UNIVERSITI MALAYA

WQD7005 Data Mining

Alternative Assessment 1 (AA1)

Name:	Woo Wai Hong
Matric No:	22065374
Occurrence:	Group 1, Thursday 6pm to 9pm
Lecturer:	Dr. Teh Ying Wah

WQD7005 Data Mining Alternative Assessment 1 (AA1)

WOO WAI HONG (22065374)

Case Study: E-Commerce Customer Behaviour Analysis

Link to GitHub repository: https://github.com/woogamanga/WQD7005-AA1-Woo-Wai-Hong-22065374

Data mining in e-commerce involves extracting valuable insights and patterns from large datasets to enhance decision-making processes within online retail environments. The benefits of data mining in e-commerce include improved customer segmentation, personalized recommendations, and optimized marketing strategies based on historical customer behaviour. Customer churn prediction is particularly crucial in e-commerce as it helps identify customers at risk of leaving, allowing businesses to implement targeted retention strategies and ultimately enhance customer loyalty and profitability.

Objective of AA1

The objectives of AA1 are as follows:

1. To understand the distribution of key attributes in the synthetic dataset created which will inform relevant data preprocessing steps needed.
2. To predict customer churn in the e-commerce domain using tree-based classifiers.
3. To determine the best performing tree-based classifier for the use case of customer churn classification in the e-commerce domain.
4. To improve my mastery over the use Talend Data Integration (DI), Talend Data Preparation (DP) and SAS Enterprise Miner (EM) tools in conducting an end-to-end data mining project for AA1 using the SEMMA methodology.

Role of Talend Data Integration (DI)

Talend Data Integration is an open-source ETL (Extract, Transform, Load) tool that enables organizations to connect, transform, and integrate data from various sources to meet their business needs.

The role of Talend DI in this assessment are as follows:

1. To perform data integration on the 2 synthetic datasets, sales_data.csv and customer_data.csv, using tools in Talend DI such as ‘tFileInputDelimited’, ‘tMap’, ‘tFileOutputDelimited’.

Role of Talend Data Preparation (DP)

Talend Data Preparation is a user-friendly, self-service data preparation tool that empowers business users to clean, enrich, and transform raw data into actionable insights, facilitating efficient data management and analysis.

The role of Talend DP in this assessment are as follows:

1. To perform data cleaning on the integrated synthetic dataset which has inconsistencies in columns such as ‘LastPurchaseDate’, ‘Gender’ and ‘Location’.

Role of SAS Enterprise Miner (EM)

SAS Enterprise Miner is an advanced analytics and data mining tool that empowers organizations to build, deploy, and refine predictive models, uncover patterns in data, and make informed, data-driven decisions.

The role of SAS EM in this assessment are as follows:

1. To perform data preprocessing by means of mode imputation of the column ‘Returns’ which has missing values.
2. To perform data preprocessing by means of dropping columns which are irrelevant to the analysis.

3. To perform exploratory data analysis in order to understand the underlying distribution of the key attributes in the integrated dataset.
4. To partition the integrated dataset into training, validation, and test sets in preparation for modeling using tree-based classifiers.
5. To train 3 tree-based classifiers to perform classification of e-commerce customer churn.
6. To evaluate the performance of the 3 tree-based classifiers using various performance metrics in order to determine the best performing classifier for the classification of customer churn in the e-commerce domain.

1 Dataset

2 synthetic datasets, sales_data.csv and customer_data.csv were created from the synthetic dataset obtained from Kaggle at https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis?select=ecommerce_customer_data_large.csv for this assessment.

The original dataset obtained from Kaggle was also a synthetic dataset created with the Python Faker library.

2 Dataset Description

Some names in the original dataset were changed to fulfil the requirements of this assessment and some additional columns were added. For those additional columns, synthetic data is used as well to populate the rows. The final set of attributes and their respective descriptions in the integrated and cleaned dataset are as follows:

Attribute	Description
CustomerID	ID of the customer
LastPurchaseDate	Last date by which customer purchased a particular product.

PurchasedProductCategory	Category of the product purchased.
ProductPrice	Price of the product purchased.
TotalPurchases	Total quantity purchased for a particular product.
TotalSpent	Total amount of money spent on a purchase.
PaymentMethods	Payment method used to complete the purchase.
Returns	Did the customer return the product or not?
CustomerName	Name of the customer
Age	Age of the customer
Gender	Sex of the customer
Location	Location at which the purchase was made
MembershipLevel	Membership level of the customer
Churn	Did the customer churn from the e-commerce marketplace?

3 Data Import

Data import corresponds to the Sample step in the SEMMA methodology and it includes in this assessment the following:

- Downloading the original synthetic dataset from Kaggle.
- Creating 2 synthetic datasets, sales_data.csv and customer_data.csv from the original dataset.
- Performing data integration to combine the 2 synthetic datasets into a single, integrated dataset called integrated_ecom.csv using Talend DI.

- Performing random sampling using SAS EM to get a representative sample for data mining.
- Performing data partition using SAS EM to partition the dataset into training, validation and test sets in preparation for Modeling.

Data Integration using Talend DI

Figure 1 below shows the workflow in Talend DI used to combine the 2 synthetic datasets into a single, integrated dataset:

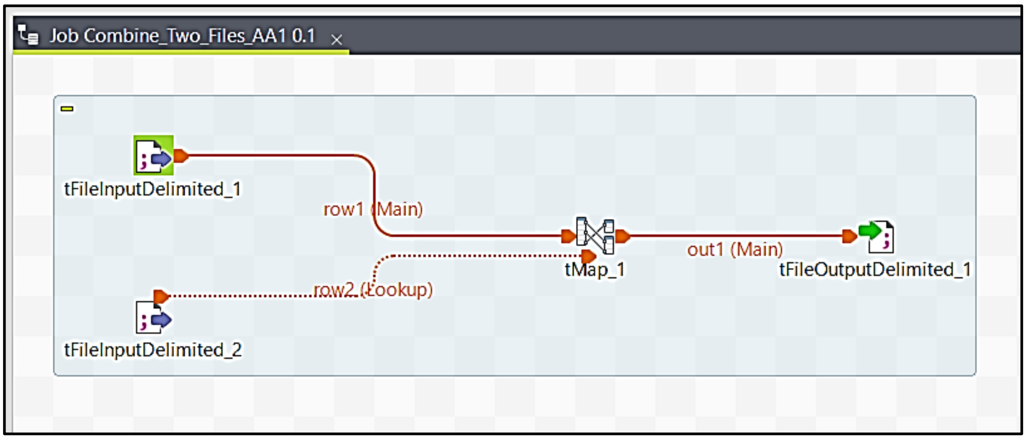


Figure 1: Data Integration Workflow in Talend DI

Figure 2 below shows the join conditions for the 2 synthetic datasets in Talend DI:

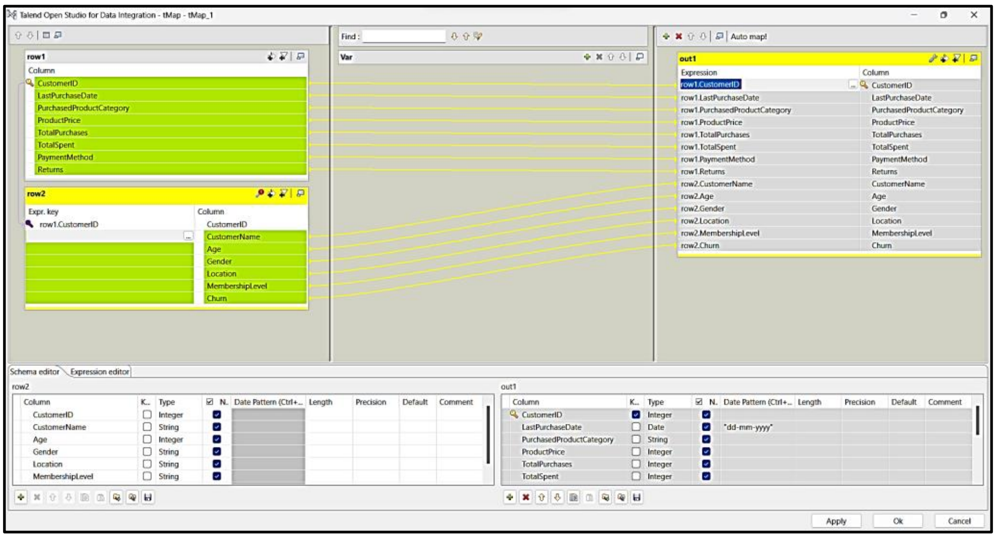


Figure 2: Mapping for join in t_Map component in Talend DI

Random Sampling using SAS EM

Figure 3 and 4 below shows the import process of data into SAS Studio and SAS EM respectively. Then Figure 5 subsequently shows the random sampling performed in SAS EM during Data Source creation where 10% of the integrated dataset is randomly selected as the representative sample for data mining in this assessment:

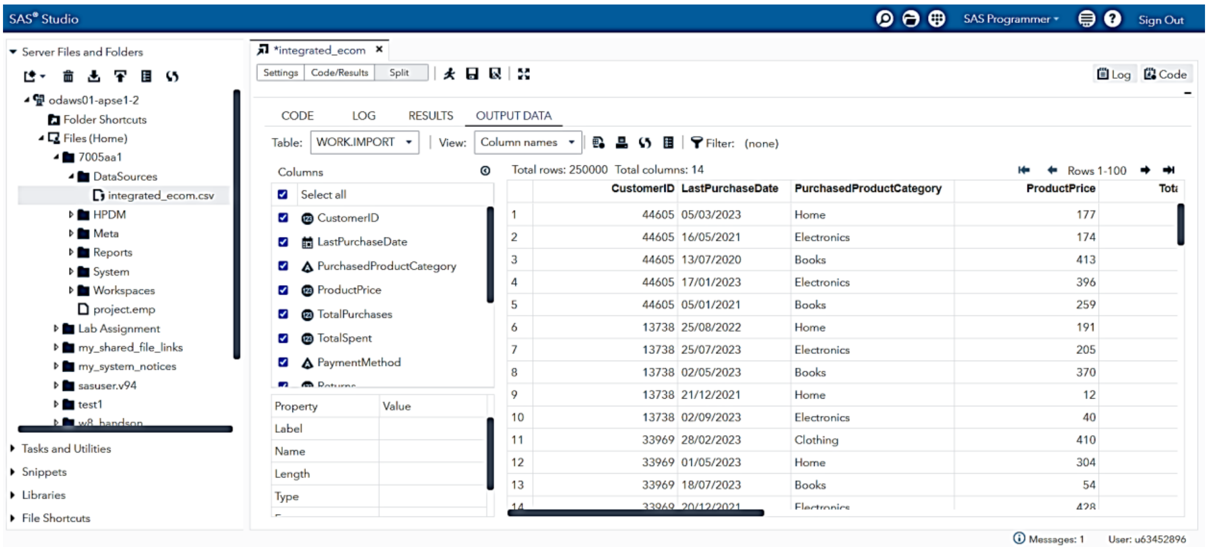


Figure 3: Import Data to SAS Studio

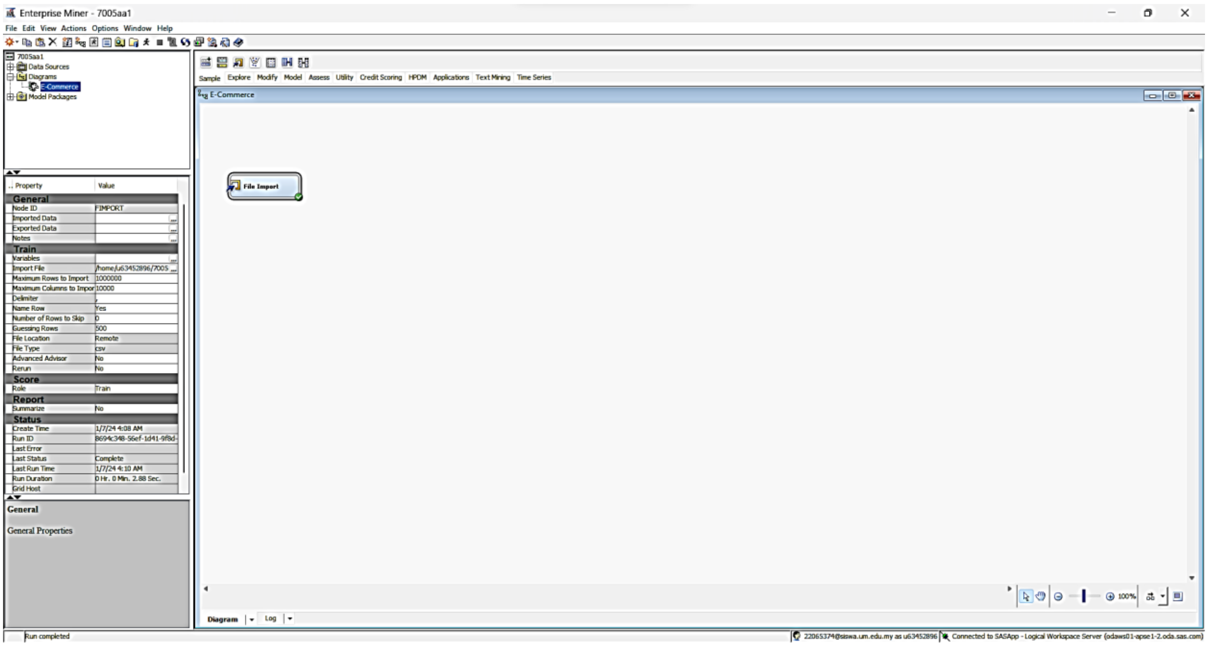


Figure 4: Import Data to SAS EM

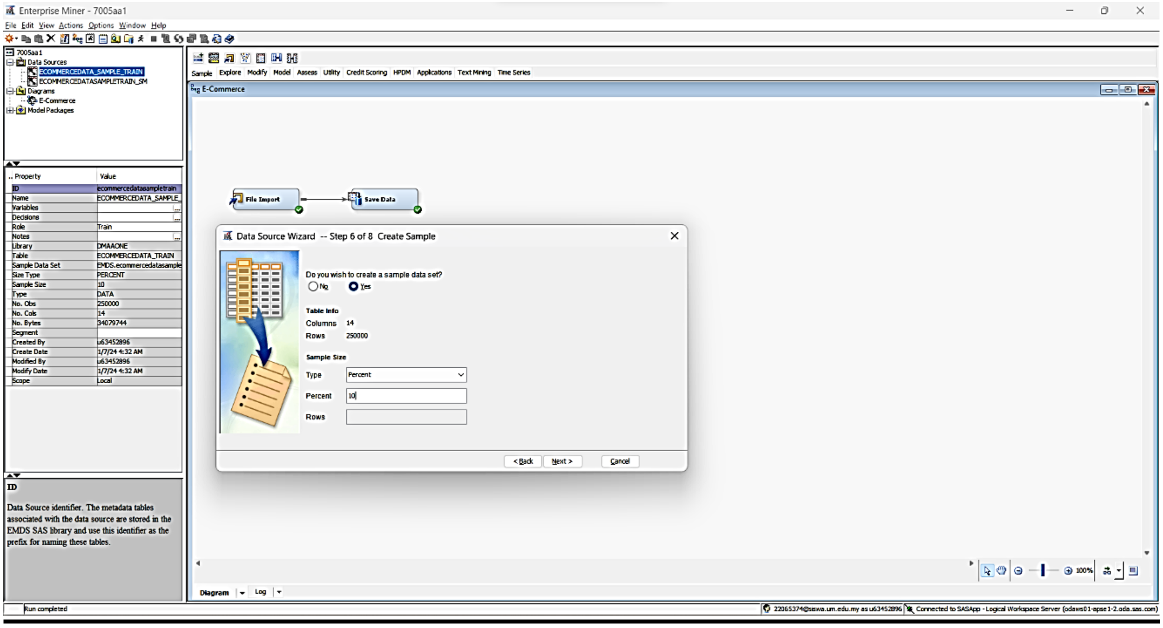


Figure 5: Random Sampling in SAS EM

Data Partition using SAS EM

Data Partition in SAS EM using 60/40/40 ratio for Train/Validate/Test

```
Partition Summary

                                        Number of
Type            Data Set              Observations

DATA        EMWS1.Impt_TRAIN              25000
TRAIN       EMWS1.Part_TRAIN              14999
VALIDATE    EMWS1.Part_VALIDATE            5000
TEST        EMWS1.Part_TEST                5001
```

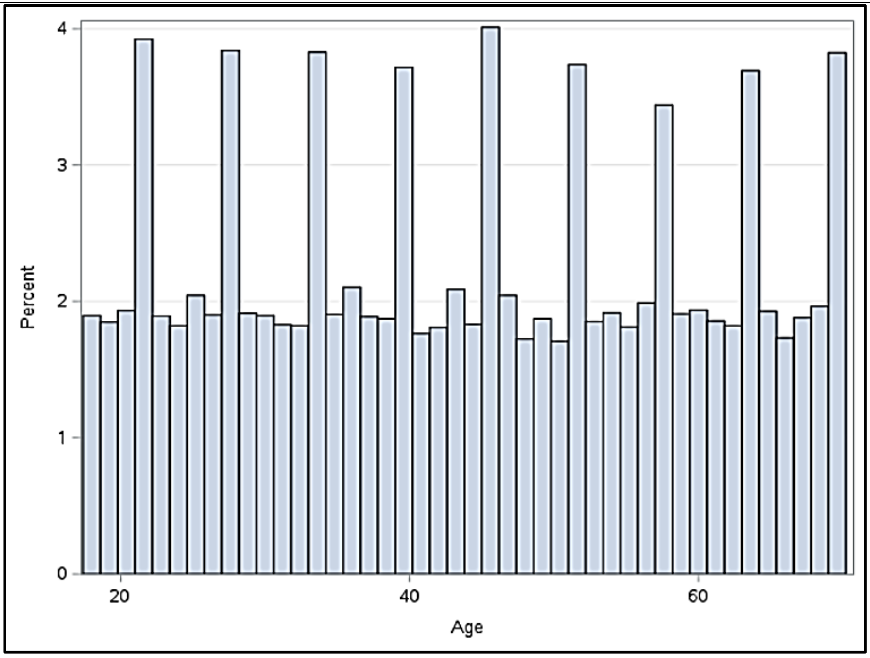
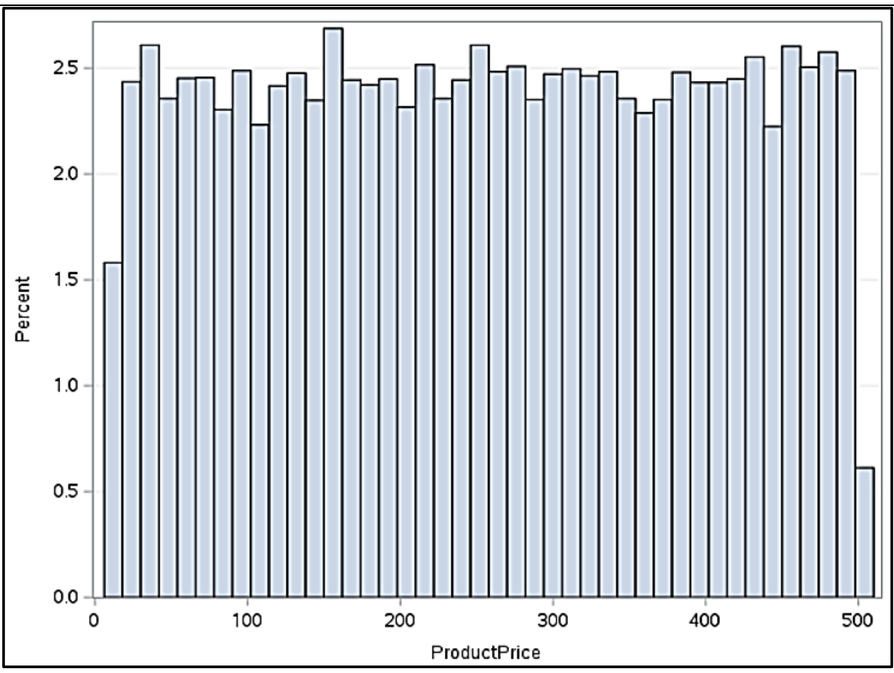
Figure 6: Data Partition on Sample in SAS EM

