PRESENTATION VIDEO LINK:
https://drive.google.com/file/d/1eeg1VH0_YouMfDY63vk087lfRYxGyC-t/view?usp=sharing

WIE3007:
DATA MINING & WAREHOUSING

GROUP MEMBER:

1. NURUL AIN BINTI KHAIRUL ANWAR (U2005370)
2. IFFAH SORFINA BINTI MOHAMAD NAIM (17203173)
3. LUBNA BINTI AHMAD NIZAM (17204040)
4. NURAUFA NATASHA BINTI AZROL (U2102739)
5. EZYAN MUNIRAH BINTI ZAINUDDIN (U2000643)

INTRODUCTION

- Focuses on Superstore's purchase history and the importance of using data mining techniques.
- Consumer choices influence broader trends, impacting product offerings, marketing, and inventory management.
- Data mining is a powerful tool for extracting insights from transactional data to anticipate future trends.

Problem Statement: To discover patterns and trends that can help the organization make strategic decisions on products to offer and improve the overall shopping experience.

OBJECTIVES

1

To understand the preferences of customers and their purchasing behavior

2

To understand the performance of the product sales

Dataset Description: Retail sales data from Superstore, “Superstore.csv” (https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting)

METHODOLOGY

Tools Used:

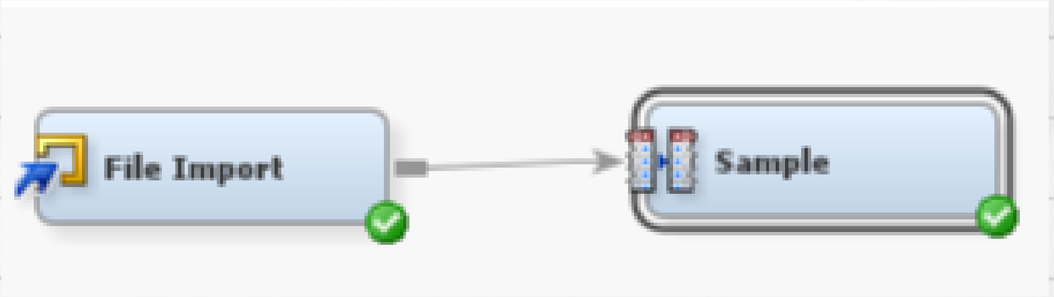
- SAS Enterprise Miner
- Talend Data Preparation
- Python

Approach

SEMMA

- Sample
- Explore
- Modify
- Model
- Assess

SAMPLE



- Using the Sample node.
- Reduce the size of huge volumes of data.
- Select 50% of the dataset.
- Random sampling.

Sampling Summary

Type	Data Set	Number of Observations
DATA	EMWS1.FIMPORT_train	9994
SAMPLE	EMWS1.Smpl_DATA	4997

