UNIVERSITI MALAYA

ALTERNATIVE ASSESSMENT 1

COURSE CODE	**WQD 7005**
COURSE	**DATA MINING**
FACULTY	**FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY**
NAME OF MEMBERS	**SOONG SING YING (S2191652)**
SECTION	**1**
SEMESTER	**SEMESTER 1 (2023/2024)**
LECTURER NAME	**PROF. DR. TEH YING WAH**
DATE OF SUBMIT	**7th JANUARY 2024**

Data Used:

The synthetic data is generated by using Python Faker. The link of python code to generate synthetic data:

https://colab.research.google.com/drive/1wCLw8JSZ24PO901fIFes4tQjewSGj0jx?usp=sharing

The dataset contains 500 rows and 12 columns. The details of the column as below:

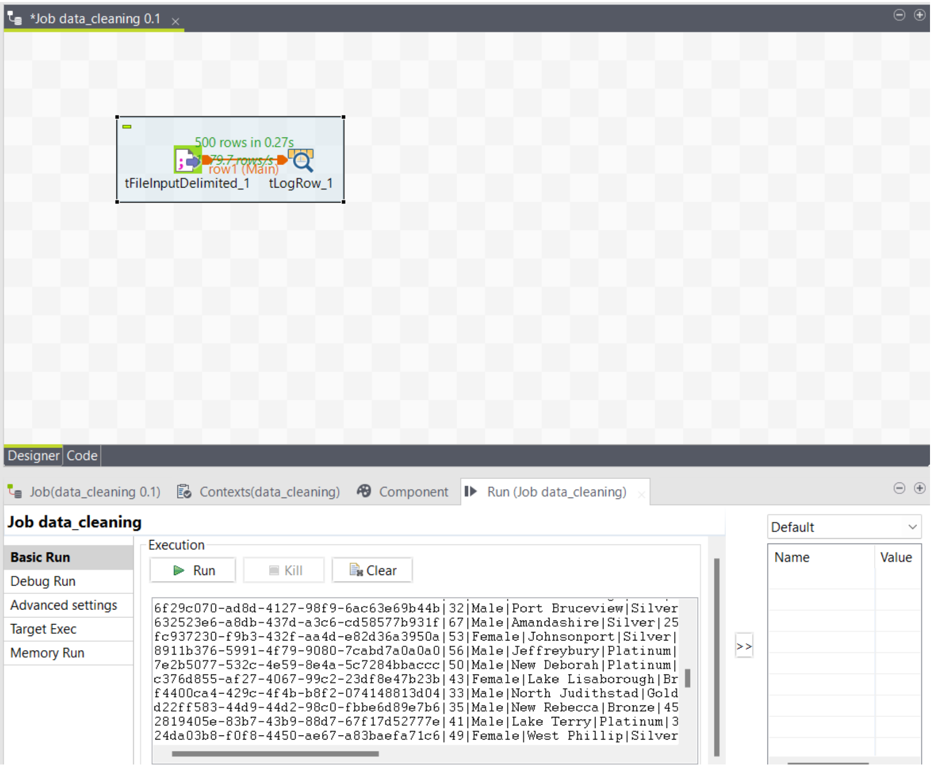
Variable	Description
CustomerID	Unique identifier for each customer.
Age	Age of the customer.
Gender	Gender of the customer.
Location	Geographic location of the customer.
MembershipLevel	Indicates the membership level (e.g., Bronze, Silver, Gold, Platinum).
TotalPurchases	Total number of purchases made by the customer.
TotalSpent	Total amount spent by the customer.
FavoriteCategory	The category in which the customer most frequently shops (Electronics, Clothing, Home Goods).
LastPurchaseDate	The date of the last purchase.
Occupation	Occupation of the customer.
FrequencyOfVisits	Frequency of the customer visit the website per month.
Churn	Indicates whether the customer has stopped purchasing (1 for churned, 0 for active)

Advantages: Can get random realistic data based on the variables required.

