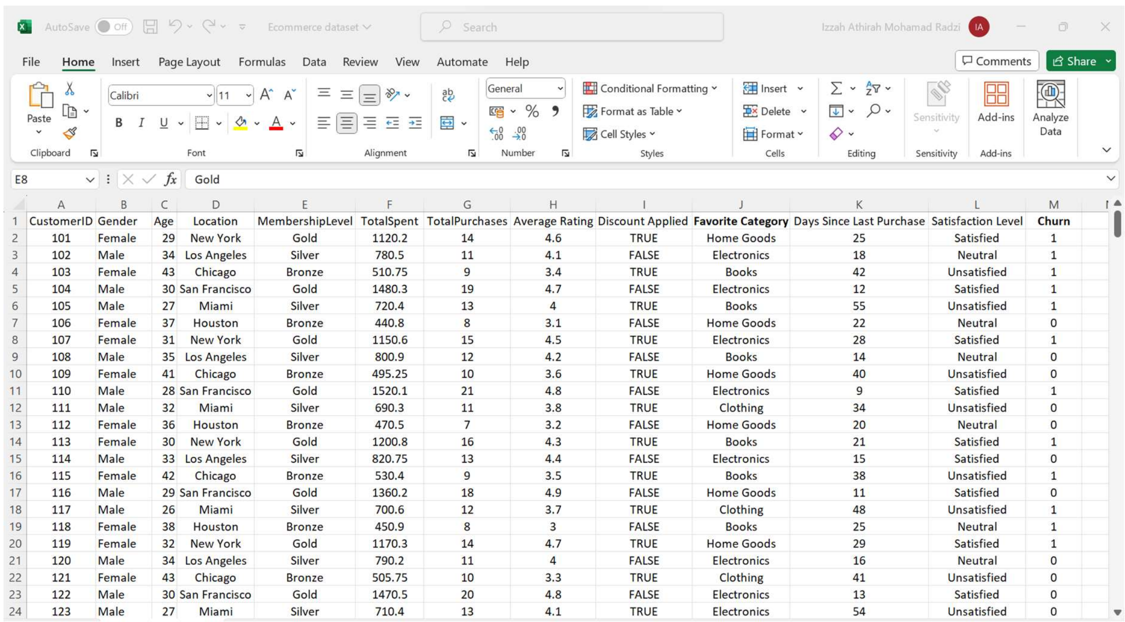
Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

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.

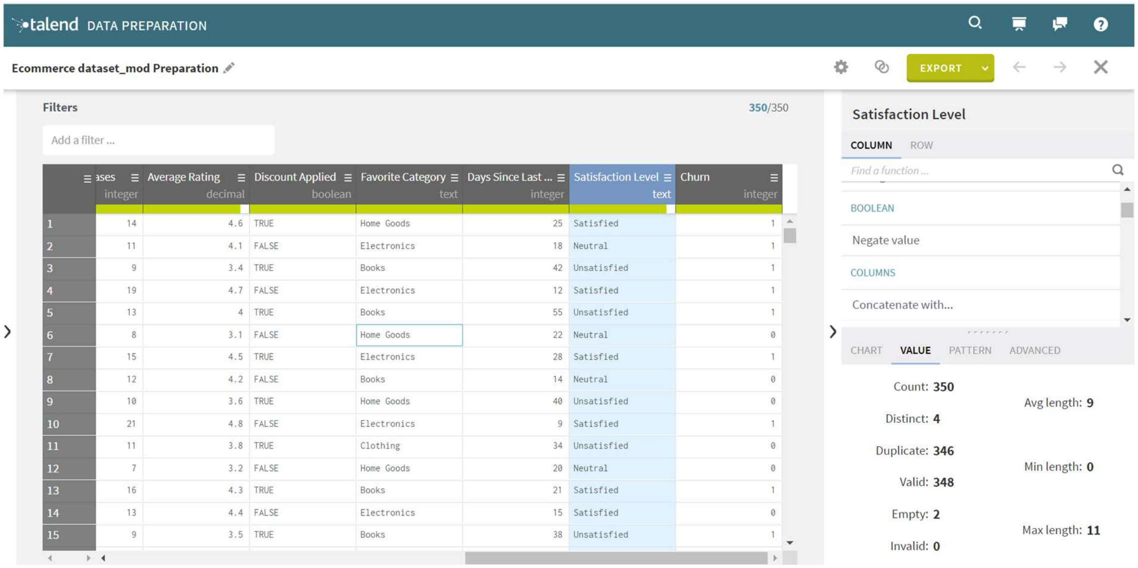
Step 1: Obtained data from Kaggle website. The dataset that I found apparently doesn't have variable and data that I can use as target variable. Hence, I added 2 new columns into the dataset which are Favorite Category as well as Churn (Column J and M).



