Alternative Assessment 1

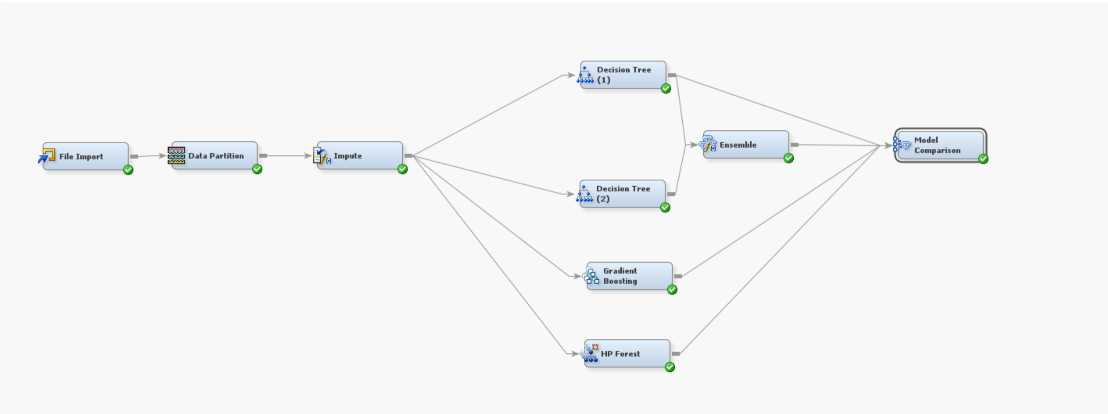
Name: Wang Ruobing

Matric No: S2163476

Github link: https://github.com/WRB-bing/WQD7005-AA1

Learning customer behaviors is crucial in e-commerce field, especially for keeping business growth and foresting customer loyalty. The dataset used in this case study is a combined dataset. Different columns are extracted from the following two datasets, https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset and https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset. This dataset includes Customer ID, Age, Gender, City, Membership Type, Items Purchased (the total number of purchased items), Total Spend, Item Purchased, Category, Last Purchase Date, Satisfaction Level, Average Rating, Subscription Status (subscribe the website or not), and the target variable, Churn.

Here is the workflow in SAS Enterprise Miner. Detail process will be shown in the following chapters.



Data import and preprocessing:

First, import the data into SAS Enterprise Miner and run it. The result shown in below.

Output

```
Label
Data Representation  SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64
Encoding             utf-8  Unicode (UTF-8)


                              Engine/Host Dependent Information

Data Set Page Size          131072
Number of Data Set Pages    1
First Data Page             1
Max Obs per Page            1090
Obs in First Data Page      352
Number of Data Set Repairs  0
Filename                    /home/u63721869/WQD7005 AA1/Workspaces/EMWS1/fimport_data.sas7bdat
Release Created             9.0401M7
Host Created                Linux
Inode Number                80217155
Access Permission           rw-r--r--
Owner Name                  u63721869
File Size                   256KB
File Size (bytes)           262144


                Alphabetic List of Variables and Attributes

 #    Variable               Type    Len    Format    Informat    Label

 1    Age                    Num       8    BEST.                 Age
13    Average_Rating         Num       8
 4    Category               Char     11    $11.      $11.        Category
 6    Churn                  Num       8    BEST.                 Churn
 3    City                   Char     13    $13.      $13.        City
 7    Customer_ID            Num       8
 5    Date                   Char     10    $10.      $10.        Date
 2    Gender                 Char      6    $6.       $6.         Gender
11    Item_Purchased         Char     10
 9    Items_Purchased        Num       8
 8    Membership_Type        Char      6
12    Satisfaction_Level     Char     11
14    Subscription_Status    Char      3
10    Total_Spend            Num       8
```

Then set the input and target value. As shown below.

Variables - FIMPORT

(none) not Equal to

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Average_Ra	Input	Interval	No		No	.	.
Category	Input	Nominal	No		No	.	.
Churn	Target	Interval	No		No	.	.
City	Input	Nominal	No		No	.	.
Customer_I	Input	Interval	No		No	.	.
Date	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Item_Purch	Input	Nominal	No		No	.	.
Items_Purc	Input	Interval	No		No	.	.
Membershi	Input	Nominal	No		No	.	.
Satisfacti	Input	Nominal	No		No	.	.
Subscripti	Input	Nominal	No		No	.	.
Total_Sper	Input	Interval	No		No	.	.

