UNIVERSITI MALAYA

WQD7005 DATA MINING

CASE STUDY

Student Name	Student Matric Number
ZENG ZITONG	22073371

Data sources

Data from Kaggle:https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset

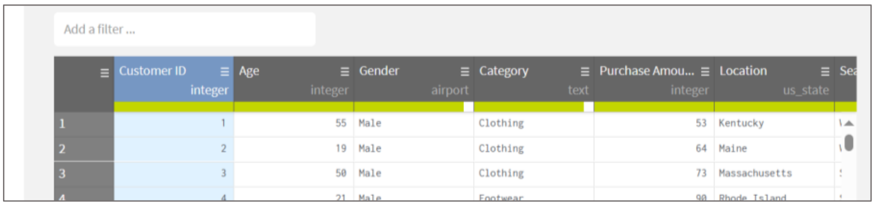
The data contains the following columns:

Column	description
Customer ID	A unique identifier assigned to each individual customer, facilitating tracking and analysis of their shopping behavior over time.
Age	The age of the customer, providing demographic information for segmentation and targeted marketing strategies.
Gender	The gender identification of the customer, a key demographic variable influencing product preferences and purchasing patterns.
Category	The broad classification or group to which the purchased item belongs (e.g., clothing, electronics, groceries).
Purchase Amount (USD)	The monetary value of the transaction, denoted in United States Dollars (USD), indicates the cost of the purchased item(s).
Location	The geographical location where the purchase was made, offering insights into regional preferences and market trends.
Season	The seasonal relevance of the purchased item (e.g., spring, summer, fall, winter), impacting inventory management and marketing strategies.
Review Rating	A numerical or qualitative assessment provided by the customer regarding their satisfaction with the purchased item.
Subscription Status	Indicates whether the customer has opted for a subscription service, offering insights into their level of loyalty and potential for recurring revenue.
Shipping Type	Specifies the method used to deliver the purchased item (e.g., standard shipping, express delivery), influencing delivery times and costs.
Discount Applied	Indicates if any promotional discounts were applied to the purchase, shedding light on price sensitivity and promotion effectiveness.
Previous Purchases	Provides information on the number or frequency of prior purchases made by the customer, contributing to customer segmentation and retention strategies.
Payment Method	Specifies the mode of payment employed by the customer (e.g., credit card, cash), offering insights into preferred payment options.
Frequency of Purchases	Indicates how often the customer engages in purchasing activities, a critical metric for assessing customer loyalty and lifetime value.

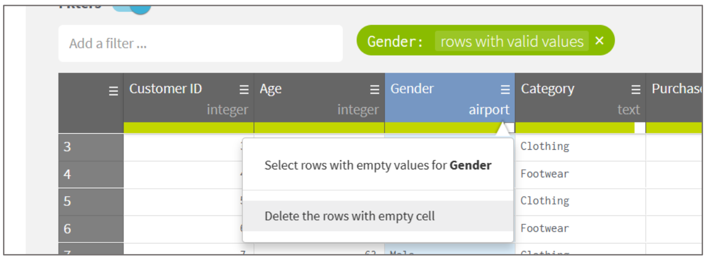
Data preprocessing use Talend

1. Remove missing values

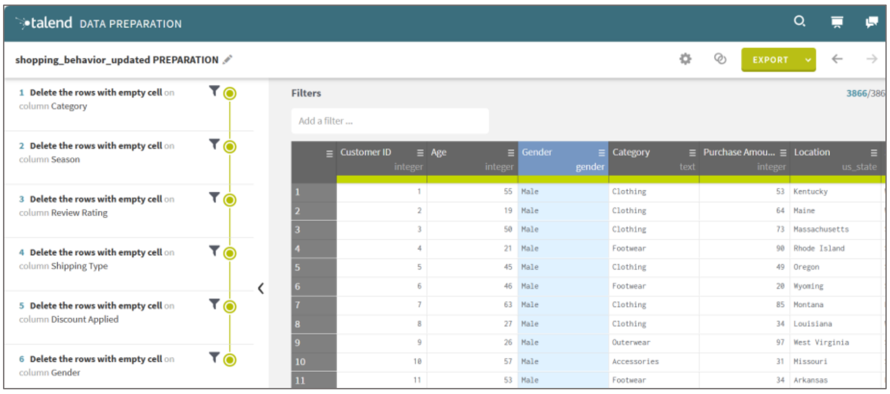
   After uploading the data to Talend, can see that some columns have missing values(with white squares).



