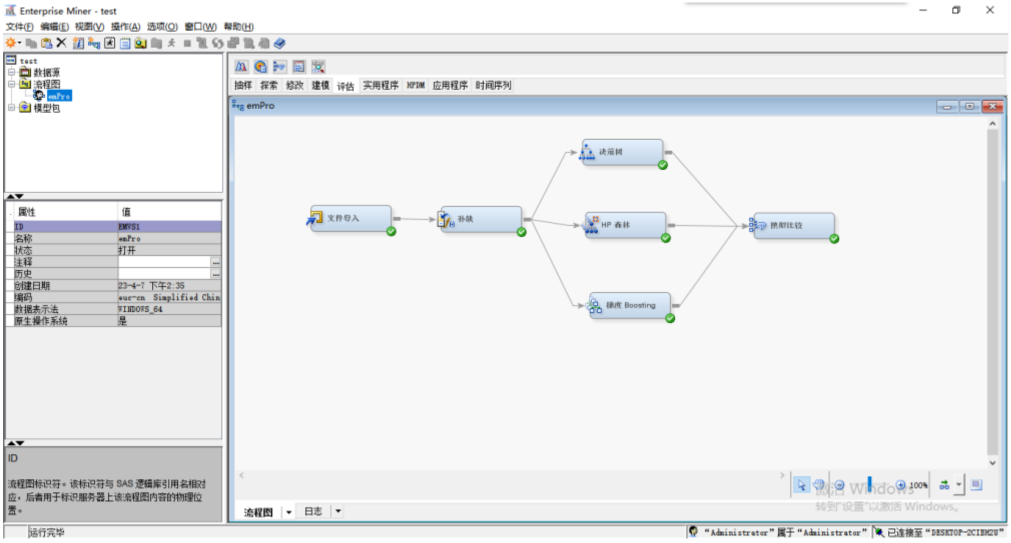
WQD7005 DATA MINING

ALTERNATIVE ASSESSMENT 1

S2121005 XU HUANDI

Case Study: E-Commerce Customer Behaviour Analysis
Data sources：https://www.kaggle.com
Dataset link：
https://drive.google.com/drive/folders/17oIfGGlBI_cSjsItHpvvMQCT6VQ7eNmm?usp=sharing

1. Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.
   —Overall Flow Chart



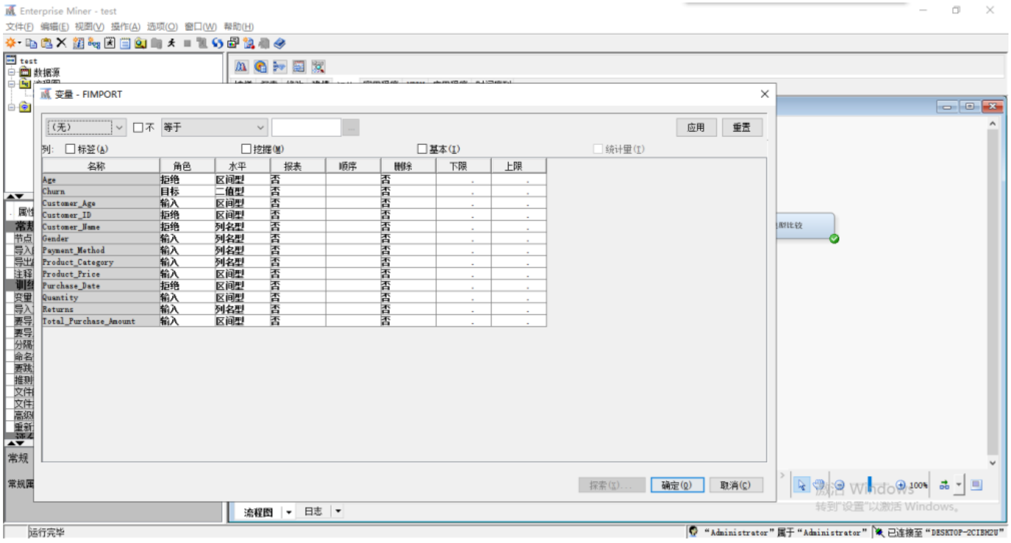
属性	值
ID	EMWS1
名称	emPro
状态	打开
注释	
历史	
创建日期	23-4-7 下午2:35
编码	euc-cn Simplified Chin
数据表示法	WINDOWS_64
原生操作系统	是

ID

流程图标识符。该标识符与 SAS 逻辑库引用名相对应，后者用于标识服务器上该流程图内容的物理位置。

"Administrator" 属于 "Administrator" 已连接至 "DESKTOP-2CIBM2U"

