WQD 7005

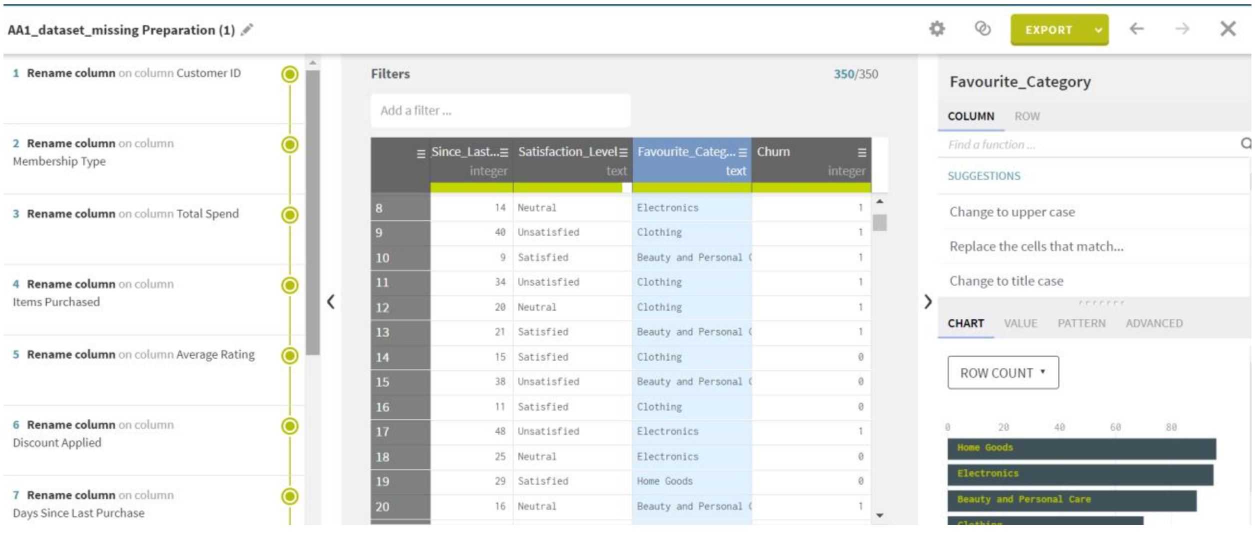
DATA MINING

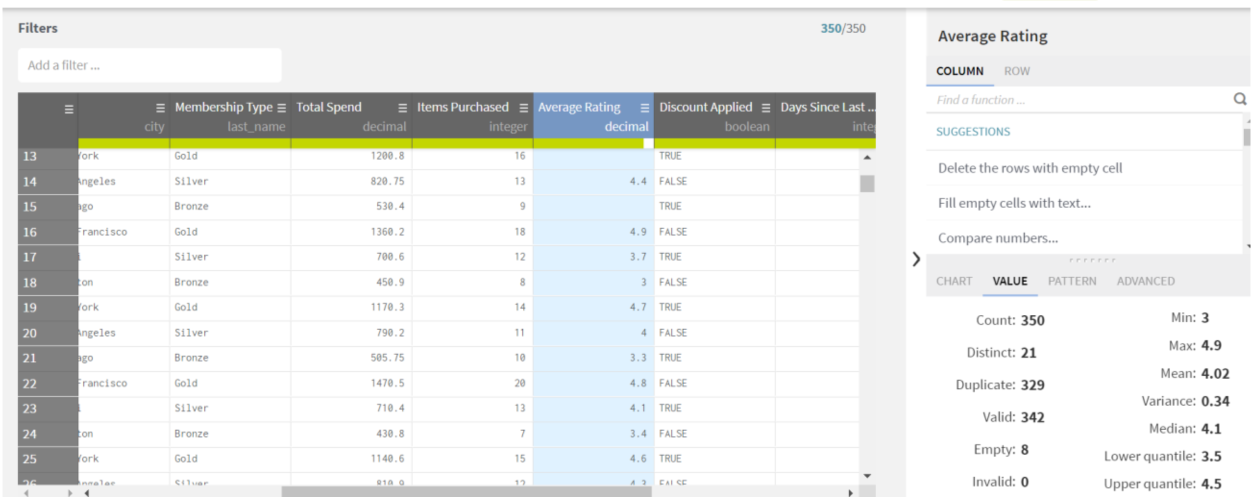
ALTERNATIVE ASSESSMENT 1

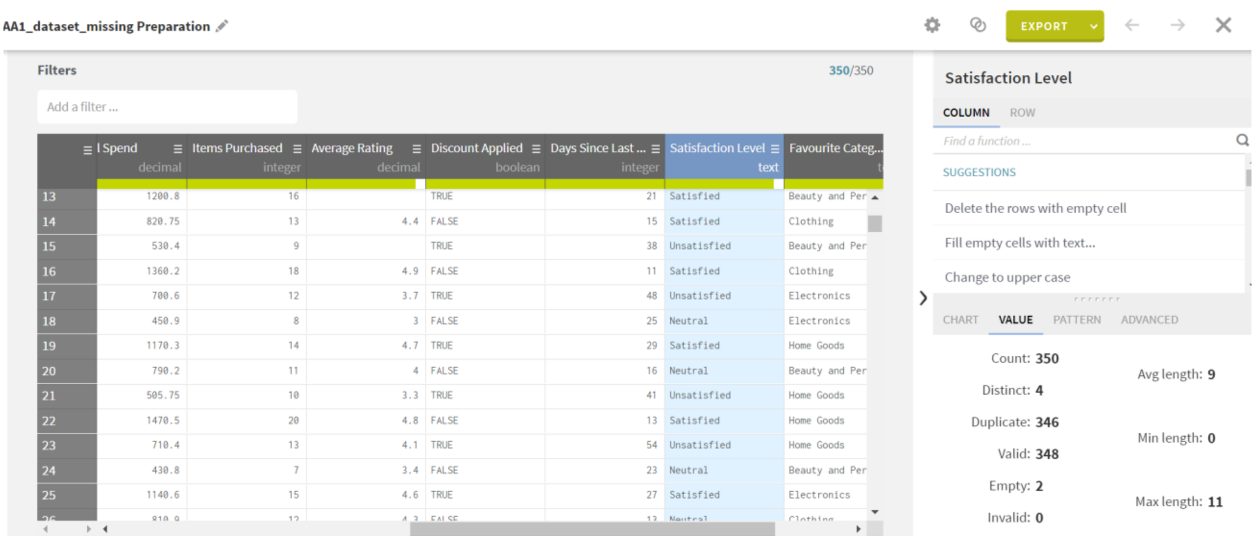
SAIDATUL HANIDA BINTI MOHD YUKHI
22082961

Github link: https://github.com/saidatulhanida/AA1

Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.







In the preprocessing task, Talend Data Preparation was utilized to rename the column, making it more readable for the subsequent creation of a data source in SAS Enterprise Miner.Additionally, preliminary checks for missing values and outliers were conducted using Talend Data Preparation. The analysis revealed the presence of missing values in the "Average Rating" and "Satisfaction Level" columns. The next step involves addressing these missing values in SAS Enterprise Miner.

