WQD7005

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

GitHub: https://github.com/GuanMengg/7005_casestudy

ALTERNATIVE ASSESSMENT 1 (50 marks) - WEEK 12

Answer the question below based on the given scenario. Submit your answer within ONE (1) DAY after the question is given in SPECTRUM. Answers should be submitted and saved with the student's name followed by matric number as the file name in the format of .pdf (e.g.Ali_s123456.pdf).

Case Study: E-Commerce Customer Behavior Analysis

Background:

You will work with a dataset of customer transactions from an e-commerce website, encompassing various customer attributes and purchase history over the last year. The structure provided below is a guideline. Feel free to enhance this dataset by adding relevant attributes that you believe will enrich your analysis. Use the structure as a foundation to create your own sample dataset that reflects realistic customer behavior.

Dataset Structure:

CustomerID: Unique identifier for each customer.

Age: Age of the customer.

Gender: Gender of the customer.

Location: Geographic location of the customer.

MembershipLevel: Indicates the membership level (e.g., Bronze, Silver, Gold, Platinum).

TotalPurchases: Total number of purchases made by the customer.

TotalSpent: Total amount spent by the customer.

FavoriteCategory: The category in which the customer most frequently shops (e.g., Electronics, Clothing, Home Goods).

LastPurchaseDate: The date of the last purchase.

[Additional Attributes]: Consider adding more attributes like customer's occupation, frequency of website visits, etc.

Churn: Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

Tasks

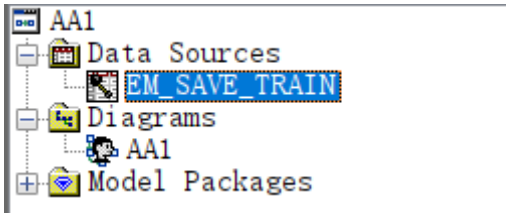
Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

[15 marks]

Project objectives:

- Determine which variables (such as age, purchase frequency, consumption level, etc.) are significantly related to customer satisfaction.
- Analyze the specific impact of these key factors on satisfaction, that is, how and to what extent they affect customer satisfaction.
- Divide customers into different groups based on key influencing factors to facilitate targeted analysis and strategy development.
- Develop differentiated marketing and service improvement strategies based on the factors that influence the satisfaction of different customer groups.

Step1: import dataset into SAS Enterprise Miner:

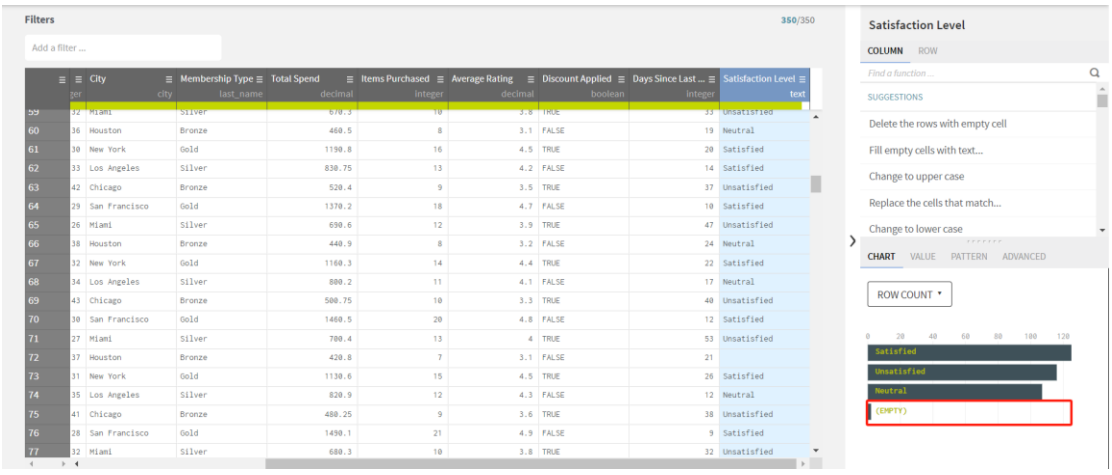


Step2: Use the StatExplore node to find missing values in dataset:

Step3: Run the StatExplore node and find that the satisfaction_level has two missing values:

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Age	INPUT	16	0	30	13.71	32	9.43
TRAIN	City	INPUT	6	0	Los Angeles	16.86	New York	16.86
TRAIN	Discount_Applied	INPUT	2	0	FALSE	50.00	TRUE	50.00
TRAIN	Gender	INPUT	2	0	Female	50.00	Male	50.00
TRAIN	Items_Purchased	INPUT	15	0	10	13.43	9	9.71
TRAIN	Membership_Type	INPUT	3	0	Gold	33.43	Silver	33.43
TRAIN	Satisfaction_Level	TARGET	4	2	Satisfied	35.71	Unsatisfied	33.14

In addition, I also imported the original data set into Talend Data Preparation, and it also showed that there were two missing values in satisfaction_level.



