

Name: Chew Chi Hsiang

Matric Number: s2153581

GitHub Link: [https://github.com/chchewmy/WQD7005\\_Data\\_Mining/](https://github.com/chchewmy/WQD7005_Data_Mining/)**ALTERNATIVE ASSESSMENT 1 (50 marks) - WEEK 12**

Answer the question below based on the given scenario. Submit your answer within ONE (1) DAY after the question is given in SPECTRUM. Answers should be submitted and saved with the student's name followed by matric number as the file name in the format of .pdf (e.g. Ali\_s123456.pdf).

**Case Study: E-Commerce Customer Behaviour Analysis****Background:**

You will work with a dataset of customer transactions from an e-commerce website, encompassing various customer attributes and purchase history over the last year. The structure provided below is a guideline. Feel free to enhance this dataset by adding relevant attributes that you believe will enrich your analysis. Use the structure as a foundation to create your own sample dataset that reflects realistic customer behaviour.

**Dataset Structure:**

CustomerID: Unique identifier for each customer.

Age: Age of the customer.

Gender: Gender of the customer.

Location: Geographic location of the customer.

MembershipLevel: Indicates the membership level (e.g., Bronze, Silver, Gold, Platinum).

TotalPurchases: Total number of purchases made by the customer.

TotalSpent: Total amount spent by the customer.

FavoriteCategory: The category in which the customer most frequently shops (e.g., Electronics, Clothing, Home Goods).

LastPurchaseDate: The date of the last purchase.

[Additional Attributes]: Consider adding more attributes like customer's occupation, frequency of website visits, etc.

Churn: Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

**Objective:**

The case study aims to assess students' ability to apply decision tree and ensemble methods in a practical context, demonstrating their understanding of the concepts and their ability to derive meaningful business insights from data analysis.

**Tasks:**

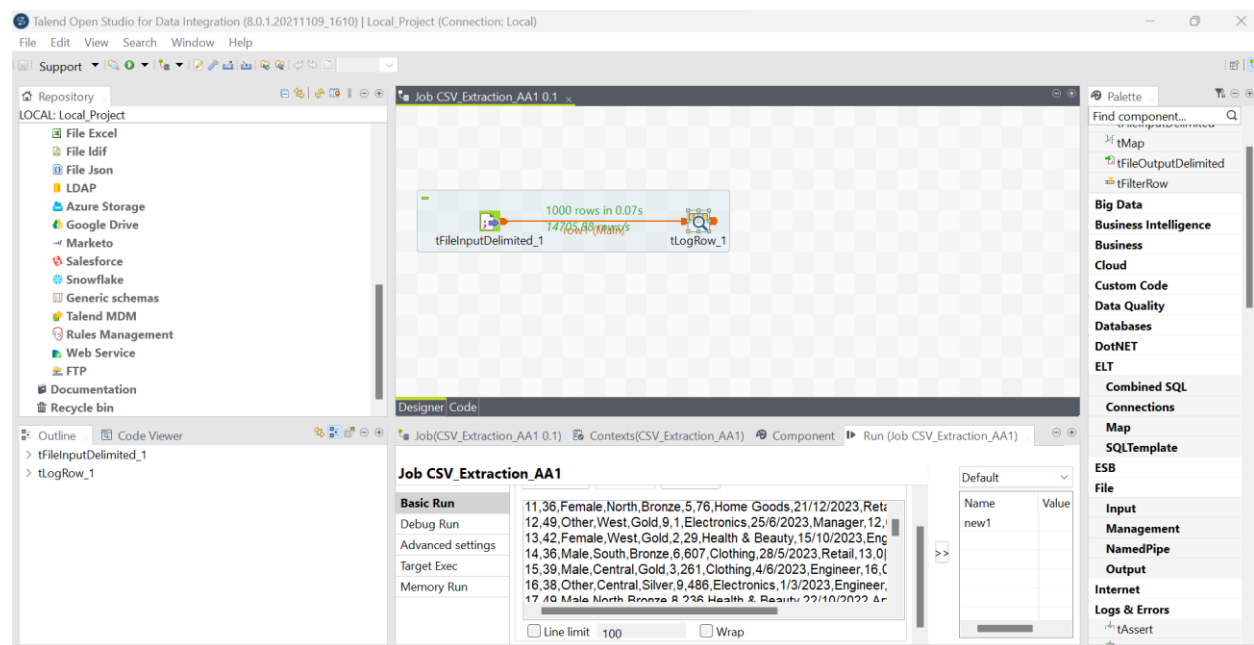
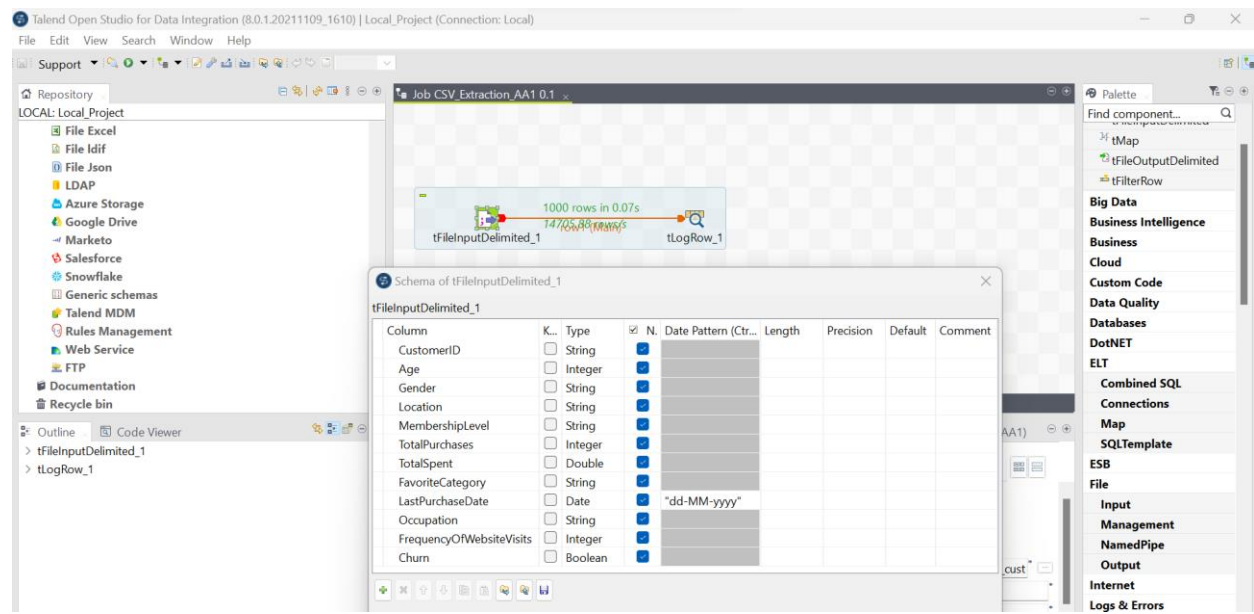
- 1. Data Import and Preprocessing:** Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

**Answer:****Sample Phase:**

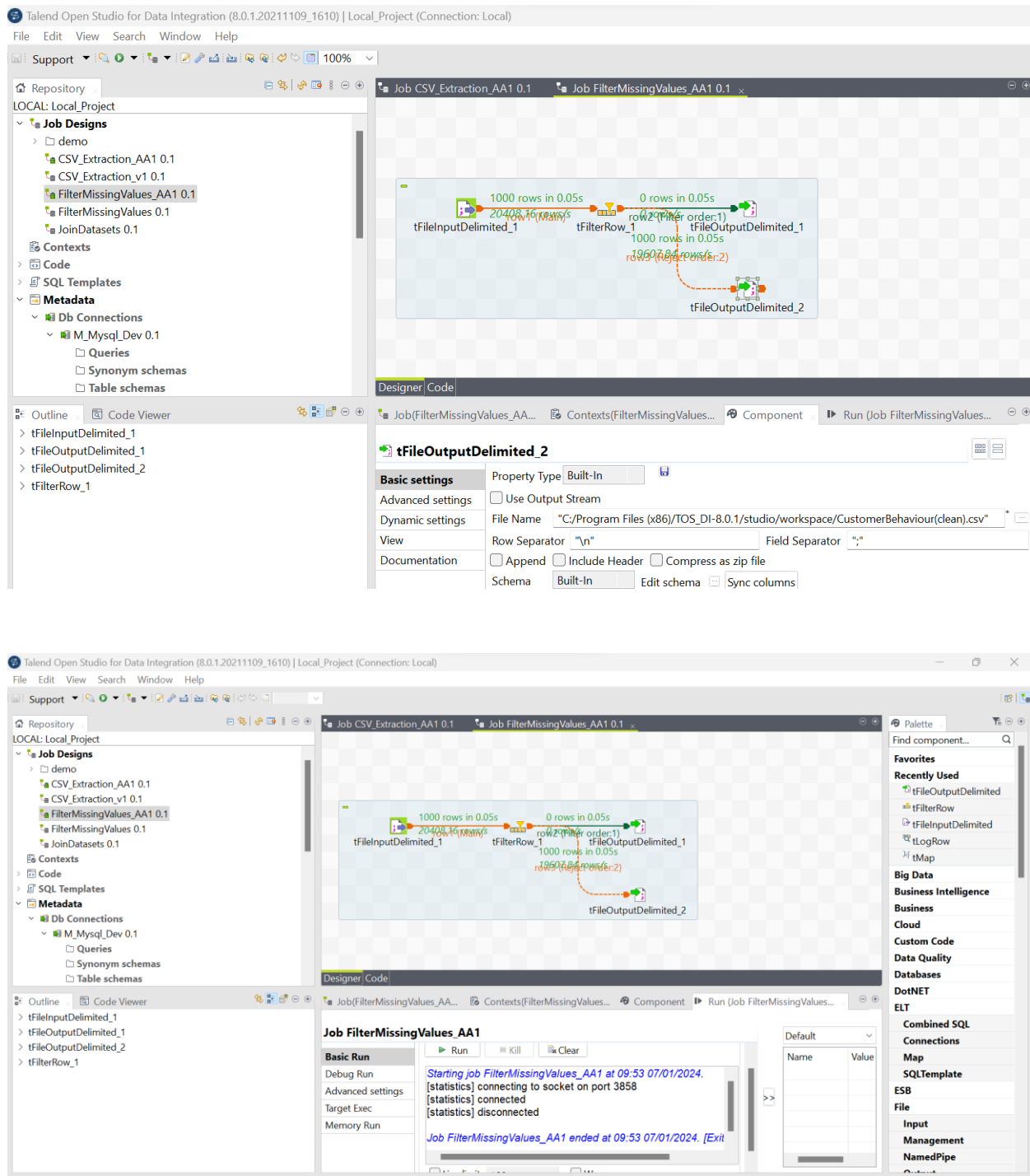
The dataset is created using a data generator to create since obtaining the real-world data is challenging to align with the specific data requirements. Hence, data generators can help in creating a dataset that closely matches the AA1's criteria and allows for a controlled environment to test various analytical methods in order to represent the actual e-commerce customer behavior to make the analysis more meaningful and applicable to the real-world scenarios.

## In Talend Data Integration:

- Extract data from the provided CSV file for the raw dataset.



- Transform the data by filtering out records with missing values and save the output under the designated folder.



## In Talend Data Preparation:

- **Data Profiling:** A technique that involves analysing data in order to gain an understanding of its structure, content, relationships, and quality abnormalities is referred to as data profiling. After completing this phase, you will have a comprehensive view of the present state of the data, which will highlight areas that require attention or improvement.
- **Data Cleaning:** The process of finding and repairing (or deleting) mistakes and inconsistencies in data in order to improve the quality of the data includes the process of data cleaning.
- Incoming or existing data must be validated to ensure that it satisfies the standards and criteria that have been established beforehand.

Profiling and understand the value and pattern of all the columns:

**talend DATA PREPARATION**

**Customer Behaviours**

Filters: Add a filter ...

