UNIVERSITI MALAYA

WQD7005 DATA MINING

SEM 1/2023/2024

ALTERNATIVE ASSESSMENT 1

GROUP 1

Github repository:

https://github.com/lowkianhaw/datamining_AA1

Matric Number:	S2190839
Name:	Low Kian Haw

Introduction

Hypermarkets play a pivotal role in providing an extensive range of products and services to a diverse customer base. However, one of the significant challenges faced by hypermarkets is the issue of customer churn, where customers discontinue their association with the hypermarket and shift their loyalty elsewhere. The ability to predict and understand customer churn is crucial for hypermarkets to implement effective retention strategies, enhance customer satisfaction, and maintain a competitive edge in the market.

This study focuses on delving into the intricate dynamics of hypermarket customer churn, utilizing decision trees to develop predictive insights. By leveraging historical transactional data, customer interactions, and demographic information, the study aims to uncover patterns and indicators that contribute to customer churn in hypermarkets.

Objectives

The objectives of the study on predicting hypermarket customer churn are as follows:

1. To explore and analyze various factors that contribute to customer churn within hypermarkets.
2. To utilize decision tree algorithms to develop accurate predictive models to predict customer churn.

Dataset description

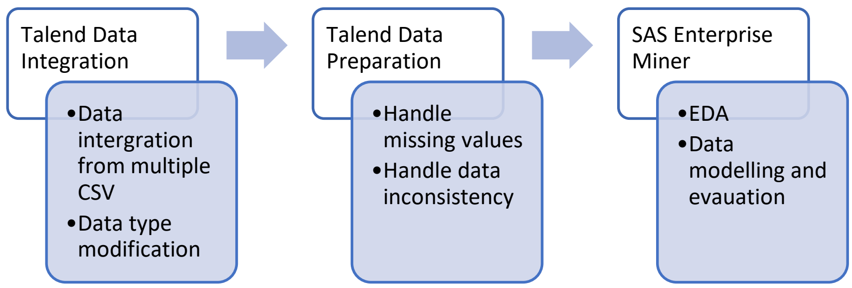


Figure 1: Overview of the roles of several tools

In the Alternative Assessment 1, there are 2 datasets being used- Customer Description.csv and Customer Purchase.csv. The description of the 2 datasets is being shown as below:

Customer Description.csv:

Total number of rows:1000

Total number of variables: 7

Variables	Description
CustomerID	A unique identifier assigned to each customer, used to distinguish one customer from another in the dataset.
Age	The age of the customer
Gender	The gender of the customer

MembershipLevel	The level of membership or loyalty status assigned to the customer
Location	The geographical location of the customer
Occupation	The type of job or profession that the customer is engaged in
MaritalStatus	The marital status of the customer

Customer Purchase.csv:

Total number of rows:1000

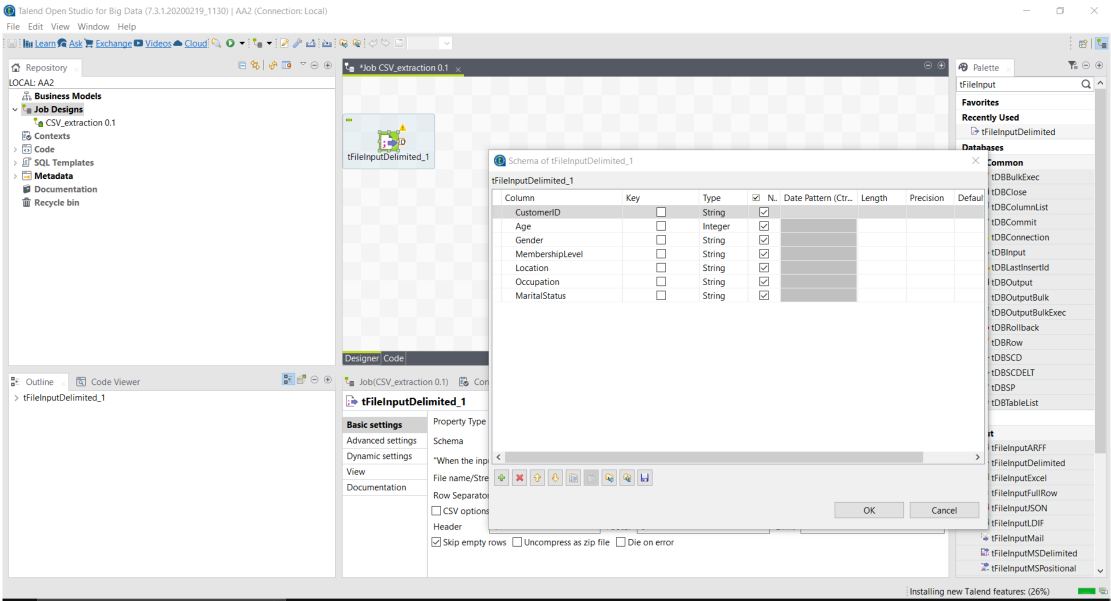
Total number of variables: 6

Variables	Description
CustomerID	A unique identifier assigned to each customer, used to distinguish one customer from another in the dataset.
TotalPurchases	The cumulative count of purchases made by the customer
TotalSpent	The total monetary amount spent by the customer across all purchases
FavoriteCategory	The product or service category that the customer frequently purchases or shows a preference for.
LastPurchaseDate	The date of the customer's most recent purchase
Churn	A binary indicator (e.g., 1 or 0) representing whether the customer has churned (last 6 months)

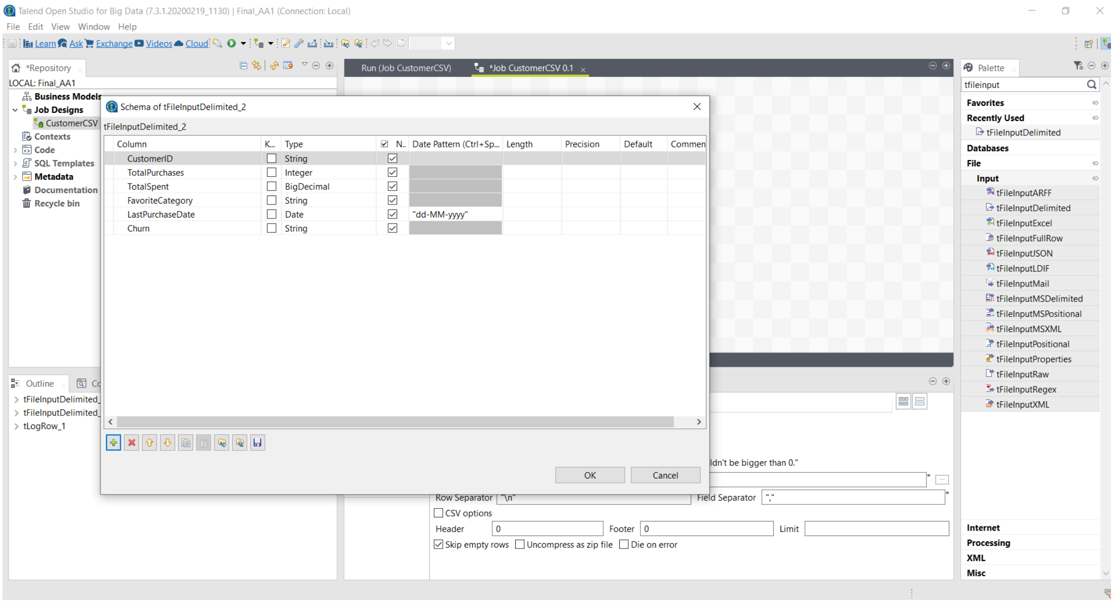
Data Integration:

The 2 datasets were firstly being imported into Talend Data Integration for data integration steps. Talend Data Integration is a popular open-source data integration and ETL (Extract, Transform, Load) tool that is widely used for various data processing tasks. It comes with a rich set of pre-built components for common data integration tasks. These components cover a wide range of functionalities, such as data cleansing, transformation, enrichment, and loading, making it easier to build complex data workflows.

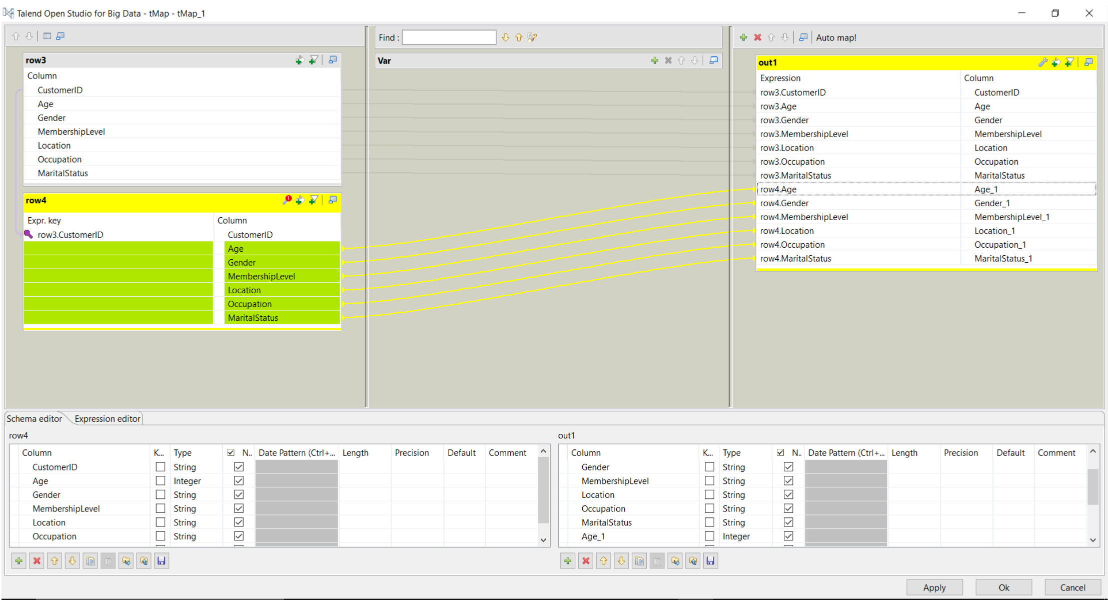
1. A job named CSV_extraction was created using Talend Data Integration. The Customer Description.csv was firstly imported into Talend Data Integration by using tFileInputDelimited_1 component. All the default data types are correct, except ‘Age’ which is labelled as String. It was corrected to Integer.



