WQD7005 Data mining Case Study
22062358 HAN DIE

Dataset Source: https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset
GitHub: https://github.com/22062358HanDie/Case-Study_data-mining

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.

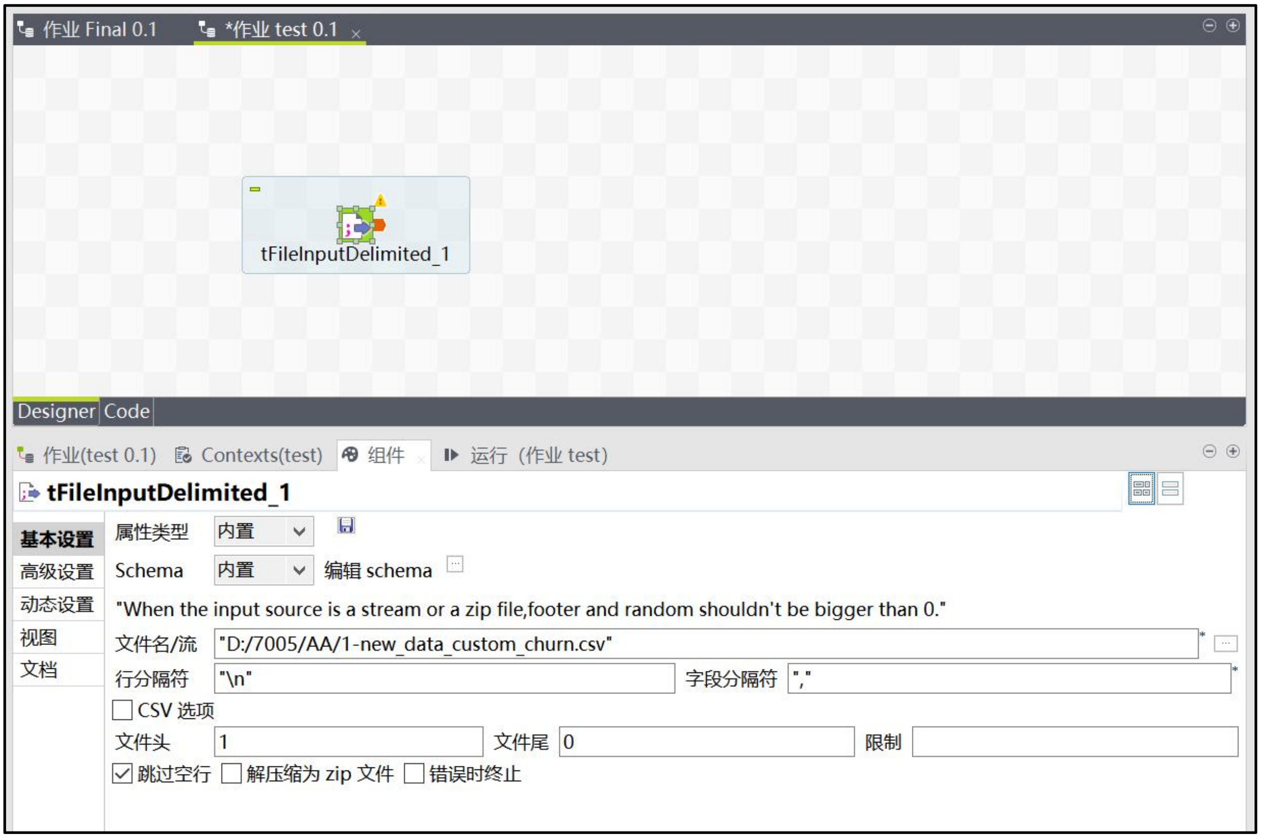
Subsequently, the analysis transitions to SAS Enterprise Miner, where the imported dataset undergoes preprocessing. This step involves handling missing values and specifying variable roles, setting the stage for robust analysis. The subsequent phases encompass Decision Tree Analysis and the application 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.



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 "D:/7005/AA/1-new_data_custom_churn.csv"

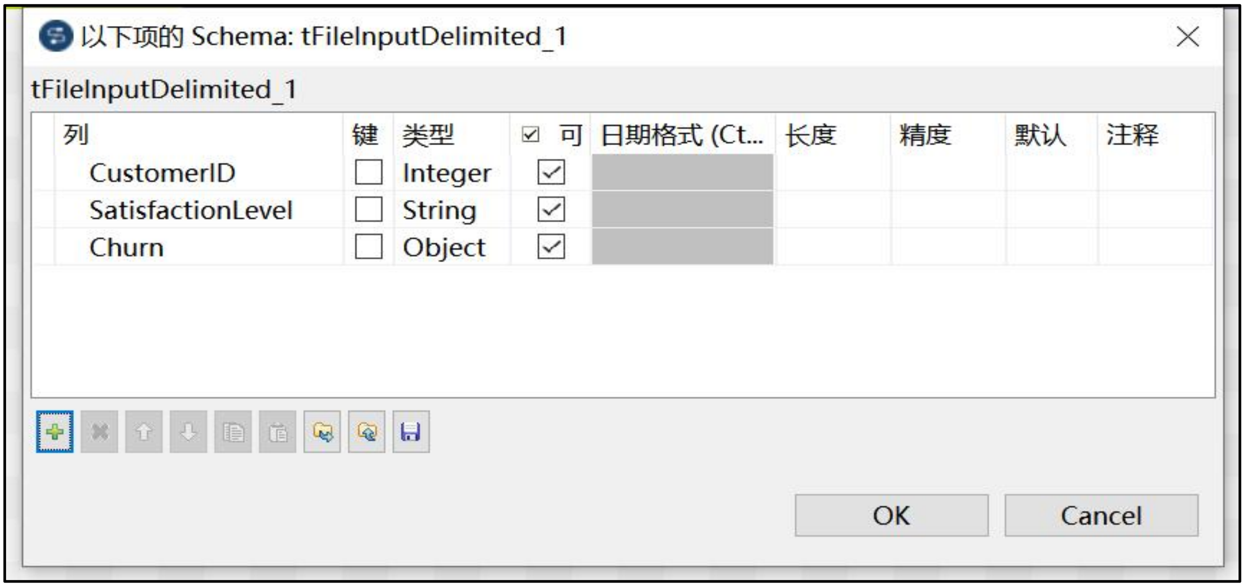
行分隔符 "\n" 字段分隔符 ","

☐ 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

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

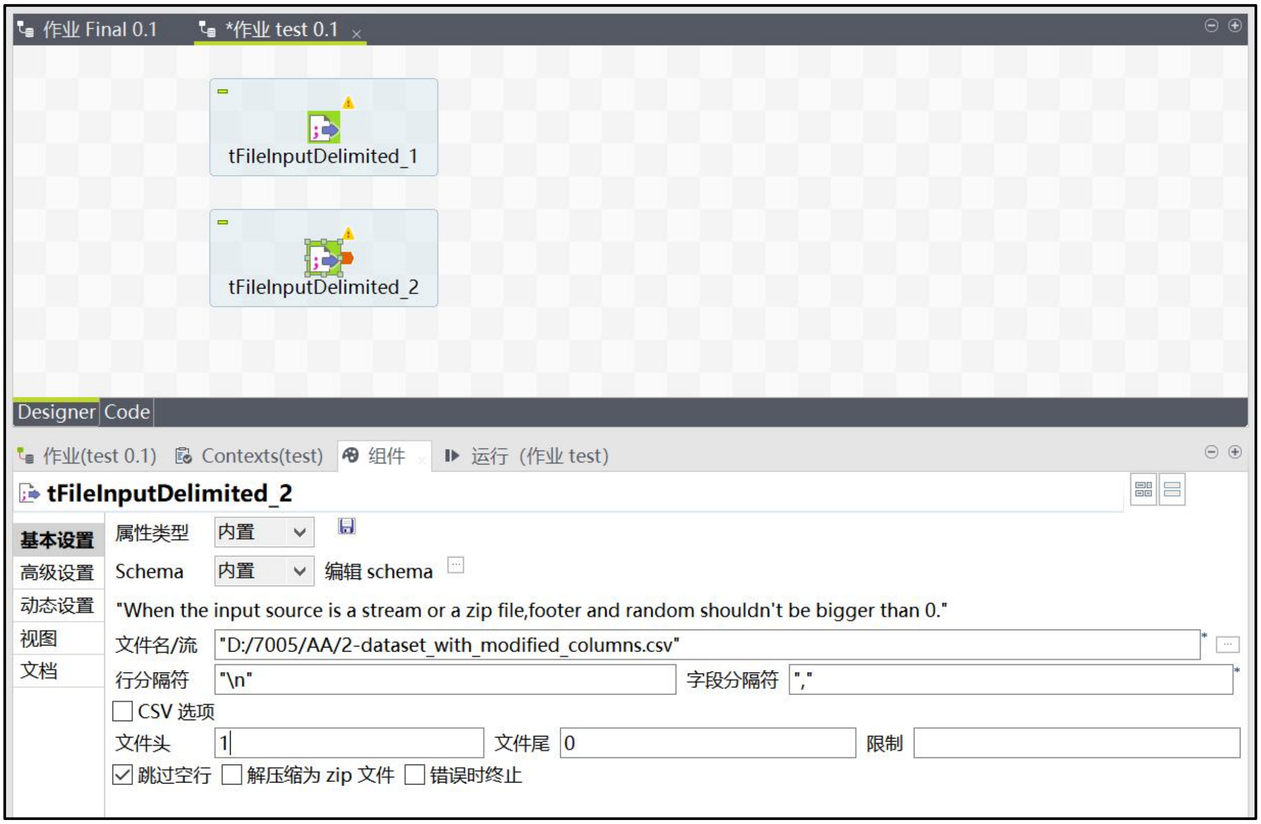


以下项的 Schema: tFileInputDelimited_1

tFileInputDelimited_1

列	键	类型	☑ 可	日期格式 (Ct...	长度	精度	默认	注释
CustomerID	☐	Integer	☑					
SatisfactionLevel	☐	String	☑					
Churn	☐	Object	☑					

Same as above



tFileInputDelimited_2

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 "D:/7005/AA/2-dataset_with_modified_columns.csv"

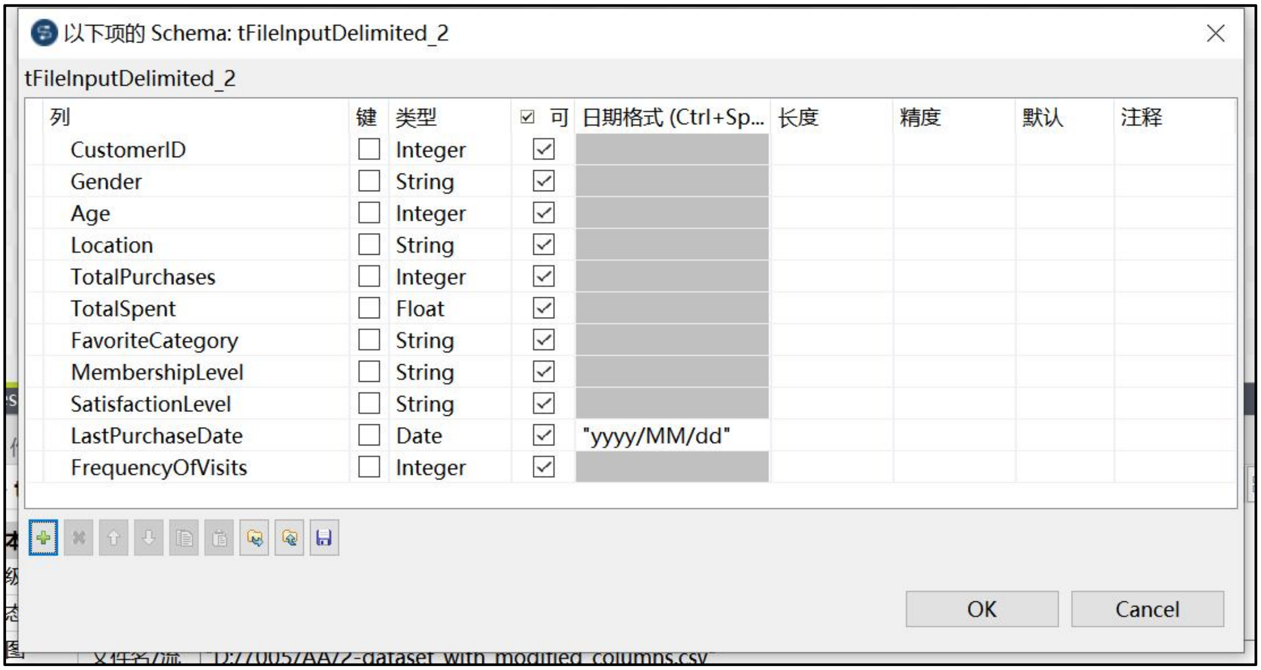
行分隔符 "\n" 字段分隔符 ","

