UNIVERSITI MALAYA

WQD7005

DATA MINING

I/2023/2024

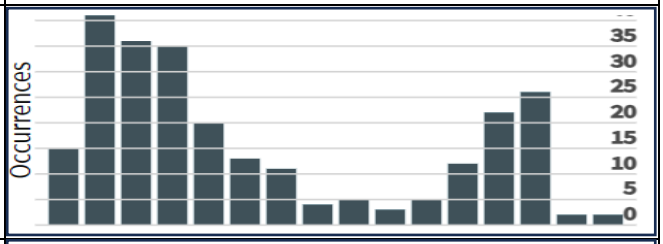
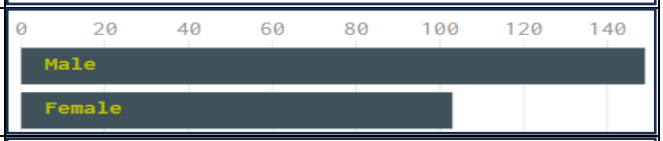
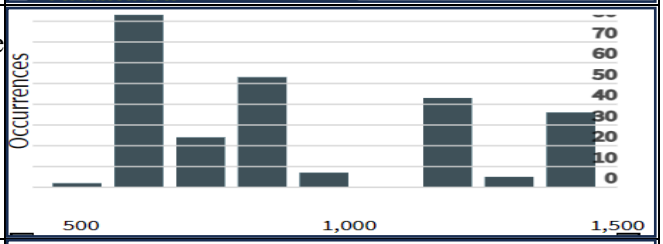
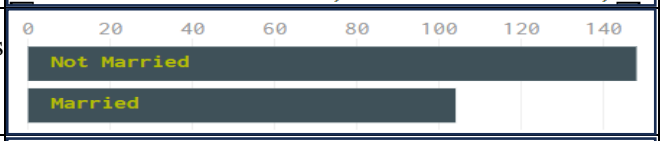
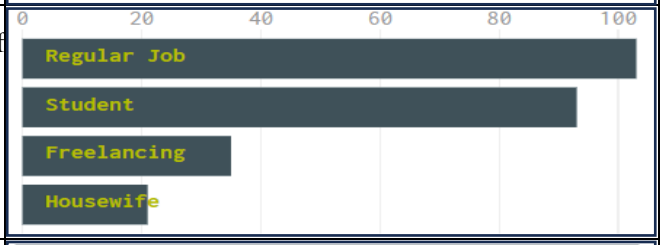
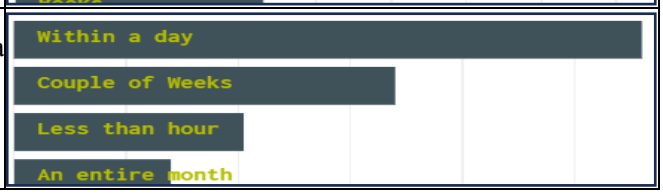
Case Study: E-Commerce Customer Behaviour Analysis

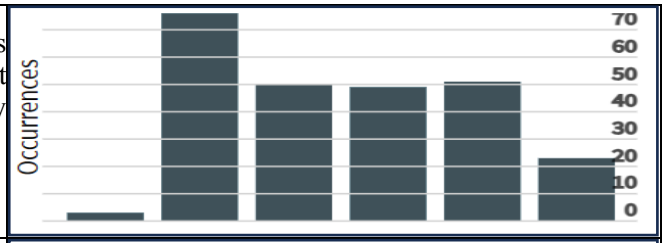
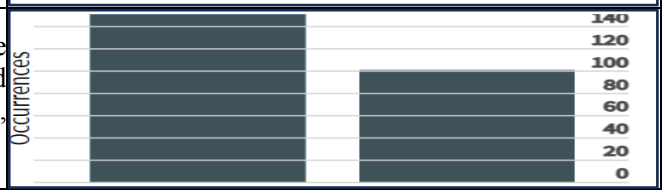
Student Name	Student Matric Number
Zhang Yimei	22063349

