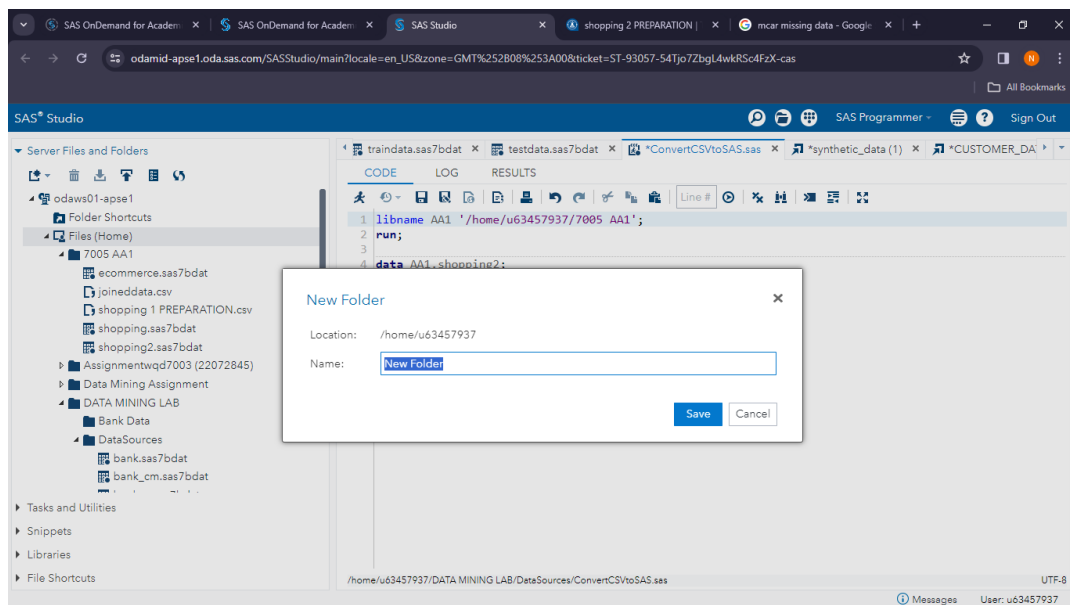


WQD7005 AA1

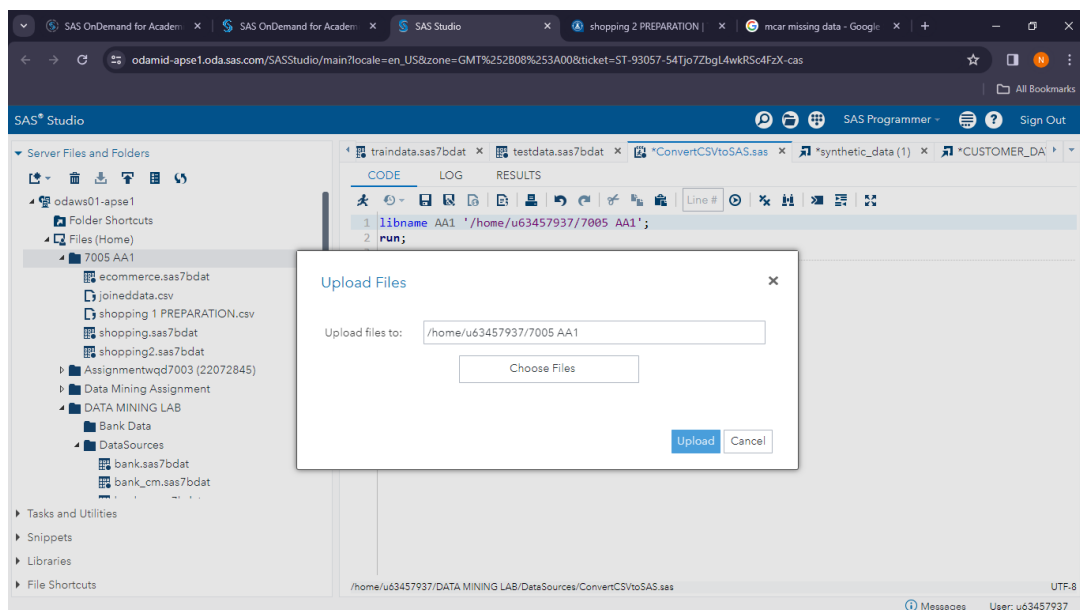
Nur Aisyah Yusof (22072845)

SAS e-Miner is used for data Preprocessing and data modelling to handle missing value, identify variable and data modelling.

