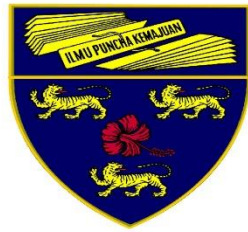


WQD 7005 DATA MINING



**UNIVERSITY
OF MALAYA**
K U A L A L U M P U R

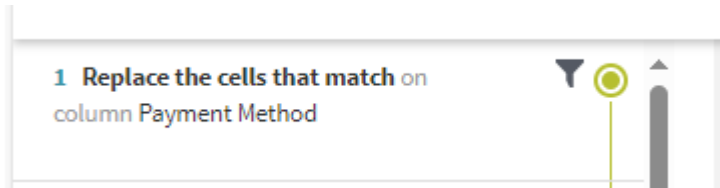
CASE STUDY –

**E-COMMERCE CUSTOMER BEHAVIOUR
ANALYSIS**

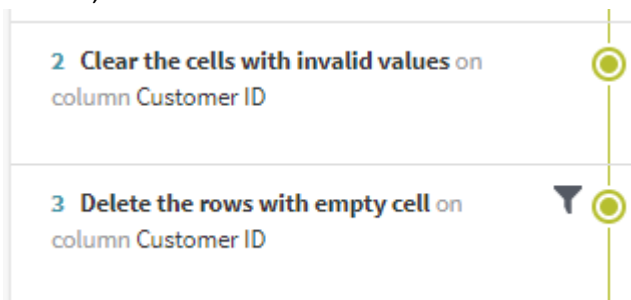
ALTERNATIVE ASSESSMENT I

Talend Data Preparation.

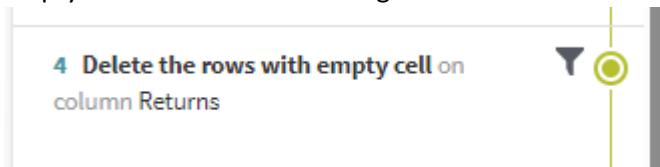
1. The dataset is downloaded from Kaggle
([🛒 E-commerce Customer Data For Behavior Analysis \(kaggle.com\)](https://www.kaggle.com/datasets/ashishpatel26/e-commerce-customer-data-for-behavior-analysis)) .
2. The dataset is being uploaded on Data Preparation stage.
3. The data cleaning start with replace the inconsistent data with the respective values
CC = Credit Card. This to ensure there gonna be standardize payment method on those column.



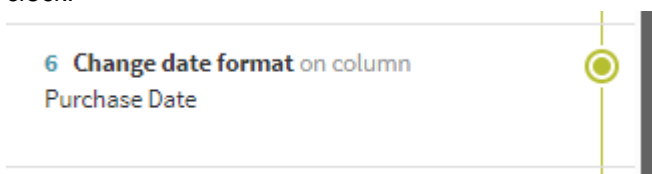
4. Next, invalid values are removed from the Customer ID. Also, from the blank customerID column, the rows are removed.



5. Empty cell on Returns also being removed the rows.



6. The Date format changed to dd-MM-yyyy . this helps to avoid the exact data without the clock.



13 Change date format on column Purchase Date

Current format:

I don't know, best guess

New format:

custom

Your format:

dd-MM-yyyy

SUBMIT

talend DATA PREPARATION

ecommerce_customer_data_DIRTY Preparation

1 Clear the cells with invalid values on column Customer ID

2 Replace the cells that match on column Payment Method

3 Delete the rows with empty cell on column Customer ID

4 Delete the rows with empty cell on column Returns

5 Extract date parts on column Purchase Date

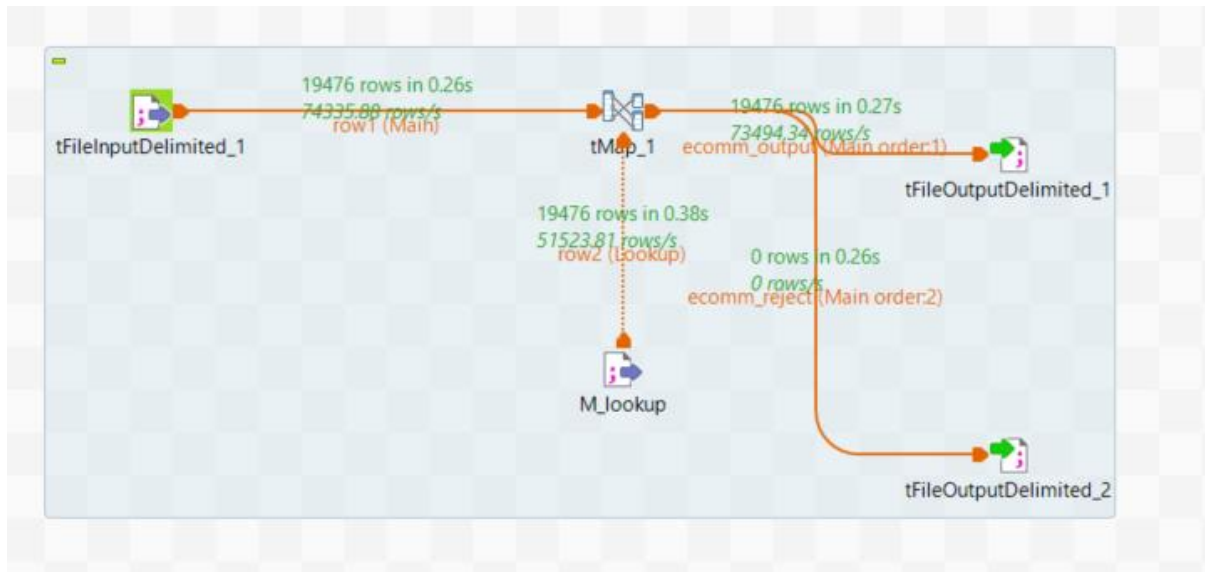
6 Change date format on column Purchase Date

Filters

Add a filter ...

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase A...	Payment Method	Customer Age	Returns
	fr_postal_code	date	text	integer	integer	integer	text	integer	int
1	44605	03-05-2023	Home	177	1	2427	PayPal	31	
2	44605	16-05-2021	Electronics	174	3	2448	PayPal	31	
3	44605	13-07-2020	Books	413	1	2345	Credit Card	31	
4	44605	17-01-2023	Electronics	396	3	937	Cash	31	
5	44605	01-05-2021	Books	259	4	2598	PayPal	31	
6	13738	25-08-2022	Home	191	3	3722	Credit Card	27	
8	13738	05-02-2023	Books	370	5	1486	Cash	27	
10	13738	09-02-2023	Electronics	40	4	4327	Cash	27	
12	33969	05-01-2023	Home	384	1	3883	PayPal	27	
13	33969	18-07-2023	Books	54	2	4187	PayPal	27	
14	33969	20-12-2021	Electronics	428	4	2289	Cash	27	
15	33969	07-03-2019	Books	781	1	3810	Cash	27	

Talend Open Studio for Data Integration



1. Combine the 2 tables. match the both table by using Customer_ID as the primary key.
Execute the new table with new column that indicates the churn:

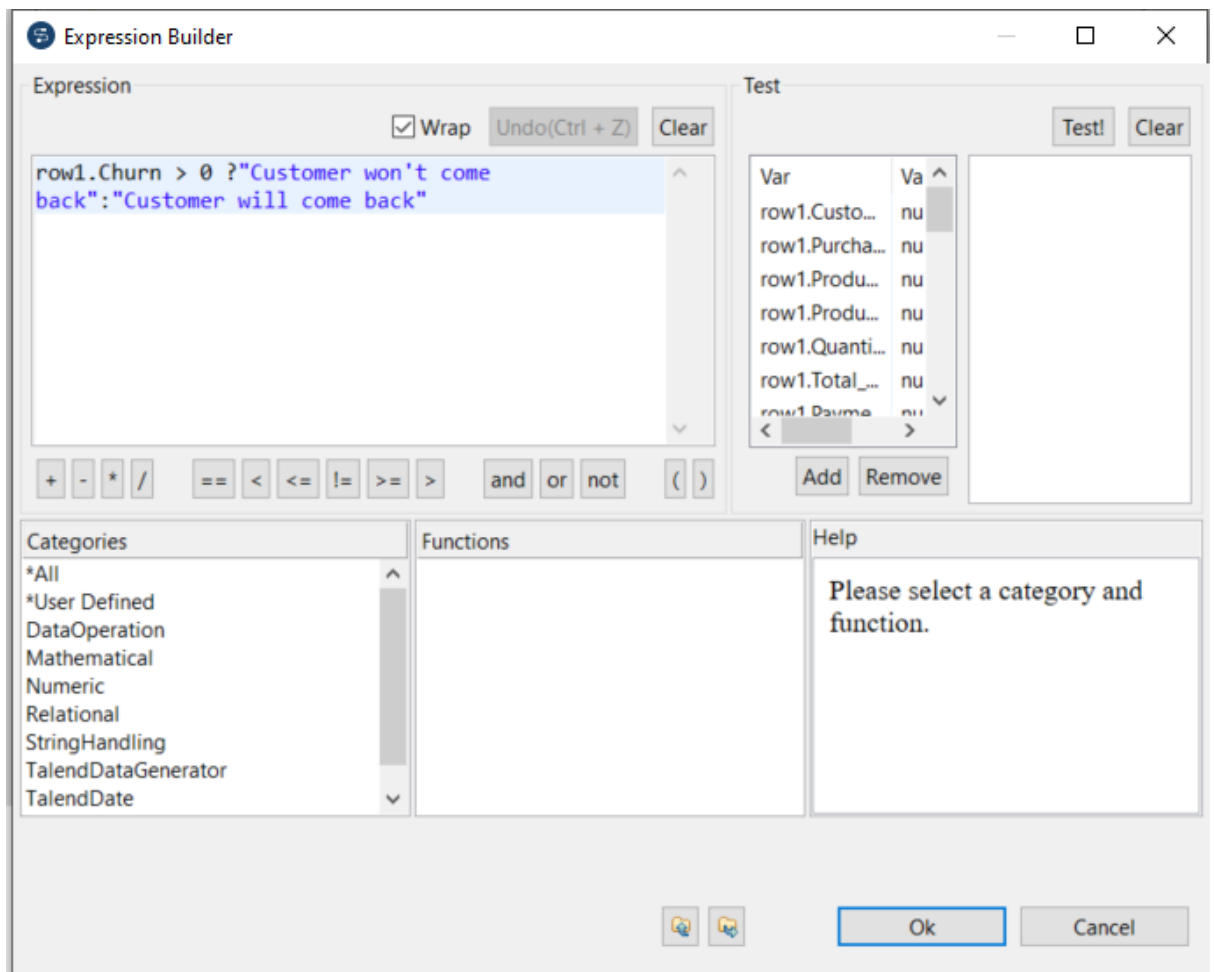
row1

Column	Type	Length	Precision	Default	Comment
Customer_ID	Integer	5	0		
Purchase_Date	Date	10	0		
Product_Category	String	11	0		
Product_Price	Integer	3	0		
Quantity	Integer	1	0		
Total_Purchase_Amount	Integer	4	0		
Payment_Method	String	11	0		
Customer_Age	Integer	2	0		
Returns	Integer	1	0		
Customer_Name	String	15	0		

ecommm_output

Column	Type	Length	Precision	Default	Comment
Customer_ID	Integer	5	0		
Purchase_Date	Date	10	0		
Product_Category	String	11	0		
Product_Price	Integer	3	0		
Quantity	Integer	1	0		
Total_Purchase_Amount	Integer	4	0		
Payment_Method	String	11	0		
Customer_Age	Integer	2	0		
Returns	Integer	1	0		
Customer_Name	String	15	0		
Churn_Category	String	30	0		

2. new column is added into the new column named as Churn_Category that portray the churn analysis (1= Customer won't come back, 0= Customer will come back)



- The new table is generated as (ecomm_output.csv) that have the City and churn category added.