	CustomerID	Age	Gender	Location	MembershipLevel	TotalPurchases	TotalSpent
	integer	integer	gender	last_name	city	integer	integer
2	2	39	Other	West	Silver	5	
3	3	44	Other	South	Gold	7	
4	4	57	Other	South	Gold	4	
5	5	53	Female	East	Gold	6	
6	6	25	Female	Central	Bronze	11	
7	7	44	Female	North	Platinum	6	
8	8	33	Female	Central	Gold	4	
9	9	33	Male	East	Gold	2	
10	10	39	Female	South	Silver	5	
11	11	36	Female	North	Bronze	5	
12	12	49	Other	West	Gold	9	
13	13	42	Female	West	Gold	2	
14	14	36	Male	South	Bronze	6	
15	15	39	Male	Central	Gold	3	

CustomerID

COLUMN ROW

Find a function ...

SUGGESTIONS

Compare numbers...

Add, multiply, subtract or divide...

BOOLEAN

CHART VALUE PATTERN ADVANCED

Count: 1000 Min: 1

Distinct: 1000 Max: 1000

Duplicate: 0 Mean: 500.5

Valid: 1000 Variance: 83416.67

Empty: 0 Median: 500.5

Invalid: 0 Lower quantile: 250.25

Upper quantile: 750.75

Apply Date Standardization: With the 'LastPurchaseDate' column selected, look for the "Actions" or "Functions" pane on the right pane. Search for or locate the "Change Date Format" or a similar function related to date transformations. In the transformation options, specify the source format(s) and the target format (DD/MM/YYYY) and apply the transformation.

**Customer Behaviours**

Filters: Add a filter ... Gender: rows with valid values

	MembershipLevel	TotalPurchases	TotalSpent	FavoriteCategory	LastPurchaseDate	Occupation	FrequencyOfWe...
	city	integer	integer	text	date	job_title	integer
1		2	124	Home Goods	15/4/2022	Doctor	
5		6	37	Health & Beauty	15/12/2023	Artist	
6		11	243	Home Goods	20/7/2022	Teacher	
7		6	577	Electronics	15/1/2023	Retail	
8		4	675	Clothing	13/6/2023	Retail	
9		2	235	Clothing	16/6/2023	Engineer	
10		5	15	Clothing	6/7/2023	Teacher	
11		5	76	Home Goods	21/12/2023	Retail	
13		2	29	Health & Beauty	15/10/2023	Engineer	
14		6	607	Clothing	28/5/2023	Retail	
15		3	261	Clothing	4/6/2023	Engineer	
17		8	236	Health & Beauty	22/10/2022	Artist	
18		0	537	Health & Beauty	30/3/2023	Doctor	
19		10	1251	Home Goods	10/12/2022	Unemployed	

LastPurchaseDate

COLUMN ROW

change date format

Define my own format

Your format: DD/MM/YYYY

Apply changes to: ☐ All rows ☒ Filtered rows

CHART VALUE PATTERN ADVANCED

200 400 600 800 1,000

d/M/yyyy

M/d/yyyy

dd/MM/yyyy

MM/dd/yyyy

After applying the date format change, validate the transformation by reviewing the 'LastPurchaseDate' column to ensure that all dates are now in the desired format and examine any anomalies or unexpected formats. These could arise if there were date entries that didn't match the source formats that specified. Export the dataset once the dataset has been completed all the required data profiling, cleansing, and validation processes.

**talend DATA PREPARATION**

**e\_commerce\_customer\_data PREPARATION**

1 Change date format on column LastPurchaseDate

Filters: Gender: rows with valid values (652/1000)

	lalPurchases	TotalSpent	FavoriteCategory	LastPurchaseDate
	integer	integer	text	date
1	2	124	Home Goods	15/4/2022
5	6	37	Health & Beauty	15/12/2023
6	11	243	Home Goods	20/7/2022
7	6	577	Electronics	15/1/2023
8	4	675	Clothing	13/6/2023
9	2	235	Clothing	16/6/2023
10	5	15	Clothing	6/7/2023
11	5	76	Home Goods	21/12/2023
13	2	29	Health & Beauty	15/10/2023
14	6	607	Clothing	28/5/2023
15	3	261	Clothing	4/6/2023
17	8	236	Health & Beauty	22/10/2022
18	0	537	Health & Beauty	30/3/2023
19	10	1251	Home Goods	10/12/2022

**LastPurchaseDate**

COLUMN ROW

Find a function ...

SUGGESTIONS

Delete these filtered rows

Keep these filtered rows

Apply changes to: ☐ All rows ☒ Filtered rows

CHART VALUE PATTERN ADVANCED

0 200 400 600 800 1,000

d/M/yyyy

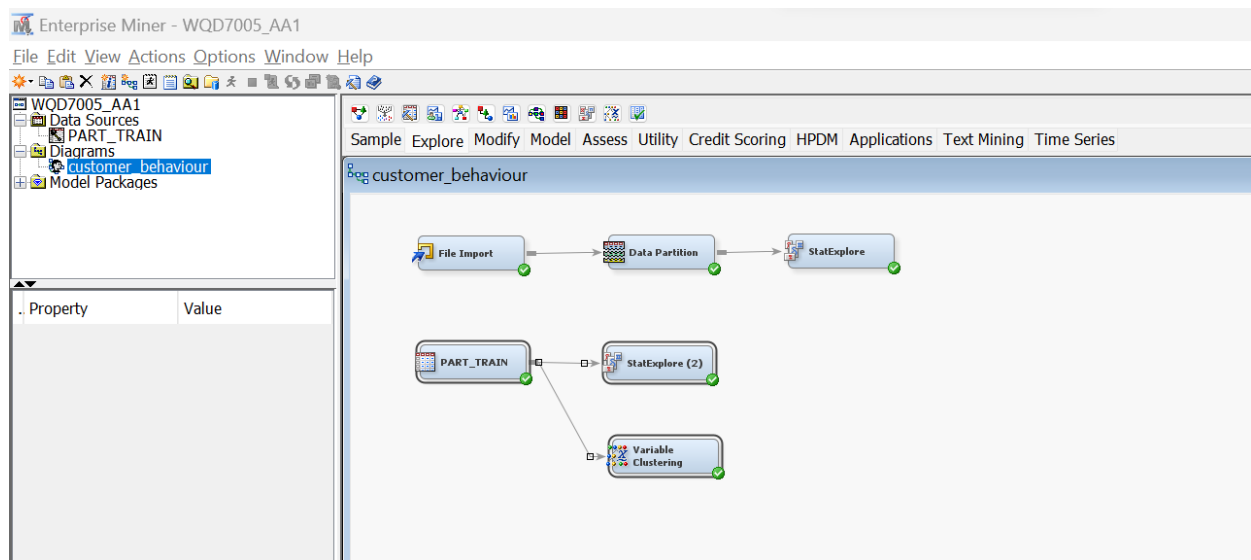
M/d/yyyy

dd/M/yyyy

M/d/yyyy

## Explore Phase:

## In SAS Enterprise Miner



Select Random at the Sample Method. Set Max for Fetch Size. This 700-rows sample now has distributional properties that are similar to the original 1000 observation table. This gives us an idea about the general characteristics of the variables. The random sampling method improves on the default method (at the top of the dataset) with output as below:

Results - Node: PART\_TRAIN Diagram: customer\_behaviour

File Edit View Window

Output

```

1 *-----*
2 User:      u63452954
3 Date:      06 January 2024
4 Time:      16:14:28
5 *-----*
6 * Training Output
7 *-----*
8
9
10
11
12 Variable Summary
13
14 Measurement Frequency
15 Role Level Count
16
17 ID NOMINAL 1
18 INPUT INTERVAL 4
19 INPUT NOMINAL 5
20 REJECTED INTERVAL 1
21 TARGET BINARY 1
22 TIMEID INTERVAL 1
23
24
25

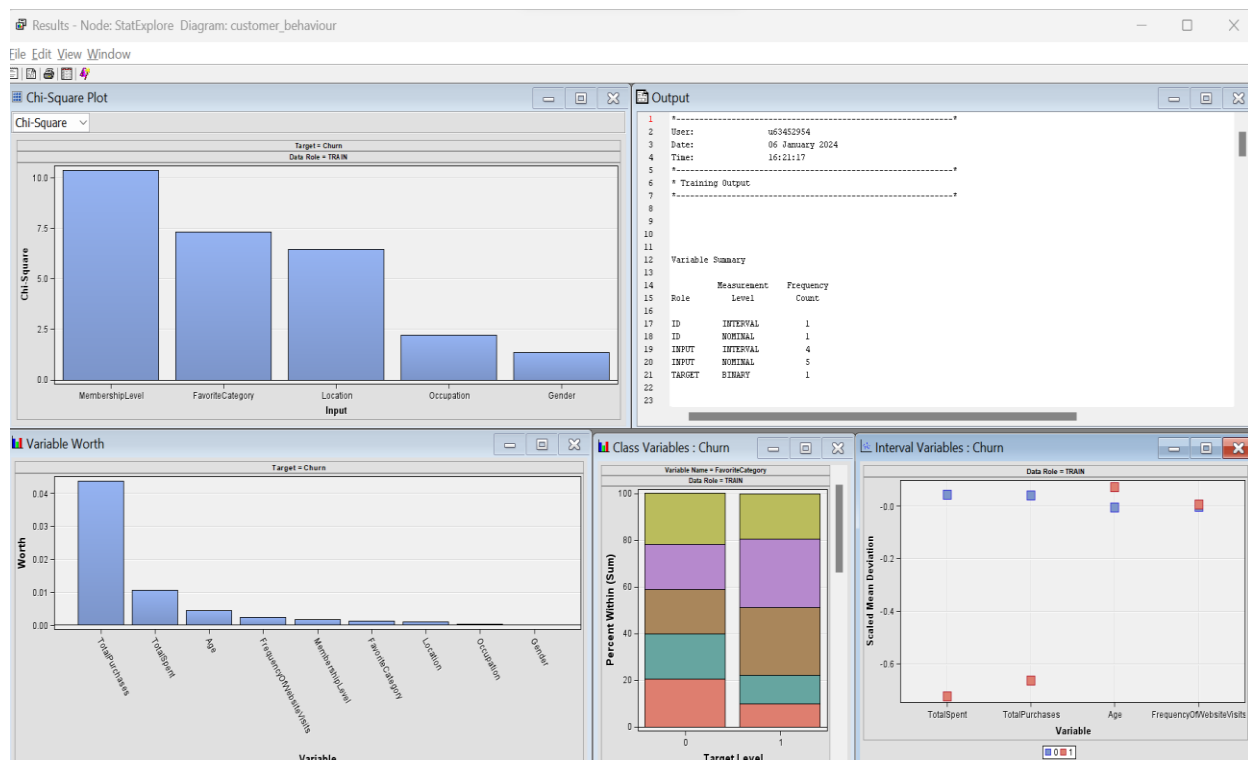
```

Variables

Variable Name	Role	Measurement Level	Order	Label	Drop
Age	Input	Interval			No
Churn	Target	Binary			No
CustomerID	ID	Nominal			No
FavoriteCategory	Input	Nominal			No
FrequencyOfWebsiteVisits	Input	Interval			No
Gender	Input	Nominal			No
LastPurchaseDate	Time ID	Interval			No
Location	Input	Nominal			No
MembershipLevel	Input	Nominal			No
Occupation	Input	Nominal			No
TotalPurchases	Input	Interval			No
TotalSpent	Input	Interval			No
dataobs	Rejected	Interval		Observation Number	Yes

## For Original Cleansed Dataset:

- StatExplore node:



The bar chart (Square Plot) likely shows the frequency or proportion of a categorical variable. If it represents churn, the bars may indicate how many customers churned versus those who didn't. Large discrepancies in the bar sizes would suggest class imbalance, which is important to address during model training.

The variable worth plot typically ranks variables based on their predictive power regarding the target variable. A higher bar indicates a variable is more important for predicting the outcome. For instance, "TotalPurchases" has the highest bar, it's a key variable for predicting churn.

The Output is probably a summary of model performance or variable statistics. It contains a confusion matrix or classification report that could look at the accuracy, precision, recall, and F1-score to assess the model. High values indicate good performance, while low values may suggest areas for improvement.

