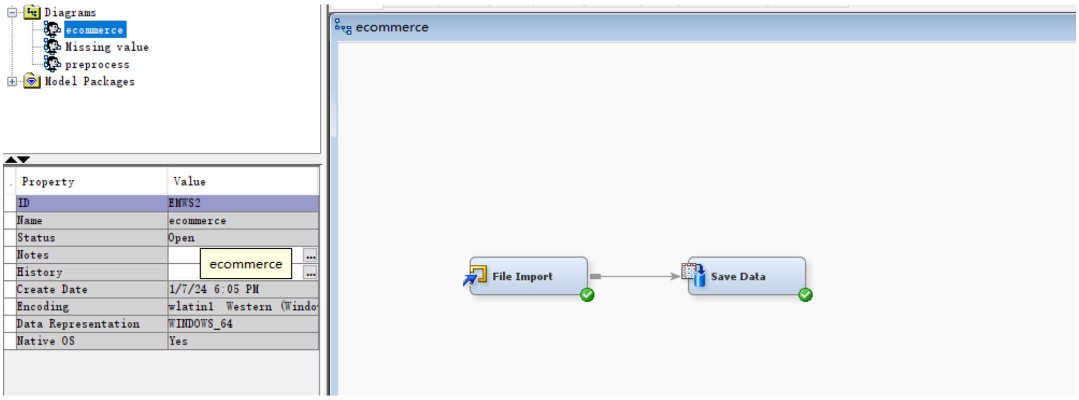
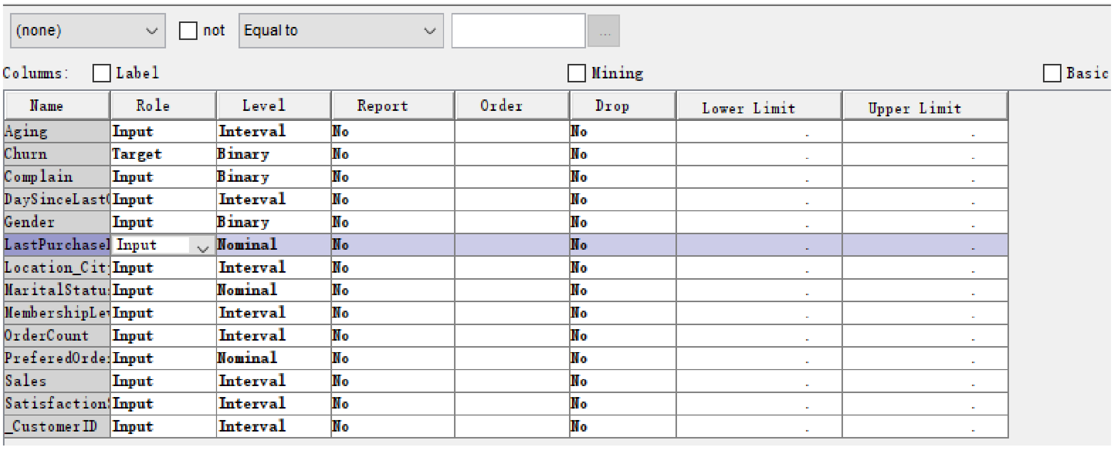
1.Import and statistic

