

Dataset Source: <https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset>

Introduction:

In the realm of e-commerce, understanding customer behavior is paramount for businesses striving to enhance their operational strategies and boost customer retention. The dataset at hand encapsulates a rich tapestry of customer transactions, spanning diverse attributes such as age, gender, location, membership level, and purchase history. This case study delves into the dynamic landscape of e-commerce customer behavior analysis, leveraging advanced analytics techniques.

This comprehensive data analysis journey begins with the utilization of Talend Data Integration, a powerful tool for joining datasets. Through the integration of customer churn and customer behavior datasets, a tMap component is employed to facilitate the merging process based on the 'CustomerID' column. This strategic integration lays the foundation for a more holistic understanding of customer interactions and transactions within an e-commerce platform.

Following the dataset integration, the focus shifts to data modification using Talend Data Preparation. The objective is to standardize categorical values within the 'Gender' and 'Churn' columns. In the 'Gender' column, the four categories (M, F, Male, Female) are harmonized to represent 'Male' and 'Female.' Similarly, in the 'Churn' column, the disparate categories (0, 1, No, Yes) are standardized to 'No' and 'Yes.' This standardization ensures consistency and facilitates more meaningful analysis.

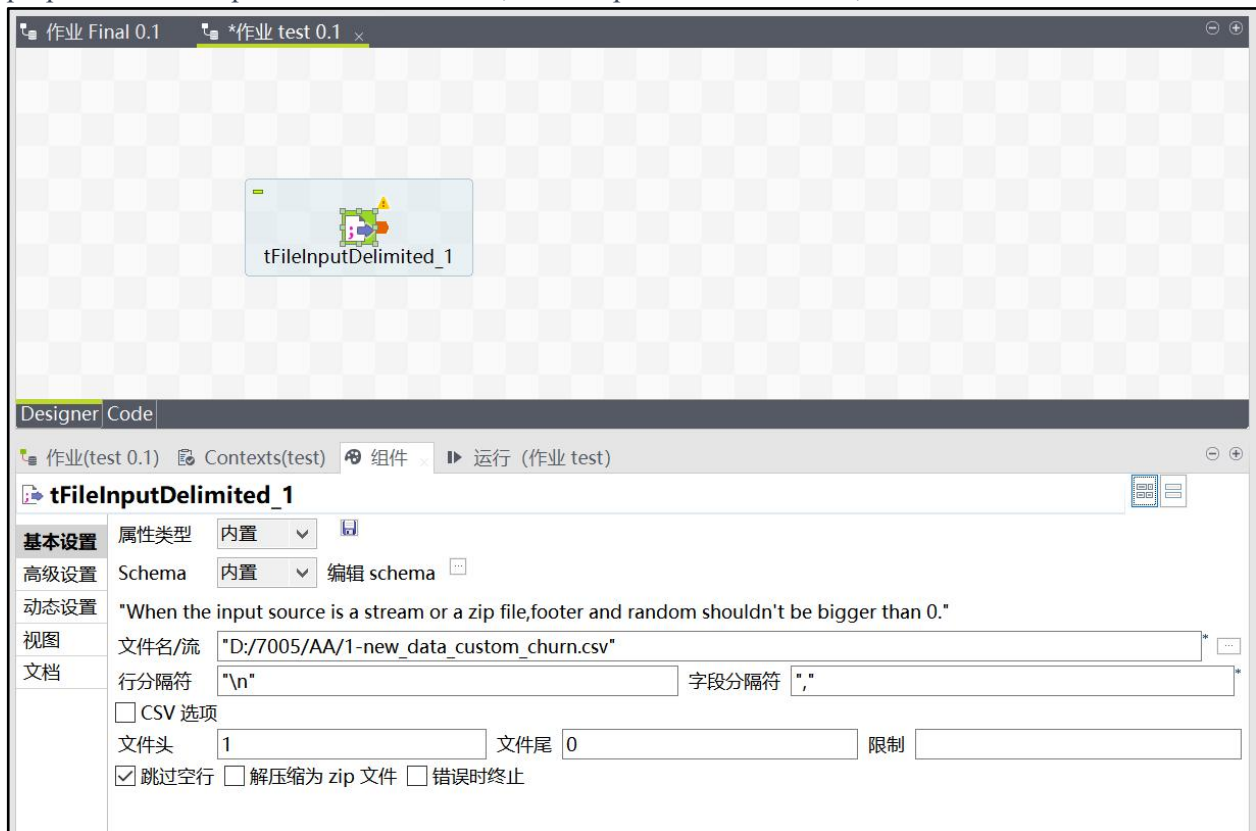
Subsequently, the analysis transitions to SAS Enterprise Miner, where the imported dataset undergoes preprocessing. This step involves handling missing values and specifying variable roles, setting the stage for robust analysis. The subsequent phases encompass Decision Tree Analysis and the application of Ensemble Methods, specifically Bagging and Boosting using the Random Forest algorithm as an illustrative example of Bagging.

Objective:

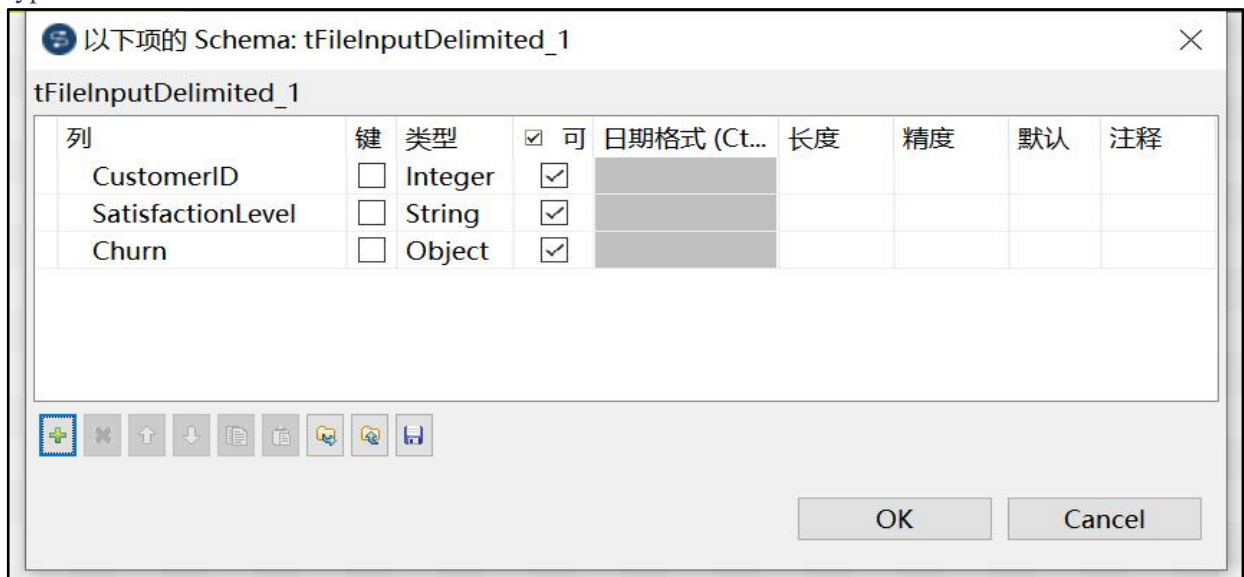
The objective of this case study is to leverage advanced analytics techniques to gain meaningful insights into e-commerce customer behavior. Through a meticulous data analysis journey, we aim to enhance operational strategies and bolster customer retention for businesses operating in the e-commerce domain.

Using Talend Data Integration to Join Datasets

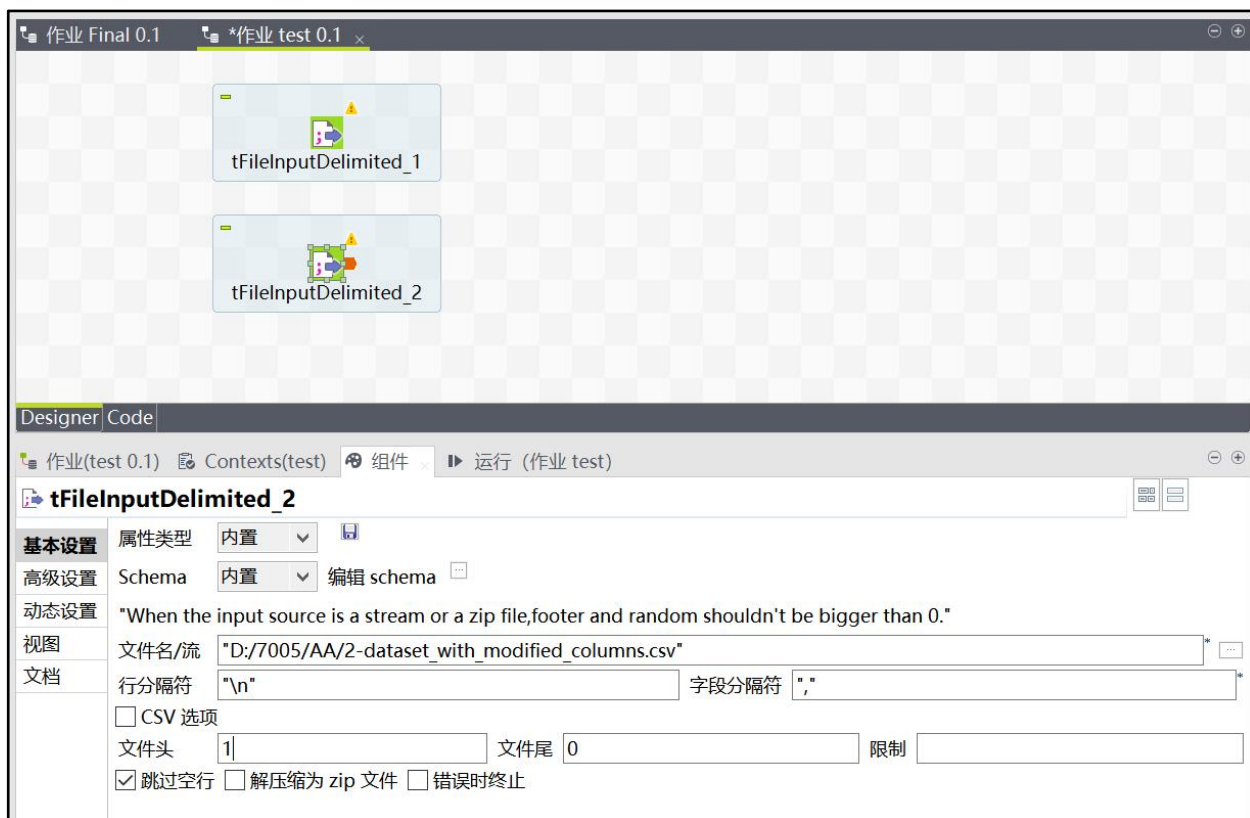
Drag the **tFileInputDelimited** component to the design workspace. In the Component panel below, set the properties: Row Separator: Set this to "\n", Field Separator: Set this to "," and Header: set this to 1.



