S2100146 Anis Sofea

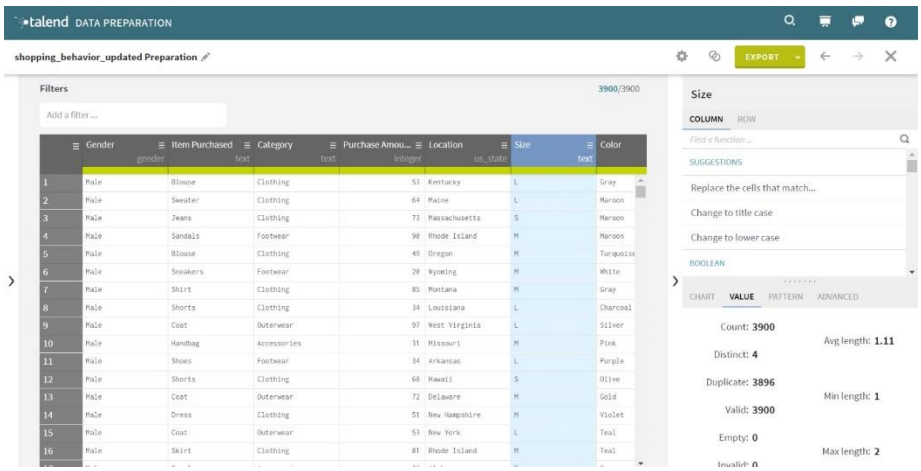
Alternative Assessment

Case Study: E-Commerce Customer Behavior Analysis

https://github.com/Anissoo56/WQD7005AA1

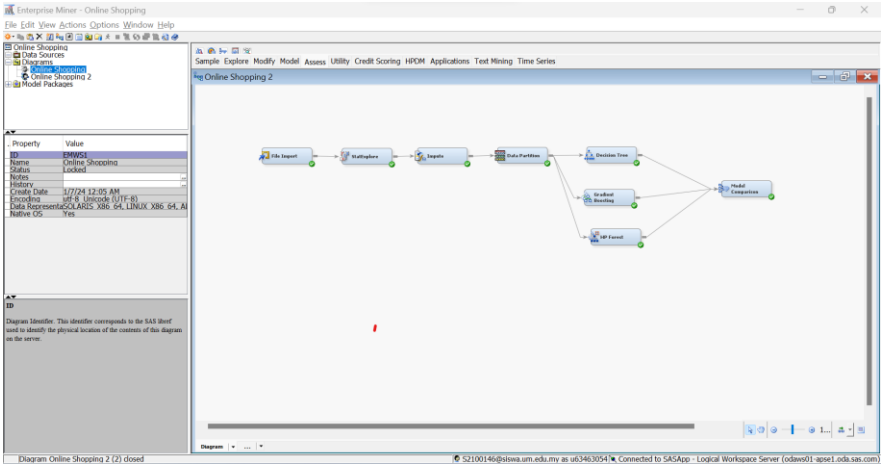
Talend Data Preparation

1) Talend Data Preparation is a tool that allows users to explore, cleanse, and manipulate data to prepare it for analysis or integration.

