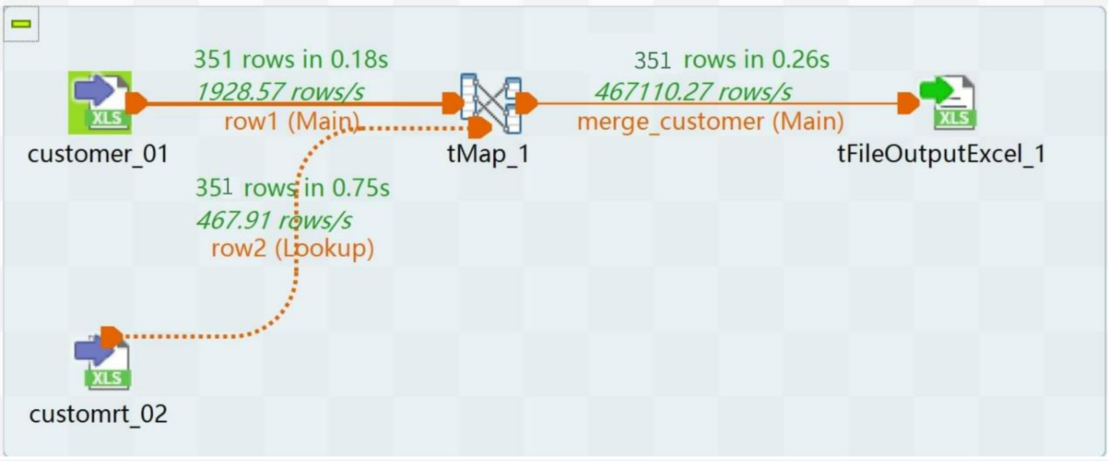
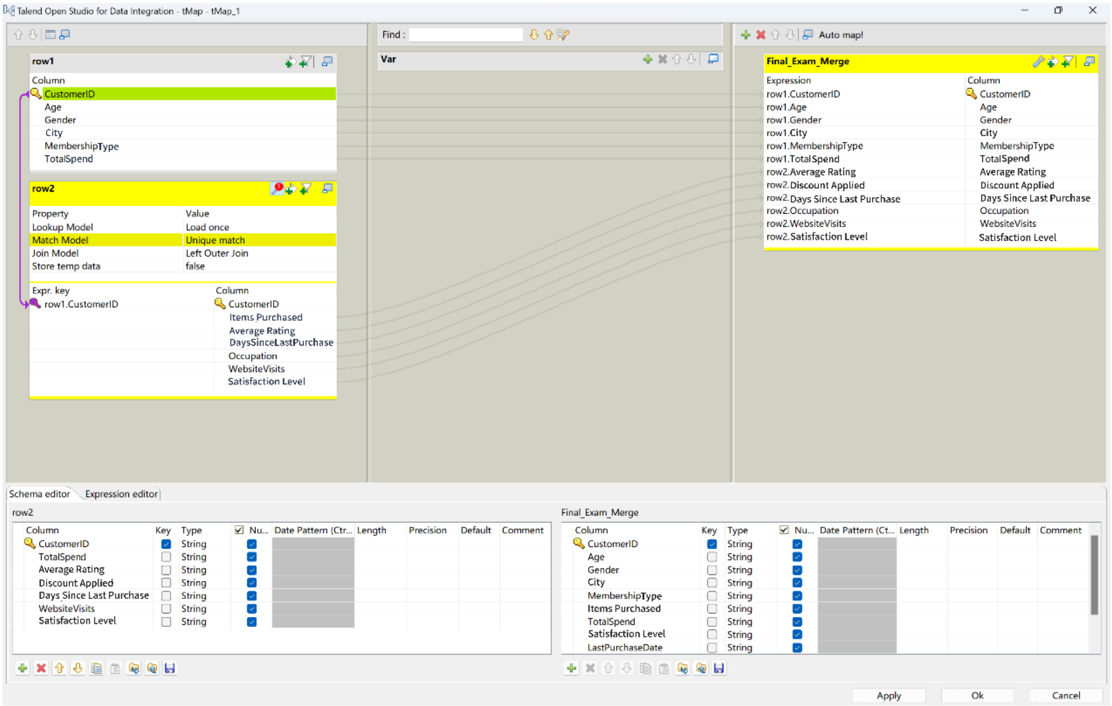
Documentation for Each Tool

Talend Data Integration：

Firstly, it is necessary to merge two datasets using Talend, both of which have 351 rows. One dataset contains attributes such as Age, Gender, City, MembershipType, and Items Purchased, while the other dataset contains attributes such as TotalSpend, Average Rating, Discount Applied, Days Since Last Purchase, Satisfaction Level. Both datasets have the primary key of Customer ID, It is also through this primary key that the two datasets are associated.





row1

Column
CustomerID
Age
Gender
City
MembershipType
TotalSpend

row2

Property	Value
Lookup Model	Load once
Match Model	Unique match
Join Model	Left Outer Join
Store temp data	false