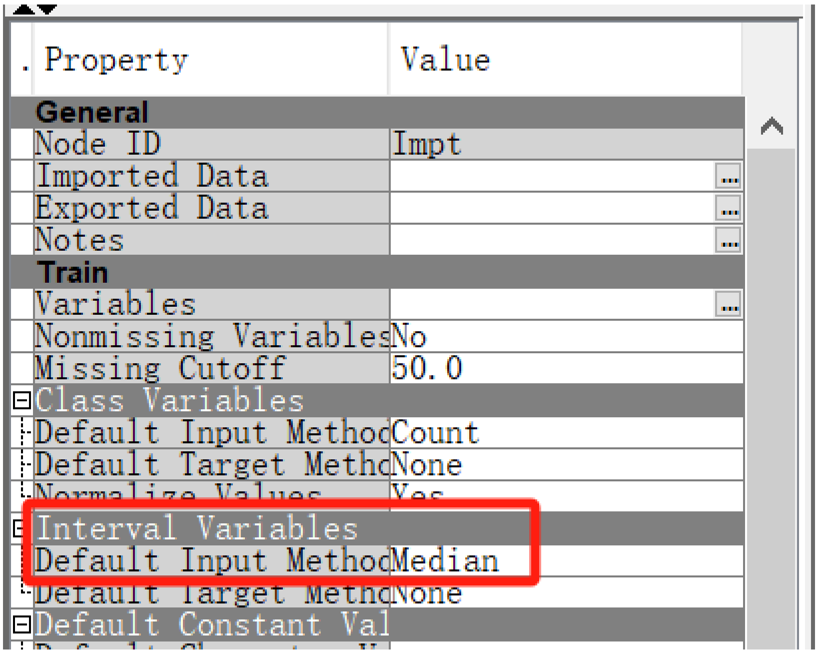
Decide the train data and test data. Divide data into train set (60%) and test set (40%).

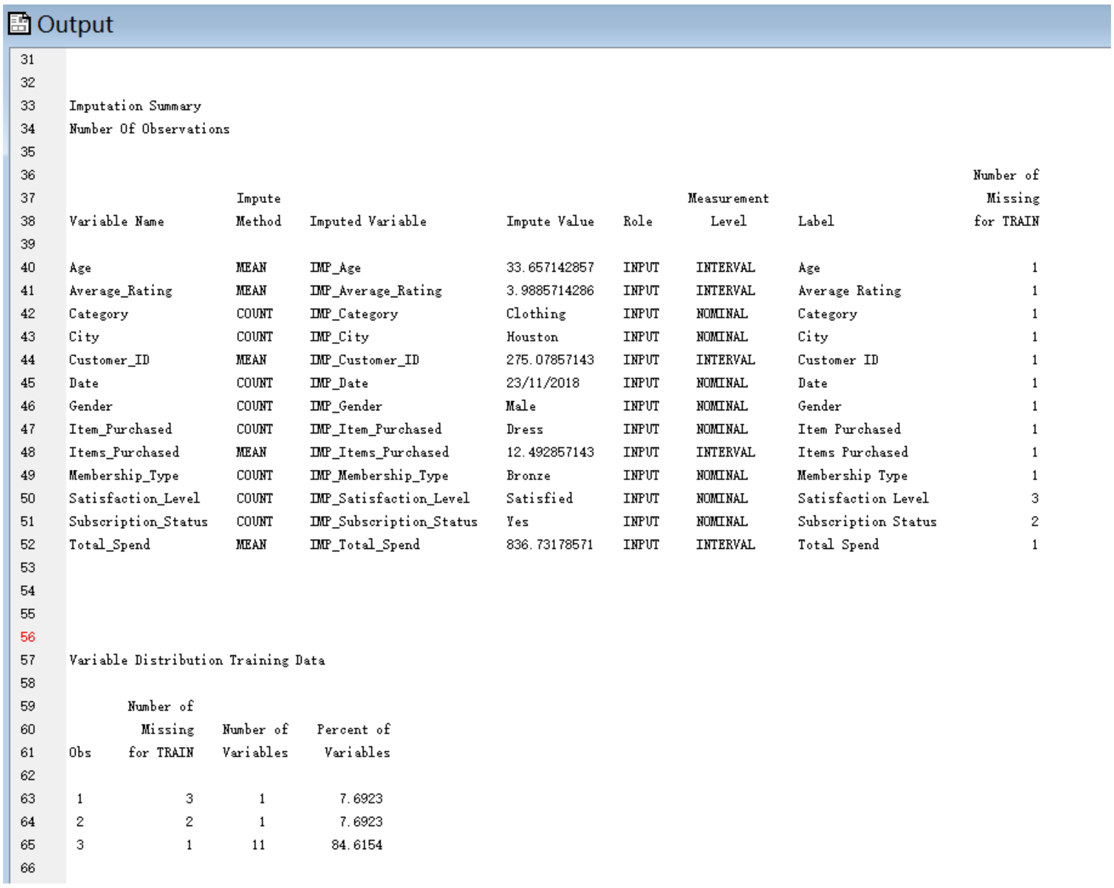
Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Met	Default
Random Seed	12345
Data Set Allocat	
Training	60.0
Validation	40.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/7/24 4:28 AM
Run ID	d14cfffe-7836-2
Last Error	
Last Status	Complete
Last Run Time	1/7/24 5:00 AM
Run Duration	0 Hr. 0 Min. 2.
Grid Host	
User-Added Node	No