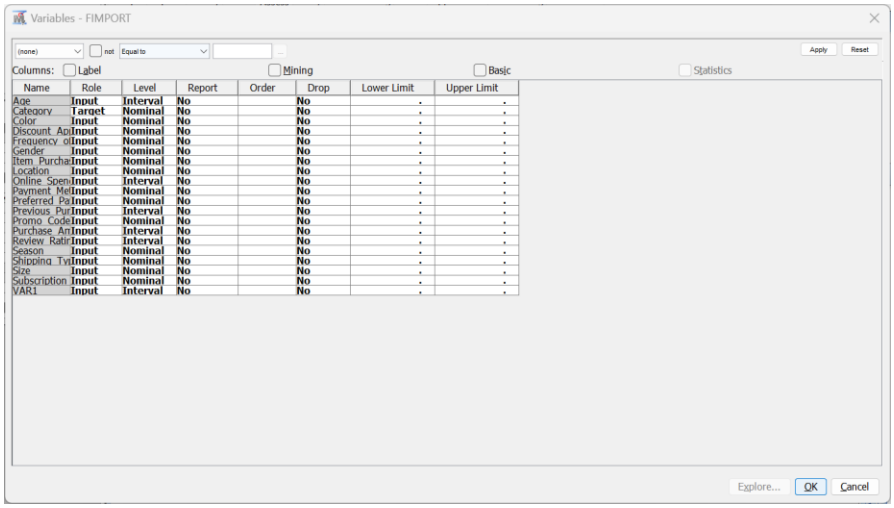


SAS Enterprise Miner

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

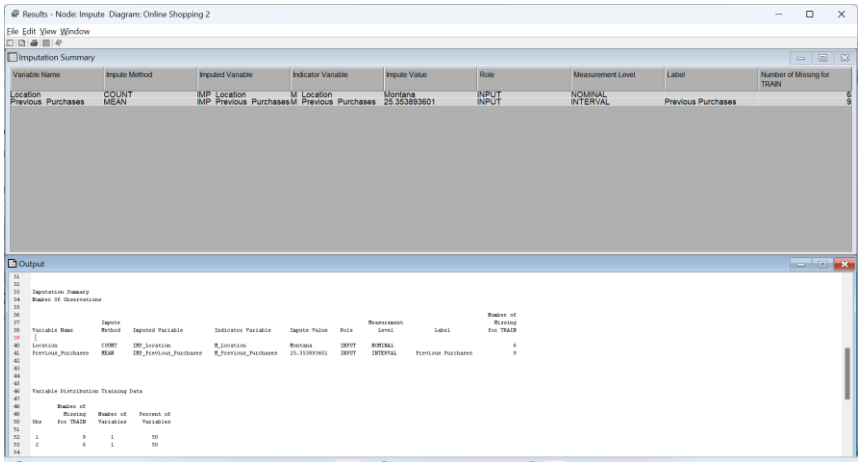


2) The role are classified corectly for each variable, The Category is specified as ‘Target’



3) Using the StatExplore, we can find the statistics missing value for each variable. Found out two variable name has missing values:

Location – 6 number of train
Previous_Purchases - 9 missing number of train



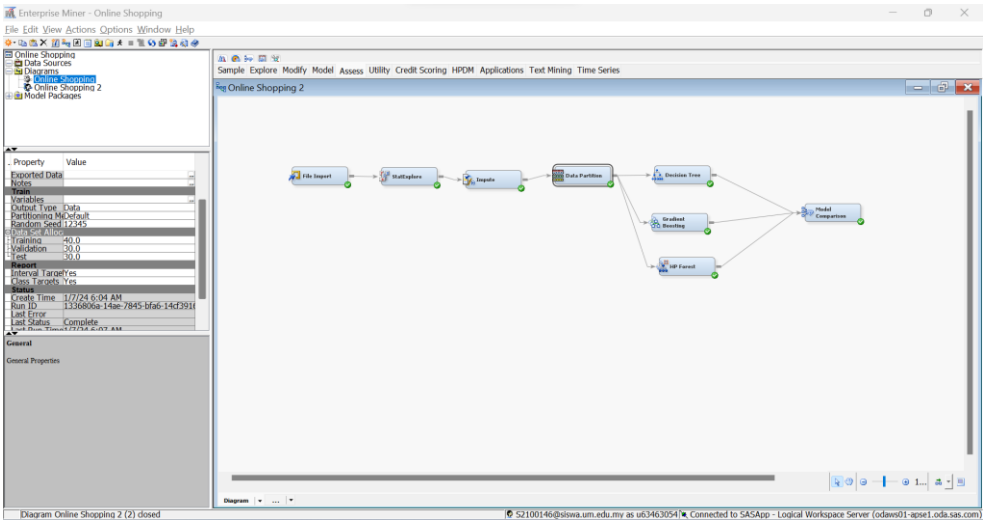
4) Data Partition

In SAS Enterprise Miner, the Data Partition node is used to divide a dataset into multiple subsets or partitions for the purpose of training and evaluating predictive models. Where the Training, Validation and Test is specified.