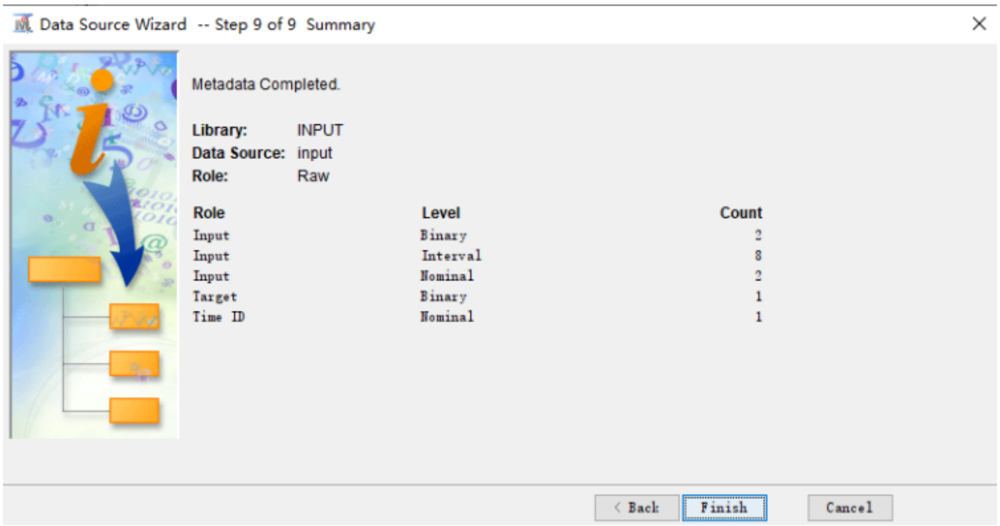
- ## Create Diagram



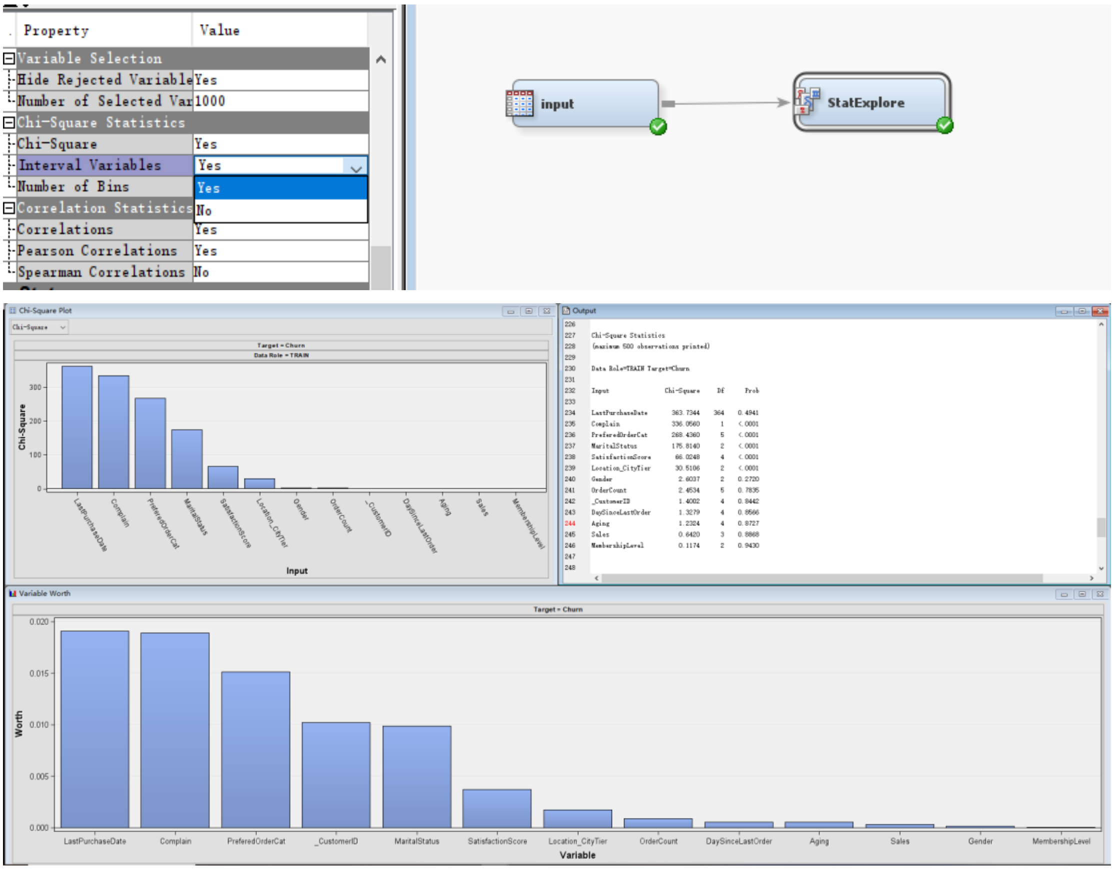
- ## Create Data Source



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Aging	Input	Interval	No		No	.	.
Churn	Target	Binary	No		No	.	.
Complain	Input	Binary	No		No	.	.
DaySinceLast(	Input	Interval	No		No	.	.
Gender	Input	Binary	No		No	.	.
LastPurchasel	Input	Nominal	No		No	.	.
Location_Cit	Input	Interval	No		No	.	.
MaritalStatu	Input	Nominal	No		No	.	.
MembershipLe	Input	Interval	No		No	.	.
OrderCount	Input	Interval	No		No	.	.
PreferedOrde	Input	Nominal	No		No	.	.
Sales	Input	Interval	No		No	.	.
Satisfaction	Input	Interval	No		No	.	.
_CustomerID	Input	Interval	No		No	.	.