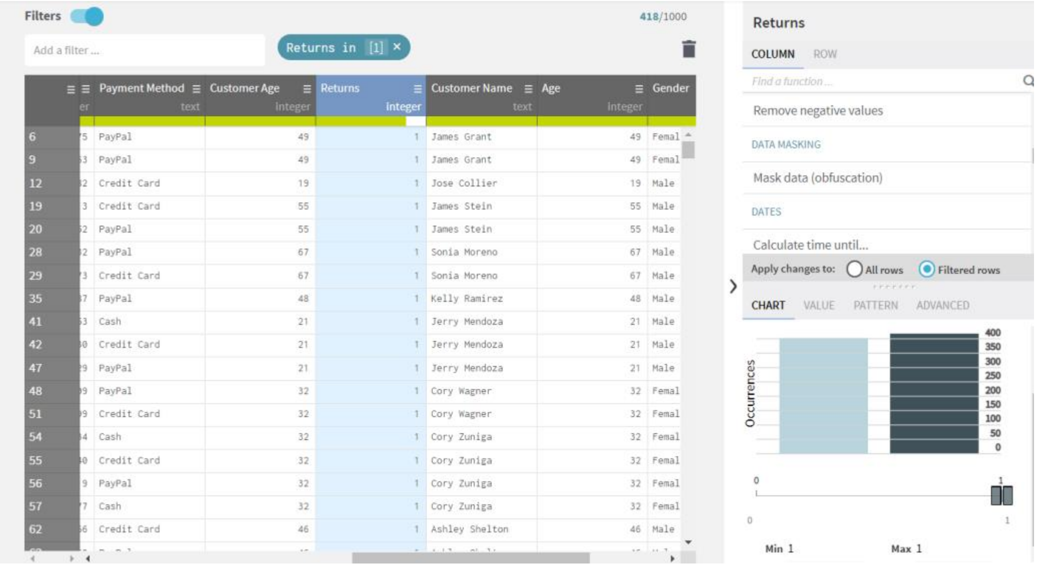
—File import



名称	角色	水平	报表	顺序	删除	下限	上限
Age	拒绝	区间型	否		否	.	.
Churn	目标	二值型	否		否	.	.
Customer_Age	输入	区间型	否		否	.	.
Customer_ID	拒绝	区间型	否		否	.	.
Customer_Name	拒绝	列名型	否		否	.	.
Gender	输入	列名型	否		否	.	.
Payment_Method	输入	列名型	否		否	.	.
Product_Category	输入	列名型	否		否	.	.
Product_Price	输入	区间型	否		否	.	.
Purchase_Date	拒绝	区间型	否		否	.	.
Quantity	输入	区间型	否		否	.	.
Returns	输入	列名型	否		否	.	.
Total_Purchase_Amount	输入	区间型	否		否	.	.

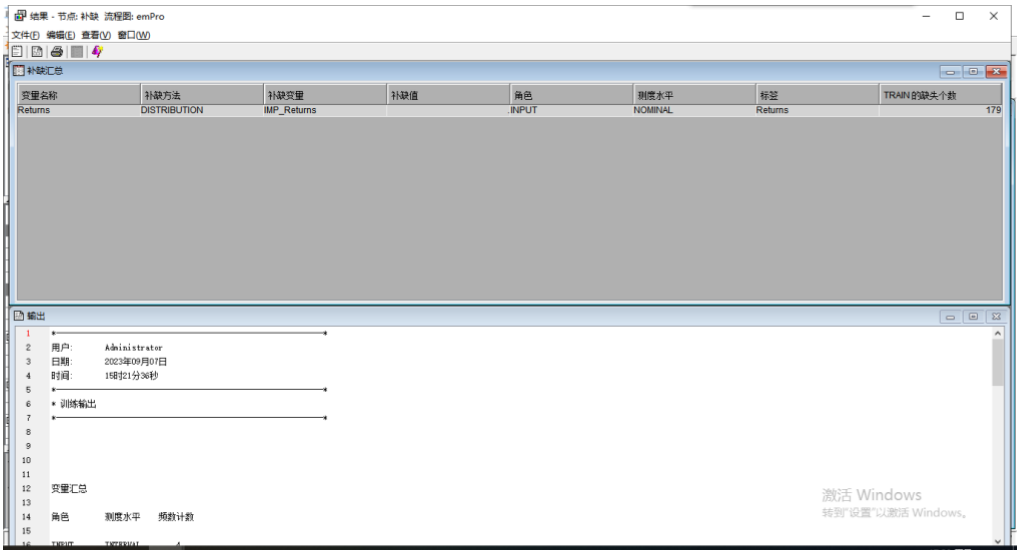
"Administrator" 属于 "Administrator" 已连接至 "DESKTOP-2CIBM2U"