- 1) First, to use the clean dataset into SAS EM, the dataset must be uploaded into SAS Studio to transform the csv format into SAS7bdat and to create a library that linked into SAS EM.



New folder is created.



Upload the dataset into folder.

```

1 /* Generated Code (IMPORT) */
2 /* Source File: shopping 1 PREPARATION.csv */
3 /* Source Path: /home/u63457937/7005 AA1 */
4 /* Code generated on: 1/8/24, 7:01 PM */
5
6 %web_drop_table(WORK.IMPORT);
7
8
9 FILENAME REFFILE '/home/u63457937/7005 AA1/shopping 1 PREPARATION.csv';
10
11 PROC IMPORT DATAFILE=REFFILE
12     DBMS=CSV
13     OUT=WORK.IMPORT;
14     GETNAMES=YES;
15 RUN;
16
17 PROC CONTENTS DATA=WORK.IMPORT; RUN;
18
19
20 %web_open_table(WORK.IMPORT);

```

Line 20, Column 30 UTF-8
Messages User: u63457937

Open the csv files and run the code to read the files.

Table: WORK.IMPORT View: Column names Filter: (none)

Columns: Select all, Customer_ID, Age, Gender, Location, Last_Purchase_Date, Item_Purchased, Categories

Property Value

Label

Name

Length

Type

Total rows: 3888 Total columns: 20 Rows 1-100

Customer_ID	Age	Gender	Location	Last_Purchase_Date	Item_Purchased	Categories
1	55	Male	Kentucky	16/06/2022		B
2	19	Male	Maine	15/08/2022		S
3	50	Male	Massachusetts	13/02/2022		J
4	21	Male	Rhode Island	21/09/2022		S
5	45	Male	Oregon	12/08/2023		B
6	46	Male	Wyoming	23/07/2022		S
7	63	Male	Montana	26/07/2022		S
8	27	Male	Louisiana	11/05/2022		S
9	26	Male	West Virginia	23/09/2023		C
10	57	Male	Missouri	20/12/2023		H
11	53	Male	Arkansas	14/02/2022		S
12	30	Male	Hawaii	04/05/2022		S

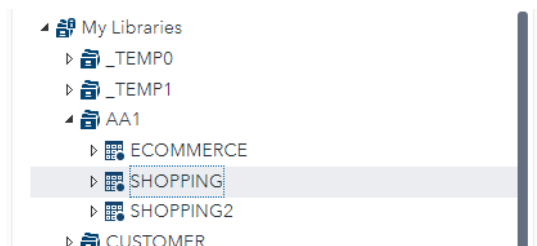
The output table shown as above.

```

1 libname AA1 '/home/u63457937/7005 AA1';
2 run;
3
4 data AA1.shopping2;
5     set import;
6 run;
7

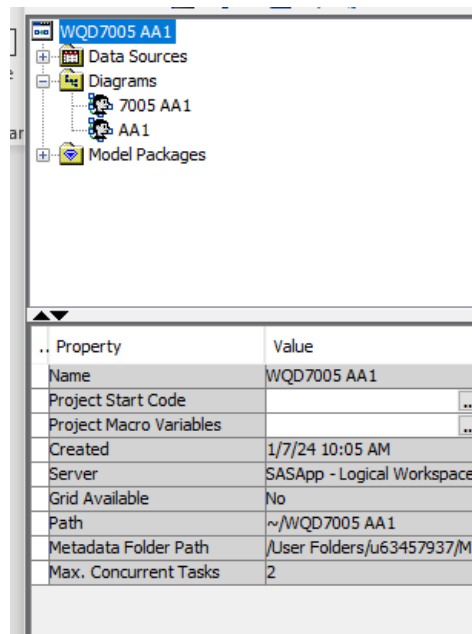
```

Then, wrote this code to create a library that linked with SAS EM and import the csv files into the library.



The metadata is created in SAS7bdat format.

