Dataset Description

The dataset is from Kaggle, containing information about customers of an e-commerce company. There are 20 columns of data in total. The following summarizes the variable names and descriptions in the dataset.

CustomerID (Categorical - Nominal): Unique identifier assigned to each customer. This variable serves as an identifier and is not used for analytical purposes other than uniquely identifying customers.

Churn (Categorical - Binary): Indicates whether the customer has churned or not. It's a binary categorical variable with values TRUE (churned) or FALSE (not churned).

Tenure (Numerical - Interval): Represents the duration of the customer's association with the business. It is a quantitative and continuous variable, indicating the length of the customer's tenure.

PreferredLoginDevice (Categorical - Nominal): Represents the preferred device for customer login. This is a categorical variable with different device categories.

CityTier (Categorical - Ordinal): Indicates the tier of the city where the customer is located. It's an ordinal categorical variable with different city tier levels.

WarehouseToHome (Numerical - Interval): Represents the distance from the warehouse to the customer's home. It is a quantitative and continuous variable.

PreferredPaymentMode (Categorical - Nominal): Specifies the preferred mode of payment chosen by the customer. It's a categorical variable with different payment mode categories.

Gender (Categorical - Nominal): Represents the gender of the customer. It's a categorical variable with two possible values: Male or Female.

HourSpendOnApp (Numerical - Interval): Indicates the number of hours the customer spends on the mobile application. This variable is quantitative and continuous.

NumberOfDeviceRegistered (Numerical - Interval): Represents the number of devices registered by the customer. It is a quantitative and discrete variable.

PreferedOrderCat (Categorical - Nominal): Indicates the preferred category for ordering. This is a categorical variable with different order category options.

SatisfactionScore (Numerical - Interval): Reflects the satisfaction score given by the customer. This variable is quantitative and continuous.

MaritalStatus (Categorical - Nominal): Represents the marital status of the customer. It's a categorical variable with values like Single, Married, etc.

NumberOfAddress (Numerical - Interval): Indicates the number of addresses associated with the customer. It is a quantitative and discrete variable.

Complain (Categorical - Binary): Indicates whether the customer has lodged a complaint. It's a binary categorical variable with values TRUE or FALSE.

OrderAmountHikeFromlastYear (Numerical - Interval): Represents the percentage increase in order amount from the last year. This variable is quantitative and continuous.

CouponUsed (Numerical - Interval): Indicates the number of coupons used by the customer. It is a quantitative and discrete variable.

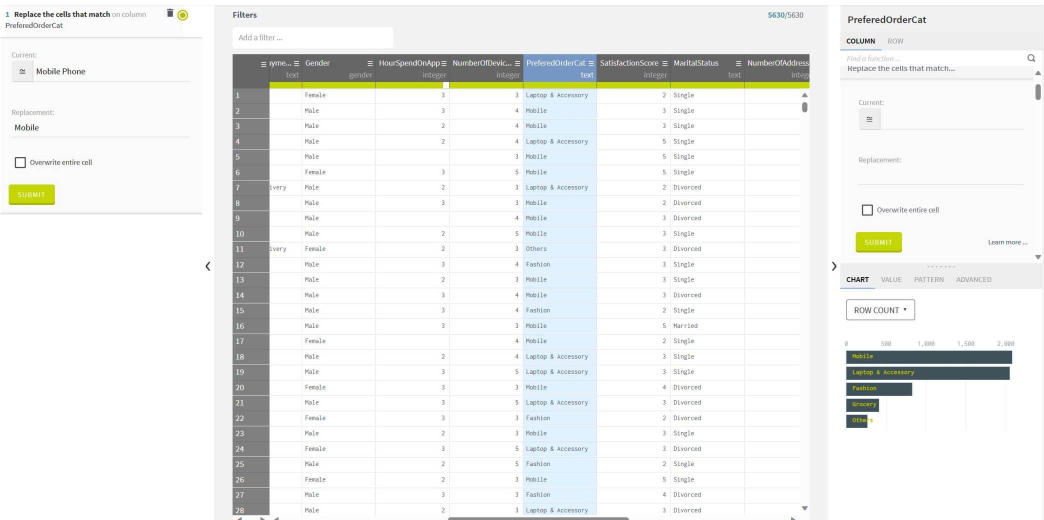
OrderCount (Numerical - Interval): Represents the count of orders placed by the customer. It is a quantitative and discrete variable.

DaySinceLastOrder (Numerical - Interval): Represents the number of days since the customer's last order. It is a quantitative and continuous variable.