Output

```
                                    Number of
Type          Data Set            Observations

DATA     EMWS1.FIMPORT_train          352
TRAIN    EMWS1.Part_TRAIN             211
TEST     EMWS1.Part_TEST              141


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


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




Summary Statistics for Interval Targets

Data=DATA

                                                 Number of                  Standard
Variable   Maximum        Mean       Minimum   Observations   Missing      Deviation     Label

 Churn        1      0.3333333333       0          348            4       0.4720832894   Churn


Data=TEST

                                                 Number of                  Standard
Variable   Maximum        Mean       Minimum   Observations   Missing      Deviation     Label

 Churn        1      0.3285714286       0          140            1       0.4713802959   Churn


Data=TRAIN

                                                 Number of                  Standard
Variable   Maximum        Mean       Minimum   Observations   Missing      Deviation     Label

 Churn        1      0.3365384615       0          208            3       0.4736654667   Churn
```

Next, handling with missing values. The result is as shown. And input the empty cell with its median.





Output

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEAN	IMP_Age	33.657142857	INPUT	INTERVAL	Age	1
Average_Rating	MEAN	IMP_Average_Rating	3.9885714286	INPUT	INTERVAL	Average Rating	1
Category	COUNT	IMP_Category	Clothing	INPUT	NOMINAL	Category	1
City	COUNT	IMP_City	Houston	INPUT	NOMINAL	City	1
Customer_ID	MEAN	IMP_Customer_ID	275.07857143	INPUT	INTERVAL	Customer ID	1
Date	COUNT	IMP_Date	23/11/2018	INPUT	NOMINAL	Date	1
Gender	COUNT	IMP_Gender	Male	INPUT	NOMINAL	Gender	1
Item_Purchased	COUNT	IMP_Item_Purchased	Dress	INPUT	NOMINAL	Item Purchased	1
Items_Purchased	MEAN	IMP_Items_Purchased	12.492857143	INPUT	INTERVAL	Items Purchased	1
Membership_Type	COUNT	IMP_Membership_Type	Bronze	INPUT	NOMINAL	Membership Type	1
Satisfaction_Level	COUNT	IMP_Satisfaction_Level	Satisfied	INPUT	NOMINAL	Satisfaction Level	3
Subscription_Status	COUNT	IMP_Subscription_Status	Yes	INPUT	NOMINAL	Subscription Status	2
Total_Spend	MEAN	IMP_Total_Spend	836.73178571	INPUT	INTERVAL	Total Spend	1

Variable Distribution Training Data

Obs	Number of Missing for TRAIN	Number of Variables	Percent of Variables
1	3	1	7.6923
2	2	1	7.6923
3	1	11	84.6154

Decision tree analysis:

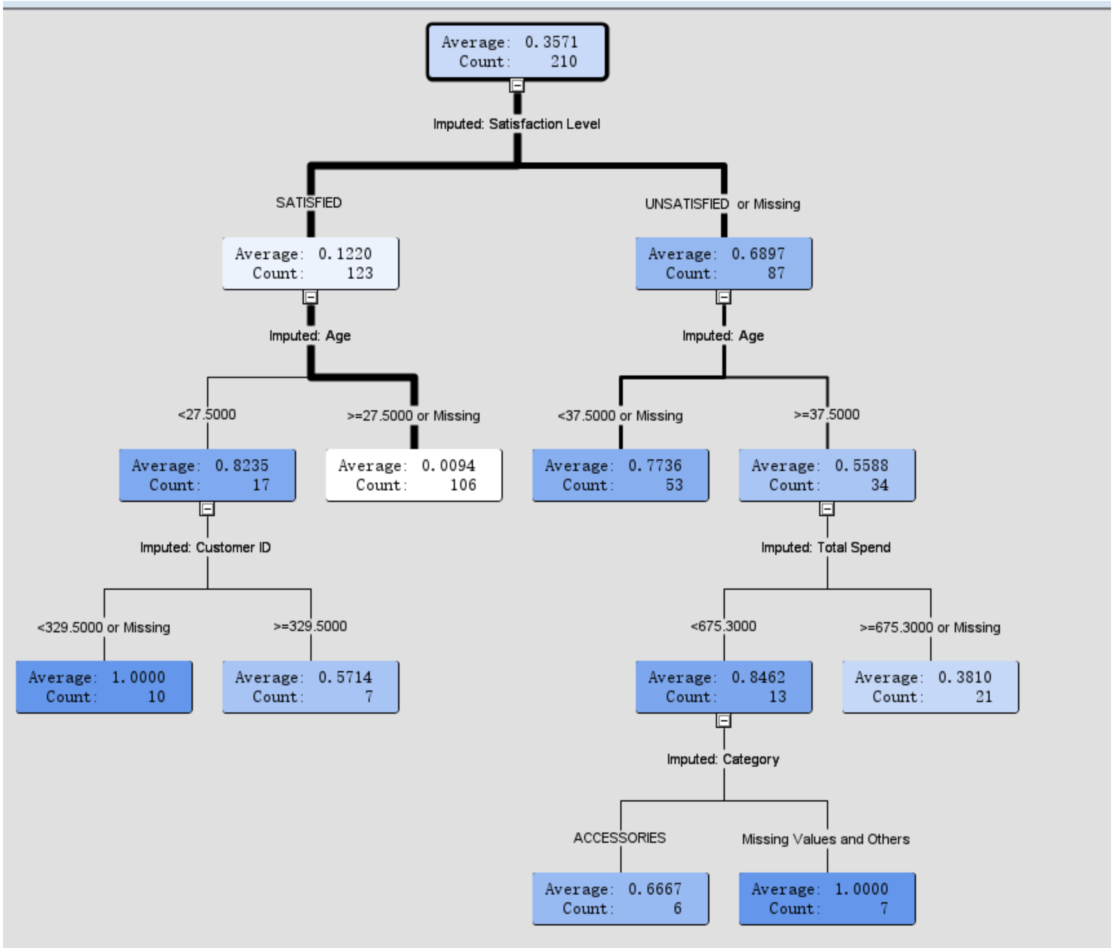
Create a Decision Tree node and use the interactive decision tree.

Here in the graph, it is shown clearly that Decision Tree 1 is divided basically depend on the satisfaction level, so it is the root node. Take it as the beginning cause customers satisfaction level will heavily influence the churn result. Data is divided into two groups, satisfied and unsatisfied, all missing values have been handled before.

The second level split is dependent on Age. This may have an impact on customers shopping satisfaction. For example, because of different expectations and service interaction experiences, younger and older customers may have different reviews and feedback, all of which may affect the customer retention.

The third level split is dependent on Customer ID and Total Spend for different groups. This is because for different group of customers, they may focus on different things. The customer ID is automatically generated by the computer and assigned based on the user's registration time, which means that the earlier registration, the smaller the ID number (e.g. 101 is registered earlier than 110), and the longer the customer has been shopping online. And the total spend money is also influence customers satisfaction. Customers spend more are expecting a better service and shopping experience which will affect their churn too.

The terminal nodes, also known as leaves, are split based on category. Whether or not there is a full range of products will affect the retention of customers.