□ 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

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

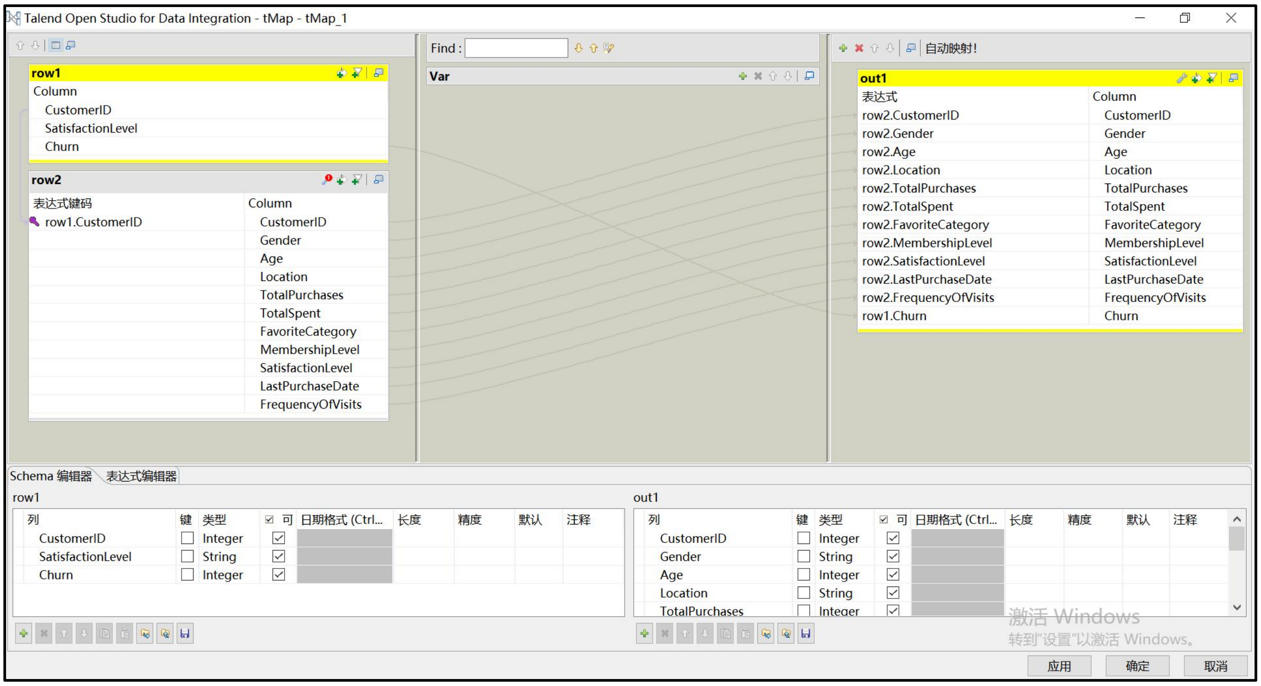


以下项的 Schema: tFileInputDelimited_2

tFileInputDelimited_2

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

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.



row1

Column
CustomerID
SatisfactionLevel
Churn

row2

表达式键码	Column
row1.CustomerID	CustomerID
	Gender
	Age
	Location
	TotalPurchases
	TotalSpent
	FavoriteCategory
	MembershipLevel
	SatisfactionLevel
	LastPurchaseDate
	FrequencyOfVisits

out1

表达式	Column
row2.CustomerID	CustomerID
row2.Gender	Gender
row2.Age	Age
row2.Location	Location
row2.TotalPurchases	TotalPurchases
row2.TotalSpent	TotalSpent
row2.FavoriteCategory	FavoriteCategory
row2.MembershipLevel	MembershipLevel
row2.SatisfactionLevel	SatisfactionLevel
row2.LastPurchaseDate	LastPurchaseDate
row2.FrequencyOfVisits	FrequencyOfVisits
row1.Churn	Churn

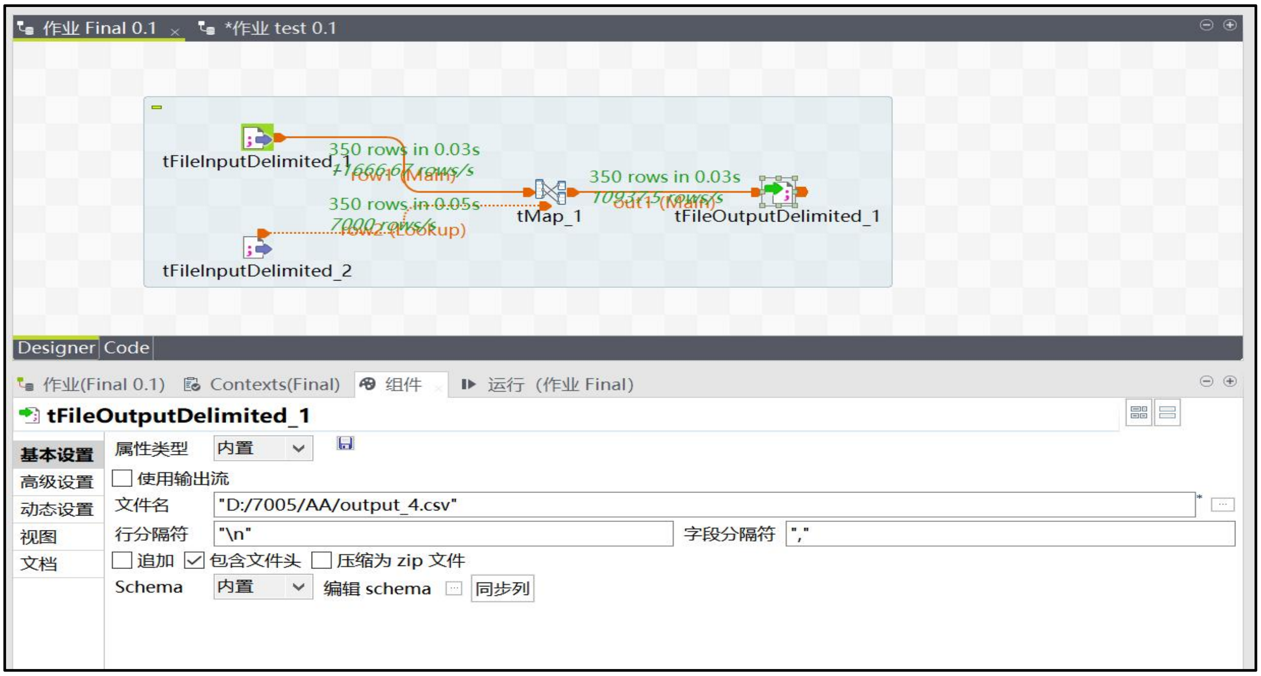
row1

列	键	类型	☑ 可	日期格式 (Ctrl...	长度	精度	默认	注释
CustomerID	□	Integer	☑					
SatisfactionLevel	□	String	☑					
Churn	□	Integer	☑					

out1

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

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.



tFileOutputDelimited_1

属性类型 内置

□ 使用输出流

文件名 "D:/7005/AA/output_4.csv"

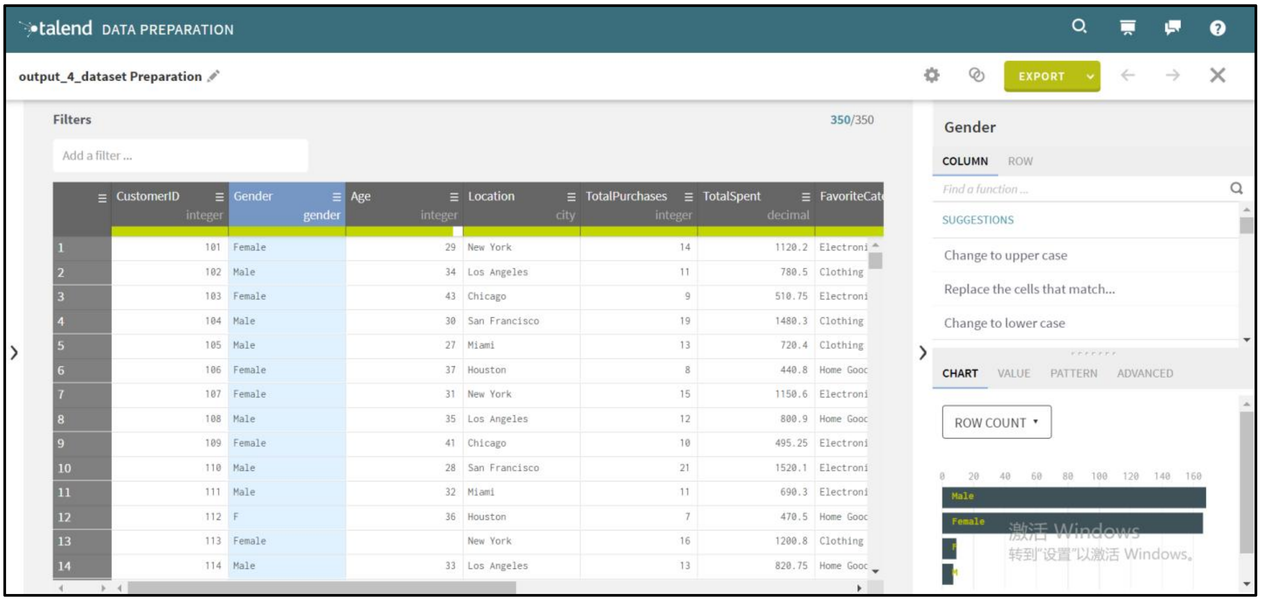
行分隔符 "\n" 字段分隔符 ","

□ 追加 ☑ 包含文件头 □ 压缩为 zip 文件

Schema 内置 编辑 schema 同步列

Data Modification Using Talend Data Preparation

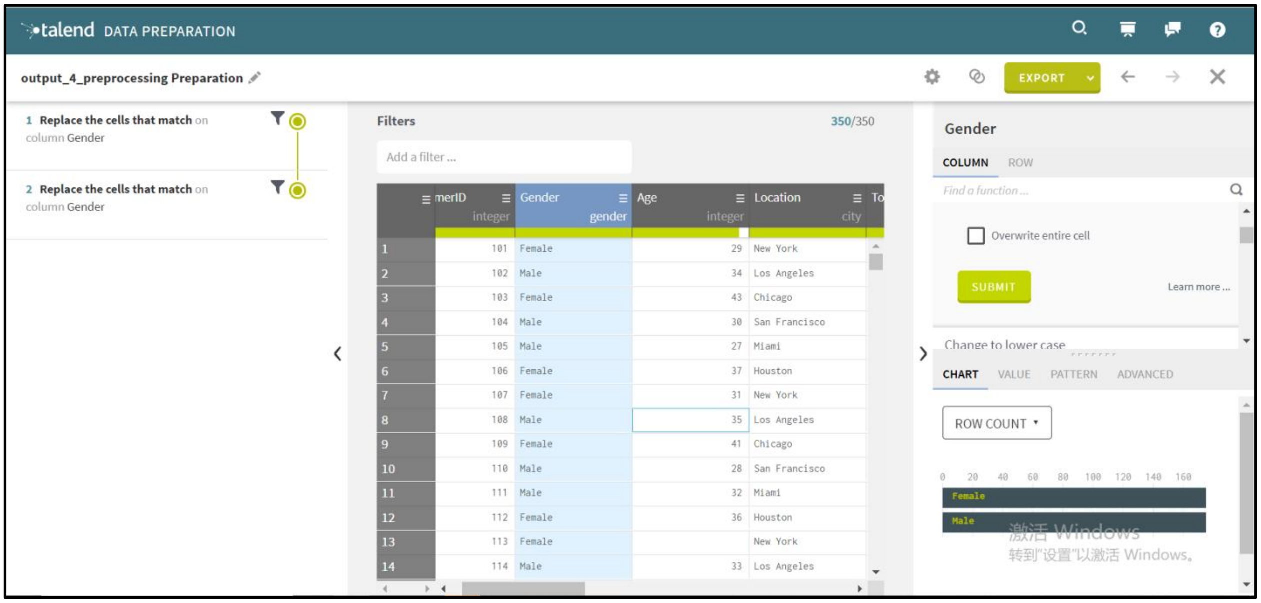
Before: In the Gender column, there are four categories with M, F, Male and Female. The inputs are not standardized because F represents Female and M represents Male.



	CustomerID integer	Gender gender	Age integer	Location city	TotalPurchases integer	TotalSpent decimal	FavoriteCate
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 Gooc
7	107	Female	31	New York	15	1150.6	Electroni
8	108	Male	35	Los Angeles	12	800.9	Home Gooc
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 Gooc
13	113	Female		New York	16	1200.8	Clothing
14	114	Male	33	Los Angeles	13	820.75	Home Gooc

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

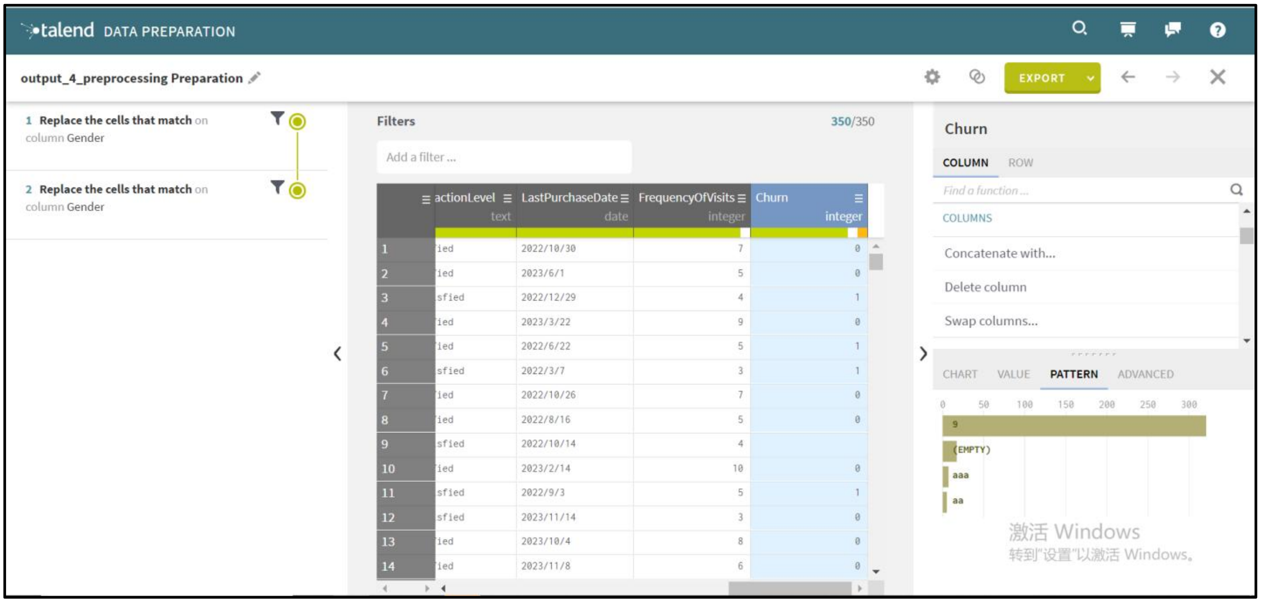
After: Replace the cells that match F to Female and M to Male



	merID integer	Gender gender	Age integer	Location city	To
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.



output_4_preprocessing Preparation

1 Replace the cells that match on column Gender

2 Replace the cells that match on column Gender

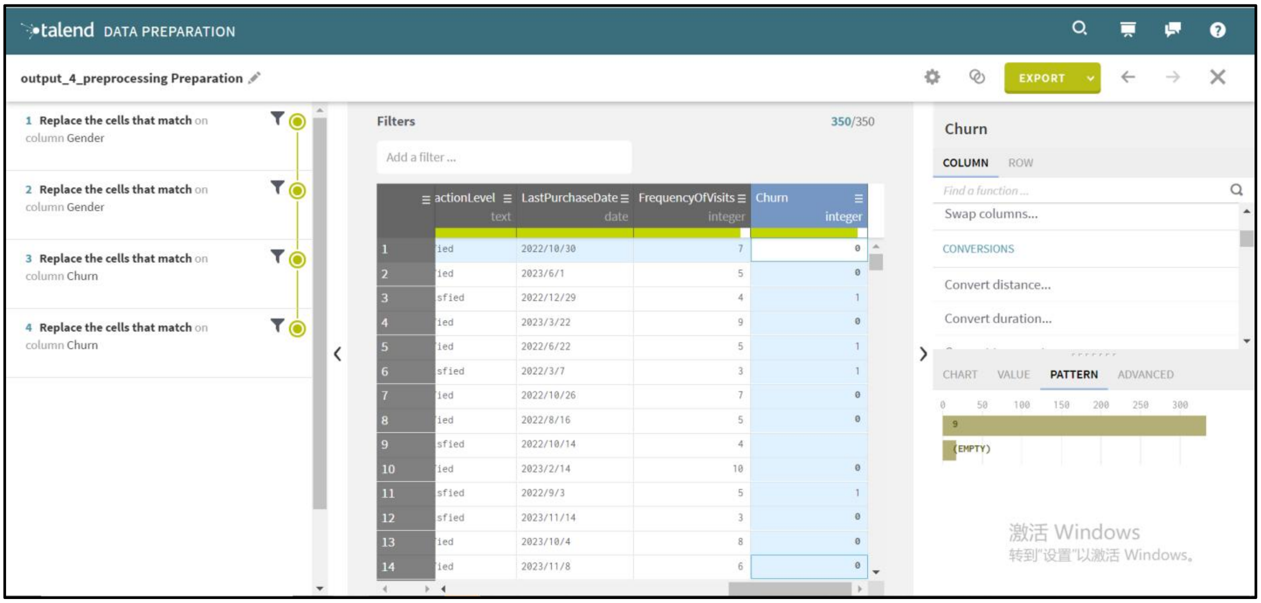
Filters 350/350

Add a filter ...

	actionLevel text	LastPurchaseDate date	FrequencyOfVisits integer	Churn 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	
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

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.



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 350/350

Add a filter ...

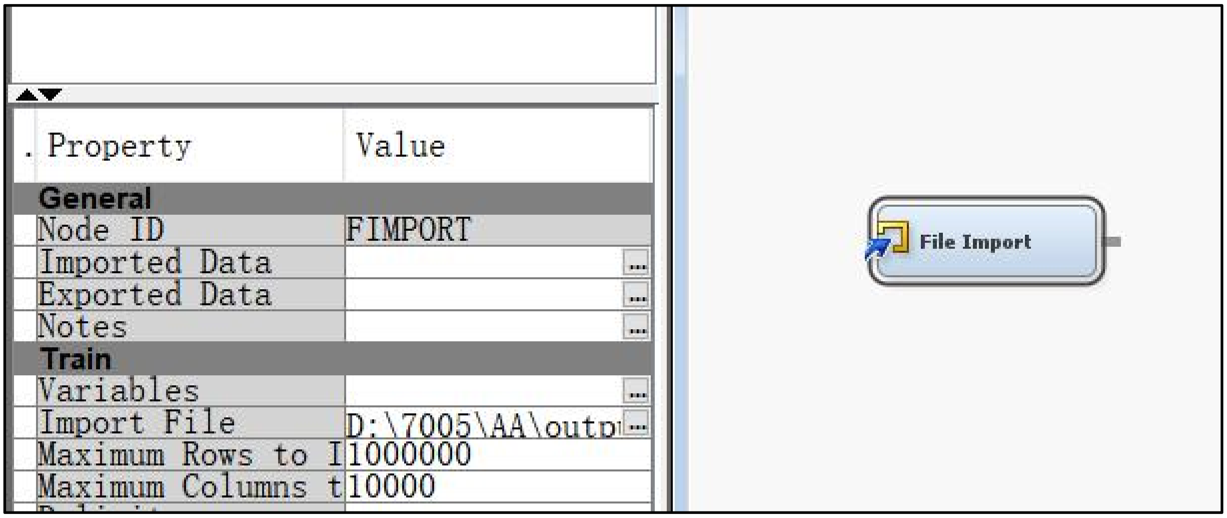
	actionLevel text	LastPurchaseDate date	FrequencyOfVisits integer	Churn integer
1	ied	2022/10/30	7	0
2	ied	2023/6/1	5	0
3	sfied	2022/12/29	4	1
4	ied	2023/3/22	9	0
5	ied	2022/6/22	5	1
6	sfied	2022/3/7	3	1
7	ied	2022/10/26	7	0
8	ied	2022/8/16	5	0
9	sfied	2022/10/14	4	
10	ied	2023/2/14	10	0
11	sfied	2022/9/3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

After this, export the file.

Tasks in SAS EM

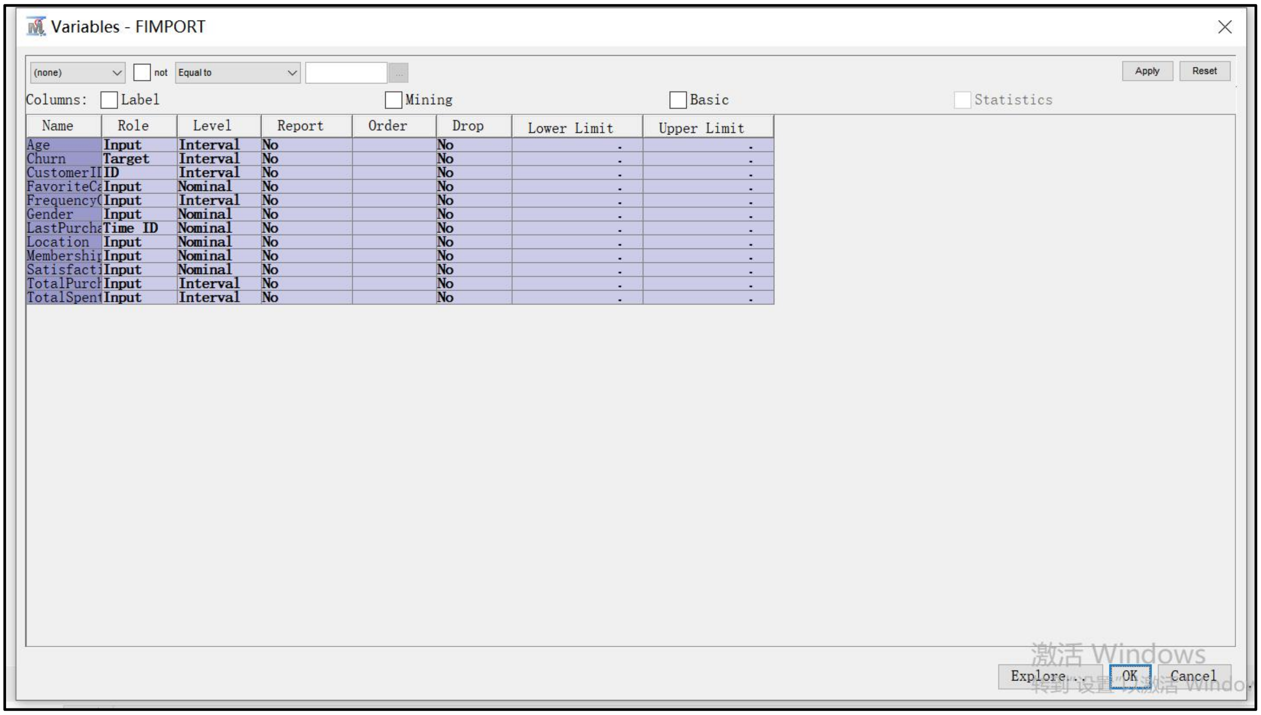
Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

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



. Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	D:\7005\AA\outpu...
Maximum Rows to I	1000000
Maximum Columns t	10000

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.