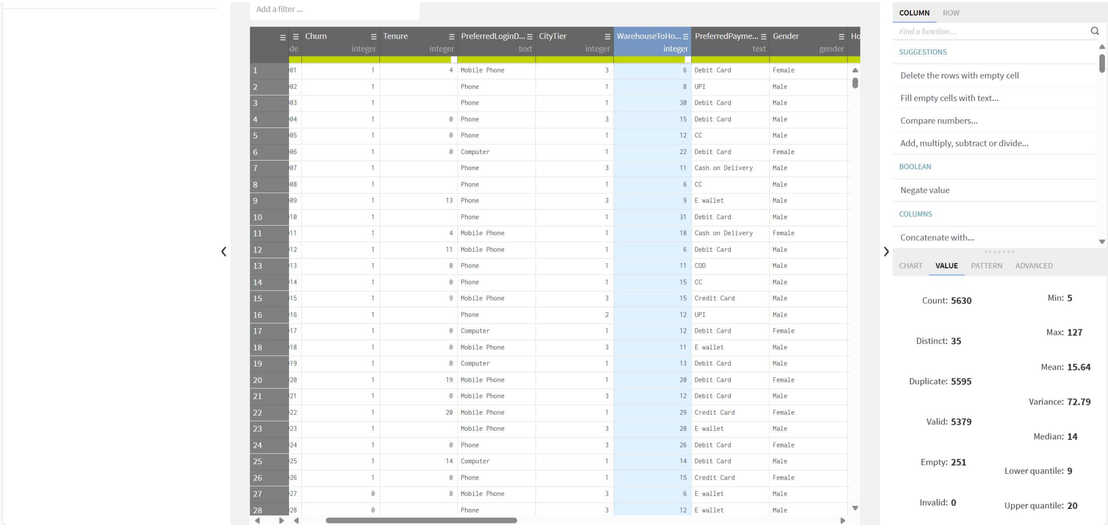
CashbackAmount (Numerical - Interval): Reflects the cashback amount received by the customer. This variable is quantitative and continuous.

1.Data Pre-processing using Talend Data Preparation

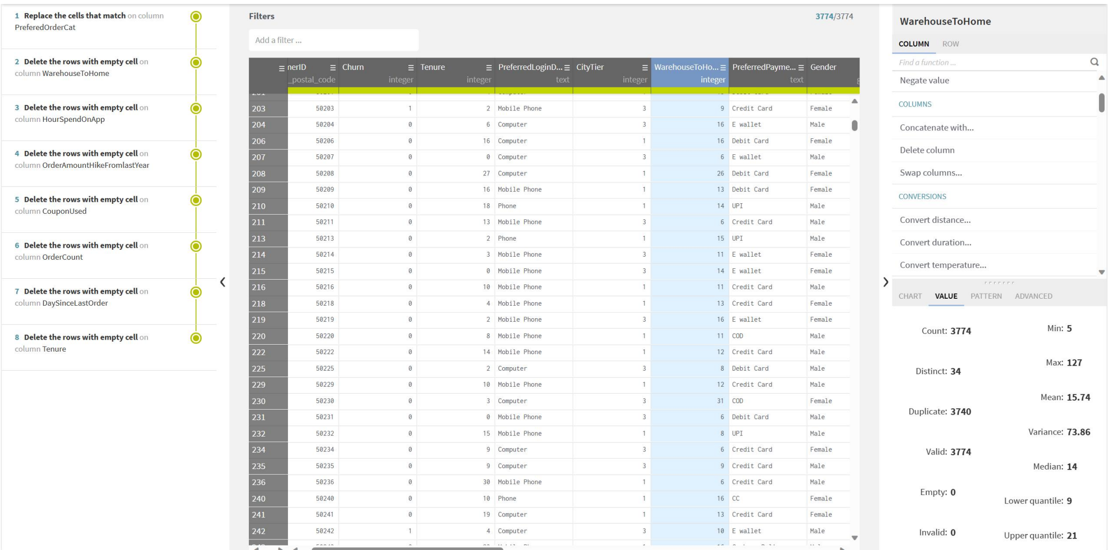
In the "PreferredOrderCat" column, "Mobile Phone" should be a subset of "Mobile". To ensure data consistency, "Mobile Phone" is changed to "Mobile".



In the "WarehouseToHome" column, there are many missing values. Use Delete the rows with empty cell to delete the missing values.

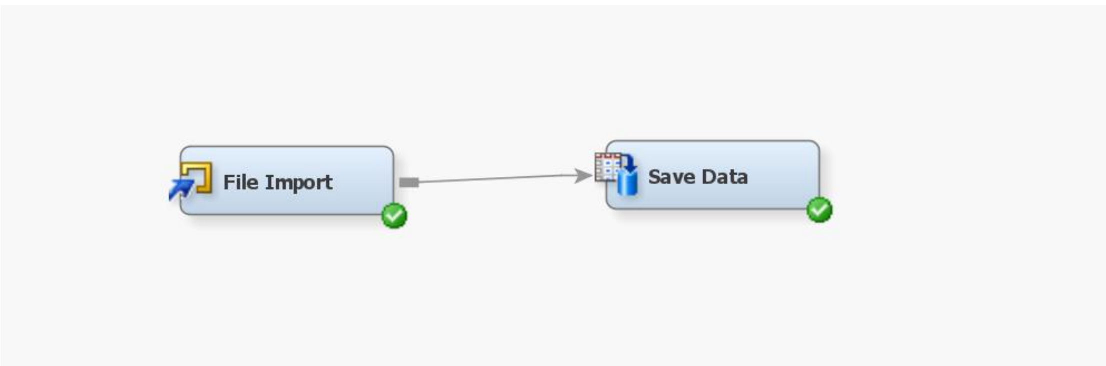


Similarly, missing values in other columns were deleted, and a total of 1856 rows of data containing missing values were deleted.



