MASTER OF DATA SCIENCE (SEMESTER 1 – 2023/2024) FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

WQD7005 DATA MINING GROUP ASSIGNMENT DATA EXPLORATION ON THE STORE SALES DATA

WQD 7005 Data Mining

INSTRUCTOR: PROF DR TEH YING WAH

https://github.com/uuqq3/FinalExam_NANXI

NAN XI

S2174013

Table of Content

◆ Dataset description:

This dataset represents customer transactions over the past year from an e-commerce platform. It contains various customer attributes, including demographic details, engagement metrics, and purchase history. The dataset has been enriched with additional relevant attributes to provide a comprehensive view of customer behavior. This allows for a multifaceted analysis of factors influencing purchase decisions, customer loyalty, and churn. It is structured to enable the identification of trends, patterns, and correlations within customer activity on the e-commerce site.

Attribute Table:

Attribute Name	Data Type	Description
CustomerID	String	A unique identifier for each customer.
Age	Integer	Age of the customer.
Gender	String	Gender of the customer (e.g., Male, Female).
Location	String	Geographical location of the customer.
MembershipLevel	String	Membership tier of the customer (e.g., Bronze, Silver, Gold).
TotalPurchases	Integer	Total number of purchases made by the customer over the last year.
TotalSpent	Float	Total amount of money spent by the customer over the last year.
FavoriteCategory	String	Most frequently purchased category by the customer.
LastPurchaseDate	Date	Date when the last purchase was made by the customer.
Occupation	String	Customer's occupation.
FrequencyOfWebsiteVisits	Integer	Number of times the customer visited the website over the last year.

Churn	Boolean	Indicator of whether the customer has churned (e.g., True, False).
Delivery_Charges	Float	Average delivery charges paid by the customer.
LastLogin	Date	The last date and time the customer logged into the website.
AccountCreatedDate	Date	The date the customer's account was created.
FeedbackScore	Integer	Customer's average feedback score for purchases.

◆ Analysis goal

The primary objective of this analysis is to understand and predict customer behaviors that lead to churn on the e-commerce platform. By examining various customer attributes and historical transaction data, the analysis aims to uncover patterns and trends that signal customer disengagement. The insights derived from this analysis will be instrumental in developing targeted strategies for customer retention, personalized marketing, and enhanced user experience. Ultimately, the goal is to leverage the data to implement data-driven decisions that can reduce customer churn rates, increase customer lifetime value, and optimize the overall customer journey on the e-commerce site.

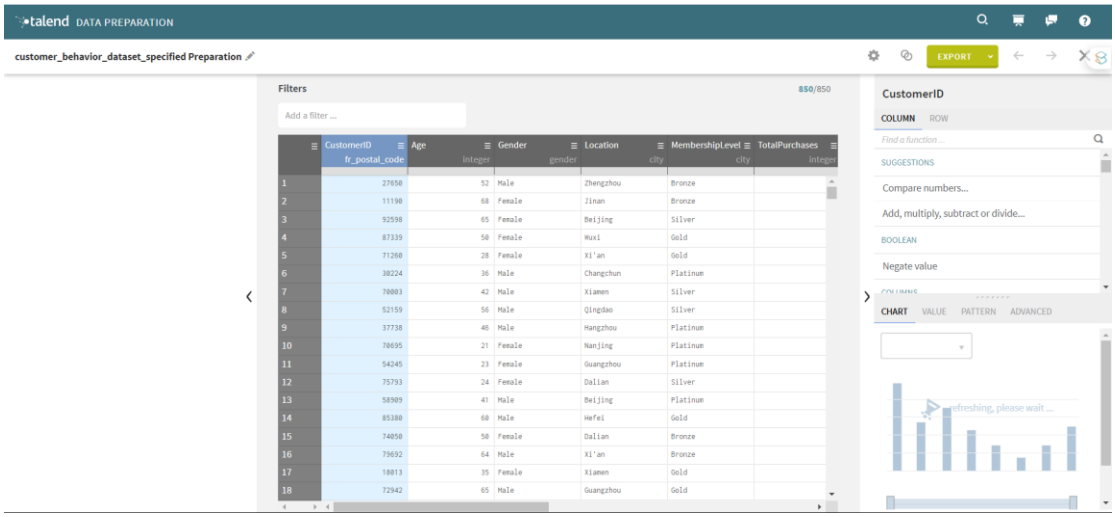
◆ Process

1. Data Import and Preprocessing

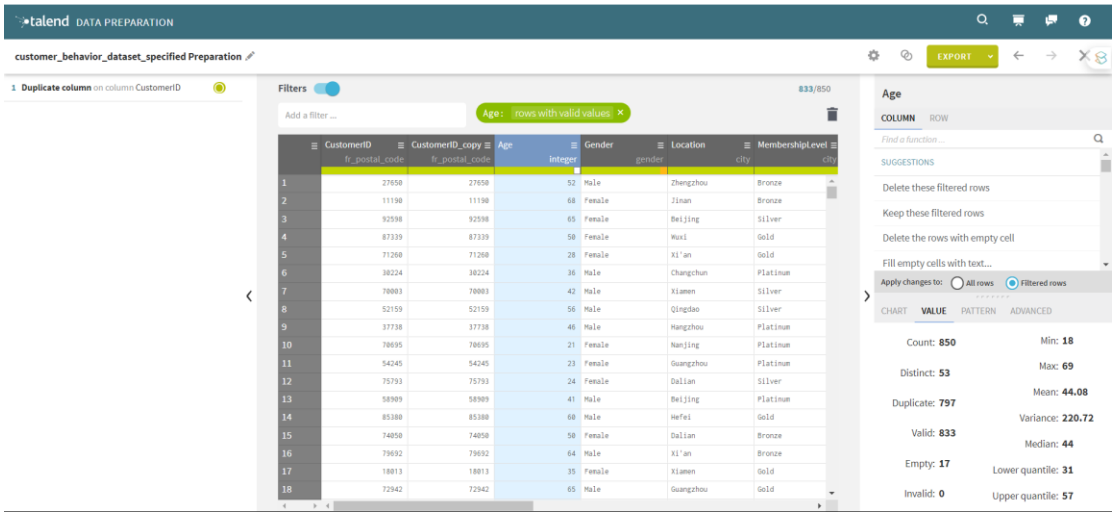
1.1 Use Talend Data Preparation (DP) to do preprocessing

1.1.1 Importing the Dataset:

- Open Talend Data Preparation.
- Create a new preparation and import your dataset (**customer_behavior_dataset_specified.csv**).



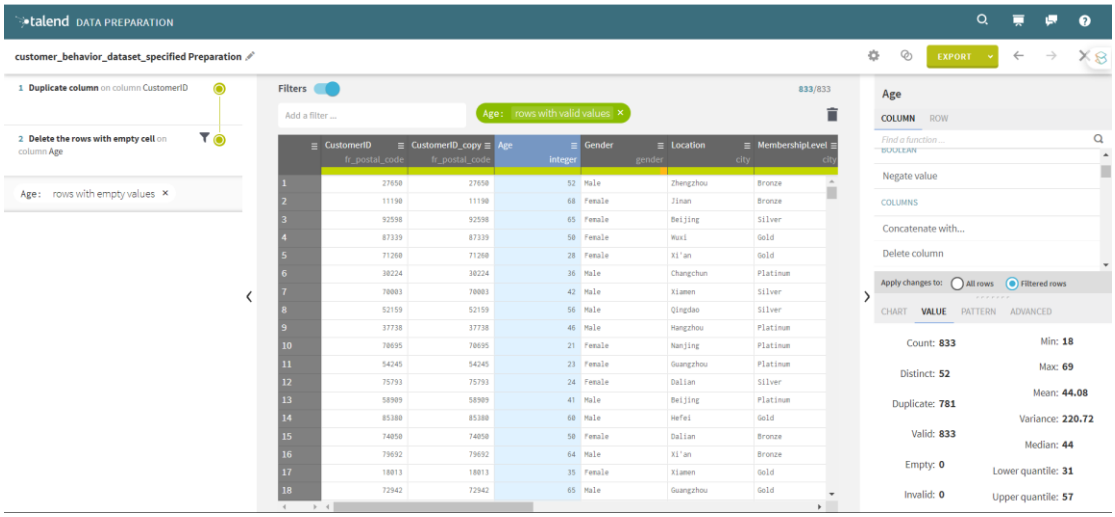
1.1.2 View all columns:



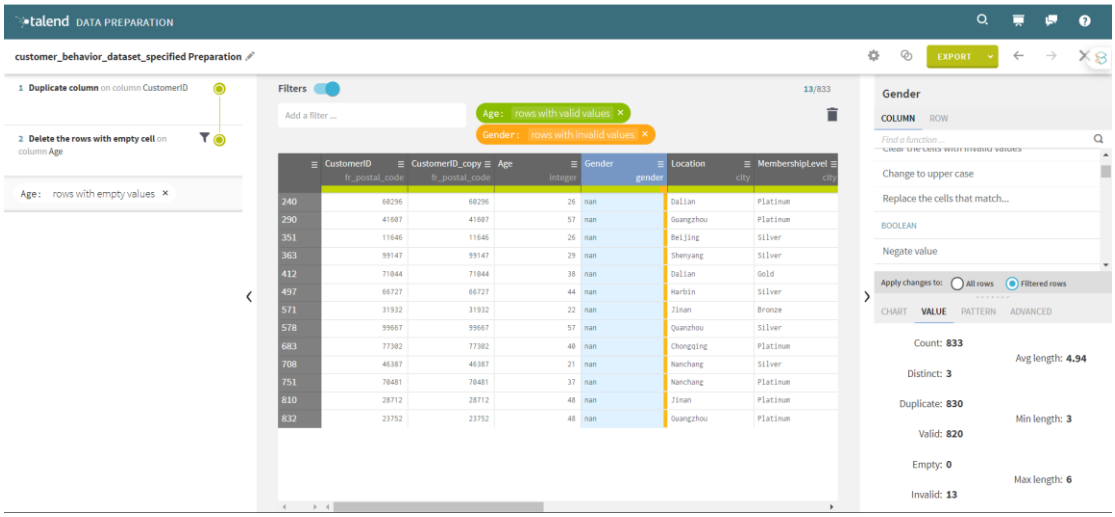
Only Age and Gender need to be preprocessed in the dataset.

