UNIVERSITY OF MALAYA

WQD7005 DATA MINING

AA1

TITLE:

E-commerce customer behavior analysis

Matric Number:	S2188090
Name:	HEI ZHANCANG

The data set of this project was established through AI and personal modification. The data set can truly represent actual e-commerce customer behavior. The data set contains 3,000 items and 12 attributes.
The dataset for an e-commerce website contains several key attributes related to customer transactions. Here is an overview of the dataset structure:

Github link：https://github.com/HEI640322/7005-AA1

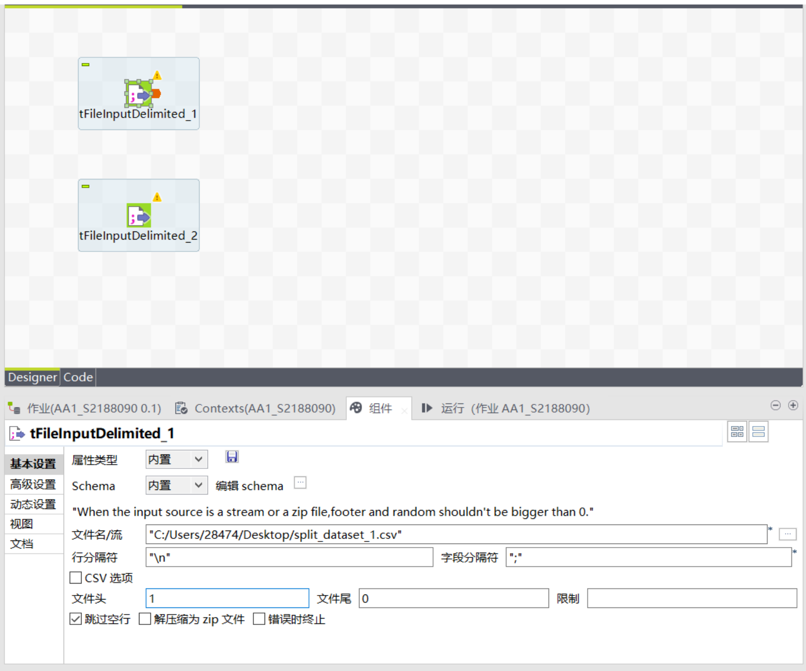
CustomerID	A unique identifier for each customer.
Age	The age of the customer.
Gender	The gender of the customer.
Location	The customer’s geographical location.
MembershipLevel	The customer's membership level (e.g., Silver, Gold, Bronze).
TotalPurchases	The total number of purchases made by a customer on the website.
TotalSpent	The total amount spent by the customer.
FavoriteCategory	The customer's favorite shopping category.
LastPurchaseDate	The date of the customer's last purchase.
Churn	Indicates whether the customer has stopped purchasing (1 means churn, 0 means active).
Occupation:	The client’s occupation.
WebsiteVisitFrequency	How often customers visit the website.

Data Integration

Talend Data Integration (Talend DI) is a tool that helps organizations manage, move and transform data. It lets you easily connect different types of data sources, clean and transform data, and move them to where you need them, all through a visual interface without writing code. This tool supports big data processing and has strong community support.

The following is how I use Talend Data Integration to integrate the AI data set with the data set I collected.

By searching for the component of tFileInputDelimited_1, this is typically used for reading CSV files. Then confirm that the line separator is correct, set the field separator, and enter 1 in the file header. After finishing, edit the schema of the tFileInputDelimited_1 file by adding the name and data type of each data field in the data set (such as integer, string, date etc.), as well as attributes such as whether to allow null values and whether to serve as a primary key. After defining the schema, Talend can process the data based on this information to ensure data accuracy and consistency. For example, the LastPurchaseDate field is set to a date type, with the date format 'yyyy-MM-dd', so Talend will expect this field in each record to represent the date in this format when processing the data.



tFileInputDelimited_1

属性类型 内置

Schema 内置 编辑 schema

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

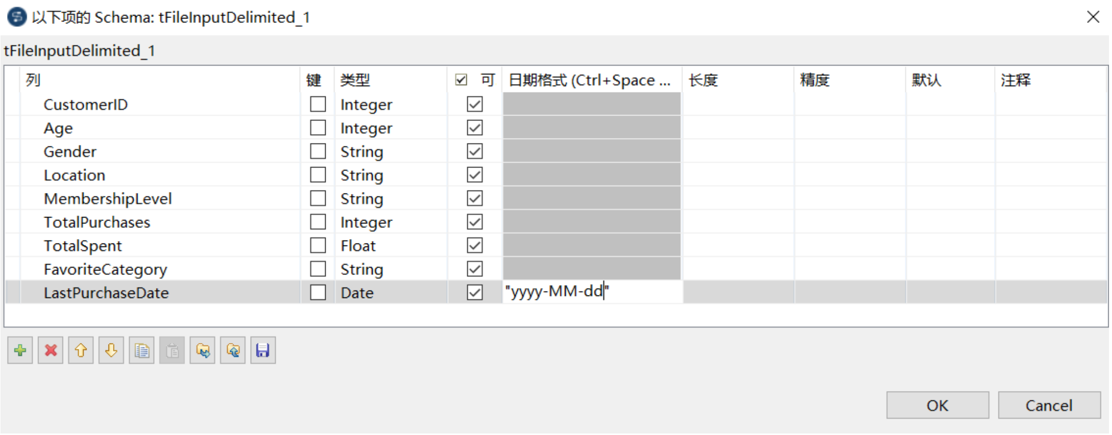
文件名/流 "C:/Users/28474/Desktop/split_dataset_1.csv"

行分隔符 "\n" 字段分隔符 ";"

☐ CSV 选项

文件头 1 文件尾 0 限制

☑ 跳过空行 ☐ 解压缩为 zip 文件 ☐ 错误时终止

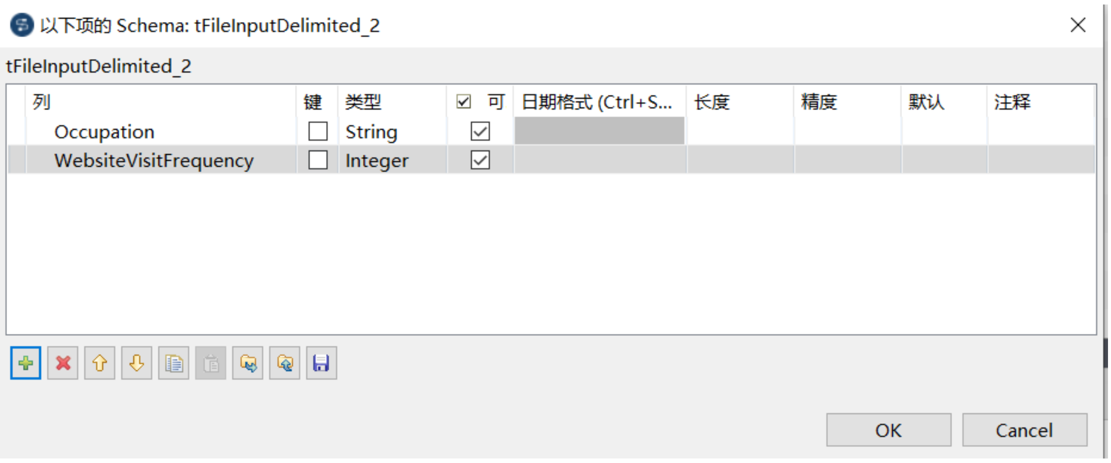


以下项的 Schema: tFileInputDelimited_1

tFileInputDelimited_1

列	键	类型	☑ 可	日期格式 (Ctrl+Space ...	长度	精度	默认	注释
CustomerID	☐	Integer	☑					
Age	☐	Integer	☑					
Gender	☐	String	☑					
Location	☐	String	☑					
MembershipLevel	☐	String	☑					
TotalPurchases	☐	Integer	☑					
TotalSpent	☐	Float	☑					
FavoriteCategory	☐	String	☑					
LastPurchaseDate	☐	Date	☑	"yyyy-MM-dd"				

In the same way, edit the schema for the data set tFileInputDelimited_2.

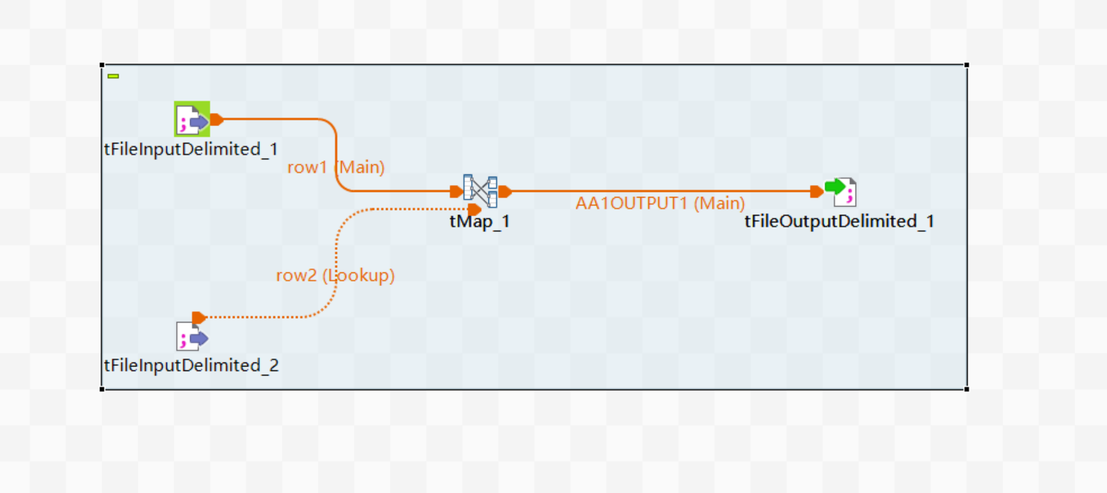


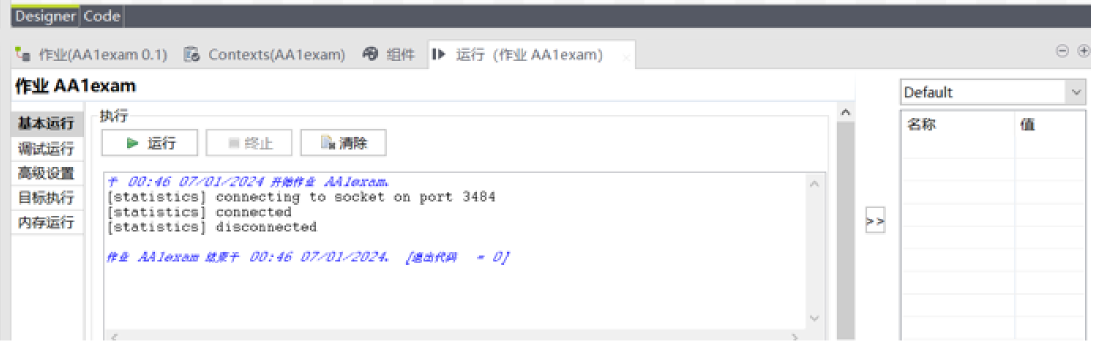
以下项的 Schema: tFileInputDelimited_2

tFileInputDelimited_2

列	键	类型	☑ 可	日期格式 (Ctrl+S...	长度	精度	默认	注释
Occupation	□	String	☑					
WebsiteVisitFrequency	□	Integer	☑					

This figure shows fileInputDelimited_1 and fileInputDelimited_2, connected through the tMap component. The tMap_1 component may be used to integrate the data in fileInputDelimited_1 and fileInputDelimited_2, and perform operations such as merging fields, filtering records, converting data types, etc. row1(Main) represents the main data flow of the first input file, while row2(lookup) represents the data flow of the second input file as a lookup. In tMap, the search flow is usually used to query or supplement data associated with the main data flow. After processing, the output is connected to tFileOutputDelimited_1, which indicates that the processed data will be written to a file in a new delimited format. The result of this process is the integrated output of the two datasets. The execution log of the job shows that the export was successful.

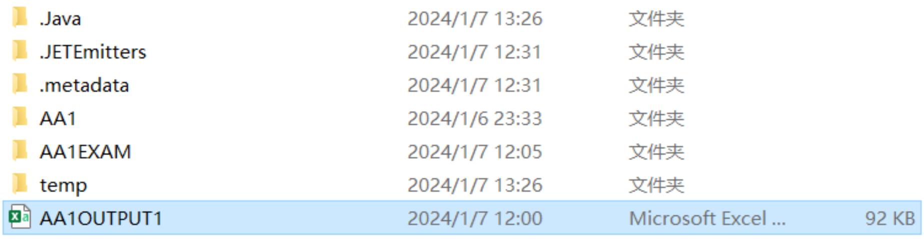




作业 AA1exam

执行

```
于 00:46 07/01/2024 开始作业 AA1exam.
[statistics] connecting to socket on port 3484
[statistics] connected
[statistics] disconnected

作业 AA1exam 结束于 00:46 07/01/2024. [退出代码 = 0]
```

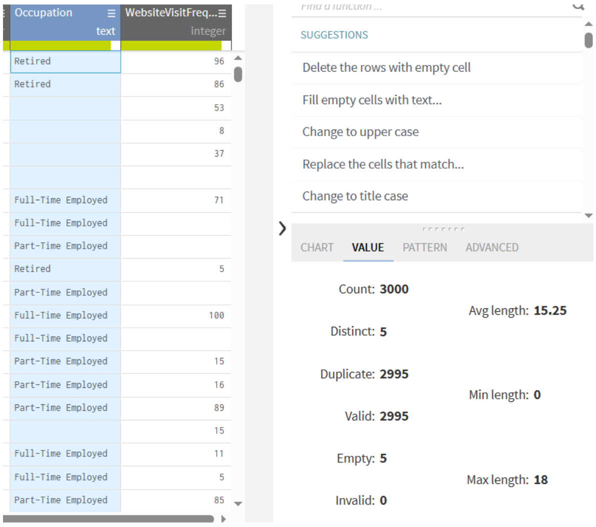


.Java	2024/1/7 13:26	文件夹	
.JETEmitters	2024/1/7 12:31	文件夹	
.metadata	2024/1/7 12:31	文件夹	
AA1	2024/1/6 23:33	文件夹	
AA1EXAM	2024/1/7 12:05	文件夹	
temp	2024/1/7 13:26	文件夹	
AA1OUTPUT1	2024/1/7 12:00	Microsoft Excel ...	92 KB

Talend Data Preparation

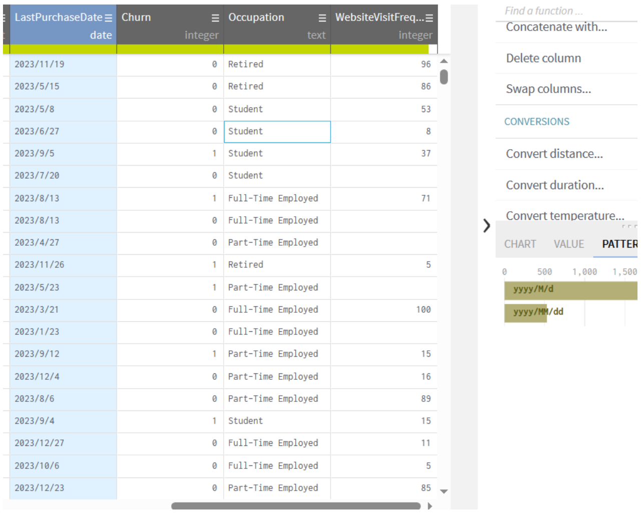
Talend DP is a data preparation tool from Talend Inc. that cleans, transforms, and prepares data for analysis and reporting. It features a visual interface, automation capabilities, multi-data source integration and advanced data analytics to help improve data quality and preparation efficiency.

By uploading the data set to the talend system, we can see that the career attribute of the data set also has a "VALUE" panel on the right, which contains some statistical information about the data: "Empty: 5" Indicates that there are 5 rows of empty data.
In addition, we can see in the date pattern that the format of the date is not yyyy-MM-dd. At the same time, we can see that the gender attribute contains 7 invalid values. We fill in the null value as stduent, change the date to the standard format, and fill in the invalid value. for female. The data set was cleaned.

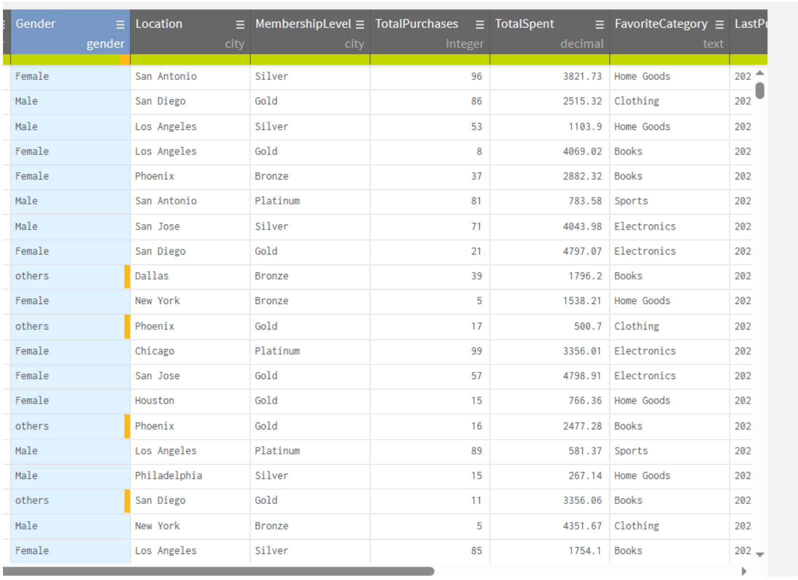


Occupation text	WebsiteVisitFreq... integer
Retired	96
Retired	86
	53
	8
	37
Full-Time Employed	71
Full-Time Employed	
Part-Time Employed	
Retired	5
Part-Time Employed	
Full-Time Employed	100
Full-Time Employed	
Part-Time Employed	15
Part-Time Employed	16
Part-Time Employed	89
	15
Full-Time Employed	11
Full-Time Employed	5
Part-Time Employed	85

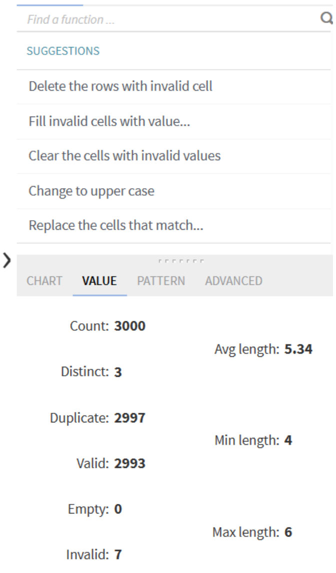
Count: 3000
Avg length: 15.25
Distinct: 5
Duplicate: 2995
Min length: 0
Valid: 2995
Empty: 5
Max length: 18
Invalid: 0



LastPurchaseDate date	Churn integer	Occupation text	WebsiteVisitFreq... integer
2023/11/19	0	Retired	96
2023/5/15	0	Retired	86
2023/5/8	0	Student	53
2023/6/27	0	Student	8
2023/9/5	1	Student	37
2023/7/20	0	Student	
2023/8/13	1	Full-Time Employed	71
2023/8/13	0	Full-Time Employed	
2023/4/27	0	Part-Time Employed	
2023/11/26	1	Retired	5
2023/5/23	1	Part-Time Employed	
2023/3/21	0	Full-Time Employed	100
2023/1/23	0	Full-Time Employed	
2023/9/12	1	Part-Time Employed	15
2023/12/4	0	Part-Time Employed	16
2023/8/6	0	Part-Time Employed	89
2023/9/4	1	Student	15
2023/12/27	0	Full-Time Employed	11
2023/10/6	0	Full-Time Employed	5
2023/12/23	0	Part-Time Employed	85



Gender	Location	MembershipLevel	TotalPurchases	TotalSpent	FavoriteCategory	LastPu
gender	city	city	integer	decimal	text	
Female	San Antonio	Silver	96	3821.73	Home Goods	202
Male	San Diego	Gold	86	2515.32	Clothing	202
Male	Los Angeles	Silver	53	1103.9	Home Goods	202
Female	Los Angeles	Gold	8	4069.02	Books	202
Female	Phoenix	Bronze	37	2882.32	Books	202
Male	San Antonio	Platinum	81	783.58	Sports	202
Male	San Jose	Silver	71	4043.98	Electronics	202
Female	San Diego	Gold	21	4797.07	Electronics	202
others	Dallas	Bronze	39	1796.2	Books	202
Female	New York	Bronze	5	1538.21	Home Goods	202
others	Phoenix	Gold	17	500.7	Clothing	202
Female	Chicago	Platinum	99	3356.01	Electronics	202
Female	San Jose	Gold	57	4798.91	Electronics	202
Female	Houston	Gold	15	766.36	Home Goods	202
others	Phoenix	Gold	16	2477.28	Books	202
Male	Los Angeles	Platinum	89	581.37	Sports	202
Male	Philadelphia	Silver	15	267.14	Home Goods	202
others	San Diego	Gold	11	3356.06	Books	202
Male	New York	Bronze	5	4351.67	Clothing	202
Female	Los Angeles	Silver	85	1754.1	Books	202


Find a function ...
SUGGESTIONS
Delete the rows with invalid cell
Fill invalid cells with value...
Clear the cells with invalid values
Change to upper case
Replace the cells that match...
CHART
VALUE
PATTERN
ADVANCED

Count: 3000

Avg length: 5.34

Distinct: 3

Duplicate: 2997

Min length: 4

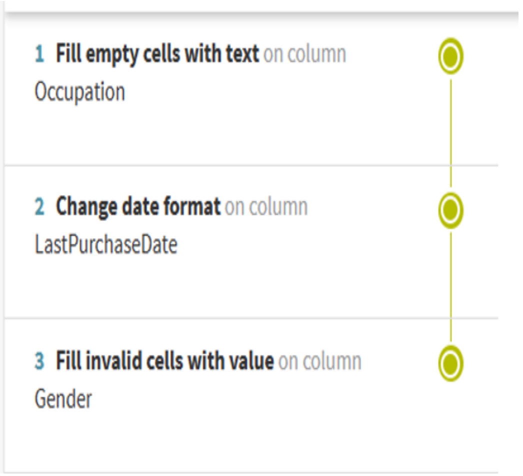
Valid: 2993

Empty: 0

Max length: 6

Invalid: 7

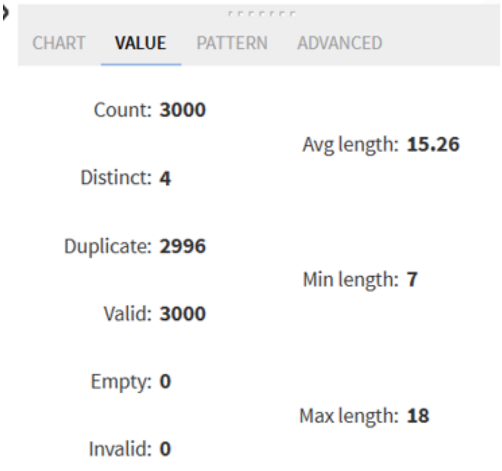
As you can see from the picture below, the data set has been successfully cleaned, there are no null values, no invalid values, and the date format has been corrected successfully.



1 **Fill empty cells with text** on column Occupation

2 **Change date format** on column LastPurchaseDate

3 **Fill invalid cells with value** on column Gender



Count: 3000

Avg length: 15.26

Distinct: 4

Duplicate: 2996

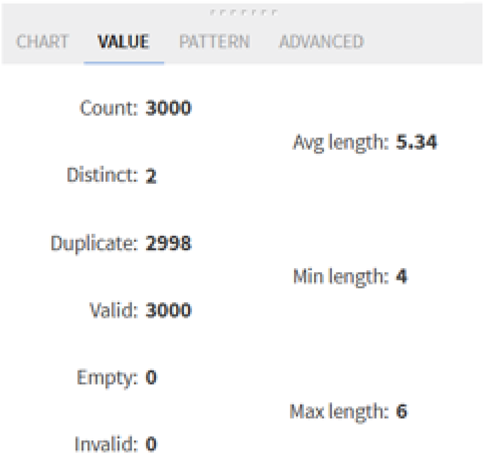
Min length: 7

Valid: 3000

Empty: 0

Max length: 18

Invalid: 0



Count: 3000

Avg length: 5.34

Distinct: 2

Duplicate: 2998

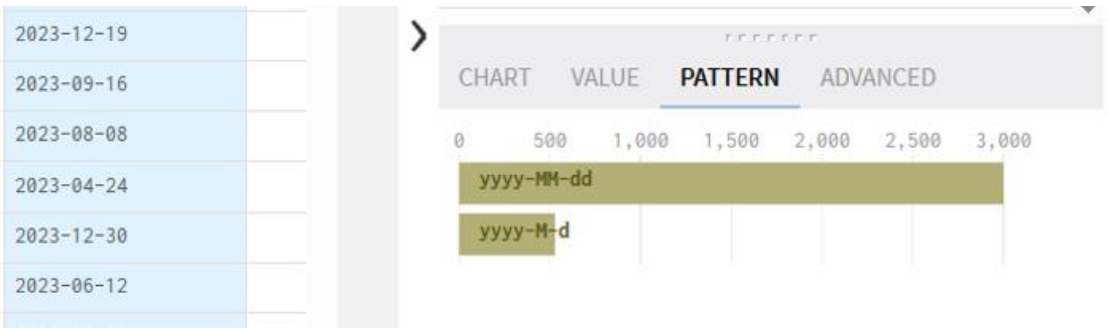
Min length: 4

Valid: 3000

Empty: 0

Max length: 6

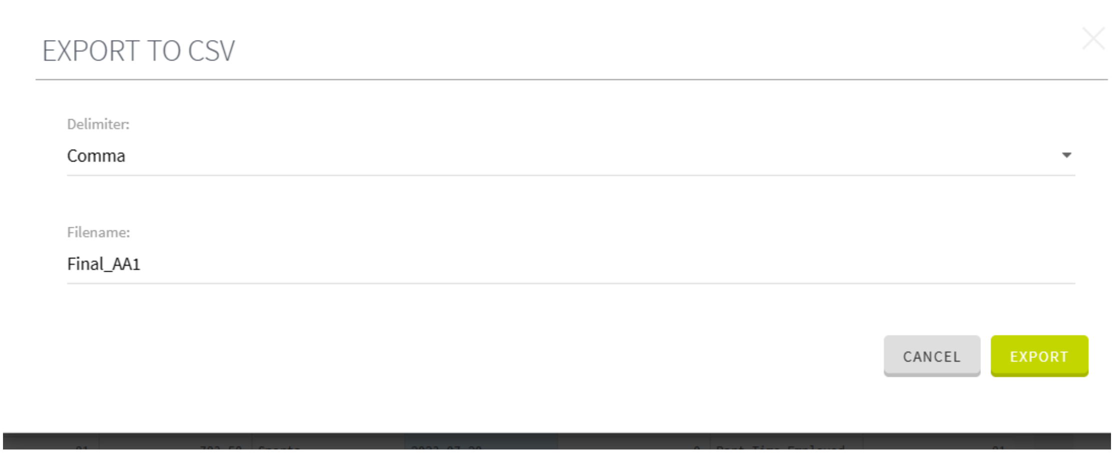
Invalid: 0



2023-12-19
2023-09-16
2023-08-08
2023-04-24
2023-12-30
2023-06-12

This interface is used to export the current data set to a CSV (comma separated values) file. Delimiter: The delimiter is set to comma to separate the data in the CSV file.

Filename: The file name has been set to "Final_AA1".
Creates a CSV file based on the selected delimiter and specified file name.



EXPORT TO CSV

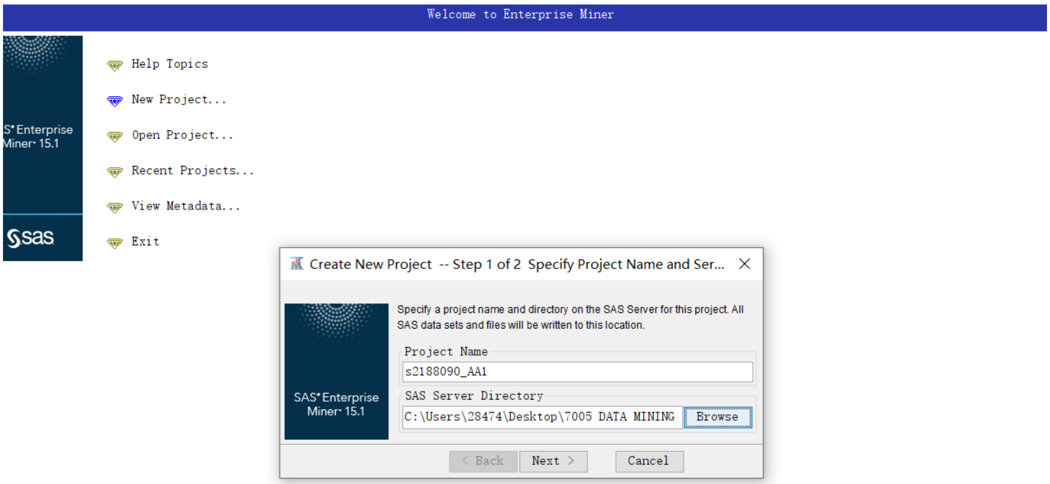
Delimiter: Comma

Filename: Final_AA1

Tasks

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

Data import part: Start SAS Enterprise Miner, create project: enter the name S2188090_AA1, select the server.



Welcome to Enterprise Miner

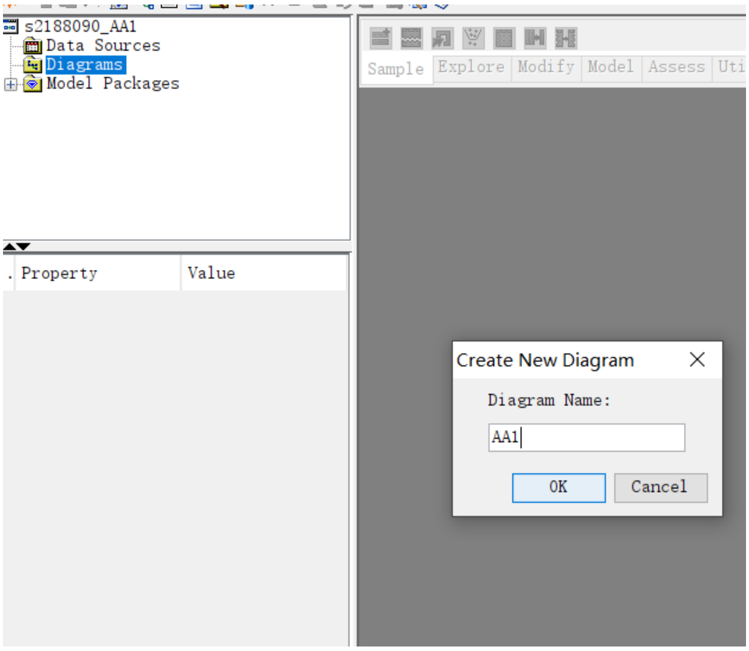
Create New Project -- Step 1 of 2 Specify Project Name and Ser...

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.

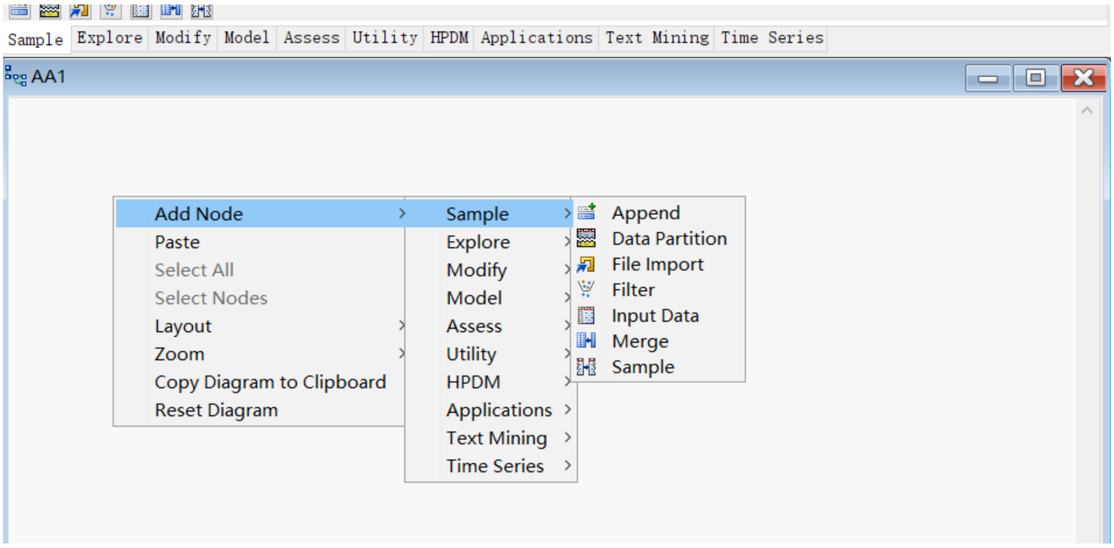
Project Name: s2188090_AA1

SAS Server Directory: C:\Users\28474\Desktop\7005 DATA MINING

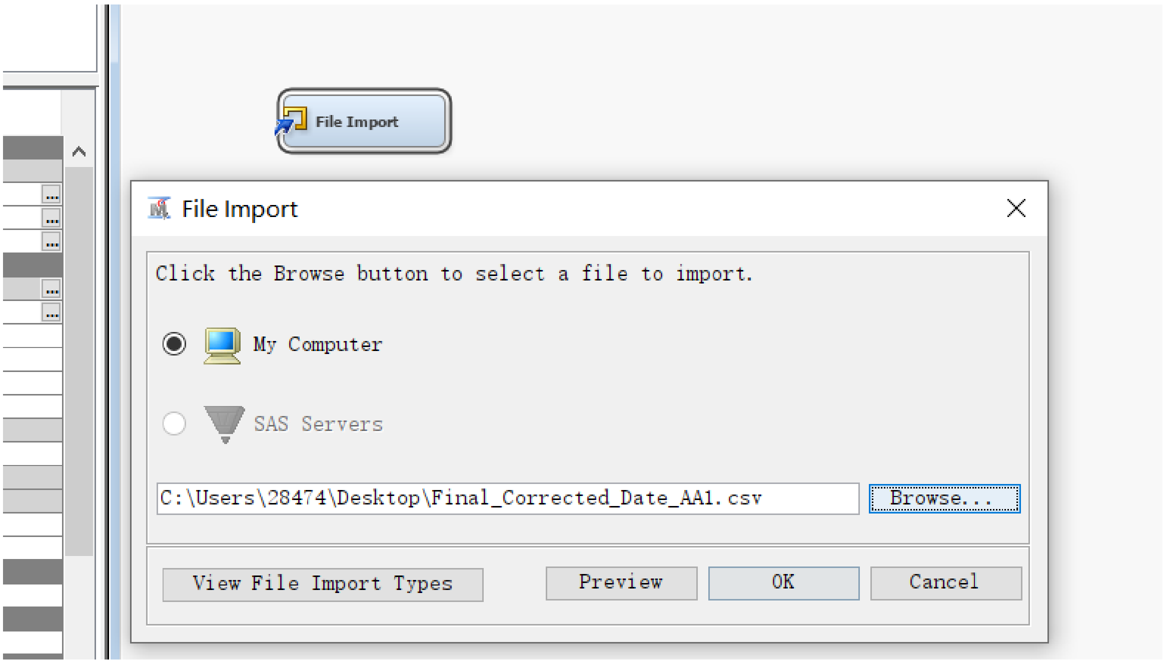
Create New Diagram. Here the user entered "AA1" as the name of the new chart and clicked "OK" to create the chart process.



