WQD7005 DATA MINING

2023/2024 S1

Results and Analysis

Matric Number: S2164046

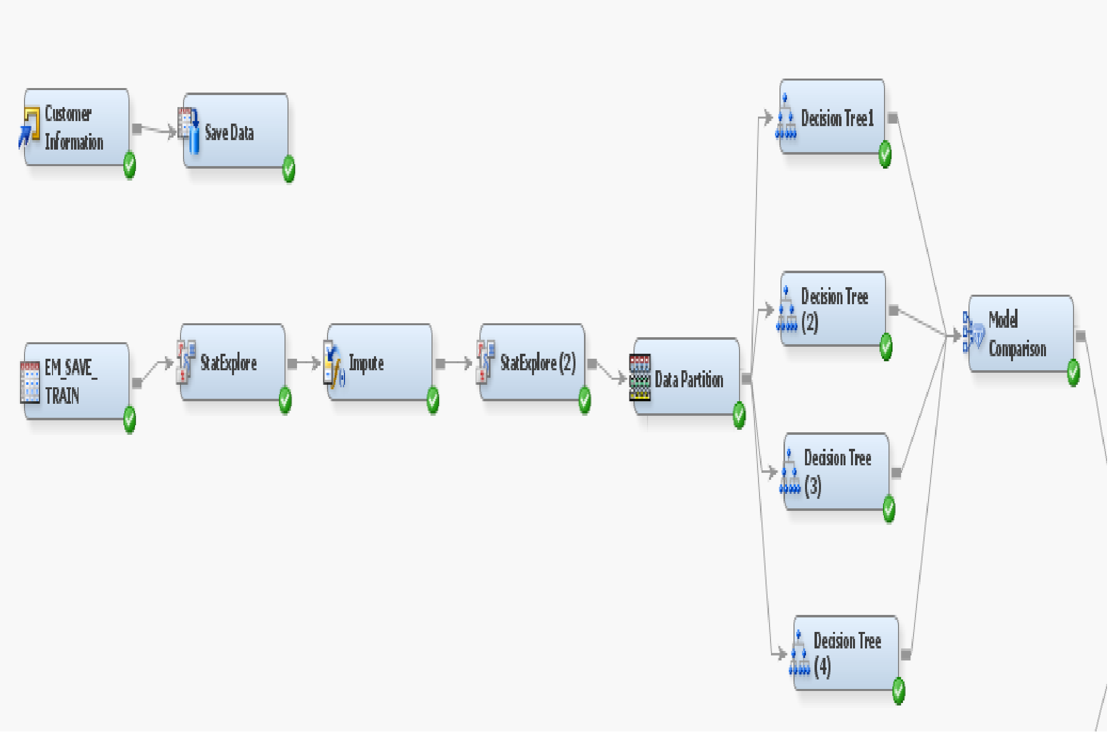
Name: HU LIANGLIANG

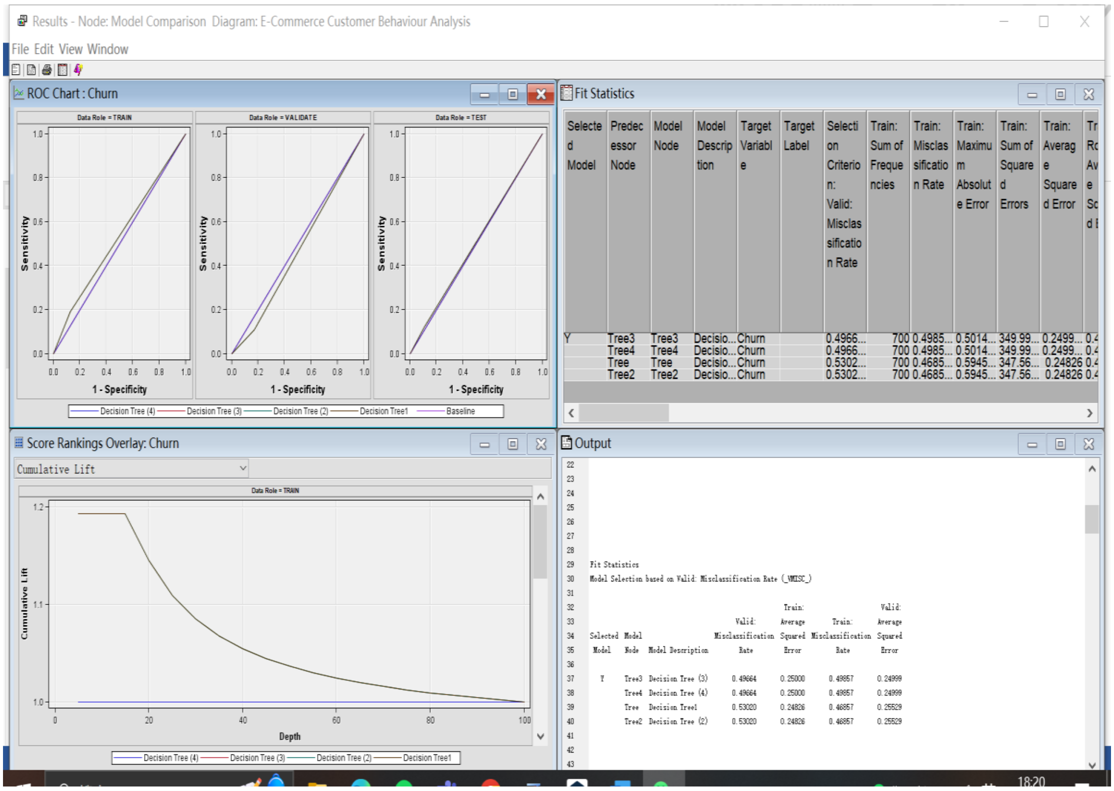
I will elaborate my conclusion according to the following structure

1. decision tree compare
2. Choosing best decision tree analysis-(tree1)
3. Because decision trees don’t perform very well, we move to Gradient Boosting and HP Forest models, Compare these three models.
4. Based on the comparison of the three models, I choose the best models(Gradient Boosting) for analysis
5.Conclusions and business recommendations

1. decision tree compare

I'll start with comparison of different ways to build decision tree





Results - Node: Model Comparison Diagram: E-Commerce Customer Behaviour Analysis

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Tr Ro Av e Sc d E
Y	Tree3	Tree3	Decisio...	Churn		0.4966...	700	0.4985...	0.5014...	349.99...	0.2499...	0.4
	Tree4	Tree4	Decisio...	Churn		0.4966...	700	0.4985...	0.5014...	349.99...	0.2499...	0.4
	Tree	Tree	Decisio...	Churn		0.5302...	700	0.4685...	0.5945...	347.56...	0.24826	0.4
	Tree2	Tree2	Decisio...	Churn		0.5302...	700	0.4685...	0.5945...	347.56...	0.24826	0.4

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

                                                  Train:                      Valid:
                                      Valid:      Average       Train:        Average
Selected  Model                   Misclassification Squared  Misclassification Squared
 Model    Node   Model Description       Rate        Error         Rate         Error

