

S2100146 Anis Sofea

Alternative Assessment

Case Study: E-Commerce Customer Behavior Analysis

<https://github.com/Anissoo56/WQD7005AA1>

Talend Data Preparation

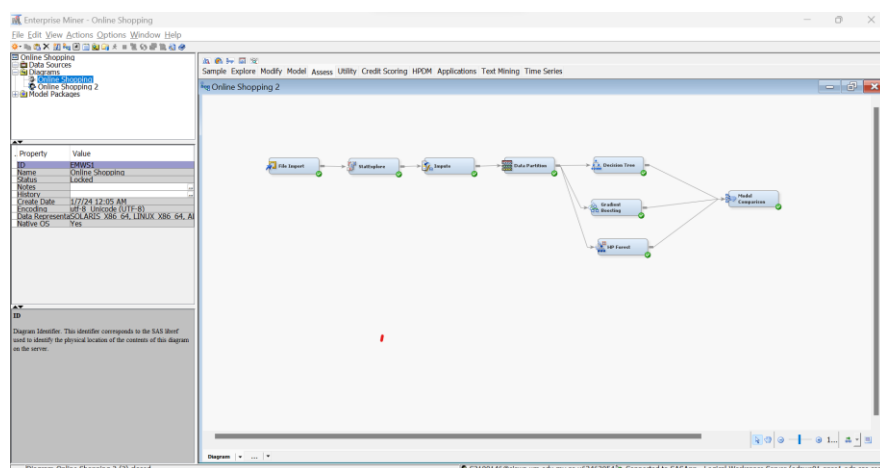
- 1) Talend Data Preparation is a tool that allows users to explore, cleanse, and manipulate data to prepare it for analysis or integration.

The screenshot shows the Talend Data Preparation interface. On the left, a table displays 16 rows of data with columns: Gender, Item Purchased, Category, Purchase Amount, Location, Size, and Color. The data includes various clothing items like Blouse, Sweater, Jeans, Sandals, Sneakers, Shirt, Shorts, Coat, Handbag, Shoes, and Dress, categorized as Clothing or Outerwear, from different US states. On the right, a 'Size' summary panel shows statistics for the 'Color' column: Count: 3900, Distinct: 4, Duplicates: 3896, Valid: 3900, Empty: 0, Innull: 0, Avg length: 1.11, Min length: 1, and Max length: 2.

Gender	Item Purchased	Category	Purchase Amount	Location	Size	Color
Male	Blouse	Clothing	53	Kentucky	L	Gray
Male	Sweater	Clothing	64	Maine	L	Maroon
Male	Jeans	Clothing	73	Massachusetts	S	Maroon
Male	Sandals	Footwear	98	Rhode Island	M	Maroon
Male	Blouse	Clothing	48	Oregon	M	Turquoise
Male	Sneakers	Footwear	28	Wyoming	M	White
Male	Shirt	Clothing	85	Montana	M	Gray
Male	Shorts	Clothing	34	Louisiana	L	Charcoal
Male	Coat	Outerwear	97	West Virginia	L	Silver
Male	Handbag	Accessories	31	Missouri	M	Pink
Male	Shoes	Footwear	34	Arkansas	L	Purple
Male	Shorts	Clothing	68	Hawaii	S	Olive
Male	Coat	Outerwear	72	Delaware	M	Gold
Male	Dress	Clothing	51	New Hampshire	M	Violet
Male	Coat	Outerwear	53	New York	L	Teal
Male	Skirt	Clothing	81	Rhode Island	M	Teal

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 correctly for each variable, The Category is specified as 'Target'

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Category	Target	Nominal	No		No	.	.
Color	Input	Nominal	No		No	.	.
Discount Amt	Input	Nominal	No		No	.	.
Frequency	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Item Purch	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
Online Spent	Input	Interval	No		No	.	.
Payment Mth	Input	Nominal	No		No	.	.
Preferred Pa	Input	Nominal	No		No	.	.
Previous Pur	Input	Interval	No		No	.	.
Promo Code	Input	Nominal	No		No	.	.
Purchase An	Input	Interval	No		No	.	.
Review Rat	Input	Interval	No		No	.	.
Season	Input	Nominal	No		No	.	.
Shipping Ty	Input	Nominal	No		No	.	.
Size	Input	Nominal	No		No	.	.
Subscription	Input	Nominal	No		No	.	.
VARI	Input	Interval	No		No	.	.