Chi-Square Statistics  
(maximum 500 observations printed)

Data Role=TRAIN Target=Churn

Input	Chi-Square	Df	Prob
MembershipLevel	10.4058	3	0.0154
FavoriteCategory	7.3122	4	0.1203
Location	6.4829	4	0.1659
Occupation	2.2231	6	0.8981
Gender	1.3514	2	0.5088

The above Chi-Square test results indicate that among the variables tested, Membership Level has a statistically significant relationship with Churn. This implies that changes in the Membership Level could relate to changes in churn behavior. For instance, customers with different membership levels may have different churn rates, which could guide the development of membership-related retention programs.

For the other variables (FavoriteCategory, Location, Occupation, Gender), the results suggest that they may not be strong predictors of churn. However, it's important to note that lack of statistical significance in a Chi-Square test does not necessarily mean these variables have no practical significance. They might have predictive power in conjunction with other variables or in more complex models.

The mosaic plot for class variables: Churn show the distribution of the target variable 'Churn' across different levels of a categorical variable. Significant patterns might indicate associations between categorical features and the likelihood of churn.

The scatter plot for interval variables against churn might be looking at the relationship between continuous variables and churn. Correlation patterns here can suggest potential predictors of churn.

## Basic Statistic Summary:

Results - Node: StatExplore Diagram: customer\_behaviour

File Edit View Window

Output

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
ID	NOMINAL	1
INPUT	INTERVAL	4
INPUT	NOMINAL	5
TARGET	BINARY	1

Variable Levels Summary  
(maximum 500 observations printed)

Variable	Role	Frequency Count
Churn	TARGET	2
CustomerID	ID	700
_dataobs_	ID	700

Class Variable Summary Statistics  
(maximum 500 observations printed)

Data Role=TRAIN

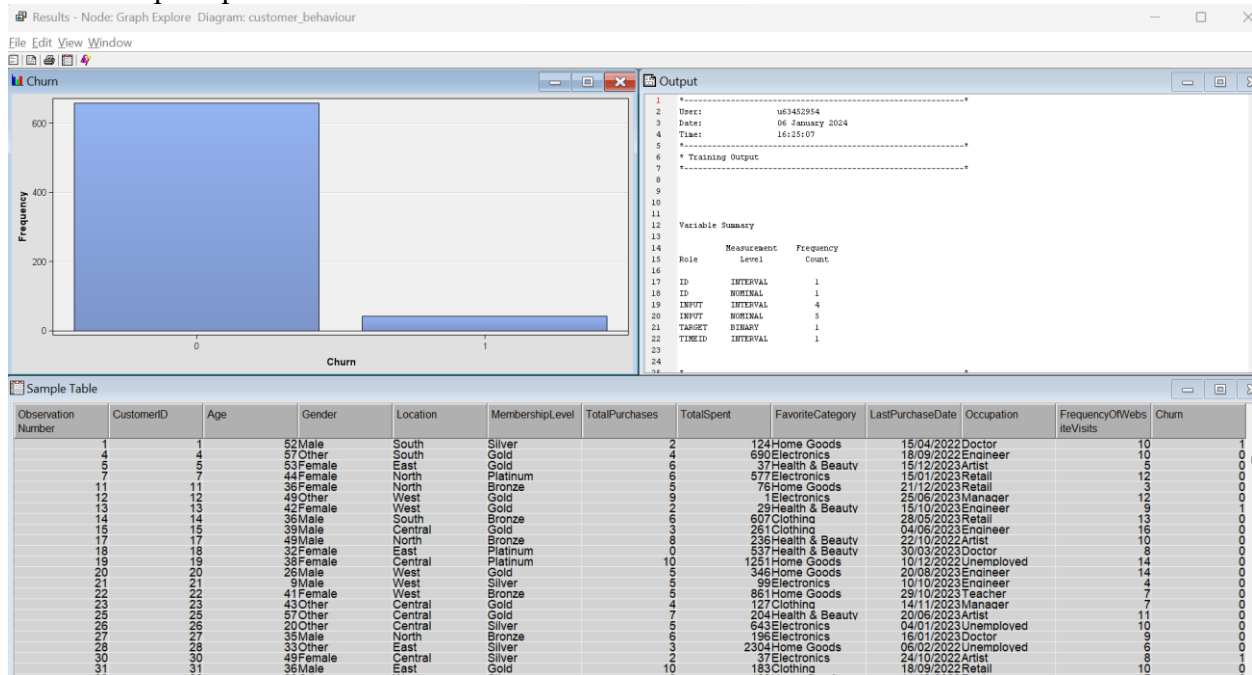
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	FavoriteCategory	INPUT	5	0	Home Goods	21.71	Health & Beauty	20.00
TRAIN	Gender	INPUT	3	0	Female	34.00	Other	33.86
TRAIN	Location	INPUT	5	0	North	21.86	Central	20.71
TRAIN	MembershipLevel	INPUT	4	0	Bronze	37.29	Silver	29.86
TRAIN	Occupation	INPUT	7	0	Artist	16.00	Doctor	15.14
TRAIN	Churn	TARGET	2	0	0	94.14	1	5.86

Distribution of Class Target and Segment Variables  
(maximum 500 observations printed)

Data Role=TRAIN

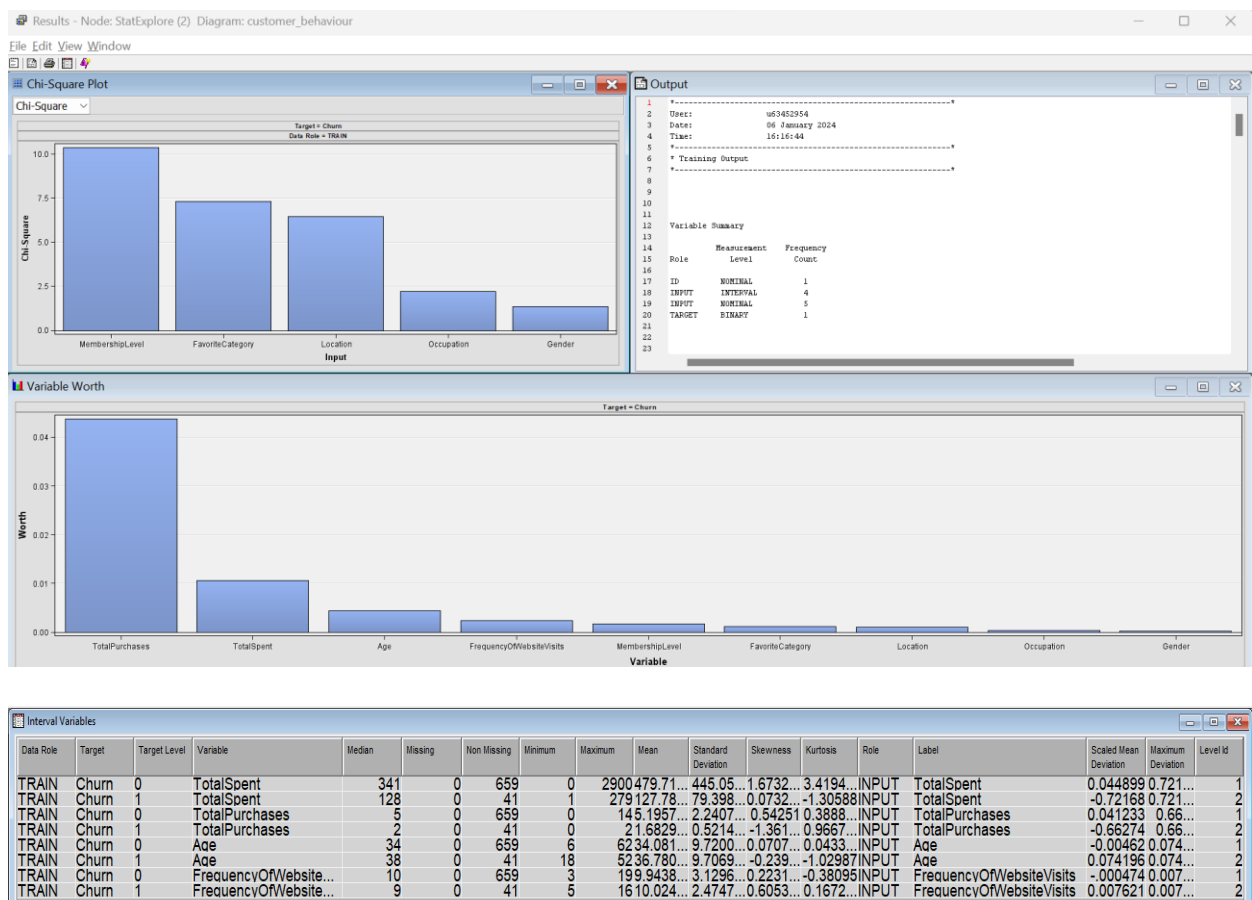
Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Churn	TARGET	0	659	94.1429
TRAIN	Churn	TARGET	1	41	5.8571

- GraphExplore Node:



For Train Dataset:

StatExplore node:



The internal variables table lists several variables that are being analyzed in relation to the target variable 'Churn'. These variables may include 'TotalSpent', 'TotalPurchases', and 'FrequencyOfWebsiteVisits'.



'Mean', 'Sum', 'Std', 'Min', 'Max', and 'Range' provide a statistical summary of each variable. For instance, a high mean in 'TotalSpent' might indicate higher average customer spending, which could be a factor in churn rate. A high standard deviation would suggest variability in spending amongst the customers. The variable 'TotalSpent' has a particularly high range, it may indicate a wide disparity in customer spending behavior, which could be a significant factor in modeling churn.

### Summary Statistic:

**Results - Node: StatExplore (2) Diagram: customer\_behaviour**

---

**File Edit View Window**

---

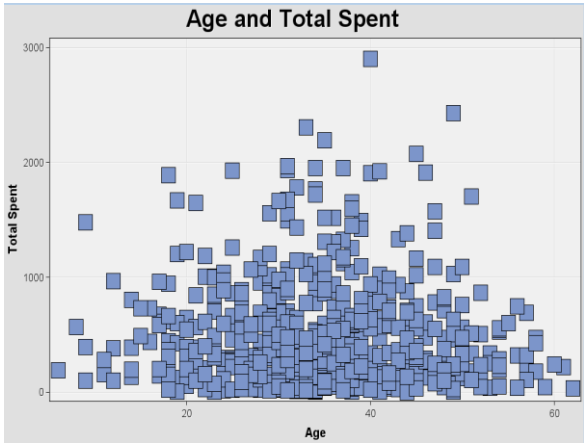

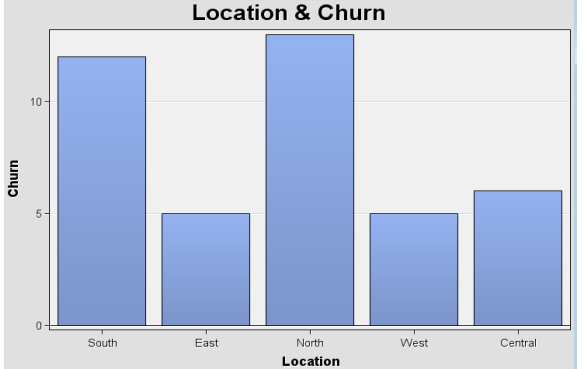
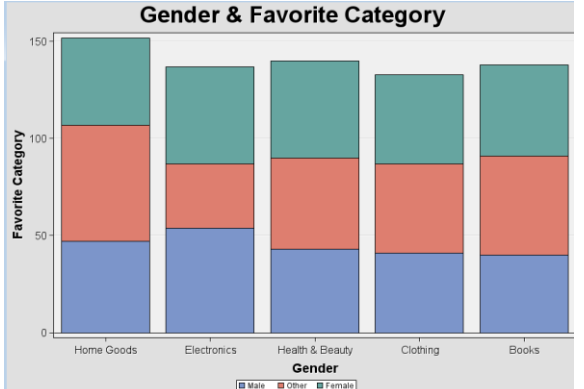
**Output**

