Case Study: Ecommerce Customer Behavior Analysis

1.0) Introduction:

Businesses seeking to acquire a competitive edge now depend heavily on the convergence of varied information and the actionable insights that can be extracted from them in the dynamic world of e-commerce. This project uses the combined power of SAS Enterprise Miner and Talend Data Integration to explore the complexities of consumer behavior analysis. While Talend Data Preparation has made thorough data understanding easier, the integration process has simplified data from several sources. The analysis that follows in SAS Enterprise Miner aims to identify trends, identify consumer behavior influencers, and generate strategic suggestions to improve e-commerce decision-making.

2.0) Objective

This case study's main goal is to use SAS Enterprise Miner to analyse an integrated e-commerce dataset in-depth, with a particular emphasis on churn prediction and customer behaviour analysis. The research attempts to get actionable insights that can guide strategic decision-making for companies involved in the digital marketplace through the use of sophisticated analytics approaches.

3.0) Dataset Structure

Column name	Description
CustomerID	It serves as a unique identifier for each customer and is not used as a target or input for modeling.
Age	Age is a continuous variable and can be used as an input for modeling.
Gender	Gender is a categorical variable and can be used as an input for modeling.
Country	Country is a categorical variable and can be used as an input for modelling
Membership level	Membership level is a categorical variable and can be used as an input for modelling
TotalPurchases	Total purchases is a continuous variable and can be used as an input for modeling.
TotalSpent	Total spent is a continuous variable and can be used as an input for modelling

FavoriteCategory	Favorite category is a categorical variable and can be used as an input for modelling
LastPurchaseDate	It can be used as an input for modeling if transformed appropriately
Occupation	Occupation is a categorical variable and can be used as an input for modelling
WebsiteVisitsFrequency	Website visits frequency is a categorical variable and can be used as an input for modelling
Item	Item is a categorical variable and can be used as an input for modelling
Churn	Churn is the target variable indicating customer churn (1 for churned, 0 for active)

4.0) Scope

1) Jupyter notebook and faker library:
   These tools are used to generate sample dataset for this case study

2) Data Talend Integration:
   Used for for the smooth integration of many datasets.

3) Talend Data Preparation:
   enables efficient preparation and understand the data.

4) SAS Enterprise Miner:
   Used for ensemble techniques application, decision tree analysis, and advanced analytics.

4.1) Process of generating data

i) Create dataset for customer details

```
[ ] import csv
    import random

    # Set the seed for reproducibility
    random.seed(42)

    # Generate a random Asian_Country data with at least 1000 rows
    data = []

    for customer_id in range(1, 1001):
        age = random.randint(18, 65)
        gender = random.choice(['Male', 'Female', 'Non-Binary'])
        country = random.choice(['China', 'India', 'Indonesia', 'Pakistan', 'Bangladesh', 'Japan', 'Philippines', 'Vietnam', 'Turkey', 'Iran'])

        # Introduce missing values intentionally
        if random.random() < 0.1:  # 10% chance of missing values
            asian_country = None

        data.append([customer_id, asian_country])

    # Write the dataset to a CSV file
    with open('asian_countries.csv', 'w', newline='') as csvfile:
        csvwriter = csv.writer(csvfile)
        # Write header
        csvwriter.writerow(['customer_id', 'asian_countries'])
        # Write data
        csvwriter.writerows(data)
```

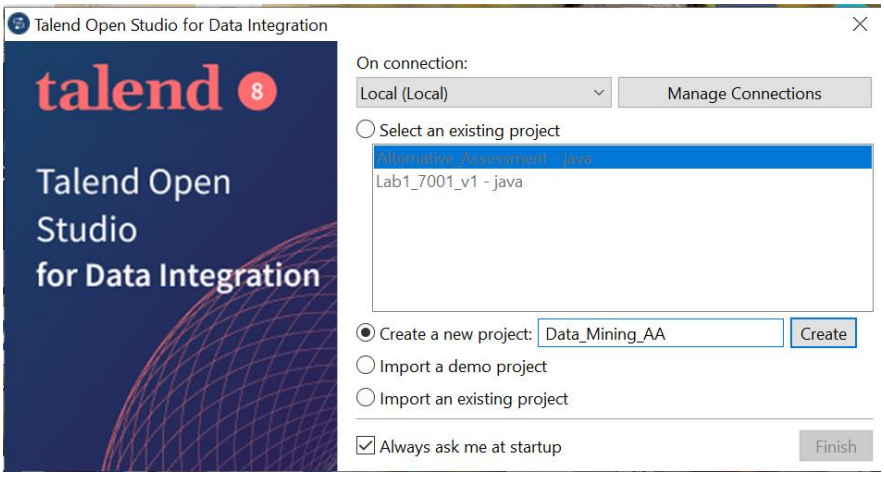
ii) Create dataset for customer behavior

```
import csv
from faker import Faker
import random
from datetime import datetime, timedelta

fake = Faker()

# Set the seed for reproducibility
random.seed(42)

# Generate a synthetic dataset with at least 1000 rows
data = []

for _ in range(1000):
    customer_id = _ + 1
    membership_level = random.choice(['Bronze', 'Silver', 'Gold', 'Platinum'])
    total_purchases = random.randint(1, 50)
    total_spent = random.uniform(50, 2000)
    favorite_category = random.choice(['Electronics', 'Clothing', 'Home Goods'])
    last_purchase_date = (datetime.now() - timedelta(days=random.randint(1, 365))).strftime('%Y-%m-%d')
    occupation = fake.job()
    website_visits_frequency = random.choice(['Daily', 'Weekly', 'Bi-weekly', 'Monthly'])

    # Introduce missing values intentionally
    if random.random() < 0.1:  # 10% chance of missing values
        age = None
        location = None
        membership_level = None

    # Determine item related to the favorite category
    if favorite_category == 'Electronics':
        item = random.choice(['Laptop', 'Smartphone', 'Headphones', 'Smartwatch'])
    elif favorite_category == 'Clothing':
        item = random.choice(['T-Shirt', 'Jeans', 'Sneakers', 'Dress'])
    elif favorite_category == 'Home Goods':
        item = random.choice(['Couch', 'Coffee Maker', 'Bedding Set', 'Lamp'])
    else:
        item = 'Unknown'
```

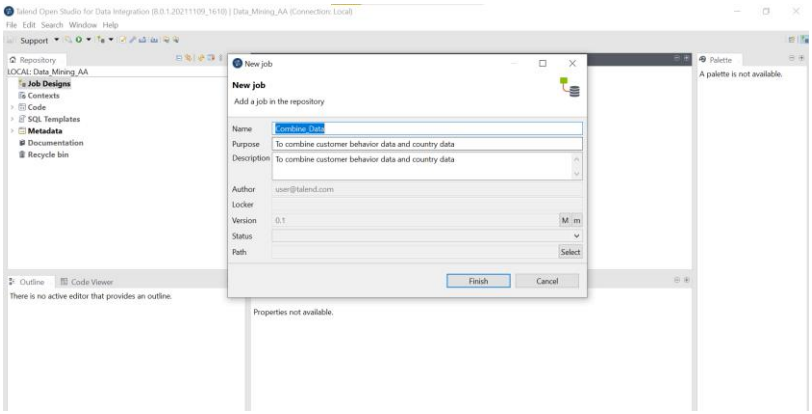
4.2) Data Talend Integration

This tool is used to merge both CSV files into once file to be used for the analysis.