EXPLORE

1. **StatExplore Node:**

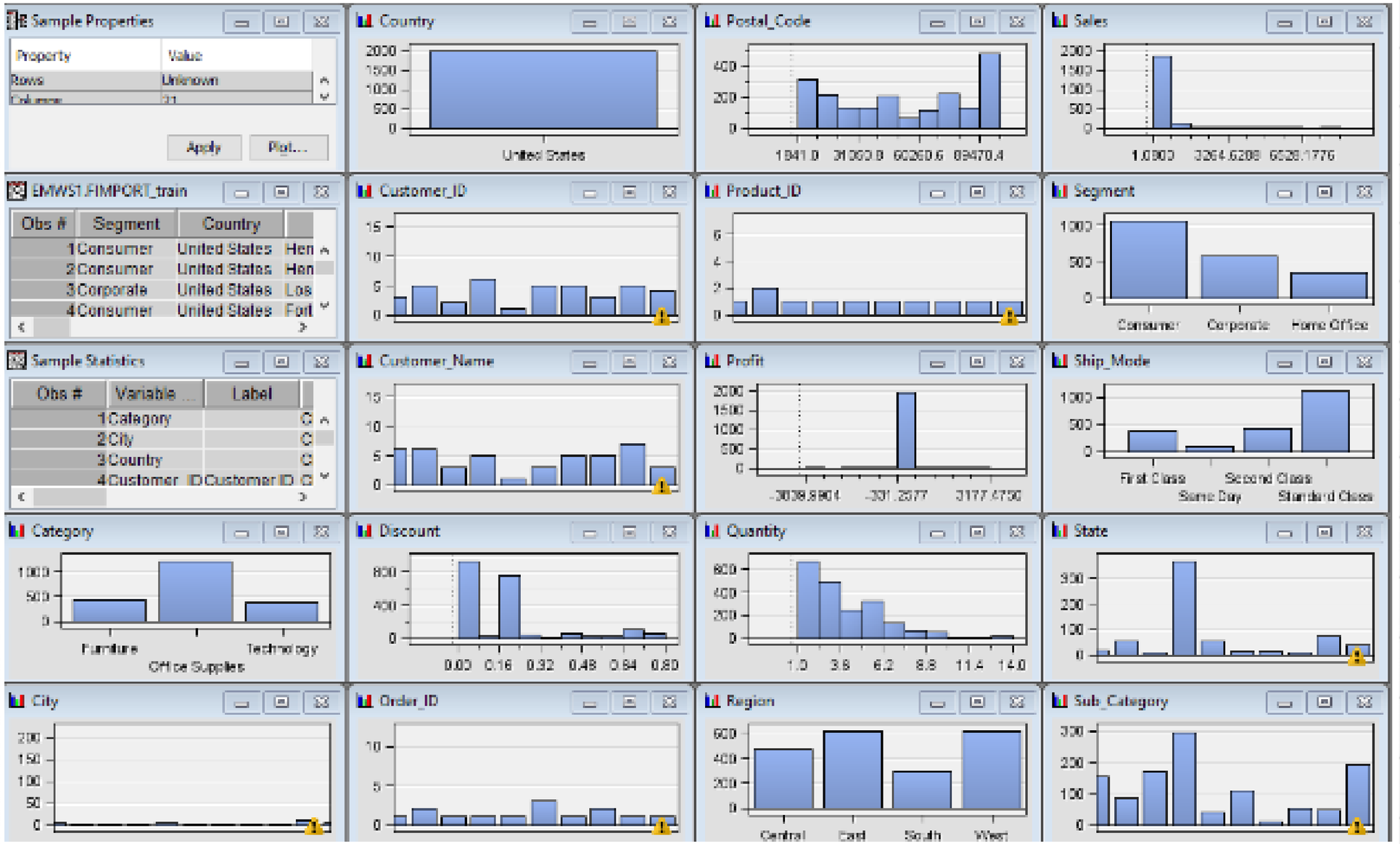
- To gain valuable insights into the underlying patterns and distributions within the dataset.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Category	INPUT	3	0	Office Supplies	60.30	Furniture	21.22
TRAIN	City	INPUT	513	0	New York City	9.16	Los Angeles	7.43
TRAIN	Country	INPUT	1	0	United States	100.0		0.00
TRAIN	Customer_ID	INPUT	513	0	ZC-21910	0.97	JE-15745	0.91
TRAIN	Customer_Name	INPUT	513	0	Zunchuss Carroll	0.97	Joel Eaton	0.91
TRAIN	Order_ID	INPUT	513	0	CA-2017-117457	0.84	CA-2014-115812	0.66
TRAIN	Product_ID	INPUT	513	0	OFF-PA-10001970	0.67	FUR-BO-10002545	0.51
TRAIN	Region	INPUT	4	0	West	32.05	East	28.50
TRAIN	Ship_Mode	INPUT	4	0	Standard Class	59.72	Second Class	19.46
TRAIN	State	INPUT	49	0	California	20.02	New York	11.29
TRAIN	Sub_Category	INPUT	17	0	Binders	15.24	Paper	13.71
TRAIN	Segment	SEGMENT	3	0	Consumer	51.94	Corporate	30.22

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Discount	INPUT	0.156203	0.206452	9994	0	0	0.2	0.8	1.684295	2.409546
Postal_Code	INPUT	55190.38	32063.69	9994	0	1040	56301	99301	-0.12853	-1.49302
Profit	INPUT	28.6569	234.2601	9994	0	-6599.98	8.662	8399.976	7.561432	397.1685
Quantity	INPUT	3.789574	2.22511	9994	0	1	3	14	1.278545	1.991889
Sales	INPUT	229.858	623.2451	9994	0	0.444	54.48	22638.48	12.97275	305.3118
VAR1	INPUT	4997.5	2885.164	9994	0	1	4997	9994	0	-1.2

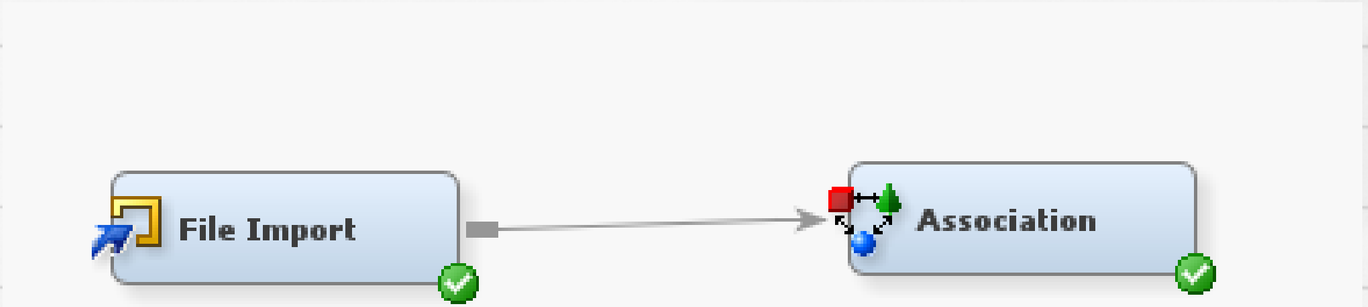

Sample Properties
Property
Value
Rows
Unknown
Apply
Plot...
Country
United States
Postal_Code
Sales
EMWS1.FIMPORT_train
Obs #
Segment
Country
1 Consumer United States
2 Consumer United States
3 Corporate United States
4 Consumer United States
Customer_ID
Product_ID
Segment
Consumer
Corporate
Home Office
Sample Statistics
Obs #
Variable
Label
1 Category
2 City
3 Country
4 Customer_ID Customer ID
Customer_Name
Profit
Ship_Mode
First Class
Second Class
Same Day
Standard Class
Category
Furniture
Office Supplies
Technology
Discount
Quantity
State
City
Order_ID
Region
Central
East
South
West
Sub_Category