Right-click the chart interface, Add node appears, then select sample, and then select the file import node.



Click file import, and then there is import file in the left column. By selecting this button, you can import data from the local computer or the connected SAS server. File path: Displays the currently selected file path, such as C:\Users\28474\Desktop\Final_Corrected_Date_AA1.csv, indicating that the CSV file named Final_Corrected_Date_AA1.csv is being imported from the desktop path.



File Import

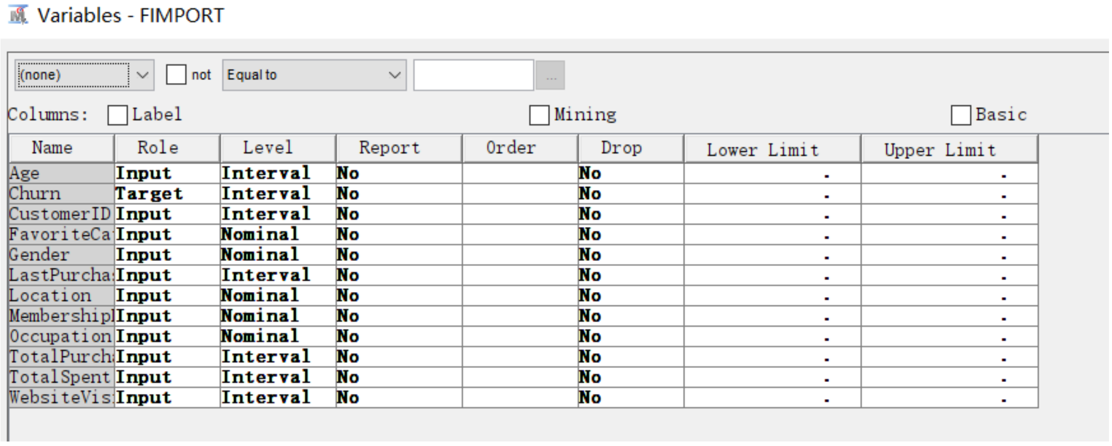
Click the Browse button to select a file to import.

● My Computer

○ SAS Servers

C:\Users\28474\Desktop\Final_Corrected_Date_AA1.csv

Variable settings section: Set churn rate as target and other roles as inputs, which means all these variables will be used in the analysis. Click OK when finished. Since the data set has already been preprocessed in talend DI and talend DP, there is no need to repeat this part.



Variables - FIMPORT

Columns: □ Label □ Mining □ Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Interval	No		No	.	.
CustomerID	Input	Interval	No		No	.	.
FavoriteCa	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurcha	Input	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
Membership	Input	Nominal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
TotalPurch	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.
WebsiteVis	Input	Interval	No		No	.	.

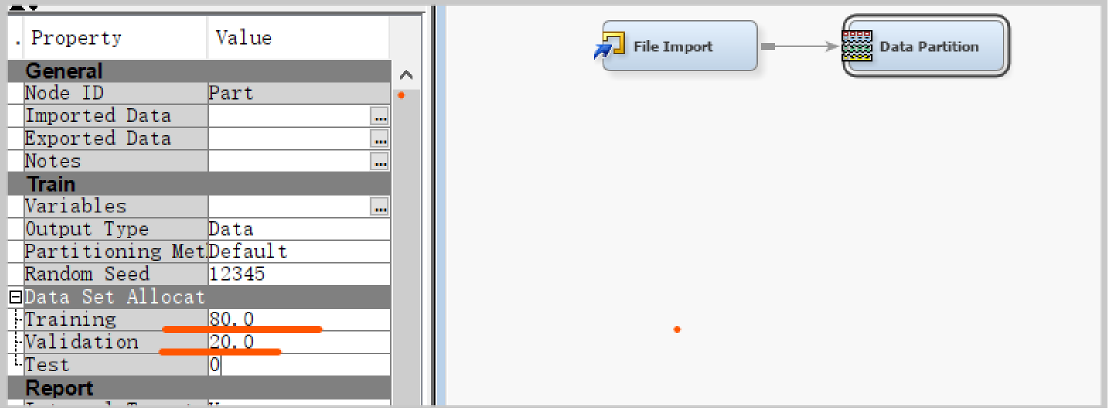
2. Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

We capture the "Data Partition" node to split the data into different sets, which is a common practice in statistical modeling and machine learning.

Data Partition node configuration: Train: Here defines the proportion of the data set allocated to the training set, validation set and test set. Training: Set to 80.0%, which means 80% of the data will be used for model training.

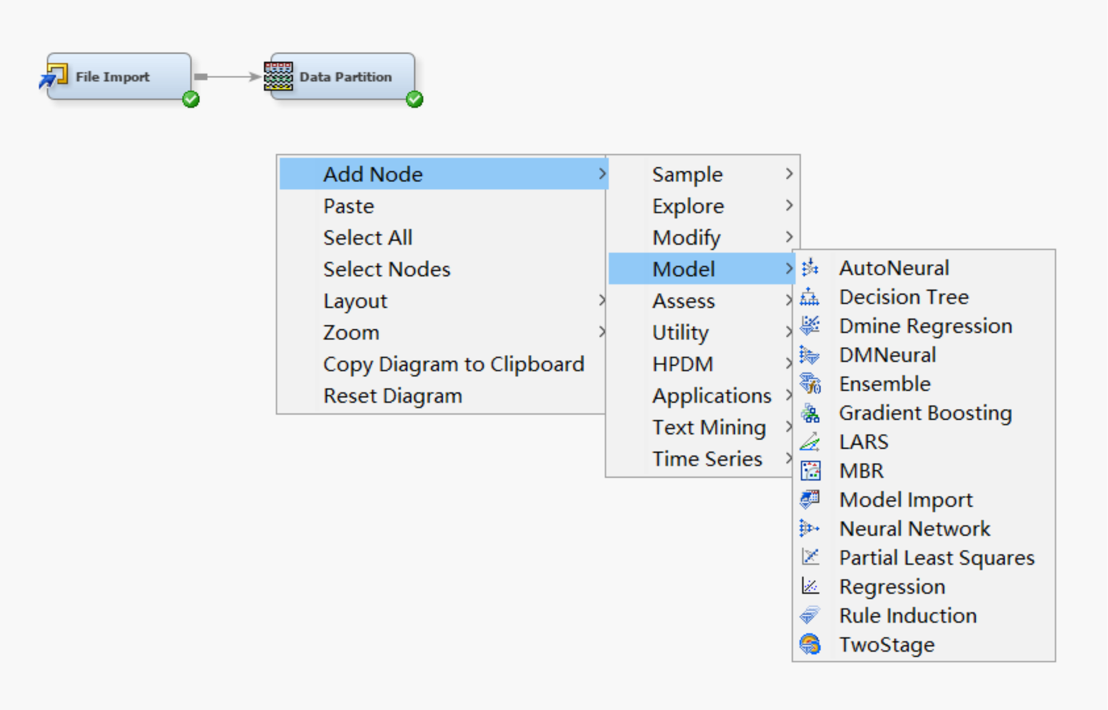
Validation: The display has been set to 20.0%, which means that the remaining 20% of the data will be used for the model validation process.

Random Seed: The random seed is set to 12345 to ensure the reproducibility of the data partition. Using the same random seed ensures the same results for each partition. Partitioning Method: Displayed as "Default", which indicates that the data partition will use the SAS Enterprise Miner default partitioning method. Then click Run.

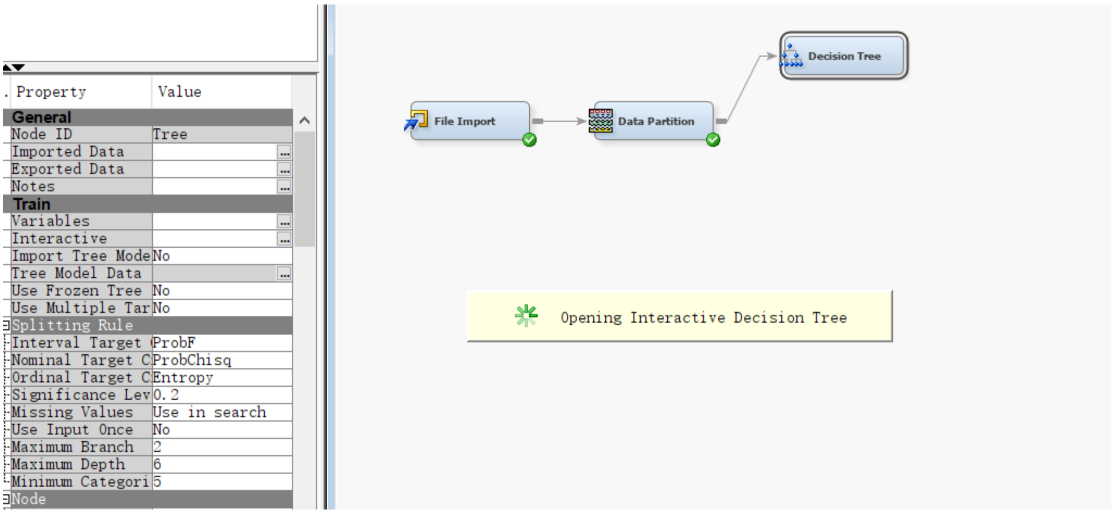


. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Met	Default
Random Seed	12345
⊟Data Set Allocat	
Training	80.0
Validation	20.0
Test	0
Report	

After running, select the decision tree model, right-click on the interface, select add node, then click model, and finally select the decision tree model.



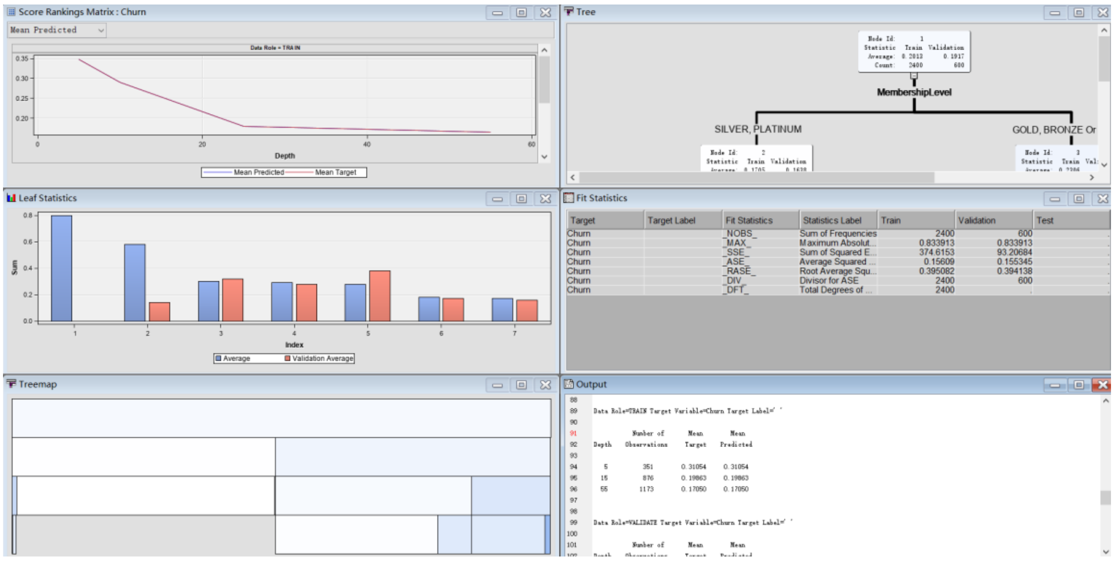
The "Interactive" option in the "Train" configuration of a decision tree model allows the user to manually participate in the decision tree building process. Users can make decisions in real time, such as selecting split points, adjusting tree sizes, applying pruning strategies, and adjusting other model parameters. This interactive approach increases user control over the model building process.



. Property	Value
General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Mode	No
Tree Model Data	...
Use Frozen Tree	No
Use Multiple Tar	No
Splitting Rule	
Interval Target	ProbF
Nominal Target C	ProbChisq
Ordinal Target C	Entropy
Significance Lev	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categori	5
Node	

Opening Interactive Decision Tree

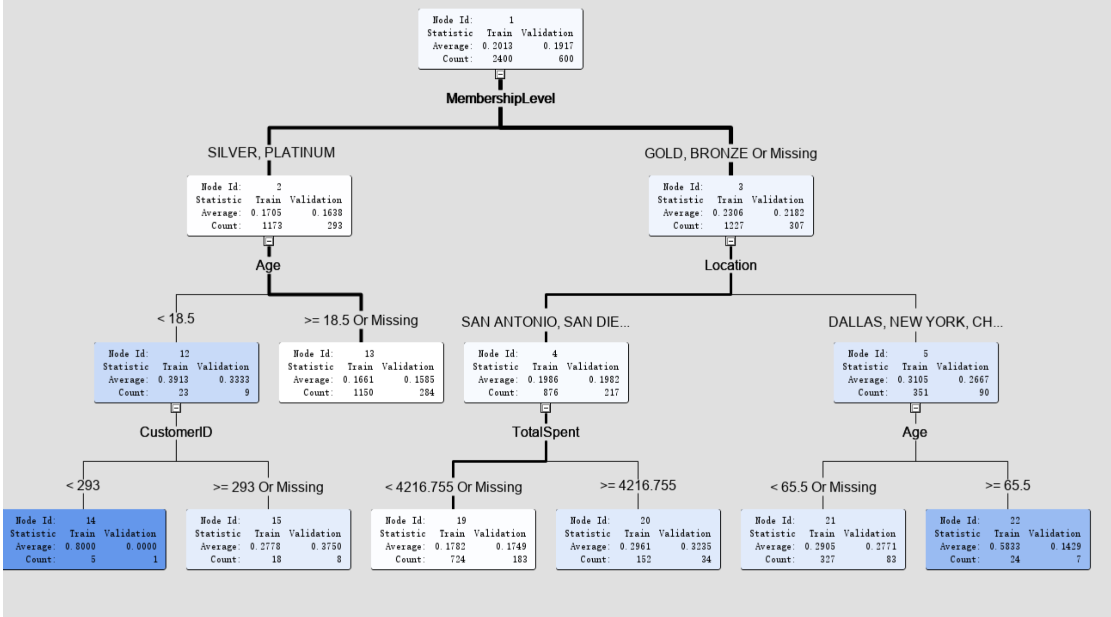
After running, view the results. Below is the overall report of the decision model.

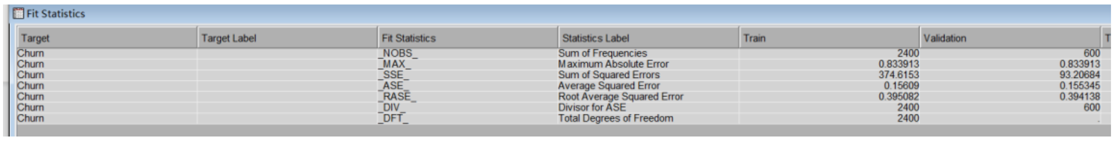


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Frequencies	2400	600	.
Churn		_MAX_	Maximum Absolut...	0.833913	0.833913	.
Churn		_SSE_	Sum of Squared E...	374.6153	93.20684	.
Churn		_ASE_	Average Squared ...	0.15609	0.155345	.
Churn		_RASE_	Root Average Squ...	0.395082	0.394138	.
Churn		_DIV_	Divisor for ASE	2400	600	.
Churn		_DFT_	Total Degrees of ...	2400	.	.

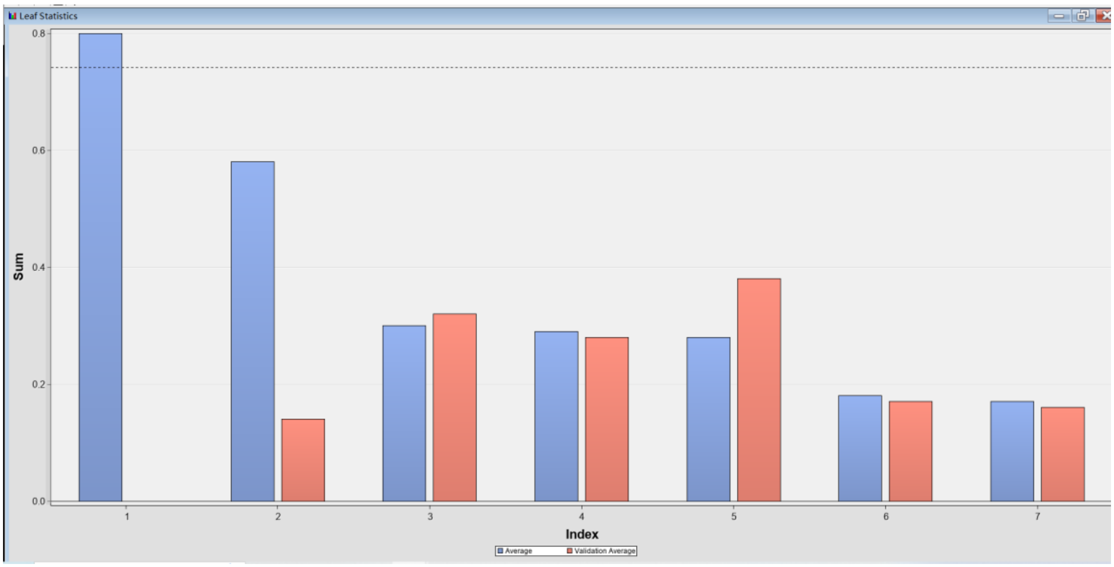
```
Data Role=TRAIN Target Variable=Churn Target Label=' '

            Number of       Mean        Mean
Depth    Observations     Target    Predicted

   5            351       0.31054     0.31054
  15            876       0.19863     0.19863
  55           1173       0.17050     0.17050


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

            Number of       Mean        Mean
```

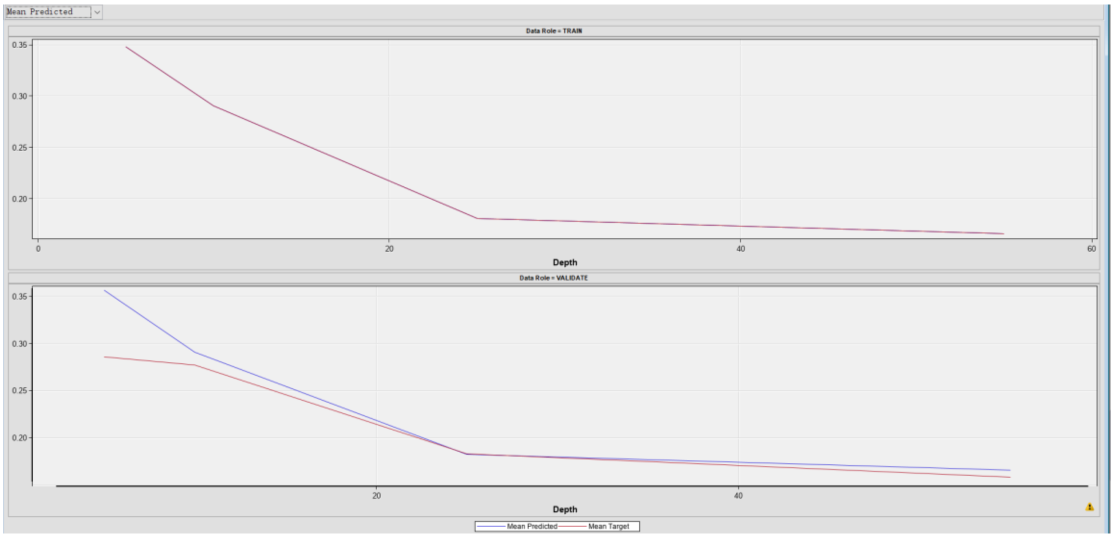




Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	T
Churn		_NOBS_	Sum of Frequencies	2400	600	
Churn		_MAX_	Maximum Absolute Error	0.833913	0.833913	
Churn		_SSE_	Sum of Squared Errors	374.6153	93.20684	
Churn		_ASE_	Average Squared Error	0.15609	0.155345	
Churn		_RASE_	Root Average Squared Error	0.395082	0.394138	
Churn		_DIV_	Divisor for ASE	2400	600	
Churn		_DFT_	Total Degrees of Freedom	2400	.	





We can see that it performs data splitting by attributes such as MembershipLevel, Age, Location, etc., where some nodes have high average values on the training data but perform poorly on the validation set, which may mean overfitting.

Signs and Strategies of Overfitting:

Inconsistent node performance: For example, if a certain leaf node has a high mean on the training data, but this value drops significantly on the validation data (such as Node Id 14), this may be a sign of overfitting. This means that the decision tree may have overlearned specific patterns in the training data that do not necessarily hold true for unseen data.

Complex tree structure: The depth of the tree and the number of branches are large, especially when some leaf nodes contain only few data points (such as Node Id 14), which may cause the model to be overly sensitive to noise in the training data.

Solution strategy:

Pruning: Simplify the model by reducing the depth of the tree or setting a higher minimum number of leaf node samples.

Ensemble methods: Use ensemble methods such as random forests or gradient boosted trees to reduce the risk of overfitting by averaging the results of multiple decision trees.

Regularization: Add regularization terms during model training to penalize complex tree structures.

Pruning the tree before the validation error increases can improve the model's generalization ability.

Cross-validation can be used to determine the optimal tree depth that balances model complexity with predictive accuracy.

Incorporating additional data or using ensemble techniques such as random forests can also help improve model performance on unseen data.

Analyze customer behavior through decision trees.

By analyzing the above report, we can derive some insights about customer behavior:

Customer churn (Churn) prediction: The target variable is Churn, which means the model aims to predict whether a customer will churn. Based on the P_Churn (predicted churn) and R_Churn (residual of churn) variables, the model attempts to score and predict the likelihood of customer churn.

Variable importance:

Location and MembershipLevel are variables of the model splitting rule, which means that these two variables are important in predicting customer churn.

The importance of Location on the training set is 1.0000, but drops to 0.6270 on the validation set, indicating that the prediction contribution of location to the model is not as significant on the validation set as on the training set.

The validation importance of MembershipLevel is 1.0000, indicating that membership level is a very important predictor variable on the validation data set. Its significance ratio on the validation set relative to the training set exceeds 1 (1.2029), possibly indicating that membership levels have a higher predictive value for predicting churn in real-world data.

Tree depth and number of observations:

As the tree depth increases, the model's predictions tend to be consistent. This may indicate that the model is able to differentiate between customers who are at high risk of churn and those who are at low risk of churn.

In the training data, nodes with depth 5 observed a churn rate of 0.31054, while in the validation set, this number was 0.26667. This suggests that within a specific group of customers, the accuracy of predicting churn decreased on the validation set.

Leaf reports and fit statistics:

The Tree Leaf Report shows the average churn rate for different nodes of the tree (splits based on specific rules).

For nodes with a depth of 2, the churn rates of training data and validation data are similar, indicating that the model's prediction of churn at this depth is consistent on the training and validation sets.

RASE (Root Average Squared Error) in Fit Statistics shows the model's error on the training and validation data. A lower error indicates that the model's predictions are more accurate.

At the same time we can also conclude:

The impact of location on churn: A customer's geographic location may be related to their likelihood of churn, with customers in certain locations being more likely to churn.

The impact of membership levels: Different membership levels may affect customer loyalty and churn rates. Higher-level members may have lower churn rates.

Possible influence of age: Although age is not mentioned directly in the report, the decision tree may have used age as one of the splitting rules since it is often one of the factors that influence customer behavior.

Relevance of consumption behavior: Consumption behavior is not explicitly mentioned in the report, but since the TotalSpent variable appears in the decision tree model, we can speculate that the customer's consumption level may be related to its churn risk.

Together, these analytics can help business teams better understand and predict which customers are more likely to churn, so they can adopt retention strategies accordingly.

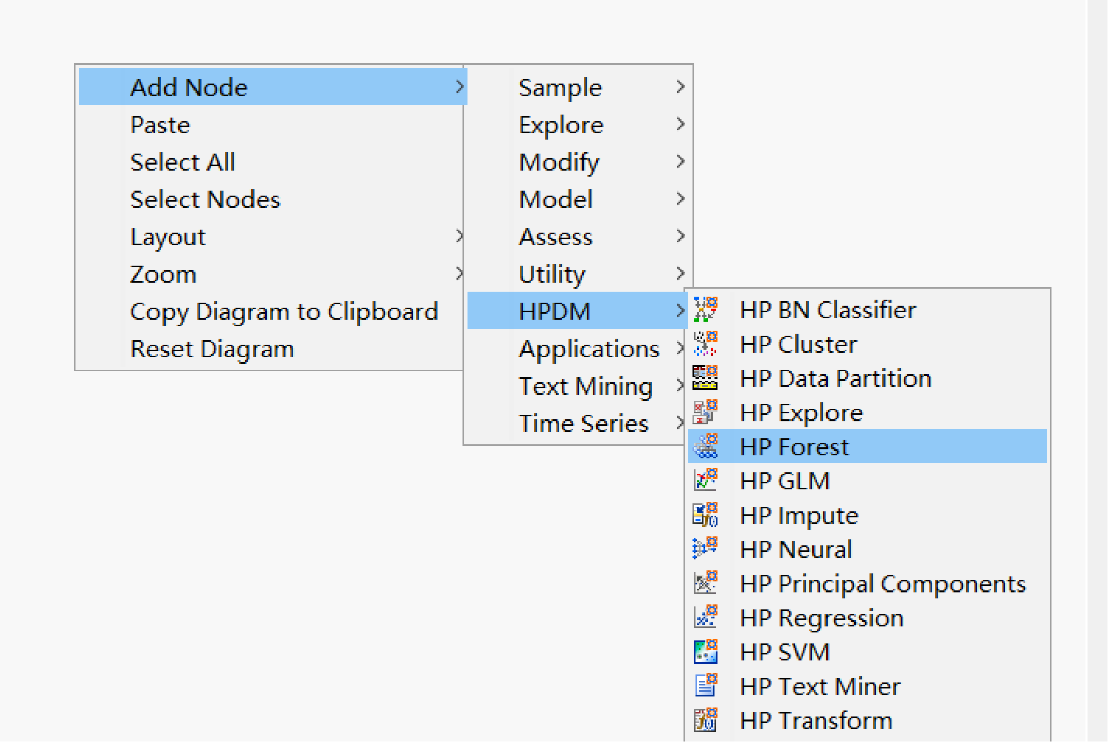
Business strategy:

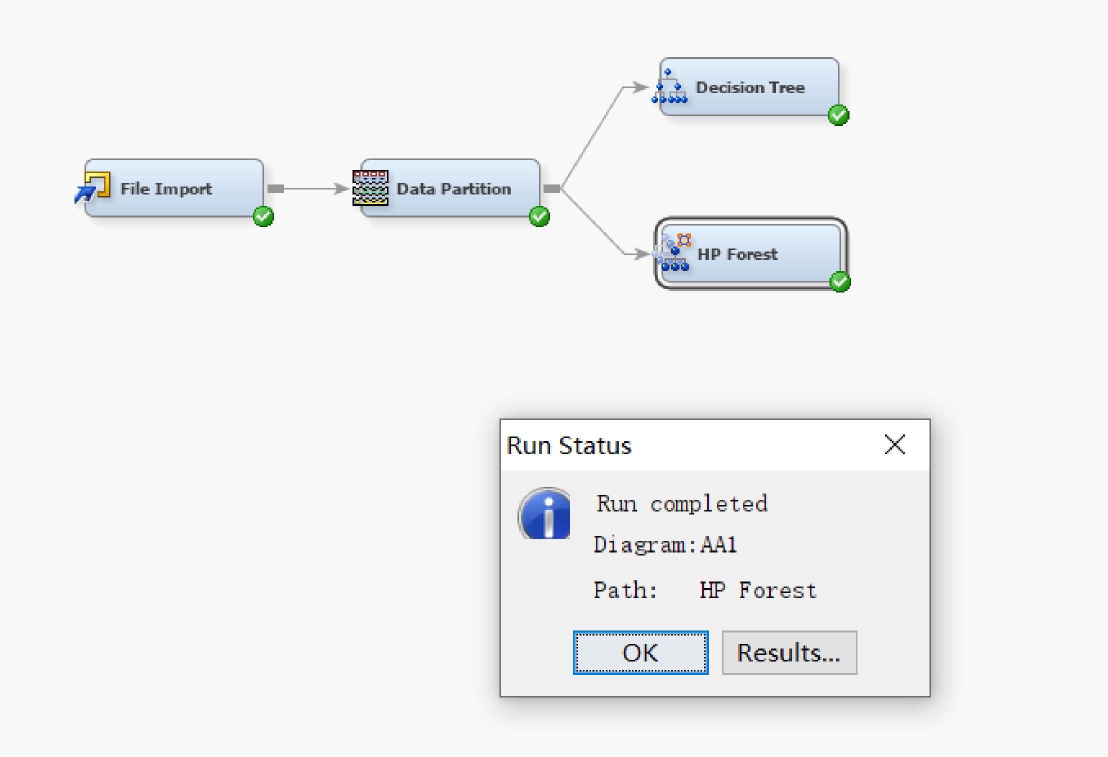
Based on the above analysis, here are some possible business strategies:
Target focus areas: Since Location appears as an important predictor, companies can develop specific marketing strategies based on location. For example, target areas with high churn rates, increase customer engagement campaigns or offer customized offers.
Membership level retention policy:
The importance of MembershipLevel indicates that customers with different membership levels may have different churn risks. Companies can offer loyalty rewards or upgrade offers to membership tiers with a high risk of churn to increase their stickiness.
For those membership levels that exhibit lower churn rates, loyalty can be further enhanced through exclusive offers or customized services.
Personalized communications and promotions:
Use data analysis to identify customer groups at high risk of churn and target them with personalized communications and promotions to increase satisfaction and retention.
Analyzing the nodes of a decision tree can reveal which customer characteristics are associated with churn, helping to design more targeted marketing messages.
Improve customer experience:
If certain nodes are showing unusually high churn rates, this could be a sign of poor customer experience. Investigate these problem areas and take steps to improve the service or product experience.
Optimize resource allocation:
Allocate more resources and attention to those customer groups most likely to churn, for example, by providing additional support or services to reduce their churn rate.
Customer lifecycle management:
Predictive models identify different stages of the customer lifecycle and then provide corresponding interventions at key moments, such as new customer welcome programs, loyal customer rewards or churn prevention programs.
Product and service improvements:
Use churn prediction data to understand why customers leave and improve your product or service accordingly.
Risk Management:
Implement a risk-based pricing strategy to provide different pricing options for customer groups at high risk of churn.
In summary, by analyzing customer behavior using predictive models, companies can more precisely target marketing and service strategies to reduce customer churn and increase customer lifetime value.

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

Random Forest

HPforest can be used instead of random forest by selecting the node and clicking HPDM.



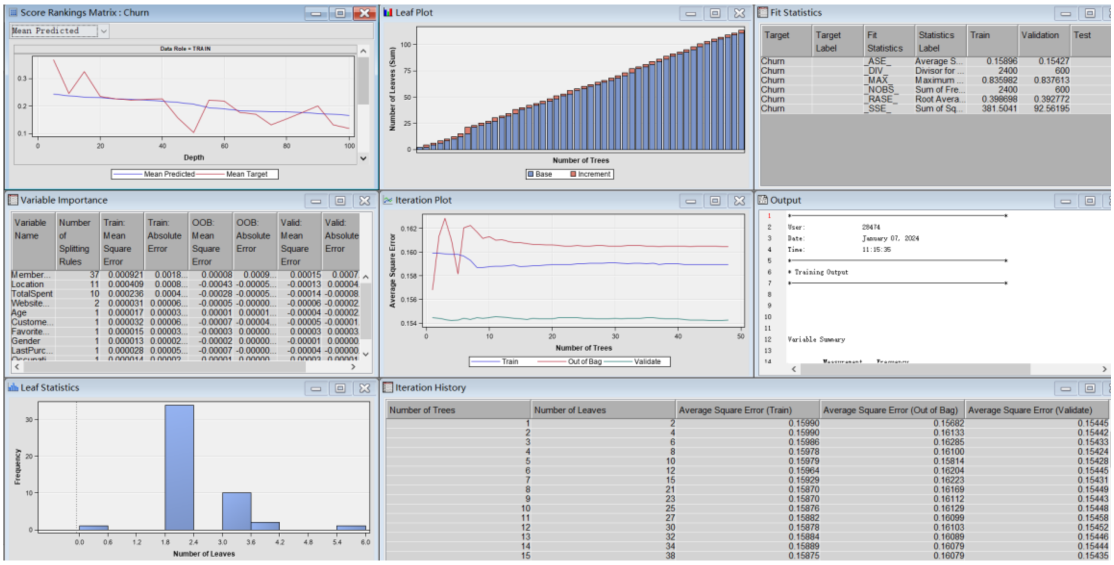


Run completed

Diagram:AA1

Path: HP Forest

Below is the result report of HPforest.