Count: **350**

Avg length: **9**

Distinct: **4**

Duplicate: **346**

Min length: **0**

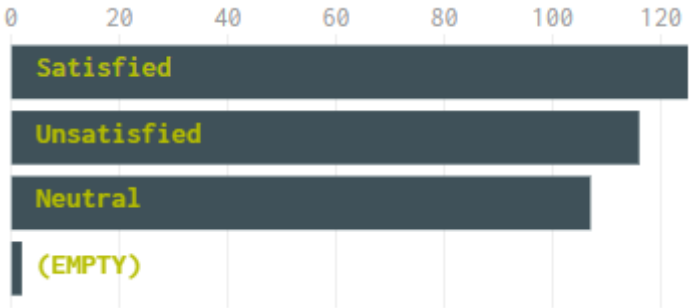
Valid: **348**

Empty: **2**

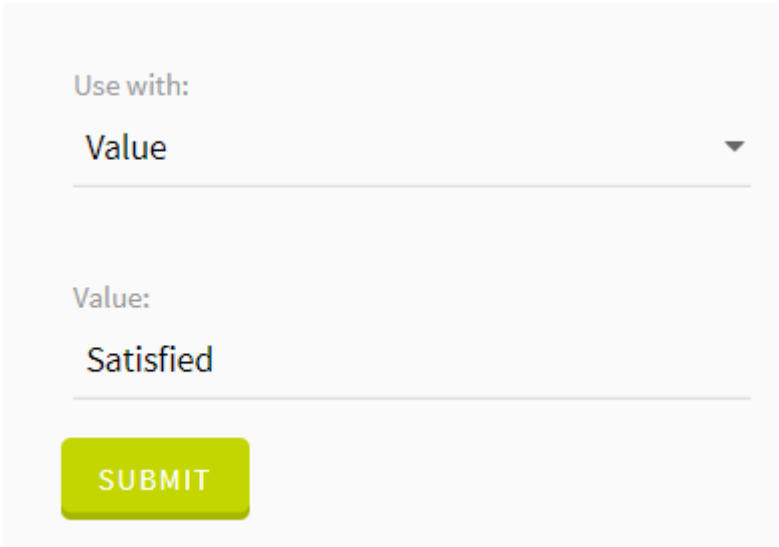
Max length: **11**

Invalid: **0**

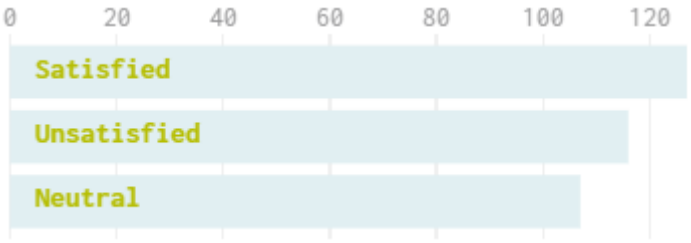
Since the most common category is satisfaction, the two missing values in the satisfaction_level are filled as "Satisfied".

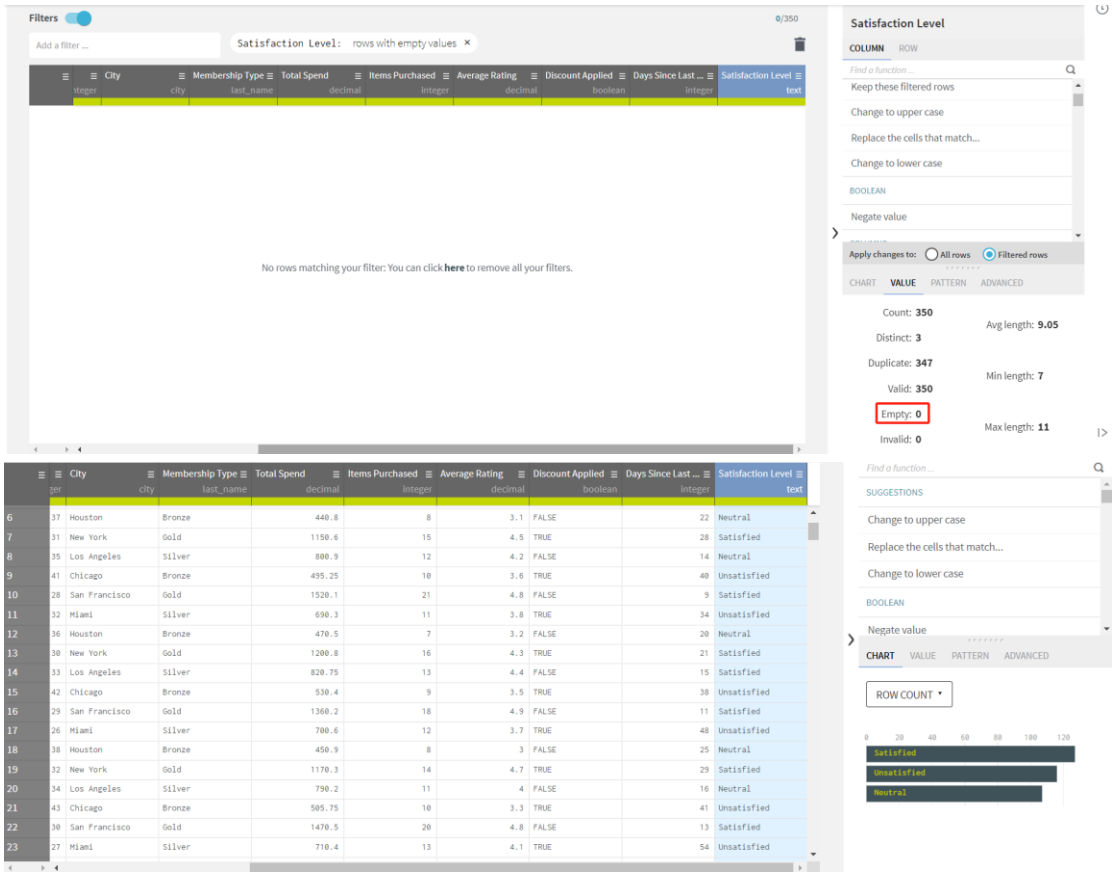


		City	Membership Type	Total Spend	Items Purchased	Average Rating	Discount Applied	Days Since Last ...	Satisfaction Level
	integer	city	last_name	decimal	integer	decimal	boolean	integer	text
72	37	Houston	Bronze	420.8	7	3.1	FALSE	21	
144	37	Houston	Bronze	430.8	7	3.4	FALSE	23	



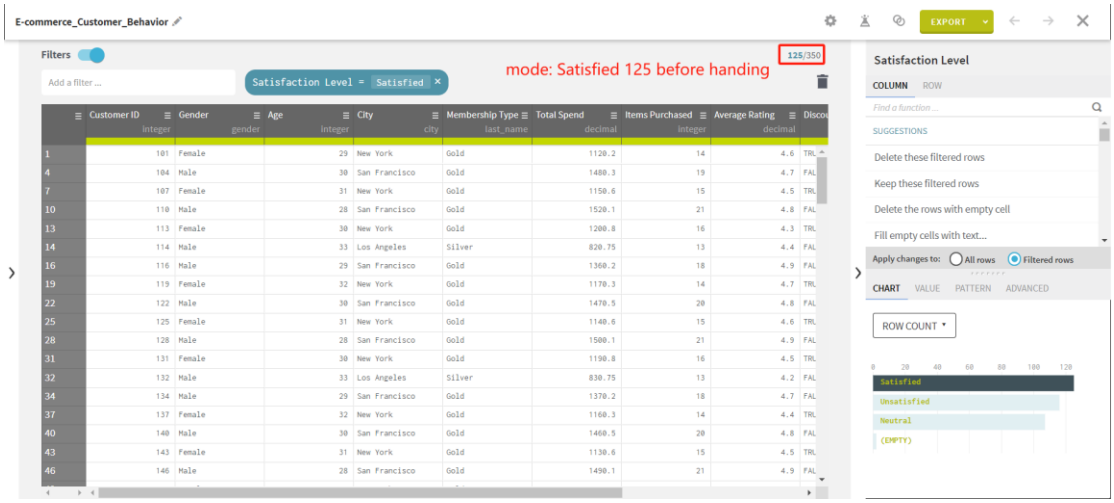
After filling in the missing values, empty no longer exists.





Missing value handling in SAS Miner:

The number of occurrences of Satisfied before missing value processing is 125.



Add impute node:

Under the class variable under the impute node, the default target method is set to count.

