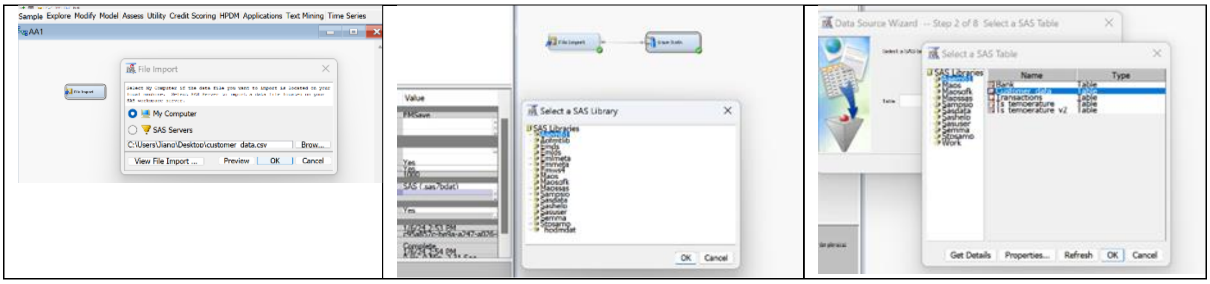
SAS Enterprise Miner

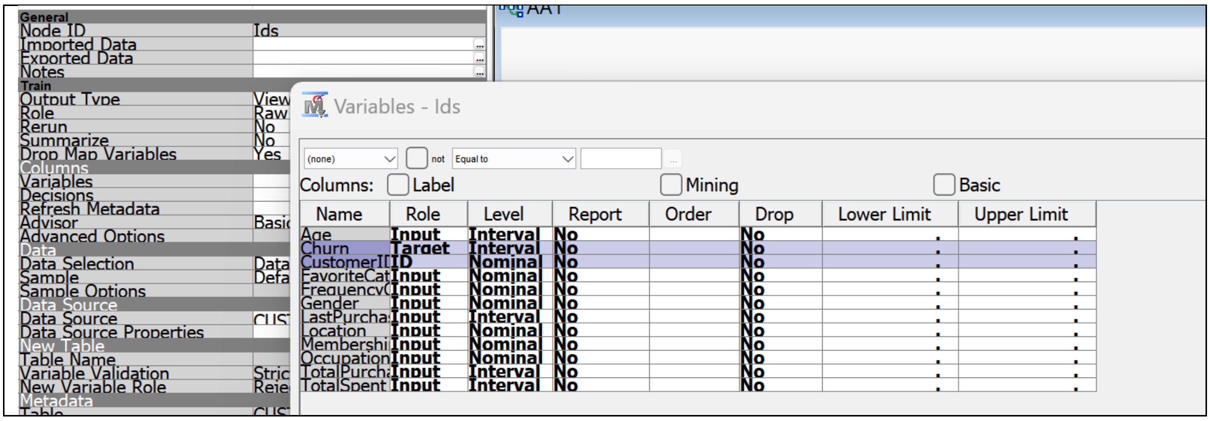
1. Import Processing:

Firstly in the SAS Enterprise Miner we create a diagram then drag the file import node upload the customer_data.csv after that drag the save file node to save the file into AAEM61 library. Then create new data source from the AAEM61 library.



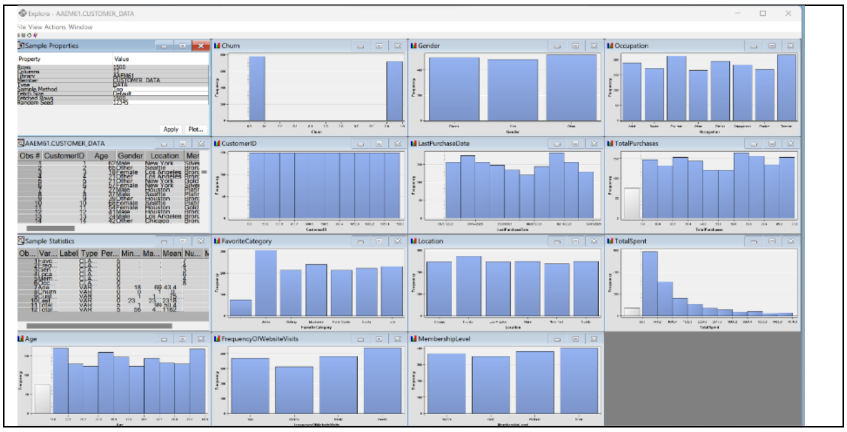
2. Variable Role Specification:

After creating the new data source, drag the data to the diagram and edit the variables set Churn as Target and Customer as ID.

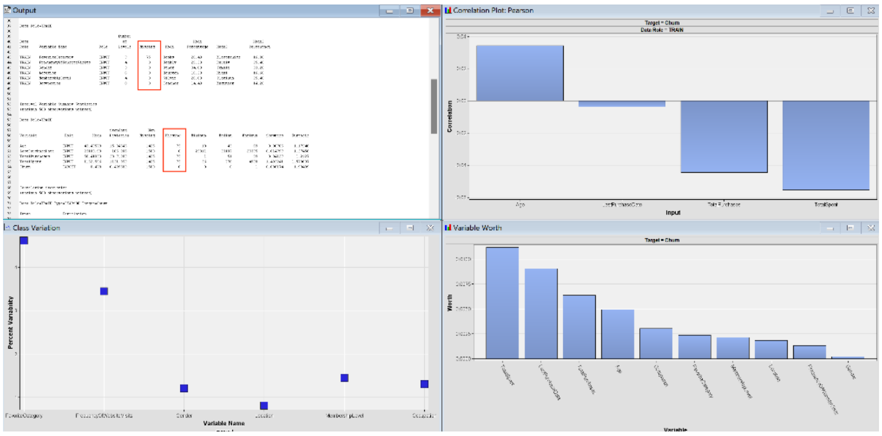


3. Dataset Explore and Check missing values

Firstly, explore all the variables, check each attributes data distribution and whether there is a missing value.