Fit Statistics

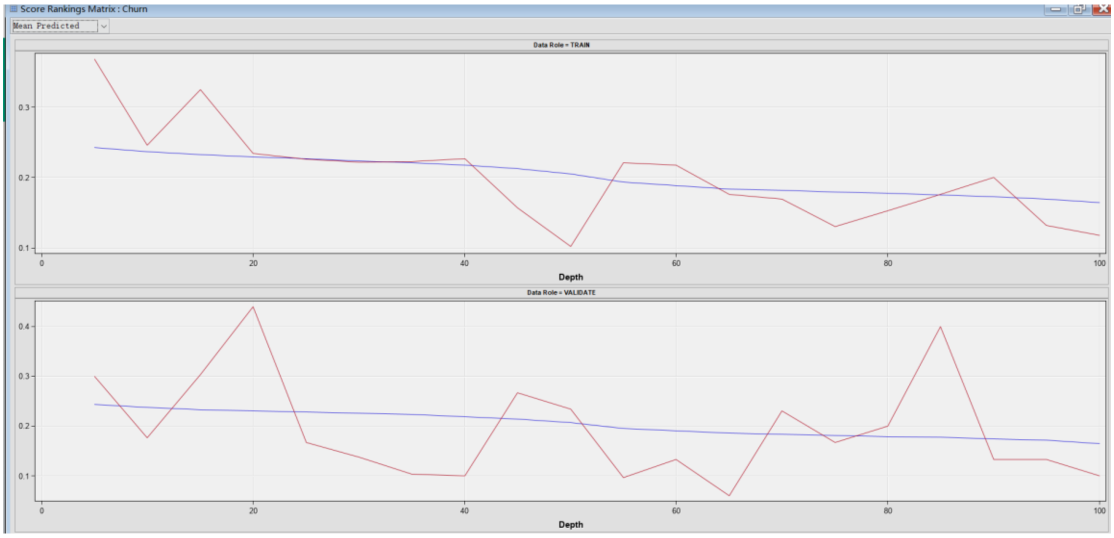
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_ASE_	Average S...	0.15896	0.15427	
Churn		_DIV_	Divisor for ...	2400	600	
Churn		_MAX_	Maximum ...	0.835982	0.837613	
Churn		_NOBS_	Sum of Fre...	2400	600	
Churn		_RASE_	Root Avera...	0.398698	0.392772	
Churn		_SSE_	Sum of Sq...	381.5041	92.56195	

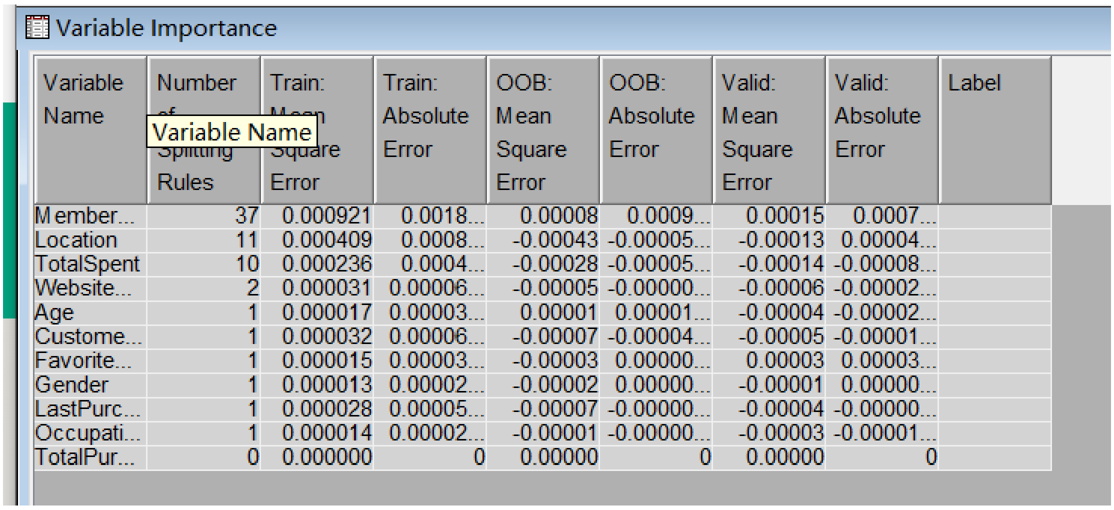
Output

```
*-----------------------------------------------------------*
User:                 28474
Date:                 January 07, 2024
Time:                 11:15:35
*-----------------------------------------------------------*
* Training Output
*-----------------------------------------------------------*




Variable Summary

         Measurement    Frequency
```

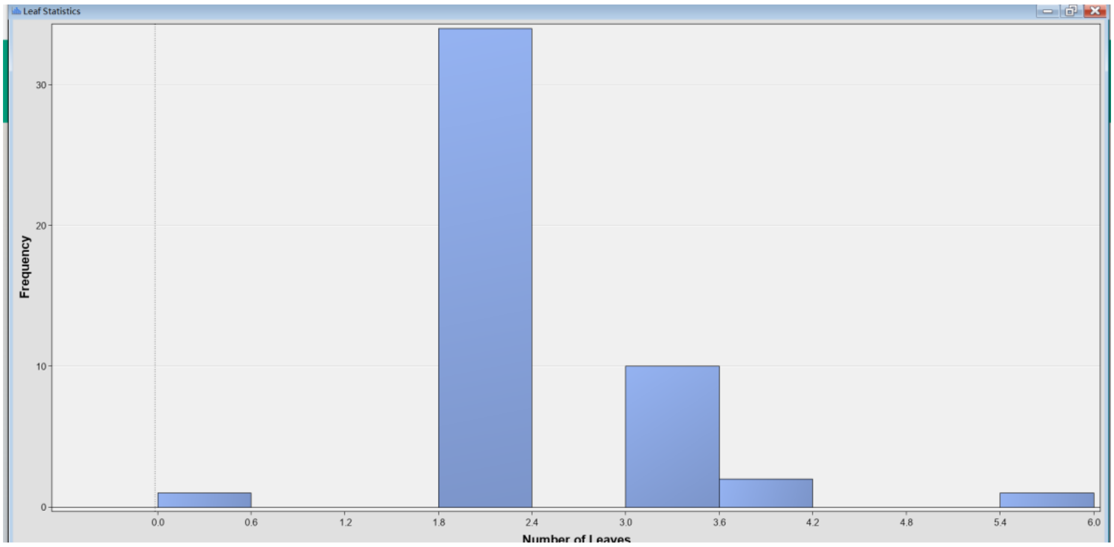
Iteration History

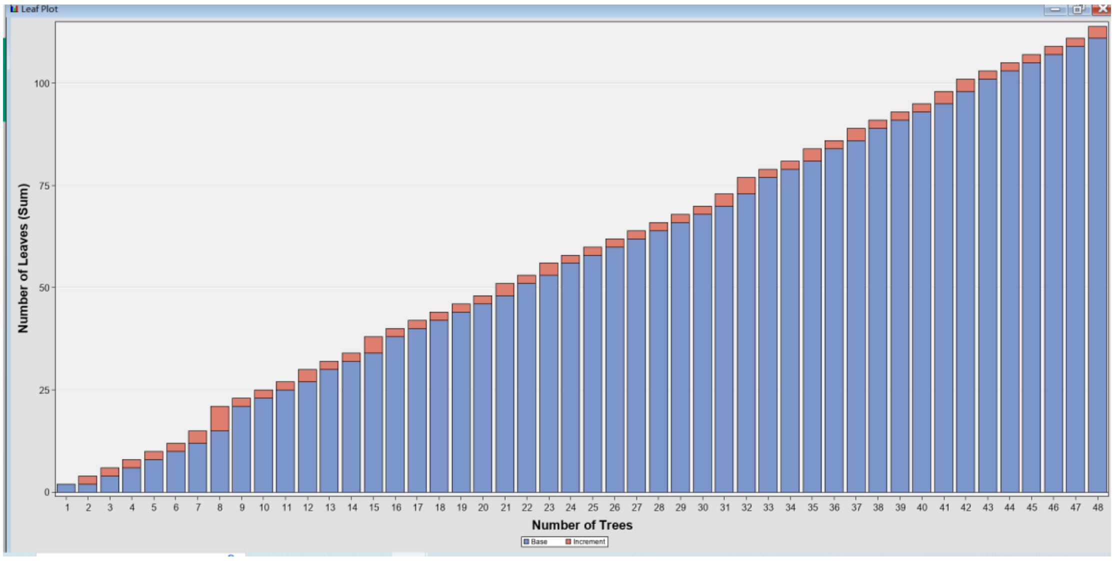
Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)
1	2	0.15990	0.15682	0.15445
2	4	0.15990	0.16133	0.15442
3	6	0.15986	0.16285	0.15433
4	8	0.15978	0.16100	0.15424
5	10	0.15979	0.15814	0.15428
6	12	0.15964	0.16204	0.15445
7	15	0.15929	0.16223	0.15431
8	21	0.15870	0.16169	0.15449
9	23	0.15870	0.16112	0.15443
10	25	0.15876	0.16129	0.15448
11	27	0.15882	0.16099	0.15458
12	30	0.15878	0.16103	0.15452
13	32	0.15884	0.16089	0.15446
14	34	0.15889	0.16079	0.15444
15	38	0.15875	0.16079	0.15435

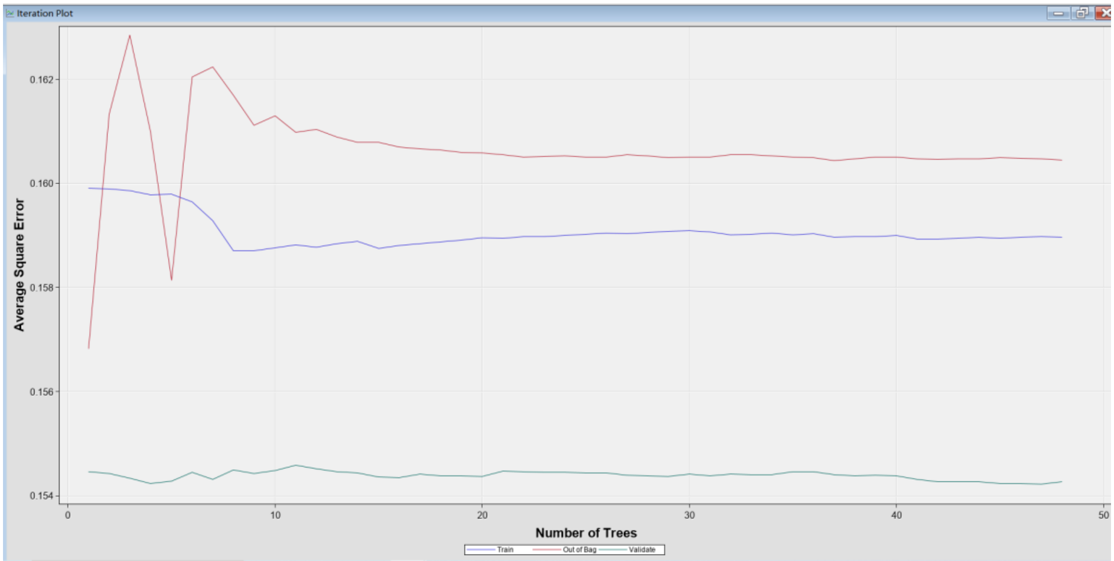


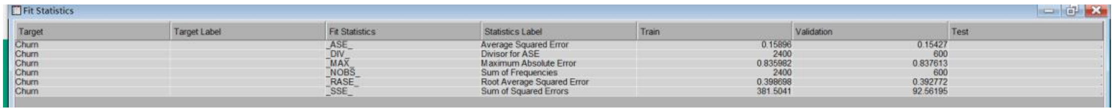

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error	Label
Member...	37	0.000921	0.0018...	0.00008	0.0009...	0.00015	0.0007...	
Location	11	0.000409	0.0008...	-0.00043	-0.00005...	-0.00013	0.00004...	
TotalSpent	10	0.000236	0.0004...	-0.00028	-0.00005...	-0.00014	-0.00008...	
Website...	2	0.000031	0.00006...	-0.00005	-0.00000...	-0.00006	-0.00002...	
Age	1	0.000017	0.00003...	0.00001	0.00001...	-0.00004	-0.00002...	
Custome...	1	0.000032	0.00006...	-0.00007	-0.00004...	-0.00005	-0.00001...	
Favorite...	1	0.000015	0.00003...	-0.00003	0.00000...	0.00003	0.00003...	
Gender	1	0.000013	0.00002...	-0.00002	0.00000...	-0.00001	0.00000...	
LastPurc...	1	0.000028	0.00005...	-0.00007	-0.00000...	-0.00004	-0.00000...	
Occupati...	1	0.000014	0.00002...	-0.00001	-0.00000...	-0.00003	-0.00001...	
TotalPur...	0	0.000000	0	0.00000	0	0.00000	0	

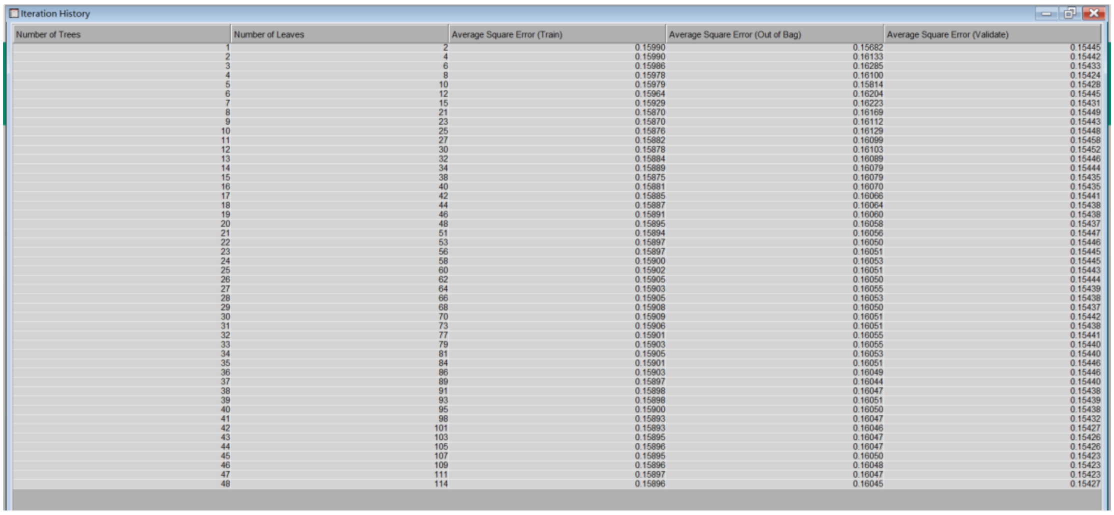







Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_ASE_	Average Squared Error	0.15896	0.15427	
Churn		_DIV_	Divisor for ASE	2400	600	
Churn		_MAX_	Maximum Absolute Error	0.835982	0.837613	
Churn		_NOBS_	Sum of Frequencies	2400	600	
Churn		_RASE_	Root Average Squared Error	0.398698	0.392772	
Churn		_SSE_	Sum of Squared Errors	381.5041	92.56195	



Iteration History

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)
1	2	0.15990	0.15682	0.15445
2	4	0.15990	0.16133	0.15442
3	6	0.15986	0.16285	0.15433
4	8	0.15978	0.16100	0.15424
5	10	0.15979	0.15814	0.15428
6	12	0.15964	0.16204	0.15445
7	15	0.15929	0.16223	0.15431
8	21	0.15870	0.16169	0.15449
9	23	0.15870	0.16112	0.15443
10	25	0.15876	0.16129	0.15448
11	27	0.15882	0.16099	0.15458
12	30	0.15878	0.16103	0.15452
13	32	0.15884	0.16089	0.15446
14	34	0.15889	0.16079	0.15444
15	38	0.15875	0.16079	0.15435
16	40	0.15881	0.16070	0.15435
17	42	0.15885	0.16066	0.15441
18	44	0.15887	0.16064	0.15438
19	46	0.15891	0.16060	0.15438
20	48	0.15895	0.16058	0.15437
21	51	0.15894	0.16056	0.15447
22	53	0.15897	0.16050	0.15446
23	56	0.15897	0.16051	0.15445
24	58	0.15900	0.16053	0.15445
25	60	0.15902	0.16051	0.15443
26	62	0.15905	0.16050	0.15444
27	64	0.15903	0.16055	0.15439
28	66	0.15905	0.16053	0.15438
29	68	0.15908	0.16050	0.15437
30	70	0.15909	0.16051	0.15442
31	73	0.15906	0.16051	0.15438
32	77	0.15901	0.16055	0.15441
33	79	0.15903	0.16055	0.15440
34	81	0.15905	0.16053	0.15440
35	84	0.15901	0.16051	0.15446
36	86	0.15903	0.16049	0.15446
37	89	0.15897	0.16044	0.15440
38	91	0.15898	0.16047	0.15438
39	93	0.15898	0.16051	0.15439
40	95	0.15900	0.16050	0.15438
41	98	0.15893	0.16047	0.15432
42	101	0.15893	0.16046	0.15427
43	103	0.15895	0.16047	0.15426
44	105	0.15896	0.16047	0.15426
45	107	0.15895	0.16050	0.15423
46	109	0.15896	0.16048	0.15423
47	111	0.15897	0.16047	0.15423
48	114	0.15896	0.16045	0.15427

Analysis of results of HPforest model.

Through HPforest's model report, we learned:
The average squared errors (ASE) of the training and validation datasets are very close (0.161 for the training dataset and 0.155 for the validation dataset), indicating that the model has good generalization ability and is not overfitting. Overfitting usually manifests itself as lower training error and higher validation error.
The average squared error for training, out-of-bag (OOB), and validation changes slightly as more trees are added, indicating that the model's performance does not improve significantly after the first few trees. This suggests that a relatively small number of trees are needed to capture patterns in the data.
variable importance
MembershipLevel appears to be the most important variable in predicting churn, contributing the most to reducing mean squared error (MSE). Its high importance in the training and validation datasets indicates that it is a strong predictor of churn.
Other variables, such as Age, TotalPurchases, Occupation, and Gender, also contribute to the model, but to a lesser extent.
fit statistics
Detailed fit statistics show that as the number of trees increases, so does the number of leaves (decision points in each tree). However, the error rate improved only slightly, suggesting that adding trees beyond a certain point does not necessarily improve the model's predictive power.
The errors during training, OOB, and validation are consistent across the number of trees, again indicating that the model is stable and not overfitting.
Assessment score ranking
This report provides the tree's in-depth performance on the training and validation sets. The average predicted values at different depths indicate that the model is consistent across depths and has no signs of overfitting or underfitting.

Customer behavior:

1. Membership Level Importance: The report indicates that 'MembershipLevel' is

highly predictive of churn. This suggests that customers' engagement and satisfaction may vary significantly across different membership tiers, influencing their likelihood to churn.

2. Location Influence: 'Location' also plays a role in predicting churn but seems to have a negative impact on the out-of-bag and validation MSE in the variable importance analysis. This could indicate that customers in certain locations are more prone to churn, or there may be regional factors at play affecting customer retention.

4. Other Variables: While 'Age', 'TotalPurchases', 'Occupation', 'Gender', 'FavoriteCategory', 'WebsiteVisitFrequency', 'LastPurchaseDate', and 'TotalSpent' have lesser importance compared to 'MembershipLevel', they still contribute to churn prediction. This suggests that demographics, purchasing behavior, and engagement with the company's online presence are relevant factors in customer churn.

Business Strategy Recommendations:

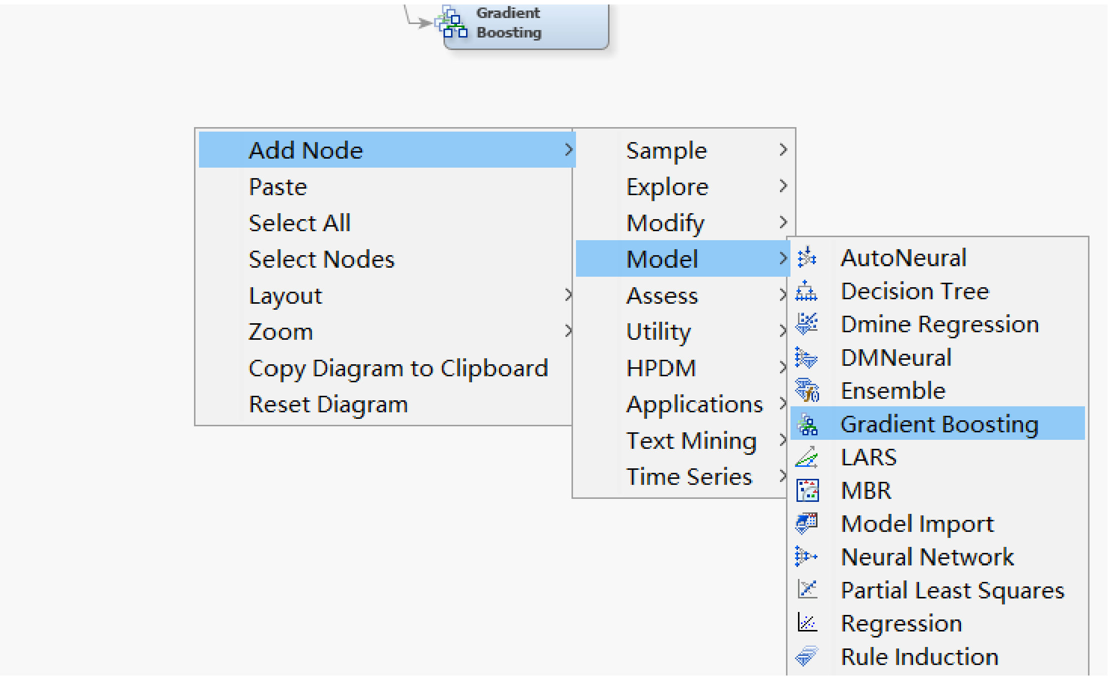
1. Tier-based Engagement Programs: Develop retention programs tailored to different membership levels, especially focusing on the tiers that contribute most to churn. This could include personalized communication, special offers, or loyalty rewards.
2. Location-specific Initiatives: Since location seems to affect churn, consider regional marketing initiatives that address the specific needs or pain points of customers in high-churn areas.
3. Customer Lifecycle Value Enhancement: For variables with less importance but still contributing to churn prediction, design initiatives to enhance the overall customer experience, such as personalized product recommendations, improved customer service, or loyalty programs.
4. Targeted Communication: Use the insights from the model to craft targeted messages for customer segments identified as at risk. This could involve outreach campaigns that address specific behaviors leading to churn.
5. Improve Online Experience: Given the inclusion of 'WebsiteVisitFrequency' and 'LastPurchaseDate', improving the online experience and simplifying the purchase process could be beneficial.
6. Data-Driven Product and Service Development: Utilize the insights from 'TotalPurchases' and 'FavoriteCategory' to refine product offerings and services that are better aligned with customer preferences.
7. Cross-Selling and Up-Selling Opportunities: Leverage information on 'TotalSpent' to identify customers who could be interested in premium offerings or additional services, using up-selling or cross-selling tactics.
8. Retention Analytics: Continuously monitor and analyze customer behavior using the model's insights to quickly identify any shifts in behavior that could signal an increased risk of churn.
9. Model Monitoring: Since the random forest model used seems to generalize well, it's

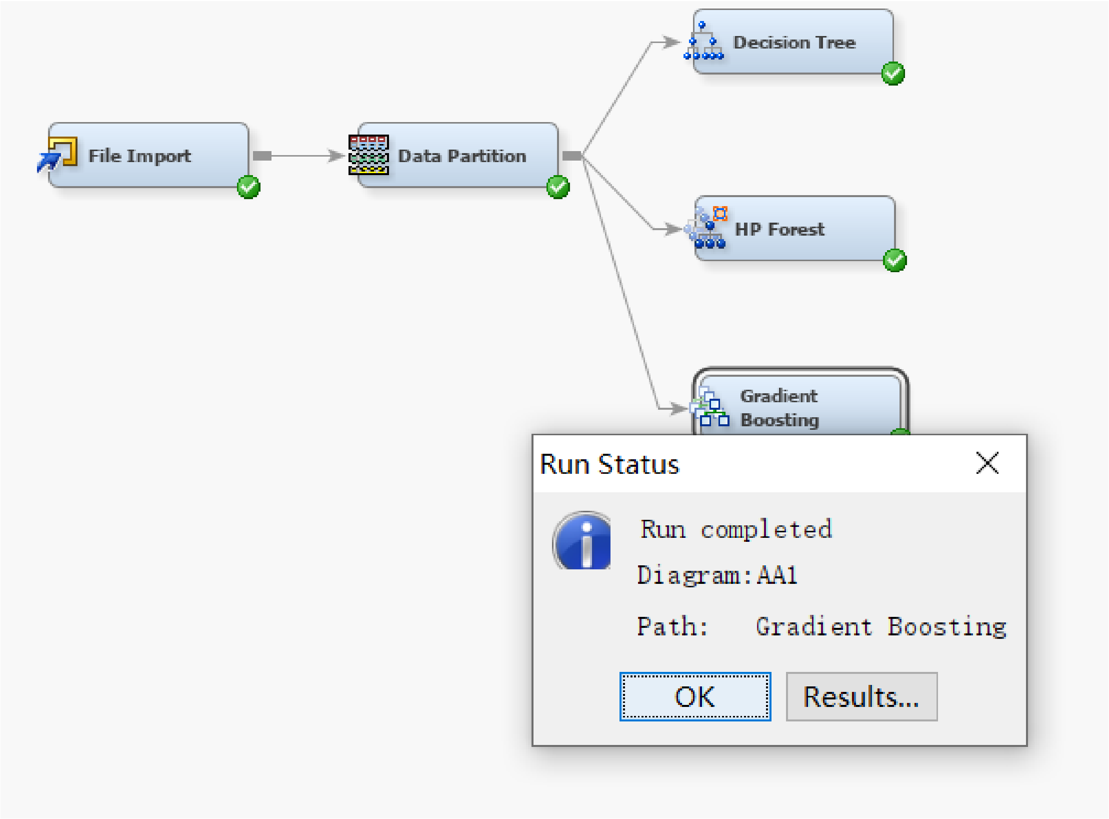
important to continuously monitor its performance over time with new data to ensure its predictions remain accurate.

Implementing these strategies can help to reduce churn rates, increase customer satisfaction, and ultimately drive higher customer lifetime value.

Gradient boosting

For Gradient boosting, we continue the previous method, grab the interface from the model, and then run it to test the performance of the model.





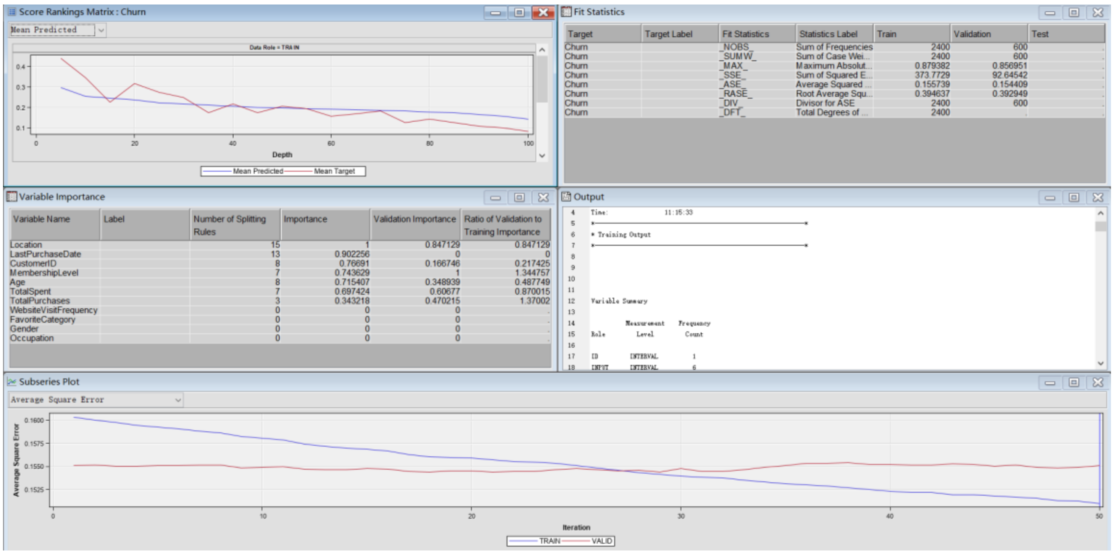
Run Status

Run completed

Diagram:AA1

Path: Gradient Boosting

The following is the result report of running the test:



Fit Statistics

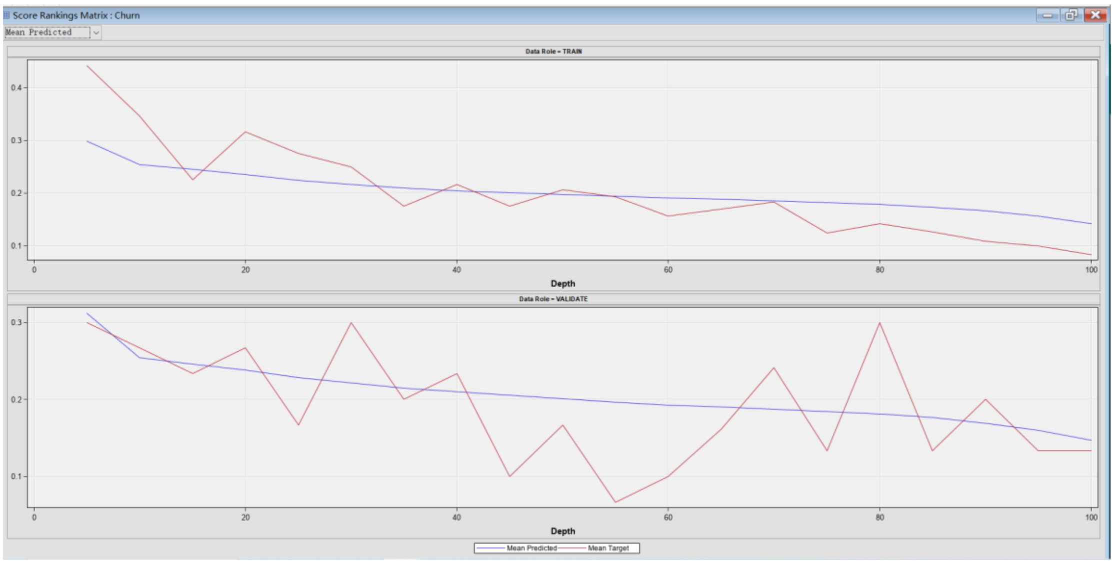
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Frequencies	2400	600	.
Churn		_SUMW_	Sum of Case Wei...	2400	600	.
Churn		_MAX_	Maximum Absolut...	0.879382	0.856951	.
Churn		_SSE_	Sum of Squared E...	373.7729	92.64542	.
Churn		_ASE_	Average Squared ...	0.155739	0.154409	.
Churn		_RASE_	Root Average Squ...	0.394637	0.392949	.
Churn		_DIV_	Divisor for ASE	2400	600	.
Churn		_DFT_	Total Degrees of ...	2400	.	.

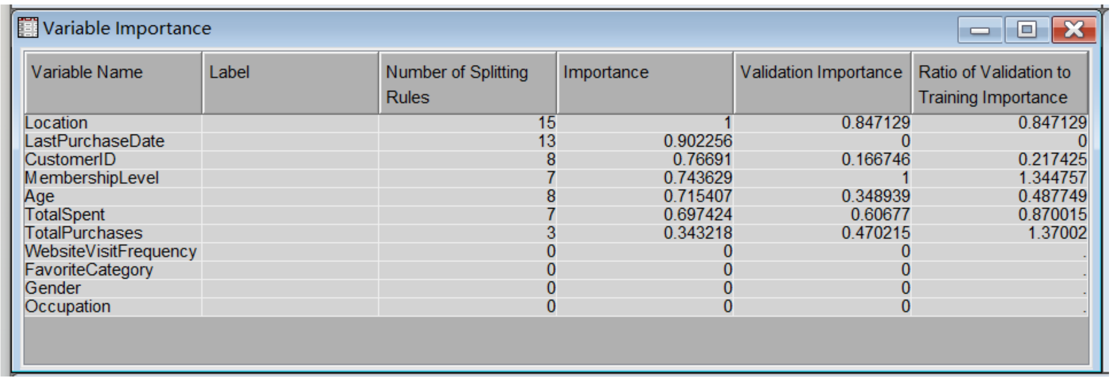
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Location		15	1	0.847129	0.847129
LastPurchaseDate		13	0.902256	0	0
CustomerID		8	0.76691	0.166746	0.217425
MembershipLevel		7	0.743629	1	1.344757
Age		8	0.715407	0.348939	0.487749
TotalSpent		7	0.697424	0.60677	0.870015
TotalPurchases		3	0.343218	0.470215	1.37002
WebsiteVisitFrequency		0	0	0	.
FavoriteCategory		0	0	0	.
Gender		0	0	0	.
Occupation		0	0	0	.

Output

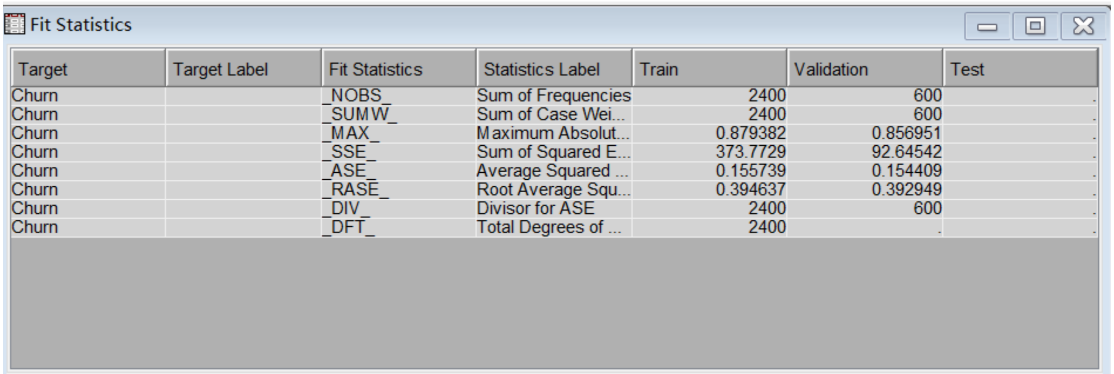
```
Time:              11:15:33
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level          Count

ID          INTERVAL            1
INPUT       INTERVAL            6
```





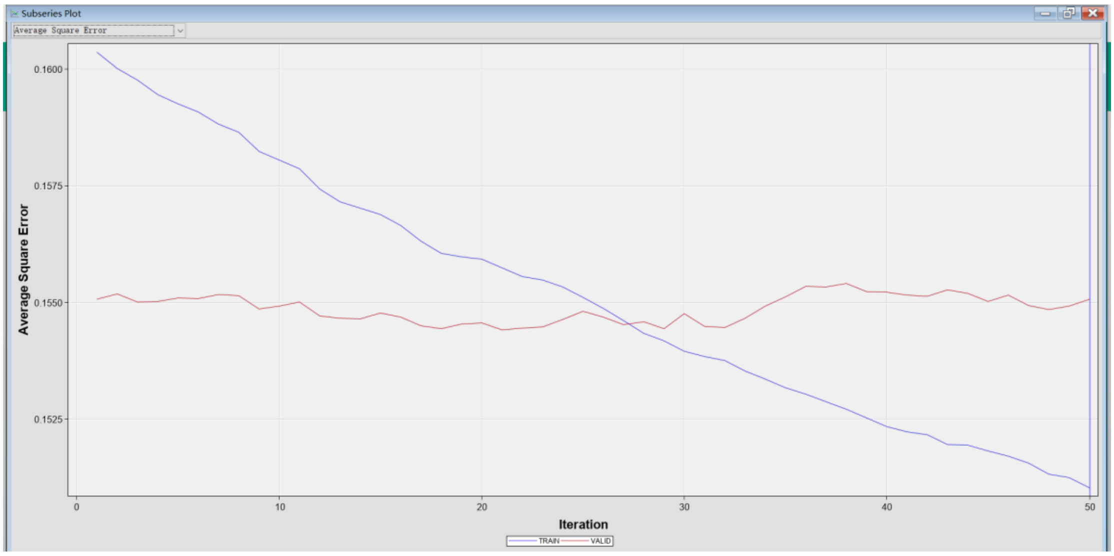
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Location		15	1	0.847129	0.847129
LastPurchaseDate		13	0.902256	0	0
CustomerID		8	0.76691	0.166746	0.217425
MembershipLevel		7	0.743629	1	1.344757
Age		8	0.715407	0.348939	0.487749
TotalSpent		7	0.697424	0.60677	0.870015
TotalPurchases		3	0.343218	0.470215	1.37002
WebsiteVisitFrequency		0	0	0	.
FavoriteCategory		0	0	0	.
Gender		0	0	0	.
Occupation		0	0	0	.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Frequencies	2400	600	.
Churn		_SUMW_	Sum of Case Wei...	2400	600	.
Churn		_MAX_	Maximum Absolut...	0.879382	0.856951	.
Churn		_SSE_	Sum of Squared E...	373.7729	92.64542	.
Churn		_ASE_	Average Squared ...	0.155739	0.154409	.
Churn		_RASE_	Root Average Squ...	0.394637	0.392949	.
Churn		_DIV_	Divisor for ASE	2400	600	.
Churn		_DFT_	Total Degrees of ...	2400	.	.



