UNIVERSITI MALAYA

# WQD7005 DATA MININGI/2023/2024

## Final exam

| Student Name | Student Matric Number |
|---|---|
| DONG XUE YING | 22074835 |

**Dataset:**

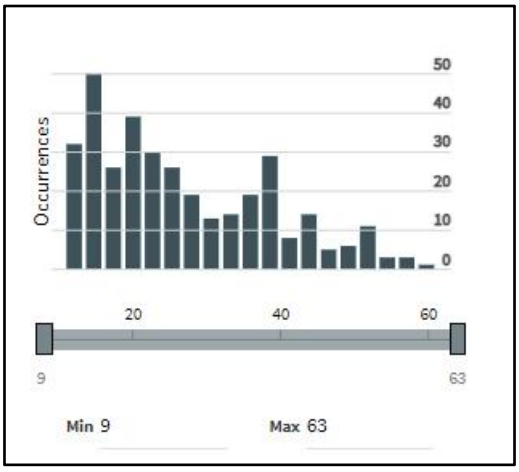
| | |
|---|---|
| Customer ID | Unique identifier for each customer. |
| Gender | Gender of customer |
| Age | age of the customer |
| City | Customer's city location |
| Membership Type | Indicates the membership level (Bronze, Silver, Gold) |
| Favorite Category | Customer's favorite product category |
| Total Spend | Total customer expenditure |
| Items Purchased | Number of items purchased per customer |
| Average Rating | Average rating of product, service or experience |
| Discount Applied | Whether discounts are applied |
| Days Since Last Purchase | Time since customer's last purchase |
| churn | Indicates whether customers are lost |
| Satisfaction Level | Customer satisfaction with products and services |

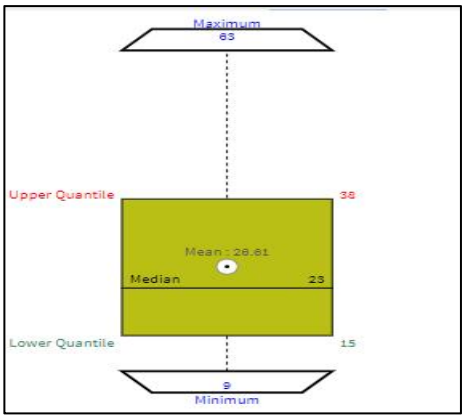
**Data Preprocessing in talent data preparation**

1. Import dataset into talend data preparation.

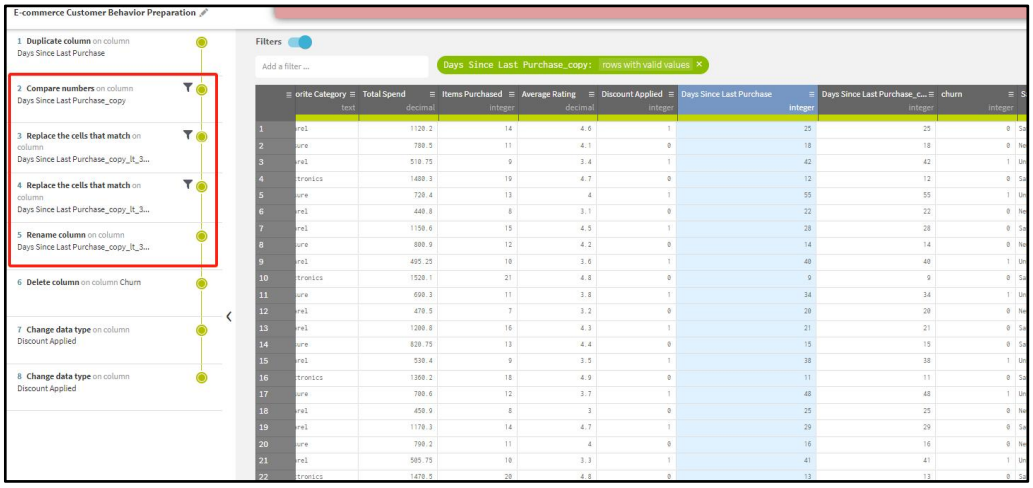

2. Using taland, add a 'churn' column

First, use the compare rule to set the compare mode to lower than and the value to 30; second, use replace the cells that match to replace 'true' with '0' and 'false' with '1'.