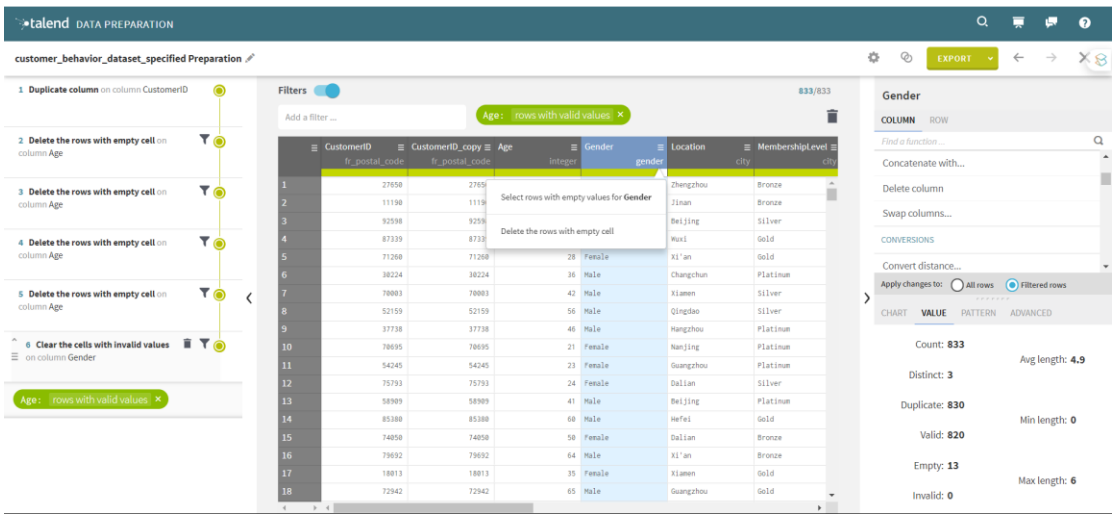
1.1.3 Age

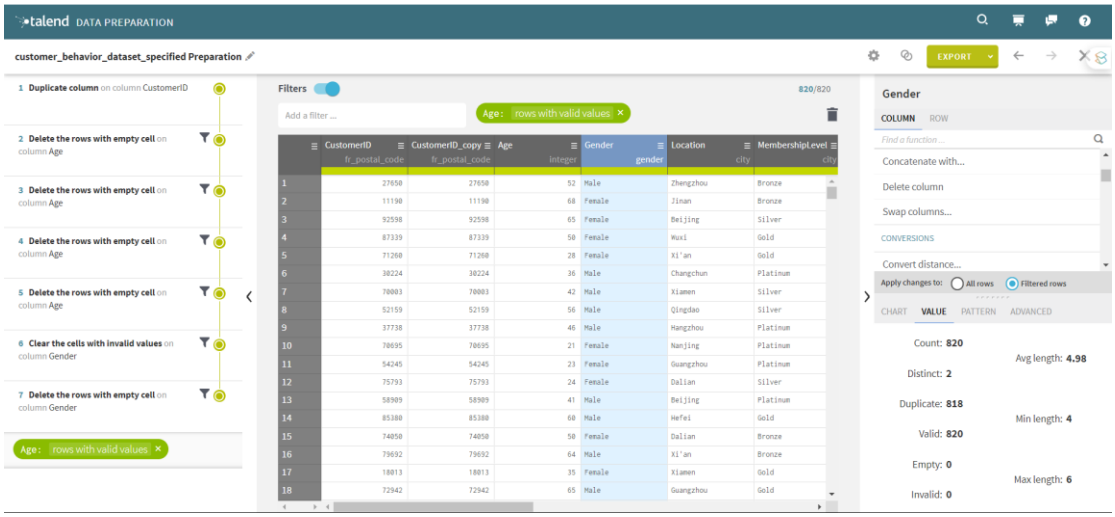
There are 17 missing values in Age. We choose to delete all.


talend DATA PREPARATION
customer_behavior_dataset_specified Preparation
EXPORT
1 Duplicate column on column CustomerID
2 Delete the rows with empty cell on column Age
Age: rows with empty values
Filters
Add a filter ...
Age: rows with valid values
833/833
CustomerID
CustomerID_copy
Age
Gender
Location
MembershipLevel
Age
COLUMN
ROW
Negate value
COLUMNS
Concatenate with...
Delete column
Apply changes to: All rows Filtered rows
CHART
VALUE
PATTERN
ADVANCED
Count: 833
Distinct: 52
Duplicate: 781
Valid: 833
Empty: 0
Invalid: 0
Min: 18
Max: 69
Mean: 44.08
Variance: 220.72
Median: 44
Lower quantile: 31
Upper quantile: 57

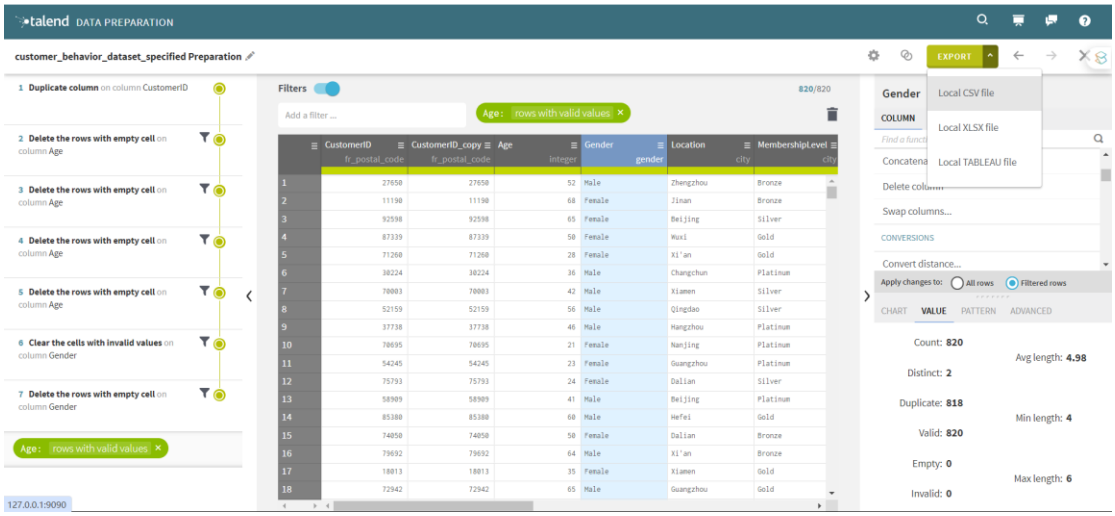
1.1.4 Gender


talend DATA PREPARATION
customer_behavior_dataset_specified Preparation
EXPORT
1 Duplicate column on column CustomerID
2 Delete the rows with empty cell on column Age
Age: rows with empty values
Filters
Add a filter ...
Age: rows with valid values
Gender: rows with invalid values
13/833
CustomerID
CustomerID_copy
Age
Gender
Location
MembershipLevel
Gender
COLUMN
ROW
Change to upper case
Replace the cells that match...
BOOLEAN
Negate value
Apply changes to: All rows Filtered rows
CHART
VALUE
PATTERN
ADVANCED
Count: 833
Distinct: 3
Duplicate: 830
Valid: 820
Empty: 0
Invalid: 13
Avg length: 4.94
Min length: 3
Max length: 6

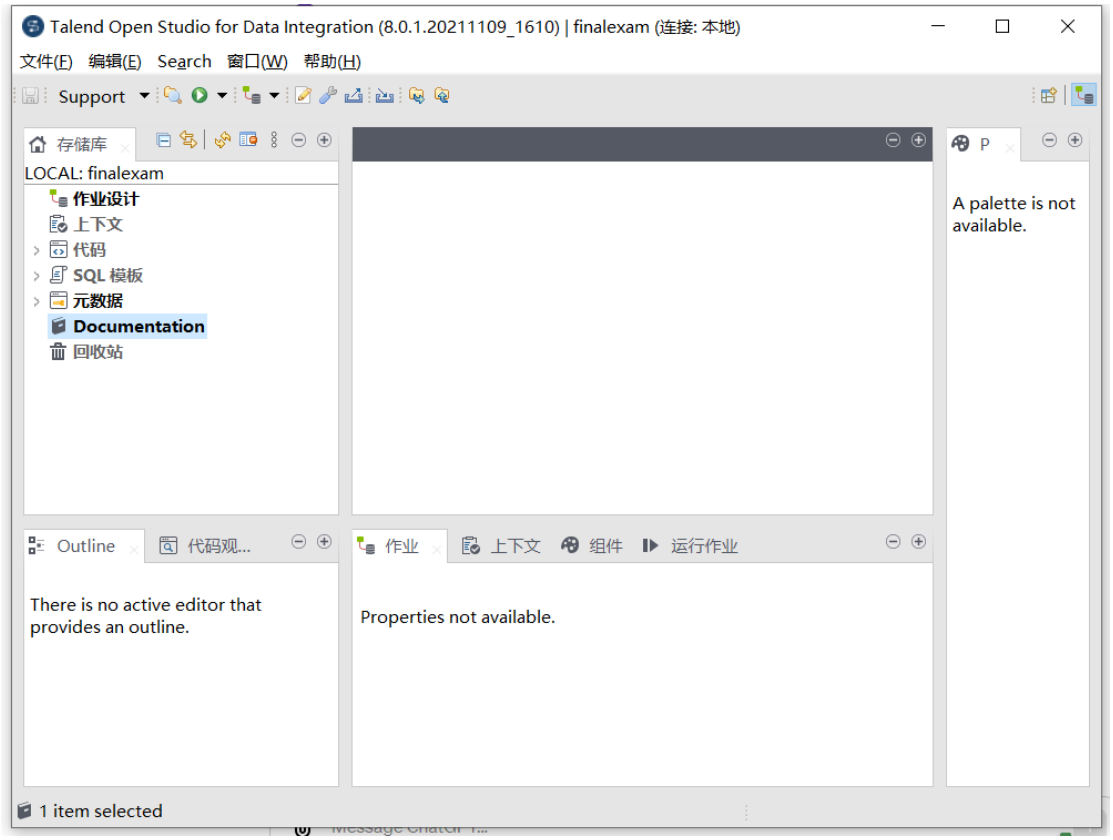

talend DATA PREPARATION
customer_behavior_dataset_specified Preparation
EXPORT
1 Duplicate column on column CustomerID
2 Delete the rows with empty cell on column Age
3 Delete the rows with empty cell on column Age
4 Delete the rows with empty cell on column Age
5 Delete the rows with empty cell on column Age
6 Clear the cells with invalid values on column Gender
Age: rows with valid values
Filters
Add a filter ...
Age: rows with valid values
833/833
CustomerID
CustomerID_copy
Age
Gender
Location
MembershipLevel
Select rows with empty values for Gender
Delete the rows with empty cell
Gender
COLUMN
ROW
Concatenate with...
Delete column
Swap columns...
CONVERSIONS
Convert distance...
Apply changes to: All rows Filtered rows
CHART
VALUE
PATTERN
ADVANCED
Count: 833
Distinct: 3
Duplicate: 830
Valid: 820
Empty: 13
Invalid: 0
Avg length: 4.9
Min length: 0
Max length: 6