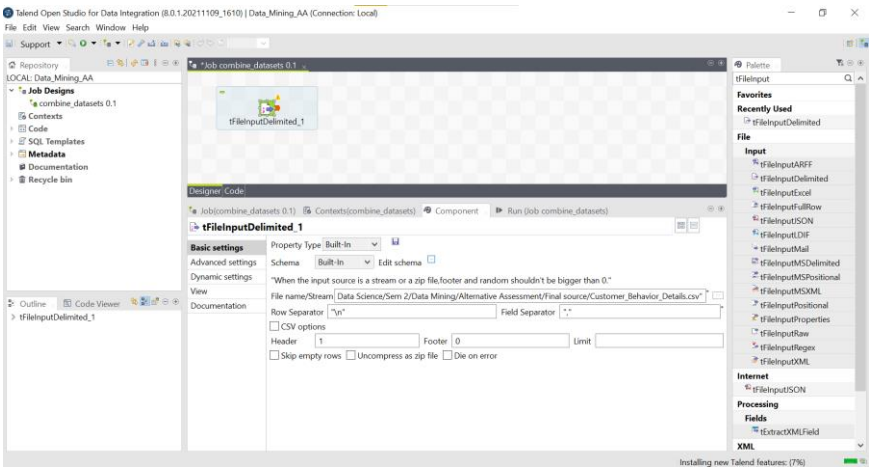
i) Create new project



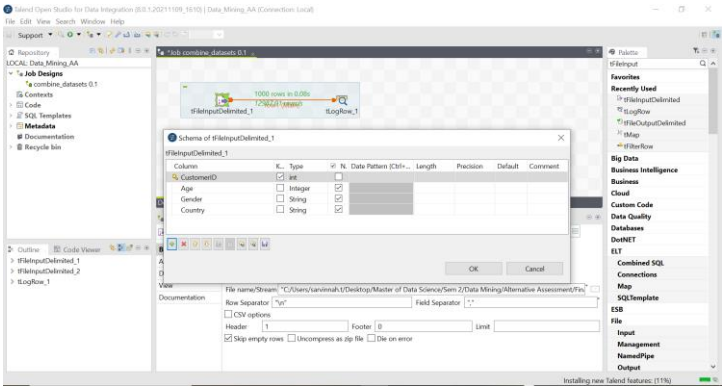
ii) Create new job to combine dataset



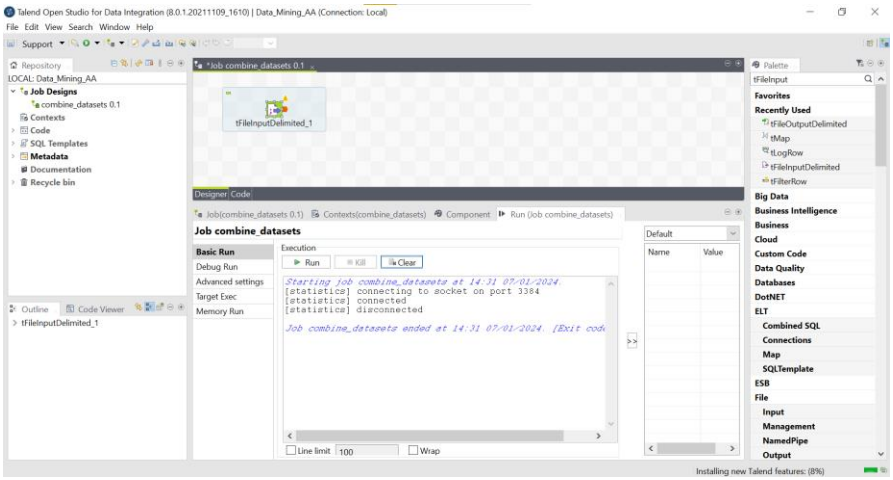
iii) Import Customer_Personal_Details dataset