Expr. key	Column
row1.CustomerID	CustomerID
	Items Purchased
	Average Rating
	DaysSinceLastPurchase
	Occupation
	WebsiteVisits
	Satisfaction Level

Final_Exam_Merge

Expression	Column
row1.CustomerID	CustomerID
row1.Age	Age
row1.Gender	Gender
row1.City	City
row1.MembershipType	MembershipType
row1.TotalSpend	TotalSpend
row2.Average Rating	Average Rating
row2.Discount Applied	Discount Applied
row2.Days Since Last Purchase	Days Since Last Purchase
row2.Occupation	Occupation
row2.WebsiteVisits	WebsiteVisits
row2.Satisfaction Level	Satisfaction Level

row2

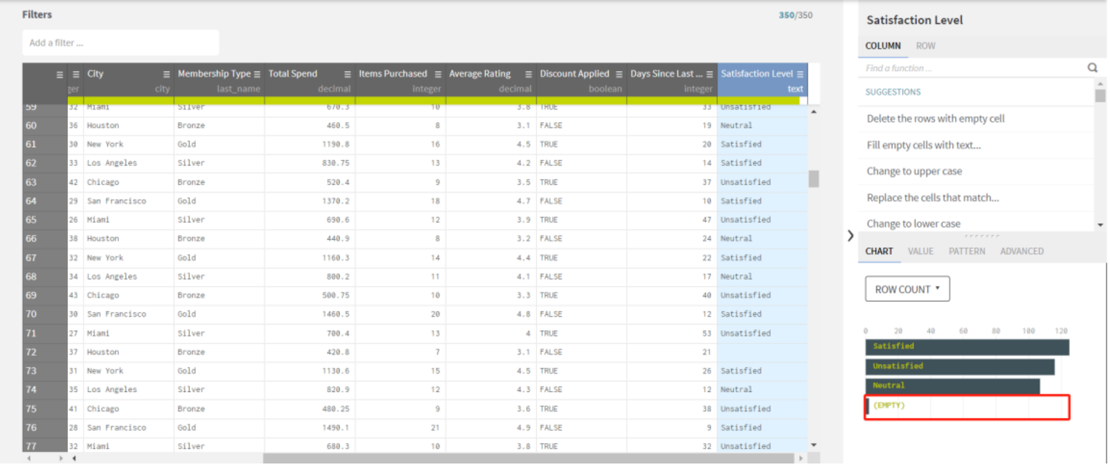
Column	Key	Type	☑ Nu...	Date Pattern (Ctr...	Length	Precision	Default	Comment
CustomerID	☑	String	☑					
TotalSpend	□	String	☑					
Average Rating	□	String	☑					
Discount Applied	□	String	☑					
Days Since Last Purchase	□	String	☑					
WebsiteVisits	□	String	☑					
Satisfaction Level	□	String	☑					

Final_Exam_Merge

Column	Key	Type	☑ Nu...	Date Pattern (Ct...	Length	Precision	Default	Comment
CustomerID	☑	String	☑					
Age	□	String	☑					
Gender	□	String	☑					
City	□	String	☑					
MembershipType	□	String	☑					
Items Purchased	□	String	☑					
TotalSpend	□	String	☑					
Satisfaction Level	□	String	☑					
LastPurchaseDate	□	String	☑					

Talend Data Prep:

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.



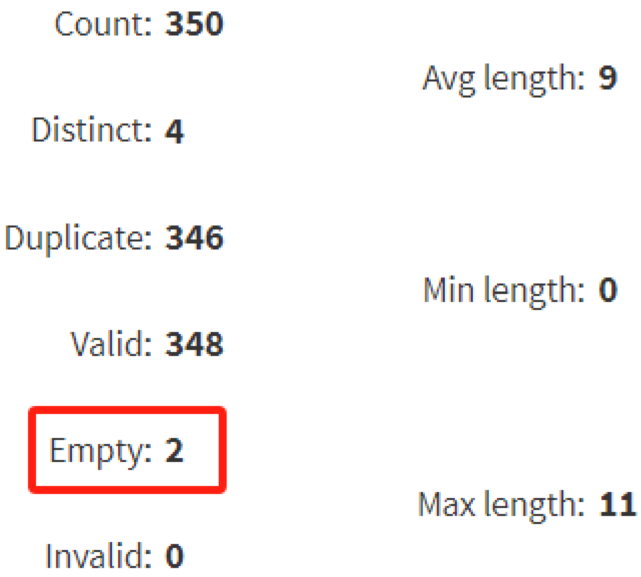
Filters

Add a filter ...