Lubna binti Ahmad Nizam (17204040/2)

EXPLORE

2. Association Node:

- To find interesting patterns between pair of items among the transactional dataset based on the rules



Minimum Confidence Level
10%

Support Percentage
5%

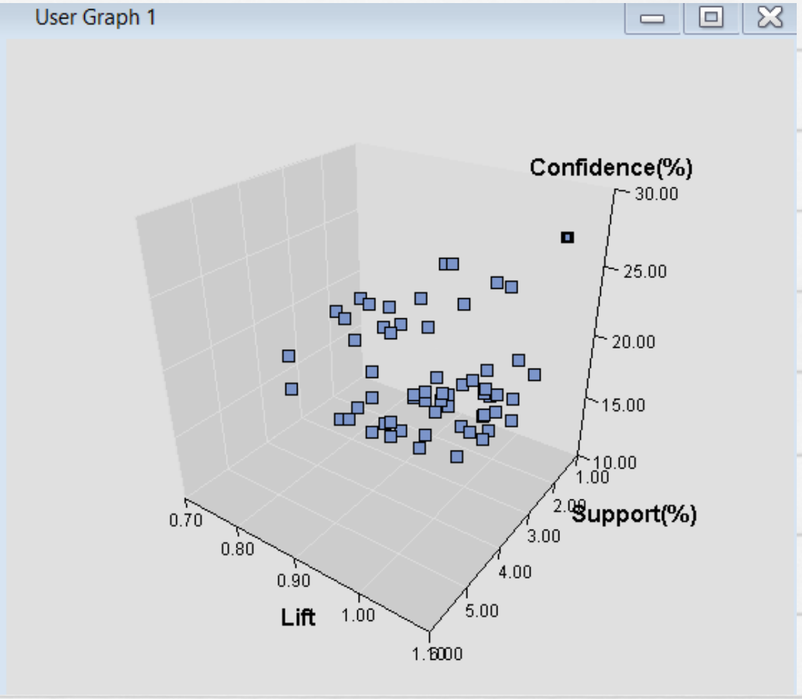
Variables

ID Role: Order_ID

Target ID: Sub_Category

Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Left Hand of Rule
2	26.27	28.82	2.60	1.10	130.00	Applianc...	Appliances E
2	14.59	15.23	2.48	1.04	124.00	Phones ...	Phones A
2	16.25	16.96	2.48	1.04	124.00	Art ==> ...	Art F
2	17.51	18.18	2.95	1.04	148.00	Phones ...	Phones F
2	16.25	16.88	2.95	1.04	148.00	Furnishin...	Furnishin... F
2	16.25	16.85	1.52	1.04	76.00	Applianc...	Appliances F
2	23.78	24.17	2.18	1.02	109.00	Applianc...	Appliances F
2	17.51	17.76	2.76	1.01	138.00	Storage ...	Storage F
2	15.51	15.74	2.76	1.01	138.00	Furnishin...	Furnishin... S
2	17.51	17.74	1.60	1.01	80.00	Applianc...	Appliances F
2	14.33	14.25	2.32	0.99	116.00	Phones ...	Phones A
2	16.25	16.16	2.32	0.99	116.00	Accessor...	Accessor... F
2	11.50	11.29	1.98	0.98	99.00	Furnishin...	Furnishin... C
2	17.51	17.19	1.98	0.98	99.00	Chairs =...	Chairs F
2	14.59	14.24	1.64	0.98	82.00	Chairs =...	Chairs A
2	11.50	11.22	1.64	0.98	82.00	Art ==> ...	Art C
2	26.27	25.61	3.97	0.97	199.00	Storage ...	Storage E
2	15.51	15.12	3.97	0.97	199.00	Binders ...	Binders S
2	23.78	23.12	1.60	0.97	80.00	Labels =...	Labels F
2	23.78	23.09	2.66	0.97	133.00	Chairs =...	Chairs F
2	11.50	11.17	2.66	0.97	133.00	Paper ==...	Paper C
2	23.78	22.91	3.55	0.96	178.00	Storage ...	Storage F