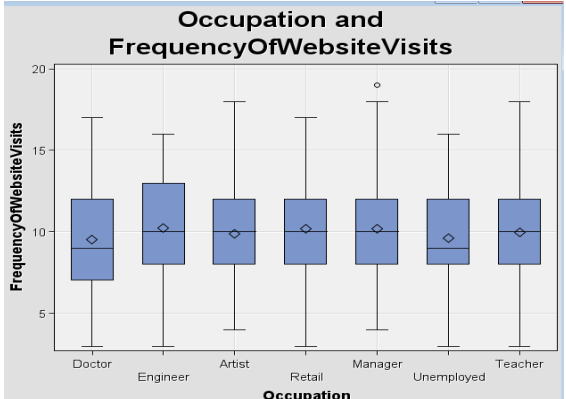
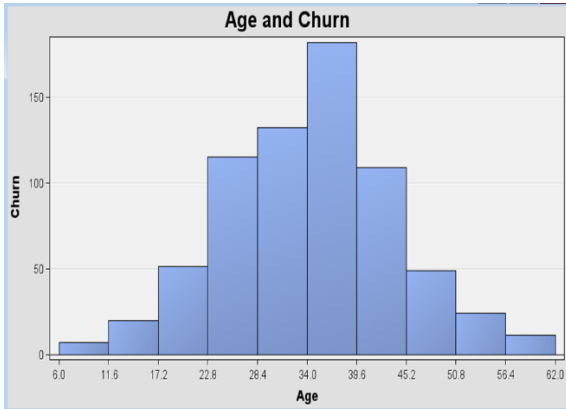
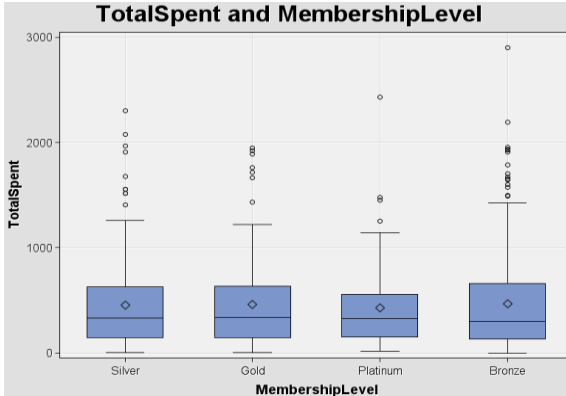
```

13
14      Measurement   Frequency
15 Role            Level    Count
16
17 ID              NOMINAL     1
18 INPUT           INTERVAL     4
19 INPUT           NOMINAL     5
20 TARGET          BINARY       1
21
22
23
24 Variable Levels Summary
25 (maximum 500 observations printed)
26
27      Frequency
28 Variable      Role    Count
29
30 Churn          TARGET     2
31 CustomerID     ID        700
32
33
34
35 Class Variable Summary Statistics
36 (maximum 500 observations printed)
37
38 Data Role=TRAIN
39
40      Number
41 Data      of
42 Role      Variable Name    Role    Levels    Missing    Mode      Mode      Mode2
43                                     Percentage    Mode2      Percentage
44 TRAIN FavoriteCategory INPUT      5         0    Home Goods    21.71    Health & Beauty    20.00
45 TRAIN Gender             INPUT      3         0    Female        34.00    Other              33.86
46 TRAIN Location           INPUT      5         0    North         21.86    Central            20.71
47 TRAIN MembershipLevel    INPUT      4         0    Bronze        37.29    Silver             29.86
48 TRAIN Occupation         INPUT      7         0    Artist        16.00    Doctor             15.14
49 TRAIN Churn              TARGET     2         0    0             94.14    1                  5.86
50
51
52
53 Distribution of Class Target and Segment Variables
54 (maximum 500 observations printed)
55
56 Data Role=TRAIN
57
58 Data      Variable
59 Role      Name      Role    Level      Frequency
60                                     Count      Percent
61 TRAIN Churn      TARGET     0         659    94.1429
62 TRAIN Churn      TARGET     1         41     5.8571
63

```

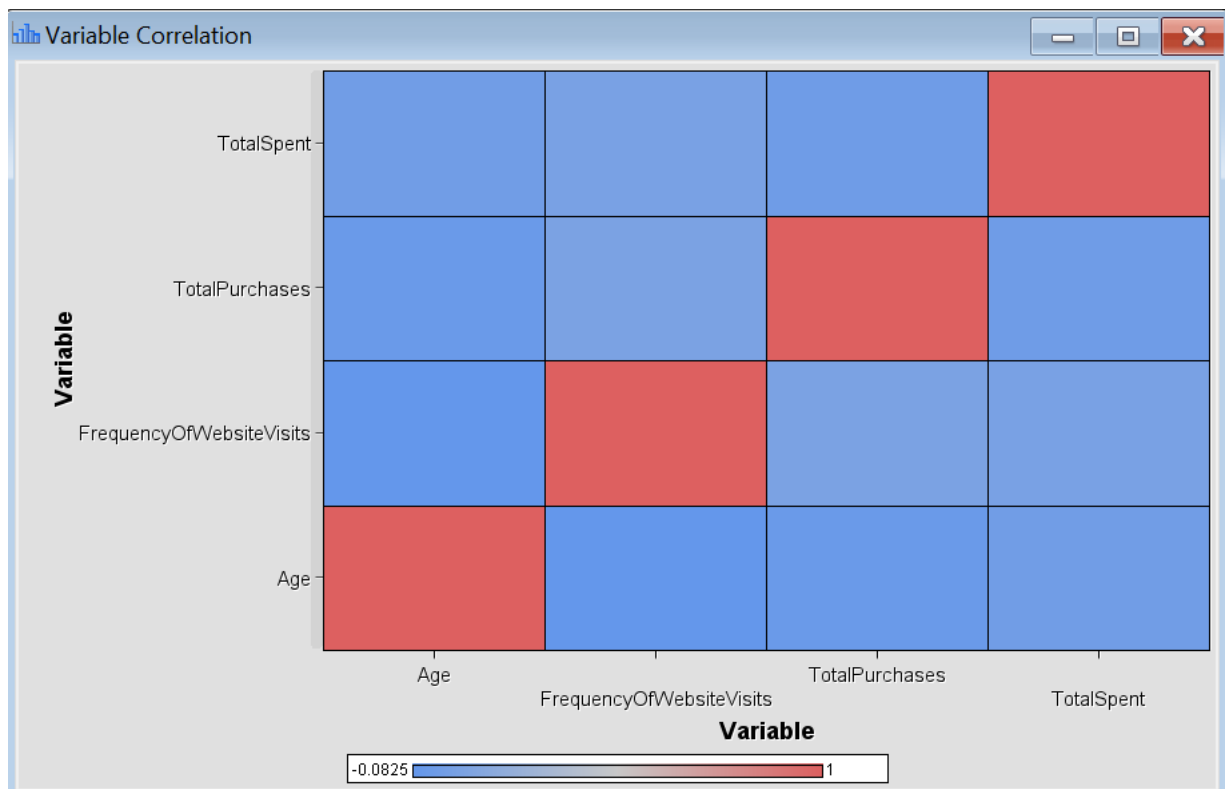
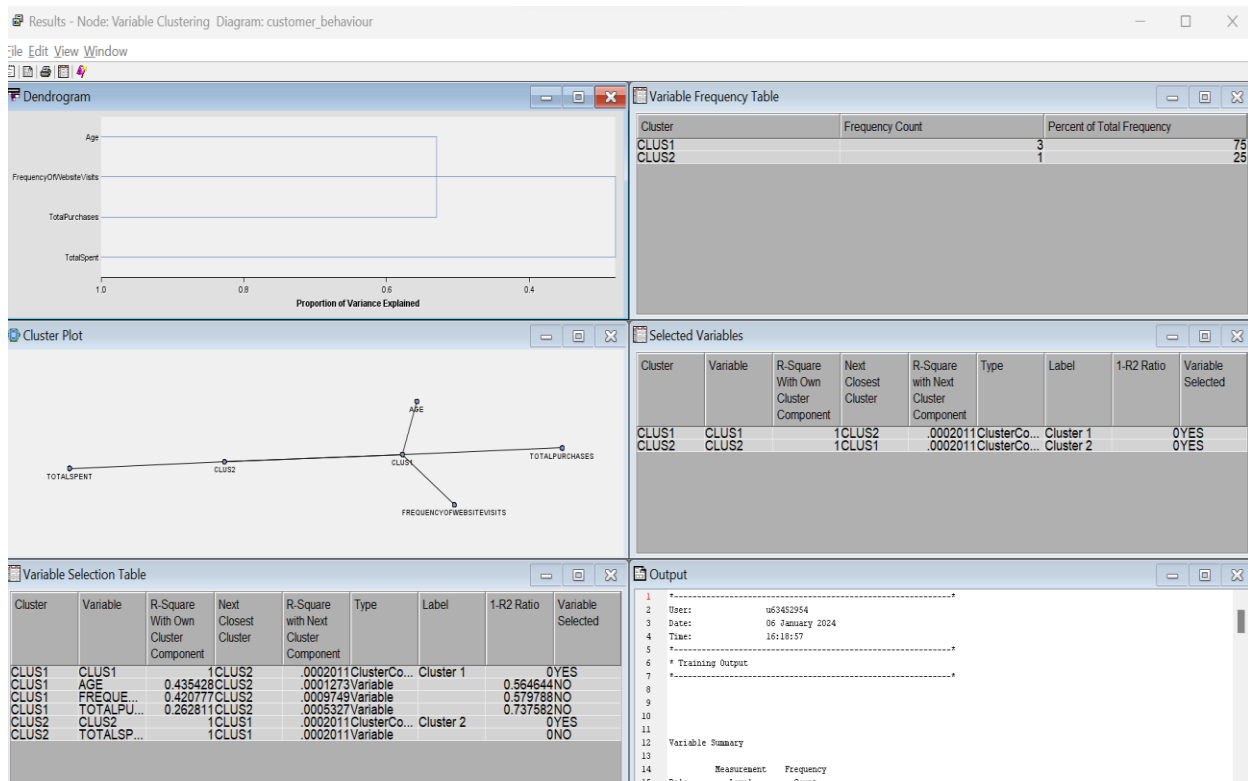
## Exploratory Data Analysis, EDA:

No	Variables & Analysis	Graph/Chart
1	<p><b>Age and Total spent:</b></p> <p><b>Chart Type:</b> Scatter Plot</p> <p><b>Purpose:</b> To explore if there's a correlation between the age of customers and the total amount they spend. This can help in understanding spending patterns across different age groups.</p> <p><b>Analysis:</b> The scatter plot does not show a clear linear relationship between age and total spent, suggesting that age may not be a strong predictor of spending behavior. The distribution of spending is consistent across different ages, with no significant outliers that suggest a particular age group spends dramatically more or less.</p>	
2	<p><b>MembershipLevel and TotalPurchases:</b></p> <p><b>Chart Type:</b> Bar Chart</p> <p><b>Purpose:</b> To analyze if higher membership levels correlate with a greater number of purchases. This can indicate the effectiveness of membership programs in encouraging purchases.</p> <p><b>Analysis:</b> It appears that silver and bronze members make more purchases than Gold and Platinum members. This could be counterintuitive as one might expect higher-tier members to make more purchases.</p>	
3	<p><b>Location and Churn:</b></p> <p><b>Chart Type:</b> Bar Chart</p> <p><b>Purpose:</b> To determine if churn rates vary significantly across different locations. This can be crucial for localized marketing strategies.</p> <p><b>Analysis:</b> Churn rates appear to be higher in the South and North regions compared to East, West, and Central. This regional variation indicates that location may influence churn behaviour.</p>	
4	<p><b>Gender and FavoriteCategory:</b></p> <p><b>Chart Type:</b> Stacked Bar Chart</p> <p><b>Purpose:</b> To examine the shopping preferences across different genders. This can inform targeted marketing strategies for each category.</p> <p><b>Analysis:</b> The preferences for 'Home Goods', 'Electronics', and 'Books' seem fairly evenly distributed across genders, while there might be a slight gender preference in categories like 'Health &amp; Beauty' and 'Clothing'.</p>	

5	<p><b>Occupation and FrequencyOfWebsiteVisits:</b>  <b>Chart Type:</b> Box Plot  <b>Purpose:</b> To see if certain occupations are associated with more frequent website visits, potentially indicating higher engagement or interest.  <b>Analysis:</b> The median frequency of website visits is relatively consistent across different occupations, but there is some variation in the interquartile range and outliers.</p>	
6	<p><b>Age and Churn:</b>  <b>Chart Type:</b> Histogram  <b>Purpose:</b> To investigate if certain age groups are more likely to churn. This can help in tailoring retention strategies for specific age demographics.  <b>Analysis:</b> The histogram shows a relatively normal distribution of churn across age groups, with a slight increase in the middle age range.</p>	
7	<p><b>TotalSpent and MembershipLevel:</b>  <b>Chart Type:</b> Box Plot  <b>Purpose:</b> To assess whether members at different levels have distinct spending patterns. This can reveal the spending habits of different membership tiers.  <b>Analysis:</b> There's quite a bit of variance in the 'TotalSpent' across all membership levels, with a few outliers, especially in the 'Silver' category. The median spend does not vary significantly by membership level.</p>	

## Variable Clustering

To understand the correlation among the variables', Variable Clustering node is added. Select Add Node-> Explore-> Variable Clustering.