—First use talend (DP) to handle missing values



	er	Payment Method text	Customer Age integer	Returns integer	Customer Name text	Age integer	Gender
6	5	PayPal	49	1	James Grant	49	Femal
9	3	PayPal	49	1	James Grant	49	Femal
12	2	Credit Card	19	1	Jose Collier	19	Male
19	3	Credit Card	55	1	James Stein	55	Male
20	2	PayPal	55	1	James Stein	55	Male
28	2	PayPal	67	1	Sonia Moreno	67	Male
29	3	Credit Card	67	1	Sonia Moreno	67	Male
35	7	PayPal	48	1	Kelly Ramirez	48	Male
41	3	Cash	21	1	Jerry Mendoza	21	Male
42	0	Credit Card	21	1	Jerry Mendoza	21	Male
47	9	PayPal	21	1	Jerry Mendoza	21	Male
48	9	PayPal	32	1	Cory Wagner	32	Femal
51	9	Credit Card	32	1	Cory Wagner	32	Femal
54	4	Cash	32	1	Cory Zuniga	32	Femal
55	0	Credit Card	32	1	Cory Zuniga	32	Femal
56	9	PayPal	32	1	Cory Zuniga	32	Femal
57	7	Cash	32	1	Cory Zuniga	32	Femal
62	6	Credit Card	46	1	Ashley Shelton	46	Male

—Noise and missing values are removed with SAS (EM).

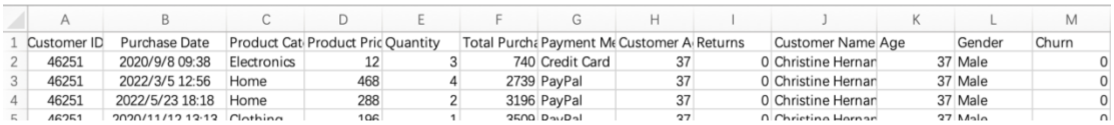

结果 - 节点: 补缺 流程图: emPro
文件(F) 编辑(E) 查看(V) 窗口(W)
补缺汇总
输出
激活 Windows
转到"设置"以激活 Windows。

变量名称	补缺方法	补缺变量	补缺值	角色	测度水平	标签	TRAIN 的缺失个数
Returns	DISTRIBUTION	IMP_Returns		.INPUT	NOMINAL	Returns	179

```
*------------------------------------------------------*
用户:        Administrator
日期:        2023年09月07日
时间:        15时21分36秒
*------------------------------------------------------*
* 训练输出
*------------------------------------------------------*



变量汇总

角色        测度水平     频数计数

INPUT       INTERVAL          4
```

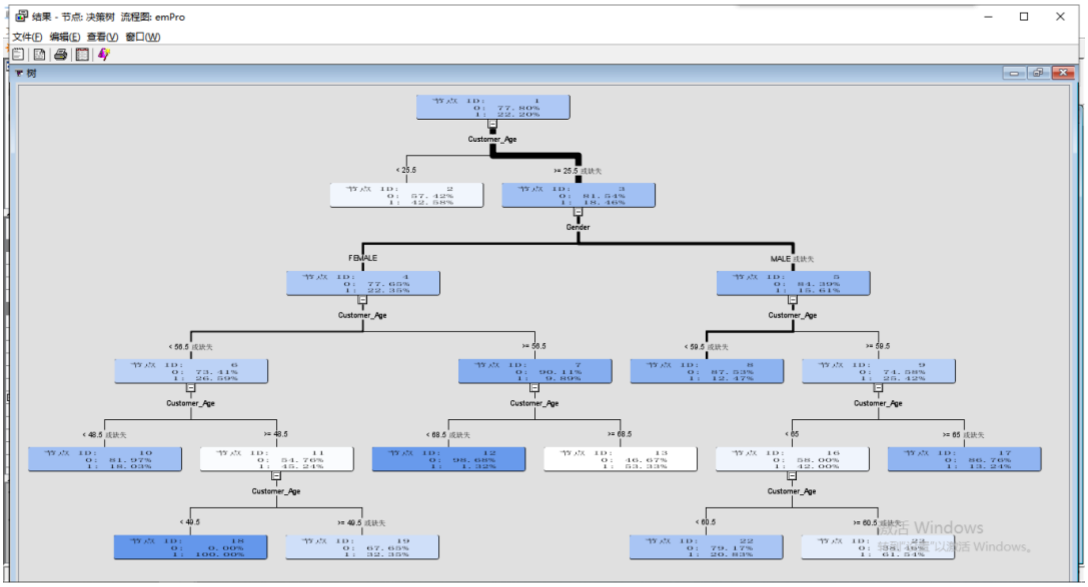
In the dataset, the dependent variable (target variable) should be "Churn" because this is the outcome I want to predict. All other variables can be used as independent variables (explanatory variables) to predict whether a customer will churn.