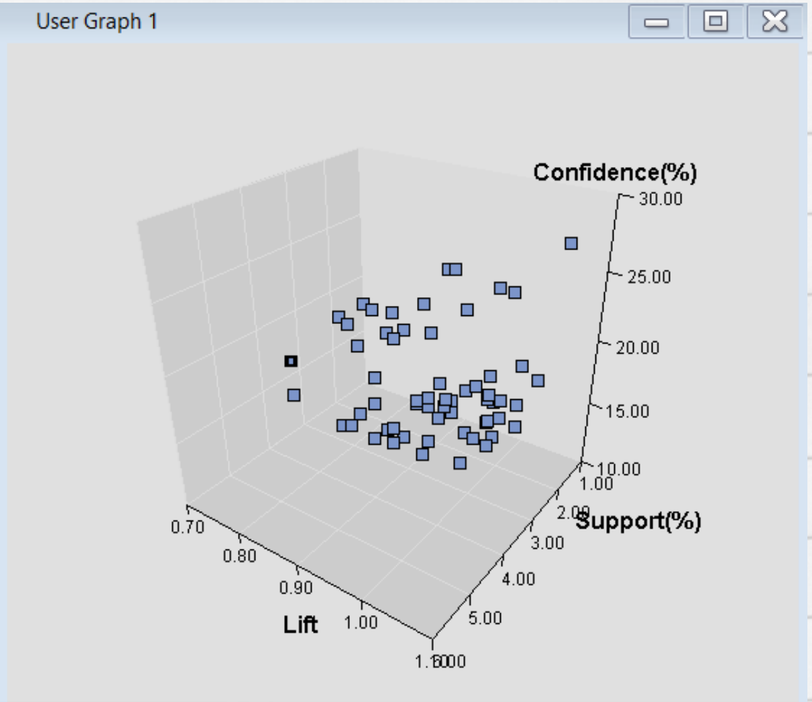
Rule Appliances==>Binders

- Highest lift at 1.10
- Confidence 28.82%

- Rule Storage==>Binders
- Lift 0.97
- Confidence 25.61%

Rules Table

Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count ▼	Rule	Left Hand of Rule	Right Hand of Rule
26.27	23.09	5.49	0.88	275.00	Paper ==...	Paper	Binders
23.78	20.90	5.49	0.88	275.00	Binders ...	Binders	Paper
26.27	25.61	3.97	0.97	199.00	Storage ...	Storage	Binders
15.51	15.12	3.97	0.97	199.00	Binders ...	Binders	Storage
26.27	24.45	3.97	0.93	199.00	Phones ...	Phones	Binders
16.25	15.12	3.97	0.93	199.00	Binders ...	Binders	Phones
26.27	22.58	3.95	0.86	198.00	Furnishin...	Furnishin...	Binders
17.51	15.05	3.95	0.86	198.00	Binders ...	Binders	Furnishin..
23.78	22.91	3.55	0.96	178.00	Storage ...	Storage	Paper
15.51	14.95	3.55	0.96	178.00	Paper ==...	Paper	Storage
23.78	20.18	3.53	0.85	177.00	Furnishin...	Furnishin...	Paper
17.51	14.86	3.53	0.85	177.00	Paper ==...	Paper	Furnishin..
23.78	21.50	3.49	0.90	175.00	Phones ...	Phones	Paper
16.25	14.69	3.49	0.90	175.00	Paper ==...	Paper	Phones
14.33	12.23	3.21	0.85	161.00	Binders ...	Binders	Accessor..
26.27	22.42	3.21	0.85	161.00	Accessor...	Accessor...	Binders
14.33	12.85	3.05	0.90	153.00	Paper ==...	Paper	Accessor..
23.78	21.31	3.05	0.90	153.00	Accessor...	Accessor...	Paper
23.78	20.52	2.99	0.86	150.00	Art ==> ...	Art	Paper
14.59	12.59	2.99	0.86	150.00	Paper ==...	Paper	Art
26.27	20.52	2.99	0.78	150.00	Art ==> ...	Art	Binders
14.59	11.40	2.99	0.78	150.00	Binders ...	Binders	Art



Rule Paper==>Binders

- Lift 0.88
- Confidence 23.09%
- Transaction count 275

MODIFY

Data Cleaning:

- Date formatting
- Profit and Sales column formatting
- Postal code formatting
- Customer ID formatting

Order ID (text)	Order Date (date)	Ship Date (date)	Ship Mode (text)	Customer ID (text)
CA-2015-146087	2015-07-06	2015-07-11	Standard Class	PP-18955
CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740
CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740
CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740
CA-2015-117961	2015-11-26	2015-11-30	Standard Class	GP-14740

	Order Date	Ship Date	Ship Mode	Customer ID	Customer	Segment
1	22/11/2015	25/11/2015	Second Cl	HD-14785	Harold Da	Home Off
1	22/11/2015	25/11/2015	Second Cl	HD-14785	Harold Da	Home Off
1	1/10/2015	4/10/2015	Second Cl	-RM19,675	Robert Ma	Home Off
1	1/10/2015	4/10/2015	Second Cl	-RM19,675	Robert Ma	Home Off
1	8/9/2016	14/9/2016	Standard (	AP-10720	Anne Pyc	Home Off

Country	City	State	Postal Code	Region
United States	Holyoke	Massachusetts	1040	East
United States	Leominster	Massachusetts	1453	East
United States	Leominster	Massachusetts	1453	East

MODIFY

Generate new column:

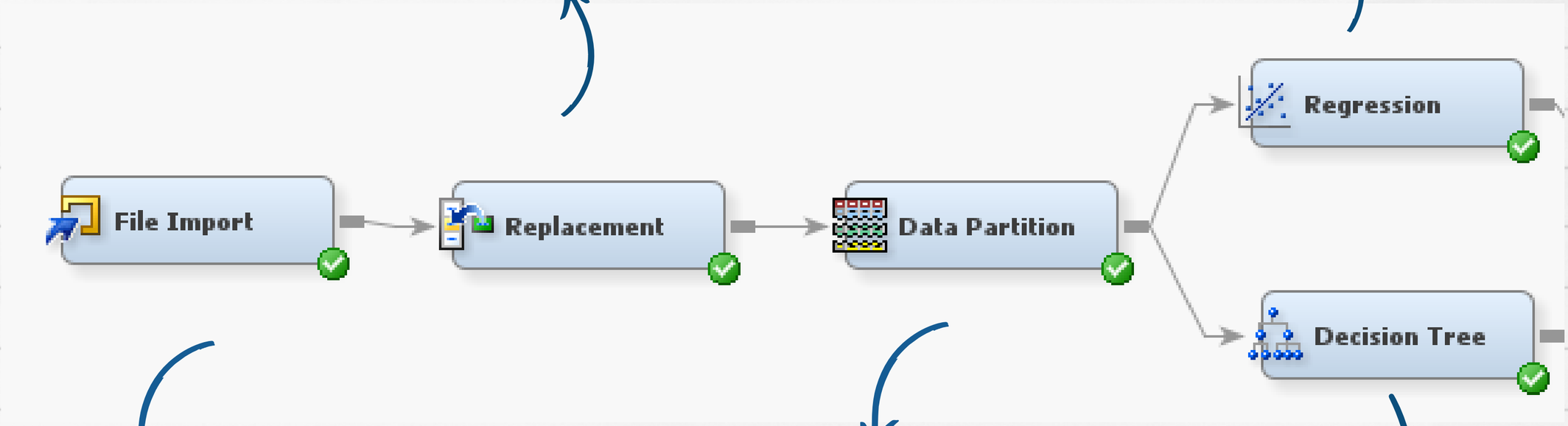
- “Shipping Duration” column was added by using 2 existing columns “Order Date” and “Ship Date”

```
import featuretools as ft
import pandas as pd

# Load the dataset
data = pd.read_csv('superstore.csv')

# Convert 'Order Date' and 'Ship Date' to datetime objects
data['Order Date'] = pd.to_datetime(data['Order Date'])
data['Ship Date'] = pd.to_datetime(data['Ship Date'])
# Calculate the shipping duration
data['Shipping Duration'] = (data['Ship Date'] - data['Order Date']).dt.days
data.to_csv('Superstore(20).csv', index=False)
```

MODELLING

to allow trimming non-missing values by replacing values that might be falsely listed

o fit regression model to predecessor dataset, forecasts the value of an interval target



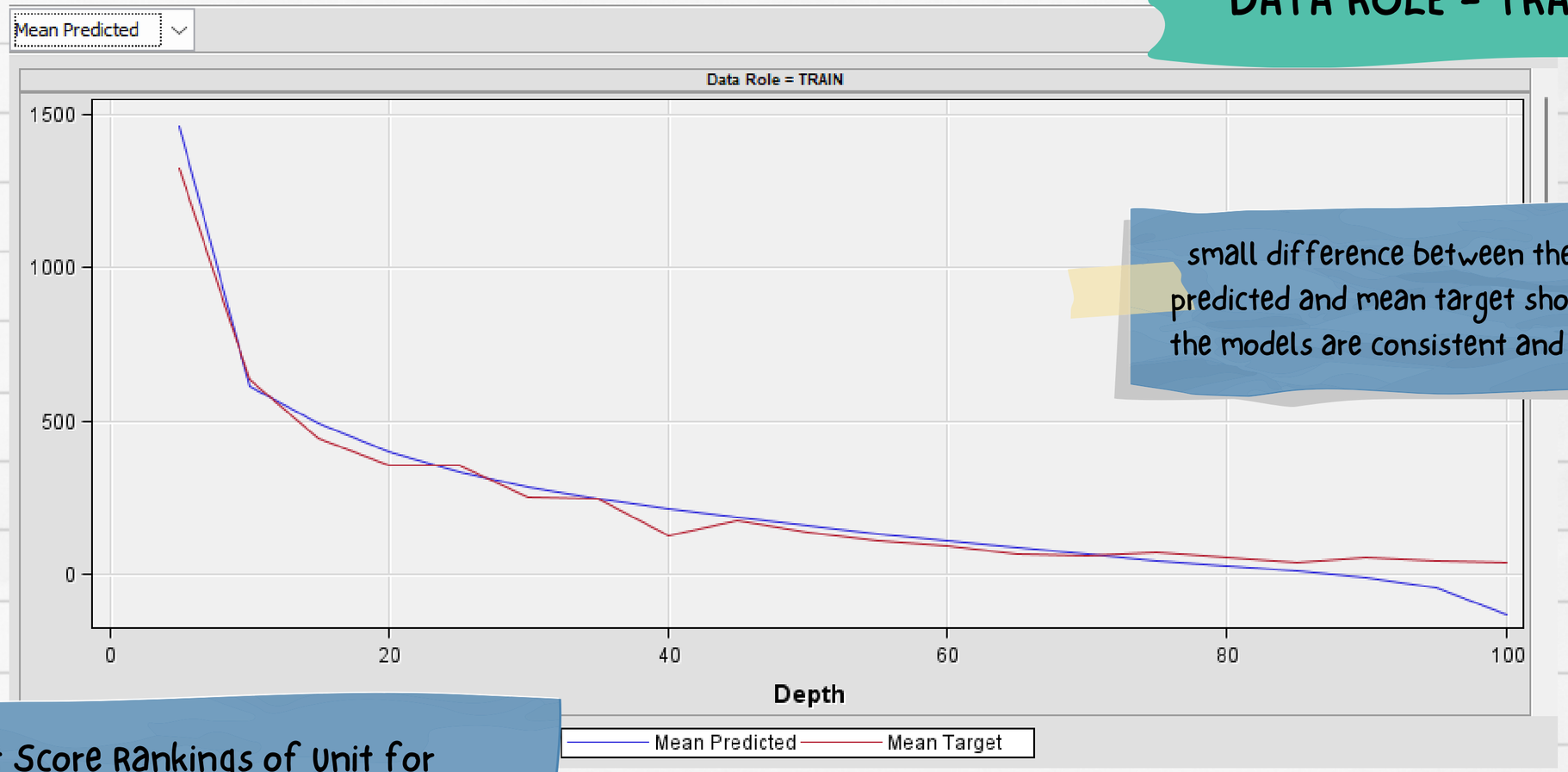
the cleaned dataset is imported into the SAS data source