   Y      Tree3  Decision Tree (3)     0.49664     0.25000      0.49857       0.24999
          Tree4  Decision Tree (4)     0.49664     0.25000      0.49857       0.24999
          Tree   Decision Tree1        0.53020     0.24826      0.46857       0.25529
          Tree2  Decision Tree (2)     0.53020     0.24826      0.46857       0.25529
```

- **Model Selection:**

Model selection is based on the Validation Misclassification Rate (VMISC), showing the performance of four different decision tree models. Tree3 and Tree4 have the lowest misclassification rates, but according to the event classification table, these models seem to have not captured any positive events (Churn=1). This suggests that the models might simply be predicting all cases as the majority class (Churn=0).

- **Fit Statistics:**

Kolmogorov-Smirnov Statistic (KS Statistic): For Tree3 and Tree4, the KS statistics for the training, validation, and test sets are all 0, indicating that these models lack discriminative power.

Roc Index: The index is approximately 0.5 for all models, implying that the predictive ability of the models is not different from random guessing.

Cumulative Percent Captured Response: In Tree3 and Tree4, this metric is 0, suggesting that the models did not capture any response from positive events.

Event Classification Table: Tree3 and Tree4 did not correctly predict any cases with Churn=1 in both the training and validation sets, while Tree and Tree2, although capturing some Churn=1 cases, still exhibit overall poor performance.

- **Based on these outputs, here is an assessment of how the decision tree models impact customer behavior analysis:**

These models struggled with handling imbalanced datasets, as accuracy is not a good metric in such cases. Even if the models predict no churn for all customers (Churn=0), accuracy may appear high.

