WQD7005 DATA MINING 2023/2024 S1
ALTERNATIVE ASSESSMENT 1

Name: Xin Dong
Matric Num: 22060696

Submitted Answer:

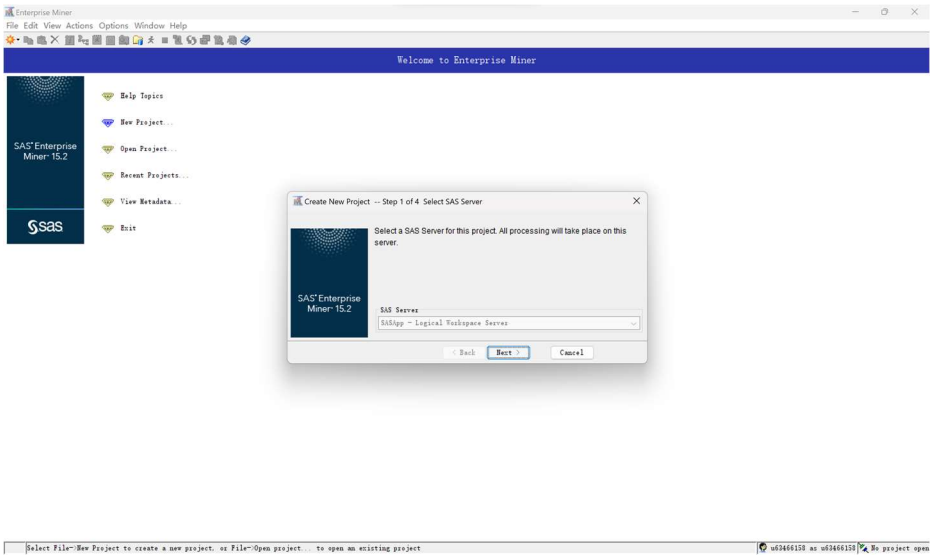
Github Link Repository: https://github.com/xinburg/Final-Assessment-1_Xin-Dong_22060696

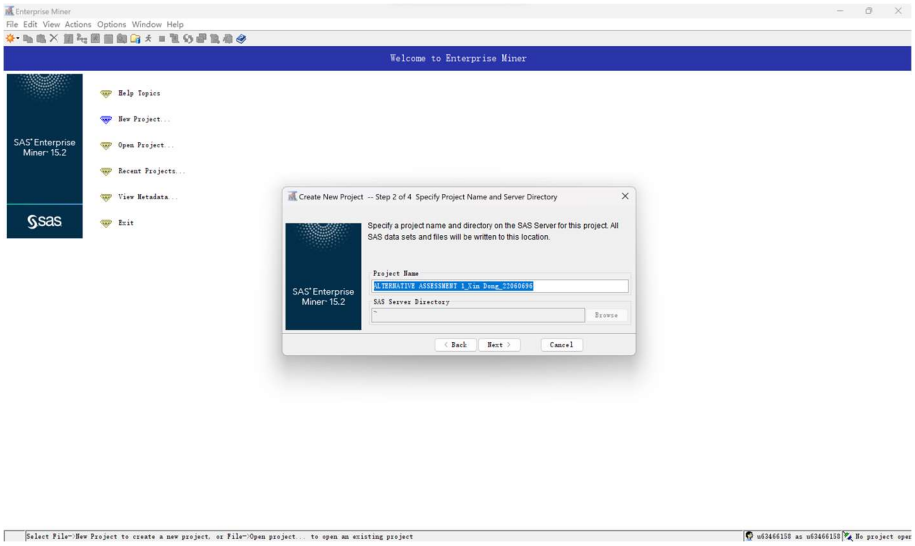
Dataset reference: https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis/data

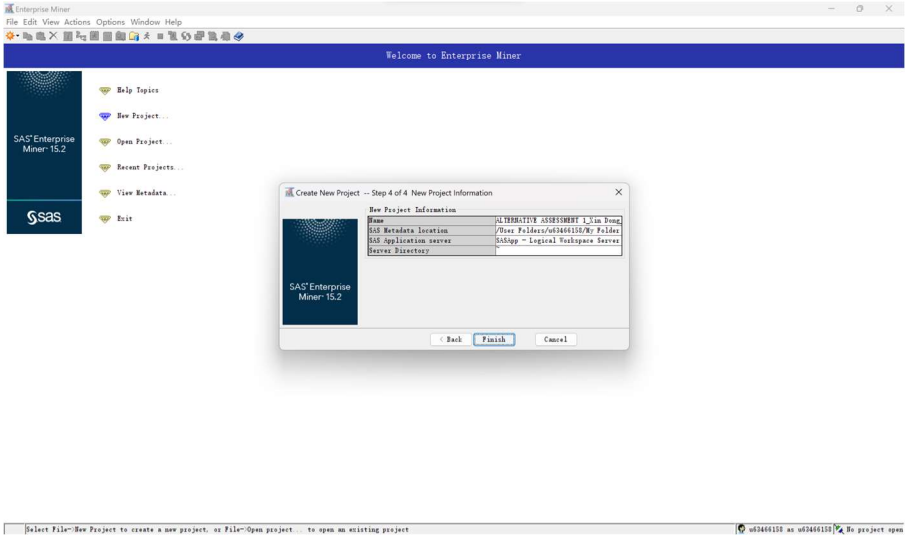
In my study, a new practical dataset was generated based on the real-world reference dataset so that it can closely match the AA1's criteria and allows for a controlled environment to test various analytical methods. The generated data is realistic and representative of actual e-commerce customer behavior.

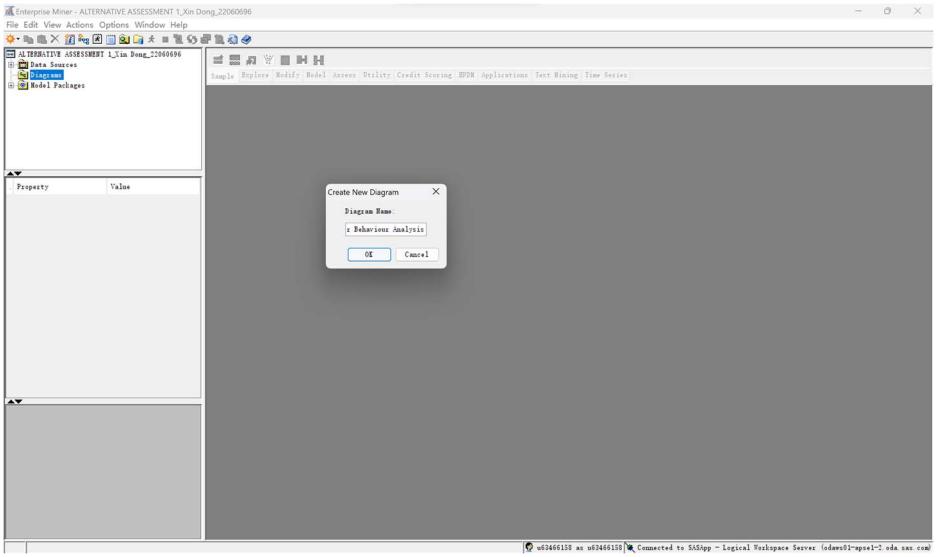
1. Create Project and Data Import in SAS Enterprise Miner

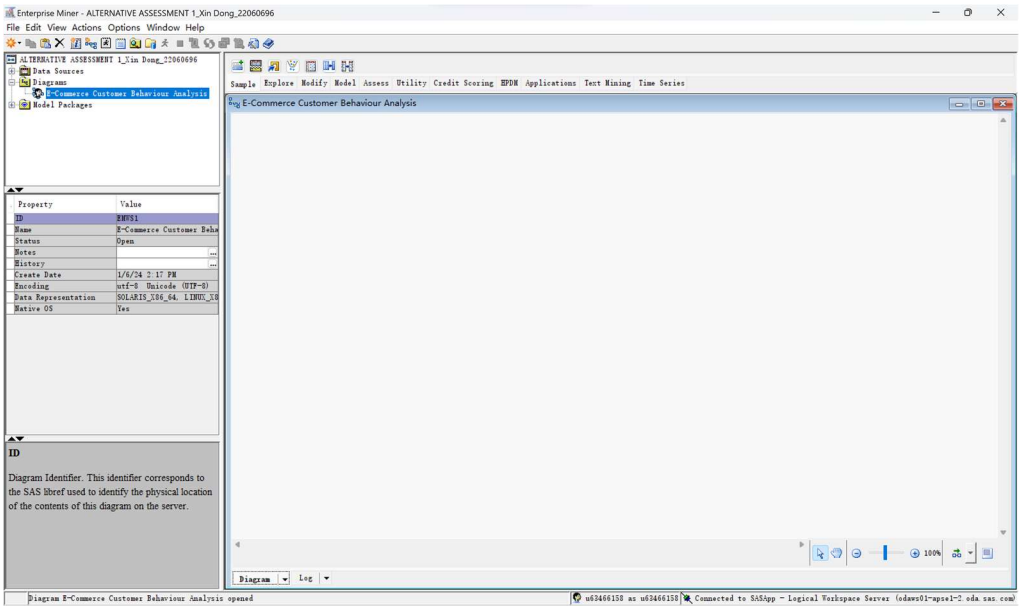
Start SAS Enterprise Miner: Open the software and create a new project named ALTERNATIVE ASSESSMENT 1_Xin Dong_22060696. And then create a new Diagram and set the name as: E-Commerce Customer Behaviour Analysis.




Enterprise Miner
File Edit View Actions Options Window Help
Welcome to Enterprise Miner
SAS Enterprise Miner 15.2
Help Topics
New Project...
Open Project...
Recent Projects...
View Metadata...
Exit
Create New Project -- Step 2 of 4 Specify Project Name and Server Directory
Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.
Project Name
ALTERNATIVE ASSESSMENT 1_Xin Dong_22060696
SAS Server Directory
Browse
< Back
Next >
Cancel
Select File->New Project to create a new project, or File->Open project... to open an existing project
u63466158 as u63466158
No project open

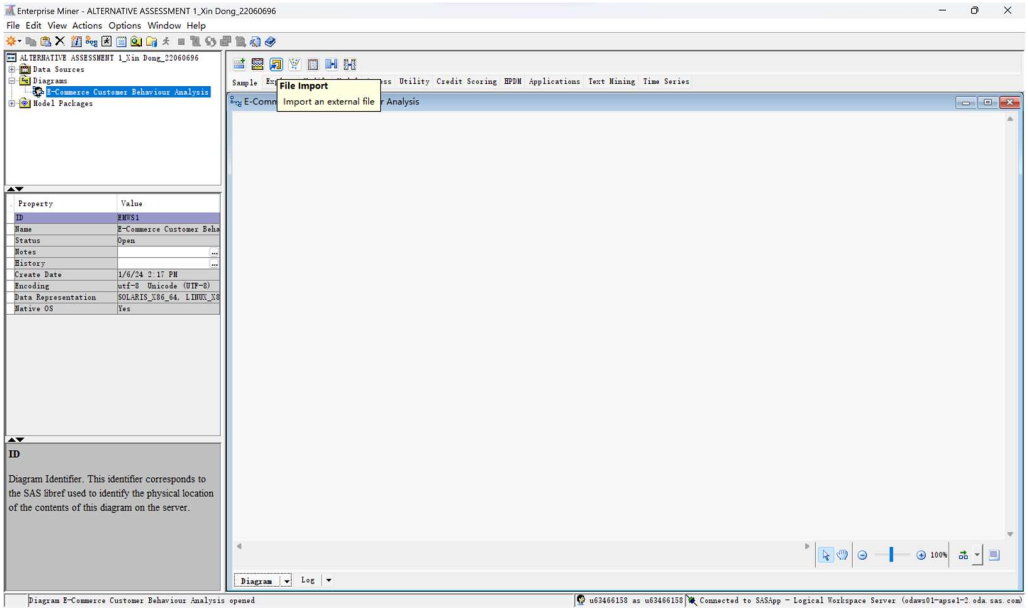

Enterprise Miner
File Edit View Actions Options Window Help
Welcome to Enterprise Miner
SAS Enterprise Miner 15.2
Help Topics
New Project...
Open Project...
Recent Projects...
View Metadata...
Exit
Create New Project -- Step 4 of 4 New Project Information
New Project Information
Name
ALTERNATIVE ASSESSMENT 1_Xin Dong
SAS Metadata location
/User Folders/u63466158/My Folder
SAS Application server
SASApp - Logical Workspace Server
Server Directory
< Back
Finish
Cancel
Select File->New Project to create a new project, or File->Open project... to open an existing project
u63466158 as u63466158
No project open

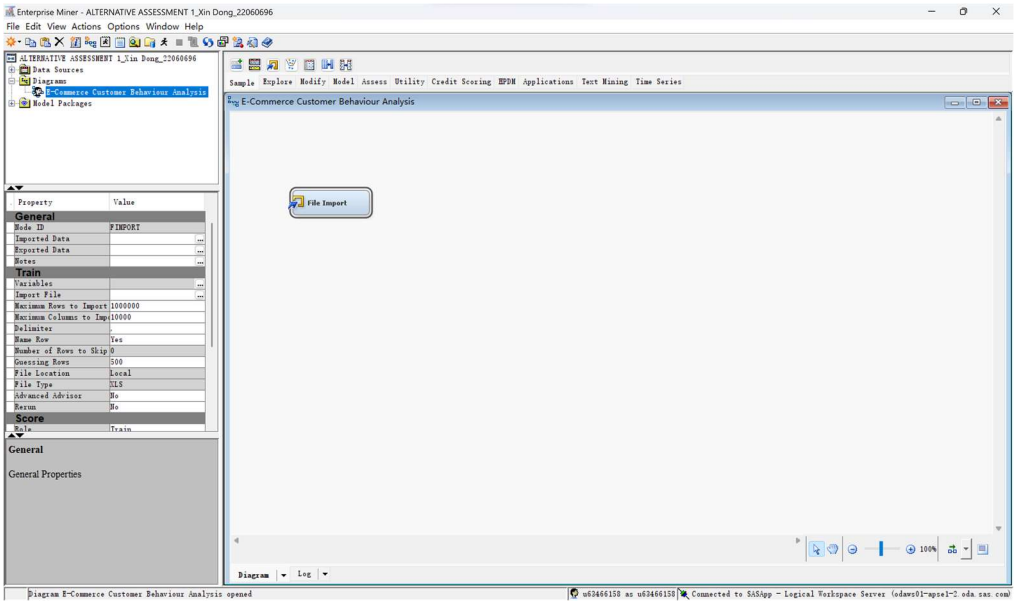

Enterprise Miner - ALTERNATIVE ASSESSMENT 1_Xin Dong_22060696
File Edit View Actions Options Window Help
ALTERNATIVE ASSESSMENT 1_Xin Dong_22060696
Data Sources
Diagrams
Model Packages
Property
Value
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
Create New Diagram
Diagram Name
x Behaviour Analysis
OK
Cancel
u63466158 as u63466158
Connected to SASApp - Logical Workspace Server (odaws01-apse1-2.oda.sas.com)



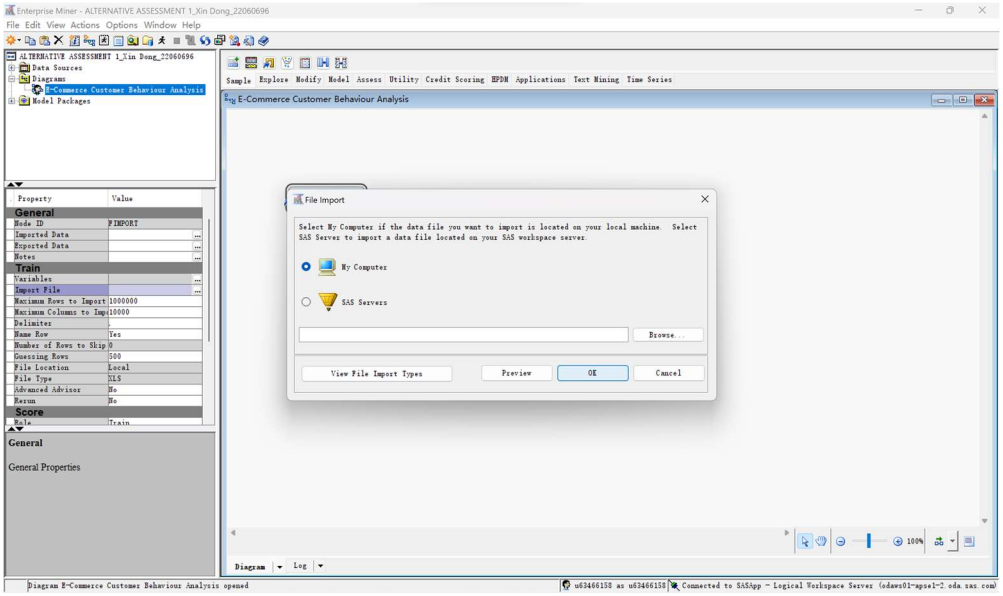
Import Dataset:

Navigate to the File menu and select Import Data.

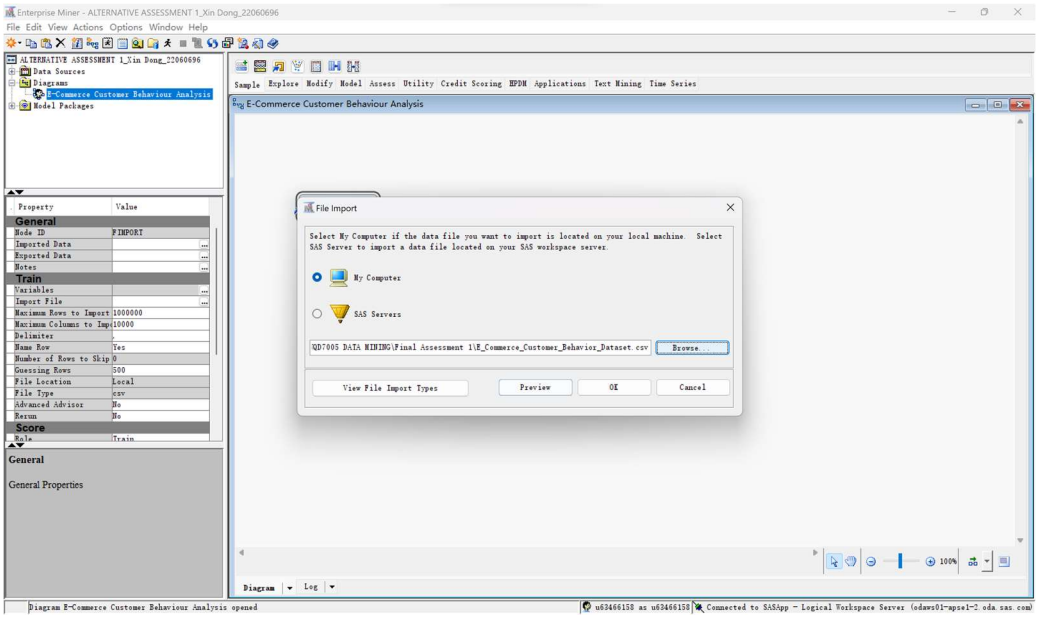




Click on File Import icon and select ‘Import the file’ from the left side configuration bar.



Browse and select the CSV file as the dataset we are going to analyze.

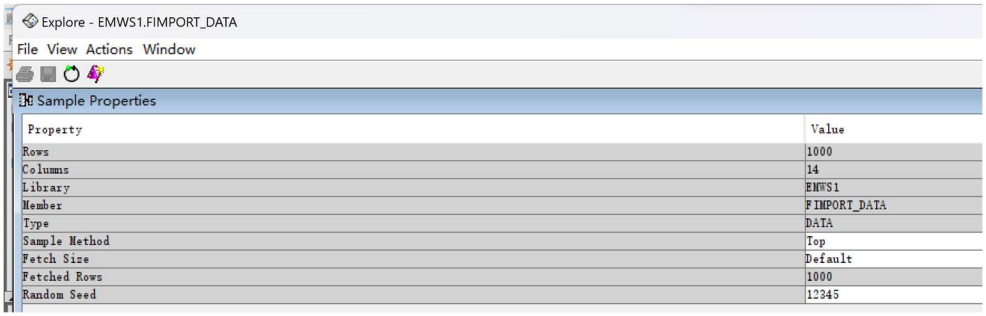


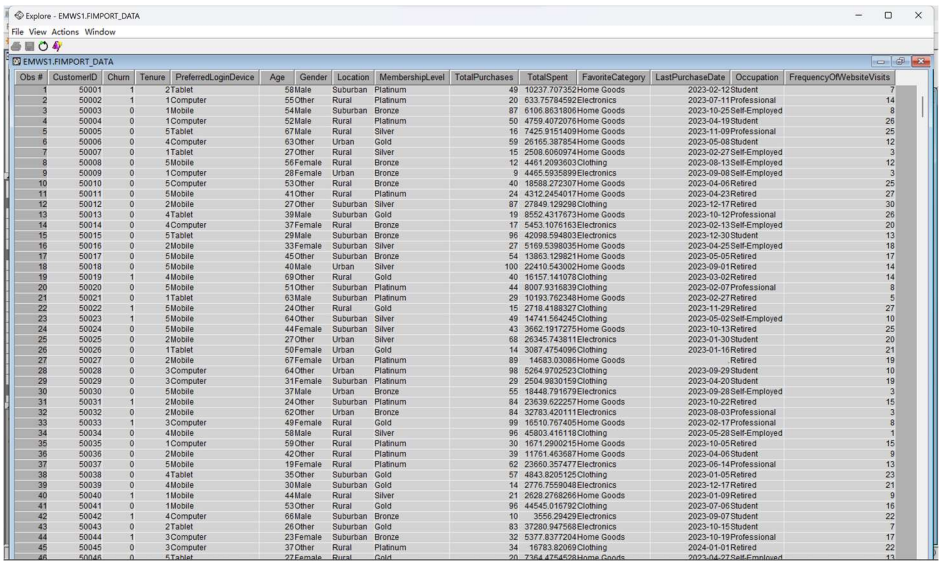
Follow the prompts in the Import Wizard to complete the import process.

2. Data Exploration

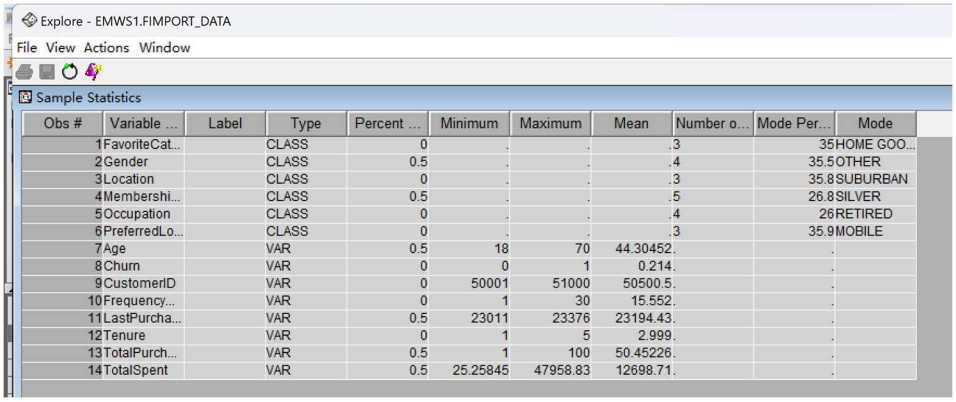
View Data: Once imported, view the dataset to ensure it's correctly loaded.
Summary Statistics: Generate summary statistics to understand distributions, averages, and other statistical measures for each variable. This step is crucial for identifying anomalies or patterns that might require addressing during preprocessing.