350/350

	≡	City ≡ city	Membership Type ≡ last_name	Total Spend ≡ decimal	Items Purchased ≡ integer	Average Rating ≡ decimal	Discount Applied ≡ boolean	Days Since Last ... ≡ integer	Satisfaction Level ≡ text
59	32	Miami	Silver	670.3	10	3.8	TRUE	33	Unsatisfied
60	36	Houston	Bronze	460.5	8	3.1	FALSE	19	Neutral
61	30	New York	Gold	1190.8	16	4.5	TRUE	20	Satisfied
62	33	Los Angeles	Silver	830.75	13	4.2	FALSE	14	Satisfied
63	42	Chicago	Bronze	520.4	9	3.5	TRUE	37	Unsatisfied
64	29	San Francisco	Gold	1370.2	18	4.7	FALSE	10	Satisfied
65	26	Miami	Silver	690.6	12	3.9	TRUE	47	Unsatisfied
66	38	Houston	Bronze	440.9	8	3.2	FALSE	24	Neutral
67	32	New York	Gold	1160.3	14	4.4	TRUE	22	Satisfied
68	34	Los Angeles	Silver	800.2	11	4.1	FALSE	17	Neutral
69	43	Chicago	Bronze	500.75	10	3.3	TRUE	40	Unsatisfied
70	30	San Francisco	Gold	1460.5	20	4.8	FALSE	12	Satisfied
71	27	Miami	Silver	700.4	13	4	TRUE	53	Unsatisfied
72	37	Houston	Bronze	420.8	7	3.1	FALSE	21	
73	31	New York	Gold	1130.6	15	4.5	TRUE	26	Satisfied
74	35	Los Angeles	Silver	820.9	12	4.3	FALSE	12	Neutral
75	41	Chicago	Bronze	480.25	9	3.6	TRUE	38	Unsatisfied
76	28	San Francisco	Gold	1490.1	21	4.9	FALSE	9	Satisfied
77	32	Miami	Silver	680.3	10	3.8	TRUE	32	Unsatisfied

Satisfaction Level



Count: 350

Avg length: 9

Distinct: 4

Duplicate: 346

Min length: 0

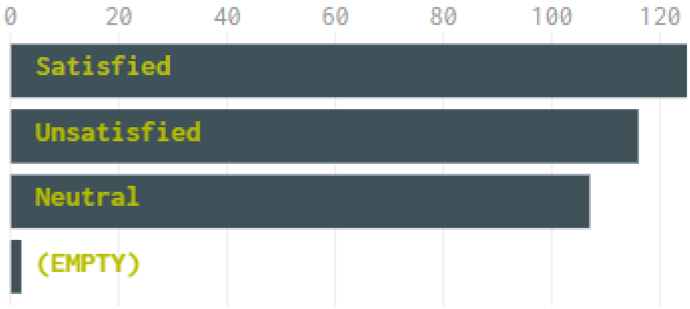
Valid: 348

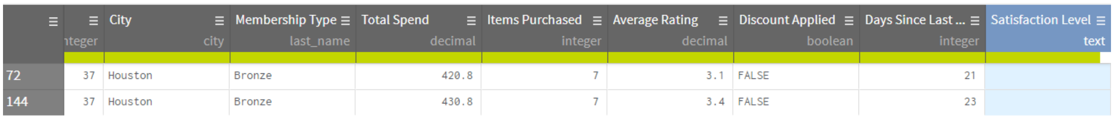
Empty: 2

Max length: 11

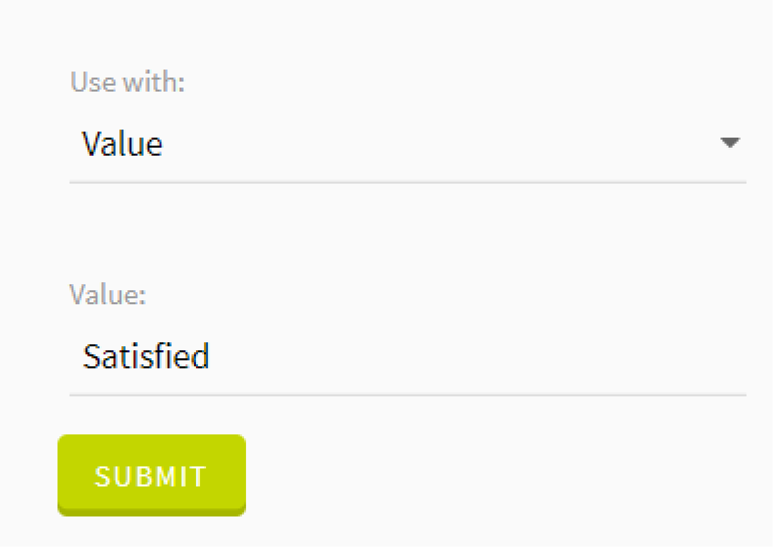
Invalid: 0

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





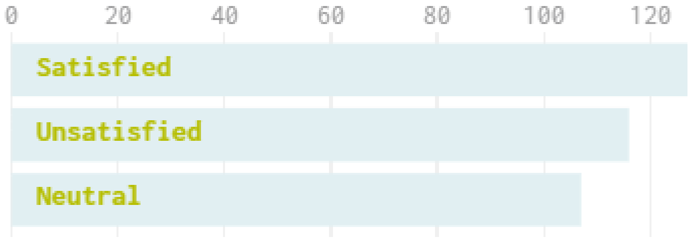
		City	Membership Type	Total Spend	Items Purchased	Average Rating	Discount Applied	Days Since Last ...	Satisfaction Level
	nteger	city	last_name	decimal	integer	decimal	boolean	integer	text
72	37	Houston	Bronze	420.8	7	3.1	FALSE	21	
144	37	Houston	Bronze	430.8	7	3.4	FALSE	23	