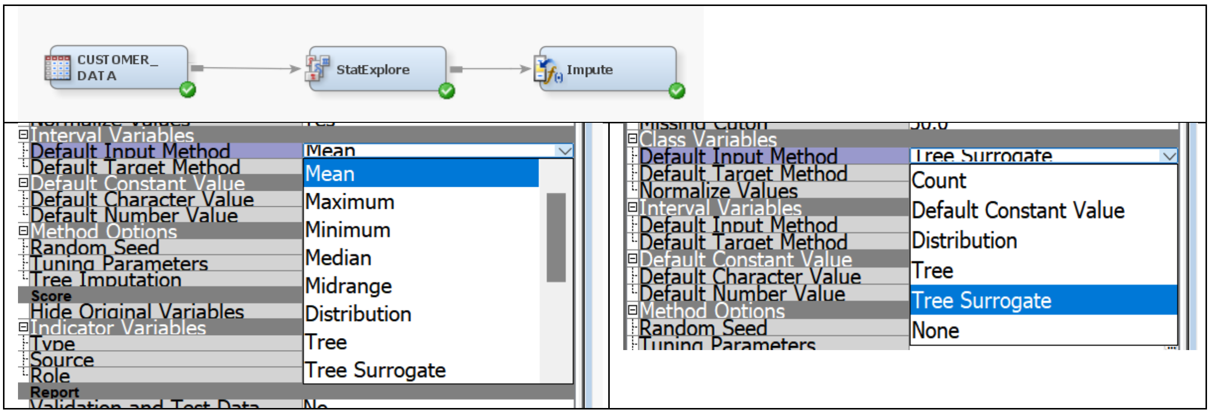


Then connect to StatExplore node check the output report then we find there have two kinds of missing values, one is class variable another is interval variable. They are FavoriteCategory (75 missing values), Age(75 missing values), TotalPurchases (75 missing values), TotalSpent(75 missing values).

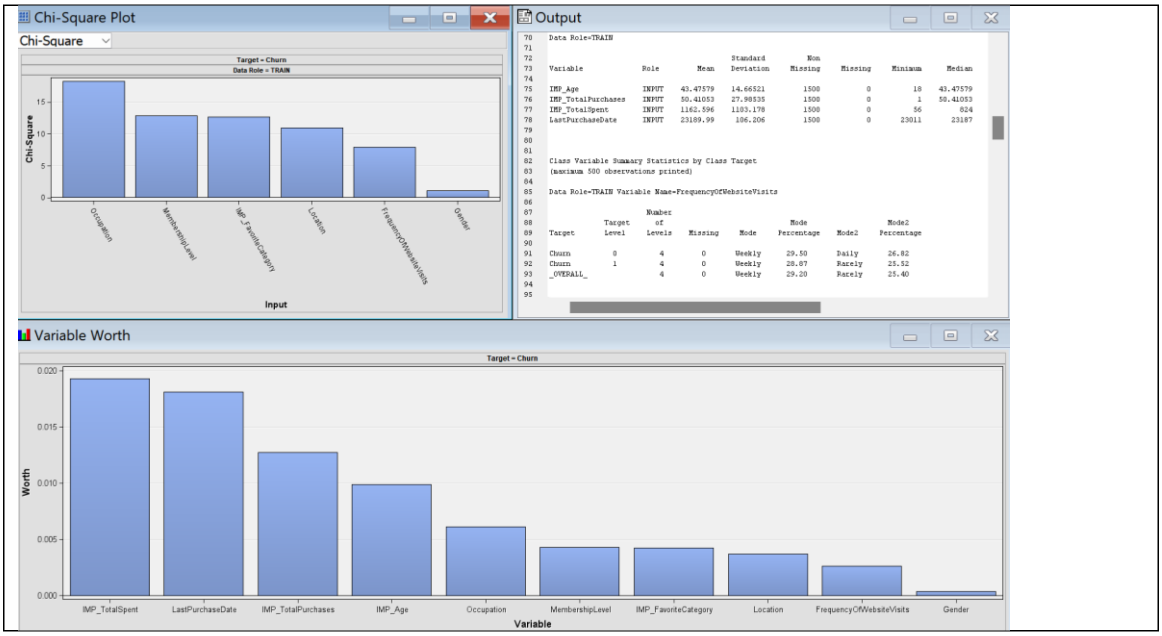


4. Handle missing value

For the missing value we drag the impute node to the diagram, for Interval missing value chose mean and Tree Surrogate for Class Variables.



After imputing the missing value, check the dataset again. We can see there has no missing value.