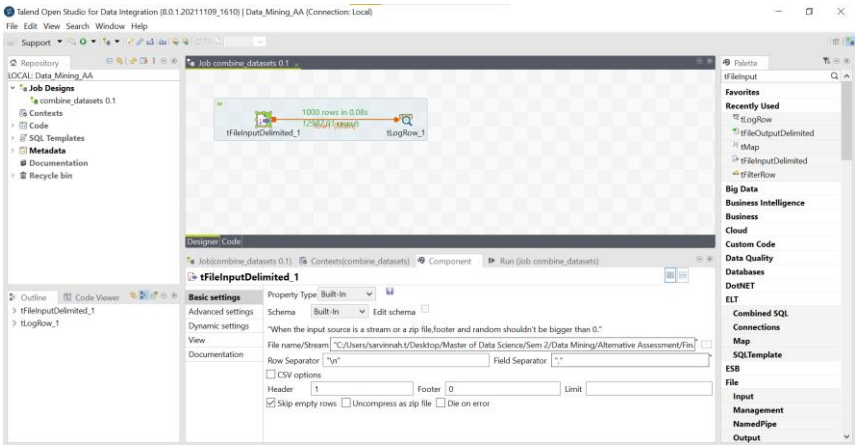
iv) Create schema for this file



v) Run the job

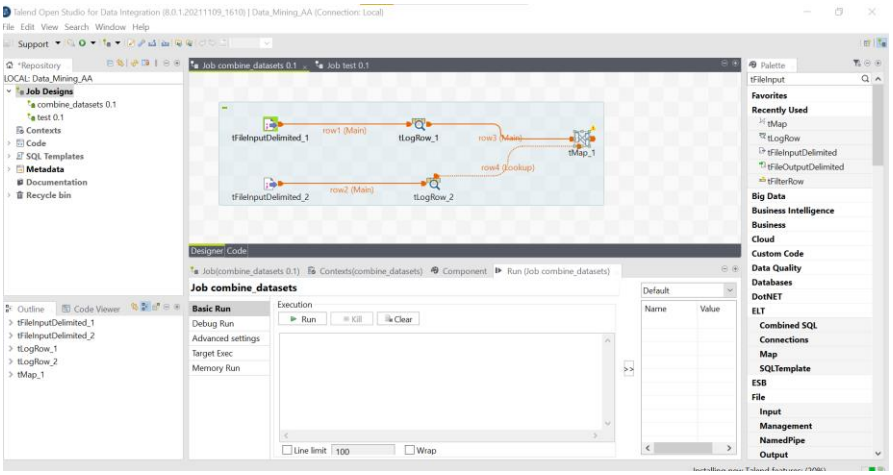


vi) View the log by importing log component

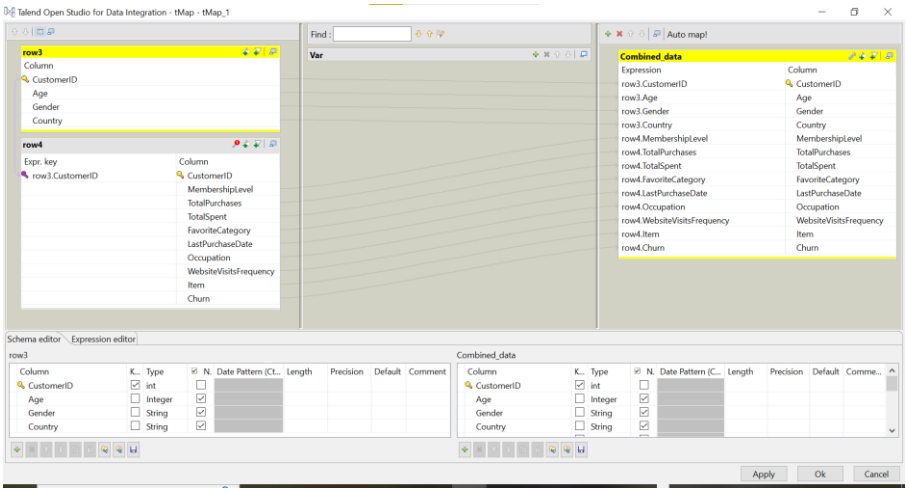


vii) Repeat the same steps for customer_behavior_dataset

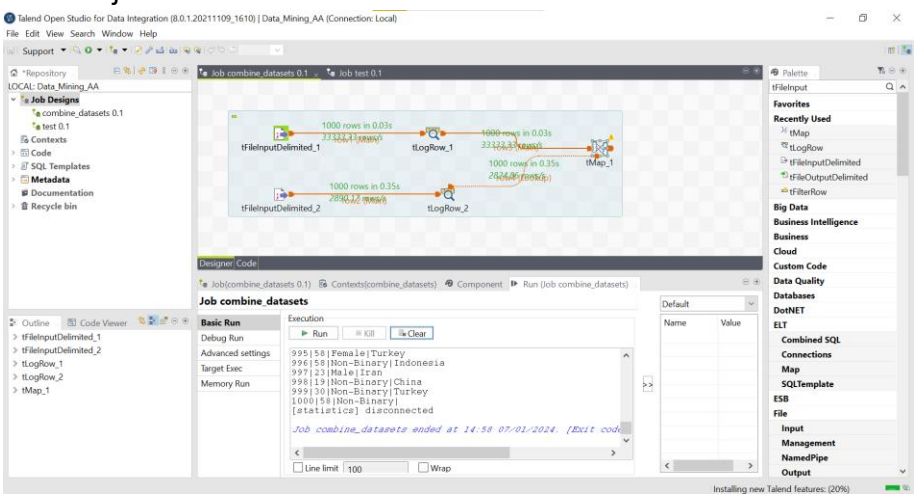
viii) Merge the datasets using the tMap component



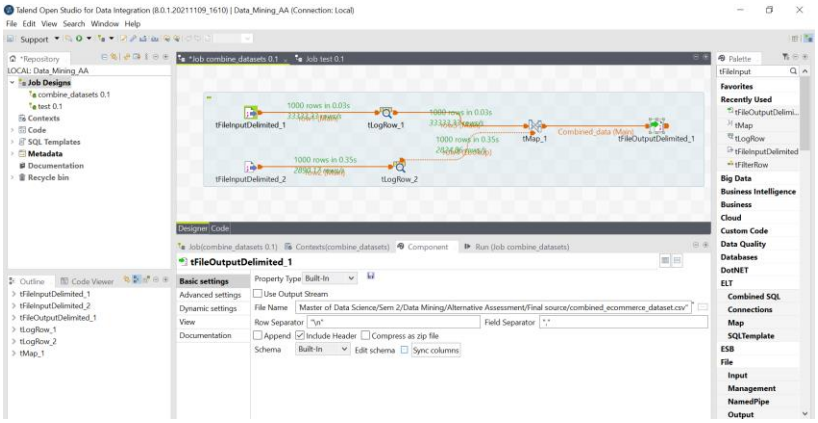
ix) Configure the tMap



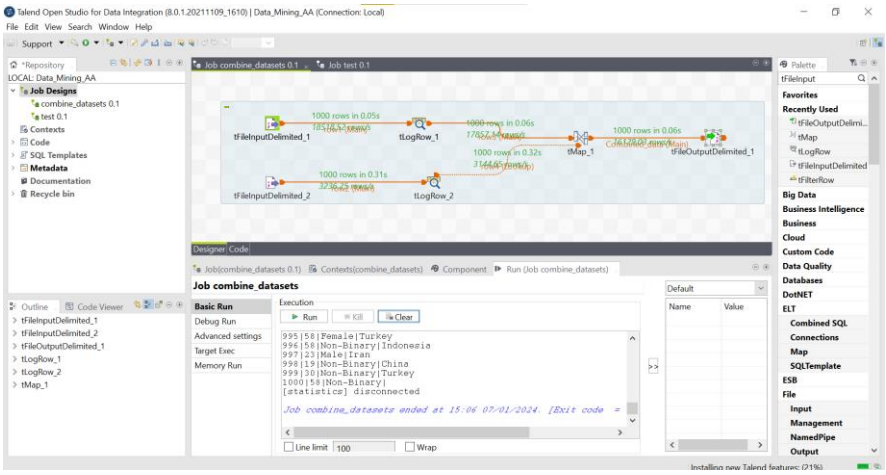
x) Run the job



xi) Export the merged file



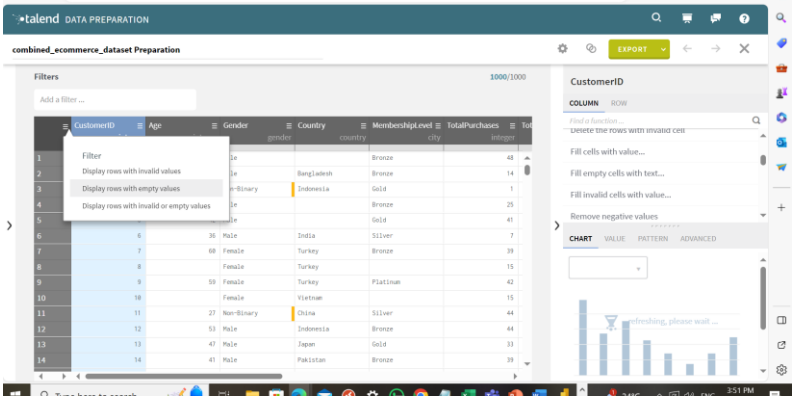
xii) Run the program



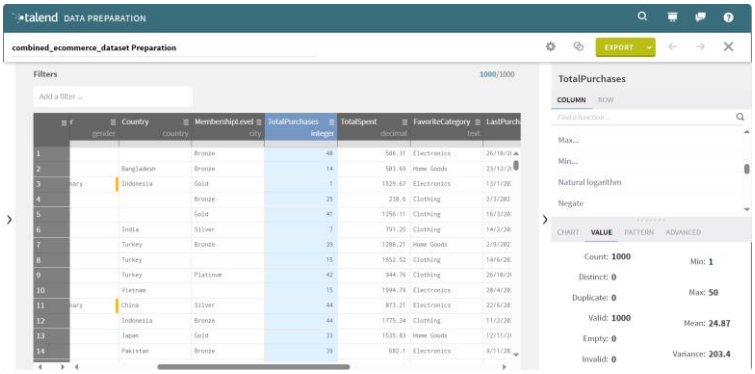
4.3) Talend Data Preparation

This tool is used to Explore the combined file that is exported from talend Data Integration. This tool helped in understanding each column better which helps during analysis phase in SAS. Below are few sample of exploration:

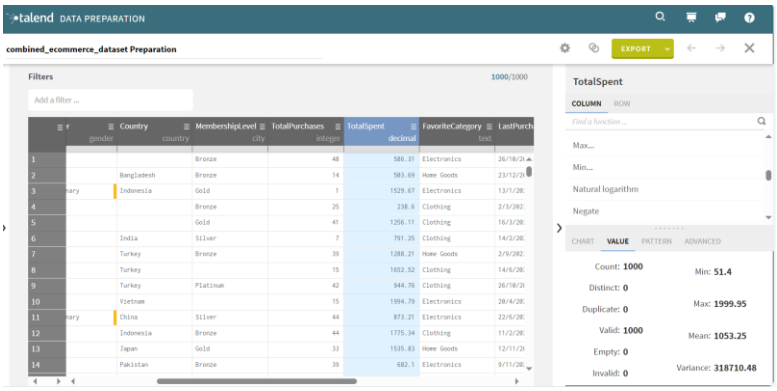
i) Explore data with empty values



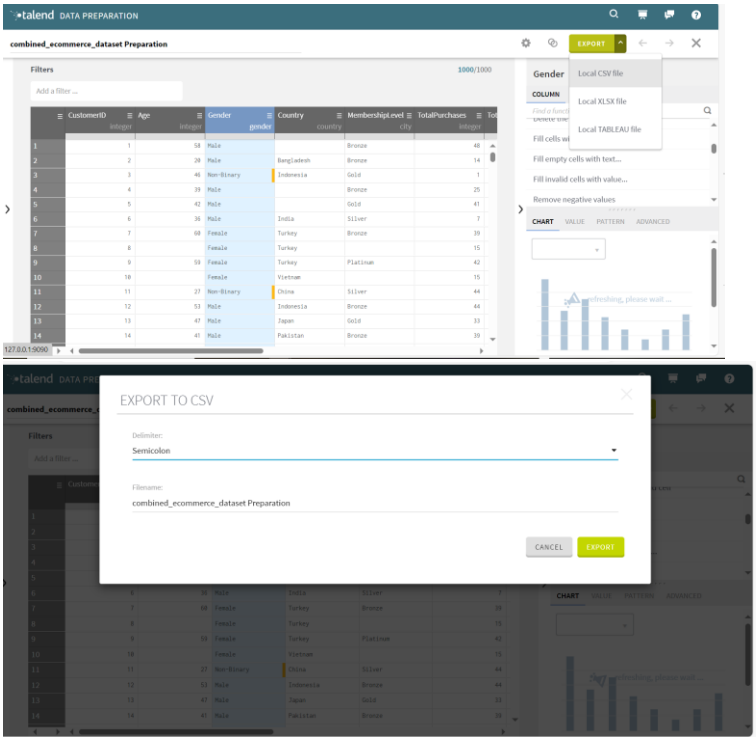
ii) Explore total purchase column



iii) Total spent column



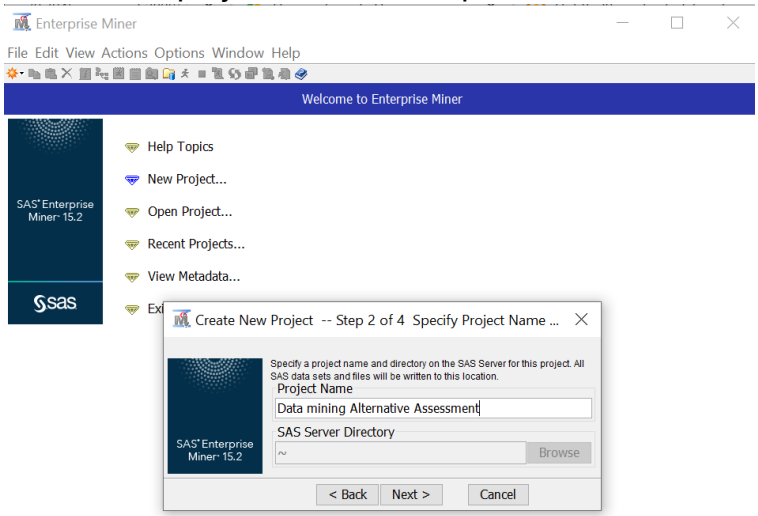
iv) Export the file to csv file



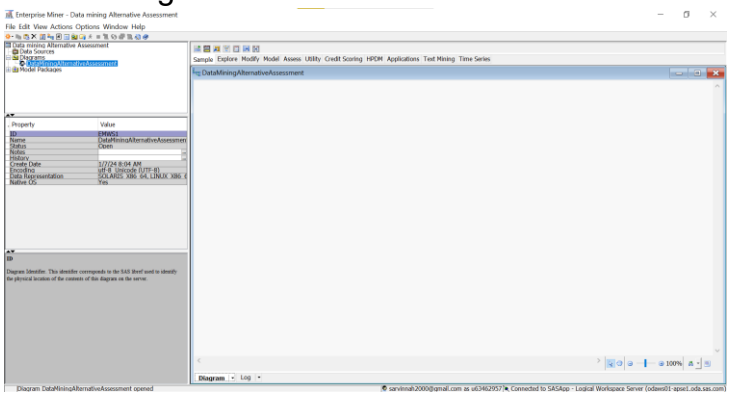
4.4) SAS Enterprise Miner

A) Prepare the environment for analysis of the ecommerce dataset.

i) Create new project in SAS enterprise miner

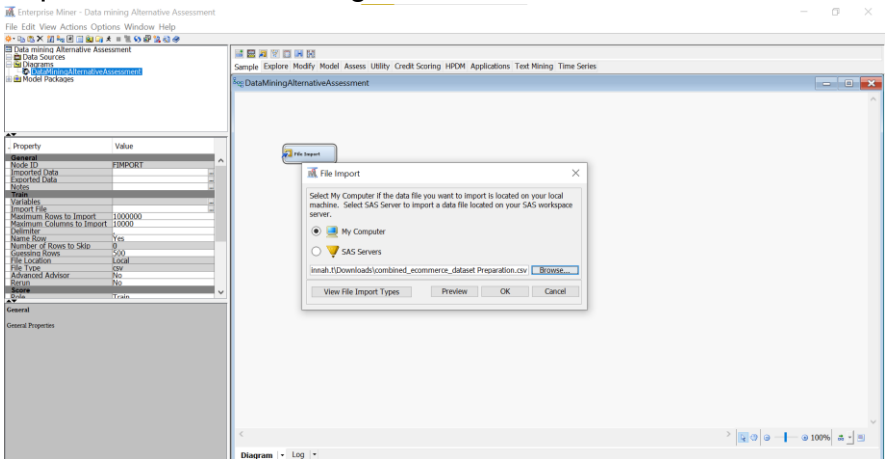


ii) Create Diagram



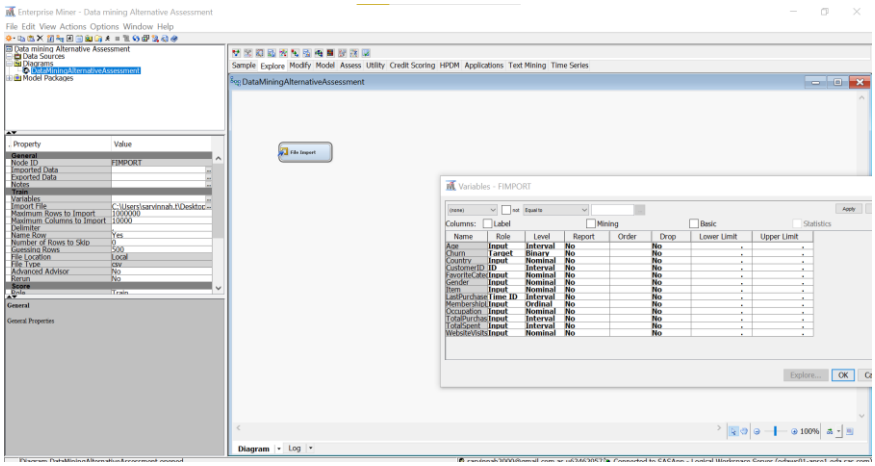
B) Data Import and Preprocessing: Import your dataset into SAS Enterprise Miner, handle missing values, and specify variable roles

i) Import dataset to the diagram



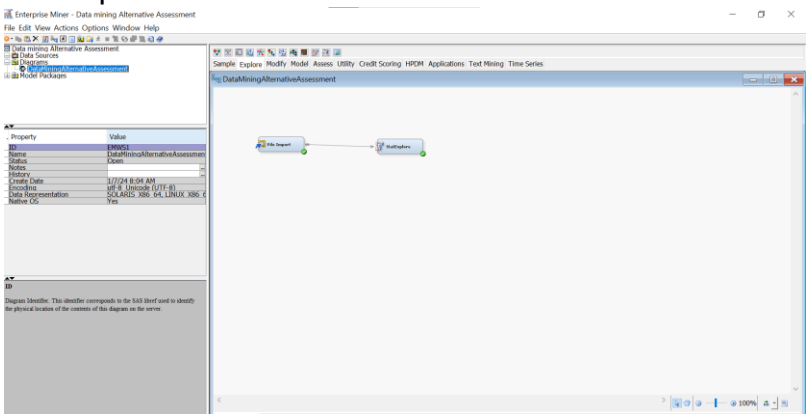
ii) Assigning role to each variable

For this case study, none of the variables were removed as each variable was required to do the analysis. Churn column is given target role for this case study to understand customer behavior.