- 2) SAS Studio can also facilitate data exploration to see the correlations between variables and further analysis. However, this step was skipped as it was not required in the assessment question.
- 3) Next, open SAS EM and a new project was created as WQD7005 AA1. Then, click on the project start code.



- 4) Wrote and ran this code to link the created library into SAS EM.

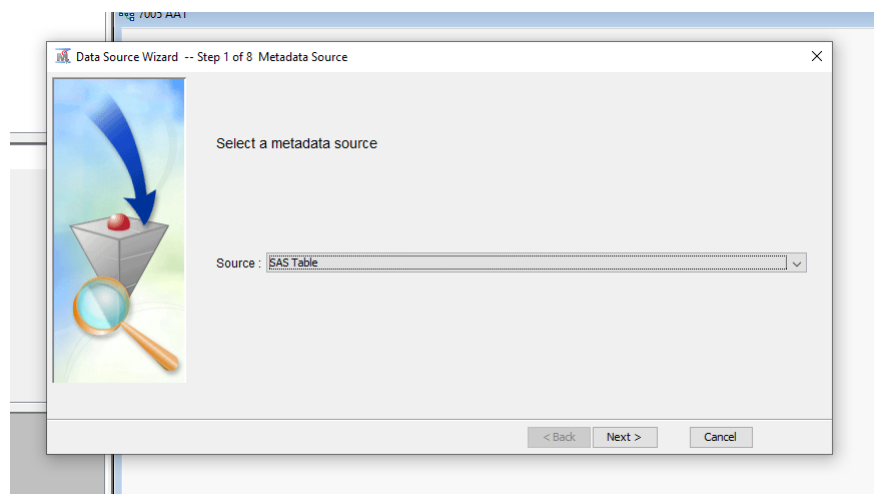
environment of the SAS code submitted by Enterprise Miner.

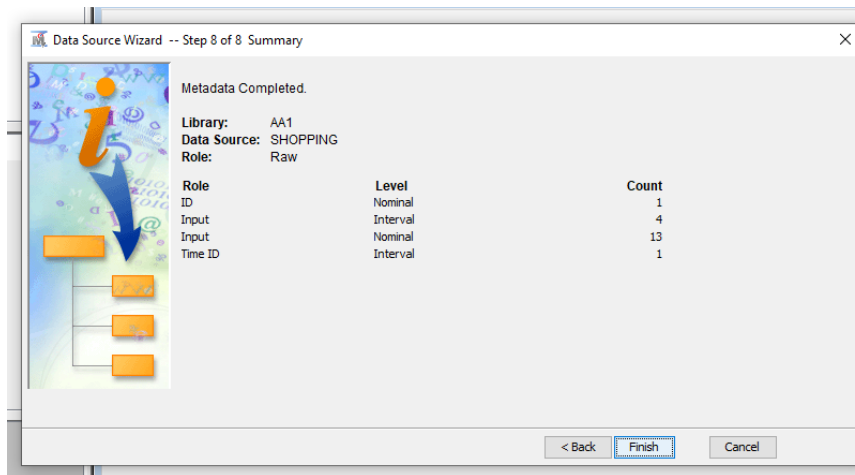
```

1 libname AA1 "/home/u63457937/7005 AA1";
2 run;
3

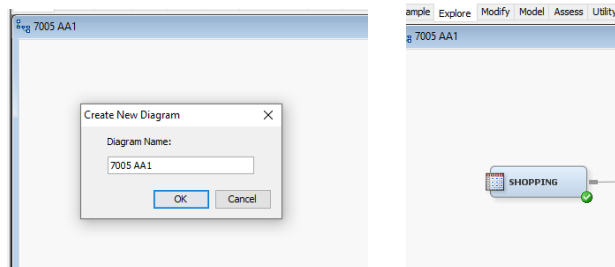
```

- 5) Then, create a new Data Source and export the files that had uploaded before.





6) Then, create the new diagram and drag the data source into the diagram.



