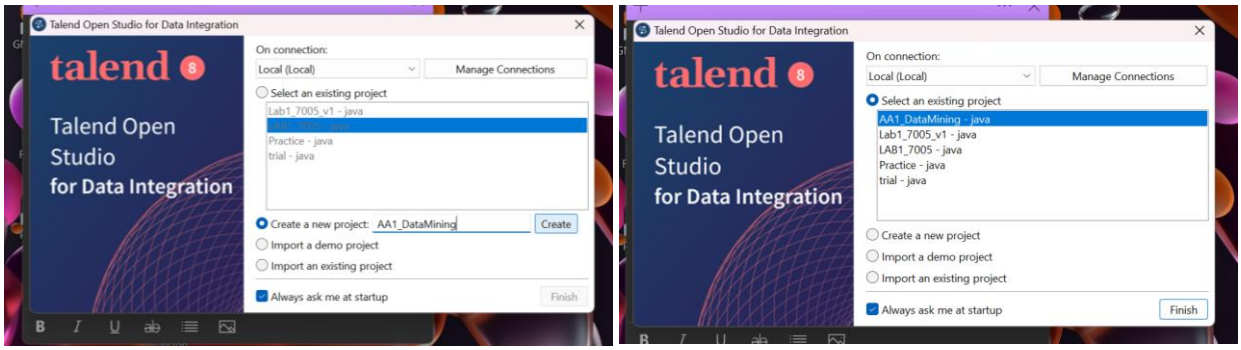
WQD7005

Alternative Assessment 1 – Case Study

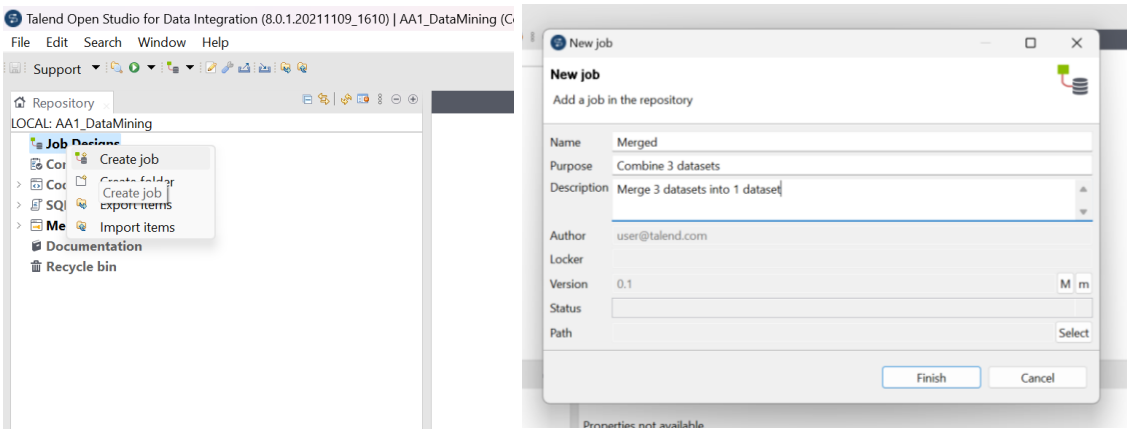
Farzana Syakira binti Bahari (22058163)

MERGING THREE DATASETS INTO ONE DATASET USING TALEND OPEN STUDIO FOR DATA INTEGRATION

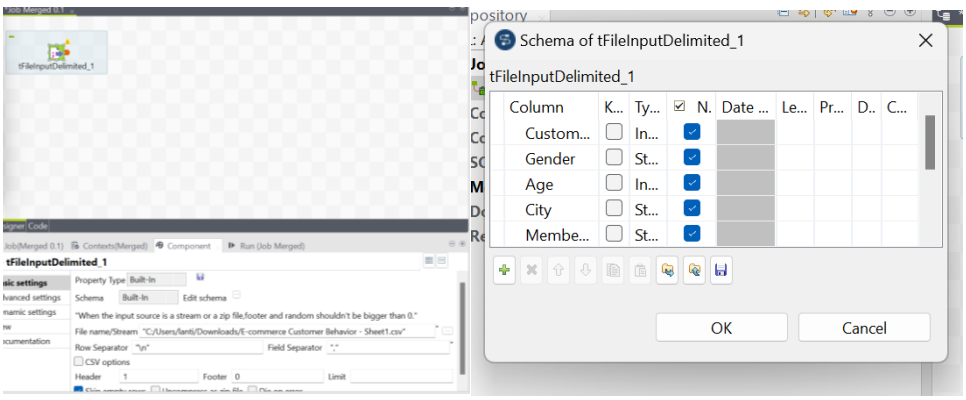
a. Create a new project in Talend Open Studio for Data Integration

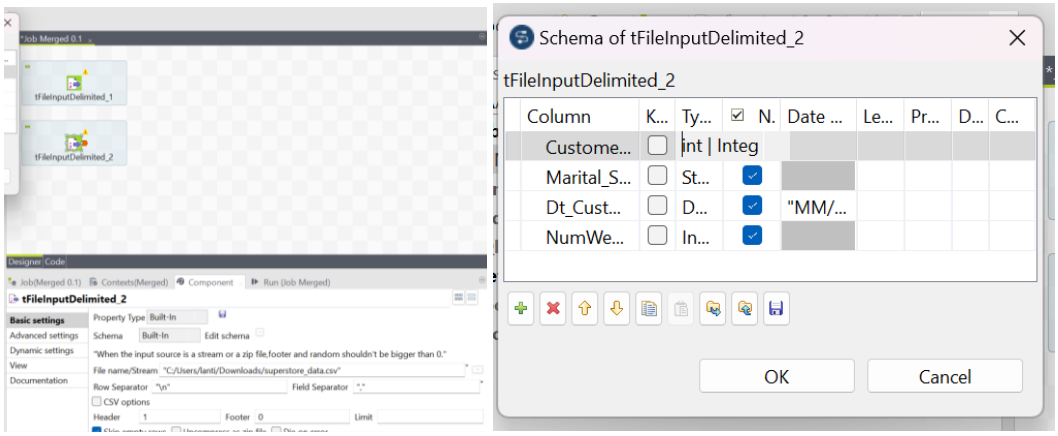


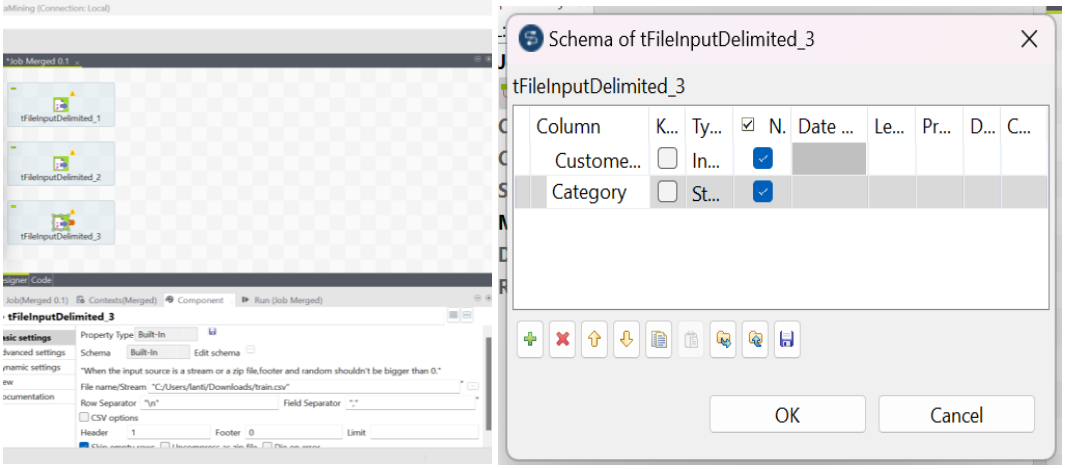
b. Create new job



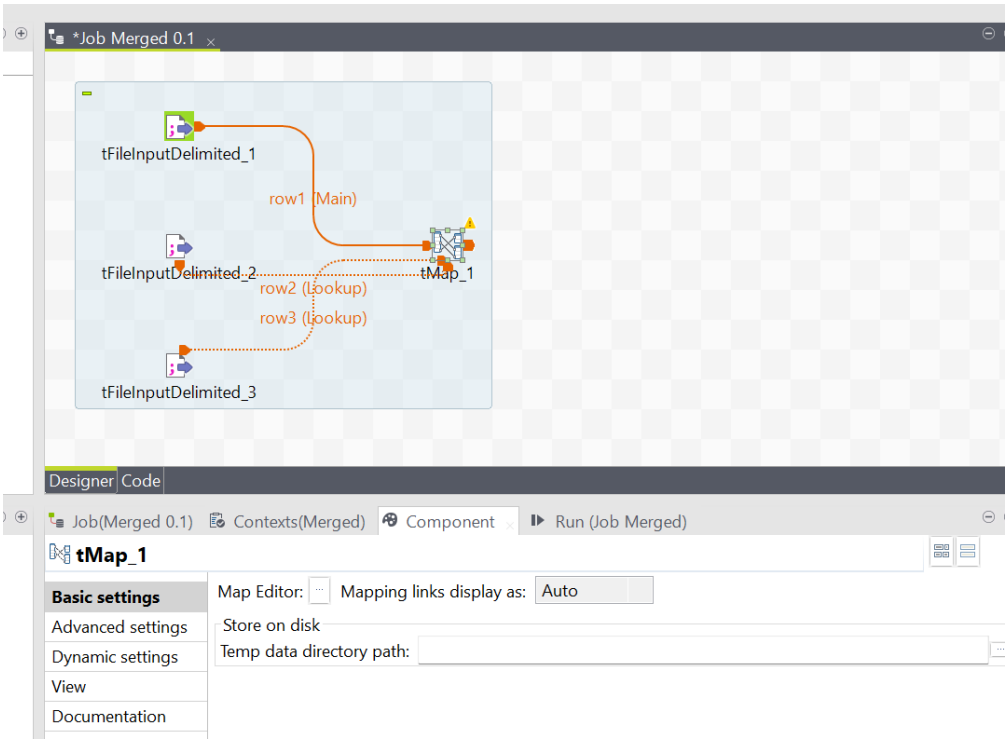
c. Ingesting the data and setting up the schema by dragging the “tFileInputDelimited” into the job space



