UNIVERSITY OF MALAYA

WQD7005 Data Mining
1/2023/2024

Alternative Assessment 1

Name	**: Soh Pei Lin**
Matrix Number	**: S2193368**
Group	**: 1**
Github Link	**:** https://github.com/Peilin-CodeVersion/Alternative-Assessment-7005.git

Background

In today's rapidly evolving e-commerce landscape, understanding customer behaviour and preferences is essential for all the business to thrive. This project study on a comprehensive dataset of customer transactions from an e-commerce which capturing with customer attributes and their purchase history within 2 years by using 3 tools which is Talend Data Integration, Talend Data Preparation and SAS Enterprise Miner. By study through the dataset, the objective of this study aims to discover the patterns and trends of customer behaviour and identify with the key factors of customer make purchasing decision.

Dataset Description

The dataset is provided with the 2000 customer transaction records within 2 years, including with personal demographics, revenue and purchasing patterns. Table 1 showed with the variables' description.

Table 1: Variable Description

Variable	**Description**
CustomerID	Unique identifier for each customer
Age	Age of the customer
Gender	Gender of the customer
Location	Geographic location of the customer
MembershipLevel	Membership level
TotalPurchases	Total number of purchases made by the customer
TotalSpent	Total amount spent by the customer
FavoriteCategory	The category in which the customer most frequently shops
LastPurchaseDate	The date of the last purchase
Occupation	Customer's job or profession
FrequencyWebsiteVisits	Frequency of customer visits the website
LoyaltyPoints	Points accumulated by the customer
EmailSubscribed	Indicates whether the customer has opted to receive marketing emails
AverageReviewRating	Average rating given by the customer in product or service reviews
Churn	Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

Description of Solution

In the research, Data mining SEMMA method has been applied as methodology to perform with the task of data import and preprocessing, decision tree analysis and ensemble methods. In SEMMA, we will cover with Sample, Explore, Modify, Model and Assess. Figure 1 shows with the summary of tools implement in each segment.

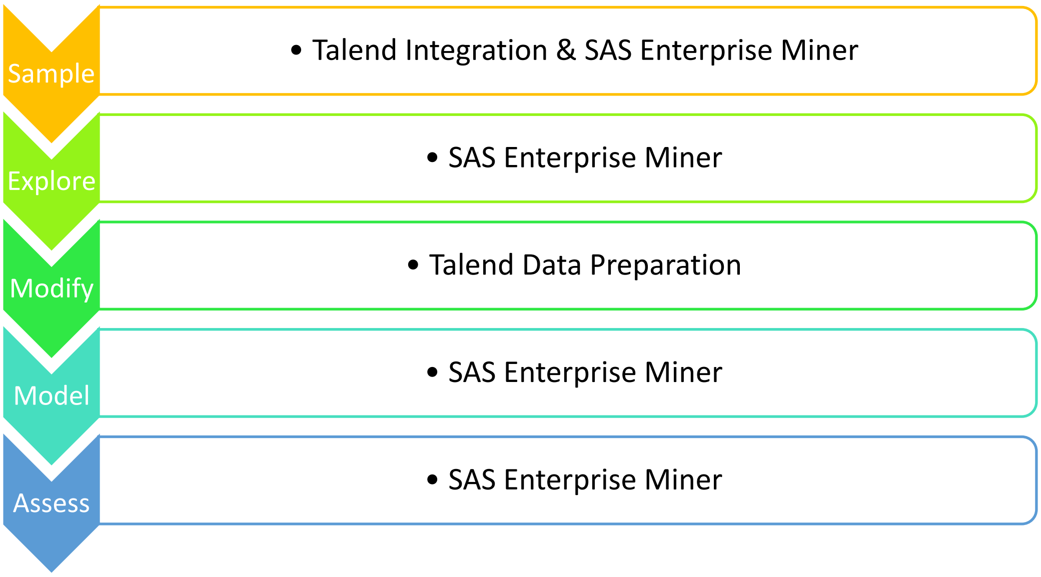


Figure 1: Summary of tools implement in each segment

SEMMA: SAMPLE

In SAMPLE stage, we collect with the data tables with essential customer transaction history. In this step, we consider with the relevant variables such as Total Purchases, Total Spent, Favourite Category, Last Purchase Date and Occupation. There are two datasets collected which is **customer's personal demographic** and **customer transaction history**. Each dataset is imported inside Talend Integration tool for data joining with unique identified CustomerID. After joining the data, the dataset is exported as CSV file. The procedure on data joining by using **Talend Integration tool** is listed in Appendix (1)(Talend Integration Tool)(i-ix). Figure 2 showed with the flow of data joining and data export.

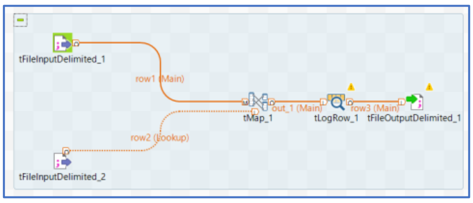
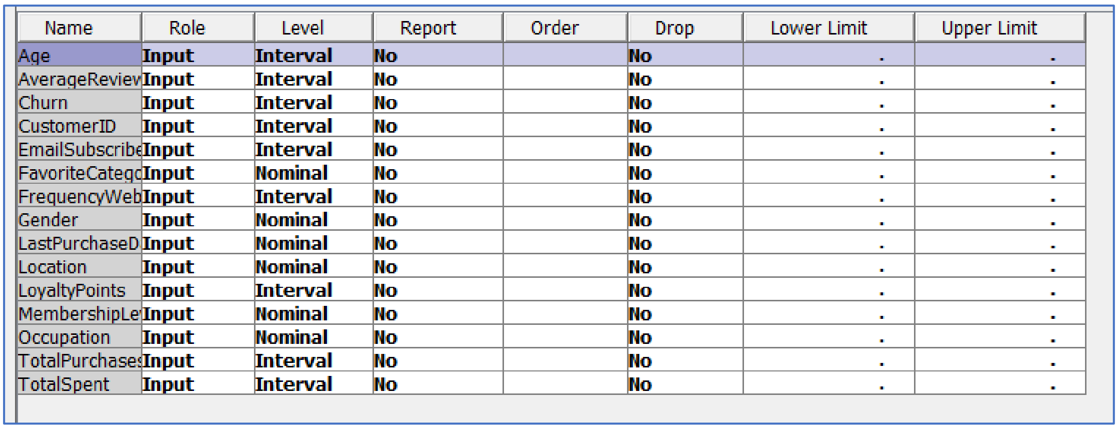


Figure 2: Talend Integration tool workflow

Next, data is saved and sample inside the **SAS Enterprise Miner**. Dataset is imported with CSV format and saved as SAS file. The procedure of data imported inside SAS Enterprise Miner, save data, create library and create data source is listed in Appendix (1)(SAS Enterprise Miner)(i-vi). Figure 2 showed the column metadata as default and noticed that by default, SAS Enterprise Miner considered the level labels as "Interval" and "Nominal are incorrect. Therefore, adjustments have been made to identify the role, level and selection of dataset by using the Advanced view. Figure 3 showed the column metadata after adjustment while Figure 4 show data type summary after adjustment.

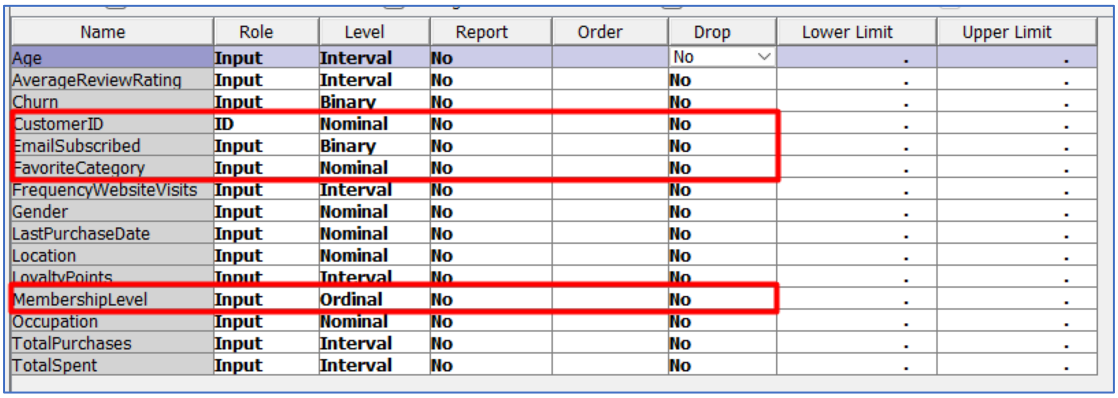
- CustomerID : Unique identifier with no meaningful order or distance, hence it is change as nominal with ID identifier.
- EmailSubscribed : This is a binary categorical variable without order, hence it is adjust as binary

- Churn : It's a binary variable indicating two distinct categories (churned or active) without any order, hence it is adjust as binary
- MembershipLevel : Membership levels have a clear order (e.g., Bronze < Silver < Gold < Platinum), hence it is adjust as ordinal.