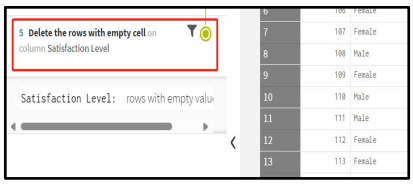



The above figure shows the histogram and box plot of the 'Days since Last Purchase' attribute, from these plots it can be observed that the number of active days is concentrated in the range of 15-38, with a median of 23. The Compare Numbers function is used to evaluate whether the customer is greater than 30 or not, and then convert the value from 'true' to '0' for no churn and 'false' to '1' for churn.




3. Handle missing value by removing vacant values from the data set.



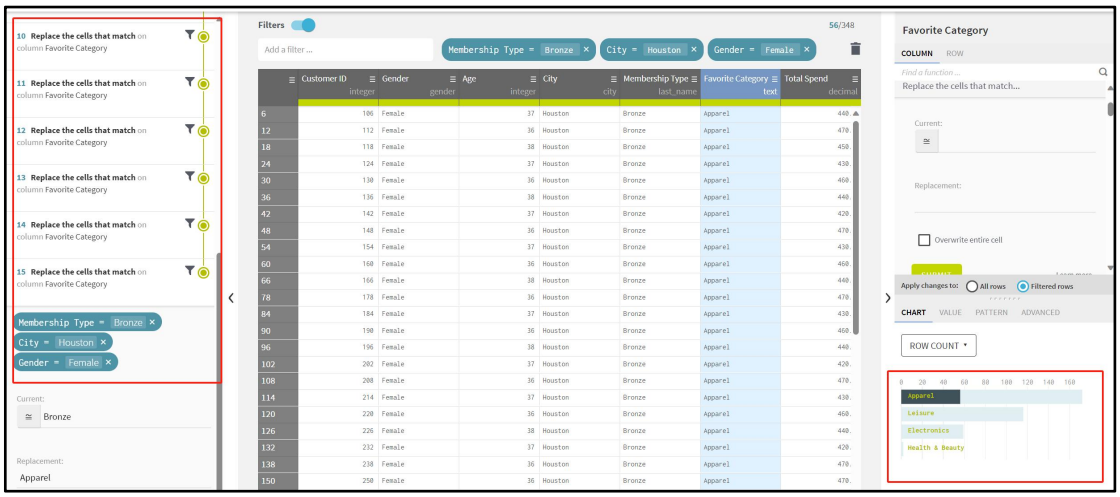

4. Adding a 'favorite category' column

Based on Gender, City, and Membership Type, set up a purchase product type categorization rule, and configure Favorite Category variables for each attribute.

| Gender | City | Membership Type | Favorite Category |
|---|---|---|---|
| Male | New York | Gold | Electronics |
| Female | New York | Gold | Apparel |
| Male | Los Angeles | Gold | Leisure |
| Female | Los Angeles | Gold | Health & Beauty |
| Male | San Francisco | Gold | Electronics |
| Female | San Francisco | Gold | Apparel |
| Male | Chicago | Gold | Home Goods |
| Female | Chicago | Gold | Apparel |
| Male | Miami | Gold | Leisure |
| Female | Miami | Gold | Health & Beauty |
| Male | Houston | Sliver | Home Goods |
| Female | Houston | Sliver | Apparel |
| Male | New York | Sliver | Electronics |
| Female | New York | Sliver | Apparel |
| Male | Los Angeles | Sliver | Leisure |
| Female | Los Angeles | Sliver | Health & Beauty |
| Male | San Francisco | Sliver | Electronics |
| Female | San Francisco | Sliver | Home Goods |
| Male | Chicago | Sliver | Home Goods |
| Female | Chicago | Sliver | Apparel |
| Male | Miami | Sliver | Leisure |