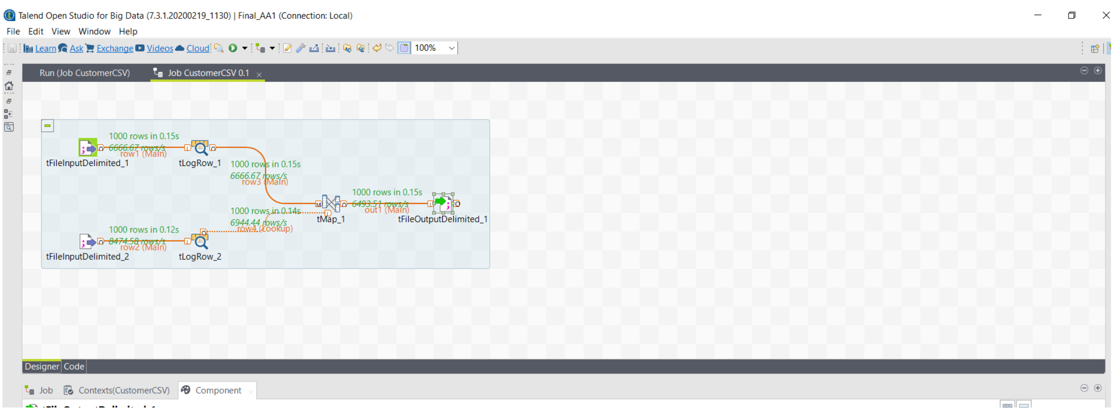
2. Similarly, Customer Purchase.csv was imported into Talend Data Integration by using tFileInputDelimited_2 component. The variables ‘Total Purchases’ was changed to Integer data type, ‘TotalSpent’ to BigDecimal data type and ‘LastPurchaseDate’ to Date.



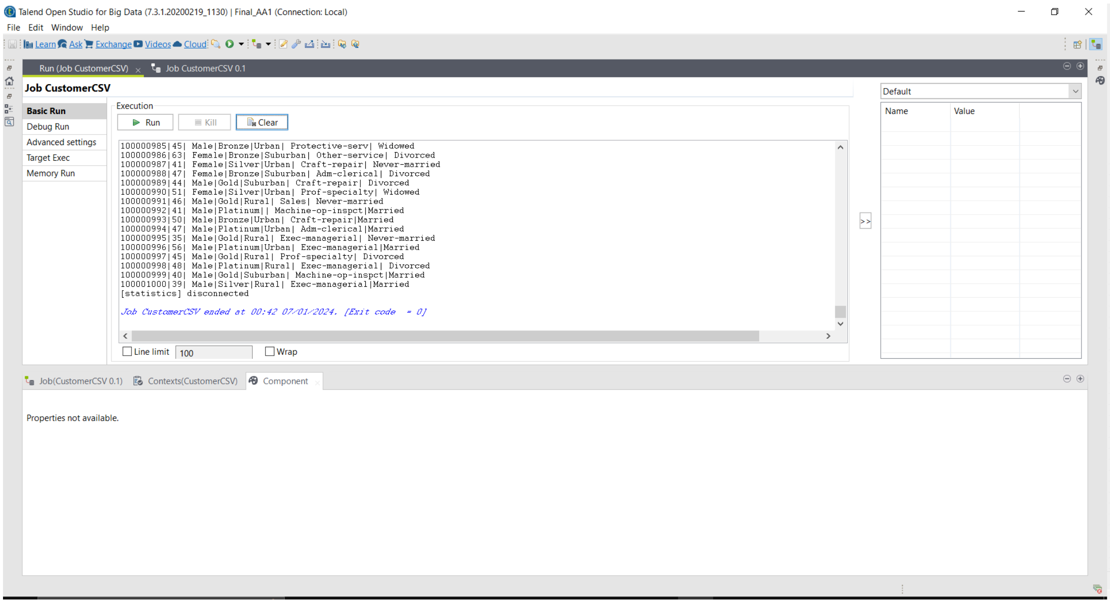
3. Both the transformed datasets were combined into a new dataset of same 1000 rows with the ‘CustomerID’ as the key column using the tMap_1.



4. The flow diagram of the data integration is shown as below:



5. Figure shows the CustomerCSV job has been run completed and a combined dataset named Customer_Combined.csv has been created.

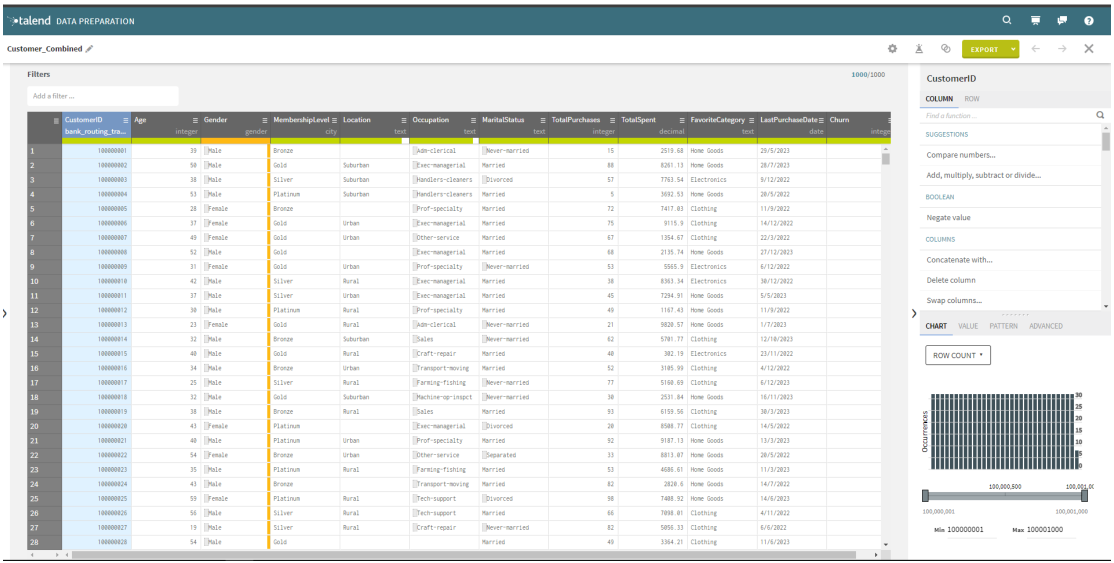


6. Overview of the Customer_Combined.csv dataset.

CustomerID	Age	Gender	Membersl	Location	Occupatio	MaritalStatus	TotalPurchase	TotalSpent	FavoriteCategory	LastPurchaseDate	Churn
100000001	39	Male	Bronze		Adm-cler	Never-married	15	2519.68	Home Goods	29/5/2023	0
100000002	50	Male	Gold	Suburban	Exec-man	Married	88	8261.13	Home Goods	28/7/2023	0
100000003	38	Male	Silver	Suburban	Handlers-	Divorced	57	7763.54	Electronics	9/12/2022	1
100000004	53	Male	Platinum	Suburban	Handlers-	Married	5	3692.53	Home Goods	20/5/2022	1
100000005	28	Female	Bronze		Prof-spec	Married	72	7417.03	Clothing	11/9/2022	1
100000006	37	Female	Gold	Urban	Exec-man	Married	75	9115.9	Clothing	14/12/2022	1
100000007	49	Female	Gold	Urban	Other-ser	Married	67	1354.67	Clothing	22/3/2022	0
100000008	52	Male	Gold		Exec-man	Married	68	2135.74	Home Goods	27/12/2023	1
100000009	31	Female	Gold	Urban	Prof-spec	Never-married	53	5565.9	Electronics	6/12/2022	0
100000010	42	Male	Silver	Rural	Exec-man	Married	38	8363.34	Electronics	30/12/2022	0
100000011	37	Male	Silver	Urban	Exec-man	Married	45	7294.91	Home Goods	5/5/2023	1
100000012	30	Male	Platinum	Rural	Prof-spec	Married	49	1167.43	Home Goods	11/9/2022	1
100000013	23	Female	Gold	Rural	Adm-cler	Never-married	21	9820.57	Home Goods	1/7/2023	0
100000014	32	Male	Bronze	Suburban	Sales	Never-married	62	5701.77	Clothing	12/10/2023	1
100000015	40	Male	Gold	Rural	Craft-rep	Married	40	302.19	Electronics	23/11/2022	1
100000016	34	Male	Bronze	Urban	Transport	Married	52	3105.99	Clothing	4/12/2022	1
100000017	25	Male	Silver	Rural	Farming-f	Never-married	77	5160.69	Clothing	6/12/2023	1
100000018	32	Male	Gold	Suburban	Machine-	Never-married	30	2531.84	Home Goods	16/11/2023	1
100000019	38	Male	Bronze	Rural	Sales	Married	93	6159.56	Clothing	30/3/2023	0
100000020	43	Female	Platinum		Exec-man	Divorced	20	8508.77	Clothing	14/5/2022	1
100000021	40	Male	Platinum	Urban	Prof-spec	Married	92	9187.13	Home Goods	13/3/2023	1
100000022	54	Female	Bronze	Urban	Other-ser	Separated	33	8813.07	Home Goods	20/5/2022	1
100000023	35	Male	Platinum	Rural	Farming-f	Married	53	4686.61	Home Goods	11/3/2023	1
100000024	43	Male	Bronze		Transport	Married	82	2820.6	Home Goods	14/7/2022	1
100000025	59	Female	Platinum	Rural	Tech-sup	Divorced	98	7408.92	Home Goods	14/6/2023	0
100000026	56	Male	Silver	Rural	Tech-sup	Married	66	7098.01	Clothing	4/11/2022	0
100000027	19	Male	Silver	Rural	Craft-rep	Never-married	82	5056.33	Clothing	6/6/2022	1
100000028	54	Male	Gold			Married	49	3364.21	Clothing	11/6/2023	0
100000029	39	Male	Silver	Urban	Exec-man	Divorced	79	1170.37	Electronics	24/7/2022	0
100000030	49	Male	Platinum	Rural	Craft-rep	Married	63	7826.93	Clothing	27/6/2023	1
100000031	23	Male	Gold	Rural	Protectiv	Never-married	41	4700.56	Clothing	28/5/2022	1
100000032	20	Male	Platinum	Suburban	Sales	Never-married	79	3730.83	Clothing	24/6/2022	0
100000033	45	Male	Silver	Rural	Exec-man	Divorced	80	1672.03	Home Goods	23/5/2022	1
100000034	30	Male	Platinum	Urban	Adm-cler	Married	93	2684	Home Goods	7/10/2023	1
100000035	22	Male	Silver	Suburban	Other-ser	Married	71	6556.31	Home Goods	12/10/2023	1
100000036	48	Male	Silver	Suburban	Machine-	Never-married	57	5648.86	Home Goods	14/2/2023	1

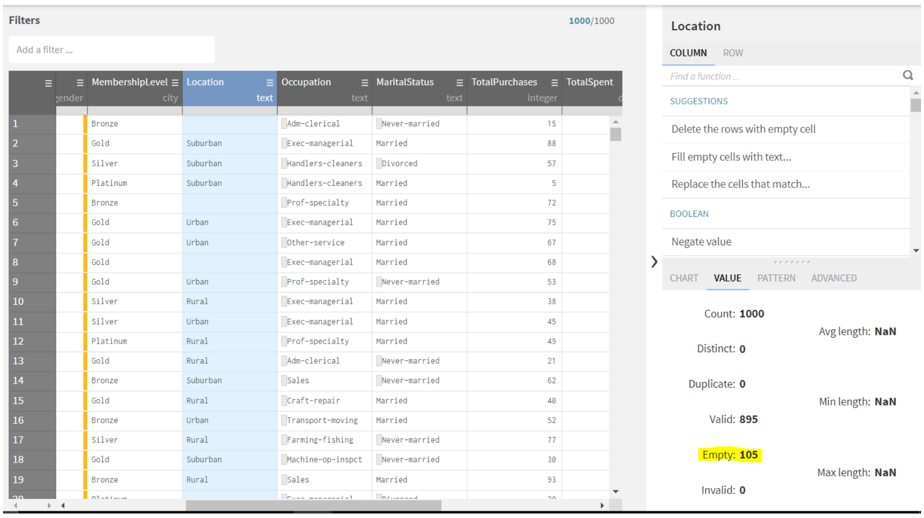
Data preprocessing:

1. Customer_Combined.csv dataset was imported into Talend Data Preparation software to undergo data preprocessing. Talend Data Preparation is a tool specifically designed for data preprocessing and data wrangling tasks. It provides a range of features that make it well-suited for preparing and cleaning data before it is used in analytical processes, machine learning models, or other downstream applications. It provides data profiling capabilities that allow users to understand the characteristics and quality of their data. Profiling helps in identifying issues such as missing values, outliers, and inconsistencies. It also offers various data cleaning and standardization functions, allowing users to address common data quality issues, such as handling missing values, correcting errors, and standardizing formats.