Tree3 and Tree4 may be too simplistic, lacking sufficient complexity to learn patterns in the data, and they might be disregarding the minority class (Churn=1).

While Tree and Tree2 attempted to distinguish between the two classes, they still need adjustments to improve their ability to identify the minority class.

SO, Simply say:

Model Comparison:

- Four models were evaluated: Tree3, Tree4, Tree, and Tree2.
- Tree3 and Tree4 demonstrated very similar performance on training, validation, and test data.
- Tree and Tree2 also exhibited similar performance, but differed from Tree3 and Tree4.

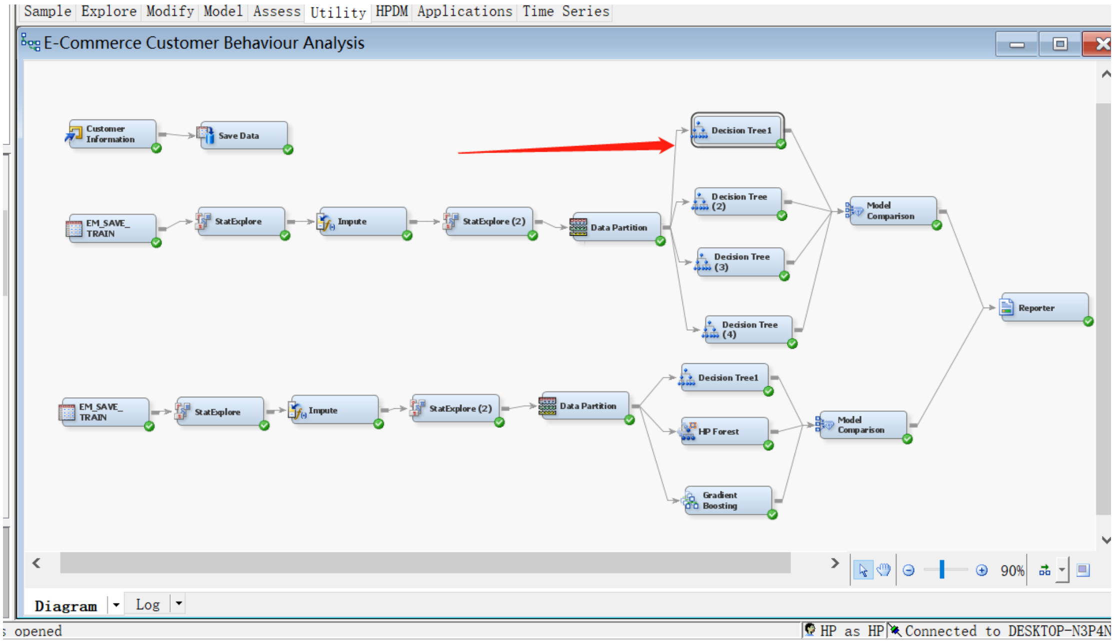
Performance Metrics:

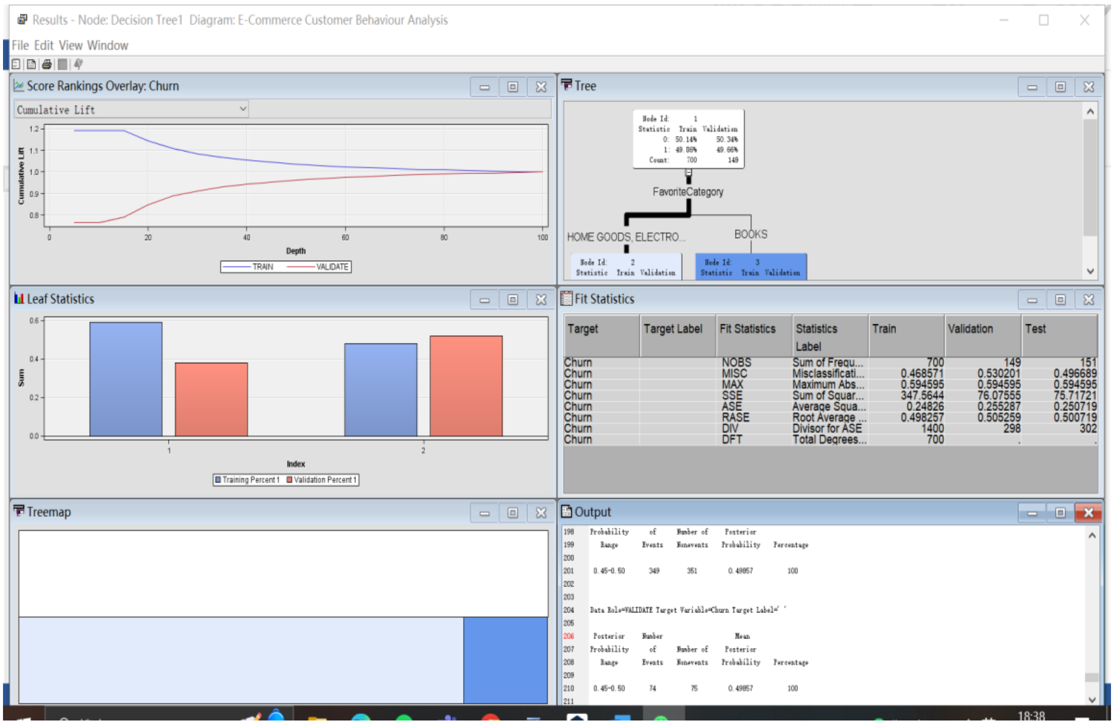
- Misclassification Rate: The misclassification rates for all models on training, validation, and test data ranged approximately between 0.47 and 0.50. This suggests an accuracy of around 50% to 53%, close to random guessing, indicating that the models might not have effectively captured the features of the data.
- ROC Index: The ROC indices for Tree and Tree2 were slightly above 0.50 but still low, indicating limited discriminatory power of the models.
- Average Squared Error: The average squared errors for all models were close to 0.25, aligning with the poorest performance in binary classification (random guessing).

Decision Tree Performance:

- Tree3 and Tree4 seemed ineffective in distinguishing between True Positives and True Negatives, as their True Positive and True Negative values were almost zero on training and validation data.
- While Tree and Tree2 showed some discrimination, their performance remained suboptimal.

2. Choosing best decision tree analysis-(tree1)