Class Variables	
Default Input Method	Count
Default Target Method	Count
Normalize Values	Yes

After processing, the number of Satisfied becomes 127, as shown in the figure below:

The number of missing values is 0.

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Age	INPUT	16	0	30	13.71	32	9.43
TRAIN	City	INPUT	6	0	Los Angeles	16.86	New York	16.86
TRAIN	Discount_Applied	INPUT	2	0	FALSE	50.00	TRUE	50.00
TRAIN	Gender	INPUT	2	0	Female	50.00	Male	50.00
TRAIN	Items_Purchased	INPUT	15	0	10	13.43	9	9.71
TRAIN	Membership_Type	INPUT	3	0	Gold	33.43	Silver	33.43
TRAIN	IMP_Satisfaction_Level	TARGET	3	0	Satisfied	36.29	Unsatisfied	33.14

specify variable roles:

This data set I use to study the factors affecting customer satisfaction level, so I set the satisfaction_level as the target variable:

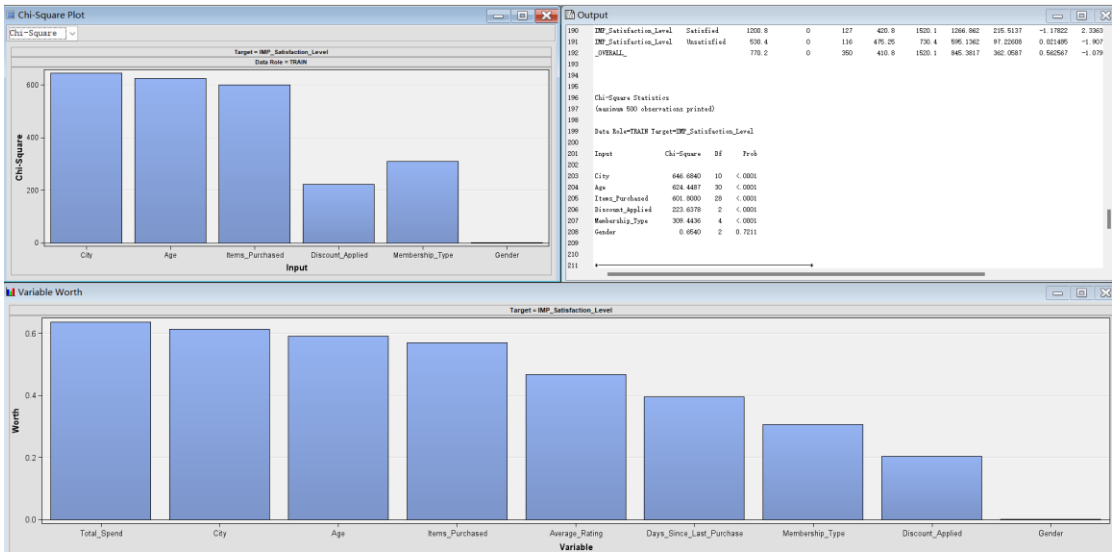
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Nominal	No		No	.	.
Average_Rating	Input	Interval	No		No	.	.
City	Input	Nominal	No		No	.	.
Customer_ID	ID	Interval	No		No	.	.
Days_Since_Last_Purchase	Input	Interval	No		No	.	.
Discount_Applied	Input	Binary	No		No	.	.
Gender	Input	Binary	No		No	.	.
Items_Purchased	Input	Nominal	No		No	.	.
Membership_Type	Input	Nominal	No		No	.	.
Satisfaction_Level	Target	Nominal	No		No	.	.
Total_Spend	Input	Interval	No		No	.	.

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

[20 marks]

Explore Data Analysis:

The result of StatExplore:

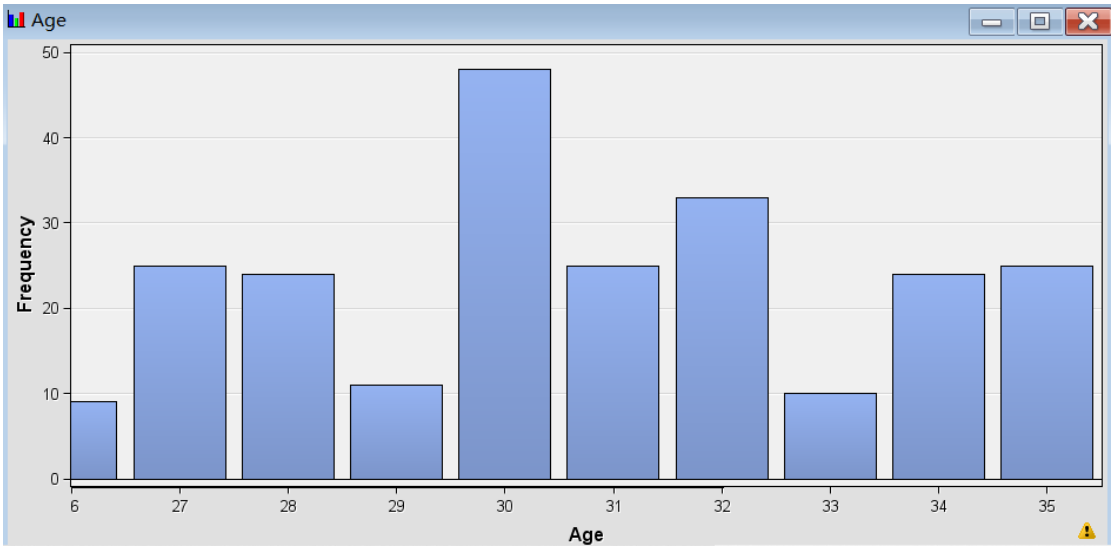


The explore for variable of age:

The age of individuals in the data set is concentrated around 30 years old.

In terms of satisfaction, the most common feedback was "satisfied."