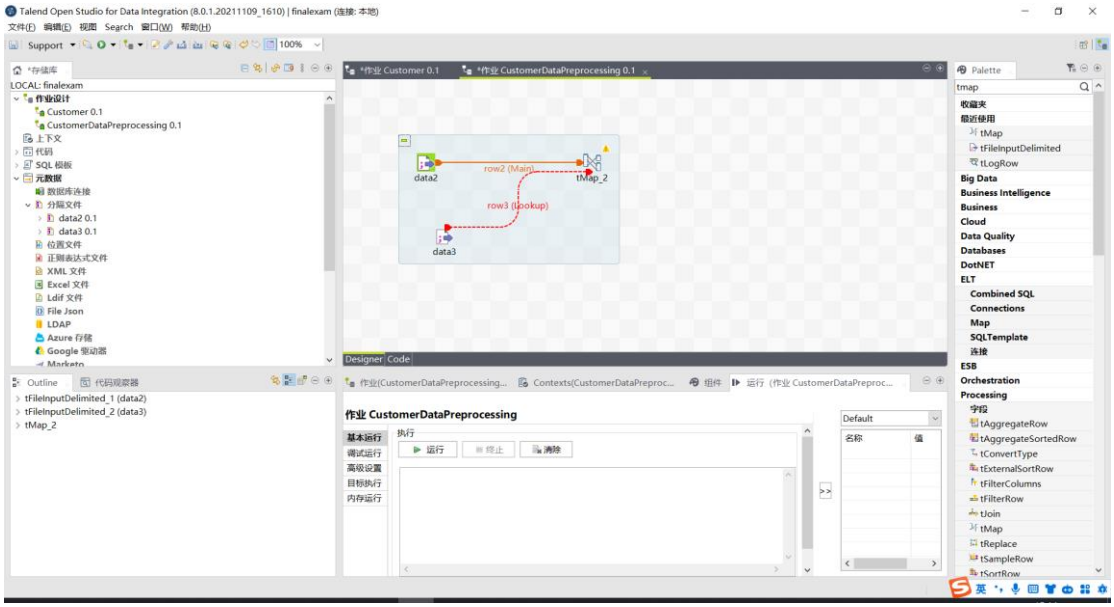


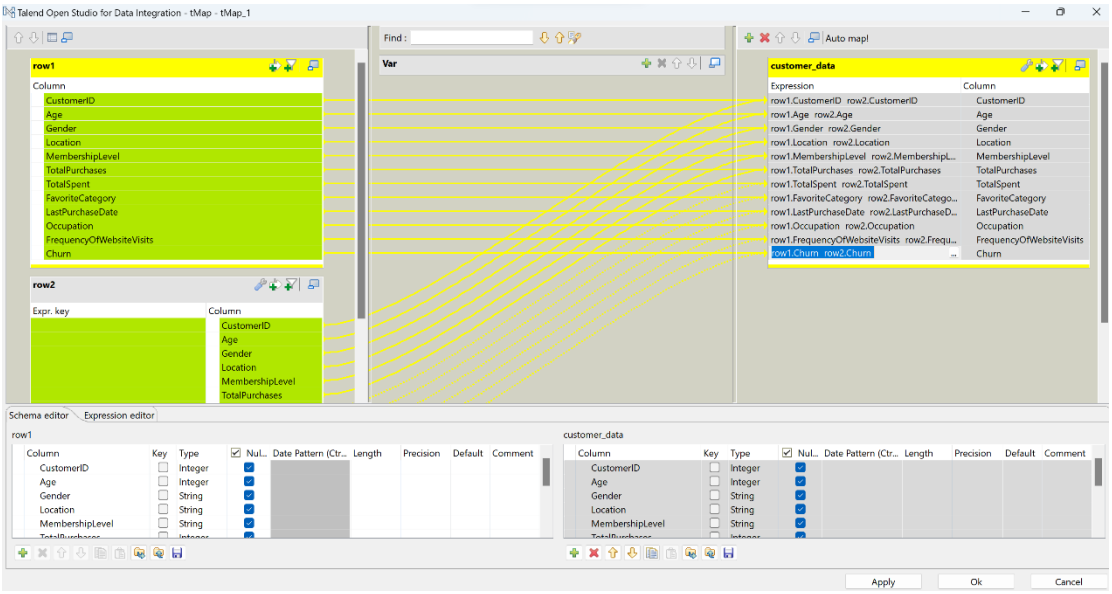
1.1.5 Export the Prepared Data:

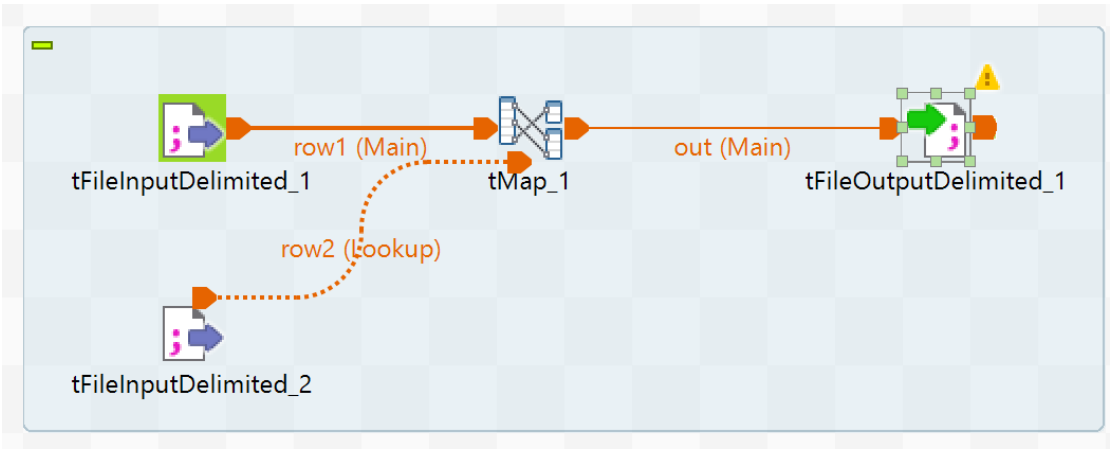


1.2 Use Talend Data Integration (DI) to do preprocessing: Integration


Talend Open Studio for Data Integration (8.0.1.20211109_1610) | finalexam (连接: 本地)
文件(F) 编辑(E) Search 窗口(W) 帮助(H)
Support
存储库
LOCAL: finalexam
作业设计
上下文
代码
SQL 模板
元数据
Documentation
回收站
P
A palette is not available.
Outline
代码观...
There is no active editor that provides an outline.
作业
上下文
组件
运行作业
Properties not available.
1 item selected