o allow trimming non-missing values by replacing values that might be falsely listed

Gives a hierarchical segmentation of the input data based on rules applied to each observation

1. REGRESSION

DATA ROLE - TRAIN

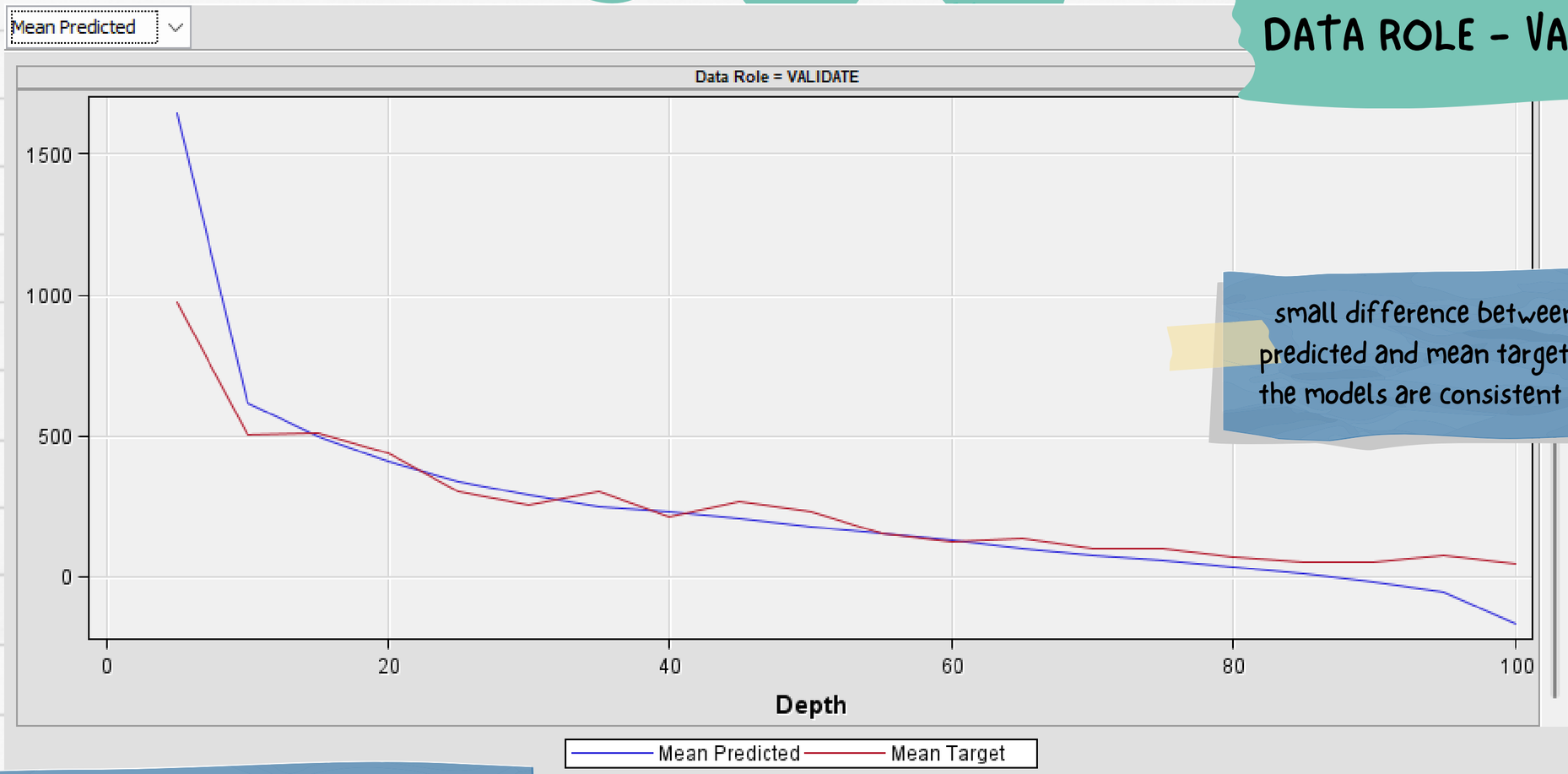


small difference between the mean predicted and mean target shows that the models are consistent and reliable

the Assessment Score Rankings of Unit for Data Role = TRAIN has the mean target and mean predicted value from depth of 5 to 100 with increment of 5.

As for mean predicted, the value is decreasing until -127.89 at depth of 100.