Specify Variable Role

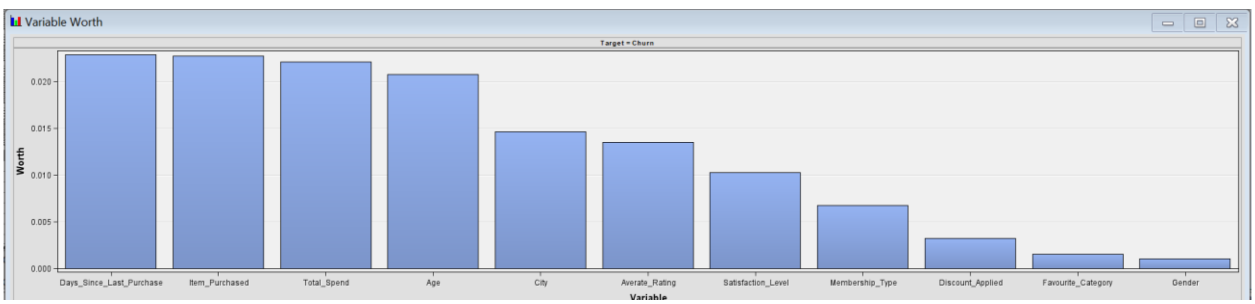
Variables - Ids

(none) not Equal to

Columns: Label Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Nominal	No		No	.	.
Averate_Rati	Input	Interval	No		No	.	.
Churn	Target	Binary	No		No	.	.
City	Input	Nominal	No		No	.	.
Customer_ID	ID	Interval	No		No	.	.
Days_Since_I	Input	Interval	No		No	.	.
Discount_Ap	Input	Binary	No		No	.	.
Favourite_Ca	Input	Nominal	No		No	.	.
Gender	Input	Binary	No		No	.	.
Item_Purcha	Input	Nominal	No		No	.	.
Membership_	Input	Nominal	No		No	.	.
Satisfaction_I	Input	Nominal	No		No	.	.
Total_Spend	Input	Interval	No		No	.	.