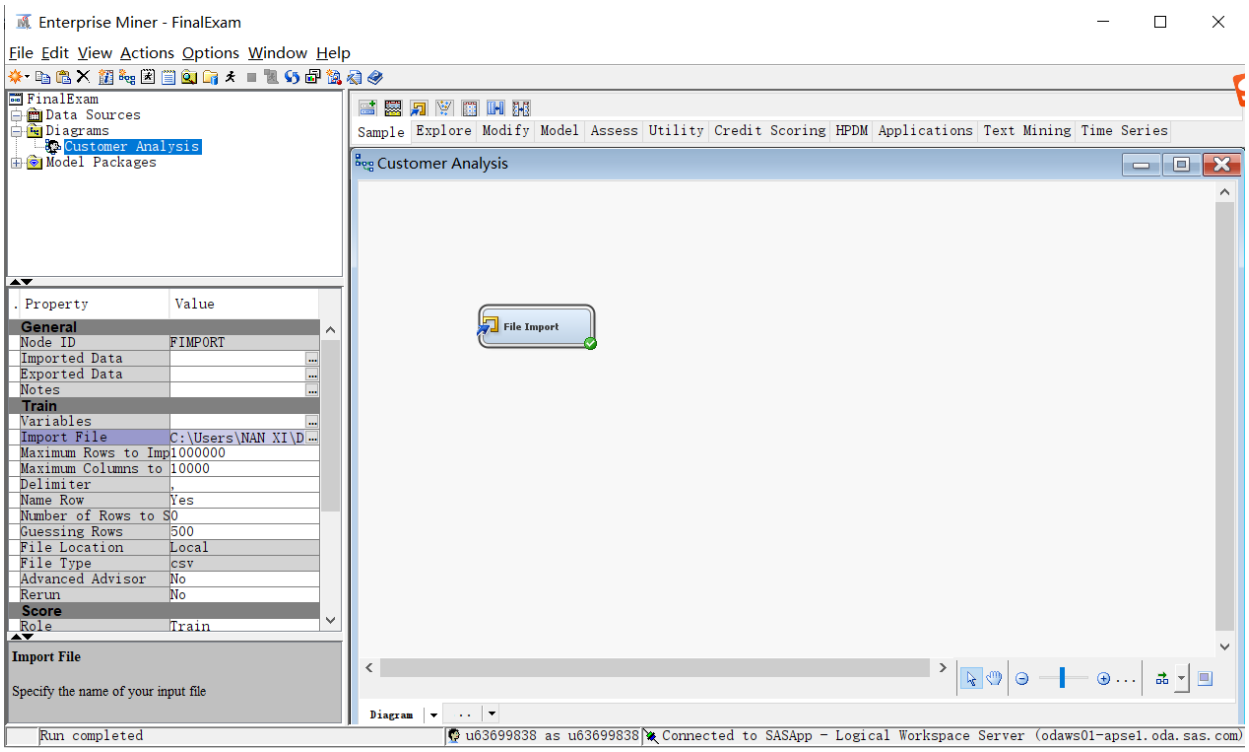




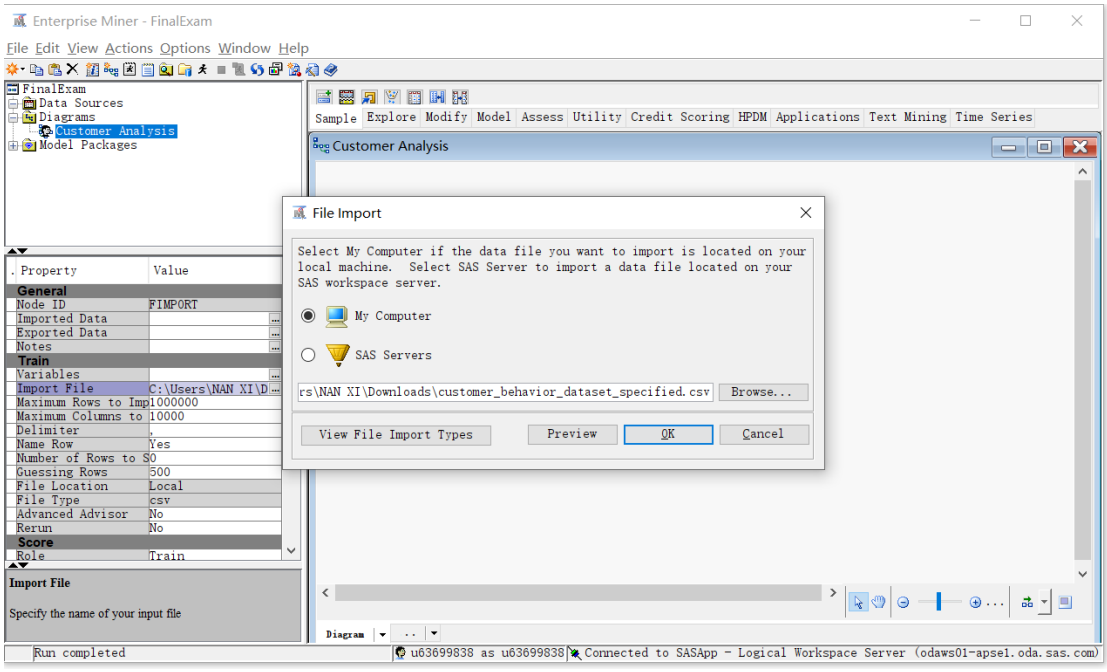
1.3 Use SEM to do preprocessing

1.3.1 Data Import in SAS Enterprise Miner:

1.3.1.1 Create “Customer Analysis” diagram:

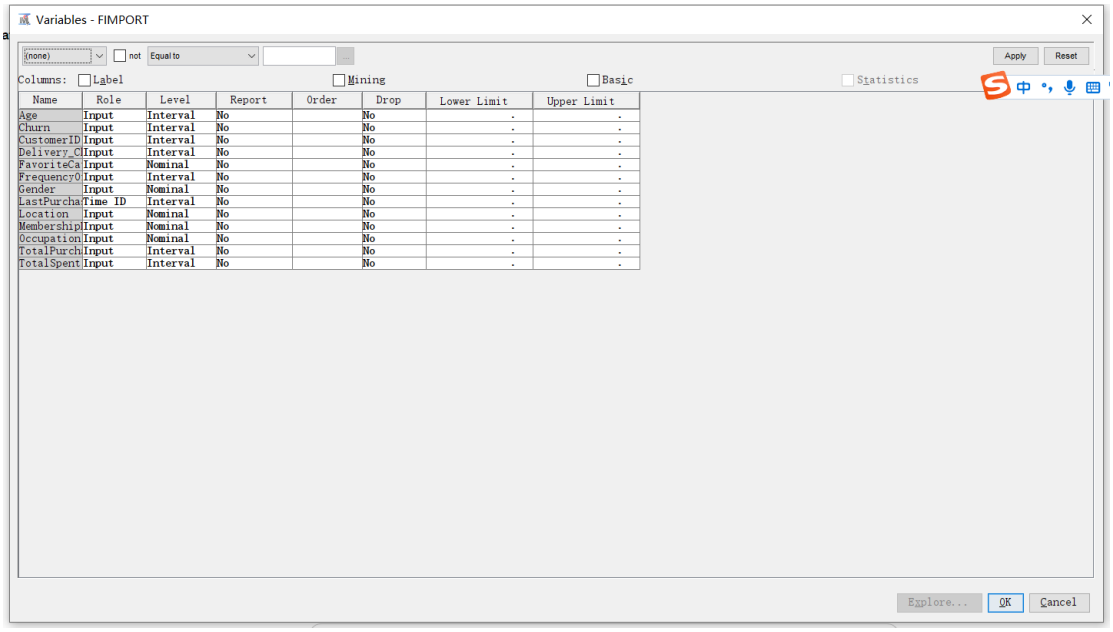


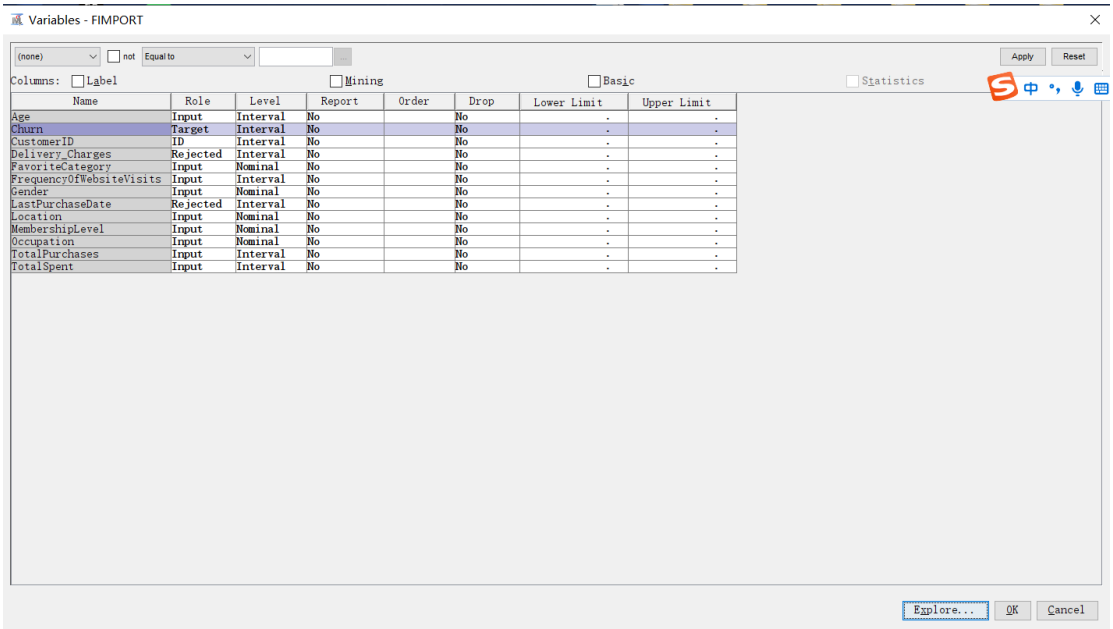
1.3.1.2 Import dataset file:



1.3.2 Specify Variable Roles:

Assign roles to each variable (input, target, ID, etc.).


Variables - FIMPORT
(none)
not
Equal to
Apply
Reset
Columns:
Label
Mining
Basic
Statistics
Name
Role
Level
Report
Order
Drop
Lower Limit
Upper Limit
Age Input Interval No No . .
Churn Input Interval No No . .
CustomerID Input Interval No No . .
Delivery_Ch Input Interval No No . .
FavoriteCa Input Nominal No No . .
Frequency0 Input Interval No No . .
Gender Input Nominal No No . .
LastPurcha Time ID 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 . .
Explore...
OK
Cancel


Variables - FIMPORT
(none)
not
Equal to
Apply
Reset
Columns:
Label
Mining
Basic
Statistics
Name
Role
Level
Report
Order
Drop
Lower Limit
Upper Limit
Age Input Interval No No . .
Churn Target Interval No No . .
CustomerID ID Interval No No . .
Delivery_Charges Rejected Interval No No . .
FavoriteCategory Input Nominal No No . .
FrequencyOfWebsiteVisits Input Interval No No . .
Gender Input Nominal No No . .
LastPurchaseDate Rejected Interval No No . .
Location Input Nominal No No . .
MembershipLevel Input Nominal No No . .
Occupation Input Nominal No No . .
TotalPurchases Input Interval No No . .
TotalSpent Input Interval No No . .
Explore...
OK
Cancel

Columns: Label

Name	Role	Level
Age	Input	Interval
Churn	Target	Interval
CustomerID	ID	Interval
Delivery_Charges	Rejected	Interval
FavoriteCategory	Input	Nominal
FrequencyOfWebsiteVisits	Input	Interval
Gender	Input	Nominal
LastPurchaseDate	Rejected	Interval
Location	Input	Nominal
MembershipLevel	Input	Nominal
Occupation	Input	Nominal
TotalPurchases	Input	Interval
TotalSpent	Input	Interval

Churn as “Target”

CustomerID as “ID”

Delivery_Charges as “Rejected”

LastPurchaseDate as “Rejected”

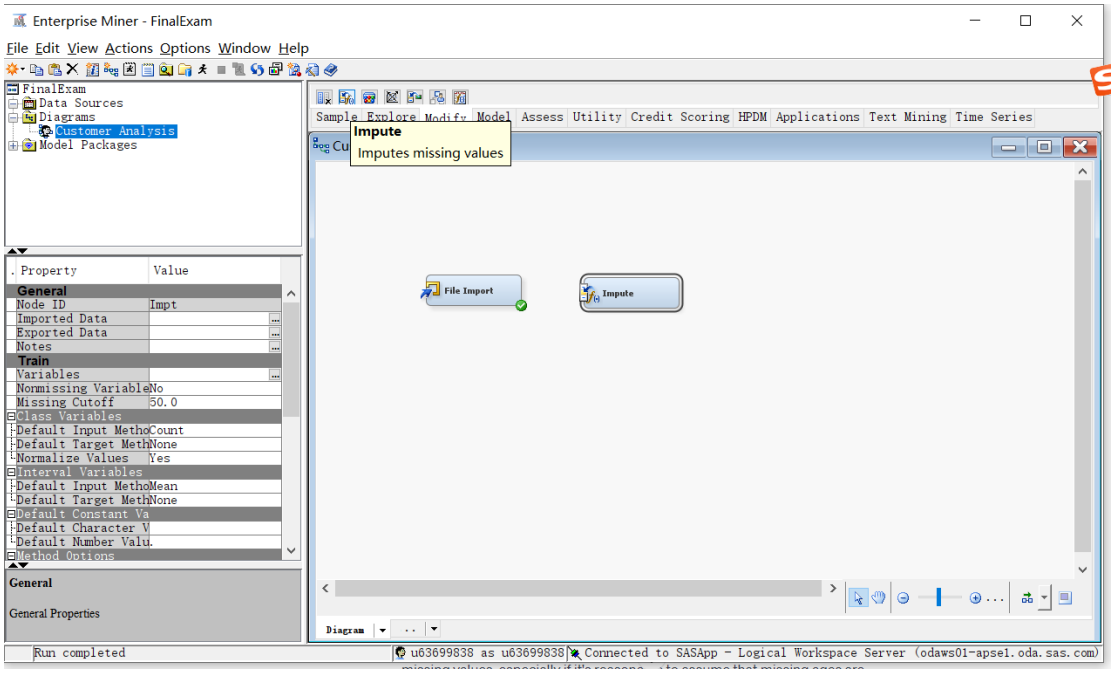
Finally, we get the dataset like this:

Exported Attributes for TRAIN Port

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	INTERVAL	4
INPUT	NOMINAL	5
REJECTED	INTERVAL	2
TARGET	INTERVAL	1

1.3.3 Handling Missing Values: Identify missing values and decide how to handle them (e.g., imputation).

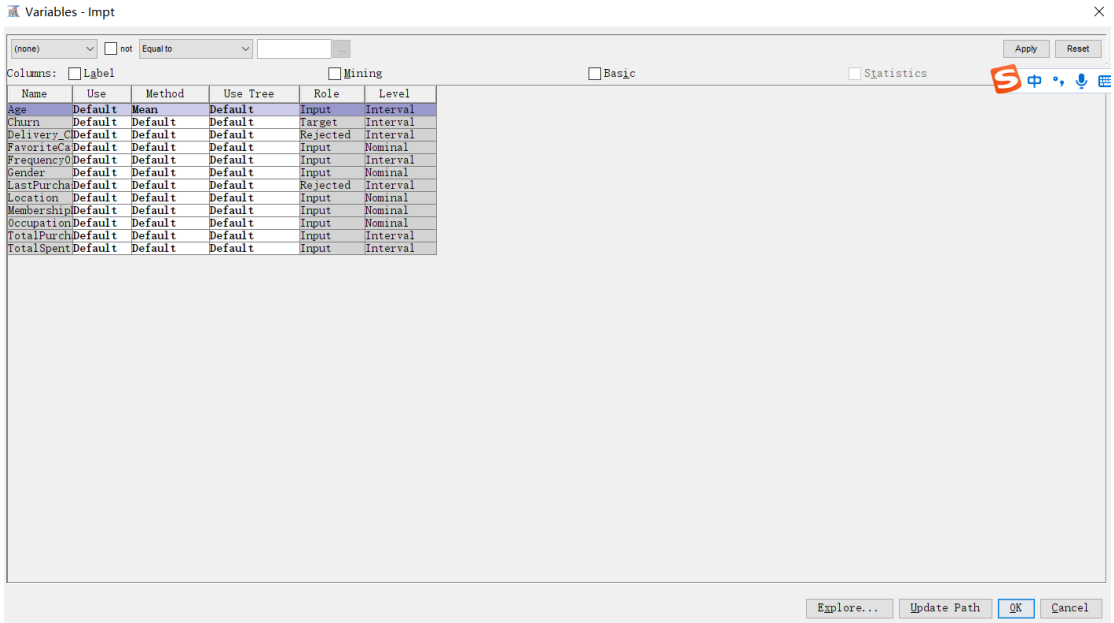
1.3.3.1 Drag Impute node:



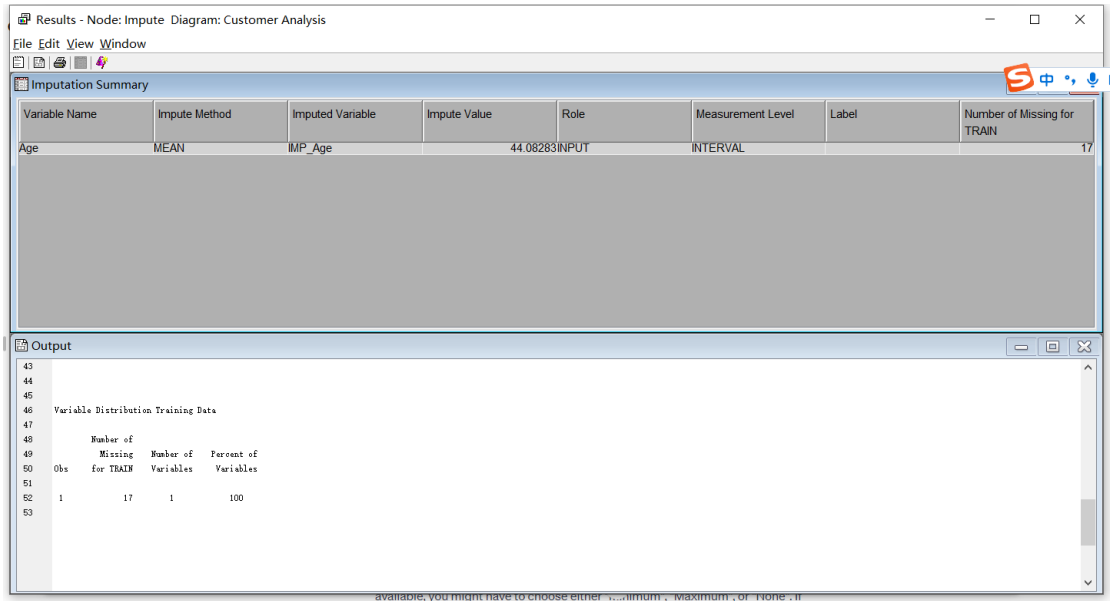
1.3.3.2 Handle missing values：

For general checking of our dataset, we can see “Age” column has some missing value, which is an interval variable, we might choose "Mean" to handle.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	44.08283	14.85666	833	17	18	44	69	-0.04532	-1.18297
FrequencyOfWebsiteVisits	INPUT	14.75647	8.418429	850	0	1	15	29	0.007883	-1.19755
TotalPurchases	INPUT	5.225882	2.224881	850	0	0	5	13	0.293118	-0.24513
TotalSpent	INPUT	1035.677	1005.276	850	0	0.67	693.59	6864.87	1.608983	3.11109
Churn	TARGET	0.304706	0.460554	850	0	0	0	1	0.850084	-1.28038

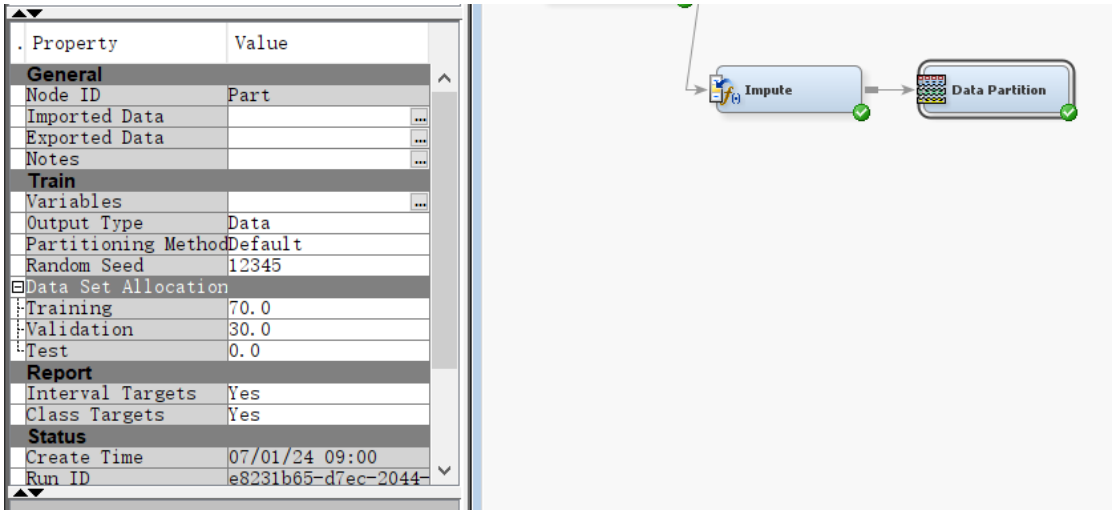


1.3.3 Results



2. Decision Tree Analysis

2.1 Data partition:



Output

Summary Statistics for Interval Targets

Data=DATA

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Churn	1	0.3047058824	0	850	0	0.4605537412	

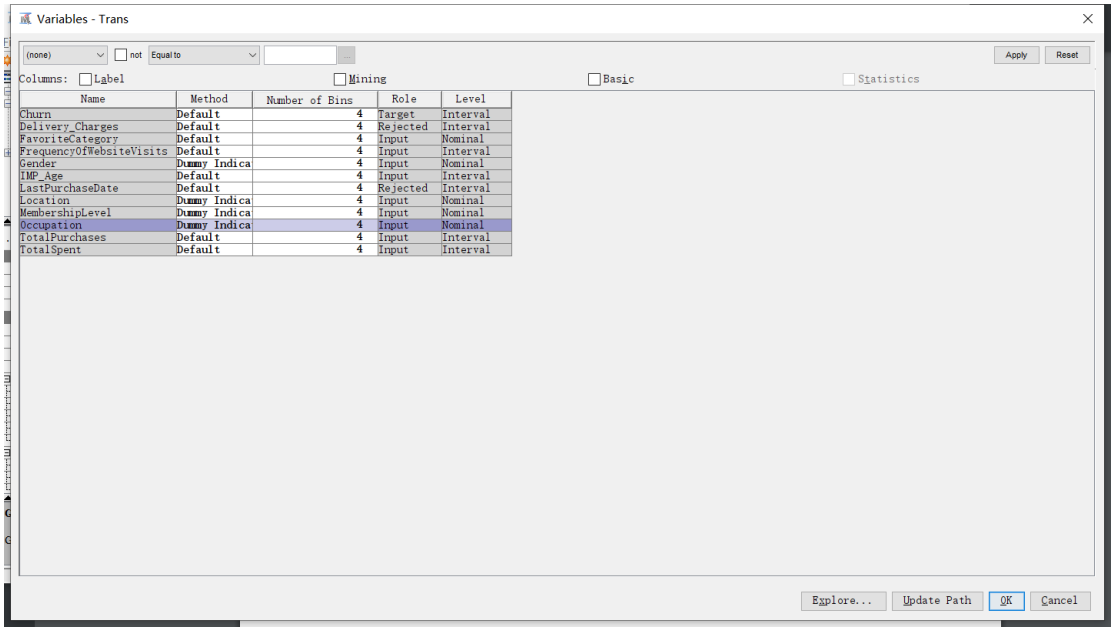
Data=TRAIN

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Churn	1	0.3042016807	0	595	0	0.4604555931	

Data=VALIDATE

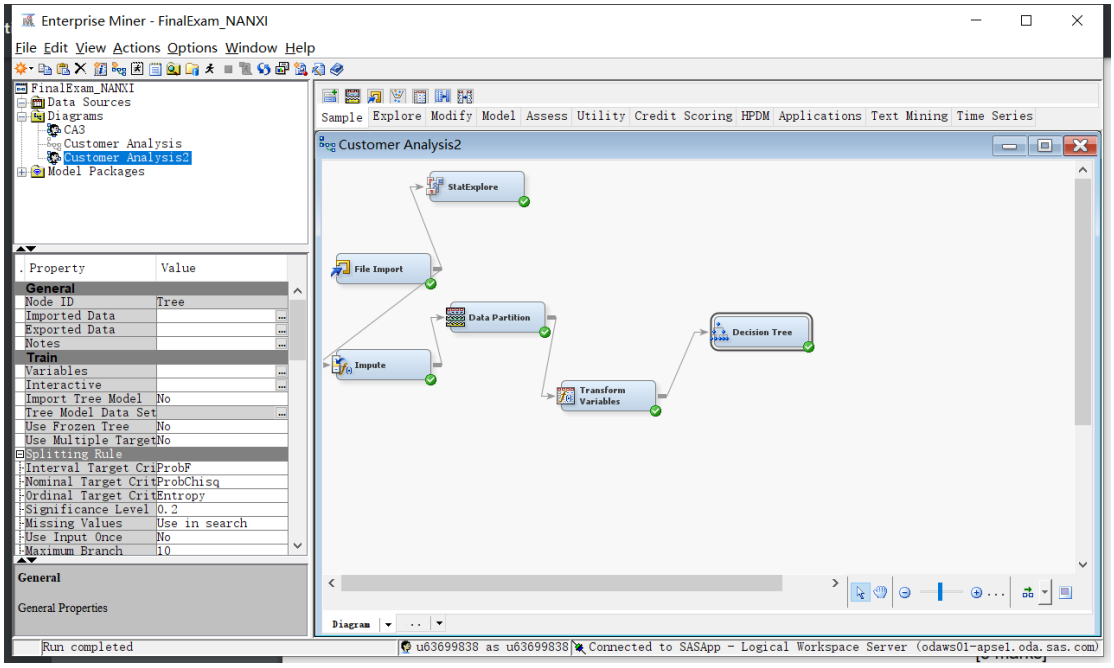
Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Churn	1	0.3058823529	0	255	0	0.4616862983	

2.2 Transformation:



2.3 Creating a Decision Tree Model: use the Decision Tree node. We need to select

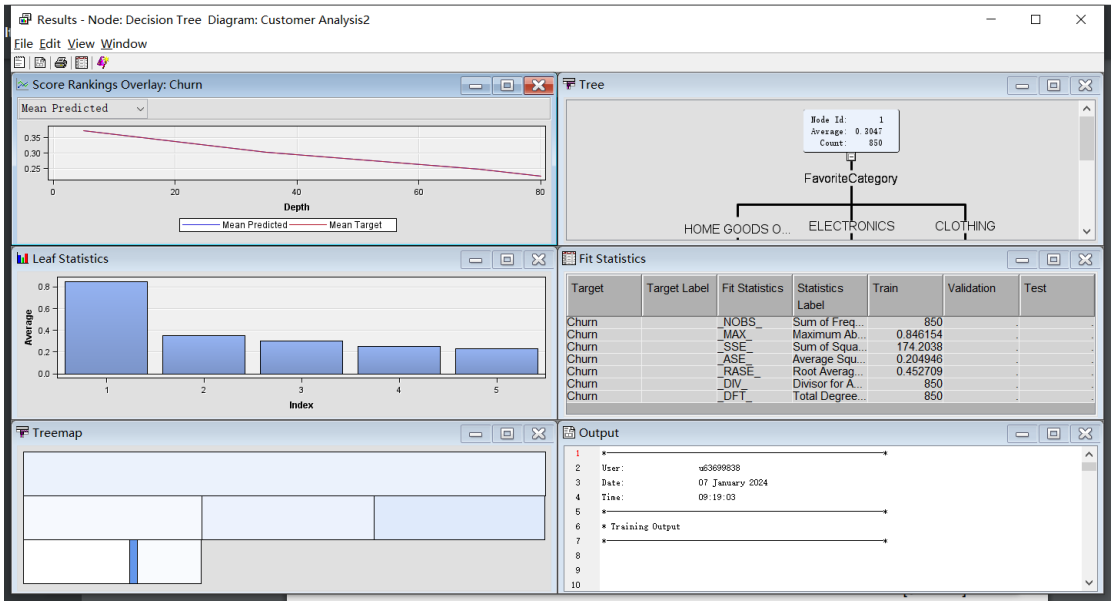
the appropriate criteria for splitting nodes and determine the depth of the tree.