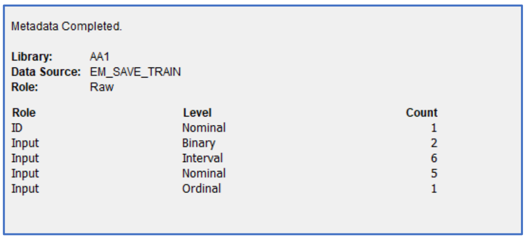
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
AverageReviev	Input	Interval	No		No	.	.
Churn	Input	Interval	No		No	.	.
CustomerID	Input	Interval	No		No	.	.
EmailSubscribe	Input	Interval	No		No	.	.
FavoriteCatego	Input	Nominal	No		No	.	.
FrequencyWeb	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchaseD	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
LoyaltyPoints	Input	Interval	No		No	.	.
MembershipLe	Input	Nominal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
TotalPurchase	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

Figure 2: Column Metadata in default



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
AverageReviewRating	Input	Interval	No		No	.	.
Churn	Input	Binary	No		No	.	.
CustomerID	ID	Nominal	No		No	.	.
EmailSubscribed	Input	Binary	No		No	.	.
FavoriteCategory	Input	Nominal	No		No	.	.
FrequencyWebsiteVisits	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchaseDate	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
LoyaltyPoints	Input	Interval	No		No	.	.
MembershipLevel	Input	Ordinal	No		No	.	.
Occupation	Input	Nominal	No		No	.	.
TotalPurchases	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

Figure 3: Column Metadata after adjustment



Metadata Completed.

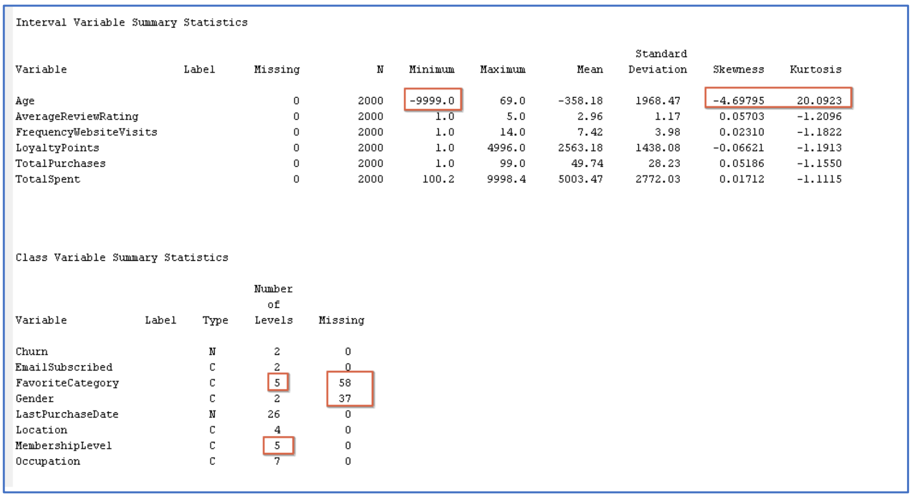
Library: AA1
Data Source: EM_SAVE_TRAIN
Role: Raw

Role	Level	Count
ID	Nominal	1
Input	Binary	2
Input	Interval	6
Input	Nominal	5
Input	Ordinal	1

Figure 4: Data Type Summary after adjustment

SEMMA: EXPLORE

In explore stage, Summary Statistics & Finding is used to explore the dataset for a deeper understanding for relationship between features. Figure 5 showed the summary statistics for the dataset. It is noticeable that for interval data, "Age" variables have minimum value -9999 which indicate that the invalid dataset exist and the outlier causing with the influence of negative skewness. For class variables, there are 58 and 37 missing values exist in "FavoriteCategory" and "Gender". "FavoriteCategory" supposedly contain with only 4 categories however there have 5 category level in the dataset. Also, “MembershipLevel” supposedly contain with only 4 categories however it also have 5 category level in the dataset.



```
Interval Variable Summary Statistics

                                                                                       Standard
Variable                  Label      Missing           N    Minimum    Maximum      Mean   Deviation    Skewness    Kurtosis

Age                                        0        2000    -9999.0       69.0   -358.18     1968.47    -4.69795     20.0923
AverageReviewRating                        0        2000        1.0        5.0      2.96        1.17     0.05703     -1.2096
FrequencyWebsiteVisits                     0        2000        1.0       14.0      7.42        3.98     0.02310     -1.1822
LoyaltyPoints                              0        2000        1.0     4996.0   2563.18     1438.08    -0.06621     -1.1913
TotalPurchases                             0        2000        1.0       99.0     49.74       28.23     0.05186     -1.1550
TotalSpent                                 0        2000      100.2     9998.4   5003.47     2772.03     0.01712     -1.1115



Class Variable Summary Statistics

                                       Number
                                         of
Variable             Label    Type     Levels    Missing

Churn                          N          2          0
EmailSubscribed                C          2          0
FavoriteCategory               C          5         58
Gender                         C          2         37
LastPurchaseDate               N         26          0
Location                       C          4          0
MembershipLevel                C          5          0
Occupation                     C          7          0
```

Figure 5: Summary statistics for the dataset

Figure 6 shows histogram for interval variable while figure 7 shows bar chart for class variable. According to the figure, the “Age” shows that “-9999” value as noisy data which required to remove. Also, “FavoriteCategory” shows that “Clothing” and “Clothings” should be existing as only one name, however it reflecting as two different category due to naming issue and should be adjust. “MembershipLevel” contain with record as “?” in the system hence reflecting with the error additional category level in summary statistics.

In conclusion, both “Age”, “FavoriteCategory”, “MembershipLevel”, “Gender” required to perform with the data cleaning action to remove the noisy data and adjust with data consistency.

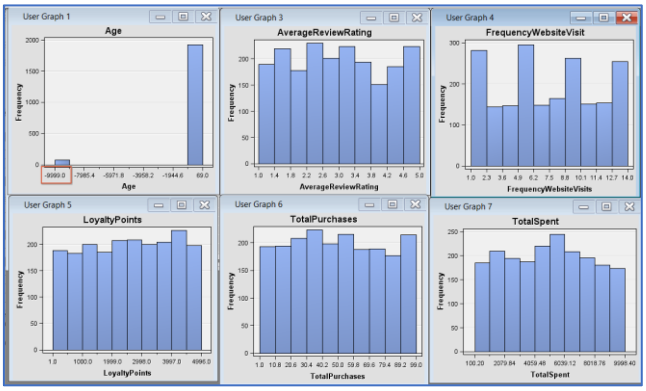


Figure 6: Histogram for Interval Variable

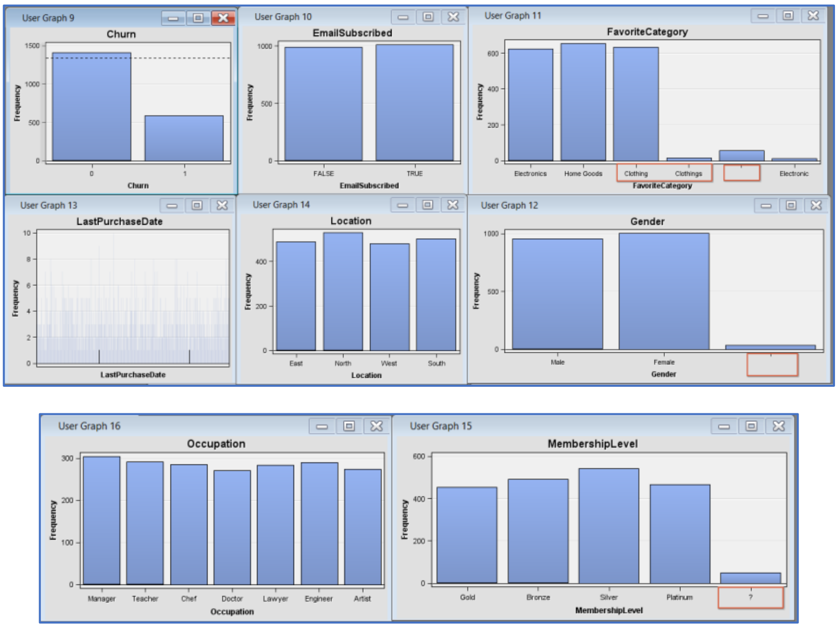
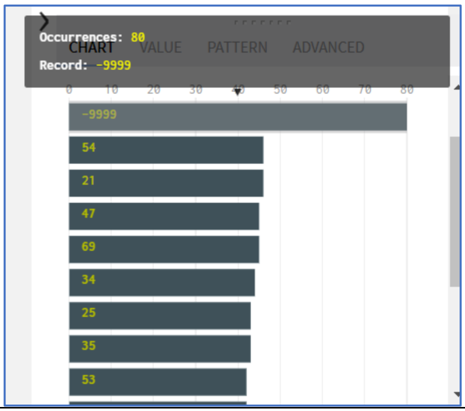
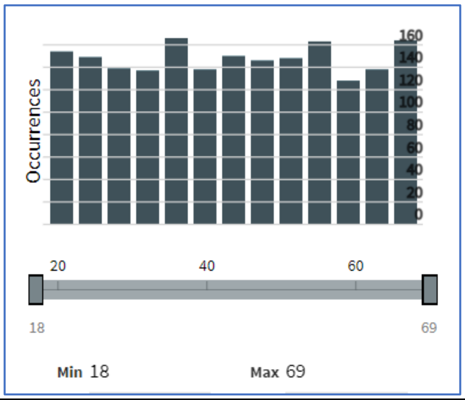