44605;04/01/2021;Electronics;174;3;2446;PayPal;31;1;John Rivera;31;Female;0;Nairobi;Customer will come back			
44605;30/12/2020;Books;413;1;2345;Credit Card;31;1;John Rivera;31;Female;0;Nairobi;Customer will come back			
44605;02/01/2023;Electronics;396;3;937;Cash;31;0;John Rivera;31;Female;0;Nairobi;Customer will come back			
44605;04/01/2021;Books;259;4;2598;PayPal;31;1;John Rivera;31;Female;0;Nairobi;Customer will come back			
13738;03/01/2022;Home;191;3;3722;Credit Card;27;1;Lauren Johnson;27;Female;0;Mumbai;Customer will come back			
13738;02/01/2023;Books;370;5;1486;Cash;27;1;Lauren Johnson;27;Female;0;Mumbai;Customer will come back			
13738;02/01/2023;Electronics;40;4;4327;Cash;27;0;Lauren Johnson;27;Female;0;Mumbai;Customer will come back			
33969;02/01/2023;Home;304;1;3883;PayPal;27;1;Carol Allen;27;Male;0;Paris;Customer will come back			
33969;02/01/2023;Books;54;2;4187;PayPal;27;0;Carol Allen;27;Male;0;Paris;Customer will come back			
33969;04/01/2021;Electronics;428;4;2289;Cash;27;0;Carol Allen;27;Male;0;Paris;Customer will come back			
33969;30/12/2020;Books;281;1;3810;Cash;27;0;Carol Allen;27;Male;0;Paris;Customer will come back			
33969;03/01/2022;Home;193;2;3198;Credit Card;27;0;Carol Allen;27;Male;0;Paris;Customer will come back			
33969;02/01/2023;Clothing;473;3;2881;Credit Card;27;1;Carol Allen;27;Male;0;Paris;Customer will come back			
42650;30/12/2020;Books;127;5;3347;Cash;20;0;Curtis Smith;20;Female;0;Nairobi;Customer will come back			
42650;30/12/2020;Home;284;2;3531;Credit Card;20;1;Curtis Smith;20;Female;0;Nairobi;Customer will come back			
42650;03/01/2022;Electronics;256;2;3548;Credit Card;20;0;Curtis Smith;20;Female;0;Nairobi;Customer will come back			

SAS Enterprise Miner

1. The variables is custom accordingly by set The Churn as the target value
2. Library is created and the data source is generated.

Variables - FIMPORT

(none) ☐ not Equal to ☐ Mining ☐ Basic

Columns: ☐ Label

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Rejected	Interval	No		Yes	.	.
Churn	Target	Interval	No		No	.	.
Customer_Age	Input	Interval	No		No	.	.
Customer_ID	ID	Nominal	No		No	.	.
Customer_Name	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Payment_Method	Input	Nominal	No		No	.	.
Product_Category	Input	Nominal	No		No	.	.
Product_Price	Input	Interval	No		No	.	.
Purchase_Date	Time ID	Interval	No		No	.	.
Quantity	Input	Interval	No		No	.	.
Returns	Input	Interval	No		No	.	.
Total_Purchase_Amount	Input	Interval	No		No	.	.

Data Source Wizard -- Step 5 of 8 Column Metadata

(none) ☐ not Equal to ☐ Mining ☐ Basic ☐ Statistics

Columns: ☐ Label

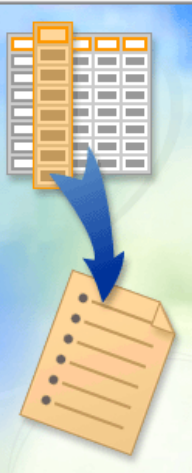
Name	Role	Level	Report	Order	Drop	Lower Limit
Age	Rejected	Interval	No		Yes	.
Churn	Target	Interval	No		No	.
Customer_Age	Input	Interval	No		No	.
Customer_ID	ID	Nominal	No		No	.
Customer_Name	Input	Nominal	No		No	.
Gender	Input	Nominal	No		No	.
Payment_Method	Input	Nominal	No		No	.
Product_Category	Input	Nominal	No		No	.
Product_Price	Input	Interval	No		No	.
Purchase_Date	Time ID	Interval	No		No	.
Quantity	Input	Interval	No		No	.
Returns	Input	Interval	No		No	.
Total_Purchase_Amount	Input	Interval	No		No	.

< >

Show code Explore Compute Summary < Back Next > Cancel

3. The sample dataset is set to 20% since the dataset is too large

Data Source Wizard -- Step 6 of 8 Create Sample



Do you wish to create a sample data set?