Property	Value
General	
Node ID	Tree
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Target	No
Splitting Rule	
Interval Target Cri	ProbF
Nominal Target Crit	ProbChisq
Ordinal Target Crit	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	10
Maximum Depth	10
Minimum Categorical	5

Node	
Leaf Size	8
Number of Rules	5
Number of Surrogate	4
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Valid	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based I	
Observation Based I	No
Number Single Var I	5
P-Value Adjustment	
Bonferroni Adjustme	Yes
Time of Bonferroni	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes

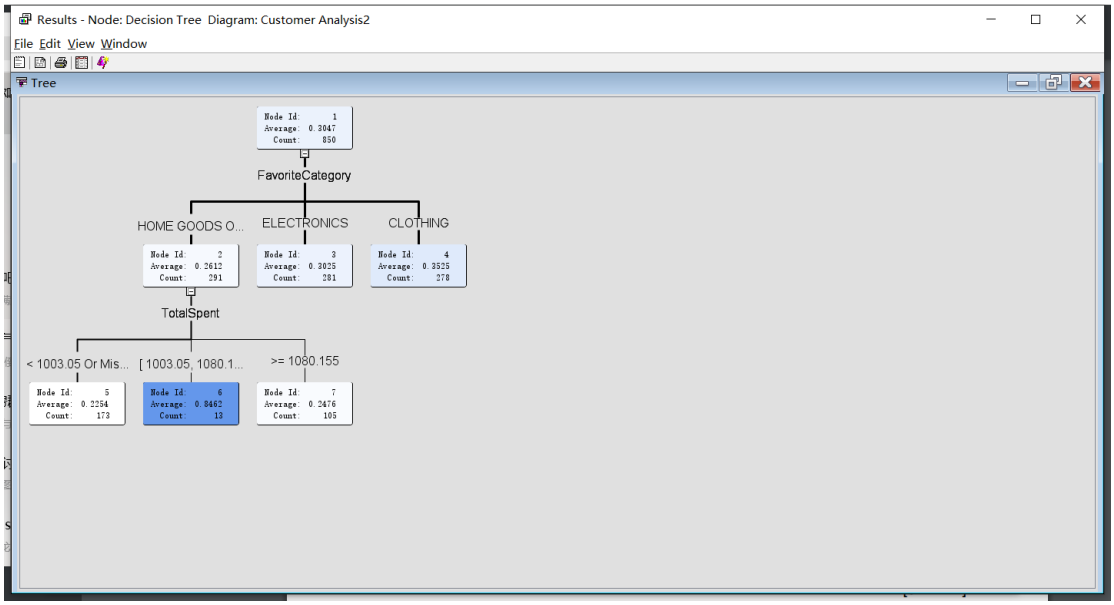
2.4 Results Analysis:



Decision Tree Score Rankings Overlay

The Score Rankings Overlay graph shows the mean predicted probability of churn against the depth of the tree. The lines represent the mean predicted values and the mean target values. The close alignment of the lines may indicate a good fit of the model to the data.

The graph is used to assess the performance of the decision tree at different depths, helping to determine the optimal complexity of the model.



Tree Overview:

- The tree starts with a root node (Node Id: 1) which represents the entire population of the dataset consisting of 850 observations.
- The churn rate for the entire dataset at this root node is 0.3047, which means approximately 30.47% of the customers are predicted to churn according to the model.

Splitting Criteria:

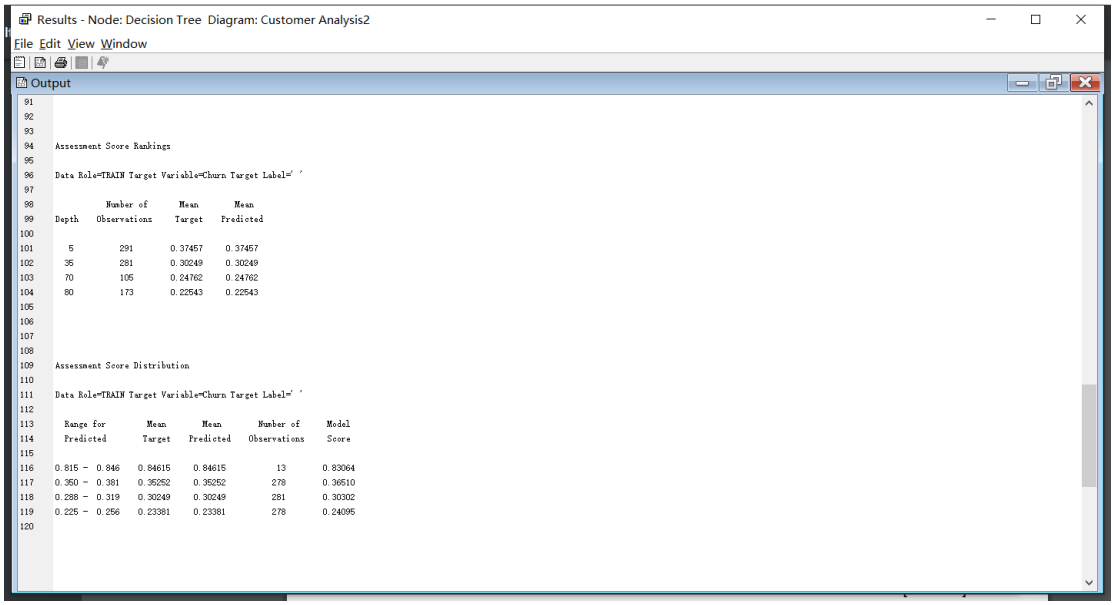
- The first split is based on the variable 'FavoriteCategory', dividing the dataset into groups based on the category of products the customers prefer.
- Each subsequent node represents a subgroup of the data and is further split based on different criteria.

Node Details:

- For example, Node Id: 2, which represents customers with a preference for HOME GOODS OR MISCELLANEOUS, has a lower average churn rate of 0.2612 and contains 291 observations. This suggests that customers who favor home goods or miscellaneous items are less likely to churn compared to the average churn rate of the entire dataset.
- Conversely, Node Id: 6 has a much higher average churn rate of 0.8462 but is based on a much smaller segment of the population (13 observations). This indicates a subset of customers with a high likelihood of churn, which might warrant further investigation for targeted retention strategies.

Analysis:

- The decision tree provides a granular view of how different customer segments have varying risks of churn. The segments are identified by their spending habits, product preferences, and other variables captured in the dataset.
- The tree can be used to inform customer retention strategies. For instance, by targeting segments identified as high-risk for churn with specific interventions.
- Additionally, the simplicity of the decision tree allows for easy interpretation and identification of key variables that contribute to churn.




```
Assessment Score Rankings

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

            Number of       Mean       Mean
Depth     Observations     Target   Predicted

    5            291    0.37457    0.37457
   35            281    0.30249    0.30249
   70            105    0.24762    0.24762
   80            173    0.22543    0.22543



Assessment Score Distribution

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

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

0.815 - 0.846    0.84615    0.84615            13       0.83064
0.350 - 0.381    0.35252    0.35252           278       0.36510
0.288 - 0.319    0.30249    0.30249           281       0.30302
0.225 - 0.256    0.23381    0.23381           278       0.24095
```

Assessment Score Rankings:

- The "Assessment Score Rankings" detail the model performance at various tree depths. For each depth, the mean target and mean predicted values are closely aligned, which suggests the model has consistent predictive accuracy across different levels of complexity.
    - At depth 5, there are 291 observations with a mean churn rate (target) and a mean predicted probability of churn of 0.37457. This indicates that at a shallow depth, the model already captures the churn rate quite well.
    - As the tree depth increases to 35, with 281 observations, the mean values

remain unchanged at 0.30249, suggesting that additional splits have not significantly changed the predictive performance for this subset.

- At depth 70, there are 105 observations with mean values of 0.24762, showing a slight decrease in both target and predicted probabilities.
- The deepest level shown, depth 80, with 173 observations, has the lowest mean values at 0.22543. This decrease in churn prediction could indicate overfitting at higher depths or could reflect a segment of the population with inherently lower churn rates.

Assessment Score Distribution:

- This section provides a distribution of the model's predictive performance across different probability ranges.
    - For the highest probability range (0.815 - 0.846), the mean predicted churn is 0.84615, which is the highest among all intervals, with a model score of 0.83064. However, this is based on only 13 observations, which suggests that while the model is confident about a high churn probability for this group, it represents a small segment.
    - The next interval (0.350 - 0.381) has a mean predicted value of 0.35252 with a model score of 0.36510, based on 278 observations. This is a larger group with moderate churn probability.
    - The interval (0.288 - 0.319), also with 281 observations, shows a mean predicted value of 0.30249 and a model score of 0.30302.
    - Finally, the lowest interval (0.225 - 0.256) presents a mean predicted churn of 0.23381 with a model score of 0.24095, again based on 278 observations, indicating the model's conservative estimate of churn probability for this segment.

Business Insights from Analysis:

- **Favorite Category Impact**: The initial split on 'FavoriteCategory' indicates this variable is influential in predicting churn. Customers preferring certain

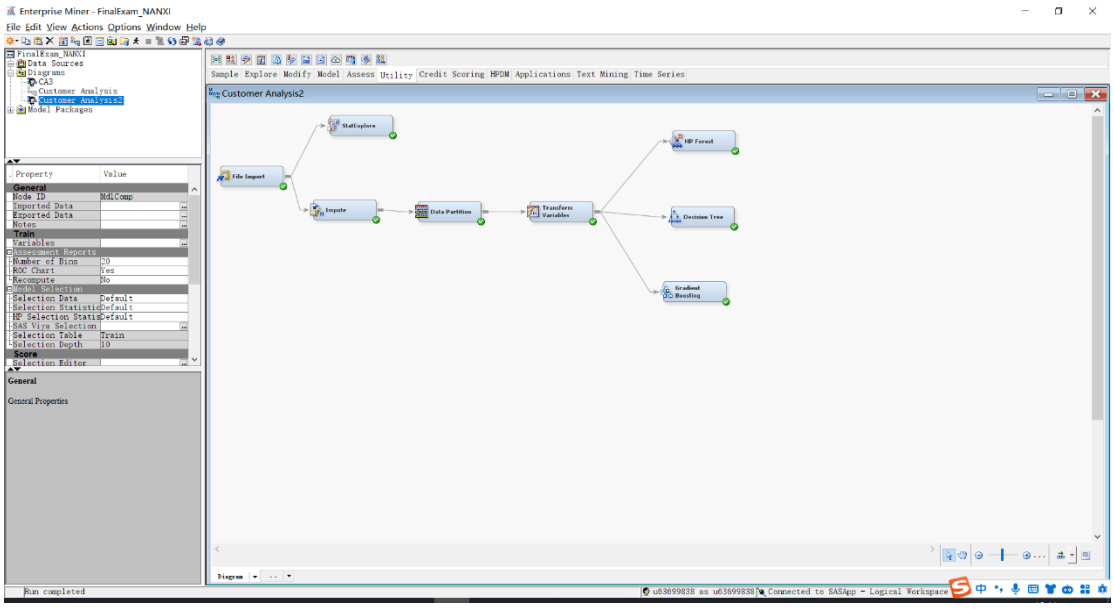
categories might have different churn rates, which suggests tailoring marketing and retention efforts based on product interest could be effective.

- **Spending Thresholds**: The tree nodes show that 'TotalSpent' is a variable used to split the data. This implies that spending levels are predictive of churn, with different churn rates at different spending thresholds.
- **Small High-Risk Groups**: The small group at the highest churn probability (Node Id: 6) suggests there are segments with very high churn risk. Although small, they require attention due to the high churn likelihood.