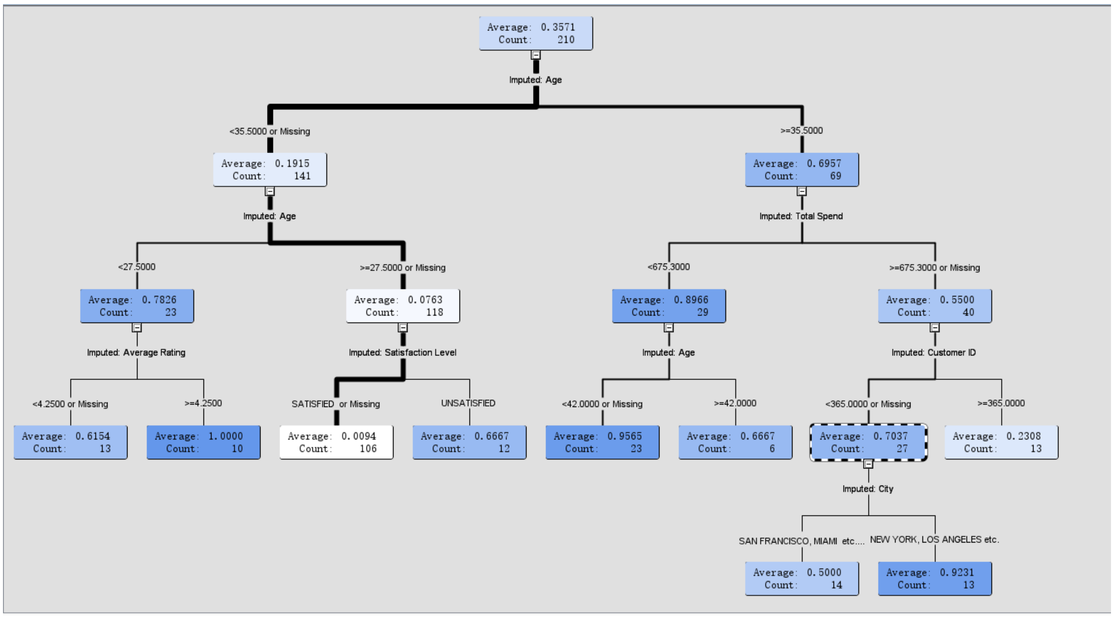


The graph below shows how Decision Tree 2 is split. Root node is Age, 35.5 is the mean. Therefore, it is divided into older than 35.5 and younger than 35.5. Customers belong to different age groups have different opinions on satisfaction factors.

The second split is also age for age under 35.5 group and total spend for those over 35.5. using age again may be due to customers have different performance in terms of satisfaction. As for total spending, its thresholds set at 765.3. It indicates that the amount of money customers spend is a strong indicator of their satisfaction.

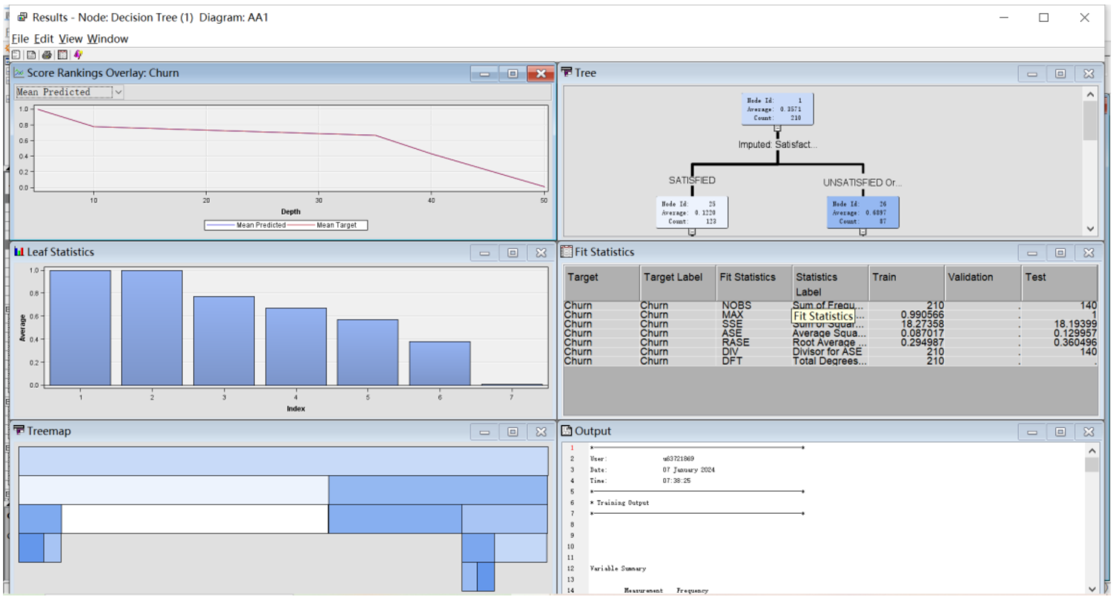
The third level split is about Average Rating, Satisfaction Level, Age, and Customer ID. These are directly related to our target variable.

