WQD7005 DATA MINING
Alternative Assessment 1

E-commerce Case Study (file in github repository can be retrieved in this link https://github.com/slayerliverave/WQD7005-E-Commerce-Dataset)

The dataset that has been used for the assessment is obtained from Kaggle titled “Customer Behaviour and Shopping Habits Dataset” (https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset) which consists of 3,900 purchases with 18 attributes that similar to the dataset structure proposed for the assessment as follow.

Required Dataset Structure	Availability in proposed dataset	Remarks
CustomerID	**Yes**	
Age	**Yes**	
Gender	**Yes**	
Location	**Yes**	
MembershipLevel	**Yes**	*Note: initial data structure has no membership level however the dataset is being modified such a way that it meets the required data structure based on Subscription Status and Frequency of Purchase*
TotalPurchases	**Yes**	Purchase Amount (USD)
TotalSpent	**Yes**	Purchase Amount (USD)
FavouriteCategory	**Yes**	Purchase category e.g. Clothing, Footwear, Outerwear, Accessories,
LastPurchaseDate	No	However other attributes that can be considered in the dataset is: • **Previous Purchases**: Provides information on the number or frequency of prior purchases made by the customer, contributing to customer segmentation and retention strategies. • **Frequency of Purchases**
Additional attributes e.g. customer’s occupation, frequency of website visits, etc	**Yes**	Color, Season, Review Rating, Shipping Type, Discount Applied, Promo Code,
Churn	**Yes**	Subscription Status (Yes/No) - *Indicates whether the customer has opted for a subscription service, offering insights into their level of loyalty and potential for recurring revenue*

WQD7005 DATA MINING | ATERNATIVE ASSESSMENT 1

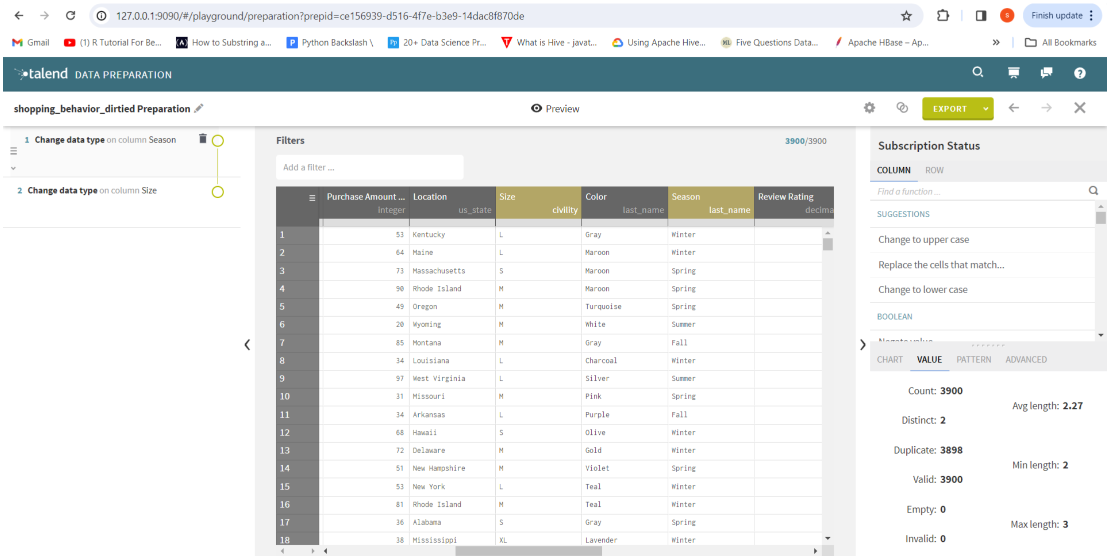
ID: 17147394 | Fathia Farhana bt Agusalim

1. Data Import and Preprocessing

1.1. Data Import in Talend Data Preparation

Prior importing the data in SAS E-Miner, the dataset is being imported to Talend Data Preparation in order to inspect the data and do necessary preprocessing to handle the missing values and any either anomalies or non-standardisation observed within the data. It is to be noted that, since the dataset in Kaggle is already cleaned data, hence the data is intentionally dirtied in order to showcase and demonstrate how data cleaning and preprocessing are executed for data mining.

Based on the dataset, there is some non-stanrdardisation of the datatype for attributes Season and Size where both datatype is changed to "text" as shown in Figure 1.1 and 1.2 below



shopping_behavior_dirtied Preparation

1 **Change data type** on column Season

2 **Change data type** on column Size

Filters

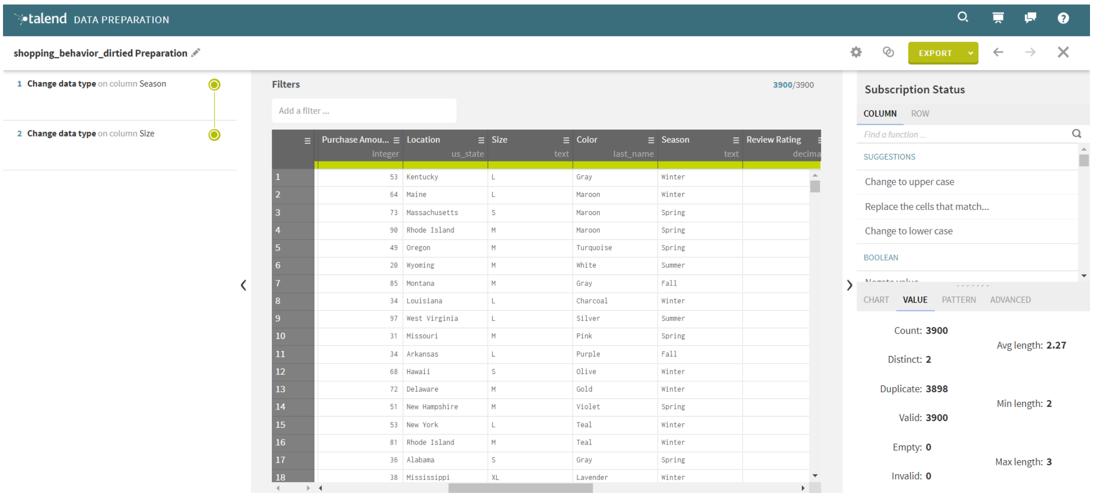
Add a filter ...

3900/3900

	Purchase Amount ... integer	Location us_state	Size civility	Color last_name	Season last_name	Review Rating decima
1	53	Kentucky	L	Gray	Winter	
2	64	Maine	L	Maroon	Winter	
3	73	Massachusetts	S	Maroon	Spring	
4	90	Rhode Island	M	Maroon	Spring	
5	49	Oregon	M	Turquoise	Spring	
6	20	Wyoming	M	White	Summer	
7	85	Montana	M	Gray	Fall	
8	34	Louisiana	L	Charcoal	Winter	
9	97	West Virginia	L	Silver	Summer	
10	31	Missouri	M	Pink	Spring	
11	34	Arkansas	L	Purple	Fall	
12	68	Hawaii	S	Olive	Winter	
13	72	Delaware	M	Gold	Winter	
14	51	New Hampshire	M	Violet	Spring	
15	53	New York	L	Teal	Winter	
16	81	Rhode Island	M	Teal	Winter	
17	36	Alabama	S	Gray	Spring	
18	38	Mississippi	XL	Lavender	Winter	

Count: 3900
Distinct: 2
Duplicate: 3898
Valid: 3900
Empty: 0
Invalid: 0
Avg length: 2.27
Min length: 2
Max length: 3

Figure 1. 1 Original datatype for Season and Size



shopping_behavior_dirtied Preparation

1 **Change data type** on column Season

2 **Change data type** on column Size

Filters

Add a filter ...

3900/3900

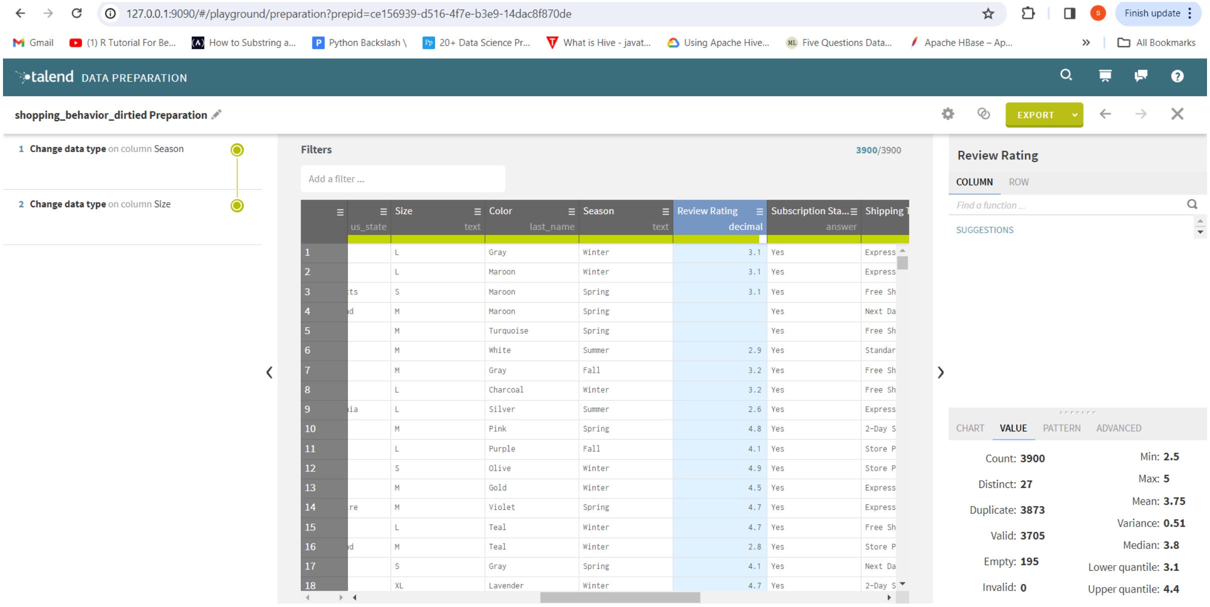
	Purchase Amou... integer	Location us_state	Size text	Color last_name	Season text	Review Rating decima
1	53	Kentucky	L	Gray	Winter	
2	64	Maine	L	Maroon	Winter	
3	73	Massachusetts	S	Maroon	Spring	
4	90	Rhode Island	M	Maroon	Spring	
5	49	Oregon	M	Turquoise	Spring	
6	20	Wyoming	M	White	Summer	
7	85	Montana	M	Gray	Fall	
8	34	Louisiana	L	Charcoal	Winter	
9	97	West Virginia	L	Silver	Summer	
10	31	Missouri	M	Pink	Spring	
11	34	Arkansas	L	Purple	Fall	
12	68	Hawaii	S	Olive	Winter	
13	72	Delaware	M	Gold	Winter	
14	51	New Hampshire	M	Violet	Spring	
15	53	New York	L	Teal	Winter	
16	81	Rhode Island	M	Teal	Winter	
17	36	Alabama	S	Gray	Spring	
18	38	Mississippi	XL	Lavender	Winter	

Count: 3900
Distinct: 2
Duplicate: 3898
Valid: 3900
Empty: 0
Invalid: 0
Avg length: 2.27
Min length: 2
Max length: 3

Figure 1. 2: Change datatype for Season and Size to "text" format

1.2. Missing Values and Data standardisation

Based on the data inspection, it can be observed that few attributed have missing values for “Review Rating” where 195 records are missing and similarly for “Payment Method” where 390 records are missing from the dataset as shown in Figure 1.3 and Figure 1.4 respectively. The remaining attributes are acceptable where no missing values have been indicated. For handling missing values, the data will then be imputed using SAS E-Miner.



shopping_behavior_dirtied Preparation

1 Change data type on column Season

2 Change data type on column Size

Filters

Add a filter ...

3900/3900

		Size	Color	Season	Review Rating	Subscription Sta...	Shipping T
	us_state	text	last_name	text	decimal	answer	
1		L	Gray	Winter	3.1	Yes	Express
2		L	Maroon	Winter	3.1	Yes	Express
3	ts	S	Maroon	Spring	3.1	Yes	Free Sh
4	d	M	Maroon	Spring		Yes	Next Da
5		M	Turquoise	Spring		Yes	Free Sh
6		M	White	Summer	2.9	Yes	Standar
7		M	Gray	Fall	3.2	Yes	Free Sh
8		L	Charcoal	Winter	3.2	Yes	Free Sh
9	ia	L	Silver	Summer	2.6	Yes	Express
10		M	Pink	Spring	4.8	Yes	2-Day S
11		L	Purple	Fall	4.1	Yes	Store P
12		S	Olive	Winter	4.9	Yes	Store P
13		M	Gold	Winter	4.5	Yes	Express
14	re	M	Violet	Spring	4.7	Yes	Express
15		L	Teal	Winter	4.7	Yes	Free Sh
16	d	M	Teal	Winter	2.8	Yes	Store P
17		S	Gray	Spring	4.1	Yes	Next Da
18		XL	Lavender	Winter	4.7	Yes	2-Day S

Review Rating

COLUMN ROW

Find a function ...

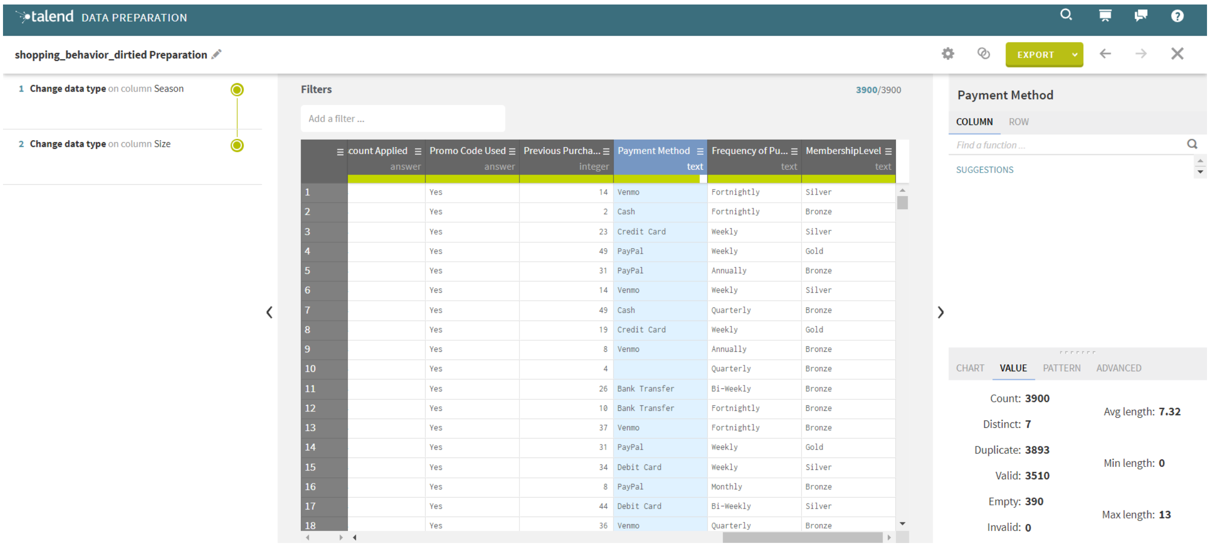
SUGGESTIONS

CHART VALUE PATTERN ADVANCED

Count: 3900
Distinct: 27
Duplicate: 3873
Valid: 3705
Empty: 195
Invalid: 0

Min: 2.5
Max: 5
Mean: 3.75
Variance: 0.51
Median: 3.8
Lower quantile: 3.1
Upper quantile: 4.4

Figure 1. 3: Total of 195 empty values in Review Rating attributes



shopping_behavior_dirtied Preparation

1 Change data type on column Season

2 Change data type on column Size

Filters

Add a filter ...

3900/3900

	count Applied	Promo Code Used	Previous Purcha...	Payment Method	Frequency of Pu...	MembershipLevel
	answer	answer	integer	text	text	text
1		Yes	14	Venmo	Fortnightly	Silver
2		Yes	2	Cash	Fortnightly	Bronze
3		Yes	23	Credit Card	Weekly	Silver
4		Yes	49	PayPal	Weekly	Gold
5		Yes	31	PayPal	Annually	Bronze
6		Yes	14	Venmo	Weekly	Silver
7		Yes	49	Cash	Quarterly	Bronze
8		Yes	19	Credit Card	Weekly	Gold
9		Yes	8	Venmo	Annually	Bronze
10		Yes	4		Quarterly	Bronze
11		Yes	26	Bank Transfer	Bi-Weekly	Bronze
12		Yes	10	Bank Transfer	Fortnightly	Bronze
13		Yes	37	Venmo	Fortnightly	Bronze
14		Yes	31	PayPal	Weekly	Gold
15		Yes	34	Debit Card	Weekly	Silver
16		Yes	8	PayPal	Monthly	Bronze
17		Yes	44	Debit Card	Bi-Weekly	Silver
18		Yes	36	Venmo	Quarterly	Bronze

Payment Method

COLUMN ROW

Find a function ...