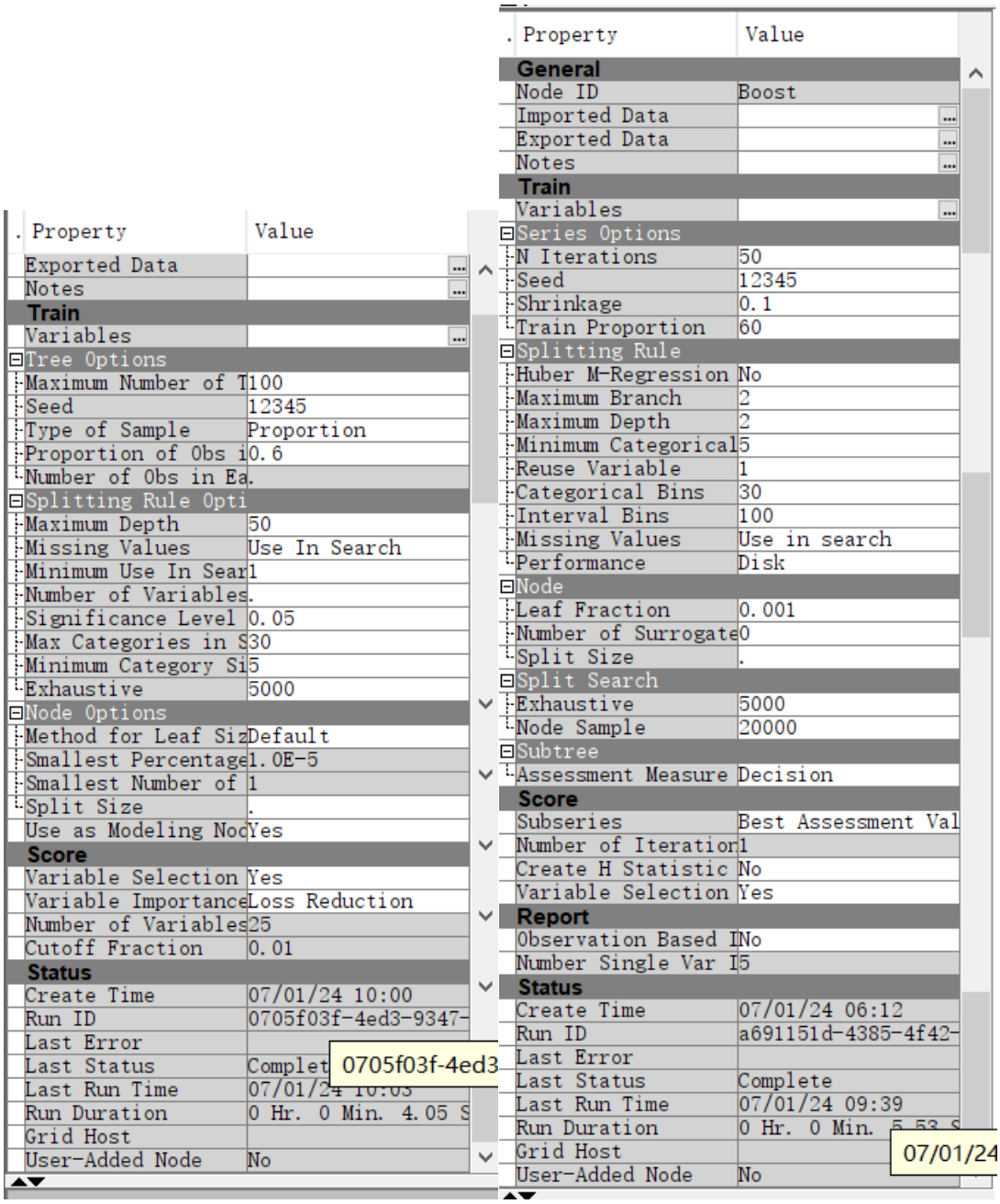
3. Ensemble Methods

3.1 Applying Bagging and Boosting:

For Bagging, we choose to use “HP Forest”; For Boosting, we choose to use “Gradient Boosting”.



“HP Forest” and “Gradient Boosting”:



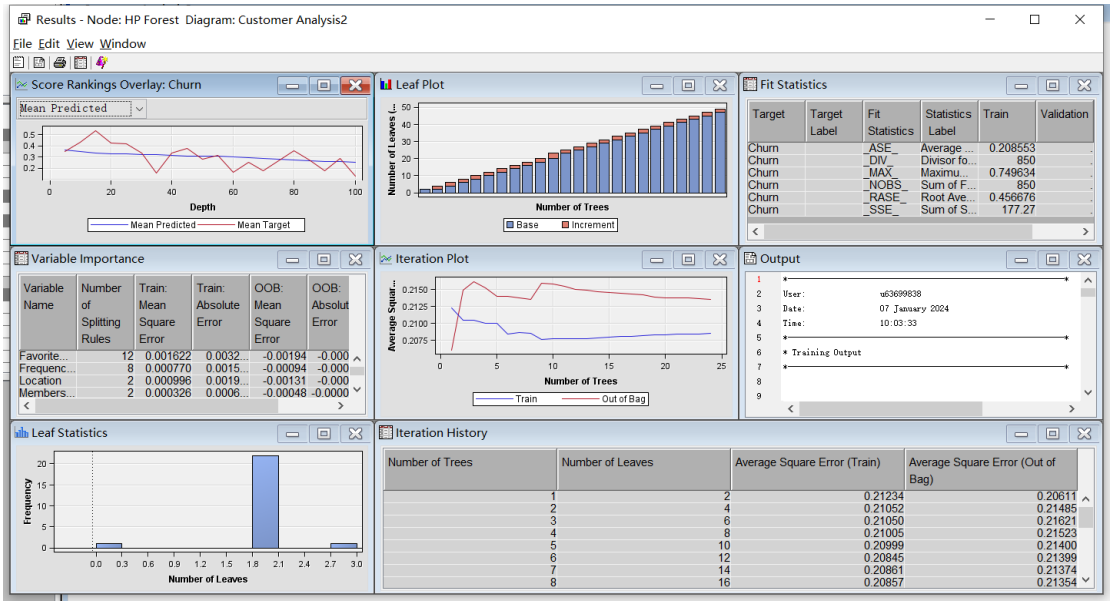
Property	Value
Exported Data	
Notes	
Train	
Variables	
Tree Options	
Maximum Number of T	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs i	0.6
Number of Obs in Ea	.
Splitting Rule Opti	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Sear	1
Number of Variables	.
Significance Level	0.05
Max Categories in S	30
Minimum Category Si	5
Exhaustive	5000
Node Options	
Method for Leaf Siz	Default
Smallest Percentage	1.0E-5
Smallest Number of	1
Split Size	.
Use as Modeling Nod	Yes
Score	
Variable Selection	Yes
Variable Importance	Loss Reduction
Number of Variables	25
Cutoff Fraction	0.01
Status	
Create Time	07/01/24 10:00
Run ID	0705f03f-4ed3-9347-
Last Error	
Last Status	Complet
Last Run Time	07/01/24 10:03
Run Duration	0 Hr. 0 Min. 4.05 S
Grid Host	
User-Added Node	No

0705f03f-4ed3

Property	Value
General	
Node ID	Boost
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Decision
Score	
Subseries	Best Assessment Val
Number of Iteration	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based I	No
Number Single Var I	5
Status	
Create Time	07/01/24 06:12
Run ID	a691151d-4385-4f42-
Last Error	
Last Status	Complete
Last Run Time	07/01/24 09:39
Run Duration	0 Hr. 0 Min. 5.53 S
Grid Host	
User-Added Node	No

07/01/24

3.2 HP Forest Results Analysis:



- **Score Rankings Overlay**: The graph plots the mean predicted versus the mean target probability of churn at different depths of the trees. The closeness of the two lines across depths suggests that the ensemble model is consistent in its predictions across different complexity levels.
- **Variable Importance**: This panel ranks the variables based on their importance in the model. ‘FavoriteCategory’ and ‘FrequencyOfWebsiteVisits’ are the most significant variables, indicating they are strong predictors of churn.
- **Leaf Plot**: It shows the number of leaves across the number of trees. The blue bars indicate the base number of leaves, and the red line shows the incremental increase, which stabilizes as more trees are added.
- **Fit Statistics**: This table provides statistical measures such as NOBS (number of observations) and SSE (sum of squares error), with the latter being a measure of the model's error.
- **Iteration Plot**: The plot shows the out-of-bag error and the training error across different numbers of trees. It helps in determining the optimal number of trees for the model by looking for the point where the out-of-bag error stabilizes or starts increasing.

Assessment Score Distribution

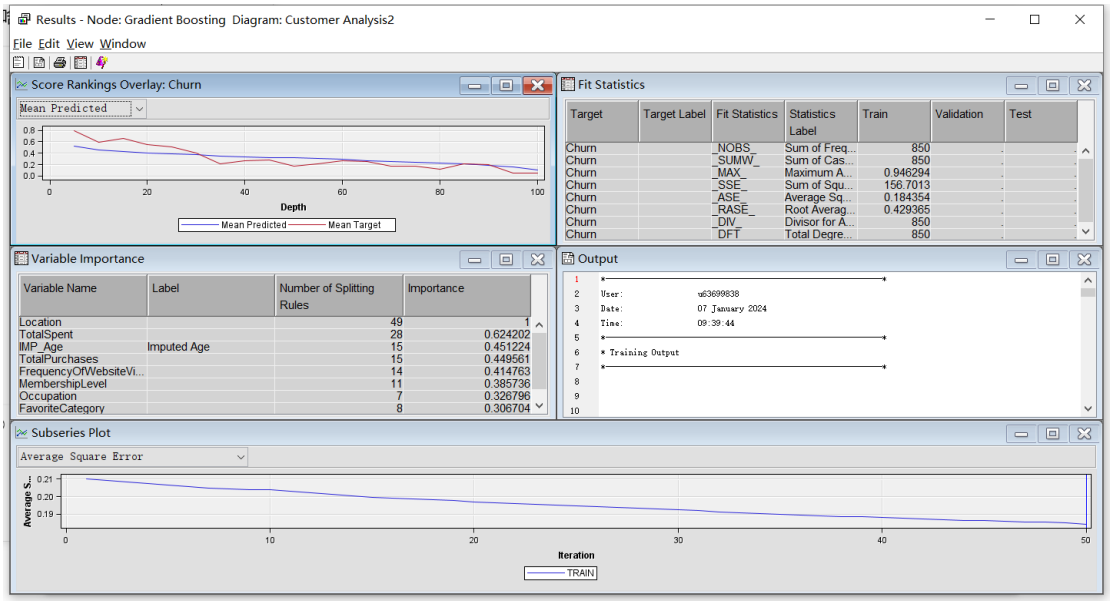
Data Role=TRAIN Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.377 - 0.384	1.00000	0.37903	5	0.38021
0.370 - 0.377	0.14286	0.37200	7	0.37355
0.364 - 0.370	0.60000	0.36792	5	0.36689
0.357 - 0.364	0.17647	0.36002	17	0.36023
0.350 - 0.357	0.36842	0.35161	19	0.35358
0.344 - 0.350	0.42308	0.34628	26	0.34692
0.337 - 0.344	0.53846	0.33973	39	0.34026
0.330 - 0.337	0.50000	0.33409	32	0.33360
0.324 - 0.330	0.40000	0.32756	80	0.32694
0.317 - 0.324	0.23810	0.31946	84	0.32028
0.310 - 0.317	0.36957	0.31393	46	0.31362
0.304 - 0.310	0.34783	0.30797	92	0.30696
0.297 - 0.304	0.19540	0.29979	87	0.30031
0.290 - 0.297	0.24242	0.29383	33	0.29365
0.284 - 0.290	0.17778	0.28801	45	0.28699
0.277 - 0.284	0.28571	0.28081	28	0.28033
0.270 - 0.277	0.34783	0.27462	23	0.27367
0.264 - 0.270	0.29333	0.26800	75	0.26701
0.257 - 0.264	0.18000	0.26045	50	0.26035
0.250 - 0.257	0.19298	0.25251	57	0.25369

- For higher ranges of predicted probabilities (rows 321-322), we observe perfect and high mean target churn rates (1.0 and 0.14286 respectively), although the mean predicted probabilities (0.37903 and 0.37200) are lower than the mean target, which indicates that the model may be underestimating the likelihood of churn for these highest-risk customers.
- The lower ranges (rows 339-341) show mean target churn rates that are relatively low (0.24871, 0.27513, and 0.19298), with the model's mean predicted probabilities being fairly close (0.28081, 0.27642, and 0.25251), suggesting the model is performing reasonably well at predicting lower-risk segments.

