UNIVERSITI MALAYA

MASTER OF DATA SCIENCE

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

WQD7005 Data Mining

ALTERNATIVE ASSESMENT

Name	Sze Hui Ling
Matric Number	22075528
Lecturer	Prof. Dr. Teh Ying Wah

Project Overview

The case study focus on data mining techniques using a Kaggle E-commerce Customer Behaviour Dataset E-commerce Customer Behavior Dataset (kaggle.com). Talend Data Preparation Tool used to handle missing values and adding attributes like "Churn" and "Age Group." The project use SAS Enterprise Miner for data import and setting variable roles. In-depth data exploration and analysis are conducted to identify key variables influencing 'Total Spend.' Various models, including regression trees, Random Forest, and Gradient Boosting, are implemented to understand and predict customer spending patterns. The goal is to enhance customer segmentation and targeting in marketing by leveraging statistical and machine learning techniques to analyze customer behaviors and spending patterns.

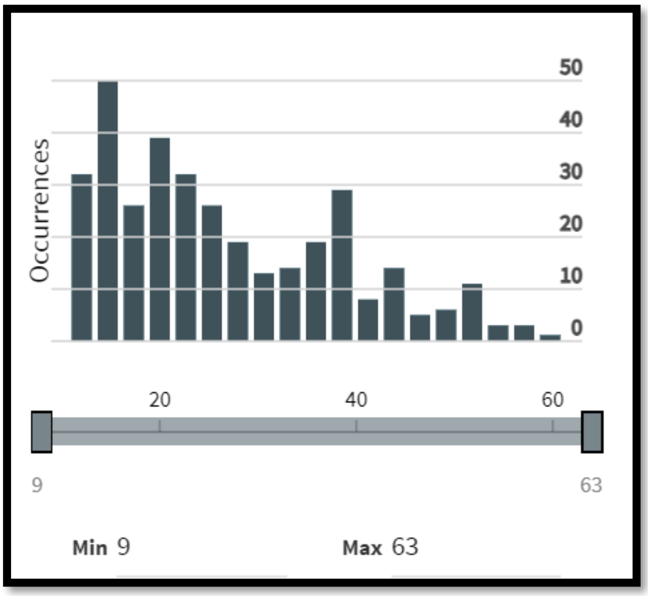
Tasks

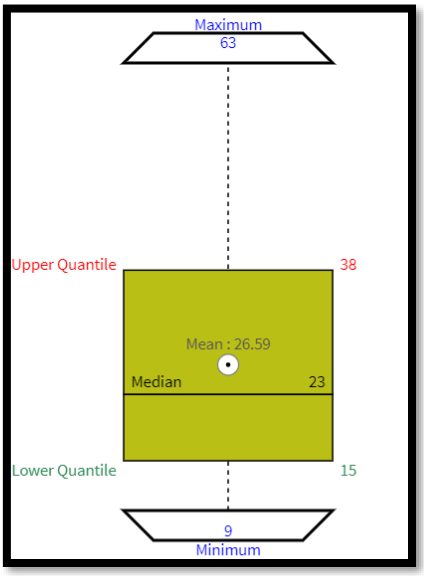
Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

[15 marks]

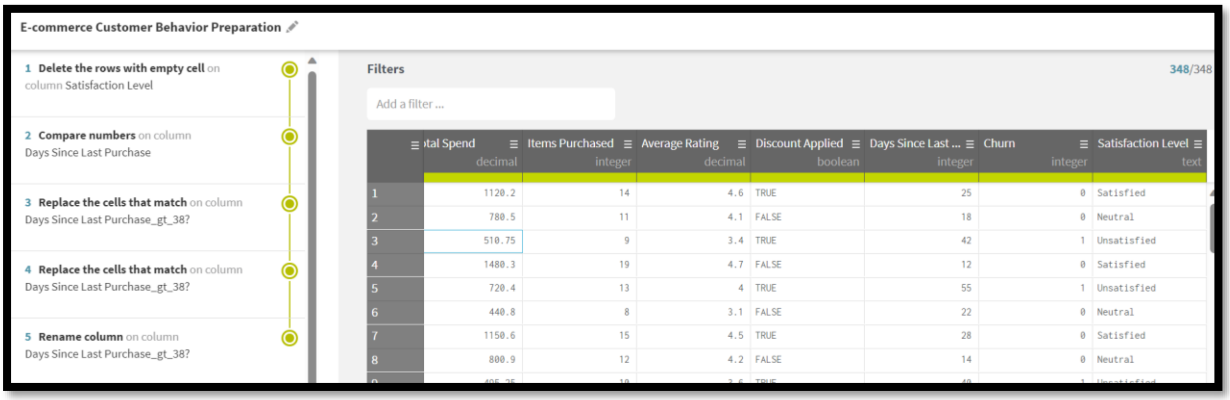
1. Handle missing value in Talend Data Preparation Tool
   Remove the row contain missing value in attribute “Satisfied level” in using function “Delete the row with empty cell”

2. Add attribute “Churn”





The figures above show the histogram chart and box plot of attribute “Days Since Last Purchase”. Define the churn based on attribute “Days Since Last Purchase”, label the upper quartile as churned. The 75 percentiles of “Days Since Last Purchase” is 38 days. This targets the 25% customers with the highest inactivity and flags them for churn. Use the "Compare numbers" function to evaluate whether the 'Days Since Last Purchase' for each customer is greater than 38 days. Then, convert the resulting Boolean values: 'True' is replaced with '1' to indicate churn, and 'False' is replaced with '0' to indicate no churn.