SUGGESTIONS

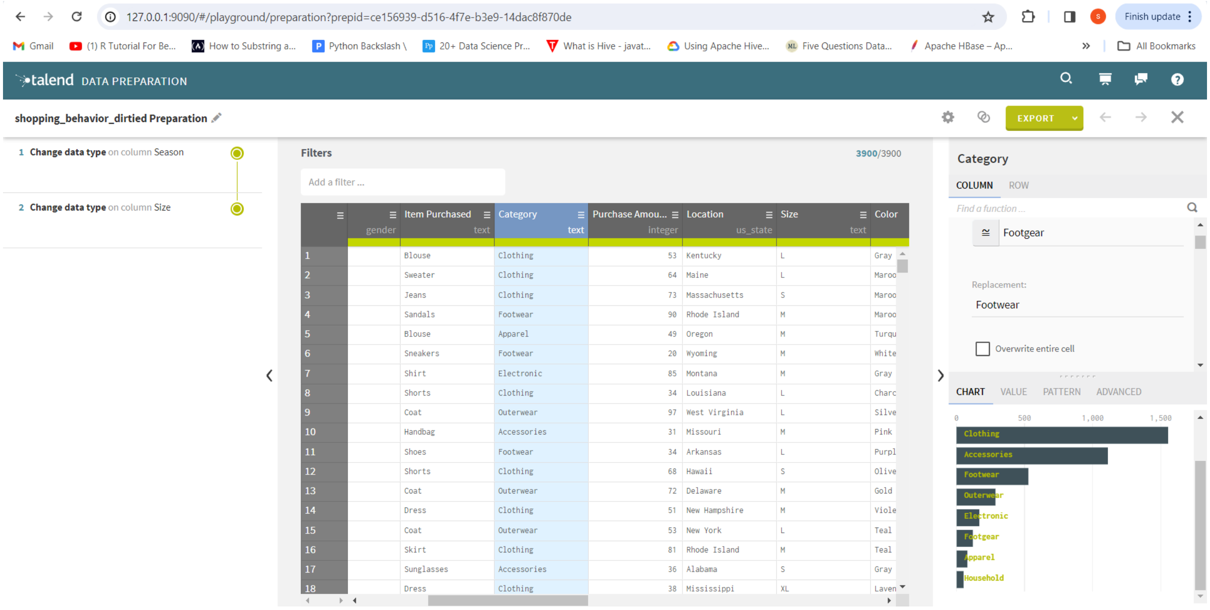
CHART VALUE PATTERN ADVANCED

Count: 3900
Distinct: 7
Duplicate: 3893
Valid: 3510
Empty: 390
Invalid: 0

Avg length: 7.32
Min length: 0
Max length: 13

Figure 1. 4: Total of 390 Empty values in Payment Method

In order to ensure that Purchase Category is standardized, and more generalisation and customisation can be made to understand the customer behaviour, the Category will be only kept to four (4) main categories such as Clothing, Footwear, Outerwear and Accessories. Hence, the Category type shown in Figure 1.5 for Apparel, Footgear, Electronic and Household will be changed according to the relevant category mentioned above.



shopping_behavior_dirtied Preparation

1 **Change data type** on column Season

2 **Change data type** on column Size

3900/3900

	gender	Item Purchased text	Category text	Purchase Amou... integer	Location us_state	Size text	Color
1		Blouse	Clothing	53	Kentucky	L	Gray
2		Sweater	Clothing	64	Maine	L	Maroo
3		Jeans	Clothing	73	Massachusetts	S	Maroo
4		Sandals	Footwear	90	Rhode Island	M	Maroo
5		Blouse	Apparel	49	Oregon	M	Turqu
6		Sneakers	Footwear	20	Wyoming	M	White
7		Shirt	Electronic	85	Montana	M	Gray
8		Shorts	Clothing	34	Louisiana	L	Charc
9		Coat	Outerwear	97	West Virginia	L	Silve
10		Handbag	Accessories	31	Missouri	M	Pink
11		Shoes	Footwear	34	Arkansas	L	Purpl
12		Shorts	Clothing	68	Hawaii	S	Olive
13		Coat	Outerwear	72	Delaware	M	Gold
14		Dress	Clothing	51	New Hampshire	M	Viole
15		Coat	Outerwear	53	New York	L	Teal
16		Skirt	Clothing	81	Rhode Island	M	Teal
17		Sunglasses	Accessories	36	Alabama	S	Gray
18		Dress	Clothing	38	Mississippi	XL	Laven

Category

COLUMN ROW

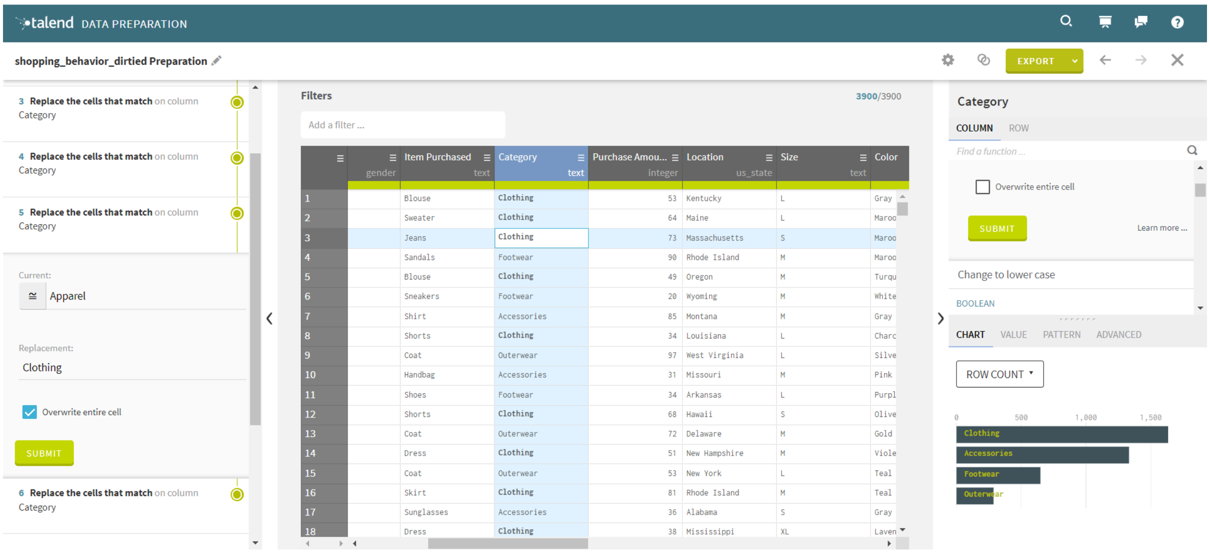
≅ Footgear

Replacement:

Footwear

□ Overwrite entire cell

Figure 1. 5: Purchase Category before standardisation



shopping_behavior_dirtied Preparation

3 **Replace the cells that match** on column Category

4 **Replace the cells that match** on column Category

5 **Replace the cells that match** on column Category

Current:

≅ Apparel

Replacement:

Clothing

☑ Overwrite entire cell

SUBMIT

6 **Replace the cells that match** on column Category

3900/3900

	gender	Item Purchased text	Category text	Purchase Amou... integer	Location us_state	Size text	Color
1		Blouse	Clothing	53	Kentucky	L	Gray
2		Sweater	Clothing	64	Maine	L	Maroo
3		Jeans	Clothing	73	Massachusetts	S	Maroo
4		Sandals	Footwear	90	Rhode Island	M	Maroo
5		Blouse	Clothing	49	Oregon	M	Turqu
6		Sneakers	Footwear	20	Wyoming	M	White
7		Shirt	Accessories	85	Montana	M	Gray
8		Shorts	Clothing	34	Louisiana	L	Charc
9		Coat	Outerwear	97	West Virginia	L	Silve
10		Handbag	Accessories	31	Missouri	M	Pink
11		Shoes	Footwear	34	Arkansas	L	Purpl
12		Shorts	Clothing	68	Hawaii	S	Olive
13		Coat	Outerwear	72	Delaware	M	Gold
14		Dress	Clothing	51	New Hampshire	M	Viole
15		Coat	Outerwear	53	New York	L	Teal
16		Skirt	Clothing	81	Rhode Island	M	Teal
17		Sunglasses	Accessories	36	Alabama	S	Gray
18		Dress	Clothing	38	Mississippi	XL	Laven

Category

COLUMN ROW

□ Overwrite entire cell

SUBMIT

Change to lower case

BOOLEAN

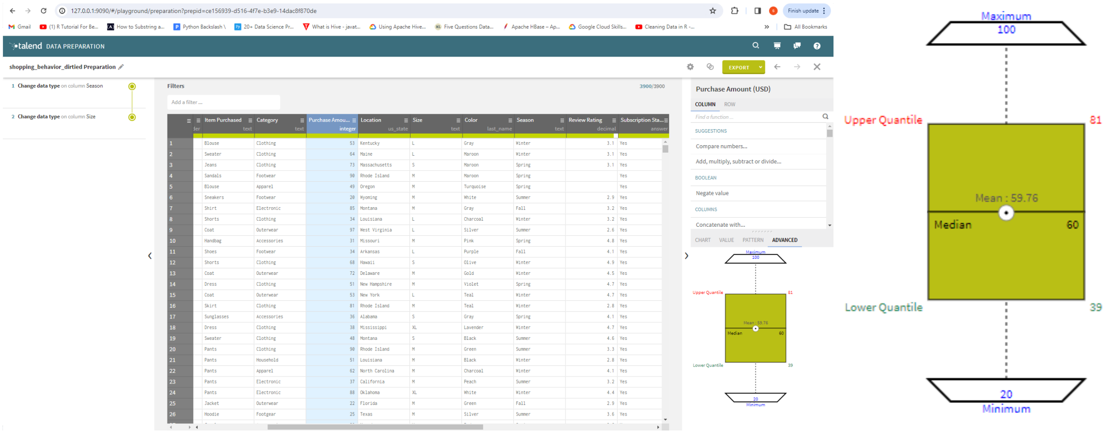
Figure 1. 6: Purchase Category after standardisation

1.3. Data Understanding

In Talend Data Preparation, we are also able to understand the data in depth which will assist in getting more insights on customer behaviour. For example, based on Purchase Amount shown in Figure 1.7, it can be summarised that the maximum purchase of the transaction is USD 100 and minimum the purchase is USD 20 whilst average purchase amount for overall records is USD 59.76. With respect to the satisfaction level, maximum Review Rating given by customer is 4.0 over 5.0 with 177 occurrences in which overall satisfaction for the purchase by customer is 3.75 over 5.0.

WQD7005 DATA MINING | ATERNATIVE ASSESSMENT 1

ID: 17147394 | Fathia Farhana bt Agusalim



127.0.0.1:9090/#/playground/preparation?prepid=ce156939-d516-4f7e-b3e9-14dac8f870de

talend DATA PREPARATION

shopping_behavior_dirtied Preparation

1 Change data type on column Season

2 Change data type on column Size

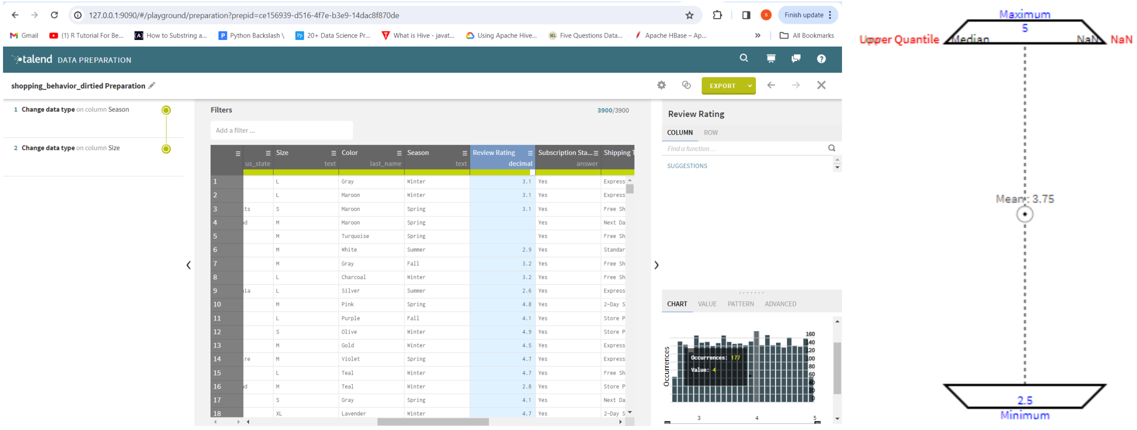
Filters

Add a filter ...

3900/3900

		Item Purchased	Category	Purchase Amou...	Location	Size	Color	Season	Review Rating	Subscription Sta...
	der	text	text	integer	us_state	text	last_name	text	decimal	answer
1		Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes
2		Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes
3		Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes
4		Sandals	Footwear	90	Rhode Island	M	Maroon	Spring		Yes
5		Blouse	Apparel	49	Oregon	M	Turquoise	Spring		Yes
6		Sneakers	Footwear	20	Wyoming	M	White	Summer	2.9	Yes
7		Shirt	Electronic	85	Montana	M	Gray	Fall	3.2	Yes
8		Shorts	Clothing	34	Louisiana	L	Charcoal	Winter	3.2	Yes
9		Coat	Outerwear	97	West Virginia	L	Silver	Summer	2.6	Yes
10		Handbag	Accessories	31	Missouri	M	Pink	Spring	4.8	Yes
11		Shoes	Footwear	34	Arkansas	L	Purple	Fall	4.1	Yes
12		Shorts	Clothing	68	Hawaii	S	Olive	Winter	4.9	Yes
13		Coat	Outerwear	72	Delaware	M	Gold	Winter	4.5	Yes
14		Dress	Clothing	51	New Hampshire	M	Violet	Spring	4.7	Yes
15		Coat	Outerwear	53	New York	L	Teal	Winter	4.7	Yes
16		Skirt	Clothing	81	Rhode Island	M	Teal	Winter	2.8	Yes
17		Sunglasses	Accessories	36	Alabama	S	Gray	Spring	4.1	Yes
18		Dress	Clothing	38	Mississippi	XL	Lavender	Winter	4.7	Yes
19		Sweater	Clothing	48	Montana	S	Black	Summer	4.6	Yes
20		Pants	Clothing	90	Rhode Island	M	Green	Summer	3.3	Yes
21		Pants	Household	51	Louisiana	M	Black	Winter	2.8	Yes
22		Pants	Apparel	62	North Carolina	M	Charcoal	Winter	4.1	Yes
23		Pants	Electronic	37	California	M	Peach	Summer	3.2	Yes
24		Pants	Electronic	88	Oklahoma	XL	White	Winter	4.4	Yes
25		Jacket	Outerwear	22	Florida	M	Green	Fall	2.9	Yes
26		Hoodie	Footgear	25	Texas	M	Silver	Summer	3.6	Yes

Figure 1. 7: Purchase Amount summary in e-commerce



127.0.0.1:9090/#/playground/preparation?prepid=ce156939-d516-4f7e-b3e9-14dac8f870de

talend DATA PREPARATION

shopping_behavior_dirtied Preparation

1 Change data type on column Season

2 Change data type on column Size

Filters

Add a filter ...