The details of the dataset as shown in below pictures:

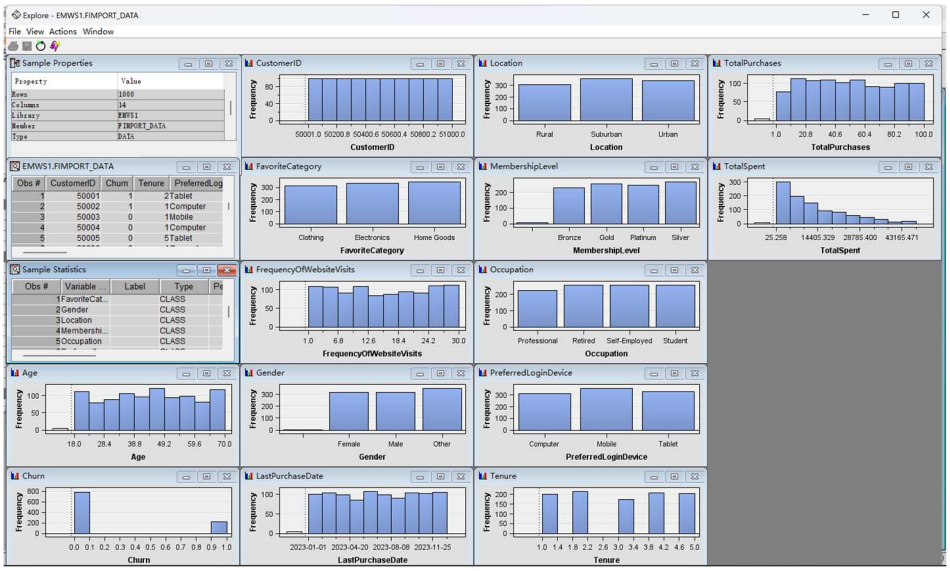




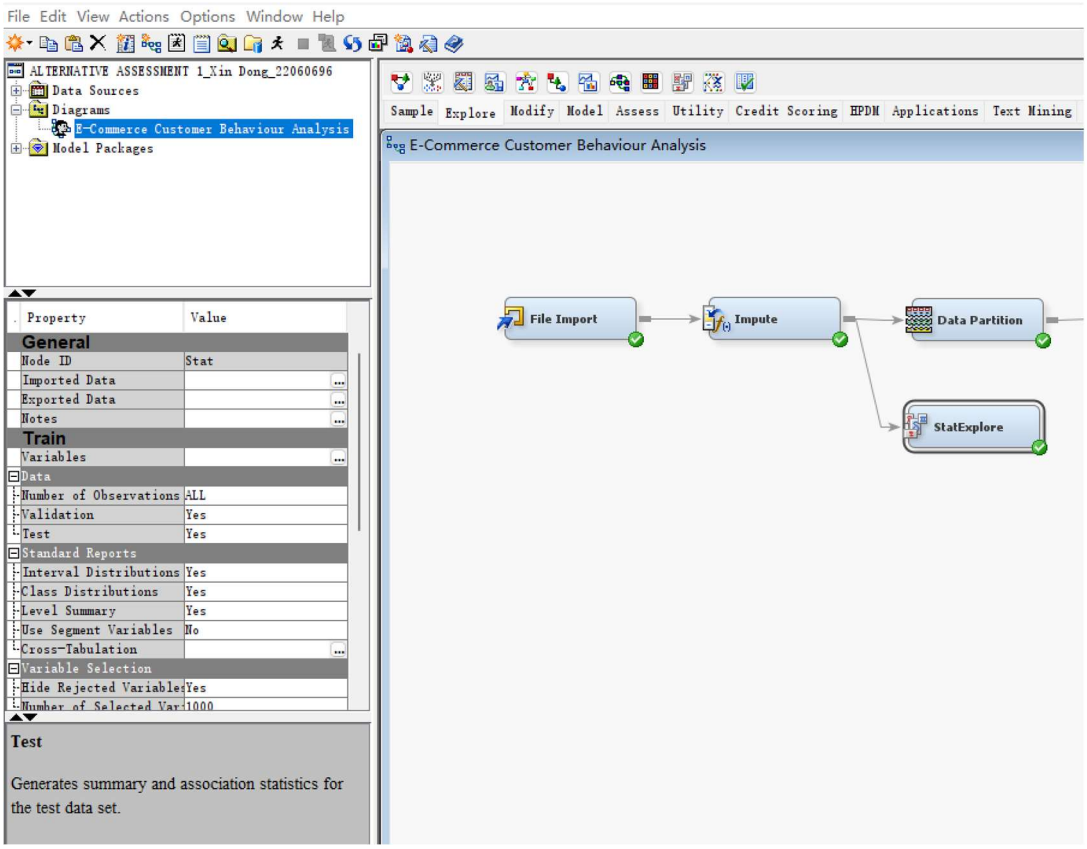
Obs #	CustomerID	Churn	Tenure	PreferredLoginDevice	Age	Gender	Location	MembershipLevel	TotalPurchases	TotalSpent	FavoriteCategory	LastPurchaseDate	Occupation	FrequencyOfWebsiteVisits
1	50001	1	2	Tablet	58	Male	Suburban	Platinum	49	10237.707352	Home Goods	2023-02-12	Student	7
2	50002	1	1	Computer	55	Other	Rural	Platinum	20	633.75784592	Electronics	2023-07-11	Professional	14
3	50003	0	1	Mobile	54	Male	Suburban	Bronze	87	6106.8631806	Home Goods	2023-10-25	Self-Employed	8
4	50004	0	1	Computer	52	Male	Rural	Platinum	50	4759.4072076	Home Goods	2023-04-19	Student	26
5	50005	0	5	Tablet	67	Male	Rural	Silver	16	7425.9151409	Home Goods	2023-11-09	Professional	25
6	50006	0	4	Computer	63	Other	Urban	Gold	59	26165.387854	Home Goods	2023-05-08	Student	12
7	50007	0	1	Tablet	27	Other	Rural	Silver	15	2508.6060974	Home Goods	2023-02-27	Self-Employed	3
8	50008	0	5	Mobile	56	Female	Rural	Bronze	12	4461.2093603	Clothing	2023-08-13	Self-Employed	12
9	50009	0	1	Computer	28	Female	Urban	Bronze	9	4465.5935899	Electronics	2023-09-08	Self-Employed	3
10	50010	0	5	Computer	53	Other	Rural	Bronze	40	18588.272307	Home Goods	2023-04-06	Retired	25
11	50011	0	5	Mobile	41	Other	Rural	Platinum	24	4312.2454017	Home Goods	2023-04-23	Retired	27
12	50012	0	2	Mobile	27	Other	Suburban	Silver	87	27849.129298	Clothing	2023-12-17	Retired	30
13	50013	0	4	Tablet	39	Male	Suburban	Gold	19	8552.4317673	Home Goods	2023-10-12	Professional	26
14	50014	0	4	Computer	37	Female	Rural	Bronze	17	5453.1076163	Electronics	2023-02-13	Self-Employed	20
15	50015	0	5	Tablet	29	Male	Suburban	Bronze	96	42098.594803	Electronics	2023-12-30	Student	13
16	50016	0	2	Mobile	33	Female	Suburban	Silver	27	5169.5398035	Home Goods	2023-04-25	Self-Employed	18
17	50017	0	5	Mobile	45	Other	Suburban	Bronze	54	13863.129821	Home Goods	2023-05-05	Retired	17
18	50018	0	5	Mobile	40	Male	Urban	Silver	100	22410.543002	Home Goods	2023-09-01	Retired	14
19	50019	1	4	Mobile	69	Other	Rural	Gold	40	16157.141078	Clothing	2023-03-02	Retired	14
20	50020	0	5	Mobile	51	Other	Suburban	Platinum	44	8007.9316839	Clothing	2023-02-07	Professional	8
21	50021	0	1	Tablet	63	Male	Suburban	Platinum	29	10193.762348	Home Goods	2023-02-27	Retired	5
22	50022	1	5	Mobile	24	Other	Rural	Gold	15	2718.4188327	Clothing	2023-11-29	Retired	27
23	50023	1	5	Mobile	64	Other	Suburban	Silver	49	14741.564245	Clothing	2023-05-02	Self-Employed	10
24	50024	0	5	Mobile	44	Female	Suburban	Silver	43	3662.1917275	Home Goods	2023-10-13	Retired	25
25	50025	0	2	Mobile	27	Other	Urban	Silver	68	26345.743811	Electronics	2023-01-30	Student	20
26	50026	0	1	Tablet	50	Female	Urban	Gold	14	3087.4754096	Clothing	2023-01-16	Retired	21
27	50027	0	2	Mobile	67	Female	Urban	Platinum	89	14683.03089	Home Goods		Retired	19
28	50028	0	3	Computer	64	Other	Urban	Platinum	98	5264.9702523	Clothing	2023-09-29	Student	10
29	50029	0	3	Computer	31	Female	Suburban	Platinum	29	2504.9830159	Clothing	2023-04-20	Student	19
30	50030	0	5	Mobile	37	Male	Urban	Bronze	55	18448.7916799	Electronics	2023-09-28	Self-Employed	3
31	50031	1	2	Mobile	24	Other	Suburban	Platinum	84	23639.622257	Home Goods	2023-10-22	Retired	15
32	50032	0	2	Mobile	62	Other	Urban	Bronze	84	32783.420111	Electronics	2023-08-03	Professional	3
33	50033	1	3	Computer	49	Female	Rural	Gold	99	16510.767405	Home Goods	2023-02-17	Professional	8
34	50034	0	4	Mobile	58	Male	Rural	Silver	96	45803.416118	Clothing	2023-05-28	Self-Employed	1
35	50035	0	1	Computer	59	Other	Rural	Platinum	30	1671.2900215	Home Goods	2023-10-05	Retired	15
36	50036	0	2	Mobile	42	Other	Rural	Platinum	39	11761.463687	Home Goods	2023-04-06	Student	9
37	50037	0	5	Mobile	19	Female	Rural	Platinum	62	23660.357477	Electronics	2023-06-14	Professional	13
38	50038	0	4	Tablet	35	Other	Suburban	Gold	57	4843.8205125	Clothing	2023-01-05	Retired	23
39	50039	0	4	Mobile	30	Male	Suburban	Gold	14	2776.7559048	Electronics	2023-12-17	Retired	21
40	50040	1	1	Mobile	44	Male	Rural	Silver	21	2628.2768266	Home Goods	2023-01-09	Retired	9
41	50041	0	1	Mobile	53	Other	Rural	Gold	96	44545.016792	Clothing	2023-07-06	Student	16
42	50042	1	4	Computer	66	Male	Suburban	Bronze	10	3556.29429	Electronics	2023-09-07	Student	22
43	50043	0	2	Tablet	26	Other	Suburban	Gold	83	37280.947568	Electronics	2023-10-15	Student	7
44	50044	1	3	Computer	23	Female	Suburban	Bronze	32	5377.8377204	Home Goods	2023-10-19	Professional	17
45	50045	0	3	Computer	37	Other	Rural	Platinum	34	16783.82069	Clothing	2024-01-01	Retired	22
46	50046	0	5	Tablet	27	Female	Rural	Gold	20	7364.4754528	Home Goods	2023-04-27	Self-Employed	13

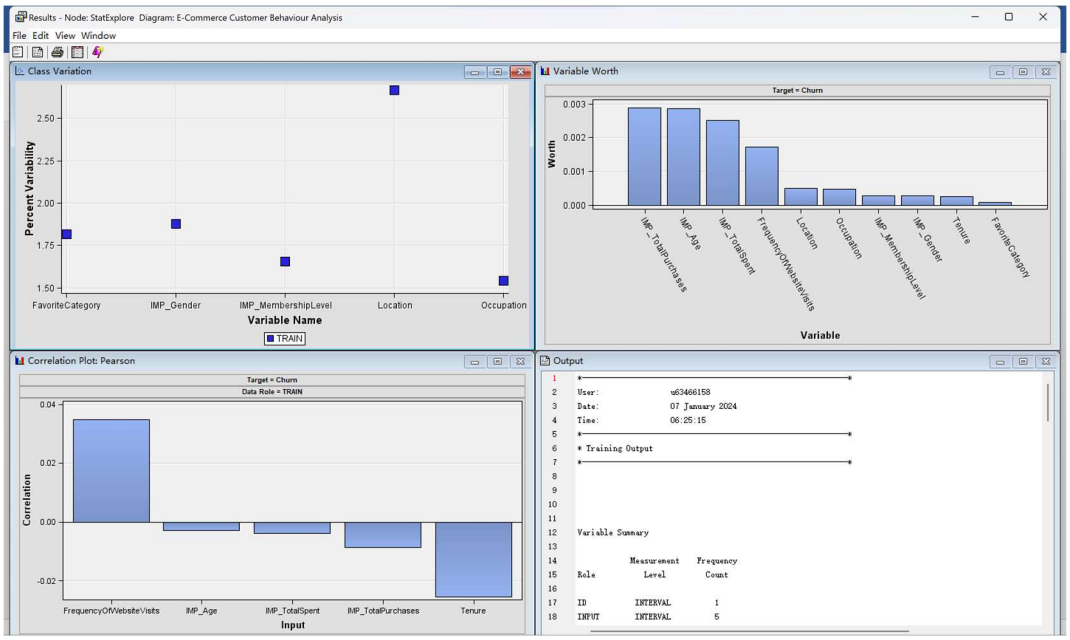


Obs #	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
1	FavoriteCat...		CLASS	0	.	.	.	3	35	HOME GOO...
2	Gender		CLASS	0.5	.	.	.	4	35.5	OTHER
3	Location		CLASS	0	.	.	.	3	35.8	SUBURBAN
4	Membershi...		CLASS	0.5	.	.	.	5	26.8	SILVER
5	Occupation		CLASS	0	.	.	.	4	26	RETIRED
6	PreferredLo...		CLASS	0	.	.	.	3	35.9	MOBILE
7	Age		VAR	0.5	18	70	44.30452	.	.	
8	Churn		VAR	0	0	1	0.214	.	.	
9	CustomerID		VAR	0	50001	51000	50500.5	.	.	
10	Frequency...		VAR	0	1	30	15.552	.	.	
11	LastPurcha...		VAR	0.5	23011	23376	23194.43	.	.	
12	Tenure		VAR	0	1	5	2.999	.	.	
13	TotalPurch...		VAR	0.5	1	100	50.45226	.	.	
14	TotalSpent		VAR	0.5	25.25845	47958.83	12698.71	.	.	



We could also Add one StatExplore node to explore the data:





From the exploration table and plots, we can summarize an attribute table:

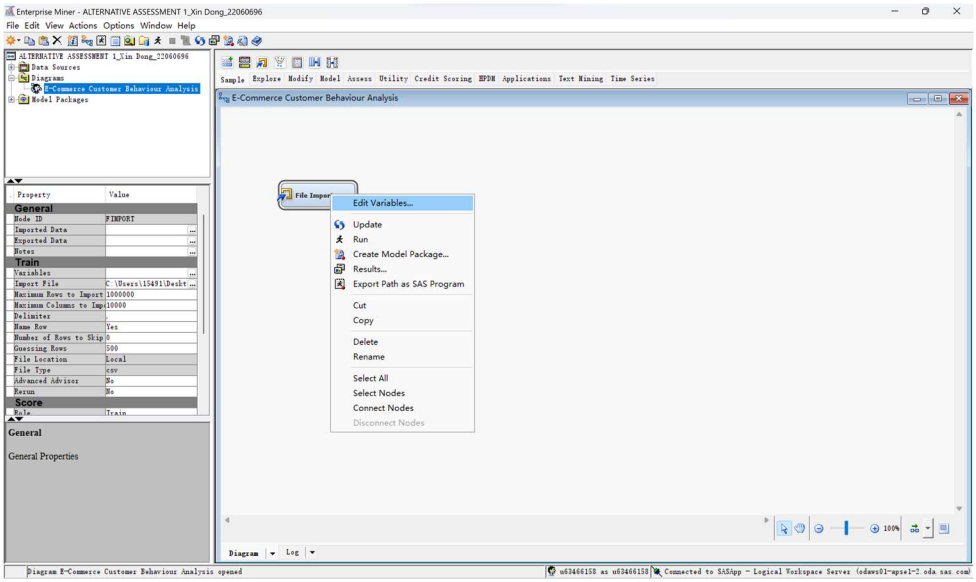
Variable	Role	Level	Rationale

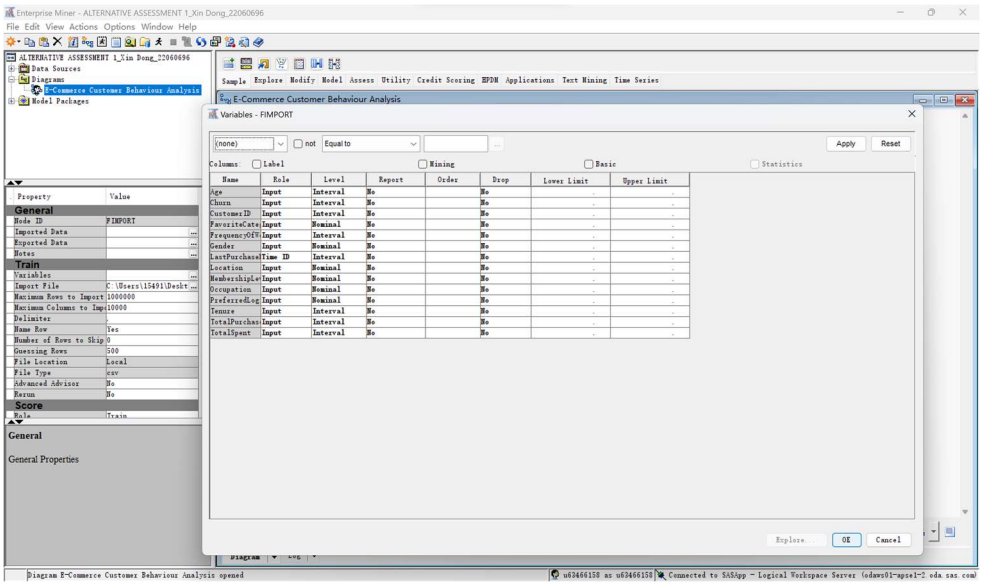
Age	Input	Interval	Age is likely to impact customer behavior and could be a predictor of churn.
Churn	Target	Binary	This is the variable you are trying to predict.
CustomerID	ID	Nominal	Unique identifier for each record, used to track customers but not for prediction.
FavoriteCategory	Input	Nominal	Shopping preferences could influence churn decisions.
FrequencyOfWebsiteVisits	Input	Interval	Frequency of visits may correlate with customer engagement and churn.
Gender	Input	Nominal	Gender may influence purchasing patterns and thus affect churn.
LastPurchaseDate	Time ID	Interval	Important for time-series analysis, if applicable, but not used directly in churn prediction.
Location	Input	Nominal	Customer location might affect purchasing habits and churn likelihood.
MembershipLevel	Input	Nominal	Membership status could impact customer loyalty and churn rate.
Occupation	Input	Nominal	Occupation might relate to disposable income and purchase frequency, affecting churn.
PreferredLoginDevice	Rejected	Nominal	May have high cardinality or low predictive power for churn.
Tenure	Input	Interval	The length of time a customer has been with the company might influence churn.
TotalPurchases	Input	Interval	Total number of purchases is directly related to customer engagement and churn.
TotalSpent	Input	Interval	Total expenditure could indicate the value of the customer and correlate with churn.

3. Specify Variable Roles and Levels

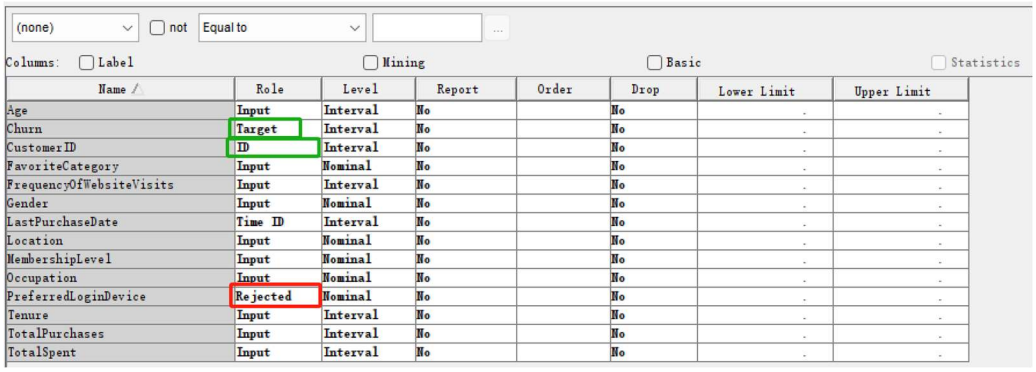
Set Target Variable: Ensure that the 'Churn' variable is set as the target.
Input Variables: Set other variables like 'Age', 'TotalSpent', etc., as input.
ID Variable: Set 'CustomerID' as the ID variable.

Click on ‘Edit Variables’, as shown in below is the original status of the variables in the dataset:





Specify the variable roles and levels, as shown in the colorful boxes below:



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Interval	No		No	.	.
CustomerID	ID	Interval	No		No	.	.
FavoriteCategory	Input	Nominal	No		No	.	.
FrequencyOfWebsiteVisits	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchaseDate	Time ID	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipLevel	Input	Nominal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
PreferredLoginDevice	Rejected	Nominal	No		No	.	.
Tenure	Input	Interval	No		No	.	.
TotalPurchases	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

Churn (Target): This variable has been set from ‘input’ into ‘target’ because it is the outcome we are trying to predict. In customer churn analysis, the goal is to determine which factors influence whether a customer will stop making purchases. Hence, it should be the dependent variable in our predictive model. And the Level should be ‘Binary’.

CustomerID (ID): Customer ID is a unique identifier for each customer and does not contain predictive information that can be generalized to unseen data. Therefore, it's appropriately set as an ‘ID’ variable to index the records without using it as part of the prediction.

PreferredLoginDevice (Rejected): There could be several reasons to reject this variable:
If there's a high proportion of missing values that are not easily imputable, it may not be reliable for use in the model. The variable may not show significant variance or impact in exploratory data analysis, suggesting it doesn't contribute much to the prediction of churn. The variable could have too many categories, leading to a sparse representation that is not useful for the model. It might be highly correlated with another variable, and to avoid multicollinearity, choose to exclude it.