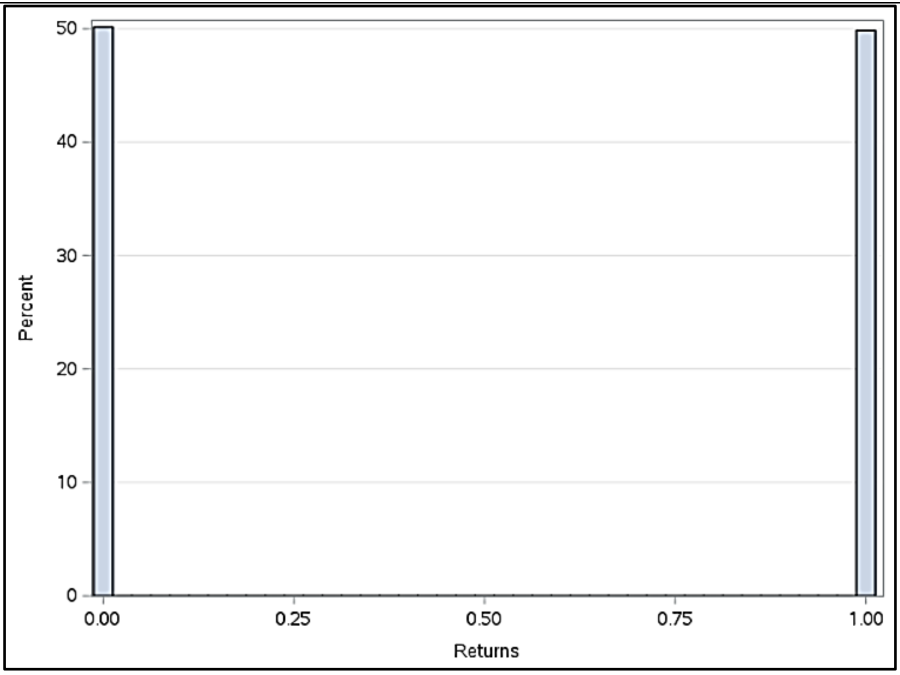
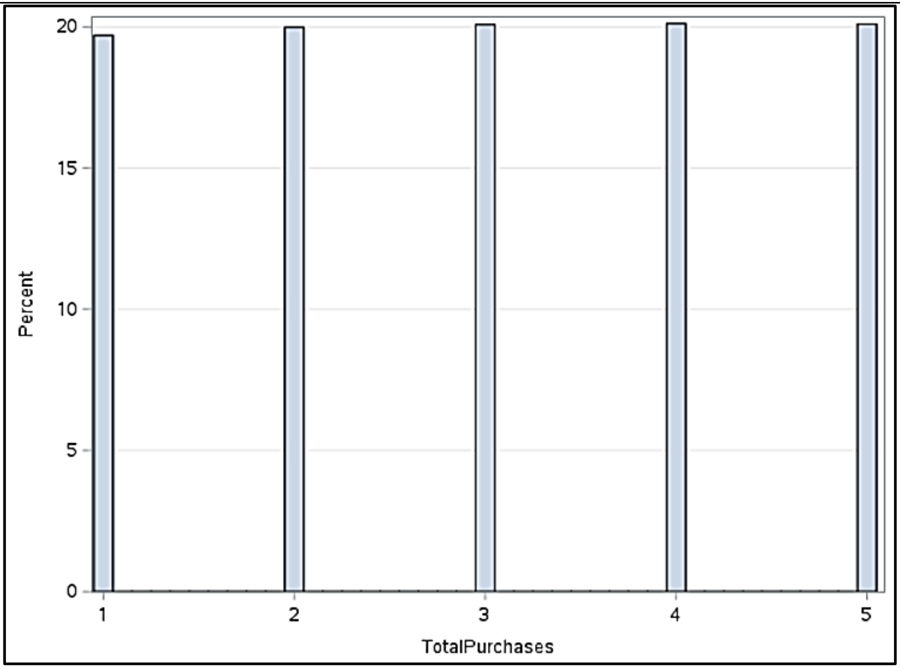
4 Exploratory Data Analysis (Visualization)

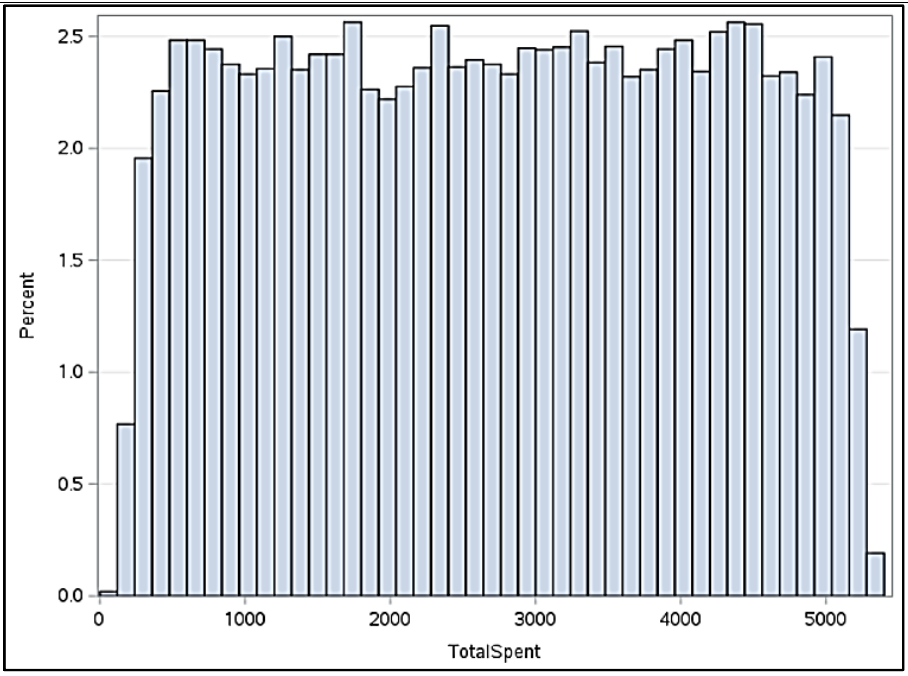
Exploratory Data Analysis or EDA corresponds to the Explore step in the SEMMA methodology and in this assessment, it includes the following:

- Histograms for numerical attributes
- Pie Charts and Bar Plots for categorical attributes
- Line Chart for temporal attribute

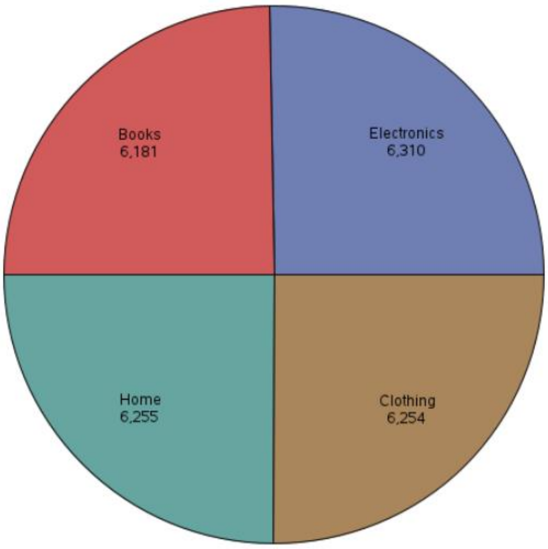
Histograms for Numerical Attributes

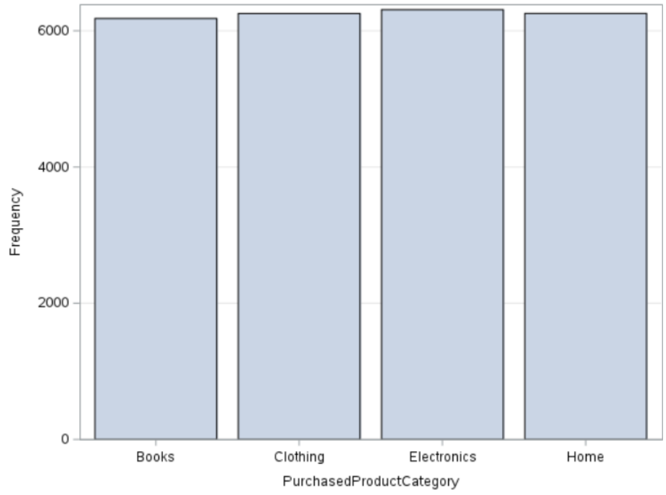
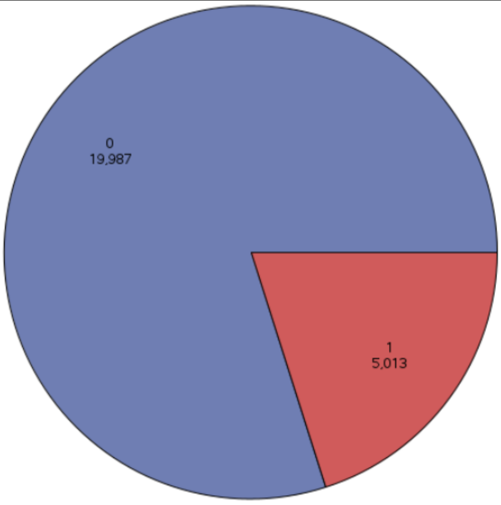
	Age attribute was found to exhibit non-normal distribution.
	ProductPrice was found to exhibit non-normal distribution.