The provided report details the outcomes of a gradient boosting model focused on predicting customer churn. Here is an analysis of the report's results:

Variable Importance:

- **Location** is the most critical predictor of churn, with the highest importance score and ratio. This suggests that customer churn is significantly influenced by geographic factors.
- **LastPurchaseDate** has a high importance score, but a ratio of 0.00000 in the validation importance, indicating it may not be as predictive on unseen data.
- **CustomerID** has a moderate level of importance, which is unusual as IDs are generally not predictive. This could indicate overfitting or data leakage.
- **MembershipLevel** also shows a high importance, suggesting that the type of membership a customer has is predictive of churn likelihood.
- **Age** and **TotalSpent** are also relevant, indicating that demographic factors and spending behavior are associated with churn risk.
- **TotalPurchases** has a lower importance score but a high validation ratio, suggesting it may become a more relevant predictor in the validation dataset.

Fit Statistics:

- The model has been trained on 2400 observations and validated on 600 observations.
- The maximum absolute error and average squared error are very close between the training and validation datasets, which suggests that the model generalizes well and is not overfitting.
- The root average squared error (RASE) is consistent between the training and validation datasets, further supporting the model's generalizability.

Assessment Score Rankings:

- The mean predicted churn decreases as the number of observations increases, suggesting the model becomes more certain about customers not churning as it has more data to learn from.
- There is a notable decrease in the mean predicted churn from the 5th to the 100th

observation in both the training and validation datasets, which could indicate that the model is effectively capturing the underlying patterns associated with churn.

Assessment Score Distribution:

- The score distribution shows a wide range of mean predicted values, especially in the training dataset, which suggests that the model is identifying different levels of churn risk across the customer base.
- The validation score distribution follows a similar pattern, though with fewer observations in each predicted range, which is expected given the smaller size of the validation dataset.

In summary, the gradient boosting model seems to be performing well, with key variables identified that influence the likelihood of churn. The close fit between training and validation datasets suggests that the model is robust and not overfitting, and the variable importance rankings provide clear indicators of which features are most predictive of churn.

Customer Behavior Insights:

1. **Geographic Trends**: Location is highly indicative of churn likelihood, which suggests geographic or regional factors significantly impact customer retention. This could be due to market saturation, local competition, or economic conditions.
2. **Engagement Levels**: Membership Level's importance indicates that customers' engagement, as defined by their membership status, plays a crucial role in their decision to stay with or leave the company. Higher or more premium membership levels might correlate with lower churn.
3. **Recency of Interaction**: The LastPurchaseDate's initial high importance score may reflect that the more recently a customer has engaged with the company's services or products, the less likely they are to churn.
4. **Customer Lifetime Value**: TotalSpent's importance signifies that customers who have spent more over time are key to focus on for retention strategies as their churn could represent a higher loss of revenue.
5. **Activity Level**: TotalPurchases, despite its lower overall importance score, is more predictive in the validation set, which might indicate that the frequency of transactions is a relevant factor in understanding churn risk.

Business Strategy Recommendations:

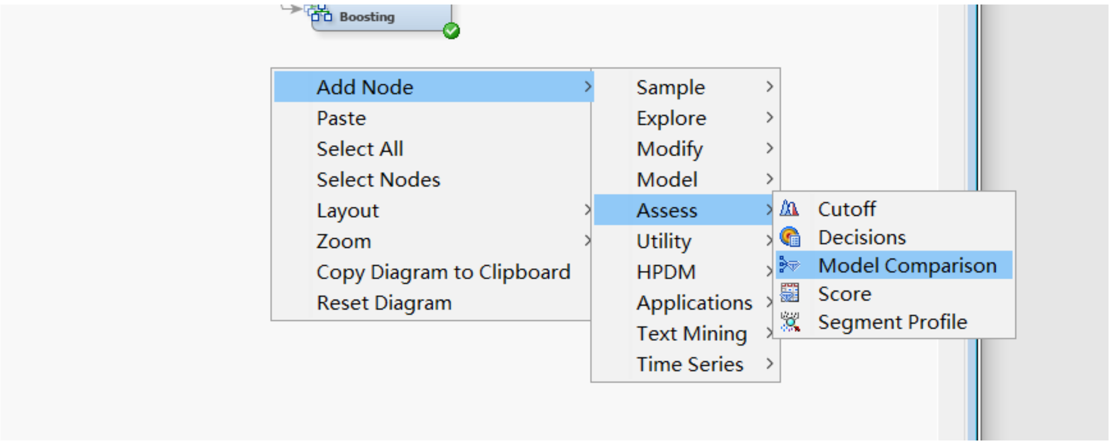
1. **Regional Marketing Initiatives**: Develop targeted marketing strategies that address the unique needs and competition in high-churn locations.
2. **Loyalty Programs**: Enhance loyalty programs to encourage higher engagement and transition customers to higher membership levels, which are less likely to churn.
3. **Customer Engagement**: Implement re-engagement campaigns targeting customers who have not made recent purchases to reduce churn.
4. **Customer Value Maximization**: Identify high-value customers based on their total spend and devise retention strategies tailored to them, such as exclusive

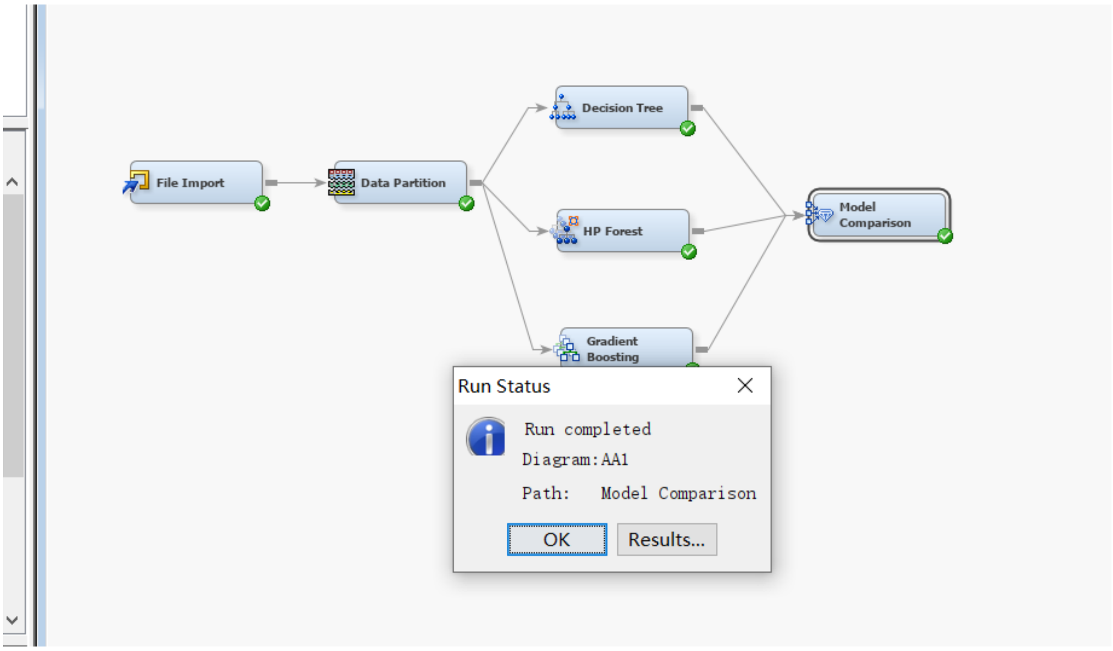
offers or personalized services.

5. **Transaction-Based Engagement**: For customers with fewer purchases, consider strategies to increase their purchase frequency, such as offering discounts on subsequent purchases or rewards for frequent shopping.
6. **Data-Driven Product Development**: Use insights from the model to inform product development and service offerings that cater to the needs of different customer segments, particularly focusing on those in high-risk churn groups.
7. **Continual Monitoring**: Regularly update and monitor the model to ensure retention strategies remain effective and adjust to new patterns as customer behavior evolves.
8. **Customer Segmentation**: Use the model's predictions to segment customers based on churn risk and tailor communication, offers, and interventions accordingly.

By acting on these insights, a business can proactively manage customer churn, tailor its marketing efforts, and ultimately improve customer loyalty and retention.

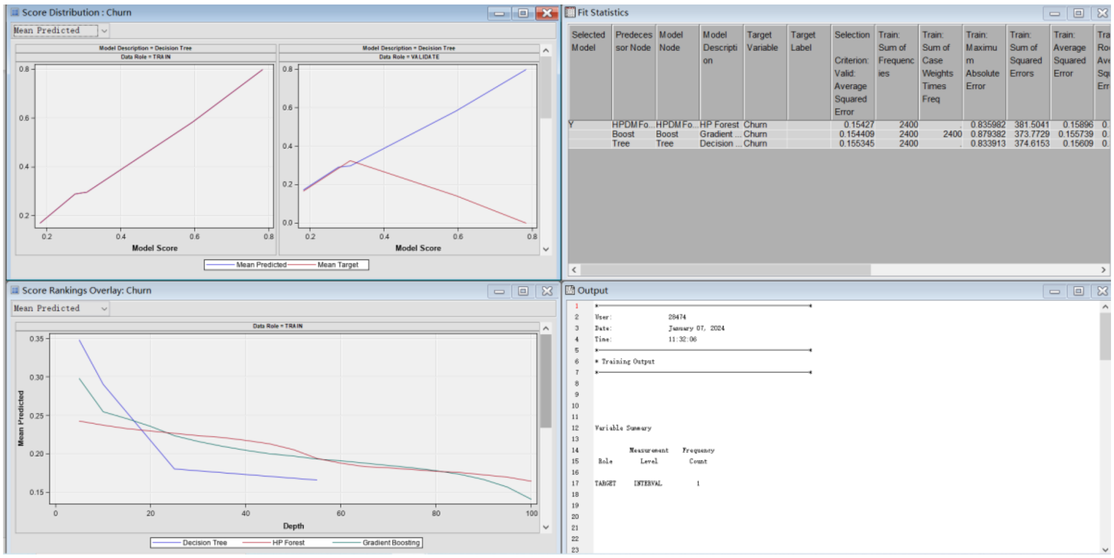
Compare these three models





Run Status

Run completed
Diagram: AA1
Path: Model Comparison



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Sum of Case Weights Times Freq	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Tra Ro Av Sq Err
Y	HPDMFo...	HPDMFo...	HP Forest	Churn		0.15427	2400	.	0.835982	381.5041	0.15896	0.
	Boost	Boost	Gradient ...	Churn		0.154409	2400	2400	0.879382	373.7729	0.155739	0.
	Tree	Tree	Decision ...	Churn		0.155345	2400	.	0.833913	374.6153	0.15609	0.

```
*------------------------------------------------------------*
User:                 28474
Date:                 January 07, 2024
Time:                 11:32:06
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

           Measurement    Frequency
 Role         Level         Count

TARGET     INTERVAL           1
```

The provided report compares three different models: HP Forest (random forest), Gradient Boosting, and Decision Tree, based on their performance statistics.

Model Comparison:

Average Squared Error (ASE):

- HP Forest: Shows slightly higher ASE on the training set (0.15896) compared to its validation set (0.15427), suggesting the model may be slightly overfitting.
- Gradient Boosting: Has a lower ASE on the training set (0.15574) than HP Forest and a very similar ASE on the validation set (0.15441). This indicates that the model is generalizing well.
- Decision Tree: Presents the highest ASE on both the training (0.15609) and validation sets (0.15534) among the three models, which may indicate it's the least complex model and potentially underfitting compared to the other models.

Maximum Absolute Error:

- HP Forest: Has the lowest maximum absolute error on the training set (0.84), but a slightly higher error on the validation set (0.838) than the Decision Tree.

- Gradient Boosting: Exhibits the highest maximum absolute error on the training set (0.88) but is still competitive on the validation set (0.857).
- Decision Tree: Shows a consistent maximum absolute error between the training (0.83) and validation sets (0.834), which is desirable.

Root Average Squared Error (RASE):

- All three models have similar RASE on the training set, around 0.39-0.40, indicating that their overall performance is quite close.
- For the validation set, RASE is also similar across all models, hovering around 0.393-0.394.

Sum of Squared Errors (SSE):

- HP Forest: Has SSE of 381.50 on the training set and 92.562 on the validation set.
- Gradient Boosting: Is slightly better with an SSE of 373.77 on the training set and 92.645 on the validation set.
- Decision Tree: Has an SSE of 374.62 on the training set, which is close to Gradient Boosting, but the highest SSE on the validation set (93.207).

Analysis:

- HP Forest and Gradient Boosting are performing similarly, with Gradient Boosting having a slight edge in terms of generalization as indicated by its lower training ASE. The random forest model is slightly overfitting, which is evident from the higher training ASE compared to validation ASE.
- The Decision Tree is the simplest model with the least ability to generalize, as indicated by its highest ASE on both the training and validation sets.

Model Selection:

- The model selection based on Valid ASE is indicating Gradient Boosting as the selected model (indicated by the 'Y'), probably due to its balance between training and validation error, showing good generalization without significant overfitting.
- While the Decision Tree model is simpler and may be faster to train and score, its slightly worse performance metrics suggest it may not capture the complexities of the data as well as the other two models.
- The HP Forest model, while slightly overfitting, may still be desirable in scenarios where the ensemble approach of multiple decision trees could provide more robust predictions, especially if the model's complexity can be tuned to reduce overfitting.

In conclusion, Gradient Boosting seems to be the preferred model out of the three due to its performance balance, followed by HP Forest and then the Decision Tree. These insights would be used to choose the best model for deployment in a production environment, considering both performance and computational efficiency.

The following are key reflections and learning outcomes:

1. Dataset Creation and Integration: I used artificial intelligence and personal modification to build a dataset that truly reflected e-commerce customer behavior. This involves using tools such as Talend Data Integration and Talend Data Preparation to integrate and clean the data.
2. Selection of analysis methods: I used decision trees, random forests (HP Forest), and gradient boosting models in SAS Enterprise Miner. These methods help extract meaningful insights from complex data sets.
3. Model evaluation and optimization: The article discusses how to identify and deal with overfitting problems, such as simplifying the model through pruning, using ensemble methods, and regularization. This demonstrates an understanding of the data and continued focus on model performance during model building.
4. Insights into customer behavior: Through the model, I was able to identify key factors that affect customer churn, such as membership level, geographical location, consumption behavior, etc. These insights help develop targeted business strategies.

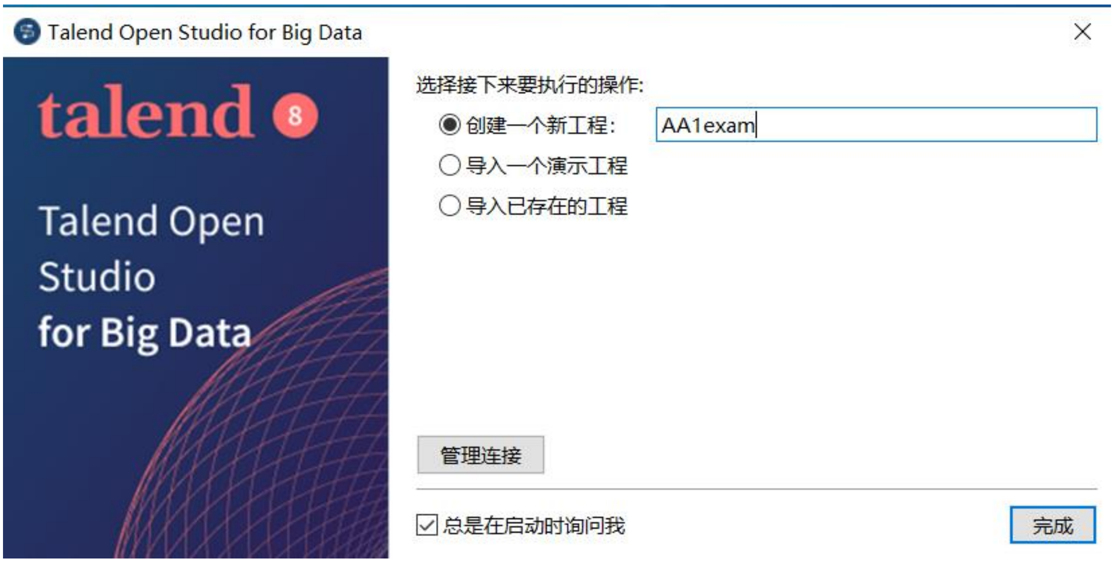
During this project I faced several major challenges and took some key steps to overcome them:

1. Data quality and integration issues: In an e-commerce environment, data is often scattered across different systems and formats. To solve this problem, I use data integration tools (such as Talend Data Integration) to unify data from different sources into a formatted dataset. Additionally, by using Talend Data Preparation, you can clean and transform your data, ensuring its quality and consistency.
2. Model selection and optimization: In the data mining process, selecting an appropriate algorithm and optimizing it is a challenge. I address this challenge by comparing the performance of decision tree, random forest, and gradient boosting models. Using SAS Enterprise Miner, I am able to test different algorithms and choose the one that works best for my data.
3. Overfitting problem: Overfitting is a common problem when building a prediction model. I combat this problem by using cross-validation and regularization techniques during model building. Additionally, I prune the model appropriately to avoid excessive complexity.
4. Understand and interpret model results: A key aspect of data mining is being able to interpret model results. I successfully explained the factors that influence customer behavior through in-depth analysis of the model output and the importance of key variables.

By working through these challenges, I not only improved my data processing and analysis skills, but also enhanced my understanding of complex data patterns. These experiences provide you with a solid foundation for future work in data science.

Appendix

Talend DI



Talend Open Studio for Big Data

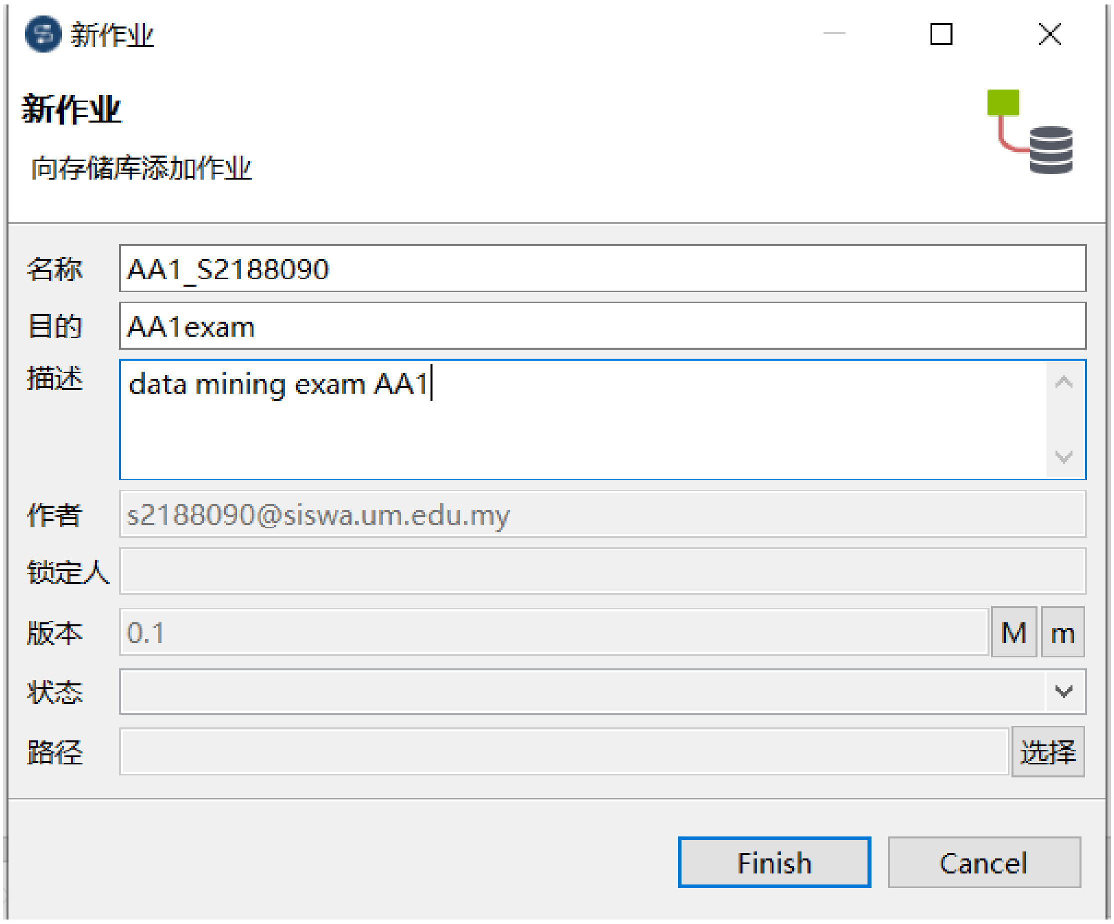
选择接下来要执行的操作:

● 创建一个新工程: AA1exam

○ 导入一个演示工程

○ 导入已存在的工程

☑ 总是在启动时询问我

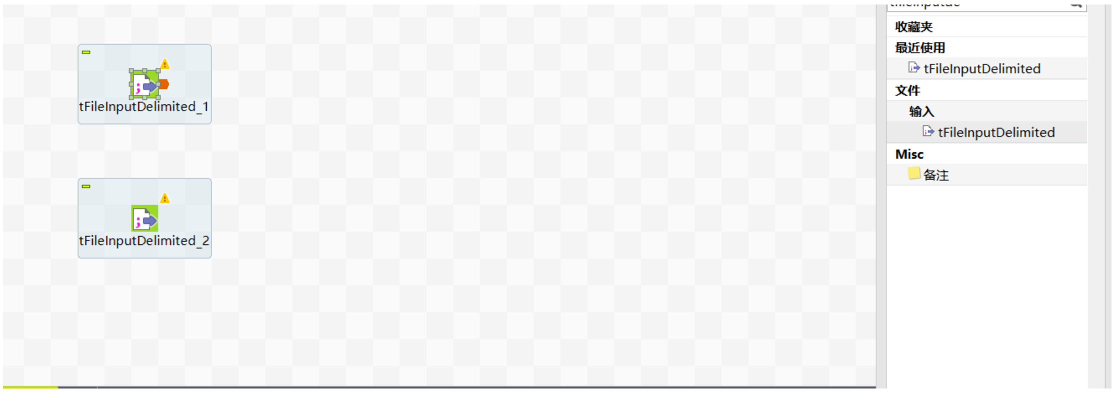


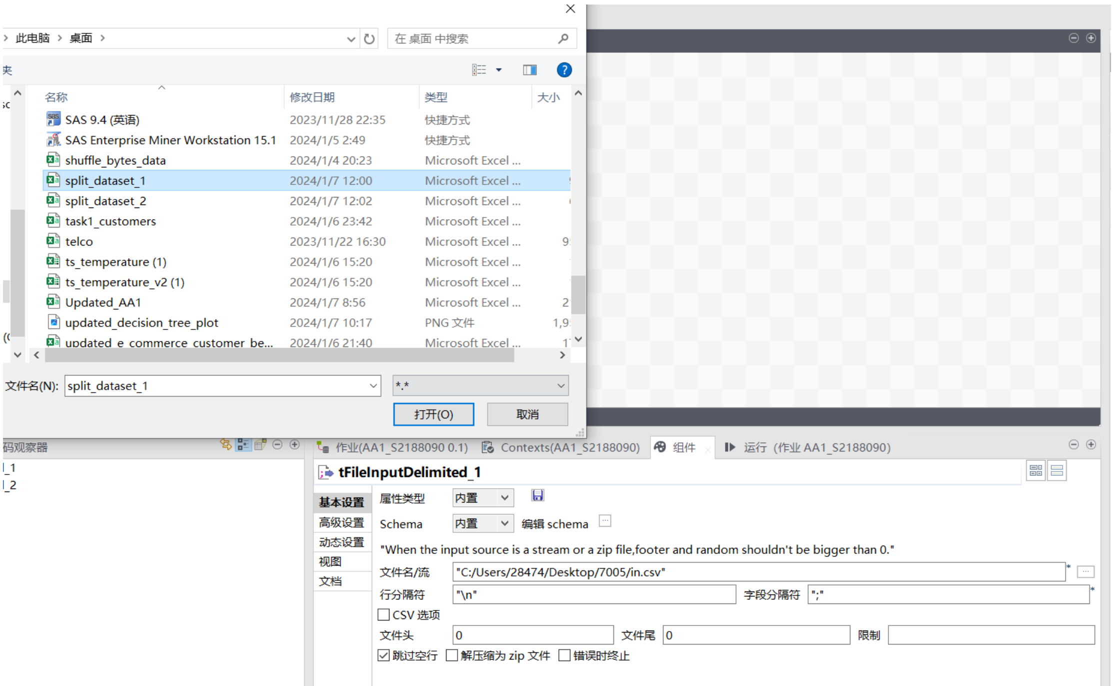
新作业

新作业

向存储库添加作业

名称	AA1_S2188090
目的	AA1exam
描述	data mining exam AA1
作者	s2188090@siswa.um.edu.my
锁定人	
版本	0.1
状态	
路径	





名称	修改日期	类型	大小
SAS 9.4 (英语)	2023/11/28 22:35	快捷方式	
SAS Enterprise Miner Workstation 15.1	2024/1/5 2:49	快捷方式	
shuffle_bytes_data	2024/1/4 20:23	Microsoft Excel ...	
split_dataset_1	2024/1/7 12:00	Microsoft Excel ...	
split_dataset_2	2024/1/7 12:02	Microsoft Excel ...	
task1_customers	2024/1/6 23:42	Microsoft Excel ...	
telco	2023/11/22 16:30	Microsoft Excel ...	9
ts_temperature (1)	2024/1/6 15:20	Microsoft Excel ...	
ts_temperature_v2 (1)	2024/1/6 15:20	Microsoft Excel ...	
Updated_AA1	2024/1/7 8:56	Microsoft Excel ...	2
updated_decision_tree_plot	2024/1/7 10:17	PNG 文件	1,9
updated_e_commerce_customer_be...	2024/1/6 21:40	Microsoft Excel ...	1

文件名(N): split_dataset_1　　*.*

tFileInputDelimited_1

属性类型 内置

Schema 内置　编辑 schema

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

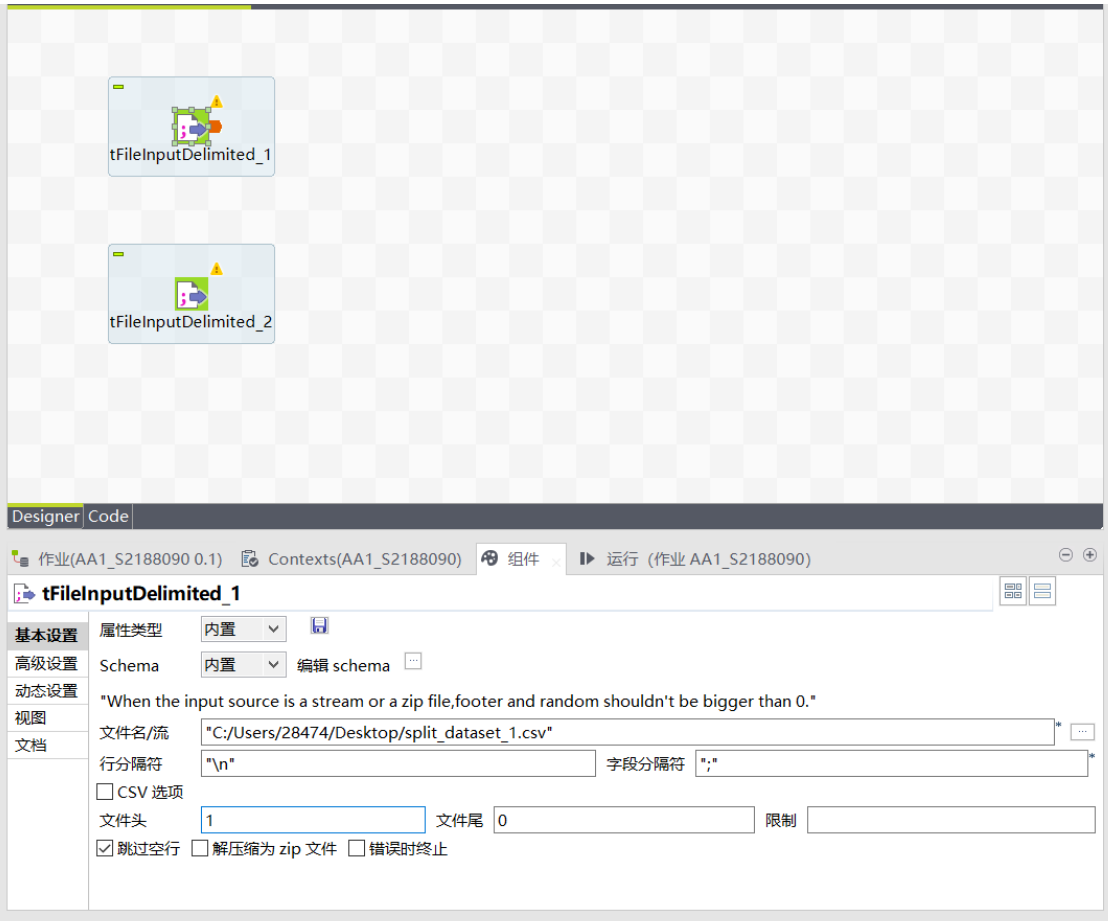
文件名/流 "C:/Users/28474/Desktop/7005/in.csv"

行分隔符 "\n"　字段分隔符 ";"

□ CSV 选项

文件头 0　文件尾 0　限制

☑ 跳过空行 □ 解压缩为 zip 文件 □ 错误时终止



tFileInputDelimited_1

属性类型 内置

Schema 内置 编辑 schema

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

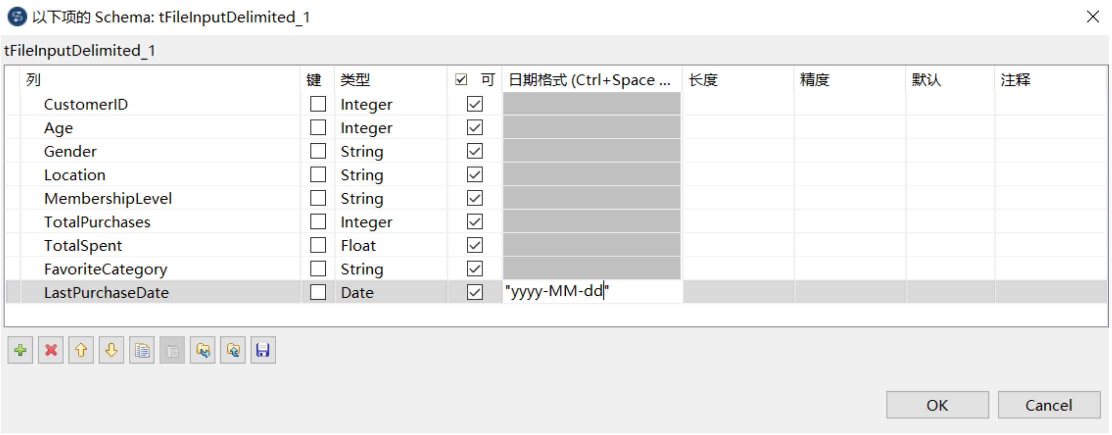
文件名/流 "C:/Users/28474/Desktop/split_dataset_1.csv"

行分隔符 "\n" 字段分隔符 ";"