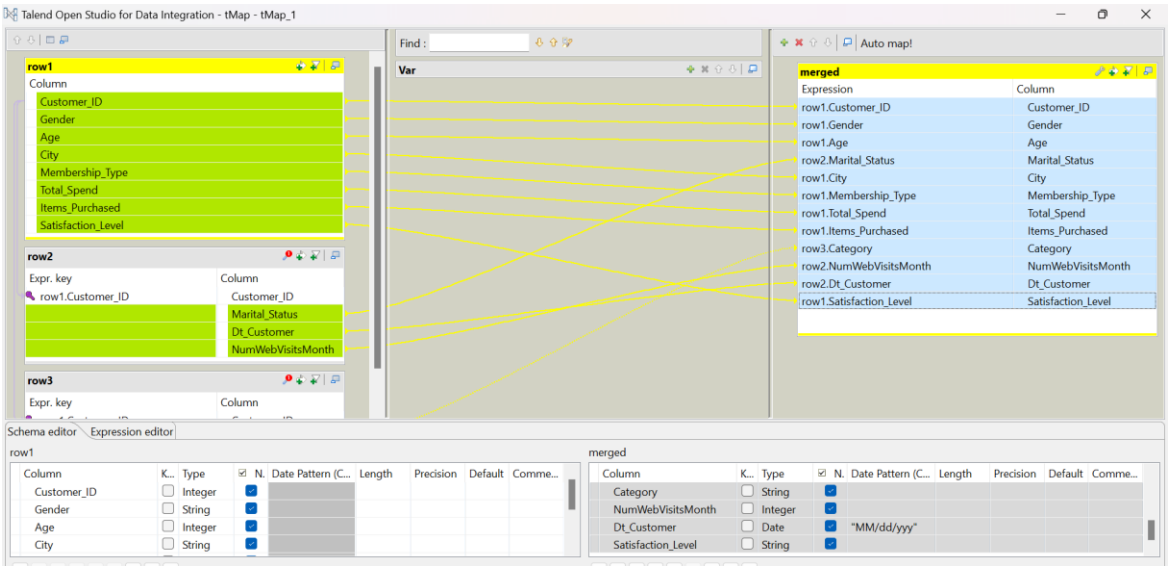




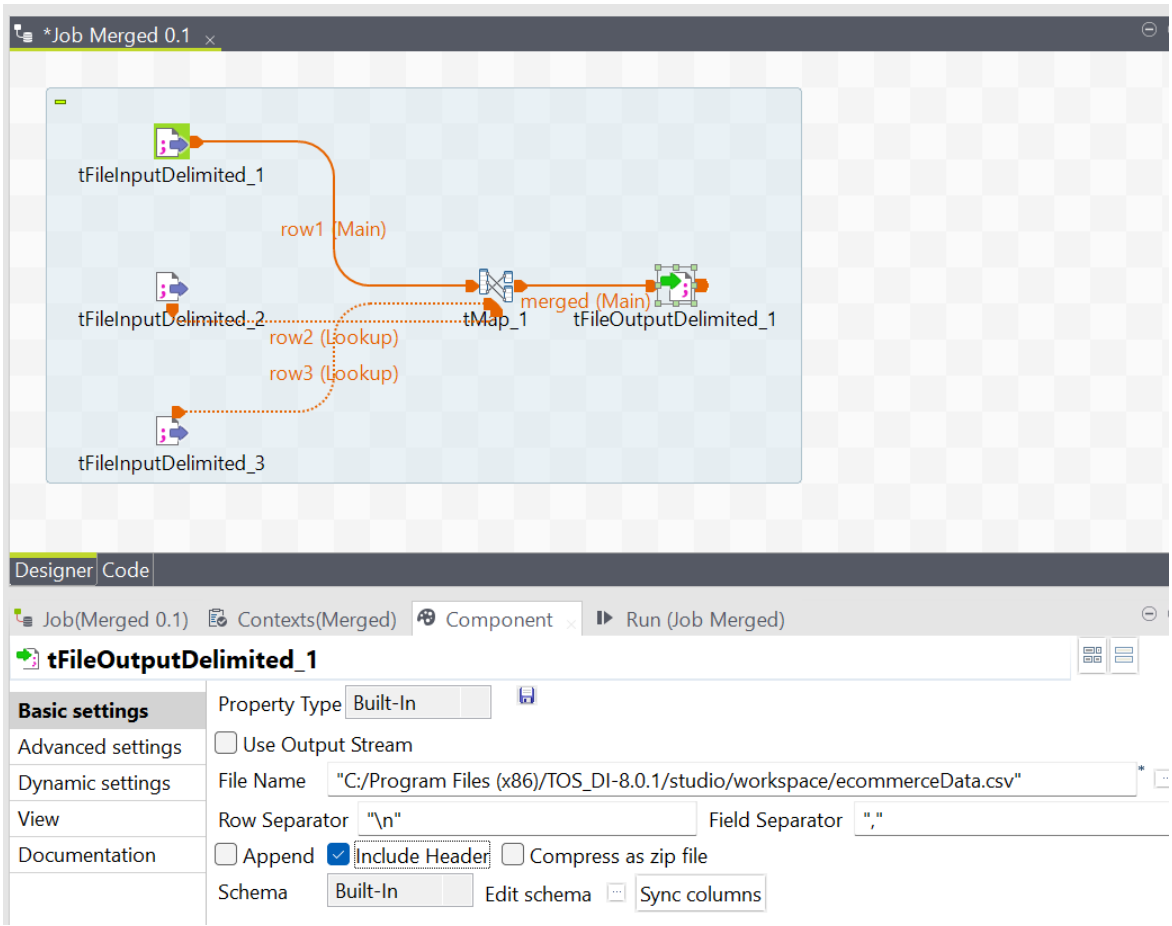
d. Combine all the datasets using “tMap_1” node by dragging it into the job space

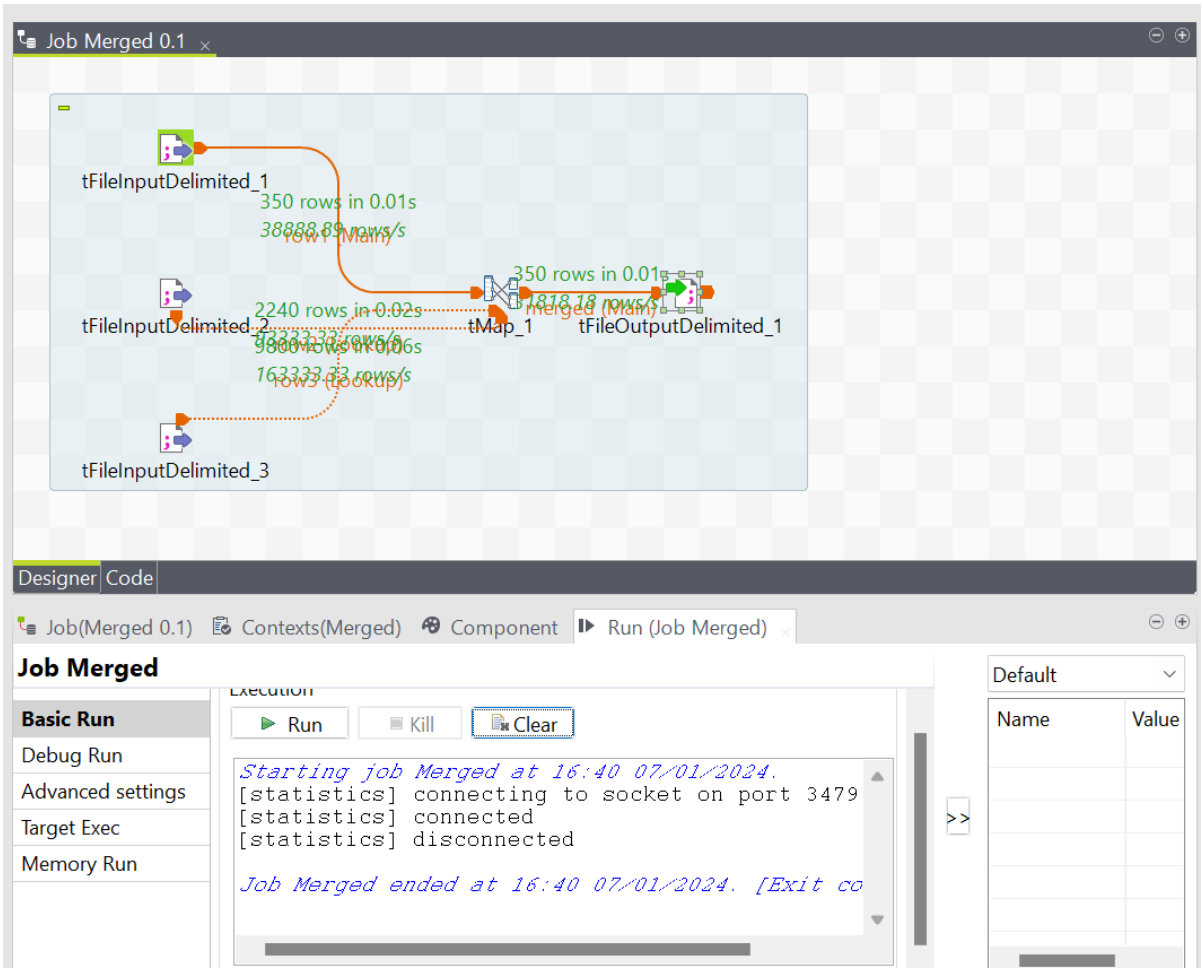


e. Connect all of the datasets using “Customer_ID” column then map all the datasets into one new dataset.