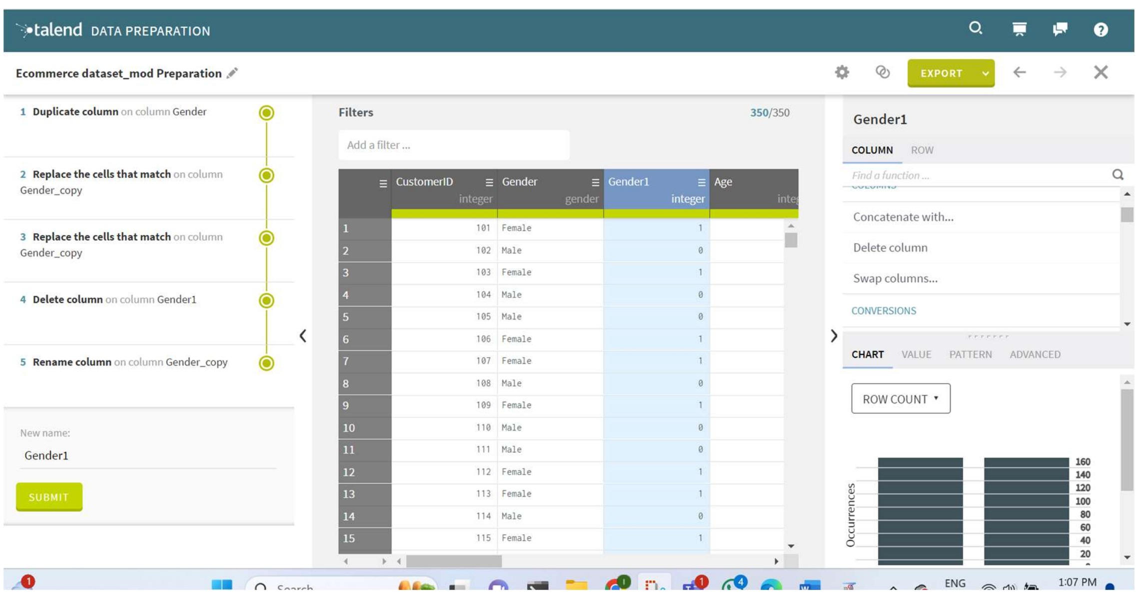
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CustomerID	Gender	Age	Location	MembershipLevel	TotalSpent	TotalPurchases	Average Rating	Discount Applied	**Favorite Category**	Days Since Last Purchase	Satisfaction Level	**Churn**
2	101	Female	29	New York	Gold	1120.2	14	4.6	TRUE	Home Goods	25	Satisfied	1
3	102	Male	34	Los Angeles	Silver	780.5	11	4.1	FALSE	Electronics	18	Neutral	1
4	103	Female	43	Chicago	Bronze	510.75	9	3.4	TRUE	Books	42	Unsatisfied	1
5	104	Male	30	San Francisco	Gold	1480.3	19	4.7	FALSE	Electronics	12	Satisfied	1
6	105	Male	27	Miami	Silver	720.4	13	4	TRUE	Books	55	Unsatisfied	1
7	106	Female	37	Houston	Bronze	440.8	8	3.1	FALSE	Home Goods	22	Neutral	0
8	107	Female	31	New York	Gold	1150.6	15	4.5	TRUE	Electronics	28	Satisfied	1
9	108	Male	35	Los Angeles	Silver	800.9	12	4.2	FALSE	Books	14	Neutral	0
10	109	Female	41	Chicago	Bronze	495.25	10	3.6	TRUE	Home Goods	40	Unsatisfied	0
11	110	Male	28	San Francisco	Gold	1520.1	21	4.8	FALSE	Electronics	9	Satisfied	1
12	111	Male	32	Miami	Silver	690.3	11	3.8	TRUE	Clothing	34	Unsatisfied	0
13	112	Female	36	Houston	Bronze	470.5	7	3.2	FALSE	Home Goods	20	Neutral	0
14	113	Female	30	New York	Gold	1200.8	16	4.3	TRUE	Books	21	Satisfied	1
15	114	Male	33	Los Angeles	Silver	820.75	13	4.4	FALSE	Electronics	15	Satisfied	0
16	115	Female	42	Chicago	Bronze	530.4	9	3.5	TRUE	Books	38	Unsatisfied	1
17	116	Male	29	San Francisco	Gold	1360.2	18	4.9	FALSE	Home Goods	11	Satisfied	0
18	117	Male	26	Miami	Silver	700.6	12	3.7	TRUE	Clothing	48	Unsatisfied	1
19	118	Female	38	Houston	Bronze	450.9	8	3	FALSE	Books	25	Neutral	1
20	119	Female	32	New York	Gold	1170.3	14	4.7	TRUE	Home Goods	29	Satisfied	1
21	120	Male	34	Los Angeles	Silver	790.2	11	4	FALSE	Electronics	16	Neutral	0
22	121	Female	43	Chicago	Bronze	505.75	10	3.3	TRUE	Clothing	41	Unsatisfied	0
23	122	Male	30	San Francisco	Gold	1470.5	20	4.8	FALSE	Electronics	13	Satisfied	0
24	123	Male	27	Miami	Silver	710.4	13	4.1	TRUE	Electronics	54	Unsatisfied	0