```
Node Id:        1
Statistic   Train  Validation
       0:  50.14%      50.34%
       1:  49.86%      49.66%
   Count:     700         149

FavoriteCategory

HOME GOODS, ELECTRO...          BOOKS

Node Id:        2               Node Id:        3
Statistic   Train  Validation   Statistic   Train  Validation
```

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Frequ...	700	149	151
Churn		MISC	Misclassificati...	0.468571	0.530201	0.496689
Churn		MAX	Maximum Abs...	0.594595	0.594595	0.594595
Churn		SSE	Sum of Squar...	347.5644	76.07555	75.71721
Churn		ASE	Average Squa...	0.24826	0.255287	0.250719
Churn		RASE	Root Average ...	0.498257	0.505259	0.500719
Churn		DIV	Divisor for ASE	1400	298	302
Churn		DFT	Total Degrees...	700	.	.

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

 0.45-0.50      349        351        0.49857         100


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

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

 0.45-0.50      74         75         0.49857         100
```

Model Performance:

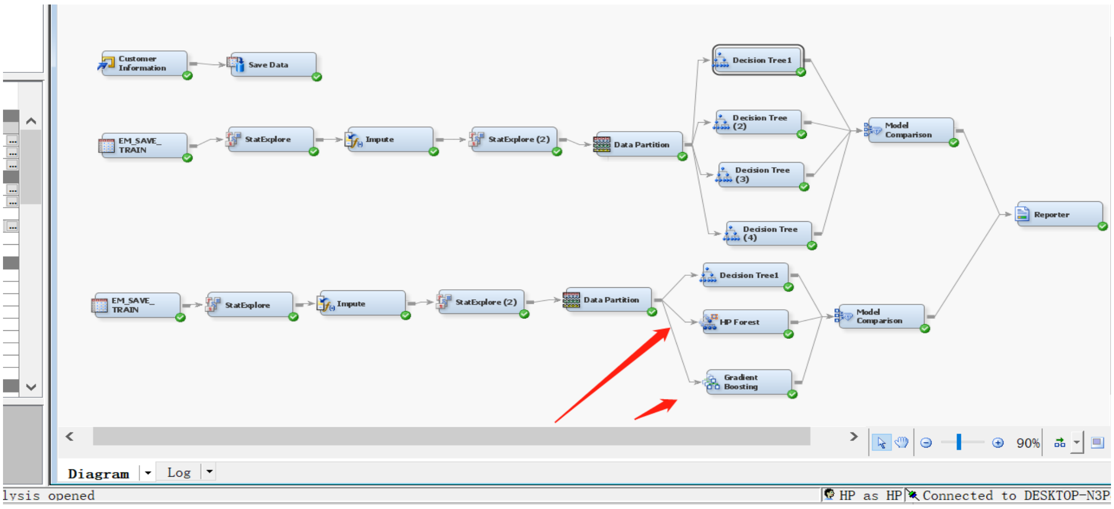
- Misclassification Rate: Approximately 0.5 on the training set, validation set, and test set, indicating poor accuracy of the model.
- Maximum Absolute Error: Close to 0.5 across all datasets, further confirming the low predictive accuracy of the model.
- Mean Squared Error: 0.25 across all datasets, representing one of the worst performances in binary classification problems.