   Click ‘delete the rows with empty cell’ to remove rows containing missing values from all columns in turn.



   From the figure below you can see that all missing values are removed.



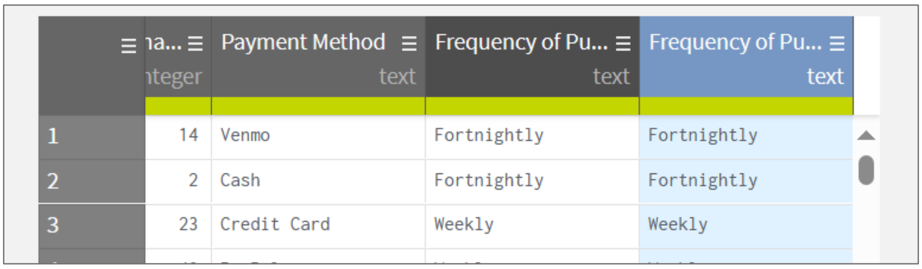
2. Adding a column using Talend

   Observations on the dataset suggest that it is possible to observe which purchasing behaviors affect the users' purchase frequency by setting the purchase frequency columns (yearly, quarterly, bi-weekly, fortnightly, monthly, trimesterly, weekly) as the target columns, which ultimately leads to the classification of the users. However, since this variable is a multivariate categorization variable, and machine learning algorithms such as decision trees are more suitable for dealing with binary classification problems. Here, we will convert a new binary variable churn based on this variable to simplify the problem and make the modeling more intuitive and manageable.

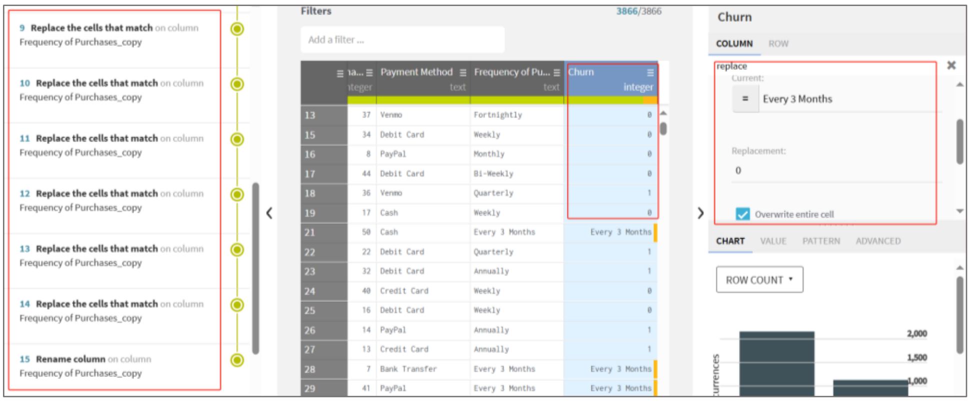
   In the churn column, churn customers are denoted by 1 and active customers by 0. Since low-frequency purchases are associated with a high risk of churn and high-frequency purchases imply customer dependency, the data will be transformed according to the table below.

Frequency of purchases	churn
Annually	1
Quarterly	1
Bi-weekly	0
Fortnightly	0
Monthly	0
Every 3 months	0
Weekly	0

Copy the Frequency of purchases column and change the copied column to churn column.

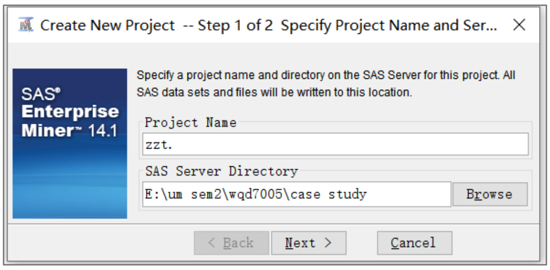


Use the replace function to replace the attributes with 0 or 1 in turn.