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:

Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

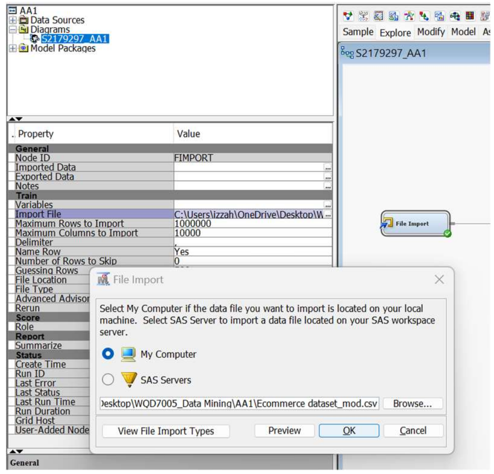


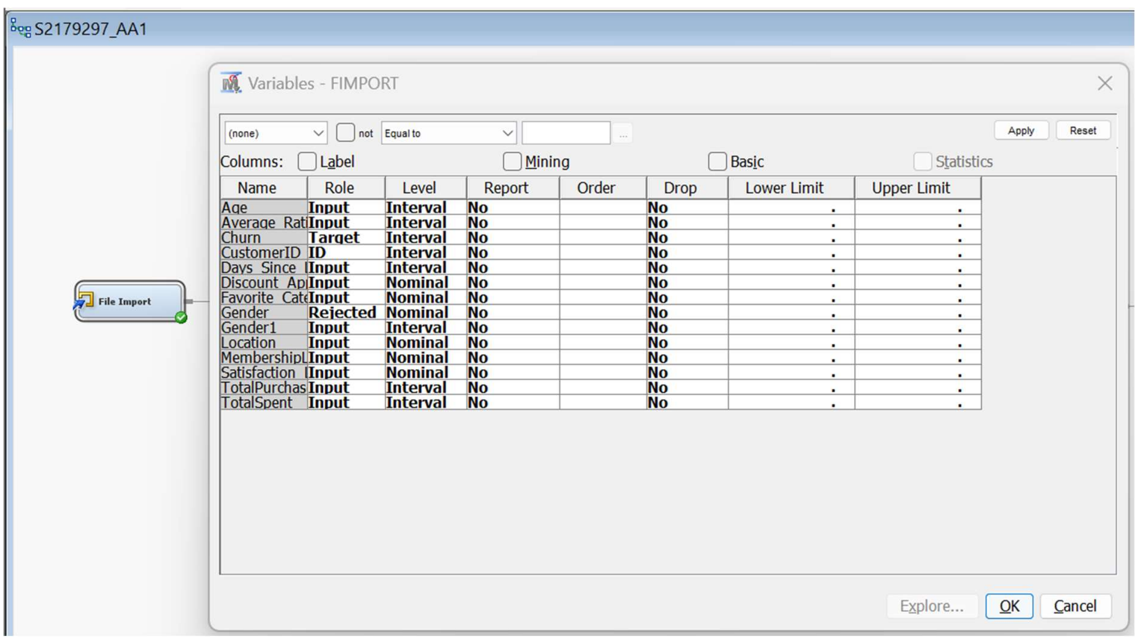
Then, duplicated the gender tab and replace the data which is 'Female' and 'Male' to 1 and 0, respectively. After that, I renamed the column to become 'Gender1'.



Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

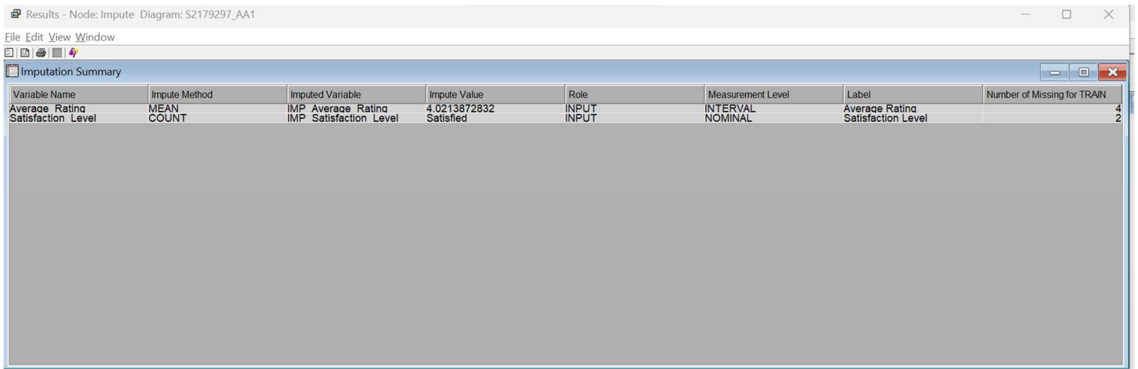
Step 3: Imported the dataset into SAS Enterprise Miner and specified the variables' roles to ID, rejected and Input.