The roles for each variable were specified as follows: "Churn" is assigned as the target variable, "Customer_ID" serves as the ID variable, and all other variables are assigned as input variables.



By using StatExplore, from the Variable Worth result, it shows that which variable is the most important or contribute the most in predicting the churn.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

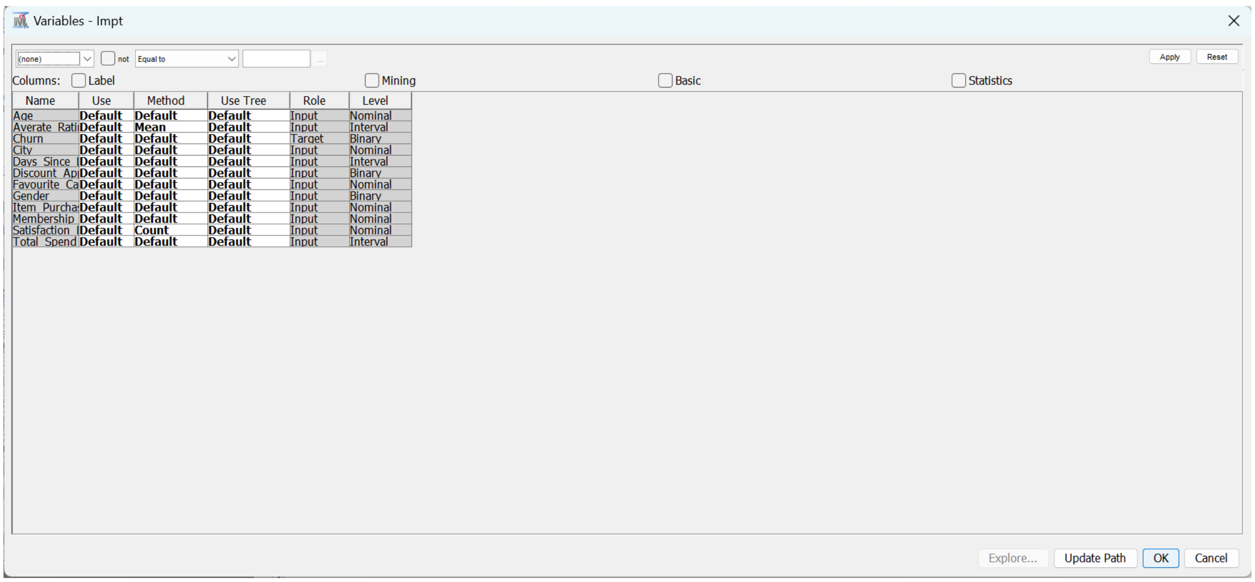
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Age	INPUT	16	0	30	13.71	32	9.43
TRAIN	City	INPUT	6	0	Los Angeles	16.86	New York	16.86
TRAIN	Discount_Applied	INPUT	2	0	FALSE	50.00	TRUE	50.00
TRAIN	Favourite_Category	INPUT	4	0	Home Goods	27.43	Electronics	27.14
TRAIN	Gender	INPUT	2	0	Female	50.00	Male	50.00
TRAIN	Item_Purchased	INPUT	15	0	10	13.43	9	9.71
TRAIN	Membership_Type	INPUT	3	0	Gold	33.43	Silver	33.43
TRAIN	Satisfaction_Level	INPUT	4	2	Satisfied	35.71	Unsatisfied	33.14
TRAIN	Churn	TARGET	2	0	0	56.57	1	43.43

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Averate_Rating	INPUT	4.022222	0.582268	342	8	3	4.1	4.9	-0.13489	-1.21476
Days_Since_Last_Purchase	INPUT	26.58857	13.44081	350	0	9	23	63	0.677545	-0.5054
Total_Spend	INPUT	845.3817	362.0587	350	0	410.8	770.2	1520.1	0.562567	-1.07986

In the summary statistics for class and interval variables, we can see that there are missing values in “Average_Rating” and “Satisfaction_Level”.