2.Data import using SAS Enterprise Miner

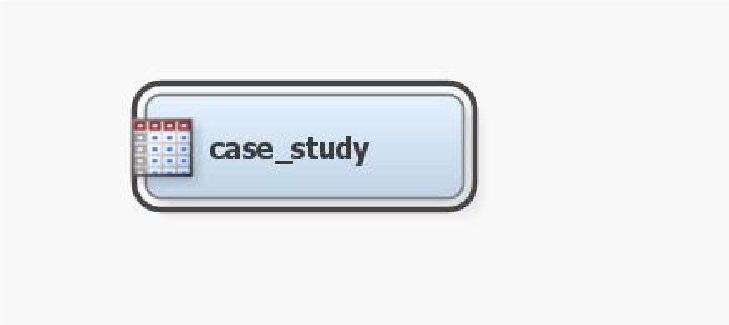
1. Import the CSV file using "File Import" node. Save it as a SAS file.



2.Creat Library and datasource

- case_study
  - Data Sources
    - case_study
  - Diagrams
    - case_study
    - Import_data
  - Model Packages

Drag the data source into a new diagram and perform operations.



Right click on case_study and select Edit Variables. In order to pay attention to customer churn, I set "Churn" as the target variable.

Variables - Ids

(none) not Equal to

Columns: Label Mining Bas

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
CashbackAn	Input	Interval	No		No	.	.
Churn	Target	Nominal	No		No	.	.
CityTier	Input	Nominal	No		No	.	.
Complain	Input	Nominal	No		No	.	.
CouponUsed	Input	Nominal	No		No	.	.
CustomerII	Input	Interval	No		No	.	.
DaySinceLa	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
HourSpend(	Input	Nominal	No		No	.	.
MaritalSta	Input	Nominal	No		No	.	.
NumberOfAd	Input	Nominal	No		No	.	.
NumberOfDe	Input	Nominal	No		No	.	.
OrderAmour	Input	Nominal	No		No	.	.
OrderCount	Input	Nominal	No		No	.	.
PreferedOr	Input	Nominal	No		No	.	.
PreferredI	Input	Nominal	No		No	.	.
PreferredF	Input	Nominal	No		No	.	.
Satisfacti	Input	Nominal	No		No	.	.
Tenure	Input	Interval	No		No	.	.
WarehouseT	Input	Interval	No		No	.	.

2.Decision Tree Modelling using SAS Enterprise Miner

Create a Data Partition node and divide the data. 70% is used for train data and 30% is used for validation data.

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	07/01/24 09:34
Run ID	3245f697-cb1d-0d4a-9563-c6b8ff0ed8
Last Error	
Last Status	Complete
Last Run Time	07/01/24 09:35
Run Duration	0 Hr. 0 Min. 1.97 Sec.
Grid Host	
User-Added Node	No

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	5
INPUT	NOMINAL	14
TARGET	NOMINAL	1

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS2.Ids_DATA	3774
TRAIN	EMWS2.Part_TRAIN	2641
VALIDATE	EMWS2.Part_VALIDATE	1133

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


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

Summary Statistics for Class Targets

Data=DATA

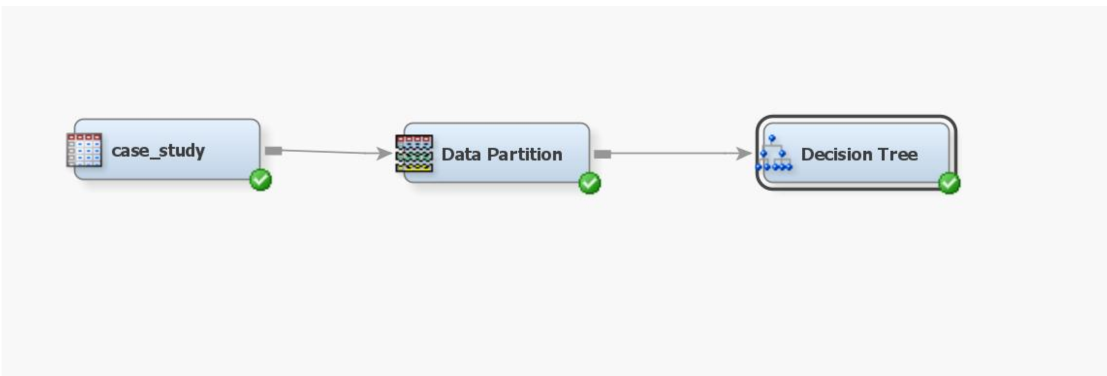
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	3143	83.2803	Churn
Churn	1	1	631	16.7197	Churn

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	2199	83.2639	Churn
Churn	1	1	442	16.7361	Churn