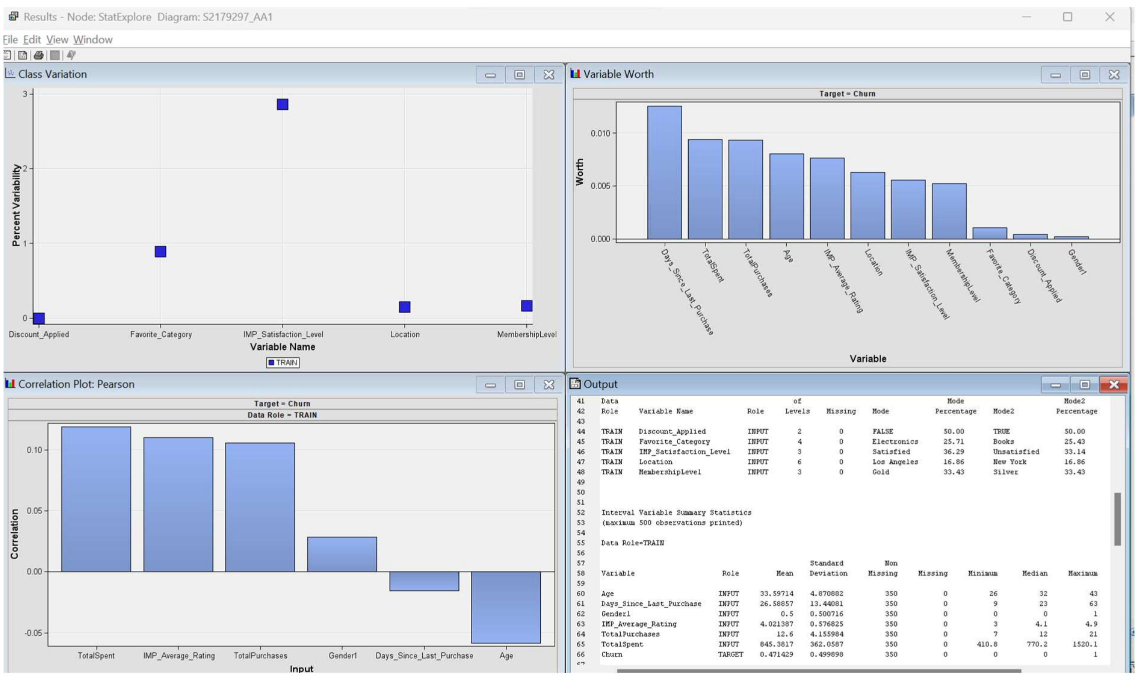




Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

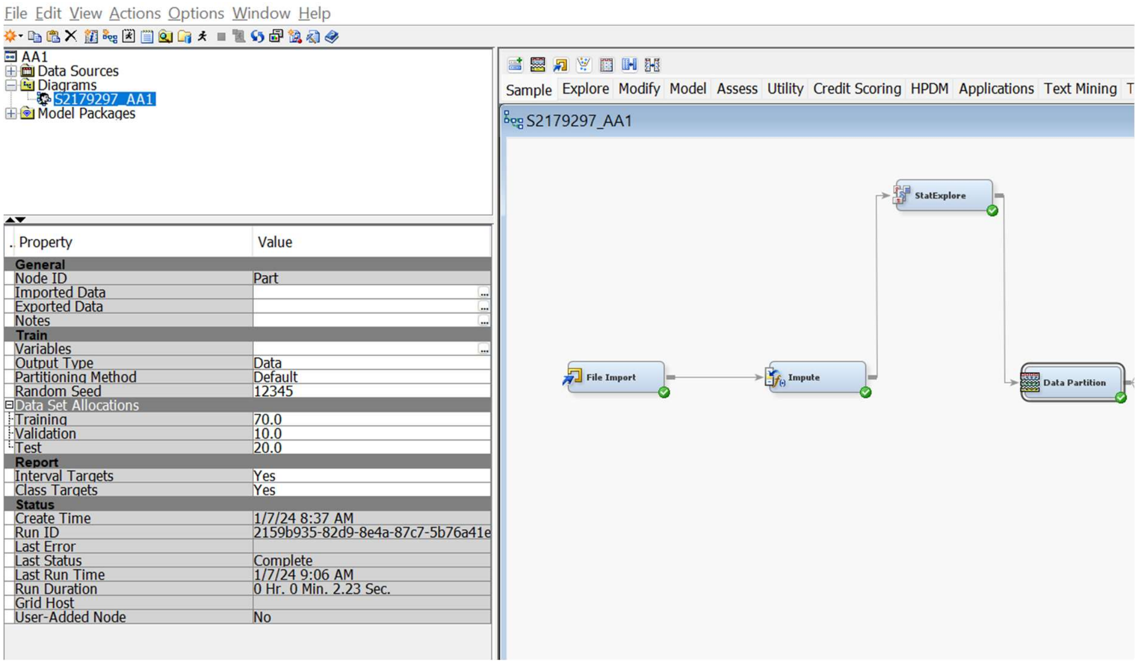
Step 4: Dragged 'Impute' function from 'Modify' tab to the workspace to impute the missing values and run the function. Used 'StatExplore' to further validate the imputation result as well as the summary of the data before doing data partitioning.