DATA ROLE - VALIDATE

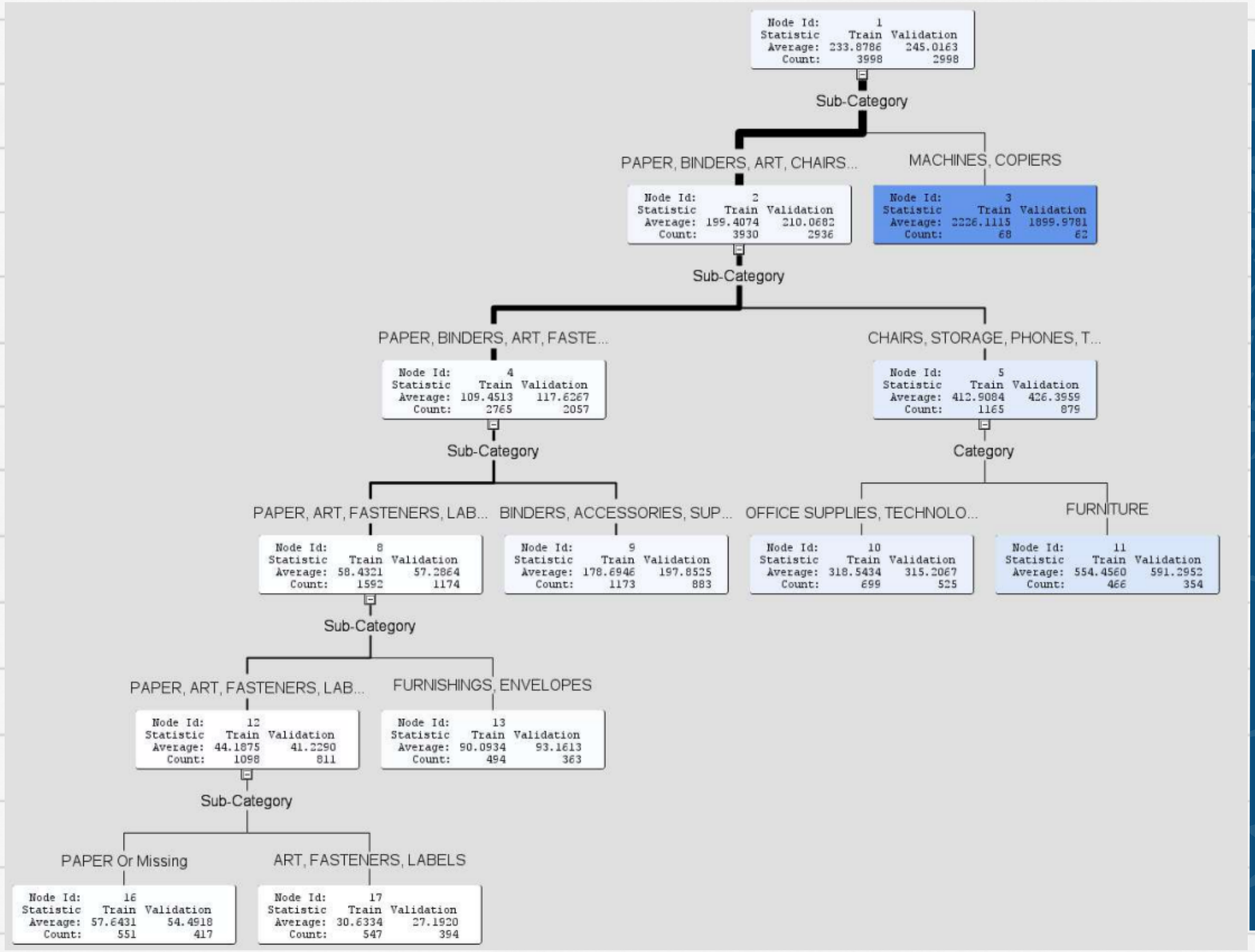


small difference between the mean predicted and mean target shows that the models are consistent and reliable

the Assessment Score Rankings of Unit for Data Role = VALIDATE has the mean target and mean predicted value from depth of 5 to 100 with increment of 5.