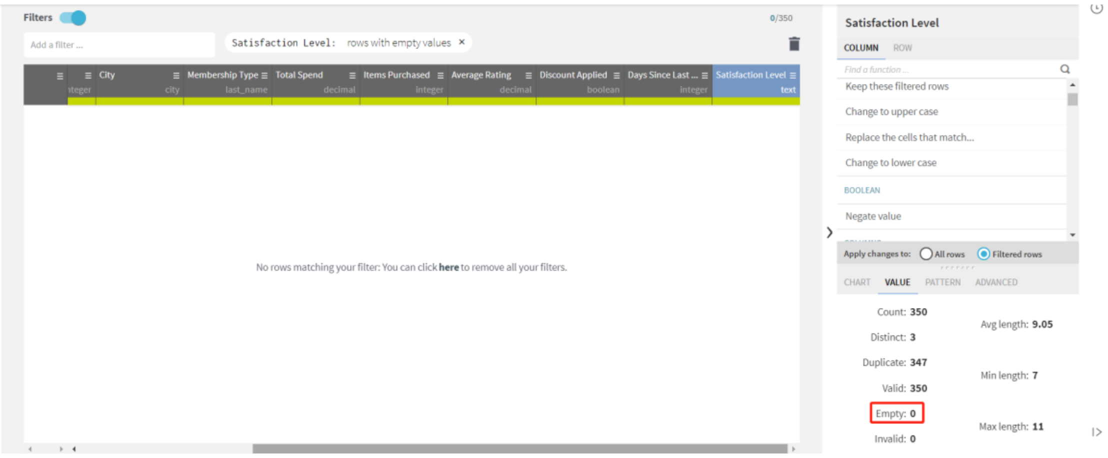


Use with: Value

Value: Satisfied

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





Filters

Add a filter ...

Satisfaction Level: rows with empty values ×

0/350

	City	Membership Type	Total Spend	Items Purchased	Average Rating	Discount Applied	Days Since Last ...	Satisfaction Level
nteger	city	last_name	decimal	integer	decimal	boolean	integer	text

No rows matching your filter: You can click **here** to remove all your filters.