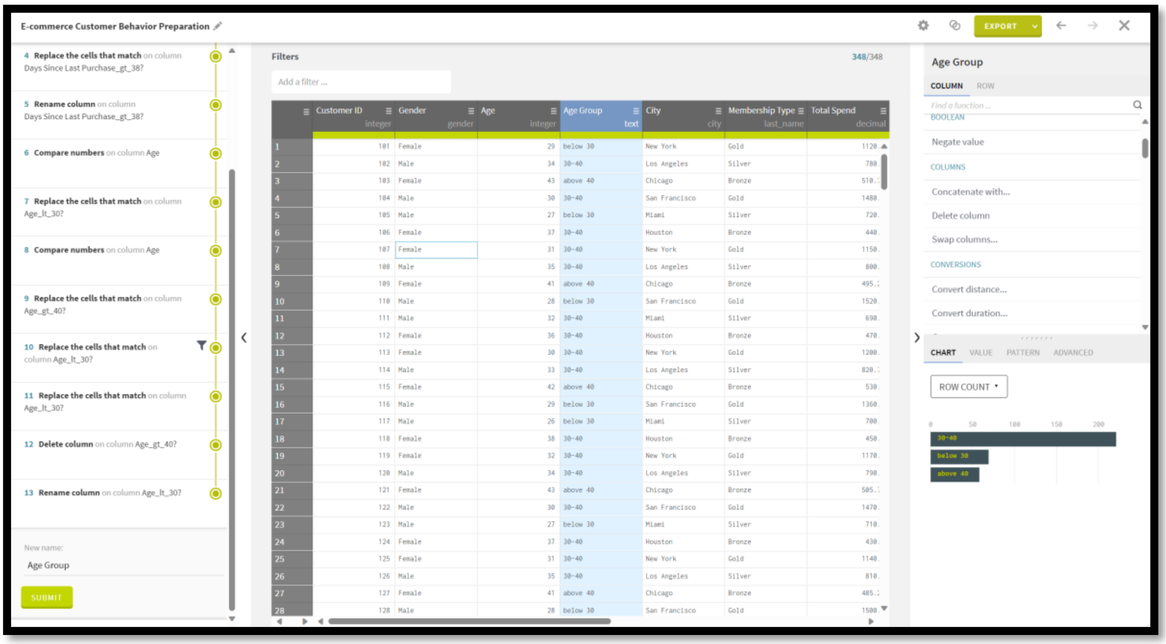
3. Add attribute "Age Group"

   Set the age group to 3 categories': below 30, 30-40 and above 40.

   Use function "compare" for cell that value less than 30. Then replace true to "below 30".

   Use function "compare" for cell that value greater than 40. Then replace true to "above 40".

   Then only view above 40, and replace in column churn, after that replace all the false to "30-40"

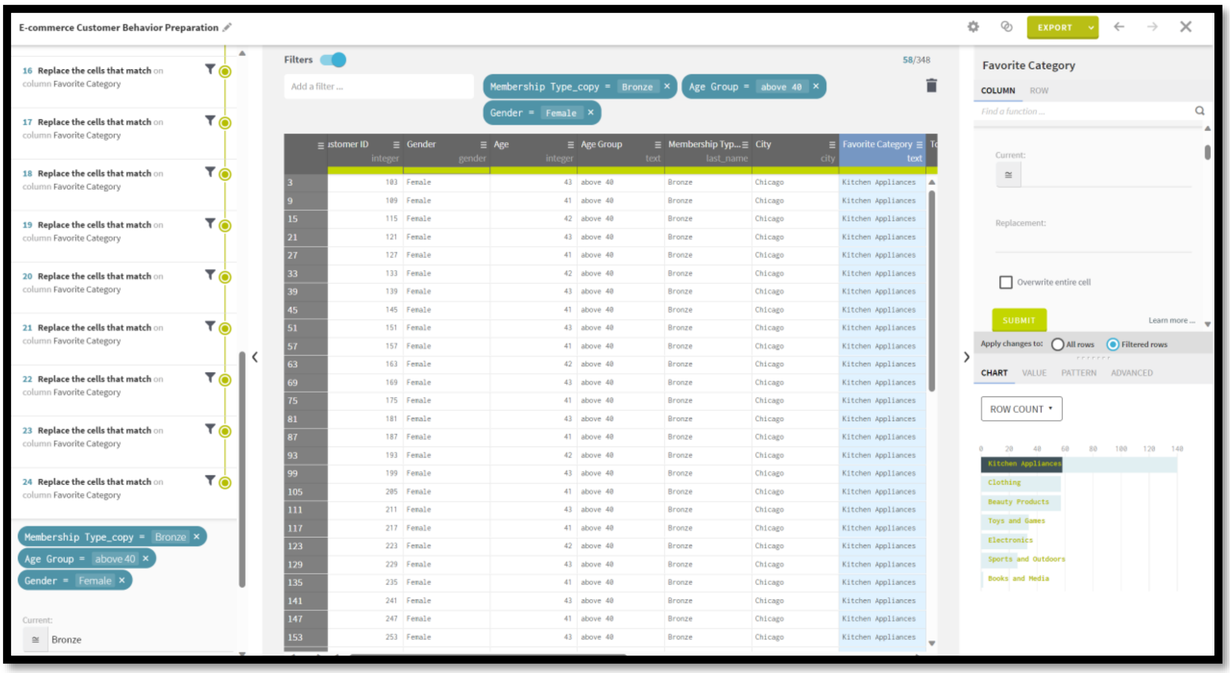


4. Add attribute “Favorite Category”

Table below is the rule that set for favorite category based on membership type, age group, and gender.