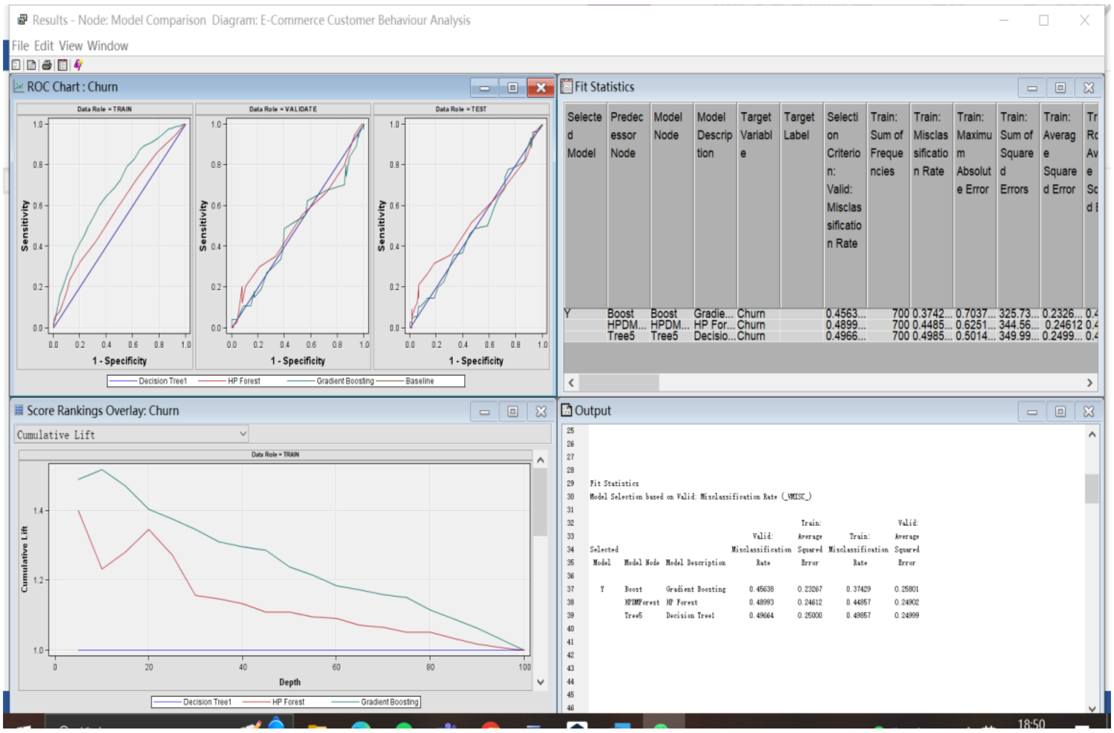
Confusion Matrix:

- The model appears to have failed to successfully distinguish between churn and non-churn customers, as the classification results on training and validation data show that all observations are assigned to the same category.

Overall, the performance is not satisfactory.

3. Because decision trees don’t perform very well, we move to Gradient Boosting and HP Forest models, Compare these three models.





Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Tr Ro Av e Sc d E
Y	Boost	Boost	Gradie...	Churn		0.4563...	700	0.3742...	0.7037...	325.73...	0.2326...	0.4
	HPDM...	HPDM...	HP For...	Churn		0.4899...	700	0.4485...	0.6251...	344.56...	0.24612	0.4
	Tree5	Tree5	Decisio...	Churn		0.4966...	700	0.4985...	0.5014...	349.99...	0.2499...	0.4

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

                                                           Train:                      Valid:
                                            Valid:        Average       Train:        Average
Selected                                 Misclassification Squared  Misclassification Squared
 Model    Model Node  Model Description       Rate          Error        Rate          Error

   Y      Boost       Gradient Boosting     0.45638       0.23267      0.37429       0.25801
          HPDMForest  HP Forest             0.48993       0.24612      0.44857       0.24902
          Tree5       Decision Tree1        0.49664       0.25000      0.49857       0.24999
```

Model Performance Analysis:

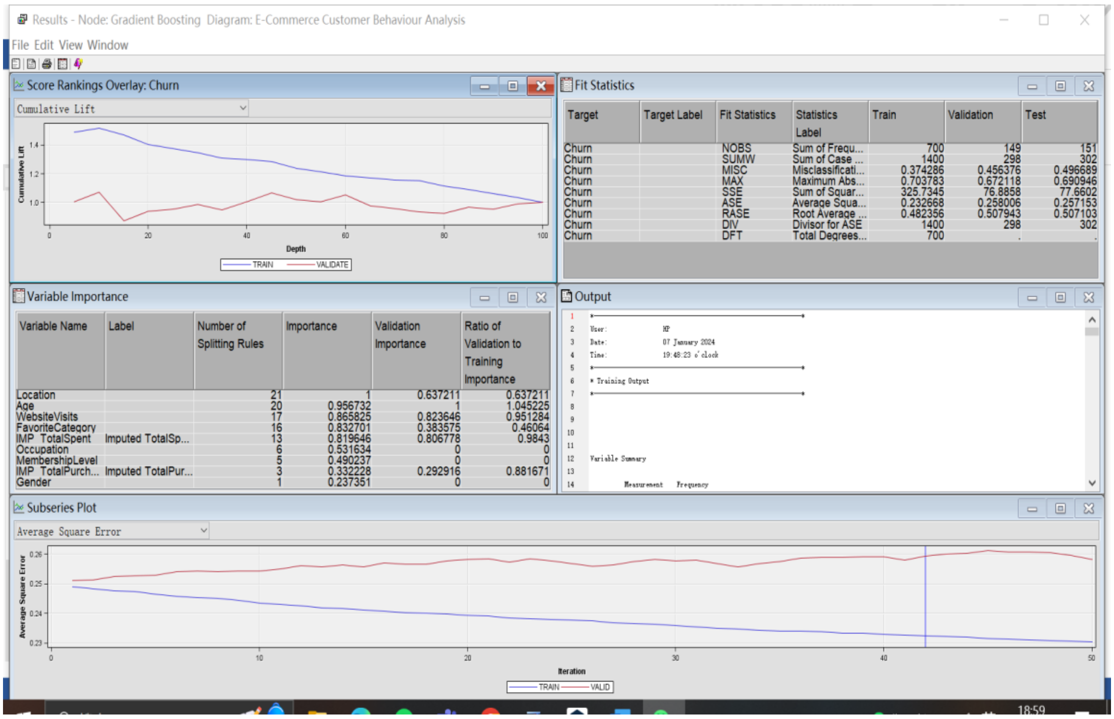
1. Gradient Boosting (Boost)

    - Misclassification Rate: 0.37 on the training data, 0.456 on the validation data, and 0.497 on the test data. This suggests relatively good performance of the model on the validation set.
    - ROC Index: 0.67 on the training data, 0.479 on the validation data, and 0.485 on the test data. The ROC index indicates the model's strong ability to differentiate between positive and negative classes, especially on the training data.
    - Other Important Metrics: Average squared error and Gini coefficient also show some level of model performance.
2. HP Forest (HPDMForest)
    - Misclassification Rate: 0.45 on the training data, 0.490 on the validation data, and 0.477 on the test data. This indicates relatively consistent performance of the model across all datasets.
    - ROC Index: 0.58 on the training data, 0.515 on the validation data, and 0.535 on the test data. These metrics suggest that the model has some classification ability but may not be as strong as the Gradient Boosting model.
3. Decision Tree (Tree5)
    - This model has a misclassification rate of around 0.5 on both training and validation data, indicating poor performance.
    - The ROC index is 0.5, further confirming this.

Overall Analysis:

- The Gradient Boosting model performs the best among these three models, especially on the training data. However, there is a slight decline in performance on the validation and test data, suggesting potential overfitting.
- The HP Forest model shows relatively consistent performance across all datasets but overall performs less well than the Gradient Boosting model.
- The Tree5 model has the poorest performance, with suboptimal results.

4. Based on the comparison of the three models, I choose the best models(Gradient Boosting) for analysis



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Frequ...	700	149	151
Churn		SUMW	Sum of Case ...	1400	298	302
Churn		MISC	Misclassificati...	0.374286	0.456376	0.496689
Churn		MAX	Maximum Abs...	0.703783	0.672118	0.690946
Churn		SSE	Sum of Squar...	325.7345	76.8858	77.6602
Churn		ASE	Average Squa...	0.232668	0.258006	0.257153
Churn		RASE	Root Average ...	0.482356	0.507943	0.507103
Churn		DIV	Divisor for ASE	1400	298	302
Churn		DFT	Total Degrees...	700	.	.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Location		21	1	0.637211	0.637211
Age		20	0.956732	1	1.045225
WebsiteVisits		17	0.865825	0.823646	0.951284
FavoriteCategory		16	0.832701	0.383575	0.46064
IMP_TotalSpent	Imputed TotalSp...	13	0.819646	0.806778	0.9843
Occupation		6	0.531634	0	0
MembershipLevel		5	0.490237	0	0
IMP_TotalPurch...	Imputed TotalPur...	3	0.332228	0.292916	0.881671
Gender		1	0.237351	0	0

Output