- Typically, the amount of the correlation coefficient is correlated with the intensity of the colour (red or blue). The darkest colour would be used to symbolise a correlation coefficient of 1 (perfect positive correlation) or -1 (perfect negative correlation).
- Positive connections between expenditures and purchases would suggest that providing incentives to augment the frequency of purchases could result in increased overall spending.

- A negative association between age and website visits may indicate the need to modify the website experience to appeal more to older demographics or direct attention towards alternative avenues of engagement for these particular clients.
- The presence of weak correlations implies that specific customer actions do not exhibit a linear relationship with the variables in question, implying the possibility that additional factors warrant investigation.

## Modify Phase:

To check for Missing Values, connect a 'Data Partition' node to the 'File Import' node, then followed by 'StatExplore' node to the 'Data Partition' node. Proceed to run the diagram and review the output of 'StatExplore' to see if there are any missing values.

If missing values are found, connect the 'Impute' node to the data source. In the properties of the 'Impute' node, select appropriate methods for imputation. For e.g., 'Mean' for continuous variables, 'Mode' for categorical variables.

The raw dataset has filtered out missing values via Talend Data Integration and hence the dataset has no missing values after having run the diagram and review the output of 'StatExplore'. Next, connect transform variables node under Modify tab and if any variables need to be dropped. Since Customer ID is used as an identifier, it will be excluded from the model as it doesn't provide predictive power as we have set as an "ID" role in the data source node.

Enterprise Miner - WQD7005\_AA1

File Edit View Actions Options Window Help

WQD7005\_AA1

- Data Sources
  - PART\_TRAIN
- Diagrams
  - customer\_behaviour
- Model Packages

customer\_behaviour

Variables - Trans

(none) ☐ not Equal to ☐ Basic

Columns: ☐ Label ☐ Mining

Name	Method	Number of Bins	Role	Level
Age	Default	4	Input	Interval
Churn	Default	4	Target	Binary
FavoriteCate	Default	4	Input	Nominal
FrequencyOf	Default	4	Input	Interval
Gender	Default	4	Input	Nominal
Location	Default	4	Input	Nominal
Membership	Default	4	Input	Nominal
Occupation	Default	4	Input	Nominal
TotalPurchas	Default	4	Input	Interval
TotalSpent	Default	4	Input	Interval

Property Value

**General**

Node ID Trans

Imported Data

Exported Data

Notes

**Train**

Variables

Formulas

Interactions

SAS Code

**Default Methods**

Interval Inputs None

Interval Targets None

Class Inputs None

Class Targets None

Treat Missing as Level

**Sample Properties**

Method First N

Size Default

Random Seed 12345

**Optimal Binning**

Number of Bins 4

Missing Values Use in Search

**Grouping Method**

Cutoff Value 0.1

Group Missing No

Number of Bins Variables

Add Minimum Value Yes

Offset Value 1

**Score**

Use Meta Transform Yes

Hide Yes

Results - Node: Transform Variables Diagram: customer\_behaviour

File Edit View Window

Output

```

1 *-----*
2 User:          u63452954
3 Date:          06 January 2024
4 Time:          15:45:49
5 *-----*
6 * Training Output
7 *-----*
8
9
10
11
12 Variable Summary
13
14      Measurement      Frequency
15      Role             Level      Count
16
17 INPUT      INTERVAL      4
18 INPUT      NOMINAL       5
19 TARGET     BINARY        1
20
21
22 *-----*
23 * Score Output
24 *-----*
25
26
27 *-----*
28 * Report Output
29 *-----*
30

```

Once the dataset has imported and pre-processed the dataset, can proceed with the Decision Tree analysis for the Modelling Phase.

[15 marks]

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

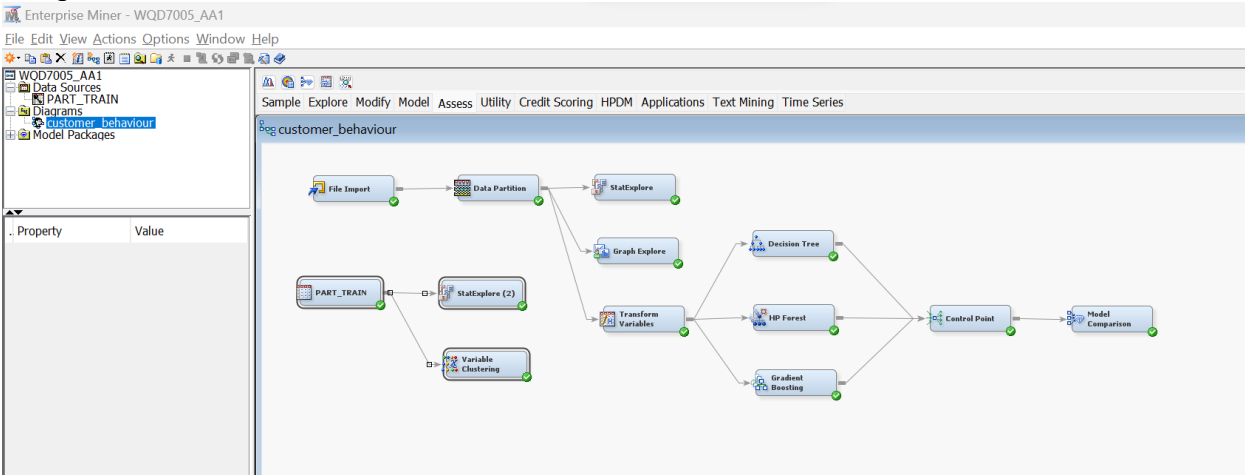
**Answer:**

### Modelling Phase:

Prior to modelling stage, we split the dataset to create both training and validations sets. The validation data is used to perform model selection from a group of candidate models. The training to validation split ratio for the dataset is 70:30 and output result from “Data Partition” node as shown in figure below:

Summary Statistics for Class Targets					
Data=DATA					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	941	94.1	
Churn	1	1	59	5.9	
Data=TRAIN					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	659	94.1429	
Churn	1	1	41	5.8571	
Data=VALIDATE					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Churn	0	0	282	94	
Churn	1	1	18	6	

The overall connection between data, data partition node, model nodes, control point and model comparison node.



The parameter is set as below to build a decision tree model.

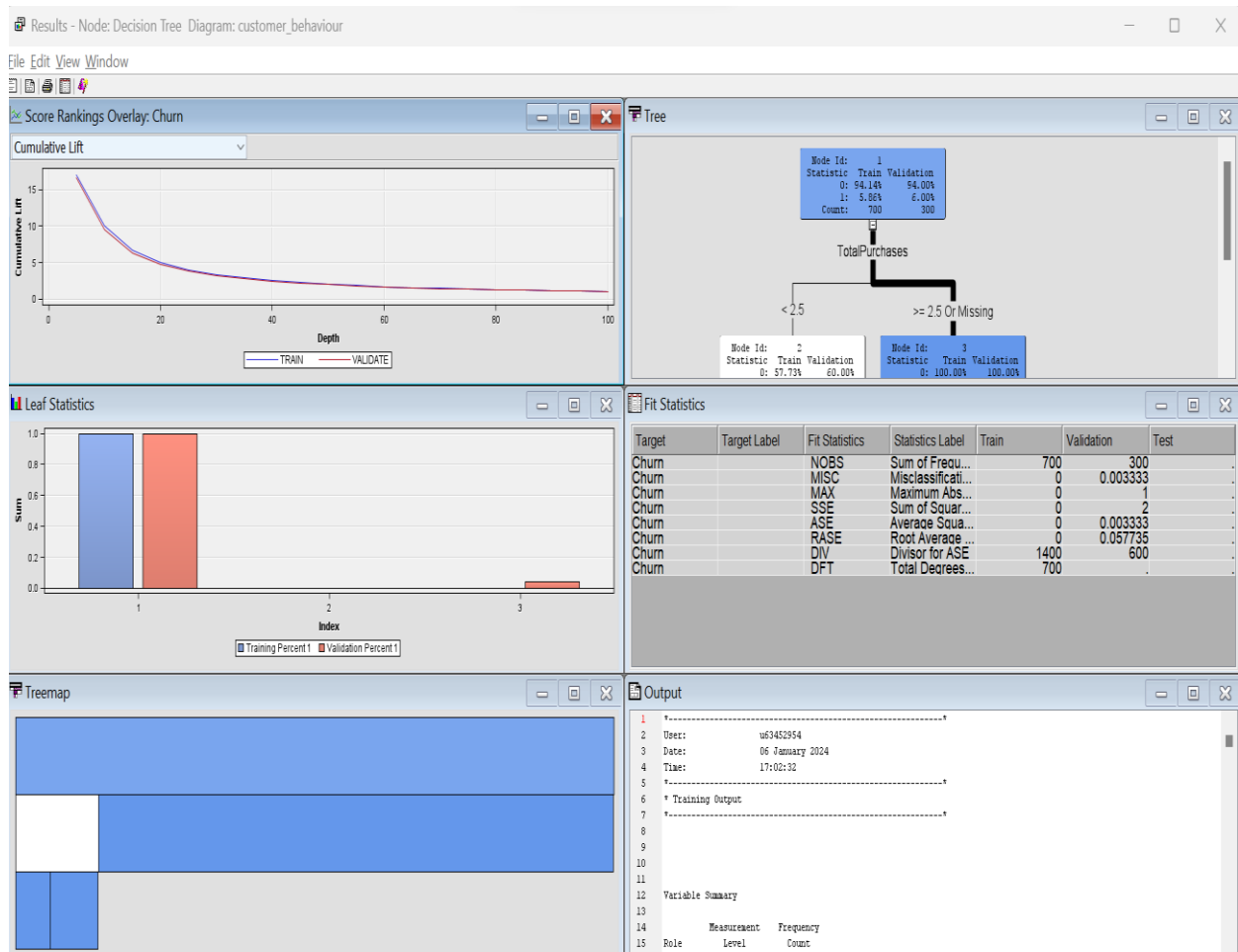
The screenshot shows the 'Variables - Tree' dialog box in Enterprise Miner. The dialog is titled 'Variables - Tree' and has a dropdown menu set to '(none)'. Below the dropdown, there are checkboxes for 'Label' and 'Mining'. The main part of the dialog is a table with columns: 'Name', 'Use', 'Report', 'Role', and 'Level'. The table lists various variables and their settings for building a decision tree model.