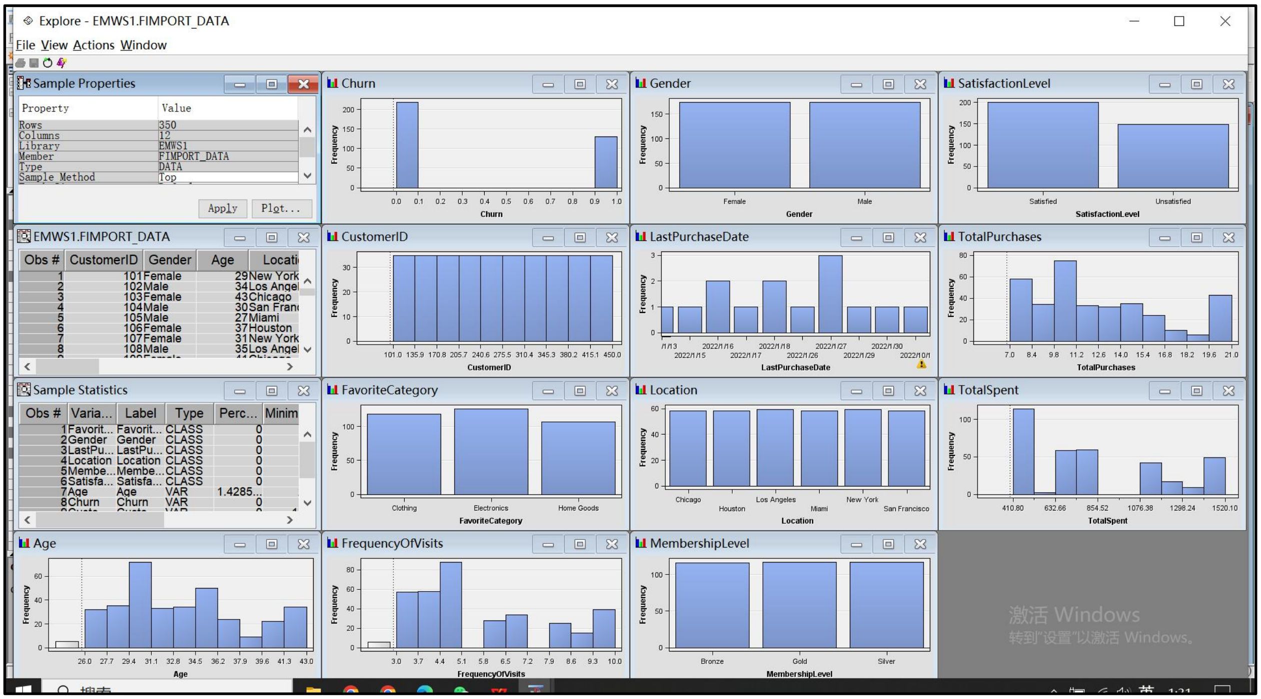
Variables - FIMPORT

(none) □ not Equal to

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	.	.
CustomerII	ID	Interval	No		No	.	.
FavoriteCa	Input	Nominal	No		No	.	.
Frequency(	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurcha	Time ID	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
Membershi	Input	Nominal	No		No	.	.
Satisfacti	Input	Nominal	No		No	.	.
TotalPurch	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

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

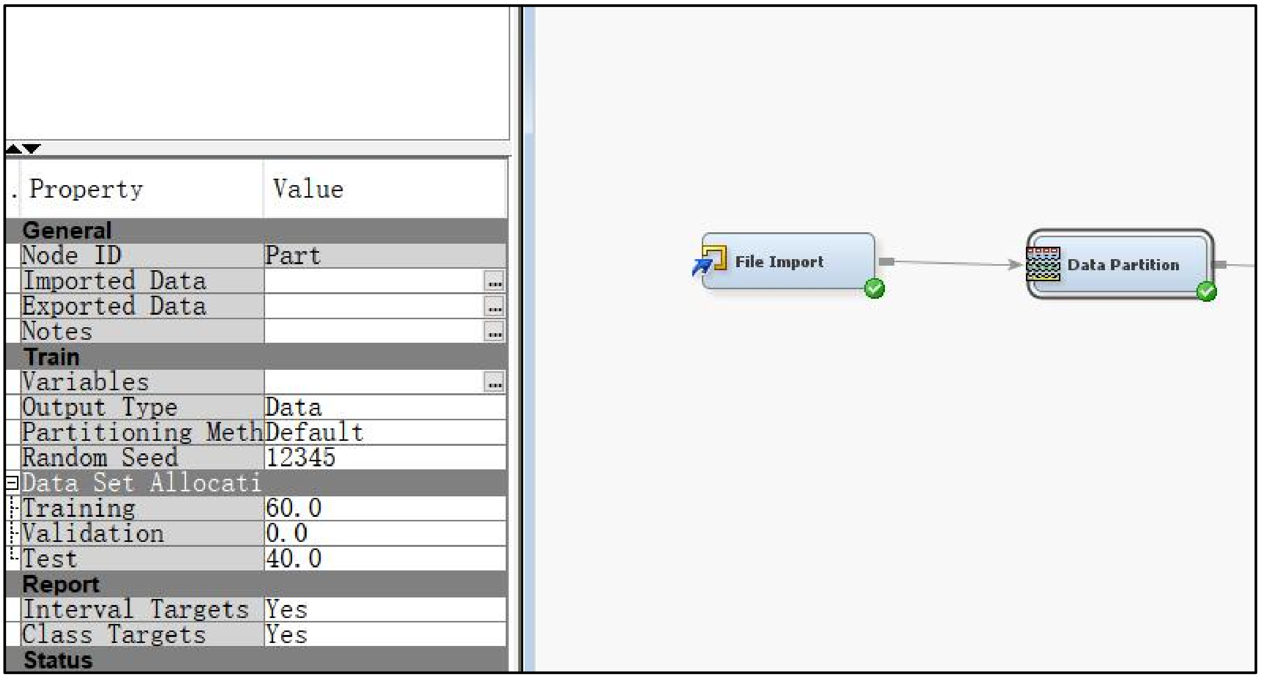


Property	Value
Rows	350
Columns	12
Library	EMWS1
Member	FIMPORT_DATA
Type	DATA
Sample Method	Top

Obs #	CustomerID	Gender	Age	Locati
1	101	Female	29	New York
2	102	Male	34	Los Angel
3	103	Female	43	Chicago
4	104	Male	30	San Franc
5	105	Male	27	Miami
6	106	Female	37	Houston
7	107	Female	31	New York
8	108	Male	35	Los Angel

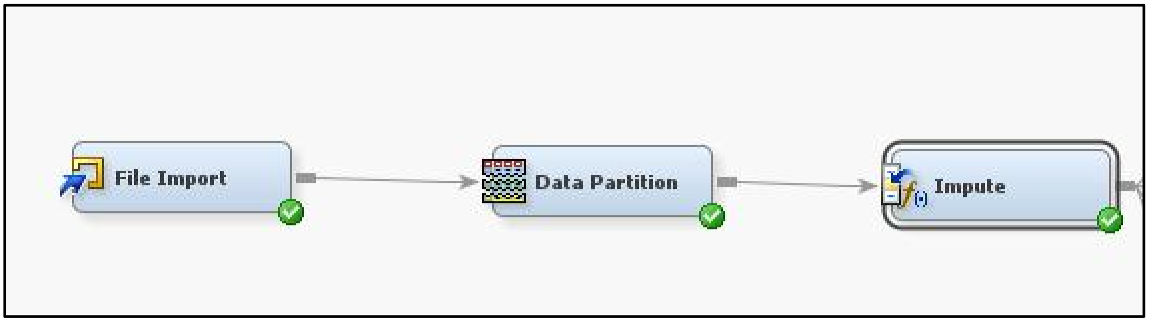
Obs #	Varia...	Label	Type	Perc...	Minim
1	Favorit...	Favorit...	CLASS	0	
2	Gender	Gender	CLASS	0	
3	LastPu...	LastPu...	CLASS	0	
4	Location	Location	CLASS	0	
5	Membe...	Membe...	CLASS	0	
6	Satisfa...	Satisfa...	CLASS	0	
7	Age	Age	VAR	1.4285...	
8	Churn	Churn	VAR	0	

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

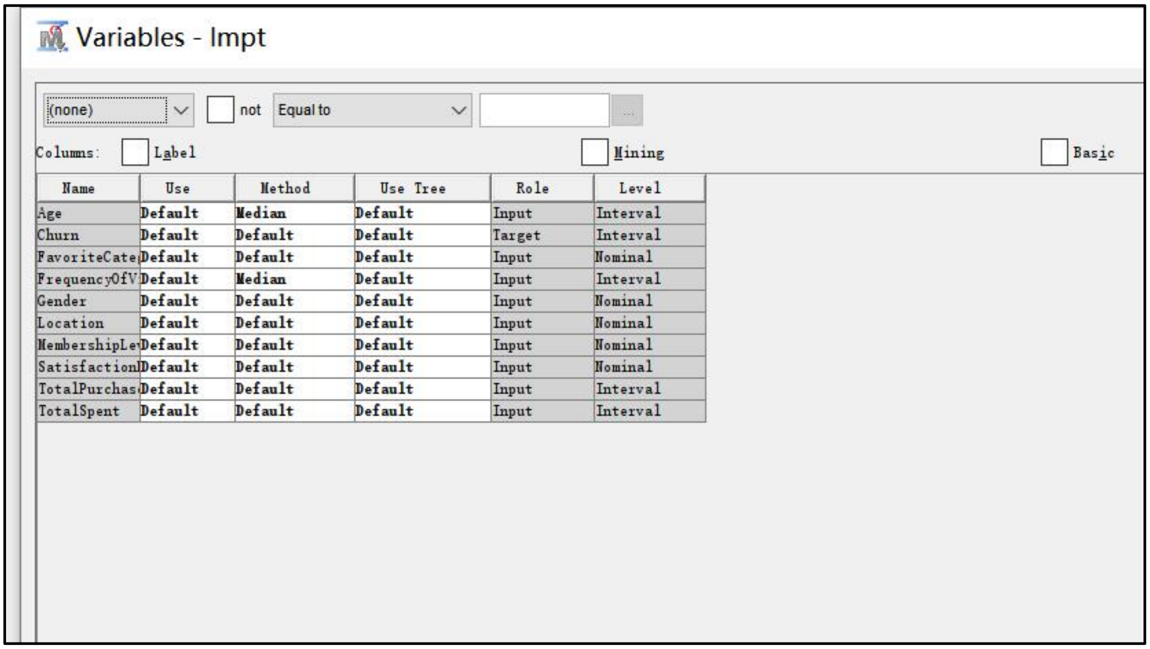


Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Meth	Default
Random Seed	12345
Data Set Allocati	
Training	60.0
Validation	0.0
Test	40.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	