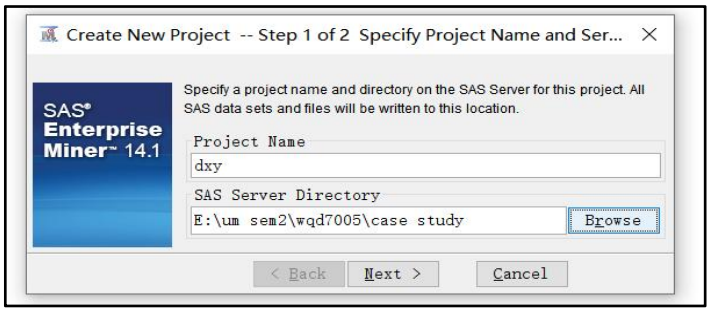
| | | | |
|---|---|---|---|
| Female | Miami | Sliver | Health & Beauty |
| Male | Houston | Sliver | Home Goods |
| Female | Houston | Sliver | Apparel |
| Male | New York | Bronze | Electronics |
| Female | New York | Bronze | Apparel |
| Male | Los Angeles | Bronze | Leisure |
| Female | Los Angeles | Bronze | Health & Beauty |
| Male | San Francisco | Bronze | Electronics |
| Female | San Francisco | Bronze | Home Goods |
| Male | Chicago | Bronze | Home Goods |
| Female | Chicago | Bronze | Apparel |
| Male | Miami | Bronze | Leisure |
| Female | Miami | Bronze | Health & Beauty |
| Male | Houston | Bronze | Electronics |
| Female | Houston | Bronze | Apparel |

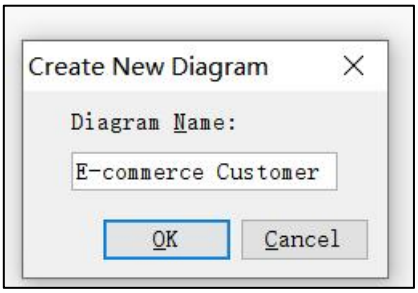
The final four appropriate types are: Apparel, Leisure, Health & Beauty, and Electronics.



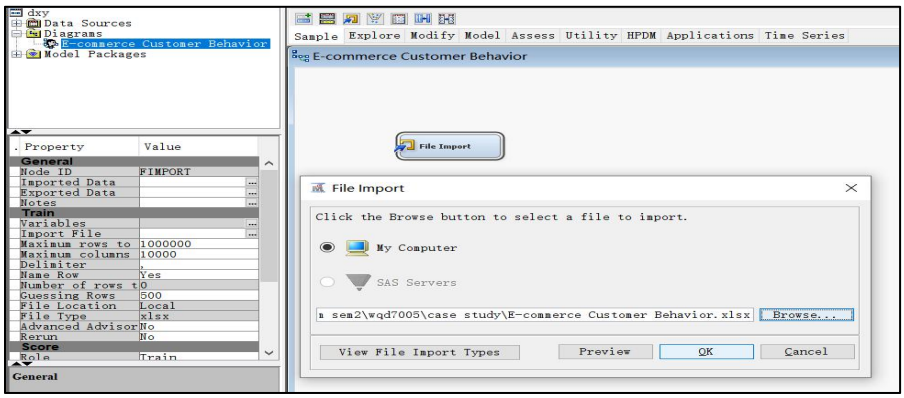

**Data Import and Preprocessing:** Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles.

1. Create new diagram，Create new diagram




2. Import dataset into SAS Enterprise Miner

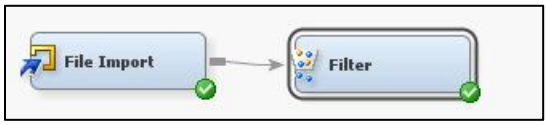



3. specify variable roles，Set Total spend to target and customer id to ID.