Training 40
Validation 30
Test 30

The partition summary are as follows:

Type	Data Set	No of Observations
DATA	EMWS1.Impt_TRAIN	3900
TRAIN	EMWS1.Part_TRAIN	1559
VALIDATE	EMWS1.Part_VALIDATE	1169
TEST	EMWS1.Part_TEST	1172



Summary Statistics for Interval Targets

Data=DATA

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Online_Spend	4055.3	1888.55	417.73	3900	0	1100.29	Online Spend

Data=TEST

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Online_Spend	4055.3	1895.17	417.73	1170	0	1137.53	Online Spend

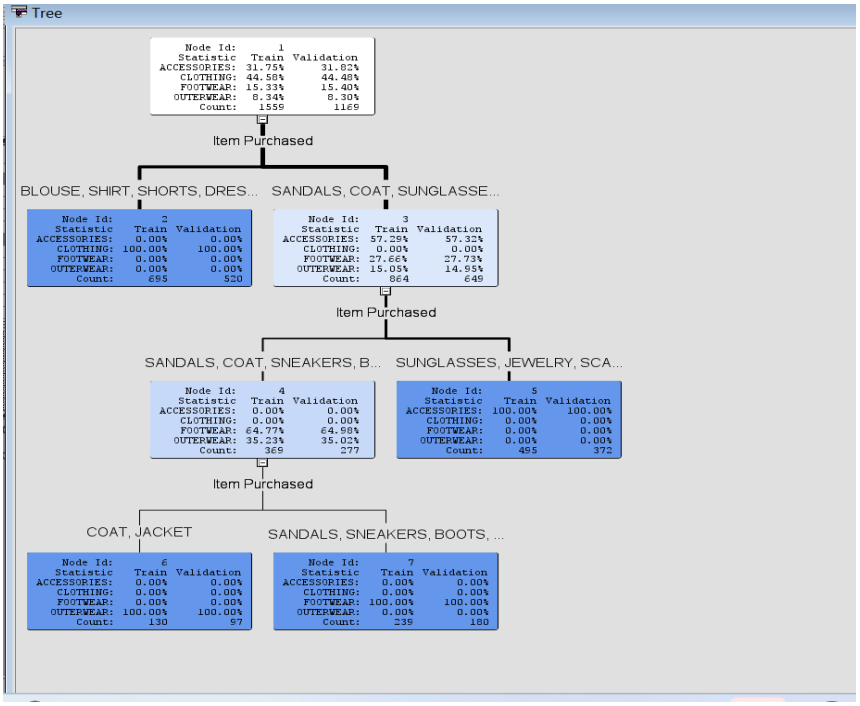
Data=TRAIN

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Online_Spend	4055.3	1889.09	417.73	1560	0	1081.14	Online Spend

Data=VALIDATE

Variable	Maximum	Mean	Minimum	Number of Observations	Missing	Standard Deviation	Label
Online_Spend	4055.3	1881.22	417.73	1170	0	1088.57	Online Spend

5) Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyses customer behavior.



Classification Table

Data Role=TRAIN Target Variable=Category Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
ACCESSORIES	ACCESSORIES	100	100	495	31.7511
CLOTHING	CLOTHING	100	100	695	44.5799
FOOTWEAR	FOOTWEAR	100	100	239	15.3303
OUTERWEAR	OUTERWEAR	100	100	130	8.3387