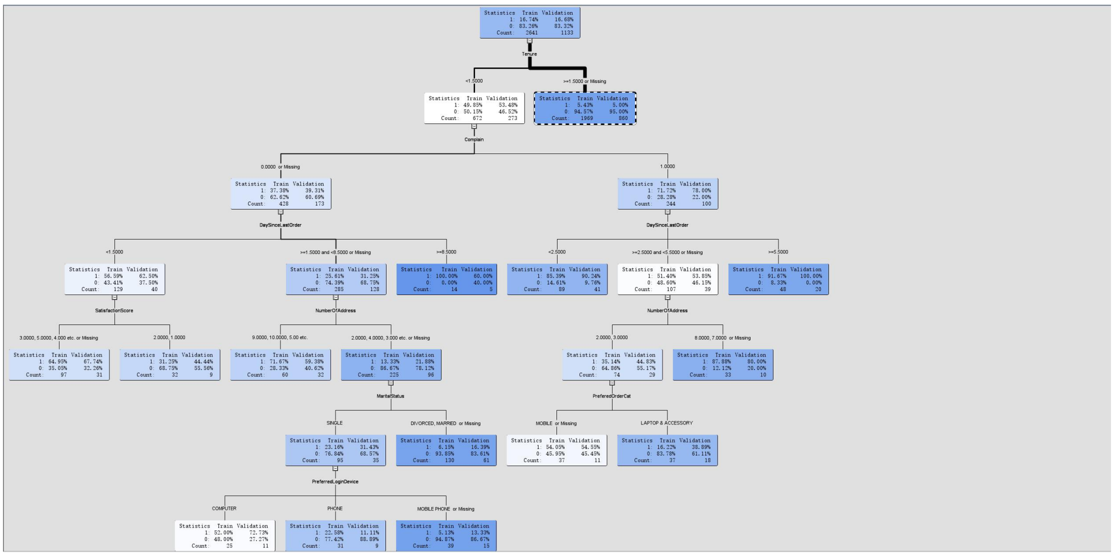
Data=VALIDATE

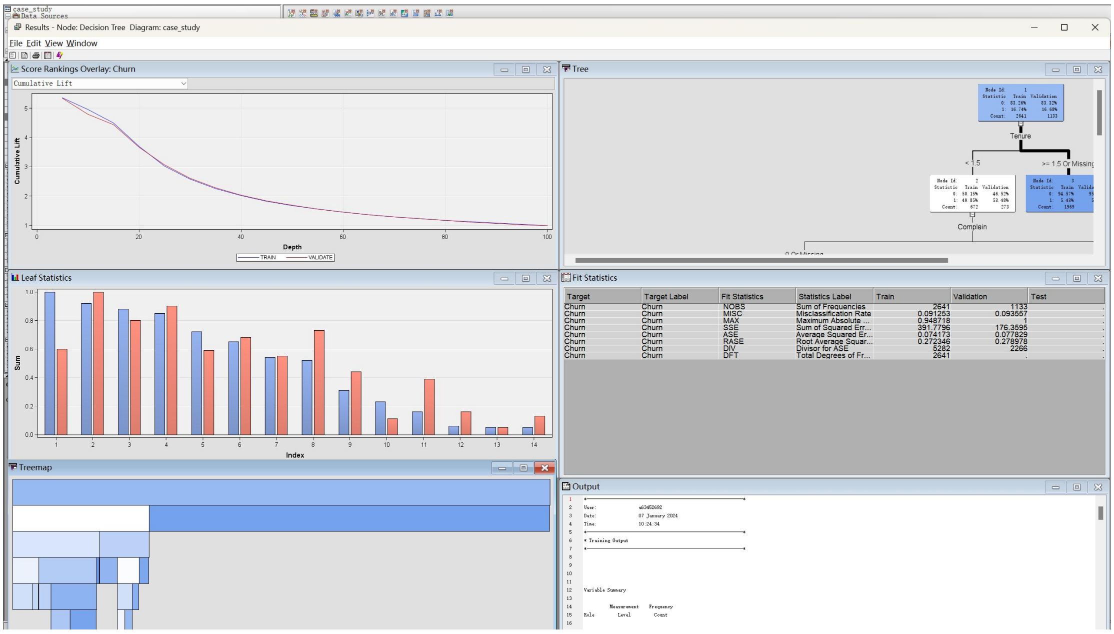
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	944	83.3186	Churn
Churn	1	1	189	16.6814	Churn

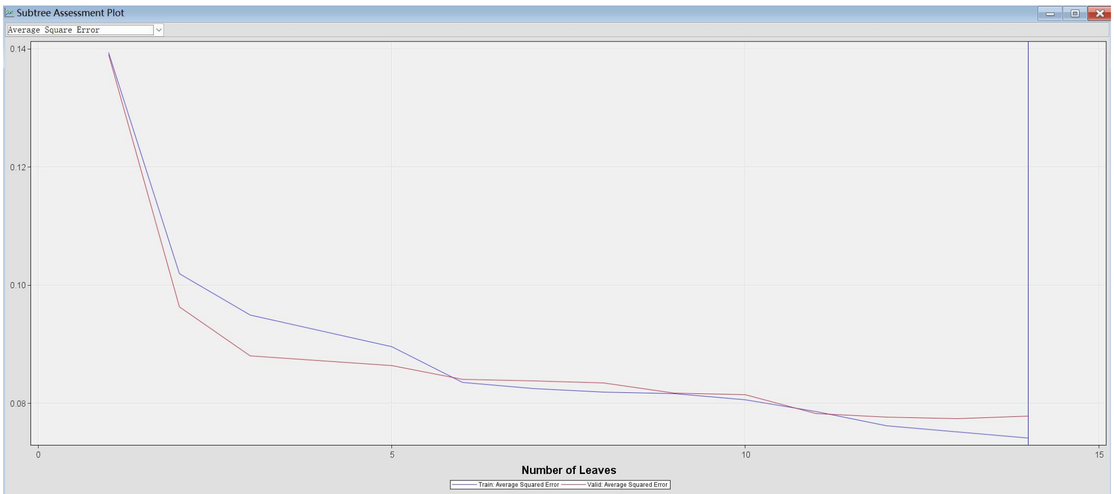


. Property	Value
General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
⊟Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
⊟Node	3
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	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
⊟Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345