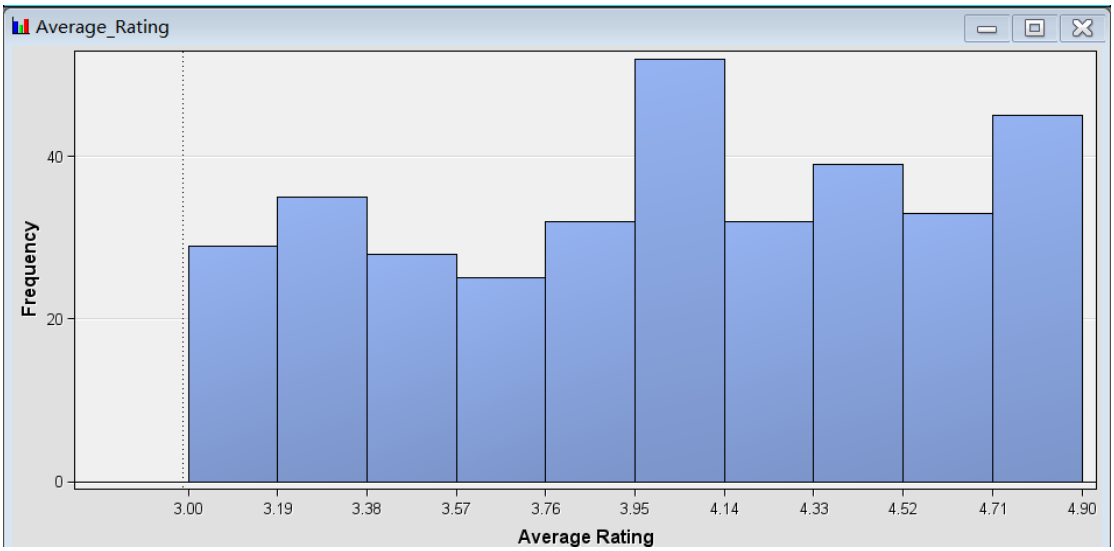
Obs #	Variable Name	Label	Type	Percent Missing	Minimum	Maximum	Mean	Number of Levels	Mode ...	Mode
1	IMP_Satisfaction_Level	Imputed: Satisfaction Level	CLASS	0	.	.	.	3	36.28571	SATISFIED
2	Age		VAR	0	26	43	33.59714	.	.	
3	Customer_ID	Customer ID	VAR	0	101	450	275.5	.	.	



The explore for variable of Average Rating:

The minimum value of the average rating is 3, the maximum value is 4.9, the mean is approximately 4.19, and the mode is 4.90. This indicates that most customers' ratings tend to be on the higher side, with greater satisfaction.

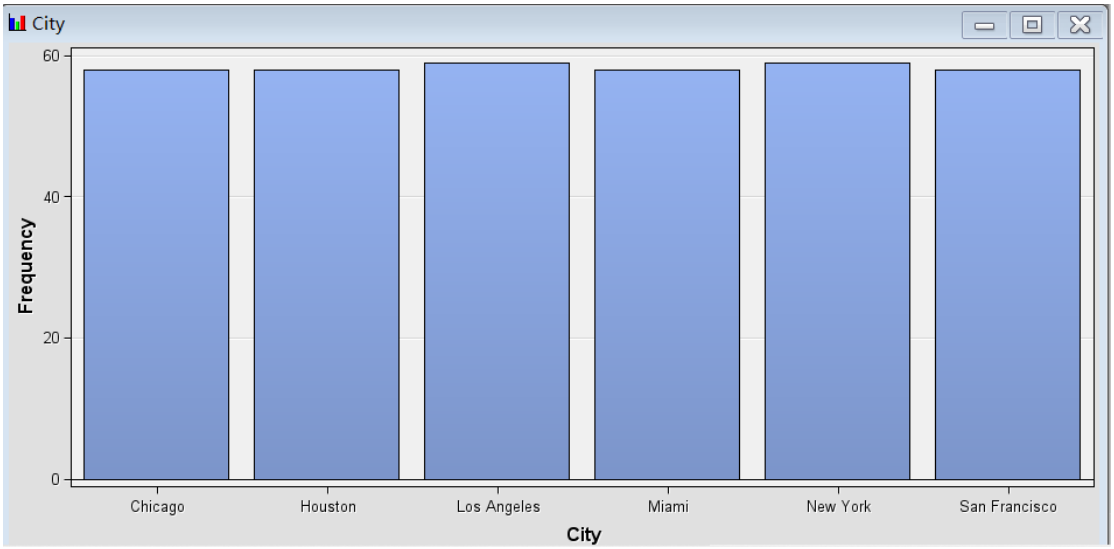
Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode
1	IMP_Satisfaction_Level	Imputed: Satisfaction Level	CLASS	0	.	.	.	3	36.285
2	Average_Rating	Average Rating	VAR	0	3	4.9	4.019143	.	
3	Customer_ID	Customer ID	VAR	0	101	450	275.5	.	



The explore for variable of City:

Since each city appears roughly equally frequently in the data set, this means it makes sense to compare satisfaction levels across these cities as there will be no bias caused by the sample size.

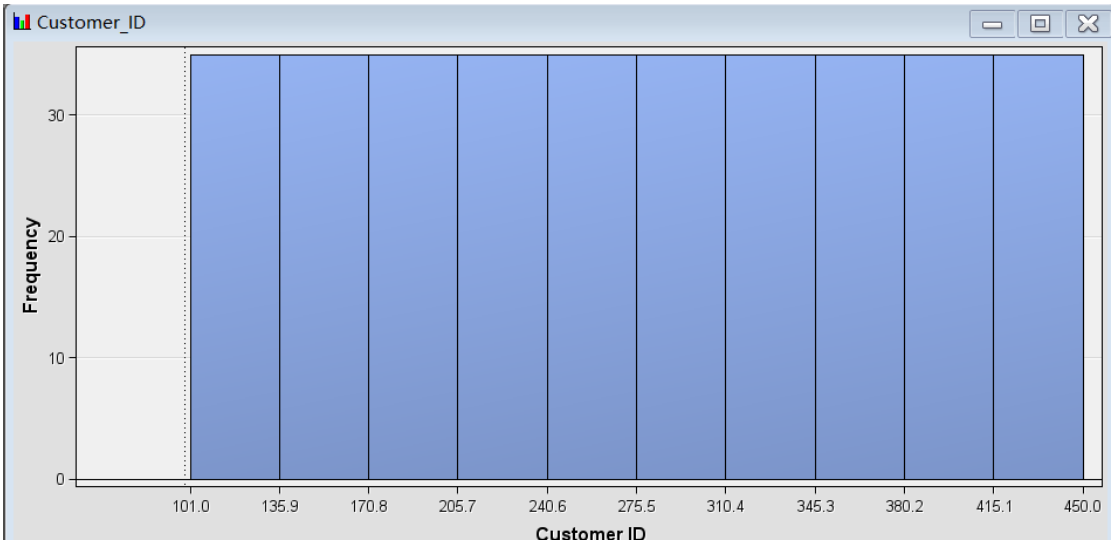
Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	City		CLASS	0	.	.	.	6	16.85714	LOS ANGELES
2	IMP_Satisfaction_Level	Imputed: ...	CLASS	0	.	.	.	3	36.28571	SATISFIED
3	Customer_ID	Custome...	VAR	0	101	450	275.5	.	.	



The explore for variable of Customer ID:

Each customer number occurs very closely, which means the data set may be evenly distributed, or there may be only one record for each customer.

Obs #	Variable Name	Label	Type	Pe...	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	IMP_Satisfaction_Level	Imputed: Satisf...	CLASS	0	.	.	.	3	36.28571	SATISFIED
2	Customer_ID	Customer ID	VAR	0	101	450	275.5	.	.	