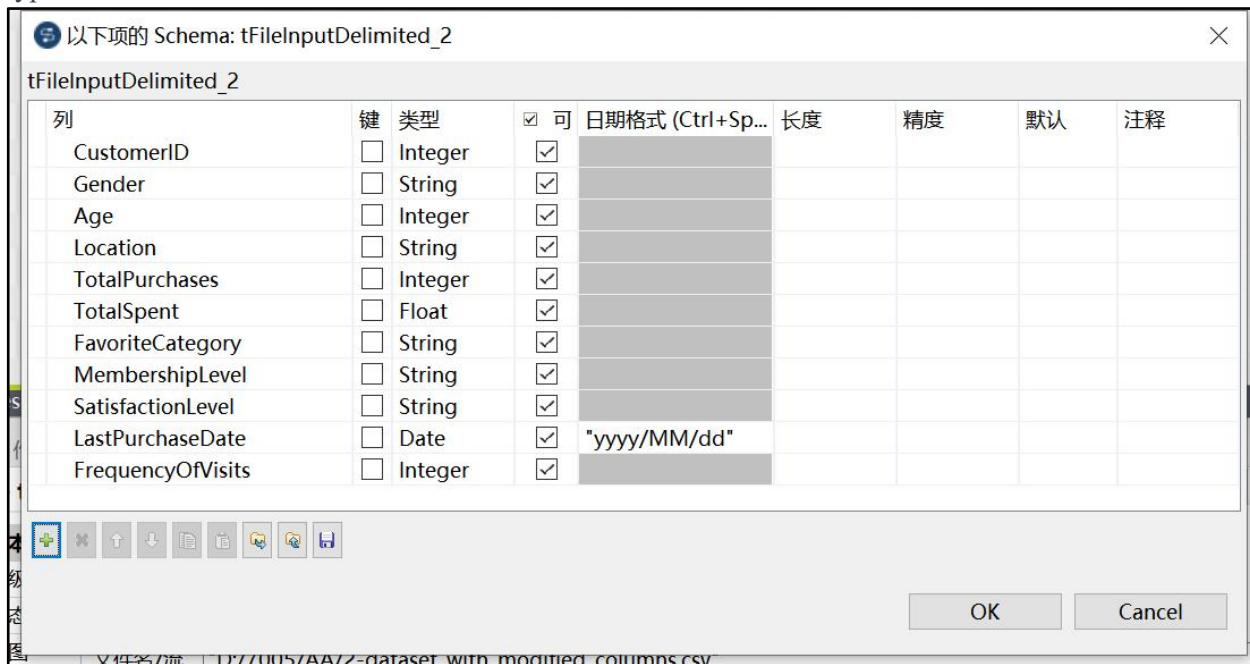
Under the **Schema** section, click on the Edit schema button. Define the columns and specify their data types.



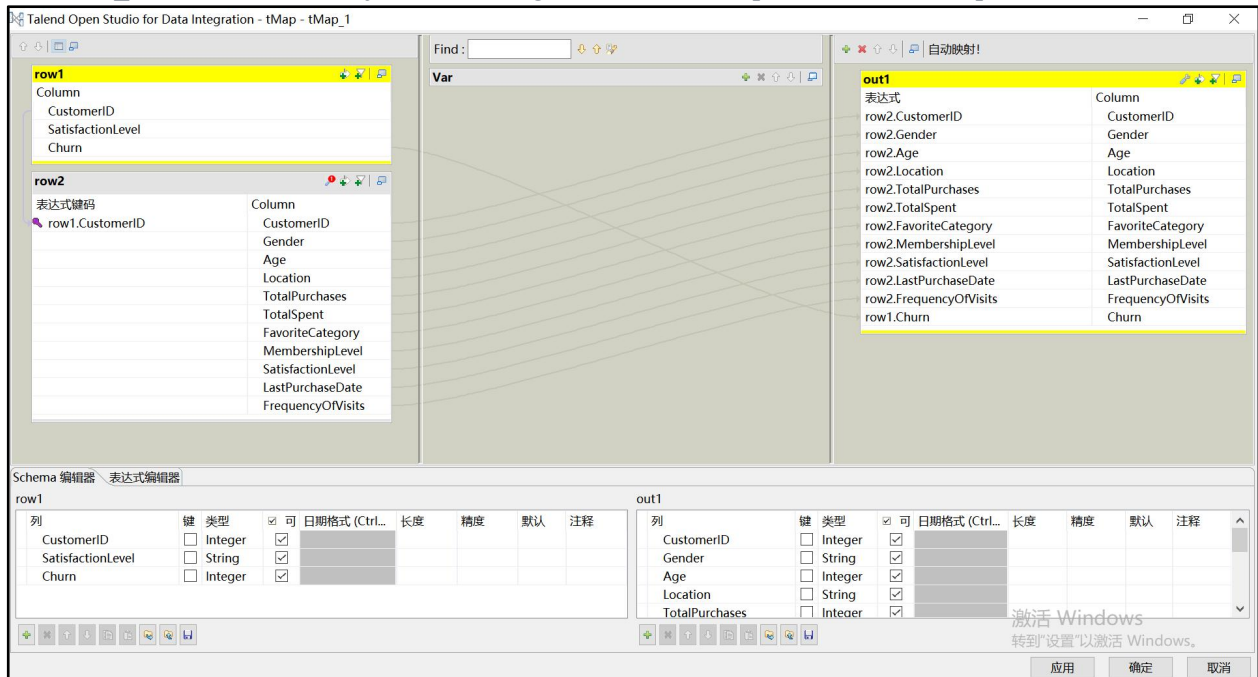
Same as above



Under the **Schema** section, click on the Edit schema button. Define the columns and specify their data types.



Drag and drop a **tMap** component onto the workspace. Link both tFileInputDelimited components to the tMap component. Drag the 'CustomerID' column from customer_churn to the 'CustomerID' column of customer_behavior to create a join. On the right side of the tMap editor, define output structure.



Drag and drop a **tFileOutputDelimited** component onto the workspace. Link the output from the tMap component to the tFileOutputDelimited component. Configure the output component to write the results to a new CSV file, say, output_4.csv.



Data Modification Using Talend Data Preparation

Before: In the Gender column, there are four categories with M, F, Male and Female.

The inputs are not standardized because F represents Female and M represents Male.

The screenshot shows the Talend Data Preparation interface for a dataset named 'output_4_dataset Preparation'. The main table displays 14 rows of data. The 'Gender' column contains values: Female, Male, Female, Male, Male, Female, Female, Male, Female, Male, Male, F, Female, and Male. The right sidebar shows the 'Gender' column selected, with a 'Find a function...' search bar and a 'SUGGESTIONS' list containing 'Change to upper case', 'Replace the cells that match...', and 'Change to lower case'. A 'CHART' section at the bottom shows a bar chart with 'Male' and 'Female' categories.

	CustomerID	Gender	Age	Location	TotalPurchases	TotalSpent	FavoriteCat
	integer	gender	integer	city	integer	decimal	
1	101	Female	29	New York	14	1120.2	Electroni
2	102	Male	34	Los Angeles	11	780.5	Clothing
3	103	Female	43	Chicago	9	510.75	Electroni
4	104	Male	30	San Francisco	19	1480.3	Clothing
5	105	Male	27	Miami	13	720.4	Clothing
6	106	Female	37	Houston	8	440.8	Home Goo
7	107	Female	31	New York	15	1150.6	Electroni
8	108	Male	35	Los Angeles	12	800.9	Home Goo
9	109	Female	41	Chicago	10	495.25	Electroni
10	110	Male	28	San Francisco	21	1520.1	Electroni
11	111	Male	32	Miami	11	690.3	Electroni
12	112	F	36	Houston	7	470.5	Home Goo
13	113	Female		New York	16	1200.8	Clothing
14	114	Male	33	Los Angeles	13	820.75	Home Goo

Function: Use **Replace the cells that match**, let Female replace F, Male replace M

After: Replace the cells that match F to Female and M to Male

The screenshot shows the Talend Data Preparation interface for a dataset named 'output_4_preprocessing Preparation'. The main table displays 14 rows of data. The 'Gender' column contains values: Female, Male, Female, Male, Male, Female, Female, Male, Female, Male, Male, Female, Female, and Male. The right sidebar shows the 'Gender' column selected, with a 'Find a function...' search bar and a 'SUGGESTIONS' list containing 'Overwrite entire cell', 'Change to lower case', and 'CHART'. A 'SUBMIT' button is visible.

	merID	Gender	Age	Location	To
	integer	gender	integer	city	
1	101	Female	29	New York	
2	102	Male	34	Los Angeles	
3	103	Female	43	Chicago	
4	104	Male	30	San Francisco	
5	105	Male	27	Miami	
6	106	Female	37	Houston	
7	107	Female	31	New York	
8	108	Male	35	Los Angeles	
9	109	Female	41	Chicago	
10	110	Male	28	San Francisco	
11	111	Male	32	Miami	
12	112	Female	36	Houston	
13	113	Female		New York	
14	114	Male	33	Los Angeles	

Before: Churn: There are four categories with 0, 1, No and Yes.