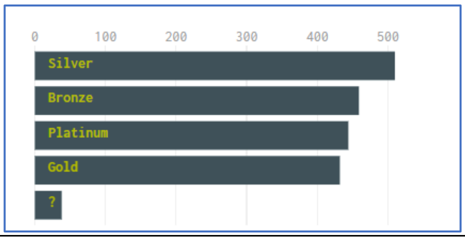
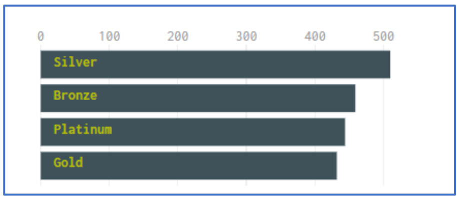
Figure 7: Bar Chart for Class Variable

SEMMA: Modify

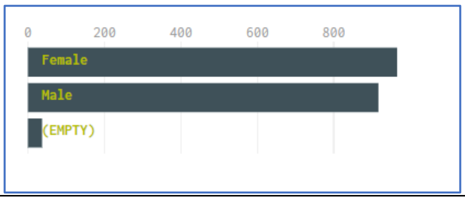
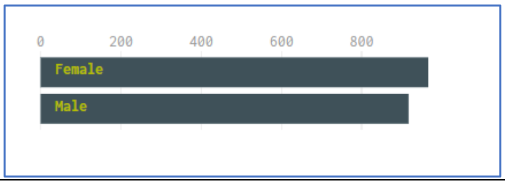
In Modify stage, data quality issues (inconsistent data, noisy data, missing value) which mentioned in Explore stage will by undergoes with task data import and preprocessing in **Talend Data Preprocessing tool.**

Data quality issue: Noisy Data

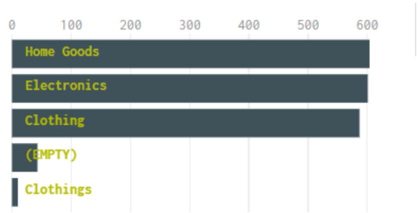
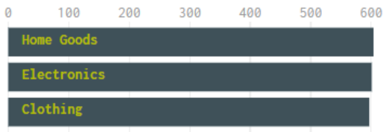
Graphs	**Description**
Before  Occurrences: 80 Record: -9999 CHART VALUE PATTERN ADVANCED 0 10 20 30 40 50 60 70 80 -9999 54 21 47 69 34 25 35 53	- "Age" shows that "-9999" value as noisy data - Since "-9999" didn't bring with useful information, the record with the noisy data is remove
After 	- After remove, the "Age" dataset looks normal

Graphs	**Description**
Before 	- "MembershipLevel" shows that "?" value as noisy data - Since "?" didn't bring with useful information, the record with the noisy data is remove
After 	- After remove, the "MembershipLevel" dataset looks normal

Data quality issue: Missing Value

Graphs	Description
Before 	- "Gender" shows that (empty) value as missing value - Since missing value didn't bring with useful information, the record with the noisy data is remove
After 	- After remove, the "Gender" dataset looks normal

Data quality issue: Inconsistence Data & Missing Value

Graphs	Description
Before 	- "FavoriteCategory" shows that (empty) value as missing value - "FavoriteCategory" shows that "Clothing" and "Clothings" should be existing as only one name, however it reflecting as two different category due to naming issue and should be adjust
After 	- "Clothings" is find and replace as "Clothing" and remove missing value - After remove, the "FavoriteCategory" dataset looks normal

SEMMA: Model

In Model stage, Decision Tree and Ensemble Methods is chosen to investigate on the customer purchasing behaviour by using **SAS Enterprise Miner** tools. Figure 8 showing with the overall flow done within the SAS Enterprise Miner.

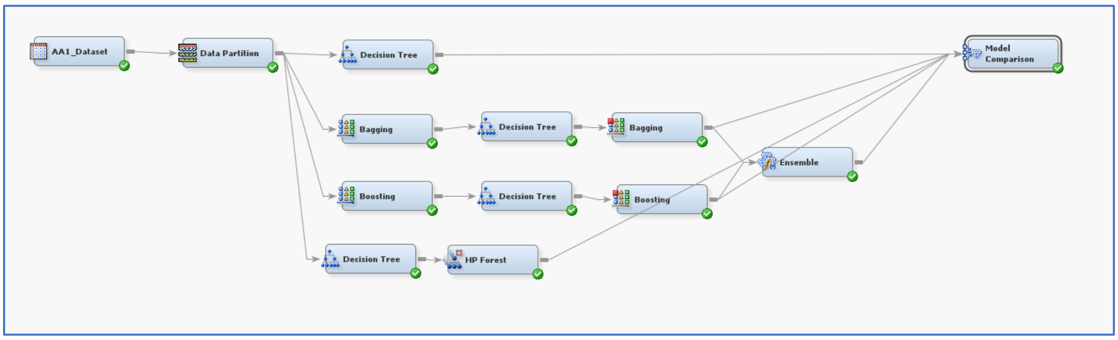


Figure 8: Overall flow done within the SAS Enterprise Miner.

Discussion on Decision Tree

The result of **decision tree** is showed in Figure 9. Based on the figure, it is observable that:

- Customers with total purchases less than or equal to 10,000 will likely churn.
- Customers with total purchases greater than 10,000 but loyalty points less than or equal to 2,363.5 and website visits less than or equal to 8.5 will likely churn.
- Customers with total purchases greater than 10,000 but loyalty points less than or equal to 2,363.5 and website visits greater than 8.5, if they are male, will likely churn.
- Customers with total purchases greater than 10,000 and loyalty points greater than 2,363.5 will likely not churn.
- Customers with total purchases less than or equal to 10,000, if they have an average review rating less than 3.0 and are not subscribed to emails, will likely churn.
- Customers with total purchases less than or equal to 10,000, if they have an average review rating less than 3.0, are subscribed to emails, and are aged 32,500 or less, will likely churn.
- Customers with total purchases less than or equal to 10,000, if they have an average review rating less than 3.0, are subscribed to emails, and are aged more than 32,500, will likely not churn.
- Customers with total purchases greater than 10,000 but loyalty points less than or equal to 2,363.5 and website visits greater than 8.5, if they are female, will likely not churn.

Rules obtained is listed as below:

- if TotalPurchases <= 10000 then CHURN=Y
- if TotalPurchases > 10000 AND LoyaltyPoints > 2363.5000 then CHURN=N
- if TotalPurchases > 10000 AND LoyaltyPoints <= 2363.5000 AND FrequencyWebsiteVisits > 8.5000 AND Gender = FEMALE then CHURN=N
- if TotalPurchases > 10000 AND LoyaltyPoints <= 2363.5000 AND FrequencyWebsiteVisits <= 8.5000 then CHURN=Y

- if TotalPurchases <= 10000 AND AverageReviewRating < 3.0000 AND EmailSubscribed = 0 then CHURN=Y
- if TotalPurchases <= 10000 AND AverageReviewRating < 3.0000 AND EmailSubscribed = 1 AND Age <= 32500 then CHURN=Y
- if TotalPurchases <= 10000 AND AverageReviewRating < 3.0000 AND EmailSubscribed = 1 AND Age > 32500 then CHURN=N
- if TotalPurchases > 10000 AND LoyaltyPoints <= 2363.5000 AND FrequencyWebsiteVisits > 8.5000 AND Gender = MALE then CHURN=Y

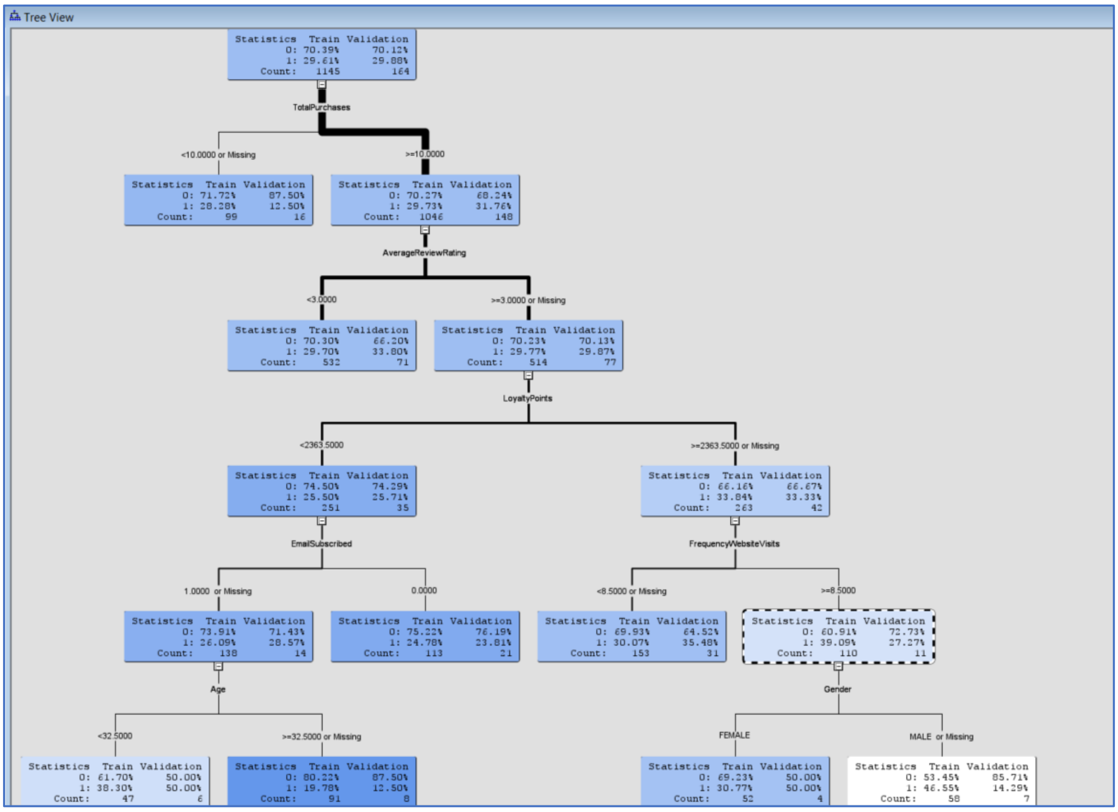
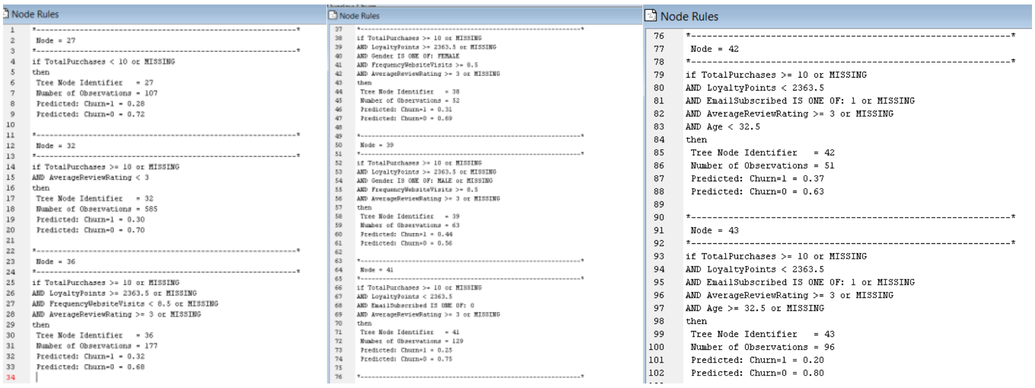


