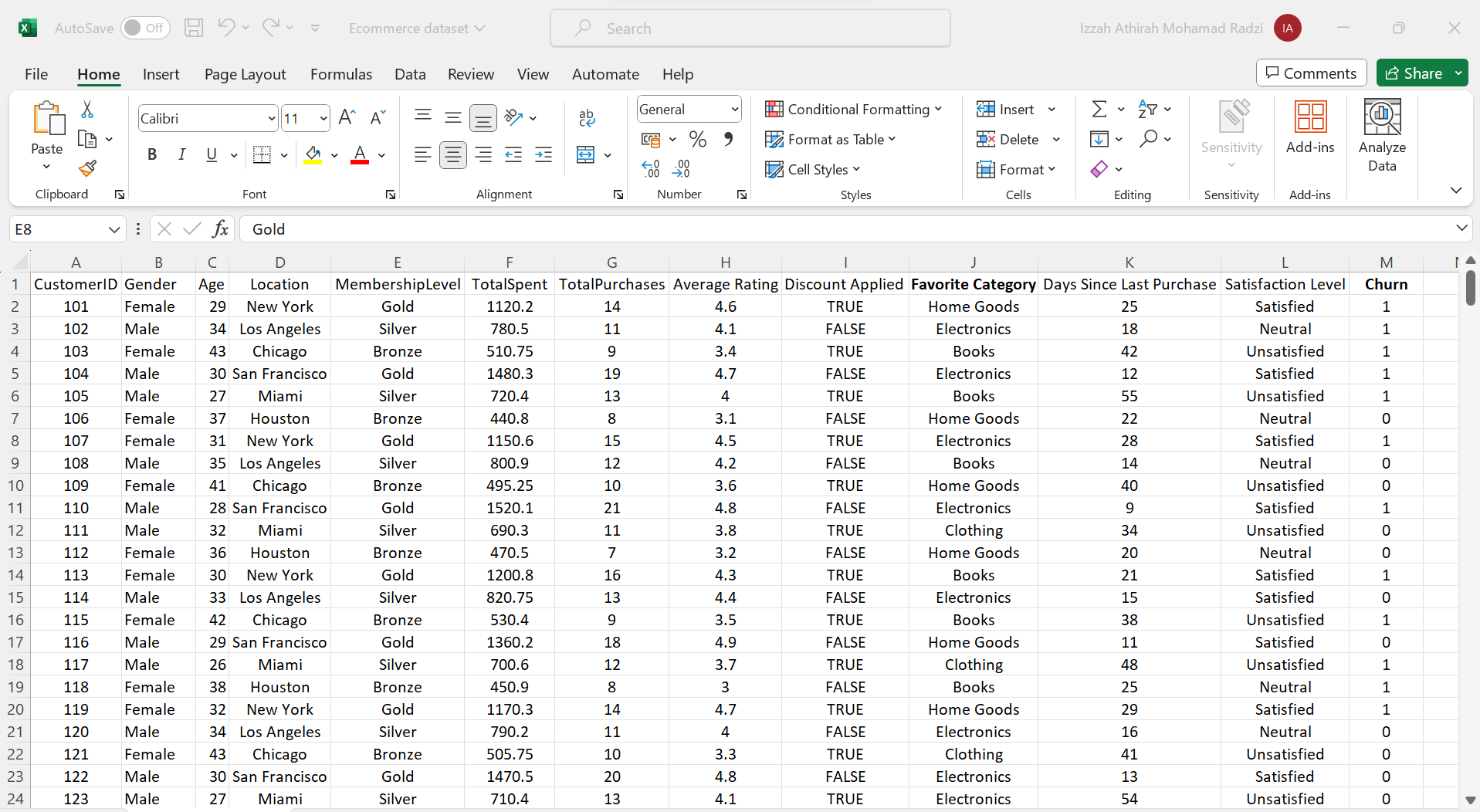
**Case Study: E-Commerce Customer Behaviour Analysis**

**Instructions/Deliverables:**

A report detailing each step of the process, including the rationale behind your choices and any challenges faced. An analysis of the decision tree and ensemble methods, with insights into customer behavior and suggestions for business strategy.

**Step 1:** Obtained data from Kaggle website. The dataset that I found apparently doesn’t have variable and data that I can use as target variable. Hence, I added 2 new columns into the dataset which are Favorite Category as well as Churn (Column J and M).