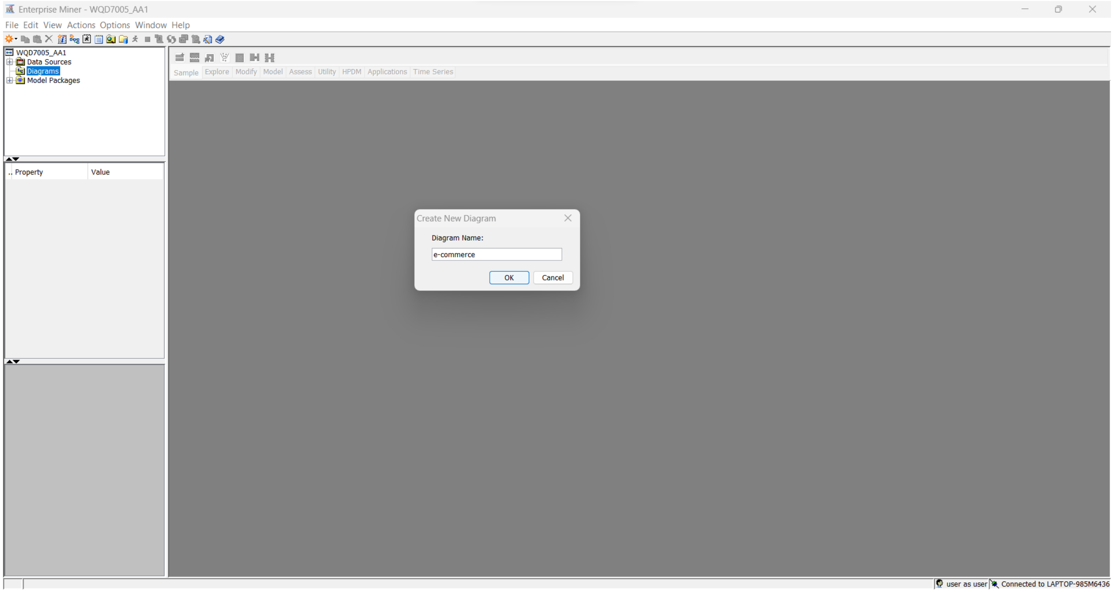
3900/3900

		Size	Color	Season	Review Rating	Subscription Sta...	Shipping T
	us_state	text	last_name	text	decimal	answer	
1		L	Gray	Winter	3.1	Yes	Express
2		L	Maroon	Winter	3.1	Yes	Express
3	ts	S	Maroon	Spring	3.1	Yes	Free Sh
4	d	M	Maroon	Spring		Yes	Next Da
5		M	Turquoise	Spring		Yes	Free Sh
6		M	White	Summer	2.9	Yes	Standar
7		M	Gray	Fall	3.2	Yes	Free Sh
8		L	Charcoal	Winter	3.2	Yes	Free Sh
9	ia	L	Silver	Summer	2.6	Yes	Express
10		M	Pink	Spring	4.8	Yes	2-Day S
11		L	Purple	Fall	4.1	Yes	Store P
12		S	Olive	Winter	4.9	Yes	Store P
13		M	Gold	Winter	4.5	Yes	Express
14	re	M	Violet	Spring	4.7	Yes	Express
15		L	Teal	Winter	4.7	Yes	Free Sh
16	d	M	Teal	Winter	2.8	Yes	Store P
17		S	Gray	Spring	4.1	Yes	Next Da
18		XL	Lavender	Winter	4.7	Yes	2-Day S

Figure 1. 8: Review Rating summary in e-commerce

1.4. Data Import and Preprocessing with SAS E-Minter

To import the dataset in SAS, an 'e-commerce' diagram will be created to New Diagram

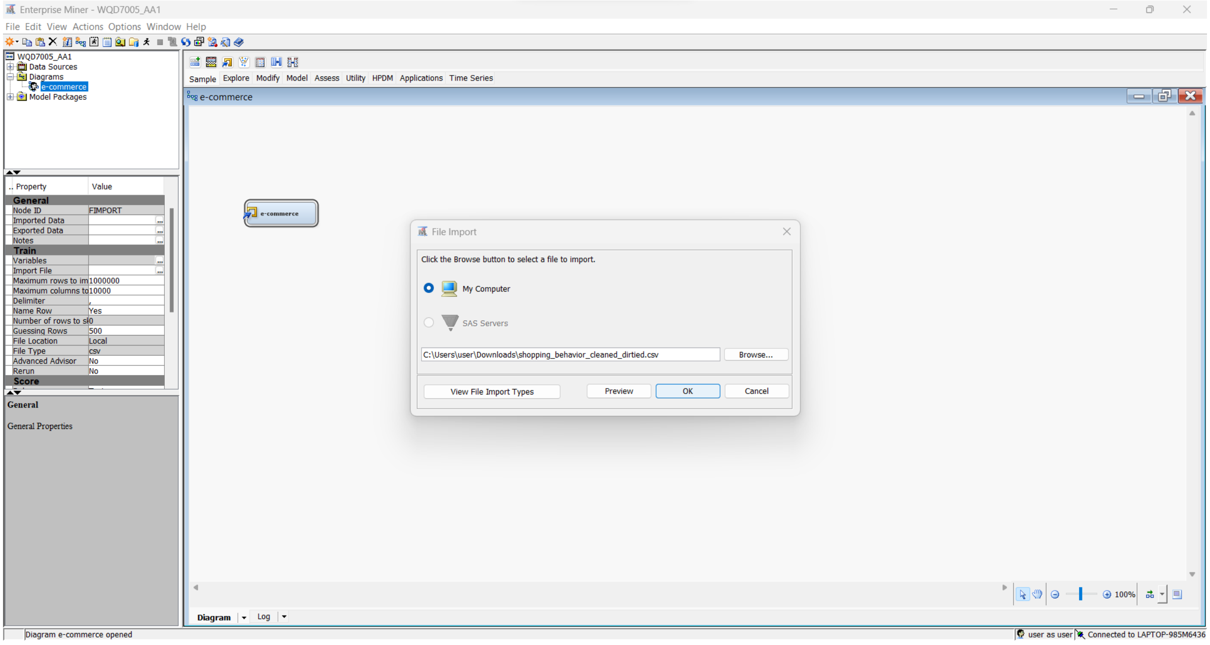

Enterprise Miner - WQD7005_AA1
File Edit View Actions Options Window Help
WQD7005_AA1
Data Sources
Diagrams
Model Packages
Property
Value
Sample Explore Modify Model Assess Utility HPDM Applications Time Series
OK
Cancel
user as user
Connected to LAPTOP-985M6436

Create New Diagram

Diagram Name: e-commerce

Figure 1. 9: New Diagram "e-commerce" is established

From the "Sample" tab, select "File Import" and drag the node into the workspace. Right click the File Import and rename the node as "e-commerce"


Enterprise Miner - WQD7005_AA1
File Edit View Actions Options Window Help
WQD7005_AA1
Data Sources
Diagrams
e-commerce
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Time Series
e-commerce
e-commerce
General
General Properties
Browse...
View File Import Types
Preview
OK
Cancel
100%
Diagram
Log
Diagram e-commerce opened
user as user
Connected to LAPTOP-985M6436

.. Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	...
Maximum rows to im	1000000
Maximum columns to	10000
Delimiter	,
Name Row	Yes
Number of rows to sk	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Rerun	No
Score	

File Import

Click the Browse button to select a file to import.

● My Computer

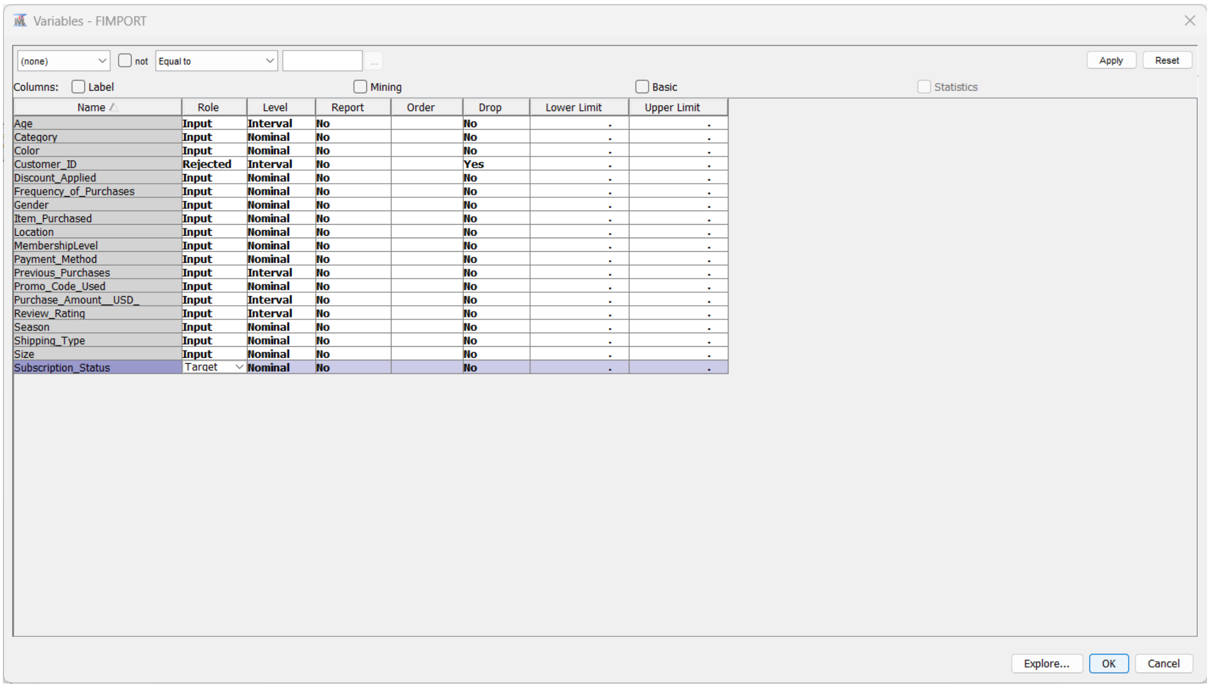
○ SAS Servers

C:\Users\user\Downloads\shopping_behavior_cleaned_dirtied.csv

Figure 1. 10: Importing e-commerce dataset into SAS E-Miner

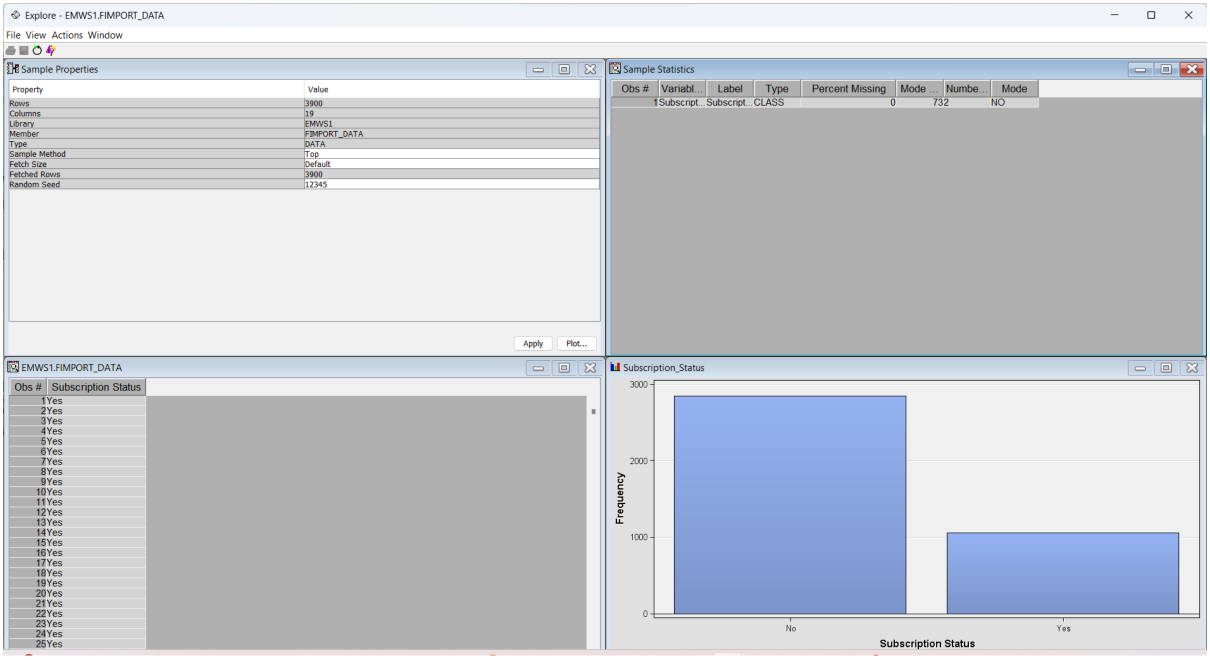
It essential to edit the variable for the dataset to understand on customer behaviour. Hence the role of the dataset is amended as follows:

- For Customer_ID, the Role is changed to "Rejected" and select "Yes" in Drop column as this attribute is insignificant for the analysis
- For the Subscription_Status, the Role is changed to "Target" to indicate whether customer will continue the subscription or not. Based on Figure 1.12, we can summarised that majority of the customer opt to "No" subscription with 2897 no. of records (73%)



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Category	Input	Nominal	No		No	.	.
Color	Input	Nominal	No		No	.	.
Customer_ID	Rejected	Interval	No		Yes	.	.
Discount_Applied	Input	Nominal	No		No	.	.
Frequency_of_Purchases	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Item_Purchased	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipLevel	Input	Nominal	No		No	.	.
Payment_Method	Input	Nominal	No		No	.	.
Previous_Purchases	Input	Interval	No		No	.	.
Promo_Code_Used	Input	Nominal	No		No	.	.
Purchase_Amount__USD_	Input	Interval	No		No	.	.
Review_Rating	Input	Interval	No		No	.	.
Season	Input	Nominal	No		No	.	.
Shipping_Type	Input	Nominal	No		No	.	.
Size	Input	Nominal	No		No	.	.
Subscription_Status	Target	Nominal	No		No	.	.

Figure 1. 11: Specifying Variable Roles in the dataset



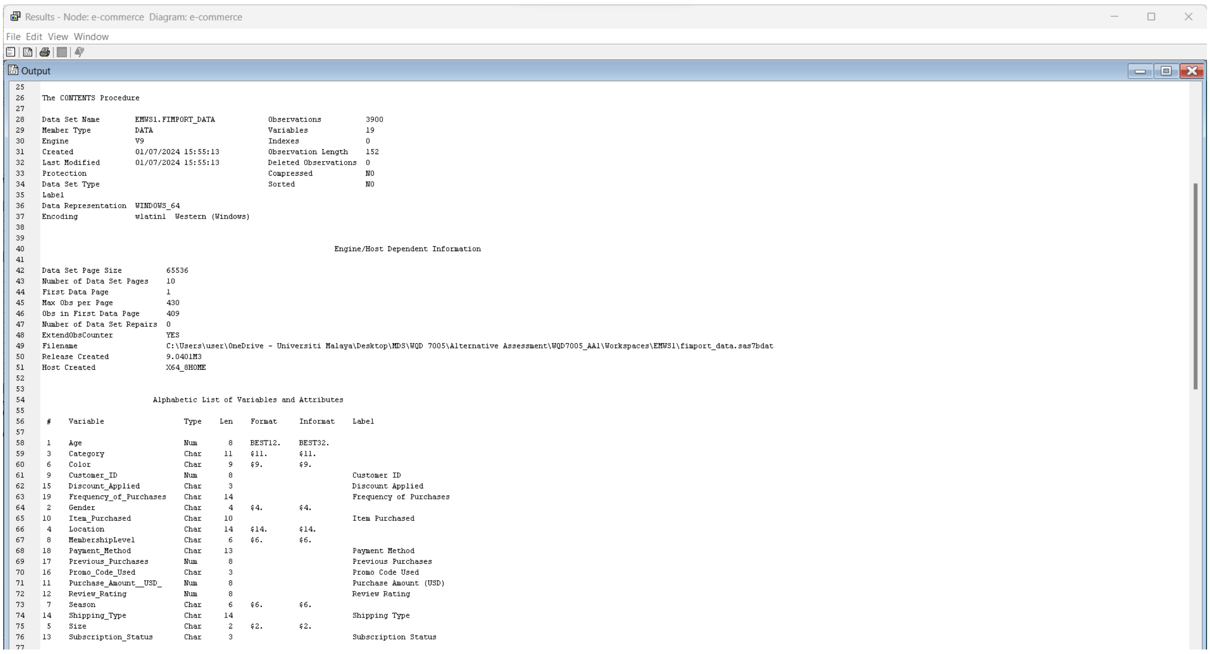
Property	Value
Rows	3900
Columns	19
Library	EMWS1
Member	FIMPORT_DATA
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	3900
Random Seed	12345

Obs #	Variabl...	Label	Type	Percent Missing	Mode ...	Numbe...	Mode
1	Subscript...	Subscript...	CLASS	0		732	NO

Obs #	Subscription Status
1	Yes
2	Yes
3	Yes
4	Yes
5	Yes
6	Yes
7	Yes
8	Yes
9	Yes
10	Yes
11	Yes
12	Yes
13	Yes
14	Yes
15	Yes
16	Yes
17	Yes
18	Yes
19	Yes
20	Yes
21	Yes
22	Yes
23	Yes
24	Yes
25	Yes