Figure 8: Decision Tree Result



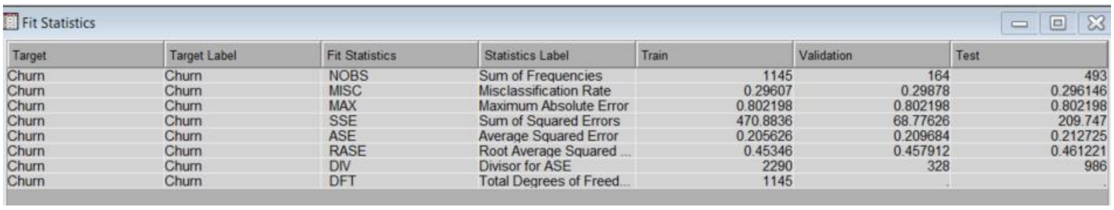
```
*--------------------------------------------------------------*
 Node = 27
*--------------------------------------------------------------*
if TotalPurchases < 10 or MISSING
then
 Tree Node Identifier   = 27
 Number of Observations = 107
 Predicted: Churn=1 = 0.28
 Predicted: Churn=0 = 0.72

*--------------------------------------------------------------*
 Node = 32
*--------------------------------------------------------------*
if TotalPurchases >= 10 or MISSING
AND AverageReviewRating < 3
then
 Tree Node Identifier   = 32
 Number of Observations = 585
 Predicted: Churn=1 = 0.30
 Predicted: Churn=0 = 0.70

*--------------------------------------------------------------*
 Node = 36
*--------------------------------------------------------------*
if TotalPurchases >= 10 or MISSING
AND LoyaltyPoints >= 2363.5 or MISSING
AND FrequencyWebsiteVisits < 8.5 or MISSING
AND AverageReviewRating >= 3 or MISSING
then
 Tree Node Identifier   = 36
 Number of Observations = 177
 Predicted: Churn=1 = 0.32
 Predicted: Churn=0 = 0.68
```

```
*--------------------------------------------------------------*
if TotalPurchases >= 10 or MISSING
AND LoyaltyPoints >= 2363.5 or MISSING
AND Gender IS ONE OF: FEMALE
AND FrequencyWebsiteVisits >= 8.5
AND AverageReviewRating >= 3 or MISSING
then
 Tree Node Identifier   = 38
 Number of Observations = 52
 Predicted: Churn=1 = 0.31
 Predicted: Churn=0 = 0.69

*--------------------------------------------------------------*
 Node = 39
*--------------------------------------------------------------*
if TotalPurchases >= 10 or MISSING
AND LoyaltyPoints >= 2363.5 or MISSING
AND Gender IS ONE OF: MALE or MISSING
AND FrequencyWebsiteVisits >= 8.5
AND AverageReviewRating >= 3 or MISSING
then
 Tree Node Identifier   = 39
 Number of Observations = 63
 Predicted: Churn=1 = 0.44
 Predicted: Churn=0 = 0.56

*--------------------------------------------------------------*
 Node = 41
*--------------------------------------------------------------*
if TotalPurchases >= 10 or MISSING
AND LoyaltyPoints < 2363.5
AND EmailSubscribed IS ONE OF: 0
AND AverageReviewRating >= 3 or MISSING
then
 Tree Node Identifier   = 41
 Number of Observations = 129
 Predicted: Churn=1 = 0.25
 Predicted: Churn=0 = 0.75

*--------------------------------------------------------------*
```

```
*--------------------------------------------------------------*
 Node = 42
*--------------------------------------------------------------*
if TotalPurchases >= 10 or MISSING
AND LoyaltyPoints < 2363.5
AND EmailSubscribed IS ONE OF: 1 or MISSING
AND AverageReviewRating >= 3 or MISSING
AND Age < 32.5
then
 Tree Node Identifier   = 42
 Number of Observations = 51
 Predicted: Churn=1 = 0.37
 Predicted: Churn=0 = 0.63

*--------------------------------------------------------------*
 Node = 43
*--------------------------------------------------------------*
if TotalPurchases >= 10 or MISSING
AND LoyaltyPoints < 2363.5
AND EmailSubscribed IS ONE OF: 1 or MISSING
AND AverageReviewRating >= 3 or MISSING
AND Age >= 32.5 or MISSING
then
 Tree Node Identifier   = 43
 Number of Observations = 96
 Predicted: Churn=1 = 0.20
 Predicted: Churn=0 = 0.80
```

Figure 9: Result Node Rules

Figure 10 shows with the fit statistics of decision tree. The model achieved a misclassification rate of 0.29607 during training, 0.29878 during validation, and 0.29614 in testing. These consistent rates across all datasets suggest the model generalizes well without significant overfitting.

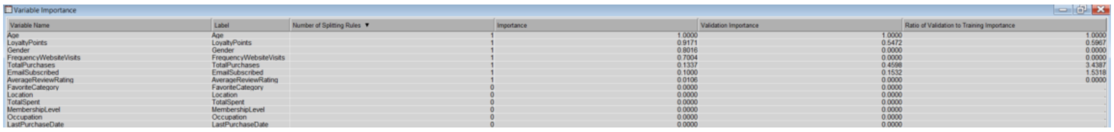


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	NOBS	Sum of Frequencies	1145	164	493
Churn	Churn	MISC	Misclassification Rate	0.29607	0.29878	0.296146
Churn	Churn	MAX	Maximum Absolute Error	0.802198	0.802198	0.802198
Churn	Churn	SSE	Sum of Squared Errors	470.8836	68.77626	209.747
Churn	Churn	ASE	Average Squared Error	0.205626	0.209684	0.212725
Churn	Churn	RASE	Root Average Squared ...	0.45346	0.457912	0.461221
Churn	Churn	DIV	Divisor for ASE	2290	328	986
Churn	Churn	DFT	Total Degrees of Freed...	1145	.	.

Figure 10: Fit statistics of decision tree

Figure 11 shows with the attribute of the decision tree accordingly with importance value. The results showed that “Age”, “LoyaltyPoint”, “Gender” and “FrequencyWebsiteVisit” variables are the most importance variables in customer churn prediction.



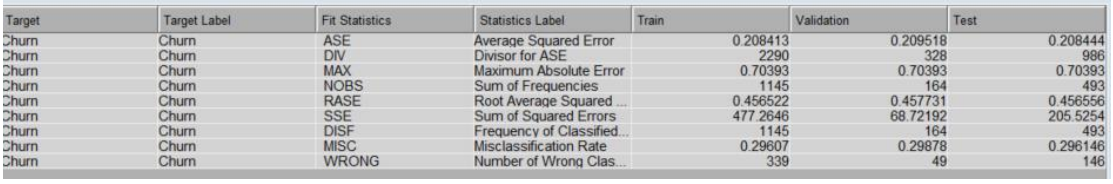
Variable Importance

Variable Name	Label	Number of Splitting Rules ▼	Importance	Validation Importance	Ratio of Validation to Training Importance
Age	Age	1	1.0000	1.0000	1.0000
LoyaltyPoints	LoyaltyPoints	1	0.9171	0.5472	0.5967
Gender	Gender	1	0.8016	0.0000	0.0000
FrequencyWebsiteVisits	FrequencyWebsiteVisits	1	0.7004	0.0000	0.0000
TotalPurchases	TotalPurchases	1	0.1337	0.4598	3.4387
EmailSubscribed	EmailSubscribed	1	0.1000	0.1532	1.5318
AverageReviewRating	AverageReviewRating	1	0.0106	0.0000	0.0000
FavoriteCategory	FavoriteCategory	0	0.0000	0.0000	.
Location	Location	0	0.0000	0.0000	.
TotalSpent	TotalSpent	0	0.0000	0.0000	.
MembershipLevel	MembershipLevel	0	0.0000	0.0000	.
Occupation	Occupation	0	0.0000	0.0000	.
LastPurchaseDate	LastPurchaseDate	0	0.0000	0.0000	.