☐ No ☒ Yes

Table Info


Columns 13
Rows 19476

Sample Size

Type Percent
Percent 20
Rows

< Back Next > Cancel

Data Source Wizard -- Step 8 of 8 Summary

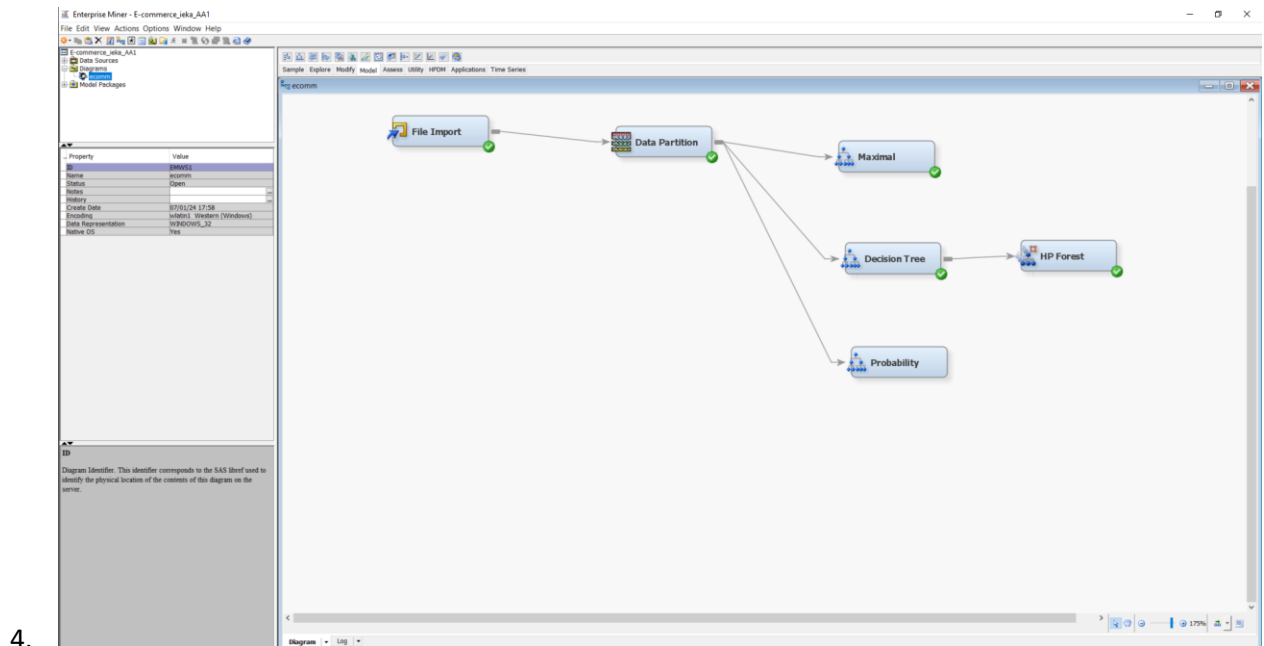


Metadata Completed.

Library: ECOMM
Data Source: FIMPORT_DATA
Role: Raw

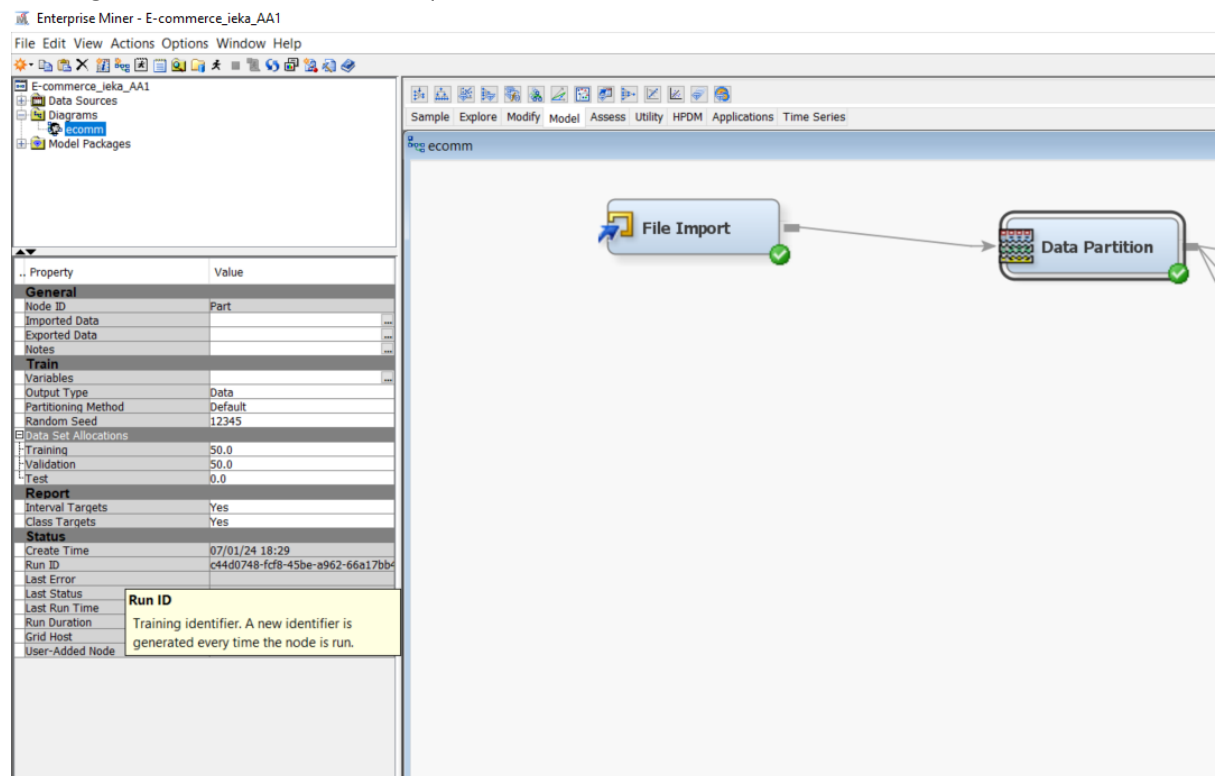
Role	Level	Count
ID	Nominal	1
Input	Interval	5
Input	Nominal	4
Rejected	Interval	1
Target	Interval	1
Time ID	Interval	1