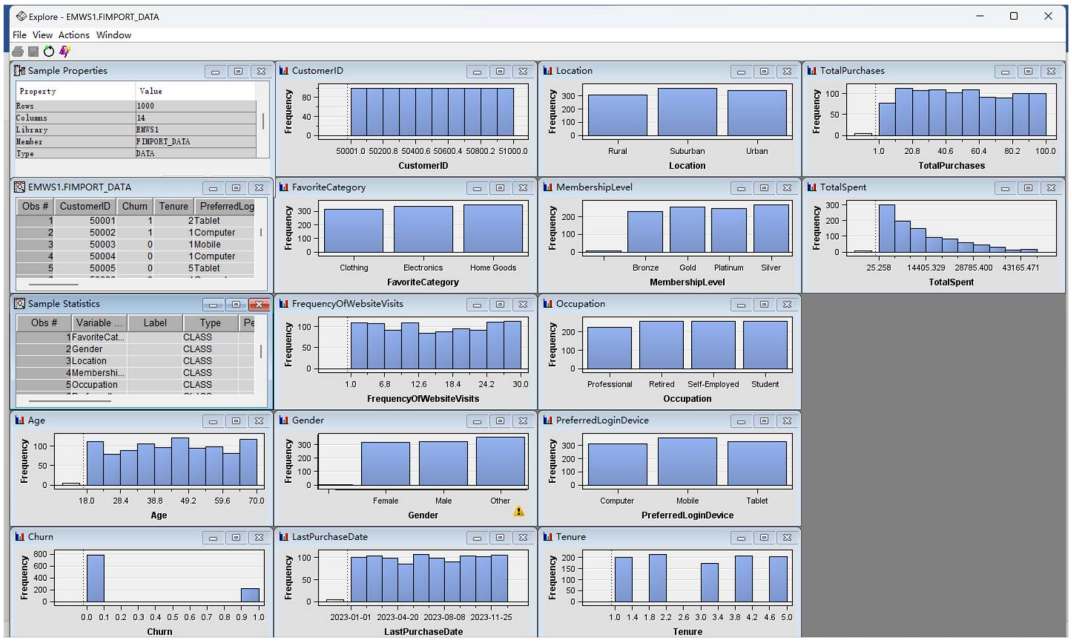
LastPurchaseDate (Time ID): This variable can be treated as a time identifier. If we're conducting a time-series analysis or the model needs to account for the time aspect of customer behavior, then it's necessary to indicate when the last purchase was made. It would not be an input variable if the model doesn't use time-series techniques.

The rest of the variables (Input): These variables are likely to contain information that can contribute to predicting customer churn. They are likely to be features that vary between customers and can be used to identify patterns associated with the churn.

4. Data Pre-processing with SAS Enterprise Miner & Talend Data Integration & Talend Data Preparation

4.1 Option one: Utilize SAS Enterprise Miner for data preprocessing

Identify Missing Values: Use the Explore node to identify the missing values in the dataset.
Decide on a Strategy: For handling missing values, decide between different approaches such as deletion, imputation, or using algorithms that can handle missing values.
Deletion: If the missing values are insignificant, we can choose to remove those rows.
Imputation: If the missing values are significant, consider imputation techniques.
For numeric variables, we can use mean/median imputation, and for categorical variables, mode imputation or more sophisticated methods like k-nearest neighbors (KNN) can be used.



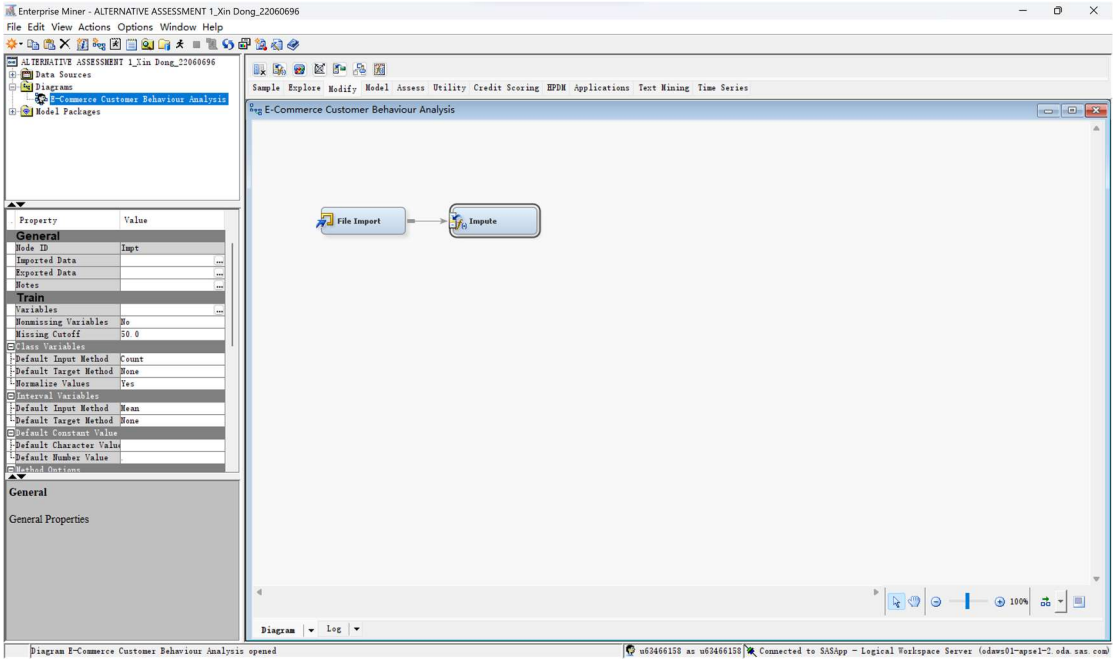
As we can observe from the table and plots and summarize the missing values and dealing methods:

Variable	Role	Level	Missing Values
Age	Input	Interval	Impute/Median
Churn	Target	Binary	None
CustomerID	ID	Nominal	None
FavoriteCategory	Input	Nominal	Impute/Distribution
FrequencyOfWebsiteVisits	Input	Interval	Impute/Mean
Gender	Input	Nominal	Impute/Distribution
LastPurchaseDate	Time ID	Interval	None
Location	Input	Nominal	Impute/Distribution

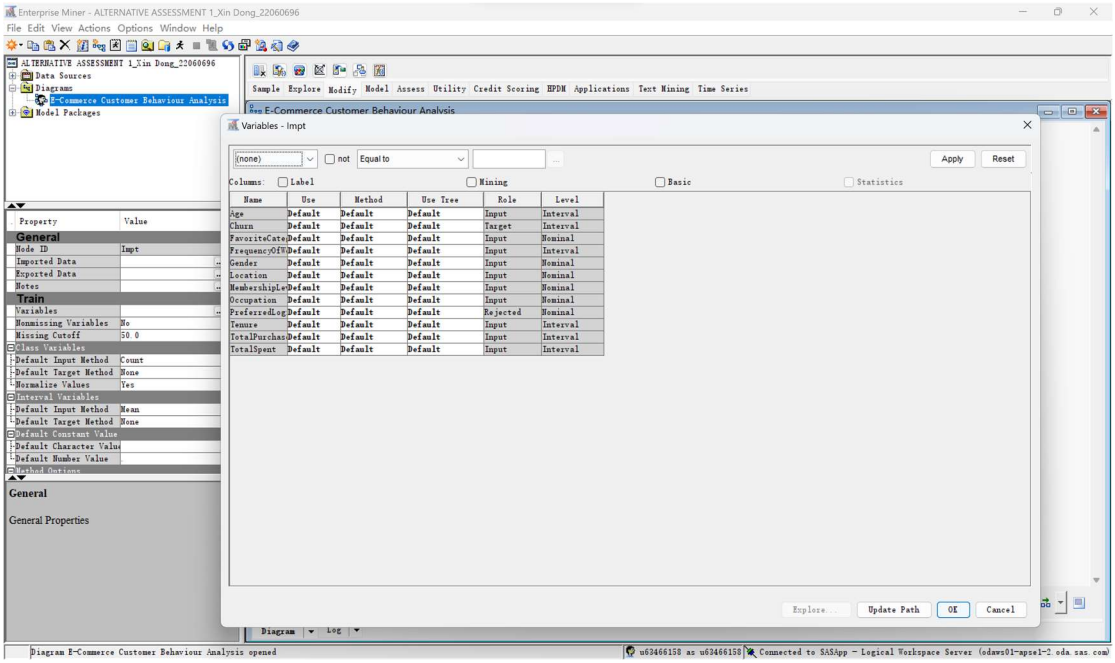
MembershipLevel	Input	Nominal	Impute/Distribution
Occupation	Input	Nominal	Impute/Distribution
PreferredLoginDevice	Rejected	Nominal	-
Tenure	Input	Interval	Impute/Median
TotalPurchases	Input	Interval	Impute/Median
TotalSpent	Input	Interval	Impute/Mean

Using the Impute Node:

Drag and drop the Impute node from the Modify tab onto the diagram workspace.
Connect our data source node to the Impute node.



Right-click the Impute node to configure the variables.



For each variable with missing values, you can select an imputation method based on the attribute table:

Mean/Median: For continuous variables like 'Age', 'TotalPurchases', and 'TotalSpent'.

Distribution: For categorical variables like 'FavoriteCategory', 'Gender', and 'MembershipLevel'.

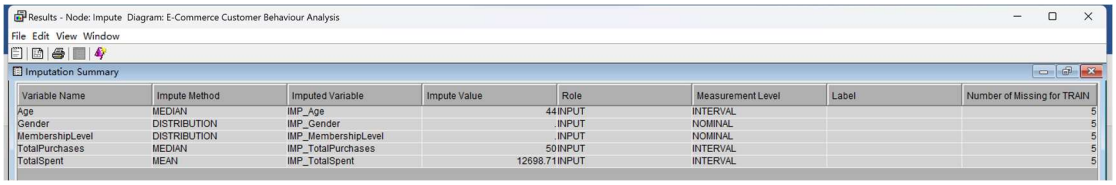
After configuring, run the Impute node to apply the changes.

(none) not Equal to ...

Columns: Label Mining B

Name	Use	Method	Use Tree	Role	Level
Age	Yes	Median	Default	Input	Interval
Churn	Default	None	Default	Target	Interval
FavoriteCategory	Yes	Distribution	Default	Input	Nominal
FrequencyOfWebsiteVisits	Yes	Mean	Default	Input	Interval
Gender	Yes	Distribution	Default	Input	Nominal
Location	Yes	Distribution	Default	Input	Nominal
MembershipLevel	Yes	Distribution	Default	Input	Nominal
Occupation	Yes	Distribution	Default	Input	Nominal
PreferredLoginDevice	No	None	Default	Rejected	Nominal
Tenure	Yes	Median	Default	Input	Interval
TotalPurchases	Yes	Median	Default	Input	Interval
TotalSpent	Yes	Mean	Default	Input	Interval

And then we can check the results of handling missing values:



Output

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEDIAN	IMP_Age	44.00	INPUT	INTERVAL		5
Gender	DISTRIBUTION	IMP_Gender	.	INPUT	NOMINAL		5
MembershipLevel	DISTRIBUTION	IMP_MembershipLevel	.	INPUT	NOMINAL		5
TotalPurchases	MEDIAN	IMP_TotalPurchases	50.00	INPUT	INTERVAL		5
TotalSpent	MEAN	IMP_TotalSpent	12698.71	INPUT	INTERVAL		5

Variable Distribution Training Data

Obs	Number of Missing for TRAIN	Number of Variables	Percent of Variables
1	5	5	100

Age:
Method: Median
Imputed Value: 44.00
Interpretation: Missing ages have been replaced with the median age value of 44, which is robust against skewed distributions and outliers.

Gender:
Method: Distribution
Interpretation: The missing values in the 'Gender' variable were filled in following the existing distribution of the 'Gender' categories in the dataset. This preserves the proportionality of each category.

MembershipLevel:
Method: Distribution
Interpretation: Similar to 'Gender', the missing values for 'MembershipLevel' were imputed based on the current distribution of membership levels among customers, maintaining the original category proportions.

TotalPurchases:
Method: Median
Imputed Value: 50.00
Interpretation: The missing values for 'TotalPurchases' were replaced with the median value of 50, which avoids the influence of extreme values and is a typical approach for count data that may be

right-skewed.

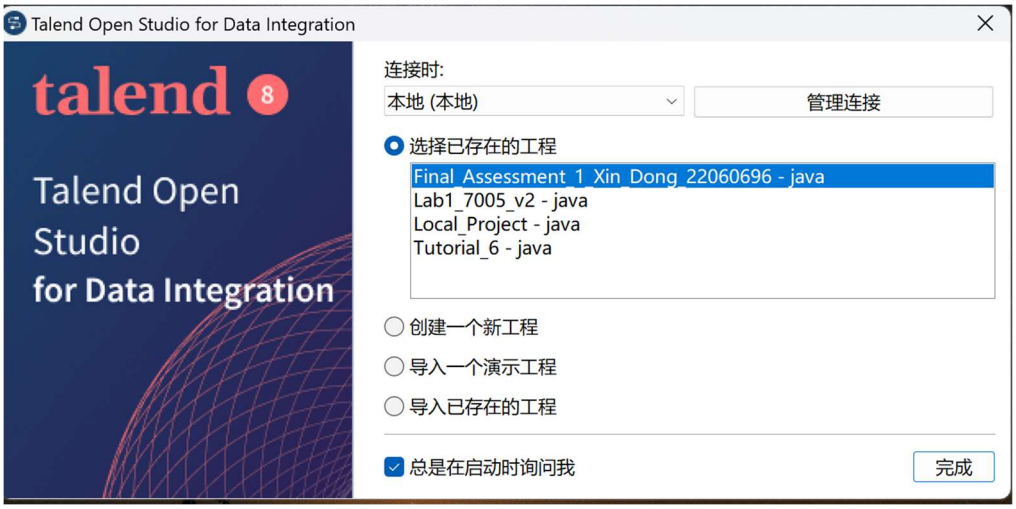
TotalSpent:
Method: Mean
Imputed Value: 12,698.71
Interpretation: The mean value of 12,698.71 was used to impute missing values in 'TotalSpent'. This suggests that on average, customers spent this amount, and it assumes that the spending is normally distributed.

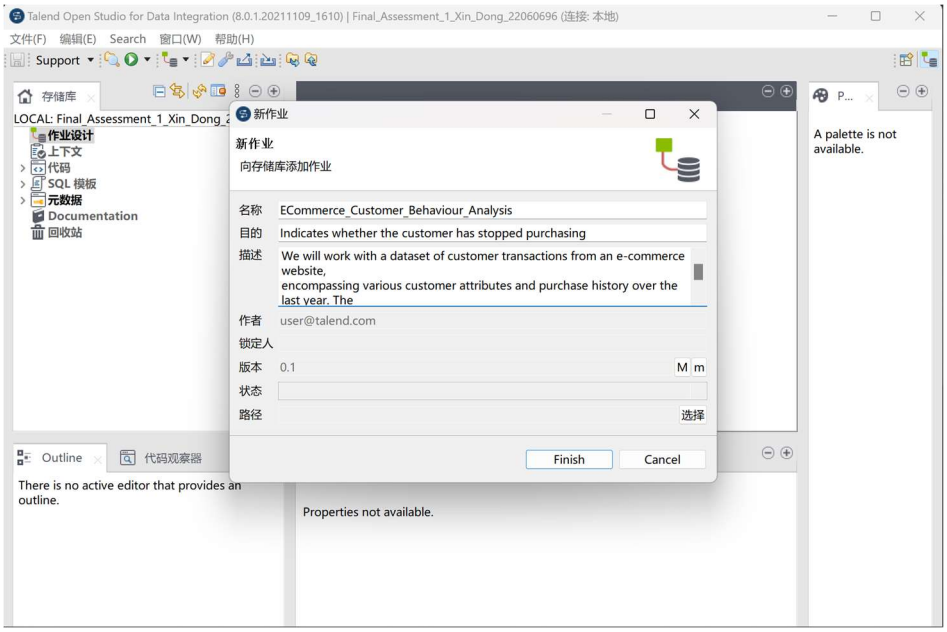
Each of these imputation methods was selected based on the nature of the variable and aimed to minimize the bias that missing data could introduce to the analysis. The number of missing values handled for each variable is consistent 5.

4.2 Another Option: Utilize Talend Data Integration & Talend Data Preparation to do Data Pre-Processing

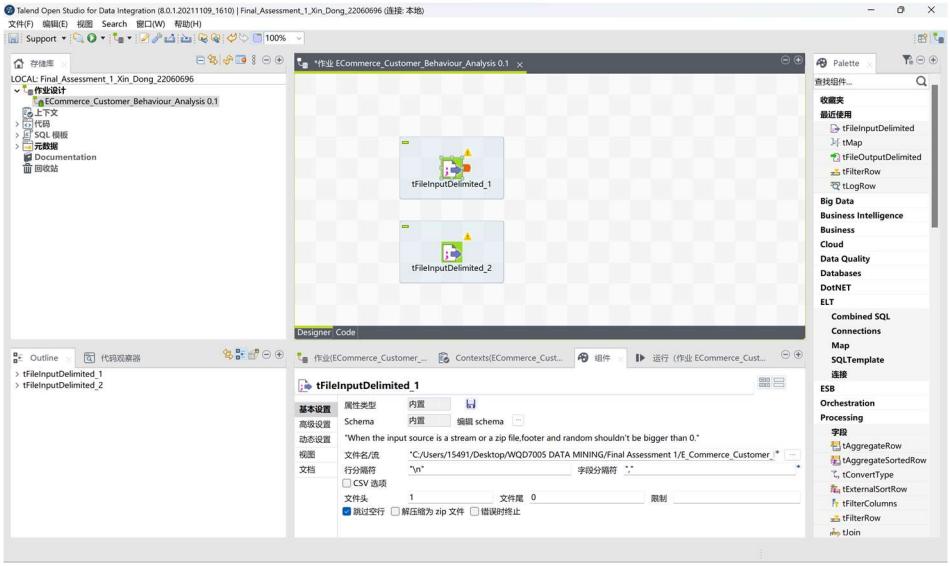
4.2.1 Talend Data Preparation:

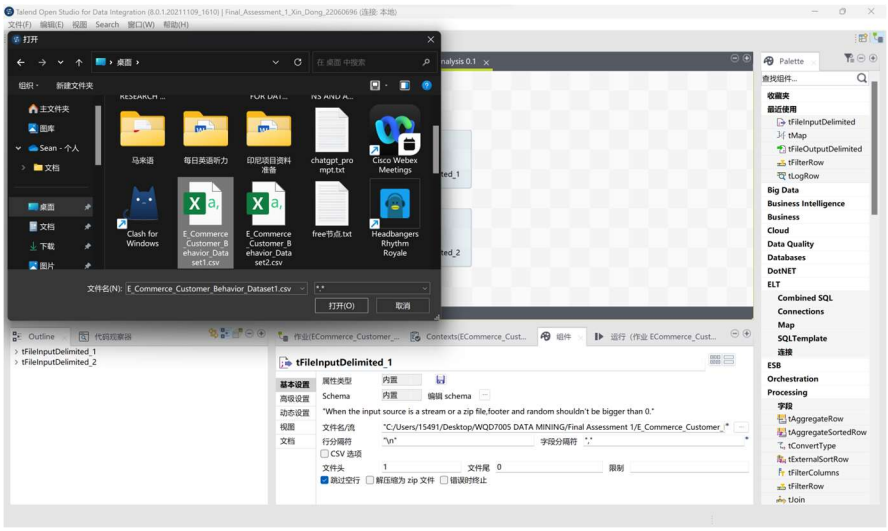
Create a project

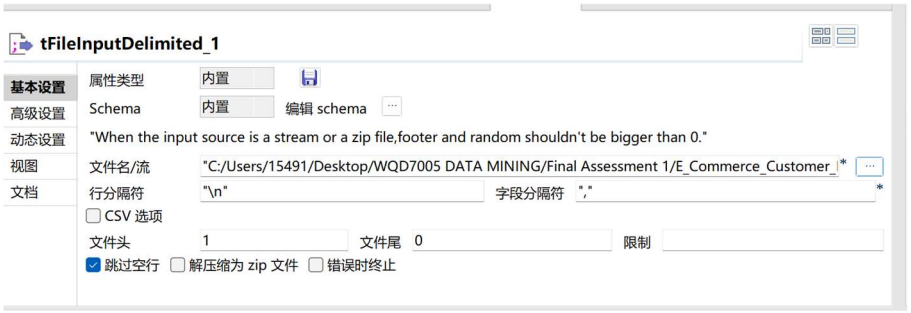


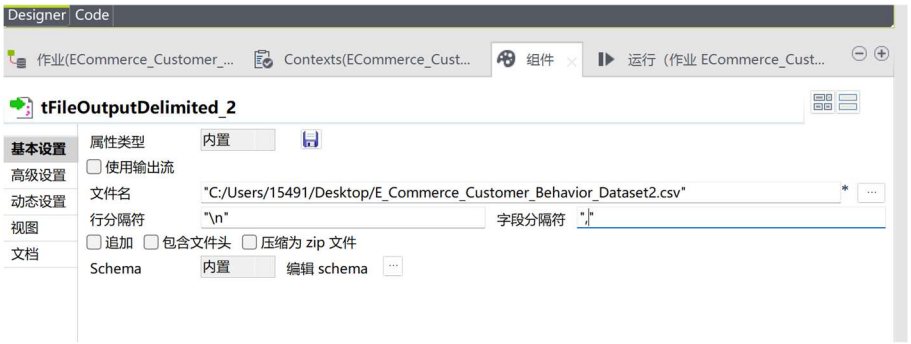


Import the 2 files to be integrated and set configuration:

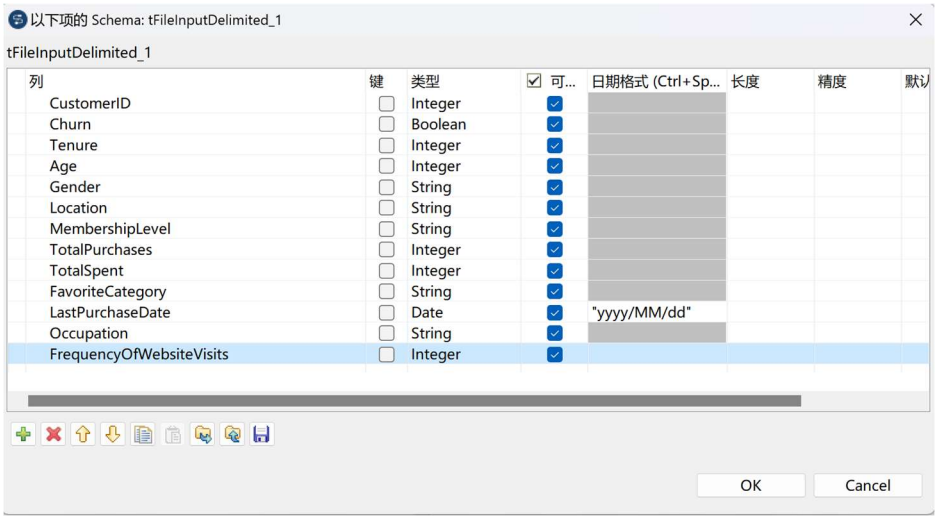


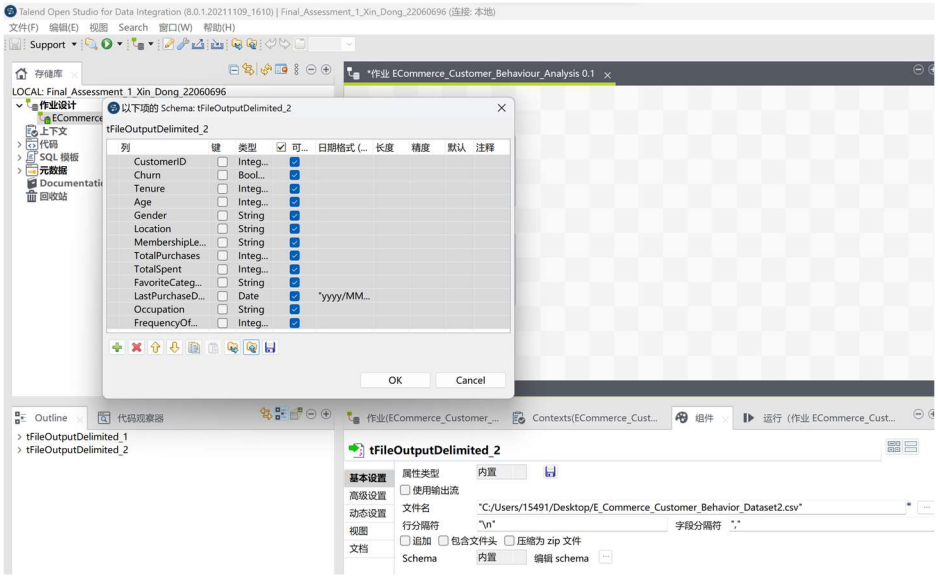




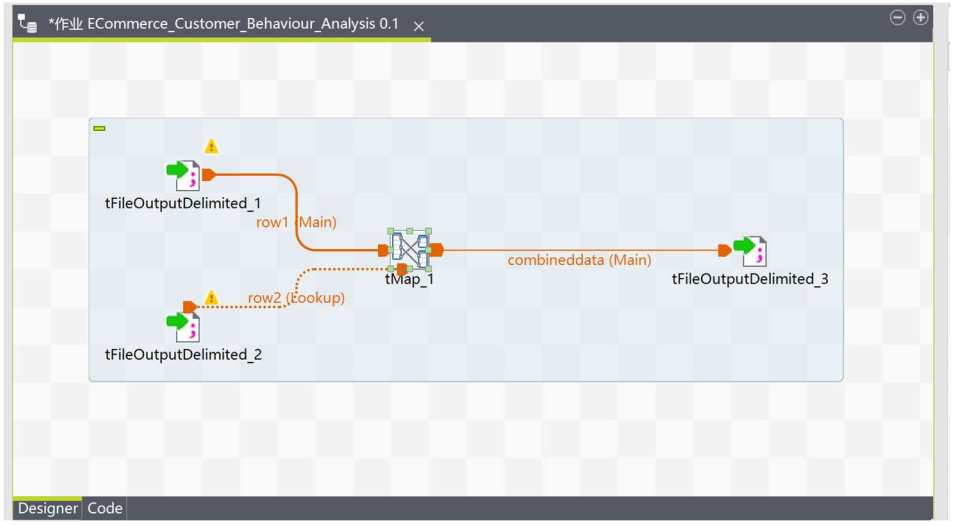


Set schema for the 2 files separately:

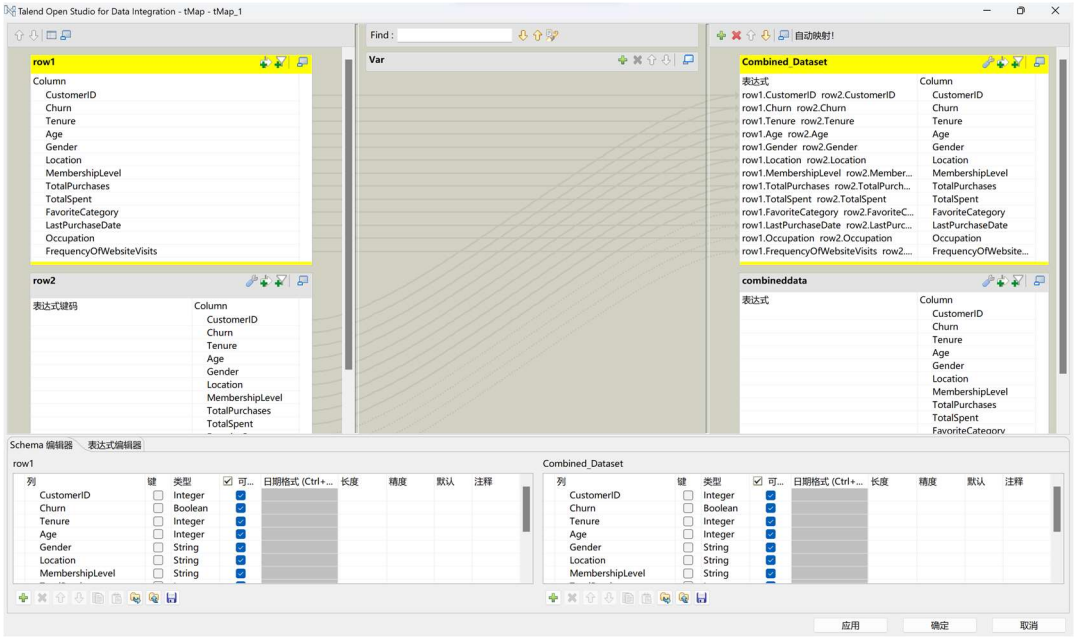




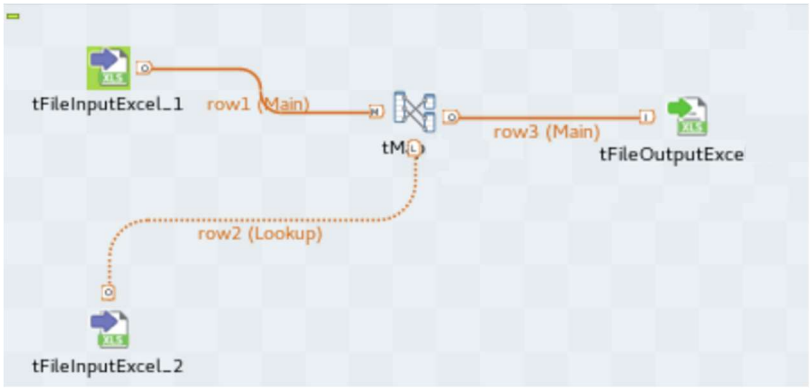
And then we put tMap and set tMap configuration, and finally set the tFileOutput.



tMap configuration:

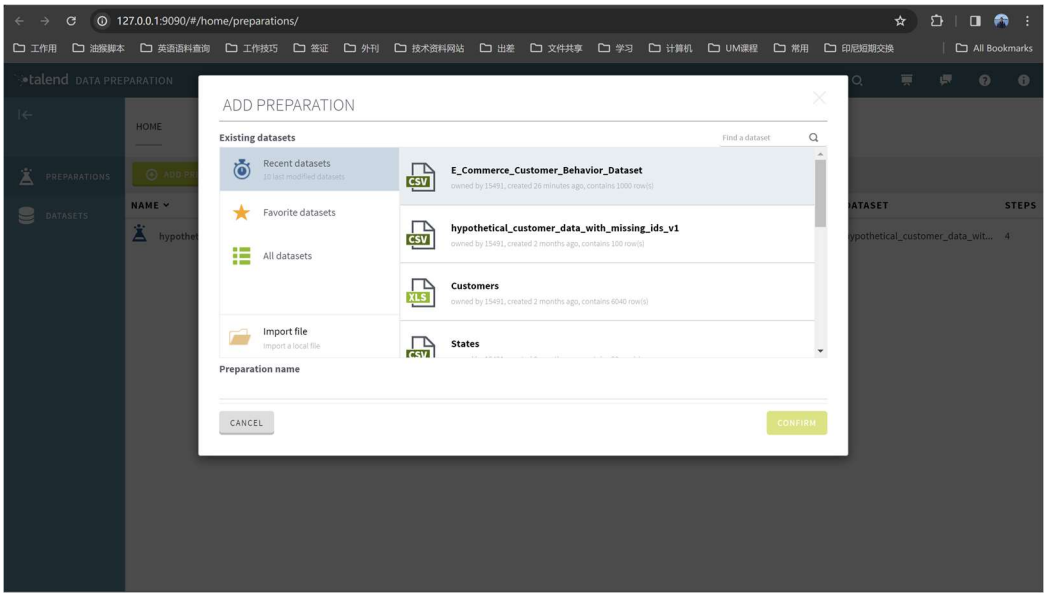


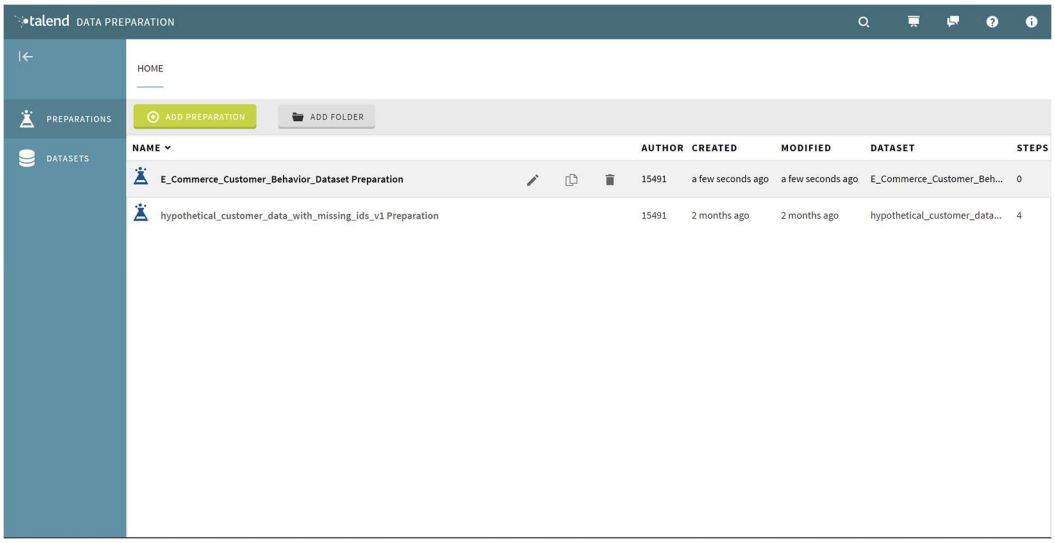
We changed the name and get the final output pic:



So we successfully combined 2 files of dataset into one.

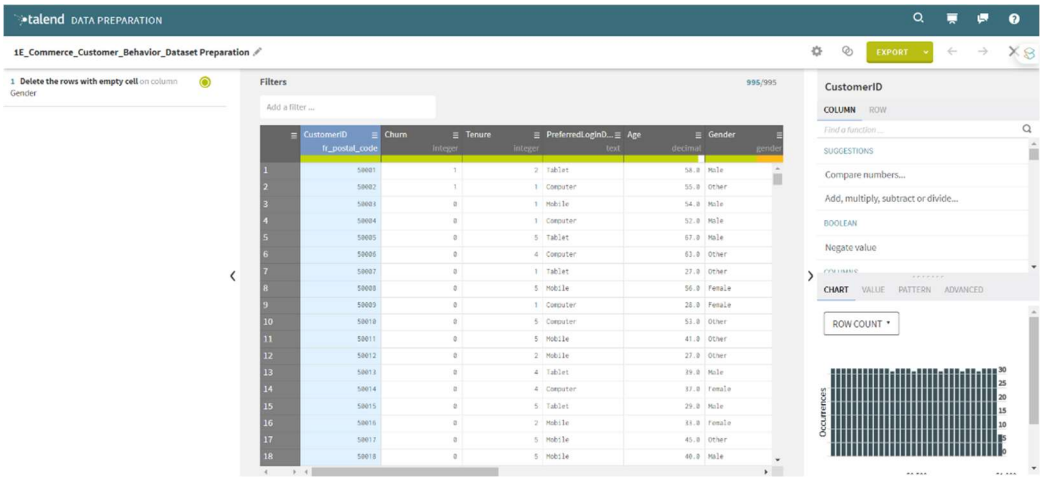
4.2.2 Talend Data Preparation:

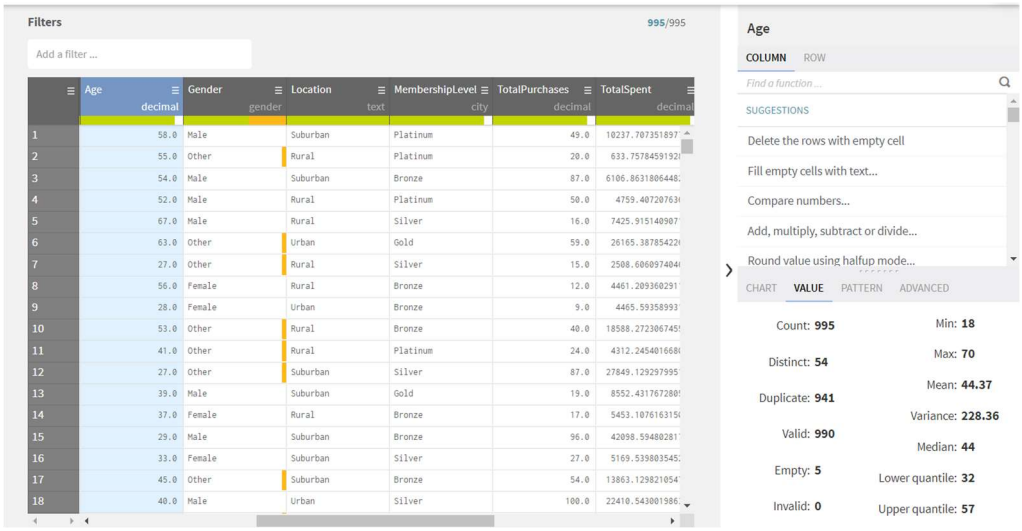


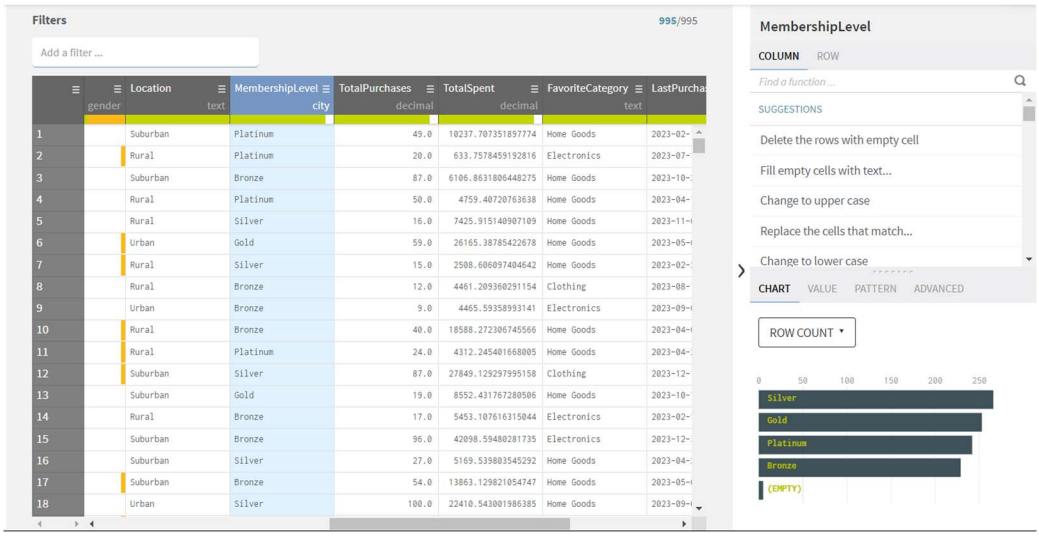


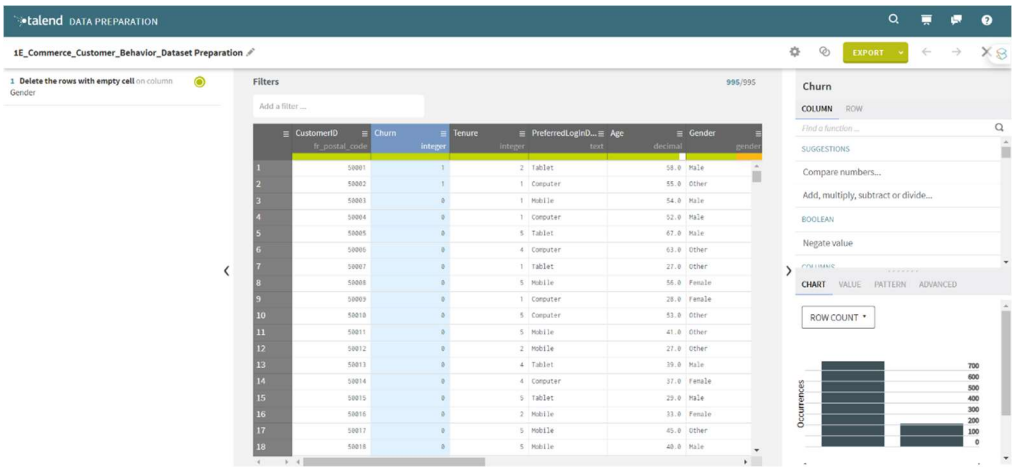
Find patterns and visualization for all the columns/variables:

We can observer the white gap for each variables represents the missing values.




Filters
995/995
Add a filter ...
Age
COLUMN
ROW
Find a function ...
SUGGESTIONS
Delete the rows with empty cell
Fill empty cells with text...
Compare numbers...
Add, multiply, subtract or divide...
Round value using halfup mode...
CHART
VALUE
PATTERN
ADVANCED
Count: 995
Distinct: 54
Duplicate: 941
Valid: 990
Empty: 5
Invalid: 0
Min: 18
Max: 70
Mean: 44.37
Variance: 228.36
Median: 44
Lower quantile: 32
Upper quantile: 57


Filters
995/995
Add a filter ...
MembershipLevel
COLUMN
ROW
Find a function ...
SUGGESTIONS
Delete the rows with empty cell
Fill empty cells with text...
Change to upper case
Replace the cells that match...
Change to lower case
CHART
VALUE
PATTERN
ADVANCED
ROW COUNT
Silver
Gold
Platinum
Bronze
(EMPTY)


talend DATA PREPARATION
1E_Commerce_Customer_Behavior_Dataset Preparation
EXPORT
1 Delete the rows with empty cell on column Gender
Filters
995/995
Add a filter ...
Churn
COLUMN
ROW
SUGGESTIONS
Compare numbers...
Add, multiply, subtract or divide...
BOOLEAN
Negate value
CHART
VALUE
PATTERN
ADVANCED
ROW COUNT
Occurrences

Filters

995/995

Add a filter ...

	gender	Location text	MembershipLevel city	TotalPurchases decimal	TotalSpent decimal	FavoriteCategory text	LastPurcha
1		Suburban	Platinum	49.0	10237.707351897774	Home Goods	2023-02-
2		Rural	Platinum	20.0	633.7578459192816	Electronics	2023-07-
3		Suburban	Bronze	87.0	6106.8631806448275	Home Goods	2023-10-
4		Rural	Platinum	50.0	4759.40720763638	Home Goods	2023-04-
5		Rural	Silver	16.0	7425.915140907109	Home Goods	2023-11-
6		Urban	Gold	59.0	26165.38785422678	Home Goods	2023-05-
7		Rural	Silver	15.0	2508.606097404642	Home Goods	2023-02-
8		Rural	Bronze	12.0	4461.209360291154	Clothing	2023-08-
9		Urban	Bronze	9.0	4465.59358993141	Electronics	2023-09-
10		Rural	Bronze	40.0	18588.272306745566	Home Goods	2023-04-
11		Rural	Platinum	24.0	4312.245401668005	Home Goods	2023-04-
12		Suburban	Silver	87.0	27849.129297995158	Clothing	2023-12-
13		Suburban	Gold	19.0	8552.431767280506	Home Goods	2023-10-
14		Rural	Bronze	17.0	5453.107616315044	Electronics	2023-02-
15		Suburban	Bronze	96.0	42098.59480281735	Electronics	2023-12-
16		Suburban	Silver	27.0	5169.539803545292	Home Goods	2023-04-
17		Suburban	Bronze	54.0	13863.129821054747	Home Goods	2023-05-
18		Urban	Silver	100.0	22410.543001986385	Home Goods	2023-09-

Location

COLUMN ROW

Find a function ...

SUGGESTIONS

Change to upper case

Replace the cells that match...

Change to lower case

BOOLEAN

Negate value

CHART VALUE PATTERN ADVANCED

ROW COUNT

Filters

995/995

Add a filter ...

	alPurchases decimal	TotalSpent decimal	FavoriteCategory text	LastPurchaseDate date	Occupation text	FrequencyOfWe... integer
1	49.0	10237.707351897774	Home Goods	2023-02-12	Student	7
2	20.0	633.7578459192816	Electronics	2023-07-11	Professional	14
3	87.0	6106.8631806448275	Home Goods	2023-10-25	Self-Employed	8
4	50.0	4759.40720763638	Home Goods	2023-04-19	Student	26
5	16.0	7425.915140907109	Home Goods	2023-11-09	Professional	25
6	59.0	26165.38785422678	Home Goods	2023-05-08	Student	12
7	15.0	2508.606097404642	Home Goods	2023-02-27	Self-Employed	3
8	12.0	4461.209360291154	Clothing	2023-08-13	Self-Employed	12
9	9.0	4465.59358993141	Electronics	2023-09-08	Self-Employed	3
10	40.0	18588.272306745566	Home Goods	2023-04-06	Retired	25
11	24.0	4312.245401668005	Home Goods	2023-04-23	Retired	27
12	87.0	27849.129297995158	Clothing	2023-12-17	Retired	30
13	19.0	8552.431767280506	Home Goods	2023-10-12	Professional	26
14	17.0	5453.107616315044	Electronics	2023-02-13	Self-Employed	20
15	96.0	42098.59480281735	Electronics	2023-12-30	Student	13
16	27.0	5169.539803545292	Home Goods	2023-04-25	Self-Employed	18
17	54.0	13863.129821054747	Home Goods	2023-05-05	Retired	17
18	100.0	22410.543001986385	Home Goods	2023-09-01	Retired	14

Occupation

COLUMN ROW

Find a function ...

SUGGESTIONS

Change to upper case

Replace the cells that match...

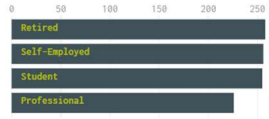
Change to title case

Change to lower case

BOOLEAN

CHART VALUE PATTERN ADVANCED

ROW COUNT



Filters

995/995

Add a filter ...

	alPurchases decimal	TotalSpent decimal	FavoriteCategory text	LastPurchaseDate date	Occupation text	FrequencyOfWe... integer
1	49.0	10237.707351897774	Home Goods	2023-02-12	Student	7
2	20.0	633.7578459192816	Electronics	2023-07-11	Professional	14
3	87.0	6106.8631806448275	Home Goods	2023-10-25	Self-Employed	8
4	50.0	4759.40720763638	Home Goods	2023-04-19	Student	26
5	16.0	7425.915140907109	Home Goods	2023-11-09	Professional	25
6	59.0	26165.38785422678	Home Goods	2023-05-08	Student	12
7	15.0	2508.606097404642	Home Goods	2023-02-27	Self-Employed	3
8	12.0	4461.209360291154	Clothing	2023-08-13	Self-Employed	12
9	9.0	4465.59358993141	Electronics	2023-09-08	Self-Employed	3
10	40.0	18588.272306745566	Home Goods	2023-04-06	Retired	25
11	24.0	4312.245401668005	Home Goods	2023-04-23	Retired	27
12	87.0	27849.129297995158	Clothing	2023-12-17	Retired	30
13	19.0	8552.431767280506	Home Goods	2023-10-12	Professional	26
14	17.0	5453.107616315044	Electronics	2023-02-13	Self-Employed	20
15	96.0	42098.59480281735	Electronics	2023-12-30	Student	13
16	27.0	5169.539803545292	Home Goods	2023-04-25	Self-Employed	18
17	54.0	13863.129821054747	Home Goods	2023-05-05	Retired	17
18	100.0	22410.543001986385	Home Goods	2023-09-01	Retired	14

FrequencyOfWebsiteVisits

COLUMN ROW

Find a function ...

SUGGESTIONS

Compare numbers...

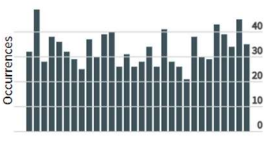
Add, multiply, subtract or divide...