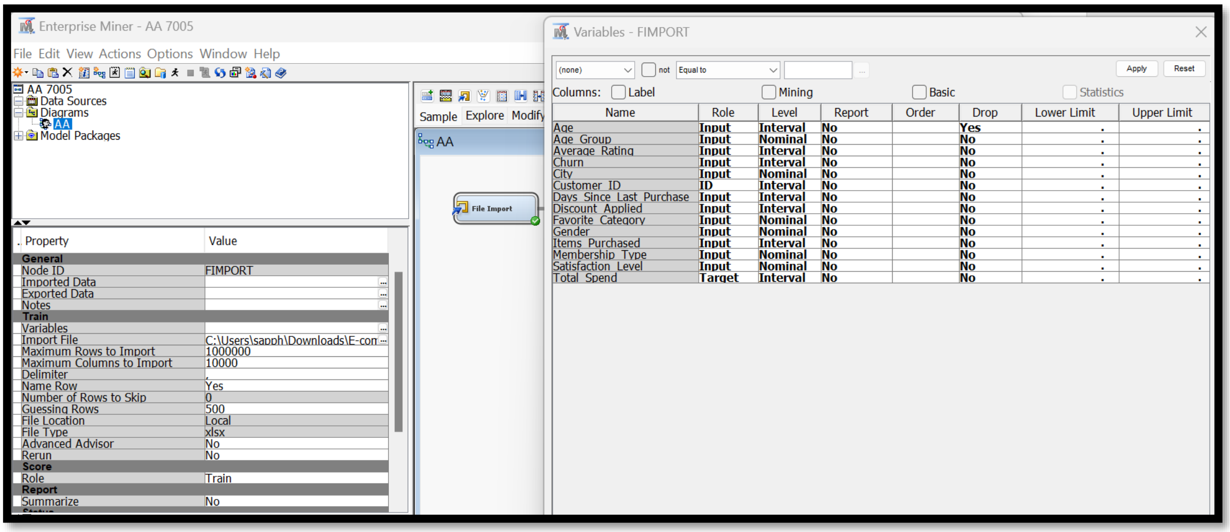
Membership Type	Age Group	Gender	Favorite Category
Gold	Below 30	Male	Electronics
Gold	Below 30	Female	Clothing
Gold	30 - 40	Male	Sports and Outdoors
Gold	30 - 40	Female	Beauty Products
Gold	Above 40	Male	Gourmet Foods
Gold	Above 40	Female	Home Goods
Silver	Below 30	Male	Toys and Games
Silver	Below 30	Female	Books and Media
Silver	30 - 40	Male	Kitchen Appliances
Silver	30 - 40	Female	Gardening and DIY
Silver	Above 40	Male	Home Goods
Silver	Above 40	Female	Gourmet Foods
Bronze	Below 30	Male	Electronics
Bronze	Below 30	Female	Beauty Products
Bronze	30 - 40	Male	Sports and Outdoors
Bronze	30 - 40	Female	Clothing
Bronze	Above 40	Male	Books and Media
Bronze	Above 40	Female	Kitchen Appliances

Then create a new column called “Favorite Category”, then replace each category one by one based on membership type, age group, and gender criteria. Finally, the favorite category got 7 category, Kitchen Application, Clothing, Beauty Products, Toys and Games, Electronics, Sport and Outdoor, Book and Media.

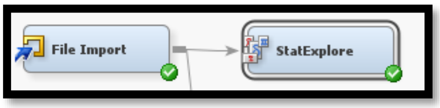


5. Import exported dataset from Talend Data Preparation Tool to SAS Enterprise Miner and set specify variable roles.

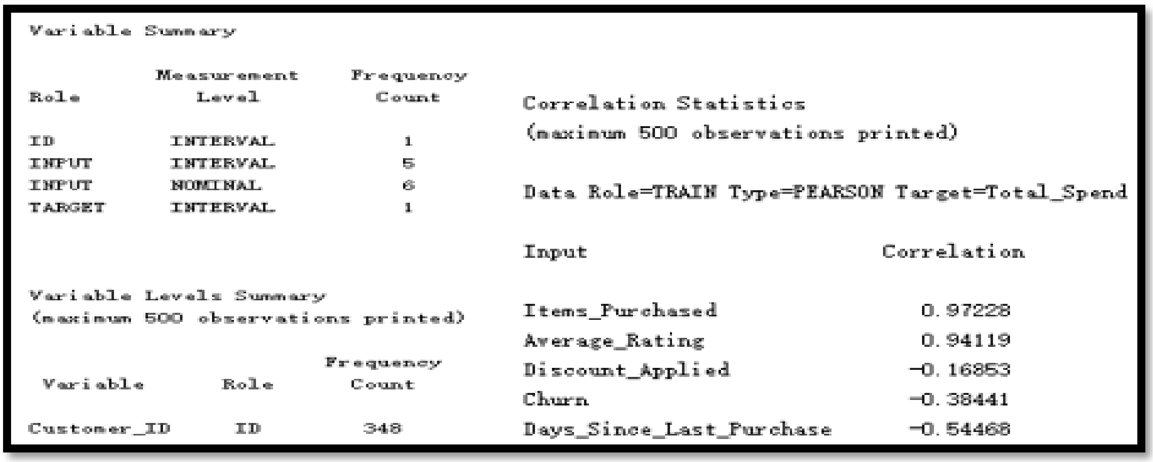
   The node ’File Import’ is drag to diagram and click import file to import exported dataset from Talend Data Preparation Tool to SAS Enterprise Miner. The 'Customer ID' is set as an 'ID' role to uniquely identify each record. The 'Total Spend' has been designated as the 'Target' variable, which allows for the identification of characteristics among high spenders. Understanding these characteristics enables more effective targeting of marketing efforts towards segments with the potential for higher revenue generation or increased sales. All other variables have been assigned the default role of 'Input', serving as predictors in the model. Besides, the age attribute is drop since the dataset already exist age group.