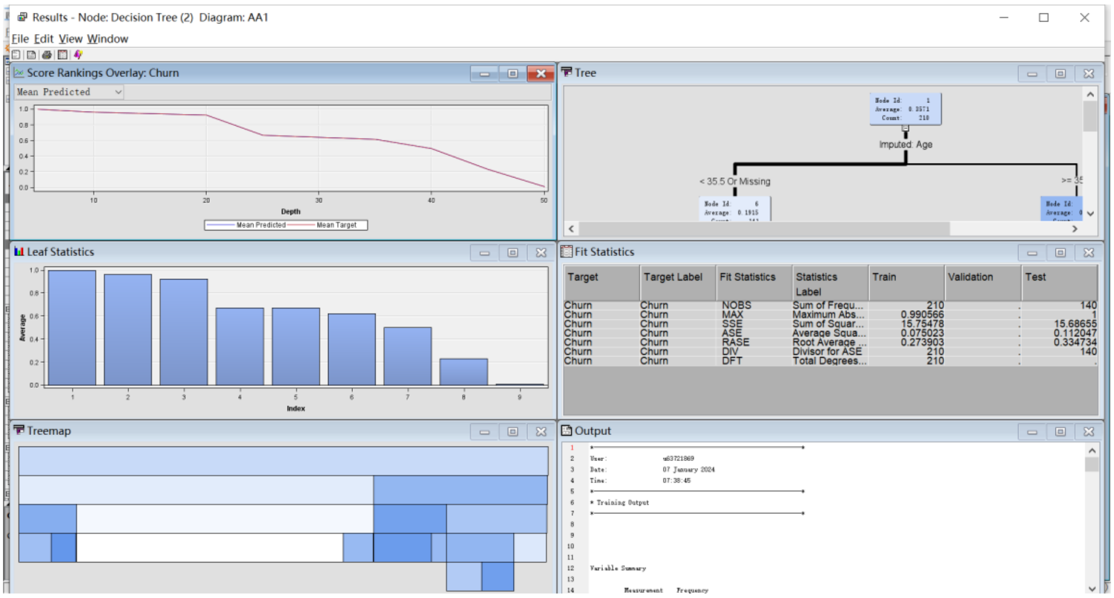
The fourth level (leaves) is based on City. Customers living in different locations may have different assessments for satisfaction. Large cities may have a more mature delivery chain, a wider range of item categories, better customer services, etc. This may influence customers criteria for satisfaction and further decide their retention.



Run the Decision Tree Node, and its result shown below.

The result shows its performance in different depth. Mean Prediction predicts the probability of churn for all node in the given depth. Mean target represents the actual percentage of customers churned. The leaf statistics shows the average churn for each leaf. Leaves with high values represent the segments of customers with a high risk of churn. Fit statistic evaluates its performance.

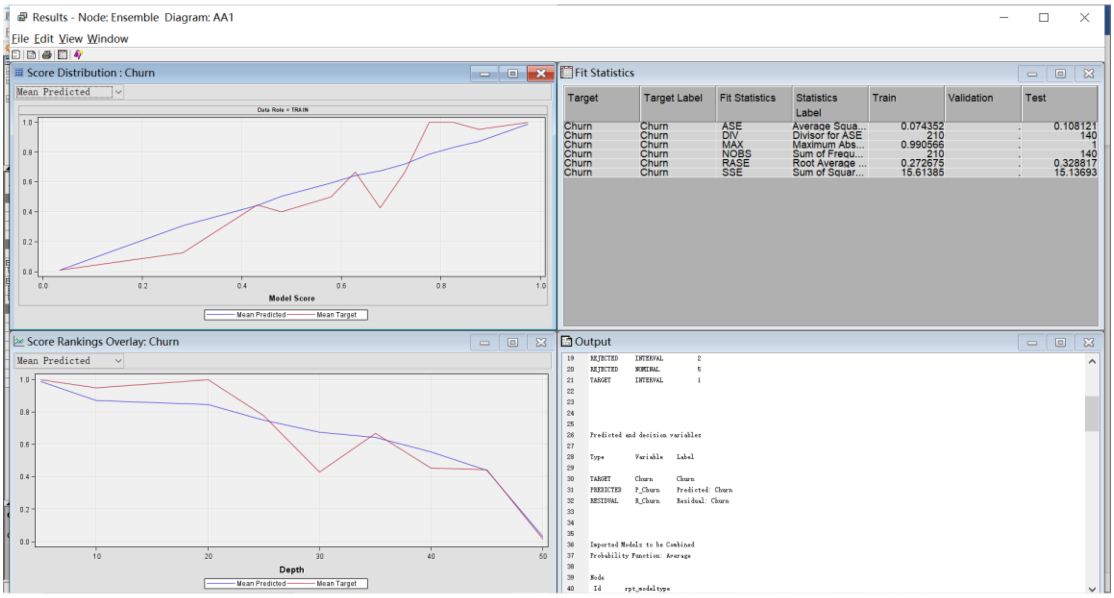




Ensemble methods:

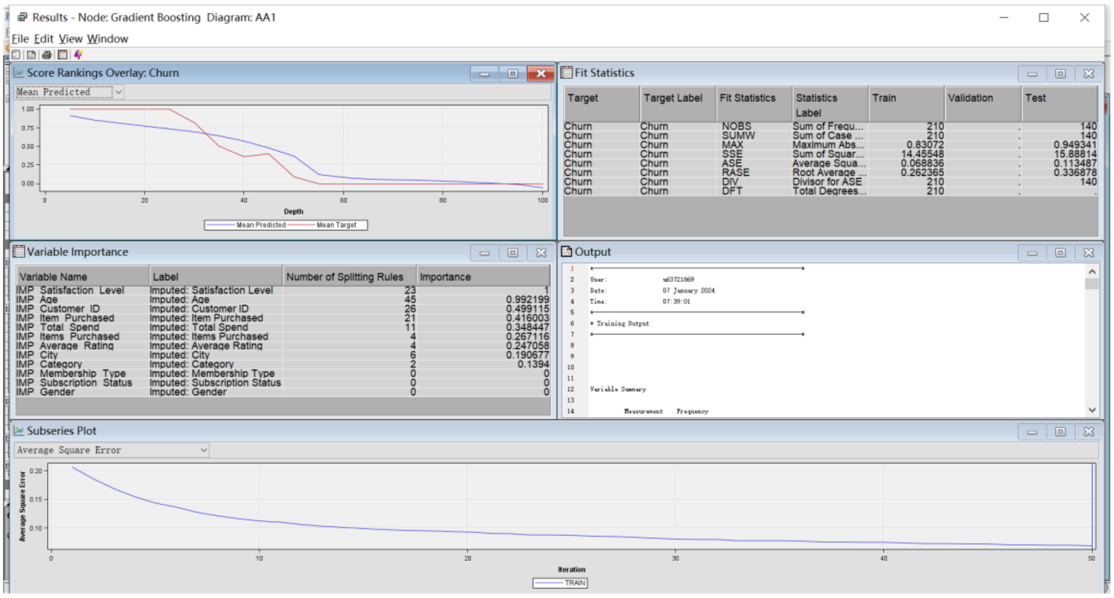
Link the Ensemble node after the Decision Tree, set its variables and target, then just run it to get the result.

In score distribution, the blue line represents mean predicted and the red one represents the mean target. Ideally, the blue line should align closely with the red line, indicating accurate predictions across all thresholds. Although they are not aligned closely, the final result seems to be the same among target and prediction. In score ranking overlay, it shows the comparison between predicted churn rate and target churn rate. As the complexity of the model increases, the extent of match between the combined predictions derived by the model ensemble's composite prediction from multiple trees matched the actual churn rate. The ASE values for training (0.074352) and test (0.108121) indicate that ensemble model performs well in predicting customer churn. The ASE value for test is relatively high may because these data are new.

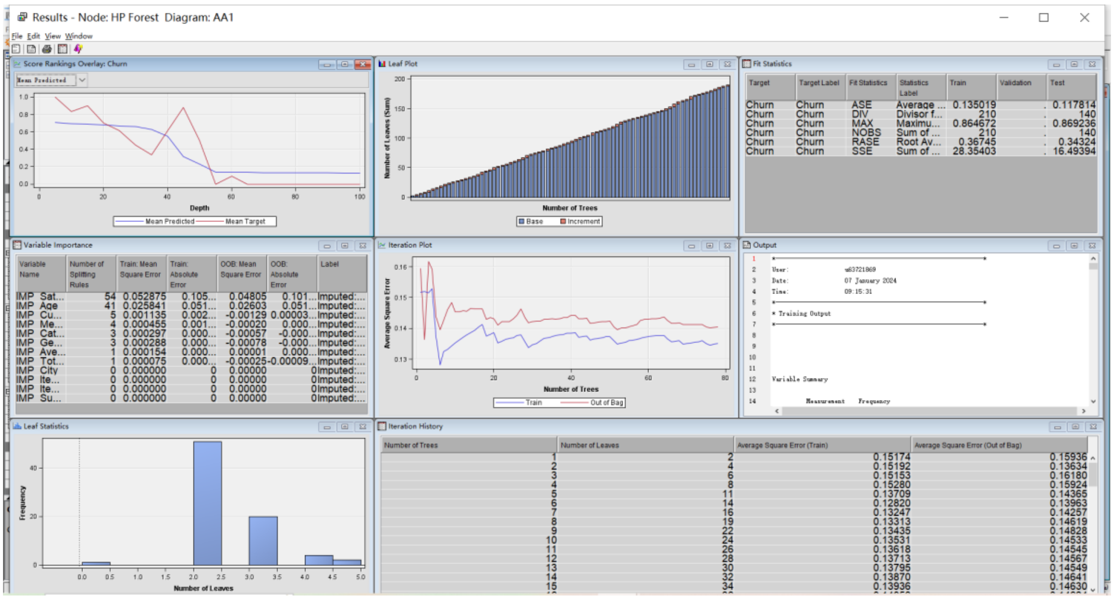


Link the Gradient Boosting node, edit its variables and run it.