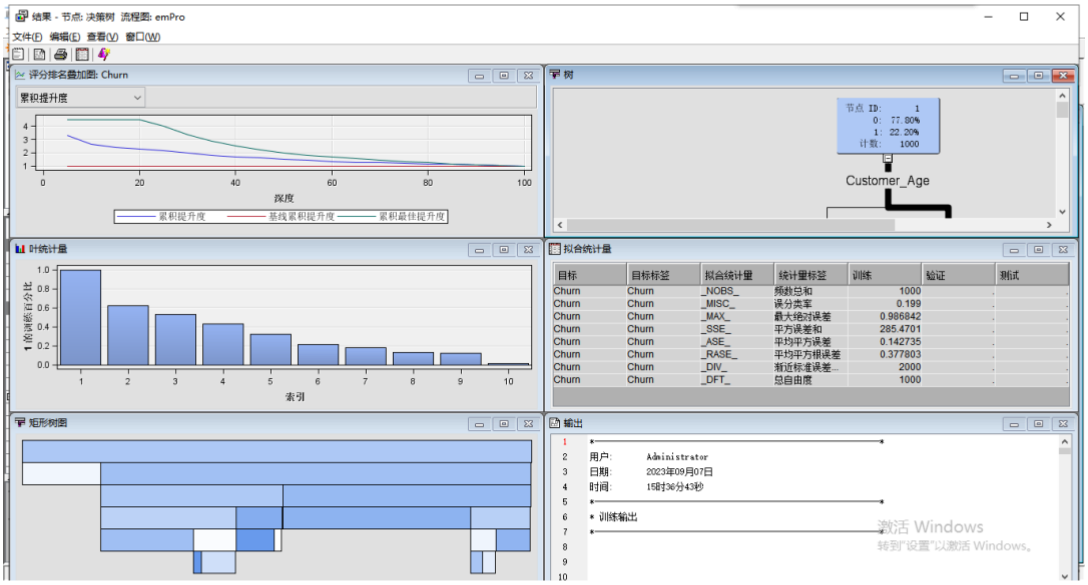
Because there are two ages in the data. So I set up a rejection for one of them, because the name, ID, and date cannot be used for modeling, so I also set up a rejection.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Customer ID	Purchase Date	Product Cat	Product Pric	Quantity	Total Purcha	Payment Me	Customer A	Returns	Customer Name	Age	Gender	Churn
2	46251	2020/9/8 09:38	Electronics	12	3	740	Credit Card	37	0	Christine Hernan	37	Male	0
3	46251	2022/3/5 12:56	Home	468	4	2739	PayPal	37	0	Christine Hernan	37	Male	0
4	46251	2022/5/23 18:18	Home	288	2	3196	PayPal	37	0	Christine Hernan	37	Male	0
5	46251	2020/11/12 13:13	Clothing	196	1	3509	PayPal	37	0	Christine Hernan	37	Male	0

2. Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour
—Create a decision


结果 - 节点: 决策树 流程图: emPro
文件(F) 编辑(E) 查看(V) 窗口(W)
树
节点 ID: 1 0: 77.80% 1: 22.20%
Customer_Age
< 25.5
>= 25.5 或缺失
节点 ID: 2 0: 57.42% 1: 42.58%
节点 ID: 3 0: 81.54% 1: 18.46%
Gender
FEMALE
MALE 或缺失
节点 ID: 4 0: 77.65% 1: 22.35%
节点 ID: 5 0: 84.39% 1: 15.61%
Customer_Age
< 56.5 或缺失
>= 56.5
< 59.5 或缺失
>= 59.5
节点 ID: 6 0: 73.41% 1: 26.59%
节点 ID: 7 0: 90.11% 1: 9.89%
节点 ID: 8 0: 87.53% 1: 12.47%
节点 ID: 9 0: 74.58% 1: 25.42%
< 48.5 或缺失
>= 48.5
< 68.5 或缺失
>= 68.5
< 65
>= 65 或缺失
节点 ID: 10 0: 81.97% 1: 18.03%
节点 ID: 11 0: 54.76% 1: 45.24%
节点 ID: 12 0: 98.68% 1: 1.32%
节点 ID: 13 0: 46.67% 1: 53.33%
节点 ID: 16 0: 58.00% 1: 42.00%
节点 ID: 17 0: 86.76% 1: 13.24%
< 49.5
>= 49.5 或缺失
< 60.5
>= 60.5 或缺失
节点 ID: 18 0: 0.00% 1: 100.00%
节点 ID: 19 0: 67.65% 1: 32.35%
节点 ID: 22 0: 79.17% 1: 20.83%
节点 ID: 23 0: 38.46% 1: 61.54%
激活 Windows
转到"设置"以激活 Windows。