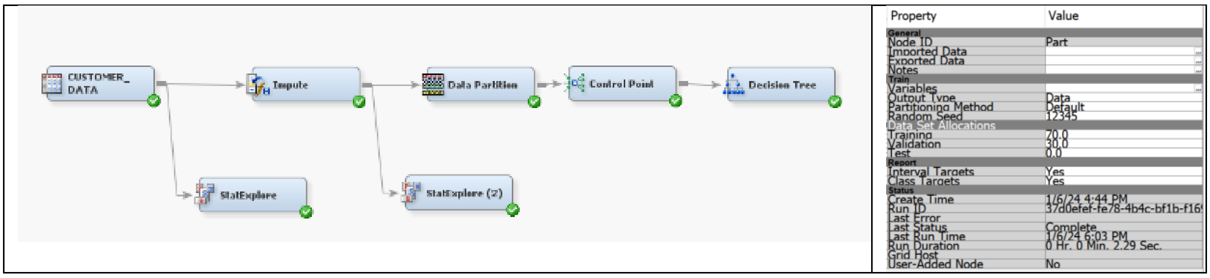
5. Decision Tree Analysis

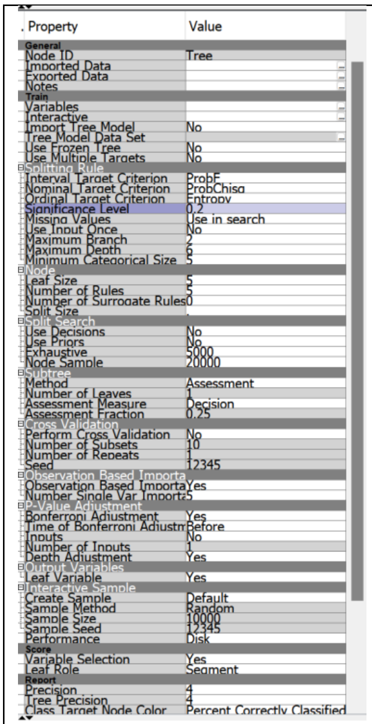
5.1 Create a decision tree model in SAS Enterprise Miner

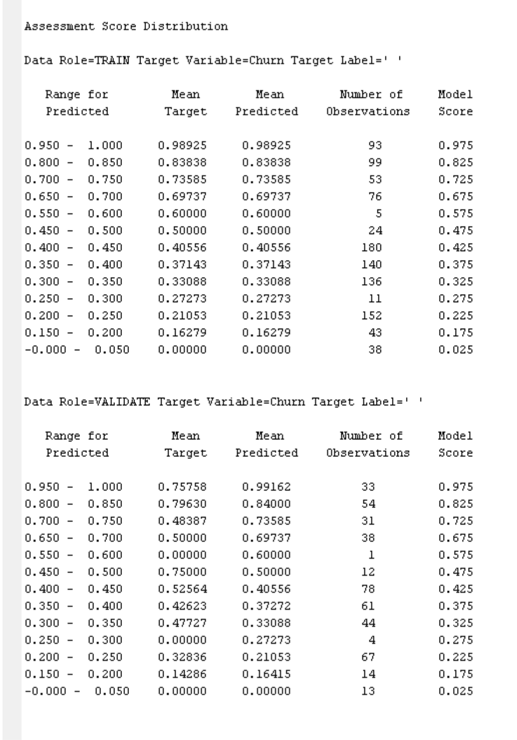
Decision tree is a versatile data mining algorithm that enables both classification and regression tasks. It models decisions and their possible consequences as a tree-like structure, where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a decision or prediction. This structure emulates the human decision-making process, making decision trees one of the most intuitive and widely used algorithms in analytics.

Before creating the decision tree model, drag the data partition node and control point. The data partition we set 70% for train and 30% for validation.

For the decision tree model, we set the max branch as 2, maximum tree depth is 6 and maximum categorical size is 5. For the node properties the leaf size is 5, number of rules is 5 and number of surrogate rules is 0.