****

**Step 2:** Used Talend Data Preparation to spot missing value as well as modifying gender data from nominal to binary. Found out there are two columns with missing values which are Satisfaction Level as well as Average Rating as below:

**A screenshot of a computer

Description automatically generated**

Then, duplicated the gender tab and replace the data which is ‘Female’ and ‘Male’ to 1 and 0, respectively. After that, I renamed the column to become ‘Gender1’.

**A screenshot of a computer

Description automatically generated**

**Step 3:** Imported the dataset into SAS Enterprise Miner and specified the variables’ roles to ID, rejected and Input.

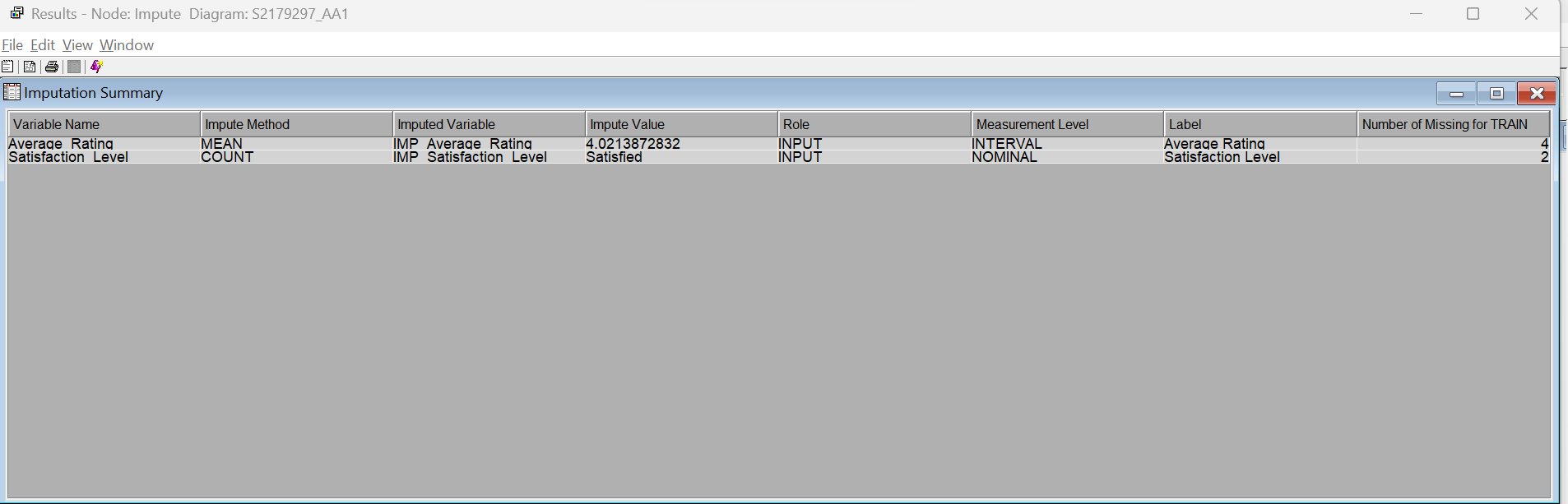
A screenshot of a computer