Figure 11: Result Variable Importance of decision tree

Discussion on Ensemble Method

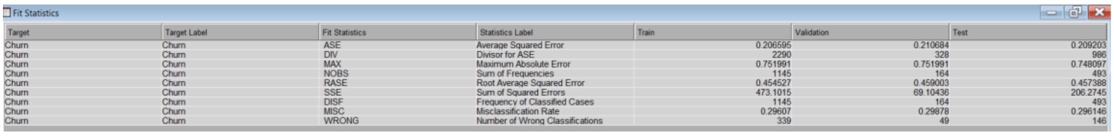
In Ensemble Method, Bagging and Boosting is applied through the decision tree and finally ensemble. Bagging is used when the model has high variance and when the dataset is large enough to ensure diverse subsets. In another side, boosting is used when the model has high bias and when we require higher predictive accuracy and are okay with slightly increased complexity. Figure 12 shows with the fit statistics of bagging and boosting decision tree. The model has a misclassification rate of 0.29607 during training, 0.29878 during validation, and 0.29614 in testing. This small variance suggests that the model's predictive accuracy is stable.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average Squared Error	0.208413	0.209518	0.208444
Churn	Churn	DIV	Divisor for ASE	2290	328	986
Churn	Churn	MAX	Maximum Absolute Error	0.70393	0.70393	0.70393
Churn	Churn	NOBS	Sum of Frequencies	1145	164	493
Churn	Churn	RASE	Root Average Squared ...	0.456522	0.457731	0.456556
Churn	Churn	SSE	Sum of Squared Errors	477.2646	68.72192	205.5254
Churn	Churn	DISF	Frequency of Classified...	1145	164	493
Churn	Churn	MISC	Misclassification Rate	0.29607	0.29878	0.296146
Churn	Churn	WRONG	Number of Wrong Clas...	339	49	146

Figure 12: fit statistics of bagging and boosting decision tree

Also, Figure 13 shows with the fit statistics of ensemble of decision tree and random Forest. The model’s misclassification rate is 0.29607 for the training set, 0.29878 for the validation set, and 0.296146 for the test set. These rates are very close, suggesting good model consistency across different datasets.

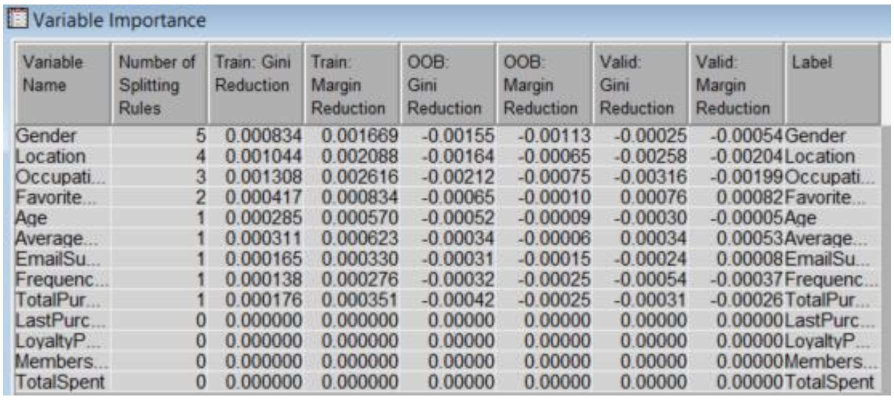


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn	Churn	ASE	Average Squared Error	0.206595	0.210684	0.209203
Churn	Churn	DIV	Divisor for ASE	2290	328	986
Churn	Churn	MAX	Maximum Absolute Error	0.751991	0.751991	0.748097
Churn	Churn	NOBS	Sum of Frequencies	1145	164	493
Churn	Churn	RASE	Root Average Squared Error	0.454527	0.459003	0.457388
Churn	Churn	SSE	Sum of Squared Errors	473.1015	69.10436	206.2745
Churn	Churn	DISF	Frequency of Classified Cases	1145	164	493
Churn	Churn	MISC	Misclassification Rate	0.29607	0.29878	0.296146
Churn	Churn	WRONG	Number of Wrong Classifications	339	49	146

Figure 13: fit statistics of ensemble of decision tree and random forest

Figure 14 shows with the attribute of the ensemble of decision tree with random forest accordingly with importance value. The results showed that “Gender”, “Location”, and “Occupation” variables are the most importance variables in customer churn prediction.



Variable Importance

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction	Label
Gender	5	0.000834	0.001669	-0.00155	-0.00113	-0.00025	-0.00054	Gender
Location	4	0.001044	0.002088	-0.00164	-0.00065	-0.00258	-0.00204	Location
Occupati...	3	0.001308	0.002616	-0.00212	-0.00075	-0.00316	-0.00199	Occupati...
Favorite...	2	0.000417	0.000834	-0.00065	-0.00010	0.00076	0.00082	Favorite...
Age	1	0.000285	0.000570	-0.00052	-0.00009	-0.00030	-0.00005	Age
Average...	1	0.000311	0.000623	-0.00034	-0.00006	0.00034	0.00053	Average...
EmailSu...	1	0.000165	0.000330	-0.00031	-0.00015	-0.00024	0.00008	EmailSu...
Frequenc...	1	0.000138	0.000276	-0.00032	-0.00025	-0.00054	-0.00037	Frequenc...
TotalPur...	1	0.000176	0.000351	-0.00042	-0.00025	-0.00031	-0.00026	TotalPur...
LastPurc...	0	0.000000	0.000000	0.00000	0.00000	0.00000	0.00000	LastPurc...
LoyaltyP...	0	0.000000	0.000000	0.00000	0.00000	0.00000	0.00000	LoyaltyP...
Members...	0	0.000000	0.000000	0.00000	0.00000	0.00000	0.00000	Members...
TotalSpent	0	0.000000	0.000000	0.00000	0.00000	0.00000	0.00000	TotalSpent

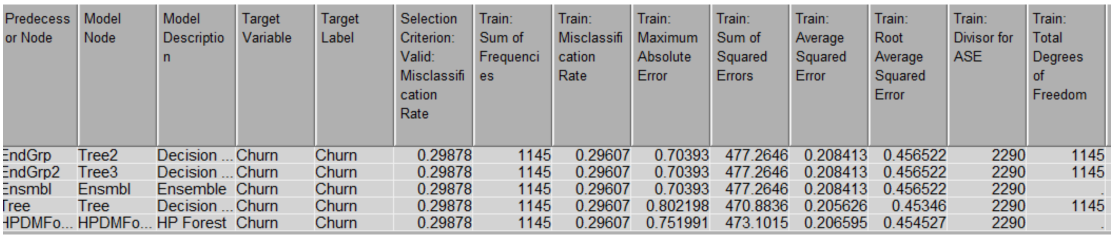
Figure 14: Result Variable Importance of ensemble of decision tree and random forest

SEMMA: Access

In Access stage, Fit Statistics is used to assess the utility and reliability of the predictive models for the customer churns by using **SAS Enterprise Miner** tools.

The Fit Statistics and Confusion Metric of overall model is showed in Figure 15. Based on the figure, it is observable that:

- All models have a churn rate criterion of 0.29878, with a sum of misclassification frequencies of 1145.
- The Train Misclassification Rate is consistent across all models at 0.29607.
- The Maximum Absolute Error varies across models, with the Tree models (including Gradient Boost and Decision Tree) showing 0.70393, and the HP Forest model (ensemble of decision tree and random forest) showing a slightly higher error at 0.751991.
- The Sum of Squared Errors for the Tree2 and Tree3 (bagging and boosting decision tree) models is 477.2646, while the Gradient Boost model shows a slightly lower error at 476.854. The Decision Tree model has a Sum of Squared Errors of 470.8386, and the HP Forest has 473.1015.
- The Average Squared Error is the same for the Tree2 and Tree3 models at 0.204813, a bit higher for the Gradient Boost model at 0.208096, and slightly lower for the Decision Tree model at 0.205626 and the HP Forest model at 0.205695.



Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom
EndGrp	Tree2	Decision ...	Churn	Churn	0.29878	1145	0.29607	0.70393	477.2646	0.208413	0.456522	2290	1145
EndGrp2	Tree3	Decision ...	Churn	Churn	0.29878	1145	0.29607	0.70393	477.2646	0.208413	0.456522	2290	1145
Ensmbl	Ensmbl	Ensemble	Churn	Churn	0.29878	1145	0.29607	0.70393	477.2646	0.208413	0.456522	2290	.
Tree	Tree	Decision ...	Churn	Churn	0.29878	1145	0.29607	0.802198	470.8836	0.205626	0.45346	2290	1145
HPDMFo...	HPDMFo...	HP Forest	Churn	Churn	0.29878	1145	0.29607	0.751991	473.1015	0.206595	0.454527	2290	.