Figure 1. 12: Explore Subscription Status data distribution

Then, Run the “e-commerce” dataset and the summary of Results is demonstrated in Figure 1.13 below:

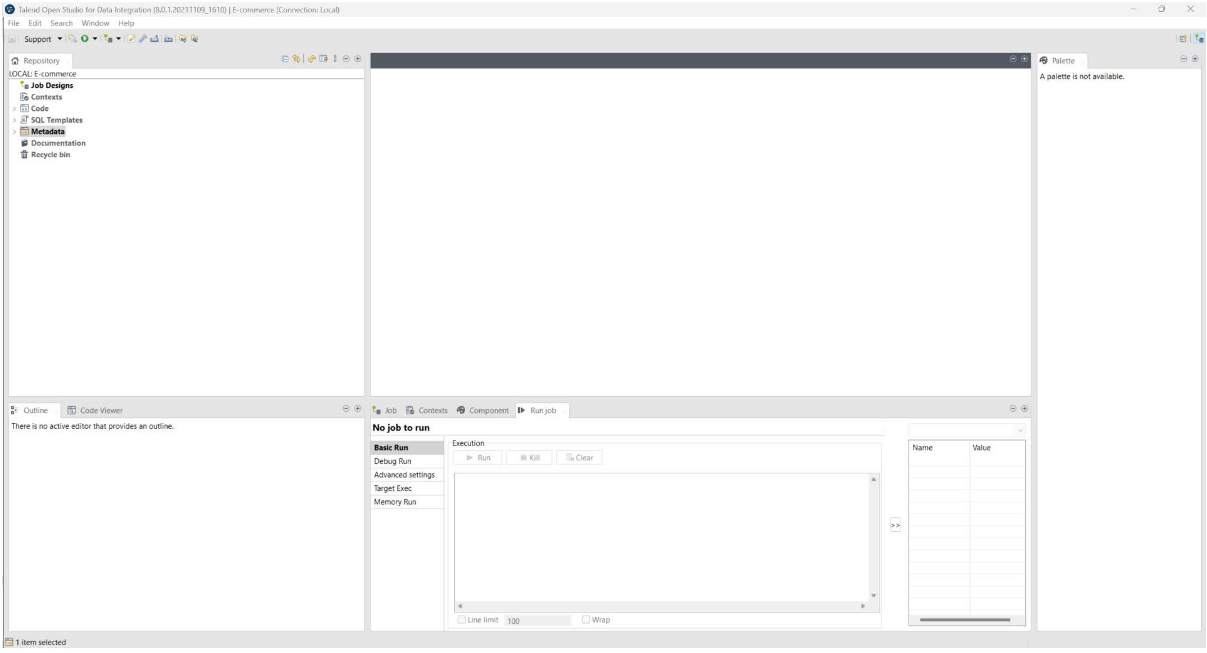


```

The CONTENTS Procedure

Data Set Name        EMWS1.FIMPORT_DATA           Observations         3900
Member Type          DATA                         Variables            19
Engine               V9                           Indexes              0
Created              01/07/2024 15:55:13          Observation Length   152
Last Modified        01/07/2024 15:55:13          Deleted Observations 0
Protection                                        Compressed           NO
Data Set Type                                     Sorted               NO
Label
Data Representation  WINDOWS_64
Encoding             wlatin1  Western (Windows)


                                             Engine/Host Dependent Information

Data Set Page Size          65536
Number of Data Set Pages    10
First Data Page             1
Max Obs per Page            430
Obs in First Data Page      409
Number of Data Set Repairs  0
ExtendObsCounter            YES
Filename                    C:\Users\user\OneDrive - Universiti Malaya\Desktop\MDS\WQD 7005\Alternative Assessment\WQD7005_AA1\Workspaces\EMWS1\fimport_data.sas7bdat
Release Created             9.0401M3
Host Created                X64_8HOME


                         Alphabetic List of Variables and Attributes

 #    Variable                 Type    Len    Format     Informat    Label

 1    Age                      Num       8    BEST12.    BEST32.
 3    Category                 Char     11    $11.       $11.
 6    Color                    Char      9    $9.        $9.
 9    Customer_ID              Num       8                           Customer ID
15    Discount_Applied         Char      3                           Discount Applied
19    Frequency_of_Purchases   Char     14                           Frequency of Purchases
 2    Gender                   Char      4    $4.        $4.
10    Item_Purchased           Char     10                           Item Purchased
 4    Location                 Char     14    $14.       $14.
 8    MembershipLevel          Char      6    $6.        $6.
18    Payment_Method           Char     13                           Payment Method
17    Previous_Purchases       Num       8                           Previous Purchases
16    Promo_Code_Used          Char      3                           Promo Code Used
11    Purchase_Amount__USD_    Num       8                           Purchase Amount (USD)
12    Review_Rating            Num       8                           Review Rating
 7    Season                   Char      6    $6.        $6.
14    Shipping_Type            Char     14                           Shipping Type
 5    Size                     Char      2    $2.        $2.
13    Subscription_Status      Char      3                           Subscription Status
```

Figure 1. 13: Summary of e-commerce dataset

1.5. Data Import and Integration in Talend Data Integration

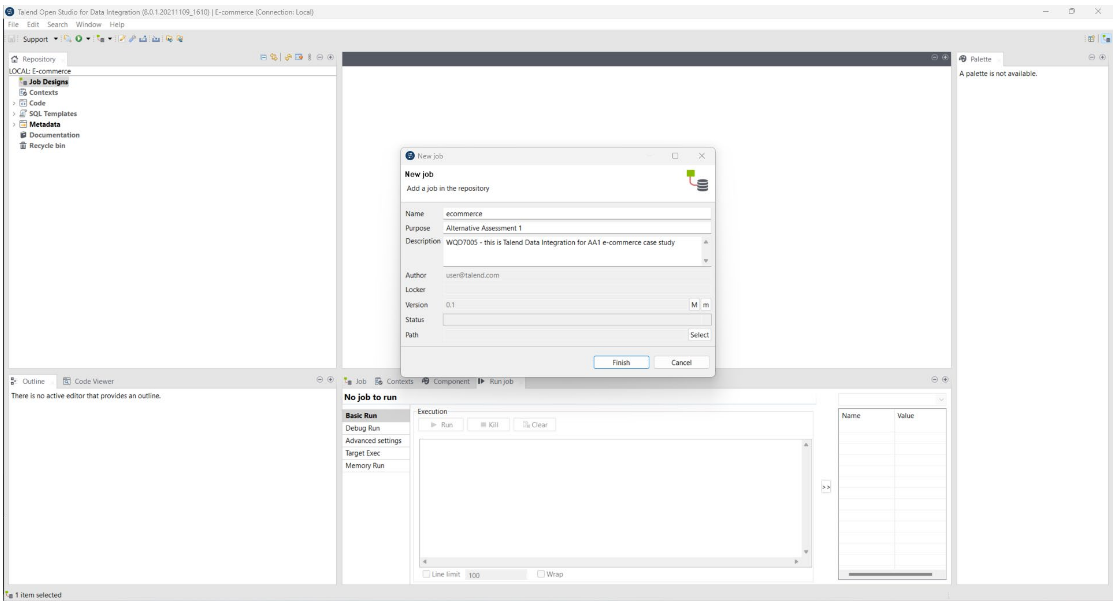
A new project is created in Talend Data Integration which title as ‘e-commerce’



WQD7005 DATA MINING | ATERNATIVE ASSESSMENT 1

ID: 17147394 | Fathia Farhana bt Agusalim

- To create a New Job:
  - Right-click in the Repository panel on the left under Job Designs.
  - Choose Create job.
  - Provide a name for the job (e.g., "ecommerce ") and a description. Click Finish.



Talend Open Studio for Data Integration (8.0.1.20211109_1610) | E-commerce (Connection: Local)

New job

Add a job in the repository

Name: ecommerce

Purpose: Alternative Assessment 1

Description: WQD7005 - this is Talend Data Integration for AA1 e-commerce case study

Author: user@talend.com

Locker:

Version: 0.1

Status:

Path:

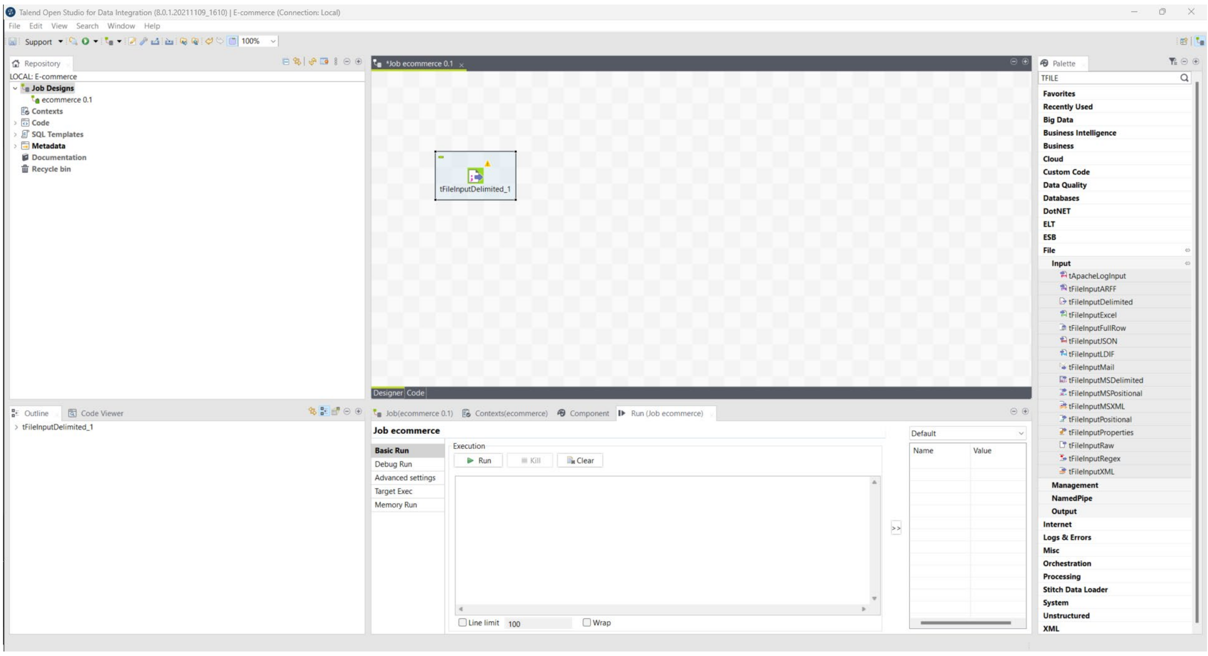
A palette is not available.

There is no active editor that provides an outline.

No job to run

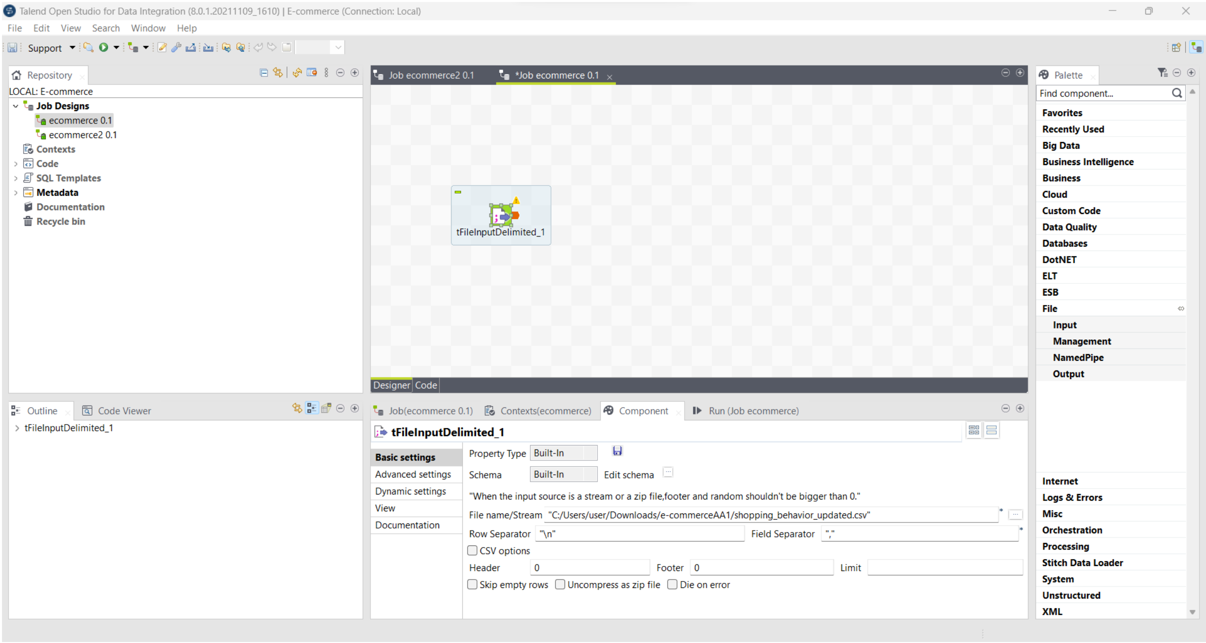
1 item selected

- Add a tFileInputDelimited Component:
  - This component is used to read delimited files like CSV.
  - From the Palette panel on the right, type "tFileInputDelimited" into the search bar.
  - Drag the tFileInputDelimited component to the design workspace.



Talend Open Studio for Data Integration (8.0.1.20211109_1610) | E-commerce (Connection: Local)

- Configure the tFileInputDelimited Component:
  - Click on the tFileInputDelimited component to select it.
  - In the Component panel below, set the properties:
    - File Name/Stream: Browse to and select your CSV file.
    - Row Separator: Set this to "\n" for new lines (most common for CSV files).
    - Field Separator: Set this to "," depending on file CSV is delimited.
    - Header: Since the dataset obtain Header, hence, set this to 1. .
    - Under the Schema section, click on the Edit schema button. Here, define the columns in the CSV file by adding columns and specifying their data types. This schema should match the structure of CSV file.



tFileInputDelimited_1

Property Type: Built-In

Schema: Built-In — Edit schema

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

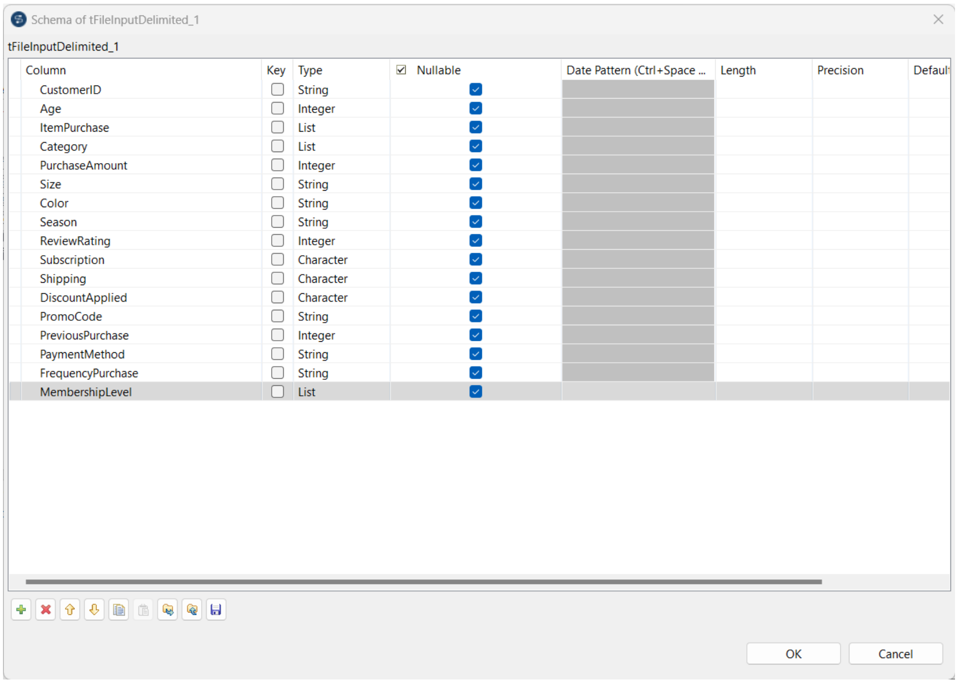
File name/Stream: "C:/Users/user/Downloads/e-commerceAA1/shopping_behavior_updated.csv"

Row Separator: "\n" Field Separator: ","

□ CSV options

Header: 0 Footer: 0 Limit:

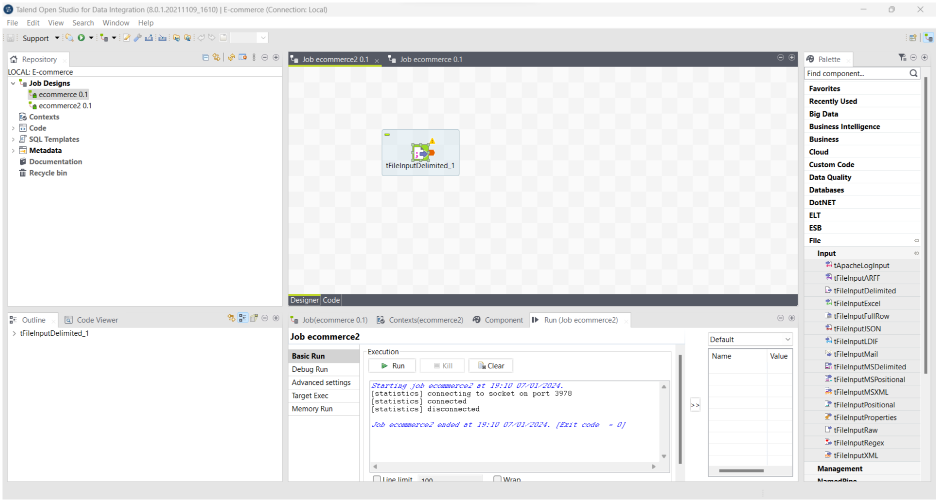
□ Skip empty rows □ Uncompress as zip file □ Die on error



tFileInputDelimited_1

Column	Key	Type	☑ Nullable	Date Pattern (Ctrl+Space ...	Length	Precision	Defaul
CustomerID	□	String	☑				
Age	□	Integer	☑				
ItemPurchase	□	List	☑				
Category	□	List	☑				
PurchaseAmount	□	Integer	☑				
Size	□	String	☑				
Color	□	String	☑				
Season	□	String	☑				
ReviewRating	□	Integer	☑				
Subscription	□	Character	☑				
Shipping	□	Character	☑				
DiscountApplied	□	Character	☑				
PromoCode	□	String	☑				
PreviousPurchase	□	Integer	☑				
PaymentMethod	□	String	☑				
FrequencyPurchase	□	String	☑				
MembershipLevel	□	List	☑				

- Run the Job to View Data:



```
Starting job ecommerce2 at 19:10 07/01/2024.
[statistics] connecting to socket on port 3978
[statistics] connected
[statistics] disconnected

Job ecommerce2 ended at 19:10 07/01/2024. [Exit code  = 0]
```

Based on the Running the job, the Talend Data Integration encounter an error. This is due to the:

- There was a connection made to a socket on port 3735, which is likely for real-time debugging or statistics gathering.
- Disconnected: The connection to the socket was disconnected normally.
- Job ended: The job ecommerce ended with an exit code of 0 which indicates that the job has completed successfully without errors. If there was an error, we would typically see a non-zero exit code and an error message in the log.

Several attempts on refreshing and creating new job to ensure that job can be successfully Run in the tools such as creating e-commerce2 new job as well as restarting the PC, however similar results achieved.

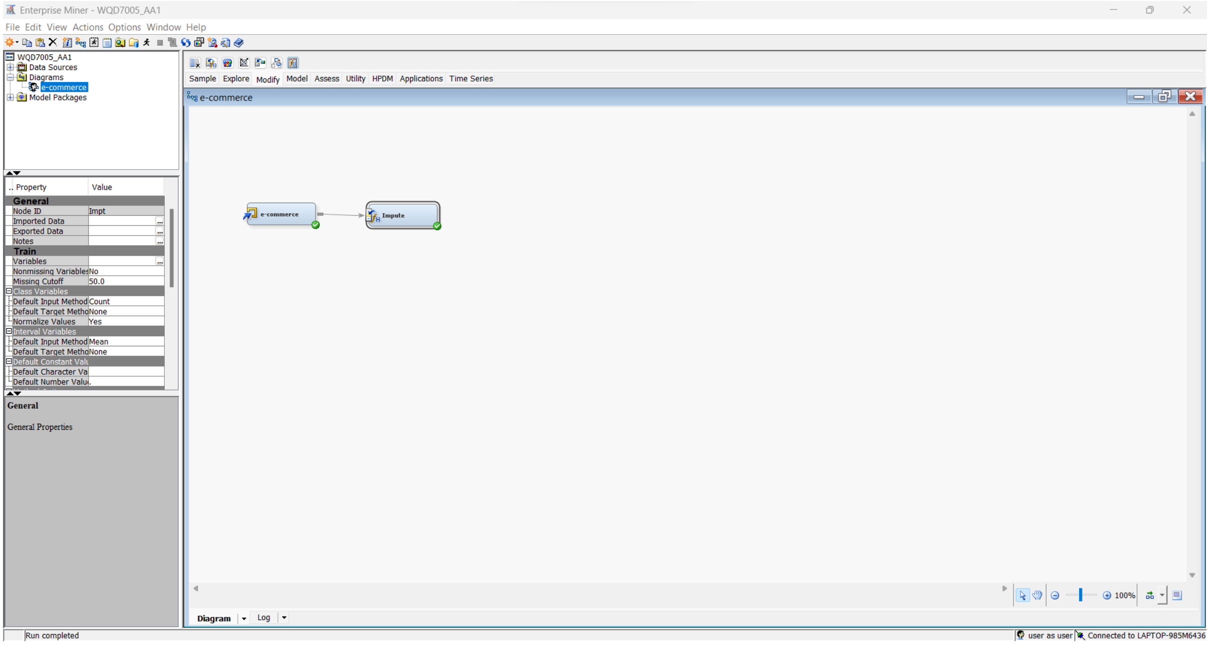
Hence for execution of Data Preprocessing using Talend Data Integration, the exercise using this tool will be excluded and assessment will be proceeded with Talend Data Preparation and SAS E-Miner.

In summary, three (3) tools have been executed to be used for this e-commerce assessment

1.6. Imputing Missing Values

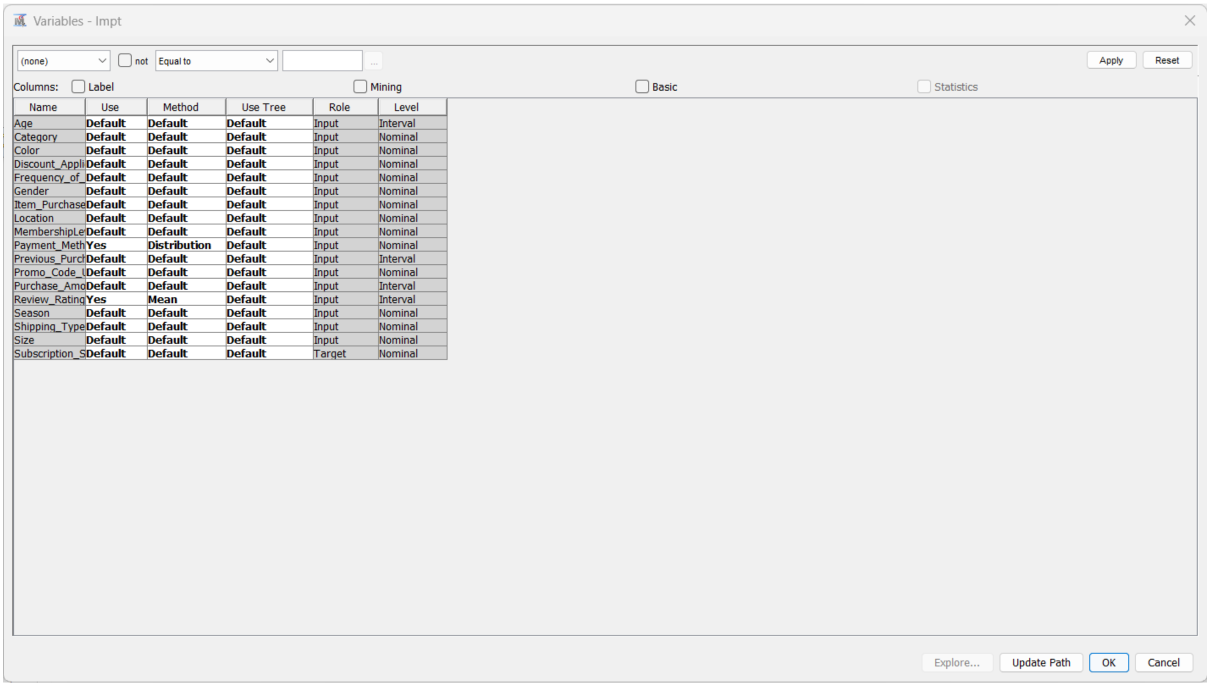
As described in the above section, there are missing values for “Review Rating” and “Payment Method” variables. Hence to handle the missing values, from “Modify” tab, select “Impute” and edit the variables for:

- “Review Rating” to “Mean”
- “Payment Method” to “Distribution”. Selecting “Distribution” for imputation means that the missing values will be imputed based on the statistical distribution of the non-missing values of that variable.


Enterprise Miner - WQD7005_AA1
File Edit View Actions Options Window Help
WQD7005_AA1
Data Sources
Diagrams
e-commerce
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Time Series
e-commerce
e-commerce
Impute
Diagram
Log
100%
General
General Properties
Run completed
user as user
Connected to LAPTOP-985M6436

Property	Value
General	
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Metho	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Metho	None
Default Constant Valu	
Default Character Va	
Default Number Valu	.

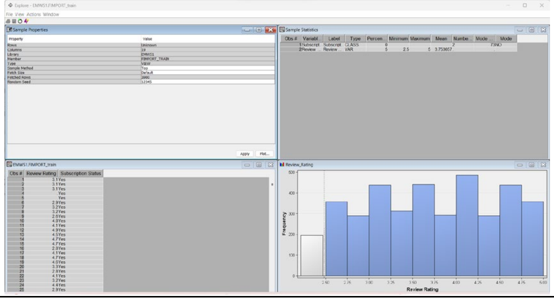
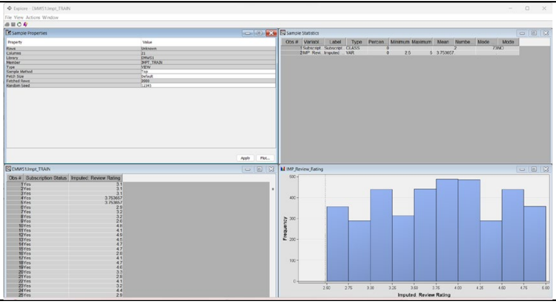
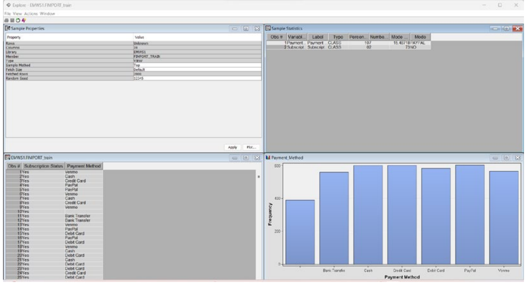
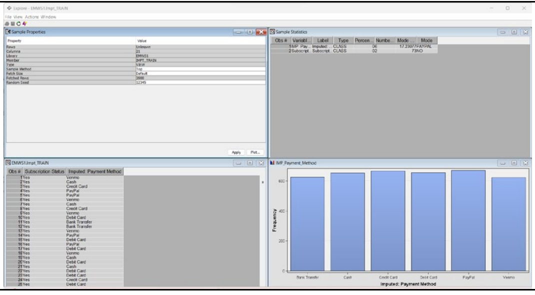
Figure 1. 14: Select Impute from Modify tab



Name	Use	Method	Use Tree	Role	Level
Age	**Default**	**Default**	**Default**	Input	Interval
Category	**Default**	**Default**	**Default**	Input	Nominal
Color	**Default**	**Default**	**Default**	Input	Nominal
Discount_Appli	**Default**	**Default**	**Default**	Input	Nominal
Frequency_of_	**Default**	**Default**	**Default**	Input	Nominal
Gender	**Default**	**Default**	**Default**	Input	Nominal
Item_Purchase	**Default**	**Default**	**Default**	Input	Nominal
Location	**Default**	**Default**	**Default**	Input	Nominal
MembershipLe	**Default**	**Default**	**Default**	Input	Nominal
Payment_Meth	**Yes**	**Distribution**	**Default**	Input	Nominal
Previous_Purch	**Default**	**Default**	**Default**	Input	Interval
Promo_Code_U	**Default**	**Default**	**Default**	Input	Nominal
Purchase_Amo	**Default**	**Default**	**Default**	Input	Interval
Review_Rating	**Yes**	**Mean**	**Default**	Input	Interval
Season	**Default**	**Default**	**Default**	Input	Nominal
Shipping_Type	**Default**	**Default**	**Default**	Input	Nominal
Size	**Default**	**Default**	**Default**	Input	Nominal
Subscription_S	**Default**	**Default**	**Default**	Target	Nominal

Figure 1. 15: Impute type for Payment Method and Review Rating

Table 1. 1: Representation of Imputation Result

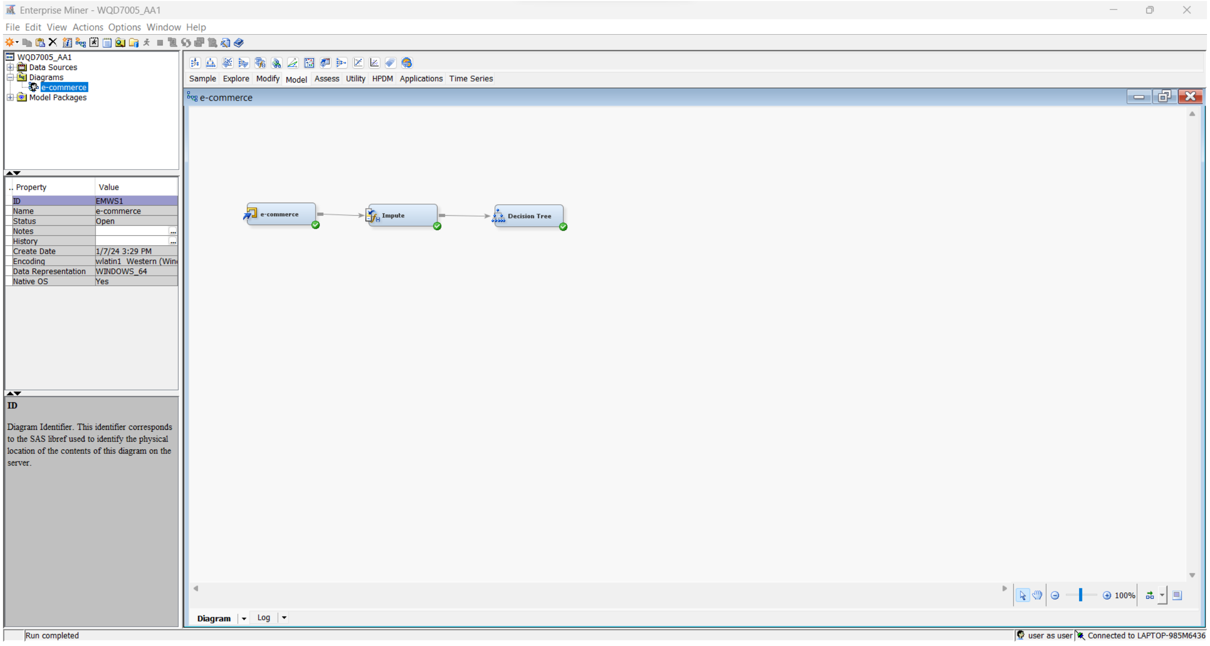
Original Dataset	After imputation
Review Rating – the missing values has been edited with Mean Values	
 Sample Statistics Obs # | Variabl... | Label | Type | Percen... | Minimum | Maximum | Mean | Numbe... | Mode ... | Mode 1 | Subscript | Subscript | CLASS | 0 | | | | 2 | 73 | NO 2 | Review | Review | VAR | 5 | 2.5 | 5 | 3.753657 | | | Review_Rating	 Sample Statistics Obs # | Variabl... | Label | Type | Percen... | Minimum | Maximum | Mean | Numbe... | Mode ... | Mode 1 | Subscript | Subscript | CLASS | 0 | | | | 2 | 73 | NO 2 | IMP_Revi... | Imputed: | VAR | 0 | 2.5 | 5 | 3.753657 | | | IMP_Review_Rating
Payment Method – the missing values has been edited based on statistical distribution	
 Sample Statistics Obs # | Variabl... | Label | Type | Percen... | Numbe... | Mode ... | Mode 1 | Payment... | Payment ... | CLASS | 10 | 7 | 15.46718 | PAYPAL 2 | Subscript | Subscript | CLASS | 0 | 2 | 73 | NO Payment_Method	 Sample Statistics Obs # | Variabl... | Label | Type | Percen... | Numbe... | Mode ... | Mode 1 | IMP_Pay... | Imputed: ... | CLASS | 0 | 6 | 17.23077 | PAYPAL 2 | Subscript | Subscript | CLASS | 0 | 2 | 73 | NO IMP_Payment_Method

2. Decision Tree Analysis

2.1. Decision Tree Model

To understand the customer behaviour, decision tree model is established. From “Model” tab, select “Decision Tree”.