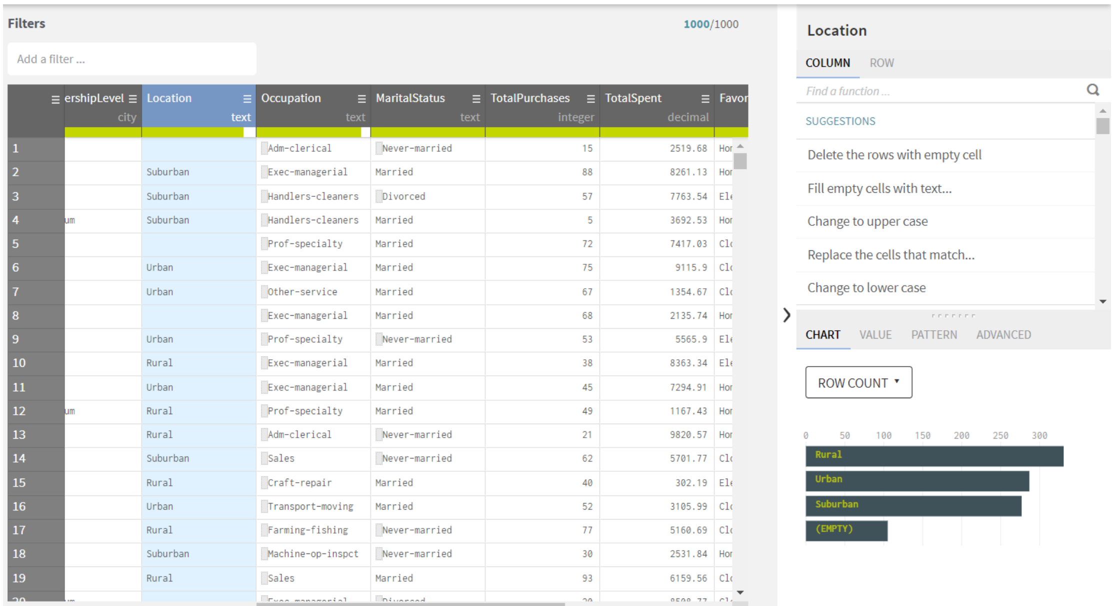
Data profiling:

2. There are some data quality issues faced in the combined datasets such as missing values and formatting issues. In the figure below, variable 'Location' contains 105 missing values.

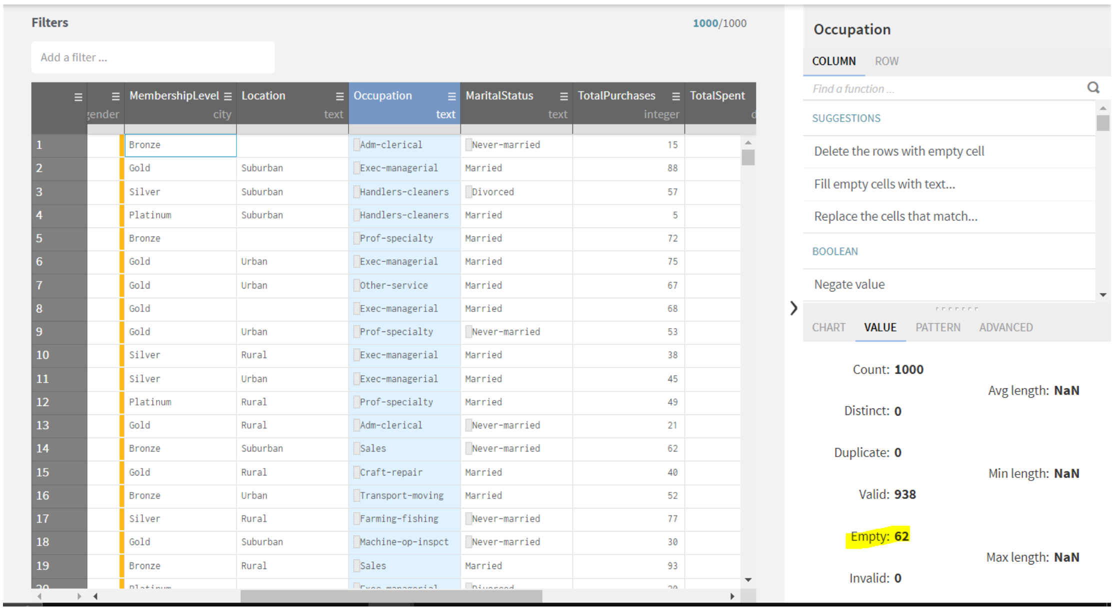


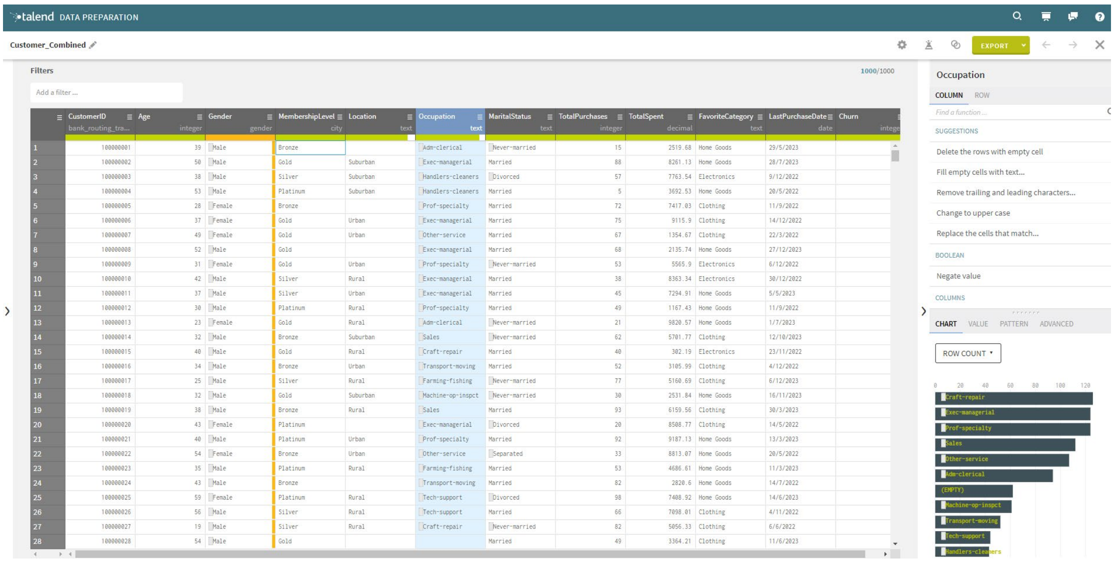
3. Since variable 'Location' is a categorical variable and the missing values were not significantly large (~10% of the total rows), it is best to be imputed with mode

imputation than using listwise deletion. Mode imputation allows us to retain all available data for analysis, even when there are missing values in the categorical variable. This helps preserve valuable information and maintains a larger sample size for analysis. Besides, listwise deletion involves removing entire observations with missing values in any variable. This can introduce bias if the missing data is not completely random and is related to the outcome or other variables. Mode imputation, on the other hand, provides a way to fill in missing values based on the observed patterns in the data, potentially reducing bias. As shown below, the mode for the variable 'Location' is 'Rural', hence the missing values were filled with the mode.

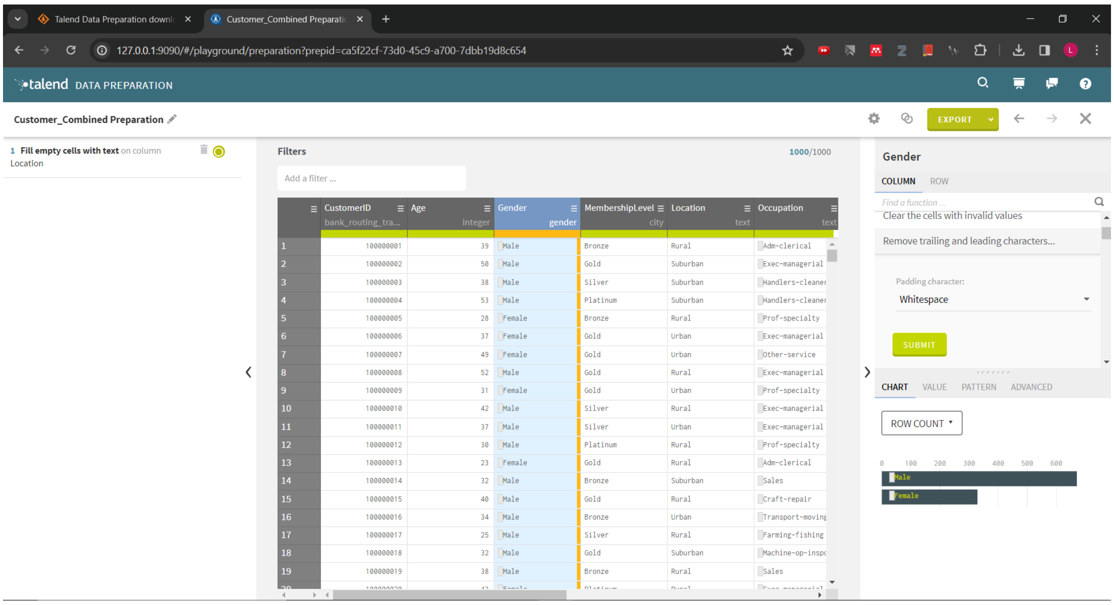


4. Similarly, the categorical variable 'Occupation' has small portion of missing values (63 cells) and was imputed by the mode - 'Craft-repair'.