- The model scores generally decrease as we move from higher to lower ranges of predicted probabilities, which might indicate that the model is more accurate at distinguishing between the highest risk and lowest risk customers.
- The number of observations in each range varies, with most intervals containing a small number of observations, which could indicate overfitting or could simply be a result of how the predicted probability ranges are defined.

3.3 Gradient Boosting Results Analysis:



- **Score Rankings Overlay**: The overlay plot indicates that the mean predicted churn probability remains relatively consistent across the different tree depths, signifying stable predictions. The plot also shows a slight divergence between mean predicted and mean target probabilities, which might point to areas where the model could be improved for accuracy.
- **Variable Importance**: The variables are ranked by their importance in predicting churn, with 'Location' being the most significant, followed by 'TotalSpent' and 'Imputed Age'. The importance values suggest these features have the most predictive power within the model.
- **Fit Statistics**: Various fit statistics are displayed, such as NOBS (number of observations), MAX (maximum absolute error), SSE (sum of squares error), and ASE (average squared error). These indicate the overall fit of the model. For instance, an SSE of 156.7013 suggests the sum of squared errors over all

training data points, and a lower ASE of 0.184354 implies that, on average, the model's predictions are close to the actual values.

- **Subseries Plot**: This plot shows the Average Squared Error (ASE) over the number of iterations during model training. The relatively flat line suggests that the model's error did not significantly change after additional boosting iterations, which could mean the model has reached its optimal point before the iterations have concluded.

Assessment Score Distribution

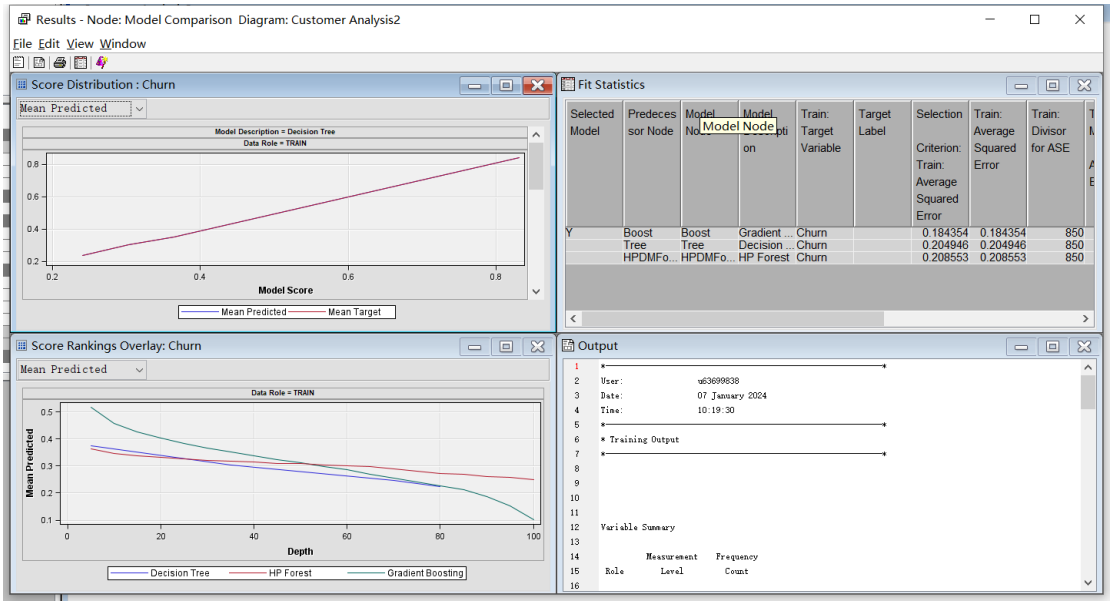
Data Role=TRAIN Target Variable=Churn Target Label=' '

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
0.575 – 0.603	1.00000	0.58715	4	0.58891
0.547 – 0.575	0.00000	0.56069	1	0.56066
0.518 – 0.547	0.80000	0.53270	15	0.53242
0.490 – 0.518	0.73333	0.50250	15	0.50418
0.462 – 0.490	0.75000	0.47609	24	0.47593
0.434 – 0.462	0.58333	0.44493	36	0.44769
0.405 – 0.434	0.65385	0.41923	52	0.41944
0.377 – 0.405	0.47619	0.39000	63	0.39120
0.349 – 0.377	0.33333	0.36271	75	0.36296
0.321 – 0.349	0.26882	0.33374	93	0.33471
0.292 – 0.321	0.18280	0.30572	93	0.30647
0.264 – 0.292	0.25974	0.27808	77	0.27822
0.236 – 0.264	0.17391	0.24920	92	0.24998
0.208 – 0.236	0.14493	0.22217	69	0.22174
0.179 – 0.208	0.24444	0.19390	45	0.19349
0.151 – 0.179	0.08824	0.16947	34	0.16525
0.123 – 0.151	0.00000	0.13513	29	0.13700
0.095 – 0.123	0.05556	0.10957	18	0.10876
0.066 – 0.095	0.00000	0.08230	11	0.08052
0.038 – 0.066	0.25000	0.04851	4	0.05227

- The model predicts a 100% churn rate for the highest probability range (0.575 – 0.603), which matches the Mean Target, albeit based on only 4 observations. This suggests that for this range, the model is highly confident and accurate.

- For the ranges (0.547 – 0.575) and (0.518 – 0.547), the Mean Target is 0, but the model predicts a probability of churn around 0.56069 and 0.53270, respectively. This discrepancy indicates overestimation of churn likelihood in these intervals.
- In the mid-range probabilities (0.462 – 0.490), the Mean Target of 0.75 shows a high actual churn rate, which the model's Mean Predicted of 0.47609 underestimates. This suggests the model may not be sensitive enough to churn risk in this segment.
- Lower ranges show a decrease in Mean Target churn rates, with the model's predictions also trending lower, which suggests the model is effectively distinguishing between higher and lower risk customers.
- The Model Score fluctuates across different ranges, which could reflect the model's varying confidence levels in its predictions across different segments.

3.4 Comparison of Models:



- **Score Distribution Plot:**

The score distribution for 'Churn' shows the Decision Tree's predicted probabilities against the actual targets. A perfect model would have a 45-degree line, indicating a perfect match between prediction and reality. The divergence from this line suggests

prediction error, common in practical models.

- **Score Rankings Overlay:**

This graph compares the mean predicted probability of churn against the depth of the tree, for each model. The depth represents the complexity of the model. Generally, as the model becomes more complex (with more depth), we expect it to fit the training data better. However, too much complexity can lead to overfitting. The Decision Tree starts to plateau quickly, suggesting limited complexity with a simpler model. The HP Forest and Gradient Boosting lines are closer together, indicating similar performance across depths. Both models appear to stabilize around a depth of 20, suggesting that adding complexity beyond this point does not yield significant gains in performance on the training data.

- **Fit Statistics:**

Fit statistics for the models show the Average Squared Error for both training and validation (if applicable) datasets. Lower values indicate a better fit. The Gradient Boosting model has the lowest Average Squared Error, suggesting it performs the best in terms of fitting to the data among the three models. The HP Forest shows a slightly higher error rate than Gradient Boosting, but still outperforms the Decision Tree.

- **Output Section:**

Provides general information about the training output, such as user, date, and time. This section is typically used for record-keeping and doesn't offer insights into model performance.

Fit Statistics Table
Target: Churn

Data Role=Train

Statistics	Boost	Tree	HPDMForest
Train: Average Squared Error	0.184	0.205	0.209
Selection Criterion: Train: Average Squared Error	0.184	0.205	0.209
Train: Total Degrees of Freedom	850.000	850.000	.
Train: Divisor for ASE	850.000	850.000	850.000
Train: Maximum Absolute Error	0.946	0.846	0.750
Train: Sum of Frequencies	850.000	850.000	850.000
Train: Root Average Squared Error	0.429	0.453	0.457
Train: Sum of Squared Errors	156.701	174.204	177.270
Train: Sum of Case Weights Times Freq	850.000	.	.

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

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

As we can see, the Gradient Boosting model appears to offer the best average performance in terms of ASE, suggesting it is the most accurate on average for the training data. However, it has a higher maximum error, indicating potential outlier predictions where it performs poorly. The HP Forest model, while having a slightly higher ASE and RASE, has the best performance on maximum error, suggesting more consistent performance across all predictions, including potentially better handling of outliers. The Decision Tree, while having the simplest model complexity, shows it is less accurate on average but doesn't perform as poorly on the worst-case predictions compared to Boosting.

4. Business Strategy Suggestions:

- **Variable Importance**: The Gradient Boosting model indicates that 'Location' and 'TotalSpent' are significant predictors of churn. This suggests that

geographical factors and customer spending habits strongly influence customer retention.

- **Model Robustness**: The Random Forest model showed the lowest maximum error, which implies that it is less likely to make extreme errors in prediction. This robustness can be important when considering business strategies that require consistent decision-making.

- **Personalized Engagement**: Since product preference categories like 'HOME GOODS OR MISCELLANEOUS', 'ELECTRONICS', 'CLOTHING' are key differentiators, create personalized engagement strategies. Offer special promotions, loyalty rewards, or new product lines tailored to these interests.
- **Spend-Based Incentives**: For segments identified by their 'TotalSpent', introduce spend-based incentives to encourage higher spending and retention. This could include volume discounts, loyalty points for certain spend thresholds, or exclusive access to premium products/services.
- **Targeted Intervention for High-Risk Segments**: For the identified high-risk segments, implement targeted intervention strategies. This could involve personalized communication, special offers, or dedicated customer service support to address their specific reasons for potential churn.
- **Continuous Monitoring and Adjustment**: Use the churn probability ranges to continuously monitor customer segments. Engage customers predicted to be at higher risk more proactively and adjust strategies based on feedback and observed behavior changes.
- **Customer Experience Improvement**: Given the importance of spending levels, review the customer experience journey to identify any pain points that might prevent higher spending. Enhancements could include streamlining the purchasing process, offering better after-sales support, or improving the overall product/service quality.

As we can see in this case study, customer behavior is still a topic worth studying.

Although the performance of the model is not good enough in this study, we can still get good business value from it, which shows the importance of data analysis.

To mitigate customer churn and bolster engagement, businesses should implement personalized engagement strategies that resonate with customers' preferred product categories, introduce spend-based incentives to promote higher expenditure, provide targeted support to high-risk segments, and continuously monitor at-risk customers to proactively address their needs. Additionally, enhancing the overall customer journey by improving the purchasing process and product quality can further solidify customer loyalty and spending.