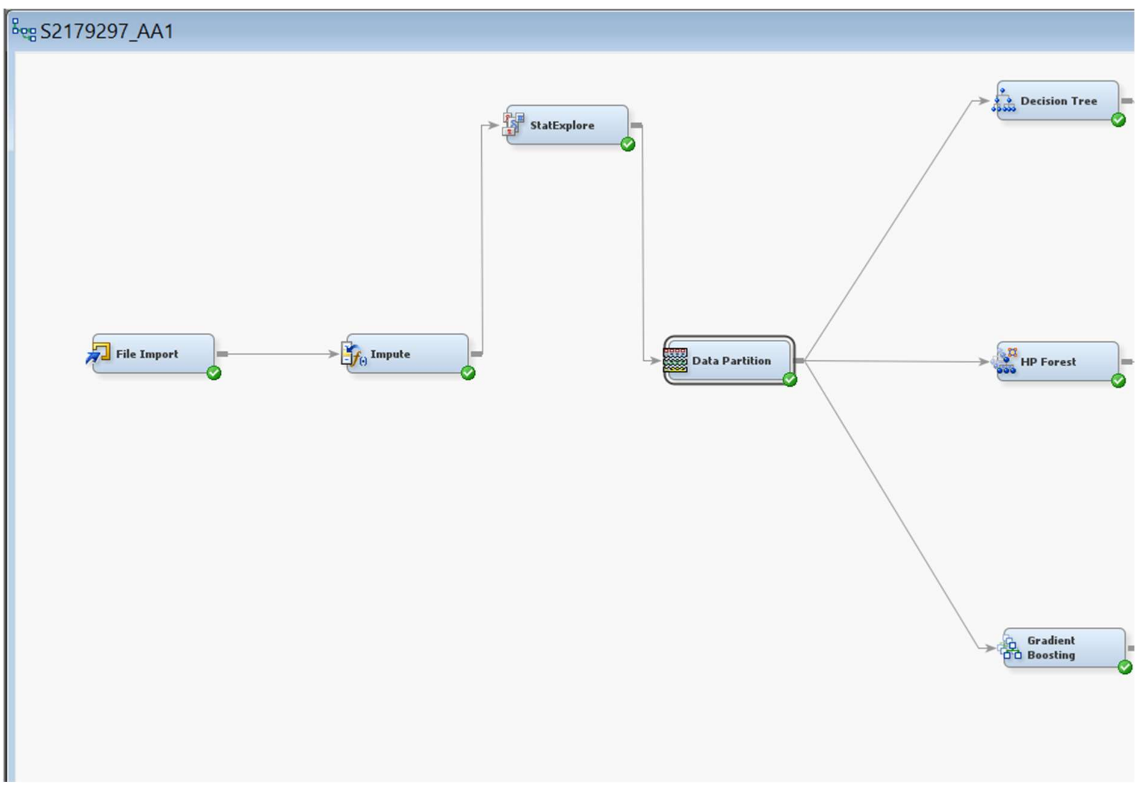


Step 5: Dragged 'Data Partition' from 'Sample' tab to split the data into training, testing and validation. I decided to go with 70:20:10 ratio for my analysis as it provides a balanced approach, ensuring enough data for model learning, tuning, and evaluation while maintaining statistical significance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

Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

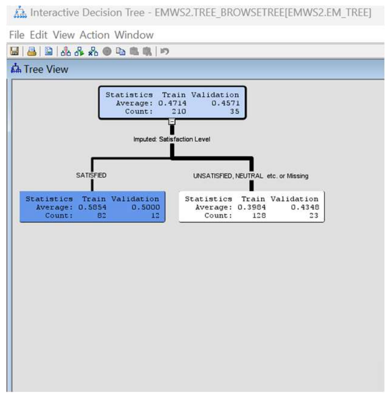


Step 6: Dragged the models which are Decision Tree, High Performance Forest as well as Gradient Boosting to analyse the dataset.