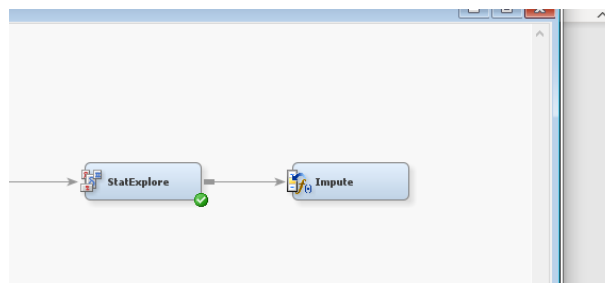
7) Choose subscription status as the target variable.

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic

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	ID	Nominal	No		No	-	-
Discount_Applied	Input	Nominal	No		No	-	-
Frequency_of_P	Input	Nominal	No		No	-	-
Gender	Input	Nominal	No		No	-	-
Item_Purchased	Input	Nominal	No		No	-	-
Last_Purchase_Time	ID	Interval	No		No	-	-
Location	Input	Nominal	No		No	-	-
Payment_Metho	Input	Nominal	No		No	-	-
Previous_Purch	Input	Interval	No		No	-	-
Promo_Code_Us	Input	Nominal	No		No	-	-
Purchase_Amou	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_Sta	Target	Nominal	No		No	-	-

8) After that, node explore and impute were put into diagram to handle any missing data.



9) After running the explore node, it is found that review rating have missing value.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	44.2074	15.28525	3110	0	18	44	70	-0.01191	-1.20917
Previous_Purchases	INPUT	25.49678	14.4425	3110	0	1	25	50	-0.0057	-1.18125
Purchase_Amount_USD	INPUT	60.01093	23.71652	3110	0	20	60	100	0.001704	-1.22917
Review_Rating	TARGET	3.755645	0.714893	3100	10	2.5	3.8	5	-0.00545	-1.17769

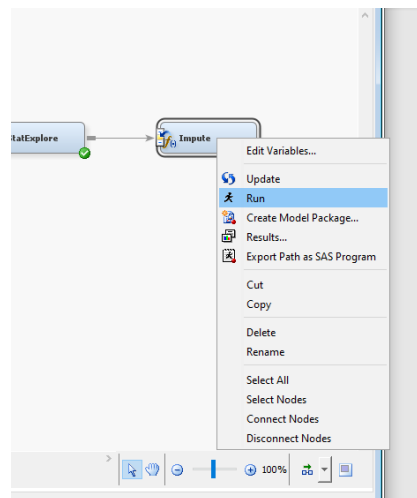
10) The missing value were then impute by the mean value of review rating and run. The result of impute value is 3.75.

Variables - Imput

(none) ☐ not Equal to ☐

Columns: ☐ Label ☐ Mining ☐ Bar

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_Applied	Default	Default	Default	Input	Nominal
Frequency_of_Purchase	Default	Default	Default	Input	Nominal
Gender	Default	Default	Default	Input	Nominal
Item_Purchased	Default	Default	Default	Input	Nominal
Location	Default	Default	Default	Input	Nominal
Payment_Method	Default	Default	Default	Input	Nominal
Previous_Purchase	Default	Default	Default	Input	Interval
Promo_Code_Used	Default	Default	Default	Input	Nominal
Purchase_Amount	Default	Default	Default	Input	Interval
Review_Rating	Default	Mean	Default	Input	Interval
Season	Default	Default	Default	Target	Nominal
Shipping_Type	Default	Default	Default	Input	Nominal
Size	Default	Default	Default	Input	Nominal
Subscription_Status	Default	Default	Default	Input	Nominal