In order to splitting Role and Node, we need to define the Maximum Branch and Maximum Depth to 2 and 6 and Leaf Size Node to 5. We can maintain all the remaining default value and click “Run” to execute the model

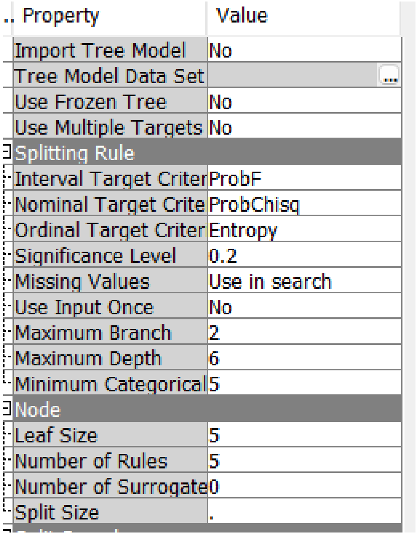


.. Property	Value
ID	EMWS1
Name	e-commerce
Status	Open
Notes	...
History	...
Create Date	1/7/24 3:29 PM
Encoding	wlatin1 Western (Win
Data Representation	WINDOWS_64
Native OS	Yes

ID

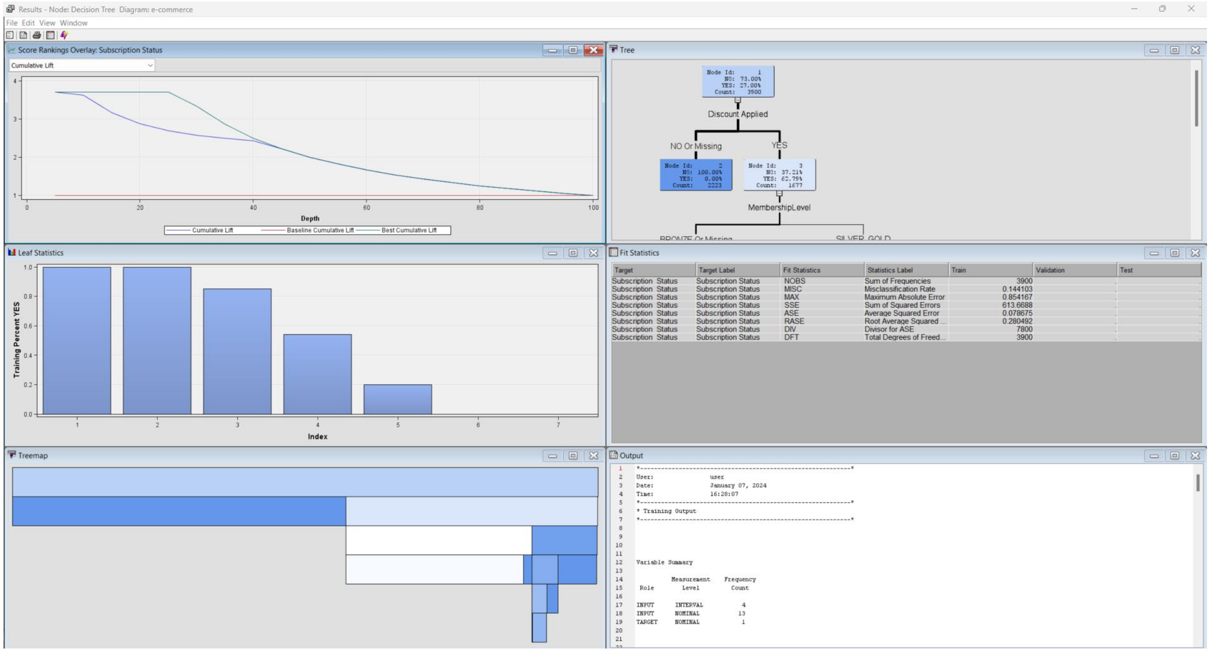
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Figure 2. 1: Decision Tree Model



.. Property	Value
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criter	ProbF
Nominal Target Crite	ProbChisq
Ordinal Target Criter	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate	0
Split Size	.

Figure 2. 2: Specifying the Property in Decision Tree Model



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Subscription Status	Subscription Status	NOBS	Sum of Frequencies	3900	.	.
Subscription Status	Subscription Status	MISC	Misclassification Rate	0.144103	.	.
Subscription Status	Subscription Status	MAX	Maximum Absolute Error	0.854167	.	.
Subscription Status	Subscription Status	SSE	Sum of Squared Errors	613.6688	.	.
Subscription Status	Subscription Status	ASE	Average Squared Error	0.078675	.	.
Subscription Status	Subscription Status	RASE	Root Average Squared ...	0.280492	.	.
Subscription Status	Subscription Status	DIV	Divisor for ASE	7800	.	.
Subscription Status	Subscription Status	DFT	Total Degrees of Freed...	3900	.	.

Output

```
*------------------------------------------------------------*
User:                    user
Date:                    January 07, 2024
Time:                    16:28:07
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role         Level         Count

INPUT      INTERVAL           4
INPUT      NOMINAL           13
TARGET     NOMINAL            1


```

Figure 2. 3: Overall Result from Decision Tree Model

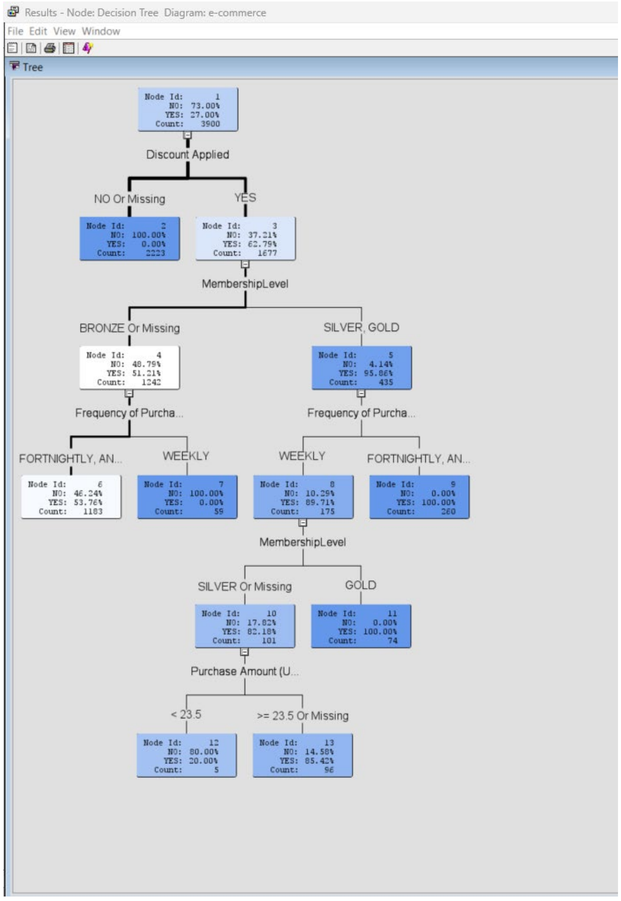
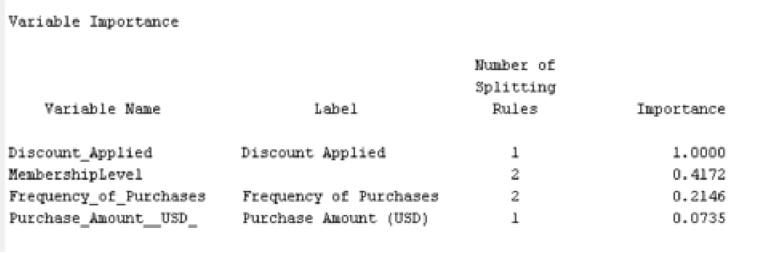


Figure 2. 4: Decision Tree Model based on Prediction Subscription Status

Based on the above result, the decision tree model appears to predict 'Subscription Status', with different paths in the tree indicating the profiles of customers who are likely to subscribe or not, based on the input variables in “e-commerce” dataset that shown in Figure 2.3 and Figure 2.4 above. It can be seen that each node (representing in the box) represents a decision rule that splits the data based on the best predictor variable at that point.

In the Decision Tree, we can see splits based on 'Discount Applied', 'MembershipLevel', 'Frequency of Purchases', and 'Purchase Amount (USD)'. The splits are binary, dividing records into two groups based on the criterion in each node.

The fact that 'Discount Applied' is the first split suggests it is an important predictor of 'Subscription Status' with important of 1.0000 followed by Membership Level with degree of important 0.4172. The Purchase_Amount_USD_ indicate least important on the customer subscription status with important 0.0735 as shown in Figure 2.5 below.



```
Variable Importance

                                                     Number of
                                                     Splitting
   Variable Name              Label                    Rules          Importance

Discount_Applied         Discount Applied                1               1.0000
MembershipLevel                                          2               0.4172
Frequency_of_Purchases   Frequency of Purchases          2               0.2146
Purchase_Amount__USD_    Purchase Amount (USD)           1               0.0735
```

Figure 2. 5: Variable Importance from Decision Tree Model

The decision tree model's predictive ability can be evaluated from the lift chart and fit statistics which can be observed from Figure 2.6 and Figure 2.7 respectively. From the chart. it seems that the predictive model to be performing better than random guessing but is not perfect.

The point where the lift line is furthest above the baseline indicates where the model is most effective. The 'Depth' on the x-axis usually represents the percentage of the dataset or population. For example, a depth of 20 could represent the top 20% of predictions from the model. The lift chart typically starts high and then decreases. This is expected because we are capturing the best predictions first (those most likely to be true positives), and as we move through the population, the effectiveness of the model diminishes.

If the lift line is below the baseline, it indicates that the model is performing worse than random chance at that point. Based on the lift chart, the lift line shows that the model has good lift at the beginning, meaning it is effective at identifying positive outcomes, but the effectiveness decreases as we move through the population. This is typical behaviour for a lift chart and indicates that using the model to prioritize the outreach or interventions would be more effective than approaching customers randomly.

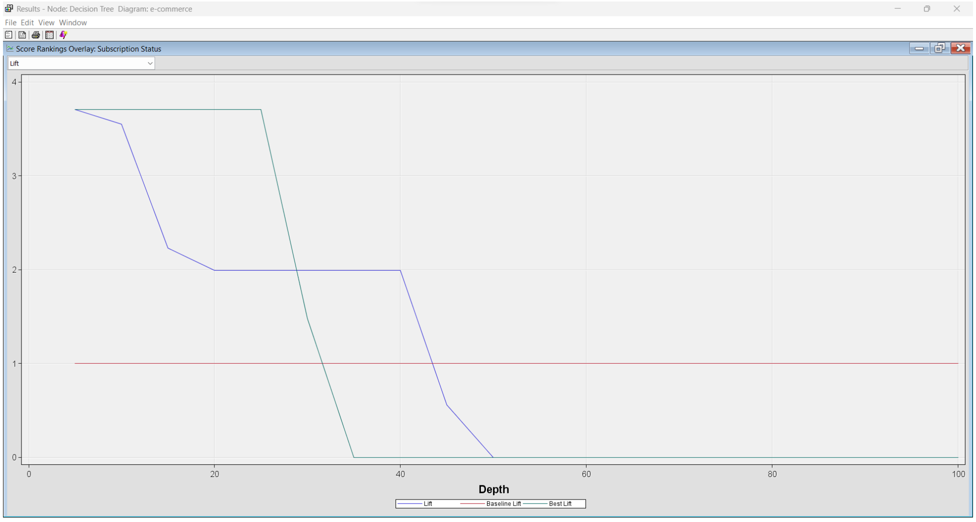
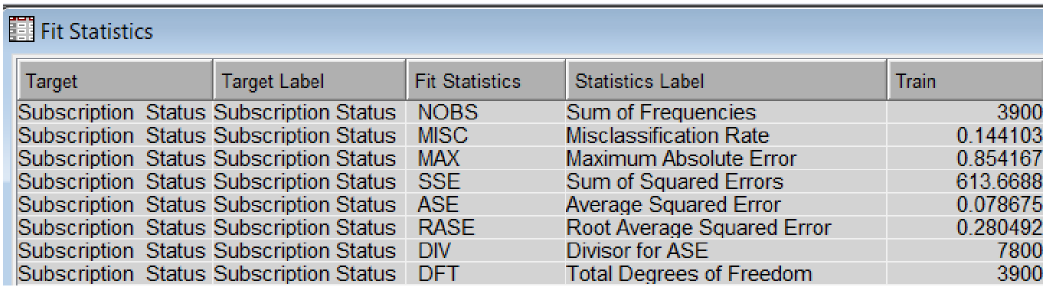


Figure 2. 6: Lift Chart from Decision Tree Model

However, it is to be noted that there is misclassification rate in the Fit Statistics with 0.144103 which indicate that the percentage of the predictions were wrong, which can be a straightforward way of evaluating the model's accuracy. The misclassification rate of about 14.41% suggests that the model is relatively accurate, correctly classifying about 85.59% of cases.

The Average Squared Error (ASE) and Root Average Squared Error (RASE) provide measures of the model's error in terms of probability estimations for the classes. Based on the result, it can be seen that the ASE and RASE is lower which indicate the model is making fewer errors on average.

The 'Total Degrees of Freedom' being equal to the 'Number of Observations' suggests that this might be a simple model without many parameters, or it might be an indication of an error in reporting or calculation

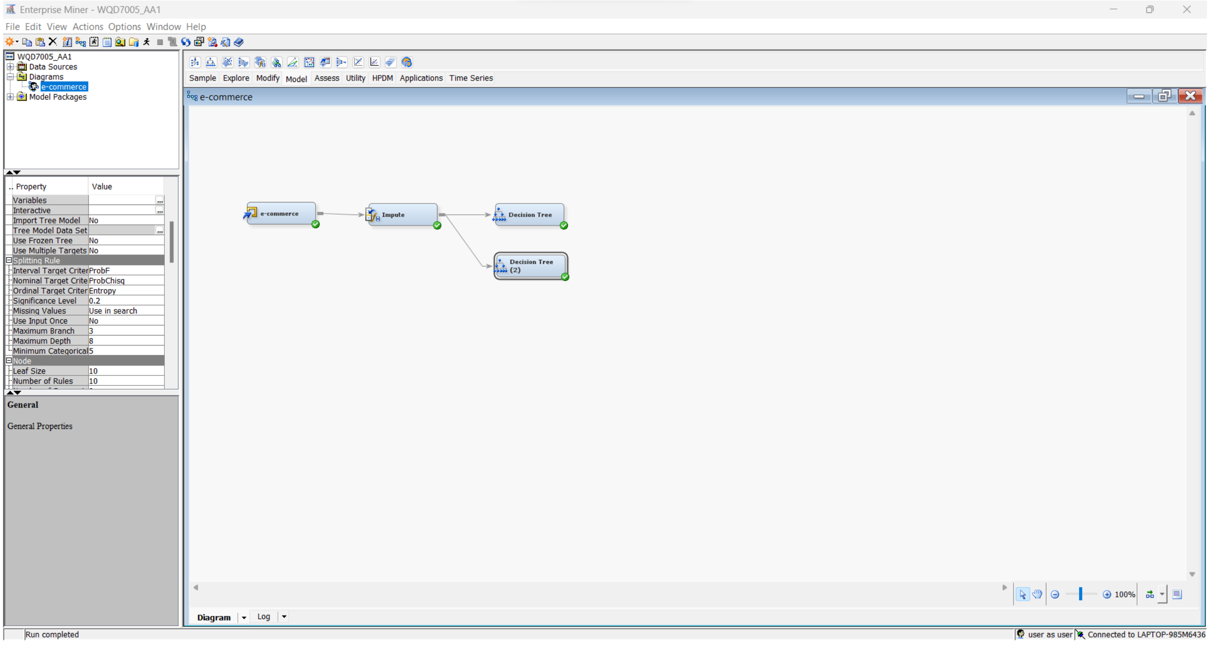

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train
Subscription Status	Subscription Status	NOBS	Sum of Frequencies	3900
Subscription Status	Subscription Status	MISC	Misclassification Rate	0.144103
Subscription Status	Subscription Status	MAX	Maximum Absolute Error	0.854167
Subscription Status	Subscription Status	SSE	Sum of Squared Errors	613.6688
Subscription Status	Subscription Status	ASE	Average Squared Error	0.078675
Subscription Status	Subscription Status	RASE	Root Average Squared Error	0.280492
Subscription Status	Subscription Status	DIV	Divisor for ASE	7800
Subscription Status	Subscription Status	DFT	Total Degrees of Freedom	3900

Figure 2. 7: Fit Statistics Result Summary

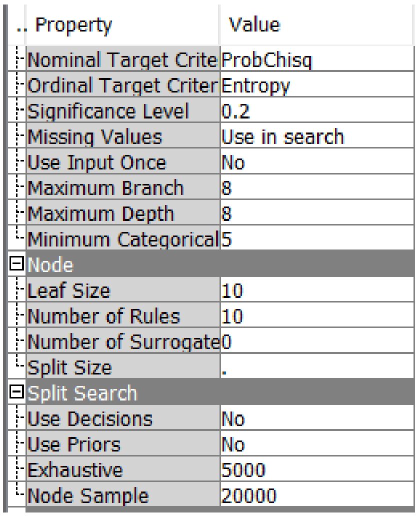
In order to improve the prediction model, it would be also essential to look at the performance on a validation set to get a sense of how the model might perform on new or unseen data.