Variables - Impt

Name	Use	Method	Use Tree	Role	Level
Age	Default	Default	Default	Input	Nominal
Averate_Rati	Default	Mean	Default	Input	Interval
Churn	Default	Default	Default	Target	Binary
City	Default	Default	Default	Input	Nominal
Days_Since_l	Default	Default	Default	Input	Interval
Discount_Ap	Default	Default	Default	Input	Binary
Favourite_Ca	Default	Default	Default	Input	Nominal
Gender	Default	Default	Default	Input	Binary
Item_Purcha	Default	Default	Default	Input	Nominal
Membership	Default	Default	Default	Input	Nominal
Satisfaction_l	Default	Count	Default	Input	Nominal
Total_Spend	Default	Default	Default	Input	Interval

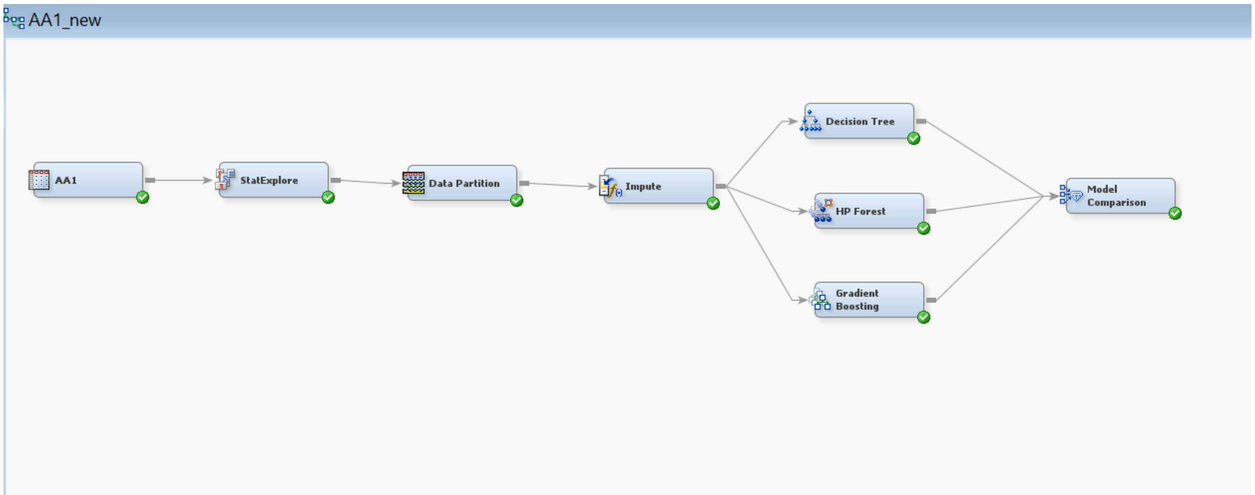
The imputation method was applied to address missing values in the variables "Average_Rating" and "Satisfaction_Level." The mean was utilized for imputing missing values in the "Average_Rating" variable, while the count method was employed for imputing missing values in “Satisfaction_Level” variable.