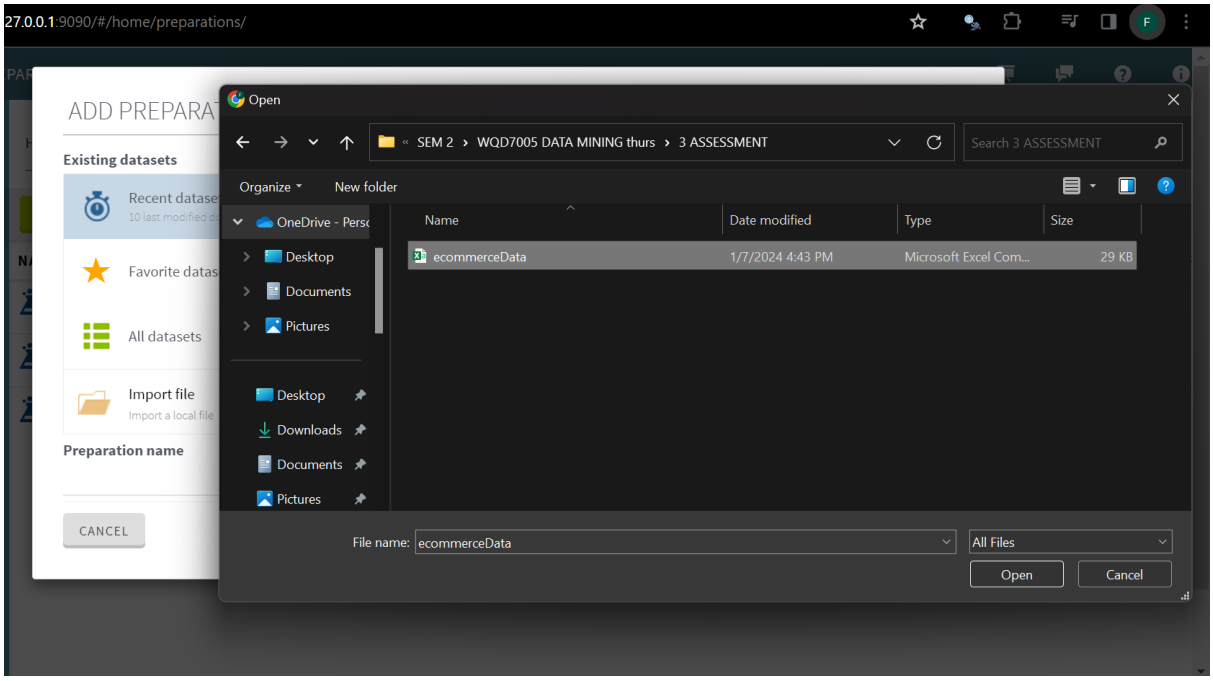
f. Export the data using “tFileOutputDelimited_1” node




Job Merged 0.1
tFileInputDelimited_1
350 rows in 0.01s
38888.89 rows/s
row1 (Main)
tFileInputDelimited_2
2240 rows in 0.02s
163333.33 rows/s
row3 (Lookup)
tMap_1
350 rows in 0.01s
1818.18 rows/s
merged (Main)
tFileOutputDelimited_1
tFileInputDelimited_3
Designer
Code
Job(Merged 0.1)
Contexts(Merged)
Component
Run (Job Merged)
Job Merged
Basic Run
Debug Run
Advanced settings
Target Exec
Memory Run
Run
Kill
Clear
Starting job Merged at 16:40 07/01/2024.
[statistics] connecting to socket on port 3479
[statistics] connected
[statistics] disconnected
Job Merged ended at 16:40 07/01/2024. [Exit co
Default
Name
Value