Satisfaction Level

Apply changes to: ○ All rows ● Filtered rows

Count: **350**

Distinct: **3**

Duplicate: **347**

Valid: **350**

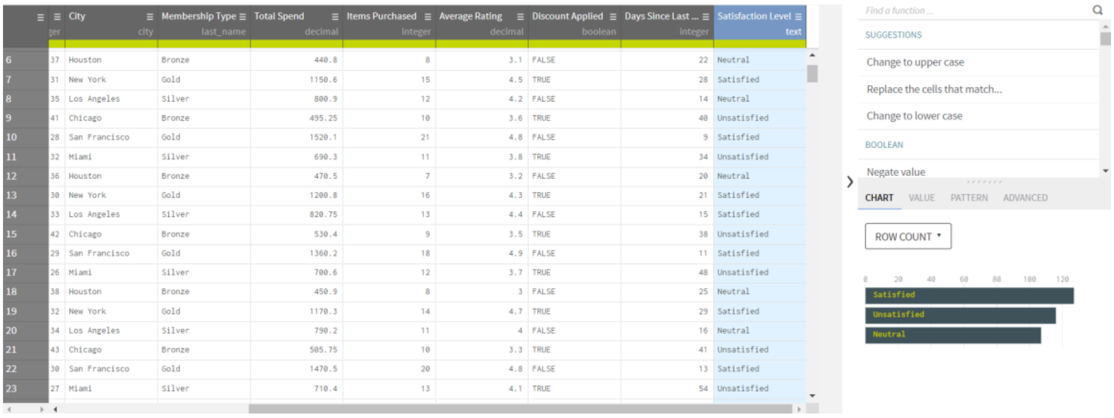
Empty: **0**

Invalid: **0**

Avg length: **9.05**

Min length: **7**

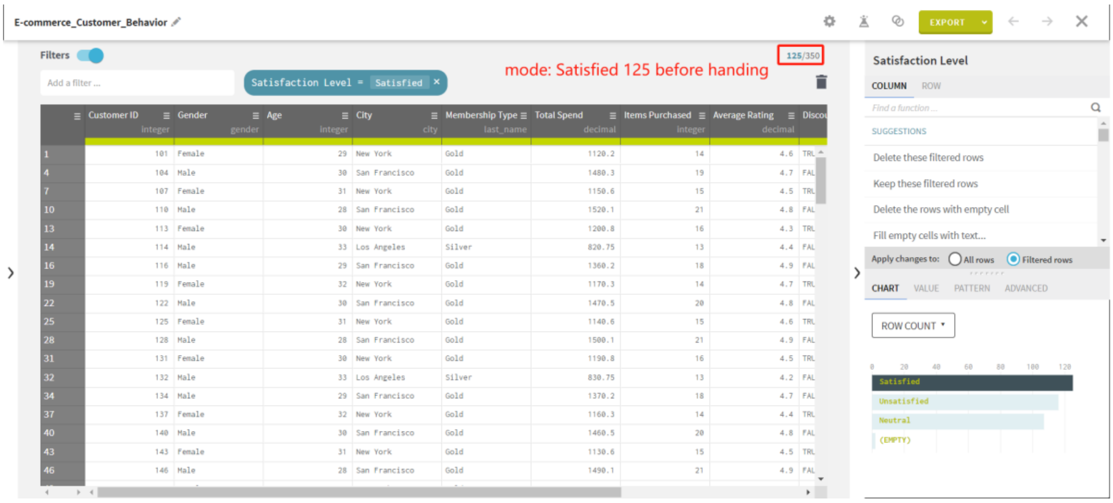
Max length: **11**



	ger	City	Membership Type	Total Spend	Items Purchased	Average Rating	Discount Applied	Days Since Last ...	Satisfaction Level
		city	last_name	decimal	integer	decimal	boolean	integer	text
6	37	Houston	Bronze	440.8	8	3.1	FALSE	22	Neutral
7	31	New York	Gold	1150.6	15	4.5	TRUE	28	Satisfied
8	35	Los Angeles	Silver	800.9	12	4.2	FALSE	14	Neutral
9	41	Chicago	Bronze	495.25	10	3.6	TRUE	40	Unsatisfied
10	28	San Francisco	Gold	1520.1	21	4.8	FALSE	9	Satisfied
11	32	Miami	Silver	690.3	11	3.8	TRUE	34	Unsatisfied
12	36	Houston	Bronze	470.5	7	3.2	FALSE	20	Neutral
13	30	New York	Gold	1200.8	16	4.3	TRUE	21	Satisfied
14	33	Los Angeles	Silver	820.75	13	4.4	FALSE	15	Satisfied
15	42	Chicago	Bronze	530.4	9	3.5	TRUE	38	Unsatisfied
16	29	San Francisco	Gold	1360.2	18	4.9	FALSE	11	Satisfied
17	26	Miami	Silver	700.6	12	3.7	TRUE	48	Unsatisfied
18	38	Houston	Bronze	450.9	8	3	FALSE	25	Neutral
19	32	New York	Gold	1170.3	14	4.7	TRUE	29	Satisfied
20	34	Los Angeles	Silver	790.2	11	4	FALSE	16	Neutral
21	43	Chicago	Bronze	505.75	10	3.3	TRUE	41	Unsatisfied
22	30	San Francisco	Gold	1470.5	20	4.8	FALSE	13	Satisfied
23	27	Miami	Silver	710.4	13	4.1	TRUE	54	Unsatisfied

Missing value handling in SAS Miner:

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



E-commerce_Customer_Behavior

Filters

Satisfaction Level = Satisfied

mode: Satisfied 125 before handing

125/350

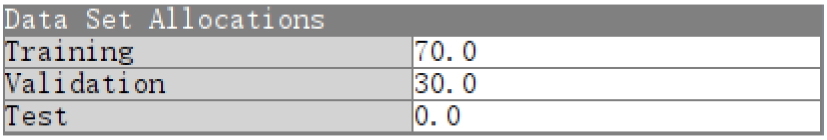
	Customer ID	Gender	Age	City	Membership Type	Total Spend	Items Purchased	Average Rating	Disco
	integer	gender	integer	city	last_name	decimal	integer	decimal	
1	101	Female	29	New York	Gold	1120.2	14	4.6	TRU
4	104	Male	30	San Francisco	Gold	1480.3	19	4.7	FAL
7	107	Female	31	New York	Gold	1150.6	15	4.5	TRU
10	110	Male	28	San Francisco	Gold	1520.1	21	4.8	FAL
13	113	Female	30	New York	Gold	1200.8	16	4.3	TRU
14	114	Male	33	Los Angeles	Silver	820.75	13	4.4	FAL
16	116	Male	29	San Francisco	Gold	1360.2	18	4.9	FAL
19	119	Female	32	New York	Gold	1170.3	14	4.7	TRU
22	122	Male	30	San Francisco	Gold	1470.5	20	4.8	FAL
25	125	Female	31	New York	Gold	1140.6	15	4.6	TRU
28	128	Male	28	San Francisco	Gold	1500.1	21	4.9	FAL
31	131	Female	30	New York	Gold	1190.8	16	4.5	TRU
32	132	Male	33	Los Angeles	Silver	830.75	13	4.2	FAL
34	134	Male	29	San Francisco	Gold	1370.2	18	4.7	FAL
37	137	Female	32	New York	Gold	1160.3	14	4.4	TRU
40	140	Male	30	San Francisco	Gold	1460.5	20	4.8	FAL
43	143	Female	31	New York	Gold	1130.6	15	4.5	TRU
46	146	Male	28	San Francisco	Gold	1490.1	21	4.9	FAL