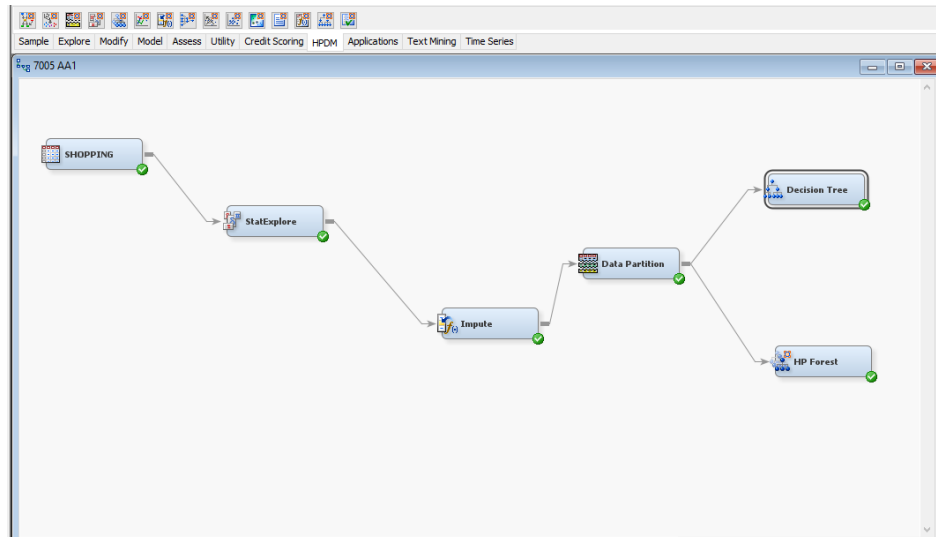


Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Review_Rating	MEAN	IMP_Review_Rating	3.752806	INPUT	INTERVAL		10

As for modelling, the decision tree and random forest is used.

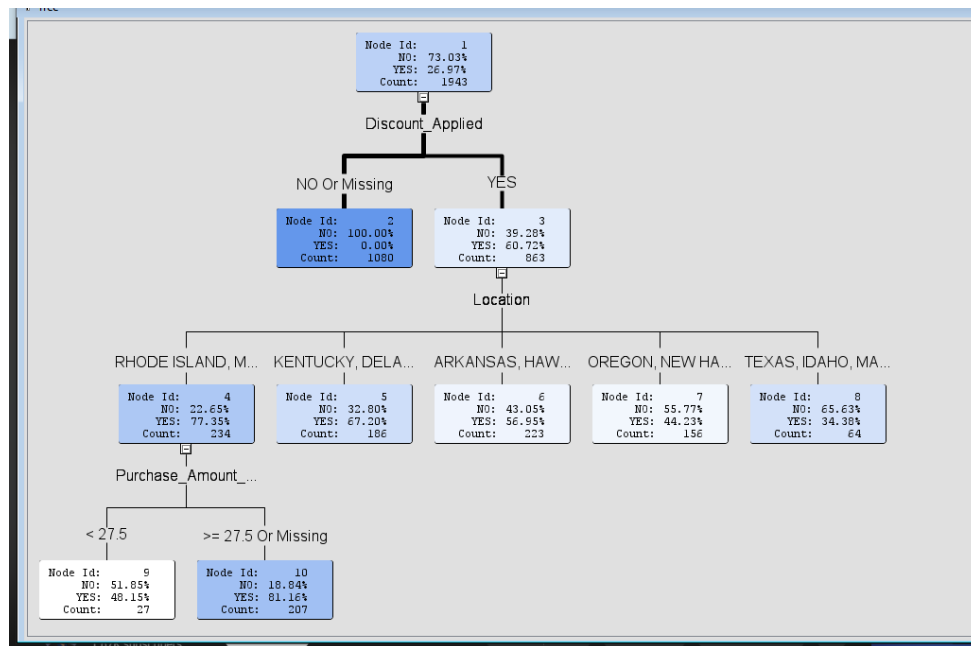
- 1) The node of Data partition, Decision Tree and Random Forest were dragged into diagram and linked with impute node as below.



- 2) The data were split into 50 percent for both train and test data.

Property	Value
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	0.0
Test	50.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	50.0
Create Time	1/7/24 7:50 AM
Run ID	ba04fa5e-4e62-df4d-b688-
Lost Error	
General	

3) Decision Tree node was then run and got this output.



4) The misclassification rate for train is 0.15 and test is 0.17, while the Average Square Error for train is 0.096 and test is 0.104.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Subscription_Status	_NOBS	Sum of Frequencies		1943		1945
Subscription_Status	_MISC	Misclassification Rate		0.1544		0.17635
Subscription_Status	_MAX	Maximum Absolute Error		0.811594		0.811594
Subscription_Status	_SSE	Sum of Squared Errors		373.9569		407.1852
Subscription_Status	_ASE	Average Squared Error		0.096232		0.104675
Subscription_Status	_RASE	Root Average Squared Error		0.310213		0.323535
Subscription_Status	_DIV	Divisor for ASE		3886		3890
Subscription_Status	_DFT	Total Degrees of Freedom		1943		

5) Here is the classification table that shows True Positive as 420, False Positive as 196, True negative as 1223 and False Negative as 104.