The inputs are not standardized because 0 represents No and 1 represents Yes.

talend DATA PREPARATION

output_4_preprocessing Preparation

1 Replace the cells that match on column Gender

2 Replace the cells that match on column Gender

Filters

Add a filter ...

	actionLevel	LastPurchaseDate	FrequencyOfVisits	Churn
	text	date	integer	integer
1	ied	2022/10/30	7	0
2	ied	2023/6/1	5	0
3	sfied	2022/12/29	4	1
4	ied	2023/3/22	9	0
5	ied	2022/6/22	5	1
6	sfied	2022/3/7	3	1
7	ied	2022/10/26	7	0
8	ied	2022/8/16	5	0
9	sfied	2022/10/14	4	0
10	ied	2023/2/14	10	0
11	sfied	2022/9/3	5	1
12	sfied	2023/11/14	3	0
13	ied	2023/10/4	8	0
14	ied	2023/11/8	6	0

Churn

COLUMN ROW

Find a function ...

COLUMNS

Concatenate with...

Delete column

Swap columns...

CHART VALUE PATTERN ADVANCED

0 50 100 150 200 250 300

9

(EMPTY)

aaa

aa

激活 Windows
转到“设置”以激活 Windows。

Function: Use **Replace the cells that match**, let 1 replace yes, 0 replace no

After: Replace the cells that match 0 to No and 1 to Yes.

talend DATA PREPARATION

output_4_preprocessing Preparation

1 Replace the cells that match on column Gender

2 Replace the cells that match on column Gender

3 Replace the cells that match on column Churn

4 Replace the cells that match on column Churn

Filters

Add a filter ...

	actionLevel	LastPurchaseDate	FrequencyOfVisits	Churn
	text	date	integer	integer
1	ied	2022/10/30	7	0
2	ied	2023/6/1	5	0
3	sfied	2022/12/29	4	1
4	ied	2023/3/22	9	0
5	ied	2022/6/22	5	1
6	sfied	2022/3/7	3	1
7	ied	2022/10/26	7	0
8	ied	2022/8/16	5	0
9	sfied	2022/10/14	4	0
10	ied	2023/2/14	10	0
11	sfied	2022/9/3	5	1
12	sfied	2023/11/14	3	0
13	ied	2023/10/4	8	0
14	ied	2023/11/8	6	0

Churn

COLUMN ROW

Find a function ...

Swap columns...

CONVERSIONS

Convert distance...

Convert duration...

CHART VALUE PATTERN ADVANCED

0 50 100 150 200 250 300

9

(EMPTY)

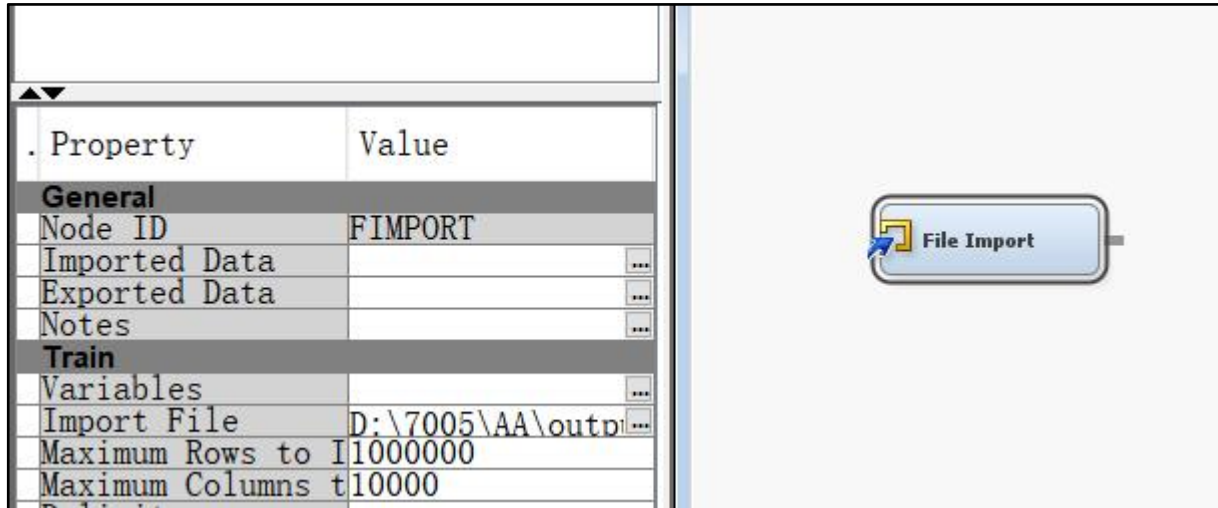
激活 Windows
转到“设置”以激活 Windows。

After this, export the file.

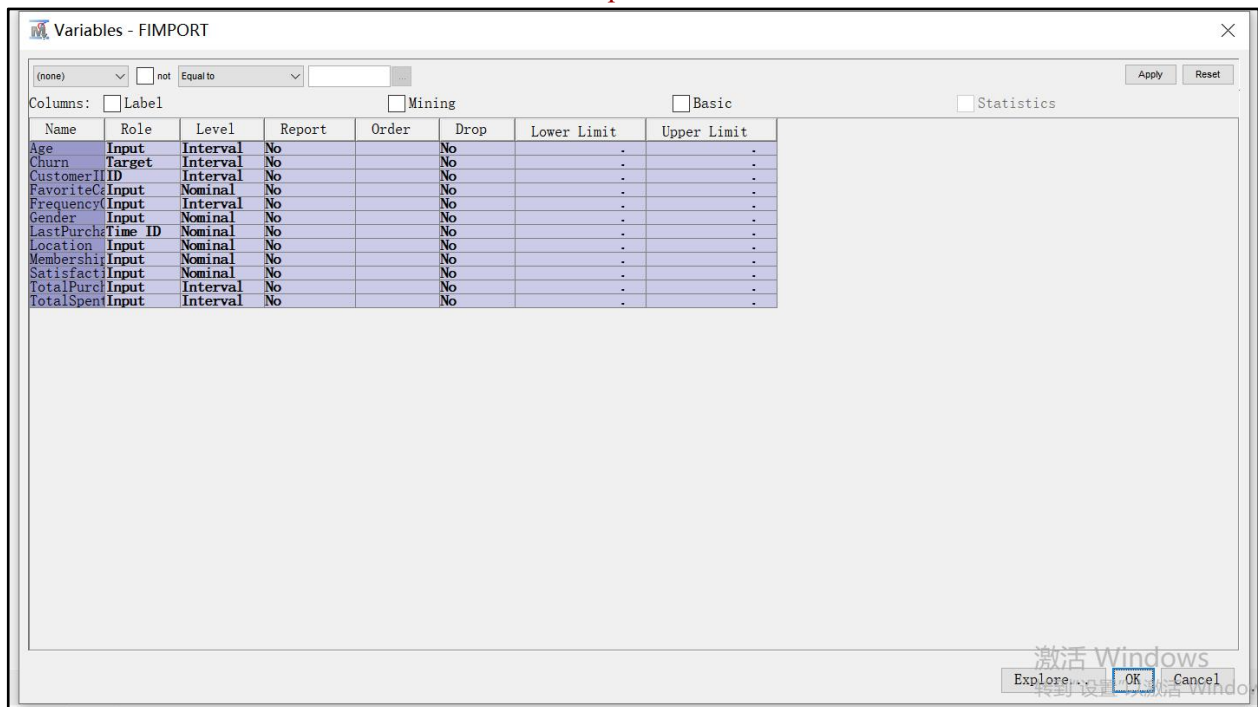
Tasks in SAS EM

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

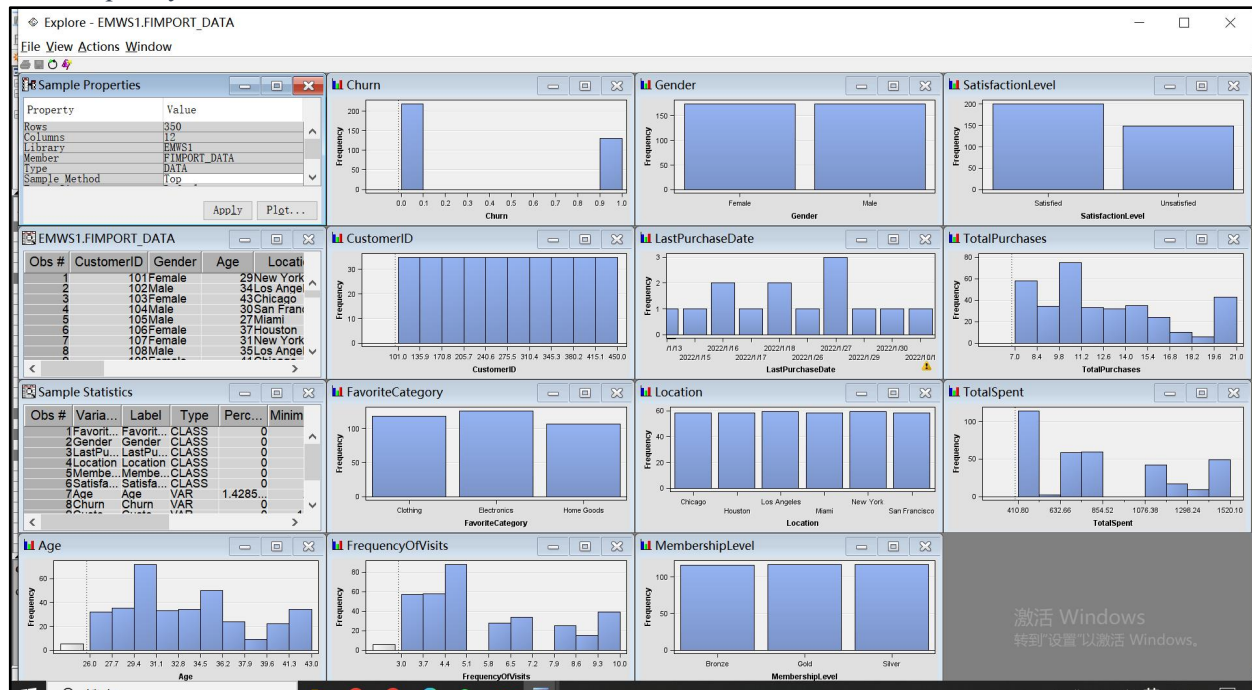
Drag and drop **File Import** and import the .csv document into SAS EM