```
*------------------------------------------------------*
User:                 HP
Date:                 07 January 2024
Time:                 19:48:23 o'clock
*------------------------------------------------------*
* Training Output
*------------------------------------------------------*




Variable Summary

          Measurement    Frequency
```

(Gradient Boosting)

1. **Model Performance**
    - **Misclassification Rate**: Compared to previous models, this model exhibits a misclassification rate of 0.37 on the training set, 0.456 on the validation set, and 0.497 on the test set, indicating a certain level of predictive capability.
    - **Maximum Absolute Error**: Slightly higher across all datasets, revealing biases in the model's predictions for certain scenarios.
2. The model output highlights the importance of the following variables (in descending order):
    - **Location**: Geographical location emerges as the most crucial factor in predicting customer churn.
    - **Age**: Age is also a significant influencing factor.
    - **WebsiteVisits**: The number of website visits indicates that customer online activity significantly affects their churn risk.
    - **FavoriteCategory and IMP_TotalSpent (Total customer spending)**: These factors also contribute significantly to the model.

These insights suggest that the model places a strong emphasis on geographic location, age, online activity, and customer spending patterns when predicting churn.

5.Conclusions and business recommendations

Based on the model outputs and dataset characteristics, we can derive the following business insights and recommendations:

Key Findings

1. **Significance of Geographic Location**: Geographic location is identified as the most critical factor influencing customer churn. This suggests a significant association between customers' region or living environment and their purchasing behavior and brand loyalty.
2. **Impact of Age**: Age is another key factor affecting customer churn. Consumers in different age groups may have distinct needs, preferences, and purchasing habits.
3. **Role of Website Interaction**: The frequency of website visits is also crucial in predicting customer churn, indicating the potential value of enhancing customer online interaction.
4. **Influence of Spending Behavior**: Total spending and customers' preferred shopping categories are identified as important factors, emphasizing the importance of monitoring and analyzing customer spending patterns.

Business Recommendations

1. **Customized Regional Marketing Strategies**: Develop marketing and service strategies tailored to specific regions or markets based on geographic location data to enhance customer satisfaction and loyalty.
2. **Age-Targeted Marketing**: Create customized products and services for different age groups to meet their unique needs.
3. **Enhance Online Experience**: Optimize website design, add engaging content, provide interactive features, and offer personalized recommendations to increase customer online engagement.
4. **Analysis of Spending Patterns**: Regularly analyze customer purchase history and preferences to identify potential high-value customers and those at risk of churn.
5. **Personalized Communication and Service**: Utilize collected data such as geographic location, age, and purchase history to implement more personalized customer communication and services.

By effectively implementing these recommendations, businesses should be able to manage customer relationships more efficiently, reduce churn, increase customer satisfaction and loyalty, thereby fostering long-term business growth and success.