. Property	Value
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
⊟Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	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
⊟Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
⊟Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
⊟P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
⊟Output Variables	
Leaf Variable	Yes
⊟Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk







Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	NOBS	Sum of Frequencies	2641	1133	.
Churn	Churn	MISC	Misclassification Rate	0.091253	0.093557	.
Churn	Churn	MAX	Maximum Absolute Error	0.948718	1	.
Churn	Churn	SSE	Sum of Squared Errors	391.7796	176.3595	.
Churn	Churn	ASE	Average Squared Error	0.074173	0.077829	.
Churn	Churn	RASE	Root Average Squared Error	0.272346	0.278978	.
Churn	Churn	DIV	Divisor for ASE	5282	2266	.
Churn	Churn	DFT	Total Degrees of Freedom	2641	.	.

Based on Fit Statistics, misclassification rate is 0.0912 for training dataset and 0.0935 for validation dataset.

Variable Importance

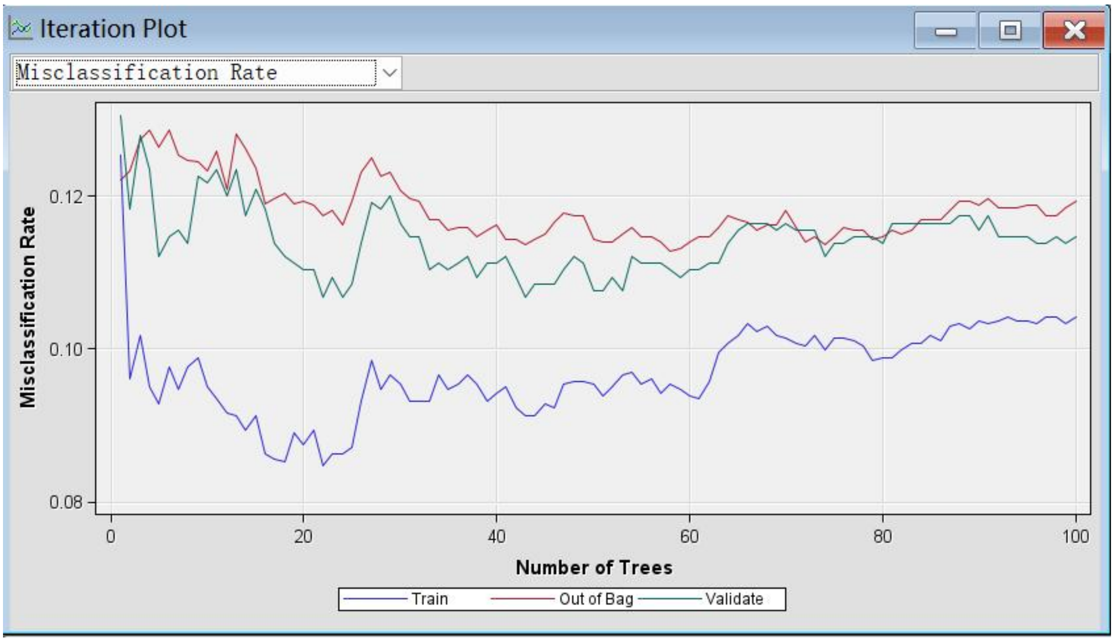
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Tenure	Tenure	1	1.0000	1.0000	1.0000
NumberOfAddress	NumberOfAddress	2	0.4768	0.2576	0.5403
DaySinceLastOrder	DaySinceLastOrder	2	0.4738	0.3496	0.7380
Complain	Complain	1	0.4306	0.4408	1.0237
PreferredLoginDevice	PreferredLoginDevice	1	0.1841	0.2175	1.1817
SatisfactionScore	SatisfactionScore	1	0.1663	0.0836	0.5026
PreferedOrderCat	PreferedOrderCat	1	0.1637	0.0000	0.0000
MaritalStatus	MaritalStatus	1	0.1267	0.0819	0.6463
WarehouseToHome	WarehouseToHome	0	0.0000	0.0000	.
HourSpendOnApp	HourSpendOnApp	0	0.0000	0.0000	.
Gender	Gender	0	0.0000	0.0000	.
CouponUsed	CouponUsed	0	0.0000	0.0000	.
CashbackAmount	CashbackAmount	0	0.0000	0.0000	.
OrderCount	OrderCount	0	0.0000	0.0000	.
NumberOfDeviceRegistered	NumberOfDeviceRegistered	0	0.0000	0.0000	.
OrderAmountHikeFromlastYear	OrderAmountHikeFromlastYear	0	0.0000	0.0000	.
CustomerID	CustomerID	0	0.0000	0.0000	.
CityTier	CityTier	0	0.0000	0.0000	.
PreferredPaymentMode	PreferredPaymentMode	0	0.0000	0.0000	.

The Variable Importance Plot displays the importance of each predictor variable in the model. Only 8 out of 18 input variables are important to the pruned decision tree model.