□ CSV 选项

文件头 1 文件尾 0 限制

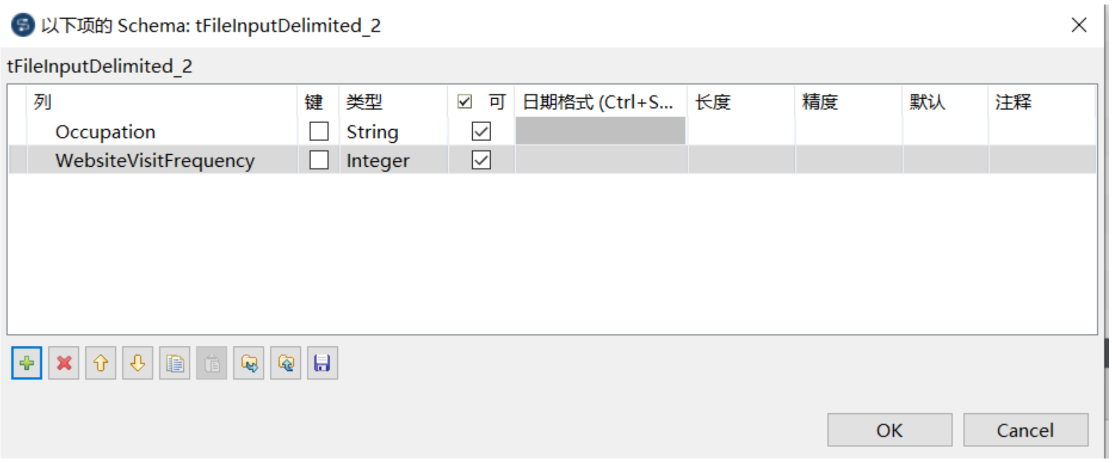
☑ 跳过空行 □ 解压缩为 zip 文件 □ 错误时终止



以下项的 Schema: tFileInputDelimited_1

tFileInputDelimited_1

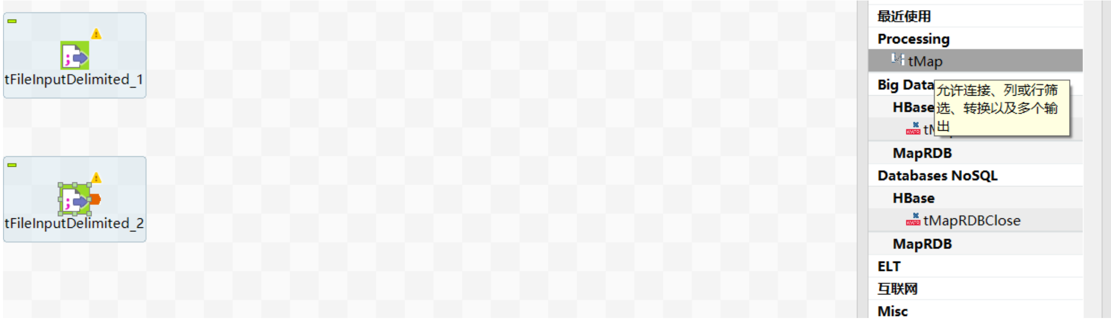
列	键	类型	☑ 可	日期格式 (Ctrl+Space ...	长度	精度	默认	注释
CustomerID	□	Integer	☑					
Age	□	Integer	☑					
Gender	□	String	☑					
Location	□	String	☑					
MembershipLevel	□	String	☑					
TotalPurchases	□	Integer	☑					
TotalSpent	□	Float	☑					
FavoriteCategory	□	String	☑					
LastPurchaseDate	□	Date	☑	"yyyy-MM-dd"				



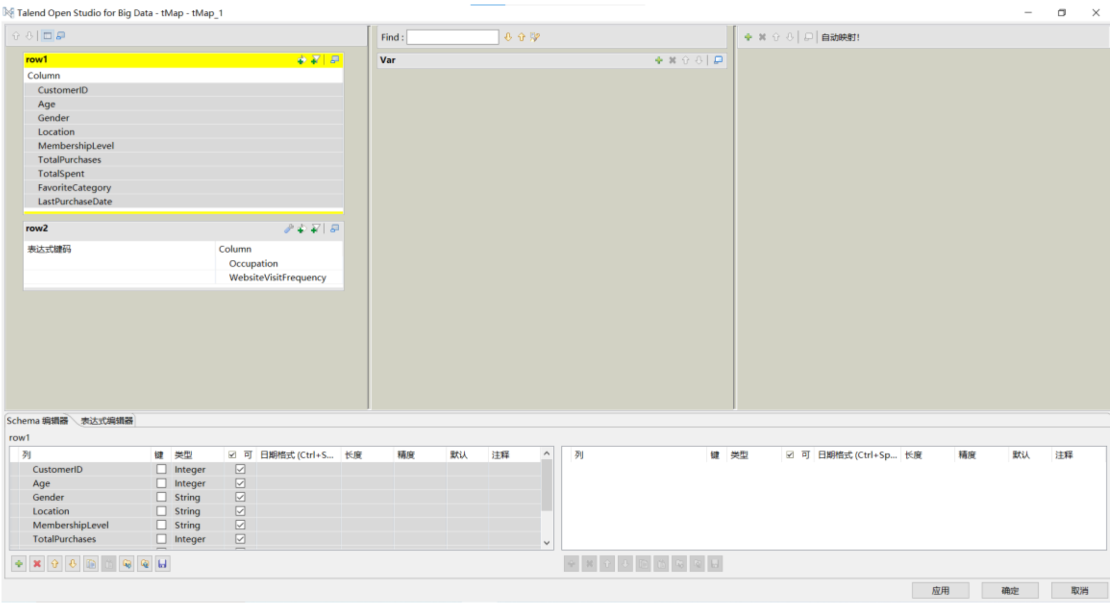
以下项的 Schema: tFileInputDelimited_2

tFileInputDelimited_2

列	键	类型	☑ 可	日期格式 (Ctrl+S...	长度	精度	默认	注释
Occupation	□	String	☑					
WebsiteVisitFrequency	□	Integer	☑					



允许连接、列或行筛选、转换以及多个输出



Talend Open Studio for Big Data - tMap - tMap_1

row1

Column
- CustomerID
- Age
- Gender
- Location
- MembershipLevel
- TotalPurchases
- TotalSpent
- FavoriteCategory
- LastPurchaseDate

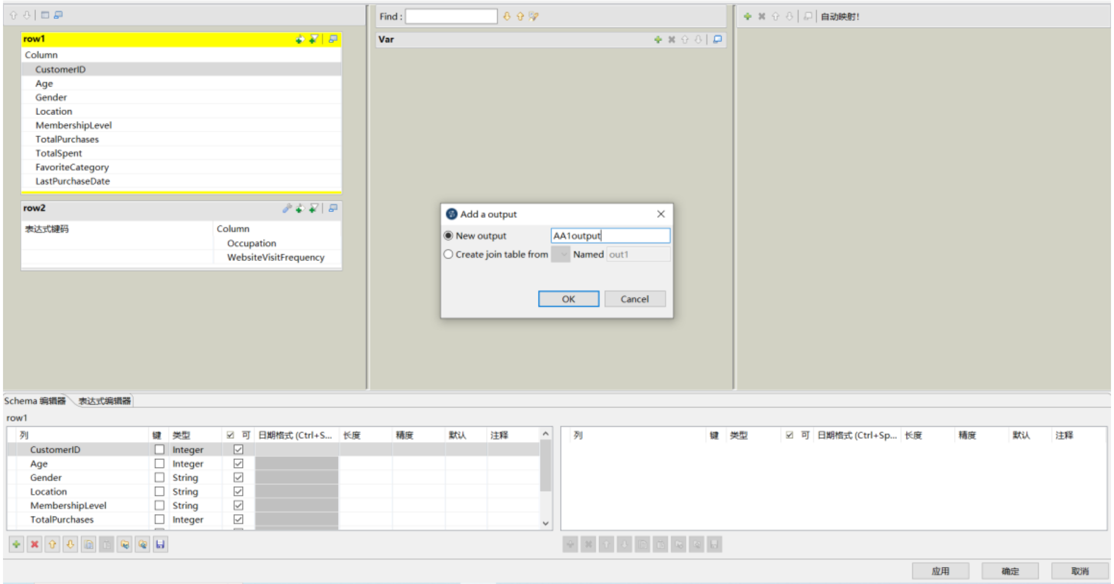
row2

表达式键码	Column
	Occupation
	WebsiteVisitFrequency

row1

列	键	类型	☑ 可	日期格式 (Ctrl+S...	长度	精度	默认	注释
CustomerID	□	Integer	☑					
Age	□	Integer	☑					
Gender	□	String	☑					
Location	□	String	☑					
MembershipLevel	□	String	☑					
TotalPurchases	□	Integer	☑					

列	键	类型	☑ 可	日期格式 (Ctrl+Sp...	长度	精度	默认	注释



row1

Column
- CustomerID
- Age
- Gender
- Location
- MembershipLevel
- TotalPurchases
- TotalSpent
- FavoriteCategory
- LastPurchaseDate

row2

表达式键码	Column
	Occupation
	WebsiteVisitFrequency

Add a output

● New output AA1output

○ Create join table from Named out1

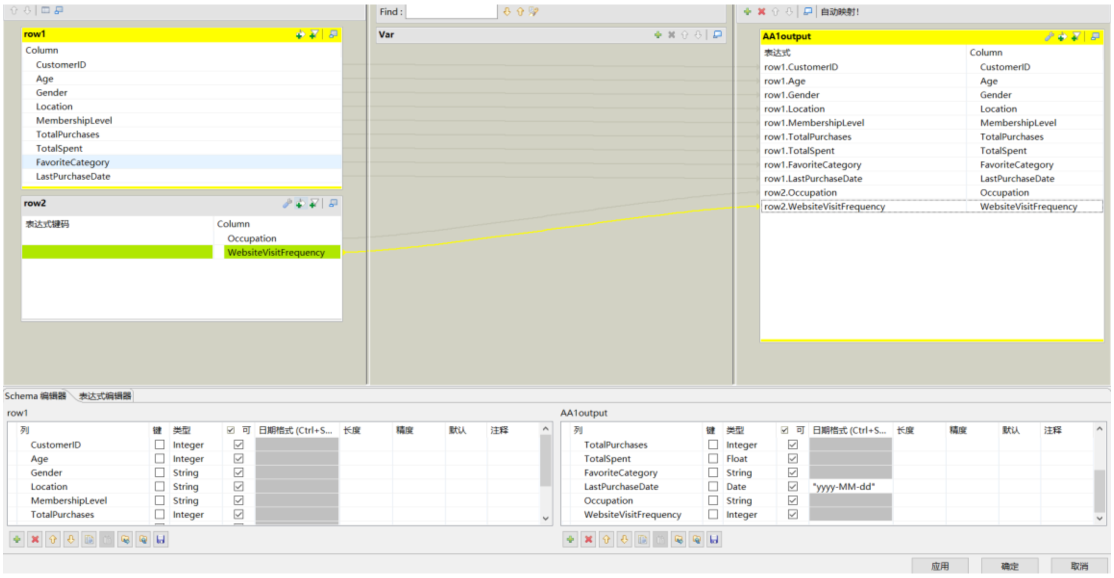
OK Cancel

Schema 编辑器 表达式编辑器

row1

列	键	类型	☑ 可	日期格式 (Ctrl+S...	长度	精度	默认	注释
CustomerID	□	Integer	☑					
Age	□	Integer	☑					
Gender	□	String	☑					
Location	□	String	☑					
MembershipLevel	□	String	☑					
TotalPurchases	□	Integer	☑					

列	键	类型	☑ 可	日期格式 (Ctrl+Sp...	长度	精度	默认	注释



row1

Column
- CustomerID
- Age
- Gender
- Location
- MembershipLevel
- TotalPurchases
- TotalSpent
- FavoriteCategory
- LastPurchaseDate

row2

表达式键码	Column
	Occupation
	WebsiteVisitFrequency

AA1output

表达式	Column
row1.CustomerID	CustomerID
row1.Age	Age
row1.Gender	Gender
row1.Location	Location
row1.MembershipLevel	MembershipLevel
row1.TotalPurchases	TotalPurchases
row1.TotalSpent	TotalSpent
row1.FavoriteCategory	FavoriteCategory
row1.LastPurchaseDate	LastPurchaseDate
row2.Occupation	Occupation
row2.WebsiteVisitFrequency	WebsiteVisitFrequency

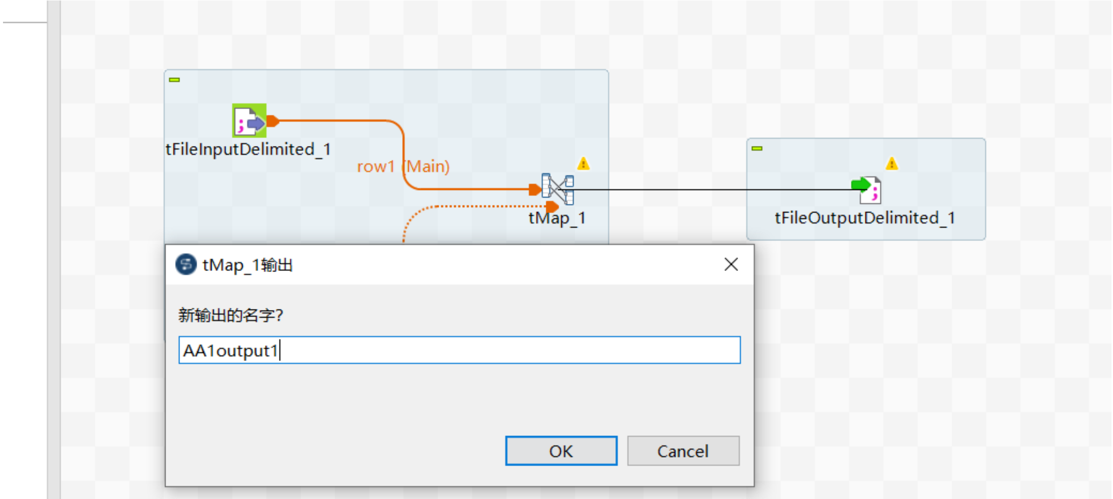
Schema 编辑器 表达式编辑器

row1

列	键	类型	☑ 可	日期格式 (Ctrl+S...	长度	精度	默认	注释
CustomerID	□	Integer	☑					
Age	□	Integer	☑					
Gender	□	String	☑					
Location	□	String	☑					
MembershipLevel	□	String	☑					
TotalPurchases	□	Integer	☑					

AA1output

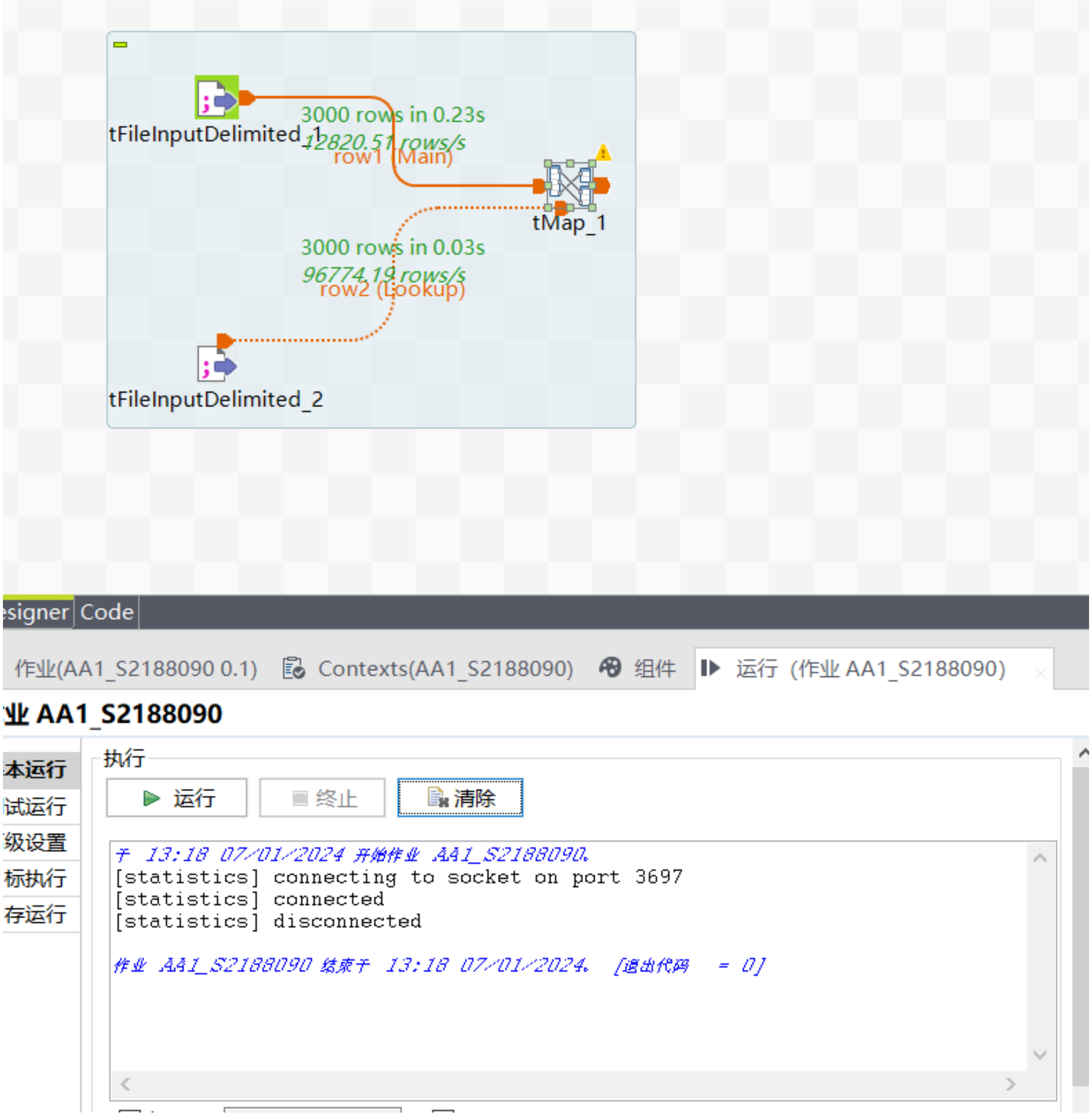
列	键	类型	☑ 可	日期格式 (Ctrl+S...	长度	精度	默认	注释
TotalPurchases	□	Integer	☑					
TotalSpent	□	Float	☑					
FavoriteCategory	□	String	☑					
LastPurchaseDate	□	Date	☑	"yyyy-MM-dd"				
Occupation	□	String	☑					
WebsiteVisitFrequency	□	Integer	☑					



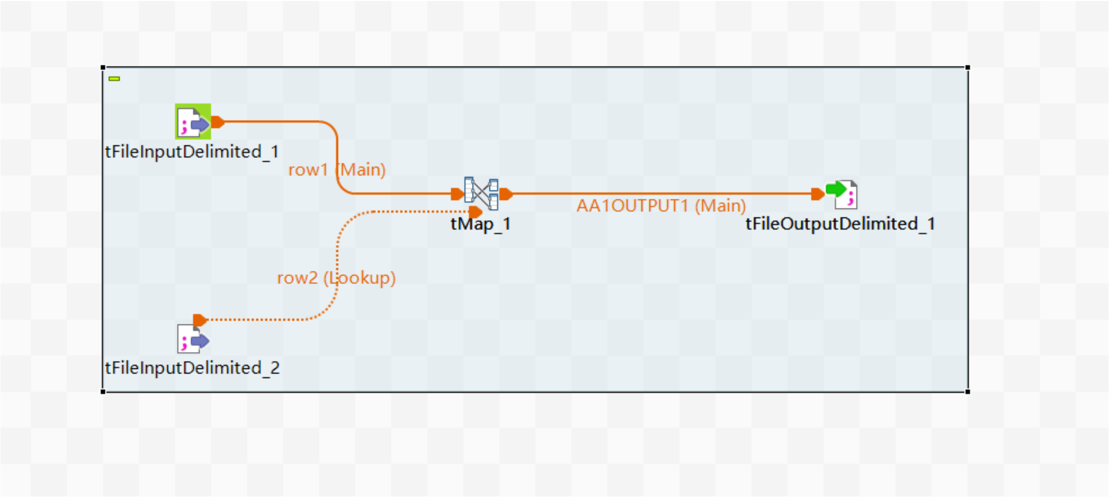
tMap_1输出

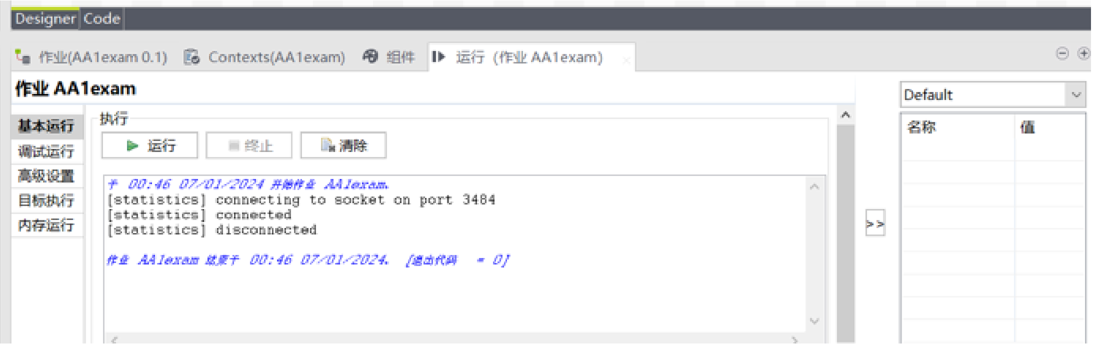
新输出的名字?

AA1output1



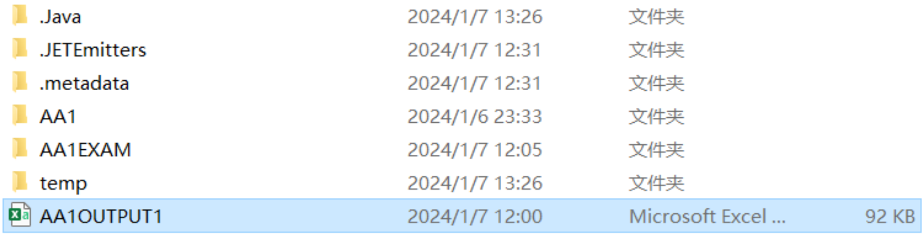
```
于 13:18 07/01/2024 开始作业 AA1_S2188090.
[statistics] connecting to socket on port 3697
[statistics] connected
[statistics] disconnected

作业 AA1_S2188090 结束于 13:18 07/01/2024。 [退出代码 = 0]
```





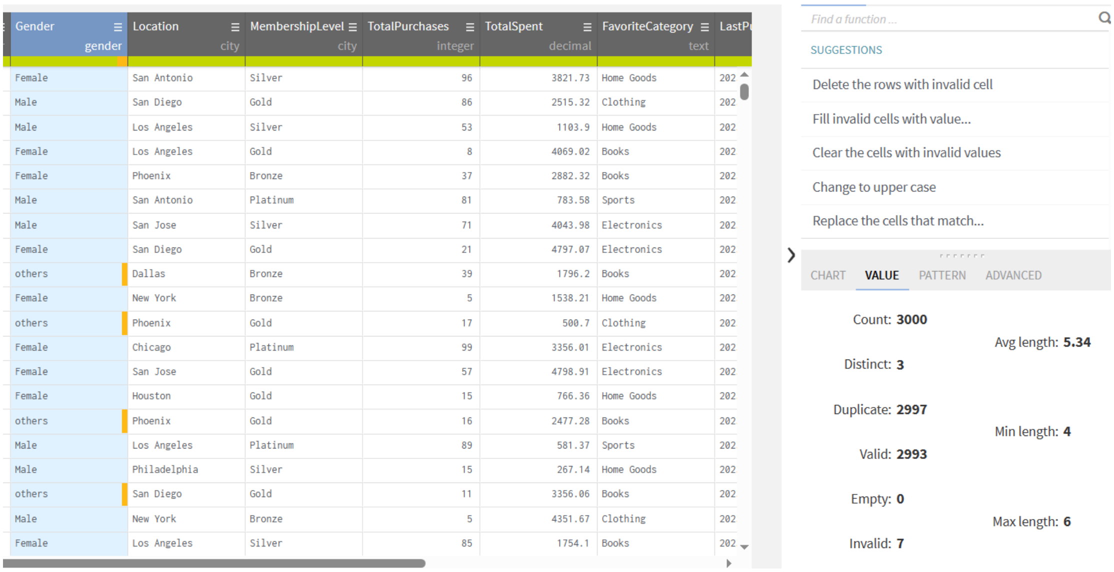
作业 AA1exam

```
于 00:46 07/01/2024 开始作业 AA1exam.
[statistics] connecting to socket on port 3484
[statistics] connected
[statistics] disconnected

作业 AA1exam 结束于 00:46 07/01/2024. [退出代码 = 0]
```



.Java	2024/1/7 13:26	文件夹	
.JETEmitters	2024/1/7 12:31	文件夹	
.metadata	2024/1/7 12:31	文件夹	
AA1	2024/1/6 23:33	文件夹	
AA1EXAM	2024/1/7 12:05	文件夹	
temp	2024/1/7 13:26	文件夹	
AA1OUTPUT1	2024/1/7 12:00	Microsoft Excel ...	92 KB

Talend DP



Gender gender	Location city	MembershipLevel city	TotalPurchases integer	TotalSpent decimal	FavoriteCategory text	LastP
Female	San Antonio	Silver	96	3821.73	Home Goods	202
Male	San Diego	Gold	86	2515.32	Clothing	202
Male	Los Angeles	Silver	53	1103.9	Home Goods	202
Female	Los Angeles	Gold	8	4069.02	Books	202
Female	Phoenix	Bronze	37	2882.32	Books	202
Male	San Antonio	Platinum	81	783.58	Sports	202
Male	San Jose	Silver	71	4043.98	Electronics	202
Female	San Diego	Gold	21	4797.07	Electronics	202
others	Dallas	Bronze	39	1796.2	Books	202
Female	New York	Bronze	5	1538.21	Home Goods	202
others	Phoenix	Gold	17	500.7	Clothing	202
Female	Chicago	Platinum	99	3356.01	Electronics	202
Female	San Jose	Gold	57	4798.91	Electronics	202
Female	Houston	Gold	15	766.36	Home Goods	202
others	Phoenix	Gold	16	2477.28	Books	202
Male	Los Angeles	Platinum	89	581.37	Sports	202
Male	Philadelphia	Silver	15	267.14	Home Goods	202
others	San Diego	Gold	11	3356.06	Books	202
Male	New York	Bronze	5	4351.67	Clothing	202
Female	Los Angeles	Silver	85	1754.1	Books	202

Delete the rows with invalid cell

Fill invalid cells with value...

Clear the cells with invalid values

Change to upper case

Replace the cells that match...

Count: 3000

Distinct: 3

Duplicate: 2997

Valid: 2993

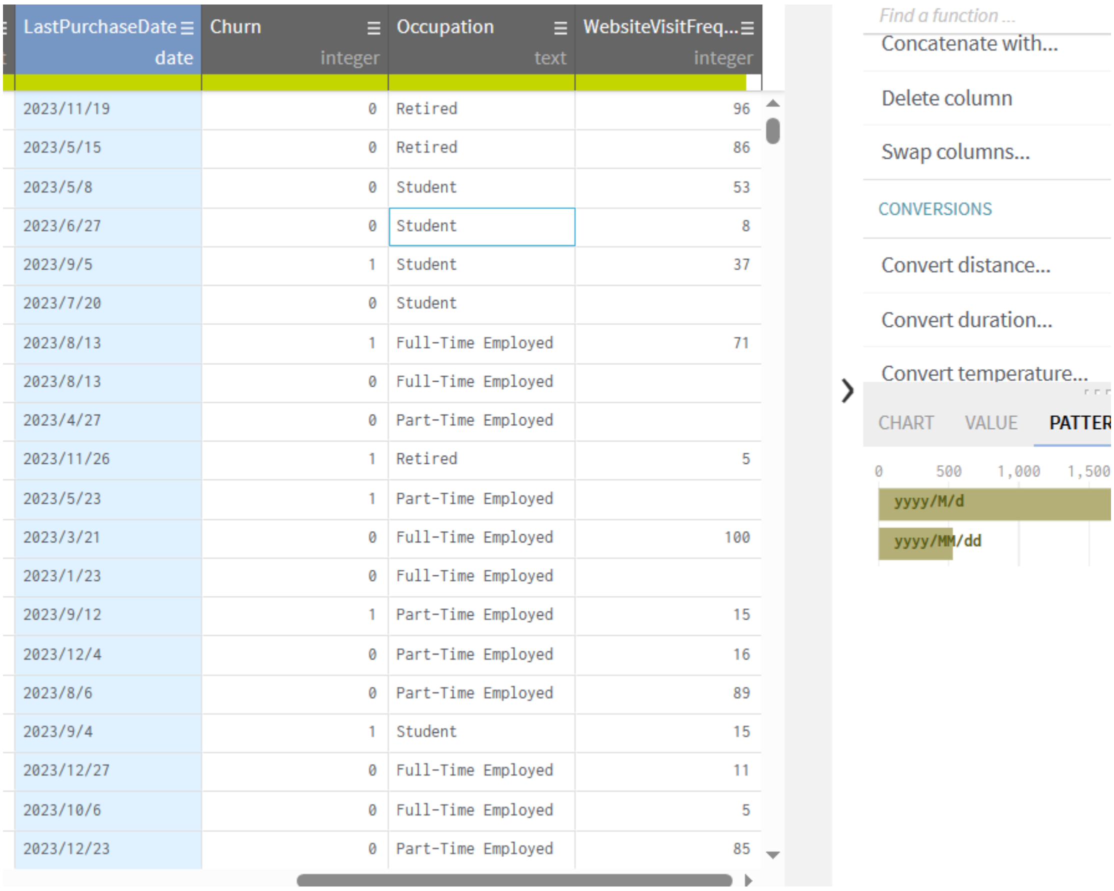
Empty: 0

Invalid: 7

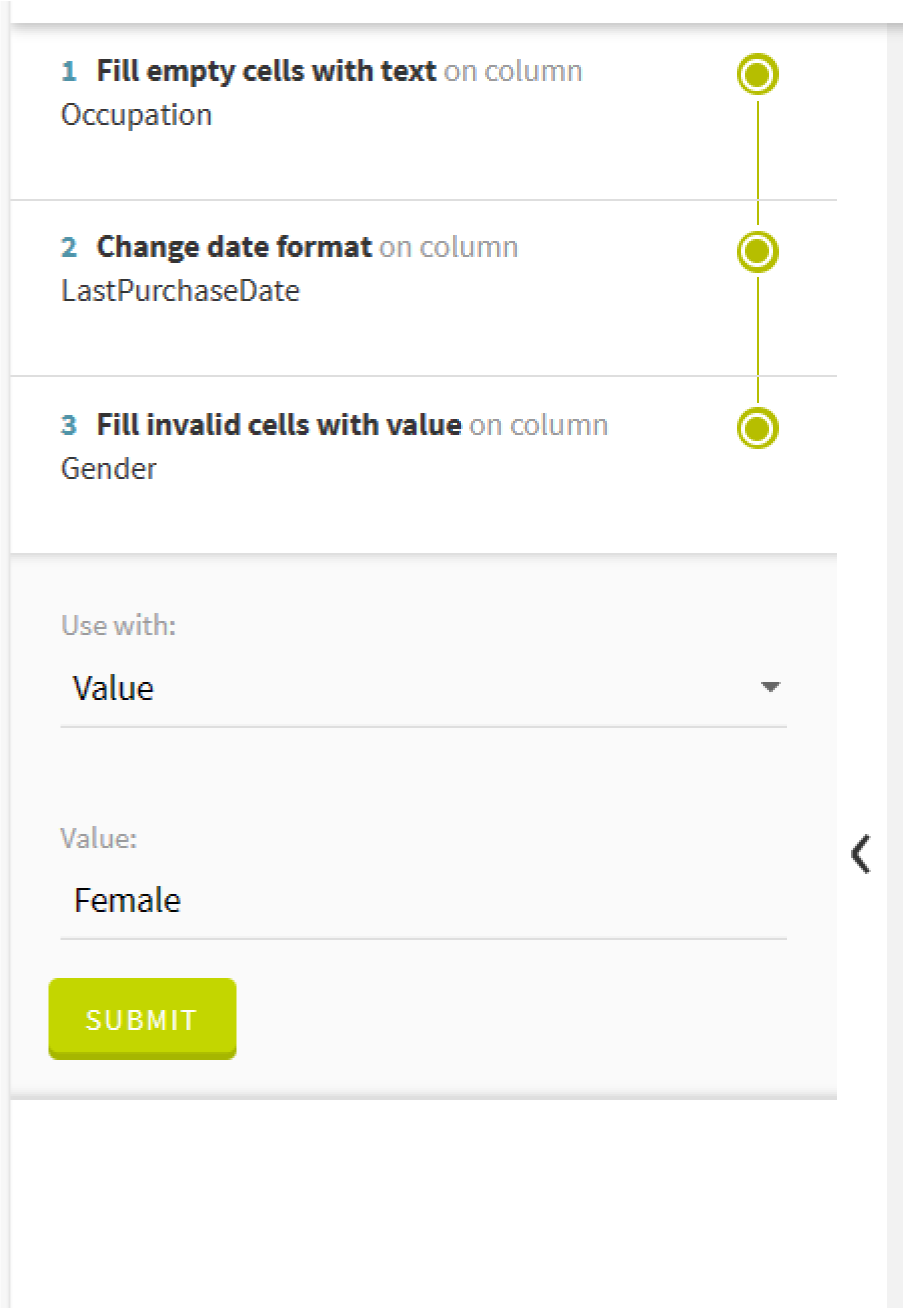
Avg length: 5.34

Min length: 4

Max length: 6



LastPurchaseDate date	Churn integer	Occupation text	WebsiteVisitFreq... integer
2023/11/19	0	Retired	96
2023/5/15	0	Retired	86
2023/5/8	0	Student	53
2023/6/27	0	Student	8
2023/9/5	1	Student	37
2023/7/20	0	Student	
2023/8/13	1	Full-Time Employed	71
2023/8/13	0	Full-Time Employed	
2023/4/27	0	Part-Time Employed	
2023/11/26	1	Retired	5
2023/5/23	1	Part-Time Employed	
2023/3/21	0	Full-Time Employed	100
2023/1/23	0	Full-Time Employed	
2023/9/12	1	Part-Time Employed	15
2023/12/4	0	Part-Time Employed	16
2023/8/6	0	Part-Time Employed	89
2023/9/4	1	Student	15
2023/12/27	0	Full-Time Employed	11
2023/10/6	0	Full-Time Employed	5
2023/12/23	0	Part-Time Employed	85