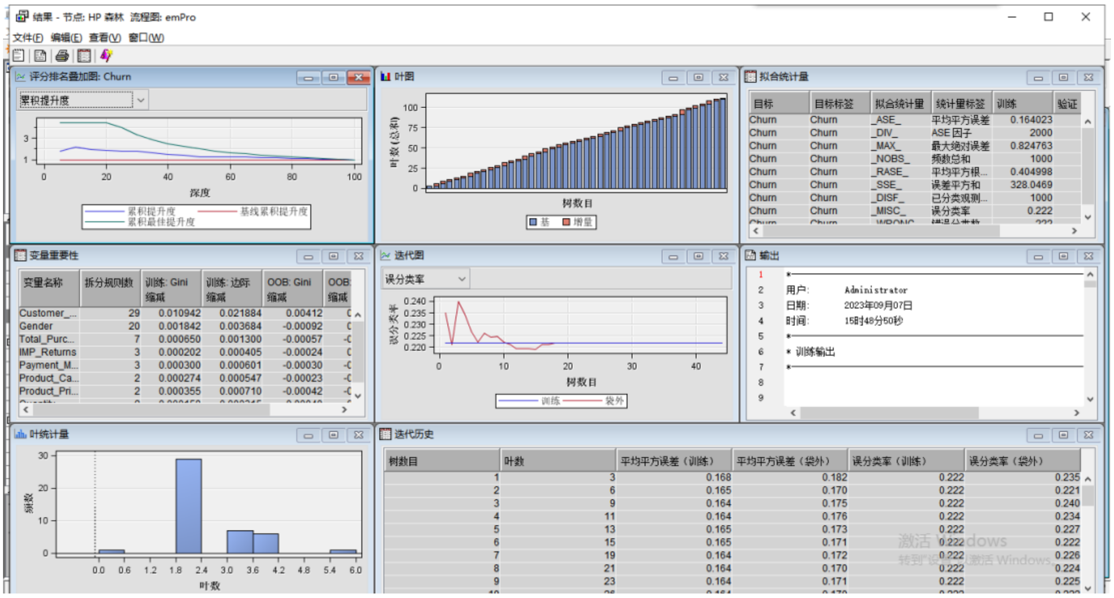
目标	目标标签	拟合统计量	统计量标签	训练	验证	测试
Churn	Churn	_NOBS_	频数总和	1000	.	.
Churn	Churn	_MISC_	误分类率	0.199	.	.
Churn	Churn	_MAX_	最大绝对误差	0.986842	.	.
Churn	Churn	_SSE_	平方误差和	285.4701	.	.
Churn	Churn	_ASE_	平均平方误差	0.142735	.	.
Churn	Churn	_RASE_	平均平方根误差	0.377803	.	.
Churn	Churn	_DIV_	渐近标准误差...	2000	.	.
Churn	Churn	_DFT_	总自由度	1000	.	.

```
*------------------------------------------------------*
用户:        Administrator
日期:        2023年09月07日
时间:        15时36分43秒
*------------------------------------------------------*
* 训练输出
```

The problem is that an overfitted model relies too much on specific features in the training data and therefore may perform poorly on new, unknown data because those features may not apply to the new data. On the contrary, underfitting models perform poorly on both training and test data because they do not fully learn the characteristics of the training data and do not understand the data well enough. To solve these problems, we can use techniques such as limiting the depth of the decision tree or the minimum number of samples at a node to prevent overfitting, or consider using ensemble methods such as random forest or gradient boosting, which work by combining multiple models. Improve generalization capabilities to achieve better performance on new data. These methods can effectively improve the performance of the model.

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

—Random Forest



拟合统计量

目标	目标标签	拟合统计量	统计量标签	训练	验证
Churn	Churn	_ASE_	平均平方误差	0.164023	
Churn	Churn	_DIV_	ASE 因子	2000	
Churn	Churn	_MAX_	最大绝对误差	0.824763	
Churn	Churn	_NOBS_	频数总和	1000	
Churn	Churn	_RASE_	平均平方根...	0.404998	
Churn	Churn	_SSE_	误差平方和	328.0469	
Churn	Churn	_DISF_	已分类观测...	1000	
Churn	Churn	_MISC_	误分类率	0.222	

变量重要性

变量名称	拆分规则数	训练: Gini 缩减	训练: 边际 缩减	OOB: Gini 缩减	OOB: 缩减
Customer_...	29	0.010942	0.021884	0.00412	C
Gender	20	0.001842	0.003684	-0.00092	C
Total_Purc...	7	0.000650	0.001300	-0.00057	-C
IMP_Returns	3	0.000202	0.000405	-0.00024	C
Payment_M...	3	0.000300	0.000601	-0.00030	-C
Product_Ca...	2	0.000274	0.000547	-0.00023	-C
Product_Pri...	2	0.000355	0.000710	-0.00042	-C