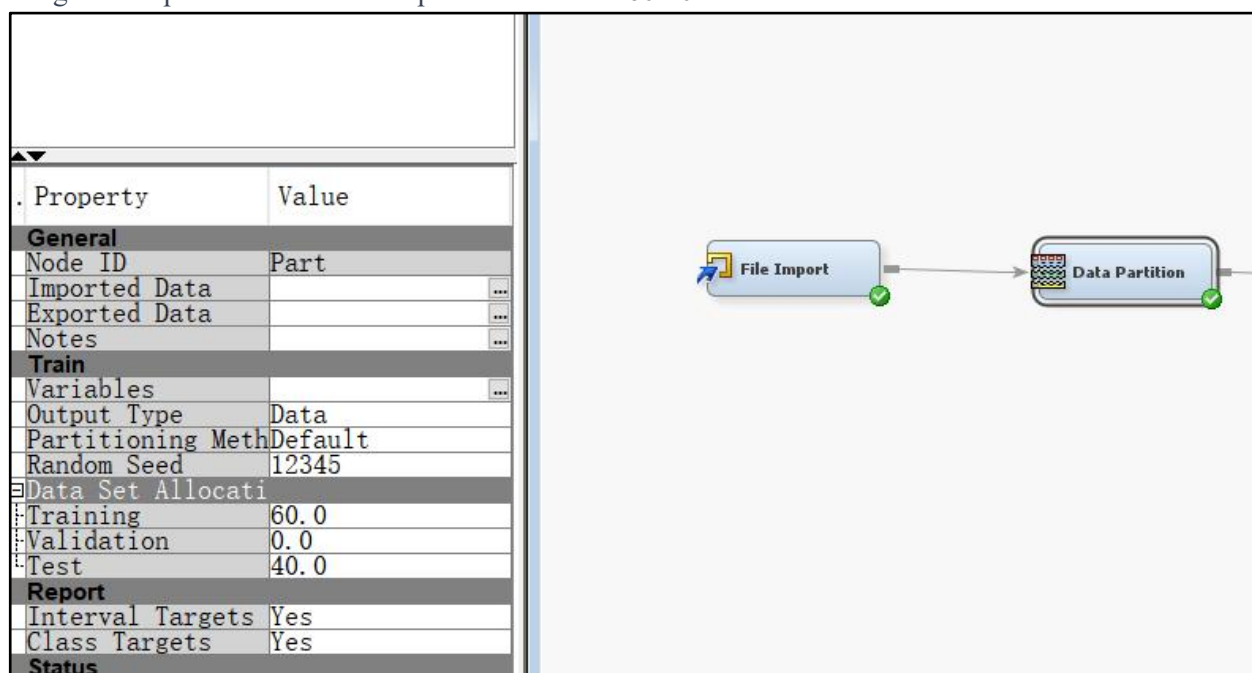
Right click to **edit variables**, and specify variable roles. The churn role is **target**, CustomerID role is **ID**, LastPurchaseDate role is **Time ID** and others are **input**.



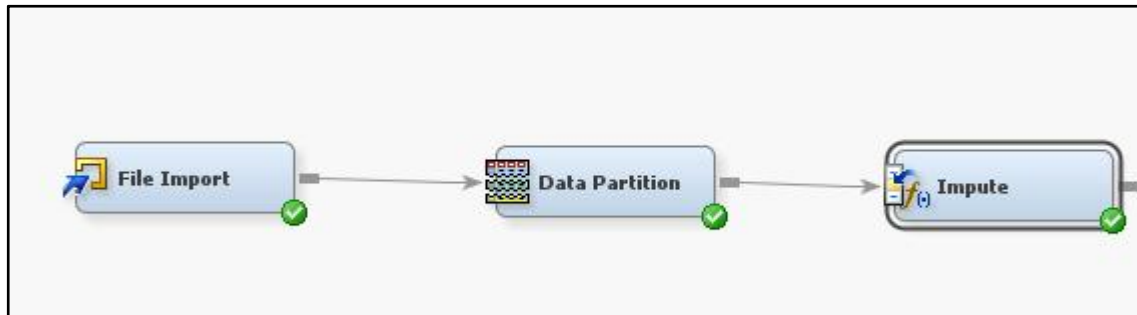
Select all variables and click **explore**. In the results, we can see that there are missing values in the Age and FrequencyOfVisits columns.



Drag and drop **Data Partition** and split train/test into 60/40



Drag and drop **Impute** and edit the variables, then run. help to **input the missing values** in the Age and FrequencyOfVisits columns.



Set the missing Age and FrequencyOfVisits values to use **median** for impute.

Variables - Impt

☐ not

Columns: ☐ Label ☐ Mining ☐ Basic

Name	Use	Method	Use Tree	Role	Level
Age	Default	Median	Default	Input	Interval
Churn	Default	Default	Default	Target	Interval
FavoriteCate	Default	Default	Default	Input	Nominal
FrequencyOfV	Default	Median	Default	Input	Interval
Gender	Default	Default	Default	Input	Nominal
Location	Default	Default	Default	Input	Nominal
MembershipLe	Default	Default	Default	Input	Nominal
Satisfaction	Default	Default	Default	Input	Nominal
TotalPurchas	Default	Default	Default	Input	Interval
TotalSpent	Default	Default	Default	Input	Interval

Results - Node: Impute Diagram: AA

File Edit View Window

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEDIAN	IMP Age	33	INPUT	INTERVAL	Age	0
FrequencyOfVisits	MEDIAN	IMP FrequencyOfVisits		INPUT	INTERVAL	FrequencyOfVisits	0

Output

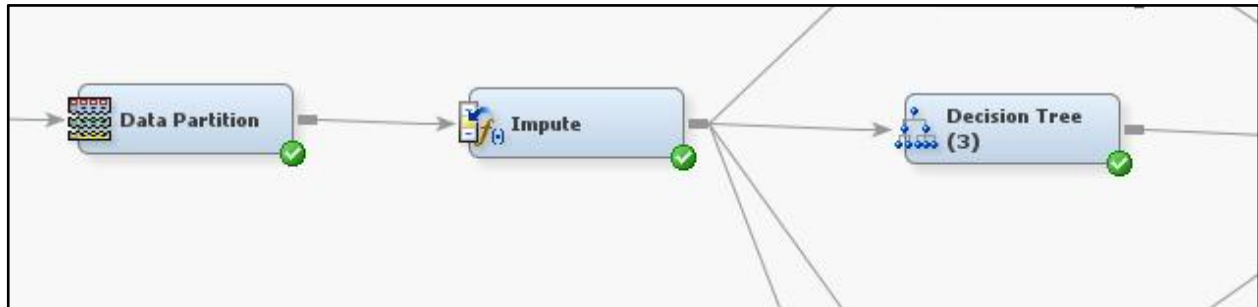
```

1
2 User: u03454190
3 Date: 08 January 2024
4 Time: 17:33:26
5
6 * Training Output
7
8
9
10
11
12 Variable Summary
13
14 Measurement Level Frequency
15 Count
16
17 INPUT INTERVAL 4
18 INPUT NOMINAL 5
19 TARGET INTERVAL 1
20
21
22
  