iii) Handle missing variables

- identify attributes with missing values using StatExplore node



Output:

Class Variable Summary Statistics
(maximum 500 observations printed)

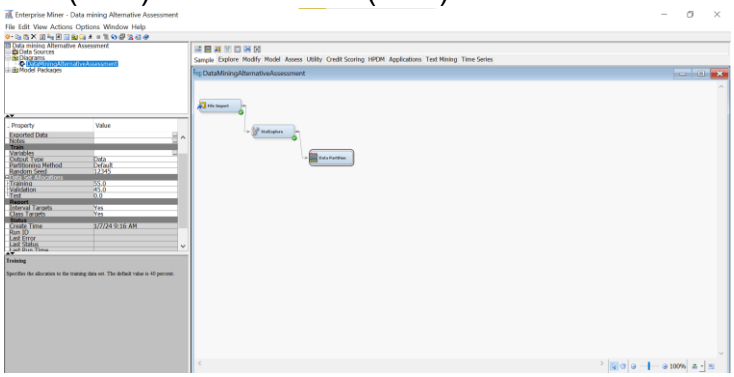
Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Country	INPUT	11	94	India	10.20	Japan	9.80
TRAIN	FavoriteCategory	INPUT	3	0	Clothing	34.40	Electronics	33.80
TRAIN	Gender	INPUT	3	0	Non-Binary	36.00	Female	32.00
TRAIN	Item	INPUT	12	0	Smartphone	9.30	Jeans	9.10
TRAIN	MembershipLevel	INPUT	5	98	Bronze	23.90	Silver	23.00
TRAIN	Occupation	INPUT	502	0	Archivist	0.60	Commissioning editor	0.60
TRAIN	WebsiteVisitsFrequency	INPUT	4	0	Daily	25.60	Weekly	25.30
TRAIN	Churn	TARGET	2	0	TRUE	50.50	FALSE	49.50

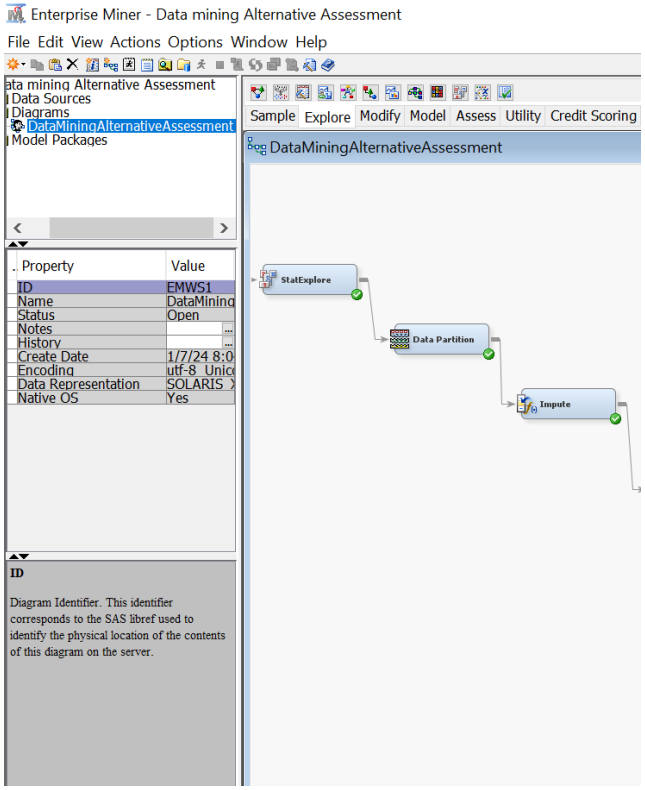
Column Name	Missing Values	
Country	94	Data Role=TRAIN Variable Name=Country Target / Target Level / Number of Levels / Missing / Mode / Mode Percentage / Mode2 / Mode2 Percentage Churn / FALSE / 11 / 39 / India / 10.71 / Indonesia / 10.10 Churn / TRUE / 11 / 55 / / 10.89 / Japan / 10.89 _OVERALL_ / / 11 / 94 / India / 10.20 / Japan / 9.80

MembershipLevel	98	Data Role=TRAIN Variable Name=MembershipLevel Target / Target Level / Number of Levels / Missing / Mode / Mode Percentage / Mode2 / Mode2 Percentage Churn / FALSE / 5 / 50 / Bronze / 24.24 / Silver / 23.03 Churn / TRUE / 5 / 48 / Bronze / 23.56 / Silver / 22.97 _OVERALL_ / 5 / 98 / Bronze / 23.90 / Silver / 23.00
Age	98	Data Role=TRAIN Variable=Age Target / Target Level / Median / Missing / Non Missing Churn / FALSE / 41 / 50 / 445 Churn / TRUE / 42 / 48 / 457 _OVERALL_ / 41 / 98 / 902

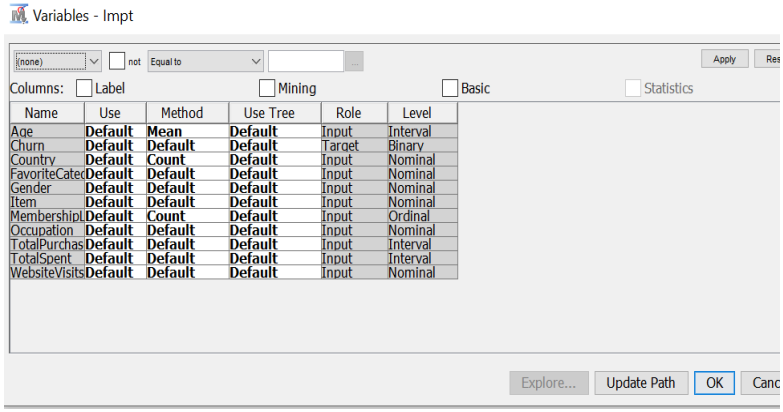
- The previous step shows that the dataset has missing values. Next, data partition is included to the determine the split test (0.0) and training data(55.0) and validation (45.0)



- Impute method is used to replace all the null values in continuous columns with mean values



Edit variables to impute:
Since the empty value is in Age, Country and Membership column only, the method for this columns are altered in which for numerical column the empty value replaced with mean and for non-numerical columns I have used count method.