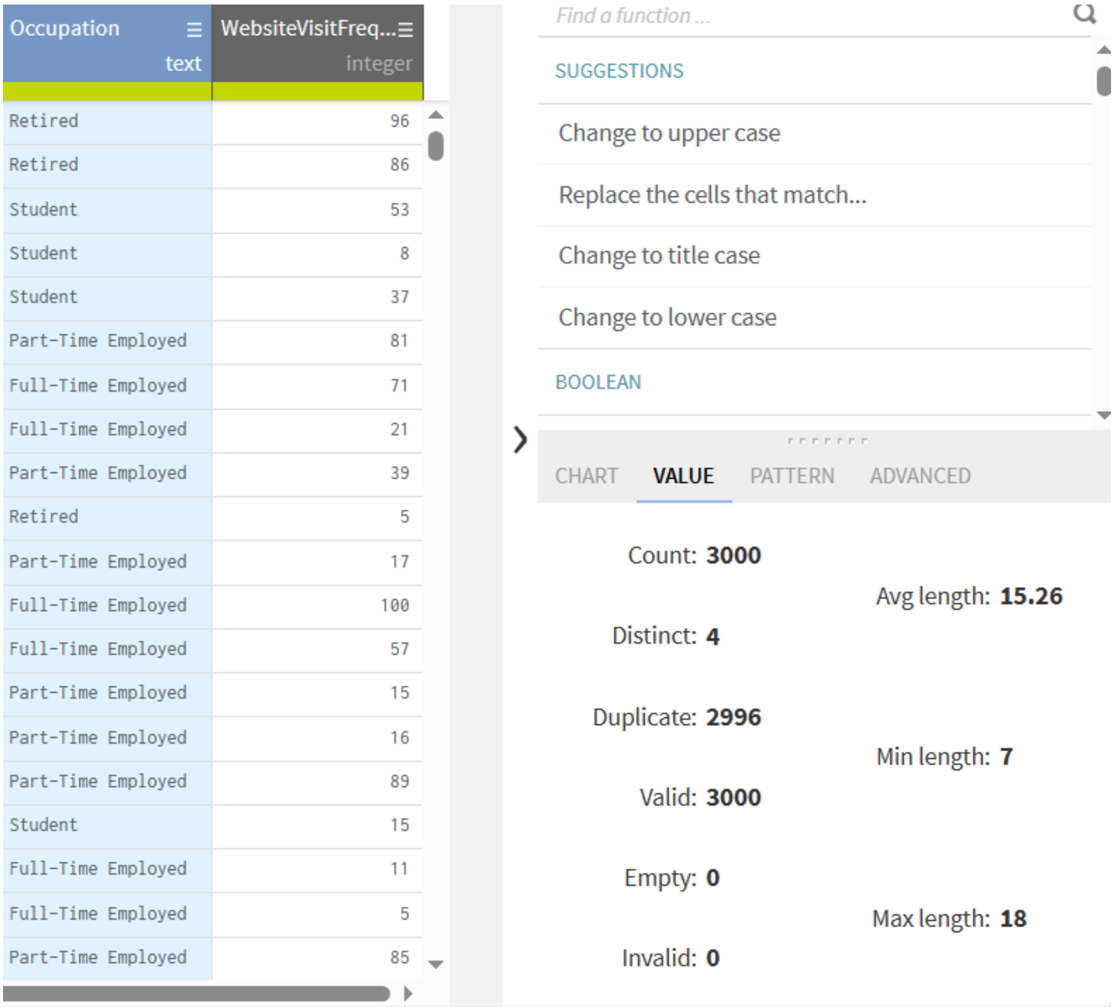
1 **Fill empty cells with text** on column Occupation

2 **Change date format** on column LastPurchaseDate

3 **Fill invalid cells with value** on column Gender

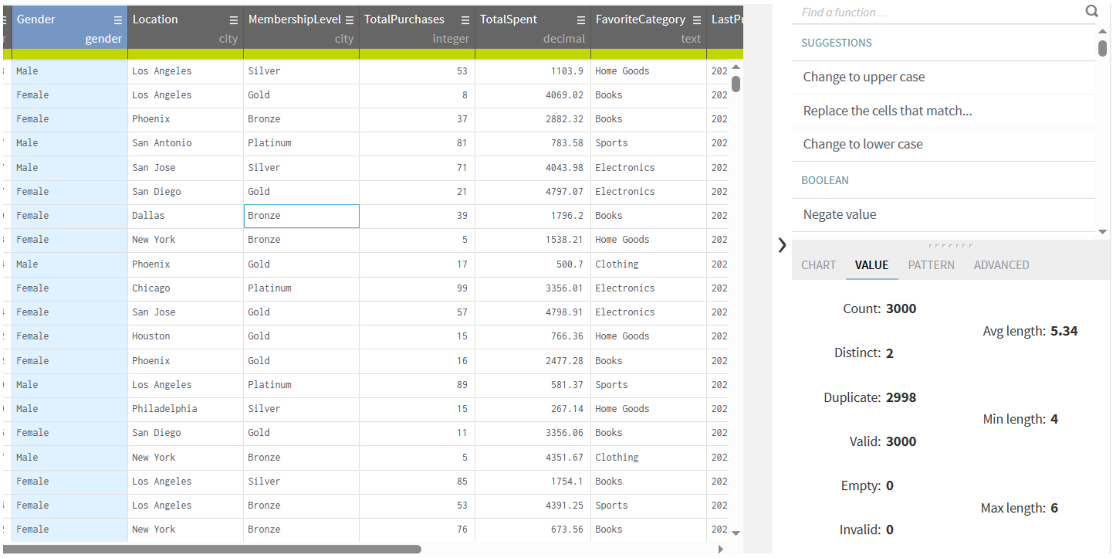
Use with: Value

Value: Female



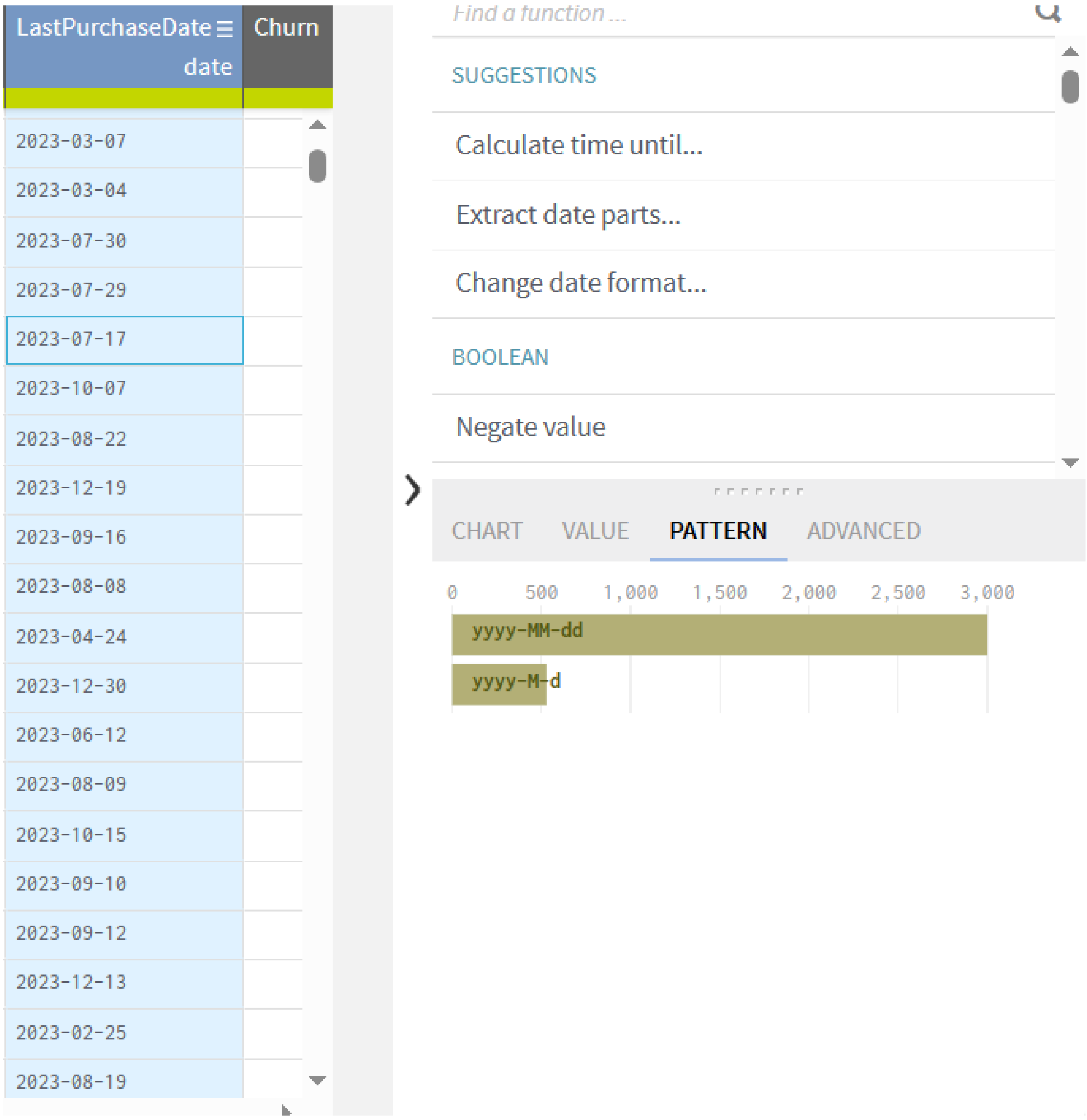
Occupation text	WebsiteVisitFreq... integer
Retired	96
Retired	86
Student	53
Student	8
Student	37
Part-Time Employed	81
Full-Time Employed	71
Full-Time Employed	21
Part-Time Employed	39
Retired	5
Part-Time Employed	17
Full-Time Employed	100
Full-Time Employed	57
Part-Time Employed	15
Part-Time Employed	16
Part-Time Employed	89
Student	15
Full-Time Employed	11
Full-Time Employed	5
Part-Time Employed	85

Count: 3000
Distinct: 4
Duplicate: 2996
Valid: 3000
Empty: 0
Invalid: 0
Avg length: 15.26
Min length: 7
Max length: 18

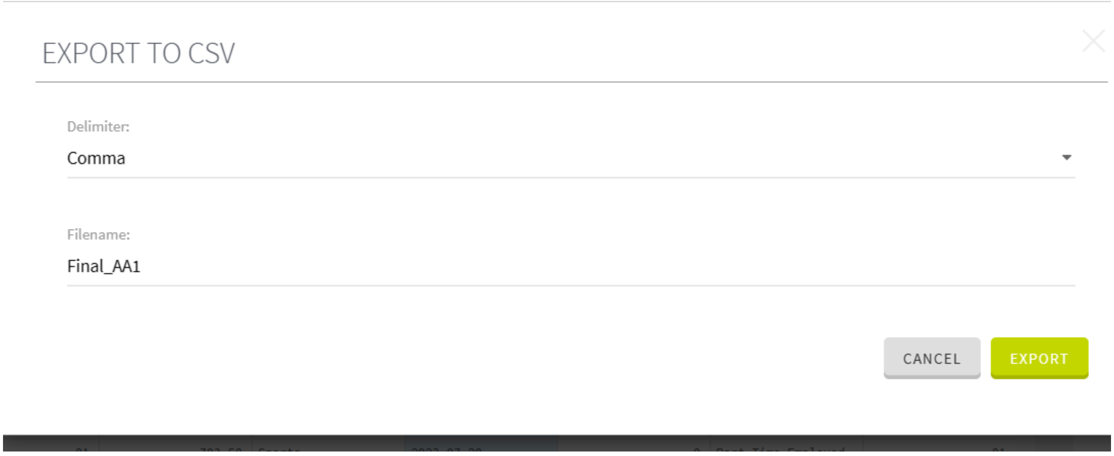


Gender gender	Location city	MembershipLevel city	TotalPurchases integer	TotalSpent decimal	FavoriteCategory text	LastP
Male	Los Angeles	Silver	53	1103.9	Home Goods	202
Female	Los Angeles	Gold	8	4069.02	Books	202
Female	Phoenix	Bronze	37	2882.32	Books	202
Male	San Antonio	Platinum	81	783.58	Sports	202
Male	San Jose	Silver	71	4043.98	Electronics	202
Female	San Diego	Gold	21	4797.07	Electronics	202
Female	Dallas	Bronze	39	1796.2	Books	202
Female	New York	Bronze	5	1538.21	Home Goods	202
Male	Phoenix	Gold	17	500.7	Clothing	202
Female	Chicago	Platinum	99	3356.01	Electronics	202
Female	San Jose	Gold	57	4798.91	Electronics	202
Female	Houston	Gold	15	766.36	Home Goods	202
Female	Phoenix	Gold	16	2477.28	Books	202
Male	Los Angeles	Platinum	89	581.37	Sports	202
Male	Philadelphia	Silver	15	267.14	Home Goods	202
Female	San Diego	Gold	11	3356.06	Books	202
Male	New York	Bronze	5	4351.67	Clothing	202
Female	Los Angeles	Silver	85	1754.1	Books	202
Female	Los Angeles	Bronze	53	4391.25	Sports	202
Female	New York	Bronze	76	673.56	Books	202

Count: 3000
Distinct: 2
Duplicate: 2998
Valid: 3000
Empty: 0
Invalid: 0
Avg length: 5.34
Min length: 4
Max length: 6



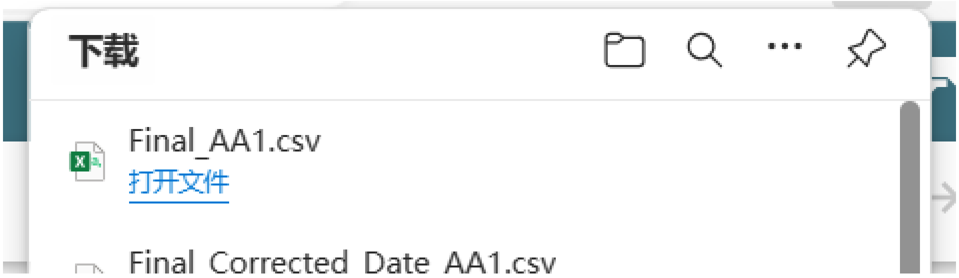
LastPurchaseDate date	Churn
2023-03-07	
2023-03-04	
2023-07-30	
2023-07-29	
2023-07-17	
2023-10-07	
2023-08-22	
2023-12-19	
2023-09-16	
2023-08-08	
2023-04-24	
2023-12-30	
2023-06-12	
2023-08-09	
2023-10-15	
2023-09-10	
2023-09-12	
2023-12-13	
2023-02-25	
2023-08-19	

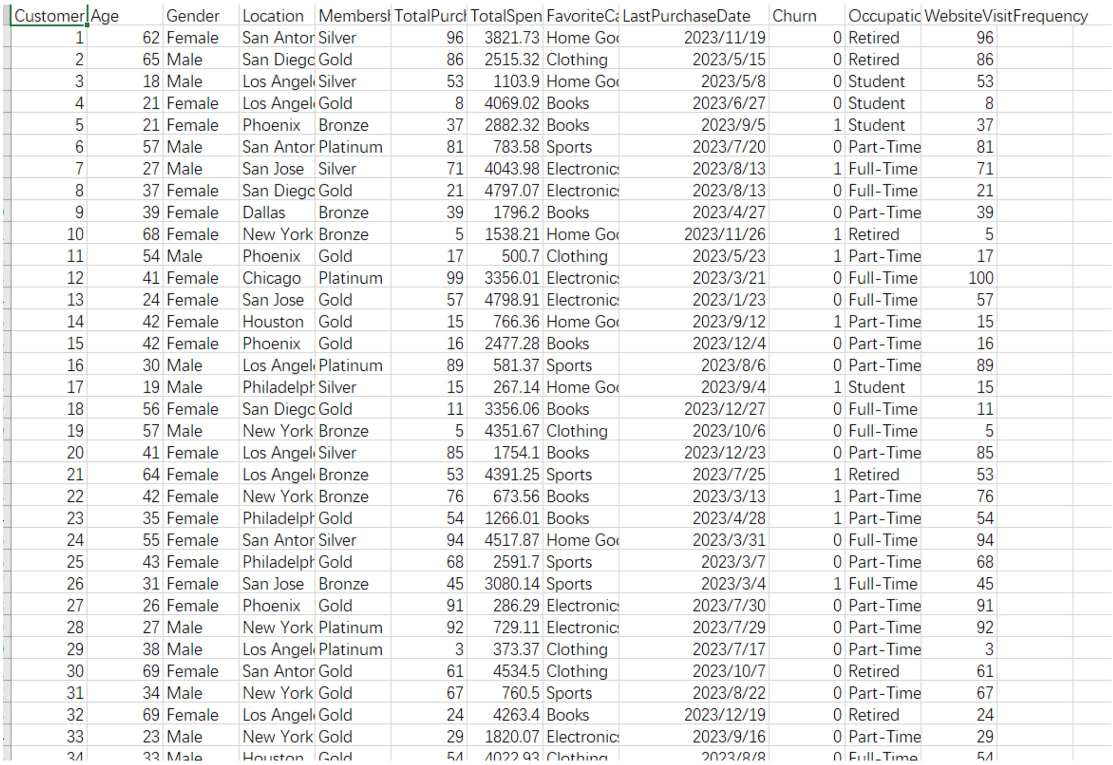


EXPORT TO CSV

Delimiter:
Comma

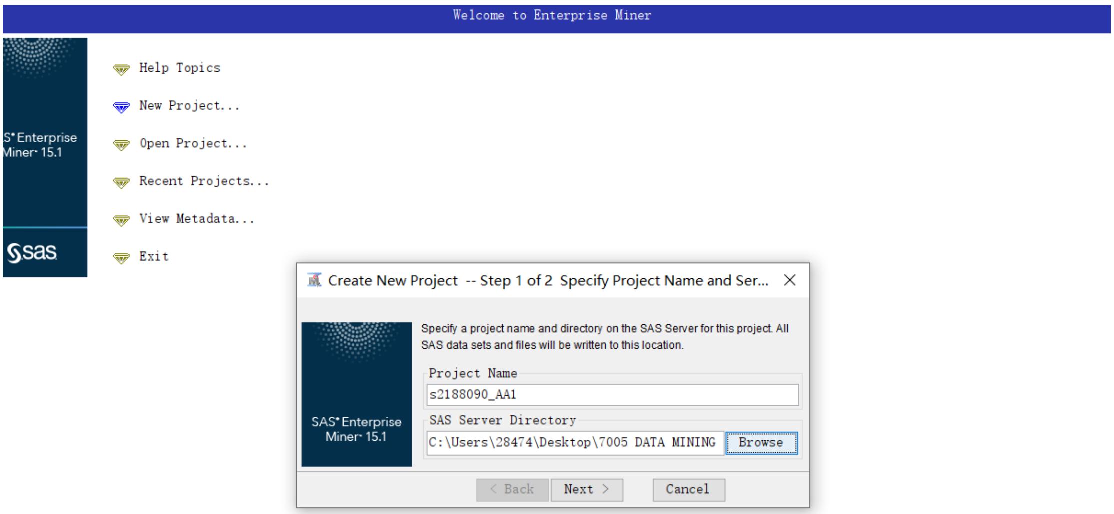
Filename:
Final_AA1





Customer	Age	Gender	Location	Membersl	TotalPurc	TotalSpen	FavoriteCa	LastPurchaseDate	Churn	Occupatic	WebsiteVisitFrequency
1	62	Female	San Antor	Silver	96	3821.73	Home Goo	2023/11/19	0	Retired	96
2	65	Male	San Diegc	Gold	86	2515.32	Clothing	2023/5/15	0	Retired	86
3	18	Male	Los Angel	Silver	53	1103.9	Home Goo	2023/5/8	0	Student	53
4	21	Female	Los Angel	Gold	8	4069.02	Books	2023/6/27	0	Student	8
5	21	Female	Phoenix	Bronze	37	2882.32	Books	2023/9/5	1	Student	37
6	57	Male	San Antor	Platinum	81	783.58	Sports	2023/7/20	0	Part-Time	81
7	27	Male	San Jose	Silver	71	4043.98	Electronic	2023/8/13	1	Full-Time	71
8	37	Female	San Diegc	Gold	21	4797.07	Electronic	2023/8/13	0	Full-Time	21
9	39	Female	Dallas	Bronze	39	1796.2	Books	2023/4/27	0	Part-Time	39
10	68	Female	New York	Bronze	5	1538.21	Home Goo	2023/11/26	1	Retired	5
11	54	Male	Phoenix	Gold	17	500.7	Clothing	2023/5/23	1	Part-Time	17
12	41	Female	Chicago	Platinum	99	3356.01	Electronic	2023/3/21	0	Full-Time	100
13	24	Female	San Jose	Gold	57	4798.91	Electronic	2023/1/23	0	Full-Time	57
14	42	Female	Houston	Gold	15	766.36	Home Goo	2023/9/12	1	Part-Time	15
15	42	Female	Phoenix	Gold	16	2477.28	Books	2023/12/4	0	Part-Time	16
16	30	Male	Los Angel	Platinum	89	581.37	Sports	2023/8/6	0	Part-Time	89
17	19	Male	Philadelph	Silver	15	267.14	Home Goo	2023/9/4	1	Student	15
18	56	Female	San Diegc	Gold	11	3356.06	Books	2023/12/27	0	Full-Time	11
19	57	Male	New York	Bronze	5	4351.67	Clothing	2023/10/6	0	Full-Time	5
20	41	Female	Los Angel	Silver	85	1754.1	Books	2023/12/23	0	Part-Time	85
21	64	Female	Los Angel	Bronze	53	4391.25	Sports	2023/7/25	1	Retired	53
22	42	Female	New York	Bronze	76	673.56	Books	2023/3/13	1	Part-Time	76
23	35	Female	Philadelph	Gold	54	1266.01	Books	2023/4/28	1	Part-Time	54
24	55	Female	San Antor	Silver	94	4517.87	Home Goo	2023/3/31	0	Full-Time	94
25	43	Female	Philadelph	Gold	68	2591.7	Sports	2023/3/7	0	Part-Time	68
26	31	Female	San Jose	Bronze	45	3080.14	Sports	2023/3/4	1	Full-Time	45
27	26	Female	Phoenix	Gold	91	286.29	Electronic	2023/7/30	0	Part-Time	91
28	27	Male	New York	Platinum	92	729.11	Electronic	2023/7/29	0	Part-Time	92
29	38	Male	Los Angel	Platinum	3	373.37	Clothing	2023/7/17	0	Part-Time	3
30	69	Female	San Antor	Gold	61	4534.5	Clothing	2023/10/7	0	Retired	61
31	34	Male	New York	Gold	67	760.5	Sports	2023/8/22	0	Part-Time	67
32	69	Female	Los Angel	Gold	24	4263.4	Books	2023/12/19	0	Retired	24
33	23	Male	New York	Gold	29	1820.07	Electronic	2023/9/16	0	Part-Time	29
34	33	Male	Houston	Gold	54	4022.93	Clothing	2023/8/8	0	Full-Time	54

Sas enterprise miner

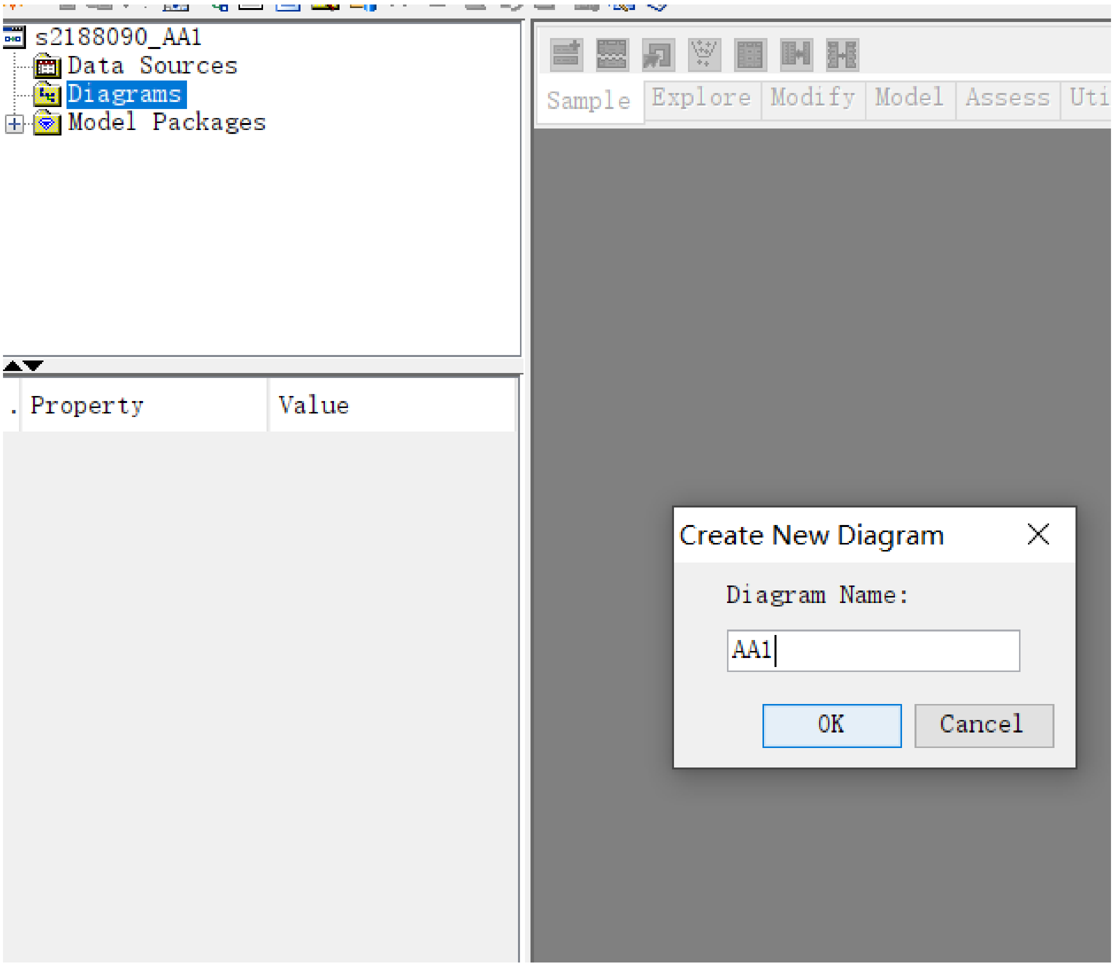


Create New Project -- Step 1 of 2 Specify Project Name and Ser...

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.

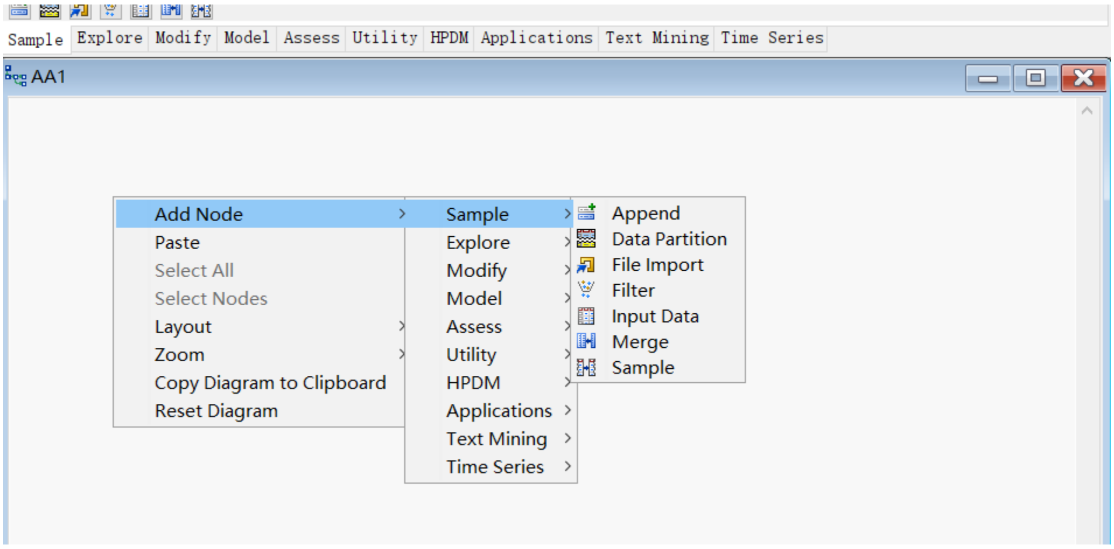
Project Name: s2188090_AA1

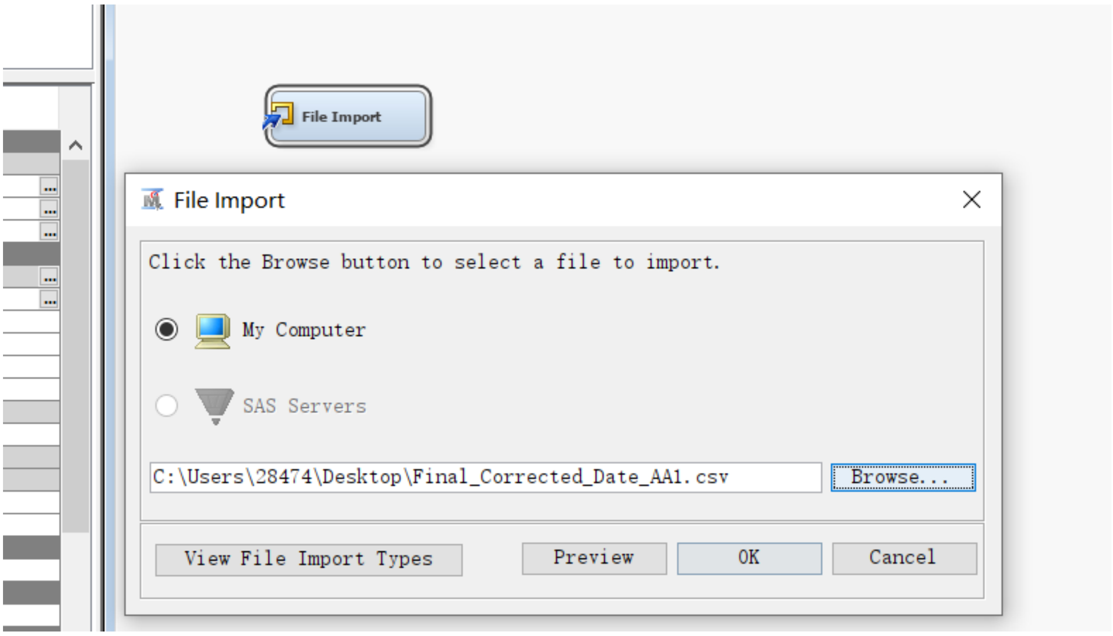
SAS Server Directory: C:\Users\28474\Desktop\7005 DATA MINING



Create New Diagram

Diagram Name: AA1





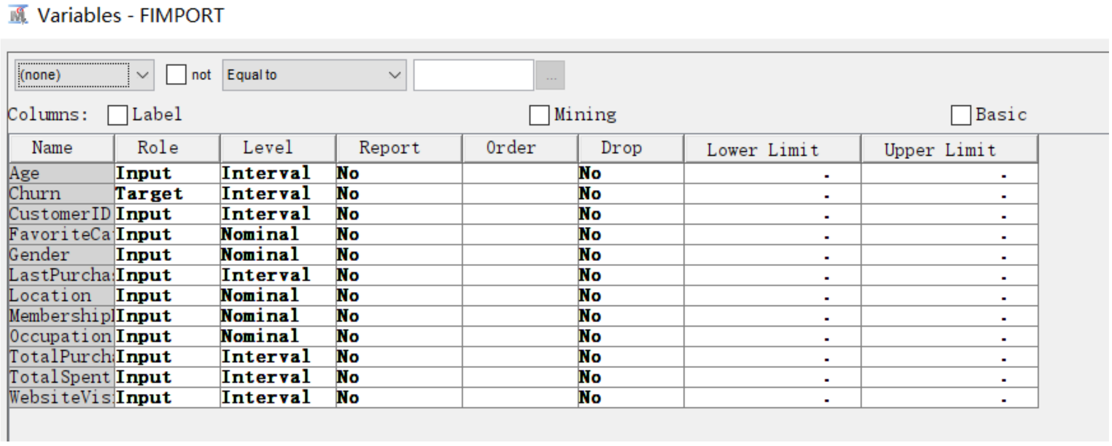
File Import

Click the Browse button to select a file to import.

● My Computer

○ SAS Servers

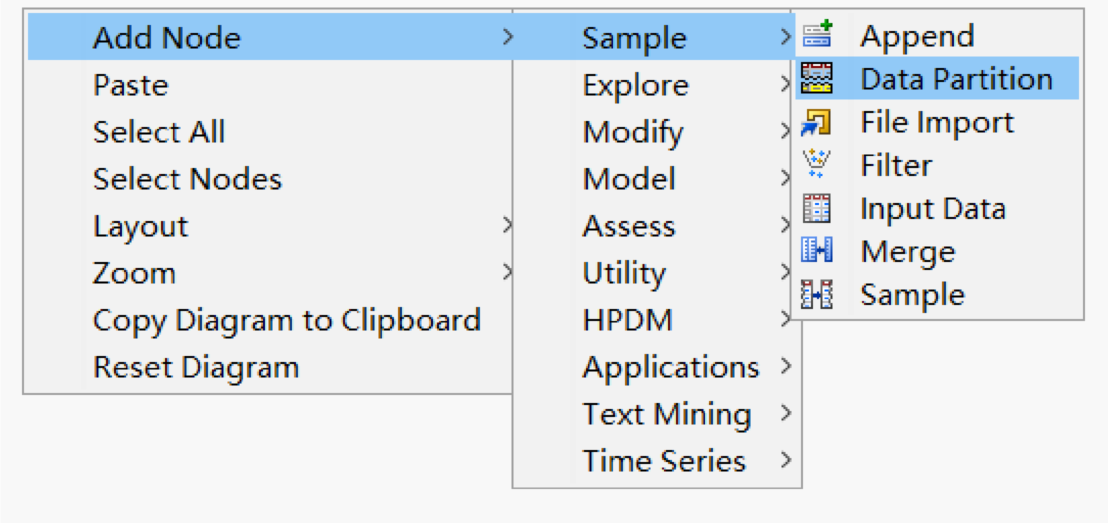
C:\Users\28474\Desktop\Final_Corrected_Date_AA1.csv

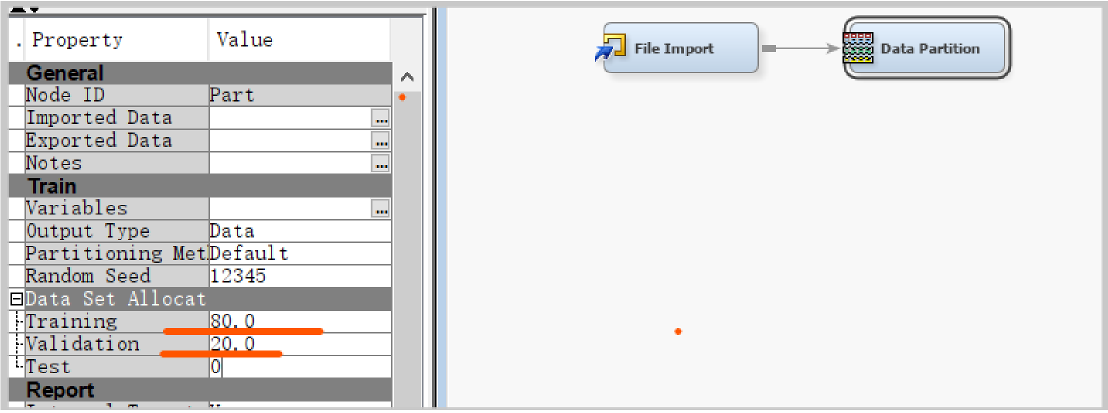

Variables - FIMPORT

(none) □ not Equal to

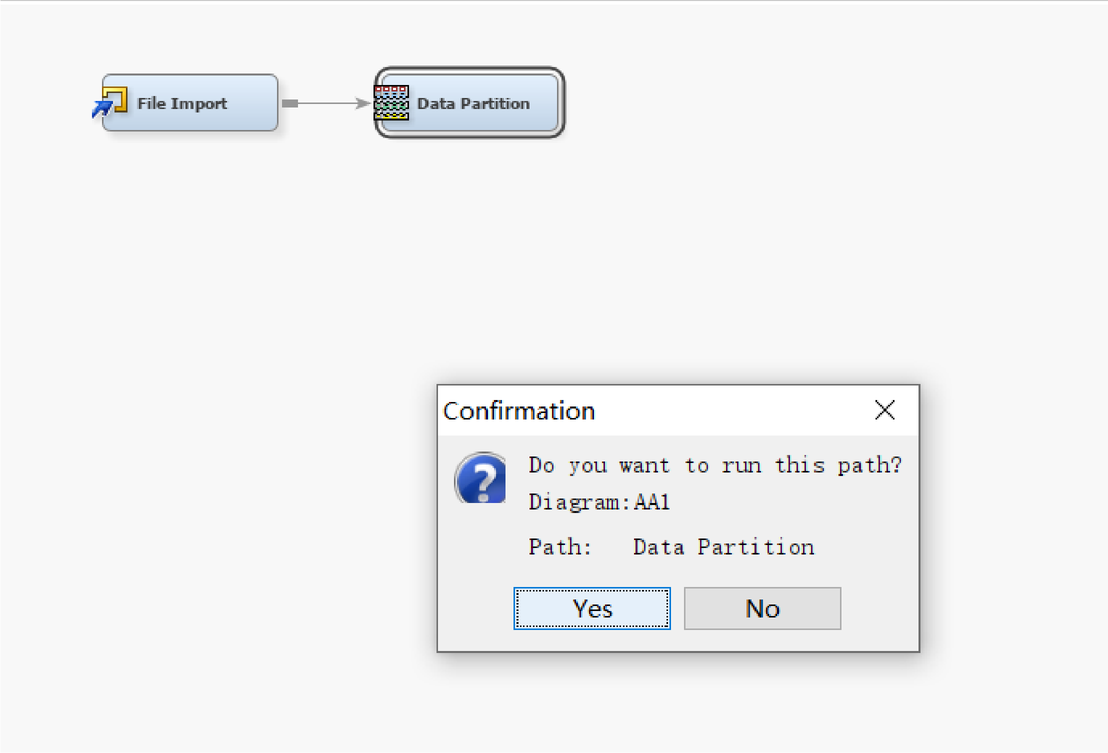
Columns: □ Label □ Mining □ Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Interval	No		No	.	.
CustomerID	Input	Interval	No		No	.	.
FavoriteCa	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurcha	Input	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
Membershipl	Input	Nominal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
TotalPurch	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.
WebsiteVis	Input	Interval	No		No	.	.