- **Statistic**



- **Find Missing Values**

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Aging	INPUT	38.53486	2.880295	5121	0	33	39	43	-0.00816	-1.21873
DaySinceLastOrder	INPUT	4.497429	3.669057	4861	260	0	3	46	1.226157	4.353751
Location_CityTier	INPUT	1.652802	0.914796	5121	0	1	1	3	0.739961	-1.39487
MembershipLevel	INPUT	1.983792	0.738407	5121	0	1	2	3	0.025605	-1.16533
OrderCount	INPUT	2.970323	2.964018	4886	235	1	2	16	2.200312	4.714572
Sales	INPUT	145.5681	65.83512	5121	0	54	118	250	0.320067	-1.26299
SatisfactionScore	INPUT	3.06776	1.376809	5121	0	1	3	5	-0.13532	-1.12023
_CustomerID	INPUT	52678.38	1549.776	5121	0	50001	52669	55369	0.008121	-1.19694

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Complain	INPUT	2	0	0	71.78	1	28.22
TRAIN	Gender	INPUT	3	1	Male	60.40	Female	39.58
TRAIN	MaritalStatus	INPUT	3	0	Married	51.92	Single	32.45
TRAIN	PreferedOrderCat	INPUT	6	0	Laptop & Accessory	37.90	Mobile Phone	22.93
TRAIN	Churn	TARGET	2	0	0	82.56	1	17.44

2. Handing missing values

There are three missing values, which will be processed separately:

- **“Gender”**

The missing value itself does not have distribution characteristics and has a relatively small relationship with other columns. The value itself does not have sensitive meaning. It can be judged as missing MCAR and the mode of the column is used to supplement it.

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Churn	0	3	1	Male	59.91	Female	40.07
Churn	1	2	0	Male	62.71	Female	37.29
OVERALL		3	1	Male	60.40	Female	39.58

• "OrderCount" & "DaySinceLastOrder"

(1) Because there is a lot of missing data, we first find out the pattern from the distribution of the missing data: they are all randomly distributed, so it maybe MCAR.
(2) By exploring their relationship with other columns, especially the relationship with the target column, it was found that the missing values are closely related to the target column. So it is MAR. Considering that this data set is large enough, slight differences will not affect the results, so "churn=0" is used. The average of "churn=0" replaces their values.

Data Role=TRAIN Variable=OrderCount

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Churn	0	2	193	4035	1	16	2.984387	2.965761	2.182136	4.616616	INPUT	OrderCount
Churn	1	2	42	851	1	16	2.903643	2.956568	2.292968	5.236085	INPUT	OrderCount
OVERALL		2	235	4886	1	16	2.970323	2.964018	2.200312	4.714572	INPUT	OrderCount

Data Role=TRAIN Variable=DaySinceLastOrder

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Churn	0	3	218	4010	0	46	4.531671	3.690728	1.276938	5.04428	INPUT	DaySinceLastOrder
Churn	1	3	42	851	0	18	4.336075	3.562861	0.954333	0.521333	INPUT	DaySinceLastOrder
OVERALL		3	260	4861	0	46	4.497429	3.669057	1.226157	4.353751	INPUT	DaySinceLastOrder

- **Replacement**

Gender= "male"

When "churn=0" the mean of "OrderCount" value is 2.984387, it is close to the mean of the overall dataset and different from the mean "2.903643" so we use "OrderCount=mean" to replace the missing values.

When "churn=0" the mean of "DaySinceLastOrder" value is 4.531671, I should use "DaySinceLastOrder=5" to replace the missing values, but because the value "5" is very different from "4.531671" and may cause changes in the whole dataset. So I use "DaySinceLastOrder=3", the median, to replace the missing values.

Because the missing values involved in this data set are MCAR and MAR, simple methods can be used to supplement them. If MNAR is involved, more complex methods may be needed to deal with missing values to avoid introducing bias.

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Churn	0		4228	N		0
Churn	1		893	N		1
Churn	_UNKNOWN_	_DEFAULT_	.	N		.
Complain	0		3676	N		0
Complain	1		1445	N		1
Complain	_UNKNOWN_	_DEFAULT_	.	N		.
Gender	Male		3093	C	Male	.
Gender	Female		2027	C	Female	.
Gender		Male	1	C		.
Gender	_UNKNOWN_	_DEFAULT_	.	C		.
MaritalStatus	Married		2659	C	Married	.
MaritalStatus	Single		1662	C	Single	.
MaritalStatus	Divorced		800	C	Divorced	.
MaritalStatus	_UNKNOWN_	_DEFAULT_	.	C		.
PreferedOrderCat	Laptop & Accessory		1941	C	Laptop & Accessory	.
PreferedOrderCat	Mobile Phone		1174	C	Mobile Phone	.
PreferedOrderCat	Mobile		808	C	Mobile	.
PreferedOrderCat	Fashion		664	C	Fashion	.
PreferedOrderCat	Grocery		281	C	Grocery	.
PreferedOrderCat	Others		253	C	Others	.