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

3.1 Using the Random Forest algorithm as a Bagging

Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	50
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs in Node	1.0E-5
Smallest Number of Obs in Node	1
Split Size	.
Use as Modeling Node	Yes
Score	
Variable Selection	Yes
Variable Importance Method	Loss Reduction
Number of Variables to Consider	25
Cutoff Fraction	0.01
Status	
Create Time	07/01/24 10:14
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

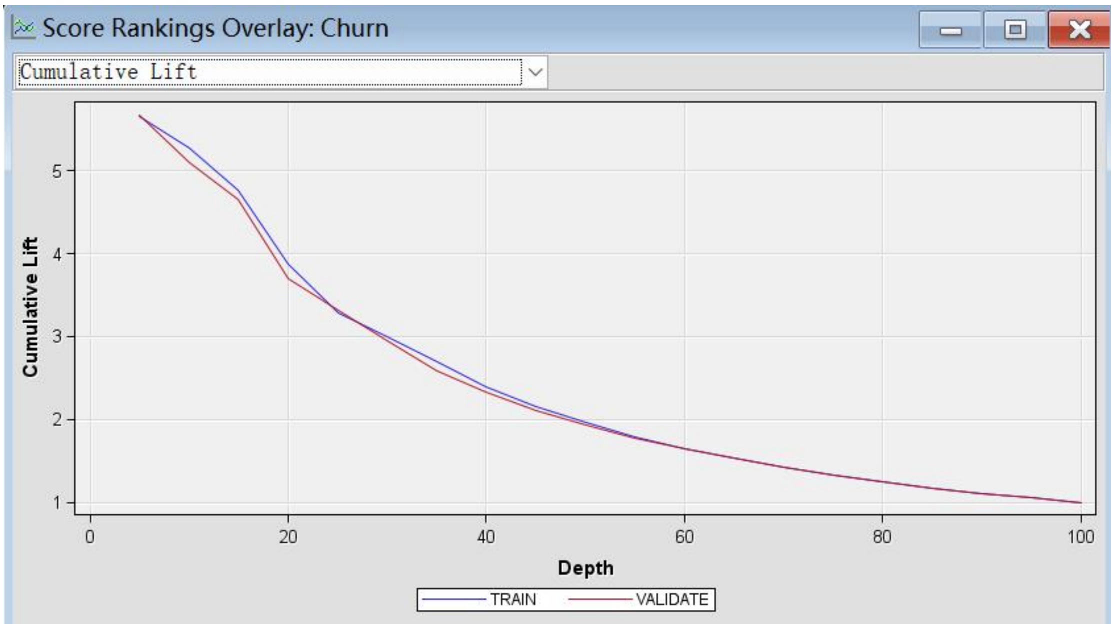


Based on Iteration Plot, misclassification rate plateaued out when number of trees reaches 25.

Fit Statistics

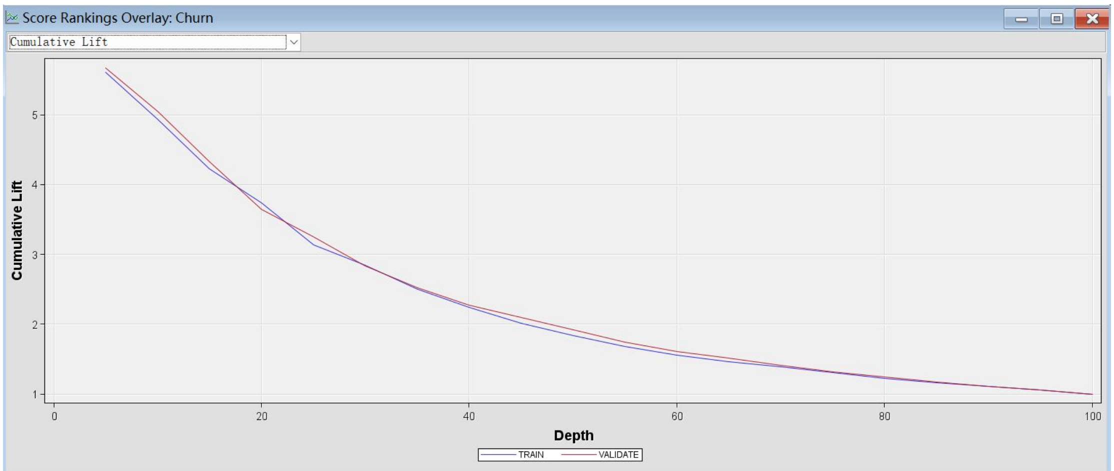
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average Squared Error	0.07725	0.082004	.
Churn	Churn	DIV	Divisor for ASE	5282	2266	.
Churn	Churn	MAX	Maximum Absolute Error	0.951645	0.952092	.
Churn	Churn	NOBS	Sum of Frequencies	2641	1133	.
Churn	Churn	RASE	Root Average Squared Error	0.277938	0.286363	.
Churn	Churn	SSE	Sum of Squared Errors	408.0329	185.821	.
Churn	Churn	DISF	Frequency of Classified Cases	2641	1133	.
Churn	Churn	MISC	Misclassification Rate	0.104127	0.11474	.
Churn	Churn	WRONG	Number of Wrong Classifications	275	130	.

Based on Fit Statistics, misclassification rate is 0.1041 for training dataset and 0.1147 for validation dataset.