```

激活 Windows
转到“设置”以激活 Windows。

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

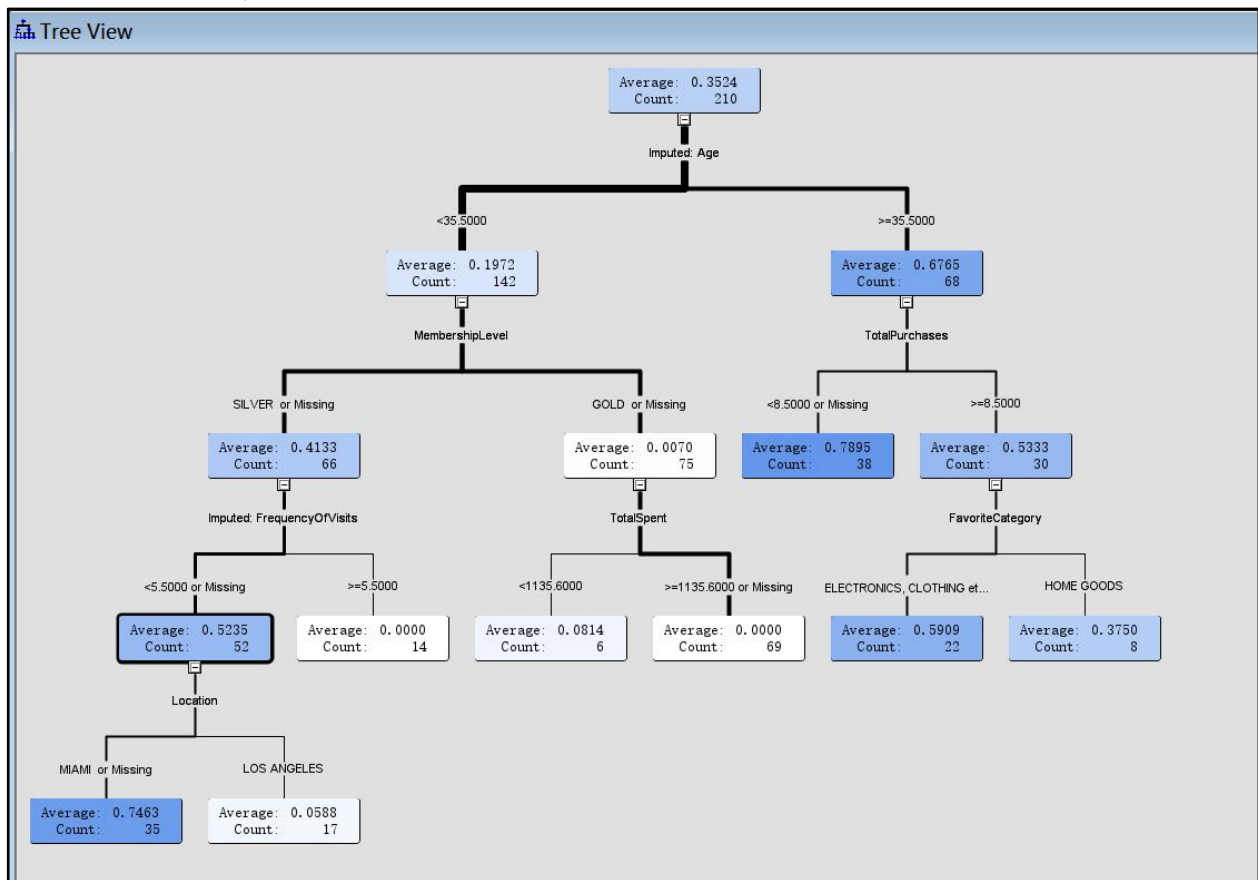
Drag and drop **Decision Tree** and run



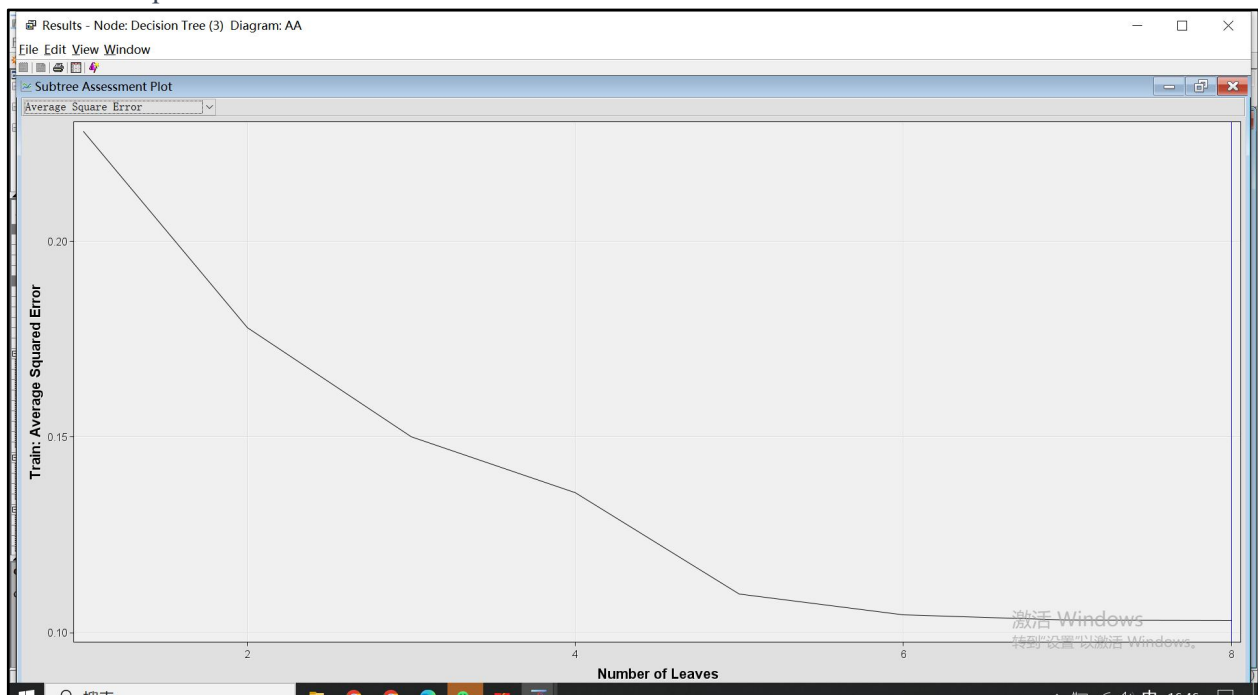
During my project, my decision tree has no nodes. Then I searched on Google and found that I could change the configuration, so I changed it and it ran successfully. **I changed the criterion to gini.**

Property	Value
General	
Node ID	Tree3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Mode	No
Tree Model Data	...
Use Frozen Tree	No
Use Multiple Target	No
Splitting Rule	
Interval Target	ProbF
Nominal Target	Gini
Ordinal Target	Gini
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categories	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogates	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

After successful run, the decision tree is as follows:



Assessment plot:



Analyze based on the results of the decision tree:

The decision tree results reveal crucial insights into customer behavior and its impact on churn. The most influential variables include IMP_Age (Imputed: Age), MembershipLevel, Location, IMP_FrequencyOfVisits (Imputed: FrequencyOfVisits), TotalPurchases, FavoriteCategory, and TotalSpent. Among these, age emerges as the most critical factor, closely followed by membership level and geographic location. The frequency of website visits, total number of purchases, favorite shopping category, and total amount spent also contribute to varying extents.

The tree leaf report categorizes the data into different segments based on the decision tree. Each node represents a specific condition or rule, and the leaves contain observations falling into those conditions. Key nodes, such as Node 7 (customers with a predicted churn probability of 0.79), Node 18 (customers with a predicted churn probability of 0.75), Node 11 (customers with a predicted churn probability of 0.59), and Node 19 (customers with a predicted churn probability of 0.06), provide critical information.

High Churn Probability Nodes (e.g., Node 7 and Node 18) represent segments where customers have a higher predicted probability of churning. Understanding the characteristics of customers in these segments is vital for devising targeted retention strategies. On the other hand, Low Churn Probability Nodes (e.g., Node 19) represent customers with a very low predicted probability of churning. Identifying what sets these customers apart can aid in reinforcing positive behaviors or pinpointing loyal segments, contributing to effective retention strategies.

suggestions for business strategy:

Drawing insights from the decision tree results, we can outline strategic recommendations to proactively address and mitigate customer churn.

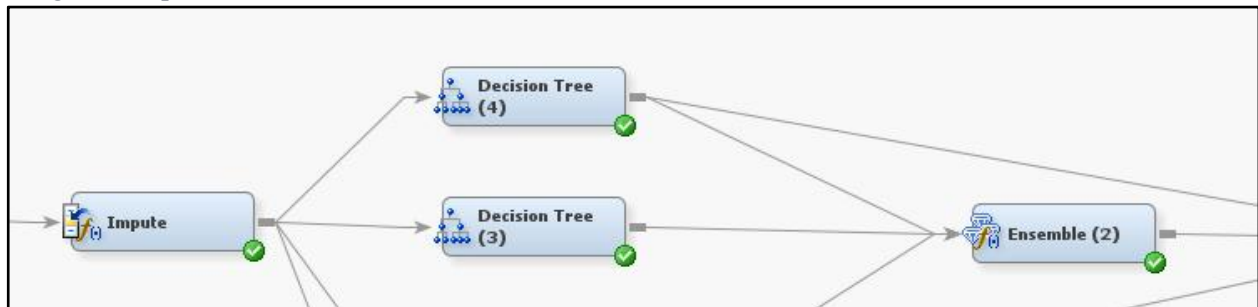
Firstly, targeted marketing campaigns are crucial for addressing the influence of customer age on churn. Businesses should tailor promotions, discounts, and product recommendations to suit the preferences and behaviors associated with different age groups. This personalized approach enhances engagement and loyalty. Secondly, recognizing the significance of membership levels is paramount. Businesses should consider introducing exclusive perks, personalized offers, or loyalty programs to encourage customers to upgrade their memberships or maintain their current status. This fosters a sense of exclusivity and loyalty among customers. Moreover, acknowledging the impact of location on churn is essential. Implementing location-specific marketing strategies tailored to regional preferences and needs can be highly effective. This may involve localized promotions, events, or partnerships that resonate with customers in specific geographic areas. While TotalSpent may have a relatively lower impact, businesses should focus on optimizing strategies to increase customer spending. Introducing tiered pricing, bundled offerings, or exclusive discounts for higher-spending customers can contribute to overall revenue growth.

Continuous monitoring and adaptation are critical components of effective churn mitigation. Regularly updating the decision tree model to align with changing customer behaviors ensures ongoing effectiveness in predicting churn. This proactive approach enables businesses to stay ahead of evolving trends and adjust strategies accordingly.

In summary, a comprehensive strategy involves personalized marketing, loyalty programs, regional targeting, spending optimization, and continuous adaptation. By integrating these recommendations, businesses can create a robust framework to reduce churn, enhance customer satisfaction, and foster long-term loyalty.