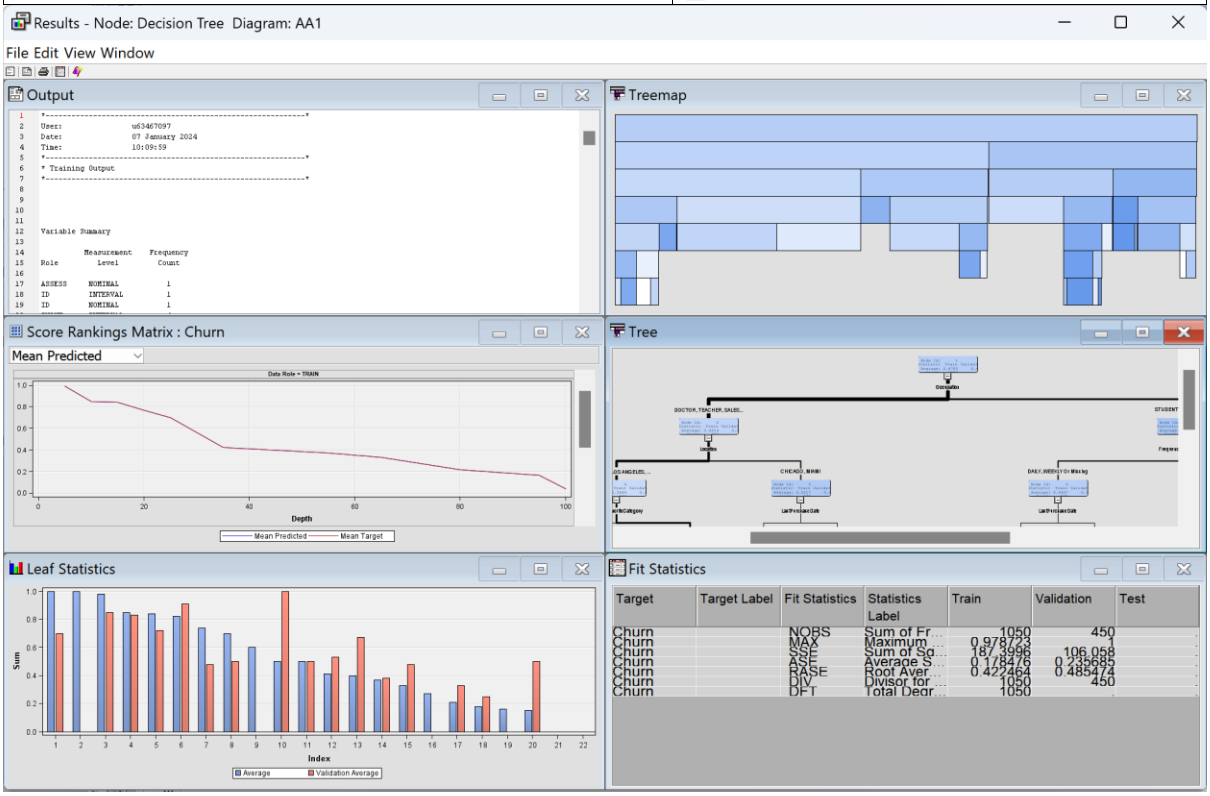

Assessment Score Distribution

Data Role=TRAIN Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.950 - 1.000	0.98925	0.98925	93	0.975
0.800 - 0.850	0.83838	0.83838	99	0.825
0.700 - 0.750	0.73585	0.73585	53	0.725
0.650 - 0.700	0.69737	0.69737	76	0.675
0.550 - 0.600	0.60000	0.60000	5	0.575
0.450 - 0.500	0.50000	0.50000	24	0.475
0.400 - 0.450	0.40556	0.40556	180	0.425
0.350 - 0.400	0.37143	0.37143	140	0.375
0.300 - 0.350	0.33088	0.33088	136	0.325
0.250 - 0.300	0.27273	0.27273	11	0.275
0.200 - 0.250	0.21053	0.21053	152	0.225
0.150 - 0.200	0.16279	0.16279	43	0.175
-0.000 - 0.050	0.00000	0.00000	38	0.025

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.950 - 1.000	0.75758	0.99162	33	0.975
0.800 - 0.850	0.79630	0.84000	54	0.825
0.700 - 0.750	0.48387	0.73585	31	0.725
0.650 - 0.700	0.50000	0.69737	38	0.675
0.550 - 0.600	0.00000	0.60000	1	0.575
0.450 - 0.500	0.75000	0.50000	12	0.475
0.400 - 0.450	0.52564	0.40556	78	0.425
0.350 - 0.400	0.42623	0.37272	61	0.375
0.300 - 0.350	0.47727	0.33088	44	0.325
0.250 - 0.300	0.00000	0.27273	4	0.275
0.200 - 0.250	0.32836	0.21053	67	0.225
0.150 - 0.200	0.14286	0.16415	14	0.175
-0.000 - 0.050	0.00000	0.00000	13	0.025



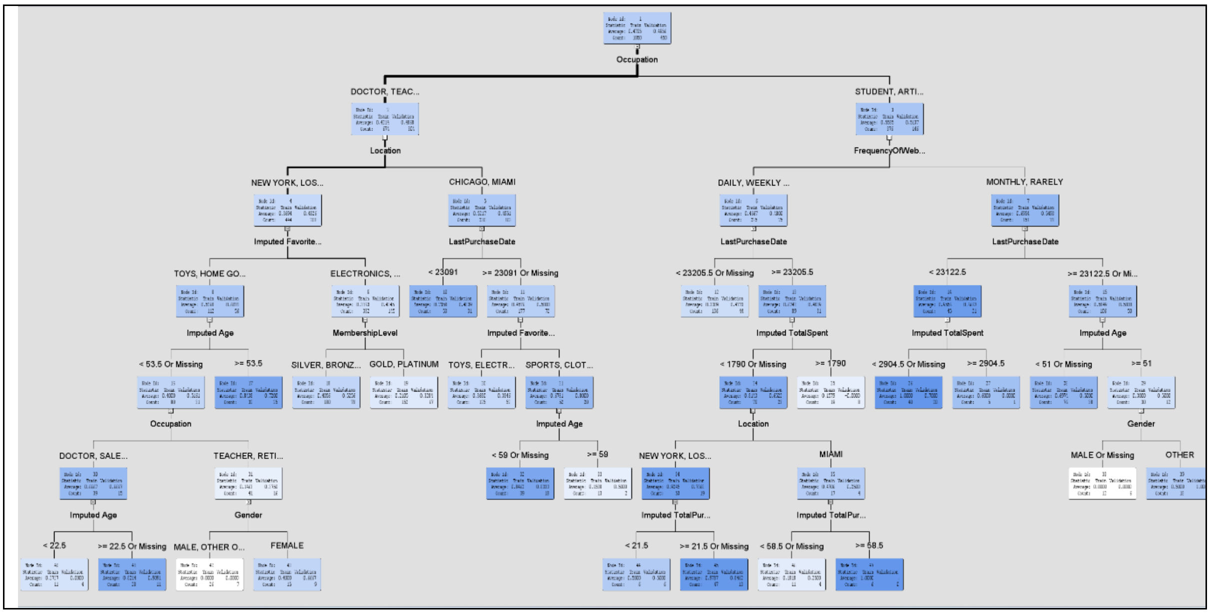
The results from the decision tree model show its performance in classifying customers' churn probabilities on both training and validation data.