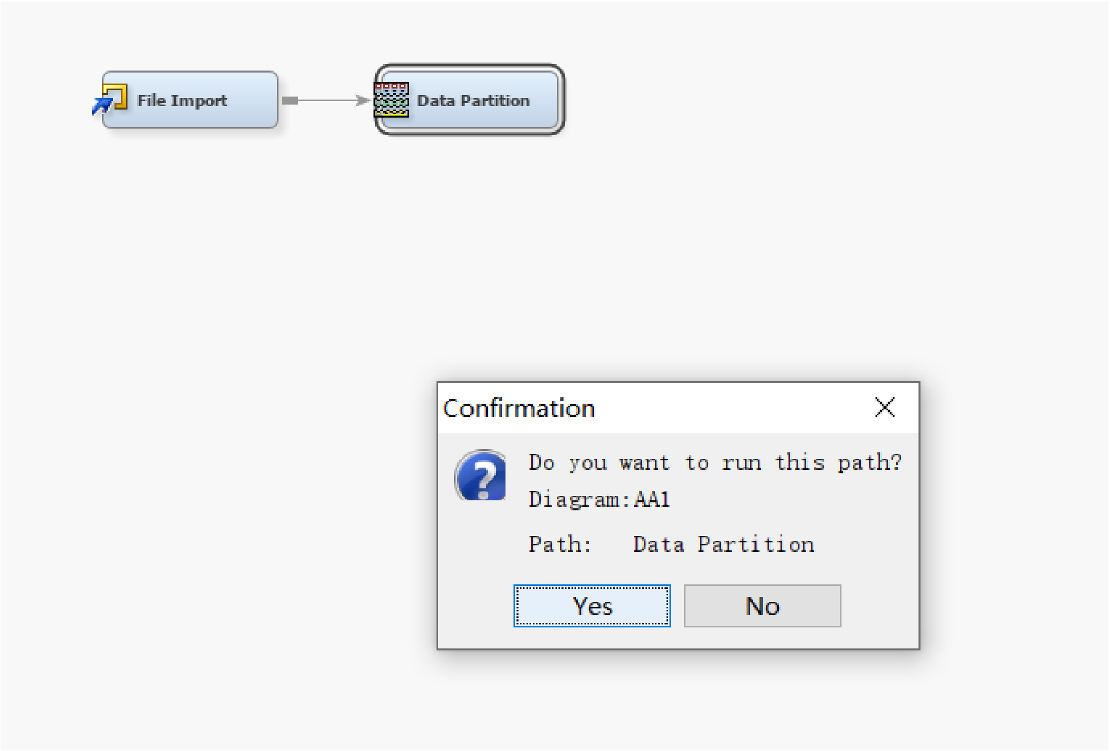


. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Met	Default
Random Seed	12345
⊟Data Set Allocat	
Training	80.0
Validation	20.0
Test	0
Report	



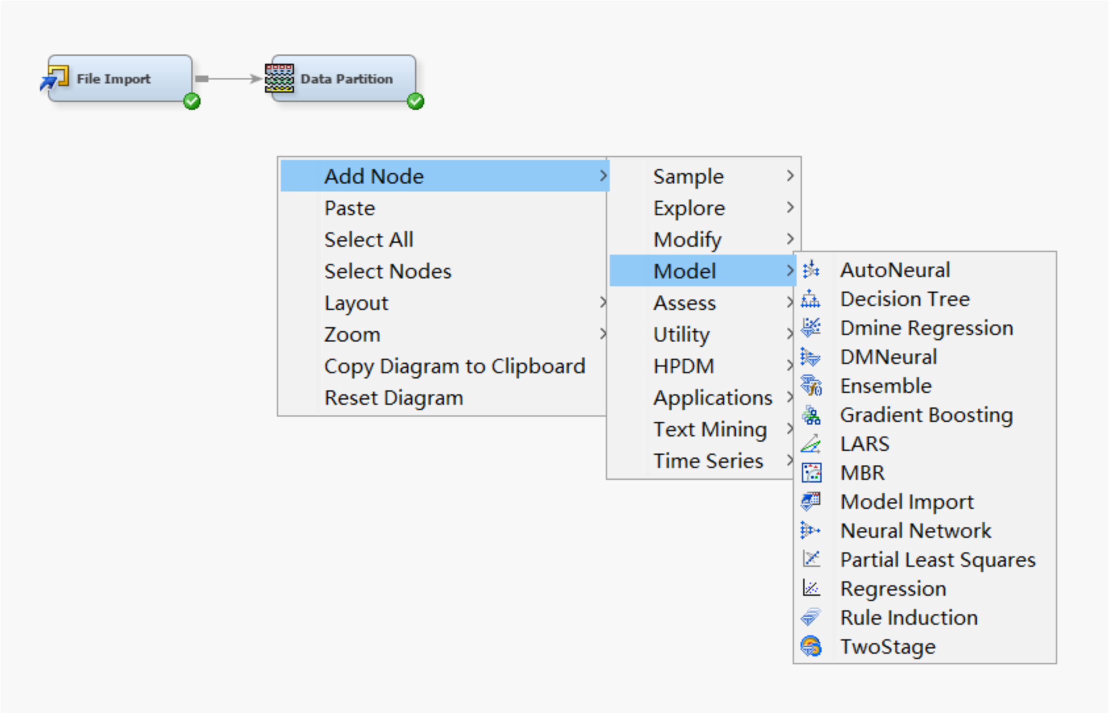
Confirmation

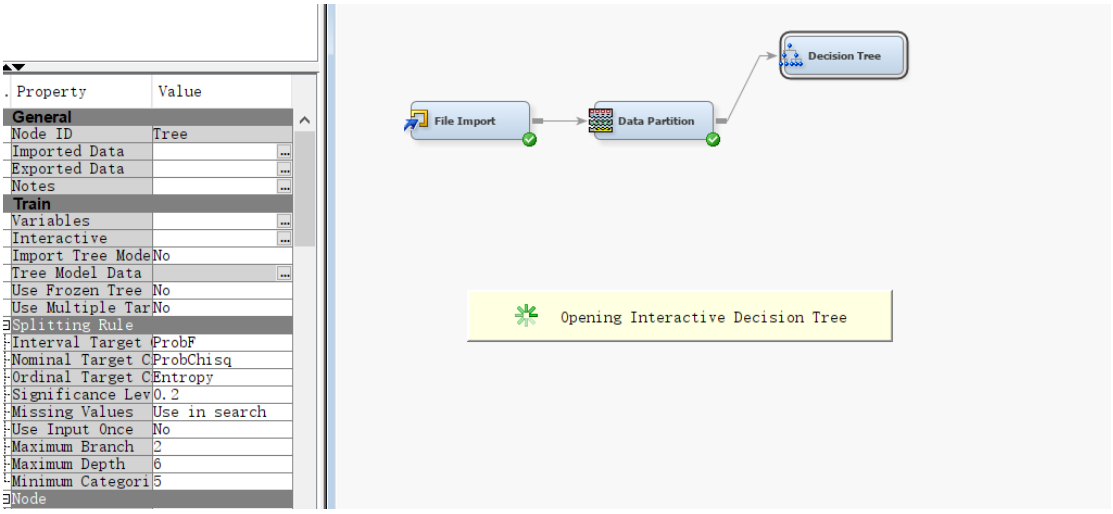
Do you want to run this path?
Diagram: AA1
Path: Data Partition



Do you want to run this path?
Diagram:AA1

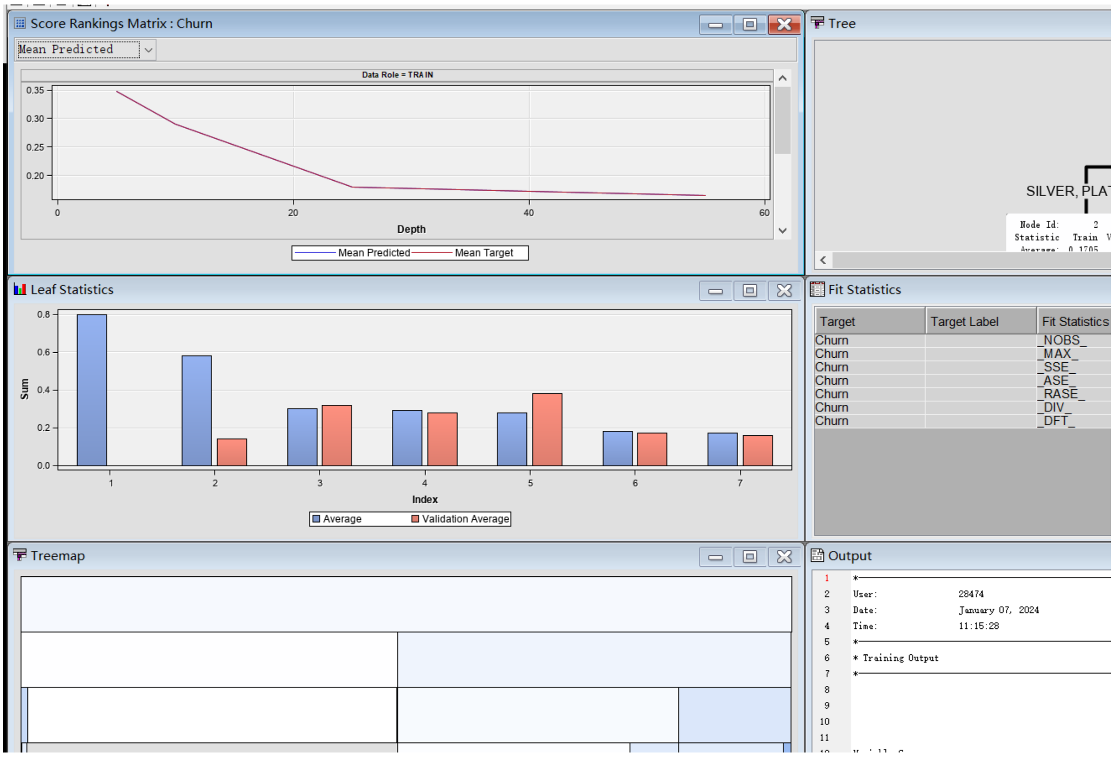
Path: Data Partition





Property	Value
General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Mode	No
Tree Model Data	...
Use Frozen Tree	No
Use Multiple Tar	No
Splitting Rule	
Interval Target C	ProbF
Nominal Target C	ProbChisq
Ordinal Target C	Entropy
Significance Lev	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categori	5
Node	

Opening Interactive Decision Tree



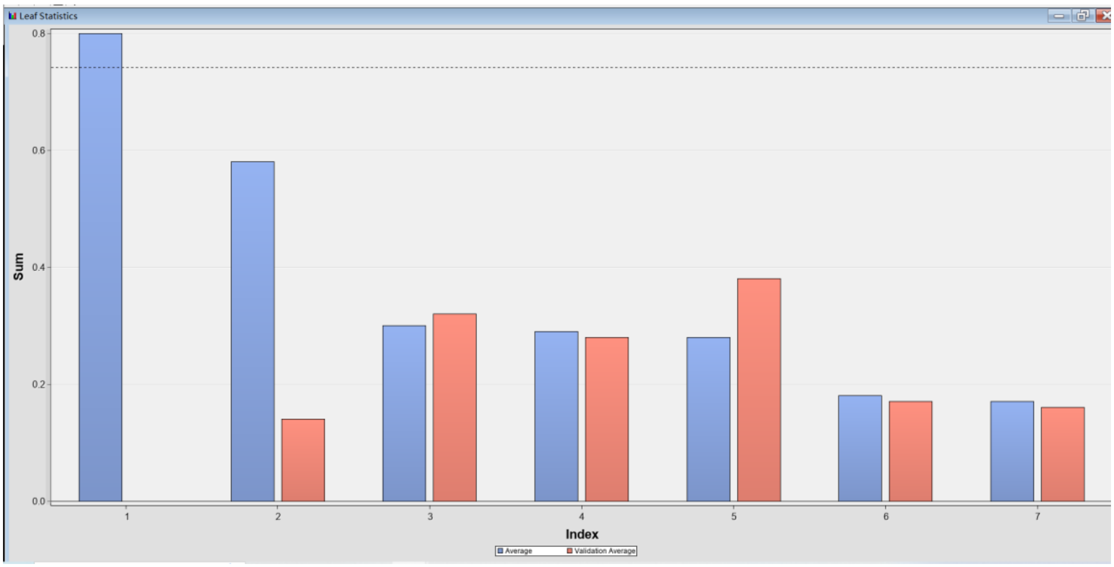
```
Node Id:       2
Statistic   Train  V
Average:  0.1705
```

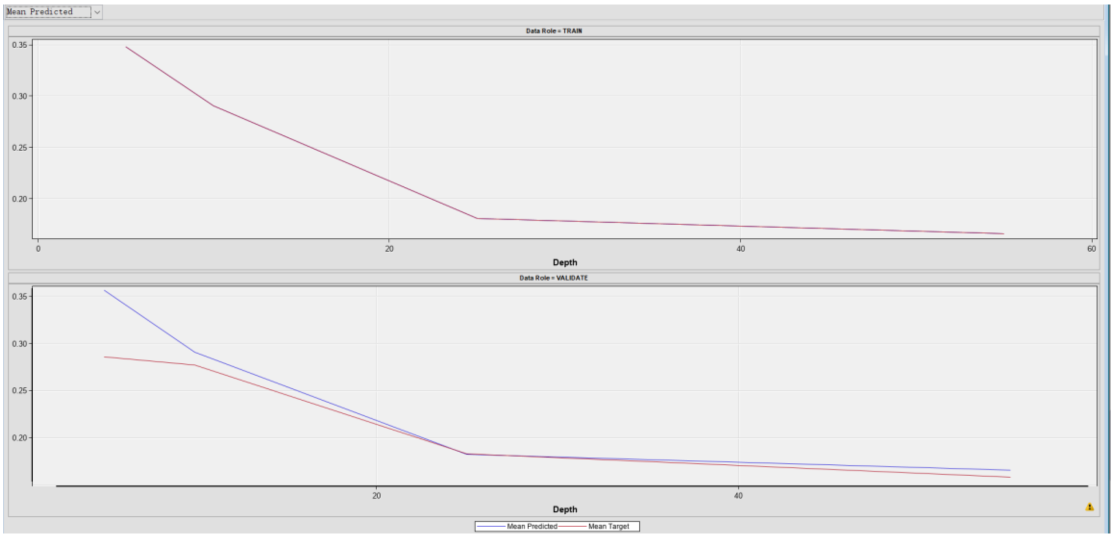
Fit Statistics

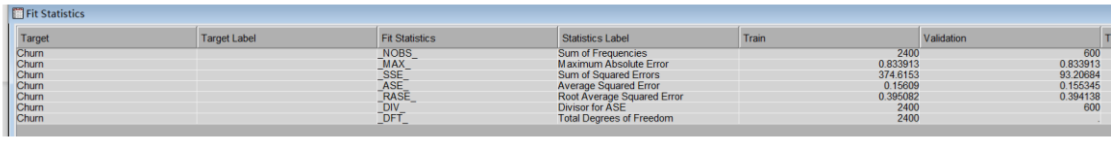
Target	Target Label	Fit Statistics
Churn		_NOBS_
Churn		_MAX_
Churn		_SSE_
Churn		_ASE_
Churn		_RASE_
Churn		_DIV_
Churn		_DFT_

Output

```
*------------------------------------------------------------
User:                 28474
Date:                 January 07, 2024
Time:                 11:15:28
*------------------------------------------------------------
* Training Output
*------------------------------------------------------------
```

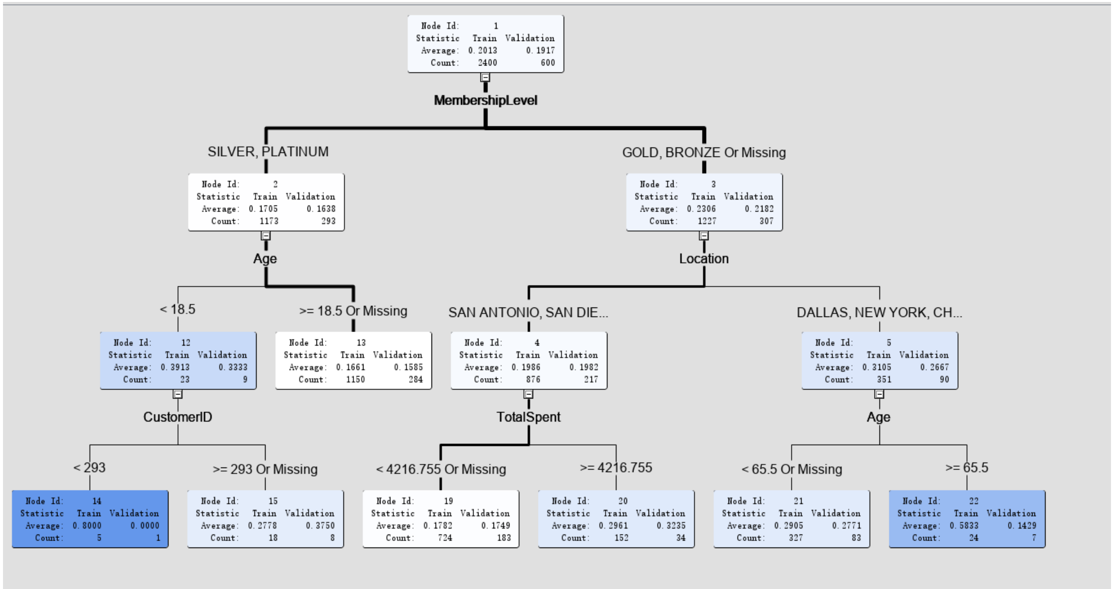


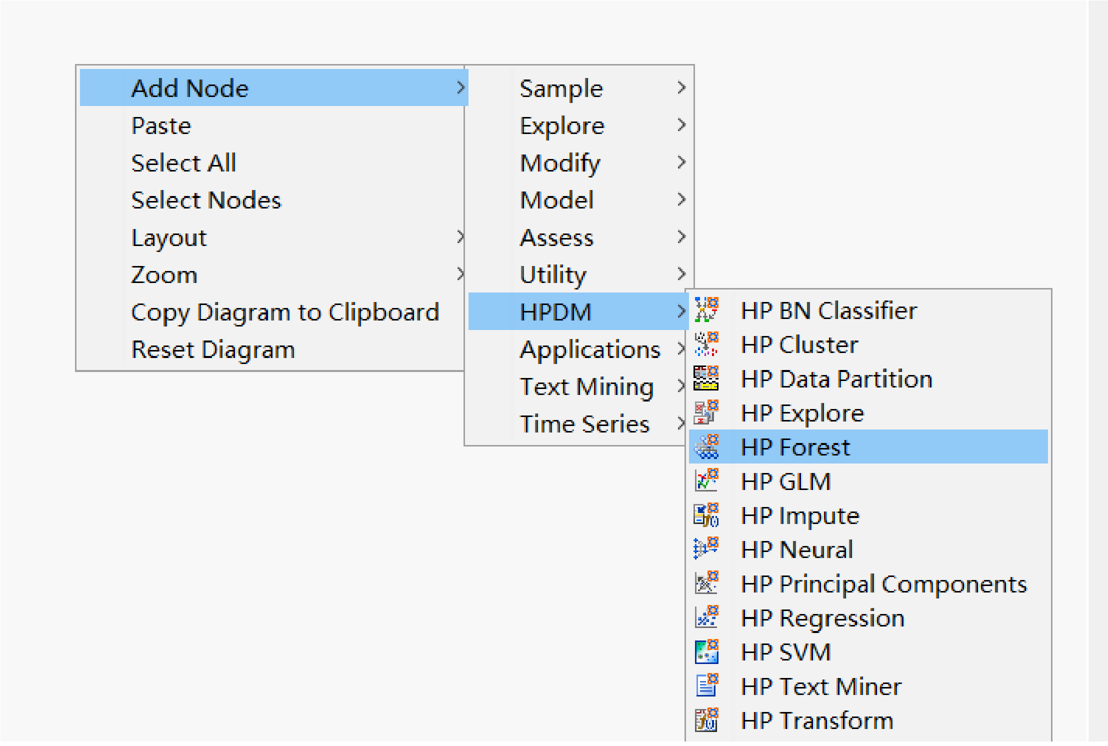


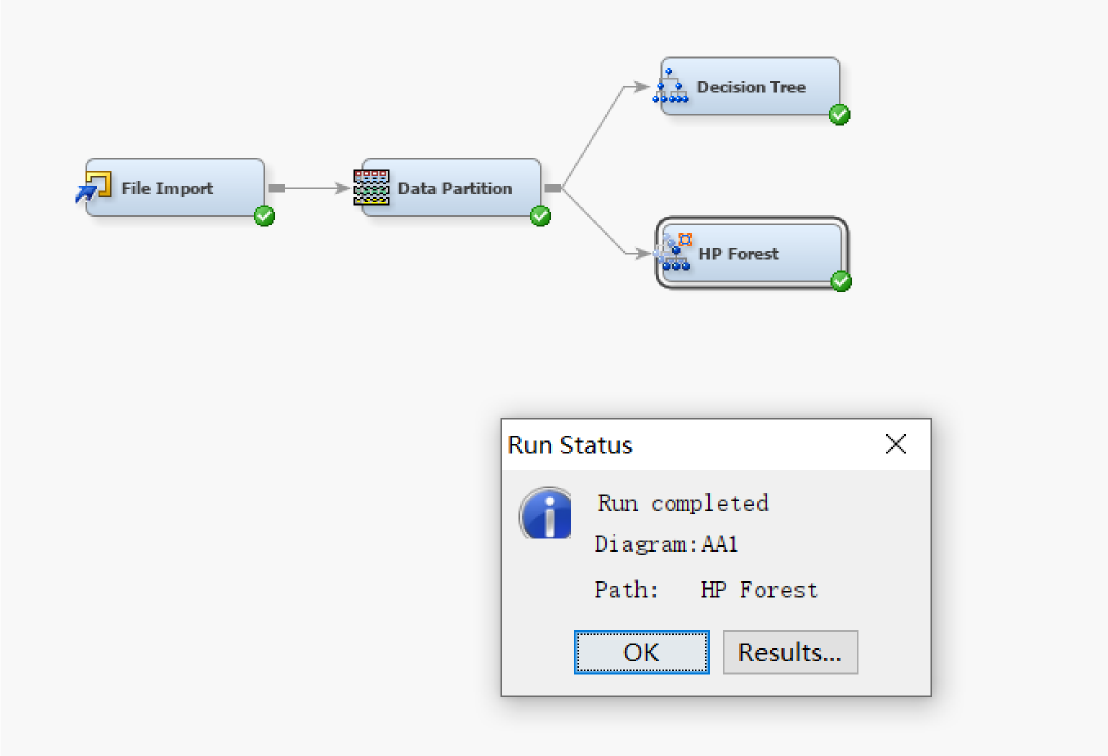


Fit Statistics

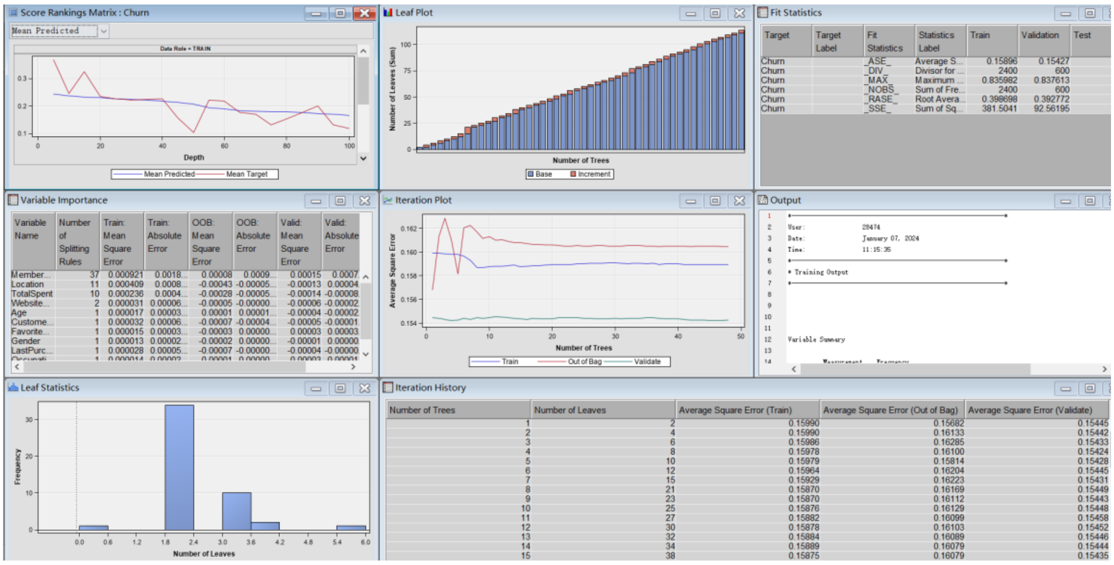
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	T
Churn		_NOBS_	Sum of Frequencies	2400	600	
Churn		_MAX_	Maximum Absolute Error	0.833913	0.833913	
Churn		_SSE_	Sum of Squared Errors	374.6153	93.20684	
Churn		_ASE_	Average Squared Error	0.15609	0.155345	
Churn		_RASE_	Root Average Squared Error	0.395082	0.394138	
Churn		_DIV_	Divisor for ASE	2400	600	
Churn		_DFT_	Total Degrees of Freedom	2400	.	







Run completed
Diagram:AA1
Path: HP Forest



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_ASE_	Average S...	0.15896	0.15427	
Churn		_DIV_	Divisor for ...	2400	600	
Churn		_MAX_	Maximum ...	0.835982	0.837613	
Churn		_NOBS_	Sum of Fre...	2400	600	
Churn		_RASE_	Root Avera...	0.398698	0.392772	
Churn		_SSE_	Sum of Sq...	381.5041	92.56195	

Output

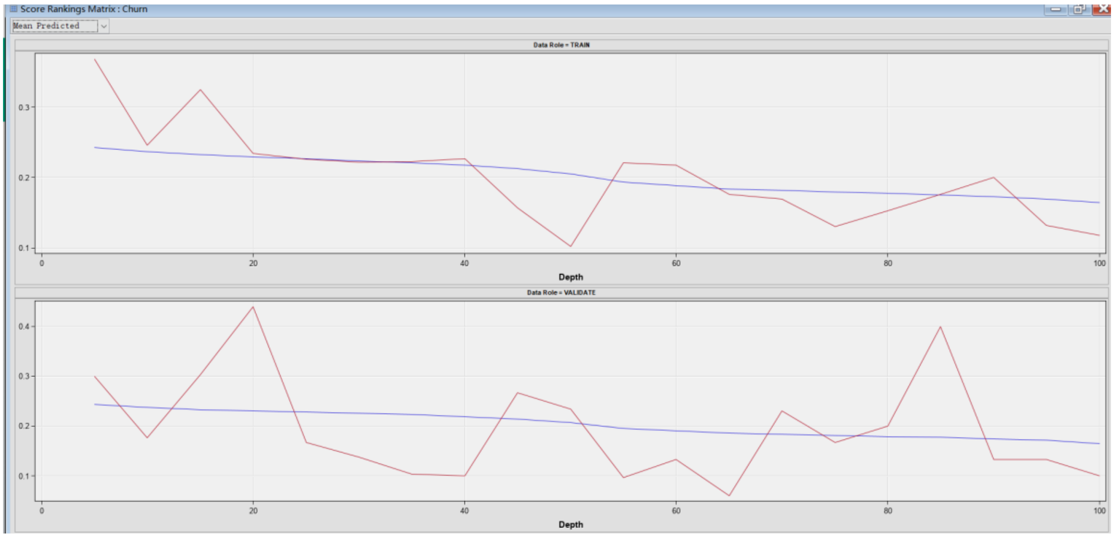
```
*------------------------------------------------------*
User:               28474
Date:               January 07, 2024
Time:               11:15:35
*------------------------------------------------------*
* Training Output
*------------------------------------------------------*




Variable Summary

         Measurement    Frequency
```

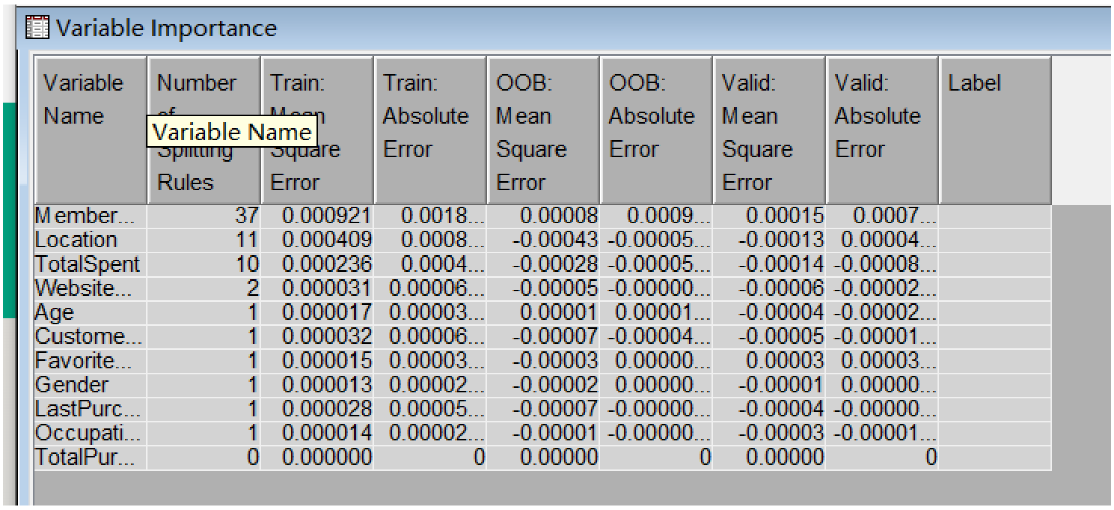
Iteration History

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)
1	2	0.15990	0.15682	0.15445
2	4	0.15990	0.16133	0.15442
3	6	0.15986	0.16285	0.15433
4	8	0.15978	0.16100	0.15424
5	10	0.15979	0.15814	0.15428
6	12	0.15964	0.16204	0.15445
7	15	0.15929	0.16223	0.15431
8	21	0.15870	0.16169	0.15449
9	23	0.15870	0.16112	0.15443
10	25	0.15876	0.16129	0.15448
11	27	0.15882	0.16099	0.15458
12	30	0.15878	0.16103	0.15452
13	32	0.15884	0.16089	0.15446
14	34	0.15889	0.16079	0.15444
15	38	0.15875	0.16079	0.15435

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error
Member...	37	0.000921	0.0018...	0.00008	0.0009...	0.00015	0.0007
Location	11	0.000409	0.0008...	-0.00043	-0.00005...	-0.00013	0.00004
TotalSpent	10	0.000236	0.0004...	-0.00028	-0.00005...	-0.00014	-0.00008
Website...	2	0.000031	0.00006...	-0.00005	-0.00000...	-0.00006	-0.00002
Age	1	0.000017	0.00003...	0.00001	0.00001...	-0.00004	-0.00002
Custome...	1	0.000032	0.00006...	-0.00007	-0.00004...	-0.00005	-0.00001
Favorite...	1	0.000015	0.00003...	-0.00003	0.00000...	0.00003	0.00003
Gender	1	0.000013	0.00002...	-0.00002	0.00000...	-0.00001	0.00000
LastPurc...	1	0.000028	0.00005...	-0.00007	-0.00000...	-0.00004	-0.00000
Occupati...	1	0.000014	0.00002...	-0.00001	0.00000...	-0.00003	-0.00001