As for mean predicted, the value is decreasing until -163.91 at depth of 100

2. DECISION TREE



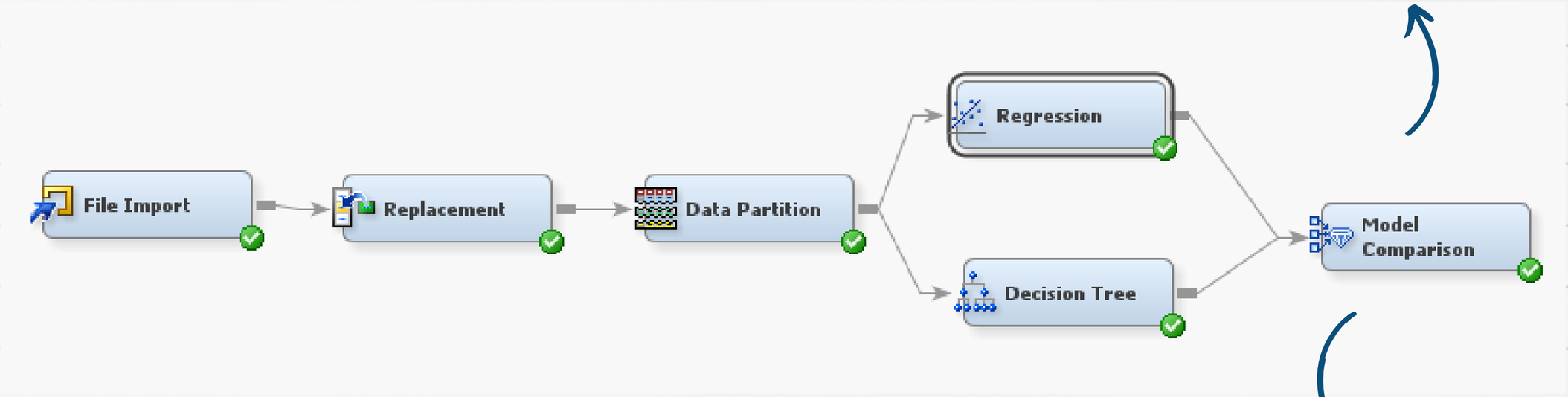
- Root node – the top node of a vertical tree that contains all observations.
- Internal nodes – non-terminal nodes that contain the splitting rule. This includes the root node.
- Leaf nodes – terminal nodes that contain the final classification for a set of observations.

The nodes are colored by the average of Unit. The line width is proportional to the ratio of the number of observations in the branch to the number of observations in the root node.

The root node reveals an average unit value for sub-category items of approximately 245.016

ASSESS

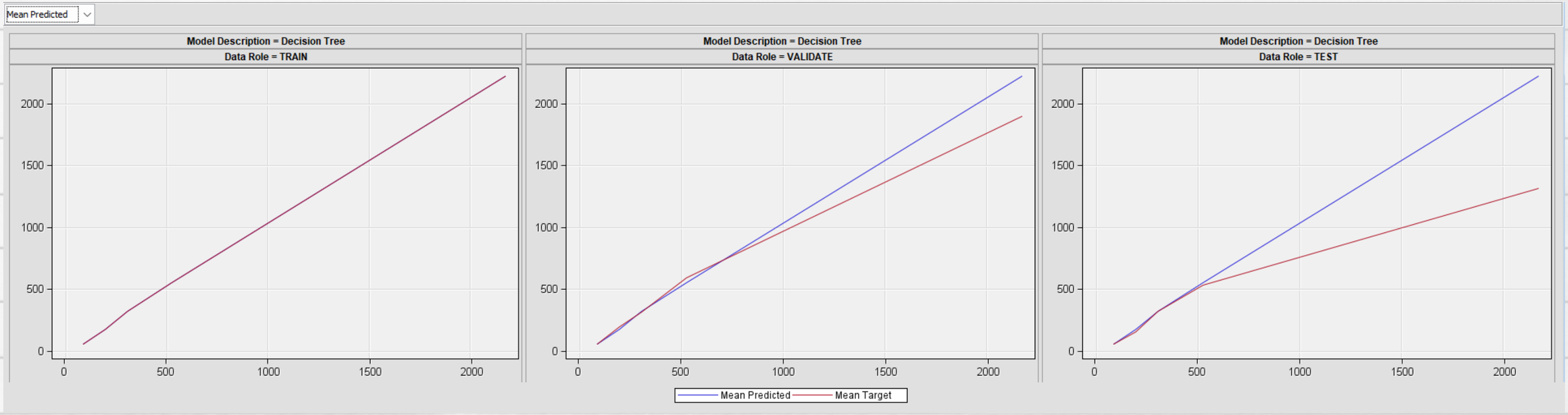
assessing the performance of different models to determine which one is most effective for your specific problem



Specify which models to compare and the evaluation metrics you want to use

SCORE DISTRIBUTION OF SALES

Decision Tree across three data roles: train, validate, and test.



Data Role - Train: both models show a linear increase in mean prediction and mean target

Data Role - Validate: both models show a linear increase but there is small difference between mean predicted and mean target

Data Role - Test: both models show a linear increase in mean prediction and mean target with a big difference between them

Regression Model across three data roles: train, validate, and test.

Data Role Train - mean predicted show linear increase but mean target show consistent increase until a drop and increasing back normally

Data Role Validate : mean target shows inconsistent of trend increasing and dropping

Data Role Test : mean target shows trend of slowly increasing and keeps growing linearly before it fall sharply

CONCLUSION & REFLECTION

CONCLUSION

It evaluates supervised and unsupervised learning methods in data analysis, demonstrating their application in SAS Enterprise Miner. Data mining tools enhance analysis depth and speed, enabling thorough examination of datasets, identification of patterns and trends, and extraction of key information. The focus was on understanding customer preferences, purchasing behavior, and product sales performance using techniques like data reduction, cleaning, regression, and various nodes. Future aims include refining supervised learning techniques for classification within SAS Enterprise Miner to improve data analysis outcomes.

REFLECTION

The reflection emphasizes adapting to the changing landscape of data analytics, with a focus on exploring new data sources and integrating advanced machine learning for more nuanced insights. It suggests a comparative study of different analytical tools to understand their unique strengths and applications. Ethical considerations, particularly in customer privacy and data security, are also highlighted. Finally, it recommends collaboration with experts in related fields to enhance the comprehensiveness and impact of the analysis.

THANK YOU VERY MUCH!