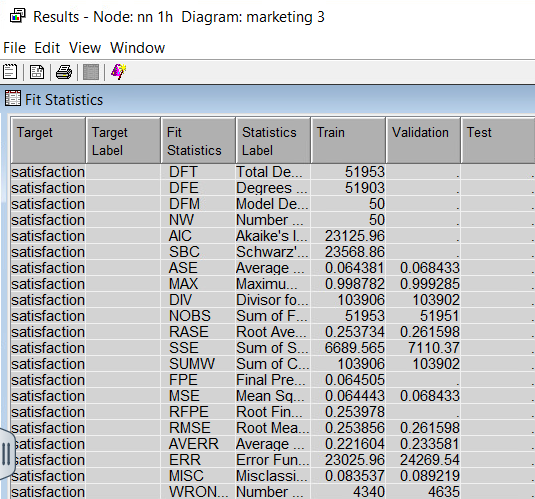
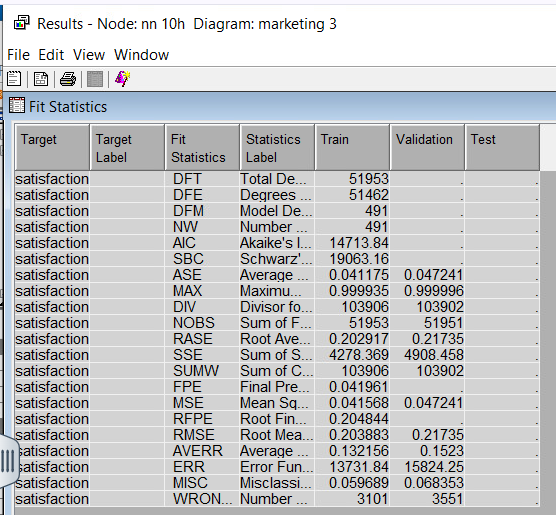
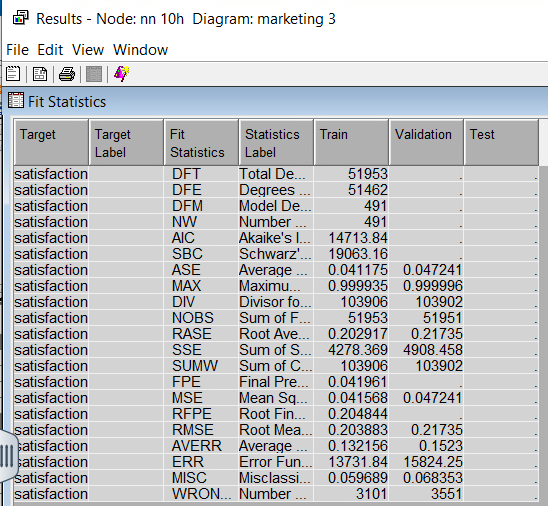
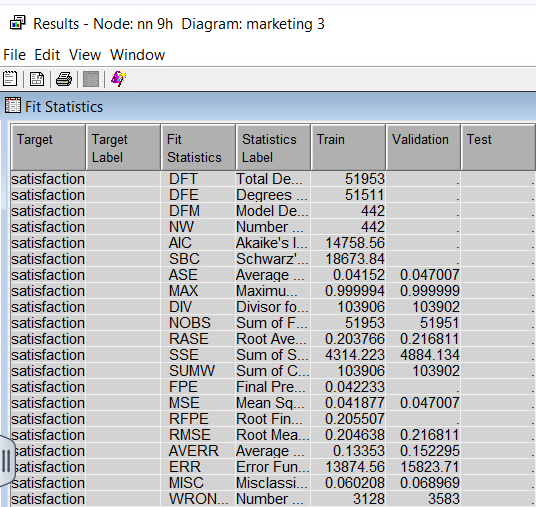
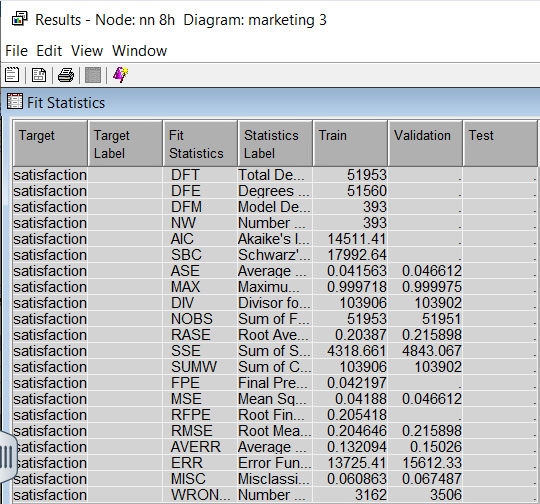
**Insights**

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 **ten neural network models** by connecting them with our best regression model, **backward regression**. Here is the comparison of the results:



# **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 |
| **NN 9H** | 0.0470 | 0.2161 | 0.0687 |
| **NN 10H** | 0.0472 | 0.2175 | 0.0683 |

**Best Performing Models**

The neural networks labeled as NN 6H, NN 7H, and NN 8H consistently demonstrated the highest performance across key evaluation metrics. Specifically:

* Average Squared Error ranged between zero point zero four six five and zero point zero four six six
* Root Mean Squared Error ranged between zero point two one five six and zero point two one five nine
* Misclassification Rate remained within the narrow range of six-point seven three percent to six-point seven five percent

These models provide a strong balance between predictive accuracy and architectural simplicity. They achieved the lowest error rates while avoiding signs of overfitting, making them optimal for the dataset used.