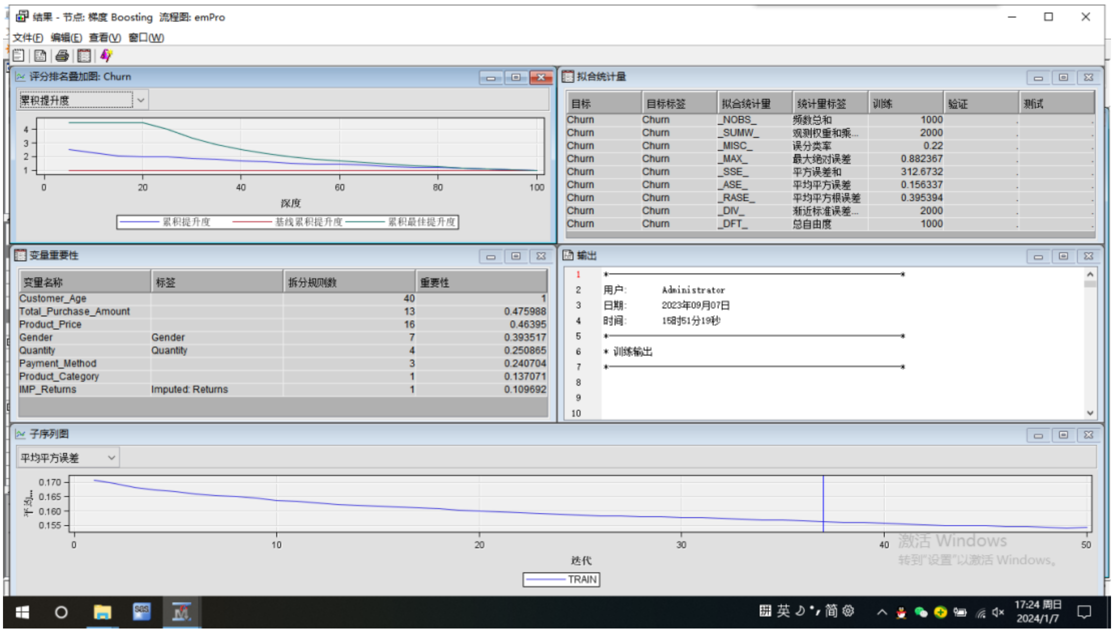
输出

```
1  *------------------------------------------------------------
2  用户:        Administrator
3  日期:        2023年09月07日
4  时间:        15时48分50秒
5  *------------------------------------------------------------
6  * 训练输出
7  *------------------------------------------------------------
8
9
```

迭代历史

树数目	叶数	平均平方误差（训练）	平均平方误差（袋外）	误分类率（训练）	误分类率（袋外）
1	3	0.168	0.182	0.222	0.235
2	6	0.165	0.170	0.222	0.221
3	9	0.164	0.175	0.222	0.240
4	11	0.164	0.176	0.222	0.234
5	13	0.165	0.173	0.222	0.227
6	15	0.165	0.171	0.222	0.222
7	19	0.164	0.172	0.222	0.226
8	21	0.164	0.170	0.222	0.224
9	23	0.164	0.171	0.222	0.225

—Boosting



拟合统计量

目标	目标标签	拟合统计量	统计量标签	训练	验证	测试
Churn	Churn	_NOBS_	频数总和	1000	.	.
Churn	Churn	_SUMW_	观测权重和乘...	2000	.	.
Churn	Churn	_MISC_	误分类率	0.22	.	.
Churn	Churn	_MAX_	最大绝对误差	0.882367	.	.
Churn	Churn	_SSE_	平方误差和	312.6732	.	.
Churn	Churn	_ASE_	平均平方误差	0.156337	.	.
Churn	Churn	_RASE_	平均平方根误差	0.395394	.	.
Churn	Churn	_DIV_	渐近标准误差...	2000	.	.
Churn	Churn	_DFT_	总自由度	1000	.	.

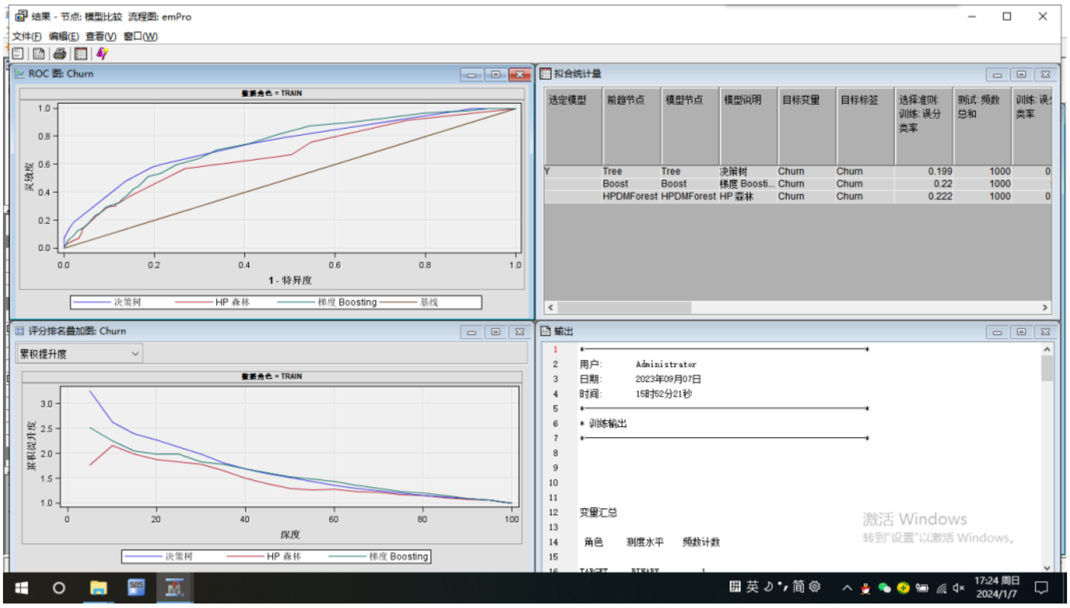
变量重要性

变量名称	标签	拆分规则数	重要性
Customer_Age		40	1
Total_Purchase_Amount		13	0.475988
Product_Price		16	0.46395
Gender	Gender	7	0.393517
Quantity	Quantity	4	0.250865
Payment_Method		3	0.240704
Product_Category		1	0.137071
IMP_Returns	Imputed: Returns	1	0.109692