Imputation Summary
Number Of Observations

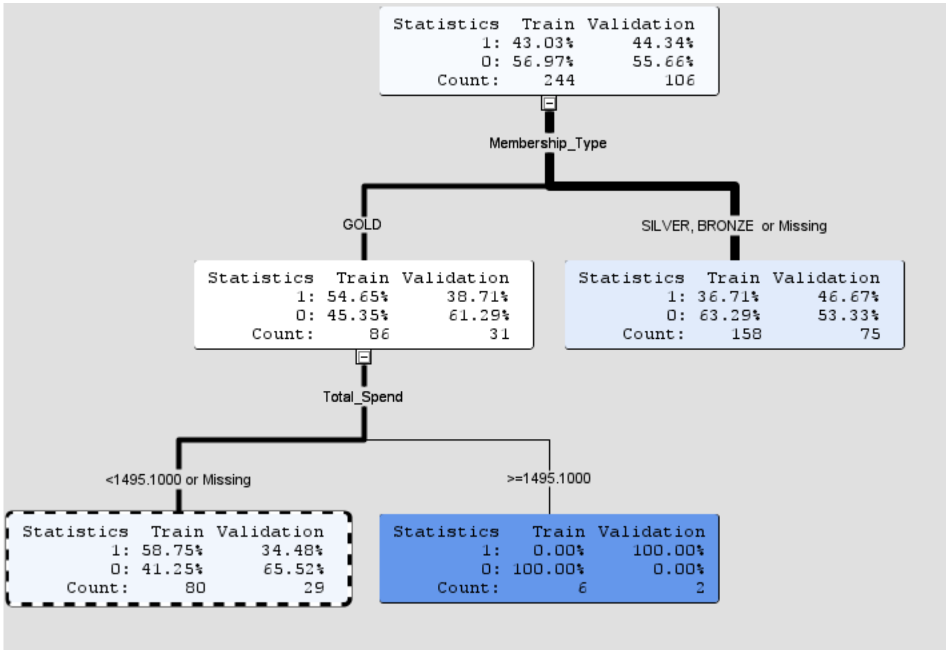
Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Averate_Rating	MEAN	IMP_Averate_Rating	4.0347280335	INPUT	INTERVAL		5
Satisfaction_Level	COUNT	IMP_Satisfaction_Level	Satisfied	INPUT	NOMINAL		2

The figures above show the summary of the imputation after using the imputation method mentioned above.

Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.



Property	Value
General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	5
Minimum Categorical Size	7
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25



Statistics	Train	Validation
Count	244	106
Prediction	0	0
% with target = 1	43.03%	44.34%
% with target = 0	56.97%	55.66%
% correctly predicted	56.97%	55.66%
Average profit with target = 1	0.4303	0.4434
Average profit with target = 0	0.5697	0.5566

The decision tree model provides valuable insights into factors influencing customer churn. Based on the identified conditions, Total_Spent < 1495.1 and Membership_Type is Gold, we observe a higher likelihood of churn within this subgroup.

```
Classification Table

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

                      Target      Outcome     Frequency     Total
Target    Outcome   Percentage   Percentage     Count     Percentage

  0          0        56.9672       100          139        56.9672
  1          0        43.0328       100          105        43.0328


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

                      Target      Outcome     Frequency     Total
Target    Outcome   Percentage   Percentage     Count     Percentage

  0          0        55.6604       100           59        55.6604
  1          0        44.3396       100           47        44.3396
```

The model predicts all instances as Churn = 0 (No Churn), resulting in misclassification of Churn = 1 (Churn) instances. This may indicate a potential issue with the model's ability to distinguish between the two classes, and further evaluation or adjustments to the model may be necessary to improve its performance.

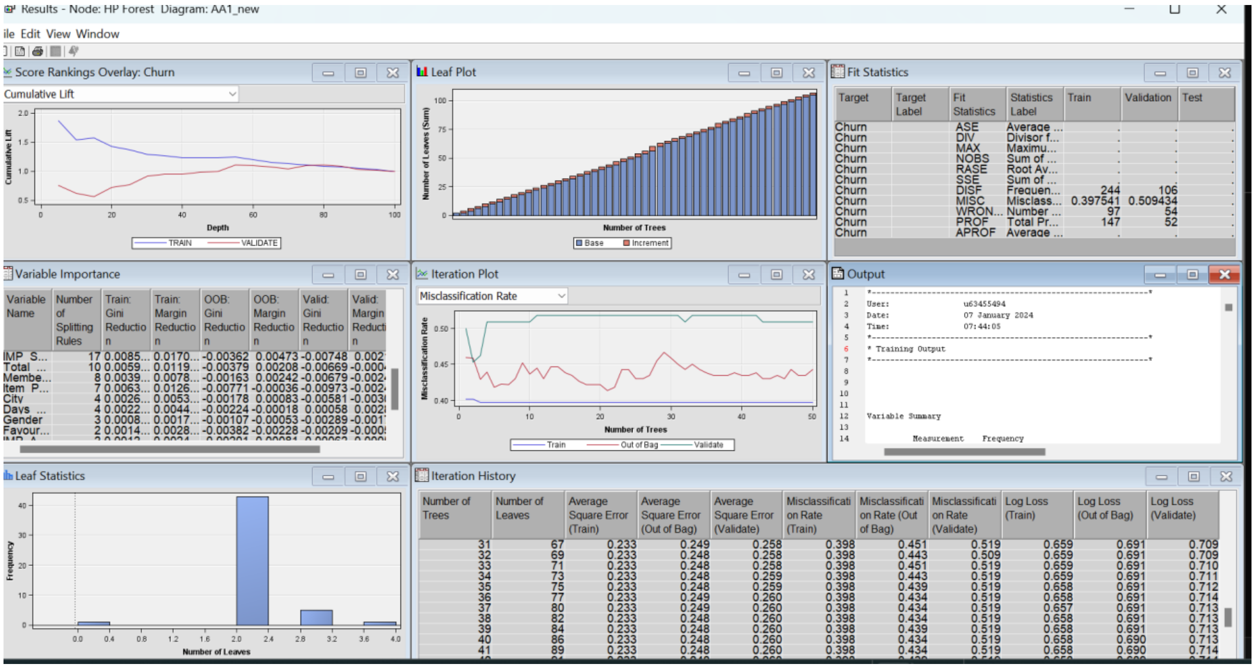
Ensemble Methods: Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.