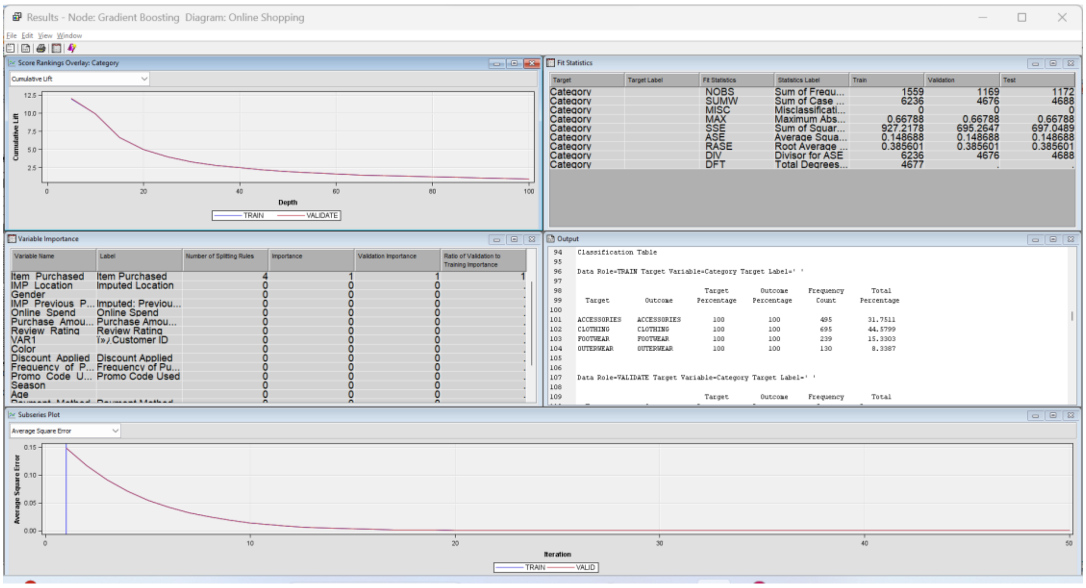
Data Role=VALIDATE Target Variable=Category Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
ACCESSORIES	ACCESSORIES	100	100	372	31.8221
CLOTHING	CLOTHING	100	100	520	44.4825
FOOTWEAR	FOOTWEAR	100	100	180	15.3978
OUTERWEAR	OUTERWEAR	100	100	97	8.2977

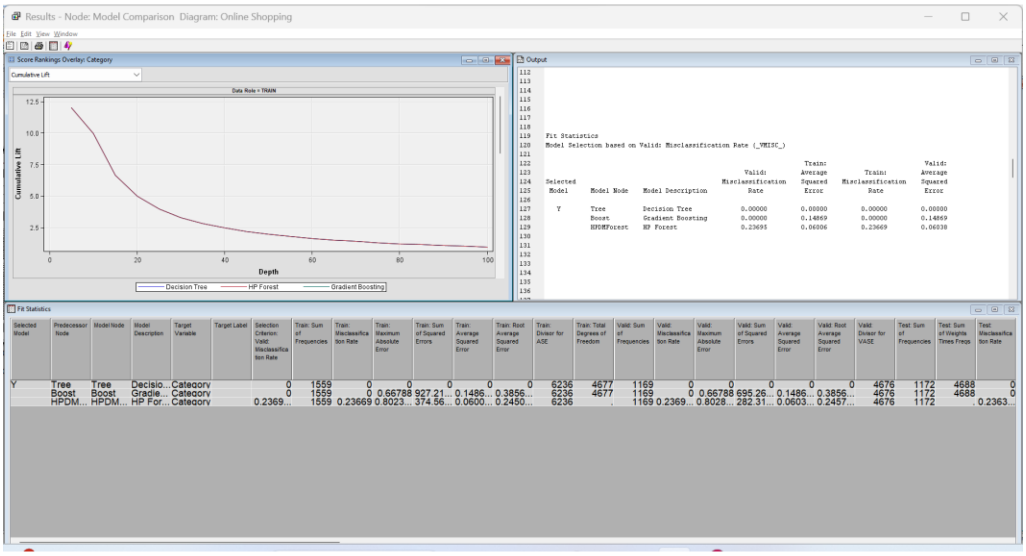
As can be seen in the classification table, Clothing is the most preferred category purchased item by the buyers which has the highest percentage 44.5%, followed by accessories 31.8%, footwear 15.4% and outwear 8.3%.