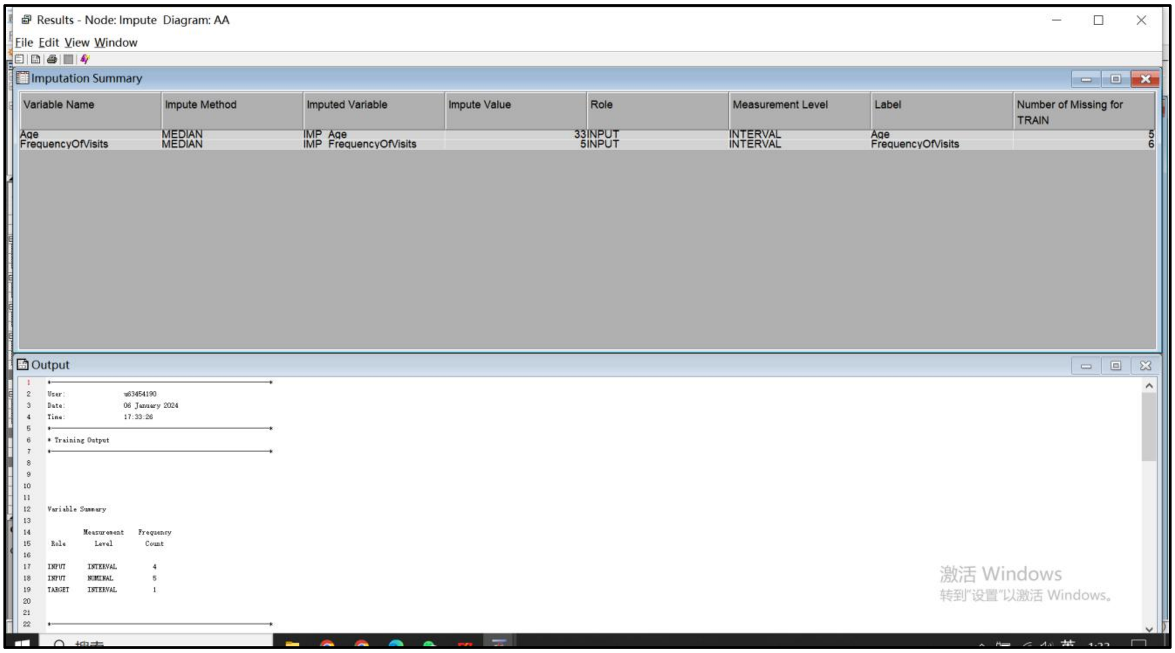
Drag and drop Impute and edit the variables, then run. help to input the missing values in the Age and FrequencyOfVisits columns.



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



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

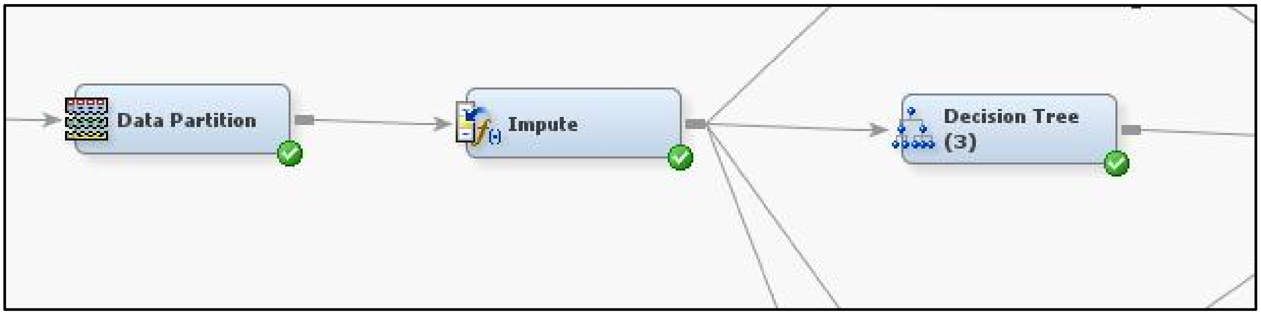


Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEDIAN	IMP_Age	33	INPUT	INTERVAL	Age	5
FrequencyOfVisits	MEDIAN	IMP_FrequencyOfVisits	5	INPUT	INTERVAL	FrequencyOfVisits	6

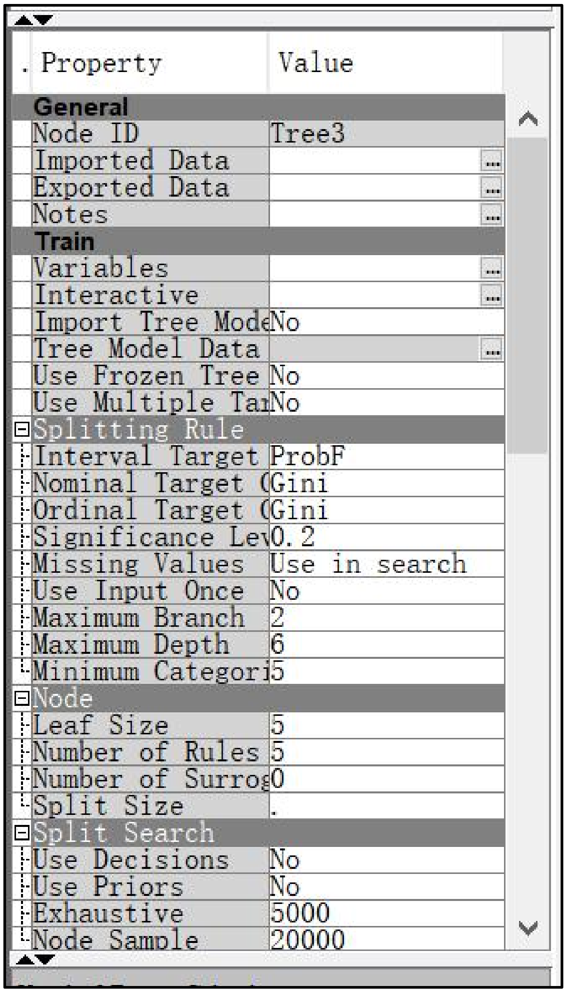
```
*------------------------------------------------------------*
User:               u63454190
Date:               06 January 2024
Time:               17:33:26
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role        Level         Count

INPUT      INTERVAL          4
INPUT      NOMINAL           5
TARGET     INTERVAL          1


*------------------------------------------------------------*
```

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

Drag and drop Decision Tree and run

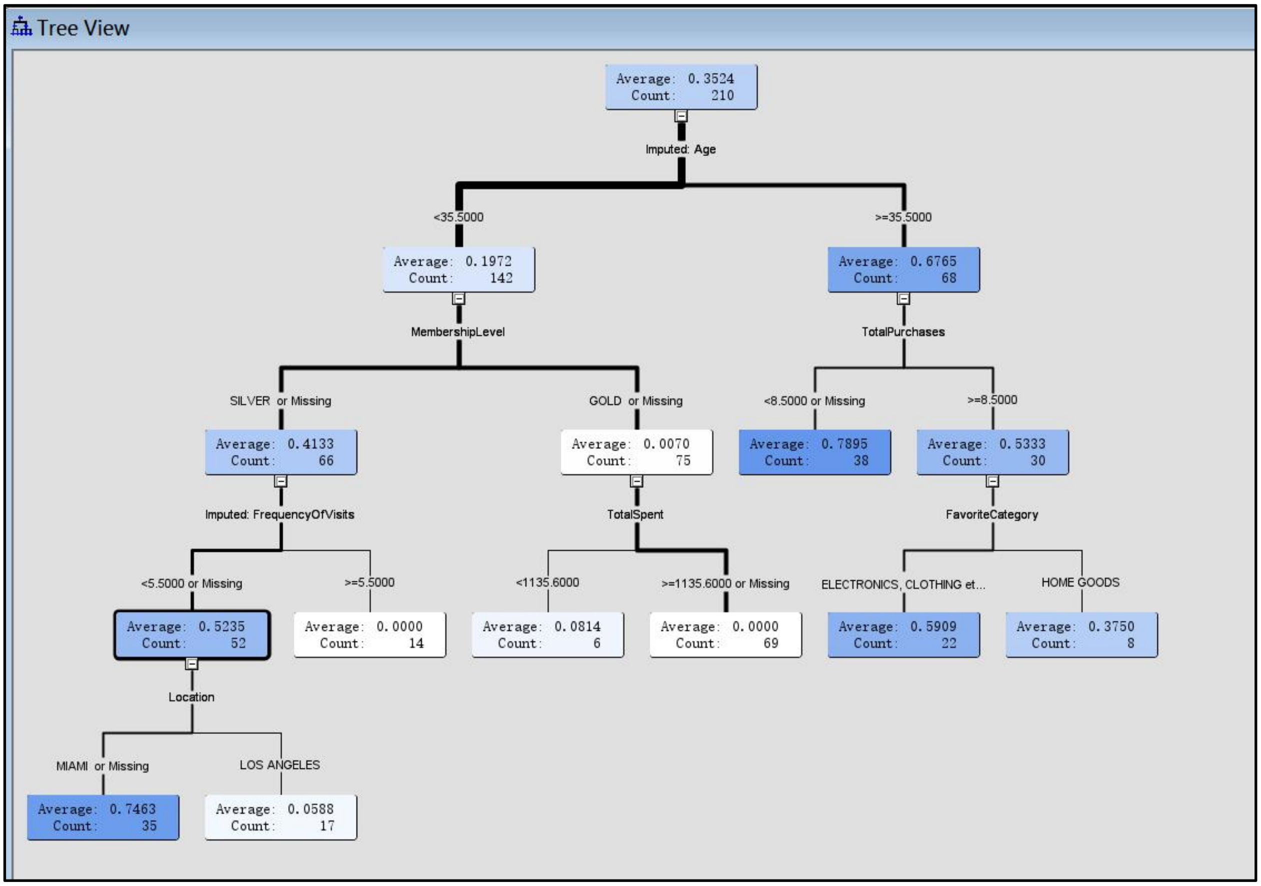


During my project, my decision tree has no nodes. Then I searched on Google and found that I could change the configuration, so I changed it and it ran successfully. I changed the criterion to gini.

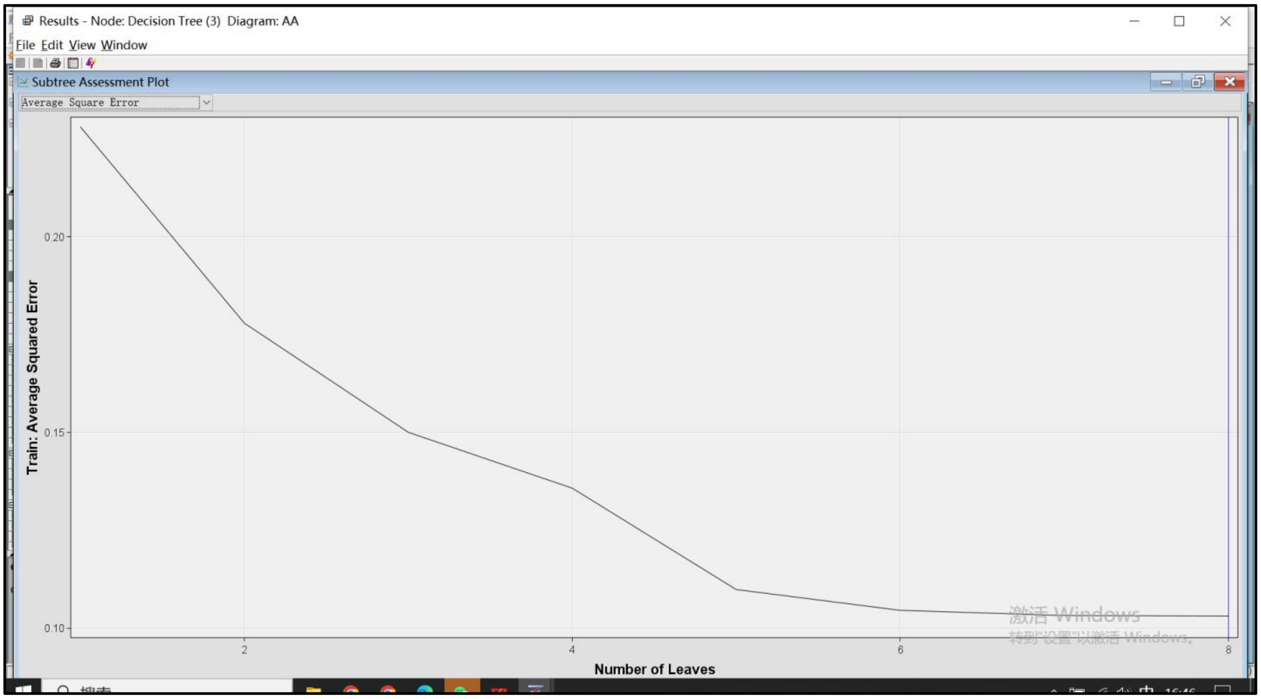


. Property	Value
General	
Node ID	Tree3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Mode	No
Tree Model Data	...
Use Frozen Tree	No
Use Multiple Tar	No
⊟Splitting Rule	
Interval Target	ProbF
Nominal Target (	Gini
Ordinal Target (	Gini
Significance Lev	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categori	5
⊟Node	
Leaf Size	5
Number of Rules	5
Number of Surrog	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.