Random Forest

Applying Random Forest for Bagging:

Firstly, connect HP Forest Node to the Impute Node. Then, specify the target variable and input variables by right click the node.

Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sam	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Conside	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs in N	1.0E-5
Smallest Number of Obs in Noc	1
Split Size	.
Use as Modeling Node	Yes
Score	
Variable Selection	Yes
Variable Importance Method	Loss Reduction
Number of Variables to Conside	25
Cutoff Fraction	0.01
Status	
Create Time	1/7/24 7:31 AM
Run ID	5ec2eb90-b1a6-d74e-b2f2-fdcc
Last Error	
Last Status	Complete
Last Run Time	1/7/24 7:44 AM



	Predicted 0	Predicted 1
Actual 0	76.26%	23.74%
Actual 1	60.95%	39.05%

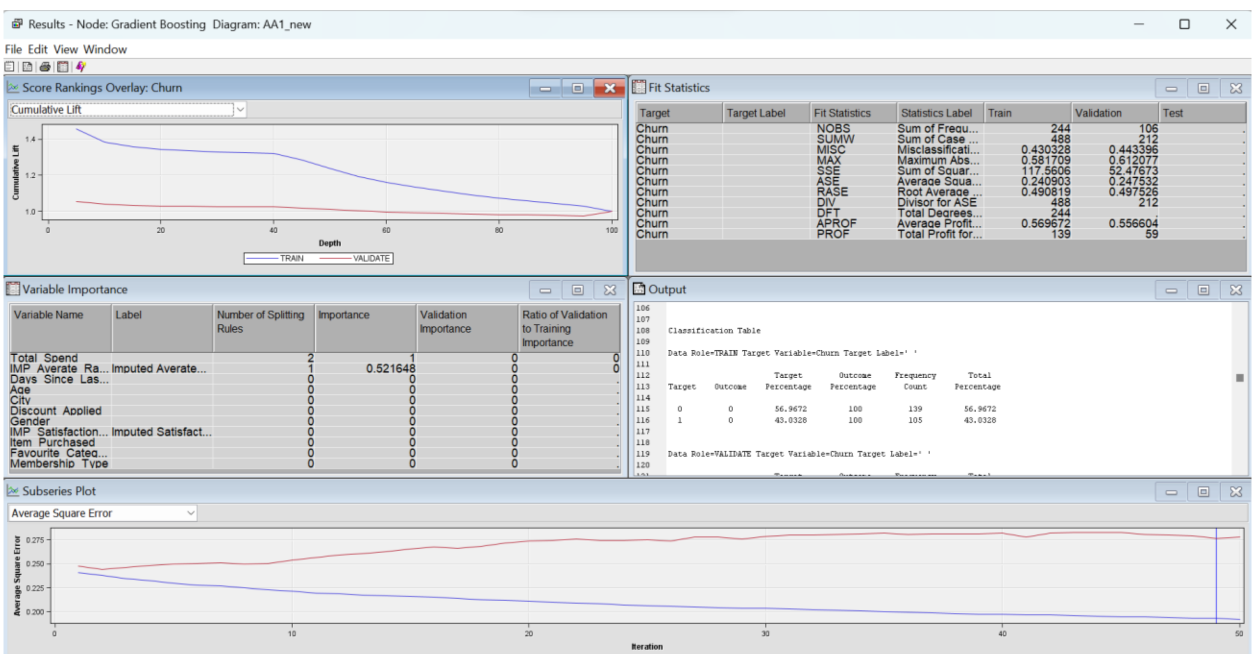
The table shows outcomes based on predictions and actual occurrences. For "TRAIN," when predicting non-churn (Target 0), the model is accurate in 76.26% of cases, and for predicting churn (Target 1), the accuracy is 39.05%.