In score ranking overlay, it shows that the increasing number of depth, the more accurate the prediction will be made. Ideally, mean predict line will be aligned close to mean target line as depth increasing. If the average predict line diverges or does not converge sufficiently with the average target, it may indicate overfitting or lack of model generalization. Variable importance indicates that Satisfaction Level is the most important one, followed by Age, Customer ID, etc., and the Gender is the most unimportant one. The ASE value for test set (0.113487) is higher than that for train set (0.068836) is common. Subseries plots show that it is a downward trend, which means that it is improving with each iteration.



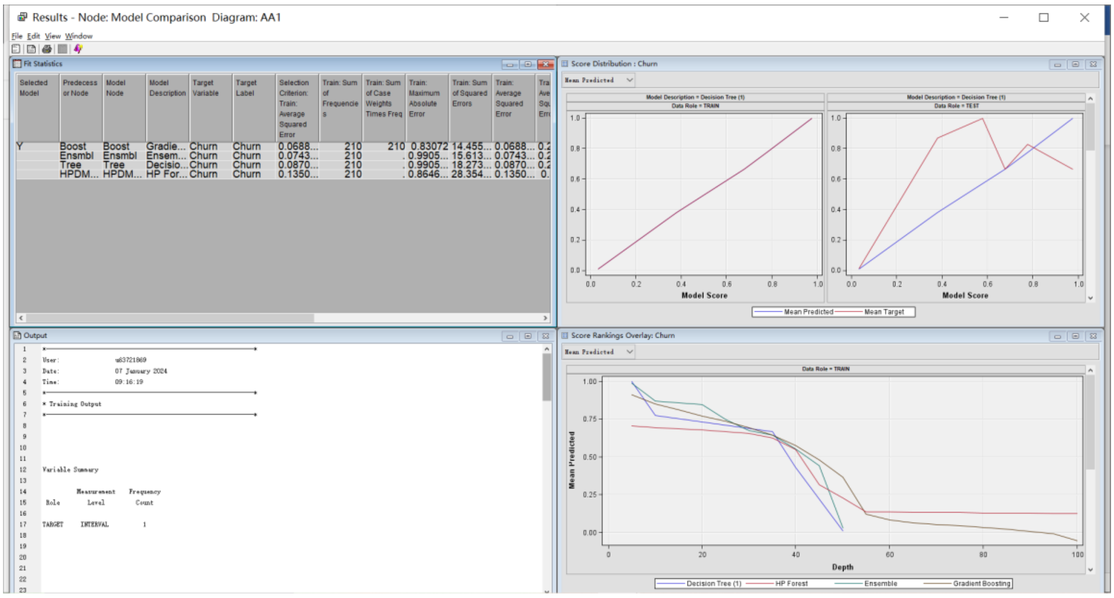
Create a HP Forest node, set its variables and run it to get the result.



Optimally, the predicted values should converge to the actual values as the depth increases. But in graph above it is not, so it may face the problem of overfitting or insufficient prediction. In the variable importance part, it tells that Satisfaction Level is the most important variable while predicting churn. As for leaf plot, with the increase of trees, more leaves will be created since each new tree will create at least two new leaves. Iteration Plot shows that as the tree increases, the error decreases and becomes stable. It is normal and ideal.

Last, create a Model Comparison node, link, and run it. Its result shows below.

According to score ranking overlay, the HP Forest and Gradient Boosting models perform better than individual decision tree. The fitting statistics also suggest that Gradient Boosting may be the better model, especially for the training set, due to the lowest ASE. In score distribution, two graphs, one for train set, one for test set. The blue line represents the predict churn and the red one represents the real churn. Ideally, the model would overlap these two lines, indicating that the predictions are consistent with the actual results. It is clear from the fit statistics table that the Gradient Boosting model has the lowest training: Average Squared Error, which may indicate that it is the most accurate of the training data comparison models.



Conclusion:

In this study, variables and models are the most importance things. The important variables such as satisfaction level and ages suggest that these variables are highly predictive of customer churn. This indicates that the younger or older age groups and their satisfaction are crucial in determining the likelihood of a customer's leaving. Models predict churn based on historical data may make some suggestion for future business. For example, the influence of service and price for customer churn, when is the high period of churn, and identify the high-risk customer groups.

For business strategies, the first thing needs to be considered is to improve customer satisfaction, such as improving customer service, personalizing interactions, and effectively handling feedback. As it shown in the models, age is one of the factors affecting churn. Companies need to develop retention strategies that target the age groups most likely to churn, implement loyalty programs or targeted promotions to attract these specific groups and increase their loyalty. Besides, companies can use the model to understand customer preferences and customize services or products to better meet those needs, which can reduce the likelihood of churn. Moreover, companies can reduce the risk of churn by maintaining open lines of communication with customers, especially around changes that may affect their perceptions and satisfaction.