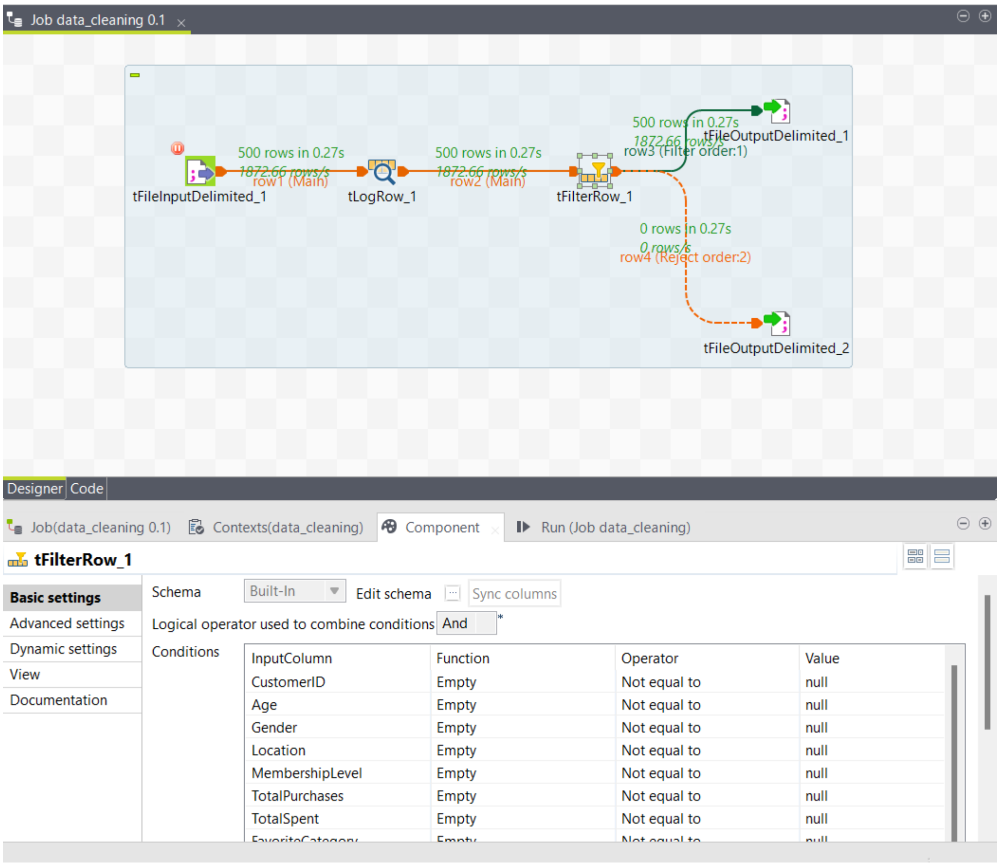
Challenges: Since the range of data is determined in the code, so the data is clean and cannot figure out the source of data if error occurred.

Data Import and Preprocessing:

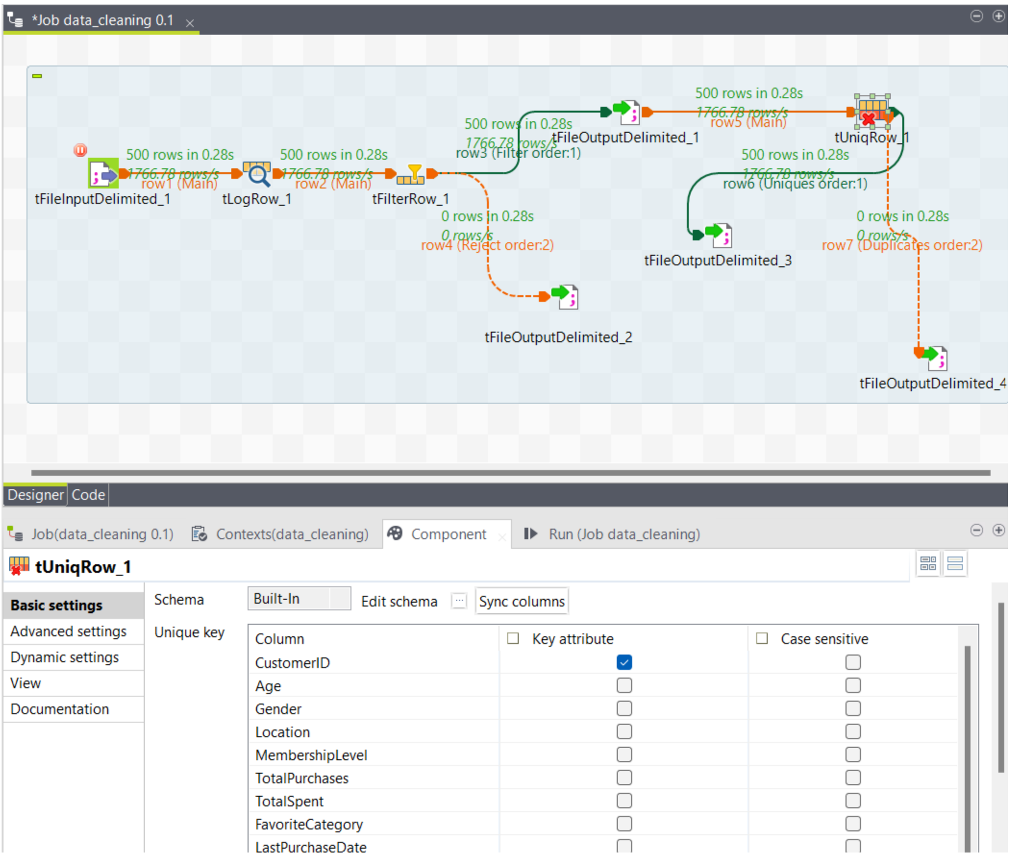
1. Extract dataset into Talend Data Integration