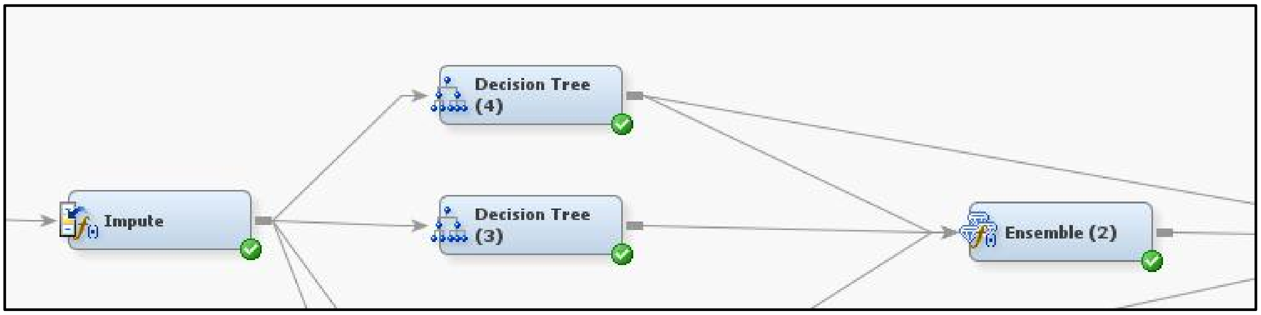
Firstly, targeted marketing campaigns are crucial for addressing the influence of customer age on churn. Businesses should tailor promotions, discounts, and product recommendations to suit the preferences and behaviors associated with different age groups. This personalized approach enhances engagement 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.

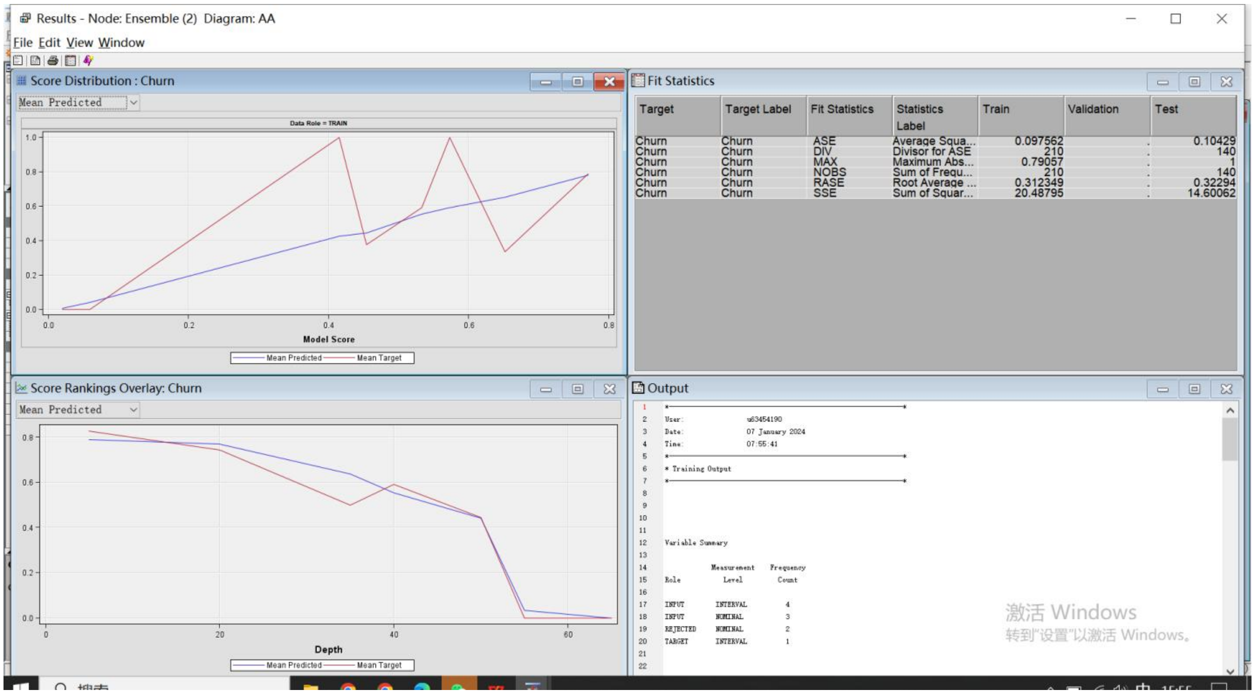
In summary, a comprehensive strategy involves personalized marketing, loyalty programs, regional targeting, spending optimization, and continuous adaptation. By integrating these recommendations, businesses can create a robust framework to reduce churn, enhance customer satisfaction, and foster long-term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



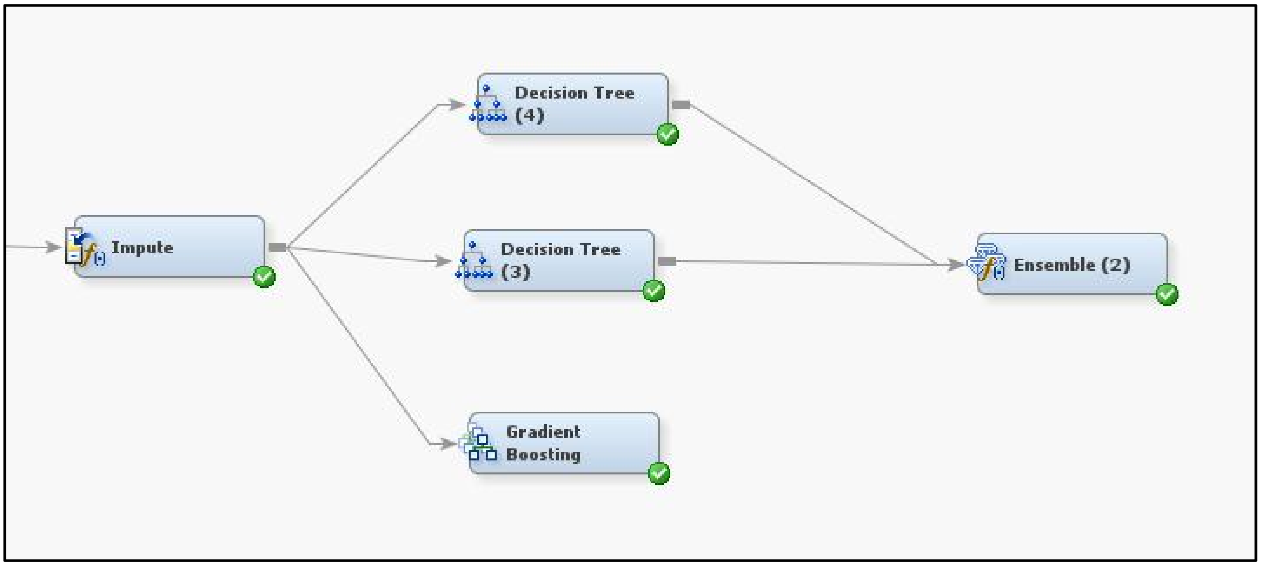
After running, the results are as follows.
The output presents results from an ensemble model combining two decision trees, TREE3 and TREE4. The model, targeting "Churn," demonstrates good accuracy with low Average Squared Error (ASE) and Root ASE values. Assessment scores at various depths and distributions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



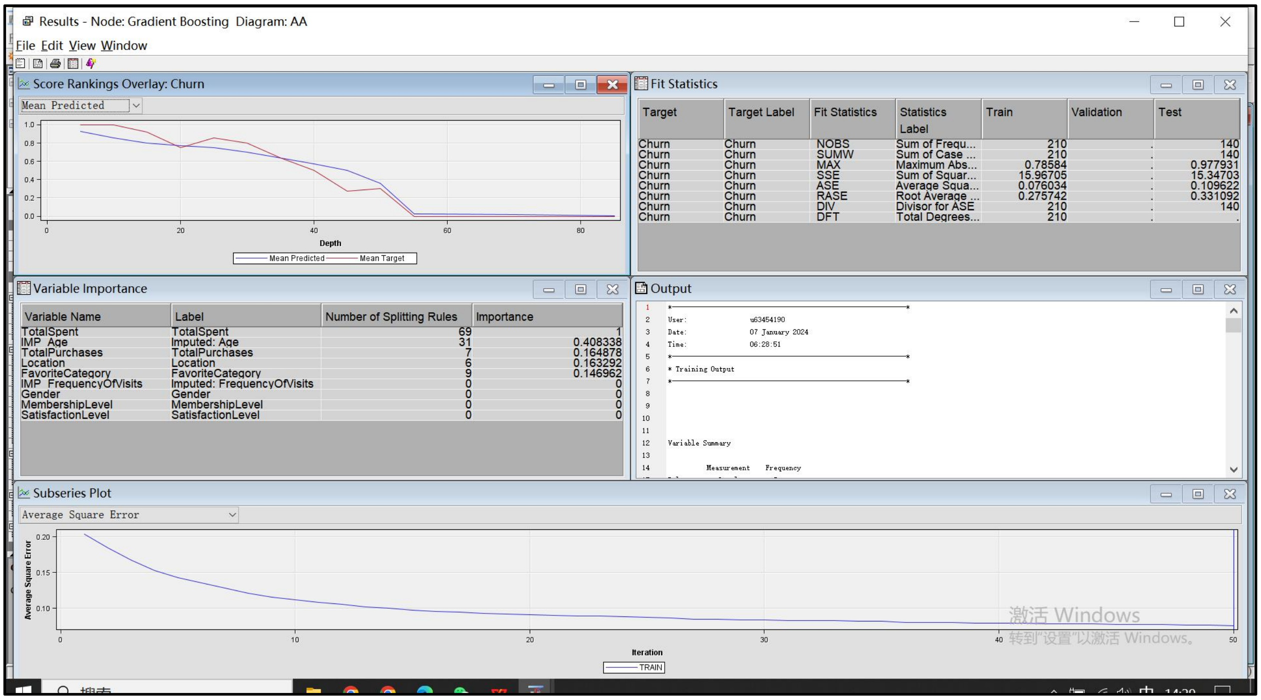
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average Squa...	0.097562	.	0.10429
Churn	Churn	DIV	Divisor for ASE	210	.	140
Churn	Churn	MAX	Maximum Abs...	0.79057	.	1
Churn	Churn	NOBS	Sum of Frequ...	210	.	140
Churn	Churn	RASE	Root Average ...	0.312349	.	0.32294
Churn	Churn	SSE	Sum of Squar...	20.48795	.	14.60062

```
*------------------------------------------------------------*
User:                   u63454190
Date:                   07 January 2024
Time:                   07:55:41
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

             Measurement    Frequency
Role            Level         Count

INPUT         INTERVAL          4
INPUT         NOMINAL           3
REJECTED      NOMINAL           2
TARGET        INTERVAL          1


```

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.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	NOBS	Sum of Frequ...	210	.	140
Churn	Churn	SUMW	Sum of Case ...	210	.	140
Churn	Churn	MAX	Maximum Abs...	0.78584	.	0.977931
Churn	Churn	SSE	Sum of Squar...	15.96705	.	15.34703
Churn	Churn	ASE	Average Squa...	0.076034	.	0.109622
Churn	Churn	RASE	Root Average ...	0.275742	.	0.331092
Churn	Churn	DIV	Divisor for ASE	210	.	140
Churn	Churn	DFT	Total Degrees...	210	.	.