E-Commerce Customer Behaviour Analysis

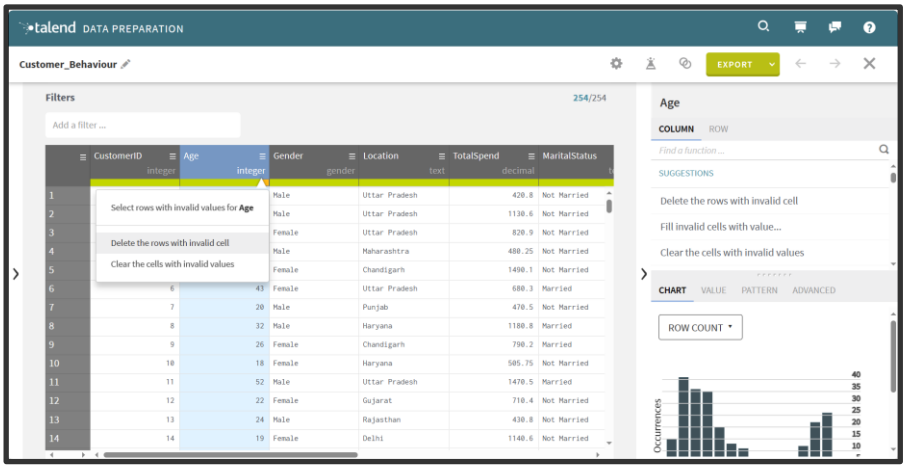
1. Data Import and Preprocessing

Import the dataset into Talend Data Preparation and understand dataset using it：

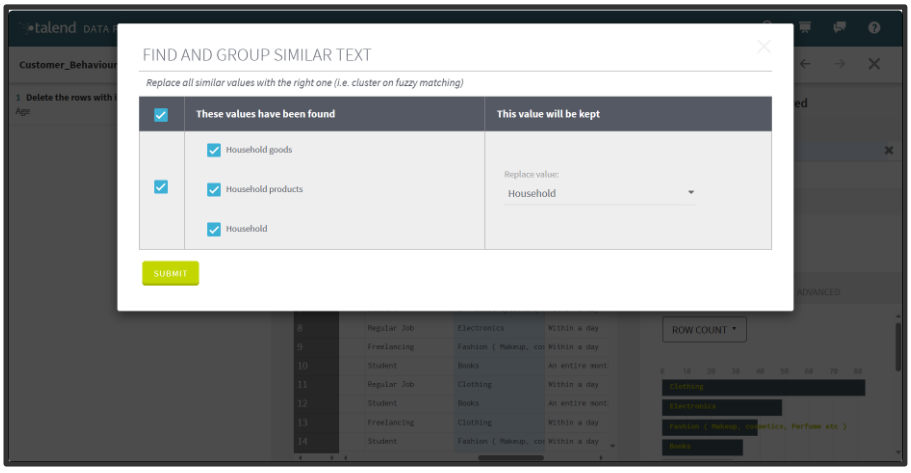
Variables	Explanation	
Age	Age of the customer.	
Gender	Gender of the customer.	
Location	Geographic location of the customer.	
Total Spend	Total amount spent by the customer.	
Marital Status	Marital status of customers	
Financial Status	Financial status of customers	
Product Category Perferred	Customers' preferred product category	
Product Selection Time	The time taken to make a product selection.	

Purchase Frequency	How often individuals make purchases. It measures customer loyalty and engagement.	
Churn	Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).	

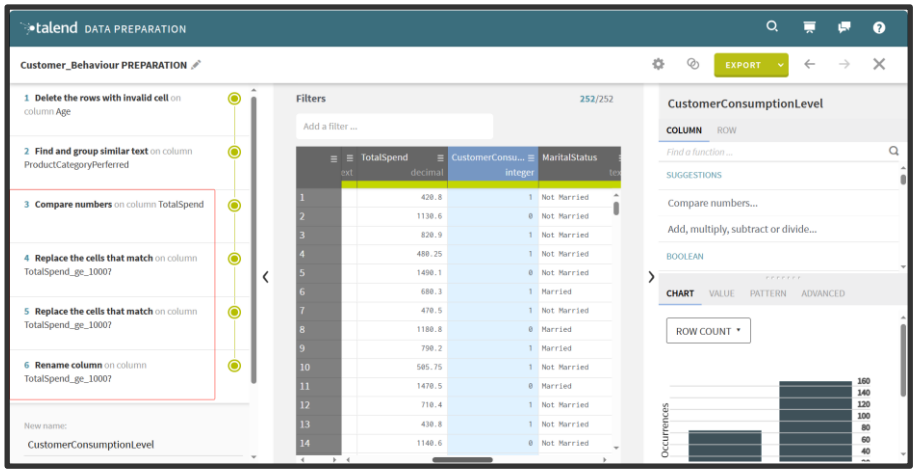
Delete the rows with invalid cell in ‘Age’ column:



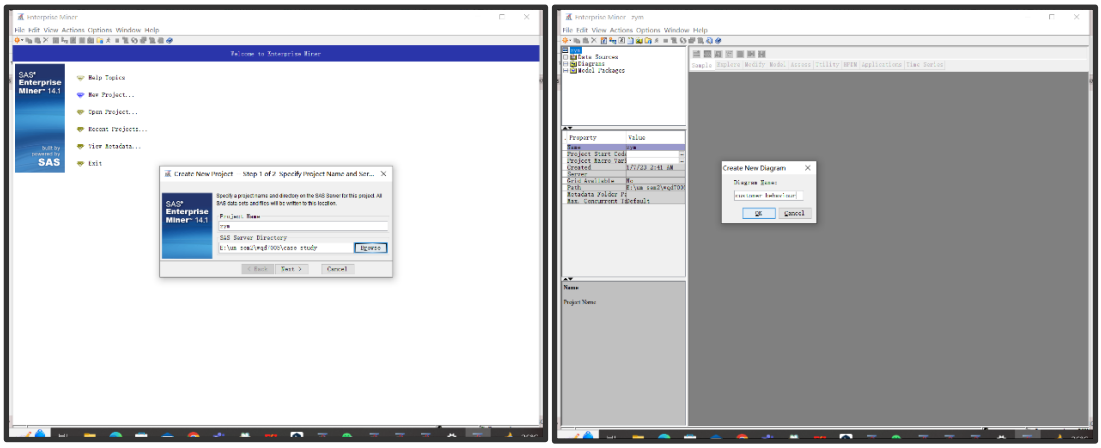
Replace all similar values with the right one in ‘ProductCategoryPerferred’ column:



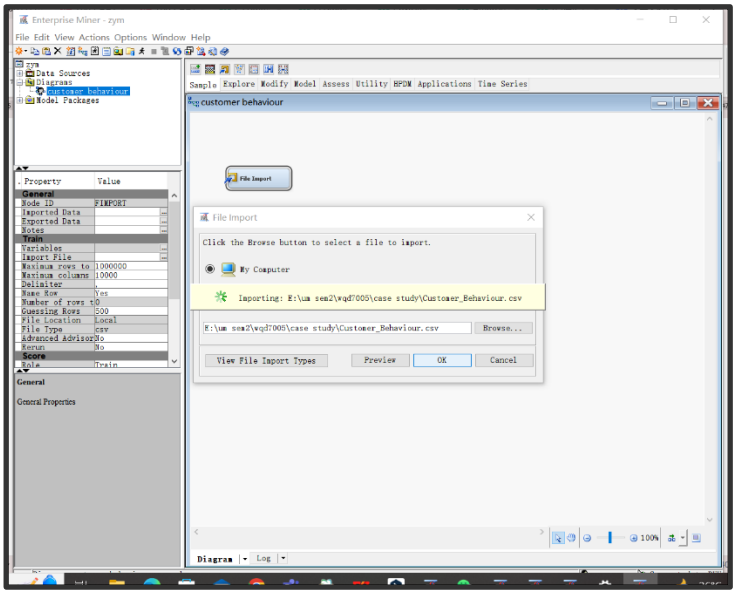
Create a new column ‘CustomerConsumptionLevel’ based on the ‘TotalSpent’ column. Customers who spend more than or equal to 1000 are marked as 0, and those who spend less than 1000 are marked as 1:



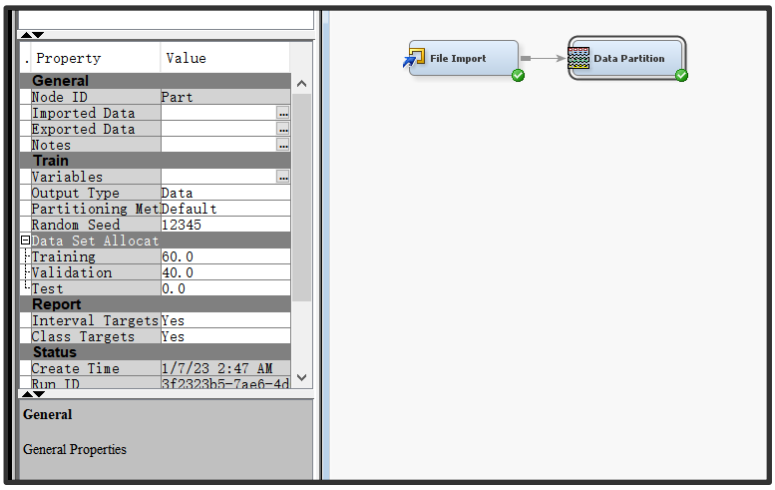
Create a Project in SAS and then create a diagram:



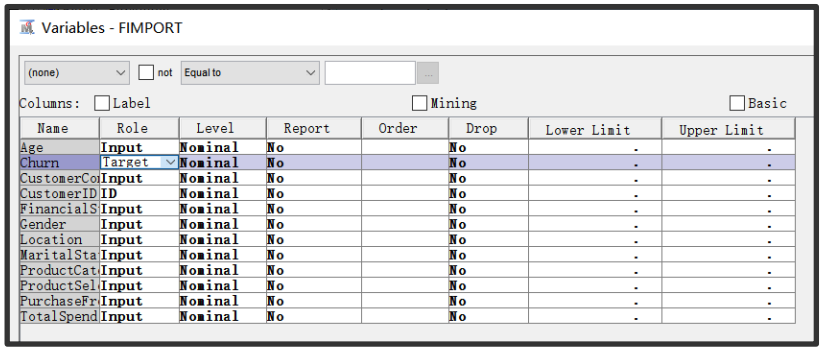
Import the dataset processed by Talend into SAS Enterprise Miner:



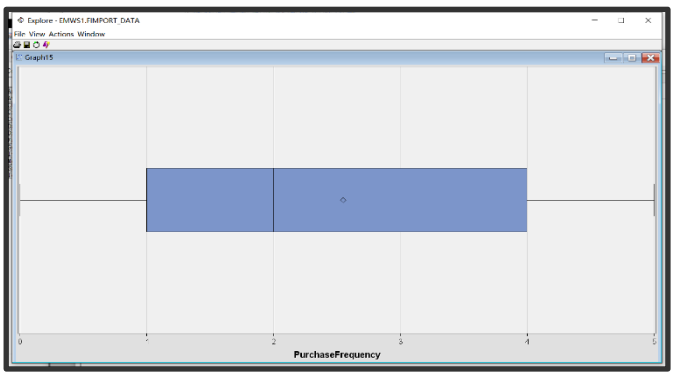
Add a data partition node to partition the data and divide the data set into 60% training set and 40% validation set:

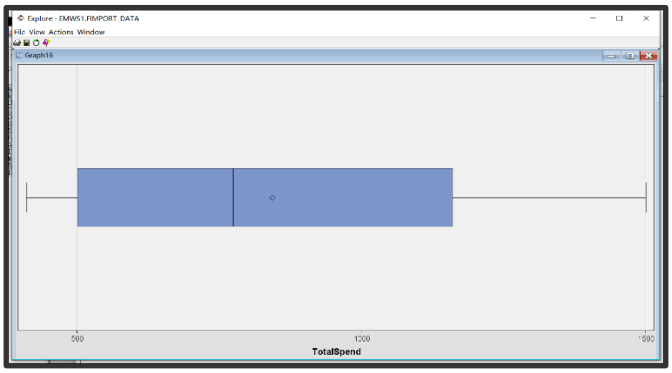


Set the churn variable as the target variable：



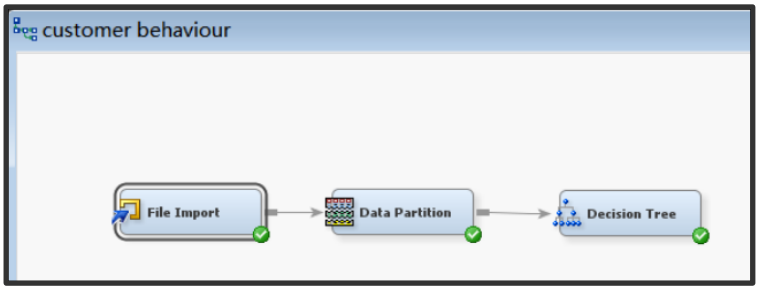
Query the data through box plots and find no outliers:



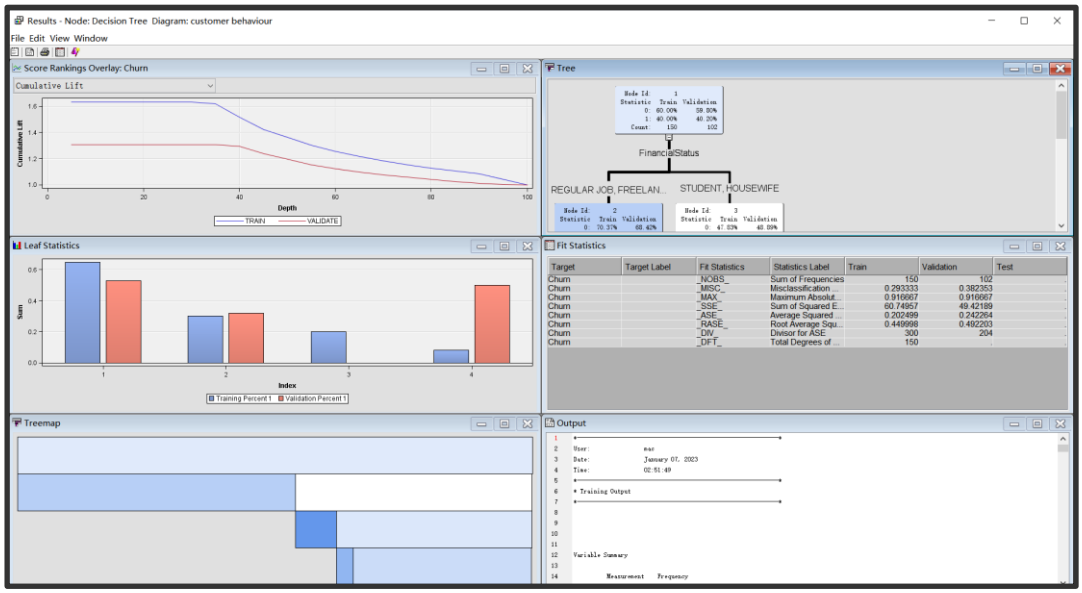


2. Decision Tree Analysis

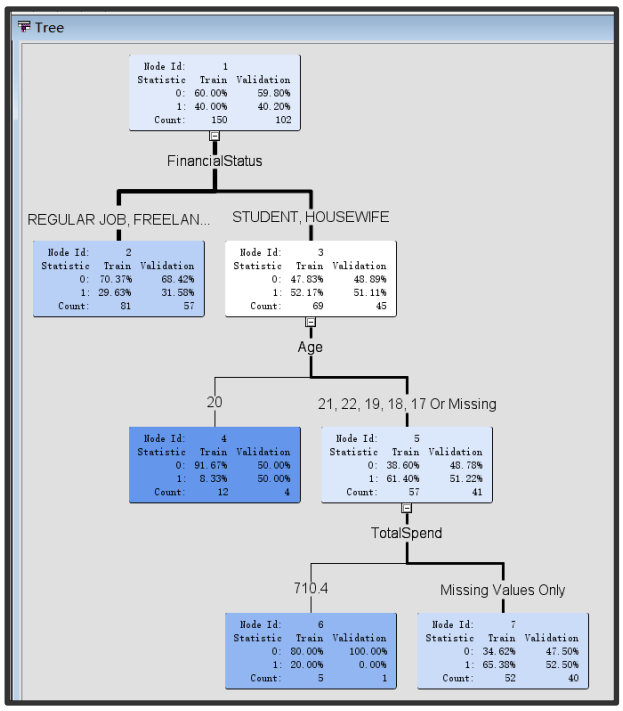
Add a Decision Tree node:



After running the decision tree model, the results are as follows:

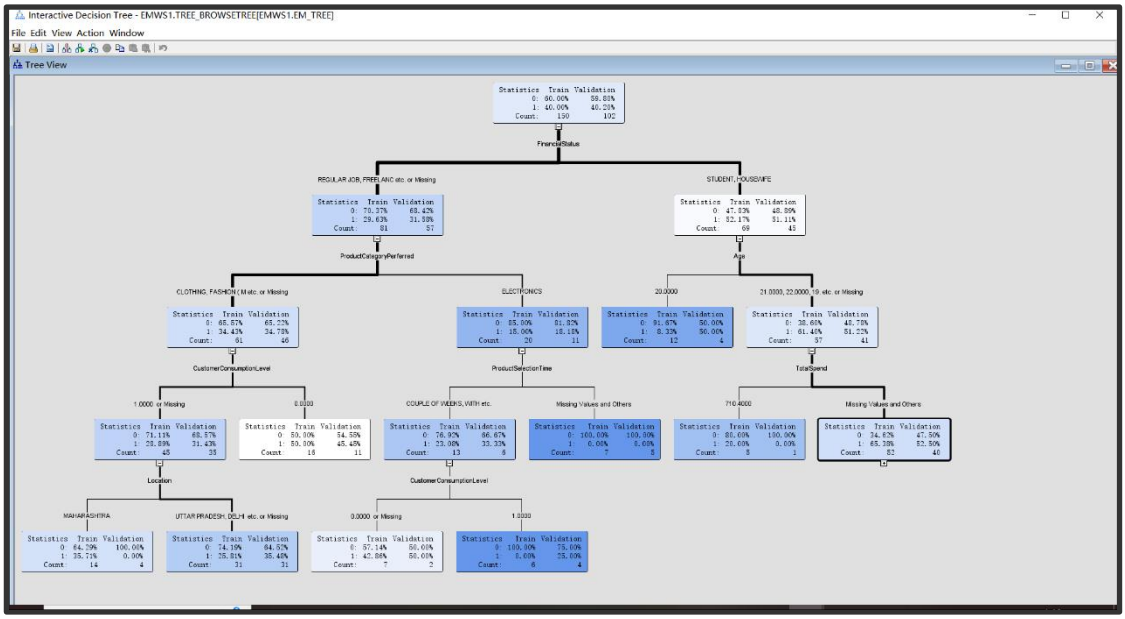


The resulting decision tree is as shown below:



- It can be seen from the decision tree that the factor that has the greatest impact on customer churn is the customer's financial status, followed by the customer's age, and the third-ranked factor is the customer's total spend. This suggests that we need to focus on customer groups with poorer financial conditions, older age, and lower total spending.
- The first level shows that the churn rate of students and housewives is greater than 50%, which is higher than that of people in other industries. This indicate that these two groups are more susceptible to attrition and require special attention in their retention strategies.
- Except for customers who are 20 years old or have a total spend of 710, the churn rate is low, and the churn rate of other customers is as high as more than 50%.

After adjusting the parameters, the resulting decision tree is as shown below:



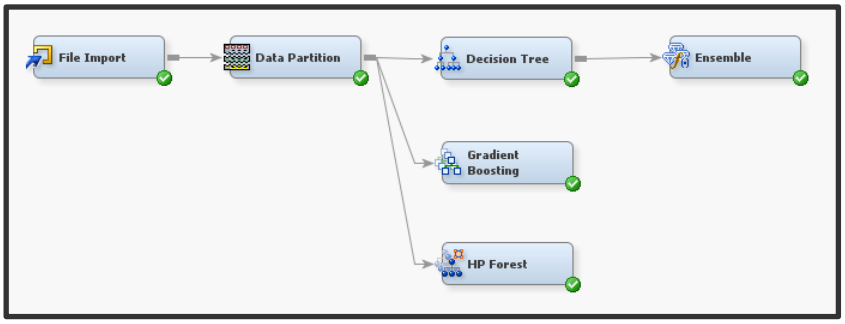
- Among other customers except students and housewives, the churn rate of customers who often buy electronic products is about 17%, which is lower than other customers who buy

clothes, fashion supplies, etc. This suggests to us that electronic products may be more popular and have a positive impact on retention.

- Customers who spend more in total are less likely to churn, and vice versa. This highlights that the key to increasing customer loyalty is to encourage customers to spend more on shopping by offering more offers, rewards or value-added services.
- People who are able to decide quickly which product to buy have a much lower churn rate than people who take a long time. This implies that we can improve decision-making speed by providing personalized recommendations, simplifying the shopping process, etc.