In the training set, the model has strong alignment between the predicted probabilities and actual churn rates, indicating good calibration. For example, in the highest probability bin (0.950 - 1.000), the model perfectly matches the Mean Predicted with the Mean Target at 0.98925, suggesting high accuracy for the most certain predictions of churn.

However, in the validation set, there are discrepancies. In the highest predicted churn range (0.950 - 1.000), the Mean Target drops to 0.75758, while the Mean Predicted stays high at

0.99162, indicating an overestimation of churn risk. This suggests the model may not generalize as well to unseen data, a common challenge known as overfitting.
The model's performance, indicated by the model score, remains high in extreme ranges (close to 0 or 1) but is lower in the middle ranges, reflecting less certainty in predictions where the churn probability is around 50%.
Overall, the decision tree demonstrates strong training performance but may require adjustments to improve its predictive accuracy on new, unseen data.

5.2 Analyse customer behaviour



To mitigate customer churn, it is essential to understand the underlying factors that contribute to a customer's likelihood of disengaging. Leveraging a decision tree analysis, we can unearth patterns and predictors within customer data that signal churn risk. This predictive model slices through layers of demographic, behavioral, and transactional data to reveal key attributes ranging from occupation and geographic location to spending habits and product preferences that are instrumental in forecasting churn. With this knowledge, we can devise targeted interventions designed to bolster retention and foster enduring customer relationships.

4.2.1 Occupational Impact on Churn:

The decision tree places occupation as a significant indicator of churn risk. The differentiation between professions such as Doctors, Teachers, and Salespeople against Students, Artists, and other occupations suggests a correlation between occupation type and customer retention. For example, busy professionals may have less time for extensive shopping and might prefer a streamlined, reliable service, potentially leading to lower churn if their expectations are met. Conversely, Students and Artists might be more price-sensitive and could churn if they find better deals elsewhere. Tailoring loyalty programs and customer service to the specific needs and behaviors of each occupational group could help mitigate churn risks.

4.2.2 Geographical and Engagement Insights:

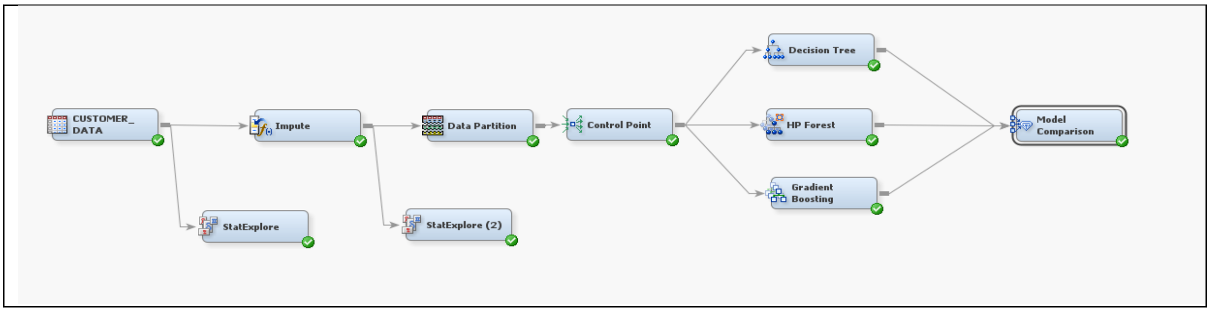
Geographical location plays a pivotal role in customer churn, with separate branches for metropolitan areas versus other cities, indicating different churn dynamics. Customers in larger cities might have access to more competitive alternatives, which could influence their loyalty. Engagement metrics, such as the recency of the last purchase and the frequency of website visits, are directly tied to churn, where infrequent visits and a longer time since the last purchase signal a higher likelihood of churn. Identifying at-risk customers through these metrics allows for timely intervention strategies, such as personalized promotions or reminders, to re-engage them.

4.2.3 Demographic and Behavioral Predictors of Churn:

The decision tree also highlights demographic factors like age and gender, which can be instrumental in predicting churn. Different age groups may have varying levels of engagement and brand loyalty, influencing their churn behavior. Additionally, spending behavior and product preferences are strong indicators of churn; customers who spend above certain thresholds or those who purchase specific product categories may exhibit different churn rates. Understanding these spending patterns can be crucial for developing targeted retention campaigns, such as offering special deals on favorite categories or appreciation rewards that encourage continued patronage.

In summary, the decision tree analysis sheds light on various factors that contribute to customer churn. By addressing these areas with focused customer retention strategies, companies can proactively reduce churn rates. This might include personalized engagement based on occupational needs, regional marketing strategies, and targeted offers that align with customer demographics and spending behaviors. Ultimately, leveraging this decision tree analysis can lead to more effective churn prevention and an overall improvement in customer loyalty.