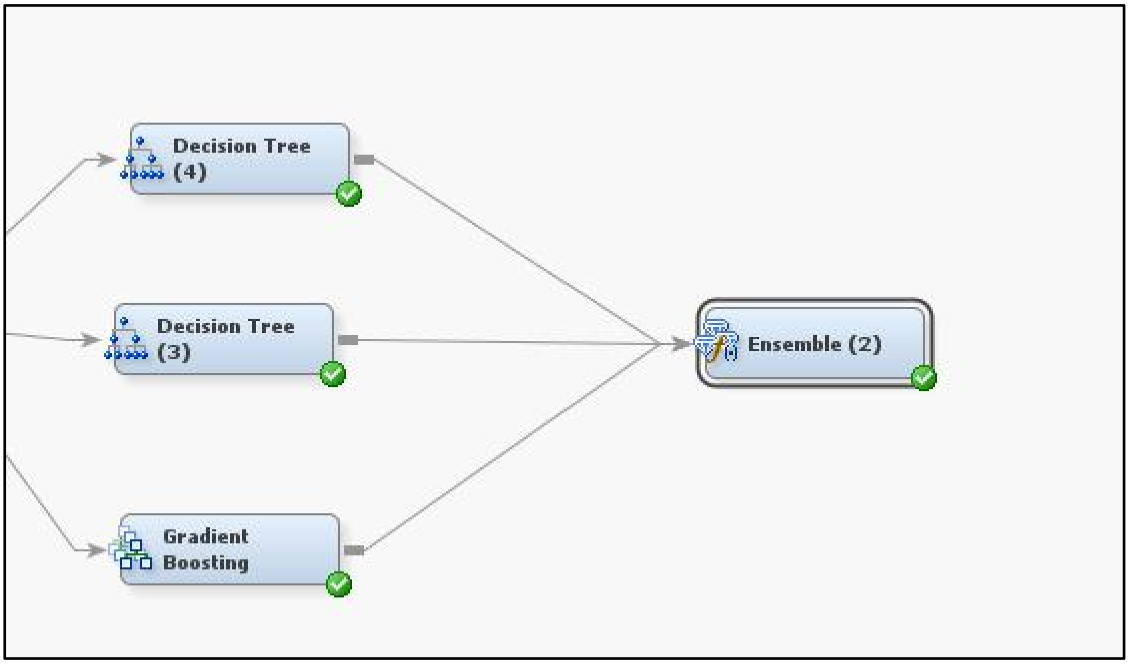
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
TotalSpent	TotalSpent	69	1
IMP_Age	Imputed: Age	31	0.408338
TotalPurchases	TotalPurchases	7	0.164878
Location	Location	6	0.163292
FavoriteCategory	FavoriteCategory	9	0.146962
IMP_FrequencyOfVisits	Imputed: FrequencyOfVisits	0	0
Gender	Gender	0	0
MembershipLevel	MembershipLevel	0	0
SatisfactionLevel	SatisfactionLevel	0	0

Output

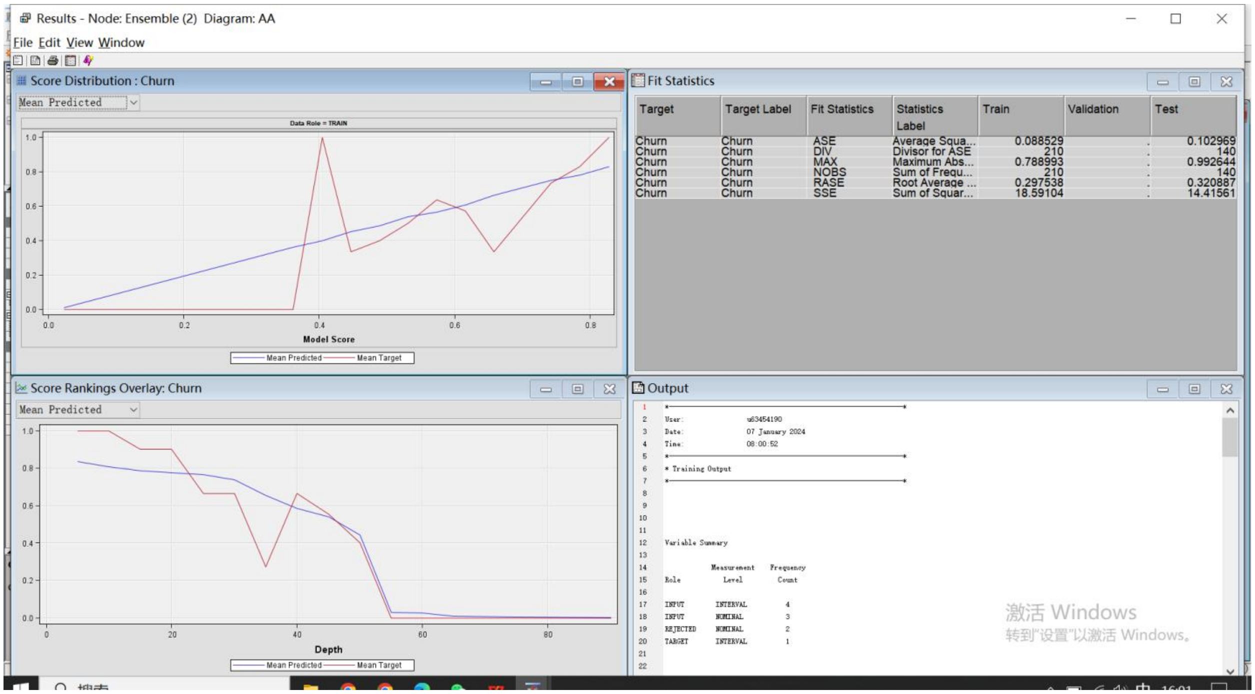
```
*------------------------------------------------------*
User:                 u63454190
Date:                 07 January 2024
Time:                 06:28:51
*------------------------------------------------------*
* Training Output
*------------------------------------------------------*




Variable Summary

         Measurement    Frequency
```

Connect Gradient Boosting to Ensemble and run



After running, the results are as follows.

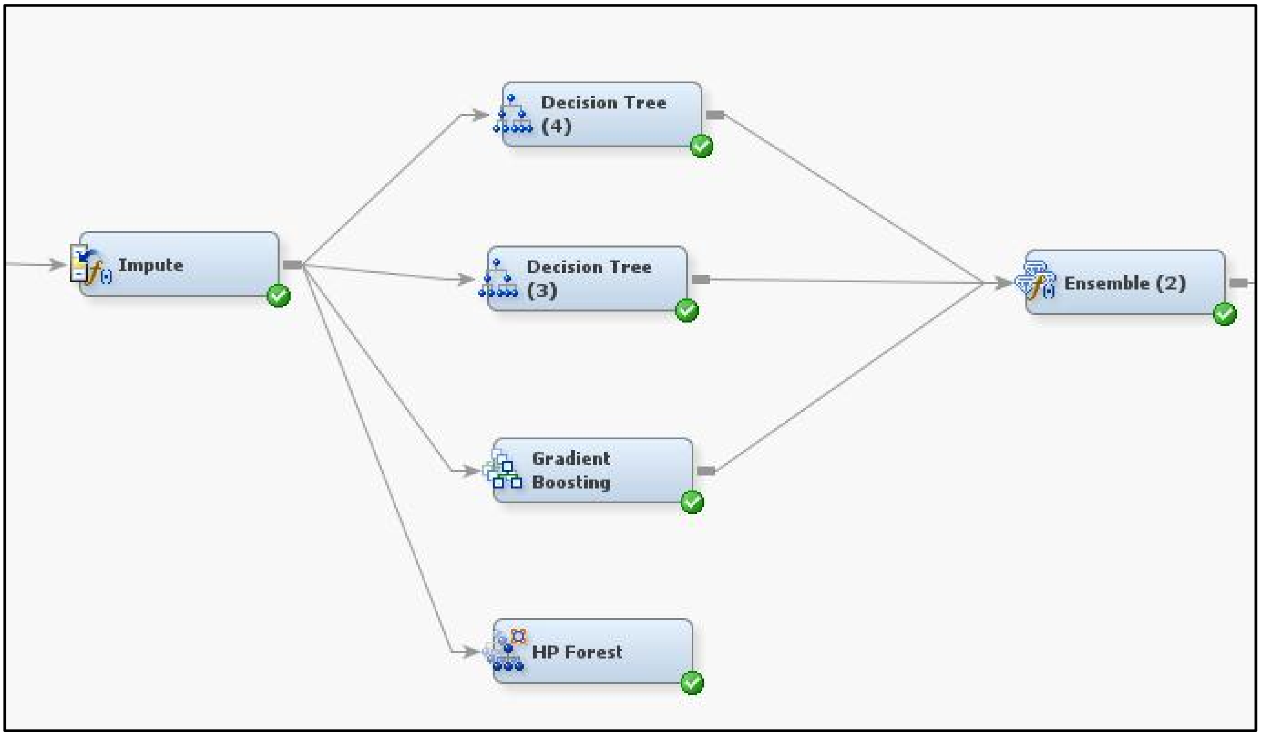
The ensemble model, combining TREE3, TREE4, and BOOST, predicts churn with an Average Squared Error of 0.089. It excels at depths 5 and 10, showing high predictability. The distribution analysis reveals strong performance in specific ranges (0.806 - 0.848, 0.764 - 0.806).