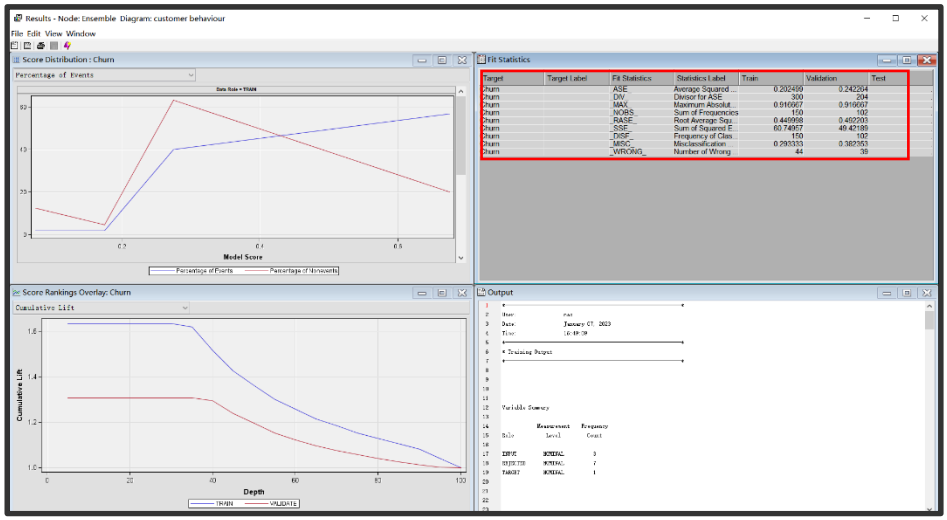
3. Ensemble Methods

Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example, using Gradient Boosting algorithm as a Boosting example.



3.1 Random Forest

The results of Random Forest operation are as follows:



As can be seen from the Fit Statistic table, the model performs relatively well on the training set, but the ASE and MISC on the validation set are higher, there are some signs of overfitting, and the performance is poor. This may be due to the model overlearning the noise or specific patterns of the data on the training set and failing to generalize well to new data.

Assessment Score Distribution

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

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.65-0.70	34	18	0.65385	34.6667
0.25-0.30	24	57	0.29630	54.0000
0.15-0.20	1	4	0.20000	3.3333
0.05-0.10	1	11	0.08333	8.0000

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.65-0.70	21	19	0.65385	39.2157
0.25-0.30	18	39	0.29630	55.8824
0.15-0.20	0	1	0.20000	0.9804
0.05-0.10	2	2	0.08333	3.9216

Classification Table

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

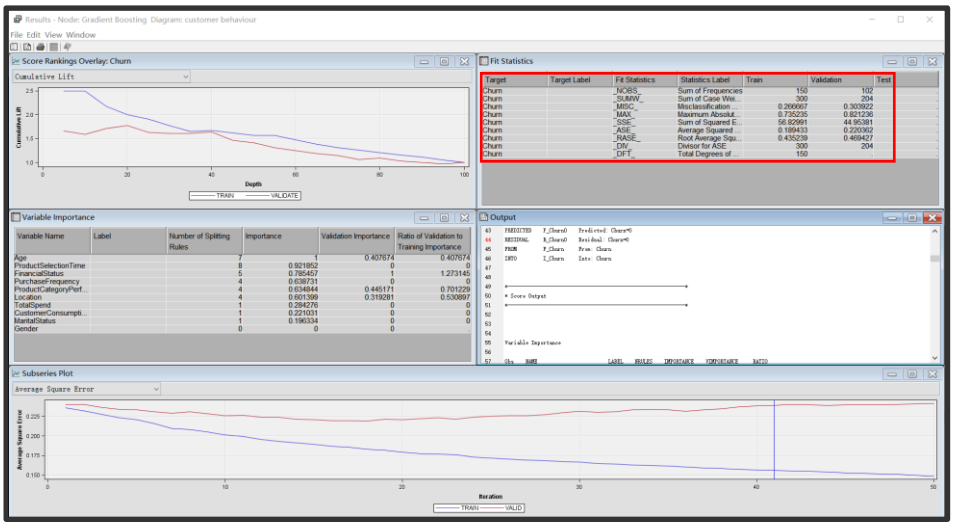
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	73.4694	80.0000	72	48.0000
1	0	26.5306	43.3333	26	17.3333
0	1	34.6154	20.0000	18	12.0000
1	1	65.3846	56.6667	34	22.6667