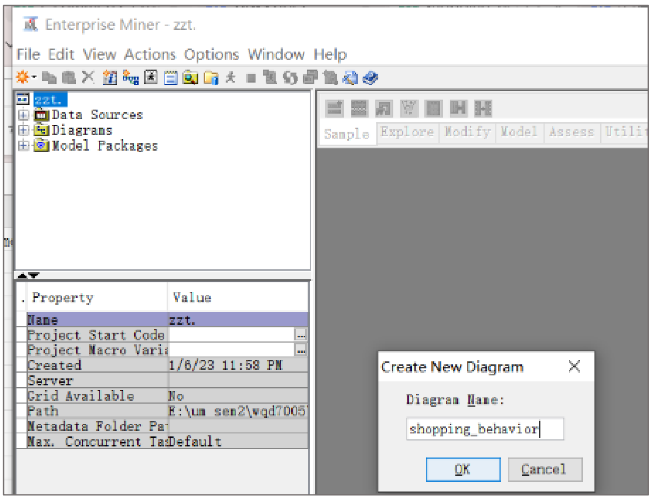


Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

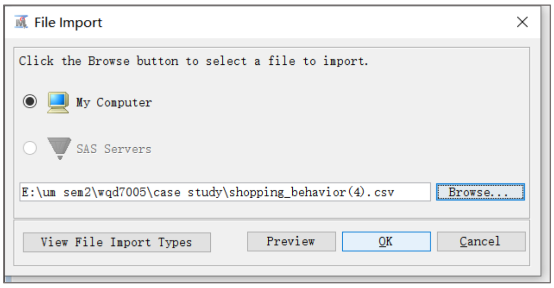
1. Create a new project



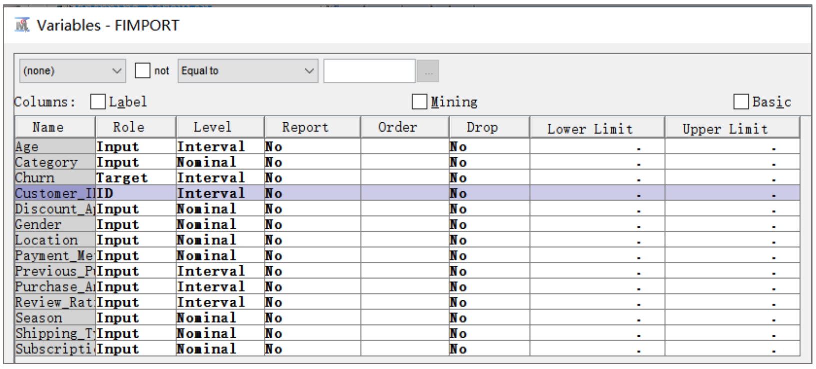
2. Create a new diagram



3. Import data in SAS Enterprise Minner

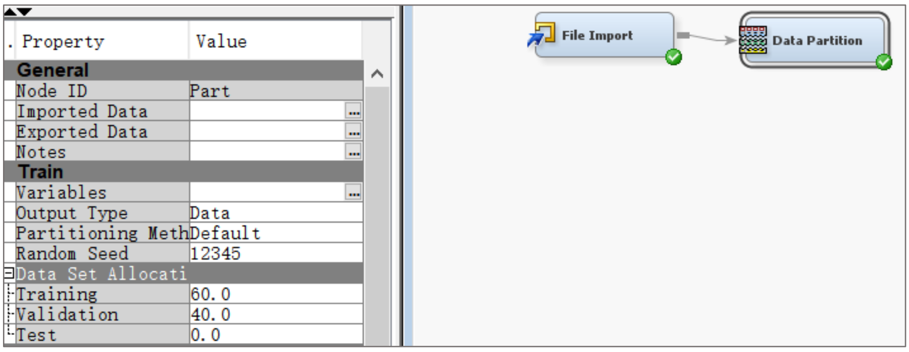


4. specify variable roles in SAS
   Before modeling, the role of each variable in the dataset needs to be defined. Here, the role of churn is set to Target; the role of Customer_ID is set to ID to prevent misuse; and the level of numeric variables are set to Interval.