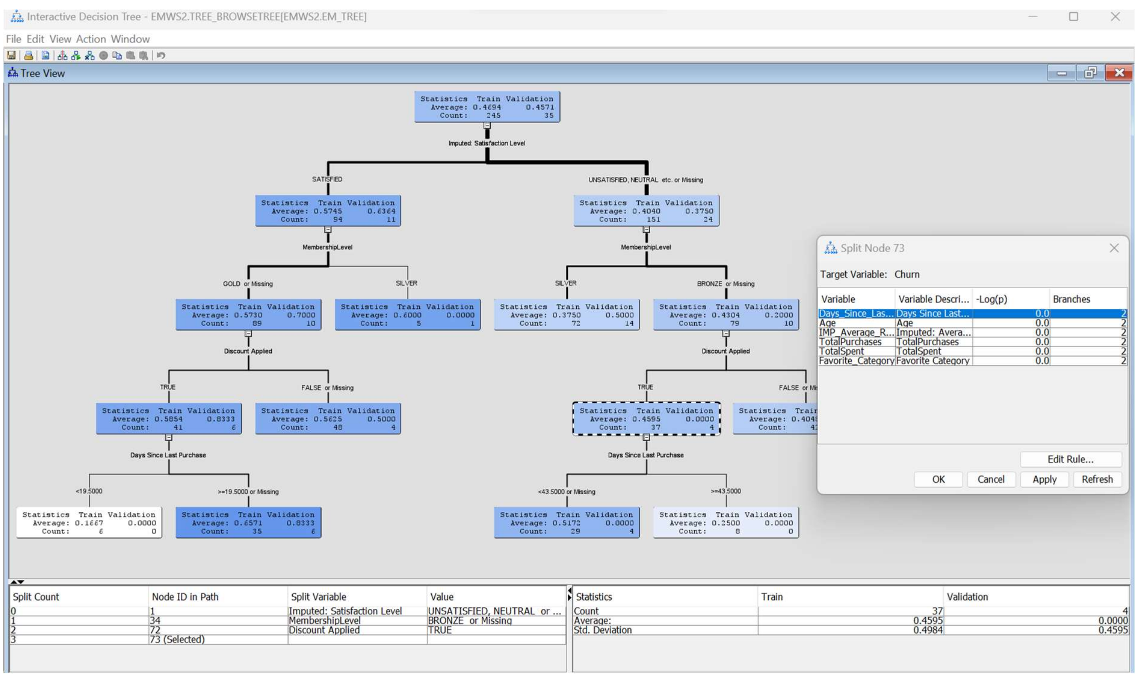


Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

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:

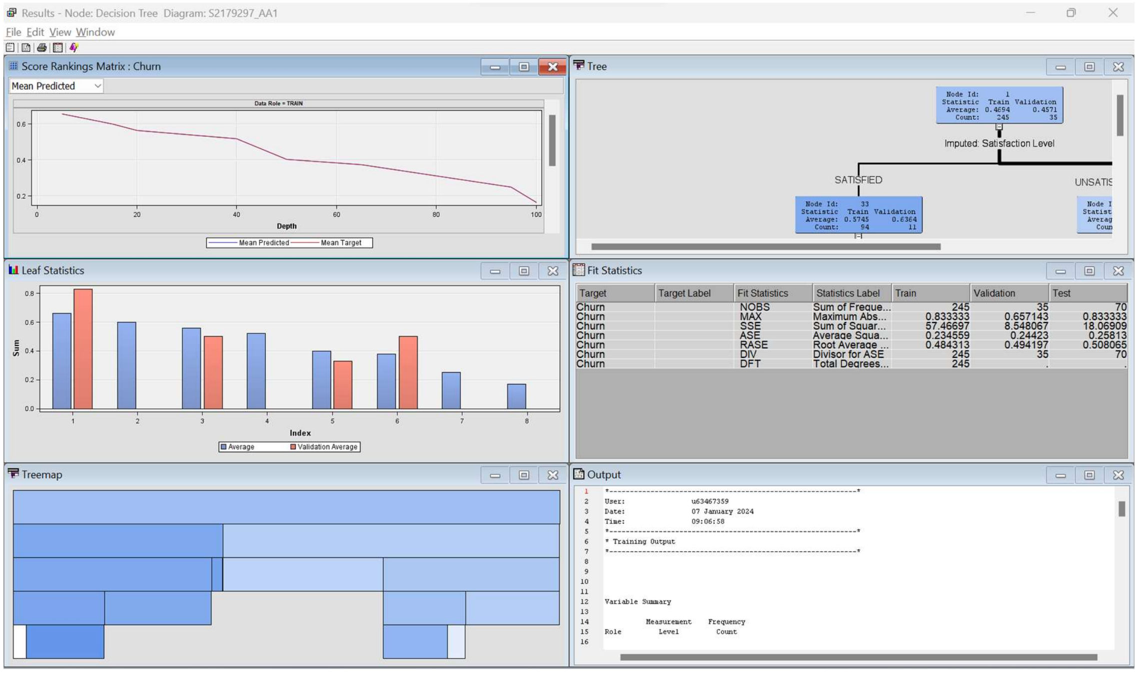


To solve the issue, I decided to do manual split node as below. I've tried to split the node for each category based on the highest -Log(p) value displayed in the split node table. I've tried to ensure that the observation node is displaying darker blue box as it indicates the percentages of correctly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

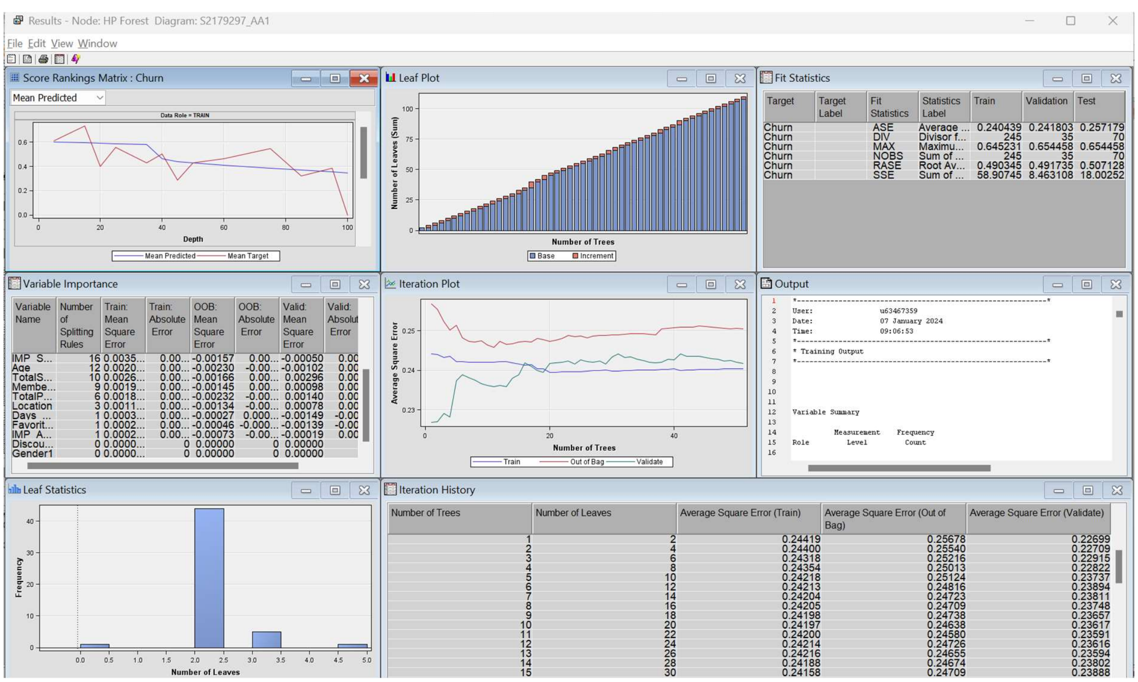


The result of Decision Tree modelling is as below. From the result, 'Satisfaction Level' is considered as the most important variable. The Fit Statistics provide evaluation metrics indicating model performance on training and validation datasets. Assessment Score Rankings and Distribution show how well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

Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

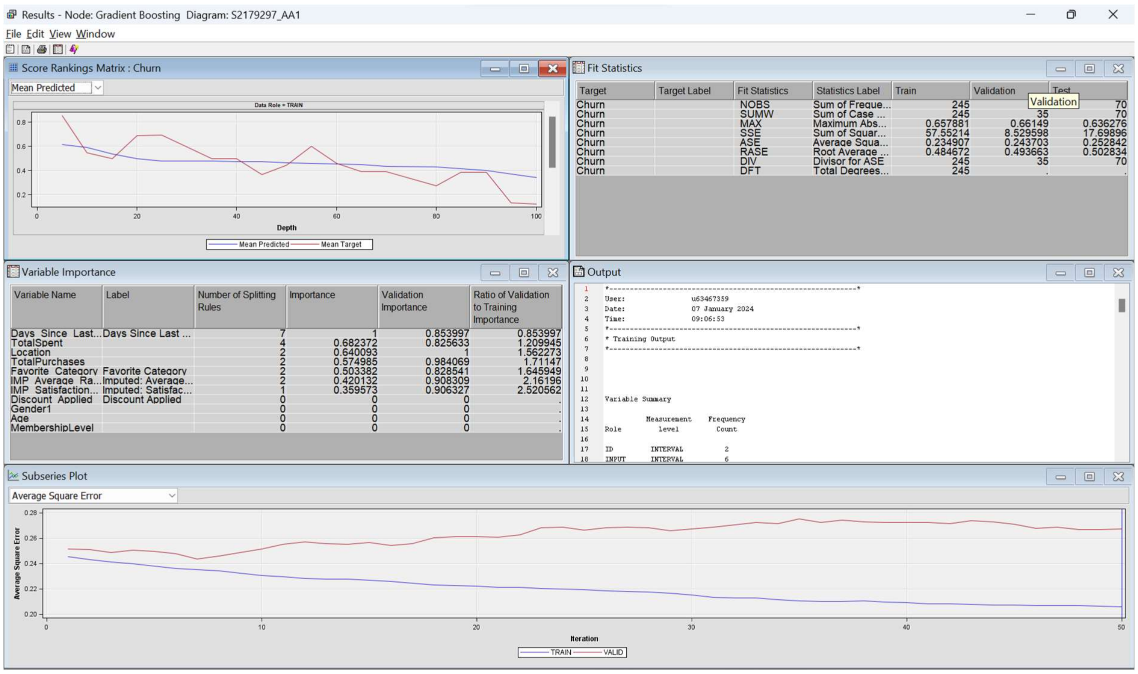


As for HP Forest, I retained the significance level of 0.05 as it represents a standard threshold for determining node splits, aiming to strike a balance between capturing meaningful patterns and avoiding overfitting in the resulting decision tree or random forest model. The result of HP Forest is as 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.



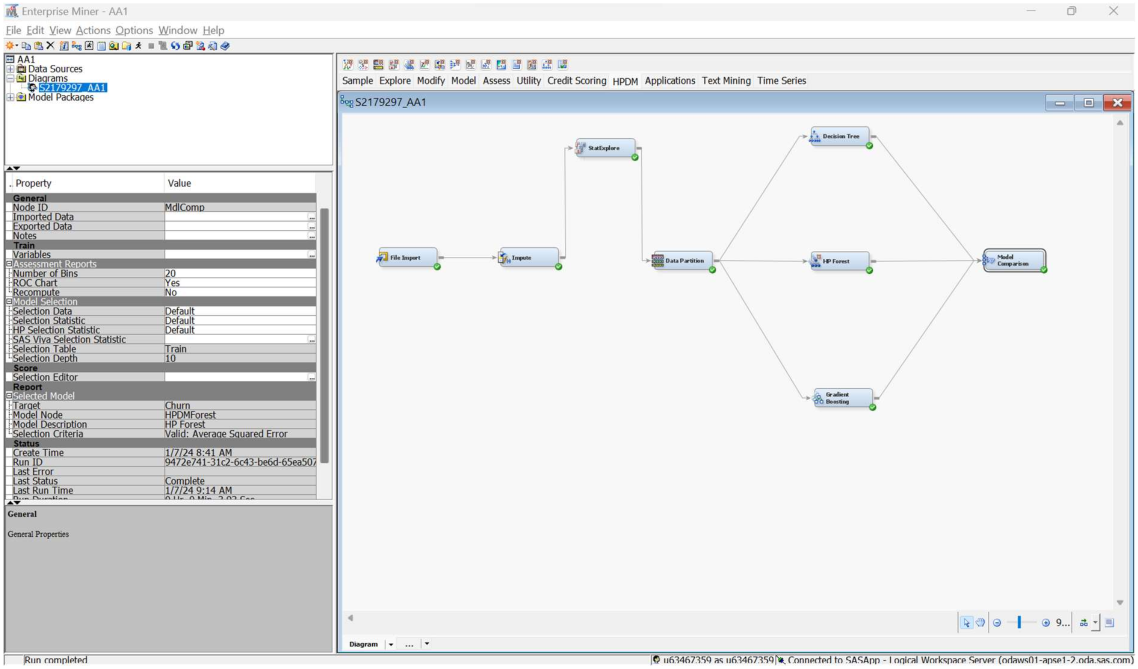
Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git

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.



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

Izzah Athirah Mohamad Radzi
S2179297
WQD7005: Alternative Assessment 1
GitHub link: https://github.com/S2179297/AA1.git



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.