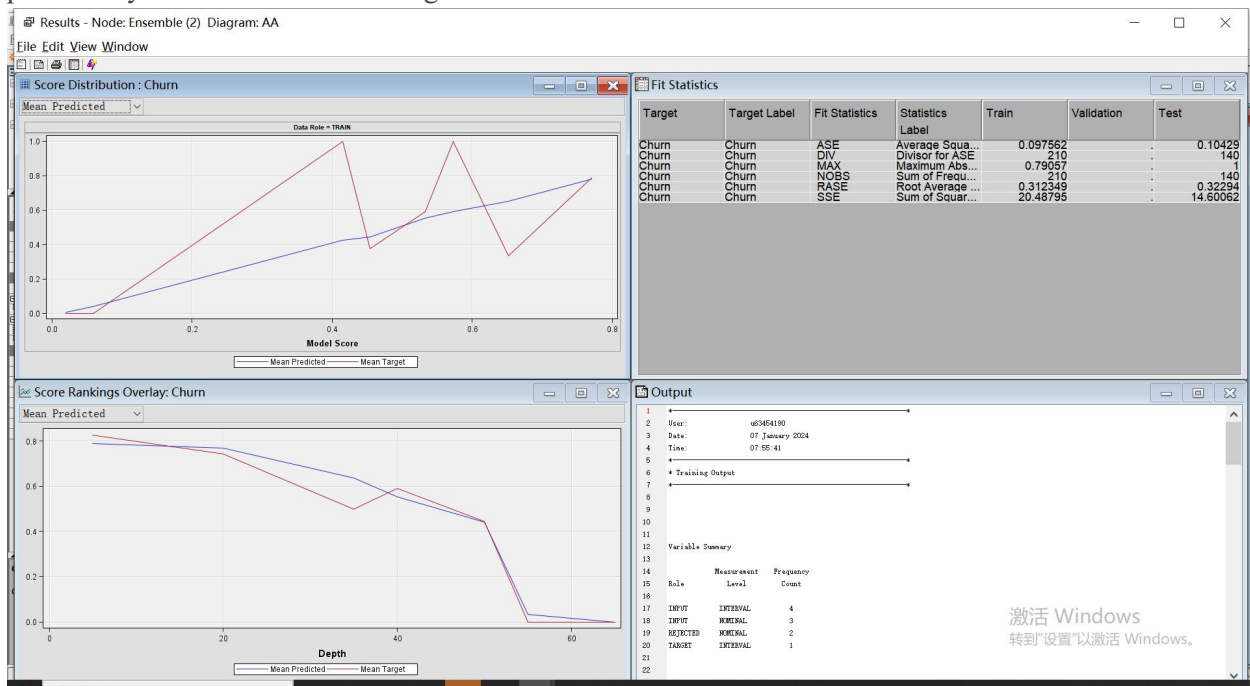
Ensemble Methods: Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.

Drag and drop one more **Decision Tree** and **Ensemble**, connect the two Decision Tree to Ensemble

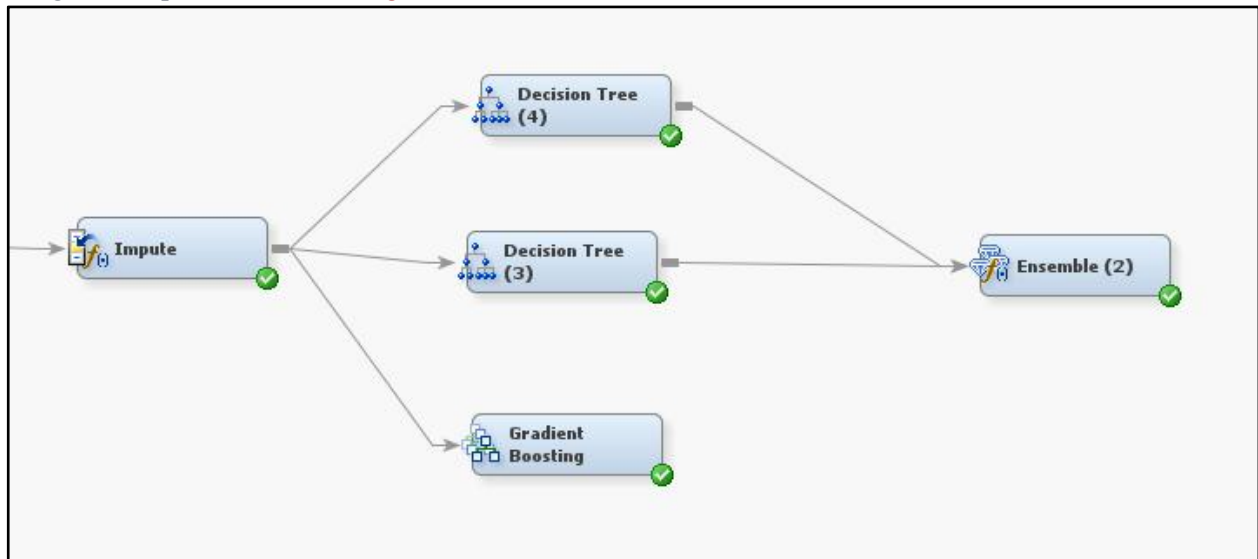


After running, the results are as follows.

The output presents results from an ensemble model combining two decision trees, TREE3 and TREE4. The model, targeting "Churn," demonstrates good accuracy with low Average Squared Error (ASE) and Root ASE values. Assessment scores at various depths and distributions highlight its predictive strength, particularly in the 0.751 - 0.791 range.

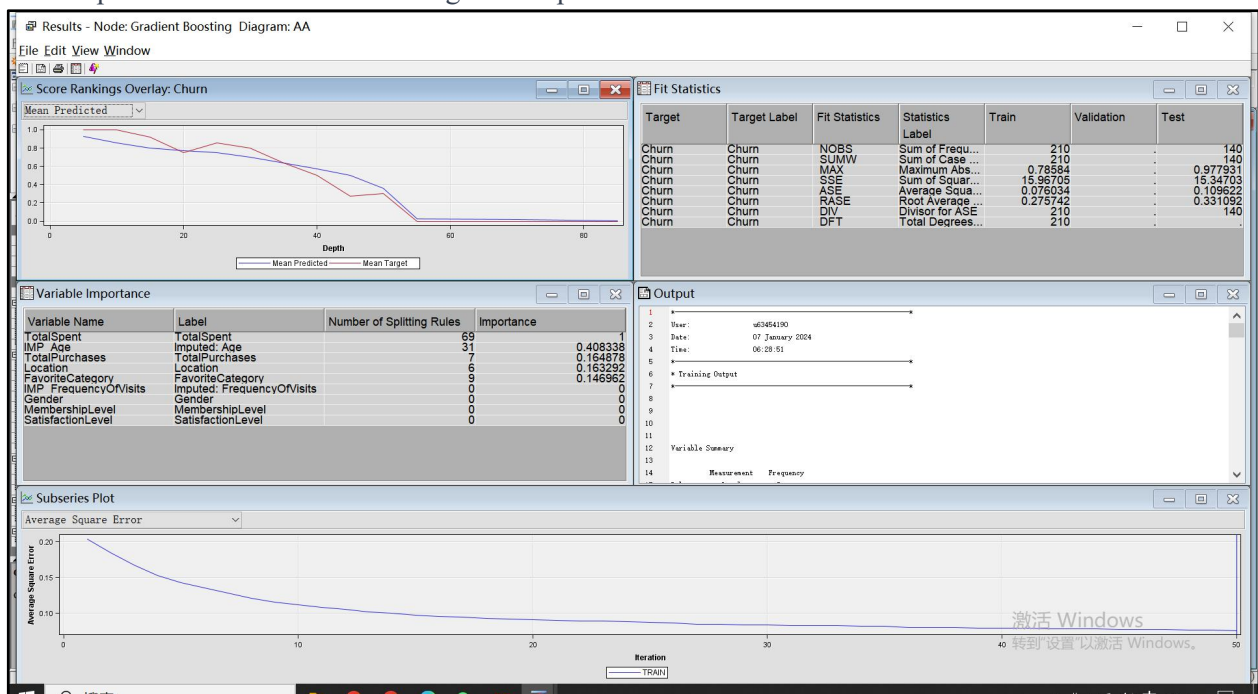


Drag and drop **Gradient Boosting** and run

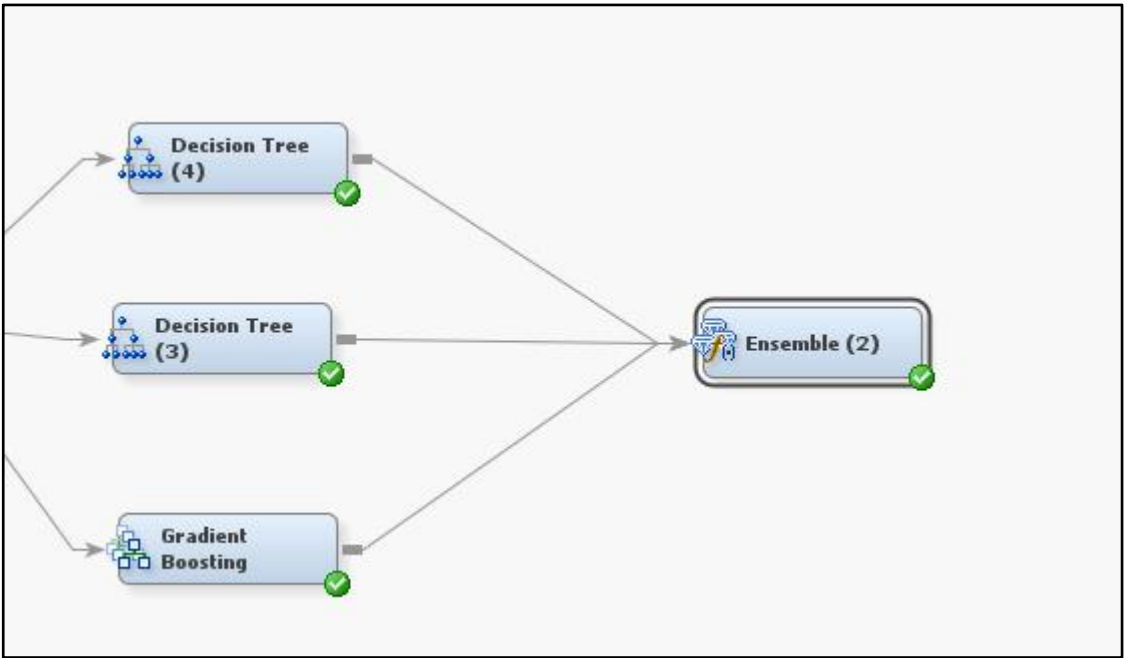


After running, the results are as follows.