- 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

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for Train
Location	COUNT	IMP_Location	I_Location	Montana	INPUT	NOMINAL	Location	6
Previous Purchases	MEAN	IMP_Previous Purchases	I_Previous Purchases	25.353853601	INPUT	INTERVAL	Previous Purchases	9

- 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.

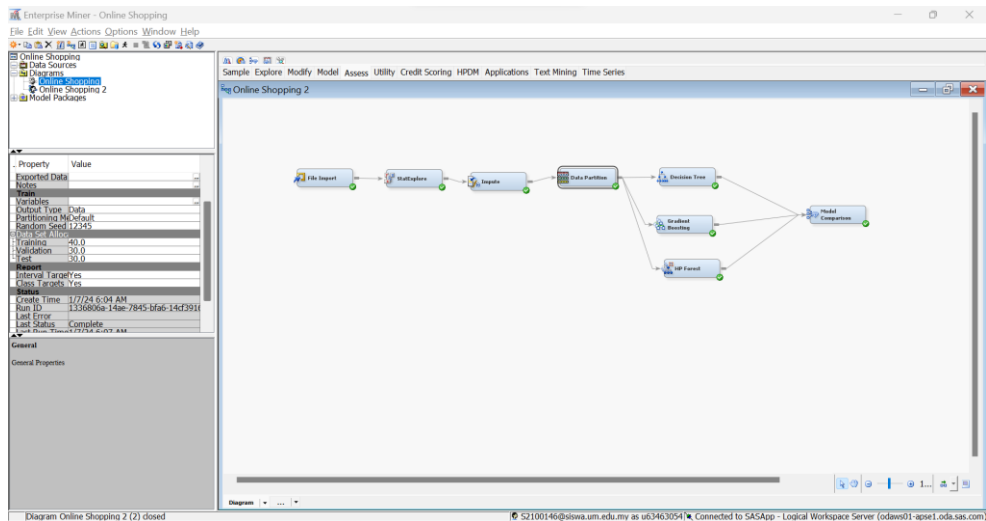
Training 40

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

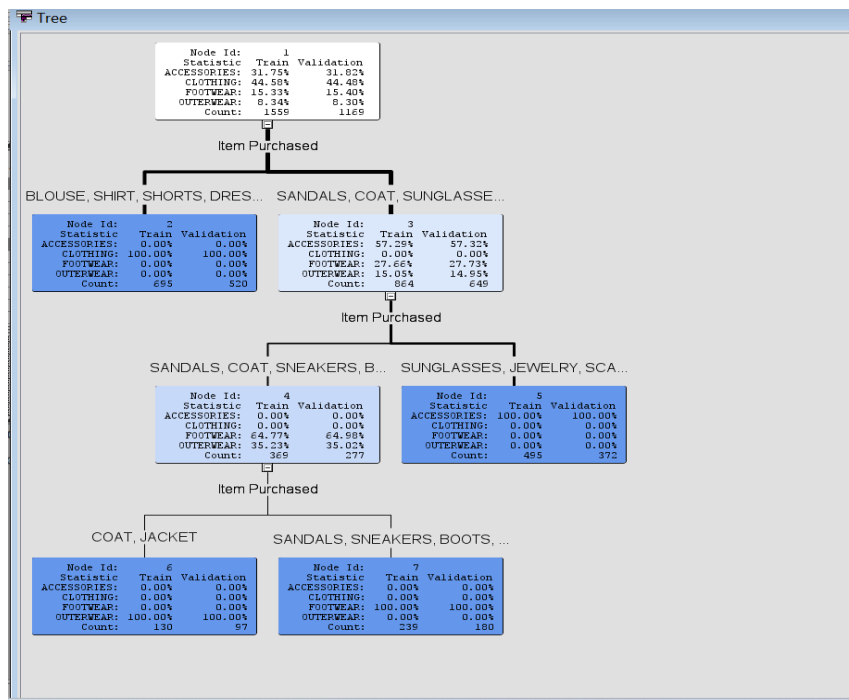
Data=TRAIN

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 analyse 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

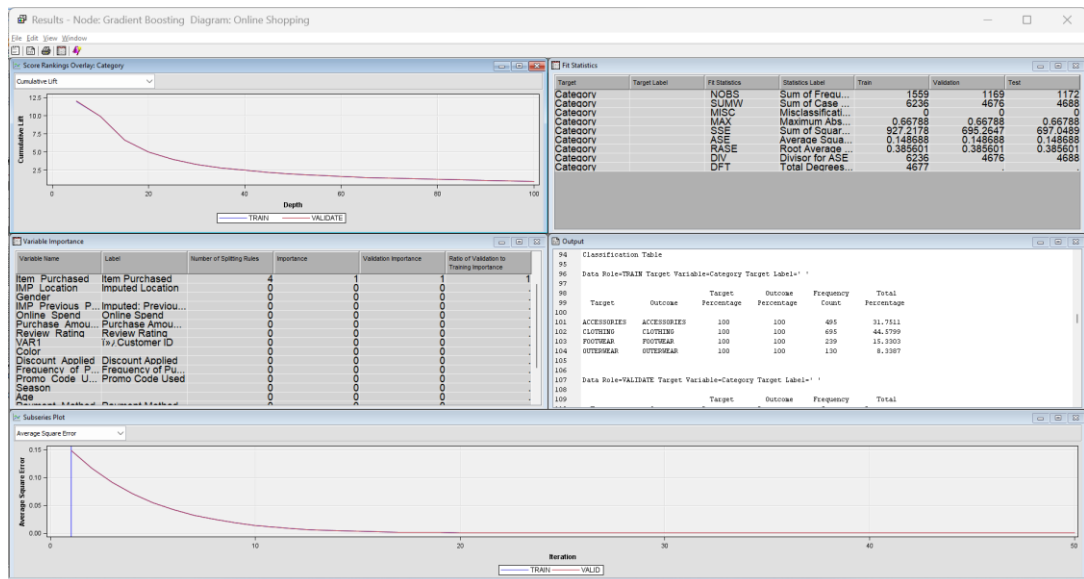
Data Role=VALIDATE Target Variable=Category Target Label=' '

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

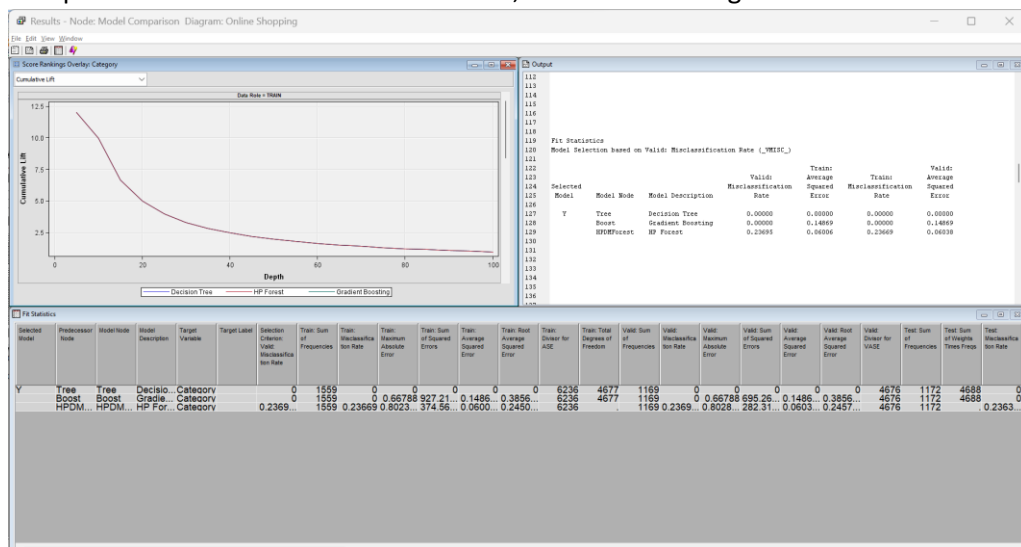
As can be seen in the classification table, Clothing is the most preferred category purchased item by the buyers which has the highest percentage 44.5%, followed by accessories 31.8%, footwear 15.4% and outwear 8.3%.

- 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

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

Fit Statistics

Model Selection based on Valid: Misclassification Rate (VMISC)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Tree	Decision Tree	0.00000	0.00000	0.00000	0.00000	0.00000
	Boost	Gradient Boosting	0.00000	0.14869	0.00000	0.14869	0.14869
	HPDMForest	HP Forest	0.23695	0.06006	0.23669	0.06038	0.2363

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.