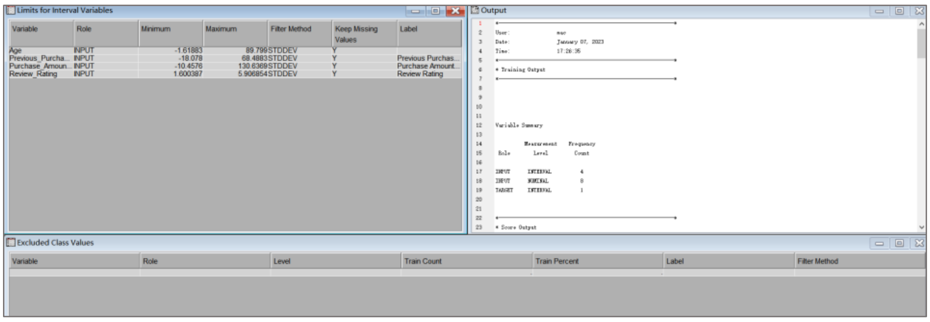


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Category	Input	Nominal	No		No	.	.
Churn	Target	Interval	No		No	.	.
Customer_I	ID	Interval	No		No	.	.
Discount_A	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
Payment_Me	Input	Nominal	No		No	.	.
Previous_P	Input	Interval	No		No	.	.
Purchase_A	Input	Interval	No		No	.	.
Review_Rat	Input	Interval	No		No	.	.
Season	Input	Nominal	No		No	.	.
Shipping_T	Input	Nominal	No		No	.	.
Subscripti	Input	Nominal	No		No	.	.

5. Data partition
   Prior to modeling, the data is divided into a training set and a validation set to prepare for the subsequent evaluation of the model and validation of the generalization performance. Since the dataset has 3900 rows, which is a sufficient amount of data, 60% of the training set and 40% of the validation set are used here.

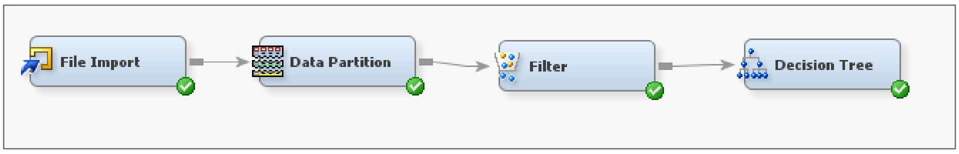


6. Viewing outliers using box plots in SAS
   View outliers through the filter node, and find that there are no outliers in the dataset.

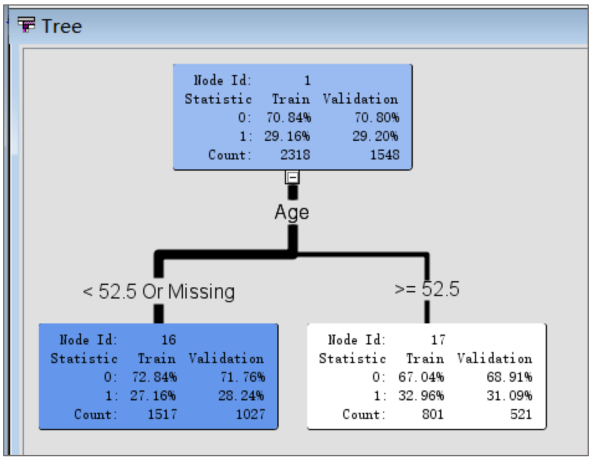


Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

Drag and drop the "Decision Tree" node from the Node Toolbox into the graph, and use the arrows to connect to the filtered data source.

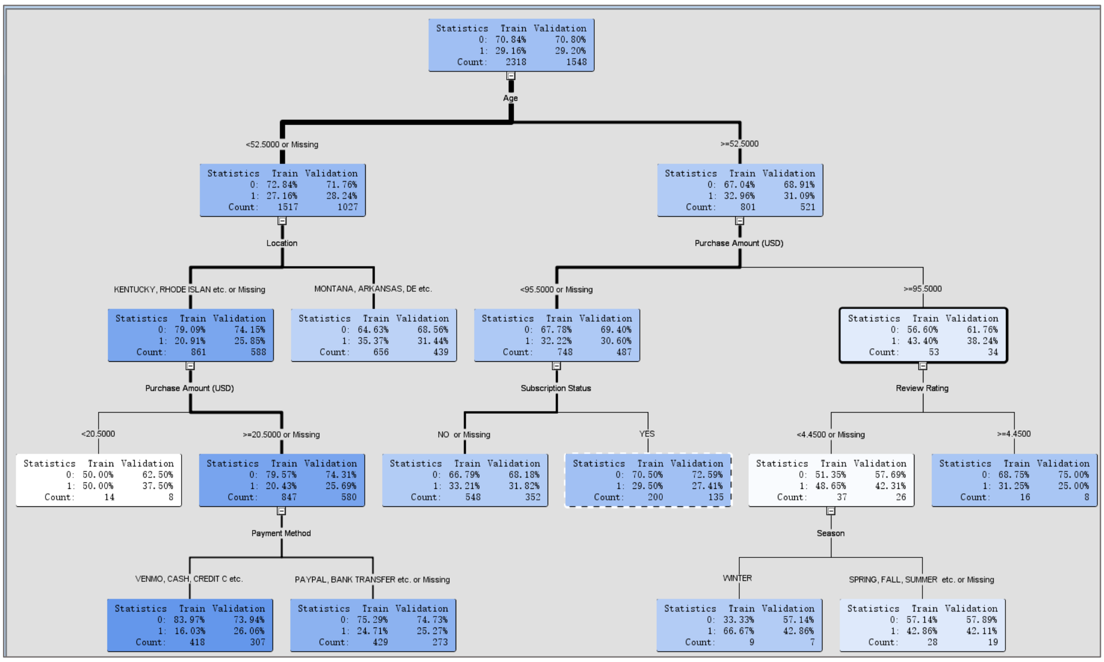


Right click on the decision tree node and click run to run the node. After the run is complete, click result to view the decision tree, as shown in the following figure.



From this decision tree, it can be seen that the data is first divided into clusters based on age, which indicates that age is the factor that most influences whether or not a user churns. When age is greater than 52.5, the percentage of churned customers is significantly higher than the group of users whose age is less than 52.5.

Expanding the decision tree with the 'split node' operation in the interactive interface to gain more observations. This is shown in the figure below.



This decision tree utilizes more features to get a more detailed look. From the tree, it can be seen that age is the factor that most influences whether a user churns or not, followed by Location and Purchase Amount(USD). The effects of Location and Purchase Amount are considered separately within each age group. When age is less than 52.5, there are more churned users living in MONTANA, ARKANSAS, etc. When age is more than 52.5, there are more churned users with Purchase Amount more than 95.5.

Secondly, in the third level, Purchase Amount(USD), Subscription Status and Review Rating are considered as the factors affecting the churn. Most of the users are active when they live in KENTUCKY, RHODE LSLAN, etc. and their Purchase Amount(USD) is more than 20.5. Considering the Payment Methods of these users, it can be seen that the users who pay by VENMO, CASH CREDIT, etc. are more active users. When Purchase Amount(USD) is less than 95.5, there are more active users when Subscription Status is YES. When Purchase Amount(USD) is greater than 95.5, review rating value is less than 4.45, obviously the proportion of churn users increases. Considering the seasonal factor further, we can see that the churn rate increases when the season is Winter.

suggestions for business strategy:

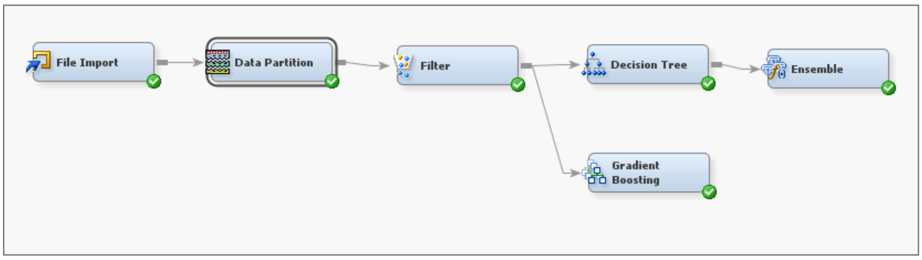
Using a decision tree analysis, here are some possible business insights and courses of action:

- Older users are more prone to churn, so consider designing more engaging marketing strategies and customized services for this age group.
- Areas such as Montana have a higher rate of user churn. This may be related to economic conditions, cultural differences, or service accessibility in the region. Market conditions in these regions can be analyzed and enhanced marketing activities in these regions can be considered.
- High purchase amount subscribers are more inclined to remain loyal, and a more personalized retention strategy could be implemented for high value subscribers.
- Users with subscriptions tend not to churn. Reinforce the value of subscription services to encourage more users to subscribe.
- Users who use specific payment methods (e.g. Venmo, cash credit, etc.) are more active. Explore the characteristics of users with these payment methods and offer targeted offers or rewards.
- Churn increases during the winter months. Launch special campaigns or offers during winter to increase user stickiness.

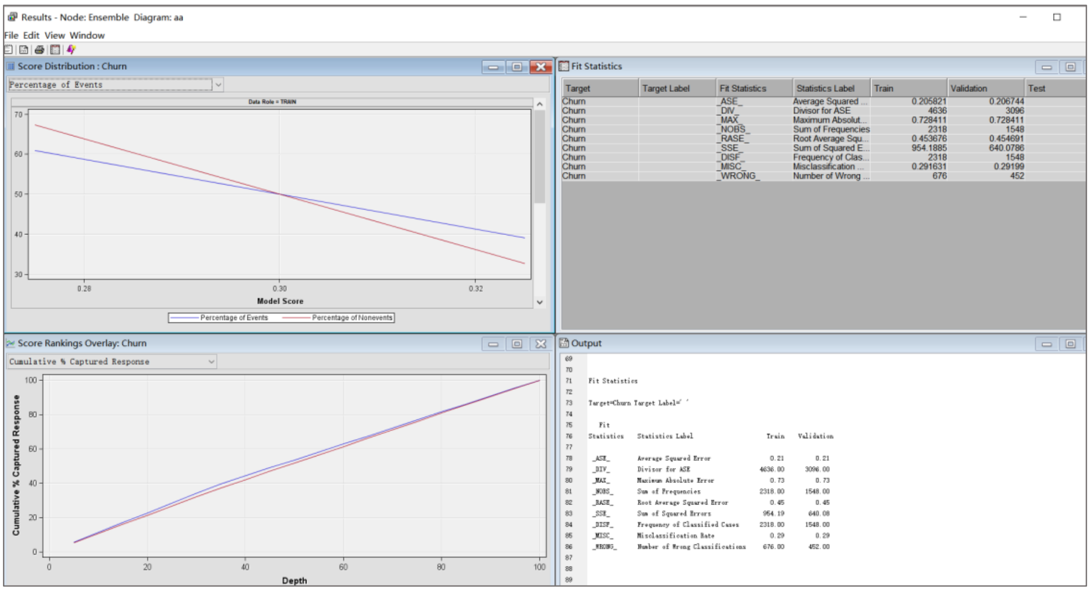
In summary, these insights can help develop more effective customer retention strategies, improve marketing effectiveness in target markets, and provide more personalized services to specific user groups. By doing so, customer satisfaction can be improved, churn can be reduced, and overall revenue can be increased.

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

Drag the ensemble node into the graph to connect to the decision tree node; drag the gradient boosting node into the graph to connect to the filter node.