3.2 Gradient Boosting

. Property	Value
General	
Node ID	Boost
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
⊟Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
⊟Node	
Leaf Fraction	0.1
Number of Surrogate Rules	0
Split Size	.
⊟Split Search	
Exhaustive	5000
Node Sample	20000
⊟Subtree	
Assessment Measure	Decision
Score	
Subseries	Best Assessment Value
Number of Iterations	1
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Importance	No
Number Single Var Importance	5
Status	
Create Time	07/01/24 10:21
Run ID	



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Tenure	Tenure	22	1	1	1
Complain	Complain	14	0.365652	0.377918	1.033547
NumberOfAddress	NumberOfAddress	23	0.288403	0.149877	0.519679
DaySinceLastOrder	DaySinceLastOrder	6	0.248545	0.217875	0.876602
PreferedOrderCat	PreferedOrderCat	7	0.181082	0.133927	0.739593
MaritalStatus	MaritalStatus	3	0.160299	0.139935	0.87296
WarehouseToHome	WarehouseToHome	5	0.155695	0	0
OrderAmountHikeFromlastYear	OrderAmountHikeFromlastYear	5	0.14942	0	0
OrderCount	OrderCount	3	0.123483	0	0
CityTier	CityTier	5	0.121687	0	0
PreferredPaymentMode	PreferredPaymentMode	2	0.110447	0	0
NumberOfDeviceRegistered	NumberOfDeviceRegistered	1	0.047227	0.068239	1.444936
CouponUsed	CouponUsed	1	0.035711	0	0
CashbackAmount	CashbackAmount	0	0	0	.
HourSpendOnApp	HourSpendOnApp	0	0	0	.
Gender	Gender	0	0	0	.
PreferredLoginDevice	PreferredLoginDevice	0	0	0	.
CustomerID	CustomerID	0	0	0	.
SatisfactionScore	SatisfactionScore	0	0	0	.

It can also be seen that the most important variable of the Boosting model is Tenure. A total of 13 variables are used by the Boosting model.

Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	INTERVAL	5
INPUT	NOMINAL	14
TARGET	NOMINAL	1

Model Events

Target	Event	Measurement Level	Number of Levels	Order	Label
Churn	1	NOMINAL	2	Descending	Churn

Predicted and decision variables

Type	Variable	Label
TARGET	Churn	Churn
PREDICTED	P_Churn1	Predicted: Churn=1
RESIDUAL	R_Churn1	Residual: Churn=1
PREDICTED	P_Churn0	Predicted: Churn=0
RESIDUAL	R_Churn0	Residual: Churn=0
FROM	F_Churn	From: Churn
INTO	I_Churn	Into: Churn

Fit Statistics

Target=Churn Target Label=Churn

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2641.00	1133.00
SUMW	Sum of Case Weights Times Freq	5282.00	2266.00
MISC	Misclassification Rate	0.10	0.10
MAX	Maximum Absolute Error	0.97	0.97
SSE	Sum of Squared Errors	426.49	182.28
ASE	Average Squared Error	0.08	0.08
RASE	Root Average Squared Error	0.28	0.28
DIV	Divisor for ASE	5282.00	2266.00
DFT	Total Degrees of Freedom	2641.00	.

Assessment Score Rankings

Data Role=TRAIN Target Variable=Churn Target Label=Churn

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	461.571	5.61571	5.61571	93.9850	93.9850	133	0.66686
10	393.792	4.25500	4.93792	71.2121	82.6415	132	0.55615
15	322.924	2.80649	4.22924	46.9697	70.7809	132	0.47382
20	273.868	2.26330	3.73868	37.8788	62.5709	132	0.37317
25	213.671	0.72426	3.13671	12.1212	52.4962	132	0.26800
30	184.063	1.35798	2.84063	22.7273	47.5410	132	0.18228
35	150.632	0.49793	2.50632	8.3333	41.9459	132	0.13070
40	123.855	0.36213	2.23855	6.0606	37.4645	132	0.10457
45	101.516	0.22633	2.01516	3.7879	33.7258	132	0.08877
50	83.641	0.22633	1.83641	3.7879	30.7343	132	0.07775
55	68.191	0.13580	1.68191	2.2727	28.1487	132	0.06842
60	55.315	0.13580	1.55315	2.2727	25.9937	132	0.06153
65	46.159	0.36213	1.46159	6.0606	24.4613	132	0.05573
70	38.633	0.40739	1.38633	6.8182	23.2017	132	0.05026
75	30.300	0.13580	1.30300	2.2727	21.8072	132	0.04457
80	22.443	0.04527	1.22443	0.7576	20.4922	132	0.03866
85	16.308	0.18106	1.16308	3.0303	19.4655	132	0.03231
90	11.106	0.22633	1.11106	3.7879	18.5949	132	0.02724
95	5.261	0.00000	1.05261	0.0000	17.6166	132	0.02326
100	0.000	0.00000	1.00000	0.0000	16.7361	132	0.01841