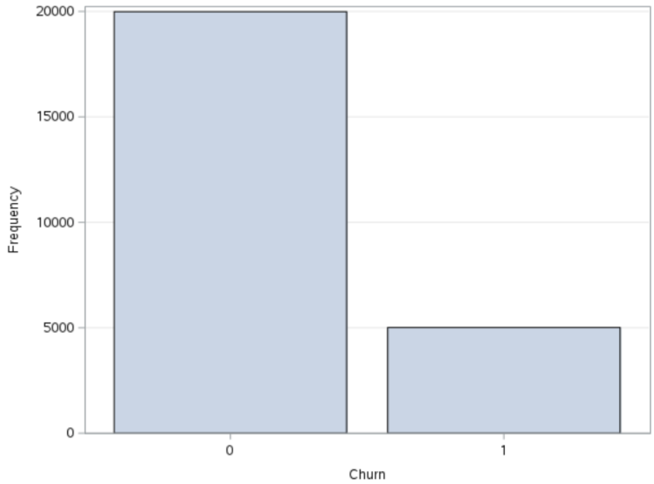
	Returns was found to exhibit multimodality.
	TotalPurchases was found to exhibit multimodality.

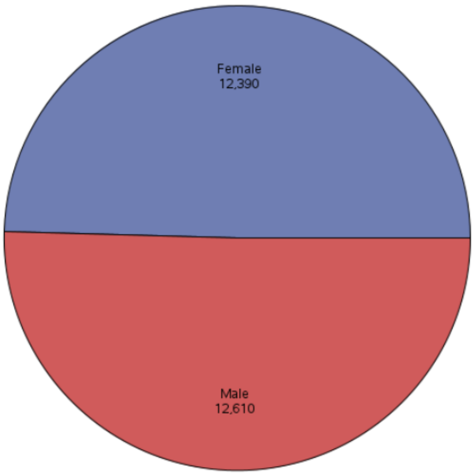
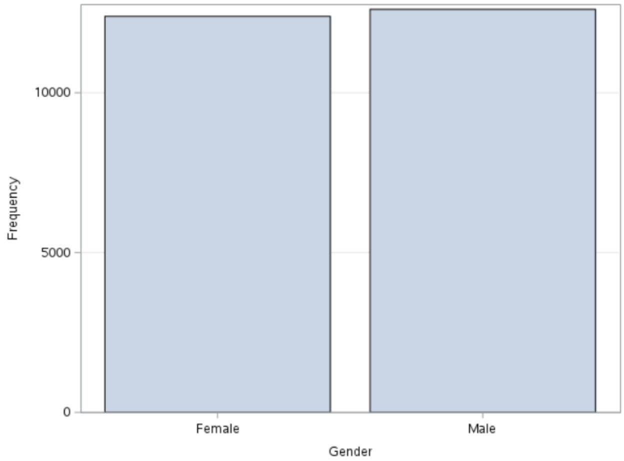
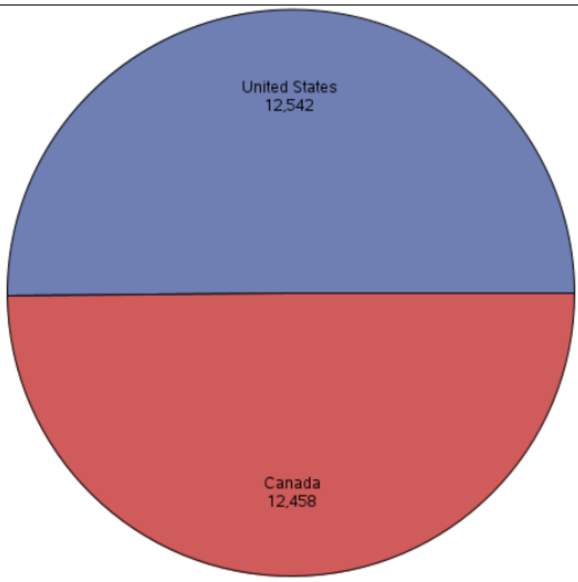
	TotalSpent was found to exhibit non-normal distribution.

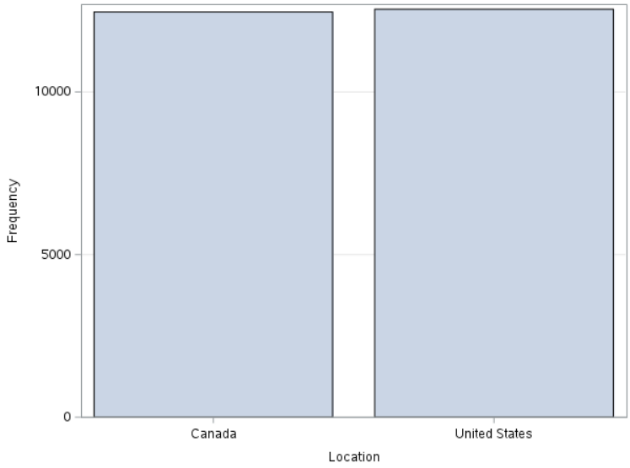
Pie Charts and Bar Plots for categorical attributes

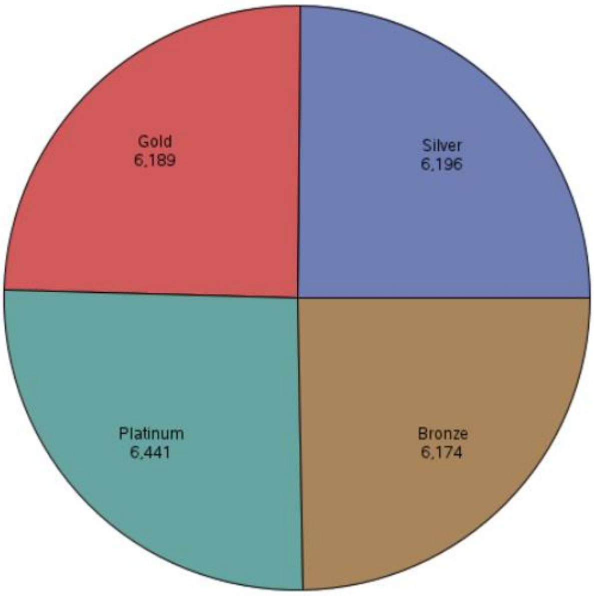
	Synthetic dataset has about the same proportion of each product category.

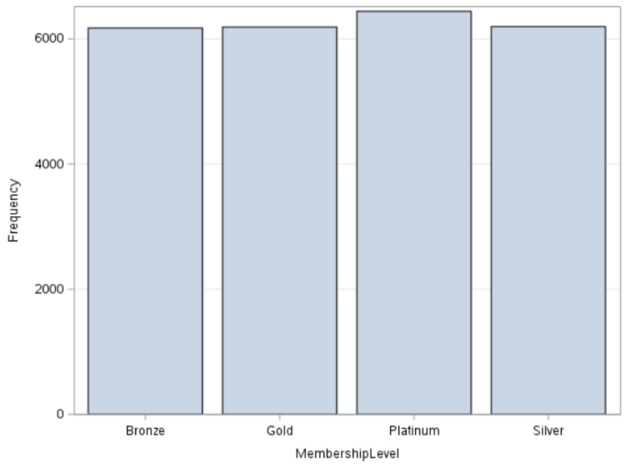
<table>
<tr>
<td>
</td>
<td></td>
</tr>
<tr>
<td>


</td>
<td>The integrated dataset is highly imbalanced with over 80% of the target being non-churn instances.</td>
</tr>
</table>

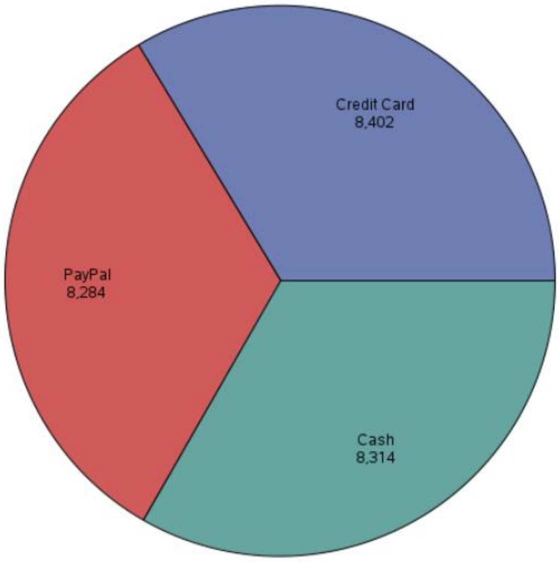
	Male and Female population are equal in the integrated dataset.
	United States and Canada have about equal proportions in the integrated dataset.

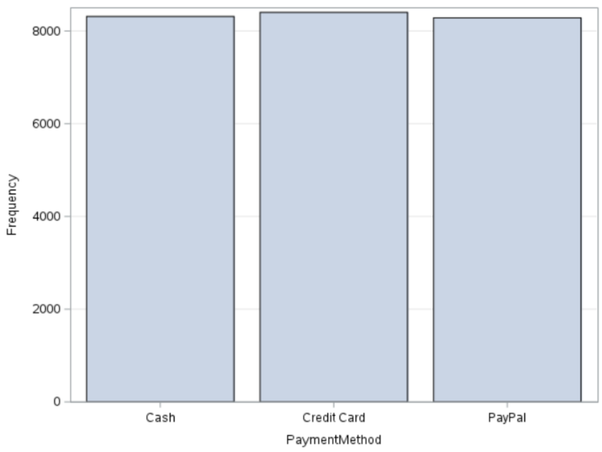
<table>
<tr>
<td>


</td>
<td></td>
</tr>
<tr>
<td>




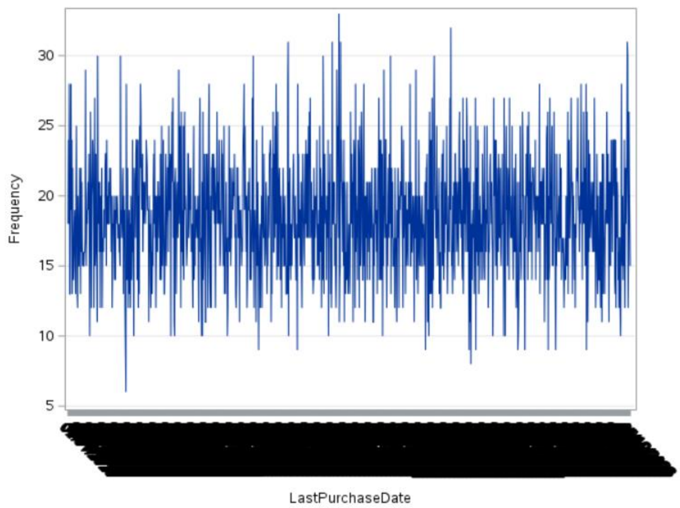
</td>
<td>Membership levels have about the same proportion for each category.</td>
</tr>
</table>





Payment methods are about equally proportioned in the integrated dataset.