BOOLEAN

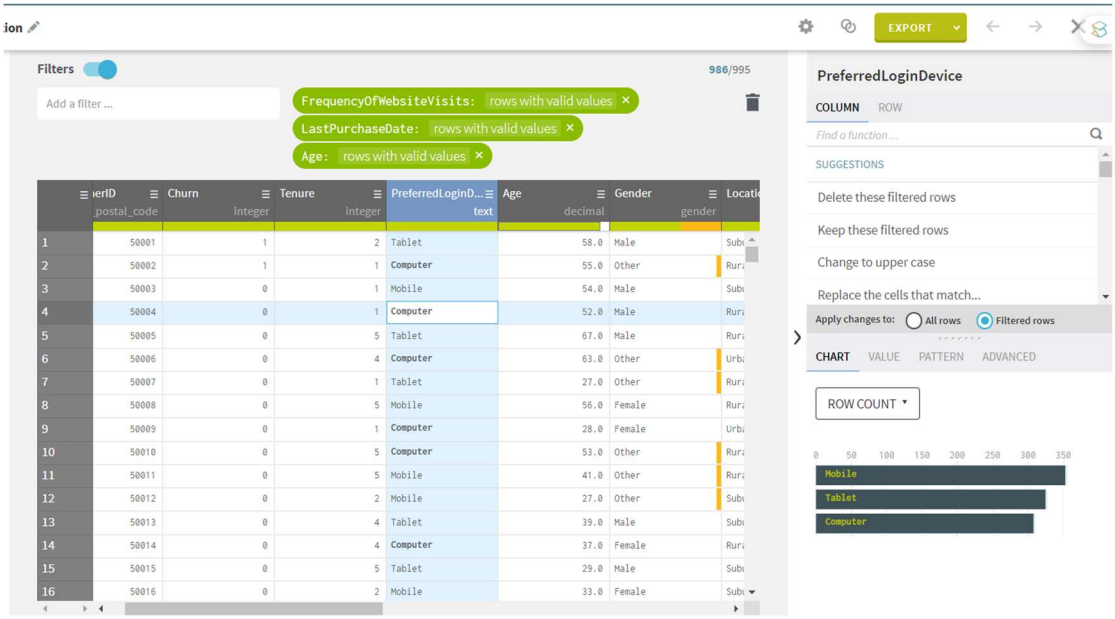
Negate value

CHART VALUE PATTERN ADVANCED

ROW COUNT

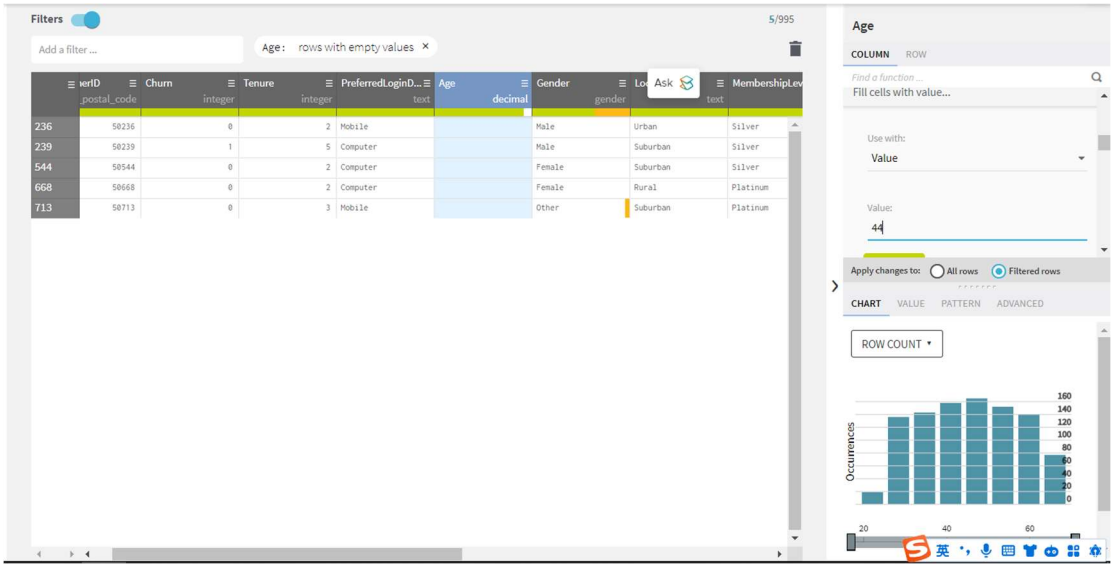


We can select the rows with valid values:

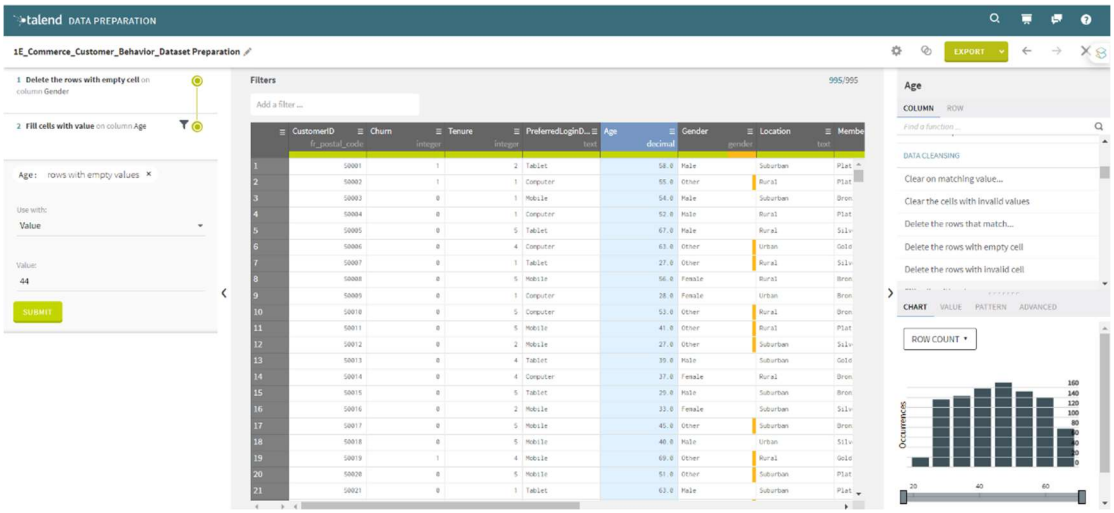


Sample of filling the missing values, including age (numeric), MembershipLevel (string), and other variables (delete unimportant rows):

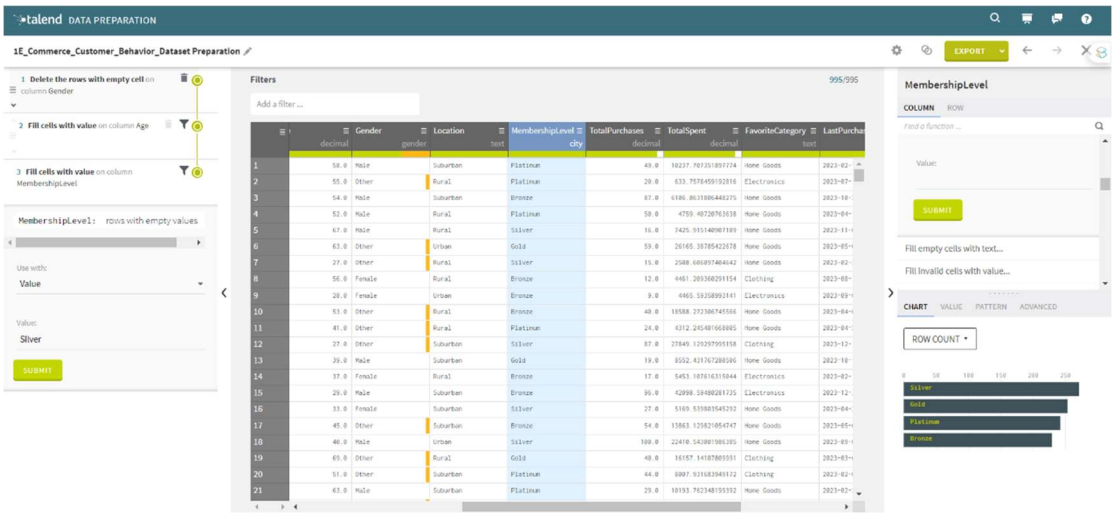
First select the rows with empty cells.



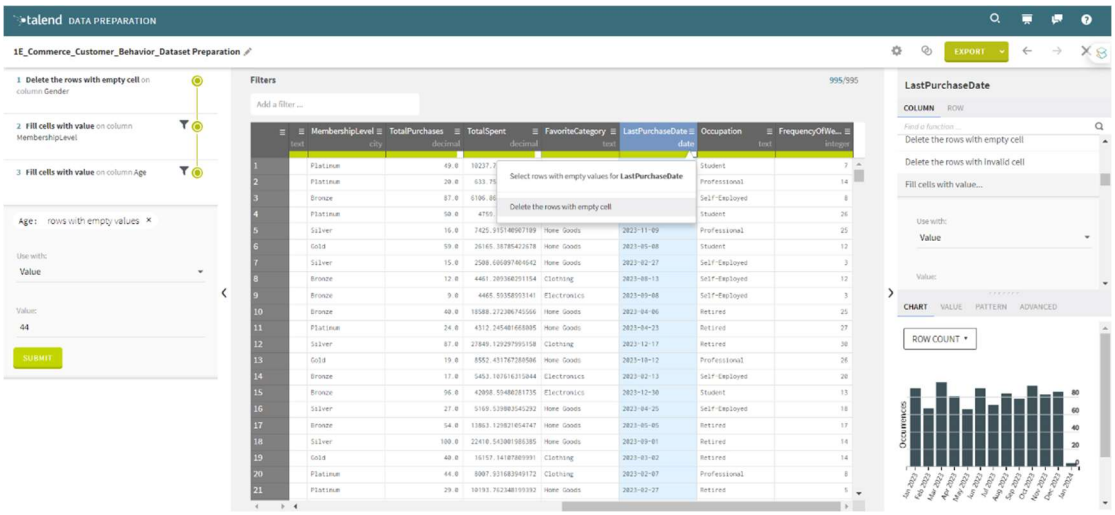
And then fill the missing values with median/mean number.



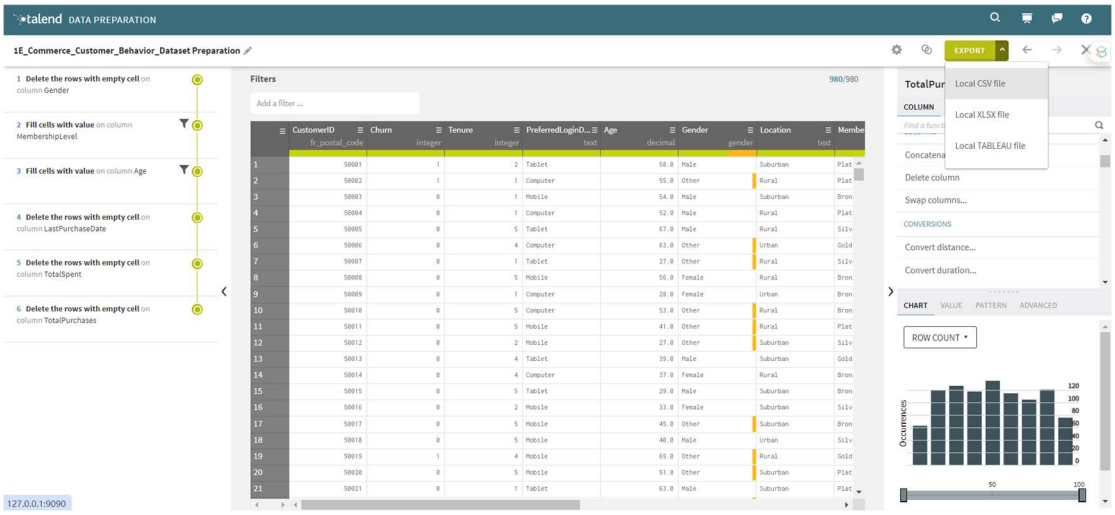
For the variables with missing value type of string/text, use the same method, impute the most frequent value based on the distribution.



Other variables, delete unimportant rows with empty values.

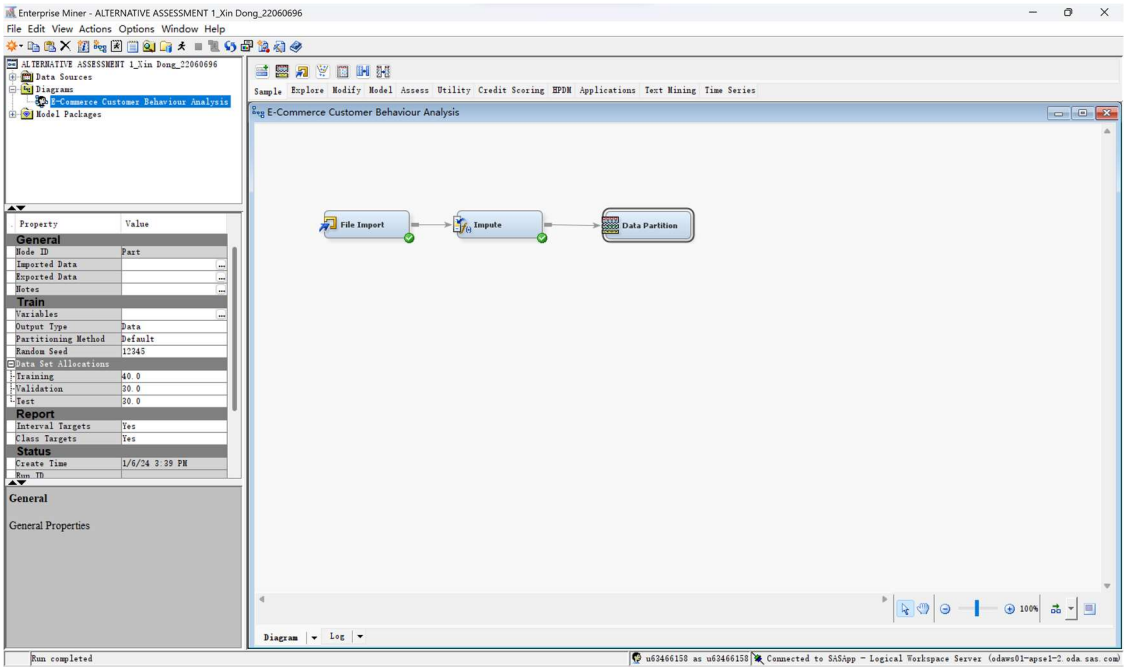


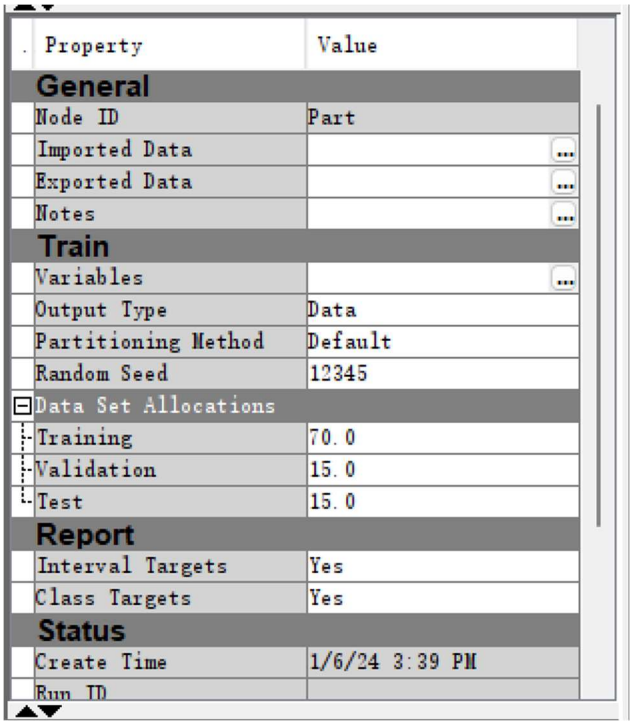
And then we finish the data pre-processing part.



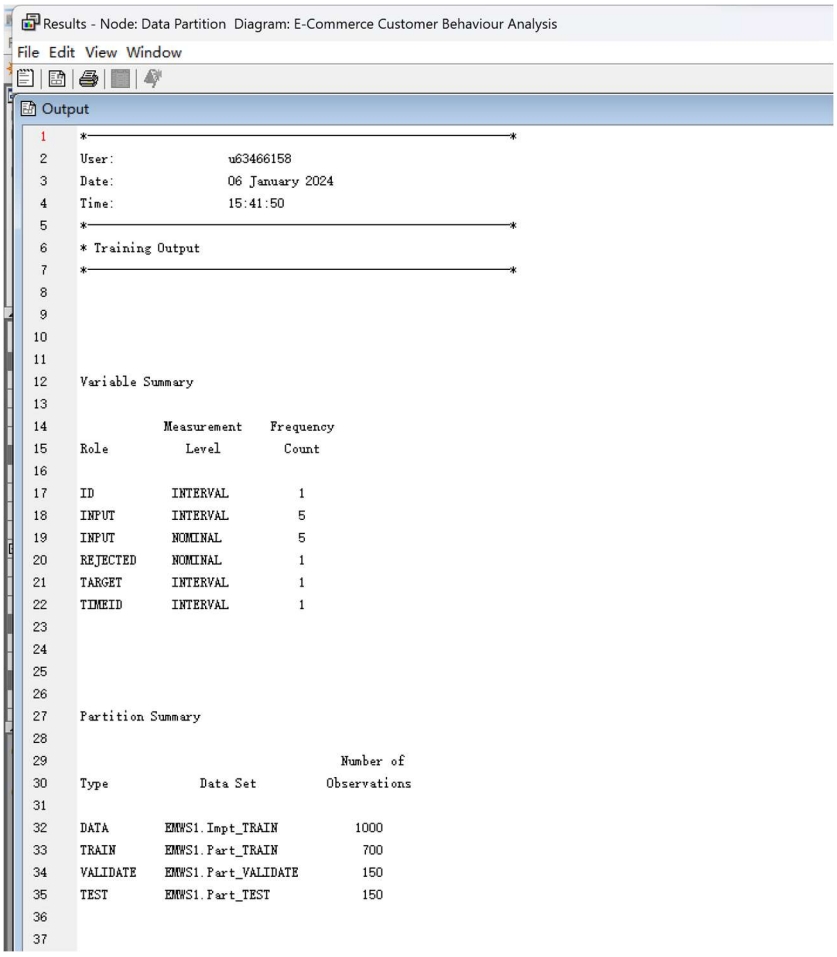
5. Data Partition

Partition the Data: Use the Data Partition node to split your data into training, validation, and testing sets. A common split is 70% training, 15% validation, and 15% testing.





As below shows the results of data partition:



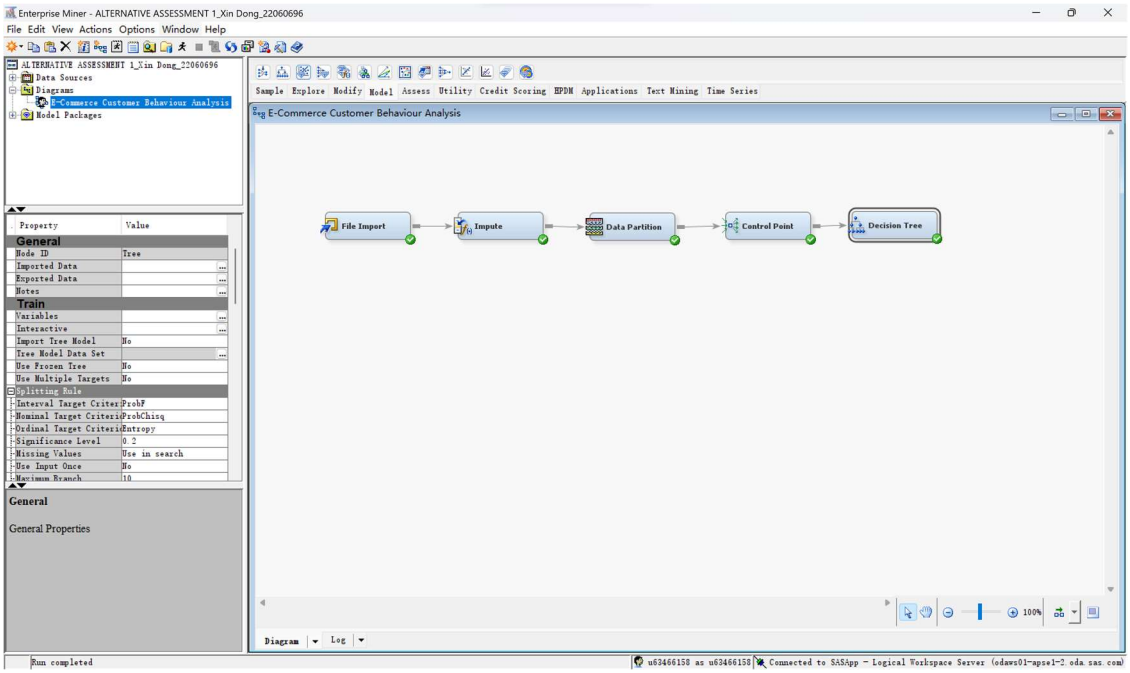

```
*------------------------------------------------------------*
User:                  u63466158
Date:                  06 January 2024
Time:                  15:41:50
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

             Measurement    Frequency
Role            Level         Count

ID           INTERVAL          1
INPUT        INTERVAL          5
INPUT        NOMINAL           5
REJECTED     NOMINAL           1
TARGET       INTERVAL          1
TIMEID       INTERVAL          1



Partition Summary

                                    Number of
Type             Data Set         Observations

DATA        EMWS1.Impt_TRAIN          1000
TRAIN       EMWS1.Part_TRAIN           700
VALIDATE    EMWS1.Part_VALIDATE        150
TEST        EMWS1.Part_TEST            150
```


```



Summary Statistics for Interval Targets

Data=DATA

                                                    Number of                    Standard
Variable        Maximum            Mean         Minimum    Observations    Missing      Deviation    Label

 Churn                1           0.214               0            1000          0       0.410332027

Data=TEST

                                                    Number of                    Standard
Variable        Maximum            Mean         Minimum    Observations    Missing      Deviation    Label

 Churn                1             0.2               0             150          0       0.4013400373

Data=TRAIN

                                                    Number of                    Standard
Variable        Maximum            Mean         Minimum    Observations    Missing      Deviation    Label

 Churn                1             0.21              0             700          0       0.4075994836

Data=VALIDATE

                                                    Number of                    Standard
Variable        Maximum            Mean         Minimum    Observations    Missing      Deviation    Label

 Churn                1    0.2466666667               0             150          0       0.4325151457
```

6. Decision Tree Analysis

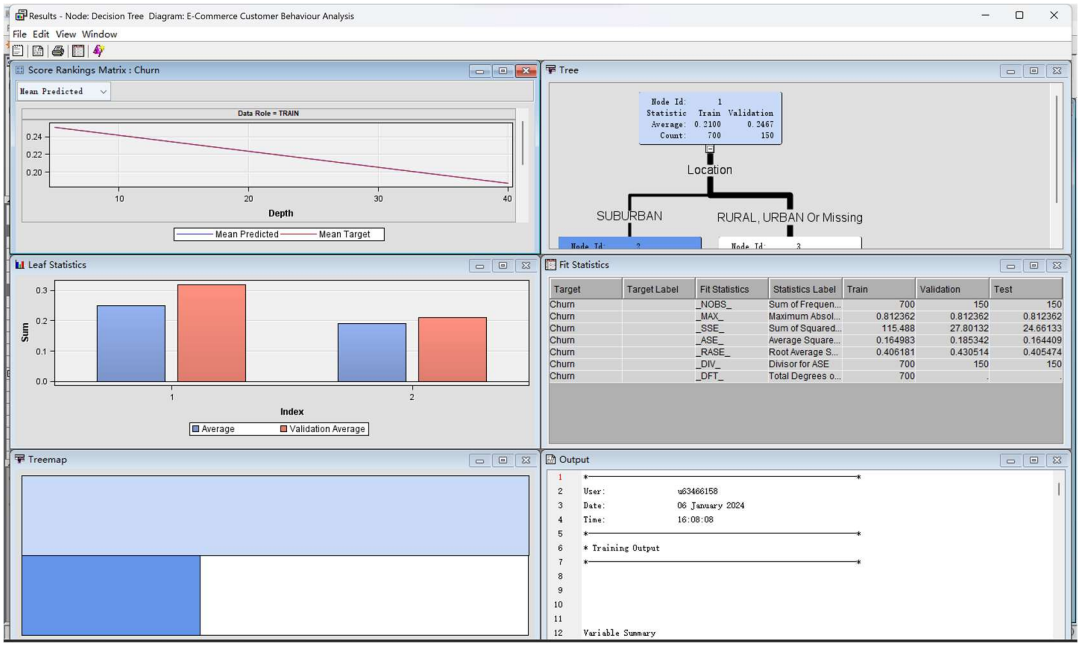
Step 1: Setting Up the Decision Tree Node
Open our project within SAS Enterprise Miner.
Drag and drop the Decision Tree node from the toolbar onto the workspace.
Connect the Data Partition node (which has the preprocessed data) to the Decision Tree node.