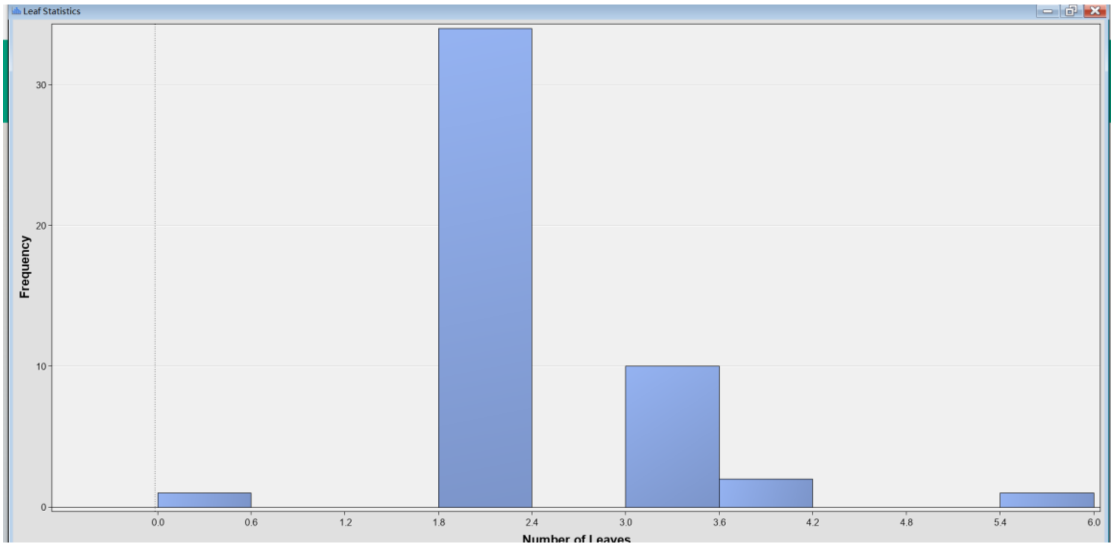


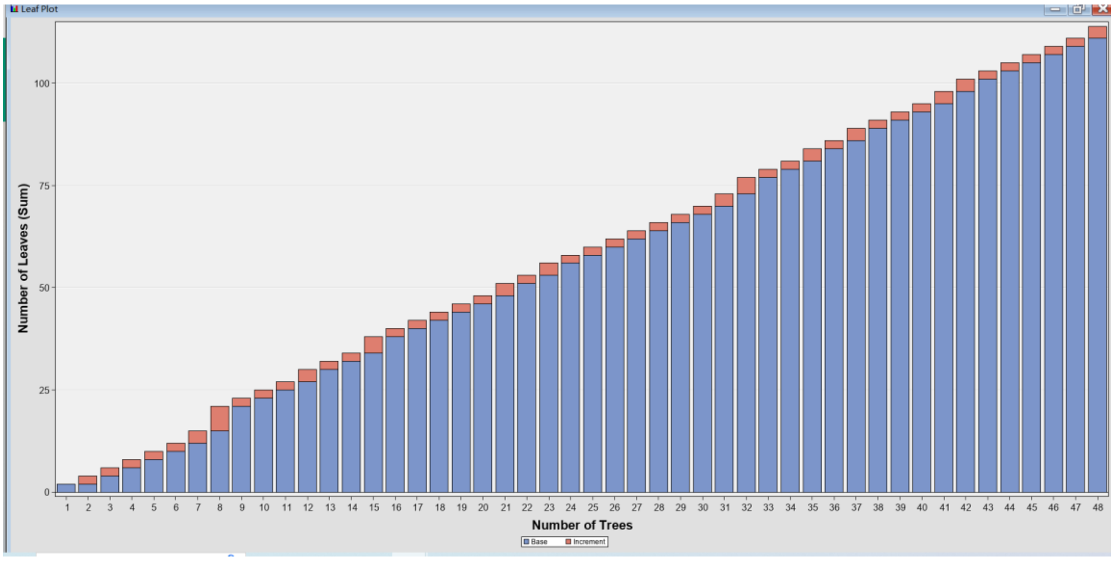


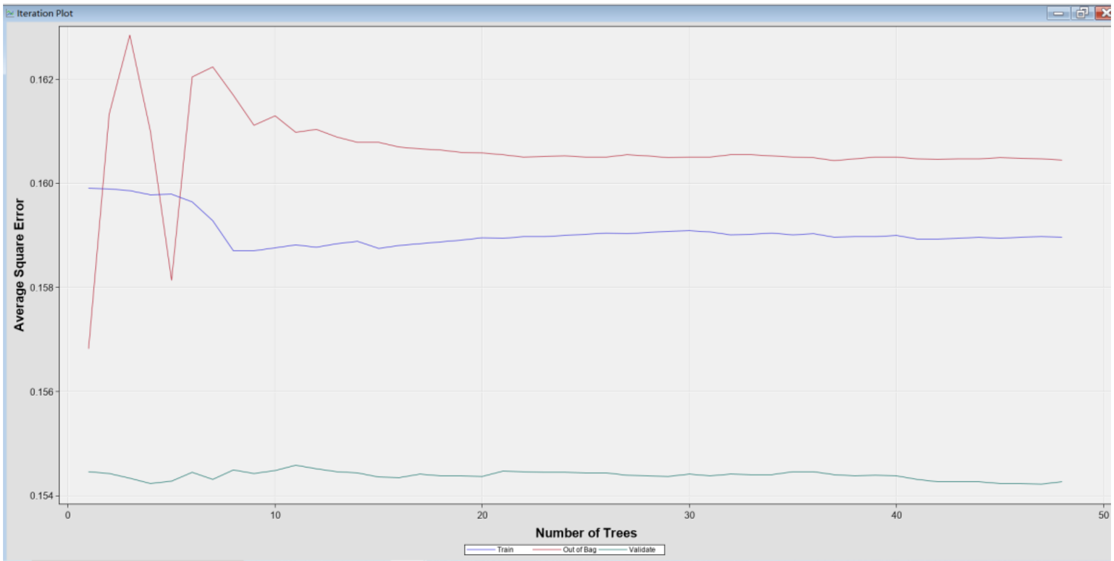
Variable Importance

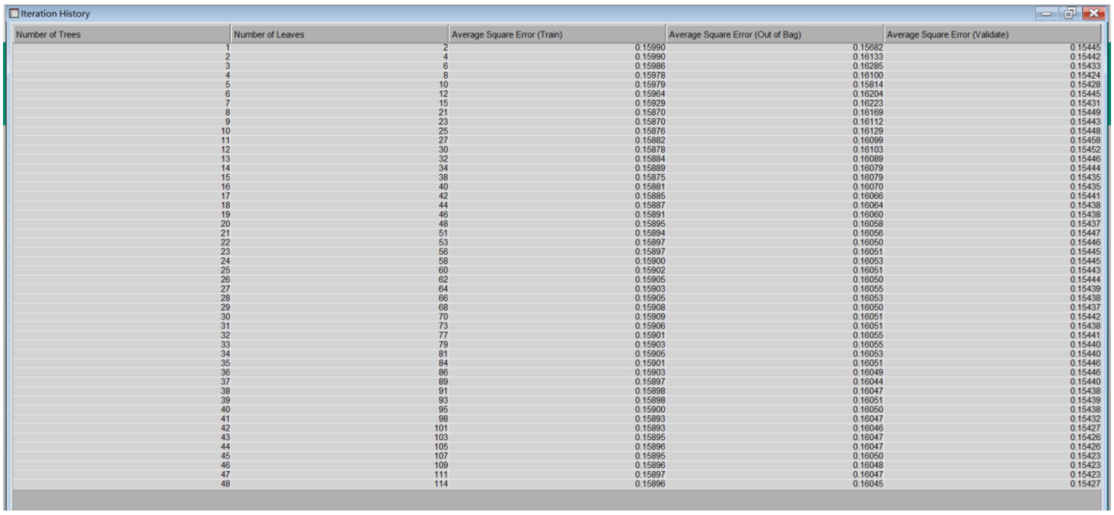
Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Valid: Mean Square Error	Valid: Absolute Error	Label
Member...	37	0.000921	0.0018...	0.00008	0.0009...	0.00015	0.0007...	
Location	11	0.000409	0.0008...	-0.00043	-0.00005...	-0.00013	0.00004...	
TotalSpent	10	0.000236	0.0004...	-0.00028	-0.00005...	-0.00014	-0.00008...	
Website...	2	0.000031	0.00006...	-0.00005	-0.00000...	-0.00006	-0.00002...	
Age	1	0.000017	0.00003...	0.00001	0.00001...	-0.00004	-0.00002...	
Custome...	1	0.000032	0.00006...	-0.00007	-0.00004...	-0.00005	-0.00001...	
Favorite...	1	0.000015	0.00003...	-0.00003	0.00000...	0.00003	0.00003...	
Gender	1	0.000013	0.00002...	-0.00002	0.00000...	-0.00001	0.00000...	
LastPurc...	1	0.000028	0.00005...	-0.00007	-0.00000...	-0.00004	-0.00000...	
Occupati...	1	0.000014	0.00002...	-0.00001	-0.00000...	-0.00003	-0.00001...	
TotalPur...	0	0.000000	0	0.00000	0	0.00000	0	

Variable Name


Leaf Statistics
Frequency
30
20
10
0
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
5.4
6.0
Number of Leaves

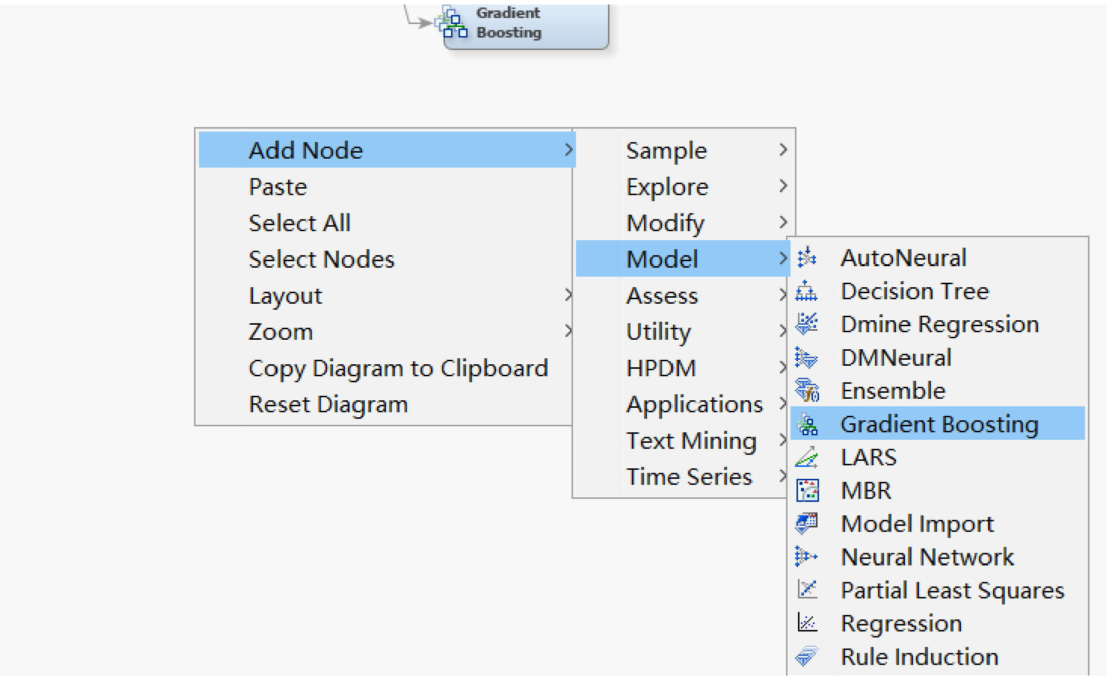

Leaf Plot
Number of Leaves (Sum)
100
75
50
25
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
Number of Trees
Base
Increment

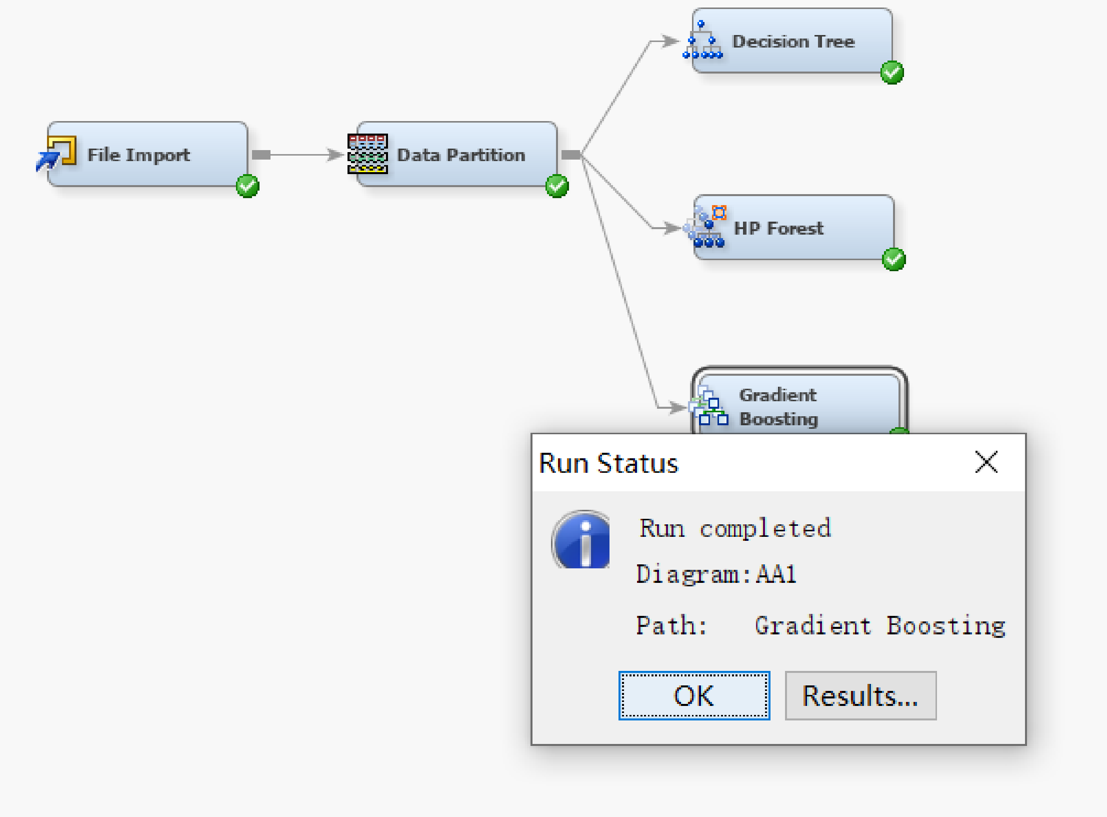

Iteration Plot
Average Square Error
0.162
0.160
0.158
0.156
0.154
0
10
20
30
40
50
Number of Trees
Train
Out of Bag
Validate



Iteration History

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)
1	2	0.15990	0.15682	0.15445
2	4	0.15990	0.16133	0.15442
3	6	0.15986	0.16285	0.15433
4	8	0.15978	0.16100	0.15424
5	10	0.15979	0.15814	0.15428
6	12	0.15964	0.16204	0.15445
7	15	0.15929	0.16223	0.15431
8	21	0.15870	0.16169	0.15449
9	23	0.15870	0.16112	0.15443
10	25	0.15876	0.16129	0.15448
11	27	0.15882	0.16099	0.15458
12	30	0.15878	0.16103	0.15452
13	32	0.15884	0.16089	0.15446
14	34	0.15889	0.16079	0.15444
15	38	0.15875	0.16079	0.15435
16	40	0.15881	0.16070	0.15435
17	42	0.15885	0.16066	0.15441
18	44	0.15887	0.16064	0.15438
19	46	0.15891	0.16060	0.15438
20	48	0.15895	0.16058	0.15437
21	51	0.15894	0.16056	0.15447
22	53	0.15897	0.16050	0.15446
23	56	0.15897	0.16051	0.15445
24	58	0.15900	0.16053	0.15445
25	60	0.15902	0.16051	0.15443
26	62	0.15905	0.16050	0.15444
27	64	0.15903	0.16055	0.15439
28	66	0.15905	0.16053	0.15438
29	68	0.15908	0.16050	0.15437
30	70	0.15909	0.16051	0.15442
31	73	0.15906	0.16051	0.15438
32	77	0.15901	0.16055	0.15441
33	79	0.15903	0.16055	0.15440
34	81	0.15905	0.16053	0.15440
35	84	0.15901	0.16051	0.15446
36	86	0.15903	0.16049	0.15446
37	89	0.15897	0.16044	0.15440
38	91	0.15898	0.16047	0.15438
39	93	0.15898	0.16051	0.15439
40	95	0.15900	0.16050	0.15438
41	98	0.15893	0.16047	0.15432
42	101	0.15893	0.16046	0.15427
43	103	0.15895	0.16047	0.15426
44	105	0.15896	0.16047	0.15426
45	107	0.15895	0.16050	0.15423
46	109	0.15896	0.16048	0.15423
47	111	0.15897	0.16047	0.15423
48	114	0.15896	0.16045	0.15427

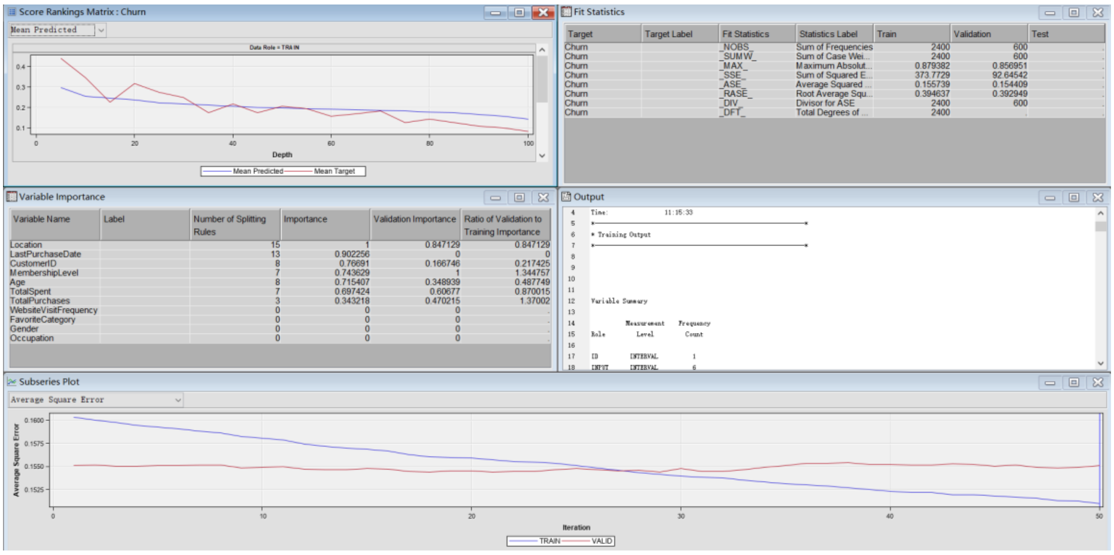




Run completed

Diagram: AA1

Path: Gradient Boosting



Fit Statistics

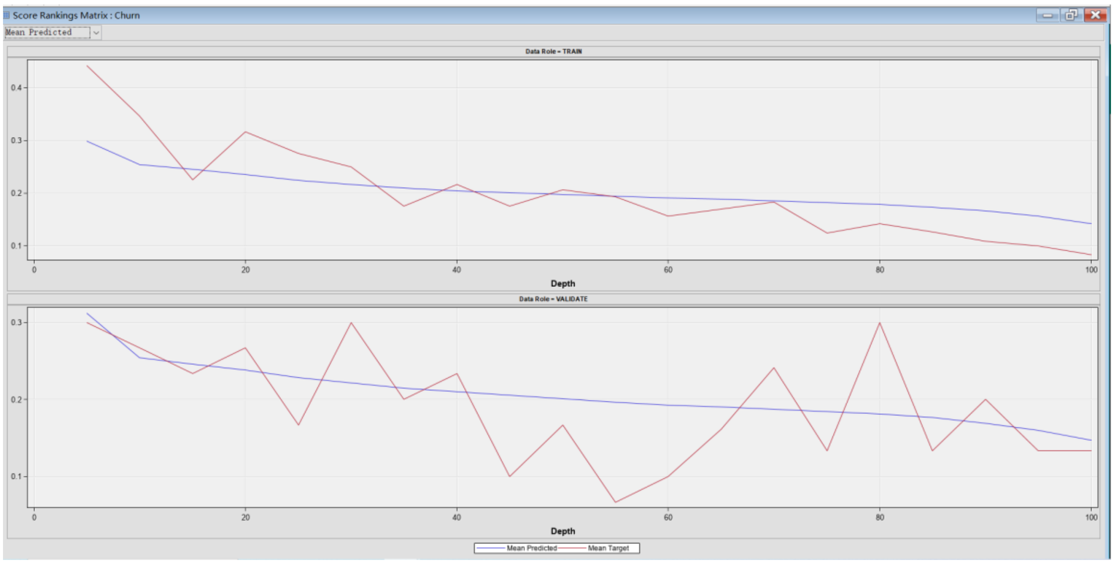
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Frequencies	2400	600	.
Churn		_SUMW_	Sum of Case Wei...	2400	600	.
Churn		_MAX_	Maximum Absolut...	0.879382	0.856951	.
Churn		_SSE_	Sum of Squared E...	373.7729	92.64542	.
Churn		_ASE_	Average Squared ...	0.155739	0.154409	.
Churn		_RASE_	Root Average Squ...	0.394637	0.392949	.
Churn		_DIV_	Divisor for ASE	2400	600	.
Churn		_DFT_	Total Degrees of ...	2400	.	.

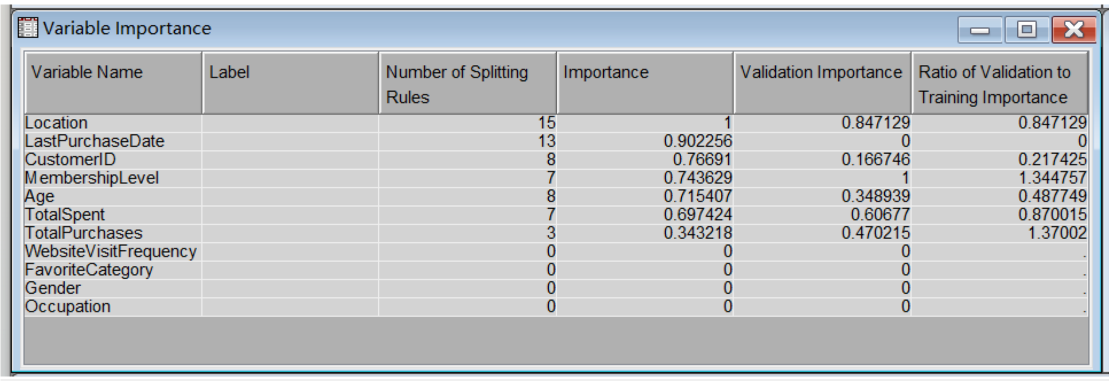
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Location		15	1	0.847129	0.847129
LastPurchaseDate		13	0.902256	0	0
CustomerID		8	0.76691	0.166746	0.217425
MembershipLevel		7	0.743629	1	1.344757
Age		8	0.715407	0.348939	0.487749
TotalSpent		7	0.697424	0.60677	0.870015
TotalPurchases		3	0.343218	0.470215	1.37002
WebsiteVisitFrequency		0	0	0	.
FavoriteCategory		0	0	0	.
Gender		0	0	0	.
Occupation		0	0	0	.

Output

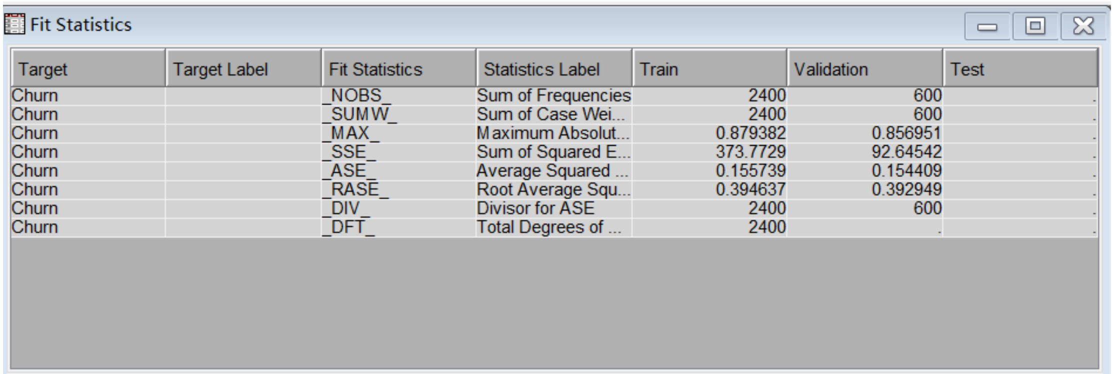
```
Time:            11:15:33
*------------------------------------------------------*
* Training Output
*------------------------------------------------------*




Variable Summary

           Measurement    Frequency
Role          Level         Count

ID          INTERVAL          1
INPUT       INTERVAL          6
```





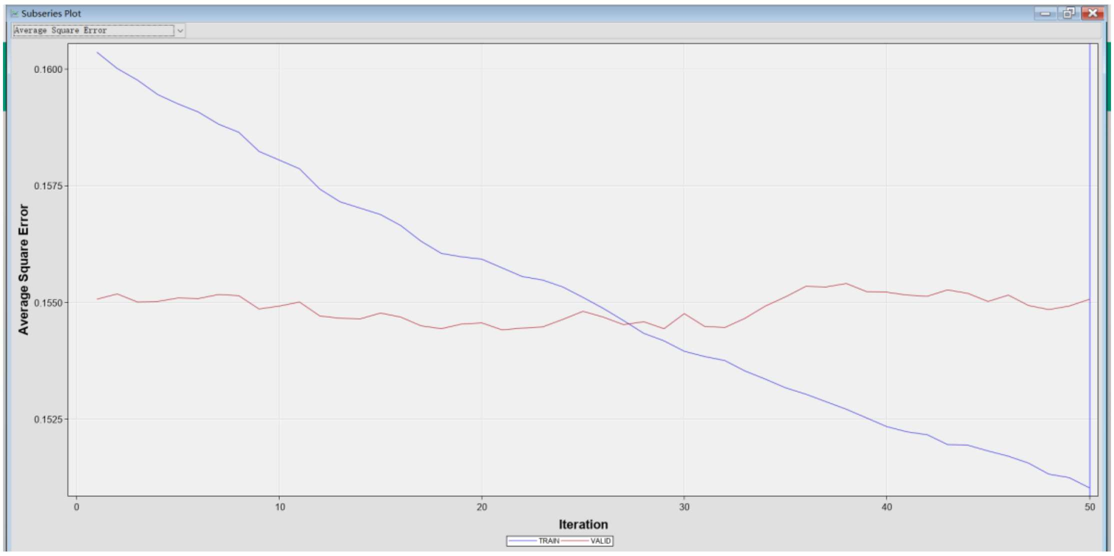
Variable Importance

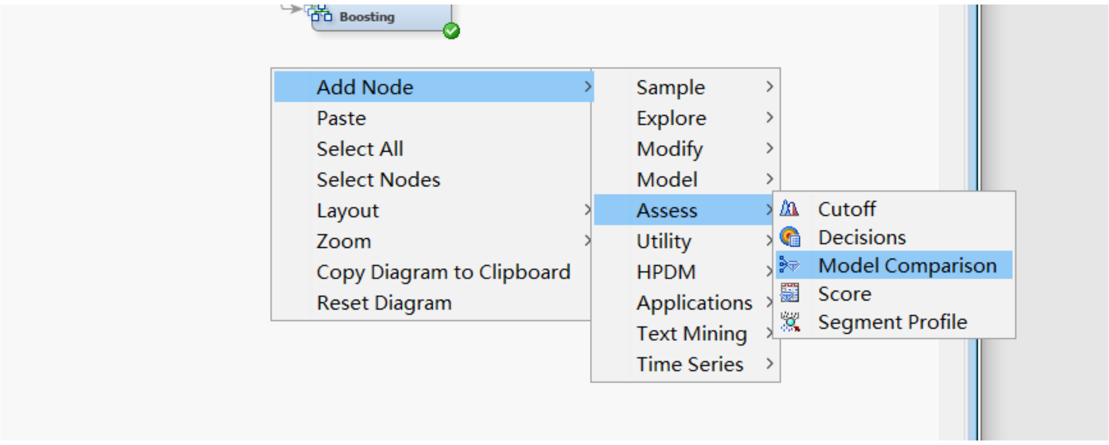
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Location		15	1	0.847129	0.847129
LastPurchaseDate		13	0.902256	0	0
CustomerID		8	0.76691	0.166746	0.217425
MembershipLevel		7	0.743629	1	1.344757
Age		8	0.715407	0.348939	0.487749
TotalSpent		7	0.697424	0.60677	0.870015
TotalPurchases		3	0.343218	0.470215	1.37002
WebsiteVisitFrequency		0	0	0	.
FavoriteCategory		0	0	0	.
Gender		0	0	0	.
Occupation		0	0	0	.

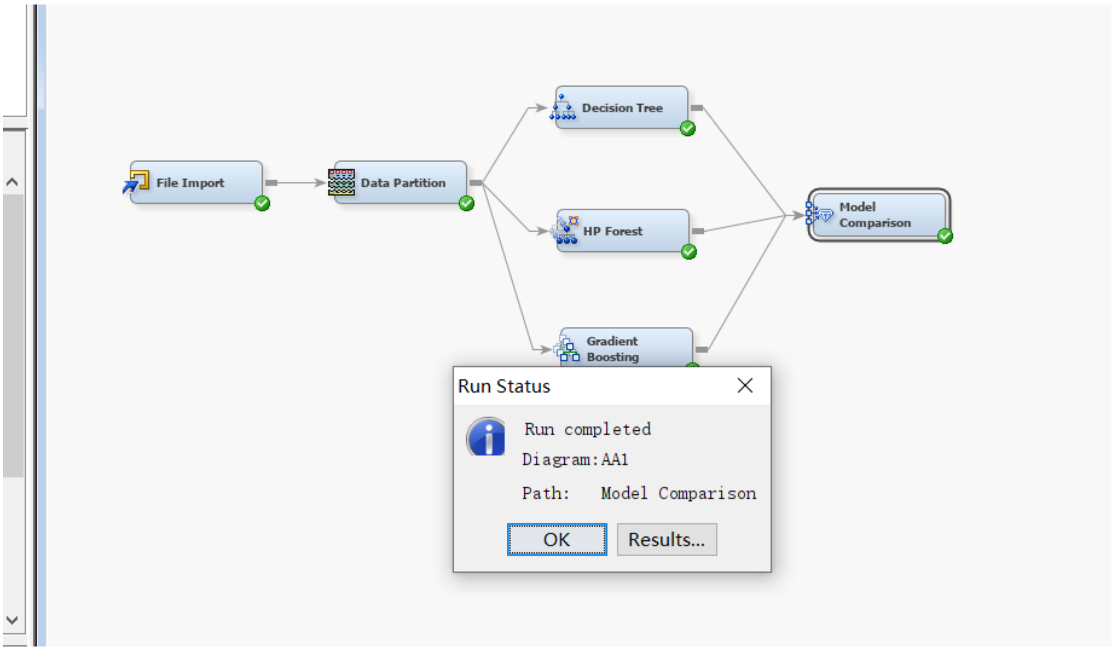


Fit Statistics

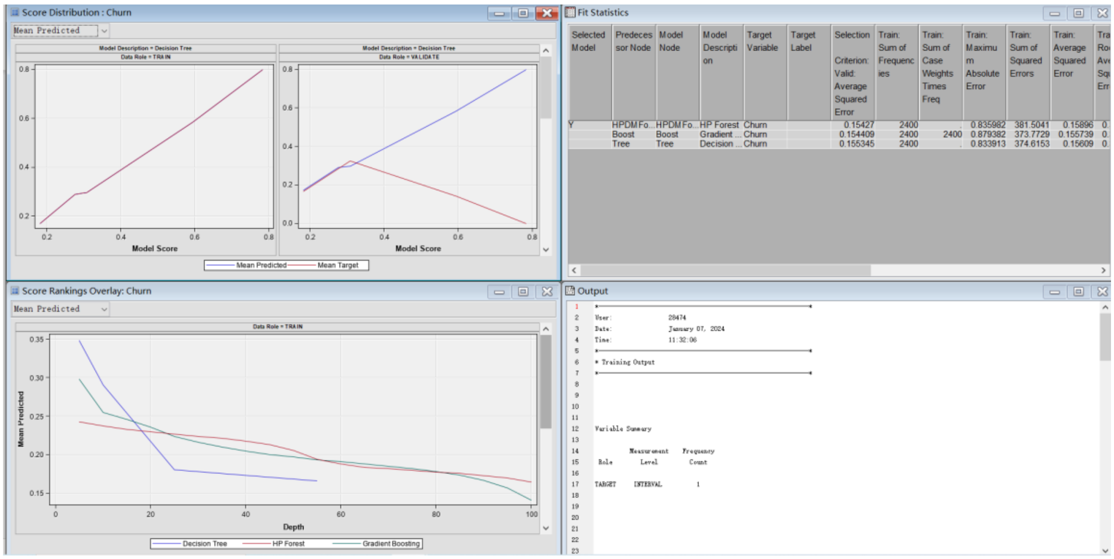
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Frequencies	2400	600	.
Churn		_SUMW_	Sum of Case Wei...	2400	600	.
Churn		_MAX_	Maximum Absolut...	0.879382	0.856951	.
Churn		_SSE_	Sum of Squared E...	373.7729	92.64542	.
Churn		_ASE_	Average Squared ...	0.155739	0.154409	.
Churn		_RASE_	Root Average Squ...	0.394637	0.392949	.
Churn		_DIV_	Divisor for ASE	2400	600	.
Churn		_DFT_	Total Degrees of ...	2400	.	.







Run completed
Diagram:AA1
Path: Model Comparison



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Sum of Case Weights Times Freq	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Tra Ro Ave Sq Err
Y	HPDMFo...	HPDMFo...	HP Forest	Churn		0.15427	2400	.	0.835982	381.5041	0.15896	0.
	Boost	Boost	Gradient ...	Churn		0.154409	2400	2400	0.879382	373.7729	0.155739	0.
	Tree	Tree	Decision ...	Churn		0.155345	2400	.	0.833913	374.6153	0.15609	0.

```
*------------------------------------------------------------*
User:                    28474
Date:                    January 07, 2024
Time:                    11:32:06
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

           Measurement    Frequency
 Role         Level         Count

TARGET     INTERVAL           1
```