Assessment Score Distribution

Data Role=TRAIN Target Variable=Churn Target Label=Churn

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.75-0.80	2	0	0.76298	0.0757
0.70-0.75	31	1	0.72228	1.2117
0.65-0.70	46	1	0.66991	1.7796
0.60-0.65	55	8	0.62213	2.3855
0.55-0.60	40	14	0.57674	2.0447
0.50-0.55	64	24	0.52597	3.3321
0.45-0.50	33	45	0.47839	2.9534
0.40-0.45	26	39	0.43024	2.4612
0.35-0.40	24	38	0.37371	2.3476
0.30-0.35	13	41	0.32747	2.0447
0.25-0.30	5	63	0.27885	2.5748
0.20-0.25	17	60	0.23024	2.9156
0.15-0.20	21	81	0.17276	3.8622
0.10-0.15	16	217	0.12050	8.8224
0.05-0.10	31	728	0.06993	28.7391
0.00-0.05	18	839	0.03212	32.4498

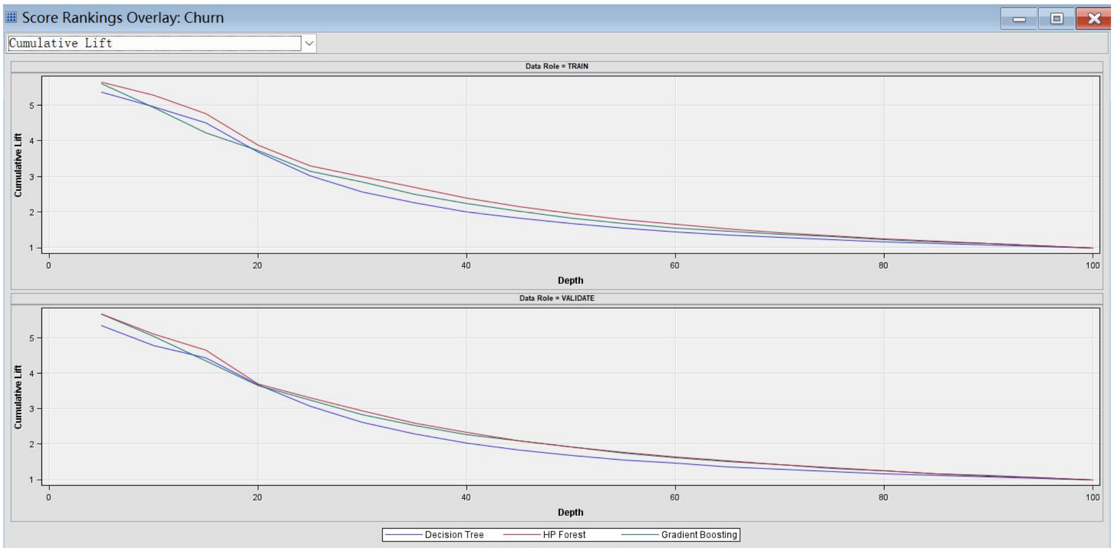
As can be seen from the above figure, the misclassification rate is only 0.1, so the model has better effect.

4.Compare models

Use the model comparison node Model Compare to compare the results of the three models. The result is as follows:

. Property	Value
General	
Node ID	MdlComp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
⊟Model Selection	
Selection Data	Default
Selection Statistic	Misclassification Rate
HP Selection Statistic	Default
SAS Viya Selection Statistic	...
Selection Table	Train
Selection Depth	10
Score	
Selection Editor	...
Report	
⊟Selected Model	
Target	Churn
Model Node	Tree
Model Description	Decision Tree
Selection Criteria	Valid: Misclassification Rate
Status	
Create Time	07/01/24 10:33
Run ID	9d98d678-f71e-2148-9890-9a15f031e5
Last Error	
Last Status	Complete
Last Run Time	07/01/24 11:07
Run Duration	0 Hr. 0 Min. 4.96 Sec.
Grid Host	
User-Added Node	No

The above is the setting process of the model comparison node. The misclassification rate is selected as the criterion for selecting the best model. The results are as follows:



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	0.09356	0.074173	0.09125	0.077829
	Boost	Gradient Boosting	0.09532	0.080744	0.09542	0.080441
	HPDMForest	HP Forest	0.11474	0.077250	0.10413	0.082004

As can be seen from the figure, the misclassification rate of Decision tree is 0.0912, boost is 0.0954, and bagging is 0.1041. The Decision Tree has the lowest misclassification rate (0.0912), making it the best-performing model among the three based on the given metric. Lower misclassification rates usually suggest better predictive performance.

Examining decision tree and ensemble models, specifically in the context of customer behavior, offers valuable insights for shaping business strategies. The Decision Tree model, boasting low Root Average Squared Error (RASE) and Sum of Squared Errors (SSE), lays a robust foundation for comprehending factors influencing customer loyalty and churn. Crucial factors like "Tenure," "Preferred Login Device," and "Satisfaction Score" emerge as pivotal in shaping customer decisions. Meanwhile, the Boosting model, despite slightly higher RASE and SSE, delves into nuanced patterns, adding depth to the analysis. On the contrary, the HPDM, likely a hyperparameter-tuned Decision Tree, presents higher complexity with an elevated RASE, prompting careful consideration.

Strategic recommendations involve prioritizing insights from the Decision Tree, harnessing the nuanced findings of the Boosting model, and thoughtfully evaluating the benefits of hyperparameter tuning. Businesses can refine customer retention strategies, tailor promotions based on factors like "Coupon Usage" and "Order Frequency," and implement targeted engagement approaches, taking into account "Days Since Last Order."

In essence, a thorough examination of decision tree and ensemble models equips businesses with actionable insights to elevate customer retention strategies, optimize promotional initiatives, and enhance overall customer engagement.