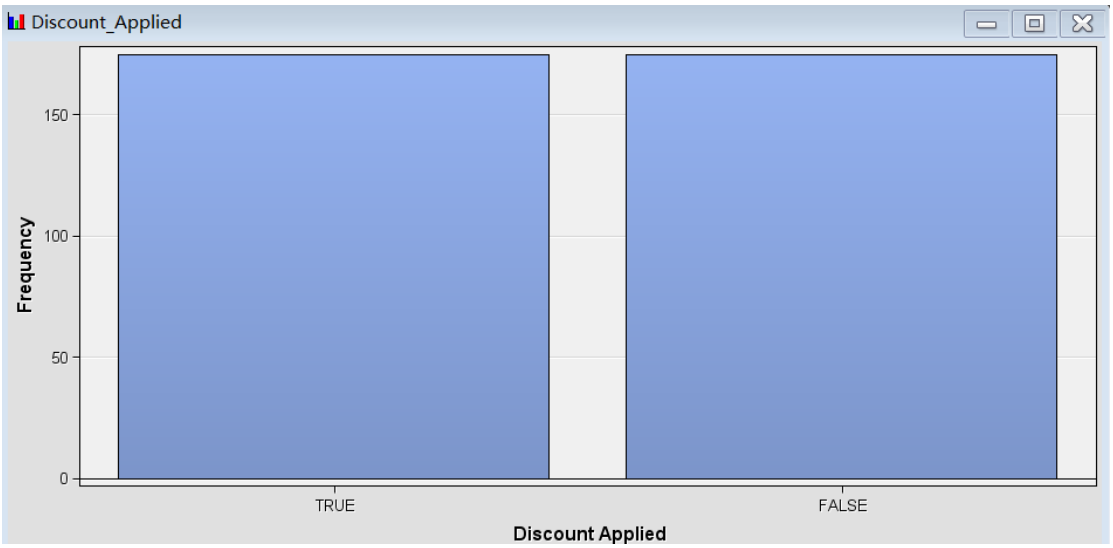


The explore for variable of Days_Since_Last_Purchase:

Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...
1	IMP_Satisfaction_Level	Imputed: Satisfaction Level	CLASS	0	.	.	.3	
2	Customer_ID	Customer ID	VAR	0	101	450	275.5.	
3	Days_Since_Last_Purchase	Days Since Last Purchase	VAR	0	9	63	26.58857.	

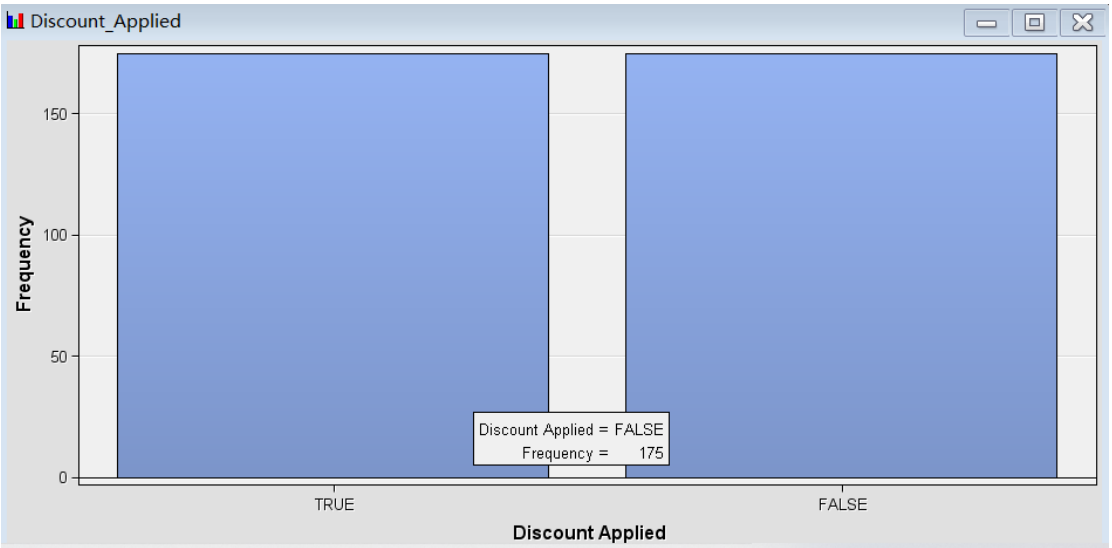
The explore for variable of Days_Since_Last_Purchase:

#	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...	I
1	Discount_Applied	Discount Applied	CLASS	0	.	.	.2		50	FA
2	IMP_Satisfaction_Level	Imputed: Satisfaction Level	CLASS	0	.	.	.3		36.28571	SA
3	Customer_ID	Customer ID	VAR	0	101	450	275.5.		.	



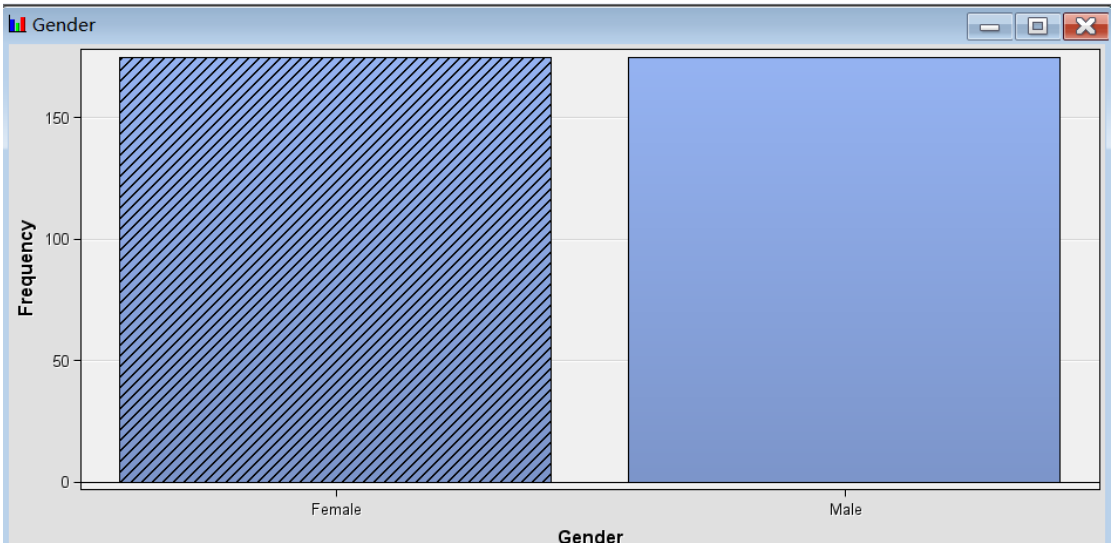
The explore for variable of Discount_Applied:

Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	Discount_Applied	Discount ...	CLASS	0	.	.	.	2	50	FALSE
2	IMP_Satisfaction_Level	Imputed: ...	CLASS	0	.	.	.	3	36.28571	SATISFIED
3	Customer_ID	Customer...	VAR	0	101	450	275.5	.	.	