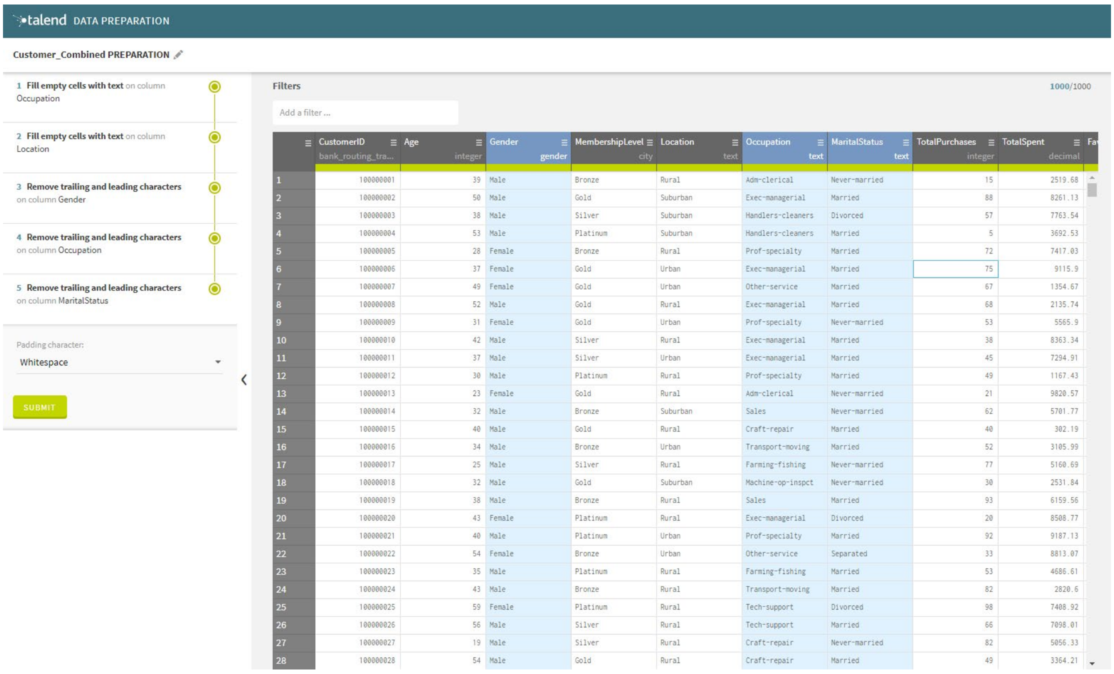




5. In term of data formatting, there are some inconsistencies such as additional whitespace in the cells. This issue was identified in variables ‘Gender’, ‘Occupation’ and ‘Marital Status’.

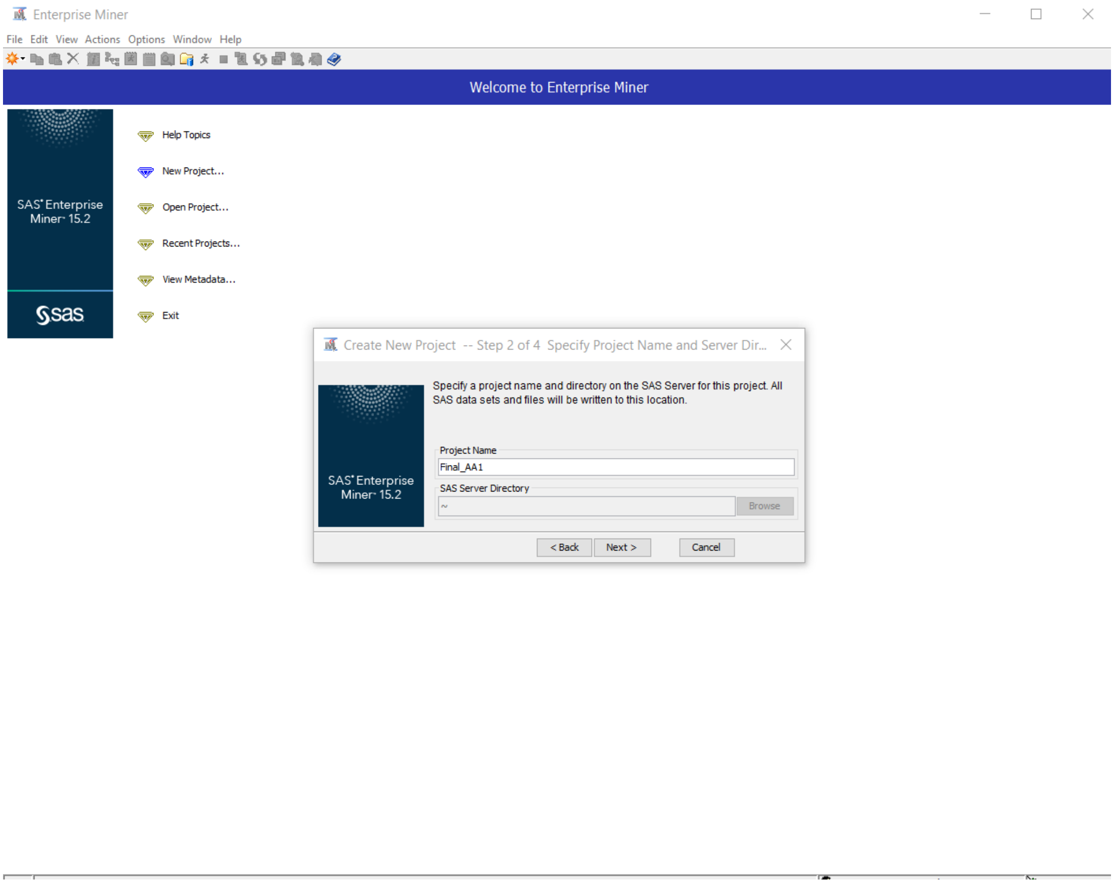


6. Therefore, the whitespaces were trimmed by using the ‘Remove trailing and leading characters’ in the 3 identified columns. The final dataset are complete as per confirmed by the green labelled on top of the variables. The dataset was exported into CSV with the name Customer_Combined Preparation.

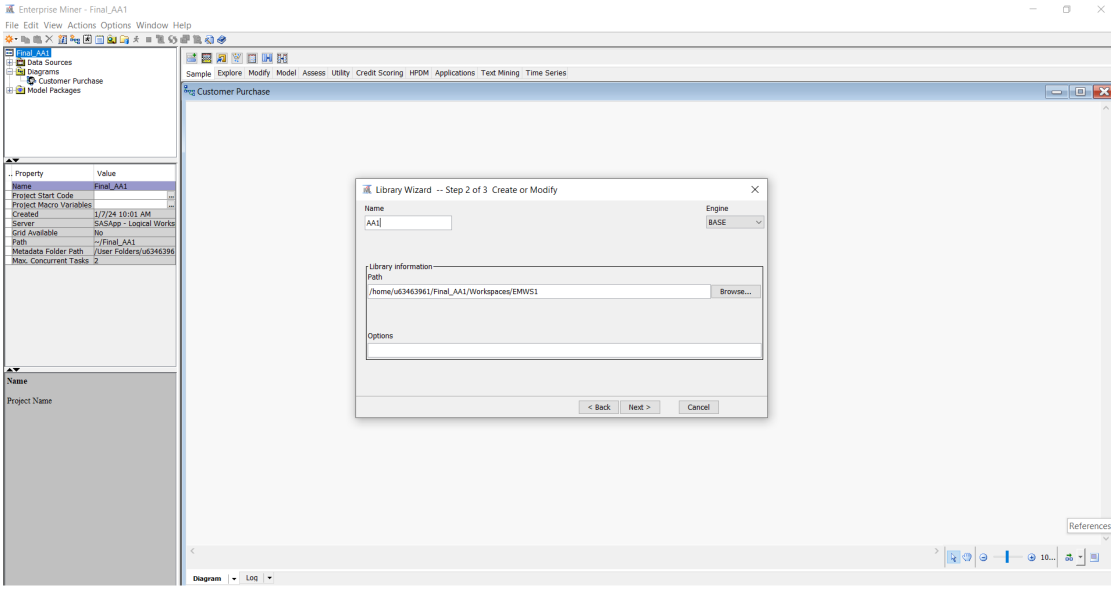

talend DATA PREPARATION
Customer_Combined PREPARATION
1 Fill empty cells with text on column Occupation
2 Fill empty cells with text on column Location
3 Remove trailing and leading characters on column Gender
4 Remove trailing and leading characters on column Occupation
5 Remove trailing and leading characters on column MaritalStatus
Padding character:
Whitespace
SUBMIT
Filters
1000/1000
Add a filter ...
CustomerID bank_routing_tra... | Age integer | Gender gender | MembershipLevel city | Location text | Occupation text | MaritalStatus text | TotalPurchases integer | TotalSpent decimal | Fa
1 100000001 39 Male Bronze Rural Adm-clerical Never-married 15 2519.68
2 100000002 50 Male Gold Suburban Exec-managerial Married 88 8261.13
3 100000003 38 Male Silver Suburban Handlers-cleaners Divorced 57 7763.54
4 100000004 53 Male Platinum Suburban Handlers-cleaners Married 5 3692.53
5 100000005 28 Female Bronze Rural Prof-specialty Married 72 7417.03
6 100000006 37 Female Gold Urban Exec-managerial Married 75 9115.9
7 100000007 49 Female Gold Urban Other-service Married 67 1354.67
8 100000008 52 Male Gold Rural Exec-managerial Married 68 2135.74
9 100000009 31 Female Gold Urban Prof-specialty Never-married 53 5565.9
10 100000010 42 Male Silver Rural Exec-managerial Married 38 8363.34
11 100000011 37 Male Silver Urban Exec-managerial Married 45 7294.91
12 100000012 30 Male Platinum Rural Prof-specialty Married 49 1167.43
13 100000013 23 Female Gold Rural Adm-clerical Never-married 21 9820.57
14 100000014 32 Male Bronze Suburban Sales Never-married 62 5701.77
15 100000015 40 Male Gold Rural Craft-repair Married 40 302.19
16 100000016 34 Male Bronze Urban Transport-moving Married 52 3105.99
17 100000017 25 Male Silver Rural Farming-fishing Never-married 77 5160.69
18 100000018 32 Male Gold Suburban Machine-op-inspct Never-married 30 2531.84
19 100000019 38 Male Bronze Rural Sales Married 93 6159.56
20 100000020 43 Female Platinum Rural Exec-managerial Divorced 20 8508.77
21 100000021 40 Male Platinum Urban Prof-specialty Married 92 9187.13
22 100000022 54 Female Bronze Urban Other-service Separated 33 8813.07
23 100000023 35 Male Platinum Rural Farming-fishing Married 53 4686.61
24 100000024 43 Male Bronze Rural Transport-moving Married 82 2820.6
25 100000025 59 Female Platinum Rural Tech-support Divorced 98 7408.92
26 100000026 56 Male Silver Rural Tech-support Married 66 7098.01
27 100000027 19 Male Silver Rural Craft-repair Never-married 82 5056.33
28 100000028 54 Male Gold Rural Craft-repair Married 49 3364.21

Exploratory Data Analysis and Data Modelling:

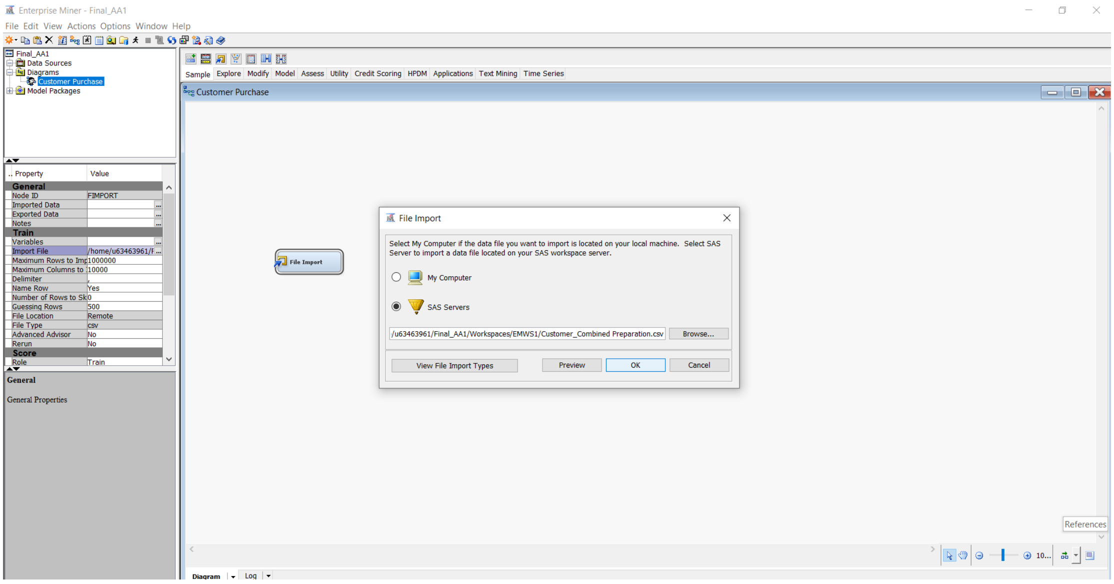
1. For exploratory data analysis and data modelling, SAS Enterprise Miner was used as it provides an integrated environment for EDA and data modelling. Users can seamlessly transition from data exploration to model development and deployment within a single platform. The tool allows us to compare multiple models and evaluate their performance using various metrics. This helps in selecting the best-performing models for Customer_Combined Preparation.csv dataset. As shown below, a new project was created and named as 'Final_AA1'.



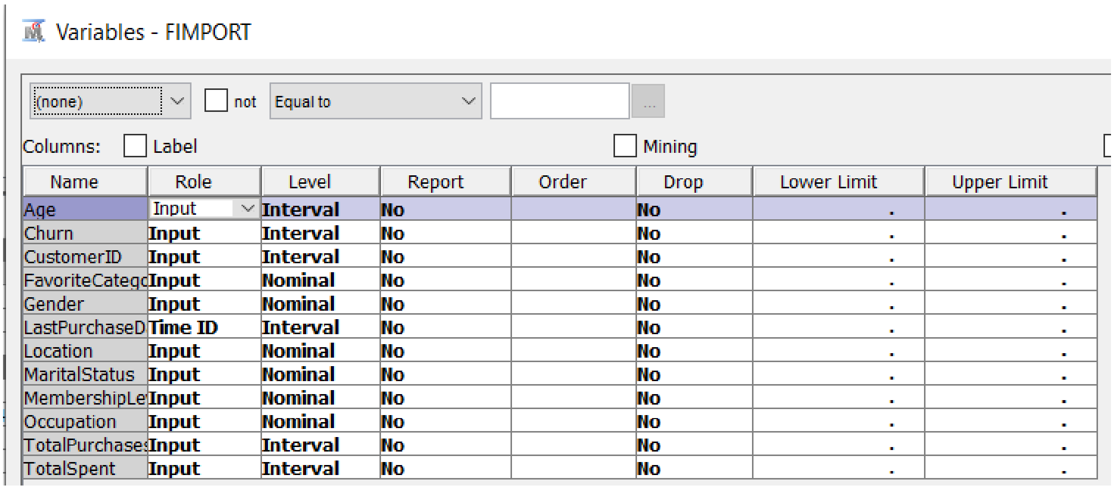
2. A new library named 'AA1' was created with the directory shown as below which contains the Customer_Combined Preparation.csv dataset.



3. The Customer_Combined Preparation.csv was imported from the new created library located in the SAS Servers.



4. All the variables are identified automatically as the input role except variable ‘LastPucchaseDate’ which was identified as ‘Time ID’ role. The levels are divided into interval and nominal level.


Variables - FIMPORT

(none) | not | Equal to

Columns: Label | Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Input	Interval	No		No	.	.
CustomerID	Input	Interval	No		No	.	.
FavoriteCatego	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchaseD	Time ID	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
MaritalStatus	Input	Nominal	No		No	.	.
MembershipLe	Input	Nominal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
TotalPurchase	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

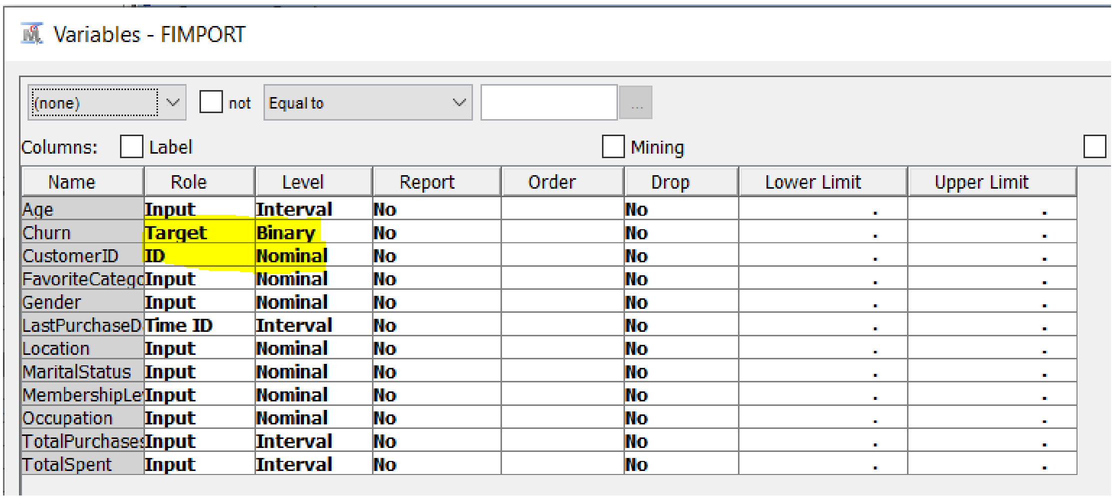
5. For each of the variable, their roles and levels are maintained with following reasons:

Variables	Role	Role Reasons	Level	Level Reasons
Age	Input	Age is a numeric variable that can be used as input for predicting customer behavior. It provides valuable information about the demographic of the customers.	Interval	Age is a numeric variable with a meaningful order and magnitude
Churn	Target	Churn is the variable to predict. It represents whether a customer has churned or not, making it the target variable for predictive modelling	Binary	Churn is typically a binary variable indicating whether a customer has churned (1) or not (0). It is nominal with two categories
CustomerID	ID	CustomerID is typically used as an identifier and does not contribute to the predictive modelling process. It helps uniquely identify each record but doesn't provide predictive information.	Nominal	CustomerID is an identifier and does not have an inherent order or magnitude.
Gender	Input	Gender is a categorical variable that can be used as input for predicting customer behavior. It may help capture gender-specific patterns.	Nominal	Gender is a categorical variable without a natural order.

MembershipLevel	Input	MembershipLevel is likely a categorical variable indicating the level of membership. It can be useful for predicting customer behavior, especially if different levels have distinct characteristics.	Nominal	MembershipLevel is a categorical variable without a natural order.
Location	Input	Location is a categorical variable that could be relevant for predicting customer behavior. It may capture geographical patterns.	Nominal	Location is a categorical variable without a natural order.
Occupation	Input	Occupation is a categorical variable that can provide insights into the type of work a customer is engaged in, which may influence their behavior.	Nominal	Occupation is a categorical variable without a natural order.
MaritalStatus	Input	MaritalStatus is a categorical variable that can be relevant for predicting customer behavior. It may capture differences in purchasing behavior based on marital status.	Nominal	MaritalStatus is a categorical variable without a natural order.
TotalPurchases	Input	TotalPurchases is a numeric variable and can be considered an input for predicting customer behavior.	Interval	TotalPurchases is a numeric variable with a meaningful order and magnitude
TotalSpent	Input	TotalSpent is a numeric variable and can be considered an input for predicting customer behavior. It represents the total amount spent by a customer.	Interval	TotalSpent is a numeric variable with a meaningful order and magnitude
FavoriteCategory	Input	FavoriteCategory is a categorical variable representing the customer's preferred product category. It can provide valuable insights into customer preferences.	Nominal	FavoriteCategory is a categorical variable without a natural order.
LastPurchaseDate	TimeID	LastPurchaseDate is a timestamp variable indicating when the last purchase was made. It can be used as an input to capture recency effects in	Interval	LastPurchaseDate is a timestamp variable.

		predicting churn or other behaviors.		