Now, we try to see how does changing the Maximum Branch and Depth and Node for Leaf Size and Number of Rules will affect the result. From the Property, we will specify the Maximum Branch and Maximum Depth to be 4 and 8 and Node for Lead Size and Number of Rules to 10 (double from the first decision tree model)



.. Property	Value
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
⊟ Splitting Rule	
Interval Target Criter	ProbF
Nominal Target Crite	ProbChisq
Ordinal Target Criter	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	8
Minimum Categorical	5
⊟ Node	
Leaf Size	10
Number of Rules	10

Figure 2. 8: Decision Tree Model 2

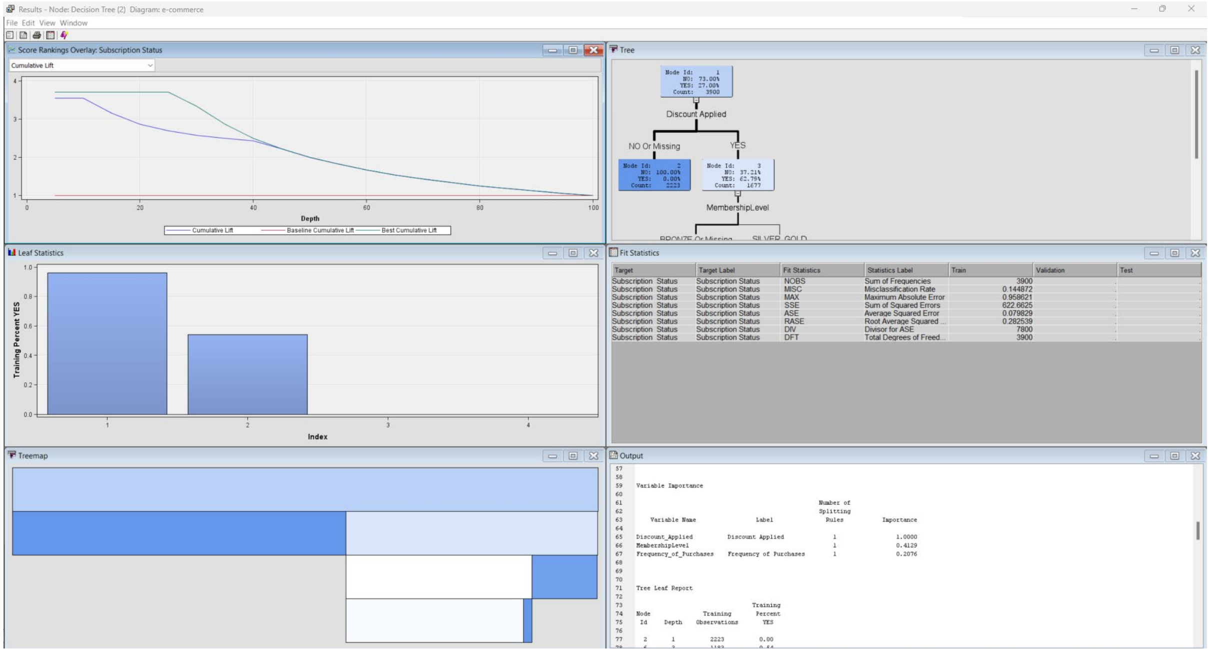


.. Property	Value
Nominal Target Crite	ProbChisq
Ordinal Target Criter	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	8
Maximum Depth	8
Minimum Categorical	5
⊟ Node	
Leaf Size	10
Number of Rules	10
Number of Surrogate	0
Split Size	.
⊟ Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

Figure 2. 9: Specifying the Property for Decision Tree Model 2

WQD7005 DATA MINING | ATERNATIVE ASSESSMENT 1

ID: 17147394 | Fathia Farhana bt Agusalim



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Subscription Status	Subscription Status	NOBS	Sum of Frequencies	3900	.	.
Subscription Status	Subscription Status	MISC	Misclassification Rate	0.144872	.	.
Subscription Status	Subscription Status	MAX	Maximum Absolute Error	0.958621	.	.
Subscription Status	Subscription Status	SSE	Sum of Squared Errors	622.6625	.	.
Subscription Status	Subscription Status	ASE	Average Squared Error	0.079829	.	.
Subscription Status	Subscription Status	RASE	Root Average Squared ...	0.282539	.	.
Subscription Status	Subscription Status	DIV	Divisor for ASE	7800	.	.
Subscription Status	Subscription Status	DFT	Total Degrees of Freed...	3900	.	.

```


Variable Importance

                                                            Number of
                                                            Splitting
   Variable Name                  Label                       Rules           Importance

Discount_Applied          Discount Applied                      1               1.0000
MembershipLevel                                                 1               0.4129
Frequency_of_Purchases    Frequency of Purchases                1               0.2076



Tree Leaf Report

                                          Training
Node                     Training          Percent
 Id      Depth         Observations          YES

  2        1               2223              0.00
```

Figure 2. 10: Overall Result from Decision Tree Model 2

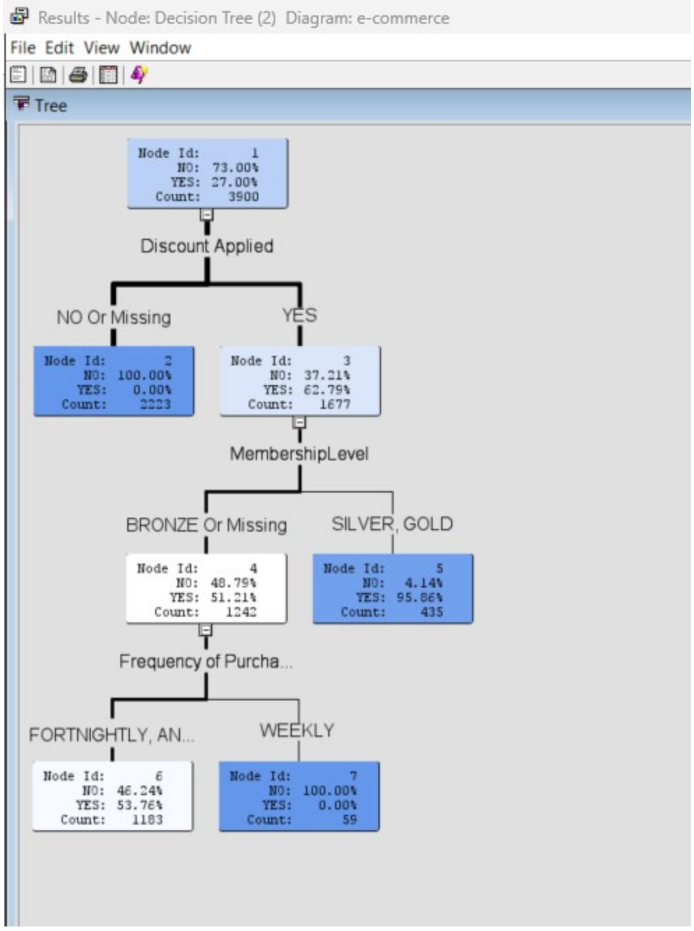


Figure 2. 11: Decision Tree Model 2 based on Prediction Subscription Status

Based on the result, the “Discount Applied” remain the first variable importance in the decision tree where it is the most significant predictor of “Subscription Status”. For those who received a discount, 'MembershipLevel' is an important next predictor, with 'Frequency of Purchases' also playing a crucial role for customers with higher membership levels. The tree suggests that customers with "SILVER, GOLD" membership and "WEEKLY" purchase frequency have a higher likelihood (95.86% "YES") of subscribing.

In terms of comparison between the first and second Decision Tree Model, the choice between the two depends on the context and the goal of the analysis.

If interpretability and simplicity are priorities (for example, if the rules are to be communicated to a non-technical audience or if the model is to be implemented in a real-world scenario where simplicity is key), the second tree might be preferable.

If the goal is to maximize prediction accuracy and the additional complexity of the first tree offers a significant improvement in performance (as assessed by validation metrics), the first tree might be the better choice. In addition, it is t's also important to consider overfitting whereby more complex trees can fit the training data too closely, capturing noise rather than underlying patterns, which can lead to poor performance on new data. This is why model validation using separate datasets is crucial.

3. Ensemble Method Analysis

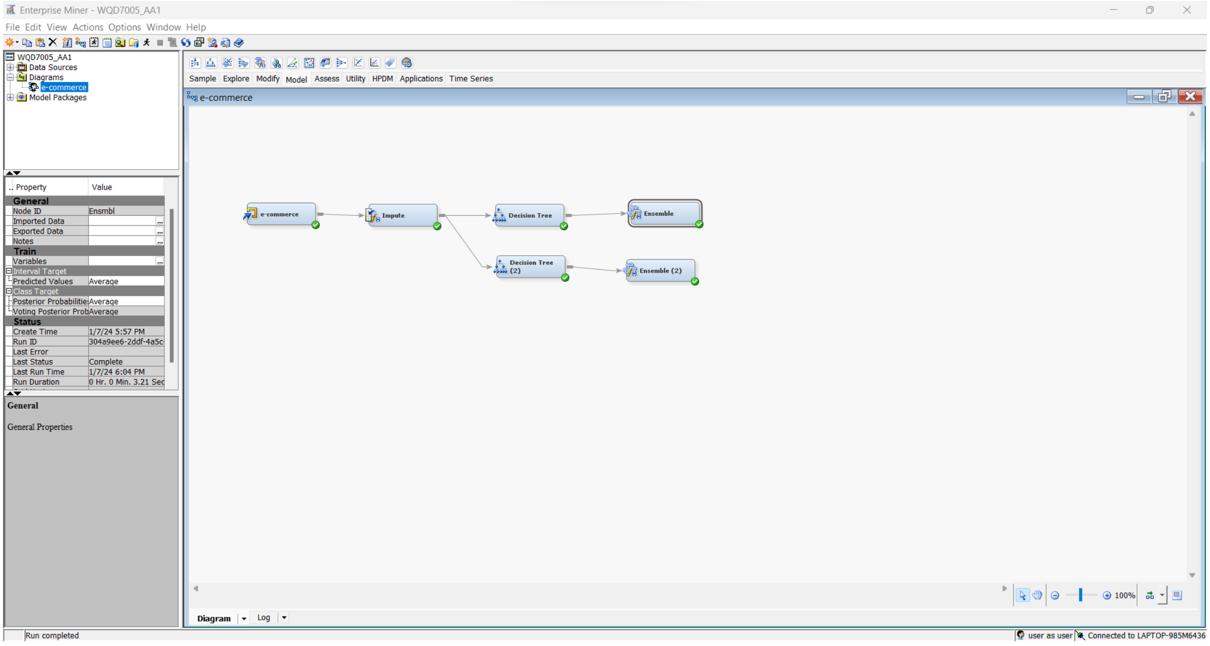
In performing Ensemble Method, below steps will be executed in the SAS E-Miner. The Ensemble Node creates a new model by taking a function of posterior probabilities (for class targets) or the predicted values (for interval target) from multiple models.

First, we need to drag and drop the "Ensemble" node from the “Model” tab into the workspace. To configure the Ensemble Node, double-click on the "Ensemble" node to configure its properties. We then specify type of ensemble method that need to be executed, in this case for Bagging (Bootstrap Aggregating), we will select "Bagging" as the method. Then, choose "Random Forest" as the algorithm within the bagging method settings. In SAS Enterprise Miner, Random Forest is a separate node, but it can also be part of the ensemble node options.

Connect the data source node to the "Ensemble" node and “Run” the result.

WQD7005 DATA MINING | ATERNATIVE ASSESSMENT 1

ID: 17147394 | Fathia Farhana bt Agusalim

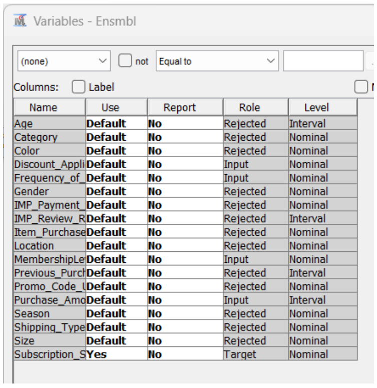

Enterprise Miner - WQD7005_AA1
File Edit View Actions Options Window Help
WQD7005_AA1
Data Sources
Diagrams
e-commerce
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Time Series
e-commerce
e-commerce
Impute
Decision Tree
Ensemble
Decision Tree (2)
Ensemble (2)
100%
Diagram
Log
Run completed
user as user
Connected to LAPTOP-985M6436

Property	Value
General	
Node ID	Ensmbl
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interval Target	
Predicted Values	Average
Class Target	
Posterior Probabilitie	Average
Voting Posterior Prob	Average
Status	
Create Time	1/7/24 5:57 PM
Run ID	304a9ee6-2ddf-4a5c
Last Error	
Last Status	Complete
Last Run Time	1/7/24 6:04 PM
Run Duration	0 Hr. 0 Min. 3.21 Sec

General

General Properties

Figure 3. 1: Ensemble Method Analysis

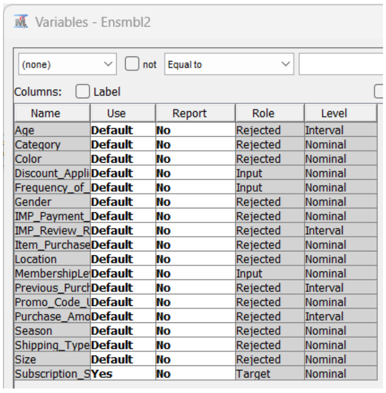


Variables - Ensmbl

(none) □ not Equal to

Columns: □ Label

Name	Use	Report	Role	Level
Age	Default	No	Rejected	Interval
Category	Default	No	Rejected	Nominal
Color	Default	No	Rejected	Nominal
Discount_Appli	Default	No	Input	Nominal
Frequency_of_	Default	No	Input	Nominal
Gender	Default	No	Rejected	Nominal
IMP_Payment_	Default	No	Rejected	Nominal
IMP_Review_R	Default	No	Rejected	Interval
Item_Purchase	Default	No	Rejected	Nominal
Location	Default	No	Rejected	Nominal
MembershipLe	Default	No	Input	Nominal
Previous_Purch	Default	No	Rejected	Interval
Promo_Code_U	Default	No	Rejected	Nominal
Purchase_Amo	Default	No	Input	Interval
Season	Default	No	Rejected	Nominal
Shipping_Type	Default	No	Rejected	Nominal
Size	Default	No	Rejected	Nominal
Subscription_S	Yes	No	Target	Nominal



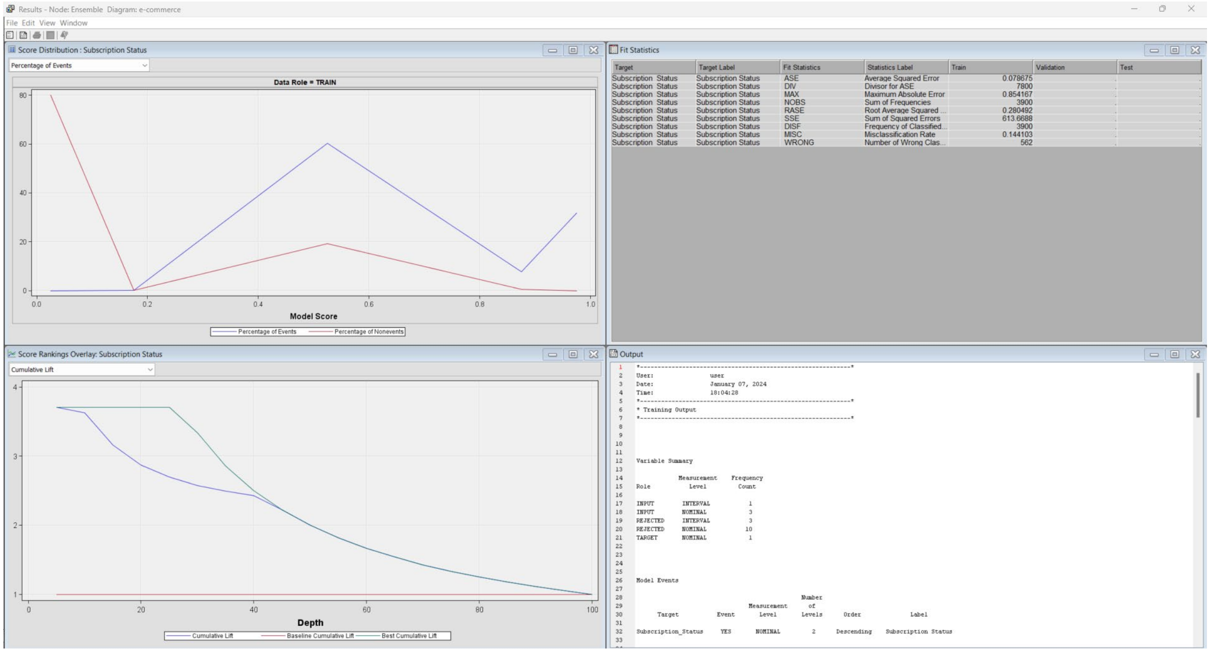
Variables - Ensmbl2

(none) □ not Equal to

Columns: □ Label

Name	Use	Report	Role	Level
Age	Default	No	Rejected	Interval
Category	Default	No	Rejected	Nominal
Color	Default	No	Rejected	Nominal
Discount_Appli	Default	No	Input	Nominal
Frequency_of_	Default	No	Input	Nominal
Gender	Default	No	Rejected	Nominal
IMP_Payment_	Default	No	Rejected	Nominal
IMP_Review_R	Default	No	Rejected	Interval
Item_Purchase	Default	No	Rejected	Nominal
Location	Default	No	Rejected	Nominal
MembershipLe	Default	No	Input	Nominal
Previous_Purch	Default	No	Rejected	Interval
Promo_Code_U	Default	No	Rejected	Nominal
Purchase_Amo	Default	No	Rejected	Interval
Season	Default	No	Rejected	Nominal
Shipping_Type	Default	No	Rejected	Nominal
Size	Default	No	Rejected	Nominal
Subscription_S	Yes	No	Target	Nominal

Figure 3. 2: Variables for Ensemble 1 and Ensemble 2



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Subscription Status	Subscription Status	ASE	Average Squared Error	0.078675	.	.
Subscription Status	Subscription Status	DIV	Divisor for ASE	7800	.	.
Subscription Status	Subscription Status	MAX	Maximum Absolute Error	0.854167	.	.
Subscription Status	Subscription Status	NOBS	Sum of Frequencies	3900	.	.
Subscription Status	Subscription Status	RASE	Root Average Squared ...	0.280492	.	.
Subscription Status	Subscription Status	SSE	Sum of Squared Errors	613.6688	.	.
Subscription Status	Subscription Status	DISF	Frequency of Classified...	3900	.	.
Subscription Status	Subscription Status	MISC	Misclassification Rate	0.144103	.	.
Subscription Status	Subscription Status	WRONG	Number of Wrong Clas...	562	.	.

```
*------------------------------------------------------------*
User:                user
Date:                January 07, 2024
Time:                18:04:28
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level         Count

INPUT        INTERVAL          1
INPUT        NOMINAL           3
REJECTED     INTERVAL          3
REJECTED     NOMINAL          10
TARGET       NOMINAL           1




Model Events

                                                  Number
                                     Measurement    of
      Target           Event        Level       Levels     Order              Label

Subscription_Status     YES       NOMINAL         2      Descending    Subscription Status

```

Figure 3. 3: Result Summary for Ensemble Method

Based on the result shown in Figure 3.3, it can be seen be summarised that the model appears to have a moderate level of predictive accuracy, with a misclassification rate of 14% (0.144103). The ensemble method, which seems to be averaging predictions from a set of models, has been able to correctly identify a good proportion of both events (subscriptions – “YES”) and non-events (non-subscriptions “NO”).

Based on the Score Distribution Plot, it can be seen that the x-axis represents the model score, which is the predicted probability that an observation is an event (YES). Ideally, we want to obtain the blue line chart (events) high on the left (indicating high model scores for actual events) and the red line (non-events) high on the right (indicating low model scores for actual non-events).

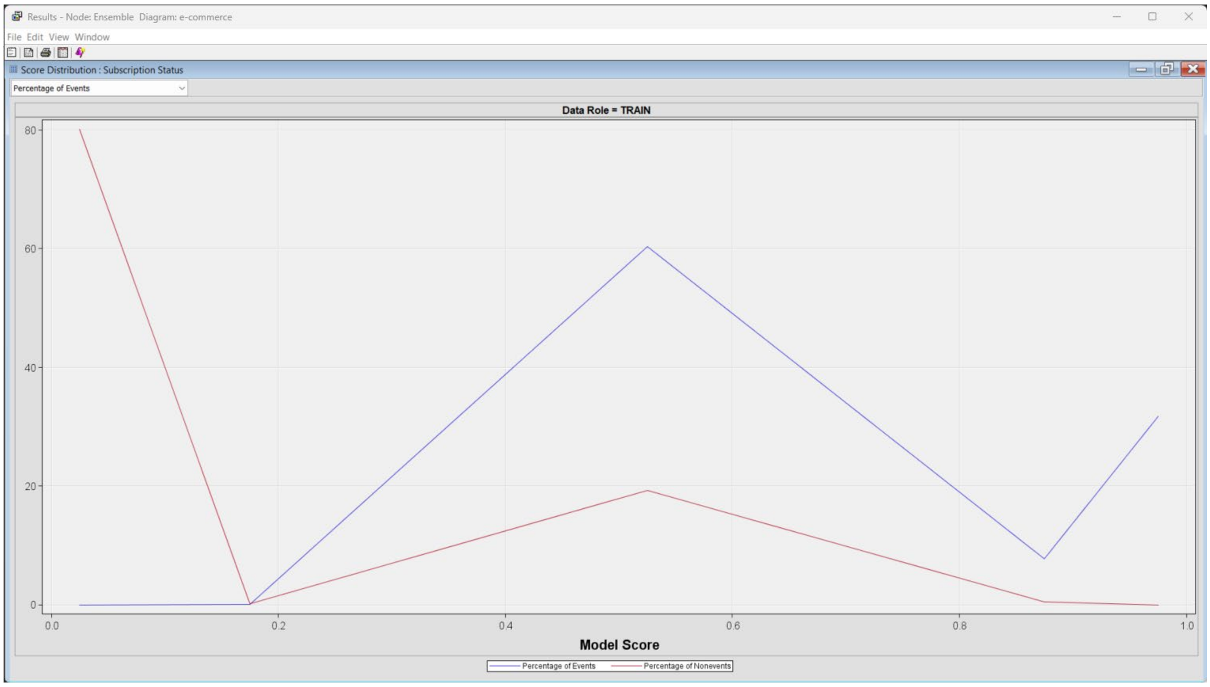


Figure 3. 4: Score Distribution Plot

In terms of Cumulative Lift Chart, it can be observed that the chart shows how much better the ensemble model is at predicting the target variable compared to a random guess. The "Cumulative Lift" line should ideally be above the "Baseline Cumulative Lift" line, which indicates the model is better than random chance.

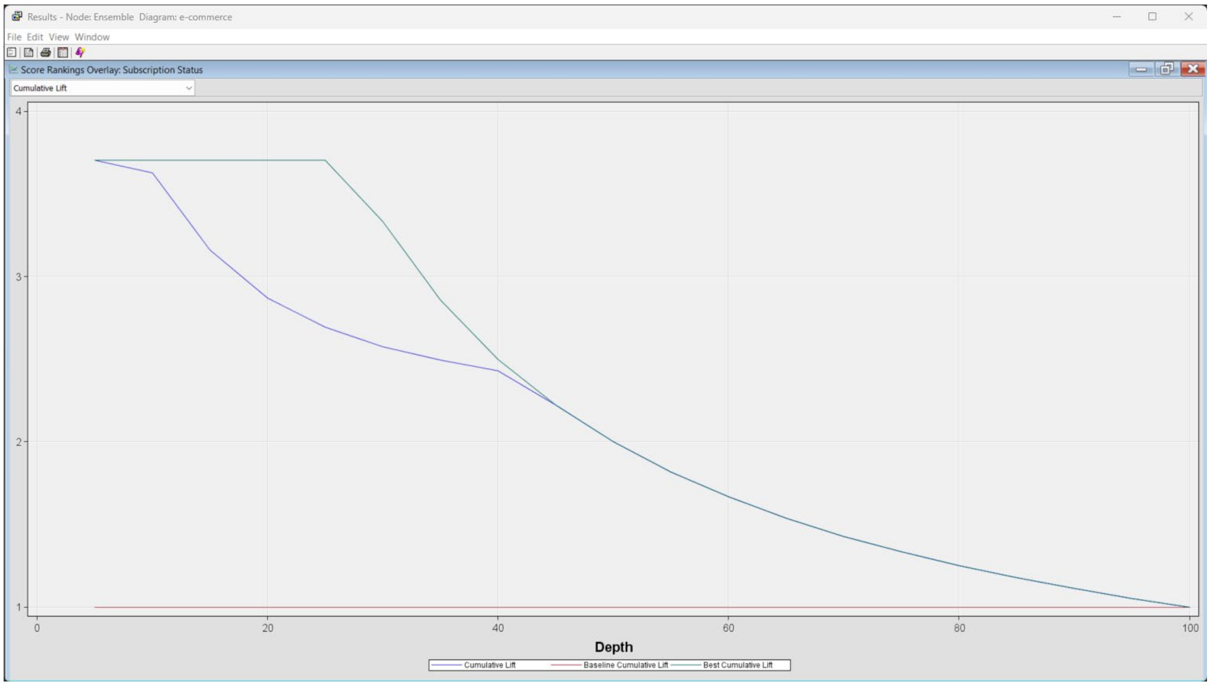


Figure 3. 5: Cumulative Lift Results

For the Gain summary result, the model starts off capturing a high number of positive cases (subscriptions) quickly. This is evident in the steep initial slope of the Gain Line.

As more cases are considered, the incremental gains decrease. The Gain Line is above the Baseline, indicating that the model is doing better than random chance at identifying positive outcomes. The Best Gain Line indicates the maximum potential for gain, which is not fully visible in the chart but would typically be at or near the top of the chart, showing the theoretical maximum gain if every positive case was identified before any negative case.

Based on the overall Gain line chart, the ensemble model is effective, especially at the top end, meaning it is good at identifying the most likely positive outcomes. However, as it reach deeper into the population, the effectiveness diminishes, which is typical of such models. The model would be most useful for targeting the top-ranked individuals according to the model's scores, as that's where the most substantial gains are realized compared to random selection.

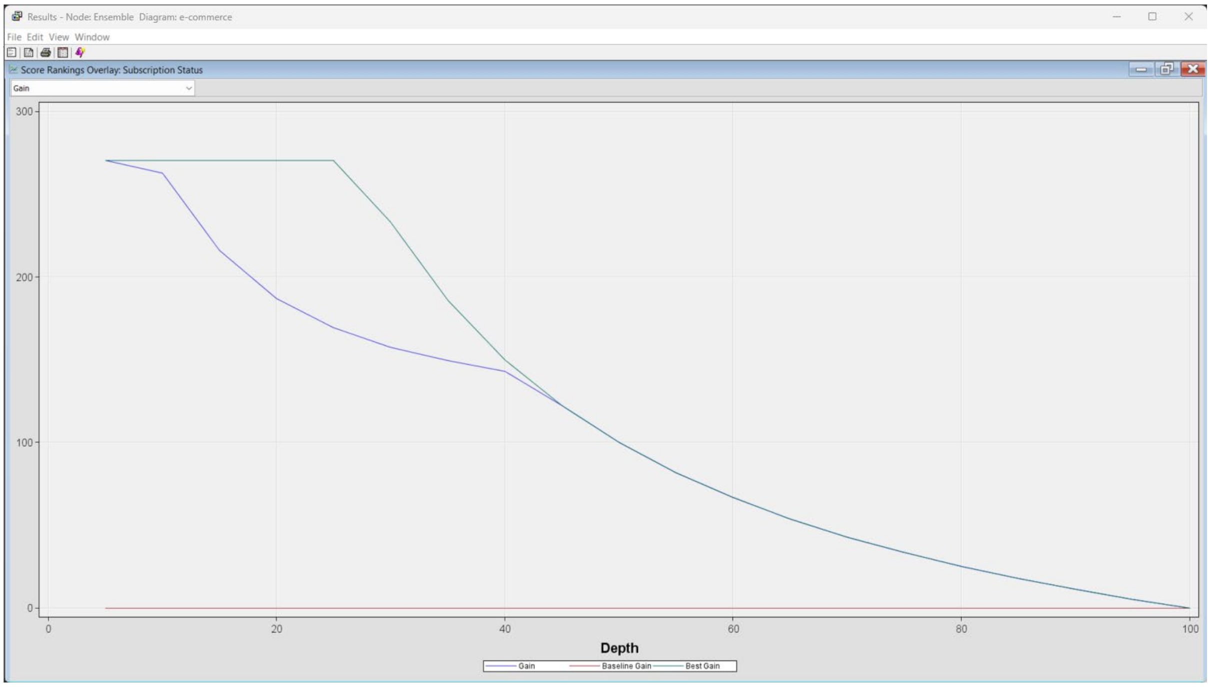
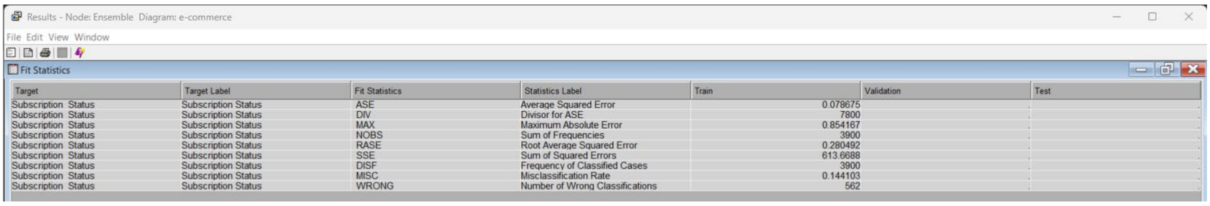


Figure 3. 6: Gain Results



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Subscription Status	Subscription Status	ASE	Average Squared Error	0.078675	.	.
Subscription Status	Subscription Status	DIV	Divisor for ASE	7800	.	.
Subscription Status	Subscription Status	MAX	Maximum Absolute Error	0.854167	.	.
Subscription Status	Subscription Status	NOBS	Sum of Frequencies	3900	.	.
Subscription Status	Subscription Status	RASE	Root Average Squared Error	0.280492	.	.
Subscription Status	Subscription Status	SSE	Sum of Squared Errors	613.6688	.	.
Subscription Status	Subscription Status	DISF	Frequency of Classified Cases	3900	.	.
Subscription Status	Subscription Status	MISC	Misclassification Rate	0.144103	.	.
Subscription Status	Subscription Status	WRONG	Number of Wrong Classifications	562	.	.

Based on the Fit Statistics, it can be summarised that the ensemble model seems to have a reasonably good fit with an Average Squared Error (ASE) of about 0.08, and a misclassification rate of about 14.41%.

The RASE suggests that the model's predicted probabilities are, on average, about 0.28 away from the actual binary outcomes, which could be considered moderate performance. The number of wrong classifications (562 out of 3900) provides a raw count of errors, which can be useful for assessing the model's performance in more tangible term.