Line Chart for Temporal Attribute



Purchasing behaviours tend to vary from customer to customer and does not follow any set patterns in terms of time as seen by the erratic line chart for LastPurchaseDate.

5 Data Preprocessing

Data preprocessing corresponds to the Modify step in the SEMMA methodology and in this assessment, it includes the following:

- Data cleaning on columns with inconsistencies such as 'LastPurchaseDate', 'Gender' and 'Location' using Talend DP.
- Mode imputation on 'Returns' column using SAS EM.
- Dropping of unnecessary columns such as 'CustomerID' and 'CustomerName using SAS EM.

<u>Data Cleaning using Talend DP</u>

Figures 7,8,9,10,11 and 12 show the outputs in Talend DP for 'LastPurchaseDate' before cleaning, 'LastPurchaseDate' after cleaning, 'Gender' before cleaning, 'Gender' after cleaning, 'Location' before cleaning and 'Location' after cleaning respectively.

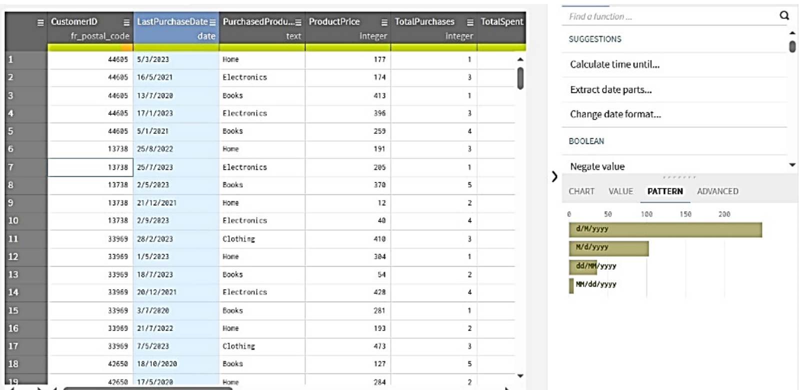


Figure 7: LastPurchaseDate before cleaning in Talend DP

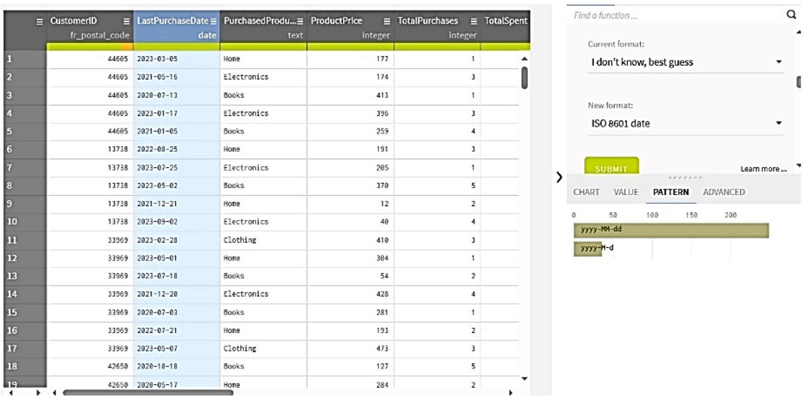


Figure 8: LastPurchaseDate after cleaning in Talend DP

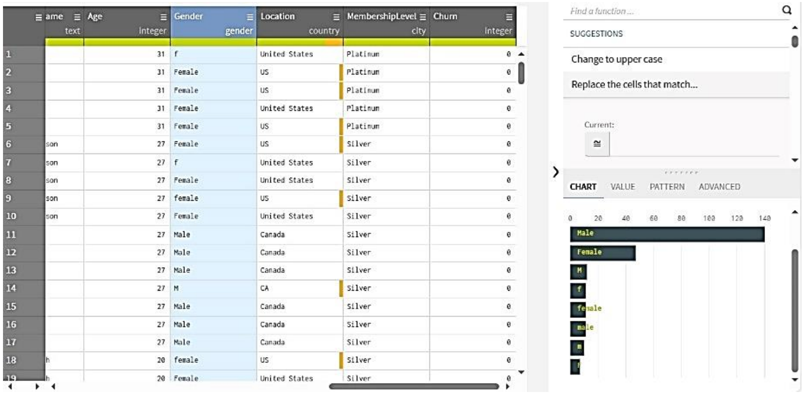


Figure 9: Gender before cleaning in Talend DP

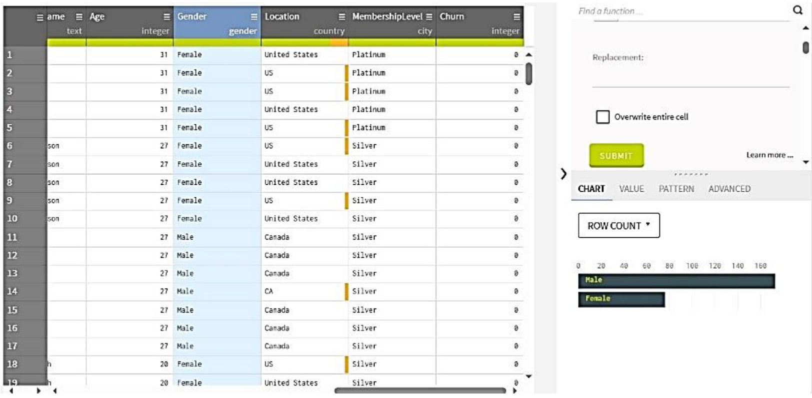


Figure 10: Gender after cleaning in Talend DP

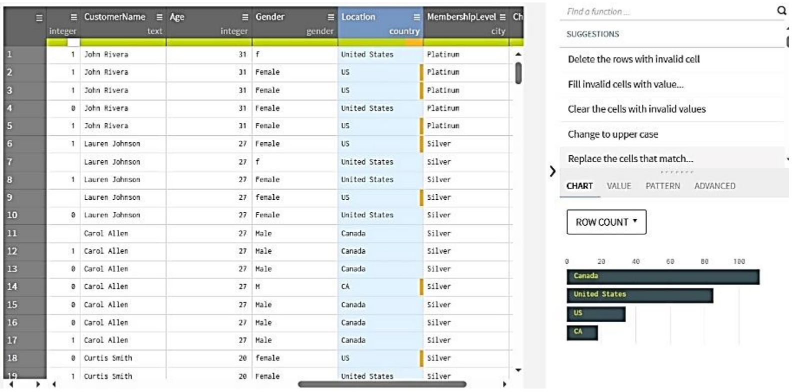


Figure 11: Location before cleaning in Talend DP

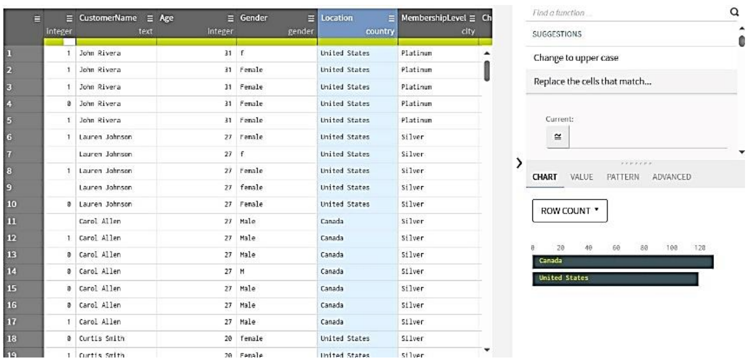


Figure 12: Location after cleaning in Talend DP

Mode Imputation in SAS EM

Figure 13 below shows how mode imputation is performed in SAS EM using the Impute node:

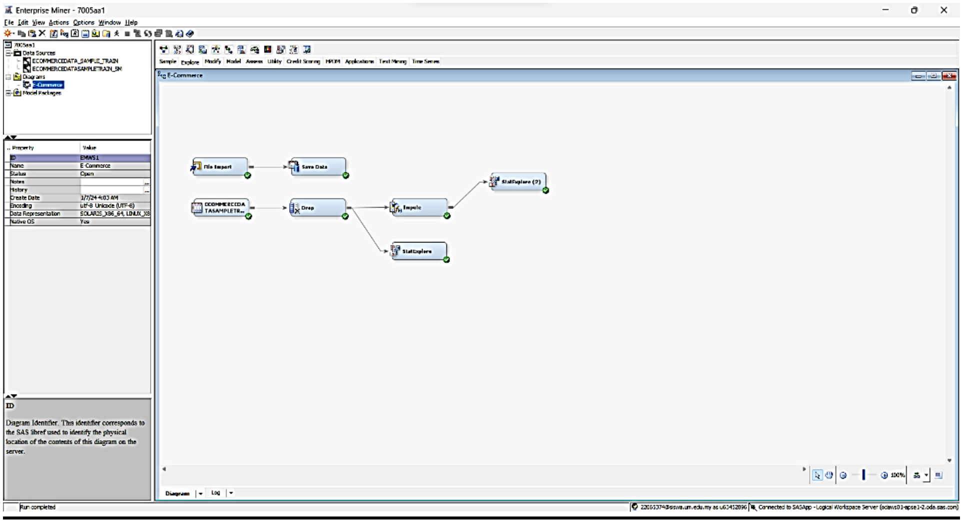


Figure 13: Mode Imputation in SAS EM

Figures 14 and 15 show the column metadata for number of missing values before mode imputation and after mode imputation in SAS EM:

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	Gender	INPUT	2	0	Male	50.44	Female	49.56
TRAIN	Location	INPUT	2	0	United States	50.17	Canada	49.83
TRAIN	MembershipLevel	INPUT	4	0	Platinum	25.76	Silver	24.78
TRAIN	PaymentMethod	INPUT	3	0	Credit Card	33.61	Cash	33.26
TRAIN	PurchasedProductCategory	INPUT	4	0	Electronics	25.24	Home	25.02
TRAIN	Returns	INPUT	3	4756	0	40.61	1	40.36
TRAIN	Churn	TARGET	2	0	0	79.95	1	20.05