Figure 15: Fit Statistic for three model

Figure 16 showing the ROC graph for the model. For training set, all Decision Tree, HP Forest (ensemble of decision tree and random forest), and Ensemble models (bagging and boosting decision tree) all perform significantly better than the Baseline, indicating that all models have predictive capability. The Ensemble model showing the highest area under the curve (AUC), followed closely by the HP Forest and then the Decision Tree. A higher AUC represents a better performance in distinguishing the positive and negative classes. In Validation and testing graph, Ensemble model still showing the best performance comparing with the rest model. By comparing across all three modelling, we can conclude that Ensemble model consistently shows a better performance, which giving the suggestion to us that combining multiple models' predictions is effective for this dataset. Despite from the good performance of the Ensemble (bagging and boosting decision tree)and HP Forest (ensemble of decision tree and random forest) models, further parameter tuning and validation are recommended to ensure that the models are not overfitting, particularly given that the ROC curves are based on the same data used to train the models.

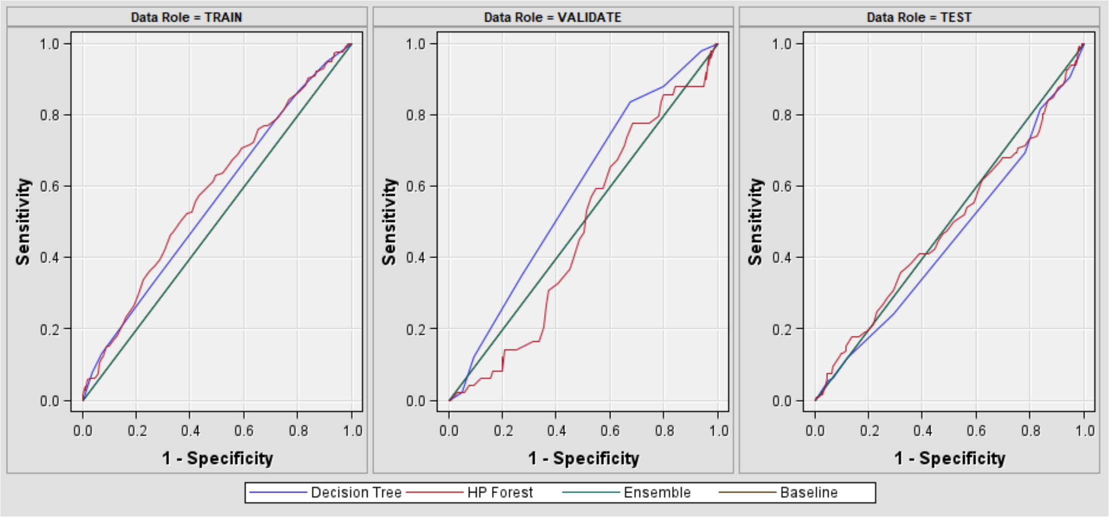


Figure 16: ROC graph for the model

Conclusion

In conclusion, with the objective of exploration of customer behavior within the e-commerce sector, we analysed a detailed two-year transaction dataset using Talend Data Integration, Talend Data Preparation, and SAS Enterprise Miner. Our objective was to discover with the patterns of customer behaviour and find out the crucial factors influencing their purchasing decisions. The predictive modelling techniques we employed included decision trees, decision tree ensembles through bagging and boosting, and a hybrid model that combined decision trees with a random forest approach. Also, the variable such as “Age”, “LoyaltyPoint”, “Gender”, “FrequencyWebsiteVisit”, “Location”, and “Occupation” could be the recommend and most important criteria that seller could use to analyst the customer behavior.

Among these, the Ensemble model was found by demonstrating with the best predictive performance among with others model. However, the Ensemble model have the potential risk of overfitting, where the model are highly fitting towards to the training dataset. Overfitting risk might causing with the risk that the model are not applicable to those unseen data. To avoid this issue, it is recommend that to employing cross-validation for more reliable performance metrics, applying regularization to prevent excessive model.

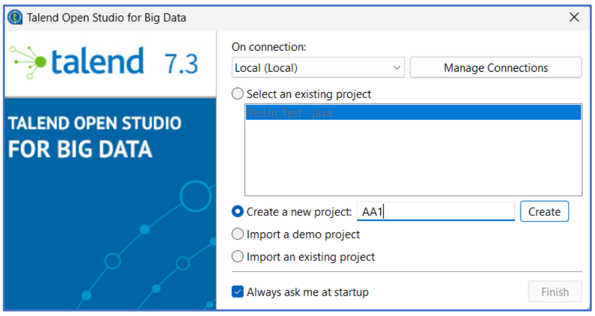
Appendix

Below listed with the procedure for conducting the step of **S**EMMA in Talend Integration, SAS Enterprise Miner, S**E**MMA in SAS Enterprise Miner, SE**M**MA in Talend Data Preparation, SEM**MA** in SAS Enterprise Miner.

1. **S**EMMA – **S**AMPLE – Talend Integration & SAS Enterprise Miner

Talend Integration

i. In talend, create new project with naming "AA1"


Talend Open Studio for Big Data
talend 7.3
TALEND OPEN STUDIO FOR BIG DATA

On connection: Local (Local)

Manage Connections

○ Select an existing project

PeiLin_Test - java

● Create a new project: AA1 Create

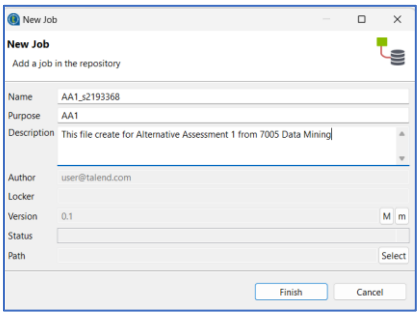
○ Import a demo project

○ Import an existing project

☑ Always ask me at startup

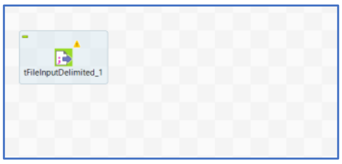
Finish

ii. Create a new project naming as "AA1"

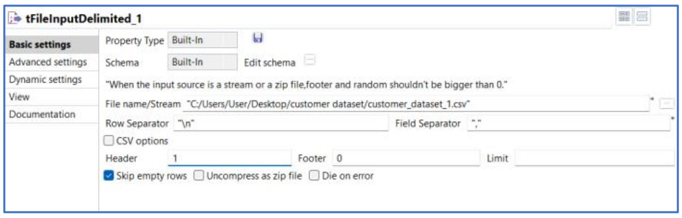


Name	AA1_s2193368
Purpose	AA1
Description	This file create for Alternative Assessment 1 from 7005 Data Mining
Author	user@talend.com
Locker	
Version	0.1
Status	
Path	

iii. Drag the tFileInputDelimited component to the design workspace



iv. In the Component panel below, set the properties



Property Type Built-In

Schema Built-In

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

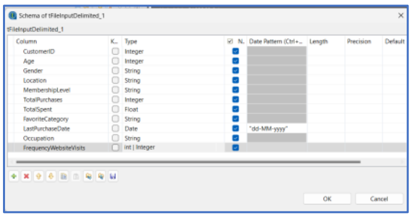
File name/Stream "C:/Users/User/Desktop/customer dataset/customer_dataset_1.csv"

Row Separator "\n" Field Separator ","

□ CSV options

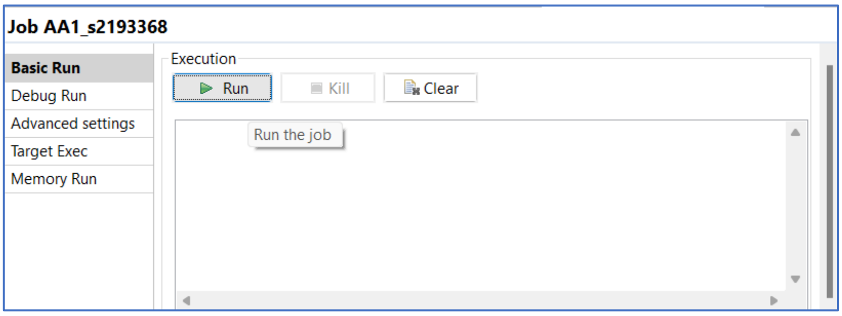
Header 1 Footer 0 Limit

☑ Skip empty rows □ Uncompress as zip file □ Die on error

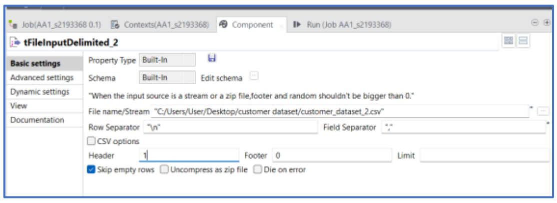


Column	K...	Type	N.	Date Pattern (Ctrl+...	Length	Precision	Default
CustomerID	□	Integer	☑				
Age	□	Integer	☑				
Gender	□	String	☑				
Location	□	String	☑				
MembershipLevel	□	String	☑				
TotalPurchases	□	Integer	☑				
TotalSpent	□	Float	☑				
FavoriteCategory	□	String	☑				
LastPurchaseDate	□	Date	☑	"dd-MM-yyyy"			
Occupation	□	String	☑				
FrequencyWebsiteVisits	□	int | Integer	☑				

v. Press "Run"



vi. Input second data file



tFileInputDelimited_2

Property Type Built-In

Schema Built-In Edit schema

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

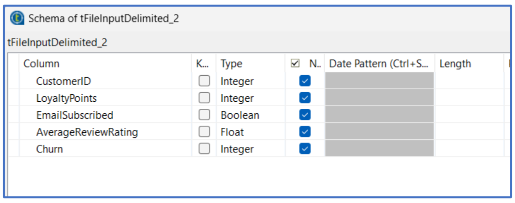
File name/Stream "C:/Users/User/Desktop/customer dataset/customer_dataset_2.csv"