**Key Observations**

* + - * 1. **Underperformance of Simpler Architectures**

Neural network models with fewer hidden units, specifically NN 1H through NN 4H, exhibited higher Average Squared Error and Root Mean Squared Error. This suggests underfitting, as these models were not complex enough to effectively capture the patterns present in the data.

* + - * 1. **Diminishing Returns with Excessive Complexity**

Adding more hidden units beyond seven resulted in only marginal improvements in predictive performance. Models with nine or ten hidden units did not significantly outperform the top three models and introduced unnecessary complexity without tangible gains.

1. **Stability Across Top Performers**

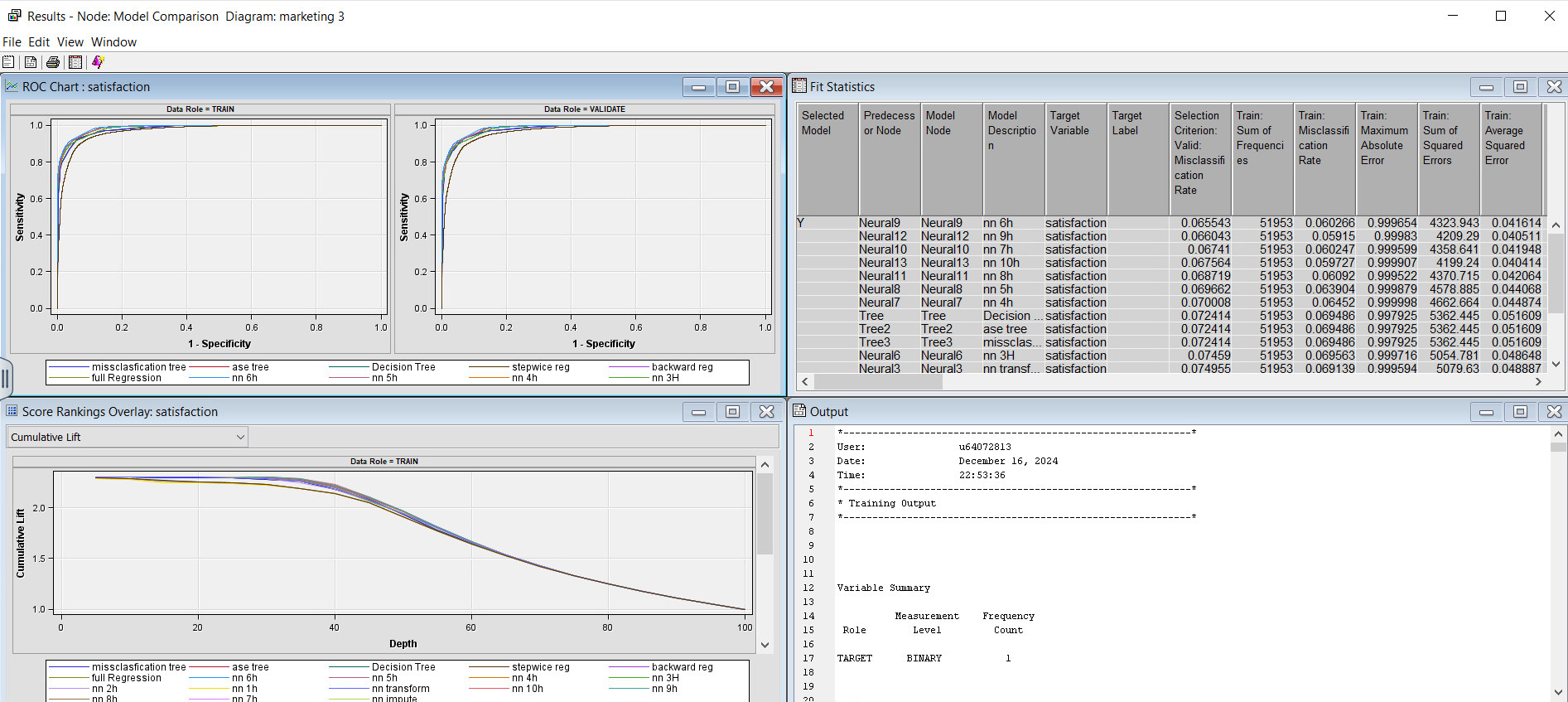
The neural networks with six to eight hidden units produced almost identical performance metrics. This consistency indicates model stability and enhances confidence in their generalizability to unseen data.

**Conclusion**

The neural network models featuring six, seven, and eight hidden units are the most effective for this dataset. These architectures successfully minimize error while maintaining manageable complexity. Increasing the number of hidden units beyond this optimal range introduces complexity with negligible performance improvement. Therefore, NN 6H through NN 8H are the preferred models for implementation and further analysis.