Name	Use	Report	Role	Level
Age	Default	No	Input	Interval
Churn	Yes	No	Target	Binary
CustomerID	Default	No	ID	Nominal
FavoriteCate	Default	No	Input	Nominal
FrequencyOf	Default	No	Input	Interval
Gender	Default	No	Input	Nominal
LastPurchase	Default	No	Time ID	Interval
Location	Default	No	Input	Nominal
Membership	Default	No	Input	Nominal
Occupation	Default	No	Input	Nominal
TotalPurchas	Default	No	Input	Interval
TotalSpent	Default	No	Input	Interval
dataobs	Default	No	ID	Interval

On the left side of the dialog, there is a 'Property' pane with a table of settings:

Property	Value
Observation Based In	No
Number Single Var	15
P-Value Adjustment	Yes
Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	Yes
Leaf Variable	Yes
Interactive Sample	Default
Create Sample	Random
Sample Method	10000
Sample Size	12345
Sample Seed	Disk
Performance	Segment
Score	Segment
Variable Selection	Segment
Leaf Role	Segment
Report	Segment
Precision	4
Tree Precision	4
Class Target Node C	Percent Correctly Cla
Interval Target Node	Average
Node Text	...
Status	...
Create Time	6/1/24 3:50 PM
Run ID	aa4c4842-105b-5642
Last Error	
Last Status	Complete
Last Run Time	6/1/24 3:53 PM

Run the decision tree node with the above parameters and the outputs are as below:

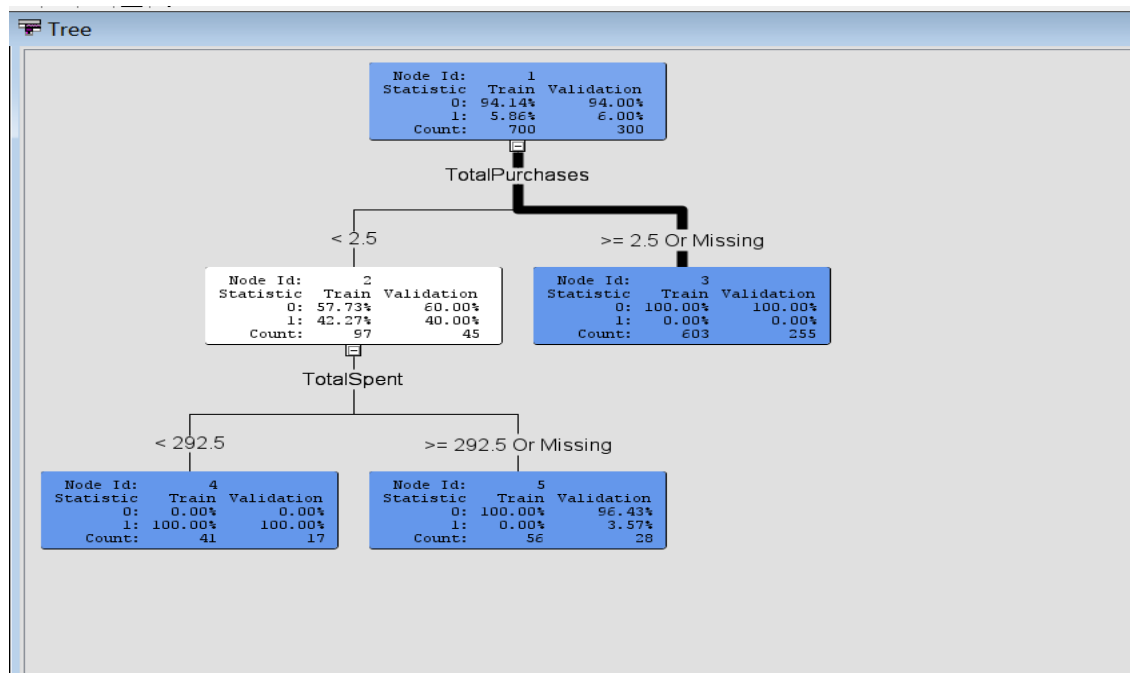


A decision tree predictive model is a type of supervised machine learning model with tree-like structure that is used to for classification and regression tasks. The base of tree is the root node. The root nodes will split into decision node and each decision node represents a questions or split point, and the leaf nodes that split from a decision node represent the possible answers. It is a type of machine learning algorithm that make decisions by splitting the data based on different feature values. However, decision trees typically work with numerical or binary features rather than categorical variables. The binary data type of “Churn” which is a format that decision trees can understand. Each category within these variables is represented as a separate binary feature, where it indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

The decision tree provided seems to focus on two primary variables “TotalPurchases” and “TotalSpent”. The tree splits first on “TotalPurchases” and then on “TotalSpent”, which suggests that these two variables are significant predictors of churn in the dataset.

- **Node 1 (Root Node):** This is the starting point of the analysis, indicating the initial churn rate across the entire dataset before any splits are made.
- **Split on TotalPurchases:** The tree splits based on whether customers made less than 2.5 number of purchases or 2.5 and above. This indicates that customers with fewer number of purchases are more likely to churn.
- **Further Splits on TotalSpent:** Within the group that made more purchases, the tree further distinguishes between customers based on their spending, with a split at \$292.5. It implies that among customers who made more purchases, those who spent less are more likely to churn.





The fit statistics table shows several metrics, but the key ones to focus on would be:

- **MISC (Misclassification Rate):** This is quite low for both training and validation, which suggests that the model is performing well in classifying customers as churned or not churned.
- **SSE (Sum of Squares for Error):** The fact that these are zero indicates a perfect fit for the training data, which could be a sign of overfitting, although the validation MISC is still low.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		NOBS	Sum of Freque...	700	300	.
Churn		MISC	Misclassificati...	0	0.003333	.
Churn		MAX	Maximum Abs...	0	1	.
Churn		SSE	Sum of Squar...	0	2	.
Churn		ASE	Average Squa...	0	0.003333	.
Churn		RASE	Root Average ...	0	0.057735	.
Churn		DIV	Divisor for ASE	1400	600	.
Churn		DFT	Total Degrees...	700	.	.

The variable importance table ranks the variables by their importance in predicting churn.

- **TotalSpent and TotalPurchases:** These are the only two variables used in the splitting rules, with TotalSpent being the most important predictor. The lack of other variables in the splits suggests that they do not add additional predictive power beyond what is provided by these two variables.
- **Other Variables:** The other variables (Age, FavoriteCategory, etc.) have zero importance, indicating that they were not used in any of the splitting rules. This could be due to the strength of the primary predictors, or it could suggest that the tree depth is not sufficient to capture the complexities of the dataset.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TotalSpent		1	1.0000	1.0000	1.0000
TotalPurchases		1	0.7942	0.7879	0.9921
FavoriteCategory		0	0.0000	0.0000	.
Age		0	0.0000	0.0000	.
FrequencyOfWebsite...		0	0.0000	0.0000	.
Location		0	0.0000	0.0000	.
Gender		0	0.0000	0.0000	.
Occupation		0	0.0000	0.0000	.
MembershipLevel		0	0.0000	0.0000	.

According to the analysis presented above, Total spent and purchases serve as crucial indicators for comprehending client behaviour. According to the data, significant markers of churn include the quantity of purchases and the overall expenditure. Customers who make fewer purchases and have lower spending patterns are more susceptible to churn. It may be advantageous for the organisation to formulate focused initiatives aimed at augmenting client involvement and expenditure amid the demographic of less frequent buyers. Given that variables such as MembershipLevel and FrequencyOfWebsiteVisits possess little significance in the tree, it could be prudent to examine the triggers that enable us to determine whether the organization's membership and engagement initiatives are efficacious in mitigating customer attrition. Further, it is recommended to assess and possibly modify loyalty programmes in order to ascertain their efficacy in stimulating the intended customer behaviours. This can be accomplished by formulating business and retention strategies that centre on customers who exhibit low purchase frequency and spend.

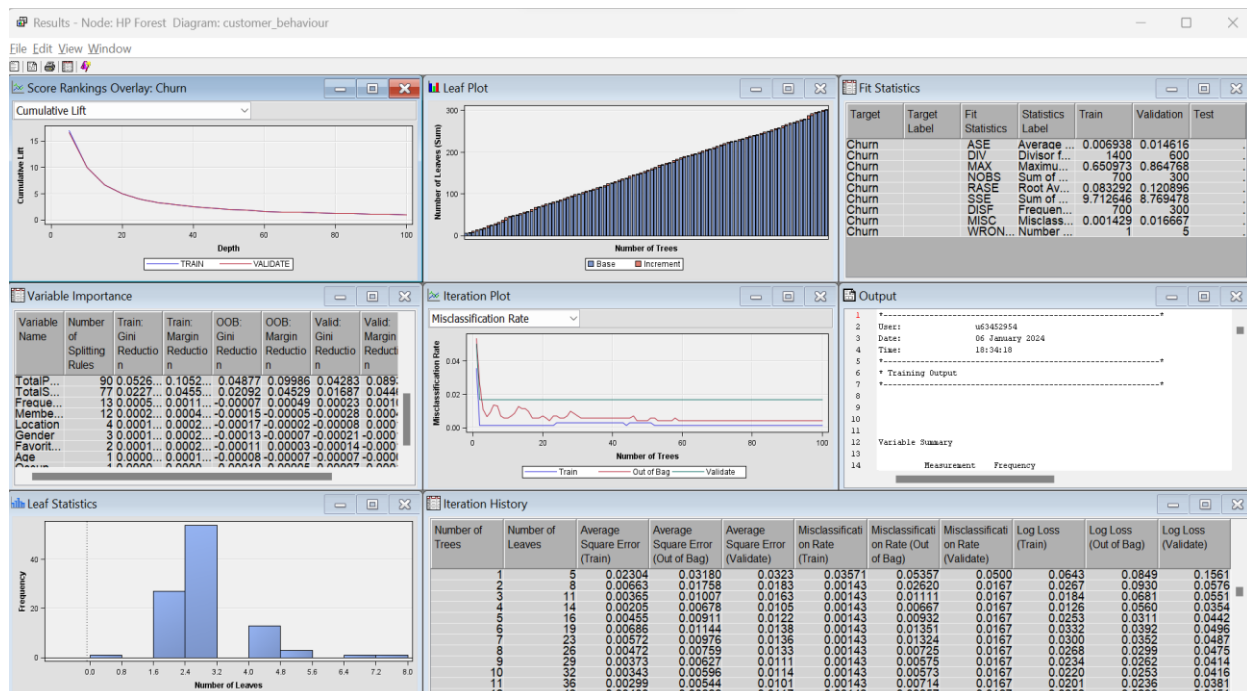
[20 marks]

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

In machine learning, bagging and boosting are both ensemble learning approaches that share a common characteristic: they both use the combination of a collection of weak learners to produce a robust learner whose performance surpasses that of any individual learner. Ensemble learning enhances the performance of machine learning models through the fusion of many models. This methodology facilitates the generation of superior prediction outcomes in comparison to employing a solitary model.

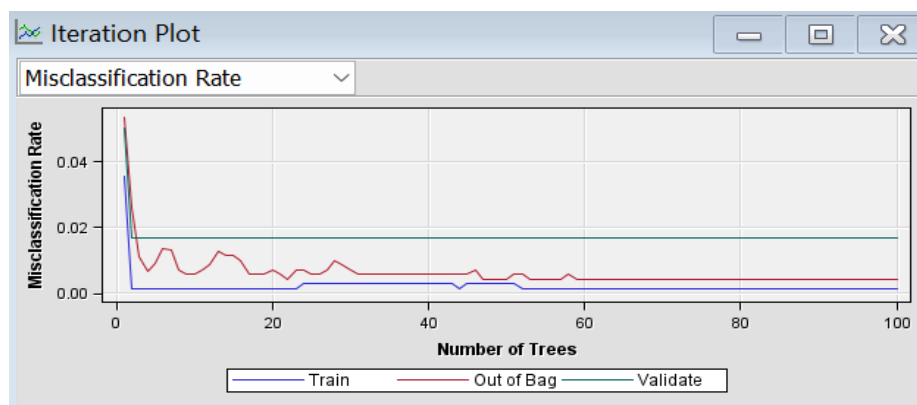
Bagging, which is an abbreviation for Bootstrap Aggregating, is an ensemble technique designed to prevent overfitting and reduce the variance of the prediction model. In the absence of the customary Random Forest node, the HP Forest node was implemented as a substitute. By creating multiple decision trees using bootstrapped data subsets and combining their predictions to improve the model's generalizability, it complies to the bagging principles. The powerful algorithm of the HP Forest is designed to deliver high-performance analytics, making it especially well-suited for handling large-scale datasets. Moreover, it operates effectively in distributed computing systems, demonstrating its scalability and efficiency.

Run the “HP Forest” as Random Forest node and the outputs are as below:



From the iteration plot, we see the following:

- **Training Misclassification Rate**: Remains consistent as the number of trees increases, which is typical for Random Forest models as they tend to stabilize after a certain number of trees.
- **Out of Bag (OOB) Misclassification Rate**: Also remains fairly stable and closely follows the training misclassification rate. OOB is an estimate of the error rate without using a separate validation set.
- **Validation Misclassification Rate**: It is very close to the OOB error, indicating that the model generalizes well.