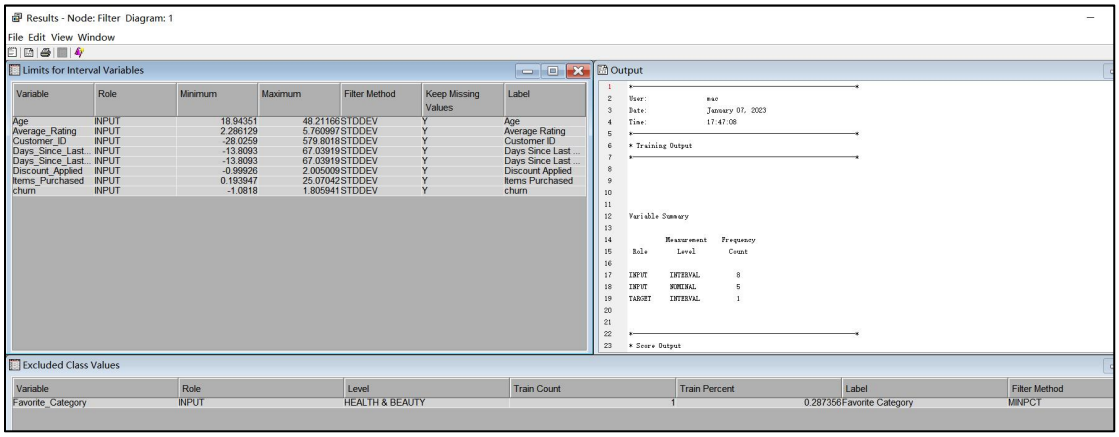
| Name | Role | Level | Report | Order | Drop | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|---|
| Age | Input | Interval | No | | No | . | . |
| Average_Ra | Input | Interval | No | | No | . | . |
| City | Input | Nominal | No | | No | . | . |
| Customer_ID | ID | Interval | No | | No | . | . |
| Days_Since_ | Input | Interval | No | | No | . | . |
| Days_Since_ | Input | Interval | No | | No | . | . |
| Discount_A | Input | Interval | No | | No | . | . |
| Favorite_C | Input | Nominal | No | | No | . | . |
| Gender | Input | Nominal | No | | No | . | . |
| Items_Purc | Input | Interval | No | | No | . | . |
| Membership | Input | Nominal | No | | No | . | . |
| Satisfacti | Input | Nominal | No | | No | . | . |
| Total_Spen | Target | Interval | No | | No | . | . |
| churn | Input | Interval | No | | No | . | . |

4. Handling of outliers

Use the 'filter' node to query and remove outliers




Run the node and find an outlier in the data.




The Excluded Class Values is ': "HEALTH & BEAUTY" in "Favorite_Category", indicating that this outlier has been excluded.

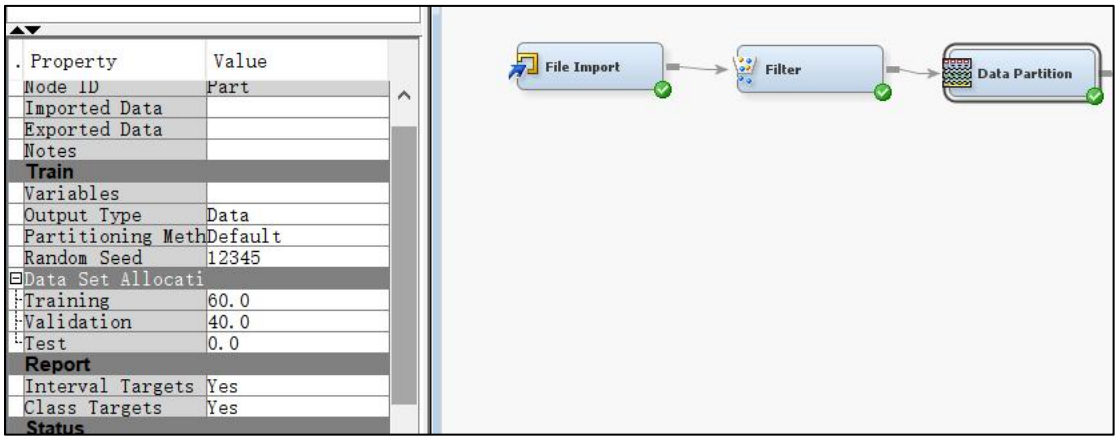
**Decision Tree Analysis:** Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

1. Using the Data Partitioning node

For model training and evaluation, the dataset is divided into 60% training set and 40% validation