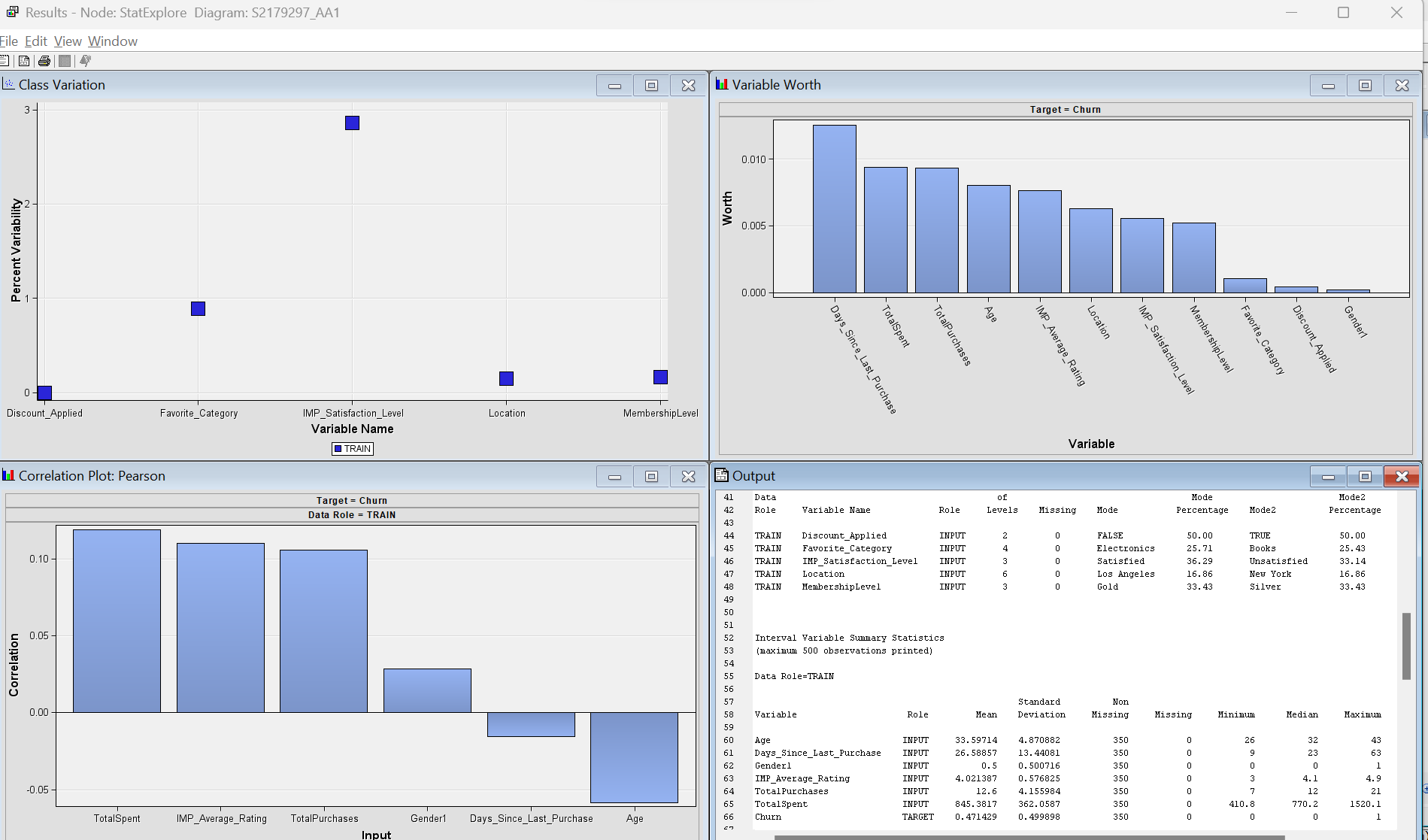
Description automatically generated

A screenshot of a computer

Description automatically generated

**Step 4:** Dragged ‘Impute’ function from ‘Modify’ tab to the workspace to impute the missing values and run the function. Used ‘StatExplore’ to further validate the imputation result as well as the summary of the data before doing data partitioning.





**Step 5:** Dragged ‘Data Partition’ from ‘Sample’ tab to split the data into training, testing and validation. I decided to go with 70:20:10 ratio for my analysis as it provides a balanced approach, ensuring enough data for model learning, tuning, and evaluation while maintaining statistical significance in each subset.