The fit statistics table provides a numerical summary:

- **MISC (Misclassification Rate)**: It's low for both training and validation datasets, which is good. A low misclassification rate in validation suggests the model is performing well on unseen data.
- **Other Metrics**: SSE, ASE, RASE, and MSE all provide additional context on the model fit. These values should be low for a good model, which seems to be the case here.

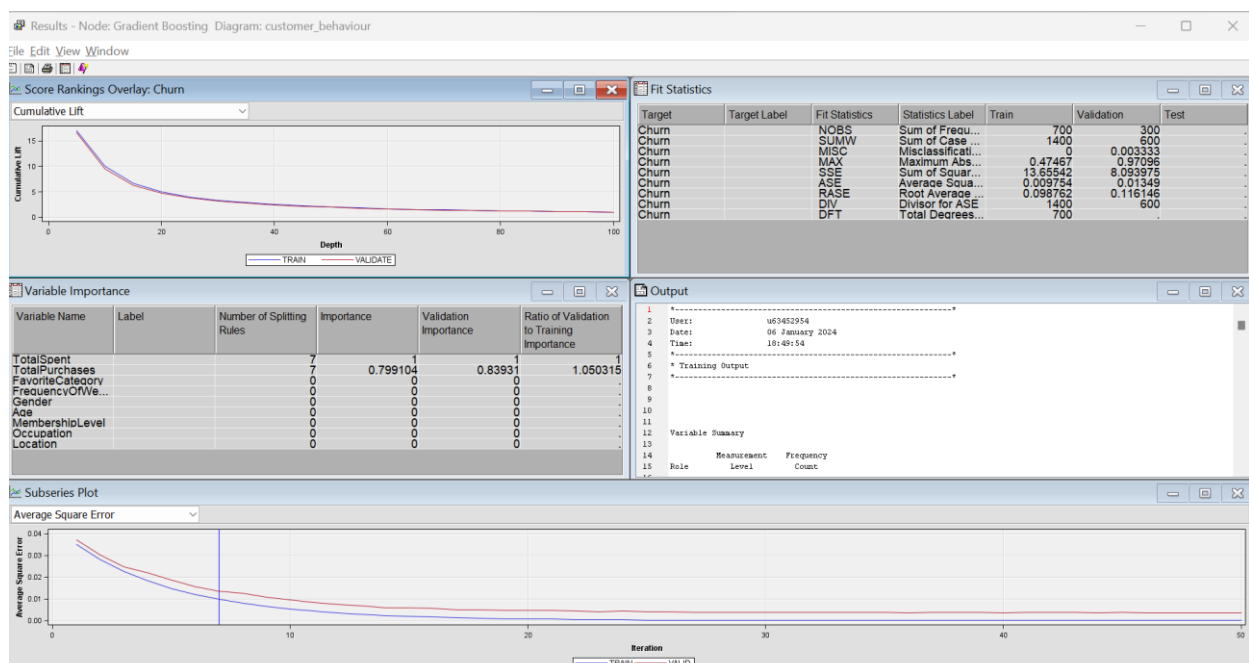
Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		ASE	Average ...	0.006938	0.014616	.
Churn		DIV	Divisor f...	1400	600	.
Churn		MAX	Maximu...	0.650973	0.864768	.
Churn		NOBS	Sum of ...	700	300	.
Churn		RASE	Root Av...	0.083292	0.120896	.
Churn		SSE	Sum of ...	9.712646	8.769478	.
Churn		DISF	Frequen...	700	300	.
Churn		MISC	Misclass...	0.001429	0.016667	.
Churn		WRON...	Number ...	1	5	.

The below table ranks variables by their importance in the model:

- **TotalSpent and TotalPurchases:** These are the most important variables, indicated by the number of splitting rules where they are used and their impact on error reduction. This is consistent with the decision tree analysis, which also highlighted these variables as key predictors.
- **Other Variables:** They appear to be less important for predicting churn. This suggests that focusing on customer spending behaviour might be the most effective strategy for understanding and preventing churn.

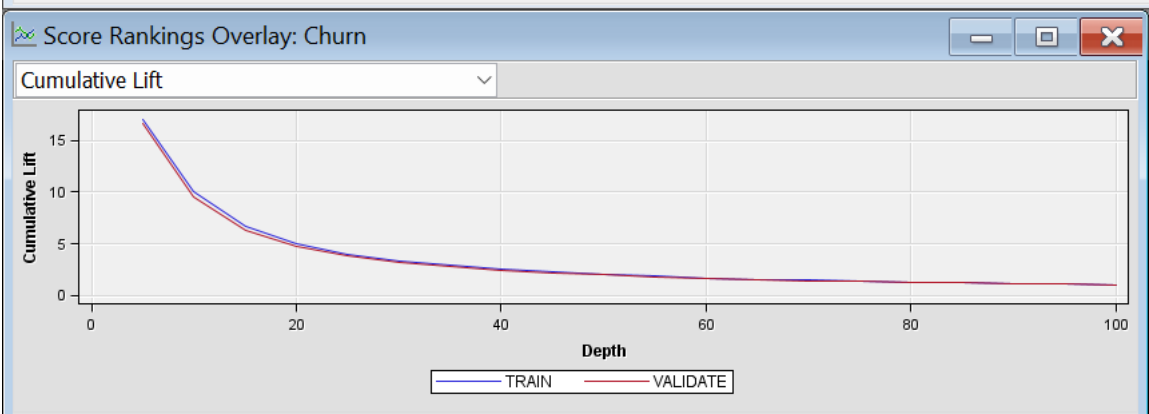
Variable Importance								
Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction	Label
TotalPurchases	90	0.052605	0.105211	0.04877	0.09986	0.04283	0.08934	
TotalSpent	77	0.022793	0.045585	0.02092	0.04529	0.01687	0.04464	
FrequencyOfWebsiteVisits	13	0.000560	0.001120	-0.00007	0.00049	0.00023	0.00103	
MembershipLevel	12	0.000248	0.000497	-0.00015	-0.00005	-0.00028	0.00042	
Location	4	0.000146	0.000292	-0.00017	-0.00002	-0.00008	0.00018	
Gender	3	0.000120	0.000240	-0.00013	-0.00007	-0.00021	-0.00012	
FavoriteCategory	2	0.000108	0.000216	-0.00011	0.00003	-0.00014	-0.00010	
Age	1	0.000056	0.000113	-0.00008	-0.00007	-0.00007	-0.00005	
Occupation	1	0.000043	0.000087	-0.00010	-0.00005	-0.00007	-0.00011	

**Gradient Boosting** is a type of boosting method. It builds trees sequentially, where each tree tries to correct the errors made by the previous ones. It typically starts with weak learners and focuses on improving them. Unlike Random Forest, which reduces variance, Gradient Boosting primarily works on reducing bias. It is often more sensitive to overfitting compared to Random Forest, especially if the data has noise. Run the “Gradient Boosting” node and the outputs are as below:



The subseries plot shows the Average Square Error (ASE) for both the training and validation datasets across iterations (number of trees used in the model).

- **Training Error:** Decreases sharply at first and then plateaus, which indicates that the model quickly captures the patterns in the training data.
- **Validation Error:** Follows a similar trend but starts to plateau slightly above the training error, suggesting that the model is generalizing well without overfitting significantly.



The Fit Statistics table provides key performance metrics for the Gradient Boosting model.

- **MISC (Misclassification Rate):** It is very low on the validation dataset, which is a good sign of model performance.
- **MAX (Maximum Absolute Error) and SSE (Sum of Squares for Error):** These are higher on validation than training, which is common as the model will naturally fit the training data better.
- **ASE (Average Square Error):** Like SSE, it's higher for validation, which suggests that there's some loss in prediction accuracy when the model is applied to unseen data.
- **RASE (Root Average Squared Error):** Provides another perspective on model accuracy, with a slightly higher error on the validation set indicating good but not perfect model fit.

Fit Statistics						
Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation	Test
Churn		ASE	Average Squared Error	0.009754	0.01349	.
Churn		DIV	Divisor for ASE	1400	600	.
Churn		MAX	Maximum Absolute Error	0.47467	0.97096	.
Churn		MISC	Misclassification Rate	0	0.003333	.
Churn		RASE	Root Average Squared Error	0.098762	0.116146	.
Churn		SUMW	Sum of Case Weights Times Freq	1400	600	.
Churn		NOBS	Sum of Frequencies	700	300	.
Churn		SSE	Sum of Squared Errors	13.65542	8.093975	.
Churn		DFT	Total Degrees of Freedom	700	.	.

The Variable Importance table identifies which variables the model found most predictive.

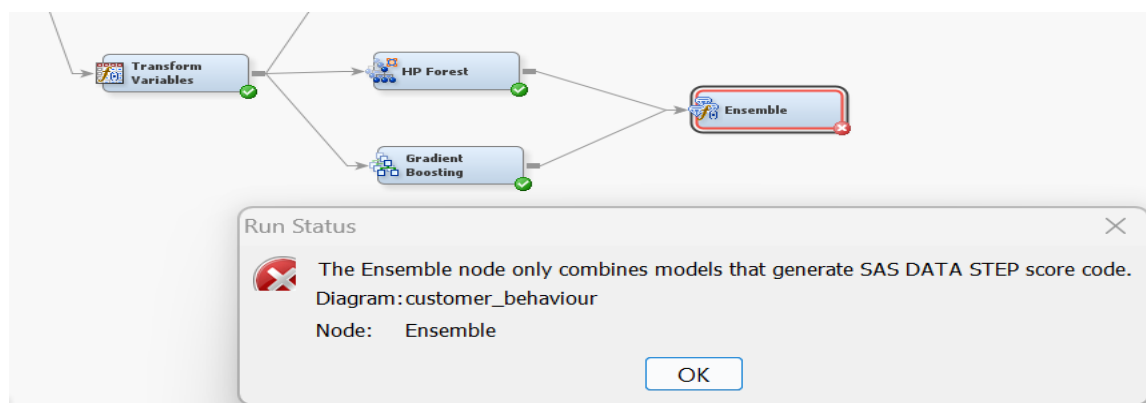
- **TotalSpent:** Continues to be the most significant predictor with the highest importance score, which aligns with previous analyses using the Random Forest model.
- **TotalPurchases:** Also significant, but less so than TotalSpent.
- **Other Variables:** They have zero importance, meaning they did not contribute to the prediction of churn in this model.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TotalSpent		7	1	1	1
TotalPurchases		7	0.799104	0.83931	1.050315
FavoriteCategory		0	0	0	.
FrequencyOfWebsiteVisits		0	0	0	.
Gender		0	0	0	.
Age		0	0	0	.
MembershipLevel		0	0	0	.
Occupation		0	0	0	.
Location		0	0	0	.

Based on the gradient boosting output, the model has identified TotalSpent and TotalPurchases as the main drivers of churn, which suggests that strategies aimed at increasing customer spend and purchase frequency might be effective for reducing churn. The model is performing well, with low misclassification rates and reasonable error metrics. The close performance between training and validation indicates that the model is not overfitting. The plateau in ASE suggests that adding more trees beyond a certain point does not significantly improve the model's performance on the validation set. It's important to find a balance in model complexity to avoid unnecessary computation without gain in accuracy.

## Ensemble Method Comparison

In comparison, both HP Forest and Gradient Boosting displayed distinct characteristics. HP Forest's strength lay in its variance reduction, driven by its bagging nature, while Gradient Boosting excelled in reducing bias through its sequential learning. Since the ensemble approach is not feasible due to HP Forest node couldn't generate SAS data step score code where have to run the models individually and use business logic or statistical reasoning to combine the insights gained from each model.