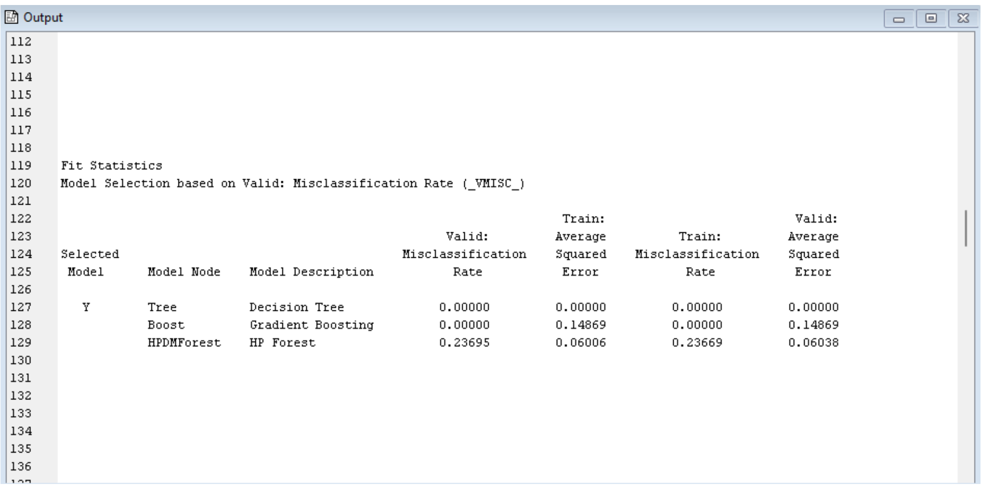
6) Ensemble Methods: Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.

Boosting are ensemble techniques used to improve the performance of machine learning models, and they can be applied to various algorithms, including the Random Forest algorithm.



7) Comparison Model between Decision Tree, Gradient Boosting and Forest.





Output

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

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

   Y       Tree         Decision Tree              0.00000        0.00000      0.00000      0.00000
           Boost        Gradient Boosting          0.00000        0.14869      0.00000      0.14869
           HPDMForest   HP Forest                  0.23695        0.06006      0.23669      0.06038
```

As Shown in the Fit Statistics table, it shows that among the Decision Tree, Gradient Boosting and Forest. The lowest misclassification rate is Decision Tree and Gradient Boosting followed by Forest.

Summary

In this study, the E commerce customer behavior is analyze, the dataset used created consist of 20 unique variable column and 3900 observations row. Talend Preparation data is used to explore, cleanse, and manipulate data to prepare it for analysis or integration. During the process found out there were missing values is some of the rows in “location” and “Item purchase” variable.

Next is the dataset is imported into SAS Miner Enterprise to handle missing values and specify variable roles. The variable category is classified correctly according to its role, The variable “Category” is specified as “Target” in this process. The missing value is then detected after StateExplore run in the proceed. The results found out the there were two variable that has missing value which are “Location” and “Previous Purchased”. Then the Data Partition node is used to divide the dataset to multiple subset for the purpose of training and evaluating predictive models, in this process the percentage of each set is allocated, Training 40, validation 30 and Test 30. The partition summary table shows that Train, Validation and Test number of observation 1559, 1169, 1172 respectively.

In the next process, the customer online shopping behavior is analyses in SAS Enterprise Miner. As can be seen in the classification table, Clothing is the most preferred category purchased item by the buyers which has the highest percentage 44.5%, followed by accessories 31.8%, footwear 15.4% and outwear 8.3%. Next is Gradient Boosting, which is to get more accurate result. After running Gradient Boosting found out that the result is similar to the result from Decision Tree which clothing, accessories, footwear and outwear. Finally, comparison is done between all three machine learning method, As shown in Fit Statistics table, it shows that among the Decision Tree, Gradient Boosting and Forest. The lowest misclassification rate is Decision Tree and Gradient Boosting followed by Forest.

In conclusion, it is very important for a business owner to understand the need of the customer and to focus on customer satisfaction and build strong relationships. Understand the customers' needs and continuously seek feedback to improve products or services.