A screenshot of a computer

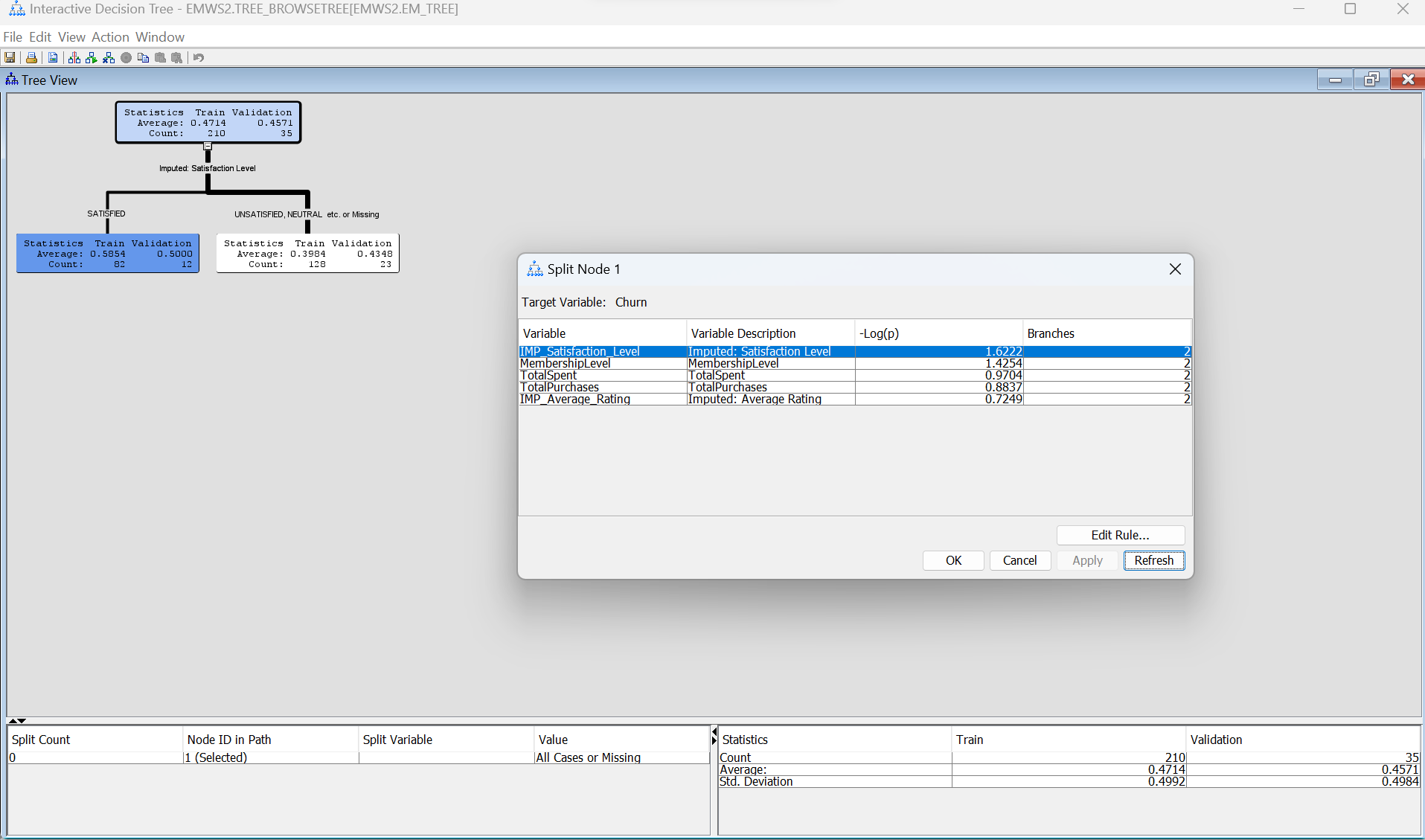
Description automatically generated

**Step 6:** Dragged the models which are Decision Tree, High Performance Forest as well as Gradient Boosting to analyse the dataset.