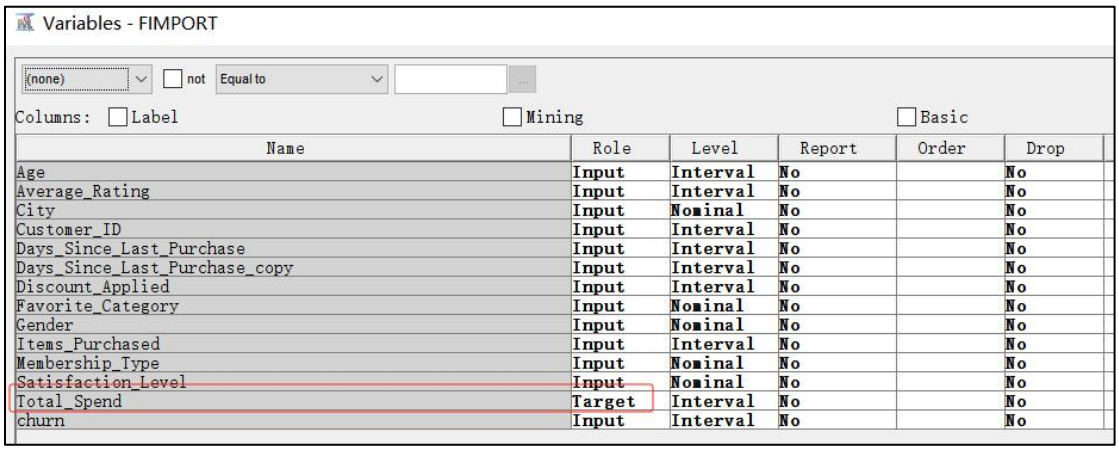
set using Data Partitioning node. The purpose of this allocation is to fully utilize the data to train the model.

The training set accounts for 60% of the overall data, providing enough data for model learning and training. The validation set accounts for 40% of the overall data and is used to evaluate the performance of the model.



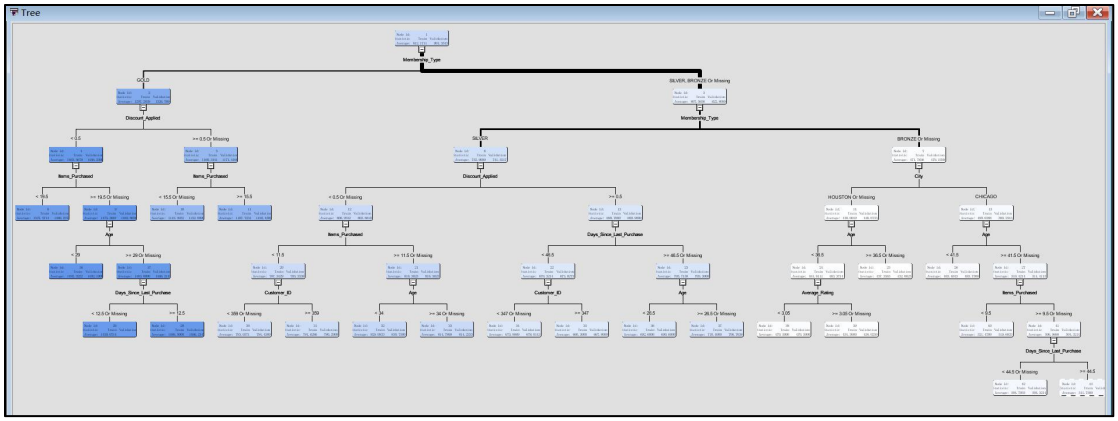

1. Creating a Decision Tree

Setting the target variable,The total spend was set as the target variable to analyze the key factors affecting the total spend amount.




| Name | Role | Level | Report | Order | Drop |
|---|---|---|---|---|---|
| Age | Input | Interval | No | | No |
| Average_Rating | Input | Interval | No | | No |
| City | Input | Nominal | No | | No |
| Customer_ID | Input | Interval | No | | No |
| Days_Since_Last_Purchase | Input | Interval | No | | No |
| Days_Since_Last_Purchase_copy | Input | Interval | No | | No |
| Discount_Applied | Input | Interval | No | | No |
| Favorite_Category | Input | Nominal | No | | No |
| Gender | Input | Nominal | No | | No |
| Items_Purchased | Input | Interval | No | | No |
| Membership_Type | Input | Nominal | No | | No |
| Satisfaction_Level | Input | Nominal | No | | No |
| Total_Spend | Target | Interval | No | | No |
| churn | Input | Interval | No | | No |