Apply changes to: ○ All rows ● Filtered rows

SAS e-Miner:

Step1:

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

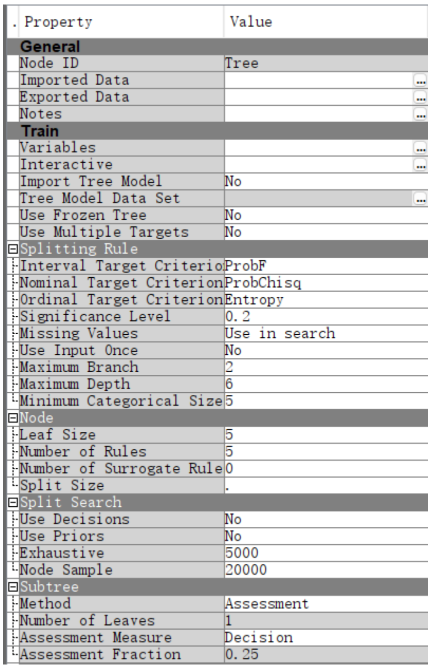


Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0

Step2:

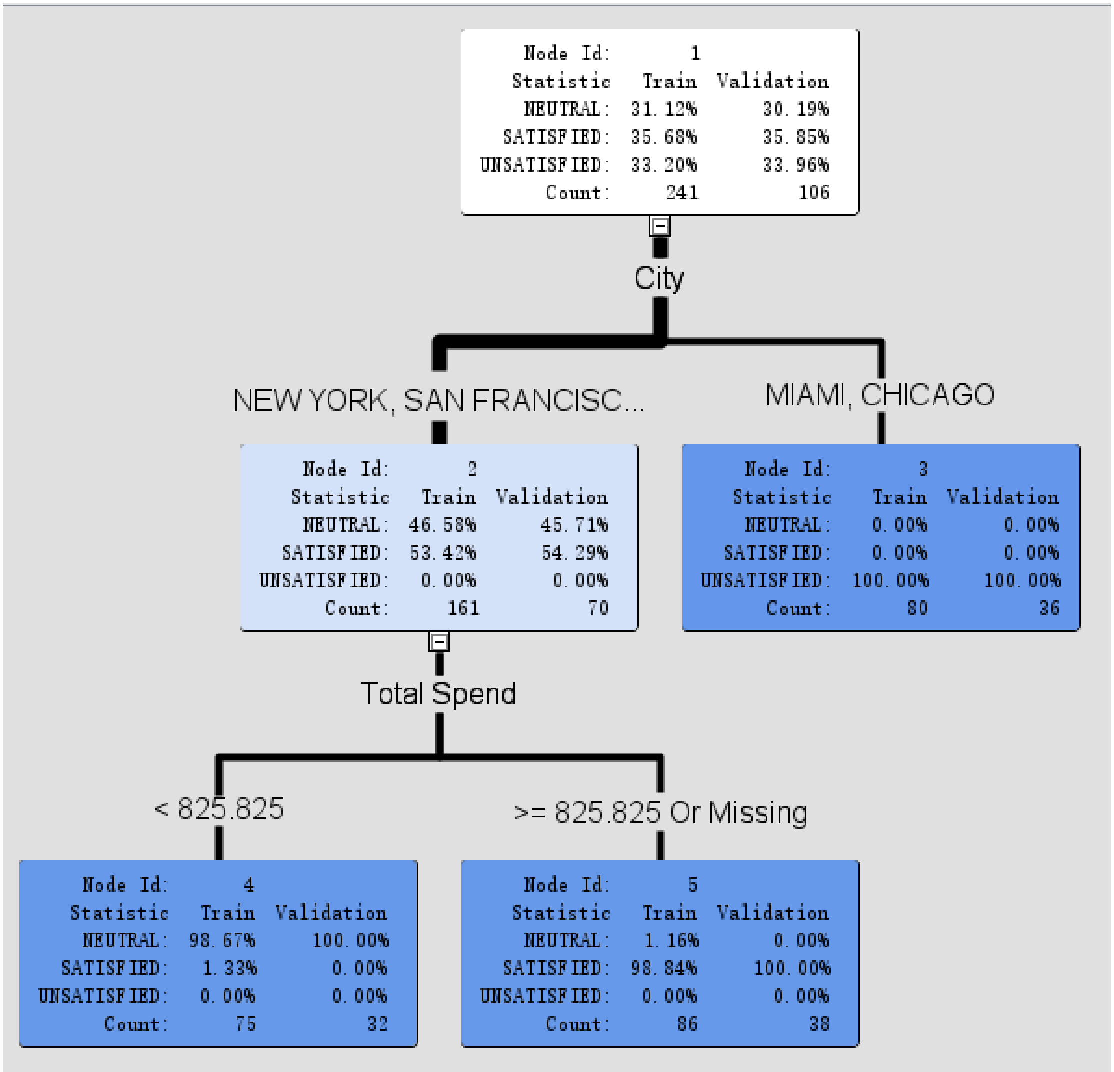
Add decision tree node:

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:



```
Node Id:        1
Statistic   Train  Validation
NEUTRAL:   31.12%      30.19%
SATISFIED:   35.68%      35.85%
UNSATISFIED:   33.20%      33.96%
Count:      241         106
```

```
Node Id:        2
Statistic   Train  Validation
NEUTRAL:   46.58%      45.71%
SATISFIED:   53.42%      54.29%
UNSATISFIED:    0.00%       0.00%
Count:      161          70
```

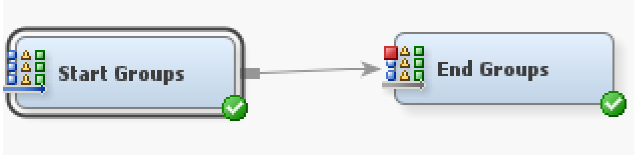
```
Node Id:        3
Statistic   Train  Validation
NEUTRAL:    0.00%       0.00%
SATISFIED:    0.00%       0.00%
UNSATISFIED:  100.00%     100.00%
Count:       80          36
```

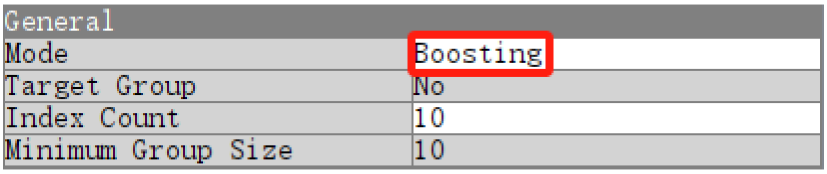
```
Node Id:        4
Statistic   Train  Validation
NEUTRAL:   98.67%     100.00%
SATISFIED:    1.33%       0.00%
UNSATISFIED:    0.00%       0.00%
Count:       75          32
```

```
Node Id:        5
Statistic   Train  Validation
NEUTRAL:    1.16%       0.00%
SATISFIED:   98.84%     100.00%
UNSATISFIED:    0.00%       0.00%
Count:       86          38
```

Ensemble Methods:

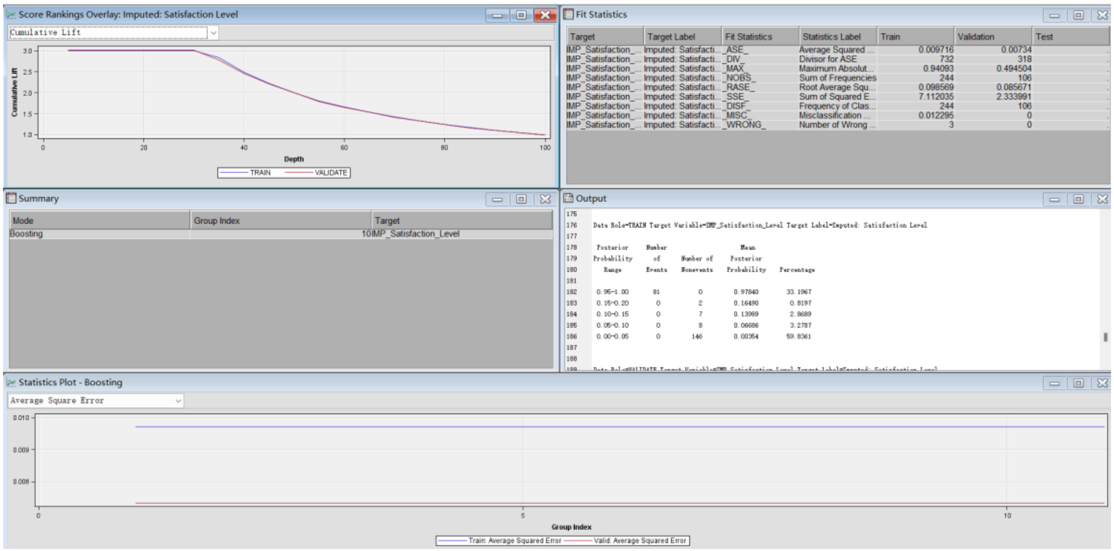
Add Start Groups and End Groups, and under the Start Groups node, set the mode to Boosting.