OK Cancel

Replacement Counts

Obs	Variable	Role	Label	Train
1	Gender	INPUT	Gender	1

Variables - Impt

(none) not Equal to ...

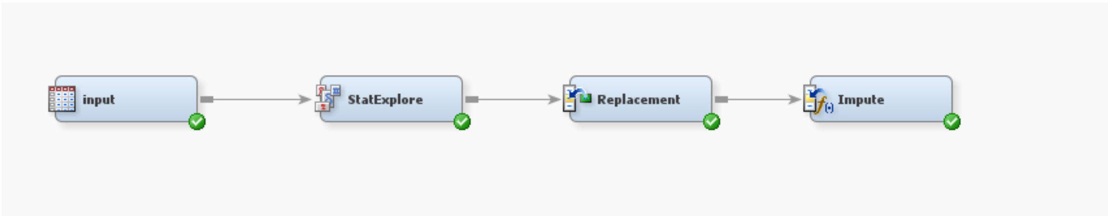
Columns: Label Mining

Name	Use	Method	Use Tree	Role	Level
Aging	Default	Default	Default	Input	Interval
Churn	Default	Default	Default	Target	Binary
Complain	Default	Default	Default	Input	Binary
DaySinceLast(	Default	Median	Default	Input	Interval
Gender	Default	Default	Default	Input	Binary
Location_Cit	Default	Default	Default	Input	Interval
MaritalStatu	Default	Default	Default	Input	Nominal
MembershipLe	Default	Default	Default	Input	Interval
OrderCount	Default	Mean	Default	Input	Interval
PreferedOrde	Default	Default	Default	Input	Nominal
Sales	Default	Default	Default	Input	Interval
Satisfaction	Default	Default	Default	Input	Interval
_CustomerID	Default	Default	Default	Input	Interval

Imputation Summary
Number Of Observations

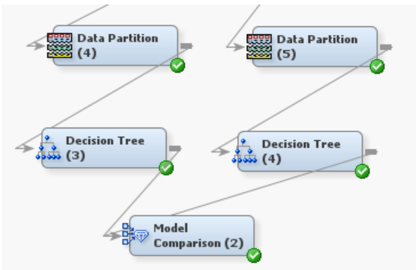
Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
DaySinceLastOrder	MEDIAN	IMP_DaySinceLastOrder	3.00000	INPUT	INTERVAL	DaySinceLastOrder	260
OrderCount	MEAN	IMP_OrderCount	2.97032	INPUT	INTERVAL	OrderCount	235

- **Workflow**



3. Decision Tree

- **Data Partition　　55：45 vs 70：30**



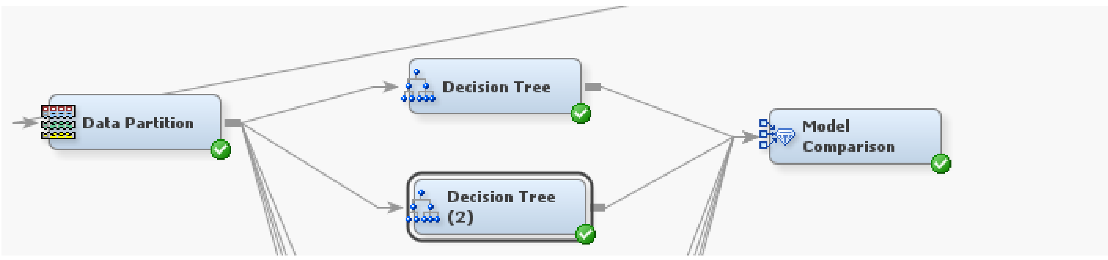
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree3	Decision Tree (3)	0.14831	0.11751	0.15169	0.11978
	Tree4	Decision Tree (4)	0.15085	0.11733	0.15211	0.12134

55：45 have a better performance.

- **Attribute selection**

Input	Chi-Square	Df	Prob
LastPurchaseDate	363.7344	364	0.4941
Complain	336.0560	1	<.0001
PreferedOrderCat	268.4360	5	<.0001
MaritalStatus	175.8140	2	<.0001
SatisfactionScore	66.0248	4	<.0001
Location_CityTier	30.5106	2	<.0001
Gender	2.6037	2	0.2720
OrderCount	2.4534	5	0.7835
_CustomerID	1.4002	4	0.8442
DaySinceLastOrder	1.3279	4	0.8566
Aging	1.2324	4	0.8727
Sales	0.6420	3	0.8868
MembershipLevel	0.1174	2	0.9430



Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Decision Tree	0.14961	0.11920	0.15382	0.12078
	Tree2	Decision Tree (2)	0.15568	0.12851	0.16448	0.12732

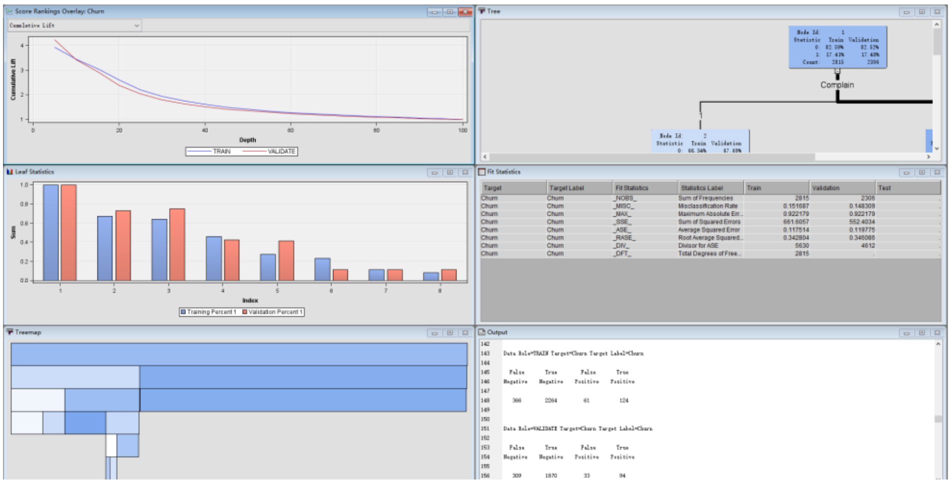
(1)Input or Drop “LastPurchaseDate”: According to the comparison of performance, I decide to input. So I choose “LastPurchaseDate”, “Complain”, “PreferedOrderCat”, “MaritalStatus”, “SatisfactionScore”, and “Location_CityTier” as input.
(2)Compare the variables I selected with the model default variables
Result: model default variables have better performance.

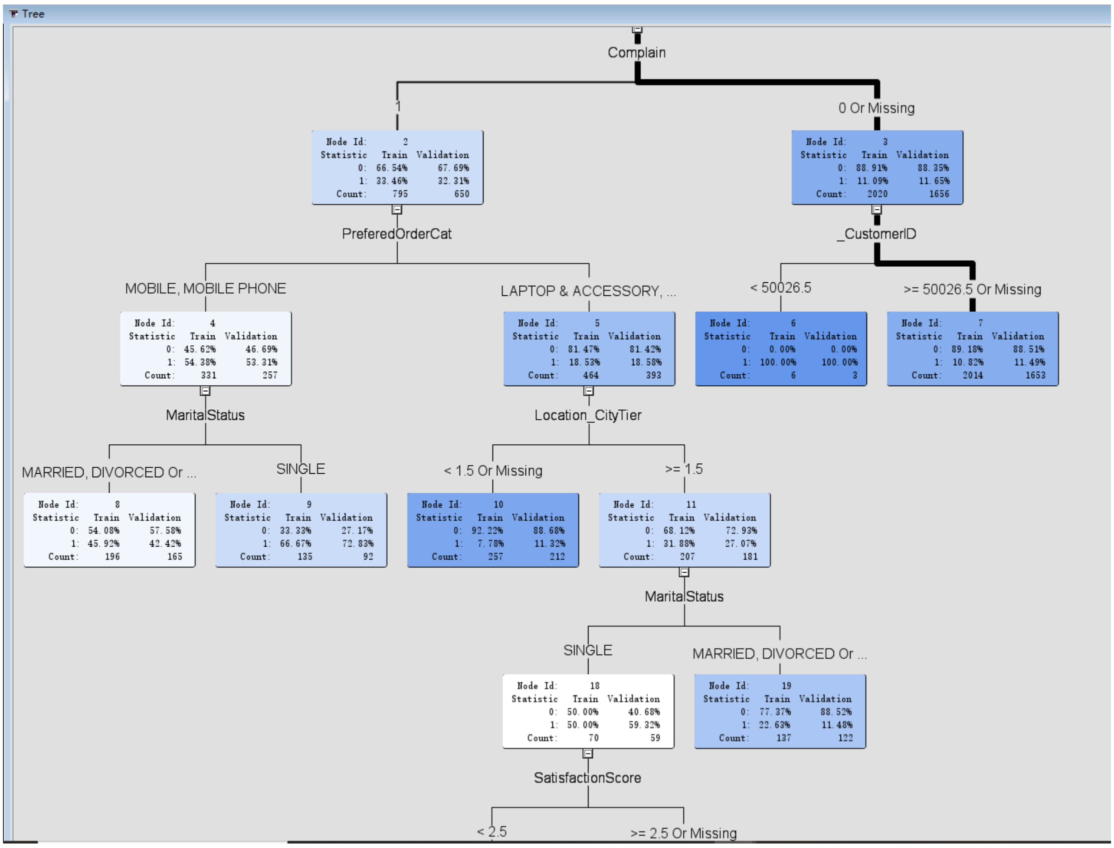
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Decision Tree (2)	0.14831	0.11751	0.15169	0.11978
	Tree	Decision Tree	0.14961	0.11920	0.15382	0.12078

- **Performance**

There is the final performance of Decision Tree.

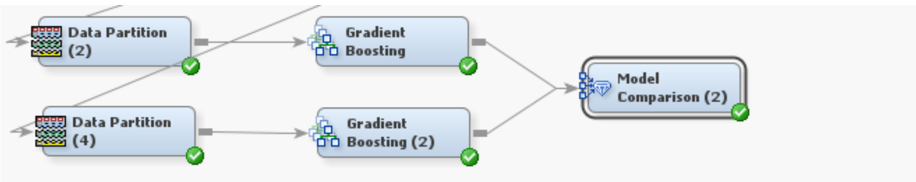
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Decision Tree (2)	0.14831	0.11751	0.15169	0.11978





4. Gradient Boosting

- **Data Partition　　55:45 vs 70:30**

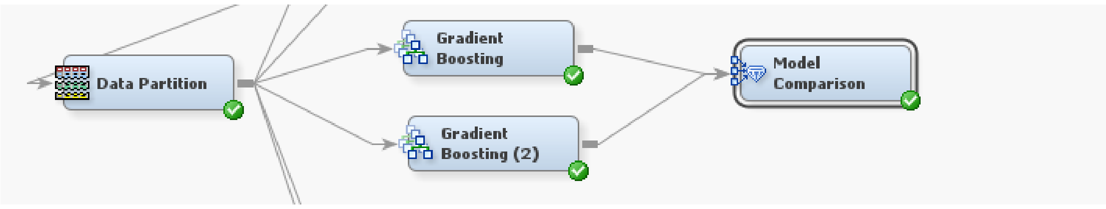



```
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)
```

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Boost	Gradient Boosting	0.15632	0.09763	0.13666	0.12023
	Boost2	Gradient Boosting (2)	0.16173	0.10268	0.14218	0.12847

55：45 have a better performance.

- **Attribute selection**



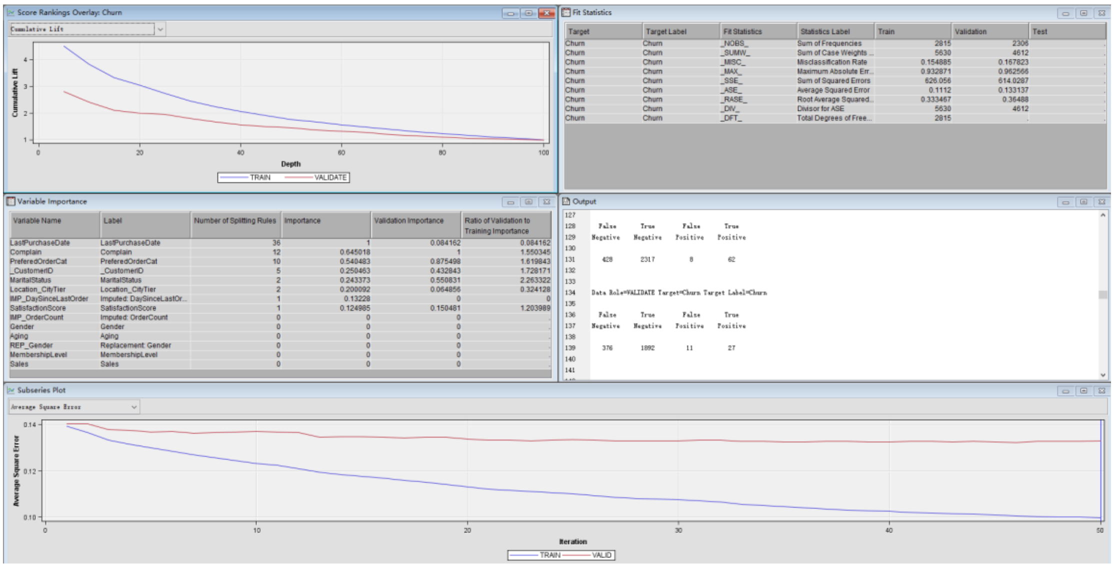
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Boost	Gradient Boosting	0.16782	0.11120	0.15488	0.13314
	Boost2	Gradient Boosting (2)	0.16782	0.11120	0.15488	0.13314

Result: using the attributes I choose will give me better model performance.
Input: "LastPurchaseDate", "Complain", "PreferedOrderCat", "MaritalStatus", "SatisfactionScore", and "Location_CityTier"

- **Performance**

There is the final performance of Gradient Boosting.



Data Role=TRAIN Target=Churn Target Label=Churn

False Negative	True Negative	False Positive	True Positive
428	2317	8	62

Data Role=VALIDATE Target=Churn Target Label=Churn

False Negative	True Negative	False Positive	True Positive
376	1892	11	27

5. HP Forest

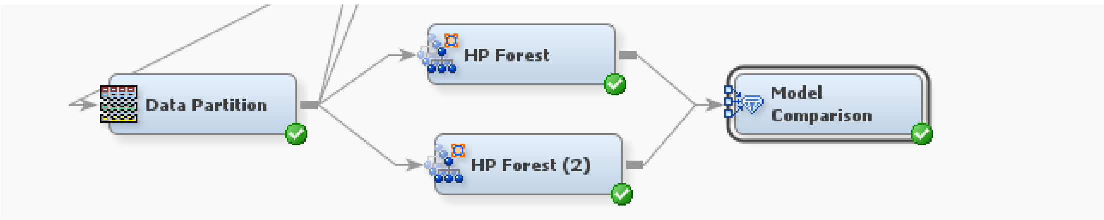
- **Data Partition 55:45 vs 70:30**

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPDMForest2	HP Forest (2)	0.15241	0.11380	0.15560	0.11327
	HPDMForest	HP Forest	0.16817	0.11796	0.16713	0.11912

55：45 have a better performance.

- **Attribute selection**



Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPDMForest2	HP Forest (2)	0.16175	0.11524	0.15950	0.11790
	HPDMForest	HP Forest	0.17086	0.11847	0.16696	0.12054

Result: model default variables have better performance.

Input Variables

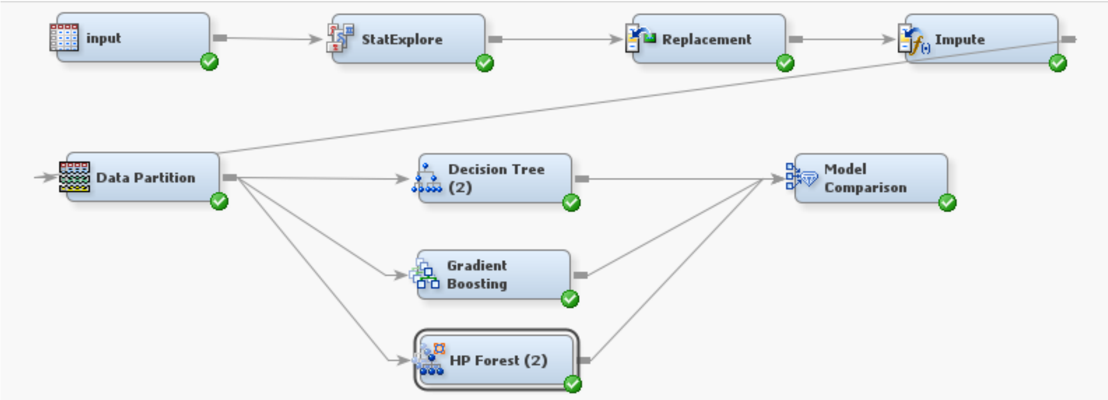
Name	Length	Role	Type	RawType	Format Name
Aging	8	Input	Interval	Num	
IMP_DaySinceLastOrder	8	Input	Interval	Num	
IMP_OrderCount	8	Input	Interval	Num	
Location_CityTier	8	Input	Interval	Num	
MembershipLevel	8	Input	Interval	Num	
Sales	8	Input	Interval	Num	
SatisfactionScore	8	Input	Interval	Num	
_CustomerID	8	Input	Interval	Num	
Complain	8	Input	Classification	Num	
REP_Gender	6	Input	Classification	Character	
MaritalStatus	8	Input	Classification	Character	
PreferedOrderCat	18	Input	Classification	Character	

- **Performance**

There is the final performance of HP Forest.

```
Event Classification Table

Data Role=TRAIN Target=Churn Target Label=Churn

  False      True      False      True
Negative   Negative   Positive   Positive

   429       2305        20         61


Data Role=VALIDATE Target=Churn Target Label=Churn

  False      True      False      True
Negative   Negative   Positive   Positive

   362       1892        11         41
```

- **Workflow**



6. Model Comparison

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Decision Tree (2)	0.14831	0.11751	0.15169	0.11978
	HPDMForest2	HP Forest (2)	0.16175	0.11524	0.15950	0.11790
	Boost	Gradient Boosting	0.16782	0.11120	0.15488	0.13314

The best model is Decision Tree.

Metrics	Decision Tree 55:45	
	Train	Validate
Precision	0.670	0.740
Recall	0.253	0.233
F1-Score	0.367	0.355
Accuracy	0.848	0.852
Specificity	0.974	0.983

Precision measures the model's ability to correctly predict the positive class. The higher precision in my model's validation dataset compared to the training dataset indicates that the model has good generalization ability.
Recall reflects the model's ability to correctly identify all positive class instances. The recall rates for both datasets are relatively low, which means my model has missed many actual positive class predictions.
The F1 score is the harmonic mean of precision and recall, providing a more comprehensive performance metric. My model's F1 scores are not high on both datasets, indicating that there needs to be a better balance between precision and recall.
Accuracy represents the proportion of correct predictions (both positive and negative classes) made by the model. My model has a relatively high accuracy on both the training and validation sets, which could be because the number of negative classes (non-churn customers) far exceeds the positive classes (churn customers).
Specificity measures the model's ability to correctly identify negative classes. Here, the model performs very well, indicating that my model is very accurate in predicting non-churn customers.

- **Performance discussion**

My model's ability to predict non-churning customers far exceeds its ability to predict churning customers, indicating that non-churning customers likely exhibit some consistent characteristics, while predicting churn may involve a broader spectrum of factors. To improve predictions, a more complex analysis incorporating additional attributes is needed.

In terms of ensemble methods, both Gradient Boosting and HP Forest models utilize collections of weak prediction models to enhance accuracy. These models usually excel in handling complex datasets and uncovering non-linear relationships. However, in my case, a single decision tree model outperformed these more complex ensemble models in predicting customer churn. This finding emphasizes that model complexity does not always lead to better outcomes, and sometimes the intuitiveness and interpretability of simpler models better meet business needs.

The decision tree model, with its principle of finding the most significant split points within data features, identified 'LastPurchaseDate', 'Complain', 'PreferedOrderCat', 'MaritalStatus', and others as significant predictors of customer churn. Relying on the efficiency of the decision tree in choosing branch points and the interpretability of the results, it is more suitable for providing clear action directions for businesses to reduce customer churn rates.

- **Business insights**

Based on the decision tree model visual, here are some specific business recommendations:

- Address Complaints Promptly: The root node of the tree is 'Complain', indicating that customer complaints are a primary factor in predicting churn. The business should implement a robust system to handle complaints quickly and effectively. This could include training customer service staff, creating a streamlined process for resolving issues, and follow-up with customers to ensure they are satisfied with the solutions provided.
- Review Ordering Preferences: 'PreferedOrderCat' suggests that the type of products customers are ordering affects churn. The company should analyze order patterns and possibly adjust inventory, offer promotions on popular items, or personalize marketing based on customers' buying preferences.
- Consider Demographics in Marketing: 'MaritalStatus' and 'Location_CityTier' are used as split points, showing the importance of demographic factors. Tailored marketing campaigns that consider these demographic details could be more effective. For instance, married or divorced customers might respond to different types of outreach or offers than single customers.
- Enhance Loyalty Programs: The presence of 'CustomerID' in the model could imply that individual customer engagement levels play a role. Businesses could develop or improve loyalty programs to increase engagement and reduce churn.
- Focus on Customer Experience: The inclusion of 'SatisfactionScore' points to the importance of customer satisfaction in retention. Regular surveys and feedback mechanisms to gauge satisfaction, and initiatives to improve the customer experience, would likely have a positive impact on reducing churn.
- Adjust Sales Strategies: For the 'Location_CityTier' variable, it would be beneficial to tailor sales and marketing strategies to different city tiers, acknowledging that customers in different locations may have different needs or preferences.
- Price Sensitivity: The tree splits on 'PreferedOrderCat' with categories like 'MOBILE_MOBILE PHONE' and 'LAPTOP & ACCESSORY', and on 'CustomerID' with a threshold value, indicating technology preference and potential price sensitivity. This suggests revising pricing strategies and offering targeted technology deals to retain customers.