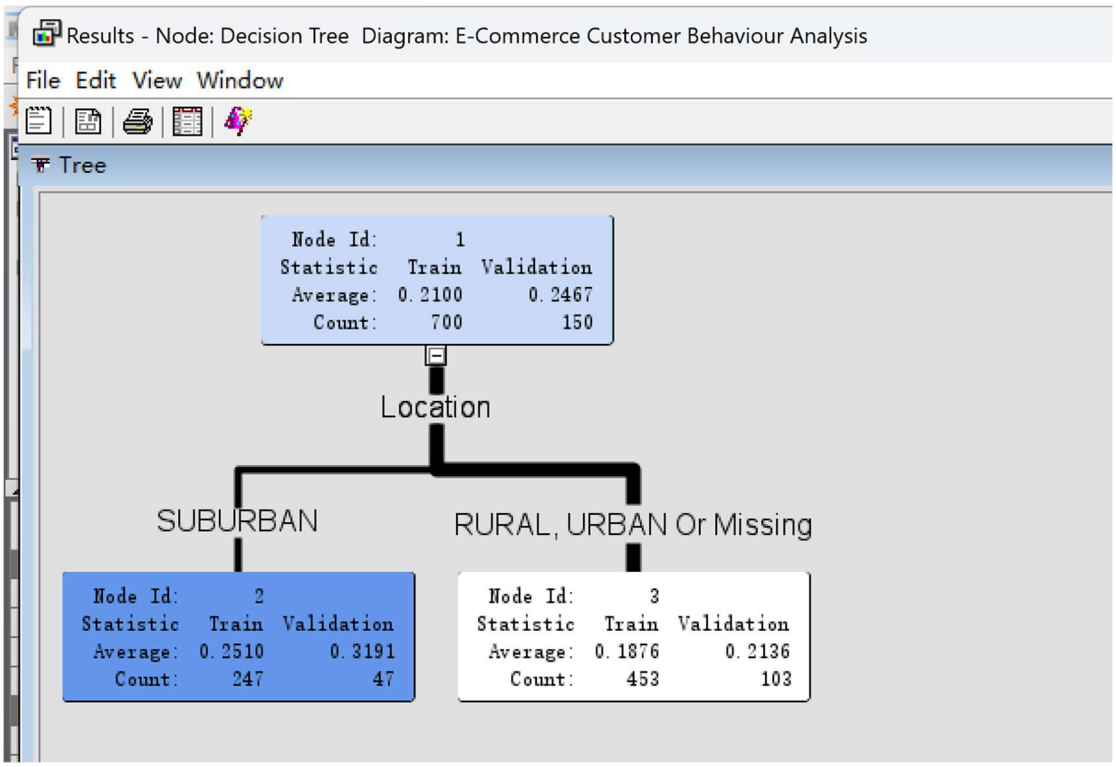
Use a Control point to prepare for more models to implement and simplify the process flow diagram.



Step 2: Configuring the Decision Tree Node
Click on the Decision Tree node to check the left side configuration bar.

Here below is the output results based on the default configure settings:


Results - Node: Decision Tree Diagram: E-Commerce Customer Behaviour Analysis
File Edit View Window
Score Rankings Matrix : Churn
Mean Predicted
Data Role = TRAIN
Depth
Mean Predicted
Mean Target
Tree
Node Id: 1
Statistic Train Validation
Average: 0.2100 0.2467
Count: 700 150
Location
SUBURBAN
RURAL, URBAN Or Missing
Leaf Statistics
Sum
Index
Average
Validation Average
Fit Statistics
Target Target Label Fit Statistics Statistics Label Train Validation Test
Churn _NOBS_ Sum of Frequen... 700 150 150
Churn _MAX_ Maximum Absol... 0.812362 0.812362 0.812362
Churn _SSE_ Sum of Squared... 115.488 27.80132 24.66133
Churn _ASE_ Average Square... 0.164983 0.185342 0.164409
Churn _RASE_ Root Average S... 0.406181 0.430514 0.405474
Churn _DIV_ Divisor for ASE 700 150 150
Churn _DFT_ Total Degrees o... 700
Treemap
Output
User: u63466158
Date: 06 January 2024
Time: 16:08:08
* Training Output
Variable Summary


Results - Node: Decision Tree Diagram: E-Commerce Customer Behaviour Analysis
File Edit View Window
Tree
Node Id: 1
Statistic Train Validation
Average: 0.2100 0.2467
Count: 700 150
Location
SUBURBAN
RURAL, URBAN Or Missing
Node Id: 2
Statistic Train Validation
Average: 0.2510 0.3191
Count: 247 47
Node Id: 3
Statistic Train Validation
Average: 0.1876 0.2136
Count: 453 103

Results - Node: Decision Tree Diagram: E-Commerce Customer Behaviour Analysis

File Edit View Window

Output

```
+-------------------------------------------------------------------------------+
User:                   u63466158
Date:                   06 January 2024
Time:                   16:08:08
+-------------------------------------------------------------------------------+
* Training Output
+-------------------------------------------------------------------------------+



Variable Summary

             Measurement    Frequency
Role           Level         Count

ID           INTERVAL          2
INPUT        INTERVAL          5
INPUT        NOMINAL           5
REJECTED     NOMINAL           1
TARGET       INTERVAL          1
TIMEID       INTERVAL          1



Predicted and decision variables

Type         Variable        Label

TARGET       Churn
PREDICTED    P_Churn     Predicted: Churn
RESIDUAL     R_Churn     Residual: Churn


+-------------------------------------------------------------------------------+
* Score Output
+-------------------------------------------------------------------------------+
```

Results - Node: Decision Tree Diagram: E-Commerce Customer Behaviour Analysis

File Edit View Window

Output

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Location		1	1.0000	1.0000	1.0000

Tree Leaf Report

Node Id	Depth	Training Observations	Training Average	Validation Observations	Validation Average	Training Root ASE	Validation Root ASE
3	1	453	0.19	103	0.21	0.39042	0.41066
2	1	247	0.25	47	0.32	0.43360	0.47110

Fit Statistics

Target=Churn Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	700.000	150.000	150.000
MAX	Maximum Absolute Error	0.812	0.812	0.812
SSE	Sum of Squared Errors	115.488	27.801	24.661
ASE	Average Squared Error	0.165	0.185	0.164
RASE	Root Average Squared Error	0.406	0.431	0.405
DIV	Divisor for ASE	700.000	150.000	150.000
DFT	Total Degrees of Freedom	700.000	.	.

Results - Node: Decision Tree Diagram: E-Commerce Customer Behaviour Analysis

File Edit View Window

Output

Assessment Score Rankings

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

Depth	Number of Observations	Mean Target	Mean Predicted
5	247	0.25101	0.25101
40	453	0.18764	0.18764

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

Depth	Number of Observations	Mean Target	Mean Predicted
5	47	0.31915	0.25101
35	103	0.21359	0.18764

Assessment Score Distribution

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

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.248 - 0.251	0.25101	0.25101	247	0.24943
0.188 - 0.191	0.18764	0.18764	453	0.18922

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

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.248 - 0.251	0.31915	0.25101	47	0.24943
0.188 - 0.191	0.21359	0.18764	103	0.18922

Summary of Results with Default Configuration:

Significance Level:
Set a lower significance level for splits to reduce complexity and overfitting, such as 0.05 or 0.01.

Leaf Size:
Increase the minimum leaf size to 10 or 20 to prevent the model from being too sensitive to noise in the training data.

Number of Rules:
Adjust the number of rules based on the complexity of the dataset. If overfitting is observed, reduce the number of rules to simplify the tree.

Maximum Depth:
Set a maximum depth based on the point where validation performance plateaus or decreases to prevent overfitting. A depth of 5 or 6 might be a good starting point.

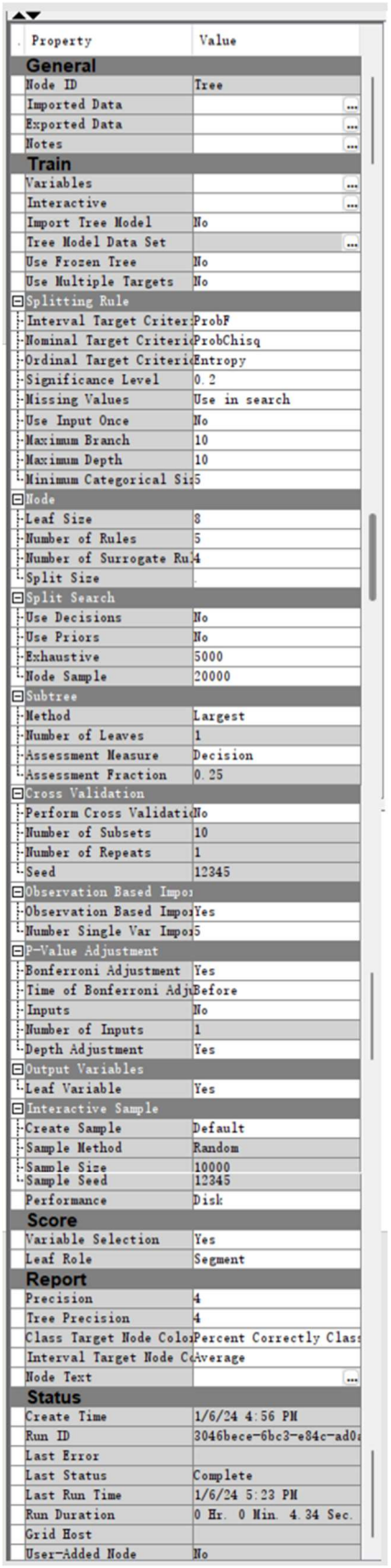
Pruning:
Implement post-pruning techniques using a validation set to remove branches that do not contribute significantly to the model's predictive power.

Variable Importance:
Review the variable importance and ensure that the tree uses a diverse set of variables for decision-making, rather than relying predominantly on one.

Max depth 6 to 10, leaf size 5 to 8, number of surrogate rules 0 to 4, make sure the assessment measure is decision, subtree method from Assessment to Largest, Observation Based Importance No to Yes, maximum branch 2 to 10.

Improvements for Better Configuration:

As updated the configuration in the screenshot below:



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 Criter	ProbF
Nominal Target Criteri	ProbChisq
Ordinal Target Criteri	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	10
Maximum Depth	10
Minimum Categorical Si	5
Node	
Leaf Size	8
Number of Rules	5
Number of Surrogate Ru	4
Split Size	
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validati	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Impo	
Observation Based Impo	Yes
Number Single Var Impo	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adj	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	
Precision	4
Tree Precision	4
Class Target Node Colo	Percent Correctly Class
Interval Target Node C	Average
Node Text	
Status	
Create Time	1/6/24 4:56 PM
Run ID	3046bece-6bc3-e84c-ad0
Last Error	
Last Status	Complete
Last Run Time	1/6/24 5:23 PM
Run Duration	0 Hr. 0 Min. 4.34 Sec.
Grid Host	
User-Added Node	No

Max Depth (6 to 10):

Increasing the maximum depth allows the tree to create more levels of decision nodes. This will

enable the model to capture more complex patterns in the data. It's important to monitor performance on validation data to find the optimal depth. I have tried that start at the lower end of this range and incrementally increase, validating performance at each step.

Maximum Branch (2 to 10):
Adjusting the maximum number of branches from a single decision node from 2 to 10 allows the tree to consider more than binary splits. This is useful for multi-class categorical variables and can provide a more nuanced understanding of the data.

Leaf Size (5 to 8):
Increasing the minimum leaf size makes the tree less sensitive to noise in the data, reducing overfitting. A leaf size of 8 means that any decision node must represent at least 8 instances in the data. This helps ensure that patterns found by the tree are statistically reliable and not due to chance variations in smaller sample sizes.

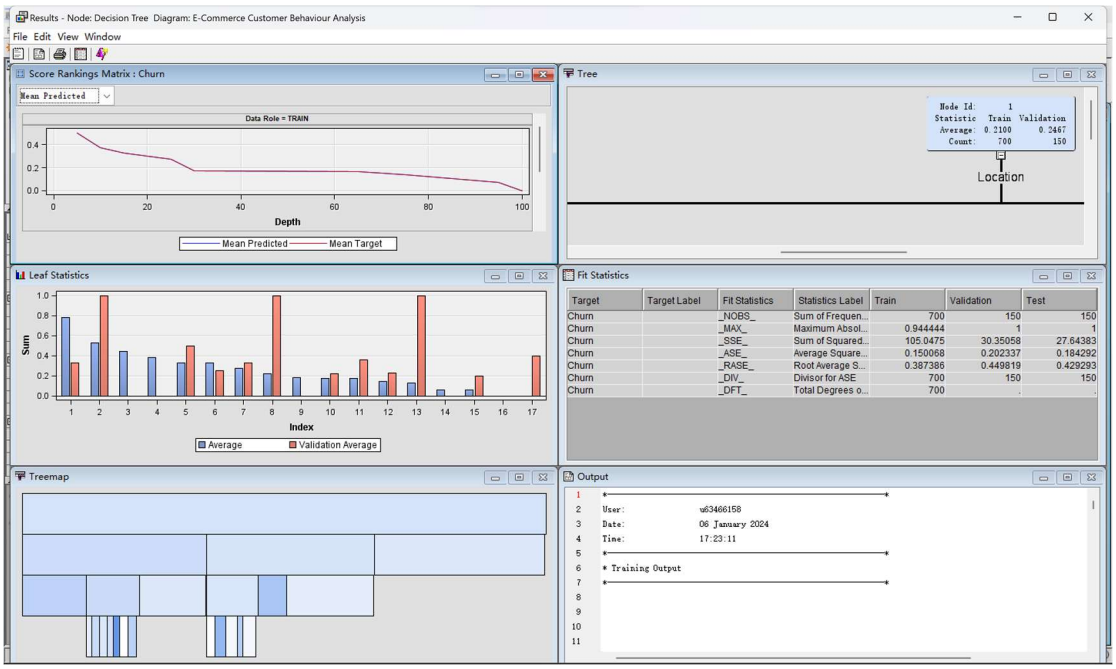
Number of Surrogate Rules (0 to 4):
Surrogate rules are backup criteria used to split a node when the primary splitting criterion is missing for some cases. This helps handle missing data and maintains the robustness of the tree structure. Introducing up to 4 surrogate rules provides a fallback for handling missing values without relying on imputation alone.

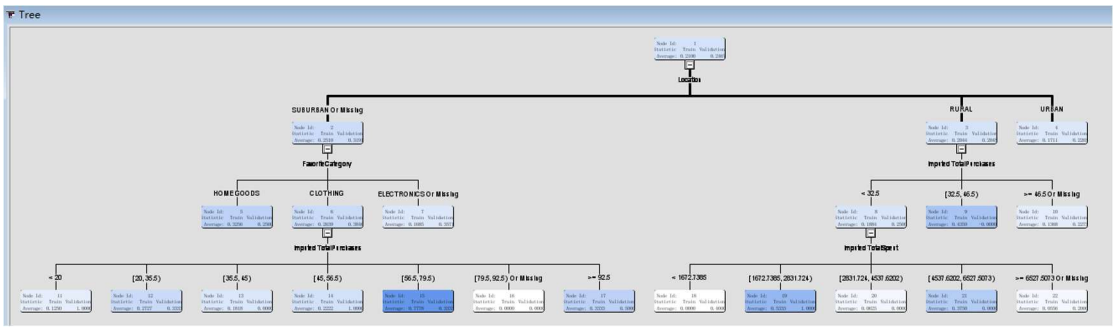
Assessment Measure (ensure it is 'decision'):
The assessment measure should be 'decision' to focus on classification accuracy. This measure will evaluate how well the tree distinguishes between the classes, which is crucial for a binary outcome like churn. This setting aligns the tree's assessment with the goal of making accurate predictions rather than simply fitting the training data.

Subtree Method (from 'Assessment' to 'Largest'):
Changing the subtree method from 'Assessment', which might focus on validation performance, to 'Largest' alters the criteria for subtree selection. 'Largest' may retain more of the tree structure, which could capture additional patterns, because the 'Assessment' method is too conservative and we want the model to retain more detail from the training data.

Observation Based Importance (No to Yes):
Enabling observation-based importance allows the tree to weight splits based on how well they improve the prediction for each observation, which can lead to a more accurate representation of variable importance. This setting is particularly useful if we have imbalanced classes or if some observations are more important than others.

And then we got the improved results as shown in the pictures below:





Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Frequencies	700	150	150
Churn		_MAX_	Maximum Absolute Error	0.944444	1	1
Churn		_SSE_	Sum of Squared Errors	105.0475	30.35058	27.64383
Churn		_ASE_	Average Squared Error	0.150068	0.202337	0.184292
Churn		_RASE_	Root Average Squared Error	0.387386	0.449819	0.429293
Churn		_DIV_	Divisor for ASE	700	150	150
Churn		_DFT_	Total Degrees of Freedom	700	.	.

Output

```
*------------------------------------------------------------*
User:                  u63466158
Date:                  06 January 2024
Time:                  17:23:11
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

                    Measurement    Frequency
Role                   Level          Count

ID                  INTERVAL           2
INPUT               INTERVAL           5
INPUT               NOMINAL            5
REJECTED            NOMINAL            1
TARGET              INTERVAL           1
TIMEID              INTERVAL           1



Predicted and decision variables

Type          Variable         Label

TARGET        Churn
PREDICTED     P_Churn      Predicted: Churn
RESIDUAL      R_Churn      Residual: Churn


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*
```

Output

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_TotalPurchases	Imputed TotalPurchases	2	3	1.0000	0.6291	0.6291
IMP_TotalSpent	Imputed TotalSpent	1	3	0.8804	0.6164	0.7002
IMP_Age	Imputed Age	0	4	0.6068	0.6153	1.0140
FrequencyOfWebsiteVisits		0	3	0.5515	0.6188	1.1219
IMP_MembershipLevel	Imputed MembershipLevel	0	3	0.5384	0.0000	0.0000
Tenure		0	1	0.3996	0.0000	0.0000
FavoriteCategory		1	0	0.3537	0.0000	0.0000
Location		1	0	0.2962	1.0000	3.3759
Occupation		0	1	0.2262	0.0000	0.0000

Output

Tree Leaf Report

Node Id	Depth	Training Observations	Training Average	Validation Observations	Validation Average	Training Root ASE	Validation Root ASE
4	1	228	0.17	59	0.22	0.37655	0.41740
10	2	117	0.14	22	0.23	0.34359	0.42874
7	2	89	0.17	14	0.36	0.37434	0.51494
5	2	86	0.33	20	0.25	0.46859	0.43956
9	2	39	0.44	6	0.00	0.49587	0.43590
22	3	18	0.06	5	0.20	0.22906	0.42528
20	3	16	0.06	3	0.00	0.24206	0.06250
19	3	15	0.53	1	1.00	0.49889	0.46667
16	3	12	0.00	2	0.00	0.00000	0.00000
17	3	12	0.33	2	0.50	0.47140	0.52705
18	3	12	0.00	5	0.40	0.00000	0.63246
12	3	11	0.27	3	0.33	0.44536	0.47528
13	3	11	0.18	1	0.00	0.38569	0.18182
14	3	9	0.22	1	1.00	0.41574	0.77778
15	3	9	0.78	3	0.33	0.41574	0.64788
11	3	8	0.13	1	1.00	0.33072	0.87500
21	3	8	0.38	2	0.00	0.48412	0.37500

Fit Statistics

Target=Churn Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	700.000	150.000	150.000
MAX	Maximum Absolute Error	0.944	1.000	1.000
SSE	Sum of Squared Errors	105.048	30.351	27.644
ASE	Average Squared Error	0.150	0.202	0.184
RASE	Root Average Squared Error	0.387	0.450	0.429
DIV	Divisor for ASE	700.000	150.000	150.000
DFT	Total Degrees of Freedom	700.000		

Output

Assessment Score Rankings

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

Depth	Number of Observations	Mean Target	Mean Predicted
5	63	0.50794	0.50794
10	8	0.37500	0.37500
15	98	0.32653	0.32653
25	11	0.27273	0.27273
30	248	0.17339	0.17339
65	89	0.16854	0.16854
75	117	0.13675	0.13675
95	42	0.07143	0.07143
100	24	0.00000	0.00000

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

Depth	Number of Observations	Mean Target	Mean Predicted
5	10	0.20000	0.54821
10	24	0.25000	0.33035
25	4	0.50000	0.26010
30	60	0.21667	0.17123
70	14	0.35714	0.16854
75	22	0.22727	0.13675
90	1	1.00000	0.12500
95	8	0.12500	0.05816
100	7	0.28571	0.00000

Output

Assessment Score Distribution

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

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.739 - 0.778	0.77778	0.77778	9	0.75833
0.506 - 0.544	0.53333	0.53333	15	0.52500
0.428 - 0.467	0.43590	0.43590	39	0.44722
0.350 - 0.389	0.37500	0.37500	8	0.36944
0.311 - 0.350	0.32653	0.32653	98	0.33056
0.272 - 0.311	0.27273	0.27273	11	0.29167
0.194 - 0.233	0.22222	0.22222	9	0.21389
0.156 - 0.194	0.17073	0.17073	328	0.17500
0.117 - 0.156	0.13600	0.13600	125	0.13611
0.039 - 0.078	0.05882	0.05882	34	0.05833
0.000 - 0.039	0.00000	0.00000	24	0.01944

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

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.739 - 0.778	0.33333	0.77778	3	0.75833
0.506 - 0.544	1.00000	0.53333	1	0.52500
0.428 - 0.467	0.00000	0.43590	6	0.44722
0.350 - 0.389	0.00000	0.37500	2	0.36944
0.311 - 0.350	0.27273	0.32629	22	0.33056
0.272 - 0.311	0.33333	0.27273	3	0.29167
0.194 - 0.233	1.00000	0.22222	1	0.21389
0.156 - 0.194	0.24324	0.17072	74	0.17500
0.117 - 0.156	0.26087	0.13624	23	0.13611
0.039 - 0.078	0.12500	0.05816	8	0.05833
0.000 - 0.039	0.28571	0.00000	7	0.01944

Improved Configuration Summary:

The tree now branches on variables other than 'Location', such as 'FavoriteCategory' and 'TotalPurchases', which suggests a more nuanced analysis of factors contributing to churn.
The tree has a greater depth, indicating that it can capture more complex relationships between variables and churn.
The leaf nodes show a wider spread in the distribution of error rates, suggesting differentiation in the predictive capability across different segments of the tree.
The use of surrogate rules has likely helped to deal with missing data and maintain the integrity of splits, improving the model's robustness.
The tree model has expanded to consider other variables and their interactions, as evidenced by the diverse branches and leaves.