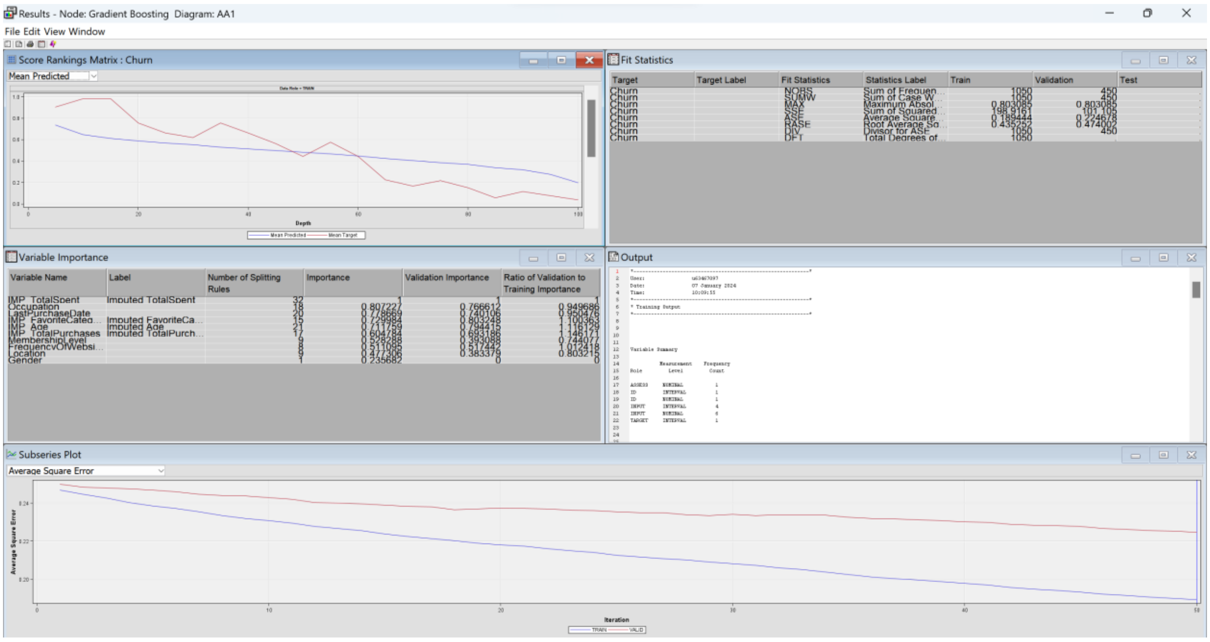
6. Ensemble Methods



In addition to decision trees, the use of ensemble methods like Boosting and Bagging with the Random Forest algorithm can significantly enhance the predictive power of your customer churn models by reducing variance, bias, or improving predictions through aggregation. In this section, we will analyze the results achieved by Boosting and Bagging with the Random Forest algorithm, assessing their respective strengths and weaknesses. Subsequently, we will perform a comparative analysis among these three models: the decision tree, Boosting, and

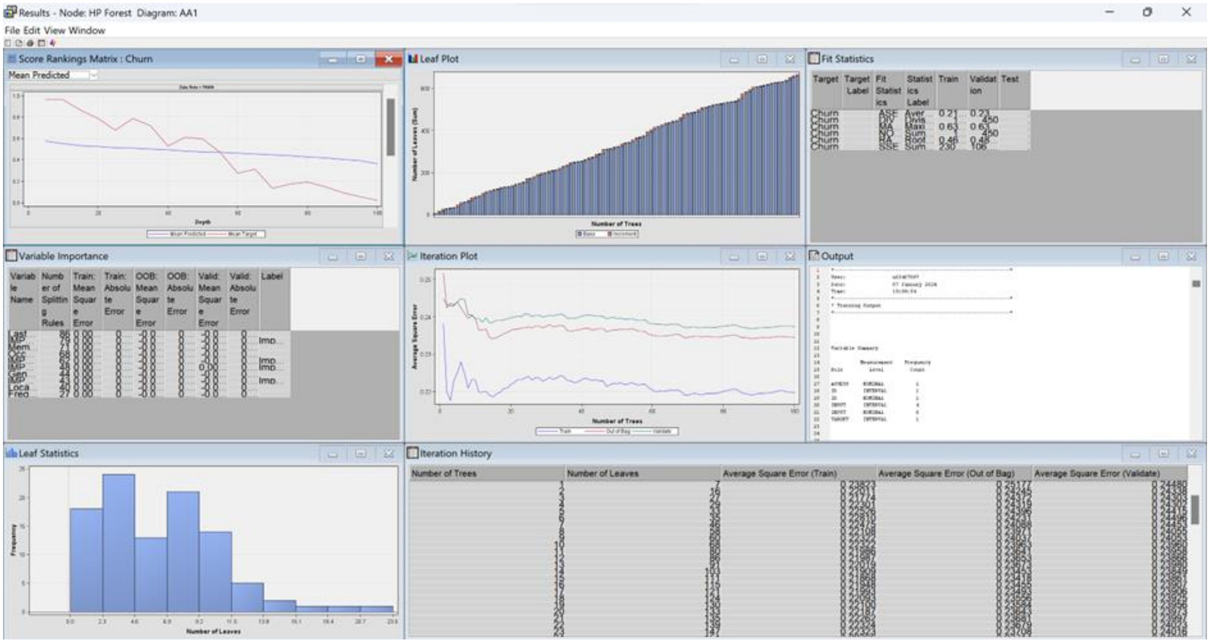
Bagging with the Random Forest algorithm, to gain insights into their performance and suitability for addressing the customer churn prediction problem.

5.1 Boosting: (Gradient Boost)



In the top probability range (0.858 - 0.898), the Gradient Boosting model predicts with high confidence, as indicated by the "Mean Predicted" value of 0.8915, which is very close to the "Mean Target" of 1.0000. This suggests that, for the 2 observations in this range, the model is almost certain they will churn.
For the mid-probability ranges, such as (0.618 - 0.658), the "Mean Predicted" value is 0.63415, with a "Mean Target" of 0.96721 for 83 observations. Here, the model is underestimating the churn risk compared to the actual outcome.
Lower probability ranges, such as (0.177 - 0.217), show a "Mean Predicted" of 0.19797 and a "Mean Target" of 0.04762 for 21 observations, indicating the model's conservative prediction for these cases.