Adding Decision Tree Nodes

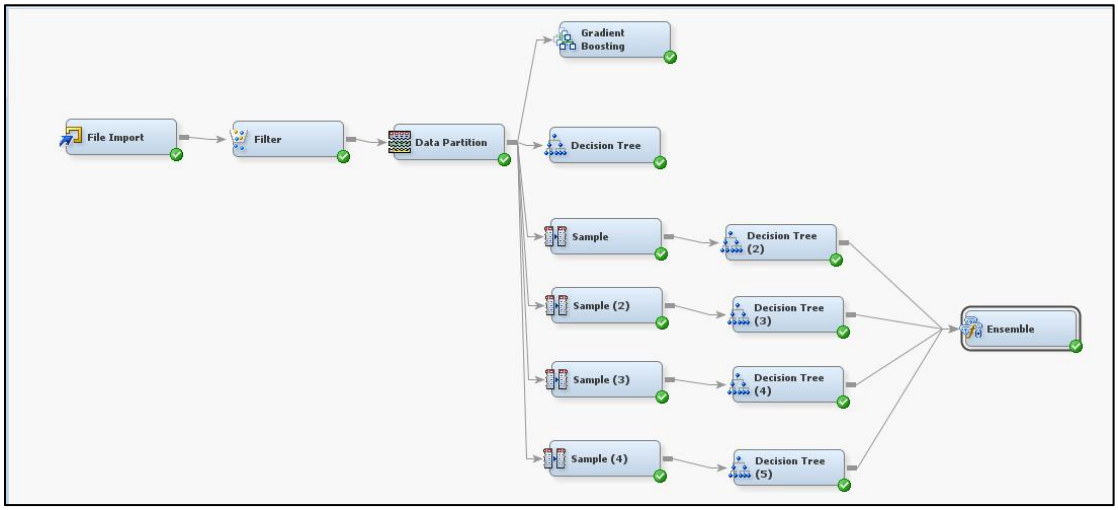



This decision tree is the original tree, set the attribute 'Total spend' as the target variable

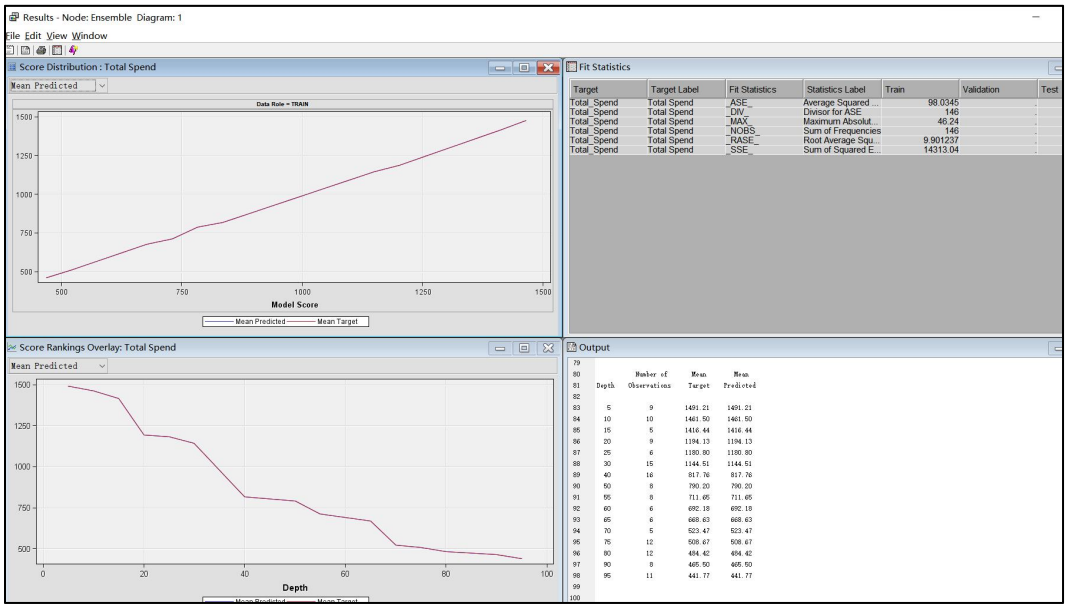
- From the decision tree, we can find that 'membership type' is the biggest factor that affects 'Total spend', when the variable is gold, the total spend is higher; when the variable is When the variable is GOLD, the total spend is more; when the variable is SILVER, BRONZE, the total spend is less.
- With GOLD as the node, the amount of total spend is higher when 'discount applied' <0.5 (0: no discount); 'discount applied' >0.5 ( 1: with discount), the total amount spent is less.
- SILVER, BRONZE are the second most influential factors.Using sliver, bronze as a node, the total amount spent is higher when the customer class is 'sliver' and lower when it is 'bronze'.
- All other variables have an impact on total spend, but to a lesser extent.

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

Use sample to randomly sample the data at 70%; add decision tree nodes after each given sample and connect all these decision trees to ensemble, run ensemble to see the fit of the model.



Run ensemble node



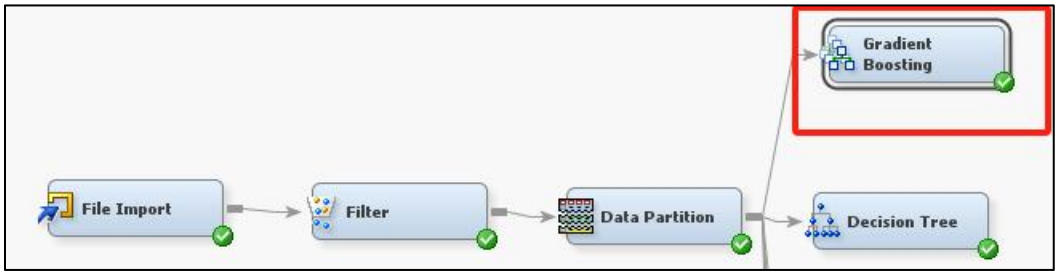

| Fit Statistics | Statistics Label | Train |
|---|---|---|
| _ASE_ | Average Squared Error | 98.03 |
| _DIV_ | Divisor for ASE | 146.00 |
| _MAX_ | Maximum Absolute Error | 46.24 |
| _NOBS_ | Sum of Frequencies | 146.00 |
| _RASE_ | Root Average Squared Error | 9.90 |
| _SSE_ | Sum of Squared Errors | 14313.04 |

The ASE is 98.03, the maximum absolute error is 46.24, the root mean squared error RASE is 9.90, while the overall sum of squared errors SSE is 14313.04.This indicates that the overall prediction error of the model is relatively small and on average the predicted values of the model are closer to the actual values.

| Depth | Number of Observations | Mean Target | Mean Predicted |
|---|---|---|---|
| 5 | 9 | 1491.21 | 1491.21 |
| 10 | 10 | 1461.50 | 1461.50 |
| 15 | 5 | 1416.44 | 1416.44 |
| 20 | 9 | 1194.13 | 1194.13 |
| 25 | 6 | 1180.80 | 1180.80 |
| 30 | 15 | 1144.51 | 1144.51 |
| 40 | 16 | 817.76 | 817.76 |
| 50 | 8 | 790.20 | 790.20 |
| 55 | 8 | 711.65 | 711.65 |
| 60 | 6 | 692.18 | 692.18 |
| 65 | 6 | 668.63 | 668.63 |
| 70 | 5 | 523.47 | 523.47 |
| 75 | 12 | 508.67 | 508.67 |
| 80 | 12 | 484.42 | 484.42 |
| 90 | 8 | 465.50 | 465.50 |
| 95 | 11 | 441.77 | 441.77 |

The values of depth range from 5 to 95, and the equal values of Mean Predicted and Mean Target indicate a good model fit.

Using Gradient Boosting Nodes




Combining decision tree predictions will iteratively fit weak learners to the residuals of previous iterations, gradually improving the overall model performance. To create a predictive model

Variable Importance

| Obs | NAME | LABEL | NRULES | IMPORTANCE | VIMPORTANCE | RATIO |
|---|---|---|---|---|---|---|
| 1 | Items_Purchased | | 54 | 1.00000 | 1.00000 | 1.00000 |
| 2 | Average_Rating | | 18 | 0.37267 | 0.35246 | 0.94579 |
| 3 | Days_Since_Last_Purchase | | 25 | 0.23094 | 0.22043 | 0.95448 |
| 4 | City | City | 20 | 0.21136 | 0.20469 | 0.96846 |
| 5 | Satisfaction_Level | | 6 | 0.19934 | 0.19391 | 0.97277 |
| 6 | Age | Age | 7 | 0.14612 | 0.13560 | 0.92806 |
| 7 | Customer_ID | | 5 | 0.02148 | 0.00058 | 0.02690 |

Items_Purchased:NRULES: 54, IMPORTANCE: 1.00000, VIMPORTANCE: 1.00000, RATIO (Relative Importance to the Most Important Variable): 1.00000, Since Items_ Purchased has the highest IMPORTANCE and VIMPORTANCE, it is the most important variable in the model. This indicates that the number of items purchased plays a key role in the decision making of the model.