6. Data Explore



   Explore the data using ‘StatExplore’, there was 5 interval and 6 nominal input variable and 1 interval target variable ‘Total Spend’. The correlation statistics shows that ‘Item Purchased’ and ‘Average Rating’ most positive correlation to target variable.



Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	INTERVAL	5
INPUT	NOMINAL	6
TARGET	INTERVAL	1

Variable Levels Summary
(maximum 500 observations printed)

Variable	Role	Frequency Count
Customer_ID	ID	348

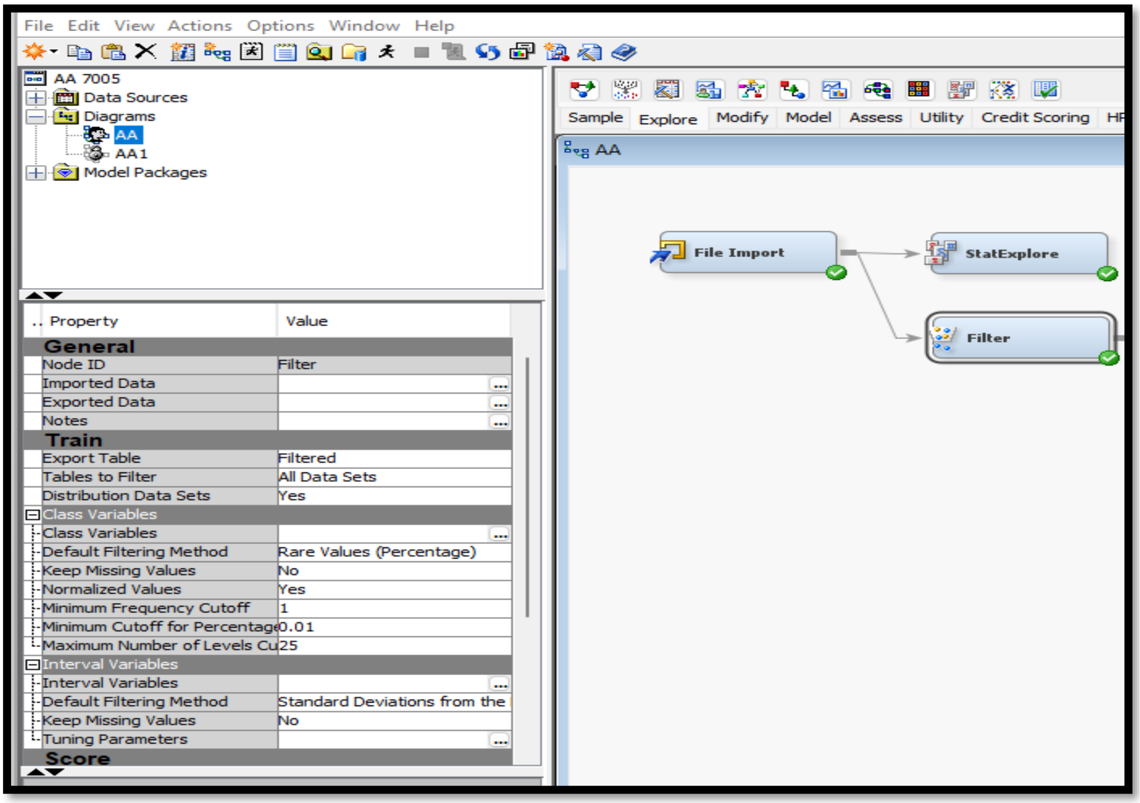
Correlation Statistics
(maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=Total_Spend

Input	Correlation
Items_Purchased	0.97228
Average_Rating	0.94119
Discount_Applied	-0.16853
Churn	-0.38441
Days_Since_Last_Purchase	-0.54468

7. Remove outlier

In SAS Enterprise Miner, “Filter” function is used to remove outlier and missing value. “Filter” setting is to 'Rare Values (Percentage)' probably means that the filter will remove values that occur below a certain threshold percentage. Choose to drop the missing values and exclude single-occurrence values as they may be noise. For interval variables, filter out values far from the mean, targeting outliers. As result, there has been 1 observation in Favourite Category been filter out and there is no any outlier presented.



Excluded Class Values
(maximum 500 observations printed)

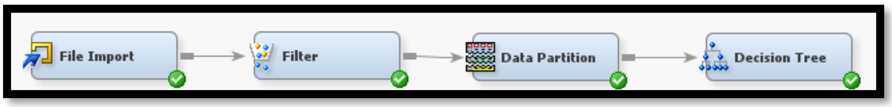
Variable	Role	Level	Train Count	Train Percent	Label	Filter Method
Favorite_Category	INPUT	BOOKS AND MEDIA	1	0.28736	Favorite Category	MINPCT

Number Of Observations

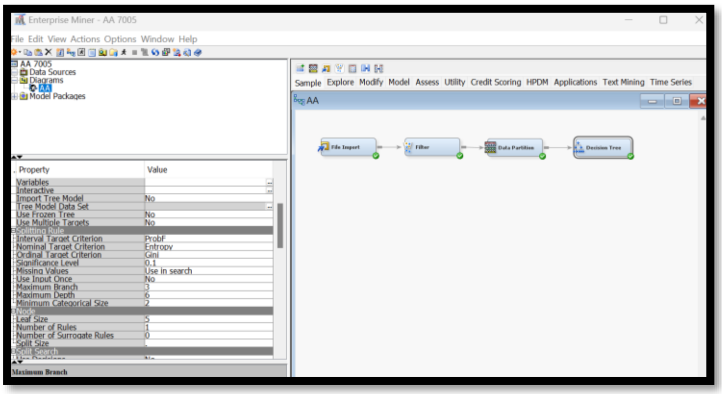
Data Role	Filtered	Excluded	DATA
TRAIN	347	1	348

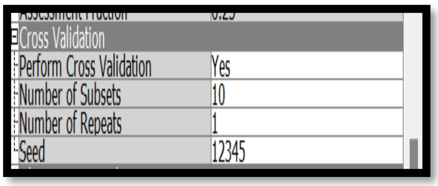
Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

[20 marks]



The data partition is set to a 70-30 split to separate the dataset into training and testing subsets. However, such a split is typically more suited for larger datasets. Therefore, to prevent overfitting and ensure the model's robustness and generalizability, I have chosen to use cross-validation instead of a simple data split. For the decision tree settings, the usual significance level for a splitting rule is 0.05, but I have selected a significance level of 0.1. This more lenient threshold permits splits that may not meet conventional standards of statistical significance but are appropriate for an exploratory model like a decision tree, which seeks to identify potential patterns. I have limited the maximum number of branches to three, aligning with the majority of categorical variables in the dataset that have 2 or 3 categories, with the exception of 'city', which has 6.





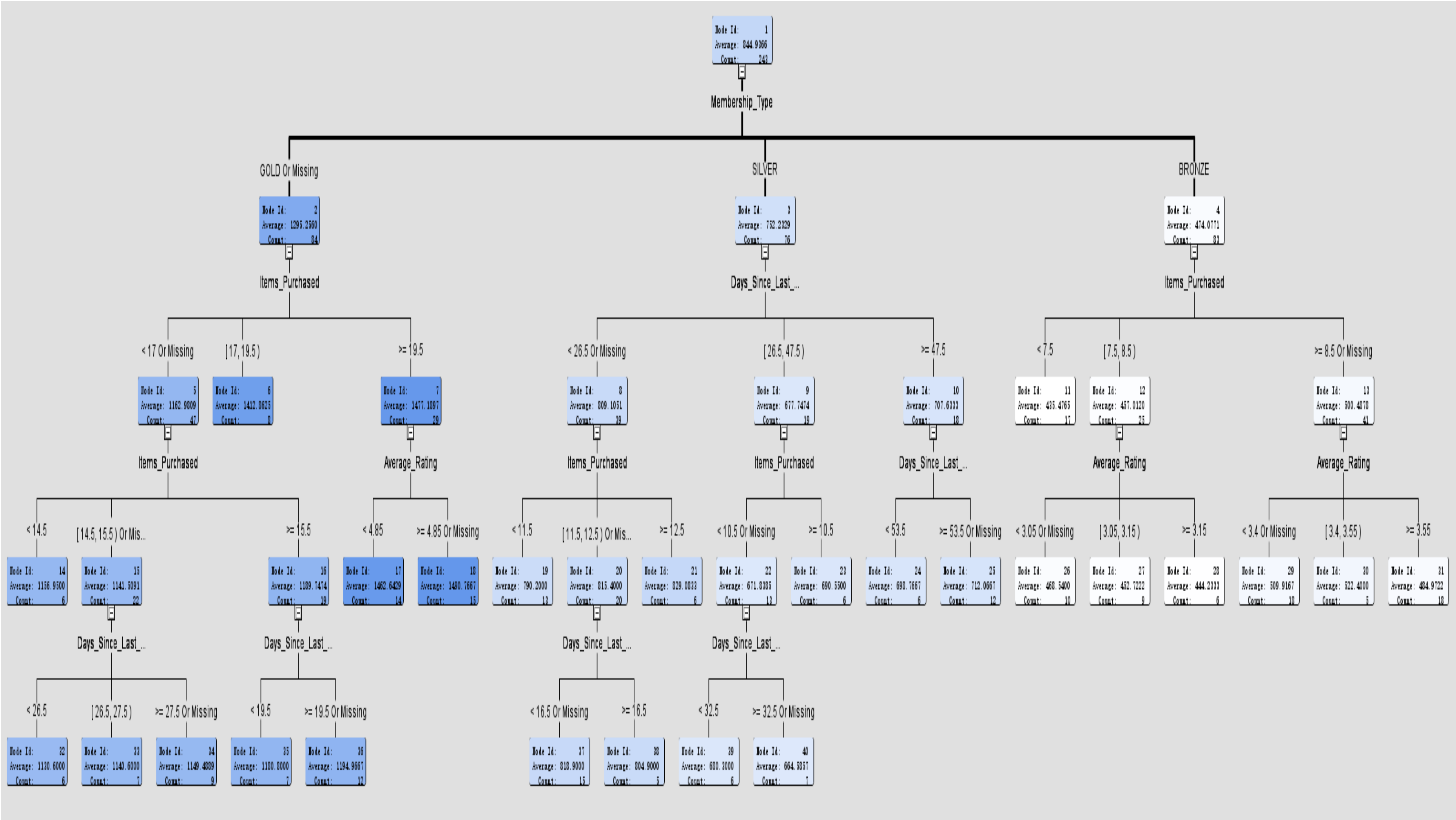
The regression tree analysis, with 'Total Spend' as the target variable, demonstrates how customer behaviors and satisfaction influence spending amounts. The model developed using key customer related features such as ‘Items Purchased’, ‘Days Since Last Purchase’ and ‘Average Rating’. The model has facilitated the identification of patterns and average spending figures within distinct customer segments, which can inform targeted marketing strategies and customer relationship management.

For example, the ‘Membership Type’ was initial split, indicating its significant influence on target variable ‘Total Spend’. In the 'GOLD' category, the regression tree reveals three spending tiers: members with fewer than 17 purchases have a lower average spend, indicating a base spending level. Those with 17 to 19.5 purchases occupy a mid-tier spending bracket, while members with over 19.5 purchases emerge as the top spenders, with the highest average spend within the group, illustrating a clear positive correlation between the frequency of purchases and spending levels among 'GOLD' members.

In the 'SILVER' category, the regression tree identifies spending patterns based on how recently they've made purchases: those buying within 26.5 days, those between 26.5 to 47.5 days, and those with more than 47.5 days since their last purchase, each with distinct average spending levels, implying more recent shoppers spend differently than less frequent ones.

In the 'BRONZE' category, fewer purchases (less than 7.5) correlate with lower spending, while higher customer ratings suggest distinct spending behaviors, indicating that both purchase frequency and satisfaction levels are key indicators of spending within this group.

Overall, the 'GOLD' category demonstrate that increased purchase frequency aligns with higher average spending within ‘Gold’ members. The 'SILVER' category demonstrates that the timing of purchases is a predictive factor for spending, with longer intervals between purchases potentially indicating a different type of spending behavior. Besides, the 'BRONZE' category demonstrate that both the quantity of items purchased and the satisfaction level are closely linked to their spending habits. These insights about 'Total Spend' suggest that spending isn't just about how much members buy or how satisfied they are, but also about how these factors interact across different membership levels. This nuanced understanding of spending behavior is vital for tailoring strategies to enhance customer engagement and increase total spending across each segment.



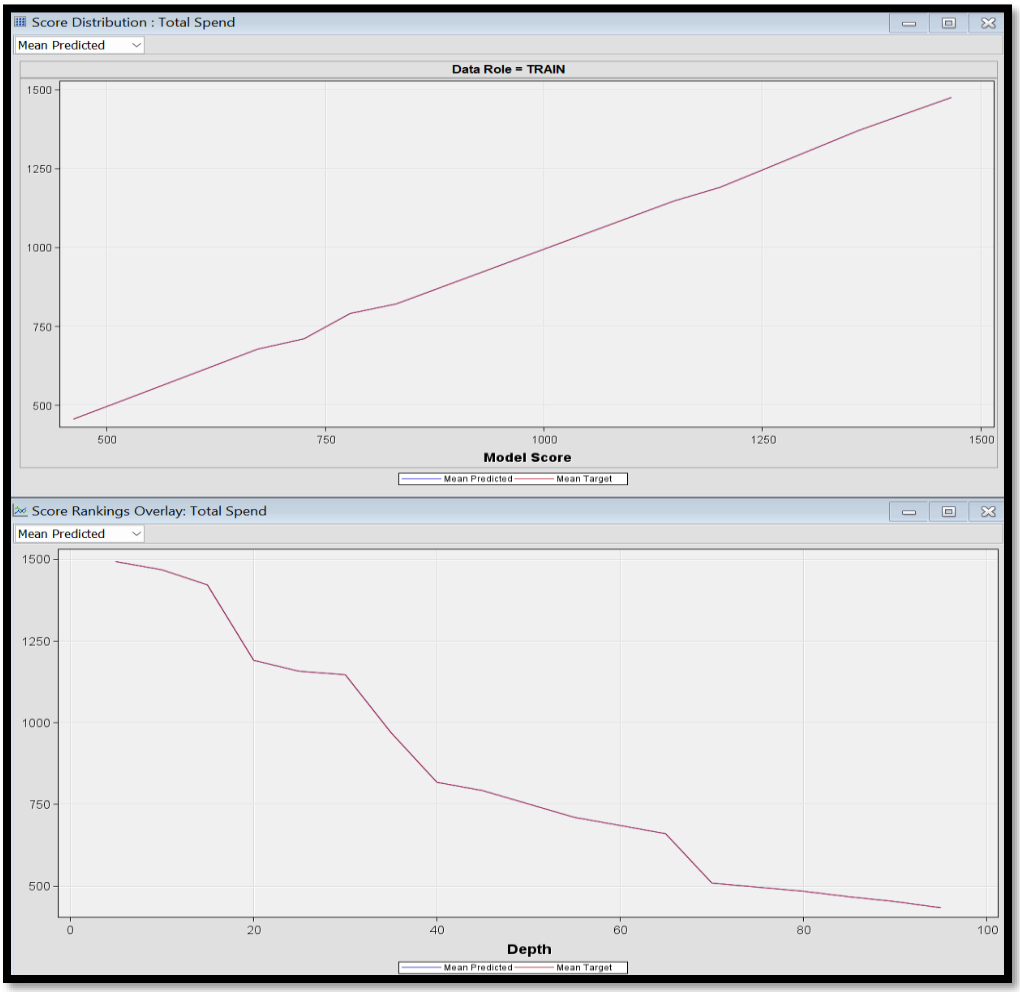
Node Id: 1
Average: 844.9366
Count: 243
Membership_Type
GOLD Or Missing
Node Id: 2
Average: 1295.2560
Count: 84
Items_Purchased
< 17 Or Missing
Node Id: 5
Average: 1162.9809
Count: 47
Items_Purchased
< 14.5
Node Id: 14
Average: 1156.9500
Count: 6
[14.5, 15.5) Or Mis...
Node Id: 15
Average: 1141.5091
Count: 22
Days_Since_Last_...
< 26.5
Node Id: 32
Average: 1130.6000
Count: 6
[26.5, 27.5)
Node Id: 33
Average: 1140.6000
Count: 7
>= 27.5 Or Missing
Node Id: 34
Average: 1149.4889
Count: 9
>= 15.5
Node Id: 16
Average: 1189.7474
Count: 19
Days_Since_Last_...
< 19.5
Node Id: 35
Average: 1180.8000
Count: 7
>= 19.5 Or Missing
Node Id: 36
Average: 1194.9667
Count: 12
[17, 19.5)
Node Id: 6
Average: 1412.8625
Count: 8
>= 19.5
Node Id: 7
Average: 1477.1897
Count: 29
Average_Rating
< 4.85
Node Id: 17
Average: 1462.6429
Count: 14
>= 4.85 Or Missing
Node Id: 18
Average: 1490.7667
Count: 15
SILVER
Node Id: 3
Average: 752.2329
Count: 76
Days_Since_Last_...
< 26.5 Or Missing
Node Id: 8
Average: 809.1051
Count: 39
Items_Purchased
< 11.5
Node Id: 19
Average: 790.2000
Count: 13
[11.5, 12.5) Or Mis...
Node Id: 20
Average: 815.4000
Count: 20
Days_Since_Last_...
< 16.5 Or Missing
Node Id: 37
Average: 818.9000
Count: 15
>= 16.5
Node Id: 38
Average: 804.9000
Count: 5
>= 12.5
Node Id: 21
Average: 829.0833
Count: 6
[26.5, 47.5)
Node Id: 9
Average: 677.7474
Count: 19
Items_Purchased
< 10.5 Or Missing
Node Id: 22
Average: 671.8385
Count: 13
Days_Since_Last_...
< 32.5
Node Id: 39
Average: 680.3000
Count: 6
>= 32.5 Or Missing
Node Id: 40
Average: 664.5857
Count: 7
>= 10.5
Node Id: 23
Average: 690.5500
Count: 6
>= 47.5
Node Id: 10
Average: 707.6333
Count: 18
Days_Since_Last_...
< 53.5
Node Id: 24
Average: 698.7667
Count: 6
>= 53.5 Or Missing
Node Id: 25
Average: 712.0667
Count: 12
BRONZE
Node Id: 4
Average: 474.0771
Count: 83
Items_Purchased
< 7.5
Node Id: 11
Average: 435.4765
Count: 17
[7.5, 8.5)
Node Id: 12
Average: 457.0120
Count: 25
Average_Rating
< 3.05 Or Missing
Node Id: 26
Average: 468.5400
Count: 10
[3.05, 3.15)
Node Id: 27
Average: 452.7222
Count: 9
>= 3.15
Node Id: 28
Average: 444.2333
Count: 6
>= 8.5 Or Missing
Node Id: 13
Average: 500.4878
Count: 41
Average_Rating
< 3.4 Or Missing
Node Id: 29
Average: 509.9167
Count: 18
[3.4, 3.55)
Node Id: 30
Average: 522.4000
Count: 5
>= 3.55
Node Id: 31
Average: 484.9722
Count: 18

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

[10 marks]

Random Forest algorithm as a Bagging, the bagging involves creating multiple decision tree, each trained on different sample of the dataset and combined output. The score distribution plot shows the relationship between the predicted total spend (Model Score) and the actual total spend (Mean Target) for the training set. The plot is a 45-degree line plot, which suggests a perfect prediction where every point's predicted value matches the target value. Score ranking overlay plot indicates the model's accuracy at different levels of confidence in predictions. The descending line shows model is less accurate at predicting spending customers. The model is better at predicting higher spenders than lower spenders.

The model achieves accurate predictions with a low ASE and RASE of 7.87, indicating precision in forecasting spending. Besides, the model captures complex spending patterns by utilizing deeper trees, enable to distinguish between different customer behaviors. The model consistently predicts both high and low spending across various depths, demonstrating its reliability in capturing diverse spending behaviors.



Fit Statistics

Target=Total_Spend Target Label=Total Spend

Fit Statistics	Statistics Label	Train
ASE	Average Squared Error	61.92
DIV	Divisor for ASE	347.00
MAX	Maximum Absolute Error	38.09
NOBS	Sum of Frequencies	347.00
RASE	Root Average Squared Error	7.87
SSE	Sum of Squared Errors	21487.14

Assessment Score Rankings

Data Role=TRAIN Target Variable=Total_Spend Target Label=Total Spend

Depth	Number of Observations	Mean Target	Mean Predicted
5	24	1492.18	1492.18
10	12	1467.17	1467.17
15	22	1420.38	1420.38
20	25	1189.99	1189.99
25	10	1158.29	1158.29
30	17	1146.48	1146.48
35	15	970.01	970.01
40	19	818.27	818.27
45	32	791.90	791.90
55	18	710.96	710.96
60	30	685.40	685.40
65	9	661.44	661.44
70	34	510.80	510.80
80	24	484.42	484.42
85	12	468.87	468.87
90	16	453.20	453.20
95	28	435.38	435.38

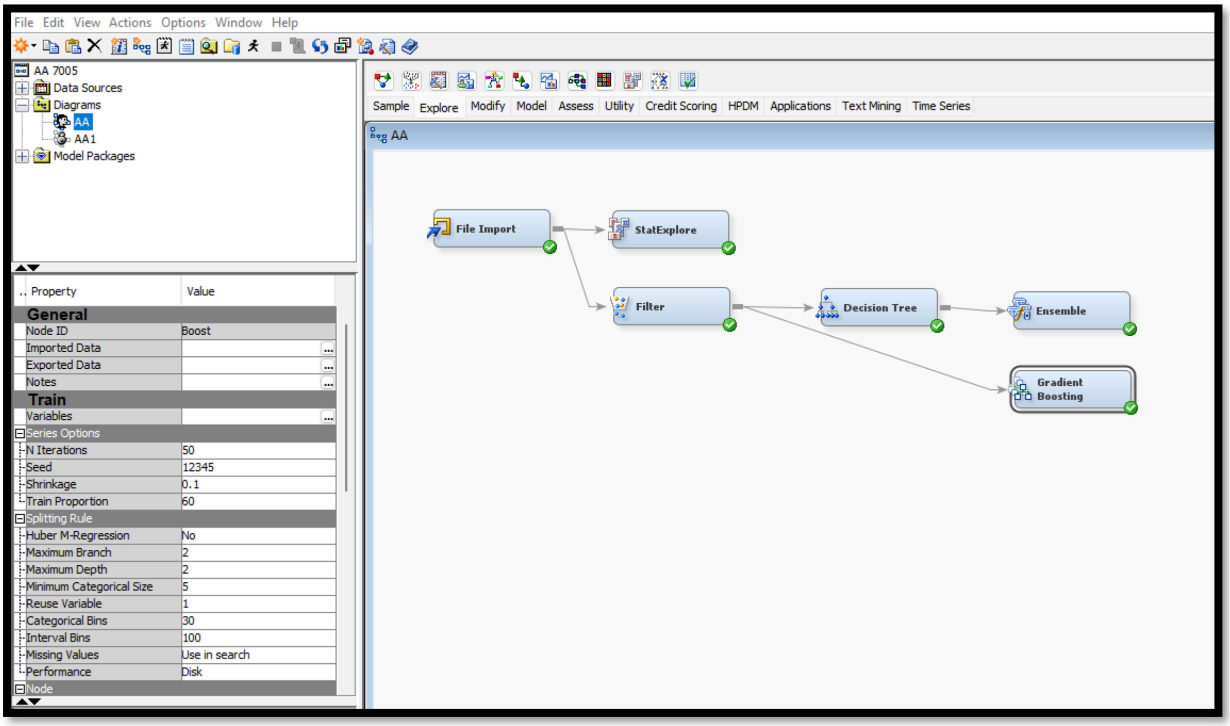
Assessment Score Distribution

Data Role=TRAIN Target Variable=Total_Spend Target Label=Total Spend

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
1439.309 - 1492.183	1476.02	1476.02	49	1465.75
1333.560 - 1386.434	1371.31	1371.31	9	1360.00
1174.936 - 1227.810	1189.99	1189.99	25	1201.37
1122.061 - 1174.936	1146.69	1146.69	34	1148.50
804.814 - 857.688	821.60	821.60	27	831.25
751.939 - 804.814	791.90	791.90	32	778.38
699.064 - 751.939	710.96	710.96	18	725.50
646.190 - 699.064	679.87	679.87	39	672.63
487.566 - 540.441	510.80	510.80	34	514.00
434.691 - 487.566	458.68	458.68	80	461.13

For Gradient Boosting model, it builds one tree at a time, where each new tree helps to correct errors made by previously trained trees. The Gradient Boosting model is set to run for 50 iterations to build enough trees for accurate predictions while avoiding overfitting and excessive computation time. A learning rate of 0.1 helps the model learn steadily without making drastic changes that could lead to overfitting. Training on 60% of the data allows for a robust learning process while reserving 40% for testing the model's ability to predict new data, ensuring it performs well not just on the training data but also on unseen data. These settings aim to create a well-generalizing model that balances learning complexity and predictive reliability.

The key findings are that 'Items Purchased' and 'Average Rating' are the most influential factors predicting customer spending. The statistical output from the boosting model demonstrates a strong performance with a low Average Squared Error, indicating that the predictions are generally close to the actual spending amounts. The model performs better at predicting higher spending, as shown by the consistent decrease in predicted values with increasing depth, suggesting higher confidence in these predictions. The model's effectiveness is further suggested by the close match between the mean predicted and actual values across different segments of the data. This consistency is crucial for the model’s reliability in practical applications. These insights can guide strategies to enhance customer spending, such as encouraging more purchases or improving product ratings.



Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
Items Purchased		67	1
Average Rating		29	0.478943
Days Since Las...		23	0.240761
City	City	17	0.208358
Satisfaction Level		6	0.165609
Discount Applied		3	0.036897
Favorite Category		4	0.023039
Churn	Churn	1	0.00344
Gender	Gender	0	0
Membership Type		0	0
Age Group		0	0

Fit Statistics

Target=Total_Spend Target Label=Total Spend

Fit Statistics	Statistics Label	Train
NOBS	Sum of Frequencies	347.00
SUMW	Sum of Case Weights Times Freq	347.00
MAX	Maximum Absolute Error	56.27
SSE	Sum of Squared Errors	47359.35
ASE	Average Squared Error	136.48
RASE	Root Average Squared Error	11.68
DIV	Divisor for ASE	347.00
DFT	Total Degrees of Freedom	347.00

Assessment Score Rankings

Data Role=TRAIN Target Variable=Total_Spend Target Label=Total Spend

Depth	Number of Observations	Mean Target	Mean Predicted
5	18	1495.10	1484.71
10	24	1469.15	1466.61
15	16	1405.96	1405.50
20	12	1187.45	1189.81
25	22	1180.14	1168.60
30	17	1146.48	1145.62
35	16	979.40	986.64
40	16	817.77	815.89
45	35	794.39	793.52
55	24	707.90	707.59
60	9	689.49	691.50
65	19	674.51	672.29
70	29	532.36	534.71
75	7	520.40	499.64
80	27	489.39	498.24
85	7	470.50	465.28
90	15	460.55	461.71
95	17	441.35	452.21
100	17	433.15	440.61

Assessment Score Distribution

Data Role=TRAIN Target Variable=Total_Spend Target Label=Total Spend

Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
1433.032 - 1485.285	1476.02	1470.62	49	1459.16
1328.526 - 1380.779	1371.31	1372.31	9	1354.65
1171.767 - 1224.020	1189.54	1187.66	24	1197.89
1119.514 - 1171.767	1148.23	1145.87	35	1145.64
858.248 - 910.501	820.75	877.02	1	884.37
805.995 - 858.248	821.63	817.68	26	832.12
753.742 - 805.995	791.90	791.58	32	779.87
701.489 - 753.742	707.90	707.59	24	727.62
649.236 - 701.489	676.44	677.27	33	675.36
492.477 - 544.730	499.88	501.78	58	518.60
440.224 - 492.477	447.65	452.87	56	466.35