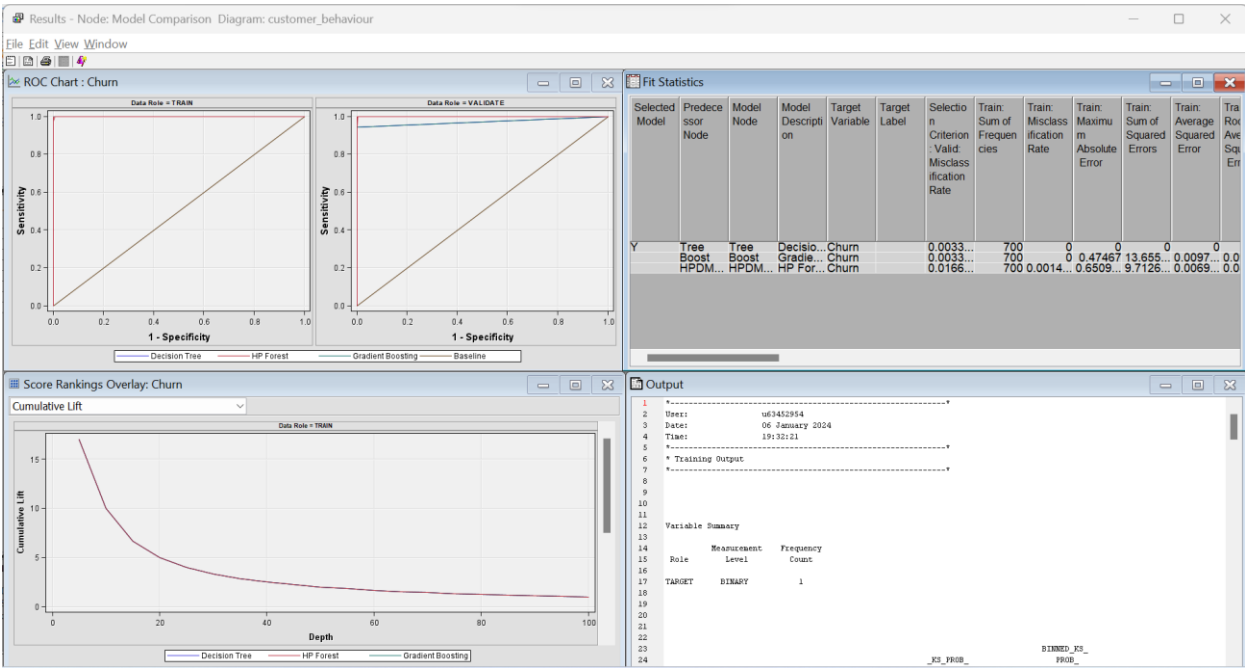


A pivotal aspect of the analysis was examining variable importance. The 'TotalSpent' and 'TotalPurchases' attributes were identified as significant predictors in both models, which highlights spending behavior as a critical factor in customer churn. This consistency across both methods reinforces the reliability of these findings.

The analytical journey with HP Forest and Gradient Boosting has illuminated the spending behavior and frequency of purchases as vital indicators for predicting churn. While the HP Forest model provides a scalable solution suitable for voluminous data, Gradient Boosting offers precision through focused learning. For businesses, these insights translate into actionable strategies that can target customer retention efforts more effectively.



In this stage, we compare the performance of customer behaviours models between Decision Tree, HP Forest and Gradient Boosting. We evaluated these models using Mean Absolute Error (MAE) and Average Square Error (ASE) metrics.



Here are the MAE and ASE values obtained for each model on the provided metrics:

Fit Statistics							
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Average Squared Error
Y	Tree	Tree	Decision Tree	Churn	0.003333	1	0.003333
	Boost	Boost	Gradient Boosting	Churn	0.003333	0.97096	0.01349
	HPDMForest	HPDMForest	HP Forest	Churn	0.016667	0.864768	0.014616

Decision Tree:

- Valid Misclassification Rate: 0.003333
- Valid Maximum Absolute Error: 1
- Valid Average Squared Error: 0.003333

Gradient Boosting:

- Valid Misclassification Rate: 0.003333
- Valid Maximum Absolute Error: 1
- Valid Average Squared Error: 0.01349

HP Forest:

- Valid Misclassification Rate: 0.016667
- Valid Maximum Absolute Error: 0.864768
- Valid Average Squared Error: 0.014616

The model incorporating decision trees and gradient boosting exhibits the lowest misclassification rate, indicating its exceptional accuracy in predicting customer attrition. The large maximum absolute error may be attributed to the highly definite prediction error made on

one or a small number of instances. The variation of the model's errors is quantified by the average squared error, which is rather little.

The HP Forest model exhibits a significantly elevated misclassification rate in contrast to the Gradient Boosting model, hence signifying a diminished capacity for precise predictions. The greatest absolute error is considerably diminished in comparison to the Gradient Boosting model, indicating that the errors produced by the former are of a less extreme nature. In terms of variance, the average squared error is extremely comparable to that of the Gradient Boosting model, suggesting comparable performance.

On the validation set, Gradient Boosting is performing exceptionally well in terms of the misclassification rate. Nonetheless, the existence of a substantial maximum absolute error suggests that the model might erroneously forecast in certain circumstances despite its high level of confidence. This may indicate that the model has been overfit to the training data or is struggling to handle the presence of outliers.

Although HP Forest exhibits a lower misclassification rate, its maximum absolute error is significantly diminished. This implies that the HP Forest model might exhibit a stricter approach towards making high-confidence wrong predictions, hence promoting conservatism in its predictions.

When minimising misclassifications is the principal consideration in model selection, the Gradient Boosting model seems to be the optimal selection. Conversely, if a more cautious stance is desired, in which forecasts that are exceedingly certain but perhaps erroneous are of lesser significance, the HP Forest model might be the superior option.

**Business Application:** In the context of customer turnover, the importance of limiting false positives (the prediction of churn when none exists) or false negatives (the failure to forecast churn when it does) may be equivalent, contingent upon the expense of intervention measures.

[10 marks]

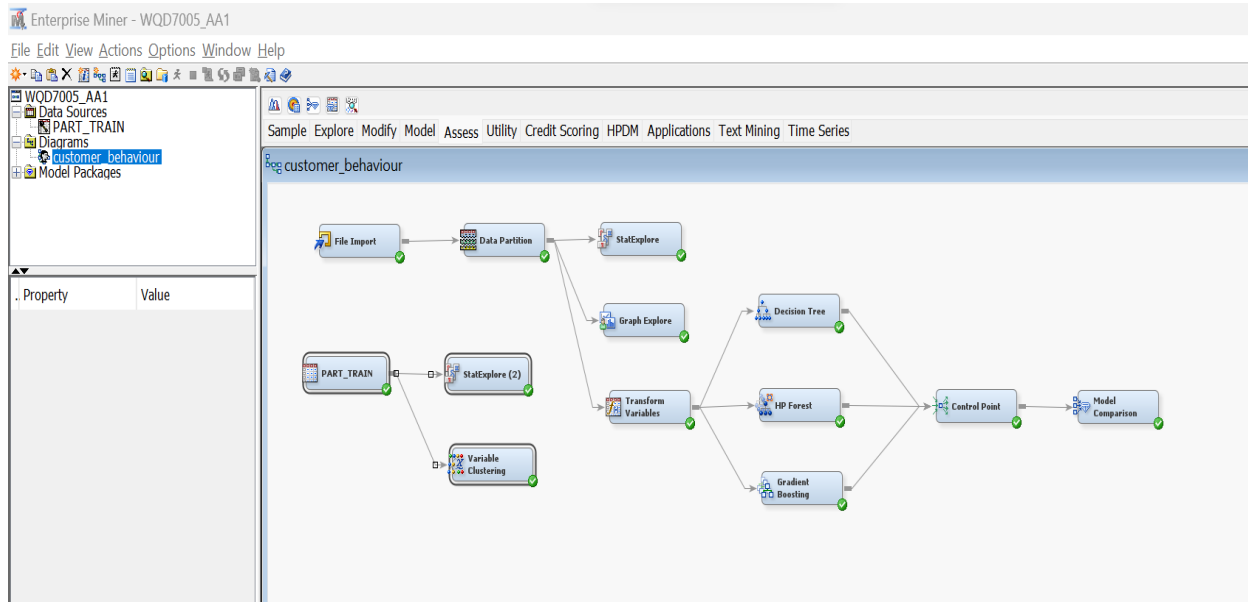
**Deliverables:** A report detailing each step of the process, including the rationale behind your choices and any challenges faced.

An analysis of the decision tree and ensemble methods, with insights into customer behavior and suggestions for business strategy.

**Answer:**

This report documents the analytical process undertaken to understand customer behaviour for an e-commerce platform. Utilizing Talend Data Integration, Talend Data Preparation, and SAS Enterprise Miner, various predictive modelling techniques have been explored to identify patterns of customer churn and to inform strategic business decisions.





## 1. Data Sampling & Preparation:

The analysis was initiated by importing the transactional data into the SAS Enterprise Miner environment using the File Import node. In addition to Age, Total Spent, and Membership Level, the dataset comprised additional consumer attributes. In order to determine the factors that influence customer churn, it is critical to have a comprehensive picture of customer interactions. Some challenges and priorities included correcting missing numbers and ensuring data quality. These were handled through data inspection and the selection of suitable imputation algorithms for missing information.

## 2. Exploratory Analysis of Data (EDA):

After importing the data, Data Partition was employed to partition the dataset in a 70:30 ratio across the training and validation sets, with the intention of ensuring a robust model evaluation. The process of partitioning the data enables us to conduct model training on a distinct subset while validating their performance on another. This ensures that our models do not overfit and can effectively extrapolate to novel data. 'Graph Explore' and 'StatExplore', two exploratory nodes, provide preliminary insights into the distributions of data and the interactions among variables. Furthermore, discovery of underlying patterns and linkages that may provide insights for the process of model construction is contingent upon EDA.

## 3. Transformation, Variable Modification, and Selection:

The data were transformed using the Transform Variables node in order to rectify skewness and generate derived variables that may have a greater predictive capacity for churn. For this reason, designing features and normalising skewed data can substantially enhance model performance. By grouping comparable variables, Variable Clustering reduced dimensionality and redundancy. Placing fewer input variables can aid in the simplification of models and provide some relief from the curse of dimensionality.

## 4. Predictive Modeling:

Three nodes for predictive modelling were utilised: Gradient Boosting, HP Forest, and Decision Tree. The decision tree structure offered a foundational model that could be easily understood in order to discover crucial predictive variables in providing a visible perspective of how input variables are utilised to generate predictions, decision trees serve as the building block for more complicated models.

HP Forest was an enhanced bagging ensemble method that was specifically designed to handle enormous datasets. The HP Forest algorithm is engineered to deliver exceptional performance and scalability, enabling it to capture intricate interrelationships among variables.

Gradient boosting was employed as a boosting technique in contrast to the bagging method. By consecutively prioritising cases that are challenging to forecast, boosting has the potential to deliver greater performance compared to bagging.

## **5. Model Evaluation, Assessment & Comparison:**

In order to guarantee that model comparisons were conducted on a consistent dataset, a Control Point node was connected. This ensures that every model is assessed using same data, hence facilitating an equitable comparison. The final stage was the Model Comparison node, which evaluated the performance of every model by considering misclassification rates and additional fit data. The utilisation of defined measurements to compare models aids in the identification and deployment of the most optimal model.

### **Outcomes and Insights:**

The results of the decision tree model indicated that membership level and total expenditure were significant predictors of turnover. Customers who maintained basic membership levels and made smaller expenditure were more susceptible to churn. The decision tree model was surpassed in performance by both HP Forest and Gradient Boosting, which exhibited reduced rates of misclassification and average squared errors. The ensemble approaches validated the predictive value of Total Purchases and Total Spending, which is consistent with the outcomes of the decision tree.

Construct retention initiatives that are specifically designed to target customers with low spending habits and basic membership tiers. Employ predictive model findings to motivate a higher frequency of purchases through the implementation of targeted marketing campaigns. Enhance membership perks as a loyalty programme incentive to encourage upgrades, hence potentially decreasing attrition among higher-tier members.

All in all, the significance of ensemble approaches in forecasting customer attrition is highlighted by the analysis. By employing a combination of bagging and boosting methodologies, we acquired a comprehensive comprehension of customer behaviour. By utilising the knowledge acquired from this study, strategic decisions may be formulated in a way that increases client loyalty and retention.

[5 marks]