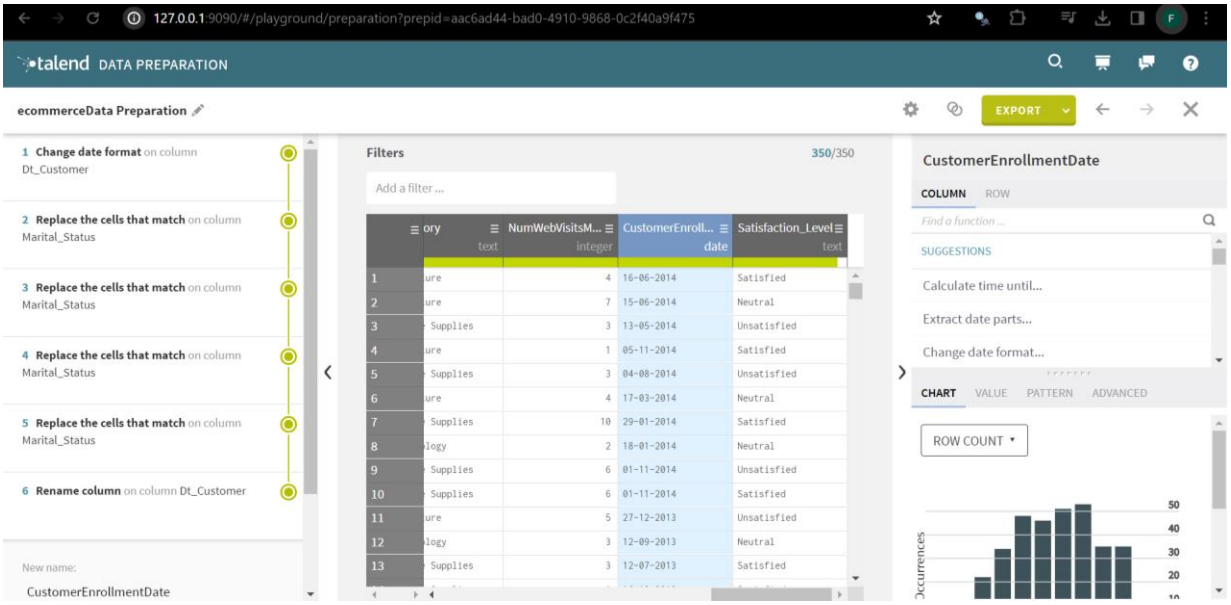
STANDARDIZING DAN CLEANING THE DATA USING TALEND DATA PREPARATION

a. Ingesting the data

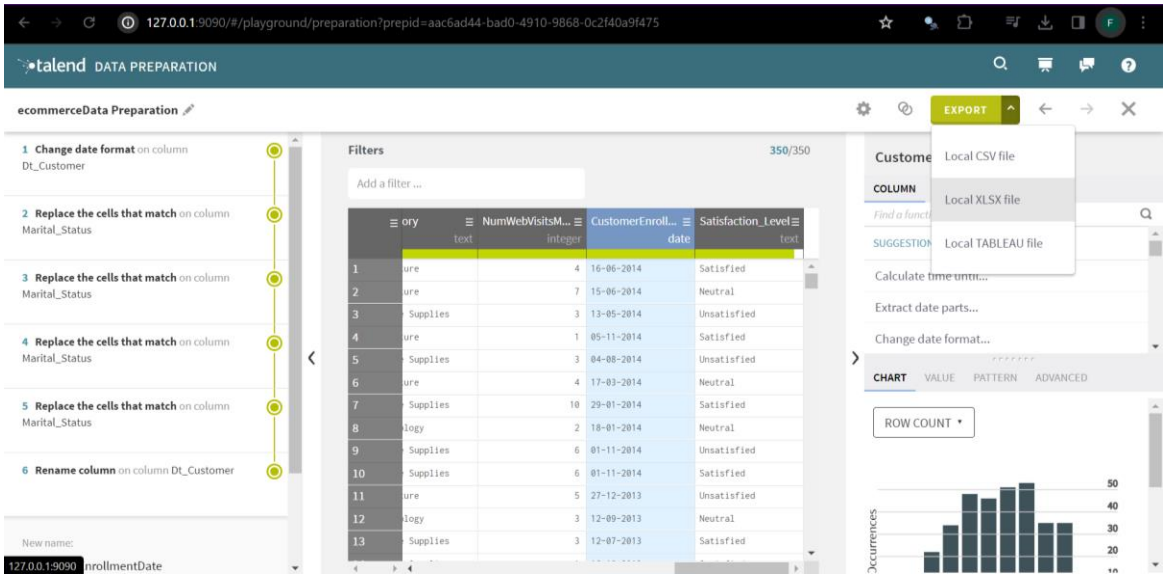


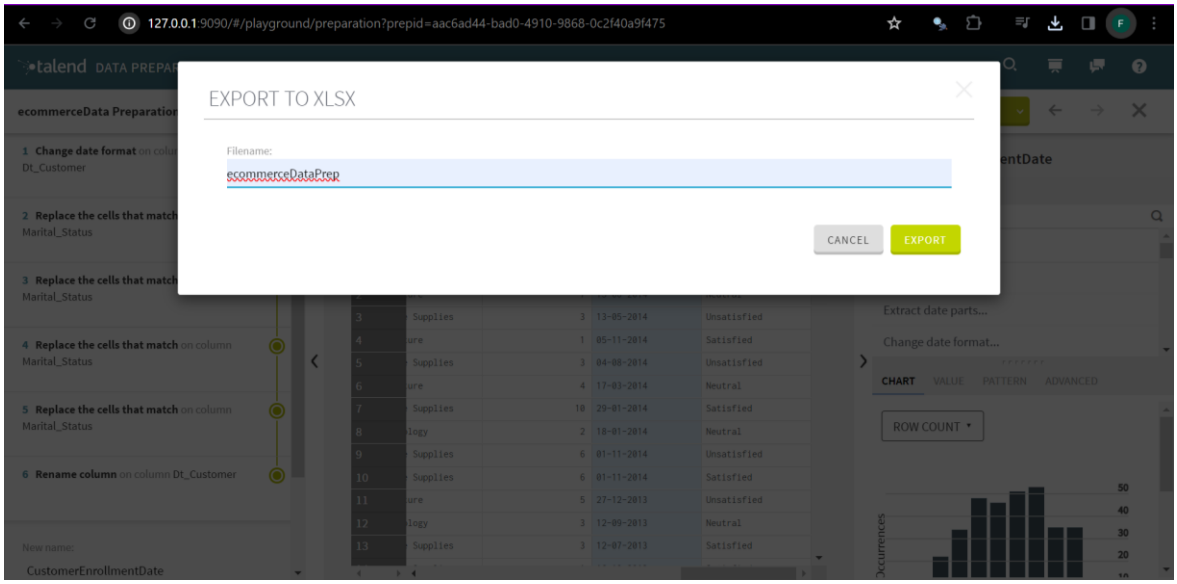
b. Pre-process the data. This includes the standardization of the format of the date and values in the Marital_Status column.

- All values in “Dt_Customer” are changed from “MM/dd/yyyy” to “dd-MM-yyyy”.
- Values in Marital_Status such as “YOLO” and “Alone” are replaced by “Single”.
- Values in Marital_Status such as “Widow” is replaced by “Divorced”.
- Values in Marital_Status such as “Together” is replaced by “Married”.
- The column name “Dt_Customer” is renamed to “CustomerEnrollmentDate”



c. Exporting the file as ecommerceDataPrep.xlsx file

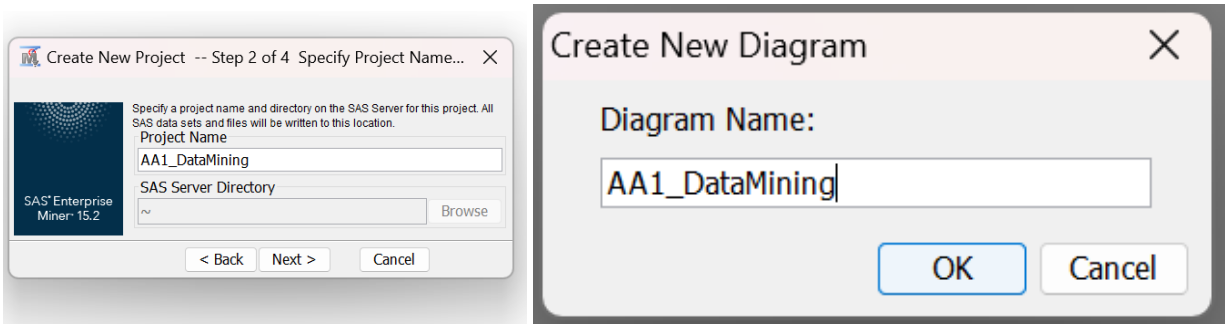




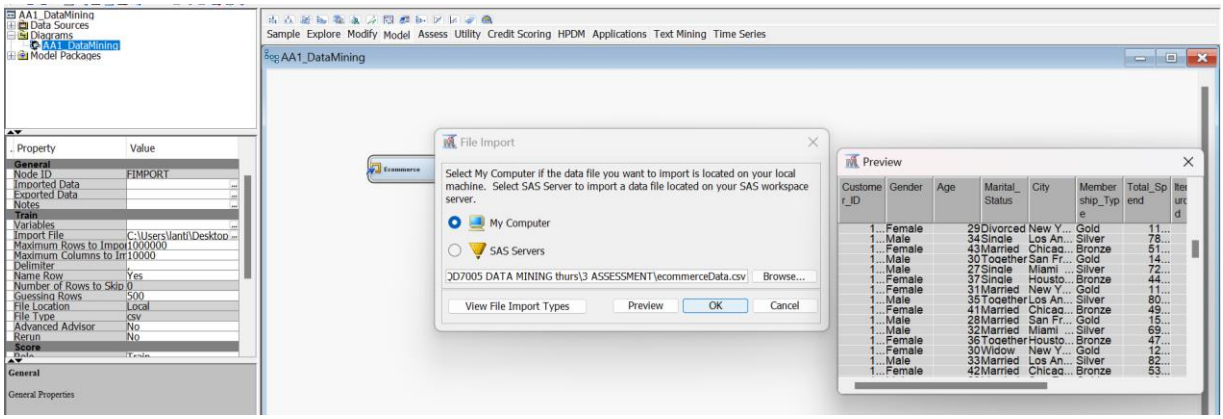
ANALYSING THE DATA USING SAS EM

Data Import and pre-processing

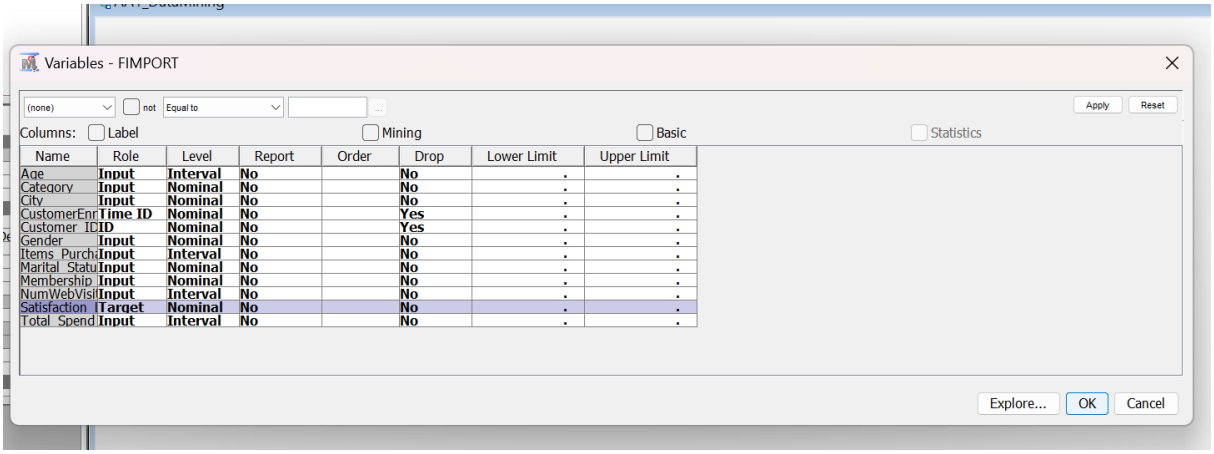
a. Creating a new project and a new diagram



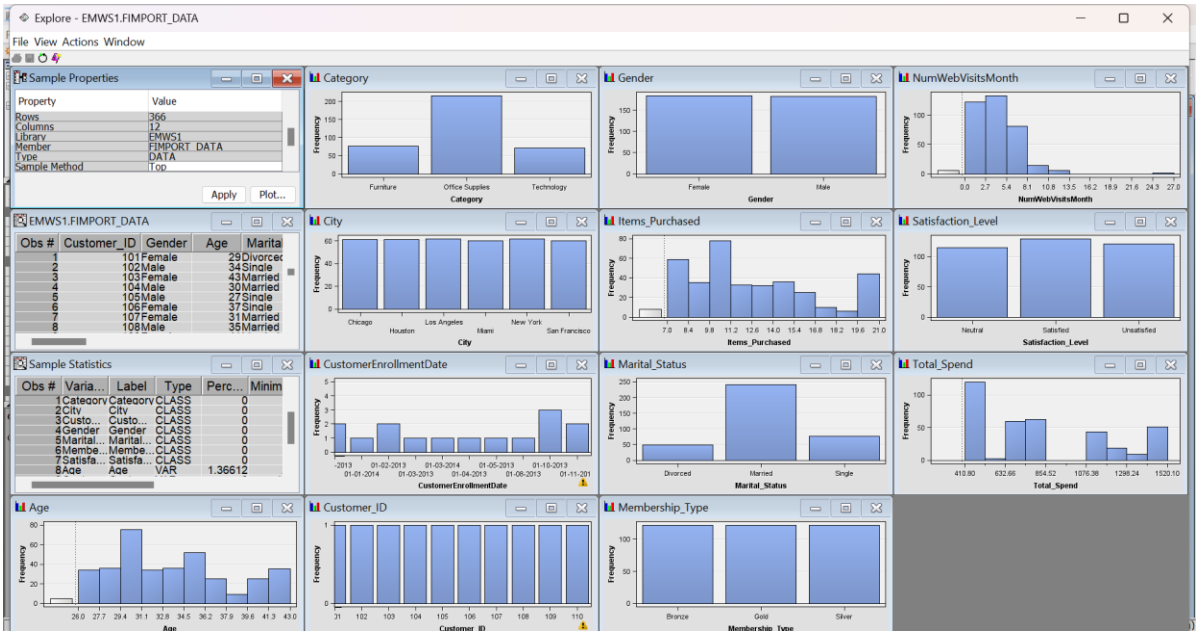
b. Ingesting the data and reviewing it to ensure that the correct data is ingested