Figure 14: Metadata Before Mode Imputation (Note: Returns)

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	Gender	INPUT	2	0	Male	50.44	Female	49.56
TRAIN	IMP_Returns	INPUT	2	0	0	59.64	1	40.36
TRAIN	Location	INPUT	2	0	United States	50.17	Canada	49.83
TRAIN	MembershipLevel	INPUT	4	0	Platinum	25.76	Silver	24.78
TRAIN	PaymentMethod	INPUT	3	0	Credit Card	33.61	Cash	33.26
TRAIN	PurchasedProductCategory	INPUT	4	0	Electronics	25.24	Home	25.02
TRAIN	Churn	TARGET	2	0	0	79.95	1	20.05

Figure 15: Metadata After Mode Imputation (Note: IMP_Returns)

Drop Unnecessary Columns in SAS EM

Figure 15 shows the dropping of unnecessary columns from the sample:

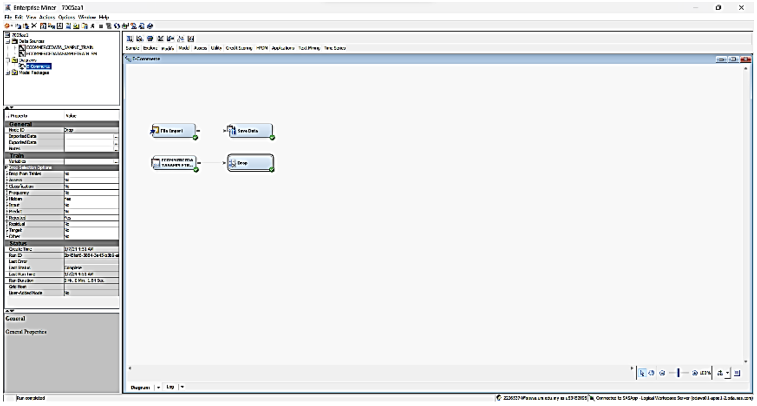


Figure 16: Drop unnecessary columns using Drop Node in SAS EM

Integrated Data Before and After Preprocessing using the Tools

Figure 17 below shows the integrated dataset before preprocessing using the tools:

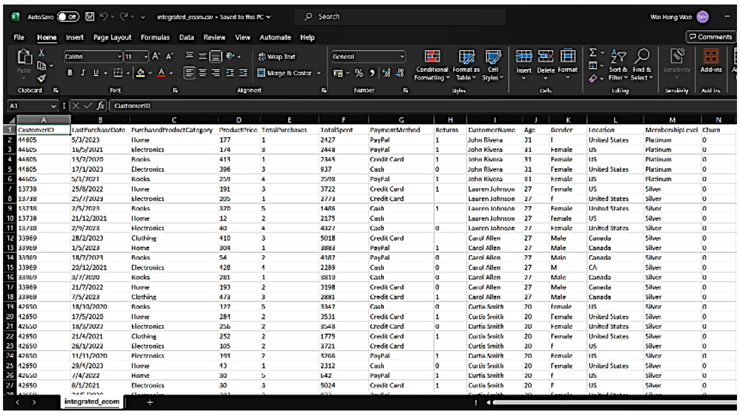


Figure 17: Integrated Dataset before Preprocessing

Figure 18 below shows the integrated dataset before preprocessing using the tools:

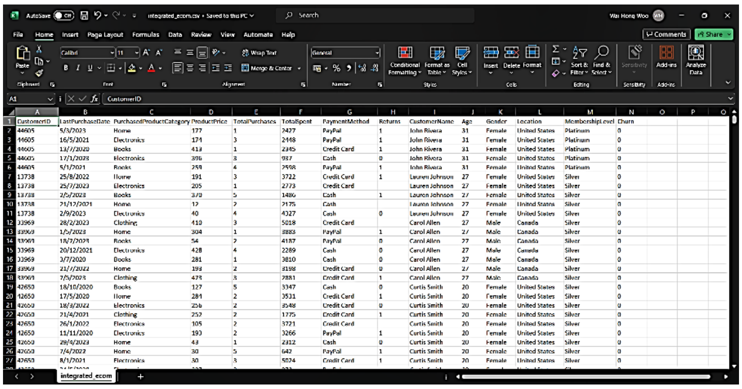


Figure 18: Integrated Dataset after Preprocessing

6 Decision Tree Analysis

Decision Tree Analysis corresponds to the Model step in the SEMMA methodology and in this assessment, it includes the following:

- Create a Decision Tree model and train it on the partitioned sample.
- Evaluate the results of the Decision Tree model

<u>Create a Decision Tree model in SAS EM</u>

Figure 19 below shows how Decision Tree model is created using the Decision Tree node in SAS EM:

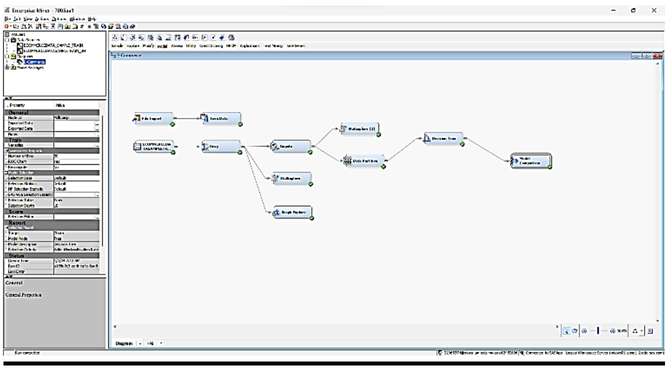


Figure 19: Decision Tree in SAS EM

<u>Results of Training of Decision Tree model in SAS EM</u>

Figure 20 below shows the results of training using Decision Tree model in SAS EM:

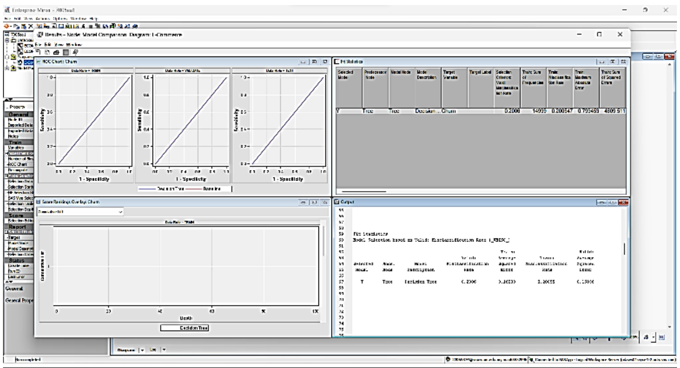


Figure 20: Results of Training Decision Tree in SAS EM

7 Ensemble Methods

Ensemble Methods analysis also corresponds to the Model step in the SEMMA methodology and in this assessment, it includes the following:

- Create 2 ensemble models, HP Forest (SAS EM implementation of Random Forest model which uses Bagging) and Gradient Boosting Classifier (which is a Boosting ensemble classifier) and train it on the partitioned sample.
- Evaluate the results of 2 ensemble models

Create 2 Ensemble models in SAS EM

Figure 21 shows how HP Forest model and Gradient Boosting classifier models are created using their specific nodes in SAS EM:

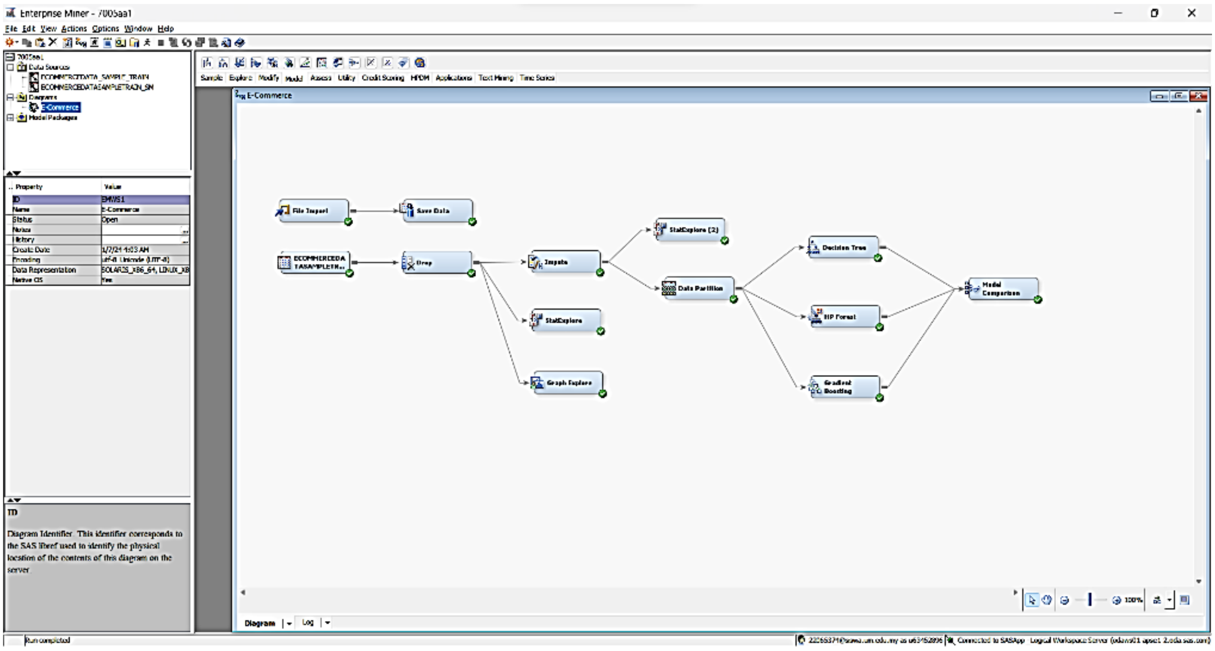


Figure 21: HP Forest and Gradient Boosting Classifier in SAS EM

Results of Training of HP Forest model in SAS EM

Figure 22 below shows the results of training of the HP Forest model in SAS EM:

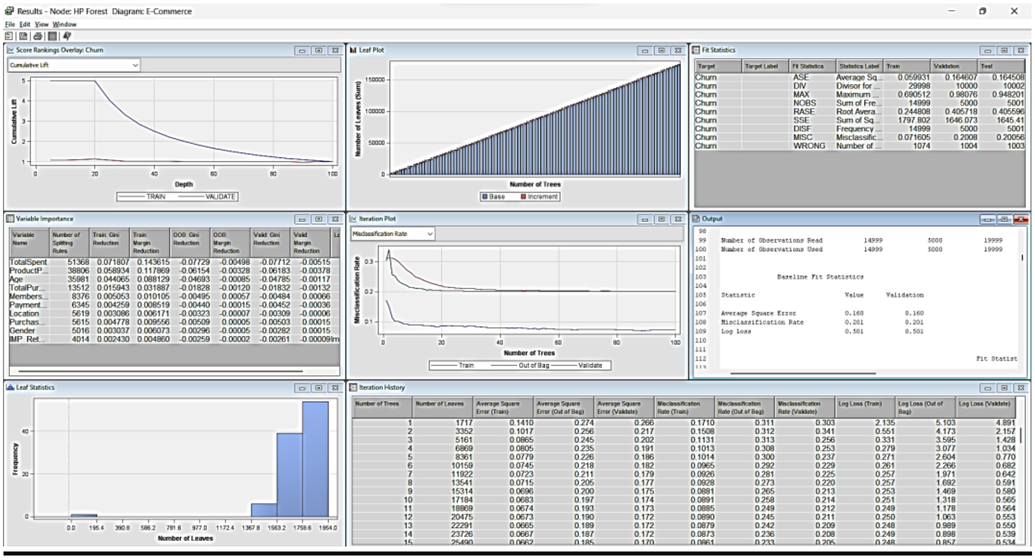


Figure 22: Results of Training of HP Forest model in SAS EM

Results of Gradient Boosting Classifier model in SAS EM

Figure 23 below shows the results of training of the Gradient Boosting classifier model in SAS EM:

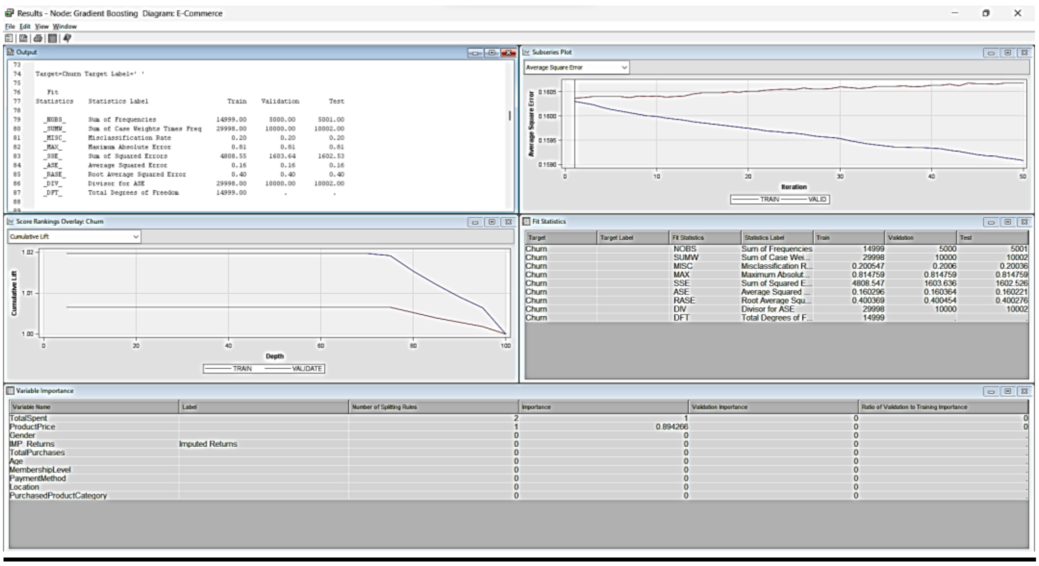


Figure 23: Results of Training of the Gradient Boosting classifier in SAS EM

8 Models' Performance Evaluation

Performance evaluation corresponds to the Assess step in the SEMMA methodology and in this assessment, it includes the following:

- Evaluate the performance of the 3 tree-based models using the misclassification rate metric on both validation and training sets in order to determine the best performing model for customer churn.

<u>Performance Results of the 3 tree-based models in SAS EM</u>

Figure 24 below shows the misclassification rates for each model in both sets:

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	Boost	Gradient Boosting	0.2006	0.16030	0.20055	0.16036
	Tree	Decision Tree	0.2006	0.16015	0.20055	0.16051
	HPDMForest	HP Forest	0.2008	0.05993	0.07160	0.16461

Figure 24: Performance Results for each model in SAS EM

<u>Confusion Matrix of the 3 tree-based models in SAS EM on Train and Validate</u>

Figure 25 below shows the confusion matrix for each model in both sets:

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
Tree	Decision Tree	TRAIN	Churn		3008	11991	0	0
Tree	Decision Tree	VALIDATE	Churn		1003	3997	0	0
HPDMForest	HP Forest	TRAIN	Churn		1074	11991	.	1934
HPDMForest	HP Forest	VALIDATE	Churn		1002	3995	2	1
Boost	Gradient Boosting	TRAIN	Churn		3008	11991	0	0
Boost	Gradient Boosting	VALIDATE	Churn		1003	3997	0	0

Figure 25: Confusion matrix for each model in SAS EM

9 Learning Outcomes, Suggestions for Business Strategy and Personal Reflection

Learning Outcomes:

Upon scrutinizing the performance metrics of the three models, the HP Forest model (or Random Forest) emerges as the optimal choice for classifying customer churn in the e-commerce domain. Notably, it achieves a significantly lower average squared error on the training set (0.05993), indicating superior predictive accuracy during training compared to both the Gradient Boosting Classifier (0.16030) and the Decision Tree (0.16015). Although the Gradient Boosting Classifier and Decision Tree share the same misclassification rate on the validation set (0.2006), the Random Forest only marginally lags with a rate of 0.2008. Furthermore, the Random Forest exhibits a notably lower misclassification rate on the training set (0.07160) compared to the Gradient Boosting Classifier and Decision Tree, both at 0.20055. Despite a slightly higher average squared error on the validation set (0.16461) than the Decision Tree (0.16051), the Random Forest's overall superior performance in training metrics positions it as the most promising model for effectively classifying customer churn in the dynamic e-commerce landscape.

Suggestions for Business Strategies:

Leveraging the predictive power of the Random Forest model, businesses in the e-commerce domain can implement targeted strategies to mitigate customer churn. The model's ability to discern patterns indicative of potential churn provides a valuable opportunity for proactive customer retention initiatives. Employing personalized marketing campaigns, tailored discounts, and exclusive promotions for identified at-risk customers can foster loyalty and incentivize continued engagement. Additionally, harnessing insights from the Random Forest, businesses can enhance customer satisfaction through improved user experiences, streamlined customer service processes, and personalized recommendations. The model's granularity enables the identification of specific pain points or areas for improvement, guiding strategic investments in product development or service enhancements. Furthermore, a real-time monitoring system, fuelled by the Random Forest's predictive capabilities, can facilitate prompt intervention when signs of churn arise, enabling businesses to implement timely retention strategies. Overall, the Random Forest model serves as a potent tool for crafting targeted and data-driven business strategies, empowering e-commerce enterprises to proactively address and mitigate customer churn.

Personal Reflection:

Embarking on this project was undoubtedly a journey marked by both challenges and growth. The intricacies of creating synthetic datasets, navigating the nuances of data integration in Talend, and meticulously cleaning data through Talend Data Preparation and SAS Enterprise Miner presented a formidable learning curve. However, the most profound challenge surfaced in the face of time constraints. Balancing the intricacies of each step with the ticking clock demanded resilience and strategic prioritization. Yet, within these challenges, I found a profound opportunity for personal and professional development. The 24-hour time constraint given for this assessment, though demanding, acted as catalysts for efficiency, forcing me to hone my problem-solving skills and embrace a mindset of continuous improvement. Ultimately, while the journey was arduous, the growth attained through overcoming these challenges leaves me with a profound sense of accomplishment and a newfound appreciation for the iterative nature of learning and problem-solving. I hope I can pass this module with flying colours.

Appendix

Customer data

CustomerID	CustomerName	Age	Gender	Location	MembershipLevel	Churn
44605	John Rivera	31	f	United States	Platinum	0
44605	John Rivera	31	Female	US	Platinum	0
44605	John Rivera	31	Female	US	Platinum	0
44605	John Rivera	31	Female	United States	Platinum	0
44605	John Rivera	31	Female	US	Platinum	0
13738	Lauren Johnson	27	Female	US	Silver	0
13738	Lauren Johnson	27	f	United States	Silver	0
13738	Lauren Johnson	27	Female	United States	Silver	0
13738	Lauren Johnson	27	female	US	Silver	0
13738	Lauren Johnson	27	Female	United States	Silver	0
33969	Carol Allen	27	Male	Canada	Silver	0
33969	Carol Allen	27	Male	Canada	Silver	0
33969	Carol Allen	27	Male	Canada	Silver	0
33969	Carol Allen	27	M	CA	Silver	0
33969	Carol Allen	27	Male	Canada	Silver	0
33969	Carol Allen	27	Male	Canada	Silver	0
33969	Carol Allen	27	Male	Canada	Silver	0
42650	Curtis Smith	20	female	US	Silver	0
42650	Curtis Smith	20	Female	United States	Silver	0
42650	Curtis Smith	20	Female	United States	Silver	0
42650	Curtis Smith	20	Female	United States	Silver	0
42650	Curtis Smith	20	f	US	Silver	0
42650	Curtis Smith	20	Female	US	Silver	0
42650	Curtis Smith	20	Female	United States	Silver	0
42650	Curtis Smith	20	f	US	Silver	0
42650	Curtis Smith	20	F	US	Silver	0

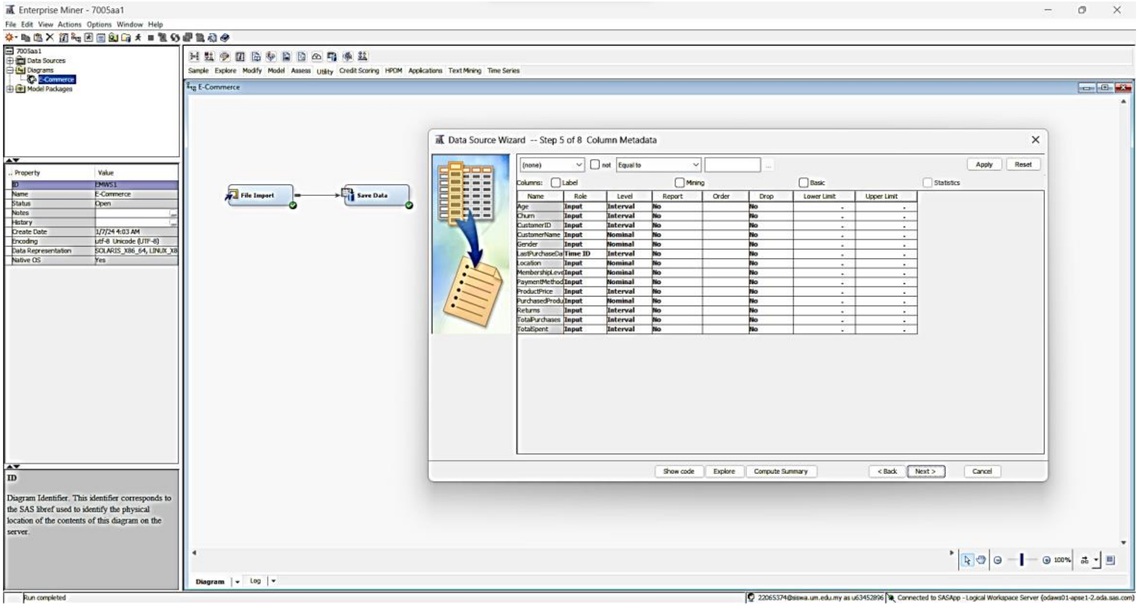
customer_data

Sales data

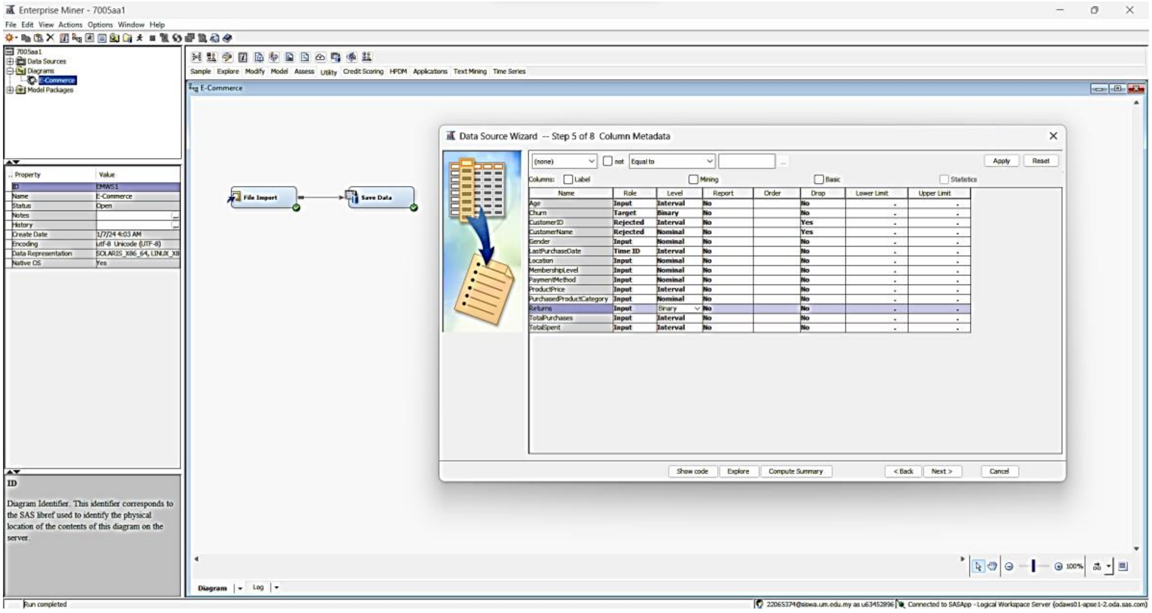
CustomerID	LastPurchaseDate	PurchasedProductCategory	ProductPrice	TotalPurchases	TotalSpent	PaymentMethod	Returns
44605	5/3/2023	Home	177	1	2427	PayPal	1
44605	16/5/2021	Electronics	174	3	2448	PayPal	1
44605	13/7/2020	Books	413	1	2345	Credit Card	1
44605	17/1/2023	Electronics	396	3	937	Cash	0
44605	5/1/2021	Books	259	4	2598	PayPal	1
13738	25/8/2022	Home	191	3	3722	Credit Card	1
13738	25/7/2023	Electronics	205	1	2773	Credit Card	
13738	2/5/2023	Books	370	5	1486	Cash	1
13738	21/12/2021	Home	12	2	2175	Cash	
13738	2/9/2023	Electronics	40	4	4327	Cash	0
33969	28/2/2023	Clothing	410	3	5018	Credit Card	
33969	1/5/2023	Home	304	1	3883	PayPal	1
33969	18/7/2023	Books	54	2	4187	PayPal	0
33969	20/12/2021	Electronics	428	4	2289	Cash	0
33969	3/7/2020	Books	281	1	3810	Cash	0
33969	21/7/2022	Home	193	2	3198	Credit Card	0
33969	7/5/2023	Clothing	473	3	2881	Credit Card	1
42650	18/10/2020	Books	127	5	3347	Cash	0
42650	17/5/2020	Home	284	2	3531	Credit Card	1
42650	18/3/2022	Electronics	256	2	3548	Credit Card	0
42650	21/4/2021	Clothing	252	2	1775	Credit Card	1
42650	26/1/2022	Electronics	105	2	3721	Credit Card	
42650	11/11/2020	Electronics	193	2	3266	PayPal	1
42650	29/4/2023	Home	43	1	2312	Cash	0
42650	7/4/2022	Home	30	5	642	PayPal	1
42650	8/1/2021	Electronics	30	3	5024	Credit Card	1

sales_data

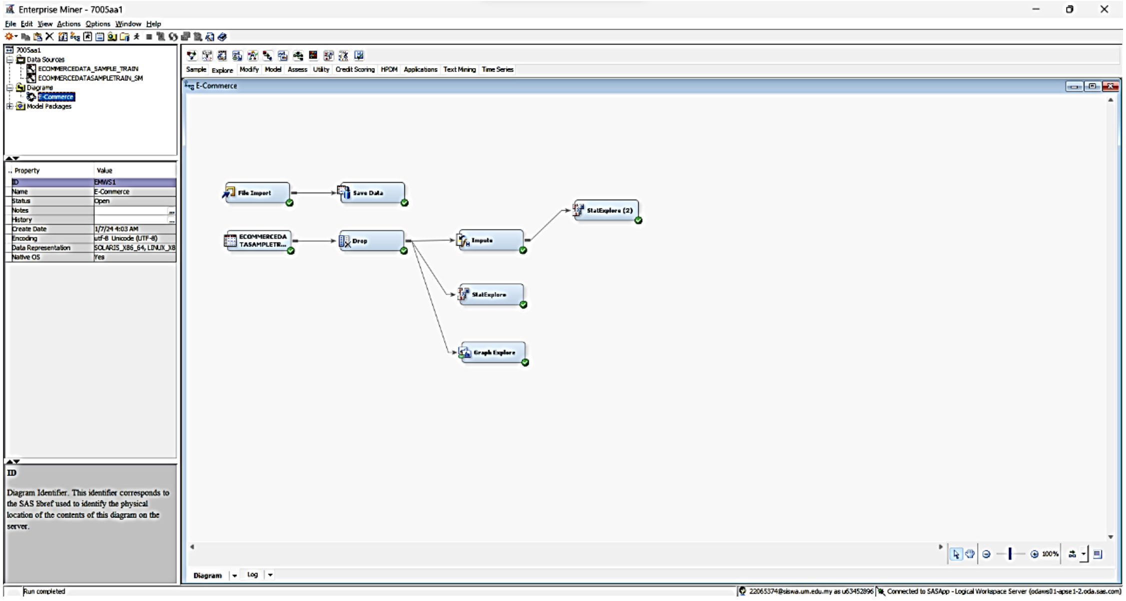
Before Reclassification of Variables in Basic Setting in SAS EM



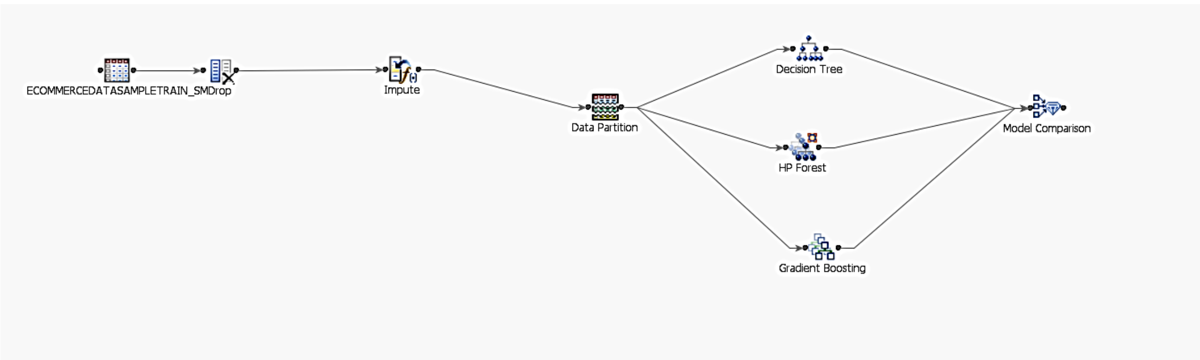
After Reclassification of Variables in Basic Setting in SAS EM



Graph Explore in SAS EM



Overall Workflow in SAS EM



Results of the 3 Tree-Based Models in SAS EM