Comparison with Default Configuration:

Model Scoring Improvement highly: The improvement in model score implies that the additional branches and depth have allowed the decision tree to capture more complex patterns that are more predictive of customer churn.
Depth and Complexity: The improved model has a greater depth, allowing for more detailed splits and potentially capturing more complex patterns than the default configuration.
Variable Interaction: Unlike the default model, which primarily used 'Location', the improved model considers other variables, such as 'FavoriteCategory' and 'TotalPurchases', indicating a better-utilized feature space.
Error Rates: The default configuration might have shown a better alignment of error rates between training and validation, but potentially at the cost of oversimplification. The improved model, while having varied error rates, may be capturing more genuine patterns at the risk of some overfitting.
Surrogate Rules: The addition of surrogate rules in the improved model suggests a strategy to handle missing values more effectively than the default setting, which could improve the model's ability to make predictions in the presence of incomplete data.
Assessment Measure and Subtree Method: Switching to 'decision' as the assessment measure and 'Largest' as the subtree method likely resulted in a more complex model that can better capture the true distribution of the target variable.

Based on the results from the updated decision tree configuration, specific inferences can be drawn regarding the factors affecting churn in the e-commerce customer dataset.

Location as a Primary Factor: The initial split on 'Location' in both the default and updated configurations implies that geographic factors are significant in predicting churn. Customers from different locations may have varying experiences or expectations that influence their likelihood to churn.

Favorite Category Significance: The appearance of 'FavoriteCategory' in subsequent splits suggests that the type of products customers frequently purchase is closely linked to their churn behavior. This could indicate satisfaction levels with product offerings or engagement with the platform.

Total Purchases Relevance: The branching on 'TotalPurchases' indicates that the frequency of transactions is a strong indicator of churn, with different purchasing behaviors leading to distinct churn probabilities. Regular purchasers might show more loyalty, while infrequent shoppers could be at higher risk of churn.

Customer Segmentation: Churn appears to be affected by customer segments defined by location and shopping preferences. Tailored strategies can be developed for these segments to address churn.

Purchase Frequency: Customers with lower purchase frequencies may need targeted engagement strategies, such as special offers or reminders, to prevent churn.

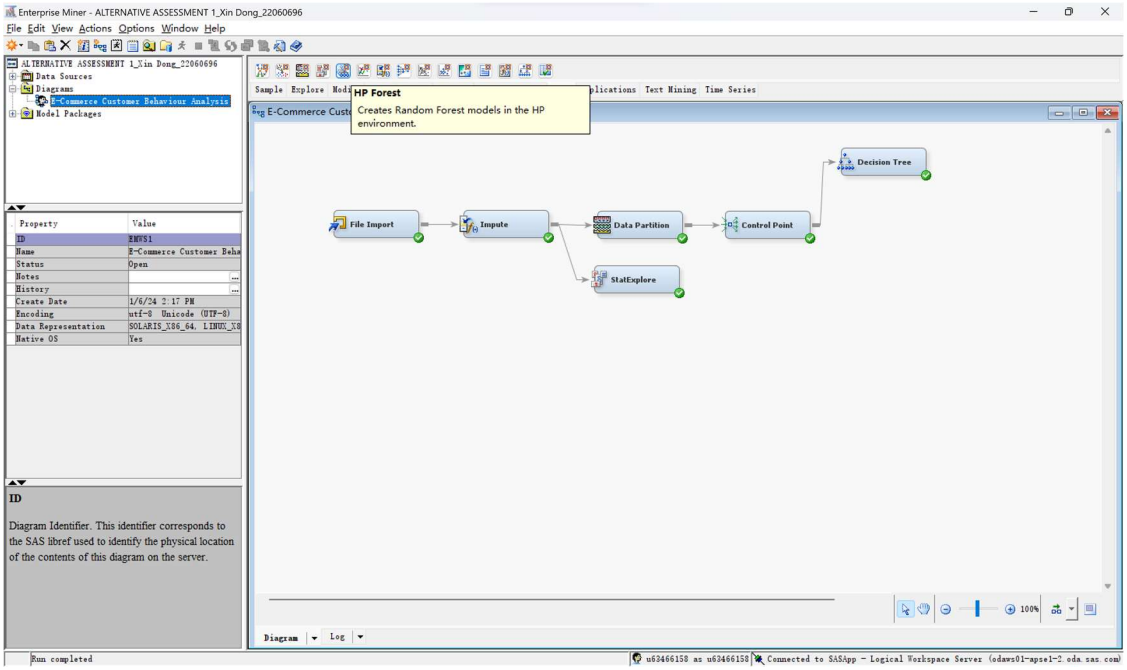
Multi-Dimensional Strategy: Churn prevention requires a multi-dimensional strategy that addresses various factors simultaneously rather than focusing on a single aspect of customer behavior.

Data Completeness: Ensuring data completeness and robust handling of missing data is crucial for accurate churn prediction and subsequent strategy formulation.

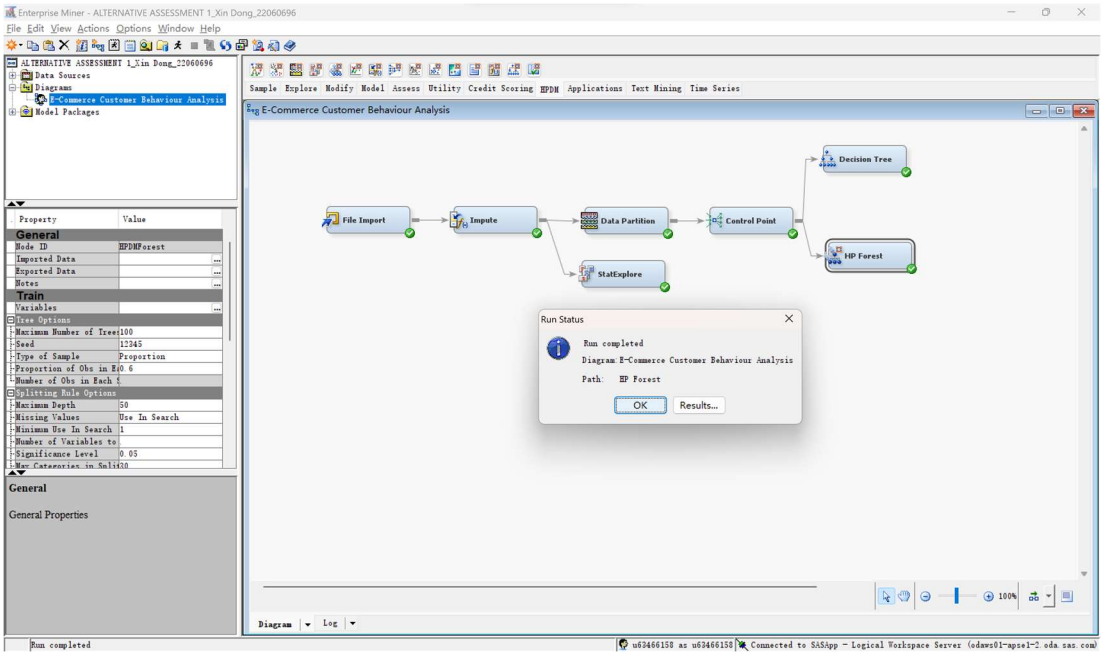
7. Ensemble Methods

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

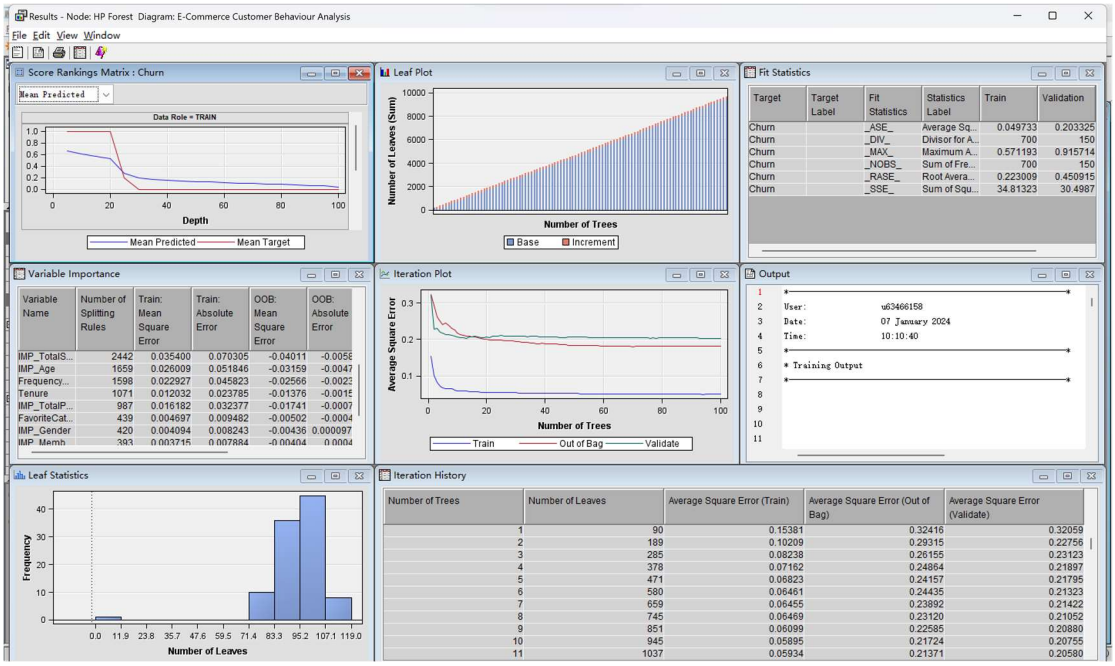
7.1 Apply Random Forest Model:

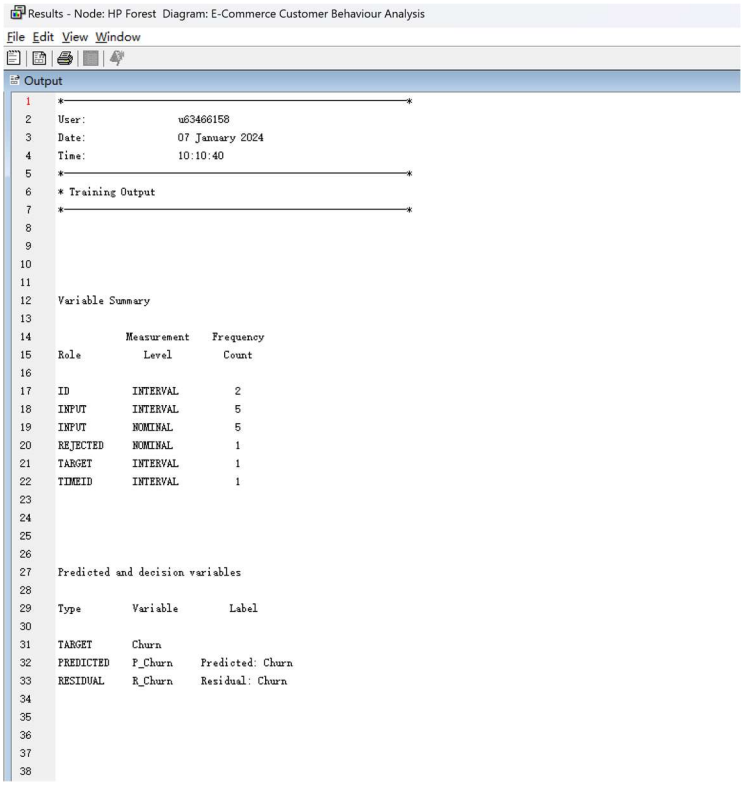


Connect the processed dataset with HP Random Forest Model by connecting Control Point node with HP Forest, and run:



Results of the Random Forest Model application:



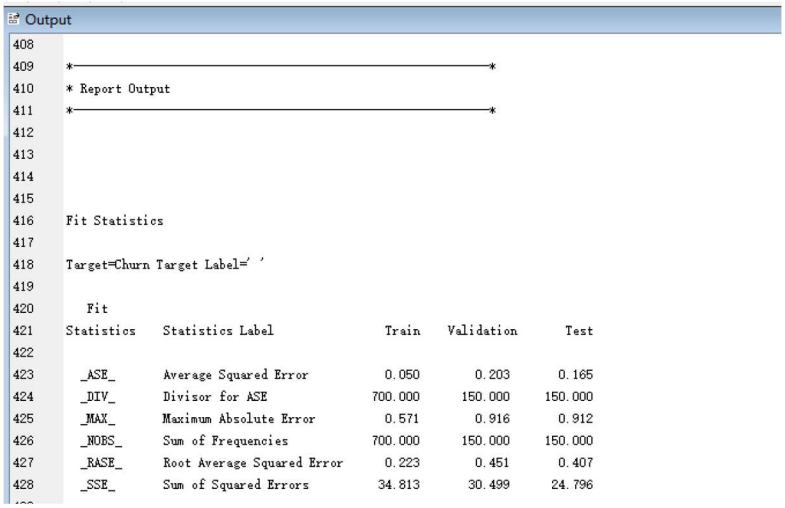


Results - Node: HP Forest Diagram: E-Commerce Customer Behaviour Analysis

File Edit View Window

Output

```
*------------------------------------------------------------*
User:                  u63466158
Date:                  07 January 2024
Time:                  10:10:40
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

             Measurement    Frequency
Role           Level         Count

ID           INTERVAL          2
INPUT        INTERVAL          5
INPUT        NOMINAL           5
REJECTED     NOMINAL           1
TARGET       INTERVAL          1
TIMEID       INTERVAL          1



Predicted and decision variables

Type         Variable        Label

TARGET       Churn
PREDICTED    P_Churn      Predicted: Churn
RESIDUAL     R_Churn      Residual: Churn
```



Output

```
*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*



Fit Statistics

Target=Churn Target Label=' '

   Fit
Statistics    Statistics Label                Train    Validation       Test

  _ASE_       Average Squared Error           0.050         0.203      0.165
  _DIV_       Divisor for ASE               700.000       150.000    150.000
  _MAX_       Maximum Absolute Error          0.571         0.916      0.912
  _NOBS_      Sum of Frequencies            700.000       150.000    150.000
  _RASE_      Root Average Squared Error      0.223         0.451      0.407
  _SSE_       Sum of Squared Errors          34.813        30.499     24.796
```

Output

Assessment Score Rankings

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

Depth	Number of Observations	Mean Target	Mean Predicted
5	35	1.0	0.65856
10	35	1.0	0.60556
15	35	1.0	0.56894
20	35	1.0	0.52559
25	35	0.2	0.27955
30	35	0.0	0.19396
35	35	0.0	0.17221
40	35	0.0	0.15505
45	35	0.0	0.14328
50	35	0.0	0.13190
55	35	0.0	0.12299
60	35	0.0	0.11443
65	35	0.0	0.10657
70	35	0.0	0.09774
75	35	0.0	0.08856
80	35	0.0	0.08201
85	35	0.0	0.07384
90	35	0.0	0.06455
95	35	0.0	0.05443
100	35	0.0	0.03883

Output

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

Depth	Number of Observations	Mean Target	Mean Predicted
5	8	0.12500	0.41979
10	7	0.00000	0.33506
15	8	0.12500	0.30641
20	7	0.57143	0.28669
25	8	0.25000	0.27065
30	7	0.28571	0.25874
35	8	0.25000	0.24628
40	7	0.00000	0.23225
45	8	0.12500	0.22084
50	7	0.14286	0.20788
55	8	0.00000	0.19683
60	7	0.28571	0.18919
65	8	0.50000	0.17962
70	7	0.00000	0.17111
75	8	0.37500	0.16408
80	7	0.28571	0.15472
85	8	0.25000	0.14685
90	7	0.85714	0.13873
95	8	0.25000	0.11855
100	7	0.28571	0.08927

Output

Assessment Score Distribution

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

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.697 - 0.732	1	0.71200	6	0.71447
0.661 - 0.697	1	0.67114	8	0.67886
0.625 - 0.661	1	0.64293	16	0.64324
0.590 - 0.625	1	0.61015	36	0.60762
0.554 - 0.590	1	0.57183	37	0.57201
0.519 - 0.554	1	0.53704	24	0.53639
0.483 - 0.519	1	0.50866	13	0.50078
0.447 - 0.483	1	0.46908	5	0.46516
0.412 - 0.447	1	0.43429	2	0.42954
0.269 - 0.305	0	0.28728	2	0.28708
0.234 - 0.269	0	0.25042	8	0.25146
0.198 - 0.234	0	0.21502	28	0.21585
0.162 - 0.198	0	0.17980	60	0.18023
0.127 - 0.162	0	0.14341	105	0.14461
0.091 - 0.127	0	0.10993	144	0.10900
0.056 - 0.091	0	0.07486	151	0.07338
0.020 - 0.056	0	0.04370	55	0.03776

Output

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

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.503 - 0.527	0.00000	0.52651	1	0.51489
0.480 - 0.503	0.00000	0.48353	1	0.49164
0.457 - 0.480	0.00000	0.47643	1	0.46840
0.387 - 0.410	0.50000	0.39491	2	0.39867
0.364 - 0.387	0.00000	0.37528	1	0.37542
0.341 - 0.364	0.00000	0.35053	4	0.35218
0.317 - 0.341	0.00000	0.33002	5	0.32893
0.294 - 0.317	0.14286	0.30849	7	0.30569
0.271 - 0.294	0.41667	0.28280	12	0.28244
0.248 - 0.271	0.33333	0.25908	15	0.25920
0.224 - 0.248	0.00000	0.23484	12	0.23596
0.201 - 0.224	0.16667	0.21627	12	0.21271
0.178 - 0.201	0.18182	0.19127	22	0.18947
0.155 - 0.178	0.22727	0.16659	22	0.16622
0.131 - 0.155	0.52941	0.14545	17	0.14298
0.108 - 0.131	0.37500	0.12132	8	0.11973
0.085 - 0.108	0.20000	0.10030	5	0.09649
0.062 - 0.085	0.33333	0.07686	3	0.07325

Key Summary and Findings from Random Forest Model results:

Fit Statistics Overview:

The model has a low Average Squared Error (ASE) for the training data, indicating good fit.

However, this error is significantly higher on the validation and test sets, suggesting the model may be overfitting and not generalizing as well to unseen data.
The Maximum Absolute Error (MAX) is relatively high on the validation and test sets, which means there are certain predictions that deviate significantly from the actual values.
The Root Average Squared Error (RASE) also increases from training to validation and test sets, which is consistent with the ASE findings and further indicates overfitting.

Assessment Score Rankings (Train):
The model predicts a high likelihood of churn (mean of 1.0) across several depths when trained, but this prediction decreases as the tree depth increases. This could mean the model is very confident in its predictions for shallow trees, but as more complexity is added (i.e., deeper trees), the confidence in the churn prediction decreases.

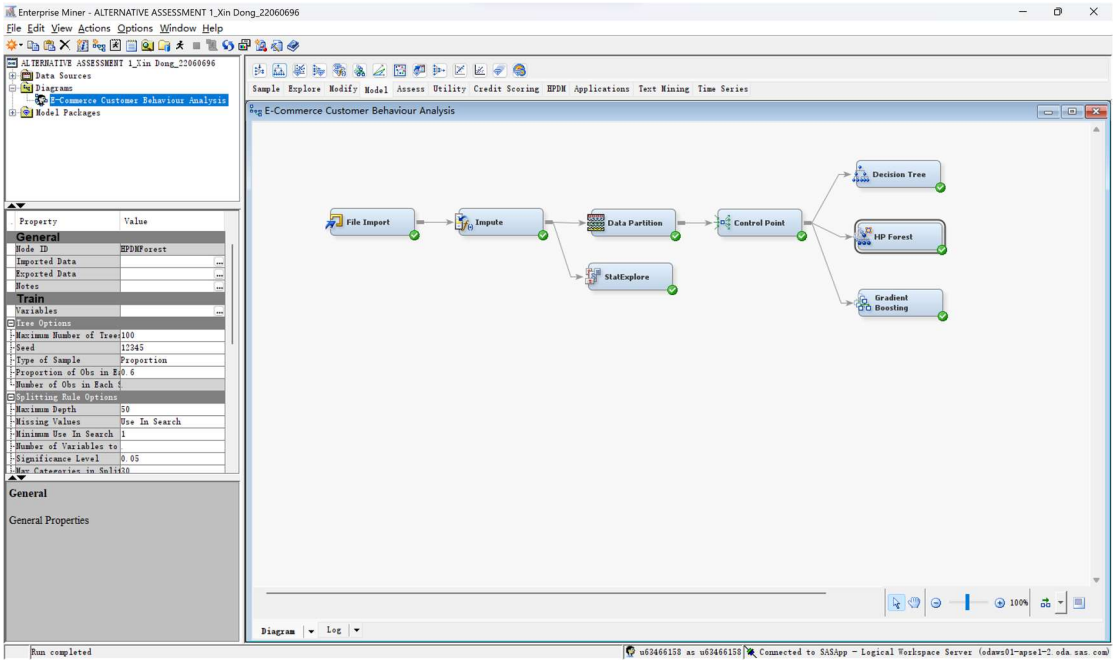
Assessment Score Rankings (Validate):
There's a significant variance in the mean predictions at different depths. For instance, at a depth of 20, the model predicts a high likelihood of churn (mean target of 0.57143), but this is not consistent across other depths. This suggests that the model may be sensitive to the depth parameter and that certain depths capture the churn behavior better than others.

Assessment Score Distribution (Train):
The model score is highest for the top range of predicted values, indicating good performance for cases with a high likelihood of churn.
As the predicted likelihood of churn decreases, so does the model score, suggesting the model is better at identifying clear cases of churn rather than borderline cases.

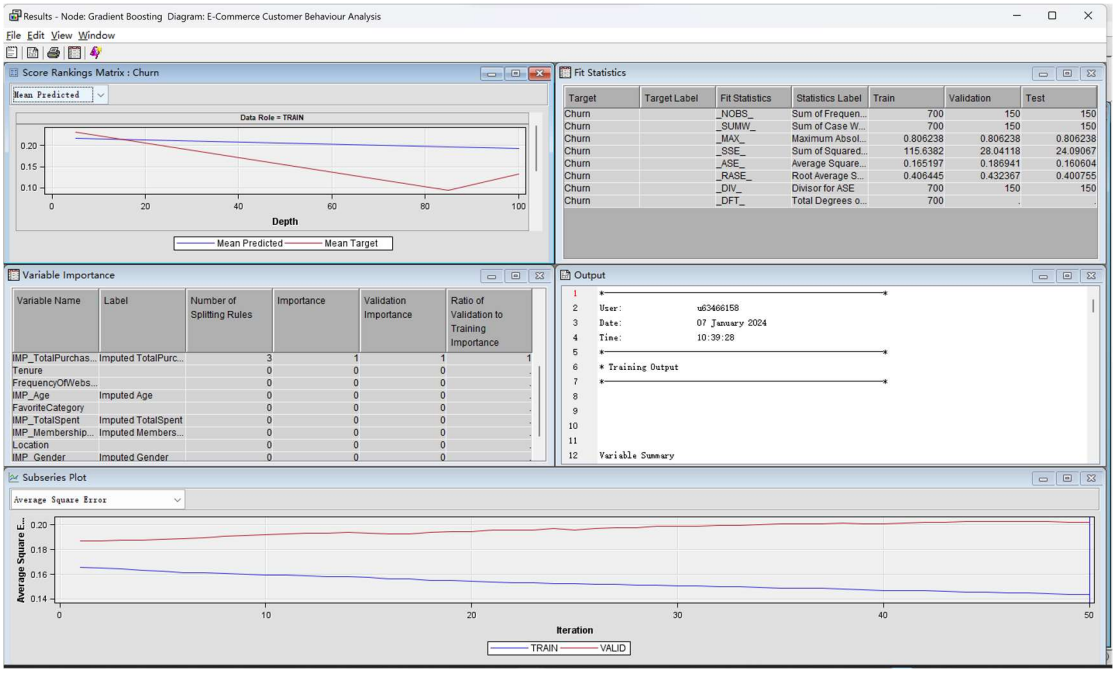
Assessment Score Distribution (Validate):
The model scores are generally lower on the validation set compared to the training set, especially for predictions with a high likelihood of churn, which indicates the model is not performing as well on unseen data.
There are instances (e.g., depth of 80 with a mean target of 0.28571) where the model's predicted likelihood of churn does not align well with the actual mean target, which may indicate issues with model bias or variance on unseen data.

7.2 Apply Gradient Boosting Model:

Connect the processed dataset with HP Random Forest Model by connecting Control Point node with Gradient Boosting, and run:



Results of the Gradient Boosting Model application:



Output

```
*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*



Fit Statistics

Target=Churn Target Label=' '

   Fit
Statistics   Statistics Label                    Train    Validation      Test

  _NOBS_     Sum of Frequencies                700.000      150.000    150.000
  _SUMW_     Sum of Case Weights Times Freq    700.000      150.000    150.000
  _MAX_      Maximum Absolute Error              0.806        0.806      0.806
  _SSE_      Sum of Squared Errors             115.638       28.041     24.091
  _ASE_      Average Squared Error               0.165        0.187      0.161
  _RASE_     Root Average Squared Error          0.406        0.432      0.401
  _DIV_      Divisor for ASE                   700.000      150.000    150.000
  _DFT_      Total Degrees of Freedom          700.000            .          .



Assessment Score Rankings

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

           Number of       Mean       Mean
Depth    Observations     Target  Predicted

   5          575        0.23304    0.21660
  85           95        0.09474    0.19684
 100           30        0.13333    0.19376
```

Output

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

           Number of       Mean       Mean
Depth    Observations     Target  Predicted

   5          135        0.23704    0.21650
  95           13        0.30769    0.19684
 100            2        0.50000    0.19376



Assessment Score Distribution

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

  Range for        Mean       Mean       Number of     Model
  Predicted       Target   Predicted   Observations    Score

0.230 - 0.232   0.25000    0.23186          12        0.23090
0.215 - 0.217   0.23268    0.21627         563        0.21567
0.196 - 0.198   0.09474    0.19684          95        0.19662
0.194 - 0.196   0.13333    0.19376          30        0.19471


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

  Range for        Mean       Mean       Number of     Model
  Predicted       Target   Predicted   Observations    Score

0.230 - 0.232   1.00000    0.23186           2        0.23090
0.215 - 0.217   0.22556    0.21627         133        0.21567
0.196 - 0.198   0.30769    0.19684          13        0.19662
0.194 - 0.196   0.50000    0.19376           2        0.19471
```

Key Summary and Findings from Gradient Boosting Model results:

Fit Statistics Overview:

The model processed 700 observations for training, and 150 each for validation and testing.

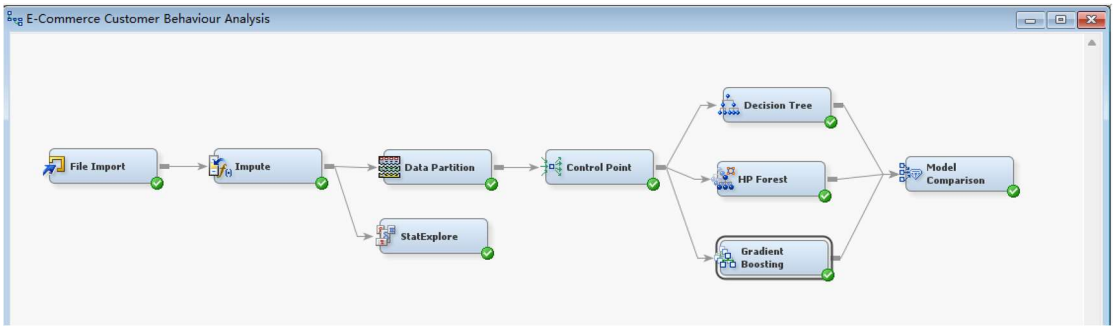
The Maximum Absolute Error (MAX) is the same across training, validation, and test datasets, which may indicate a consistent outlier or extreme value that the model is uniformly missing across all datasets.
The Average Squared Error (ASE) and Root Average Squared Error (RASE) are relatively low, indicating that on average, the model's predictions are close to the actual values. However, there is a slight increase in ASE from training to validation, which may suggest a small degree of overfitting.

Assessment Score Rankings:
For the training data, the model has a reasonable performance with a slight discrepancy between the mean target and mean predicted values at various depths. This discrepancy increases slightly at higher depths, which could be indicative of the model's increasing uncertainty or decreased performance as the complexity (depth) increases.
On the validation set, there are fewer observations at various depths, but the model shows a higher discrepancy between the mean target and mean predicted, particularly at depths of 95 and 100. This could indicate that the model's performance is less stable on validation data for those particular depths.

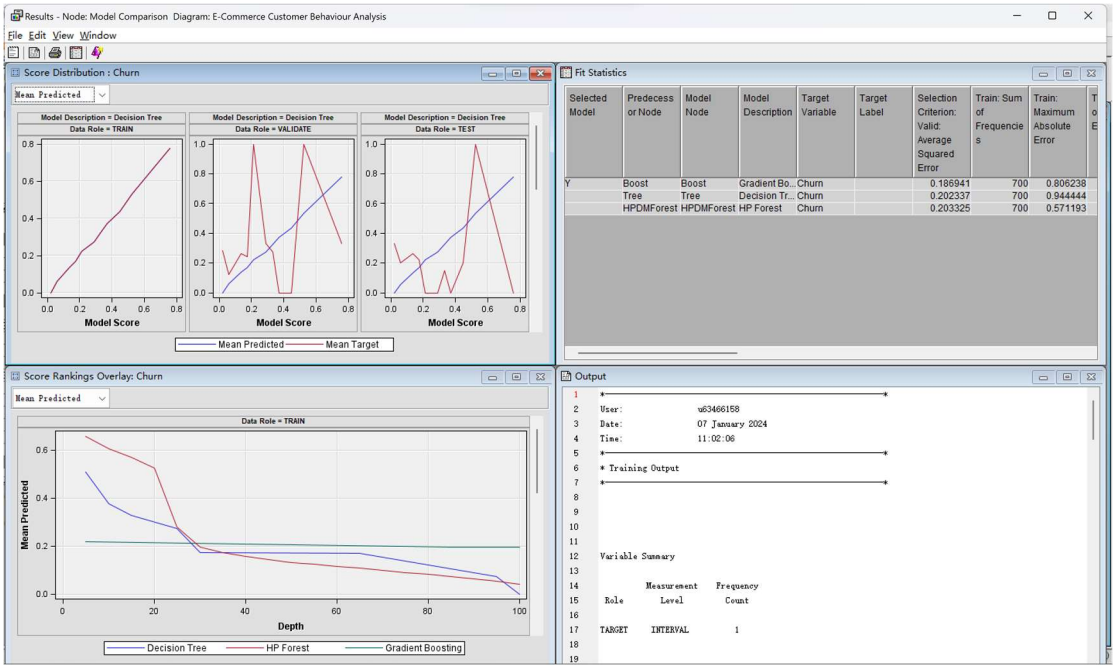
Assessment Score Distribution:
The model scores are closest for predicted ranges around 0.215 - 0.217, showing the smallest gap between the mean target and predicted values for a large number of observations in the training set. This suggests the model is most accurate around this predicted churn probability.
In the validation set, the highest model score occurs at the extreme range of 0.230 - 0.232, where there are very few observations. This could suggest the model is overfitting to specific cases in the validation set.

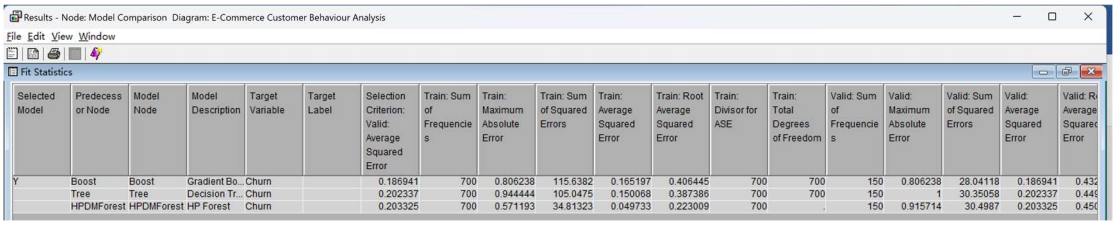
8. Model Comparison and Conclusions

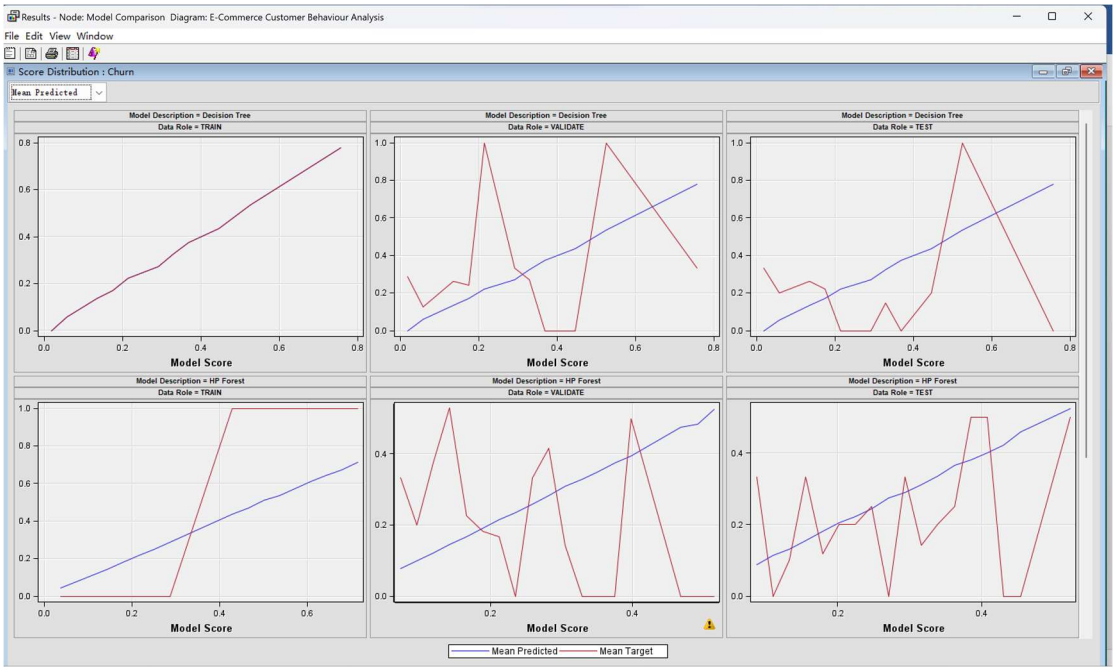
Finally, we use the Model Comparison to compare the three different models and drawn study conclusions

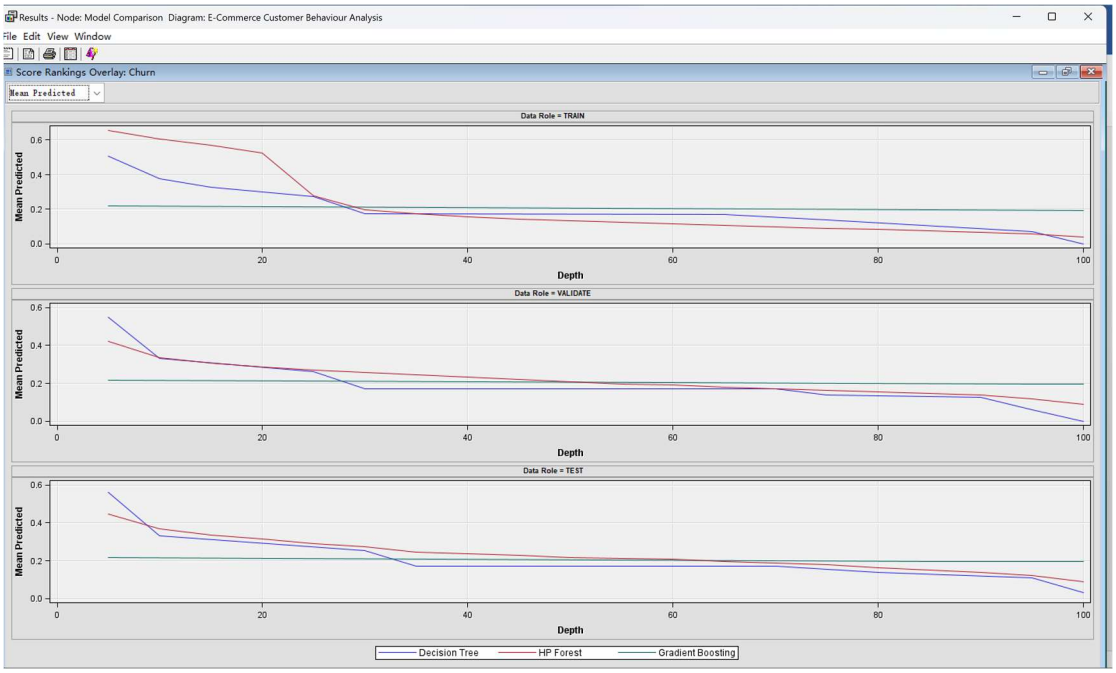


Output Results:


Results - Node: Model Comparison Diagram: E-Commerce Customer Behaviour Analysis
File Edit View Window
Score Distribution : Churn
Mean Predicted
Model Description = Decision Tree
Data Role = TRAIN
Data Role = VALIDATE
Data Role = TEST
Model Score
Mean Predicted
Mean Target
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Average Squared Error
Train: Sum of Frequencies
Train: Maximum Absolute Error
Y Boost Boost Gradient Bo... Churn 0.186941 700 0.806238
Tree Tree Decision Tr... Churn 0.202337 700 0.944444
HPDMForest HPDMForest HP Forest Churn 0.203325 700 0.571193
Score Rankings Overlay: Churn
Mean Predicted
Data Role = TRAIN
Depth
Decision Tree
HP Forest
Gradient Boosting
Output
User: w63466158
Date: 07 January 2024
Time: 11:02:06
* Training Output
Variable Summary
Role Measurement Level Frequency Count
TARGET INTERVAL 1


Results - Node: Model Comparison Diagram: E-Commerce Customer Behaviour Analysis
File Edit View Window
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Average Squared Error
Train: Sum of Frequencies
Train: Maximum Absolute Error
Train: Sum of Squared Errors
Train: Average Squared Error
Train: Root Average Squared Error
Train: Divisor for ASE
Train: Total Degrees of Freedom
Valid: Sum of Frequencies
Valid: Maximum Absolute Error
Valid: Sum of Squared Errors
Valid: Average Squared Error
Y Boost Boost Gradient Bo... Churn 0.186941 700 0.806238 115.6382 0.165197 0.406445 700 700 150 0.806238 28.04118 0.186941
Tree Tree Decision Tr... Churn 0.202337 700 0.944444 105.0475 0.150068 0.387386 700 700 150 1 30.35058 0.202337
HPDMForest HPDMForest HP Forest Churn 0.203325 700 0.571193 34.81323 0.049733 0.223009 700 150 0.915714 30.4987 0.203325


Results - Node: Model Comparison Diagram: E-Commerce Customer Behaviour Analysis
File Edit View Window
Score Distribution : Churn
Mean Predicted
Model Description = Decision Tree
Data Role = TRAIN
Data Role = VALIDATE
Data Role = TEST
Model Description = HP Forest
Data Role = TRAIN
Data Role = VALIDATE
Data Role = TEST
Model Score
Mean Predicted
Mean Target



Output

```
28
29  Fit Statistics
30  Model Selection based on Valid: Average Squared Error (_VASE_)
31
32                                              Valid:      Train:
33                                             Average     Average
34  Selected                                   Squared     Squared
35   Model      Model Node    Model Description   Error       Error
36
37     Y        Boost         Gradient Boosting   0.18694     0.16520
38              Tree          Decision Tree       0.20234     0.15007
39              HPDMForest    HP Forest           0.20332     0.04973
40
41
42
43
44
45
46
47
48
49
50
51  Fit Statistics Table
52  Target: Churn
53
54  Data Role=Train
55
56  Statistics                                          Boost       Tree     HPDMForest
57
58  Train: Average Squared Error                        0.165      0.150        0.050
59  Selection Criterion: Valid: Average Squared Error   0.187      0.202        0.203
60  Train: Total Degrees of Freedom                   700.000    700.000            .
61  Train: Divisor for ASE                            700.000    700.000      700.000
62  Train: Maximum Absolute Error                       0.806      0.944        0.571
63  Train: Sum of Frequencies                         700.000    700.000      700.000
64  Train: Root Average Squared Error                   0.406      0.387        0.223
65  Train: Sum of Squared Errors                      115.638    105.048       34.813
66  Train: Sum of Case Weights Times Freq             700.000          .            .
67
```

Output

Fit Statistics Table

Target: Churn

Data Role=Train

Statistics	Boost	Tree	HPDMForest
Train: Average Squared Error	0.165	0.150	0.050
Selection Criterion: Valid: Average Squared Error	0.187	0.202	0.203
Train: Total Degrees of Freedom	700.000	700.000	.
Train: Divisor for ASE	700.000	700.000	700.000
Train: Maximum Absolute Error	0.806	0.944	0.571
Train: Sum of Frequencies	700.000	700.000	700.000
Train: Root Average Squared Error	0.406	0.387	0.223
Train: Sum of Squared Errors	115.638	105.048	34.813
Train: Sum of Case Weights Times Freq	700.000	.	.

Data Role=Valid

Statistics	Boost	Tree	HPDMForest
Valid: Average Squared Error	0.187	0.202	0.203
Valid: Divisor for VASE	150.000	150.000	150.000
Valid: Maximum Absolute Error	0.806	1.000	0.916
Valid: Sum of Frequencies	150.000	150.000	150.000
Valid: Root Average Squared Error	0.432	0.450	0.451
Valid: Sum of Squared Errors	28.041	30.351	30.499
Valid: Sum of Case Weights Times Freq	150.000	.	.

Output

Data Role=Test

Statistics	Boost	Tree	HPDMForest
Test: Average Squared Error	0.161	0.184	0.165
Test: Divisor for TASE	150.000	150.000	150.000
Test: Maximum Absolute Error	0.806	1.000	0.912
Test: Sum of Frequencies	150.000	150.000	150.000
Test: Root Average Squared Error	0.401	0.429	0.407
Test: Sum of Squared Errors	24.091	27.644	24.796
Test: Sum of Weights Times Freqs	150.000	150.000	.

——————————————————————————
* Score Output
——————————————————————————

——————————————————————————
* Report Output
——————————————————————————

Based on the model comparison output results as shown in the above pics, here is a summary and key findings:

Model Performance Comparison:

The Gradient Boosting model (Boost) was selected as the best model based on the Valid: Average Squared Error (VASE) with a value of 0.18694, which is the lowest among the three models. The Decision Tree model (Tree) has a slightly higher VASE of 0.20234, and the HP Random Forest model (HPDMForest) is close behind with a VASE of 0.20332.

Training Data Performance:

The HP Random Forest model shows a remarkably lower Average Squared Error (ASE) of 0.050 on training data compared to the Gradient Boosting (0.165) and Decision Tree (0.150) models. However, the HP Random Forest model seems to have overfitted the training data as indicated by the large discrepancy between its train ASE and validation ASE.

Validation Data Performance:

All models have increased ASE in the validation set compared to the training set, but Gradient Boosting has the smallest increase, indicating better generalization.
The Maximum Absolute Error is notably lower for the HP Forest model (0.916) compared to the Decision Tree model (1.000) on validation data.

Testing Data Performance:

The Gradient Boosting model again performs best on the test data with an ASE of 0.161, followed closely by the HP Forest model with an ASE of 0.165, and lastly the Decision Tree model with an ASE of 0.184.
The Maximum Absolute Error on the test data is lowest for the Gradient Boosting model (0.806), which suggests it may be better at handling outliers or extreme cases.

General Findings:

The Gradient Boosting model not only performed best on the validation set (which was the selection criterion) but also maintained consistent performance on the test set, indicating it has a good balance between bias and variance.
The HP Forest model, despite its exceptional performance on the training data, showed signs of overfitting with a much higher error rate on the validation and test data.
The Decision Tree model, while not outperforming the other models, still holds a respectable performance considering its simplicity compared to the ensemble methods.

In conclusion, the **Gradient Boosting** model is recommended for its strong and consistent performance across all datasets, indicating a robust prediction capability for the churn target variable. It strikes a good balance between fitting the training data and generalizing to unseen data.

Derive Business insights from our analysis:

- Customer Churn Prediction: The Gradient Boosting model has shown to be the most effective at predicting e-commerce customer churn. This insight can be leveraged by the business to identify at-risk customers early and to deploy retention strategies proactively to reduce churn.

- Focus on Model Generalization: The lower validation and test errors of the Gradient Boosting model suggest that it generalizes well to new, unseen data. This implies that the business can expect a reliable performance from this model when applied to real-world e-commerce customer behavior, thus making informed decisions about future marketing and customer service strategies.

- Resource Allocation for Customer Retention: The precision of the Gradient Boosting model allows the e-commerce business to allocate resources more efficiently. Rather than applying broad retention strategies across the entire customer base, the business can target those specific customers identified by the model as being at high risk of churn, optimizing both the impact and cost-effectiveness of these strategies.

- Identifying Key Factors for Churn: The variable importance measures from the models, particularly from the Gradient Boosting model, can highlight the most influential factors contributing to e-commerce customer churn. Understanding these factors can inform the business on which areas (e.g., customer service, product features, pricing strategies) require improvement to enhance customer satisfaction and loyalty.

- Overfitting Awareness: The HP Forest model's overfitting on the training data serves as a reminder to the business that complex models, while powerful, need to be carefully validated to ensure they perform well in practice and not just on historical data.

- Simplicity vs. Complexity in Model Selection: While the Gradient Boosting model outperformed others, the simpler Decision Tree model still provided respectable results. This suggests that for rapid deployment or interpretation, simpler models can still offer valuable insights, especially when time or computational resources are limited.

- Strategic Timing for Interventions: The ability to predict churn at various levels of certainty (as indicated by the range of predictions and mean predicted values) allows the e-commerce business to time their customer interventions more strategically. For example, they can prioritize immediate action for customers predicted with high certainty to churn soon, while planning longer-term strategies for those with a lower immediate risk.

- Tailored Customer Experience: Insights from the model can guide the development of more personalized customer experiences. For example, if frequent purchases are a key variable, loyalty programs or personalized offers can be designed to encourage repeat e-commerce business.

- Feedback Loop for Continuous Improvement: The use of these predictive models creates a feedback loop for the business. By monitoring the outcomes of the strategies implemented based on the model's predictions, the business can further refine its models and strategies, leading to continuous improvement in customer retention efforts.