6. Figure below shows the variable roles after modification. The modified roles and levels are highlighted in yellow:



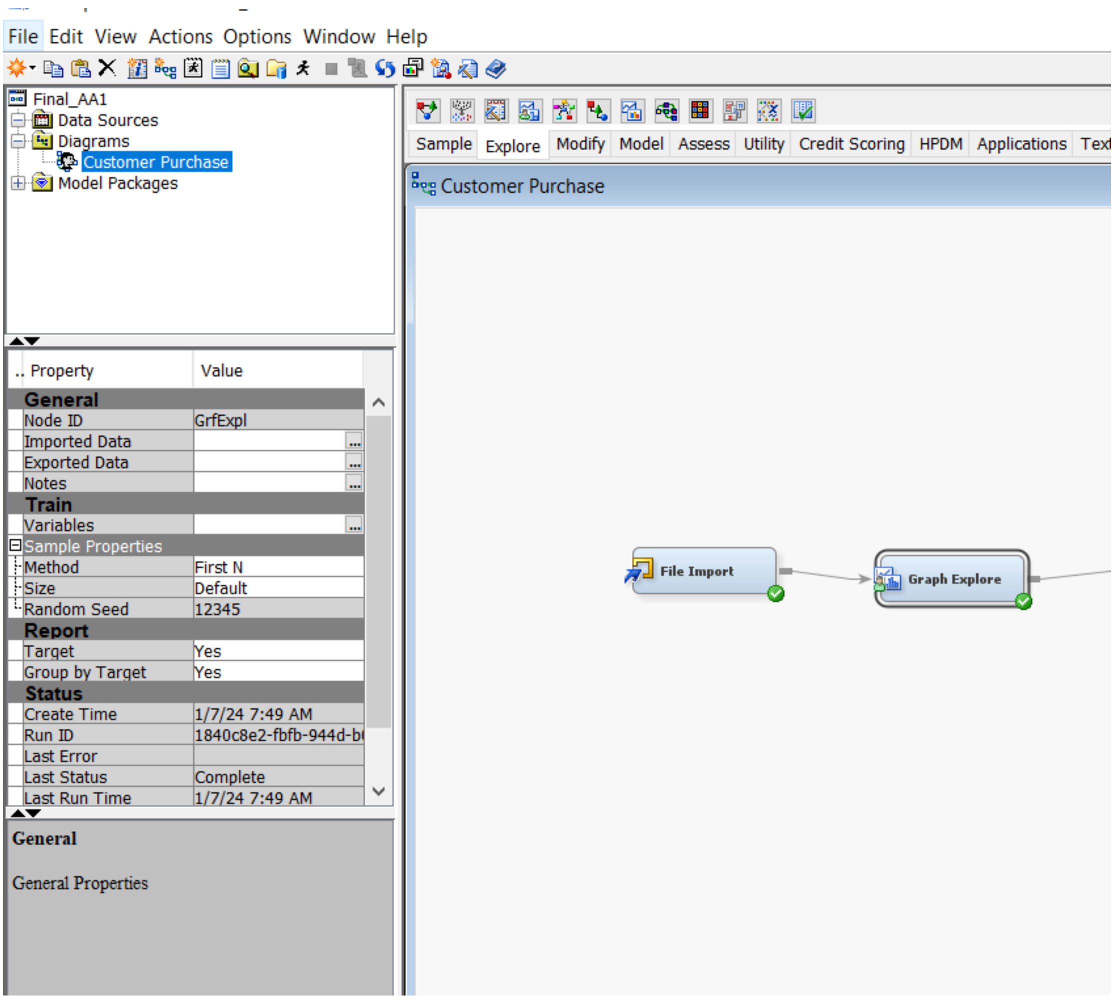
Variables - FIMPORT

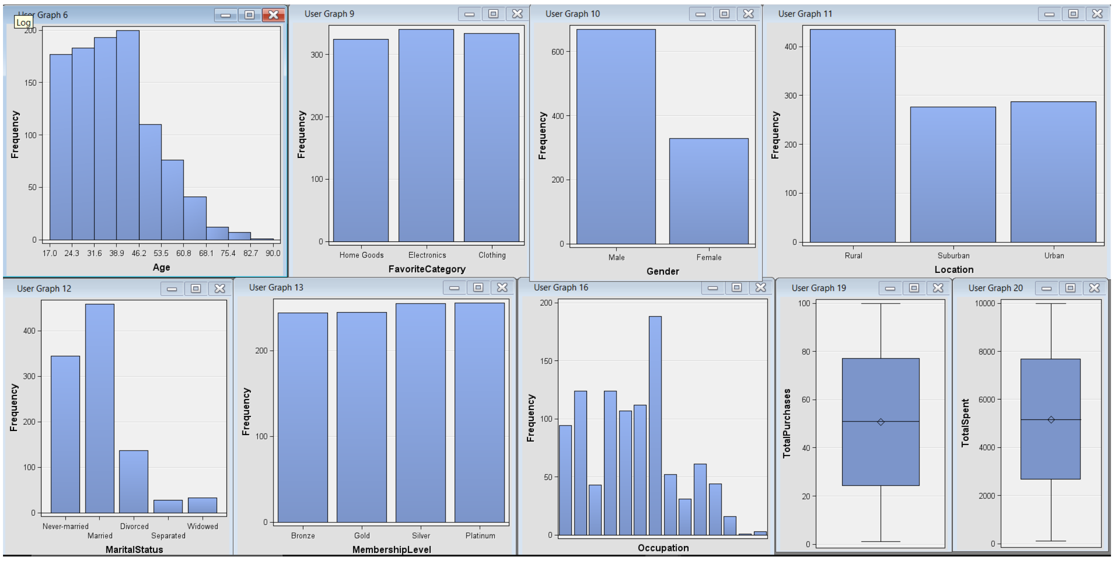
(none) not Equal to

Columns: Label Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Binary	No		No	.	.
CustomerID	ID	Nominal	No		No	.	.
FavoriteCategc	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchaseD	Time ID	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
MaritalStatus	Input	Nominal	No		No	.	.
MembershipLe	Input	Nominal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
TotalPurchase	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

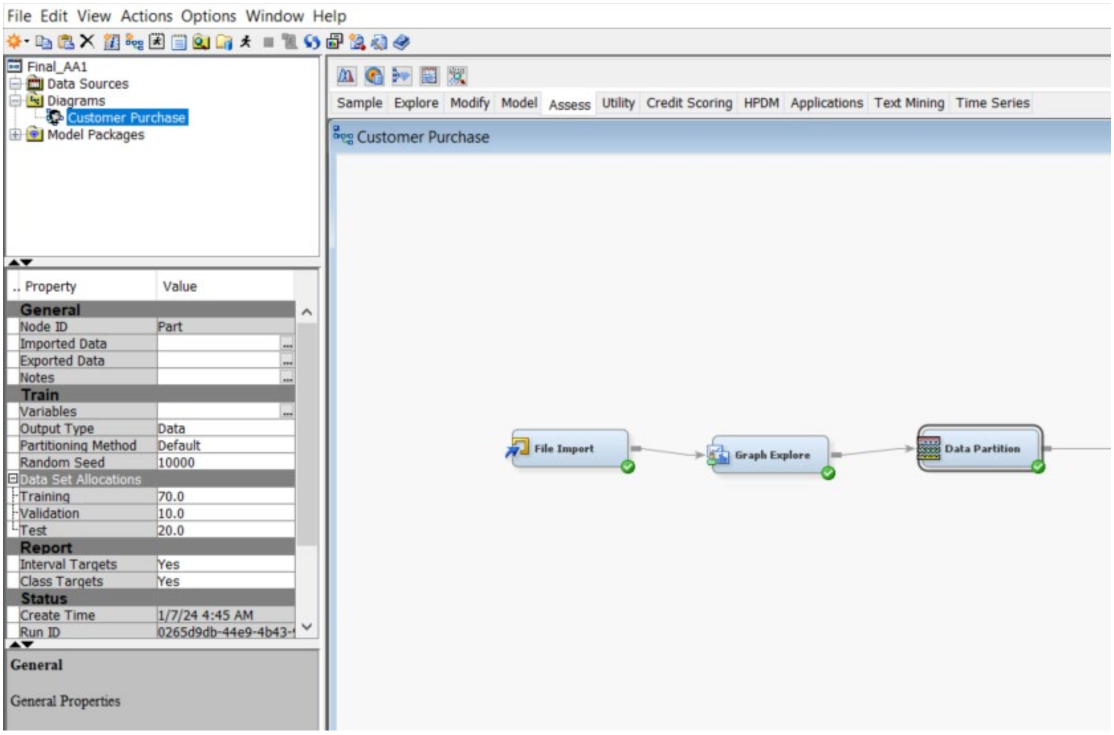
7. The distributions of each variable were shown in figures below by using Graph Explore node.





Data Modelling:

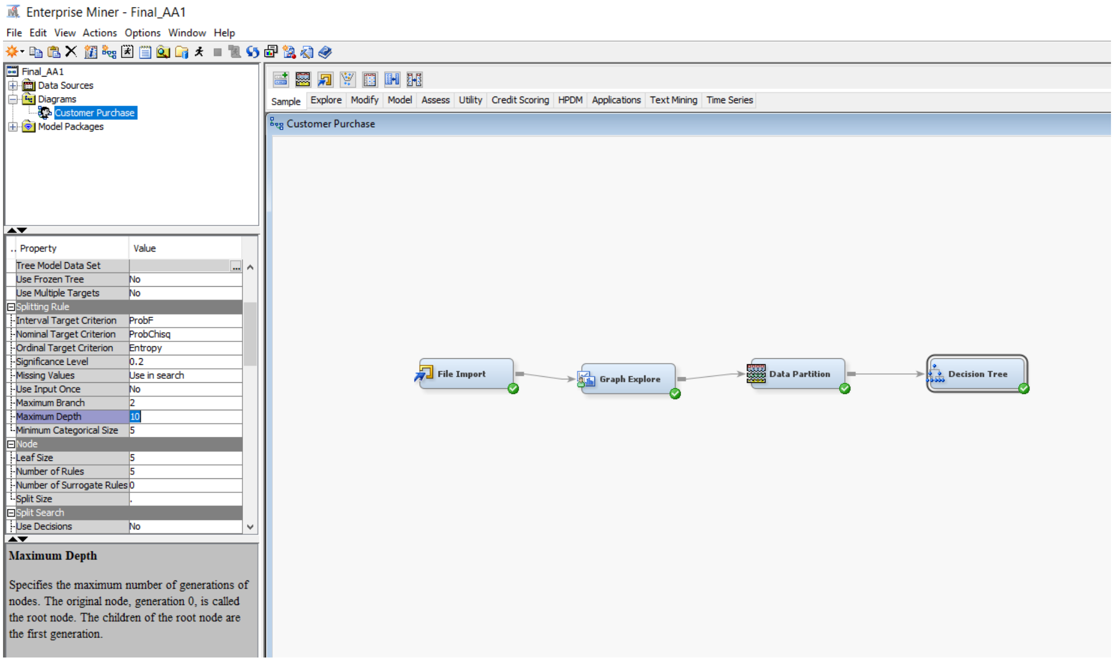
1. Before applying decision tree model, the 'Data Partition' node was added to the diagram to separate the dataset into training, validation and testing sets with ratios of 0.70, 0.10 and 0.20.



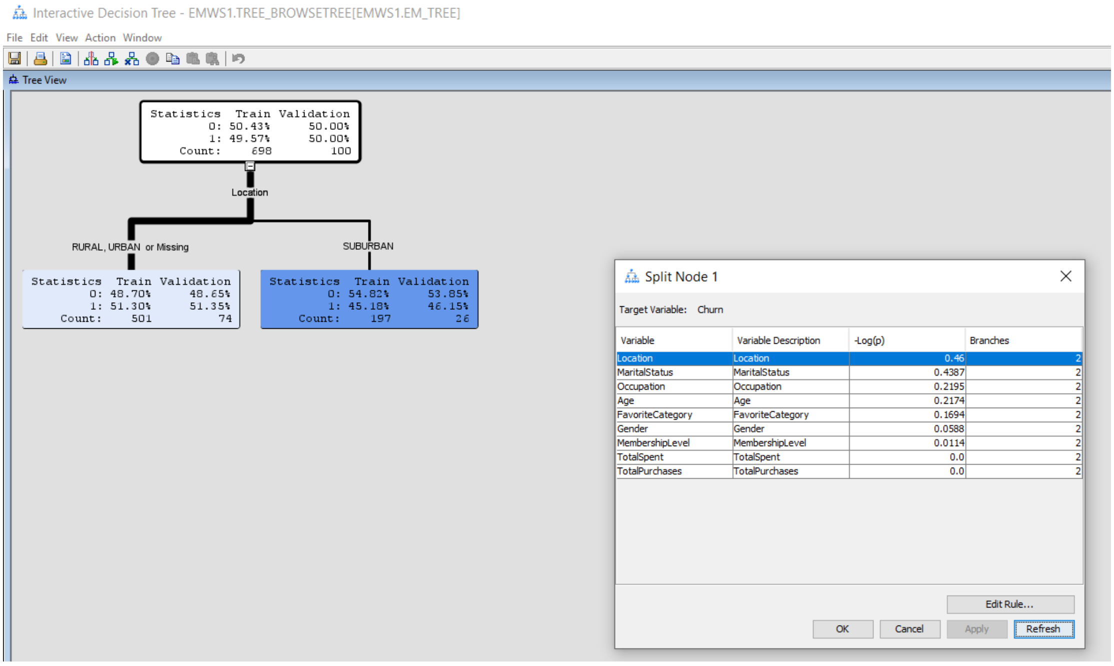
2. The dataset partition summary is shown as below:

```
Partition Summary

                                     Number of
Type             Data Set          Observations

DATA        EMWS1.GrfExpl_TRAIN        1000
TRAIN       EMWS1.Part_TRAIN            698
VALIDATE    EMWS1.Part_VALIDATE         100
TEST        EMWS1.Part_TEST             202


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


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

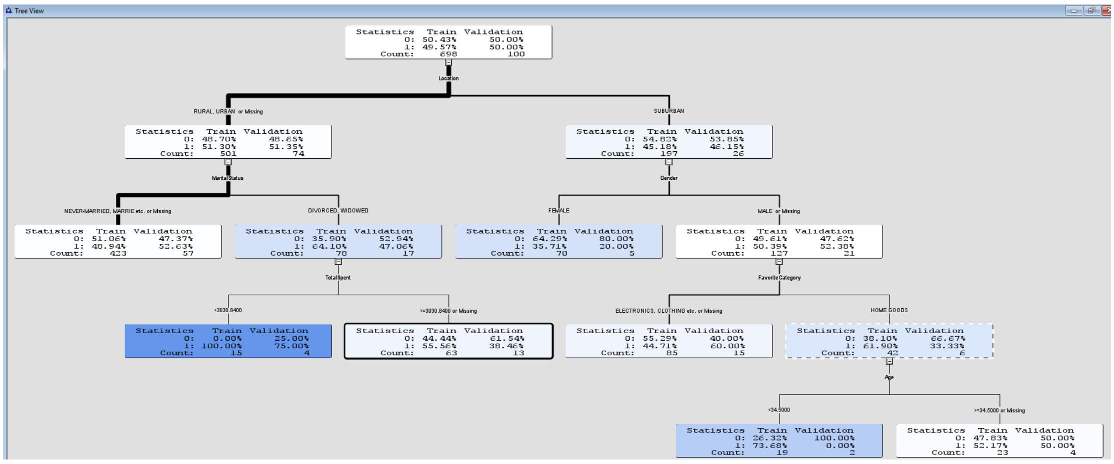
3. A decision tree node was connected to the partition dataset. The maximum depth of the decision tree was set to 10 levels to avoid overfitting.



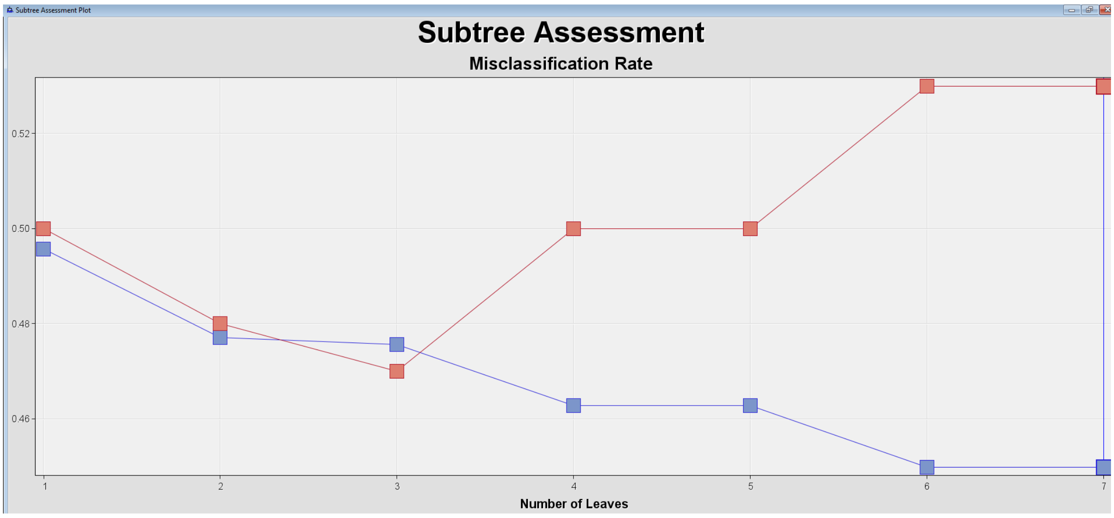
4. The goal in decision tree algorithms is to split the data in a way that minimizes entropy. Therefore, in the interactive decision tree, the node was split manually by referring to the largest values of -log(p).



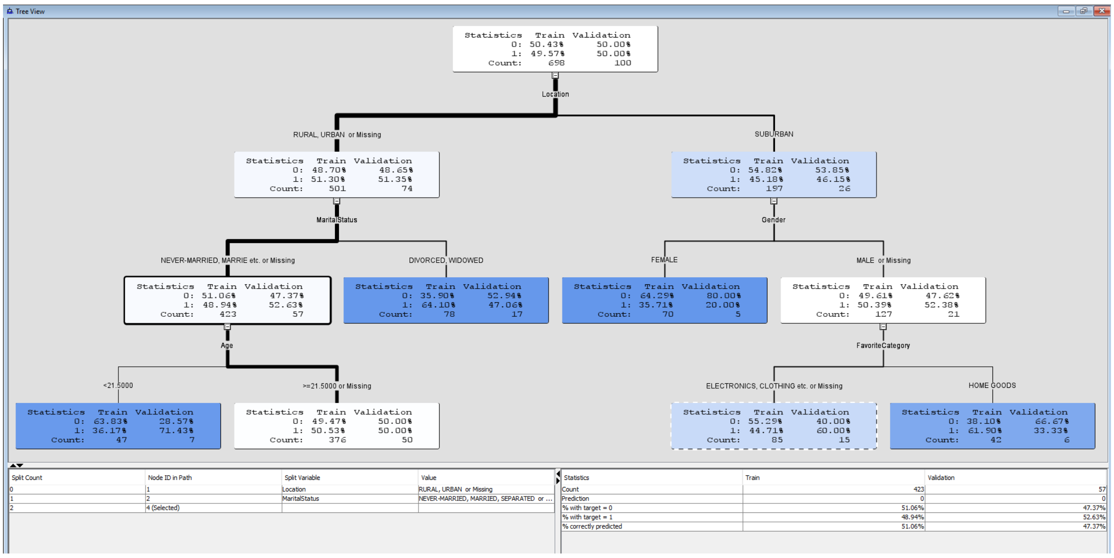
5. The nodes were split until there was no more information gain or the sample is too small. The branches from the tree that do not provide significant improvement in predictive accuracy were being removed. The final decision tree was shown as below:



6. As shown in the figure below, increasing the number of leaves will significantly increase the misclassification rate in the validation set. The divergence in the misclassification rate of training and validation sets introduced overfitting problem. Therefore, the number of leaves has to be reduced.



7. The leaves in which the prediction of training and validation sets differ significantly were removed. The new decision tree developed as figure below:



From the decision tree table created, there are several observations that could be made:

Leaf 1: Customers who are located in rural or urbans, never married or married, and age below 21.5 has 36.17% probability of churn in 6 months.

Leaf 2: Customers who are located in rural or urbans, never married or married, and age above 21.5 has 50.53% probability of churn in 6 months.

Leaf 3: Customers who are located in rural or urbans, divorced or widowed has 64.10% probability of churn in 6 months.

Leaf 4: Customers who are located in suburban area and female has 35.71% probability of churn in 6 months.

Leaf 5: Customers who are located in suburban area, male, and favourite categories are Electronics and Clothing has 44.71% probability of churn in 6 months.

Leaf 6: Customers who are located in suburban area, male, and favourite category is home goods has 61.90% probability of churn in 6 months.

Business insights from the graphs:

1. **Demographic Factors Impact Churn:**
   - The location (rural, urban, suburban) and marital status are identified as factors influencing churn. Understanding these demographics can help in tailoring marketing strategies or retention efforts based on geographical and marital characteristics.
2. **Age as a Churn Predictor:**
   - The age threshold of 21.5 appears to be a significant factor in predicting churn. Younger customers (below 21.5) have a lower probability of churn compared to older customers (above 21.5). This information can guide marketing campaigns or promotions targeted at specific age groups.
3. **Marital Status and Churn:**

- Divorced or widowed customers have a higher probability of churn. This group might require special attention and retention strategies. Understanding the reasons behind their churn could provide insights into service or product adjustments.

4. **Gender and Location Influence:**

   - There is a specific leaf for customers located in suburban areas who are male, indicating that gender and location may be combined factors in predicting churn. Tailoring retention strategies for this group may be necessary.

5. **Product Category Preferences Impact Churn:**

   - Customers who prefer Electronics and Clothing have different probabilities of churn compared to those who prefer Home Goods. This insight suggests that understanding product preferences can be valuable in predicting and mitigating churn.

6. **Fine-Tuning Marketing Efforts:**

   - By knowing the probability of churn for different customer segments, marketing efforts can be fine-tuned. For example, efforts to retain customers with a higher likelihood of churn can be prioritized, and promotions can be customized based on the identified characteristics.

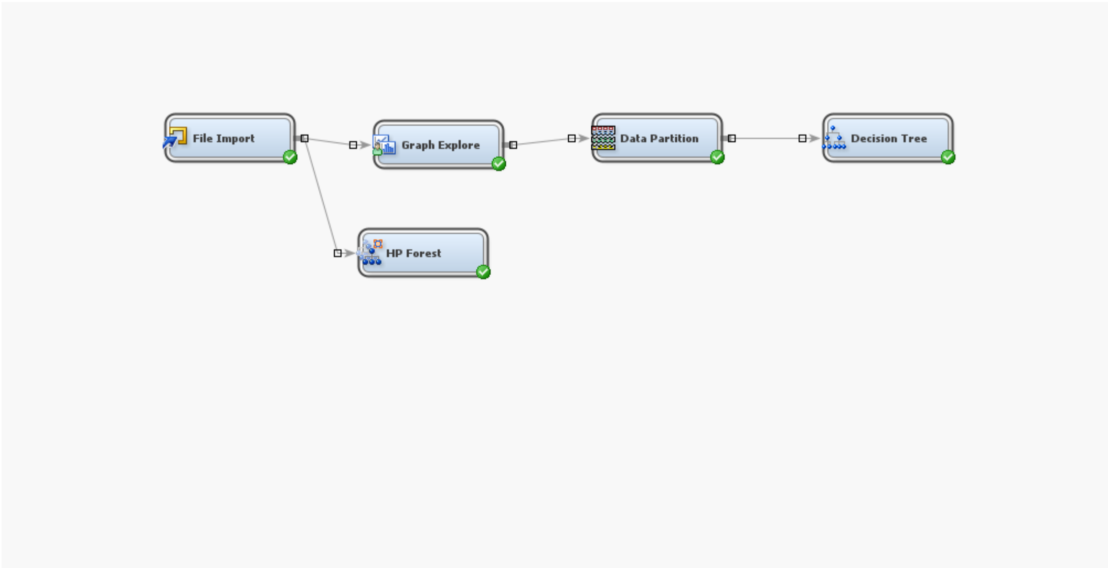
7. **Segment-Specific Retention Strategies:**

   - The information allows for the development of segment-specific retention strategies. For instance, strategies for retaining divorced or widowed customers may differ from those targeting younger, unmarried individuals.

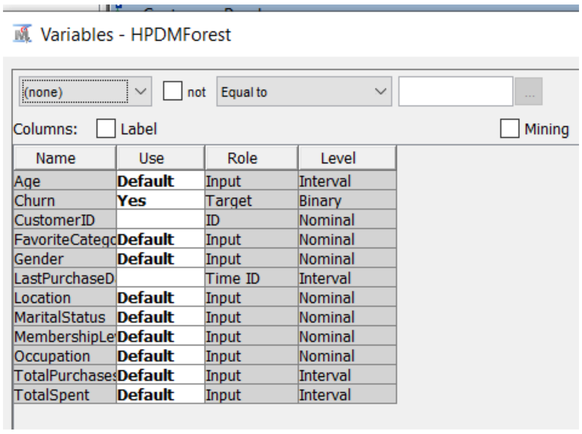
Ensemble methods:

1. To apply Bagging and Boosting, using the Random Forest algorithm

as a Bagging example, a new path was created from the data source on to the diagram. A Forest Node was dropped to the diagram and connected to the data source.



2. The target variable is 'Churn' and the other features are remained as default.



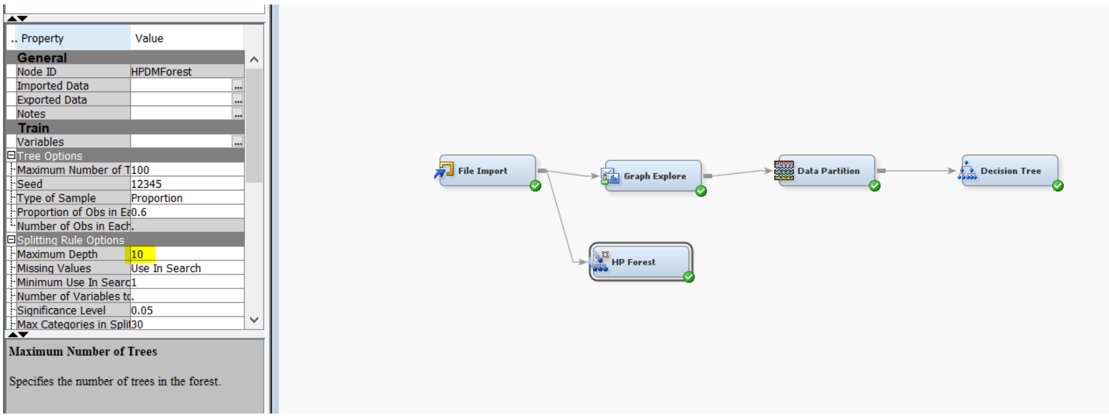
Variables - HPDMForest

(none) not Equal to

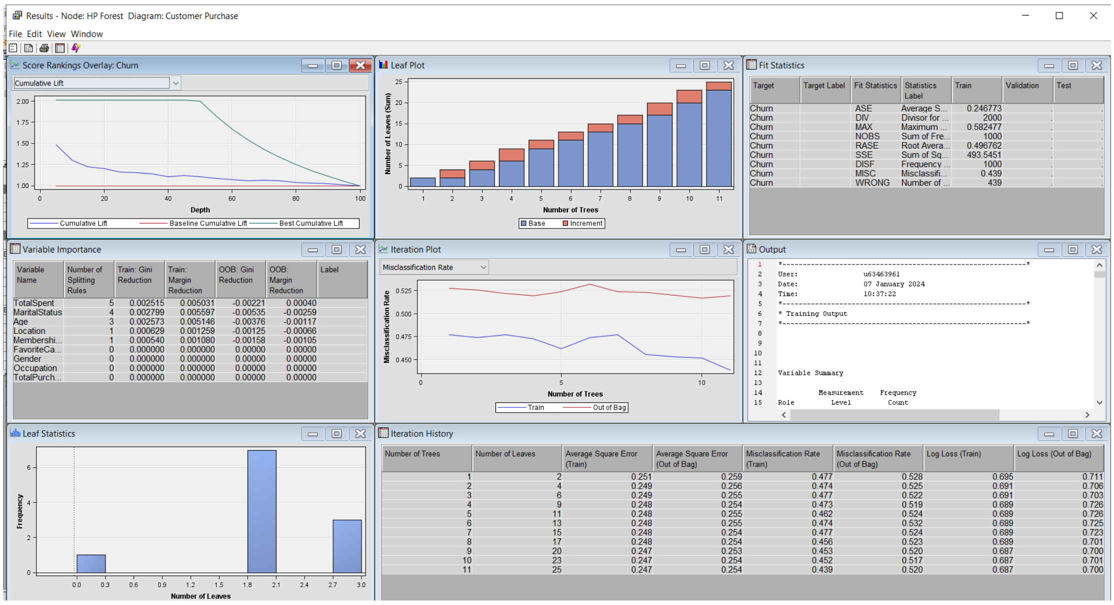
Columns: Label Mining

Name	Use	Role	Level
Age	**Default**	Input	Interval
Churn	**Yes**	Target	Binary
CustomerID		ID	Nominal
FavoriteCateg	**Default**	Input	Nominal
Gender	**Default**	Input	Nominal
LastPurchaseD		Time ID	Interval
Location	**Default**	Input	Nominal
MaritalStatus	**Default**	Input	Nominal
MembershipLe	**Default**	Input	Nominal
Occupation	**Default**	Input	Nominal
TotalPurchase	**Default**	Input	Interval
TotalSpent	**Default**	Input	Interval

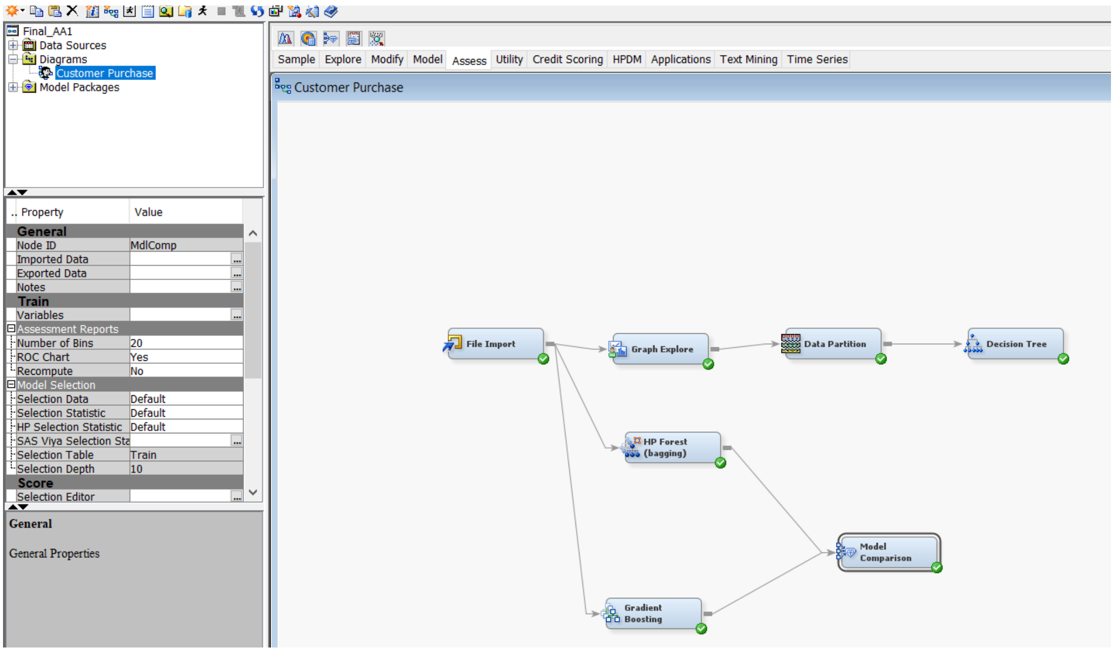
3. The other settings such as the number of trees, maximum depth, and other parameters are set as follows:



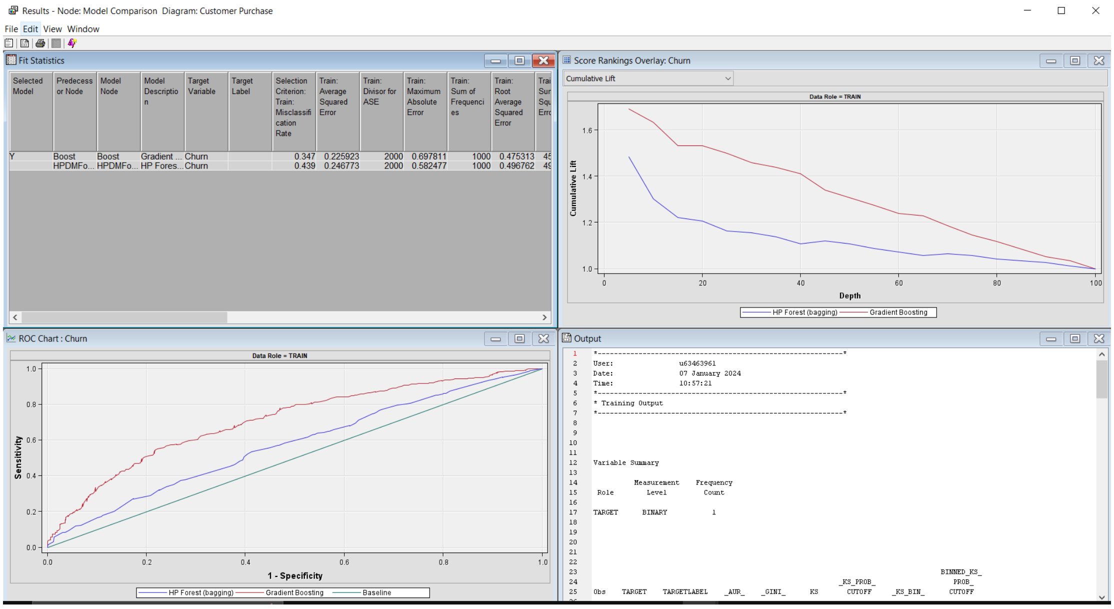
4. The results of the random forest train run are as shown in figure below:



5. To apply boosting, a gradient boosting node was added into the diagram. The setting was set as the left panel.



6. The result of the model comparison is shown as follow:



The ideal point on the ROC chart is the top-left corner (0,1), where the True Positive Rate is 1 (100% sensitivity) and the False Positive Rate is 0 (0% false positives). Achieving a point closer to this corner is indicative of better model performance. The Gradient boosting slightly overperform as compared to HP Forest, however, both performances are not too satisfactory due to low AUC Score.