输出

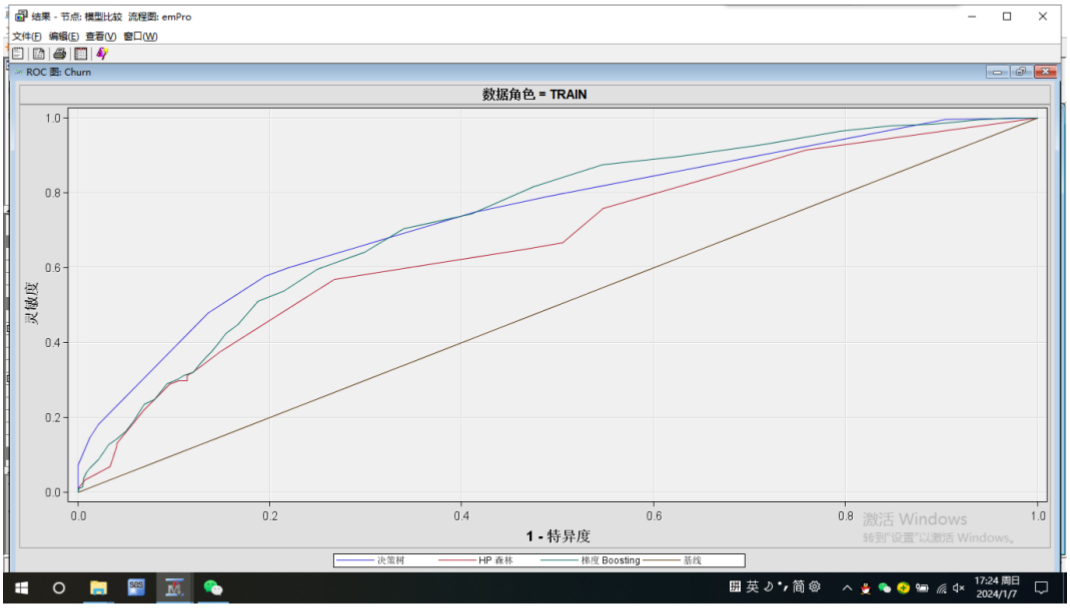
```
1   *------------------------------------------------*
2   用户:        Administrator
3   日期:        2023年09月07日
4   时间:        15时51分19秒
5   *------------------------------------------------*
6   * 训练输出
7   *------------------------------------------------*
8
9
10
```