< Back Finish Cancel

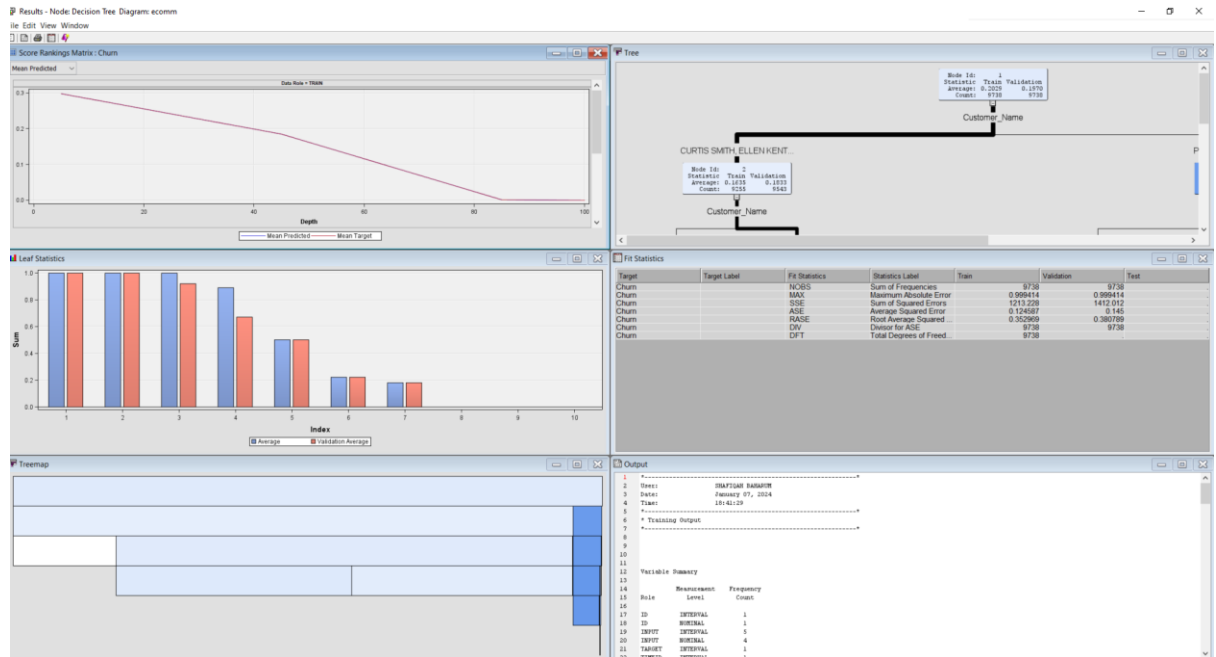


this is how it looks like on my SAS. Use the decision tree and HP Forest to generate the analysis.

5. Data partition properties:
training and validation are set as equals.



6. Results of Decision Tree:



7. End result of Random Forest:

