Master of Data Science

Faculty of Computer Science & Information Technology

WQD7005 Data Mining

Alternative Assessment 1

Name: Yong Shen Woo

ID:s2175268

GROUP: 1

Lecturer: Dr. Teh Ying Wah

Case Study: E-commerce customer behaviour analysis

E-commerce customer behaviour case study helps in building the understanding of customers, such as their preferences, spending patterns, favourite products, and churn rate. The analysis can generate insights and helps e-commerce businesses to tailor their strategies to fulfil their customer needs, leading to increased satisfaction, loyalty while increasing return rate and reducing churn rate.

GitHub Link: https://github.com/yswooum/WQD7005-AA1-s2175268

Objectives:

1. To understand key demographic segments in e-commerce
2. To identify key features that lead to customer churn
3. To predict churn using machine learning model

Role of Talend Data Integration

Talend Data Integration is a powerful data integration tool that allows users to connect, transform and manage data from various sources. In this case study, Talend Data Integration was used to integrate the two datasets ‘customer info’ and ‘customer purchase info’ by the common column which is customer ID.

Role of Talend Data Preparation

Talend Data Preparation is a user-friendly data preparation tools that is specifically designed for data processing, such as data cleaning, transformation and normalization of the data. In this case study, Talend Data Preparation was used to handle the inconsistencies in the dataset. Specifically, the location, which has UK and United Kingdom, US and United States which referred to the same country. Talend Data Preparation was used to convert UK to United Kingdom and US to United States.

Role of SAS Enterprise Miner

SAS Enterprise Miner offers a wide range of advanced analytics and machine learning techniques that can be used to model and predict the trends and hidden relationships which can provide insights into ecommerce business. SEMMA

(Sample, explore, modify, model and access) methodology was applied in this case study. Random sampling was applied to the dataset to obtain a sample that is 10% of the original dataset. Then, data analysis and exploration were performed to inspect the distribution, presence of outliers or missing values in the dataset. In the modify step, missing column (return rate) which was a binary column was imputed with mode. The model that was used to predict churn in this case study was decision trees, random forest and gradient boosting. Lastly, the performance of the 3 models were compared and analysed.

Description of the dataset

The original dataset: E-commerce Customer Behaviour Dataset was obtained from Kaggle: https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset/data . 2 synthetic dataset was generated based on the original dataset, with the generation and addition of columns such as the membership level and location to ensure the dataset is similar to the required dataset structure.

Dataset 1 (customer_info.csv)

Attributes	Description
Customer ID	ID of the customer
Customer Age	Age of the customer
Returns	Return rate, binary (1:yes, 0:no)
Customer Name	Name of the customer
Gender	Gender of the customer, (Male, Female)
Churn	Churn rate, binary (1:yes, 0:no)
Membership Level	Membership (bronze, silver, gold, platinum)
Location	Location of the customer when purchase was made

Dataset 2 (customer_purchase_info.csv)

Attributes	Description
Customer ID	ID of the customer
Purchase Date	Date of the purchase made (‘yyyy-mm-dd’)
Product Category	Category of the products purchased (home, clothing, electronics, books)
Product Price	Price of the product purchased
Total Purchased	Total amount of product purchased
Total Spent	Total amount spent
Payment Method	Method of the payment (credit card, PayPal, cash, crypto)

Methodology

1. Talend Data Integration

The two datasets had a common key, which is the ID. So, the datasets were integrated by using Talend Data Integration.

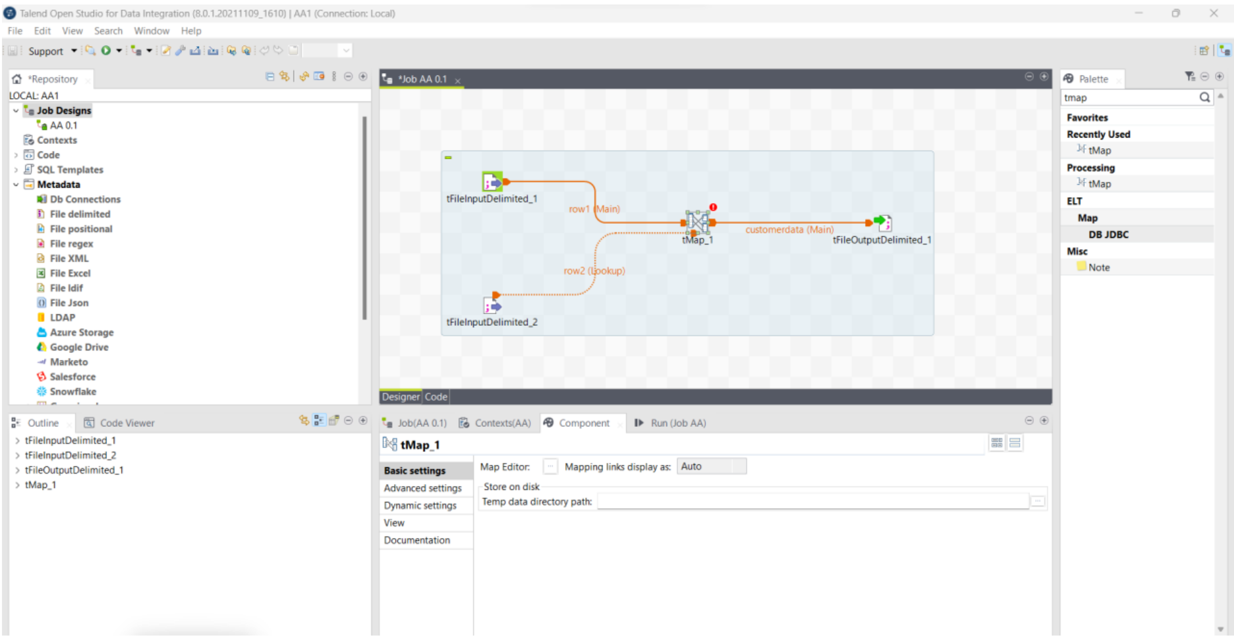


Fig 1. Integrating two datasets with Talend Data Integration. tFileInputDelimited: to import dataset into Talend Data Integration, the schema was set so that it was tally with the column

name. tMap: to integrate the datasets by the key, which is the customer ID. tFileOutputDelimited: to export the csv file (‘customer_data.csv’),

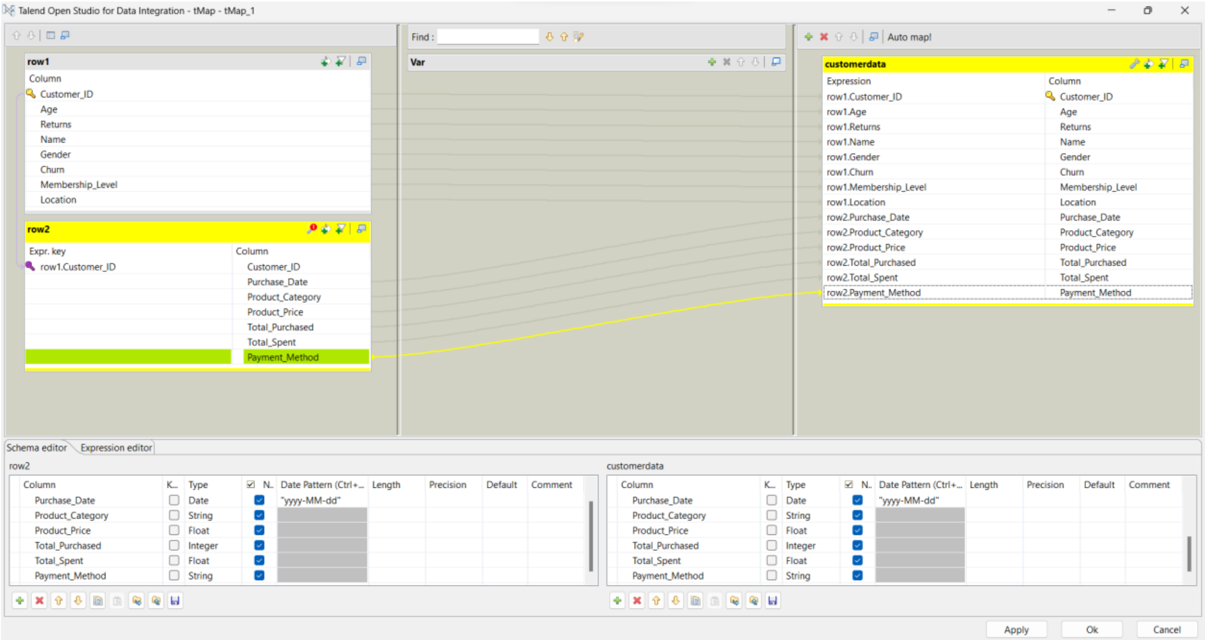
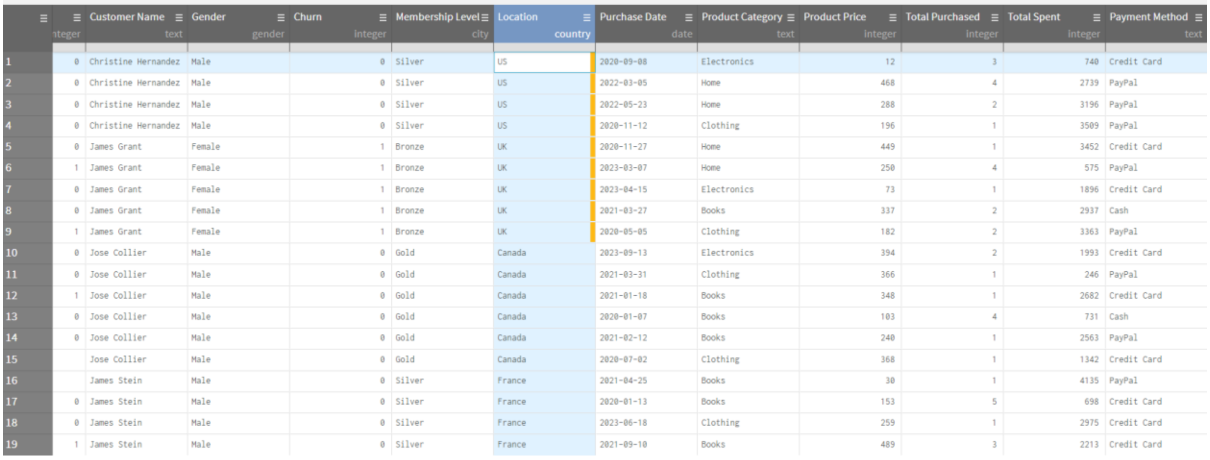


Fig. 2 Interface of tMap, the datasets were integrated by the key (customer ID).

stomer ID	Customer Age	Returns	Customer Name	Gender	Churn	Membership Level	Location	urchase Dat	Product Category	Product Price	Total Purchased	Total Spent	yment Method
46251	37	0	Christine Hernand	Male	0	Silver	US	2020-09-08	Electronics	12	3	740	Credit Card
46251	37	0	Christine Hernand	Male	0	Silver	US	2022-03-05	Home	468	4	2739	PayPal
46251	37	0	Christine Hernand	Male	0	Silver	US	2022-05-23	Home	288	2	3196	PayPal
46251	37	0	Christine Hernand	Male	0	Silver	US	2020-11-12	Clothing	196	1	3509	PayPal
13593	49	0	James Grant	Female	1	Bronze	UK	2020-11-27	Home	449	1	3452	Credit Card
13593	49	1	James Grant	Female	1	Bronze	UK	2023-03-07	Home	250	4	575	PayPal
13593	49	0	James Grant	Female	1	Bronze	UK	2023-04-15	Electronics	73	1	1896	Credit Card
13593	49	0	James Grant	Female	1	Bronze	UK	2021-03-27	Books	337	2	2937	Cash
13593	49	1	James Grant	Female	1	Bronze	UK	2020-05-05	Clothing	182	2	3363	PayPal
28805	19	0	Jose Collier	Male	0	Gold	Canada	2023-09-13	Electronics	394	2	1993	Credit Card
28805	19	0	Jose Collier	Male	0	Gold	Canada	2021-03-31	Clothing	366	1	246	PayPal
28805	19	1	Jose Collier	Male	0	Gold	Canada	2021-01-18	Books	348	1	2682	Credit Card
28805	19	0	Jose Collier	Male	0	Gold	Canada	2020-01-07	Books	103	4	731	Cash
28805	19	0	Jose Collier	Male	0	Gold	Canada	2021-02-12	Books	240	1	2563	PayPal
28805	19		Jose Collier	Male	0	Gold	Canada	2020-07-02	Clothing	368	1	1342	Credit Card
28961	55		James Stein	Male	0	Silver	France	2021-04-25	Books	30	1	4135	PayPal
28961	55	0	James Stein	Male	0	Silver	France	2020-01-13	Books	153	5	698	Credit Card
28961	55	0	James Stein	Male	0	Silver	France	2023-06-18	Clothing	259	1	2975	Credit Card
28961	55	1	James Stein	Male	0	Silver	France	2021-09-10	Books	489	3	2213	Credit Card
28961	55	1	James Stein	Male	0	Silver	France	2023-06-01	Books	232	3	4452	PayPal
12163	67		Sonia Moreno	Male	0	Gold	Sweden	2021-12-18	Clothing	255	2	1642	Credit Card
12163	67		Sonia Moreno	Male	0	Gold	Sweden	2020-06-20	Books	227	3	887	PayPal
12163	67	0	Sonia Moreno	Male	0	Gold	Sweden	2023-07-08	Clothing	288	2	1405	PayPal
12163	67		Sonia Moreno	Male	0	Gold	Sweden	2023-07-01	Books	60	3	4130	Cash
12163	67		Sonia Moreno	Male	0	Gold	Sweden	2022-11-16	Home	285	4	384	Crypto
12163	67	0	Sonia Moreno	Male	0	Gold	Sweden	2020-04-07	Books	100	5	1273	Cash

Fig. 3 The output of the integrated datasets (customer_data.csv)

2. Talend Data Preparation



	integer	Customer Name (text)	Gender (gender)	Churn (integer)	Membership Level (city)	Location (country)	Purchase Date (date)	Product Category (text)	Product Price (integer)	Total Purchased (integer)	Total Spent (integer)	Payment Method (text)
1	0	Christine Hernandez	Male	0	Silver	US	2020-09-08	Electronics	12	3	740	Credit Card
2	0	Christine Hernandez	Male	0	Silver	US	2022-03-05	Home	468	4	2739	PayPal
3	0	Christine Hernandez	Male	0	Silver	US	2022-05-23	Home	288	2	3196	PayPal
4	0	Christine Hernandez	Male	0	Silver	US	2020-11-12	Clothing	196	1	3509	PayPal
5	0	James Grant	Female	1	Bronze	UK	2020-11-27	Home	449	1	3452	Credit Card
6	1	James Grant	Female	1	Bronze	UK	2023-03-07	Home	250	4	575	PayPal
7	0	James Grant	Female	1	Bronze	UK	2023-04-15	Electronics	73	1	1896	Credit Card
8	0	James Grant	Female	1	Bronze	UK	2021-03-27	Books	337	2	2937	Cash
9	1	James Grant	Female	1	Bronze	UK	2020-05-05	Clothing	182	2	3363	PayPal
10	0	Jose Collier	Male	0	Gold	Canada	2023-09-13	Electronics	394	2	1993	Credit Card
11	0	Jose Collier	Male	0	Gold	Canada	2021-03-31	Clothing	366	1	246	PayPal
12	1	Jose Collier	Male	0	Gold	Canada	2021-01-18	Books	348	1	2682	Credit Card
13	0	Jose Collier	Male	0	Gold	Canada	2020-01-07	Books	103	4	731	Cash
14	0	Jose Collier	Male	0	Gold	Canada	2021-02-12	Books	240	1	2563	PayPal
15		Jose Collier	Male	0	Gold	Canada	2020-07-02	Clothing	368	1	1342	Credit Card
16		James Stein	Male	0	Silver	France	2021-04-25	Books	30	1	4135	PayPal
17	0	James Stein	Male	0	Silver	France	2020-01-13	Books	153	5	698	Credit Card
18	0	James Stein	Male	0	Silver	France	2023-06-18	Clothing	259	1	2975	Credit Card
19	1	James Stein	Male	0	Silver	France	2021-09-10	Books	489	3	2213	Credit Card

Fig. 4 Interface of Talend DataPrep. Upon inspection, there was inconsistency in the location column.

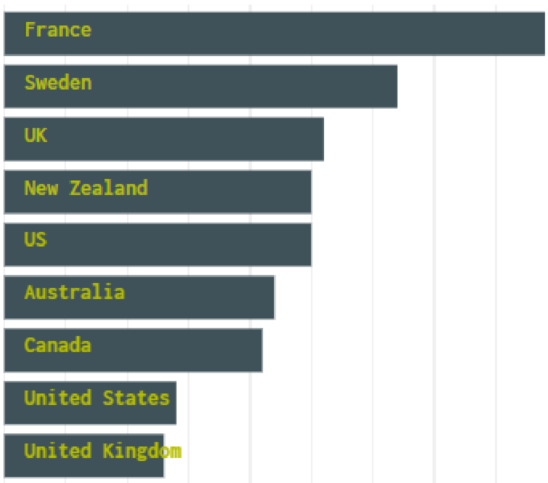


Fig. 5 The country name in the Location column. There was UK and United Kingdom as well as US and United States which were representing the same country, but in short form.

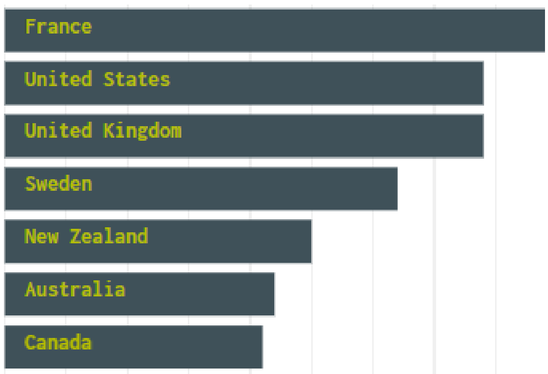


Fig. 6 Replacing UK with United Kingdom and US with United States.

3. SAS ENTREPRISE MINER

The summary of SEMMA methodology applied in the case study:

Sample: 10% of the dataset was randomly sampled.

Explore: Explored and analysed the dataset, identify outliers and missing values. No outliers were identified in the dataset, but there is a column (Return) that had missing values.

Modify: The column with the missing values (Return) was imputed with mode. No transformation and normalization were done to the dataset. Then, the dataset was split into 60 % training, 20 %validation and 20% testing sets for modelling.

Model: Three models were used to predict the churn, namely decision tree, random forest and gradient boosting.

Access: The performance of the model was compared.

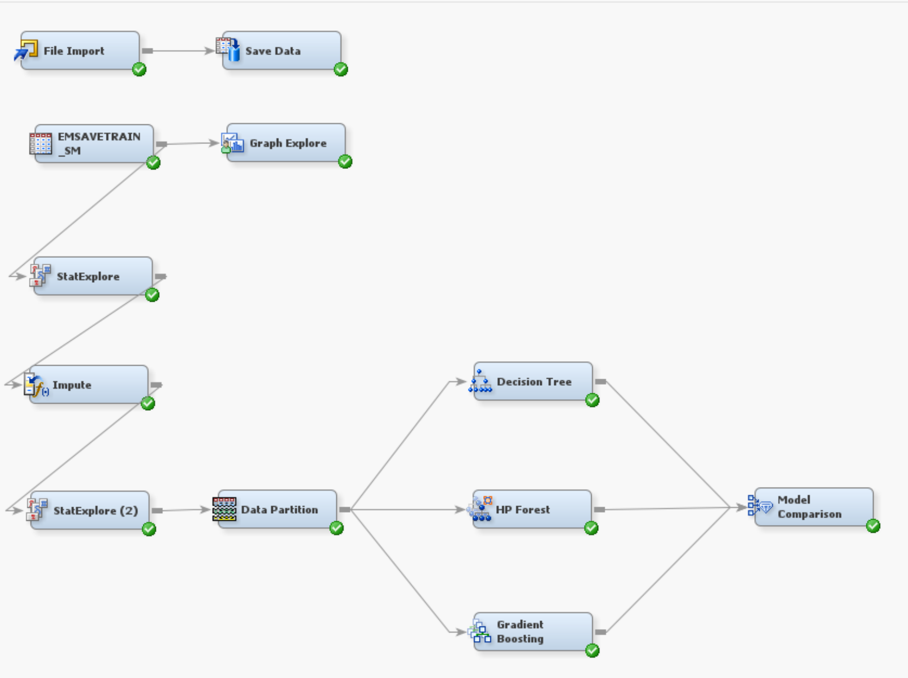


Fig 7. Summary of the workflow using SAS Enterprise Miner.

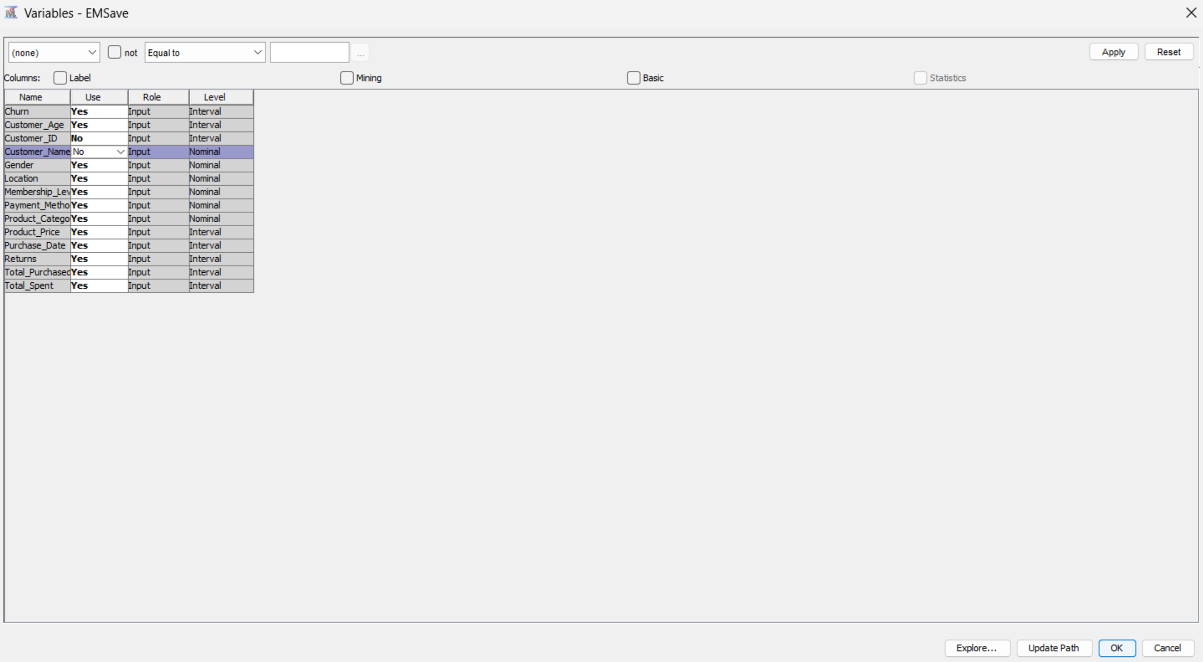


Fig. 8 Dropping ID and Name as they are irrelevant to the analysis and prediction.

SEMMA: Sample

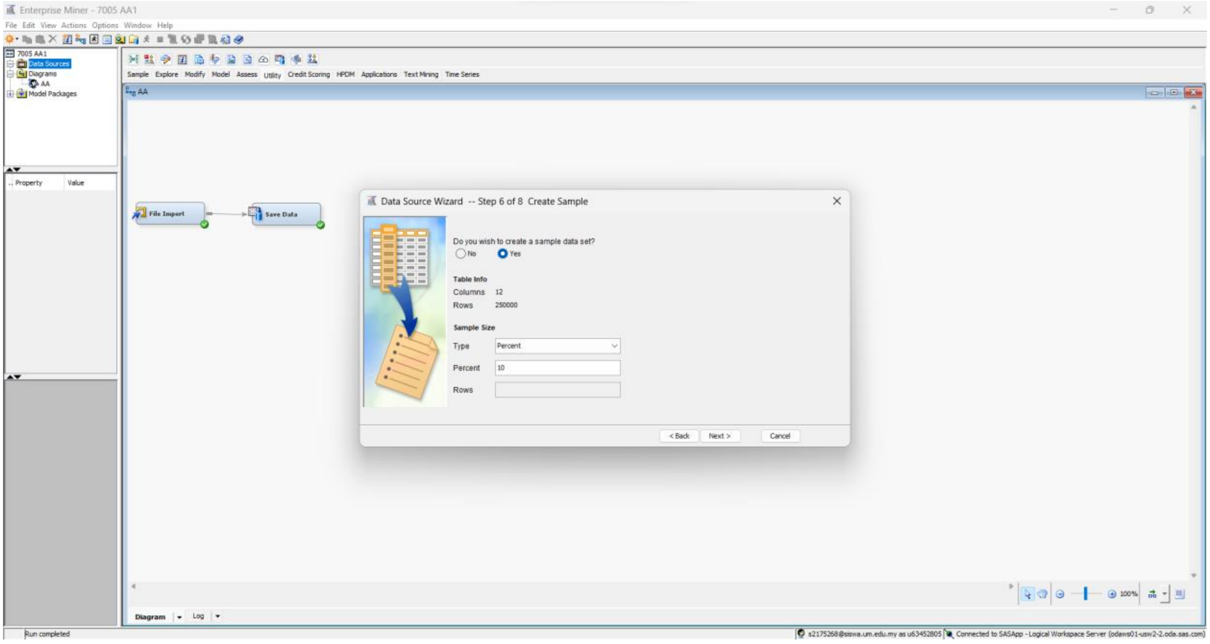


Fig. 9 Sampling step. 10% of the data was randomly sampled from the dataset.

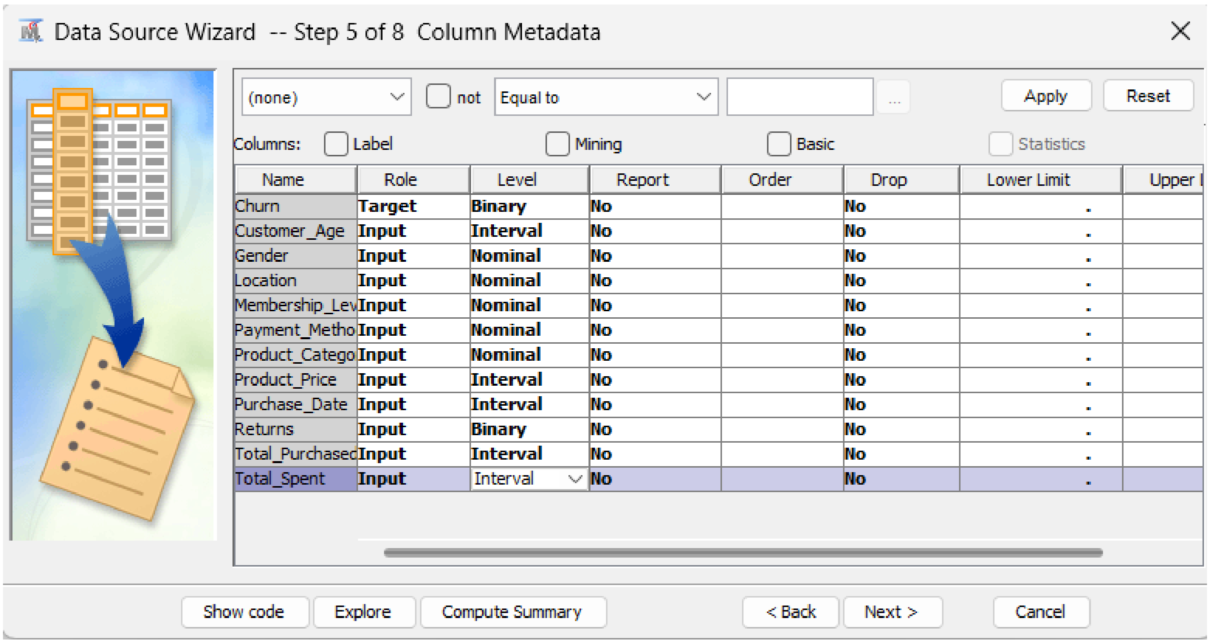
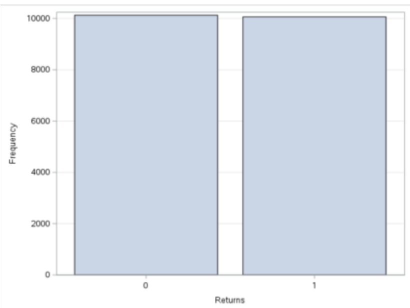
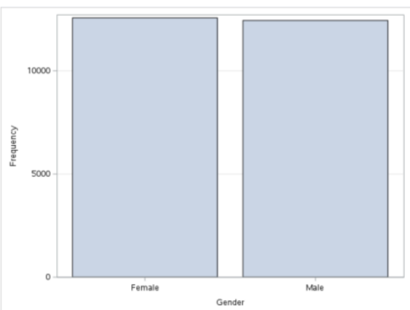
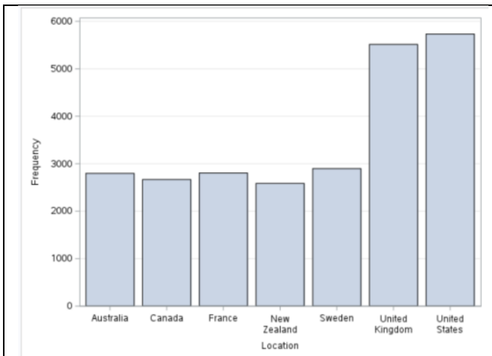
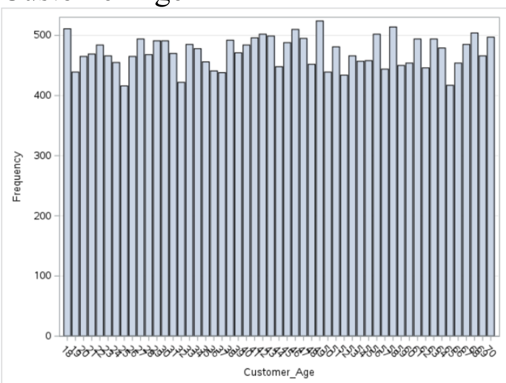
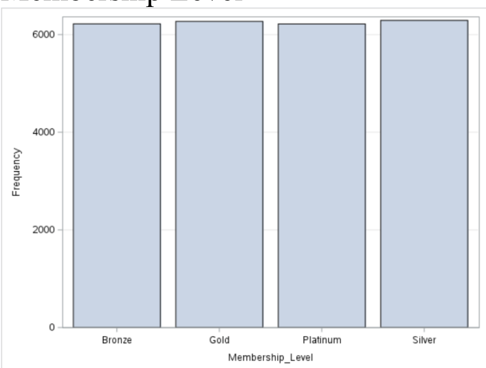
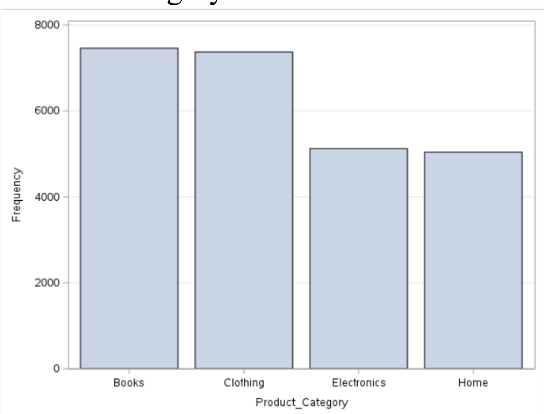


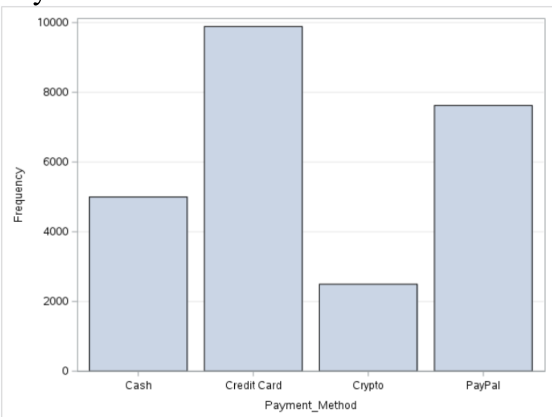
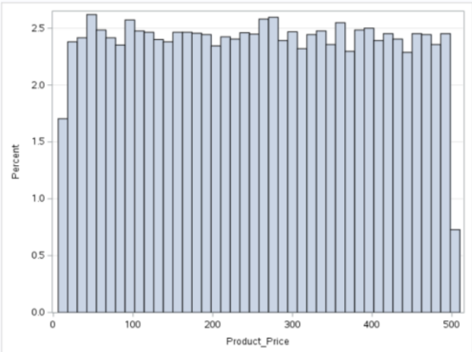
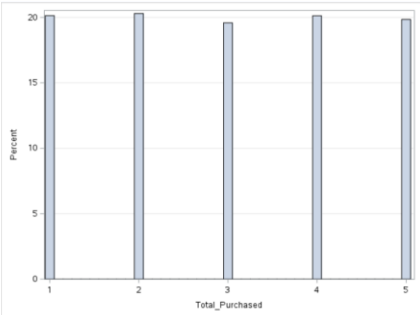
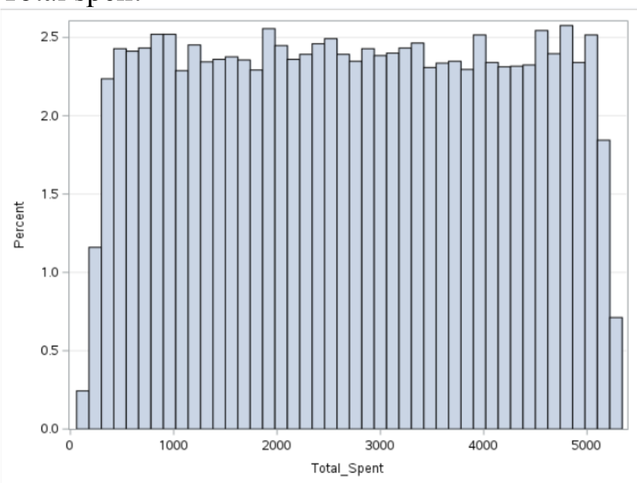
Fig. 10 Assigning the role and data type of the columns. The target is churn, which is a binary column.

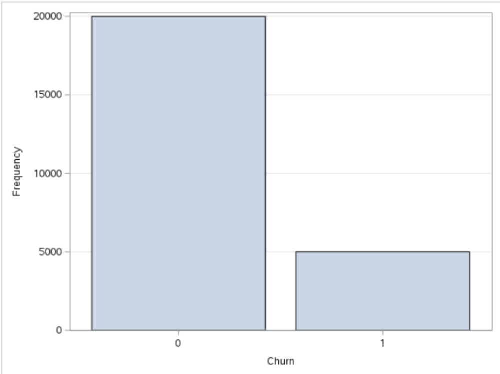
SEMMA: Explore

Explore

Attributes	Findings
Returns 	The number for customers that did not return and returned are close.
Gender 	There is no significant difference in the distribution between male and female in e-commerce customer.
Location	

	United States and United Kingdom has more customers than other countries.
Customer Age 	The distribution of age is not normally distributed. It is close to uniform distribution.
Membership Level 	The membership level is uniformly distributed.
Product Category 	Books and clothing are purchased more than electronics and home appliances.

Payment Method  	Credit card is the most frequently chosen payment method.
Product Price  	The product price is close to uniform distribution.
Total Purchased  	The number of item purchased is uniformly distributed.
Total spent  	The total spent is very similar to uniform distribution.

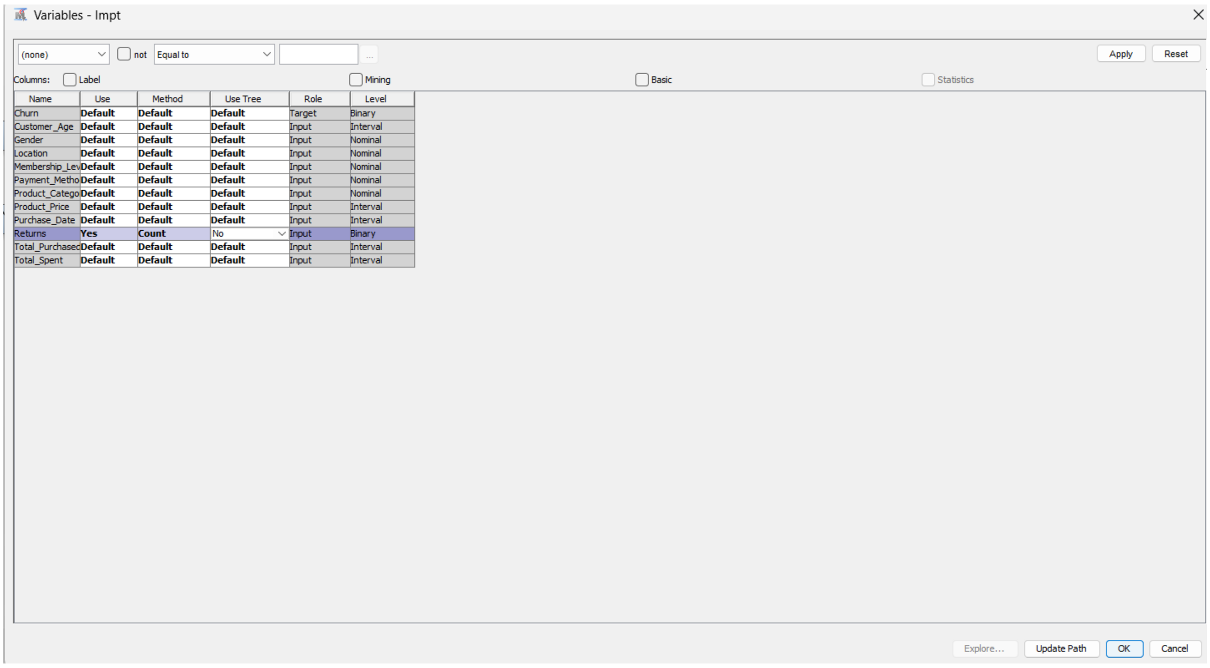
<table>
<tr>
<td>Churn
</td>
<td>Churn = 0 is higher than churn = 1</td>
</tr>
</table>

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Gender	INPUT	2	0	Female	50.25	Male	49.75
TRAIN	Location	INPUT	7	0	United States	22.93	United Kingdom	22.06
TRAIN	Membership_Level	INPUT	4	0	Silver	25.16	Gold	25.08
TRAIN	Payment_Method	INPUT	4	0	Credit Card	39.55	PayPal	30.48
TRAIN	Product_Category	INPUT	4	0	Books	29.84	Clothing	29.49
TRAIN	Returns	INPUT	3	4811	0	40.51	1	40.25
TRAIN	Churn	TARGET	2	0	0	79.98	1	20.02

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness
Customer_Age	INPUT	44.052	15.28281	25000	0	18	44	70	-0.00286
Product_Price	INPUT	254.328	141.8547	25000	0	10	255	500	0.006561
Total_Purchased	INPUT	2.99232	1.415522	25000	0	1	3	5	0.007308
Total_Spent	INPUT	2730.574	1445.849	25000	0	101	2721	5338	0.004814

Fig. 11 Checking for missing values, there was 4811 missing values for column Returns, so imputation was done.

<u>SEMMA: Modify</u>



Name	Use	Method	Use Tree	Role	Level
Churn	Default	Default	Default	Target	Binary
Customer_Age	Default	Default	Default	Input	Interval
Gender	Default	Default	Default	Input	Nominal
Location	Default	Default	Default	Input	Nominal
Membership_Lev	Default	Default	Default	Input	Nominal
Payment_Metho	Default	Default	Default	Input	Nominal
Product_Catego	Default	Default	Default	Input	Nominal
Product_Price	Default	Default	Default	Input	Interval
Purchase_Date	Default	Default	Default	Input	Interval
Returns	Yes	Count	No	Input	Binary
Total_Purchased	Default	Default	Default	Input	Interval
Total_Spent	Default	Default	Default	Input	Interval

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Returns	COUNT	IMP_Returns	0	INPUT	BINARY	Returns	4811

Fig. 12 Imputation and summary of the missing values (Returns) with mode.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Gender	INPUT	2	0	Female	50.25	Male	49.75
TRAIN	IMP_Returns	INPUT	2	0	0	59.75	1	40.25
TRAIN	Location	INPUT	7	0	United States	22.93	United Kingdom	22.06
TRAIN	Membership_Level	INPUT	4	0	Silver	25.16	Gold	25.08
TRAIN	Payment_Method	INPUT	4	0	Credit Card	39.55	PayPal	30.48
TRAIN	Product_Category	INPUT	4	0	Books	29.84	Clothing	29.49
TRAIN	Churn	TARGET	2	0	0	79.98	1	20.02

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness
Customer_Age	INPUT	44.052	15.28281	25000	0	18	44	70	-0.00286
Product_Price	INPUT	254.328	141.8547	25000	0	10	255	500	0.006561
Purchase_Date	INPUT	44505.79	391.8271	25000	0	43831	44500	45184	0.008158
Total_Purchased	INPUT	2.99232	1.415522	25000	0	1	3	5	0.007308
Total_Spent	INPUT	2730.574	1445.849	25000	0	101	2721	5338	0.004814

Fig. 13 Inspection of the missing values after imputation. There were no missing values after the imputation.

<u>SEMMA: Model</u>

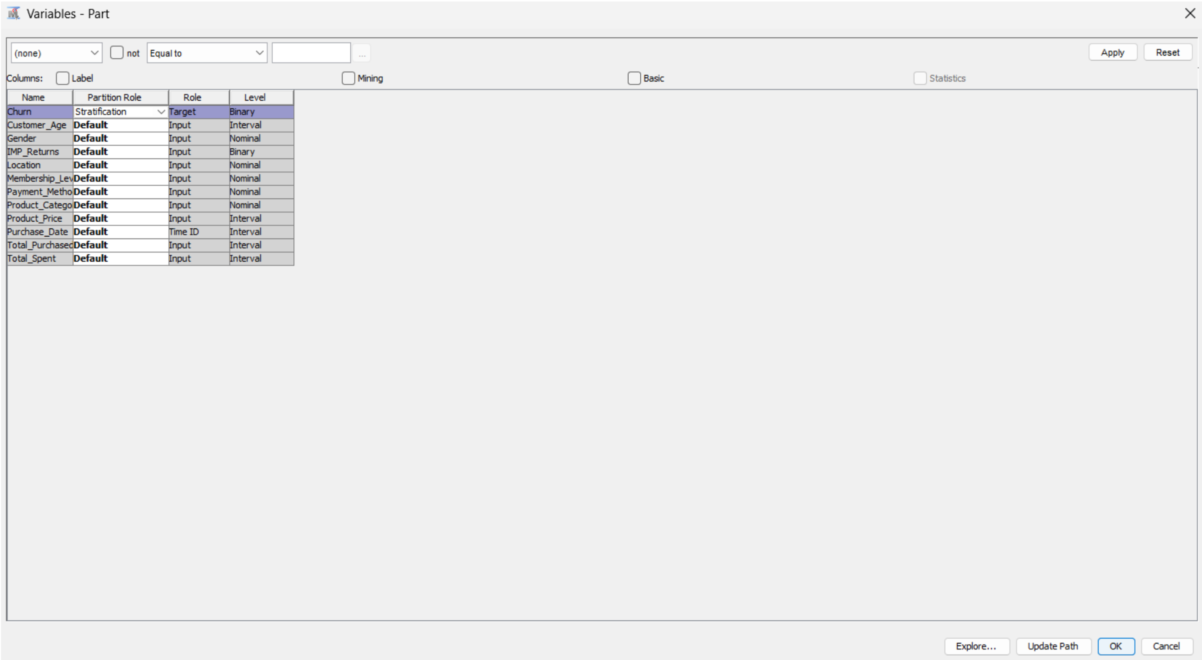


Fig.14 Assigning the partition roles before splitting the data.

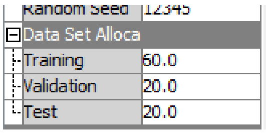


Fig.15 Split the data into 60% training, 20% validation and 20% testing.

```
Partition Summary

                                  Number of
Type            Data Set         Observations

DATA        EMWS1.Stat2_TRAIN         25000
TRAIN       EMWS1.Part_TRAIN          14999
VALIDATE    EMWS1.Part_VALIDATE        4999
TEST        EMWS1.Part_TEST            5002


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*



Summary Statistics for Class Targets

Data=DATA

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent     Label

 Churn         0           0          19994        79.976     Churn
 Churn         1           1           5006        20.024     Churn


Data=TEST

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent     Label

 Churn         0           0           4000       79.9680     Churn
 Churn         1           1           1002       20.0320     Churn


Data=TRAIN

            Numeric    Formatted    Frequency
Variable     Value       Value        Count       Percent     Label

 Churn         0           0          11996       79.9787     Churn
 Churn         1           1           3003       20.0213     Churn
```

Fig. 16 Data partition report

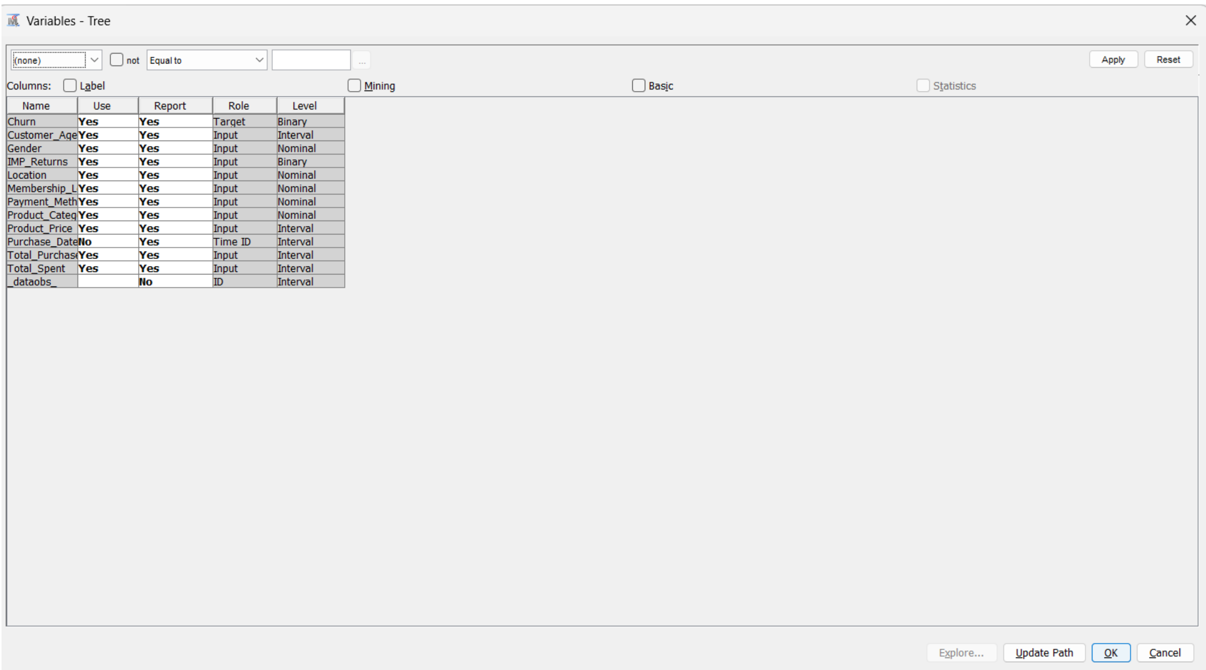


Fig. 17 Assigning roles for decision trees.

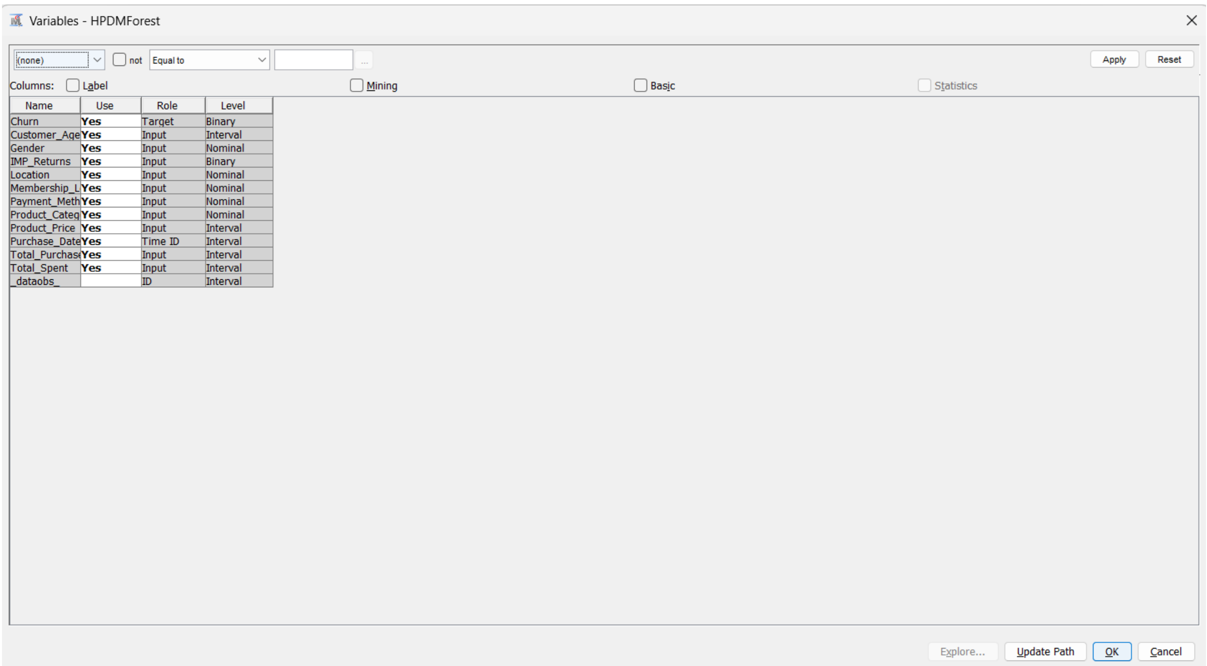
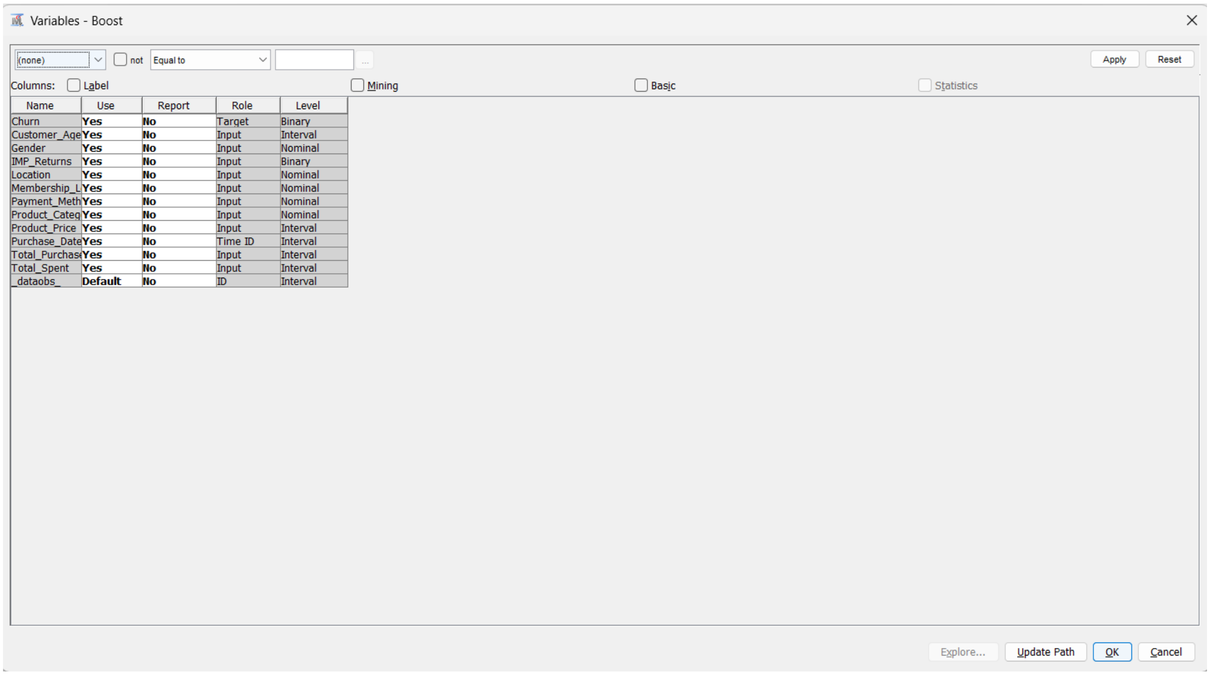


Fig. 18 Assigning roles for random forest.



Name	Use	Report	Role	Level
Churn	Yes	No	Target	Binary
Customer_Age	Yes	No	Input	Interval
Gender	Yes	No	Input	Nominal
IMP_Returns	Yes	No	Input	Binary
Location	Yes	No	Input	Nominal
Membership_L	Yes	No	Input	Nominal
Payment_Meth	Yes	No	Input	Nominal
Product_Categ	Yes	No	Input	Nominal
Product_Price	Yes	No	Input	Interval
Purchase_Date	Yes	No	Time ID	Interval
Total_Purchas	Yes	No	Input	Interval
Total_Spent	Yes	No	Input	Interval
dataobs	Default	No	ID	Interval

Fig. 19 Assigning roles for gradient boosting.

<u>SEMMA: Assess</u>

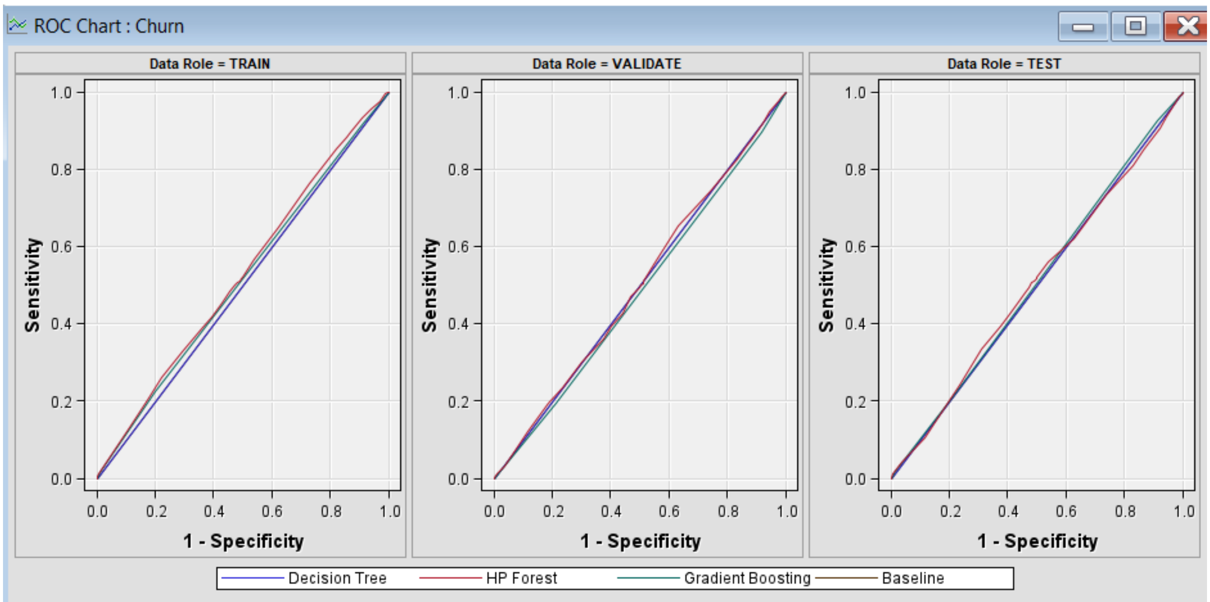


Fig.20 Receiver operating curve (ROC) for decision trees, random forest and gradient boosting for training, validation and testing sets. The ROC curve is very similar for 3 models across training, validation and testing sets which suggests that the models are not overfitting.

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	Tree	Decision Tree	0.20024	0.16013	0.20021	0.16014
	Boost	Gradient Boosting	0.20024	0.16010	0.20021	0.16018
	HPDMForest	HP Forest	0.20024	0.16000	0.20021	0.16022

Fig. 21 Misclassification rate and average squared error for decision tree, random forest and gradient boosting. The misclassification rate and average squared error across 3 models are very similar, with decision trees has slightly lower average squared error in validation set.

Important features in model prediction

Feature importance can be generated from random forest to study the most important features in the prediction.

Variable Name	Number of Splitting Rules
Gender	20
Membership Level	5
Total Purchased	4
IMP Returns	2
Location	2
Customer Age	1
Total Spent	1
Payment Method	0
Product Category	0
Product Price	0

Fig. 22 The feature importance generated from random forest in SAS E-Miner. Gender was the most important features, followed by membership level and total amount of product purchased.

Reflection

In this case study, decision tree slightly outperformed random forest and gradient boosting. Usually, the bagging (random forest) and boosting (gradient boosting) method are expected to have better performance than decision trees. This is because random forest and gradient boosting are ensemble models based on decision trees, which allow for more accurate and robust prediction. However, in a simple and straightforward dataset, decision trees can achieve better performance than random forest and gradient boosting, because random forest and gradient boosting require careful and precise hyperparameter tuning, which was not done in this case study. This is the limitation of this study, due to time constraints, the proper tuning of the models was not performed. From the feature importance in random forest, gender, membership level and total amount of product purchased were the top 3 most important features in predicting customer churn. Since gender was the most important features in predicting churn, gender-specific retention actions should be taken by the business. Tailored marketing and retention strategies for different genders can be effective in reducing customer churn.