1. Random forest applying bagging



From the graphs, it can be seen that the overlap between the training set and the validator is high, indicating that the predictive ability of the model is good.

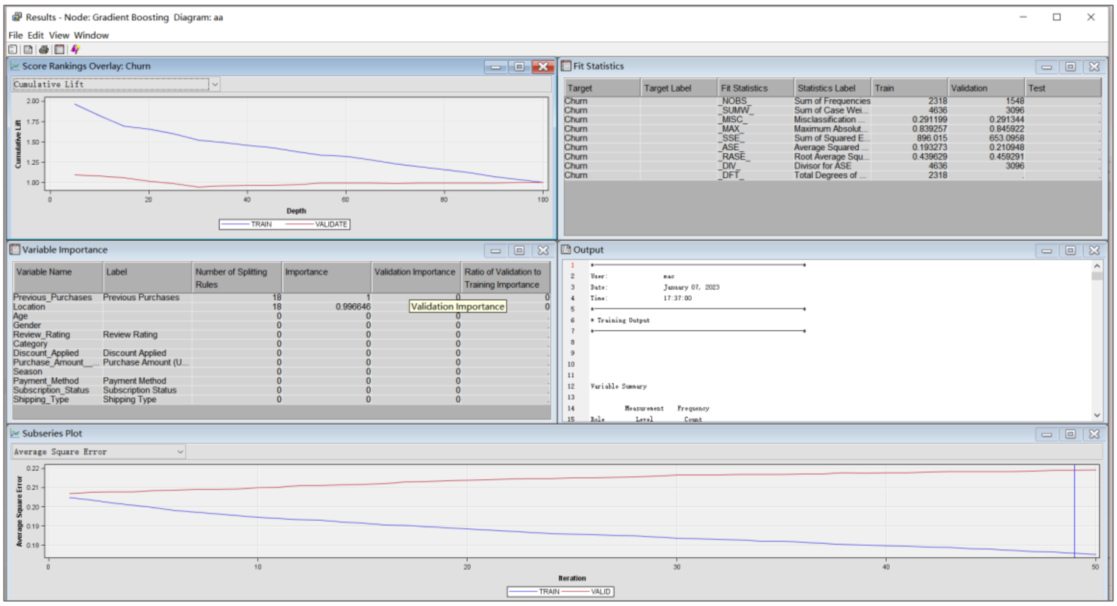
Fit Statistics

Target=Churn Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.21	0.21
DIV	Divisor for ASE	4636.00	3096.00
MAX	Maximum Absolute Error	0.73	0.73
NOBS	Sum of Frequencies	2318.00	1548.00
RASE	Root Average Squared Error	0.45	0.45
SSE	Sum of Squared Errors	954.19	640.08
DISF	Frequency of Classified Cases	2318.00	1548.00
MISC	Misclassification Rate	0.29	0.29
WRONG	Number of Wrong Classifications	676.00	452.00

From the fit statistics table, it can be seen that the random forest model shows similar error and accuracy on both the training and validation sets, which is a good sign, meaning that the model is not overfitting. The error metrics (e.g. ASE and RaSE) show that the model has some predictive power, but the 29% misclassification rate suggests that there is room for improvement.

2. Gradient Boosting



From the figure, it can be seen that the performance difference between the two lines is large, which is a manifestation of model instability, and needs to be optimized by subsequent tuning parameters.

Fit Statistics

Target=Churn Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2318.00	1548.00
SUMW	Sum of Case Weights Times Freq	4636.00	3096.00
MISC	Misclassification Rate	0.29	0.29
MAX	Maximum Absolute Error	0.84	0.85
SSE	Sum of Squared Errors	896.02	653.10
ASE	Average Squared Error	0.19	0.21
RASE	Root Average Squared Error	0.44	0.46
DIV	Divisor for ASE	4636.00	3096.00
DFT	Total Degrees of Freedom	2318.00	.

These statistics indicate that the model has an error rate of close to 30% on both the training and validation sets.The values of ASE and RASE indicate that the model has some degree of error in its predictions.