Data Role=VALIDATE Target Variable=Churn Target Label=' '

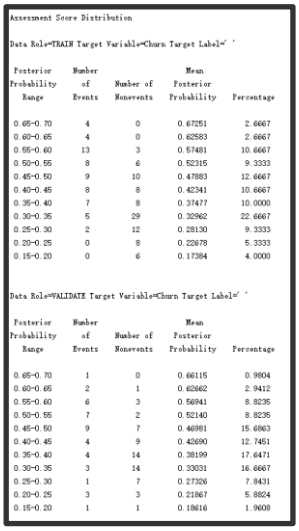
Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	67.7419	68.8525	42	41.1765
1	0	32.2581	48.7805	20	19.6078
0	1	47.5000	31.1475	19	18.6275
1	1	52.5000	51.2195	21	20.5882

3.1 Gradient Boosting

The results of Gradient Boosting operation are as follows:



As can be seen from the Fit Statistic table, the fitting effect of the model on the training set is good, but it should be noted that the misclassification rate is 26.67%; on the verification set, the misclassification rate rises to 30.392%, which may indicate that the model's generalization ability needs to be improved.



Assessment Score Distribution

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

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.65-0.70	4	0	0.67251	2.6667
0.60-0.65	4	0	0.62583	2.6667
0.55-0.60	13	3	0.57481	10.6667
0.50-0.55	8	6	0.52315	9.3333
0.45-0.50	9	10	0.47883	12.6667
0.40-0.45	8	8	0.42341	10.6667
0.35-0.40	7	8	0.37477	10.0000
0.30-0.35	5	29	0.32962	22.6667
0.25-0.30	2	12	0.28130	9.3333
0.20-0.25	0	8	0.22678	5.3333
0.15-0.20	0	6	0.17384	4.0000

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.65-0.70	1	0	0.66115	0.9804
0.60-0.65	2	1	0.62662	2.9412
0.55-0.60	6	3	0.56941	8.8235
0.50-0.55	7	2	0.52140	8.8235
0.45-0.50	9	7	0.46981	15.6863
0.40-0.45	4	9	0.42690	12.7451
0.35-0.40	4	14	0.38199	17.6471
0.30-0.35	3	14	0.33031	16.6667
0.25-0.30	1	7	0.27326	7.8431
0.20-0.25	3	3	0.21867	5.8824
0.15-0.20	1	1	0.18616	1.9608

Classification Table

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	72.3214	90.0000	81	54.0000
1	0	27.6786	51.6667	31	20.6667
0	1	23.6842	10.0000	9	6.0000
1	1	76.3158	48.3333	29	19.3333

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	68.7500	90.1639	55	53.9216
1	0	31.2500	60.9756	25	24.5098
0	1	27.2727	9.8361	6	5.8824
1	1	72.7273	39.0244	16	15.6863

4. Suggestion and Conclusion:

4.1 Suggestions for Business Strategy

- For customers with poor financial status, more financial discounts, installment payments and other services can be provided to increase their loyalty.
- For students and housewives, two groups with high attrition rates, special customized strategies can be developed, such as exclusive discounts, regular promotions, etc., to improve their satisfaction and loyalty.
- Since the churn rate of customers who purchase electronic products is relatively low, sales of this type of products can be increased by launching more electronic products and holding special events.
- For customers with higher total spending, incentives such as member-specific benefits and regular cash back activities can be launched to maintain their high level of shopping spending.
- By optimizing the user experience of the website or application and providing personalized recommendations, you can help customers make shopping decisions more quickly and reduce the churn rate.

4.2 Conclusion

Through decision tree analysis, the factors that have the greatest impact on churn are clarified, providing direction for formulating targeted business strategies. Random Forest and Gradient Boosting algorithms perform feature analysis to gain a deeper understanding of customer behavior. The comprehensive use of algorithms such as decision trees, random forests, and Gradient Boosting can more comprehensively analyze and predict customer behavior, providing strong support for the operation and promotion of e-commerce platforms.