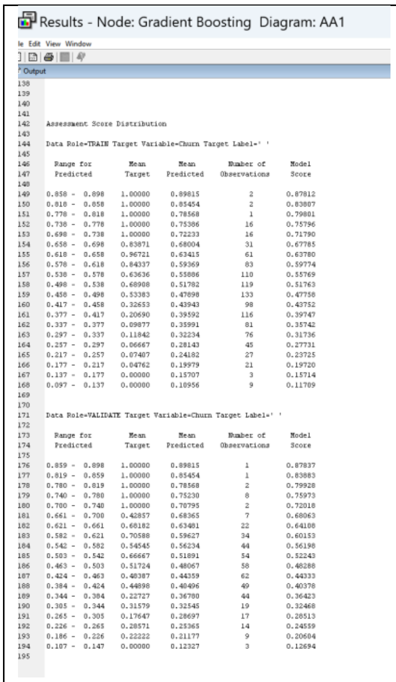
5.2 Bagging (Random Forest model in HP Environment)



The Random Forest model, in a similar top probability range (0.560 - 0.597), has a "Mean Predicted" of 0.58914 and a "Mean Target" of 0.86667 for 15 observations, which shows that the model is moderately confident about the churn prediction.
In a lower range (0.435 - 0.472), the "Mean Predicted" value is 0.45322 with a "Mean Target" of 0.30976 for 86 observations, here the model is slightly overestimating the churn likelihood.
For the lower probability range (0.347 - 0.360), the "Mean Predicted" is 0.35253, with a "Mean Target" of 0.00000 for 18 observations, suggesting that in this range, the model predicts no churn, which aligns with the actual outcome.

5.3 Model Comparison Analysis

Gradient Boosting Result	HP Forest Result	3 Model Comparison Result


Results - Node: Gradient Boosting Diagram: AA1

le Edit View Window

Output

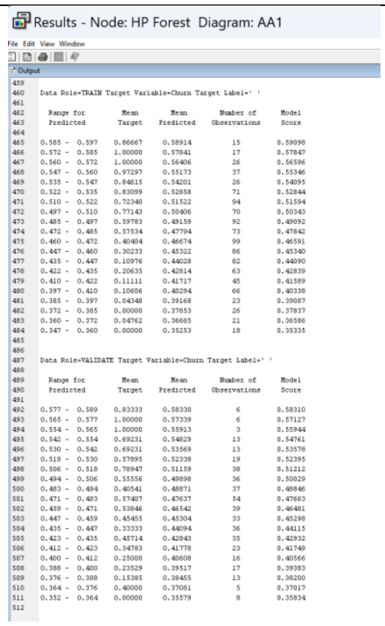
Assessment Score Distribution

Data Role=TRAIN Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.858 - 0.898	1.00000	0.89815	2	0.87812
0.818 - 0.858	1.00000	0.85454	2	0.83807
0.778 - 0.818	1.00000	0.78568	1	0.79801
0.738 - 0.778	1.00000	0.75386	16	0.75796
0.698 - 0.738	1.00000	0.72233	16	0.71790
0.658 - 0.698	0.83871	0.68004	31	0.67785
0.618 - 0.658	0.96721	0.63415	61	0.63780
0.578 - 0.618	0.84337	0.59369	83	0.59774
0.538 - 0.578	0.63636	0.55886	110	0.55769
0.498 - 0.538	0.68908	0.51782	119	0.51763
0.458 - 0.498	0.53383	0.47898	133	0.47758
0.417 - 0.458	0.32653	0.43943	98	0.43752
0.377 - 0.417	0.20690	0.39592	116	0.39747
0.337 - 0.377	0.09877	0.35991	81	0.35742
0.297 - 0.337	0.11842	0.32234	76	0.31736
0.257 - 0.297	0.06667	0.28143	45	0.27731
0.217 - 0.257	0.07407	0.24182	27	0.23725
0.177 - 0.217	0.04762	0.19979	21	0.19720
0.137 - 0.177	0.00000	0.15707	3	0.15714
0.097 - 0.137	0.00000	0.10956	9	0.11709

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.859 - 0.898	1.00000	0.89815	1	0.87837
0.819 - 0.859	1.00000	0.85454	1	0.83883
0.780 - 0.819	1.00000	0.78568	2	0.79928
0.740 - 0.780	1.00000	0.75230	8	0.75973
0.700 - 0.740	1.00000	0.70795	2	0.72018
0.661 - 0.700	0.42857	0.68365	7	0.68063
0.621 - 0.661	0.68182	0.63481	22	0.64108
0.582 - 0.621	0.70588	0.59627	34	0.60153
0.542 - 0.582	0.54545	0.56234	44	0.56198
0.503 - 0.542	0.66667	0.51891	54	0.52243
0.463 - 0.503	0.51724	0.48067	58	0.48288
0.424 - 0.463	0.48387	0.44359	62	0.44333
0.384 - 0.424	0.44898	0.40496	49	0.40378
0.344 - 0.384	0.22727	0.36780	44	0.36423
0.305 - 0.344	0.31579	0.32545	19	0.32468
0.265 - 0.305	0.17647	0.28697	17	0.28513
0.226 - 0.265	0.28571	0.25365	14	0.24559
0.186 - 0.226	0.22222	0.21177	9	0.20604
0.107 - 0.147	0.00000	0.12327	3	0.12694