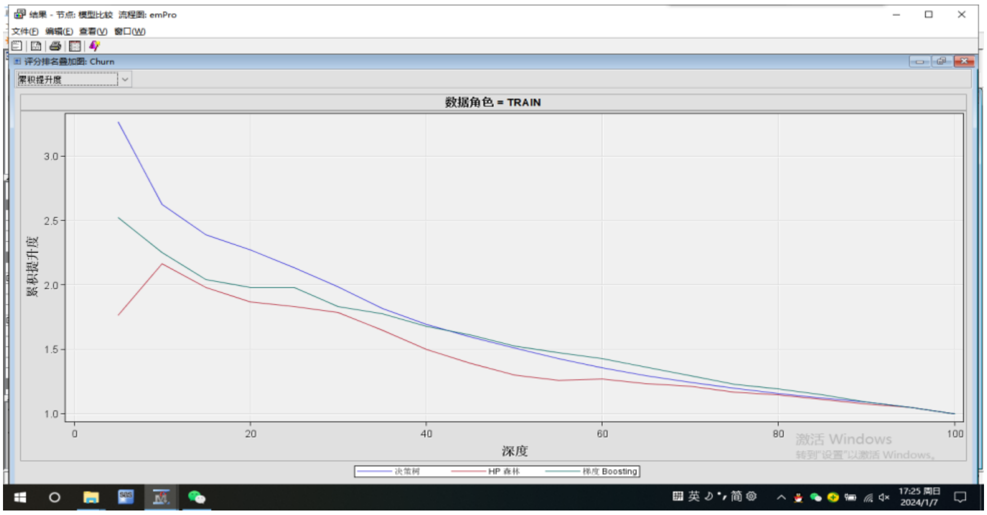
4. Model comparison

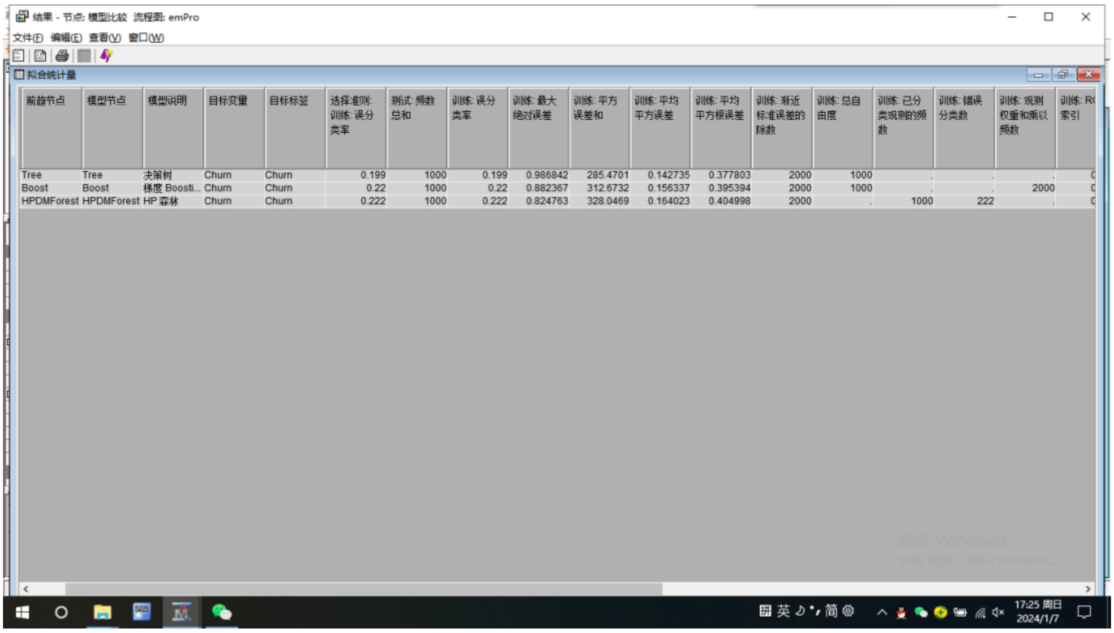


选定模型	前趋节点	模型节点	模型说明	目标变量	目标标签	选择准则: 训练: 误分类率	测试: 频数总和	训练: 误分类率
Y	Tree	Tree	决策树	Churn	Churn	0.199	1000	0
	Boost	Boost	梯度 Boosti...	Churn	Churn	0.22	1000	
	HPDMForest	HPDMForest	HP 森林	Churn	Churn	0.222	1000	0

```
*------------------------------------------------------------*
用户:        Administrator
日期:        2023年09月07日
时间:        15时52分21秒
*------------------------------------------------------------*
* 训练输出
*------------------------------------------------------------*




变量汇总

 角色      测度水平     频数计数

TARGET    BINARY         1
```






结果 - 节点: 模型比较 流程图: emPro
文件(F) 编辑(E) 查看(V) 窗口(W)
拟合统计量
英 简
17:25 周日
2024/1/7

前趋节点	模型节点	模型说明	目标变量	目标标签	选择准则: 训练: 误分类率	测试: 频数总和	训练: 误分类率	训练: 最大绝对误差	训练: 平方误差和	训练: 平均平方误差	训练: 平均平方根误差	训练: 渐近标准误差的除数	训练: 总自由度	训练: 已分类观测的频数	训练: 错误分类数	训练: 观测权重和乘以频数	训练: R(索引
Tree	Tree	决策树	Churn	Churn	0.199	1000	0.199	0.986842	285.4701	0.142735	0.377803	2000	1000	.	.	.	0
Boost	Boost	梯度 Boosti...	Churn	Churn	0.22	1000	0.22	0.882367	312.6732	0.156337	0.395394	2000	1000	.	.	2000	0
HPDMForest	HPDMForest	HP 森林	Churn	Churn	0.222	1000	0.222	0.824763	328.0469	0.164023	0.404998	2000	.	1000	222	.	0

SAS helped me figure out which factors were most important in predicting whether a customer would leave us. The age of our customers, how much they spend, and the price of our products are particularly critical. It's interesting that older customers seem to have different departure patterns than younger customers. And it seems that customers who spend more are more likely to stay. Gender seems to have little impact on whether a customer will churn.
The models we used, especially ensemble methods like random forests and gradient boosting, worked well, especially in terms of accuracy and generalization, compared to decision trees alone.
With this information, I think we can do this:

1. Create some holiday or specific discounts: For those customers who may be about to leave, especially young people or those who don't spend much overall, we can create some discounts or loyalty programs to make them feel more valued and increase their trust in the brand. of loyalty.
2. Refining customer groups: We can use decision tree analysis to divide customers into different groups , and customize different marketing strategies based on their shopping habits and basic information.
3. Re-examine pricing: Since product price is so important, we may need to rethink our pricing strategy to ensure that the price is both competitive and that customers feel they are worth their money.
4. More feedback and engagement: We need to understand why customers choose to leave and improve based on their feedback. This way we can better retain customers and increase their satisfaction.

Dr, I'm sorry about the Chinese in the software, because the Mac I use can only be downloaded using a virtual machine. I don't know why it turned into Chinese, but I don't have much time to download it again. I apologize to you again.