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

The result of End Groups in Boosting:



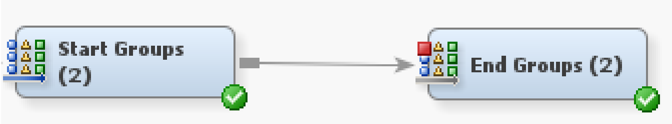
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
IMP_Satisfaction_...	Imputed: Satisfacti...	_ASE_	Average Squared ...	0.009716	0.00734	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_DIV_	Divisor for ASE	732	318	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_MAX_	Maximum Absolut...	0.94093	0.494504	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_NOBS_	Sum of Frequencies	244	106	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_RASE_	Root Average Squ...	0.098569	0.085671	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_SSE_	Sum of Squared E...	7.112035	2.333991	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_DISF_	Frequency of Clas...	244	106	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_MISC_	Misclassification ...	0.012295	0	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_WRONG_	Number of Wrong ...	3	0	.

Mode	Group Index	Target
Boosting	10	IMP_Satisfaction_Level

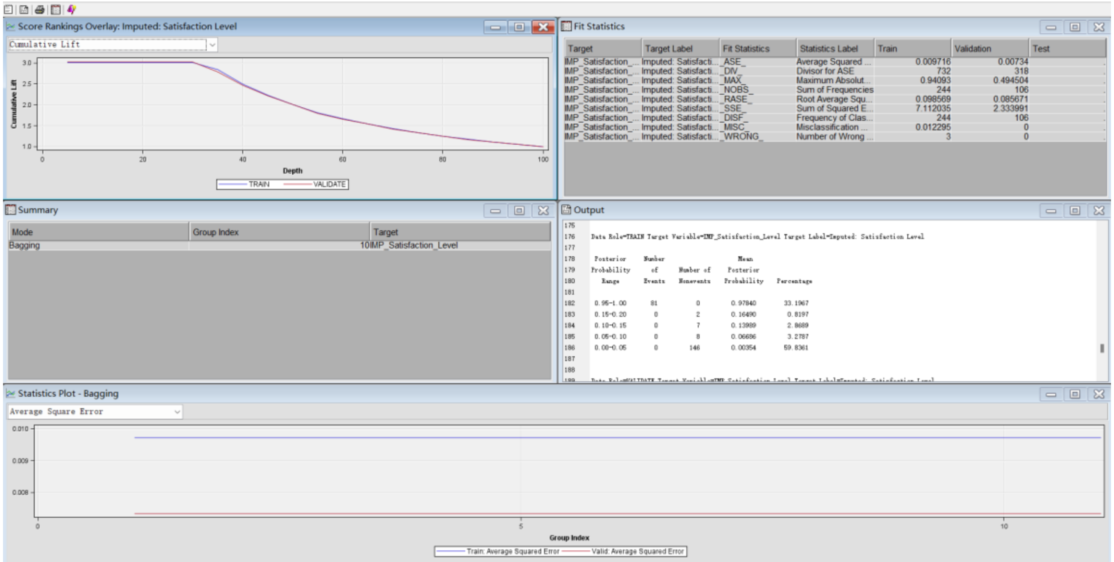
```
Data Role=TRAIN Target Variable=IMP_Satisfaction_Level Target Label=Imputed: Satisfaction Level

 Posterior     Number                   Mean
Probability      of      Number of    Posterior
   Range       Events    Nonevents   Probability    Percentage

 0.95-1.00       81           0        0.97840        33.1967
 0.15-0.20        0           2        0.16490         0.8197
 0.10-0.15        0           7        0.13989         2.8689
 0.05-0.10        0           8        0.06686         3.2787
 0.00-0.05        0         146        0.00354        59.8361


Data Role=VALIDATE Target Variable=IMP_Satisfaction_Level Target Label=Imputed: Satisfaction Level
```

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:



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
IMP_Satisfaction_...	Imputed: Satisfacti...	_ASE_	Average Squared ...	0.009716	0.00734	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_DIV_	Divisor for ASE	732	318	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_MAX_	Maximum Absolut...	0.94093	0.494504	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_NOBS_	Sum of Frequencies	244	106	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_RASE_	Root Average Squ...	0.098569	0.085671	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_SSE_	Sum of Squared E...	7.112035	2.333991	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_DISF_	Frequency of Clas...	244	106	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_MISC_	Misclassification ...	0.012295	0	.
IMP_Satisfaction_...	Imputed: Satisfacti...	_WRONG_	Number of Wrong ...	3	0	.

Mode	Group Index	Target
Bagging	10	IMP_Satisfaction_Level

```
Data Role=TRAIN Target Variable=IMP_Satisfaction_Level Target Label=Imputed: Satisfaction Level

 Posterior     Number                   Mean
Probability      of      Number of    Posterior
   Range       Events    Nonevents   Probability    Percentage

 0.95-1.00       81           0        0.97840        33.1967
 0.15-0.20        0           2        0.16490         0.8197
 0.10-0.15        0           7        0.13989         2.8689
 0.05-0.10        0           8        0.06686         3.2787
 0.00-0.05        0         146        0.00354        59.8361


Data Role=VALIDATE Target Variable=IMP_Satisfaction_Level Target Label=Imputed: Satisfaction Level
```