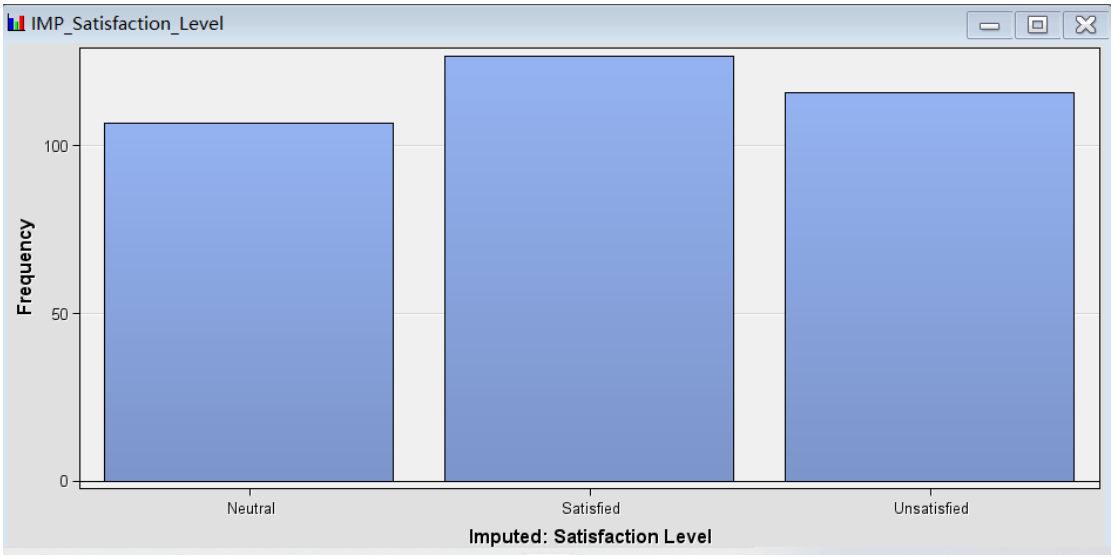
The explore for variable of Gender:

Obs #	Variable Name	Label	Type	Perc...	Minimum	Maximum	Mean	Numbe...	Mod...
1	Gender		CLASS	0	.	.	.	2	50
2	IMP_Satisfaction_Level	Imputed: Satisfaction Level	CLASS	0	.	.	.	3	36.28...
3	Customer_ID	Customer ID	VAR	0	101	450	275.5	.	.



The explore for variable of Satisfication_Level:

Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	IMP_Satisfaction_Level	Imputed: ...	CLASS	0	.	.	.	3	36.28571	SATISFIED
2	Customer_ID	Custome...	VAR	0	101	450	275.5	.	.	

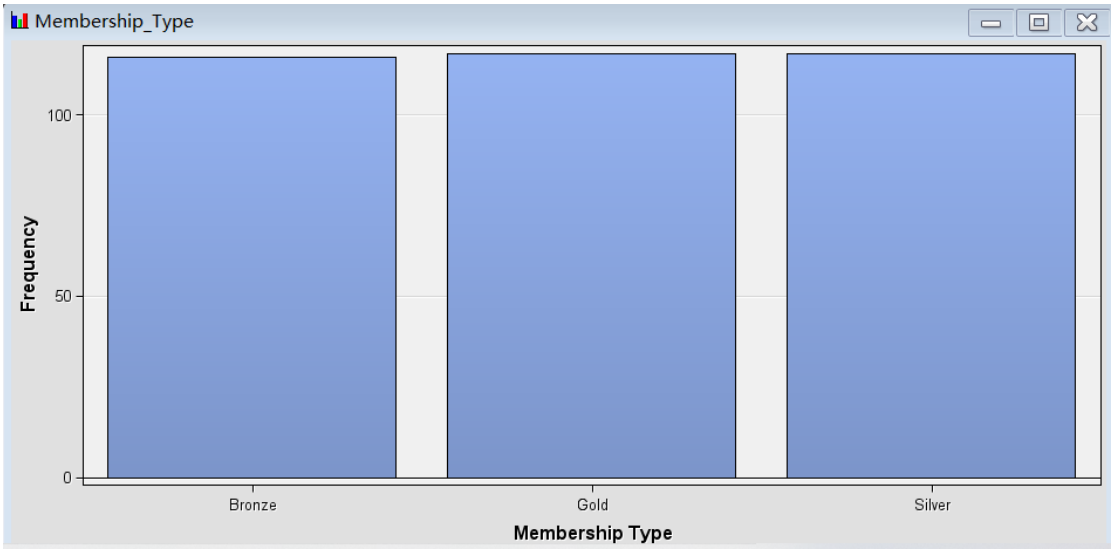


The explore for variable of Items_Purchased:

Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	IMP_Satisfaction_Level	Imputed: S...	CLASS	0	.	.	.	3	36.28571	SATISFIED
2	Customer_ID	Customer ID	VAR	0	101	450	275.5	.	.	
3	Items_Purchased	Items Purc...	VAR	0	7	21	12.6	.	.	

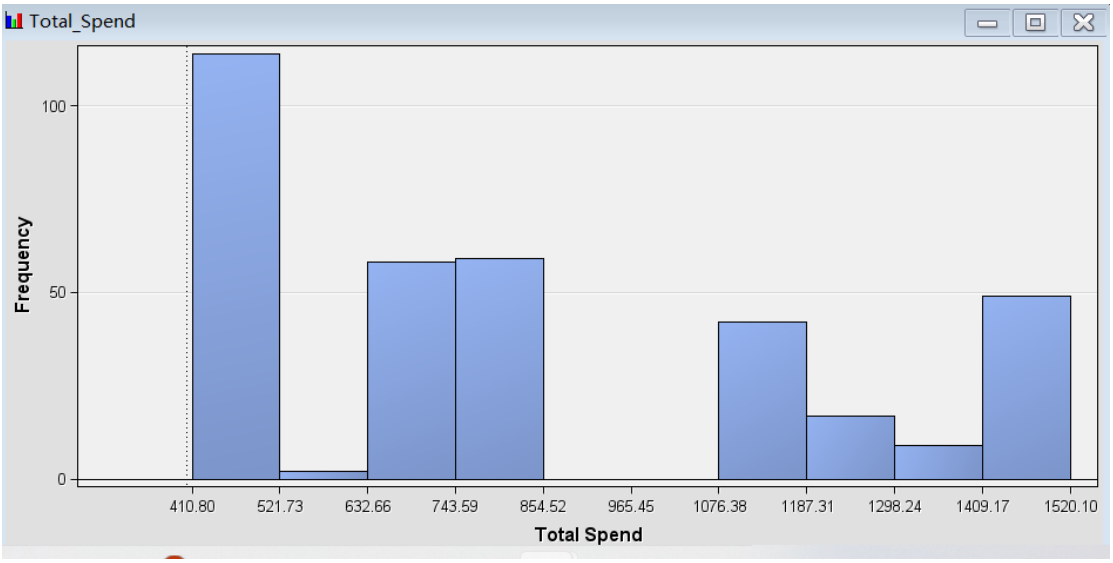
The explore for variable of Membership_Type:

Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...	Mode
1	IMP_Satisfaction_Level	Imputed: Satis...	CLASS	0	.	.	.	3	36.28571	SATISFIED
2	Membership_Type	Membership T...	CLASS	0	.	.	.	3	33.42857	GOLD
3	Customer_ID	Customer ID	VAR	0	101	450	275.5	.	.	



The explore for variable of Total_Spend:

Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Numbe...	Mode ...
1	IMP_Satisfaction_Level	Imputed: Satisfaction Level	CLASS	0	.	.	.	3	36.28571
2	Customer_ID	Customer ID	VAR	0	101	450	275.5	.	.
3	Total_Spend	Total Spend	VAR	0	410.8	1520.1	845.3817	.	.



Step1:

Before running the decision tree node, we first divide the data according to 70 (training): 30 (validation)**:**

Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0

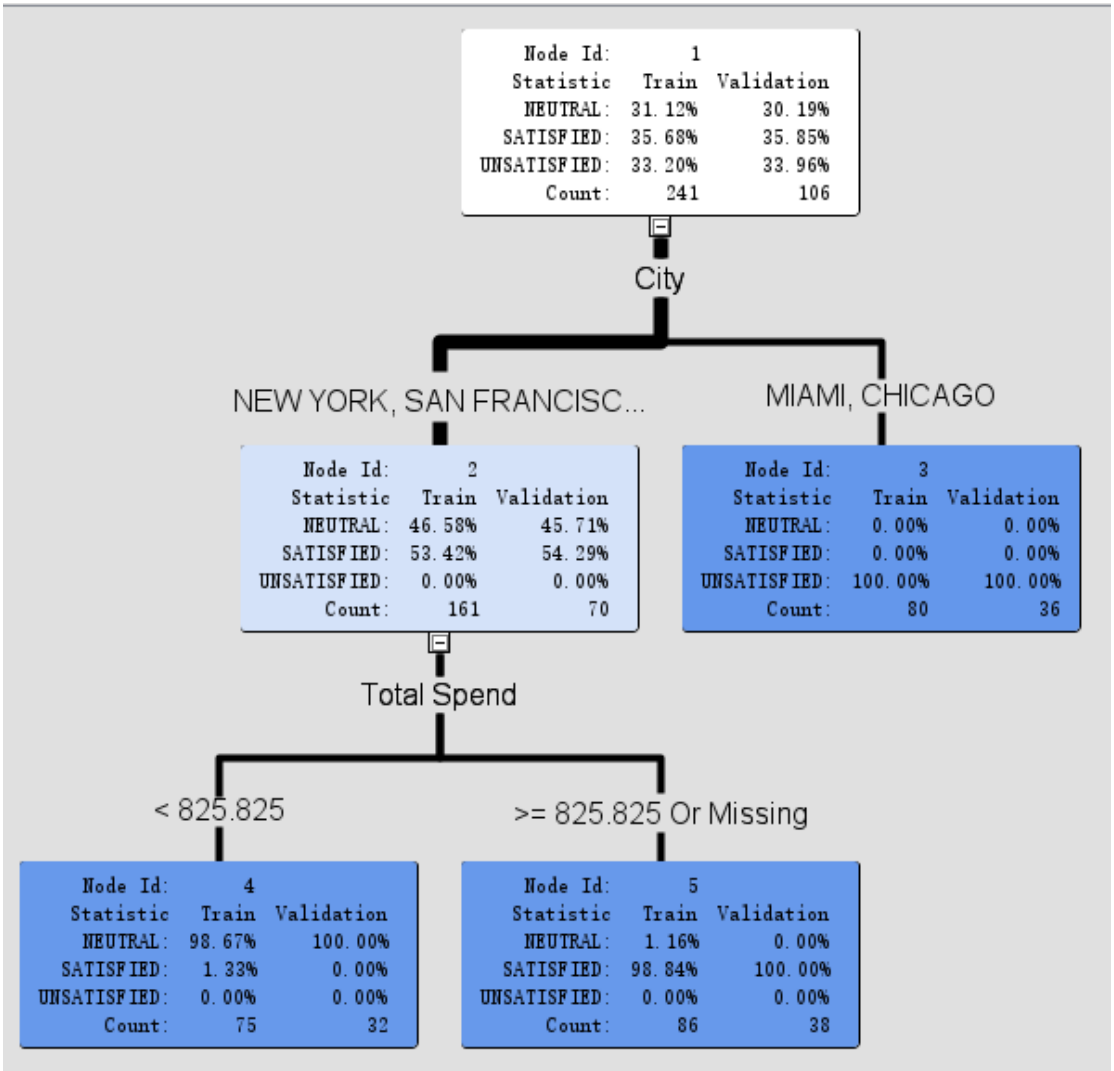
Step2:

Add decision tree node:

Configure decision tree parameters**:**

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 Targets	No
Splitting Rule	
Interval Target Criterio	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rule	0
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

Run decision tree node, the results are as follows:



Analyze customer behavior:

The impact of city on satisfaction:

The first branch of the decision tree is the city-based variable, which shows that the city where the user is located has a significant impact on their satisfaction. For example, users from Chicago and Miami were completely dissatisfied, while users from New York and Los Angeles were highly satisfied.

The impact of total consumption on satisfaction:

After the city variable, the decision tree considers the total consumption of the user. Users who spend less than 825.825 tend to show extremely high

satisfaction or a neutral attitude, while users who spend more than this threshold are completely satisfied.

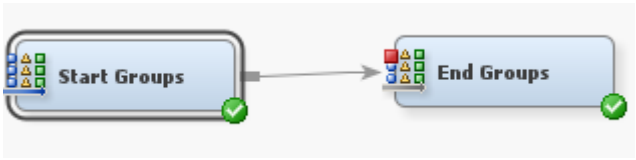
User behavior details:

Among low-consumption users, almost all are satisfied or neutral with the service, which may indicate that price has a great impact on the satisfaction of this group of users.