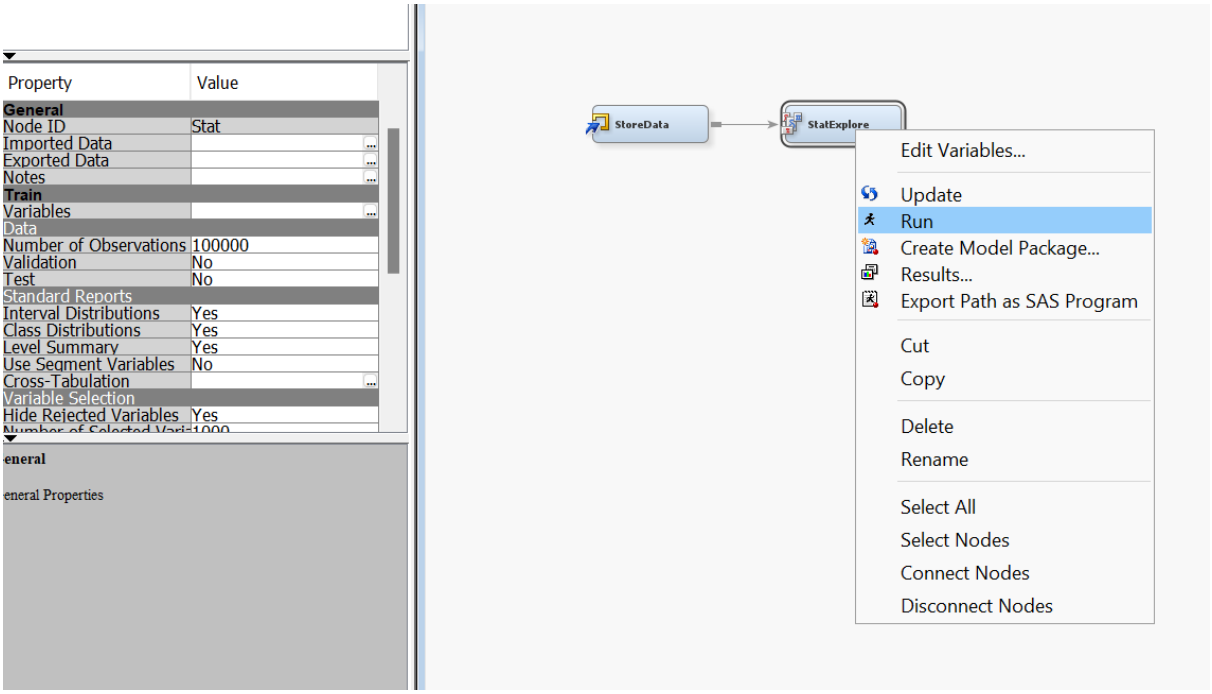
c. Edit the variables to assign the roles for each of the variables. The target variable will be the Satisfaction_Level. Customer_ID is assigned as the ID and the CustomerEnrollmentDate is assigned as Time ID. The data is explored to see every variable's distributions.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Category	Input	Nominal	No		No	.	.
City	Input	Nominal	No		No	.	.
CustomerEnr	Time ID	Nominal	No		Yes	.	.
Customer_ID	ID	Nominal	No		Yes	.	.
Gender	Input	Nominal	No		No	.	.
Items_Purch	Input	Interval	No		No	.	.
Marital_Statu	Input	Nominal	No		No	.	.
Membership_	Input	Nominal	No		No	.	.
NumWebVisi	Input	Interval	No		No	.	.
Satisfaction_	Target	Nominal	No		No	.	.
Total_Spend	Input	Interval	No		No	.	.



d. The StatExplore is attached to the diagram to see if there are missing values in the data. The categorical variables do not have any missing values whereas the interval variables such as Age, Items_purchased and NumWebVisitMonth have missing values each.



Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Category	INPUT	3	0	Office Supplies	59.29	Furniture	21.04
TRAIN	City	INPUT	6	0	Los Angeles	16.94	New York	16.94
TRAIN	Gender	INPUT	2	0	Female	50.27	Male	49.73
TRAIN	Marital_Status	INPUT	3	0	Married	65.85	Single	21.04
TRAIN	Membership_Type	INPUT	3	0	Bronze	33.33	Gold	33.33
TRAIN	Satisfaction_Level	TARGET	3	0	Satisfied	35.52	Unsatisfied	33.06

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

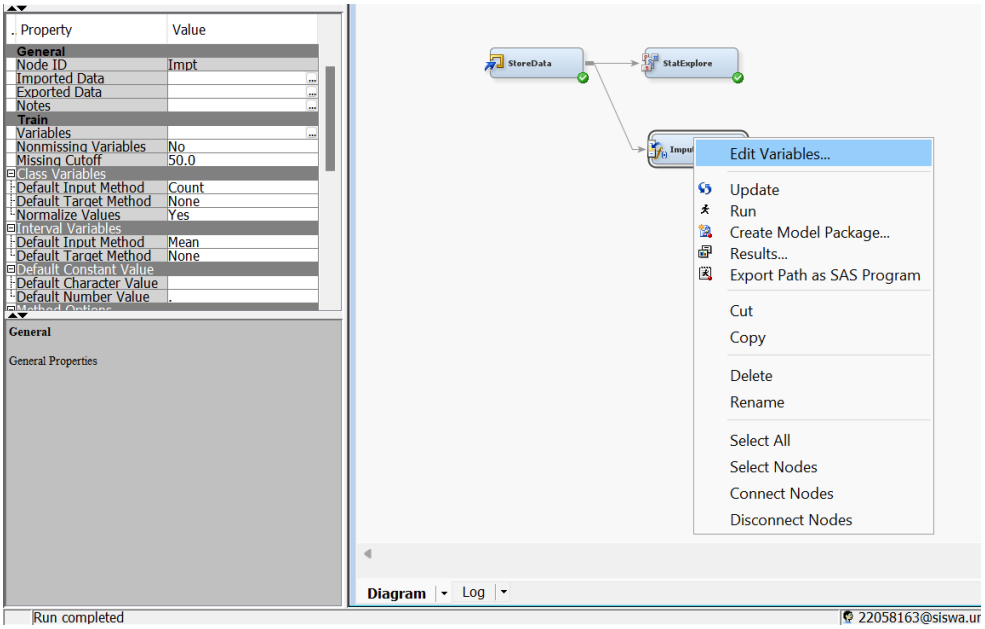
Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Satisfaction_Level	TARGET	Satisfied	130	35.5191
TRAIN	Satisfaction_Level	TARGET	Unsatisfied	121	33.0601
TRAIN	Satisfaction_Level	TARGET	Neutral	115	31.4208

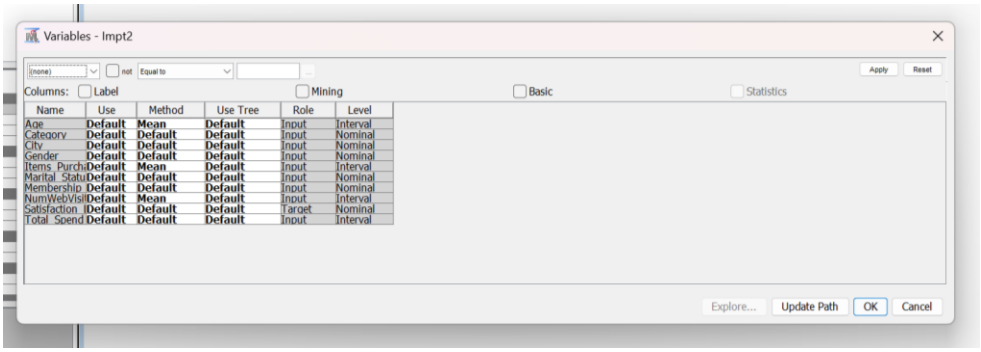
Interval Variable Summary Statistics
(maximum 500 observations printed)

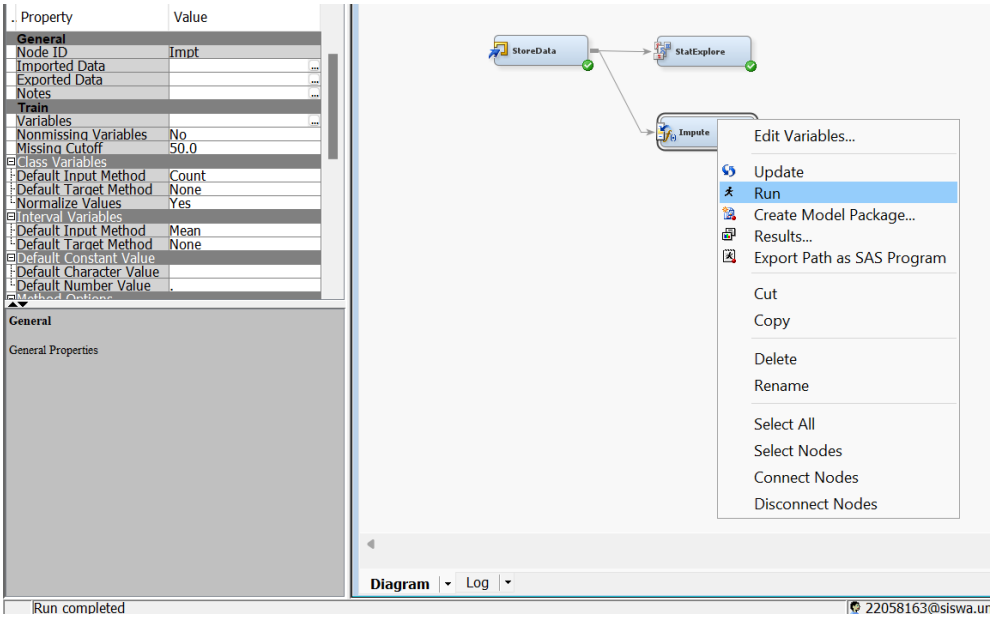
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	33.62881	4.858058	361	5	26	33	43	0.463465	-0.76948
Items_Purchased	INPUT	12.60056	4.154536	358	8	7	12	21	0.653491	-0.61776
NumWebVisitsMonth	INPUT	4.141667	3.027163	360	6	0	4	27	2.410881	13.94905
Total_Spend	INPUT	844.1173	361.6808	366	0	410.8	770.2	1520.1	0.566578	-1.07491

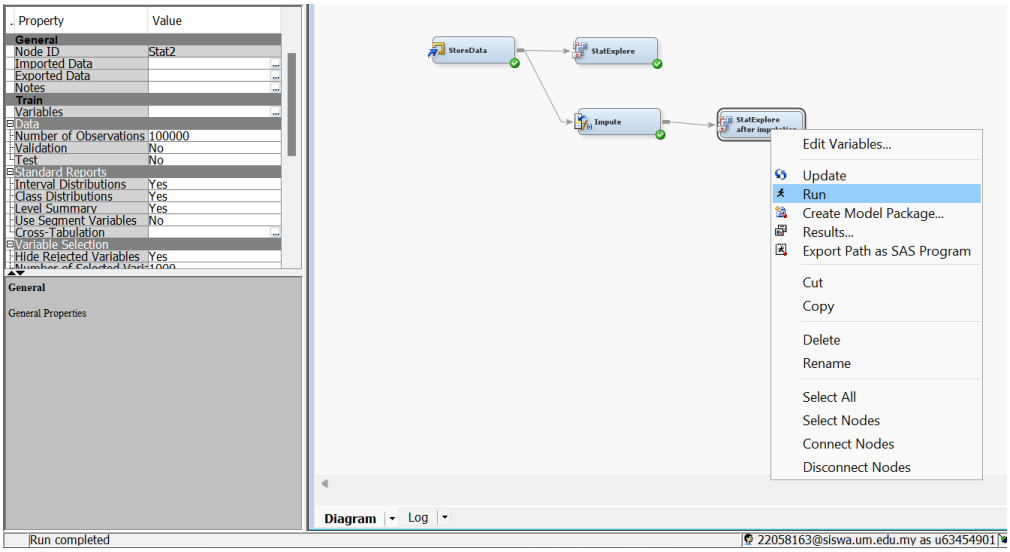
e. Imputing the missing values using Mean







f. Check if the missing values have been imputed using the “StatExplore” node.



(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Category	INPUT	3	0	Office Supplies	59.29	Furniture	21.04
TRAIN	City	INPUT	6	0	Los Angeles	16.94	New York	16.94
TRAIN	Gender	INPUT	2	0	Female	50.27	Male	49.73
TRAIN	Marital_Status	INPUT	3	0	Married	65.85	Single	21.04
TRAIN	Membership_Type	INPUT	3	0	Bronze	33.33	Gold	33.33
TRAIN	Satisfaction_Level	TARGET	3	0	Satisfied	35.52	Unsatisfied	33.06

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	Satisfaction_Level	TARGET	Satisfied	130	35.5191
TRAIN	Satisfaction_Level	TARGET	Unsatisfied	121	33.0601
TRAIN	Satisfaction_Level	TARGET	Neutral	115	31.4208

Interval Variable Summary Statistics
(maximum 500 observations printed)

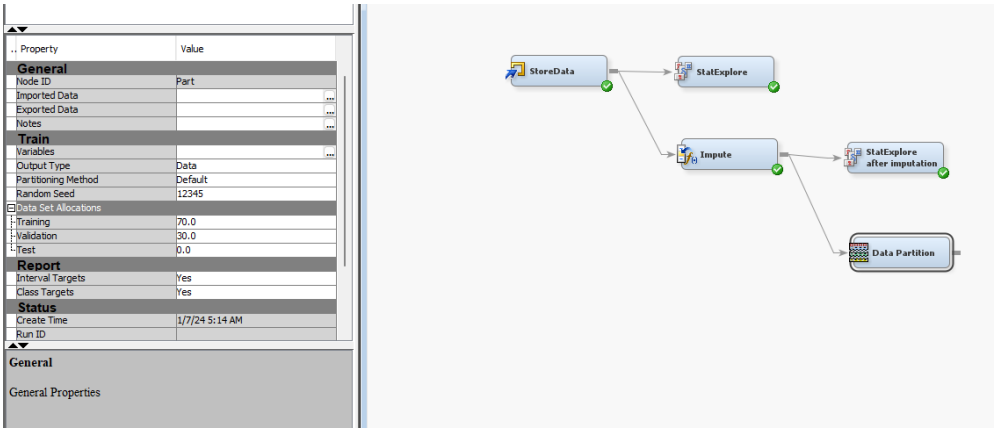
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
IMP_Age	INPUT	33.62881	4.824668	366	0	26	33	43	0.466637	-0.73833
IMP_Items_Purchased	INPUT	12.60056	4.108754	366	0	7	12	21	0.660691	-0.56415
IMP_NumWebVisitsMonth	INPUT	4.141667	3.00218	366	0	0	4	27	2.430721	14.22842
Total_Spend	INPUT	844.1173	361.6808	366	0	410.8	770.2	1520.1	0.566578	-1.07491

Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=Category

g. Attach “Data Partition” node to separate 70% of the data into testing set and the remaining into validation set. Run it and see the results to ensure that each value of the target variable has almost the same frequency.



Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Satisfaction_Level	.	Neutral	115	31.4208	Satisfaction_Level
Satisfaction_Level	.	Satisfied	130	35.5191	Satisfaction_Level
Satisfaction_Level	.	Unsatisfied	121	33.0601	Satisfaction_Level

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Satisfaction_Level	.	Neutral	80	31.3725	Satisfaction_Level
Satisfaction_Level	.	Satisfied	90	35.2941	Satisfaction_Level
Satisfaction_Level	.	Unsatisfied	85	33.3333	Satisfaction_Level

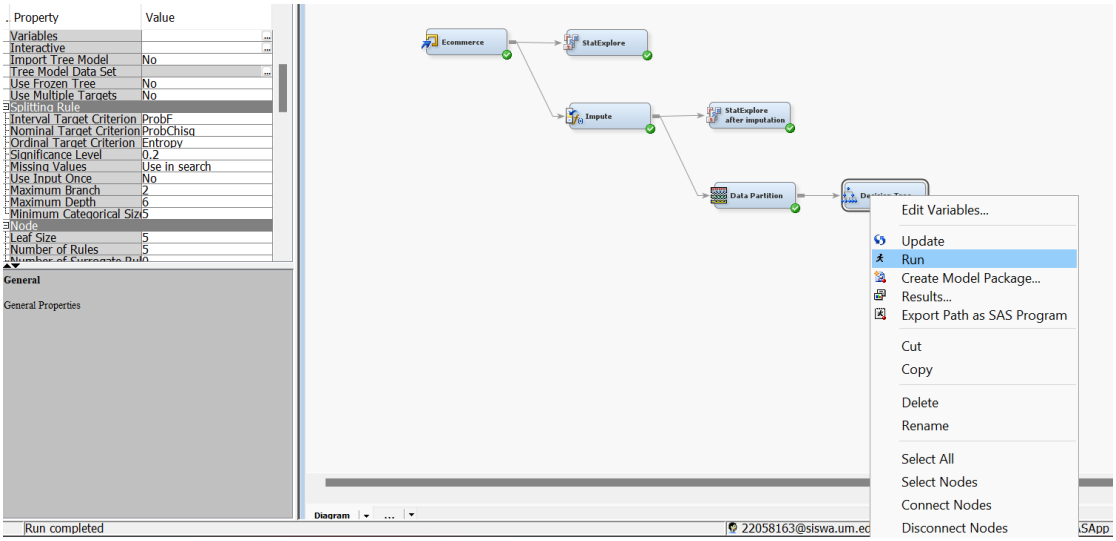
Data=VALIDATE

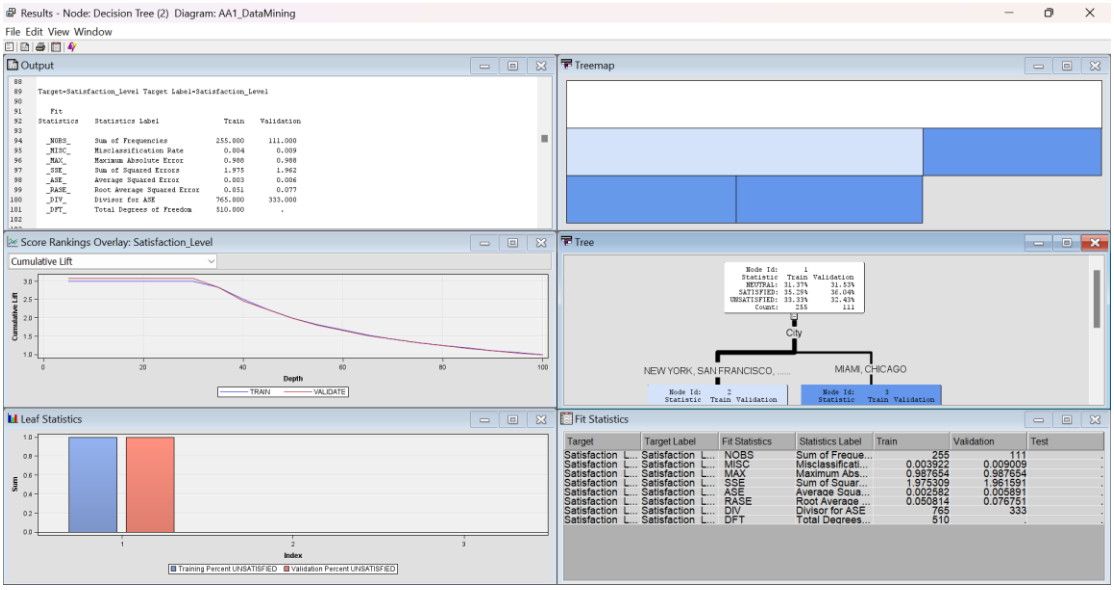
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Satisfaction_Level	.	Neutral	35	31.5315	Satisfaction_Level
Satisfaction_Level	.	Satisfied	40	36.0360	Satisfaction_Level
Satisfaction_Level	.	Unsatisfied	36	32.4324	Satisfaction_Level

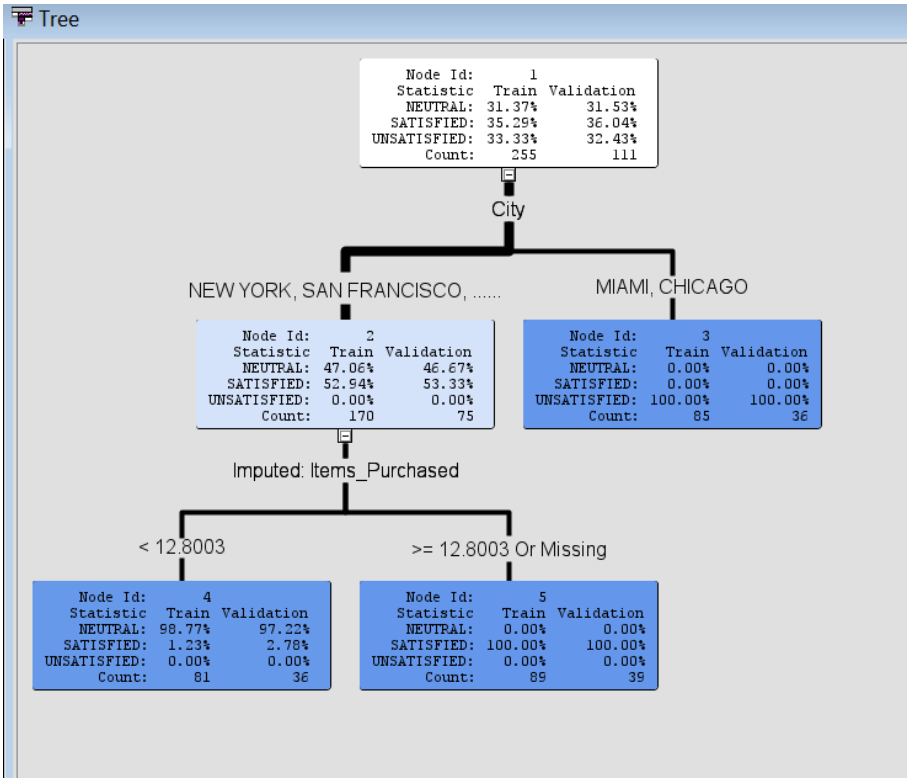
In my dataset, each value of the target variable has almost the same frequency. “Neutral” values take up 31%, “Satisfied” values take up 35% whereas “Unsatisfied” values take up 32% of the target variable. The target variable must have the same amount of different values to ensure that it is balanced.

Decision Tree Analysis

a. Attach “Decision Tree” node into the diagram and configure the “Impute” node to the “decision Tree” node, then run it.







Event Classification Table

Data Role=TRAIN Target=Satisfaction_Level Target Label=Satisfaction_Level

False Negative	True Negative	False Positive	True Positive
0	170	0	85

Data Role=VALIDATE Target=Satisfaction_Level Target Label=Satisfaction_Level

False Negative	True Negative	False Positive	True Positive
0	75	0	36

Fit Statistics

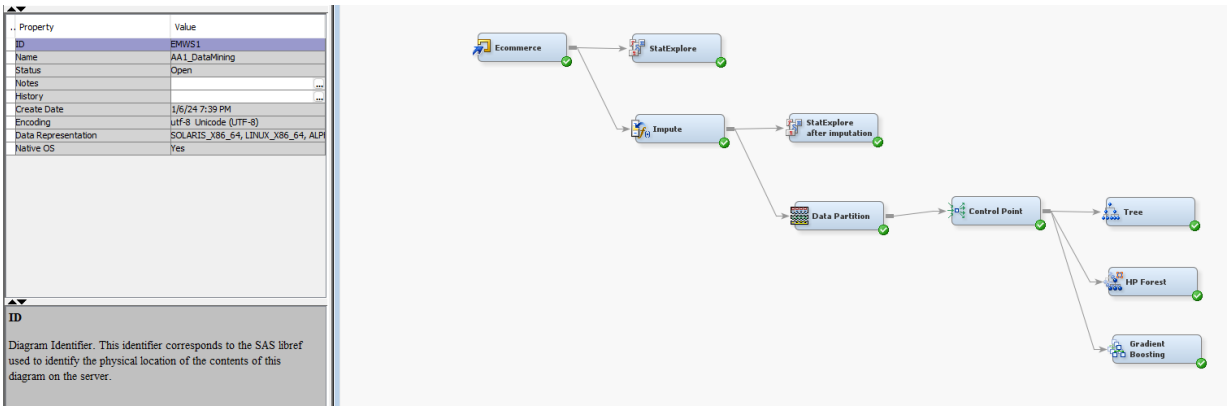
Target=Satisfaction_Level Target Label=Satisfaction_Level

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	255.000	111.000
MISC	Misclassification Rate	0.004	0.009
MAX	Maximum Absolute Error	0.988	0.988
SSE	Sum of Squared Errors	1.975	1.962
ASE	Average Squared Error	0.003	0.006
RASE	Root Average Squared Error	0.051	0.077
DIV	Divisor for ASE	765.000	333.000
DFT	Total Degrees of Freedom	510.000	.

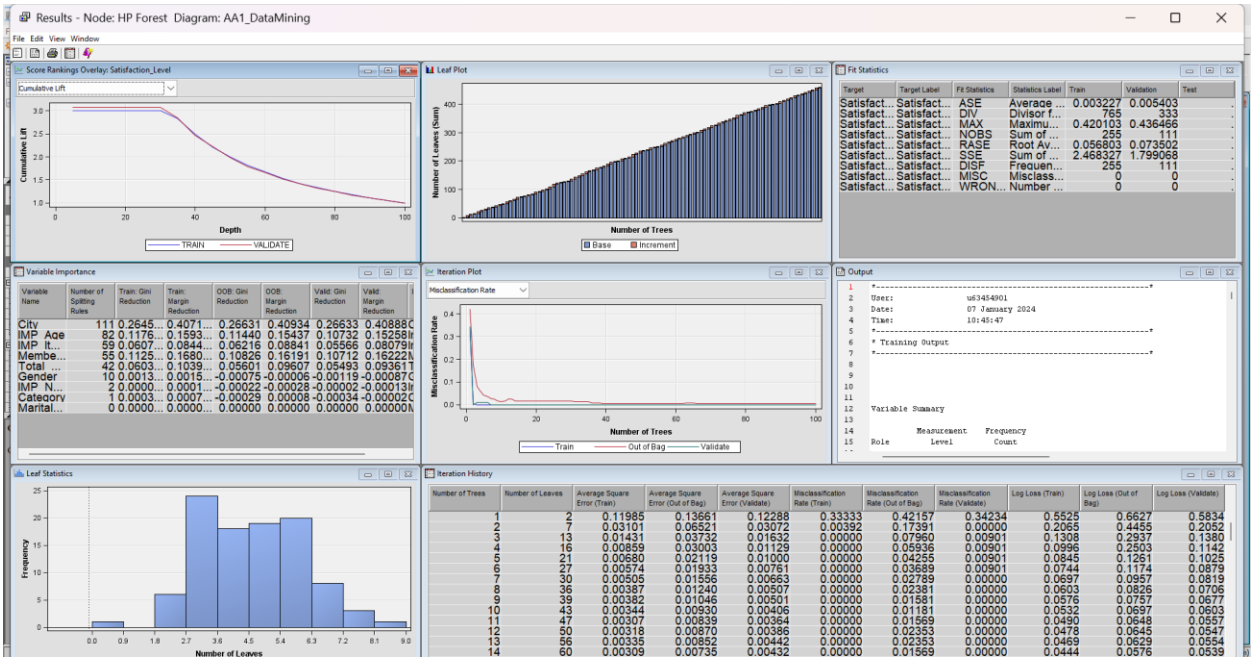
From the above results, it shows that the sum of squared errors in validation set is lower compared to training set whereas misclassification rate, average squared error and root average squared error in validation set is higher compared to training set. This shows that there is a slightly overfitting problem.