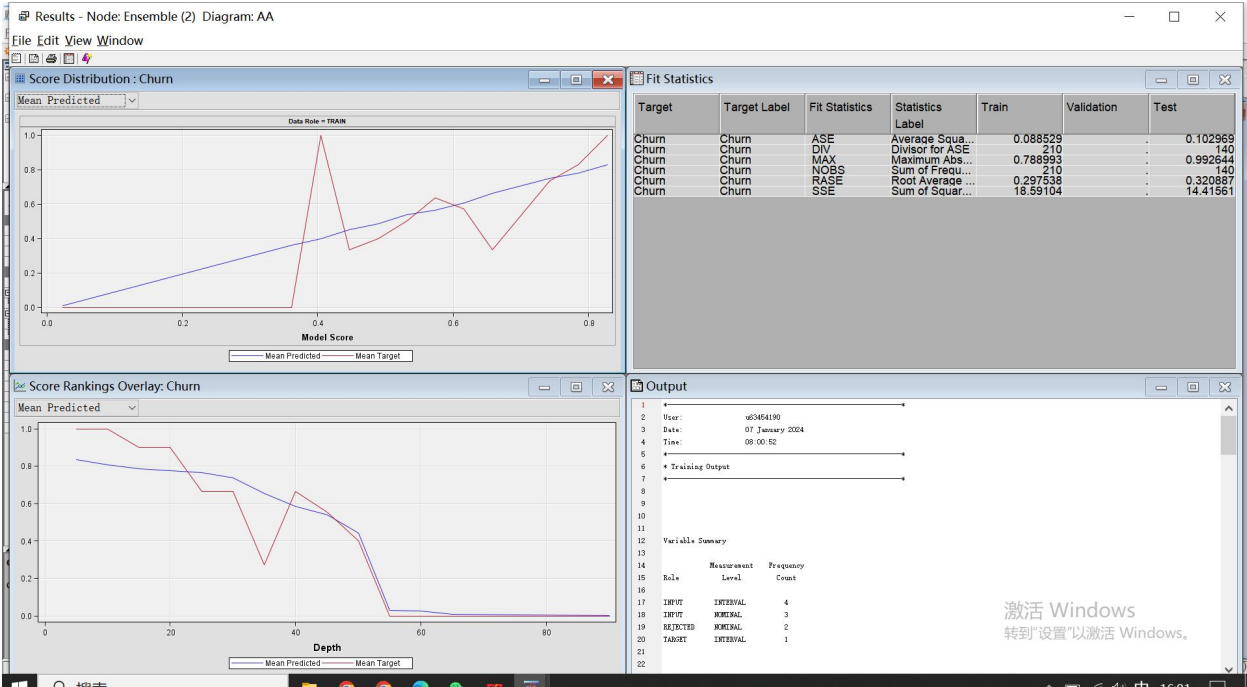
The Gradient Boosting model in SAS Enterprise Miner predicts Churn using five variables, with "TotalSpent" identified as the most significant predictor.



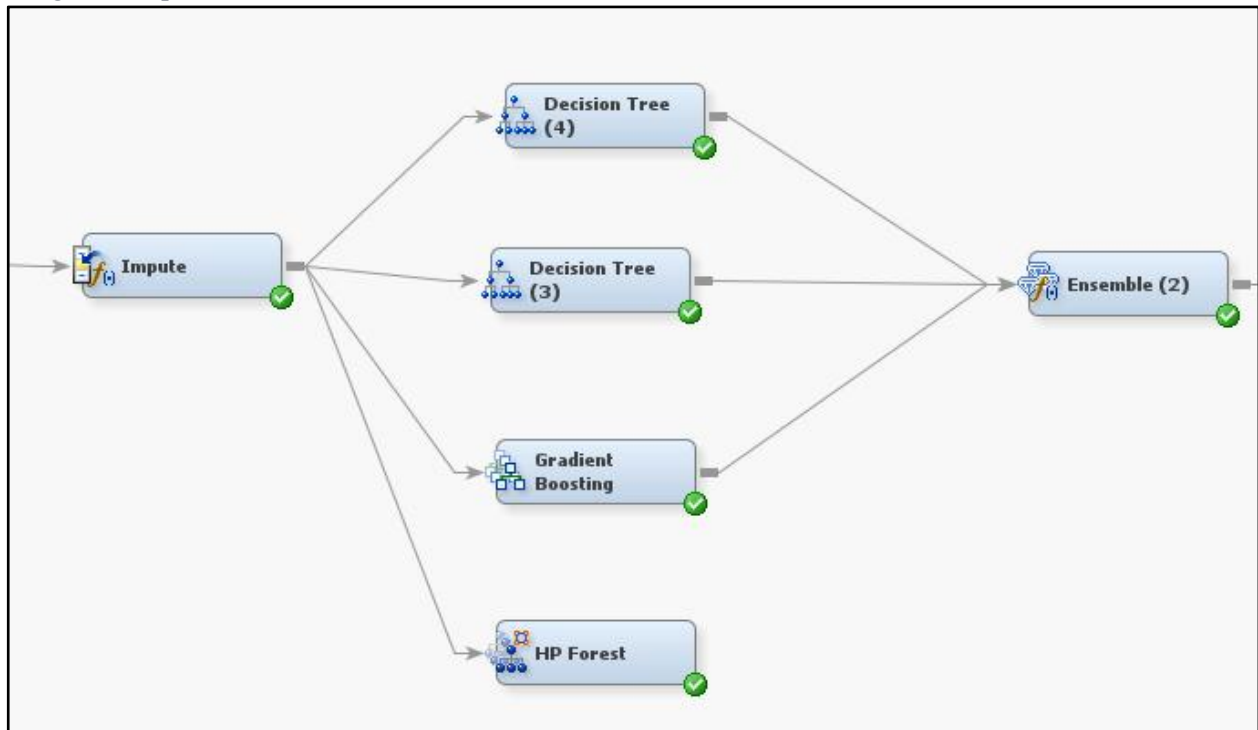
Connect Gradient Boosting to Ensemble and run



After running, the results are as follows.
The ensemble model, combining TREE3, TREE4, and BOOST, predicts churn with an Average Squared Error of 0.089. It excels at depths 5 and 10, showing high predictability. The distribution analysis reveals strong performance in specific ranges (0.806 - 0.848, 0.764 - 0.806).