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

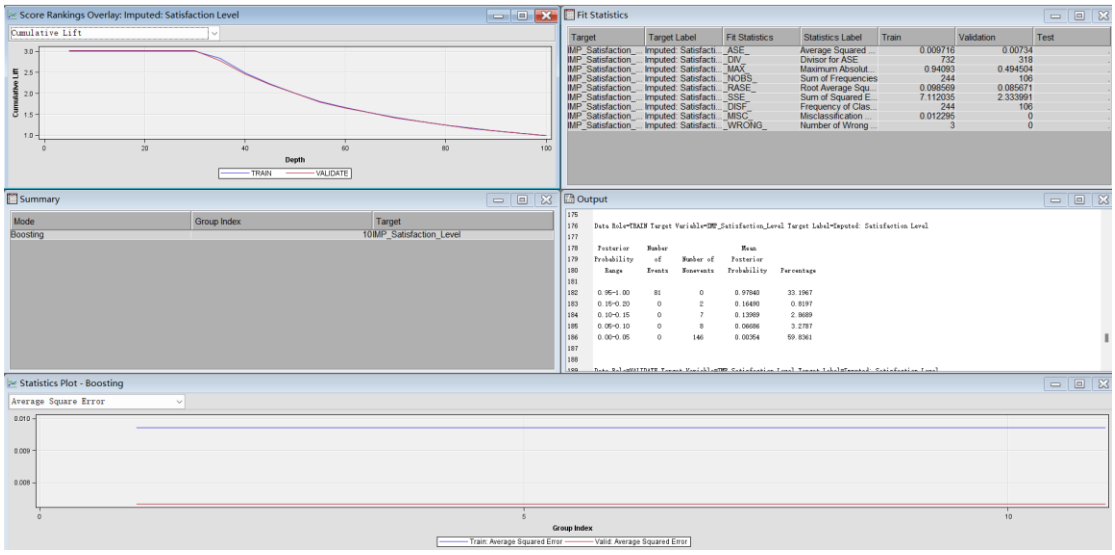
[10 marks]

Add Start Groups and End Groups, and under the Start Groups node, set the mode to Boosting.



General	
Mode	Boosting
Target Group	No
Index Count	10
Minimum Group Size	10

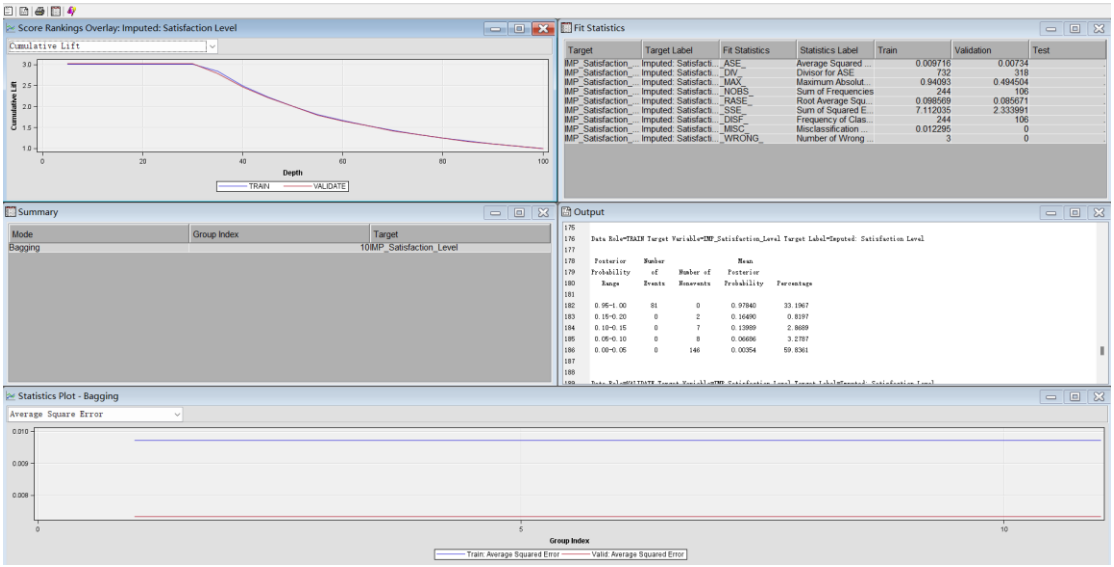
The result of End Groups in Boosting:



Add Start Groups and End Groups, and under the Start Groups node, set the mode to Bagging.

General	
Mode	Bagging
Target Group	No
Index Count	10
Minimum Group Size	10
Bagging	
Type	Percentage
Observations	.
Percentage	10.0
Random Seed	12345

The result of End Groups in Bagging:

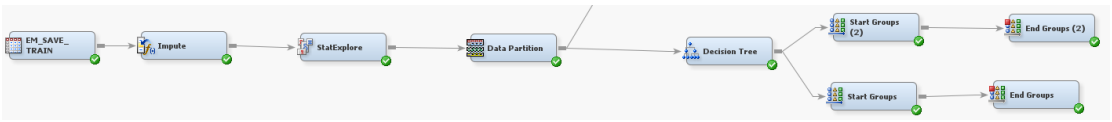


Deliverables:

A report detailing each step of the process, including the rationale behind your choices and any challenges faced.

An analysis of the decision tree and ensemble methods, with insights into customer behavior and suggestions for business strategy.

[5 marks]



Step1:

Import the dataset into Talend Data Preparation

Use the mode to fill in missing values.

Step2:

Import the original data set through the import file node, and then save it as a sas file format.

Step3:

Create data source and load data source.

Step4:
Perform exploratory analysis of data in SAS Miner by importing the StatExplore node.
Step5:
Process missing values in SAS Miner, import the Impute node, and use the mode to fill in the missing values.
Step6:
Split the data according to 70（training）:30(validation).
Step7:
Configure decision tree parameters and run decision tree node.
Step8:
View the results of a decision tree run.
Step9:
Apply Bagging and Boosting, using the Random Forest algorithm.
Step10:
View the results of a Bagging and Boosting, using the Random Forest run (by start groups and end groups).
Challenges faced:
1、I don't know how to change the missing values for the classification problem in sas miner, and finally found the solution.
2、Regarding question 3, I am not very sure how to solve this problem with Apply Bagging and Boosting, using the Random Forest algorithm. I have tried many methods, but I don't know which one is the correct solution.
insights into customer behavior：
Can help identify customer groups most likely to be satisfied or dissatisfied, allowing companies to target specific improvements or marketing campaigns to these groups.
Satisfaction is linked to specific buying patterns, consumption habits or customer feedback, allowing targeted improvements to products and services.
suggestions for business strategy：
Personalize marketing messages and promotions based on customer satisfaction levels and purchasing behavior.
Improve service processes based on factors that lead to dissatisfaction.
Pay more attention to and analyze customer feedback, understand the specific reasons for customer dissatisfaction, and take measures to solve these problems.