Results - Node: HP Forest Diagram: AA1

File Edit View Window

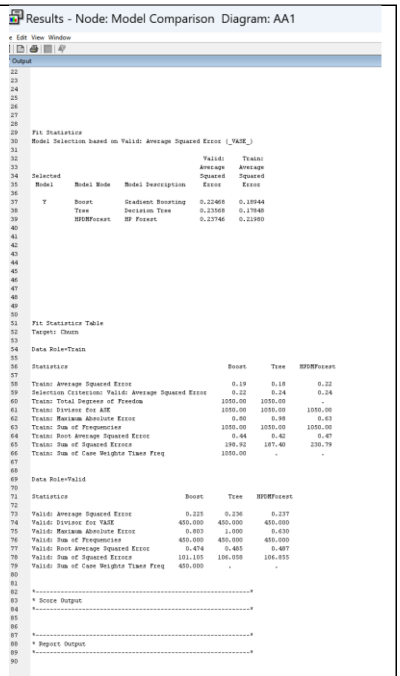
Output

Data Role=TRAIN Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.585 - 0.597	0.86667	0.58914	15	0.59098
0.572 - 0.585	1.00000	0.57841	17	0.57847
0.560 - 0.572	1.00000	0.56406	26	0.56596
0.547 - 0.560	0.97297	0.55173	37	0.55346
0.535 - 0.547	0.84615	0.54201	26	0.54095
0.522 - 0.535	0.83099	0.52858	71	0.52844
0.510 - 0.522	0.72340	0.51522	94	0.51594
0.497 - 0.510	0.77143	0.50406	70	0.50343
0.485 - 0.497	0.59783	0.49159	92	0.49092
0.472 - 0.485	0.57534	0.47794	73	0.47842
0.460 - 0.472	0.40404	0.46674	99	0.46591
0.447 - 0.460	0.30233	0.45322	86	0.45340
0.435 - 0.447	0.10976	0.44028	82	0.44090
0.422 - 0.435	0.20635	0.42814	63	0.42839
0.410 - 0.422	0.11111	0.41717	45	0.41589
0.397 - 0.410	0.10606	0.40294	66	0.40338
0.385 - 0.397	0.04348	0.39168	23	0.39087
0.372 - 0.385	0.00000	0.37853	26	0.37837
0.360 - 0.372	0.04762	0.36665	21	0.36586
0.347 - 0.360	0.00000	0.35253	18	0.35335

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.577 - 0.589	0.83333	0.58330	6	0.58310
0.565 - 0.577	1.00000	0.57339	6	0.57127
0.554 - 0.565	1.00000	0.55913	3	0.55944
0.542 - 0.554	0.69231	0.54829	13	0.54761
0.530 - 0.542	0.69231	0.53569	13	0.53578
0.518 - 0.530	0.57895	0.52338	19	0.52395
0.506 - 0.518	0.78947	0.51159	38	0.51212
0.494 - 0.506	0.55556	0.49898	36	0.50029
0.483 - 0.494	0.40541	0.48871	37	0.48846
0.471 - 0.483	0.57407	0.47637	54	0.47663
0.459 - 0.471	0.53846	0.46542	39	0.46481
0.447 - 0.459	0.45455	0.45304	33	0.45298
0.435 - 0.447	0.33333	0.44094	36	0.44115
0.423 - 0.435	0.45714	0.42843	35	0.42932
0.412 - 0.423	0.34783	0.41778	23	0.41749
0.400 - 0.412	0.25000	0.40608	16	0.40566
0.388 - 0.400	0.23529	0.39517	17	0.39383
0.376 - 0.388	0.15385	0.38455	13	0.38200
0.364 - 0.376	0.40000	0.37081	5	0.37017
0.352 - 0.364	0.00000	0.35579	8	0.35834


Results - Node: Model Comparison Diagram: AA1

e Edit View Window

Output

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Boost	Gradient Boosting	0.22468	0.18944
	Tree	Decision Tree	0.23568	0.17848
	HPDMForest	HP Forest	0.23746	0.21980

Fit Statistics Table
Target: Churn

Data Role=Train

Statistics	Boost	Tree	HPDMForest
Train: Average Squared Error	0.19	0.18	0.22
Selection Criterion: Valid: Average Squared Error	0.22	0.24	0.24
Train: Total Degrees of Freedom	1050.00	1050.00	.
Train: Divisor for ASE	1050.00	1050.00	1050.00
Train: Maximum Absolute Error	0.90	0.98	0.63
Train: Sum of Frequencies	1050.00	1050.00	1050.00
Train: Root Average Squared Error	0.44	0.42	0.47
Train: Sum of Squared Errors	198.92	187.40	230.79
Train: Sum of Case Weights Times Freq	1050.00	.	.

Data Role=Valid

Statistics	Boost	Tree	HPDMForest
Valid: Average Squared Error	0.225	0.236	0.237
Valid: Divisor for VASE	450.000	450.000	450.000
Valid: Maximum Absolute Error	0.803	1.000	0.630
Valid: Sum of Frequencies	450.000	450.000	450.000
Valid: Root Average Squared Error	0.474	0.485	0.487
Valid: Sum of Squared Errors	101.105	106.058	106.855
Valid: Sum of Case Weights Times Freq	450.000	.	.

--
* Score Output
--

--
* Report Output
--