Row Separator "\n" Field Separator ","

☐ CSV options

Header 1 Footer 0 Limit

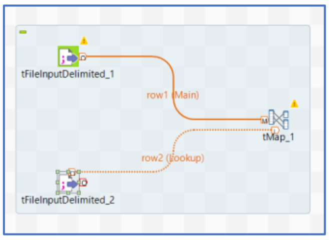
☑ Skip empty rows ☐ Uncompress as zip file ☐ Die on error


Schema of tFileInputDelimited_2

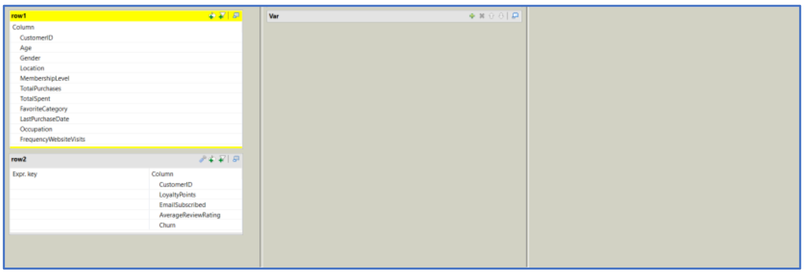
tFileInputDelimited_2

Column	K...	Type	☑ N..	Date Pattern (Ctrl+S...	Length
CustomerID	☐	Integer	☑		
LoyaltyPoints	☐	Integer	☑		
EmailSubscribed	☐	Boolean	☑		
AverageReviewRating	☐	Float	☑		
Churn	☐	Integer	☑		

vii. Drag and drop a tMap component onto the workspace & link both tFileInputDelimited components to the tMap component



viii. Joining Data

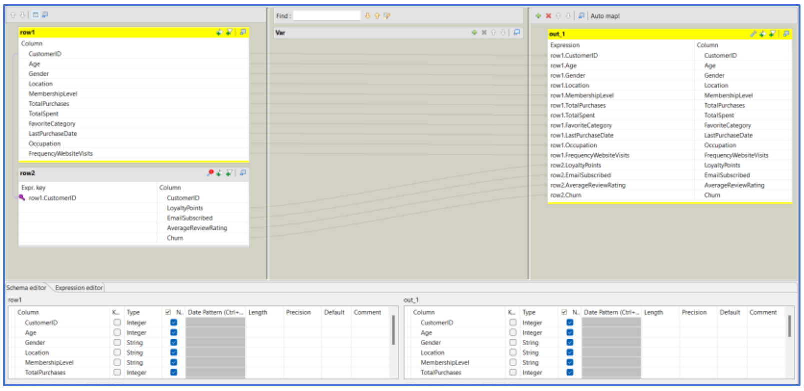


row1

Column
CustomerID
Age
Gender
Location
MembershipLevel
TotalPurchases
TotalSpent
FavoriteCategory
LastPurchaseDate
Occupation
FrequencyWebsiteVisits

row2

Expr. key	Column
	CustomerID
	LoyaltyPoints
	EmailSubscribed
	AverageReviewRating
	Churn



row1

Column
CustomerID
Age
Gender
Location
MembershipLevel
TotalPurchases
TotalSpent
FavoriteCategory
LastPurchaseDate
Occupation
FrequencyWebsiteVisits

row2

Expr. key	Column
row1.CustomerID	CustomerID
	LoyaltyPoints
	EmailSubscribed
	AverageReviewRating
	Churn

out_1

Expression	Column
row1.CustomerID	CustomerID
row1.Age	Age
row1.Gender	Gender
row1.Location	Location
row1.MembershipLevel	MembershipLevel
row1.TotalPurchases	TotalPurchases
row1.TotalSpent	TotalSpent
row1.FavoriteCategory	FavoriteCategory
row1.LastPurchaseDate	LastPurchaseDate
row1.Occupation	Occupation
row1.FrequencyWebsiteVisits	FrequencyWebsiteVisits
row2.LoyaltyPoints	LoyaltyPoints
row2.EmailSubscribed	EmailSubscribed
row2.AverageReviewRating	AverageReviewRating
row2.Churn	Churn

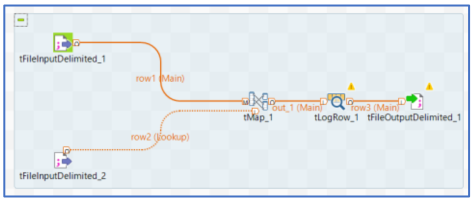
row1

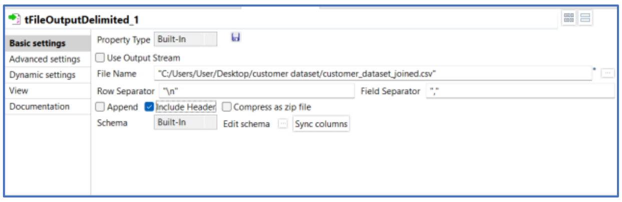
Column	K...	Type	☑ N..	Date Pattern (Ctrl+...	Length	Precision	Default	Comment
CustomerID	☐	Integer	☑					
Age	☐	Integer	☑					
Gender	☐	String	☑					
Location	☐	String	☑					
MembershipLevel	☐	String	☑					
TotalPurchases	☐	Integer	☑					

out_1

Column	K...	Type	☑ N..	Date Pattern (Ctrl+...	Length	Precision	Default	Comment
CustomerID	☐	Integer	☑					
Age	☐	Integer	☑					
Gender	☐	String	☑					
Location	☐	String	☑					
MembershipLevel	☐	String	☑					
TotalPurchases	☐	Integer	☑					

ix. Export out the dataset as CSV




tFileOutputDelimited_1

Property Type Built-In

☐ Use Output Stream

File Name "C:/Users/User/Desktop/customer dataset/customer_dataset_joined.csv"

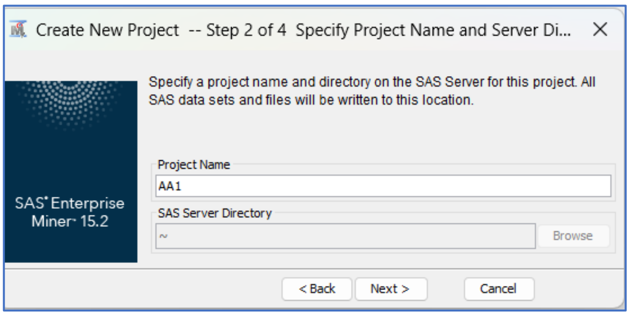
Row Separator "\n" Field Separator ","

☐ Append ☑ Include Header ☐ Compress as zip file

Schema Built-In Edit schema Sync columns

SAS Enterprise Miner

i. Create new project with naming "AA1"


Create New Project -- Step 2 of 4 Specify Project Name and Server Di...

Specify a project name and directory on the SAS Server for this project. All SAS data sets and files will be written to this location.

Project Name: AA1

SAS Server Directory: ~

ii. Create new diagram

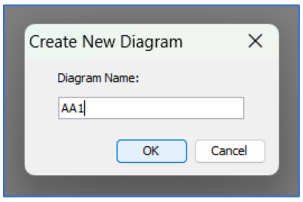
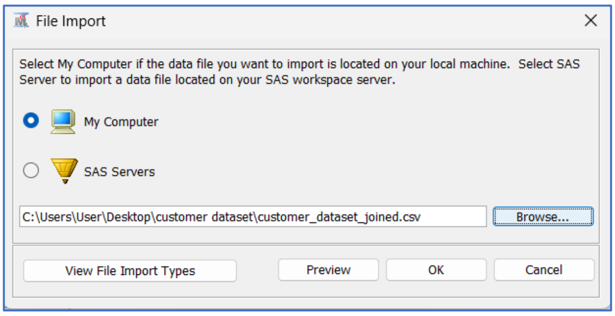

Create New Diagram

Diagram Name: AA1

iii. Import data from local machine and run


File Import

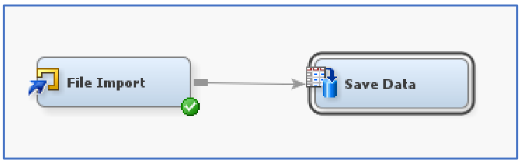
Select My Computer if the data file you want to import is located on your local machine. Select SAS Server to import a data file located on your SAS workspace server.