Ensemble Methods

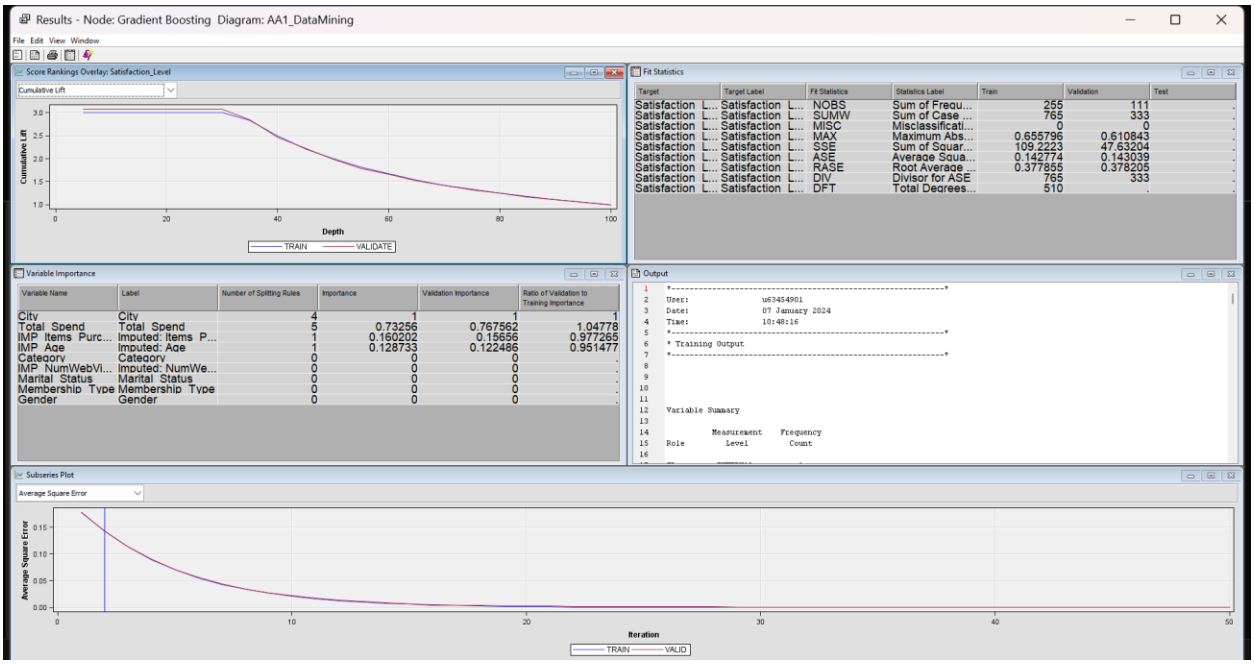
The ensemble methods used are random forest and gradient boosting. This is to compare which models will reveal the highest accuracy.



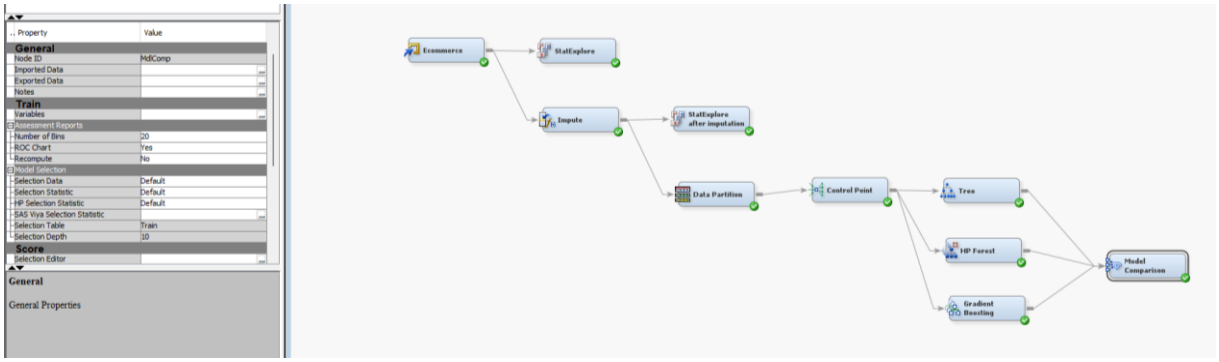
Random Forests output



Gradient boosting output



Model Comparison

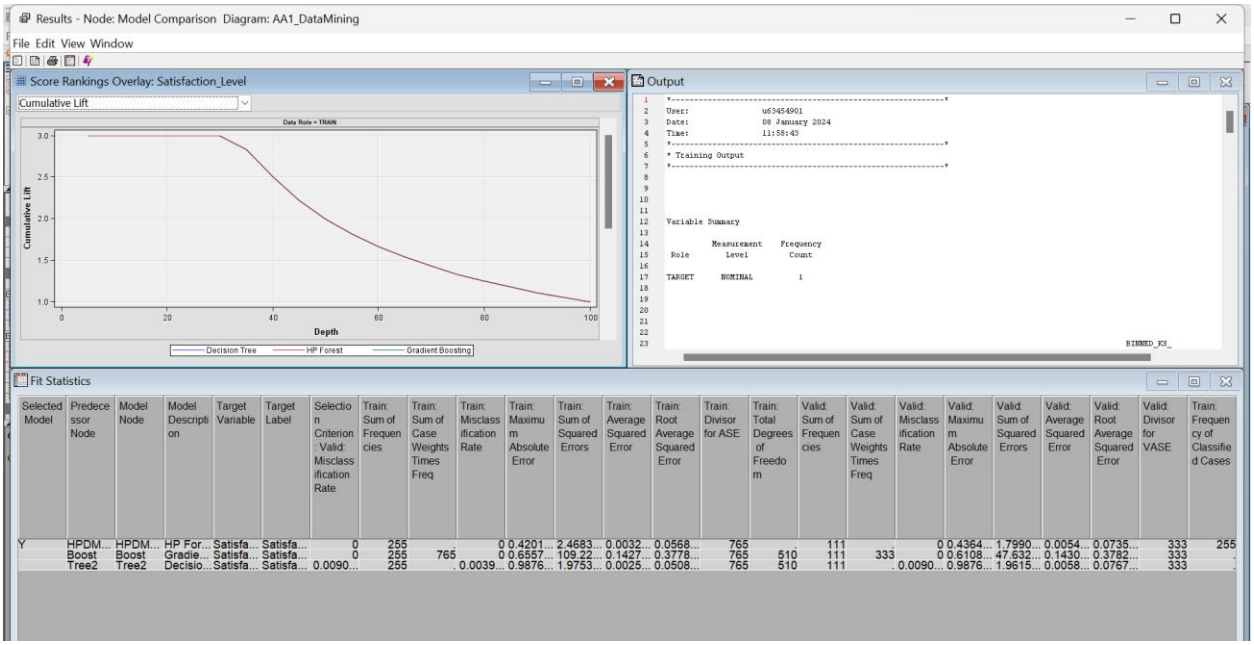


Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	HPDMForest	HP Forest	.000000000	0.00323	.000000000	0.00540
	Boost	Gradient Boosting	.000000000	0.14277	.000000000	0.14304
	Tree2	Decision Tree (2)	.009009009	0.00258	.003921569	0.00589

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Boost	Gradient Boosting	TRAIN	Satisfaction_Level	Satisfaction_Level	0	170	0	85
Boost	Gradient Boosting	VALIDATE	Satisfaction_Level	Satisfaction_Level	0	75	0	36
HPDMForest	HP Forest	TRAIN	Satisfaction_Level	Satisfaction_Level	0	170	0	85
HPDMForest	HP Forest	VALIDATE	Satisfaction_Level	Satisfaction_Level	0	75	0	36
Tree2	Decision Tree (2)	TRAIN	Satisfaction_Level	Satisfaction_Level	0	170	0	85
Tree2	Decision Tree (2)	VALIDATE	Satisfaction_Level	Satisfaction_Level	0	75	0	36



From the above images, it seems that Decision tree shows the highest accuracy as it has the lowest average squared error and misclassification rate in both training and validation set. Though it displays an

overfitting behaviour yet, it is the most suitable model to be used as the other models show overfitting behaviour as well as providing a slightly higher misclassification rate, average squared error and root average squared error.