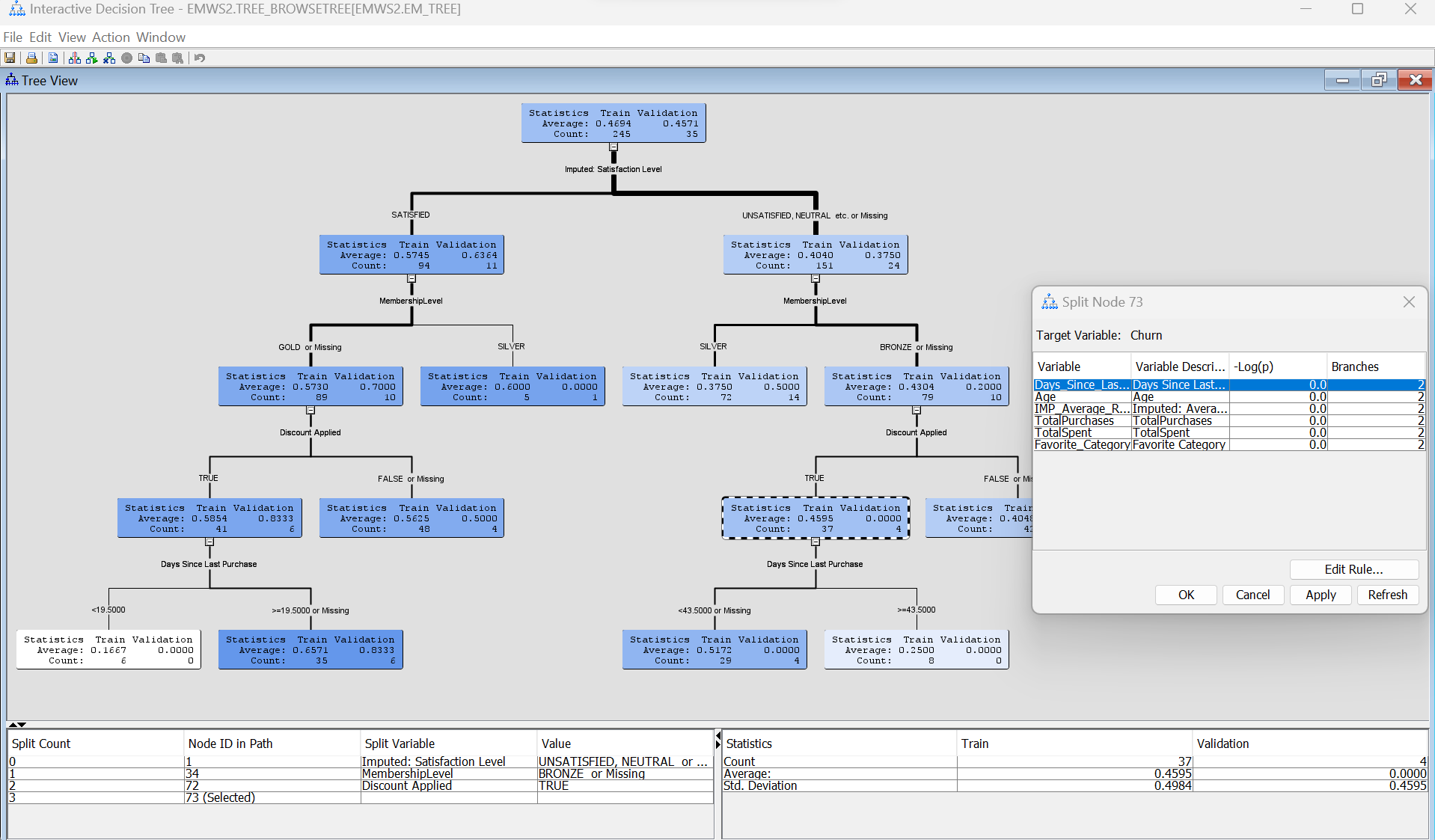
A screenshot of a computer

Description automatically generated

For Decision Tree, I set the maximum branch to 2, maximum depth to 6 and set the assessment measure to Decision. I’ve tried to further explore and edit the properties and variables for this function, but my result only appeared as below:

****

To solve the issue, I decided to do manual split node as below. I’ve tried to split the node for each category based on the highest -Log(p) value displayed in the split node table. I’ve tried to ensure that the observation node is displaying darker blue box as it indicates the percentages of correctly classified node.

****

The result of Decision Tree modelling is as below. From the result, ‘Satisfaction Level’ is considered as the most important variable. The Fit Statistics provide evaluation metrics indicating model performance on training and validation datasets. Assessment Score Rankings and Distribution show how well the model predicts churn across different depths and ranges.

A screenshot of a computer

Description automatically generated

As for HP Forest, I retained the significance level of 0.05 as it represents a standard threshold for determining node splits, aiming to strike a balance between capturing meaningful patterns and avoiding overfitting in the resulting decision tree or random forest model. The result of HP Forest is as below. Based on the result, the model seems to perform reasonably well based on the provided metrics such as ASE and MSE. It was trained and validated using 10-fold cross-validation. Variable importance suggests how certain input variables contribute to reducing prediction errors.

A screenshot of a computer

Description automatically generated

As for Gradient Boosting model, I retained the setting as default and run the function. The result is as below. Reading the result, I can see that variables like 'Days\_Since\_Last\_Purchase', 'TotalSpent', and 'Location' appear highly influential in predicting churn. As per the Fit Statistics, the model shows some errors (ASE, RASE) in predicting churn across different datasets (train, validation, test) while in Assessment Score Rankings & Distribution, I can see how well the model predicts churn across different observation depths and predicted value ranges.

A screenshot of a computer

Description automatically generated

**Step 7:** Finally, I did model comparison to compare all the models that I have run and here are the results:

A screenshot of a computer

Description automatically generated

Three models are being compared: HP Forest, Boost (Gradient Boosting), and Tree (Decision Tree).

As for Model Selection based on Valid: Average Squared Error (VASE), the selection criterion favors the model with the lowest Average Squared Error, where HPDMForest appears to have the lowest error (0.2418), followed by Boost (0.2437), and then Tree (0.2442).

As for Fit Statistics, each model's performance is assessed across different datasets (train, validation, test) using metrics like Average Squared Error, Maximum Absolute Error, Root Average Squared Error, Sum of Frequencies, etc.

The choice of the best model could be based on its performance on the validation and test datasets. In this case, HP Forest appears marginally better, but I think further analysis needs to be done to include other relevant metrics and potential overfitting.