● My Computer

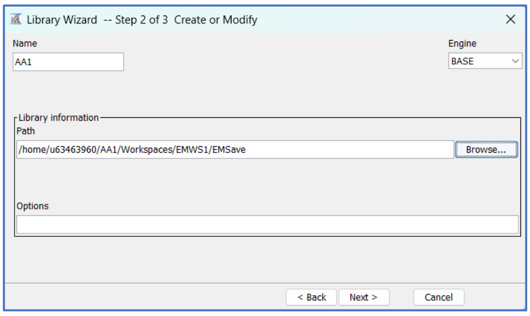
○ SAS Servers

C:\Users\User\Desktop\customer dataset\customer_dataset_joined.csv

iv. Save Data



v. Create library for imported SAS file and named as "AA1"



Library Wizard -- Step 2 of 3 Create or Modify

Name: AA1

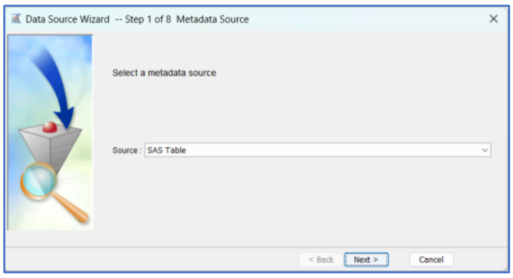
Engine: BASE

Library information

Path: /home/u63463960/AA1/Workspaces/EMWS1/EMSave

Options:

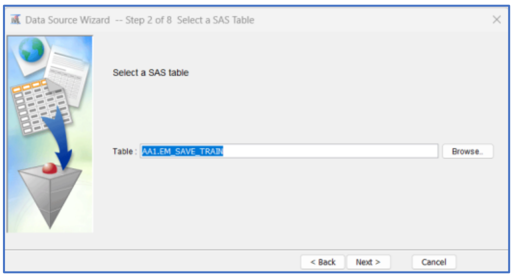
vi. Create data source and extract the imported SAS file



Data Source Wizard -- Step 1 of 8 Metadata Source

Select a metadata source

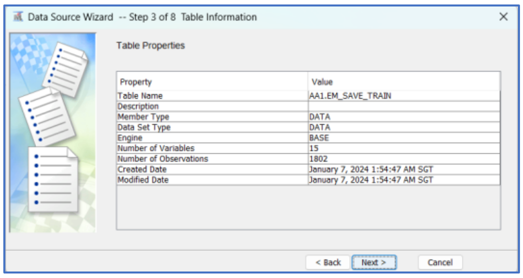
Source : SAS Table



Data Source Wizard -- Step 2 of 8 Select a SAS Table

Select a SAS table

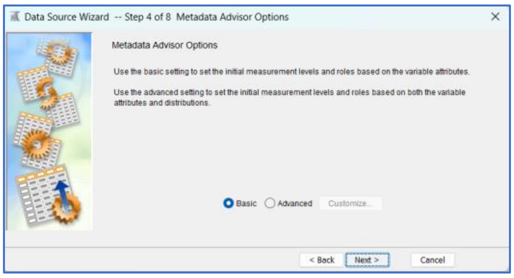
Table : AA1.EM_SAVE_TRAIN



Data Source Wizard -- Step 3 of 8 Table Information

Table Properties

Property	Value
Table Name	AA1.EM_SAVE_TRAIN
Description	
Member Type	DATA
Data Set Type	DATA
Engine	BASE
Number of Variables	15
Number of Observations	1802
Created Date	January 7, 2024 1:54:47 AM SGT
Modified Date	January 7, 2024 1:54:47 AM SGT



Data Source Wizard -- Step 4 of 8 Metadata Advisor Options

Metadata Advisor Options

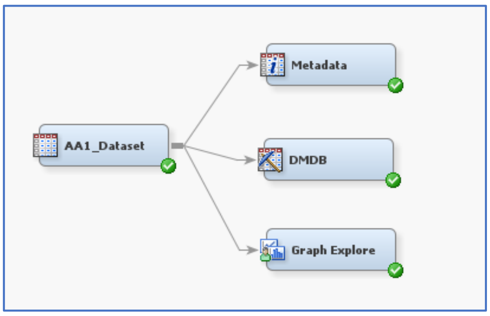
Use the basic setting to set the initial measurement levels and roles based on the variable attributes.

Use the advanced setting to set the initial measurement levels and roles based on both the variable attributes and distributions.

● Basic ○ Advanced

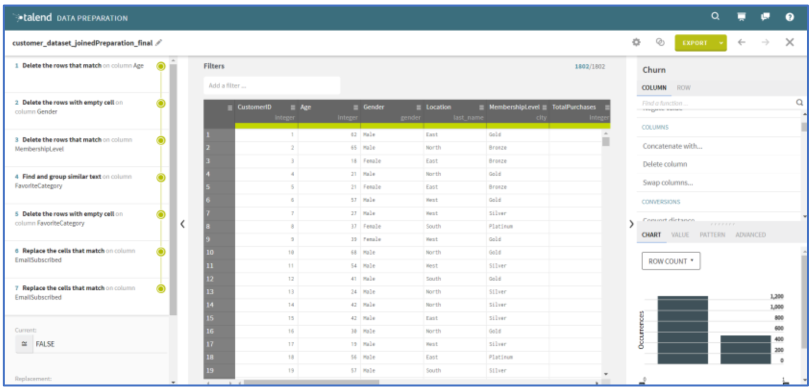
2. S**E**MMA – **E**XPLORE – SAS Enterprise Miner

i. Graph exploration by using Metadata, DMDB and graph explore node



3. SE**M**MA – **M**ODIFY – Talend Data Preparation

i. Data cleaning and pre-processing using Talend Data Preparation



customer_dataset_joinedPreparation_final

1 **Delete the rows that match** on column Age

2 **Delete the rows with empty cell** on column Gender

3 **Delete the rows that match** on column MembershipLevel

4 **Find and group similar text** on column FavoriteCategory

5 **Delete the rows with empty cell** on column FavoriteCategory

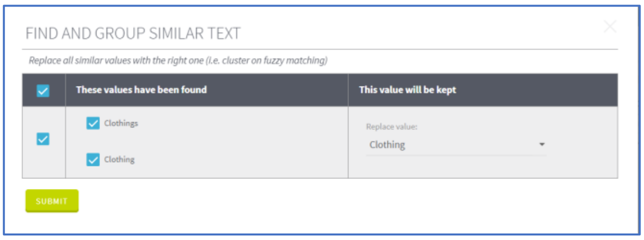
6 **Replace the cells that match** on column EmailSubscribed

7 **Replace the cells that match** on column EmailSubscribed

Current: FALSE

	CustomerID	Age	Gender	Location	MembershipLevel	TotalPurchases
	integer	integer	gender	last_name	city	integer
1	1	62	Male	East	Gold	
2	2	65	Male	North	Bronze	
3	3	18	Female	East	Bronze	
4	4	21	Male	North	Gold	
5	5	21	Female	East	Bronze	
6	6	57	Male	West	Gold	
7	7	27	Male	West	Silver	
8	8	37	Female	South	Platinum	
9	9	39	Female	West	Gold	
10	10	68	Male	North	Gold	
11	11	54	Male	West	Silver	
12	12	41	Male	South	Gold	
13	13	24	Male	North	Silver	
14	14	42	Male	North	Silver	
15	15	42	Male	East	Silver	
16	16	38	Male	North	Gold	
17	17	19	Male	West	Silver	
18	18	56	Male	East	Platinum	
19	19	57	Male	South	Silver	

ii. Find and group with similar text



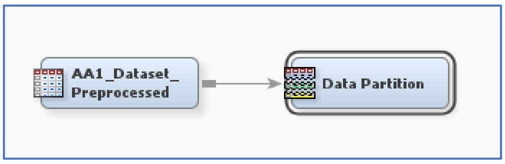
FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

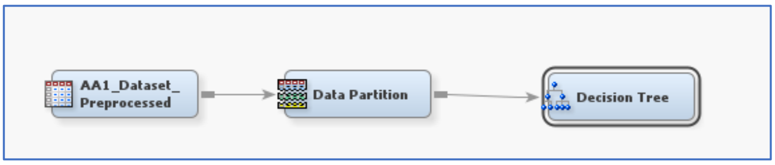
☑	These values have been found	This value will be kept
☑	☑ Clothings ☑ Clothing	Replace value: Clothing

4. SEMM**A** – **M**odelling – Decision Tree analysis, Random Forest – SAS Enterprise Miner

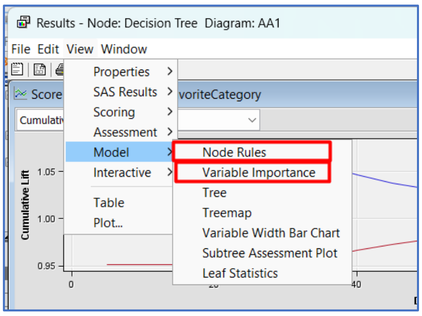
i. Split data by using data partition node



ii. Applying decision tree node

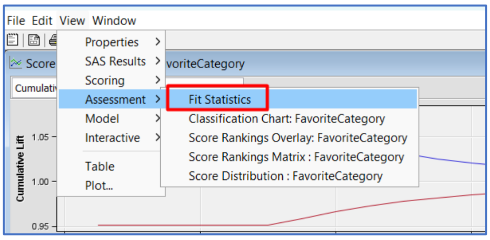


iii. Get modelling results by using View>Model>Node Rules; View>Model>Variable Importance



5. SEMM**A** – **A**ssess – SAS Enterprise Miner

i. Get modelling performance by using View>Assessment>Fit Statisics


File Edit View Window
Score
voriteCategory
Cumulati
Properties
SAS Results
Scoring
Assessment
Model
Interactive
Table
Plot...
Fit Statistics
Classification Chart: FavoriteCategory
Score Rankings Overlay: FavoriteCategory
Score Rankings Matrix : FavoriteCategory
Score Distribution : FavoriteCategory
Cumulative Lift
1.05
1.00
0.95