Gradient Boosting

Applying Random Forest for Bagging:

Firstly, connect Gradient Boosting Node to the Impute Node. Then, specify the target variable and input variables by right click the node. The number of maximum branch, maximum depth, etc., used are the default setting.

Property	Value
Exported Data	...
Notes	...
Train	
Variables	...
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Decision
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Importance	No
Number Single Var Importance	5
Status	
Create Time	1/7/24 7:32 AM



Classification Table

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

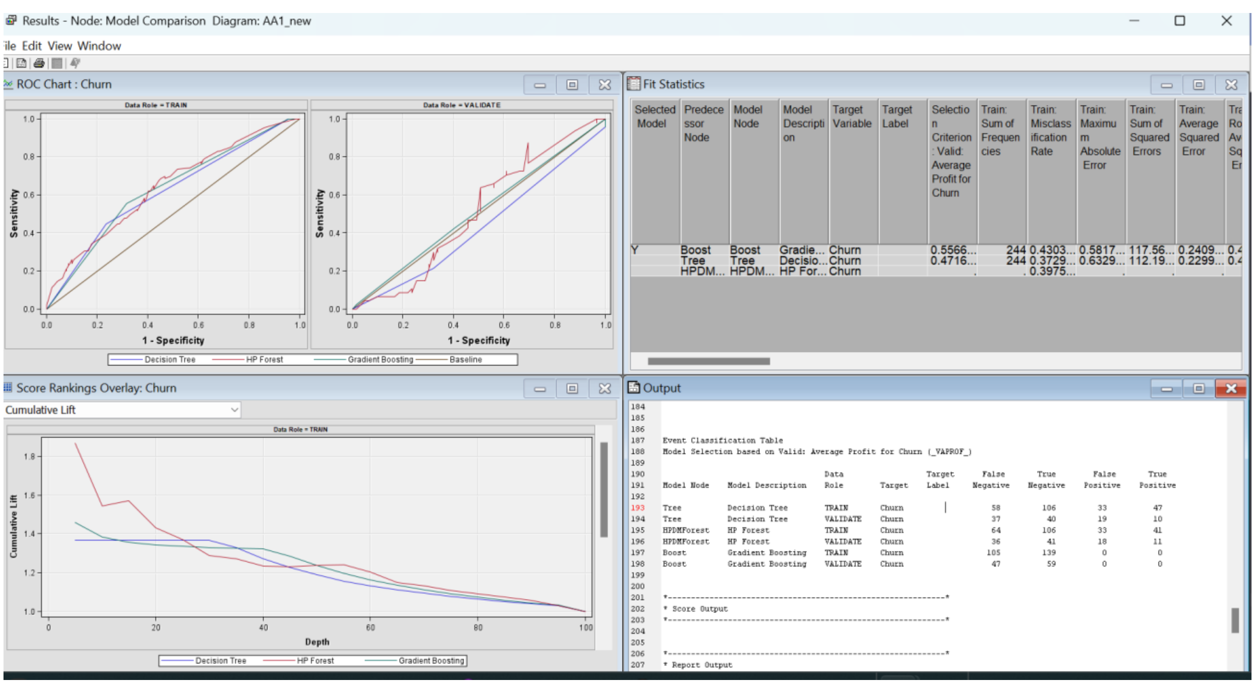
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	56.9672	100	139	56.9672
1	0	43.0328	100	105	43.0328

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	55.6604	100	59	55.6604
1	0	44.3396	100	47	44.3396

The model predicts all instances as Churn = 0 (No Churn), resulting in misclassification of Churn = 1 (Churn) instances. This may indicate a potential issue with the model's ability to distinguish between the two classes, and further evaluation or adjustments to the model may be necessary to improve its performance.

Model Comparison



Based on the average profit for Churn, the Decision Tree and Gradient Boosting models both have the same score in the VALIDATE dataset, while the Gradient Boosting model has a higher score in the TRAIN dataset.