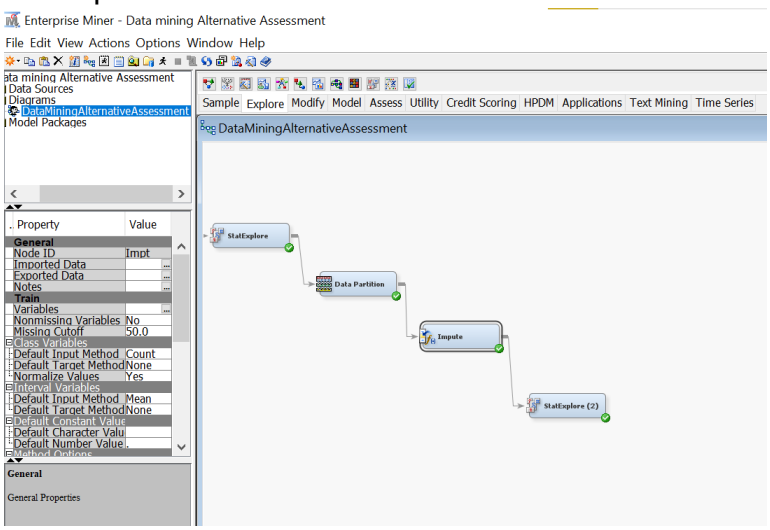


Output:

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	MEAN	IMP_Age	41.252988048	INPUT	INTERVAL		48
Country	COUNT	IMP_Country	Philippines	INPUT	NOMINAL		52
MembershipLevel	COUNT	IMP_MembershipLevel	Bronze	INPUT	ORDINAL		48

- Verify the missing value after impute method using statExplorer



Output:
Stat shows that there are no missing values

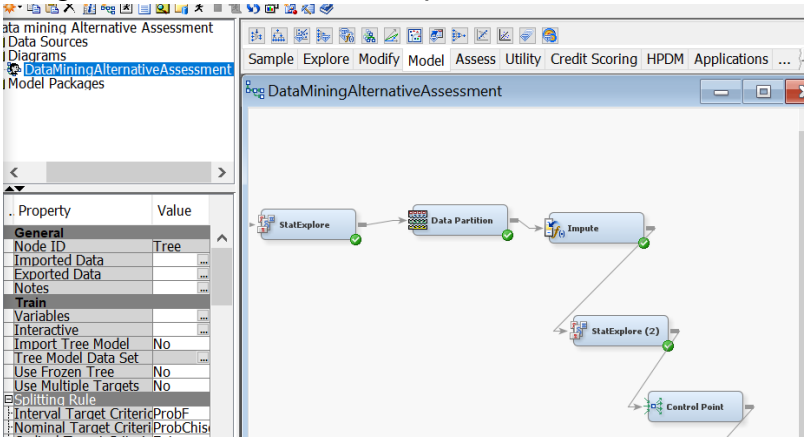
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

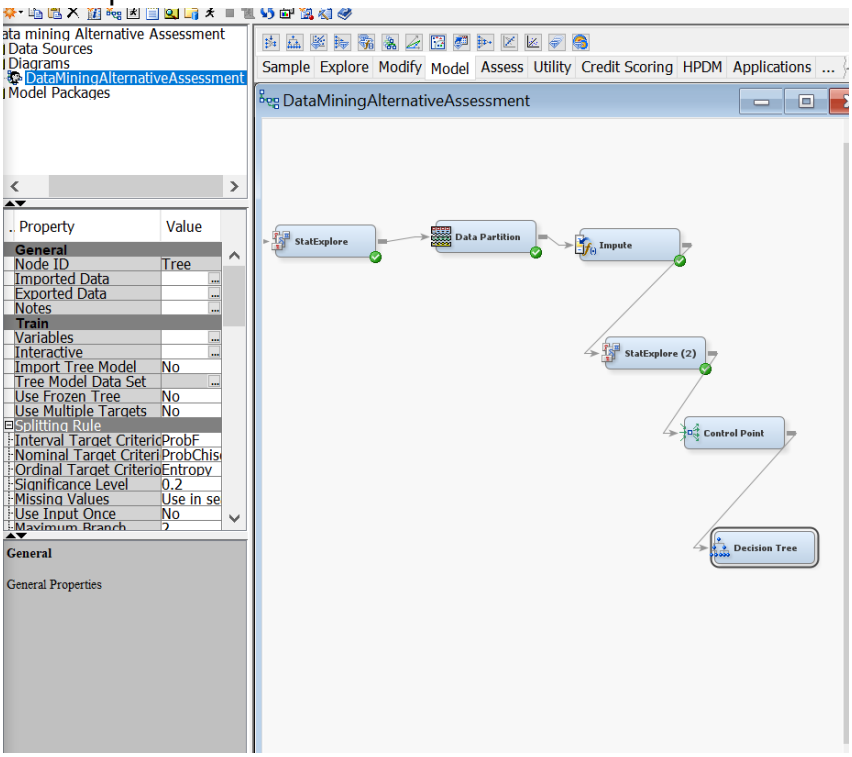
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	FavoriteCategory	INPUT	3	0	Clothing	36.55	Electronics	33.45
TRAIN	Gender	INPUT	3	0	Non-Binary	37.45	Male	31.64
TRAIN	IMP_Country	INPUT	10	0	Philippines	19.64	China	9.82
TRAIN	IMP_MembershipLevel	INPUT	4	0	Bronze	33.64	Silver	24.00
TRAIN	Item	INPUT	12	0	Jeans	11.09	Dress	9.82
TRAIN	Occupation	INPUT	374	0	Artist	0.91	Archivist	0.73
TRAIN	WebsiteVisitsFrequency	INPUT	4	0	Weekly	26.55	Daily	26.18
TRAIN	Churn	TARGET	2	0	TRUE	50.55	FALSE	49.45

C) Decision Tree Analysis: Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.

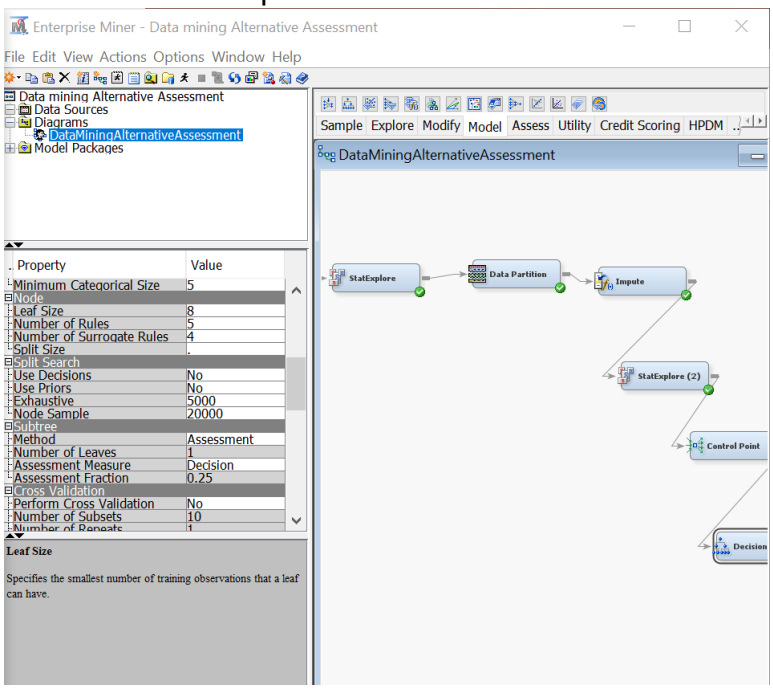
- Drag and connect the control point



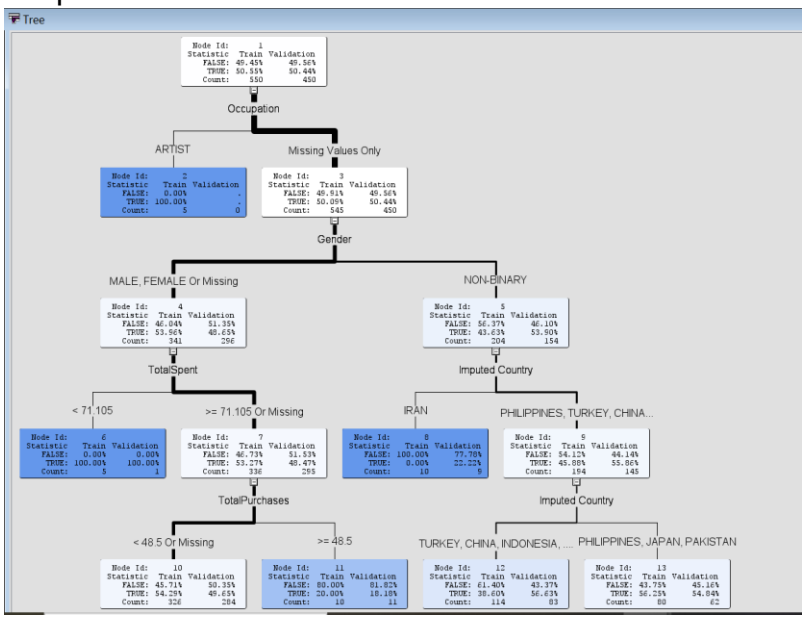
- Then, decision tree node is connected to the central point node



Configure properties for decision tree: maximum depth property to 10 to allow the tree to be trained with up to 10 generations of splits. Next the leaf size to 8 which means each leaf will require minimum of 8 training observations. Finally, set the number of surrogate rules to 4. This is to ensure that if the splitting relies on a input with missing values, then the software is configured to us up to four surrogate rules to determine splits in non-leaf nodes. The assessment measure property is set to Decision in which the decision tree will seek to maximize profit in the validation data.