Event Classification Table

Data Role=TRAIN Target=Subscription_Status Target Label=' '

False Negative	True Negative	False Positive	True Positive
104	1223	196	420

- 6) The feature importance results from decision tree are discount applied, location and purchase amount USD.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
Discount_Applied		1	1.0000
Location		1	0.3032
Purchase_Amount_USD		1	0.1213

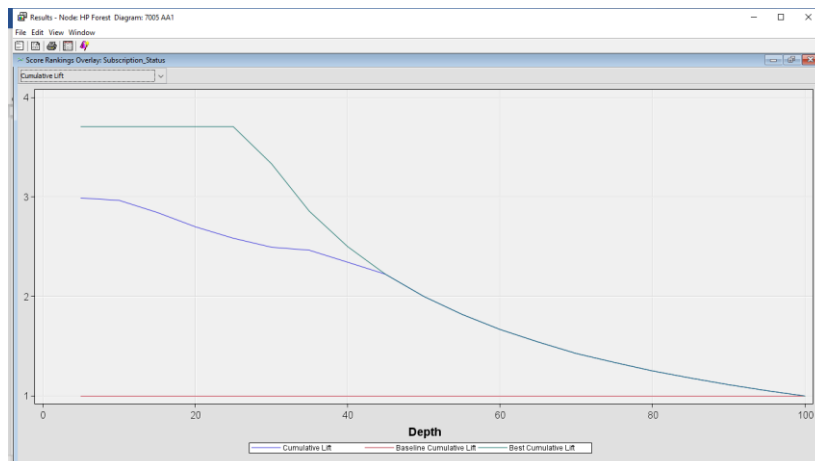
- 7) Then Ensemble method; Random Forest is used for the modelling. The node for Random Forest was then run to get the analysis results.



- 8) From the results, we know the misclassification rate for train is 0.17 and test is 0.14. The Average Square Error for train is 0.105 and test is 0.097.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Subscription...	_ASE_	Average Sq...		0.105435		0.096779
Subscription...	_DIV_	Divisor for A...		3886		3890
Subscription...	_MAX_	Maximum A...		0.604997		0.611067
Subscription...	_NOBS_	Sum of Fre...		1943		1945
Subscription...	_RASE_	Root Avera...		0.324707		0.311093
Subscription...	_SSE_	Sum of Squ...		409.7185		376.4709
Subscription...	_DISF_	Frequency ...		1943		1945
Subscription...	_MISC_	Misclassific...		0.174472		0.146015
Subscription...	_WRONG_	Number of ...		339		284

9) The cumulative lift shows the performance almost fitted with the best cumulative lift.



10) The results also show the feature importance of the analysis.

Variable Importance						
Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Label
Promo_Co...	55	0.086732	0.173463	0.08737	0.17493	
Discount_A...	43	0.065630	0.131259	0.06554	0.13128	
Gender	25	0.016143	0.032285	0.01598	0.03146	
Age	9	0.000341	0.000682	-0.00057	-0.00024	
Location	9	0.000871	0.001743	-0.00085	0.00060	
IMP_Revie...	7	0.000113	0.000226	-0.00009	0.00003	Imputed Re...
Previous_P...	7	0.000129	0.000258	-0.00023	-0.00010	
Shipping_T...	2	0.000082	0.000164	-0.00008	-0.00001	
Category	1	0.000030	0.000060	-0.00007	-0.00002	
Item_Purch...	1	0.000085	0.000171	-0.00012	-0.00006	
Purchase_...	1	0.000022	0.000044	-0.00004	-0.00002	
Size	1	0.000035	0.000071	-0.00005	-0.00002	
Color	0	0.000000	0.000000	0.00000	0.00000	
Frequency_...	0	0.000000	0.000000	0.00000	0.00000	
Payment_M...	0	0.000000	0.000000	0.00000	0.00000	
Season	0	0.000000	0.000000	0.00000	0.00000	

11) The model was then compare and found that random forest is better than decision tree.