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average Squa...	0.088529	.	0.102969
Churn	Churn	DIV	Divisor for ASE	210	.	140
Churn	Churn	MAX	Maximum Abs...	0.788993	.	0.992644
Churn	Churn	NOBS	Sum of Frequ...	210	.	140
Churn	Churn	RASE	Root Average ...	0.297538	.	0.320887
Churn	Churn	SSE	Sum of Squar...	18.59104	.	14.41561

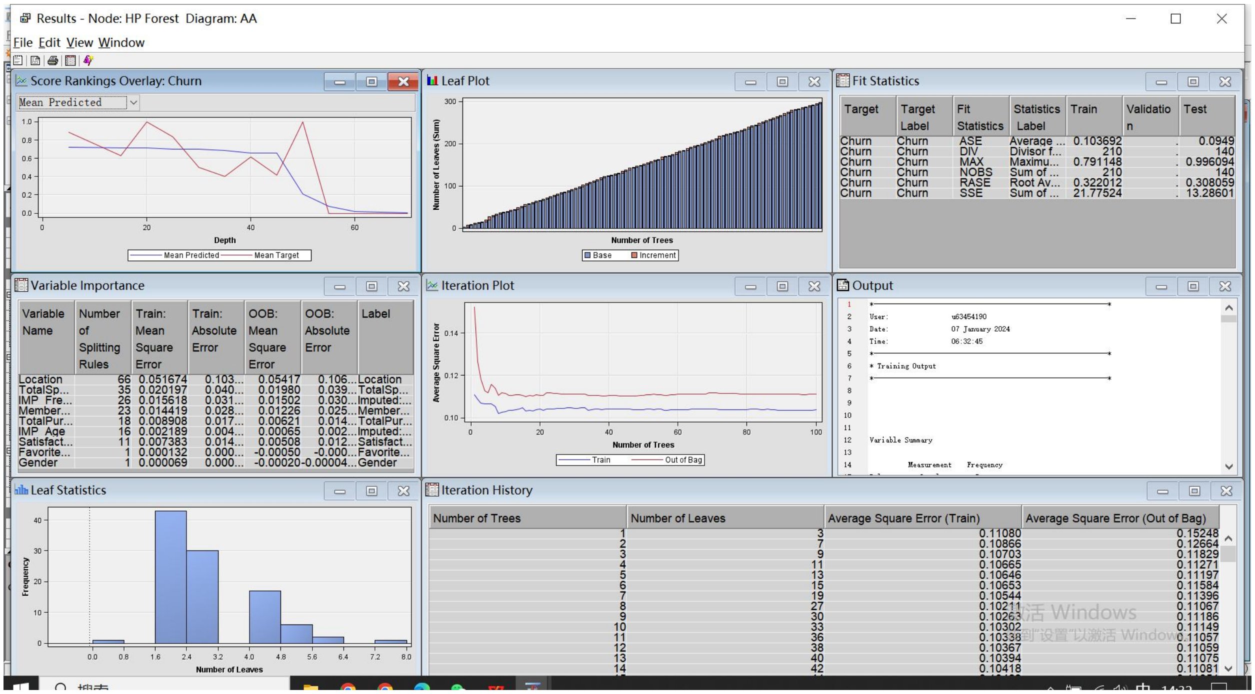
```
*------------------------------------------------------------*
User:                    u63454190
Date:                    07 January 2024
Time:                    08:00:52
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

              Measurement    Frequency
Role             Level         Count

INPUT          INTERVAL          4
INPUT          NOMINAL           3
REJECTED       NOMINAL           2
TARGET         INTERVAL          1


```

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."



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average ...	0.103692	.	0.0949
Churn	Churn	DIV	Divisor f...	210	.	140
Churn	Churn	MAX	Maximu...	0.791148	.	0.996094
Churn	Churn	NOBS	Sum of ...	210	.	140
Churn	Churn	RASE	Root Av...	0.322012	.	0.308059
Churn	Churn	SSE	Sum of ...	21.77524	.	13.28601

Variable Importance

Variable Name	Number of Splitting Rules	Train: Mean Square Error	Train: Absolute Error	OOB: Mean Square Error	OOB: Absolute Error	Label
Location	66	0.051674	0.103...	0.05417	0.106...	Location
TotalSp...	35	0.020197	0.040...	0.01980	0.039...	TotalSp...
IMP_Fre...	26	0.015618	0.031...	0.01502	0.030...	Imputed:...
Member...	23	0.014419	0.028...	0.01226	0.025...	Member...
TotalPur...	18	0.008908	0.017...	0.00621	0.014...	TotalPur...
IMP_Age	16	0.002189	0.004...	0.00065	0.002...	Imputed:...
Satisfact...	11	0.007383	0.014...	0.00508	0.012...	Satisfact...
Favorite...	1	0.000132	0.000...	-0.00050	-0.000...	Favorite...
Gender	1	0.000069	0.000...	-0.00020	-0.00004...	Gender

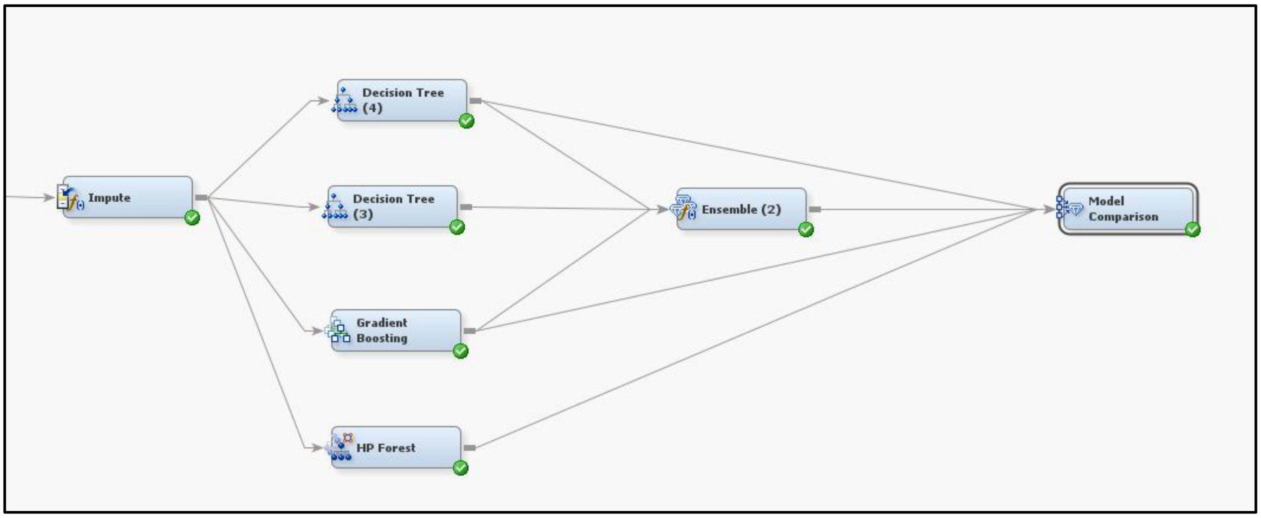
Output

```
*--------------------------------------------------*
User:                u63454190
Date:                07 January 2024
Time:                06:32:45
*--------------------------------------------------*
* Training Output
*--------------------------------------------------*




Variable Summary

          Measurement    Frequency
```

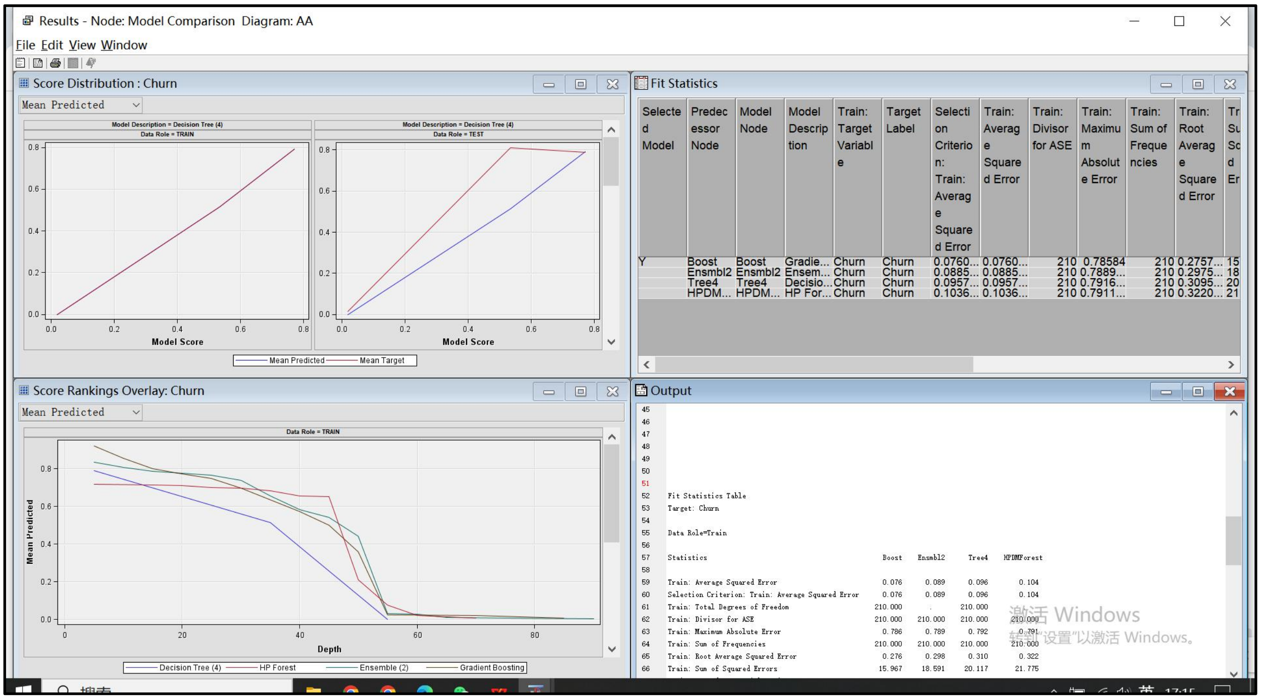
Iteration History

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)
1	3	0.11080	0.15248
2	7	0.10866	0.12664
3	9	0.10703	0.11829
4	11	0.10665	0.11271
5	13	0.10646	0.11197
6	15	0.10653	0.11584
7	19	0.10544	0.11396
8	27	0.10211	0.11067
9	30	0.10263	0.11186
10	33	0.10302	0.11149
11	36	0.10338	0.11057
12	38	0.10367	0.11059
13	40	0.10394	0.11075
14	42	0.10418	0.11081

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.



Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Train: Average Squared Error	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Tr Su Sq d Er
Y	Boost	Boost	Gradie...	Churn	Churn	0.0760...	0.0760...	210	0.78584	210	0.2757...	15
	Ensmbl2	Ensmbl2	Ensem...	Churn	Churn	0.0885...	0.0885...	210	0.7889...	210	0.2975...	18
	Tree4	Tree4	Decisio...	Churn	Churn	0.0957...	0.0957...	210	0.7916...	210	0.3095...	20
	HPDM...	HPDM...	HP For...	Churn	Churn	0.1036...	0.1036...	210	0.7911...	210	0.3220...	21

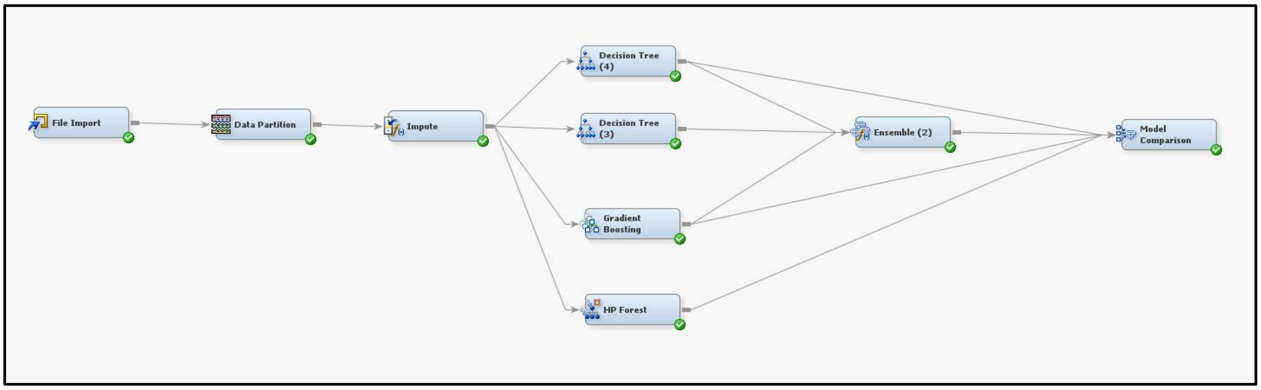
```
Fit Statistics Table
Target: Churn

Data Role=Train

Statistics                                      Boost    Ensmbl2      Tree4   HPDMForest

Train: Average Squared Error                    0.076      0.089      0.096        0.104
Selection Criterion: Train: Average Squared Error  0.076   0.089      0.096        0.104
Train: Total Degrees of Freedom               210.000          .    210.000
Train: Divisor for ASE                        210.000    210.000    210.000      210.000
Train: Maximum Absolute Error                   0.786      0.789      0.792        0.791
Train: Sum of Frequencies                     210.000    210.000    210.000      210.000
Train: Root Average Squared Error               0.276      0.298      0.310        0.322
Train: Sum of Squared Errors                   15.967     18.591     20.117       21.775
```

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 (HPDMForest) 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.