Output:



Results:

Event Classification Table

Data Role=TRAIN Target=Churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
46	88	184	232

Data Role=VALIDATE Target=Churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
51	52	171	176

Data Role=TRAIN Target Variable=Churn Target Label=' '

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	42.1987	1.42199	1.42199	71.8750	71.8750	28	0.71875
10	27.0234	1.11286	1.27023	56.2500	64.2045	27	0.56250
15	21.7144	1.11286	1.21714	56.2500	61.5211	28	0.56250
20	18.4513	1.08420	1.18451	54.8015	59.8717	27	0.54801
25	16.2124	1.07417	1.16212	54.2945	58.7401	28	0.54294
30	14.7732	1.07417	1.14773	54.2945	58.0126	27	0.54294
35	13.7060	1.07417	1.13706	54.2945	57.4732	28	0.54294
40	12.9342	1.07417	1.12934	54.2945	57.0831	27	0.54294
45	12.3113	1.07417	1.12311	54.2945	56.7683	28	0.54294
50	11.8308	1.07417	1.11831	54.2945	56.5254	27	0.54294
55	11.4229	1.07417	1.11423	54.2945	56.3192	28	0.54294
60	11.0952	1.07417	1.11095	54.2945	56.1536	27	0.54294
65	10.8075	1.07417	1.10808	54.2945	56.0082	28	0.54294
70	10.5697	1.07417	1.10570	54.2945	55.8880	27	0.54294
75	10.3560	1.07417	1.10356	54.2945	55.7799	28	0.54294
80	8.4816	0.79811	1.08482	40.3407	54.8325	27	0.40341
85	6.5598	0.76360	1.06560	38.5965	53.8611	28	0.38596
90	4.9126	0.76360	1.04913	38.5965	53.0285	27	0.38596
95	3.3839	0.76360	1.03384	38.5965	52.2559	28	0.38596
100	0.0000	0.34452	1.00000	17.4139	50.5455	27	0.17414

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Depth	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations	Mean Posterior Probability
5	12.6036	1.12604	1.12604	56.8022	56.8022	23	0.58152
10	10.7006	1.08711	1.10701	54.8387	55.8423	22	0.56250
15	9.2710	1.06474	1.09271	53.7103	55.1212	23	0.55825
20	6.6188	0.98421	1.06619	49.6479	53.7833	22	0.54294
25	4.9502	0.98421	1.04950	49.6479	52.9415	23	0.54294
30	3.8862	0.98421	1.03886	49.6479	52.4048	22	0.54294
35	3.0906	0.98421	1.03091	49.6479	52.0035	23	0.54294
40	2.5199	0.98421	1.02520	49.6479	51.7156	22	0.54294
45	2.0554	0.98421	1.02055	49.6479	51.4813	23	0.54294
50	1.7001	0.98421	1.01700	49.6479	51.3020	22	0.54294
55	1.3960	0.98421	1.01396	49.6479	51.1486	23	0.54294
60	1.1535	0.98421	1.01154	49.6479	51.0263	22	0.54294
65	0.9390	0.98421	1.00939	49.6479	50.9181	23	0.54294
70	0.7632	0.98421	1.00763	49.6479	50.8294	22	0.54294
75	0.6038	0.98421	1.00604	49.6479	50.7490	23	0.54294
80	0.9700	1.06596	1.00970	53.7716	50.9337	22	0.45018
85	1.6477	1.12255	1.01648	56.6265	51.2756	23	0.38596
90	2.2239	1.12255	1.02224	56.6265	51.5663	22	0.38596

Assessment Score Distribution

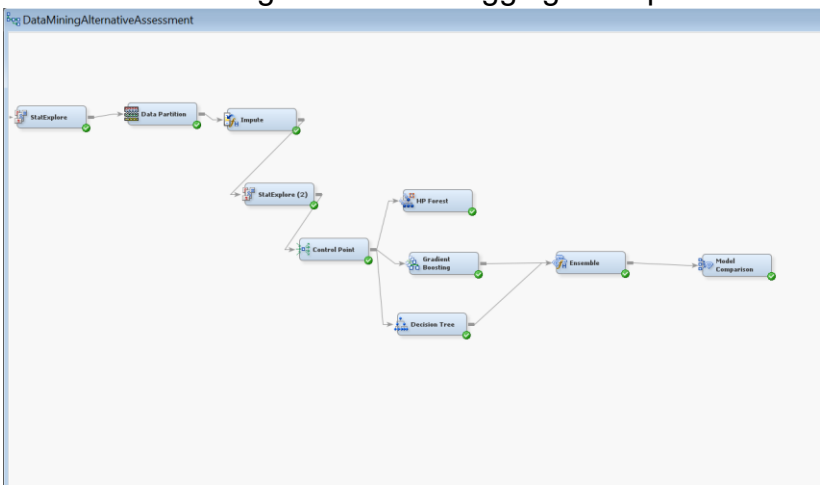
Data Role=TRAIN Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	10	0	1.00000	1.8182
0.55-0.60	45	35	0.56250	14.5455
0.50-0.55	177	149	0.54294	59.2727
0.35-0.40	44	70	0.38596	20.7273
0.15-0.20	2	8	0.20000	1.8182
0.00-0.05	0	10	0.00000	1.8182

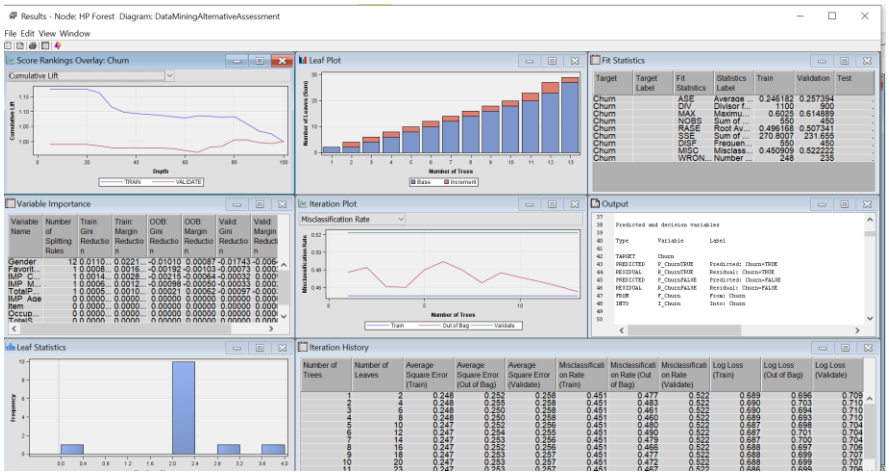
Data Role=VALIDATE Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.95-1.00	1	0	1.00000	0.2222
0.55-0.60	34	28	0.56250	13.7778
0.50-0.55	141	143	0.54294	63.1111
0.35-0.40	47	36	0.38596	18.4444
0.15-0.20	2	9	0.20000	2.4444
0.00-0.05	2	7	0.00000	2.0000