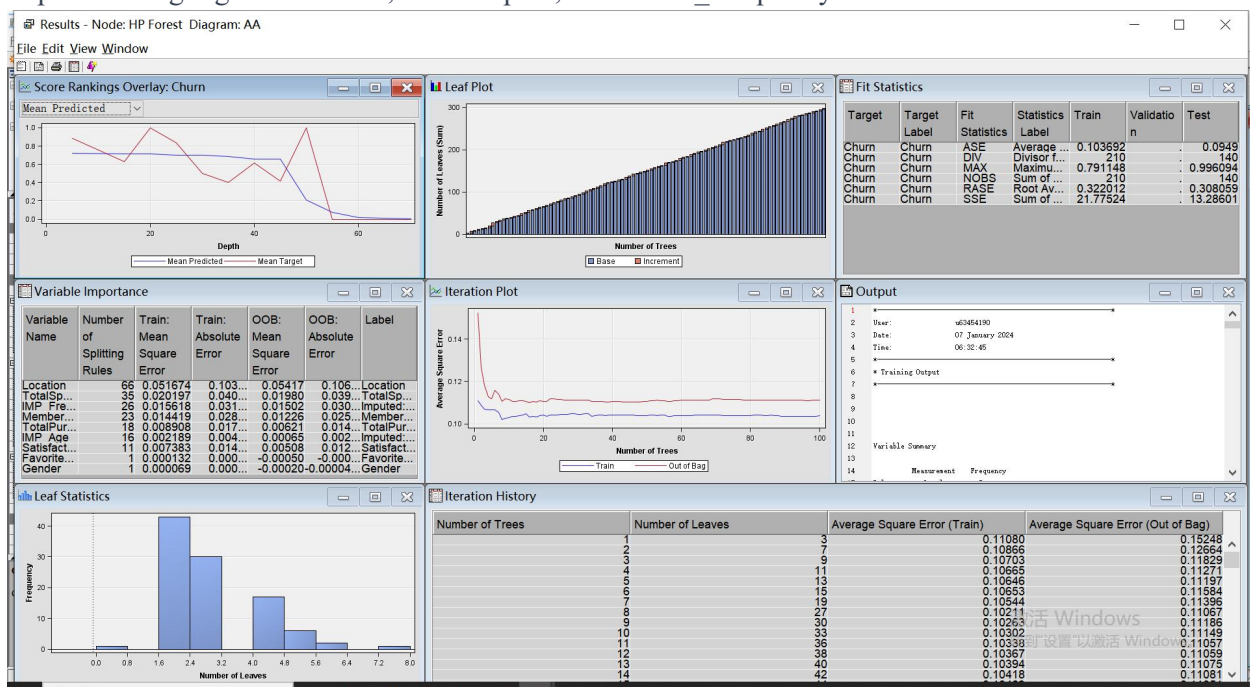


Drag and drop **HP Forest** and run

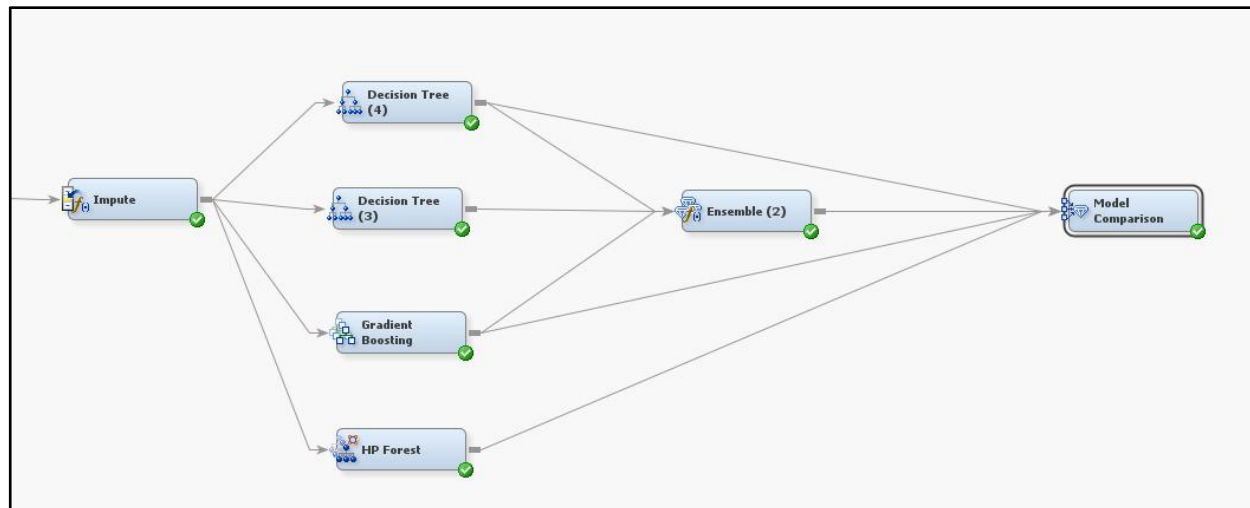


After running, the results are as follows.

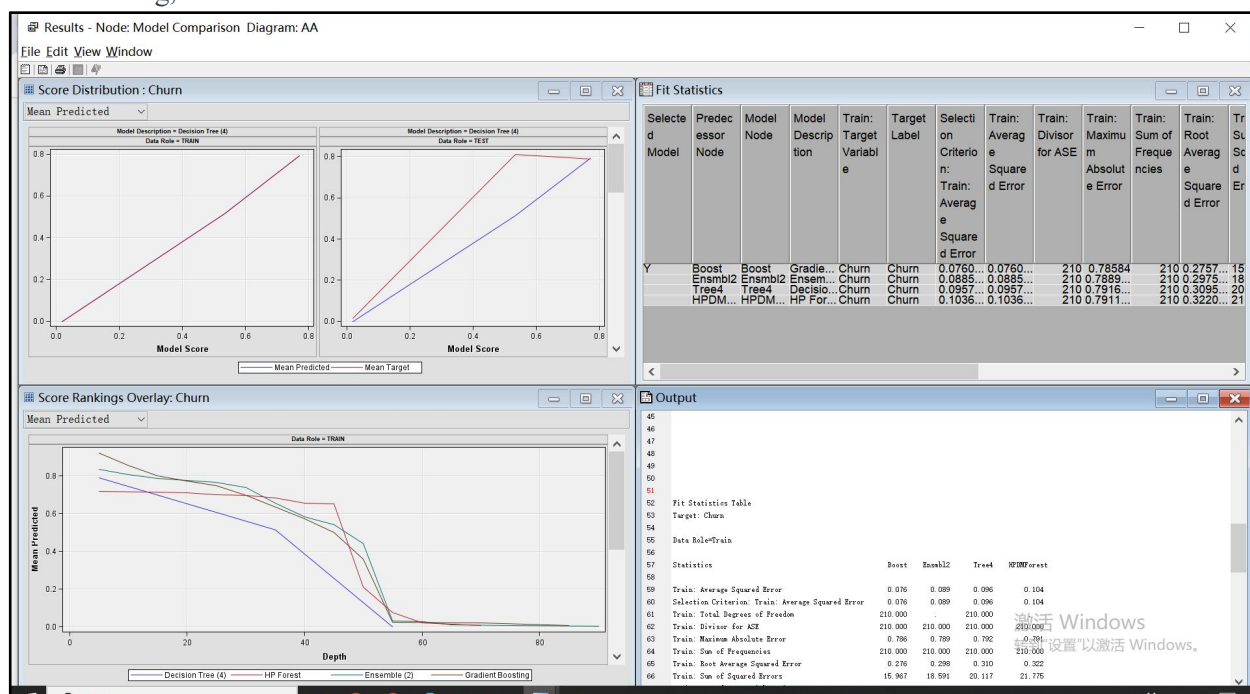
The HP Forest model in SAS Enterprise Miner was trained on 210 observations and evaluated on both training and test datasets. The average squared error (ASE) for training is 0.104, and for testing is 0.095. The model tends to underestimate churn at a depth of 5, suggesting room for improvement. Variable importance highlights "Location," "TotalSpent," and "IMP_FrequencyOfVisits."



Drag and drop **Model Comparison** and connect with Decision Tree, Ensemble, Gradient Boosting and HP Forest then run



After running, the results are as follows.

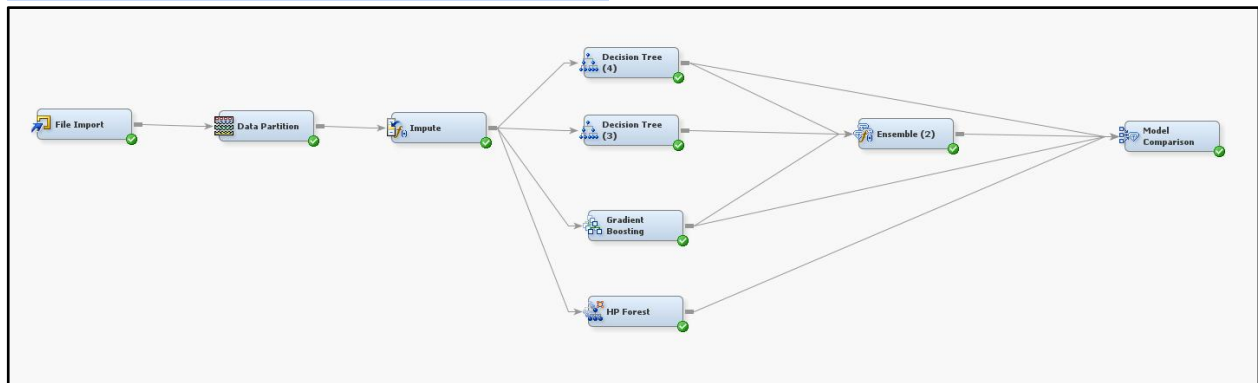


In Model Comparison:

The performance evaluation of each model provides valuable insights into their effectiveness in predicting churn. The Boost (Gradient Boosting) model demonstrates strong performance during training, achieving the lowest Average Squared Error (_ASE_) of 0.07603. However, a slight increase in the test _ASE_ to 0.110 suggests a potential risk of overfitting, necessitating further adjustments for improved generalization. The Ensemble model (Ensmbl2) stands out with a well-balanced performance, displaying a train _ASE_ of 0.08853 and a test _ASE_ of 0.103. The Root Average Squared Error (RASE) values of

0.298 (train) and 0.321 (test) indicate robustness, positioning it as a promising choice for predicting churn. While the Decision Tree (Tree4) model performs reasonably well, a slight increase in the test _ASE_ (0.107) indicates potential overfitting, prompting optimization efforts for enhanced generalization. The HP Forest (HPDPMForest) model shows adequacy in performance, with a test _ASE_ of 0.095 lower than its training _ASE_ of 0.103, making it a viable candidate for deployment. In summary, the Ensemble model (Ensmbl2) emerges as a robust and balanced choice, with continuous monitoring and updates recommended for all models to adapt to evolving data patterns and ensure sustained effectiveness in predicting customer churn.

The overall workflow of SAS EM is as follows:



Reflections and Learning Outcomes:

1. Dataset Integration Outcome: The process of integrating datasets using Talend Data Integration resulted in a successful merge based on the 'CustomerID' column. While initial challenges were faced, such as ensuring a seamless integration, troubleshooting techniques and effective utilization of Talend Data Integration tools led to a positive outcome. The final integrated dataset now serves as a foundation for more comprehensive insights into customer interactions and transactions within the e-commerce platform.
2. Data Standardization Importance: The significance of data standardization using Talend Data Preparation became evident in enhancing the consistency of categorical values. Recognizing the impact of standardized data on subsequent analyses underscored the importance of meticulous data preprocessing for meaningful insights.
3. SAS Enterprise Miner Proficiency: Engaging with SAS Enterprise Miner for data preprocessing and decision tree analysis provided hands-on experience with a powerful analytics tool. Overcoming the learning curve involved a combination of online tutorials, documentation exploration, and practical application, contributing to enhanced proficiency.

Challenges:

Challenge 1 - Dataset Integration: During the Dataset Integration process, I encountered difficulties related to specific issues in the merging steps. To address these challenges, I referred to the materials provided in Chapter 3 slides, which offered clear and detailed steps. Following the guidelines, I successfully navigated through the integration process, ensuring a smooth merge based on the 'CustomerID' column.

Challenge 2 - Decision Tree Configuration: When running the decision tree, I faced a hurdle as there were no nodes in the tree. To overcome this obstacle, I proactively searched for solutions on Google. I discovered that configuring certain parameters, changing the criterion to 'gini,' was necessary for optimal performance. Implementing this adjustment enabled the decision tree to generate more insightful and accurate results.