Average_Rating:NRULES: 18, IMPORTANCE: 0.37267, VIMPORTANCE: 0.35246, RATIO: 0.94579, Average_Rating has a relatively high importance in the model, but its importance is low compared to Items_Purchased.

Days_Since_Last_Purchase: NRULES: 25, IMPORTANCE: 0.23094, VIMPORTANCE: 0.22043, RATIO: 0.95448, Days_Since_Last_Purchase also contributes to the model's predictions, but is relatively low.

City: NRULES: 20, IMPORTANCE: 0.21136, VIMPORTANCE: 0.20469, RATIO: 0.96846, City is an important feature in the model and has an impact on the prediction goal.

Satisfaction_Level:NRULES: 6, IMPORTANCE: 0.19934
VIMPORTANCE: 0.19391, RATIO: 0.97277, Satisfaction_Level also has some impact on the output of the model and its importance is high.

Age:NRULES: 7, IMPORTANCE: 0.14612, VIMPORTANCE: 0.13560
RATIO: 0.92806, Age contributes to the prediction of the model but is relatively not the most

important variable.

Assessment Score Rankings

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

| Depth | Number of Observations | Mean Target | Mean Predicted |
|---|---|---|---|
| 5 | 12 | 1476.93 | 1469.45 |
| 10 | 13 | 1475.73 | 1463.83 |
| 15 | 7 | 1431.86 | 1438.05 |
| 20 | 12 | 1188.30 | 1205.30 |
| 25 | 13 | 1160.63 | 1153.51 |
| 30 | 6 | 1147.17 | 1141.72 |
| 35 | 10 | 1112.53 | 1106.91 |
| 40 | 16 | 820.86 | 820.78 |
| 45 | 7 | 807.94 | 803.85 |
| 50 | 8 | 790.20 | 793.89 |
| 55 | 14 | 708.97 | 710.36 |
| 60 | 7 | 691.99 | 693.02 |
| 65 | 17 | 585.24 | 584.66 |
| 70 | 7 | 505.75 | 508.25 |
| 75 | 13 | 497.25 | 496.97 |
| 80 | 5 | 497.31 | 496.65 |
| 85 | 16 | 469.19 | 468.76 |
| 90 | 5 | 458.50 | 454.31 |
| 95 | 11 | 438.99 | 451.08 |
| 100 | 9 | 431.94 | 449.73 |

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

| Depth | Number of Observations | Mean Target | Mean Predicted |
|---|---|---|---|
| 5 | 8 | 1488.90 | 1468.38 |
| 10 | 9 | 1483.61 | 1463.56 |
| 15 | 4 | 1405.35 | 1429.44 |
| 20 | 7 | 1320.34 | 1354.18 |
| 25 | 7 | 1189.37 | 1180.21 |
| 30 | 10 | 1148.48 | 1135.69 |
| 35 | 5 | 824.81 | 823.92 |
| 40 | 6 | 799.00 | 804.10 |
| 45 | 7 | 790.30 | 794.70 |
| 50 | 9 | 793.57 | 793.14 |
| 55 | 5 | 706.40 | 710.69 |
| 60 | 7 | 699.03 | 706.80 |
| 65 | 7 | 674.63 | 674.45 |
| 70 | 9 | 671.41 | 669.56 |
| 75 | 8 | 500.00 | 508.61 |
| 80 | 5 | 497.38 | 497.03 |
| 85 | 6 | 498.72 | 496.65 |
| 90 | 9 | 466.62 | 461.26 |
| 95 | 7 | 436.53 | 450.97 |
| 100 | 4 | 430.83 | 450.07 |

Based on the results of the evaluation score ranking, the performance of the model on TRAIN and VALIDATE can be observed for different Depths. In the training set, when the depth is 5 and 10, the model performs better and the predicted values are less different from the actual values. However, as the depth increases, the performance of the model on the training set becomes progressively worse and overfitting occurs, which is manifested as an overfitting to the training set and a relatively poor performance on the validation set. In the validation set, the model also performs better at depths of 5 and 10, but after a depth of 15, the model's performance gradually deteriorates.