D) Ensemble Methods: Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.



i) Random Forest (Bagging)



Output:

Classification Table

Data Role=TRAIN Target Variable=Churn Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FALSE	FALSE	55.8252	42.2794	115	20.9091
TRUE	FALSE	44.1748	32.7338	91	16.5455
FALSE	TRUE	45.6395	57.7206	157	28.5455
TRUE	TRUE	54.3605	67.2662	187	34.0000

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FALSE	FALSE	46.1039	31.8386	71	15.7778
TRUE	FALSE	53.8961	36.5639	83	18.4444
FALSE	TRUE	51.3514	68.1614	152	33.7778
TRUE	TRUE	48.6486	63.4361	144	32.0000

Event Classification Table

Data Role=TRAIN Target=Churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
91	115	157	187

Data Role=VALIDATE Target=Churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
83	71	152	144

Assessment Score Distribution

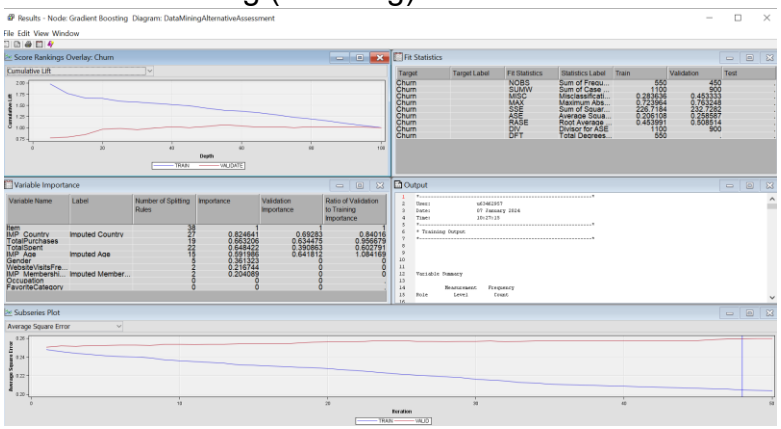
Data Role=TRAIN Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.55-0.60	154	126	0.56837	50.9091
0.50-0.55	33	31	0.54888	11.6364
0.40-0.45	85	95	0.42973	32.7273
0.35-0.40	6	20	0.39464	4.7273

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.55-0.60	118	121	0.56754	53.1111
0.50-0.55	26	31	0.54888	12.6667
0.40-0.45	70	61	0.42949	29.1111
0.35-0.40	13	10	0.39373	5.1111

ii) Gradient Boosting (Boosting)



Output:

Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	Item		38	1.00000	1.00000	1.00000
2	IMP_Country	Imputed Country	27	0.82464	0.69283	0.84016
3	TotalPurchases		19	0.66321	0.63448	0.95668
4	TotalSpent		22	0.64842	0.39086	0.60279
5	IMP_Age	Imputed Age	15	0.59199	0.64181	1.08417
6	Gender		5	0.36132	0.00000	0.00000
7	WebsiteVisitsFrequency		2	0.21674	0.00000	0.00000
8	IMP_MembershipLevel	Imputed MembershipLevel	2	0.20409	0.00000	0.00000

--
* Report Output
--

Fit Statistics

Target=Churn Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	550.00	450.000
SUMW	Sum of Case Weights Times Freq	1100.00	900.000
MISC	Misclassification Rate	0.28	0.453
MAX	Maximum Absolute Error	0.72	0.763
SSE	Sum of Squared Errors	226.72	232.728
ASE	Average Squared Error	0.21	0.259
RASE	Root Average Squared Error	0.45	0.509
DIV	Divisor for ASE	1100.00	900.000
DFT	Total Degrees of Freedom	550.00	.

Classification Table

Data Role=TRAIN Target Variable=Churn Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FALSE	FALSE	71.4815	70.9559	193	35.0909
TRUE	FALSE	28.5185	27.6978	77	14.0000
FALSE	TRUE	28.2143	29.0441	79	14.3636
TRUE	TRUE	71.7857	72.3022	201	36.5455

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
FALSE	FALSE	54.6341	50.2242	112	24.8889
TRUE	FALSE	45.3659	40.9692	93	20.6667
FALSE	TRUE	45.3061	49.7758	111	24.6667
TRUE	TRUE	54.6939	59.0308	134	29.7778

Event Classification Table

Data Role=TRAIN Target=Churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
77	193	79	201

Data Role=VALIDATE Target=Churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
93	112	111	134

Assessment Score Distribution

Data Role=TRAIN Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.75-0.80	2	0	0.77329	0.3636
0.70-0.75	19	0	0.72083	3.4545
0.65-0.70	28	6	0.66982	6.1818
0.60-0.65	40	12	0.62591	9.4545
0.55-0.60	61	24	0.57225	15.4545
0.50-0.55	51	37	0.52389	16.0000
0.45-0.50	50	52	0.47592	18.5455
0.40-0.45	21	62	0.42843	15.0909
0.35-0.40	5	37	0.37964	7.6364
0.30-0.35	0	19	0.32869	3.4545
0.25-0.30	1	18	0.28020	3.4545
0.20-0.25	0	5	0.23274	0.9091

Data Role=VALIDATE Target Variable=Churn Target Label=' '

Posterior Probability Range	Number of Events	Number of Nonevents	Mean Posterior Probability	Percentage
0.75-0.80	1	1	0.75927	0.4444
0.70-0.75	3	8	0.71884	2.4444
0.65-0.70	11	11	0.67421	4.8889
0.60-0.65	27	21	0.62324	10.6667
0.55-0.60	39	38	0.57432	17.1111
0.50-0.55	53	32	0.52710	18.8889
0.45-0.50	28	41	0.47774	15.3333
0.40-0.45	35	34	0.42721	15.3333
0.35-0.40	19	18	0.37780	8.2222
0.30-0.35	7	10	0.32484	3.7778
0.25-0.30	3	8	0.27966	2.4444
0.20-0.25	1	1	0.24945	0.4444

Conclusion:

To sum up, the E-Commerce Customer Behaviour Analysis carried out using SAS Enterprise Miner has revealed important information that is necessary for negotiating the complexities of the online marketplace. In addition to improving our knowledge of consumer behaviour, the study's emphasis on product affinity analysis, churn prediction, and customer segmentation has given organisations useful information. Businesses may now take a more customer-centric approach, proactively addressing possible churn indicators and capitalising on personalised advice. This has important consequences for strategic decision-making. The study has the potential to influence future initiatives, cultivate consumer loyalty, and ultimately support the long-term survival of e-commerce firms in a fast changing market, which will have a lasting impact on organisations who use these findings.

The model comparison between Gradient Boosting and Random Forest is presented through the Assessment Score Distribution for both the training and validation datasets. The table displays the posterior probability ranges along with the number of events (Churn) and nonevents in each range. The mean posterior probability represents the average probability within each range, offering insights into the

model's predictive behavior. In the training dataset, both models exhibit higher probabilities in the 0.50-0.60 range, indicating a focus on instances with moderate likelihoods of churn. The validation dataset demonstrates a similar trend, emphasizing the stability of the models across different datasets. Careful examination of the percentage distribution provides a nuanced understanding of the models' performance across varying probability ranges. This comparison aids in gauging the effectiveness of both Gradient Boosting and Random Forest in predicting customer churn and informs decisions on selecting the most suitable model for deployment in real-world scenario