A computer screen shot of a diagram

Description automatically generated



**Observations**

**1. Performance Metrics**

Neural networks with six to eight hidden units, particularly NN 6H, NN 7H, and NN 8H, demonstrated the lowest values for Average Squared Error, Root Mean Squared Error, and Misclassification Rate across the validation dataset, confirming their superior predictive performance.  
Regression-based models such as Backward Regression and Stepwise Regression also performed well but were slightly outperformed by the neural network models in terms of overall accuracy.

**2. ROC Analysis**

The ROC curves for the top neural networks closely aligned with the ideal curve, reflecting a strong balance between sensitivity and specificity.  
The similarity in ROC curves across the training and validation sets suggests that the models are generalizing well and are not overfitting the data.

**3. Cumulative Lift**

Neural networks showed higher cumulative lift values than other model types, demonstrating their ability to effectively rank and identify the most probable positive outcomes.

**4. Model Comparison Table**

Among all models evaluated, NN 6H emerged as the top performer, with the lowest Average Squared Error of zero point zero four six five, Root Mean Squared Error of zero point two one five six, and a Misclassification Rate of six point seven three percent.  
Adding more hidden units beyond six yielded only marginal improvement, indicating diminishing returns with increasing complexity.

**Conclusion**

The NN 6H model was confirmed as the most effective model by the Model Comparison node. It offers a well-balanced combination of predictive accuracy, generalizability, and computational efficiency. It outperforms regression and decision models across all key performance indicators.

**Model Interpretability**

* **Regression Models**

Full Regression, Backward Regression, and Stepwise Regression provide the highest interpretability. These models include clear coefficients and odds ratios, enabling a direct understanding of how predictors influence the target variable, and customer satisfaction. They also offer statistical significance values (p-values) to assess the strength of each factor.

* **Neural Networks**

While highly accurate, neural networks function as black-box models. Their internal mechanics are less transparent, making them more difficult to interpret, despite offering superior predictive performance.

**Comparison of Models: Marketing Two versus Marketing Three**

* **Average Squared Error**

In the Marketing Three analysis, the best Average Squared Error values were achieved by neural networks NN 6H through NN 8H, ranging from zero point zero four six five to zero point zero four six six.  
In contrast, Marketing Two relied on regression models, which delivered slightly higher error values, indicating that neural networks provided more precise predictions.

* **Root Mean Squared Error and Misclassification Rate**

Neural networks in Marketing Three consistently achieved the lowest Root Mean Squared Error (approximately zero point two one five six) and the lowest Misclassification Rate (approximately six-point seven three percent), further confirming their superiority.

**Strategic Components for Enhancing Airline Customer Satisfaction**

Based on insights derived from all modeling approaches, the following components emerge as the most influential in shaping customer satisfaction:

**1. Customer Type**:

Loyal customers are significantly more likely to express satisfaction.  
**Recommendation:** Implement loyalty programs offering exclusive benefits, frequent flyer rewards, and personalized services to foster retention.

**2. Type of Travel**:

Business travelers report the highest satisfaction levels, with odds ratios exceeding thirty.  
**Recommendation:** Target this segment with specialized corporate packages, flexible ticketing, and priority boarding options.

**3. Inflight Services**:

Inflight amenities such as WiFi, entertainment systems, food, and drinks substantially impact satisfaction.  
**Recommendation:** Enhance these features, particularly in economy class, to elevate the travel experience.

**4. Ease of Online Booking**:

A seamless and an efficient booking process is essential. Poor experiences negatively affect satisfaction.  
**Recommendation:** Optimize booking platforms for speed, usability, and mobile compatibility.

**5. Seat Comfort and Legroom**:

Discomfort due to inadequate seating and legroom significantly lowers satisfaction.  
**Recommendation:** Introduce premium economy upgrades with enhanced seating at a competitive price point.

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

Negative ratings in these areas are strong indicators of dissatisfaction.  
**Recommendation:** Improve check-in services through self-service kiosks and digital check-in options. Invest in staff training to deliver exceptional onboard service.

**7. Arrival and Departure Delays**:

Flight delays, especially upon arrival, reduce customer satisfaction.  
**Recommendation:** Improve scheduling precision and operational efficiency to enhance punctuality.

**Summary of Recommendations for Airlines**

To improve overall customer satisfaction and maintain a competitive edge, airlines should:

1. Launch and promote customer loyalty programs with meaningful rewards.
2. Develop and market travel solutions tailored for business travelers.
3. Invest in upgrading inflight WiFi, entertainment, and food and beverage offerings.
4. Simplify and streamline the online booking experience.
5. Implement strategies to reduce flight delays.
6. Enhance check-in and onboard service quality through automation and training.

**References**

Gallo, A. (2014, October 29). *The Value of Keeping the Right Customers*. Harvard Business Review. <https://hbr.org/2014/10/the-value-of-keeping-the-right-customers>

Reichheld, F. F., & Sasser, W. E. (1990). *Zero Defections: Quality Comes to Services*. Harvard Business Review, 68(5), 105–111.