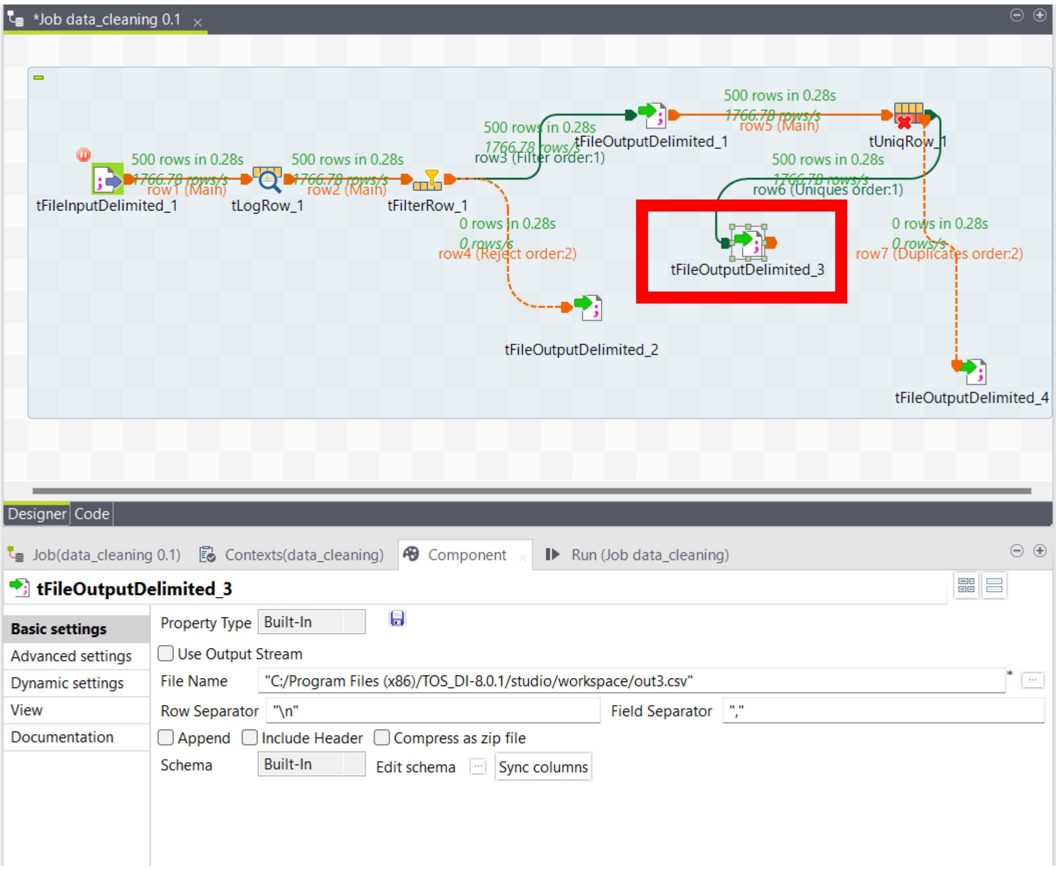
2. Carry out data cleaning in Talend Data Integration.
   a. Check missing values – No null data is found.
   Advantages: Can output the dataset into remained dataset and rejected dataset.
   Challenges: Need to take attention on whether empty and zero in the dataset considered as null data or not.



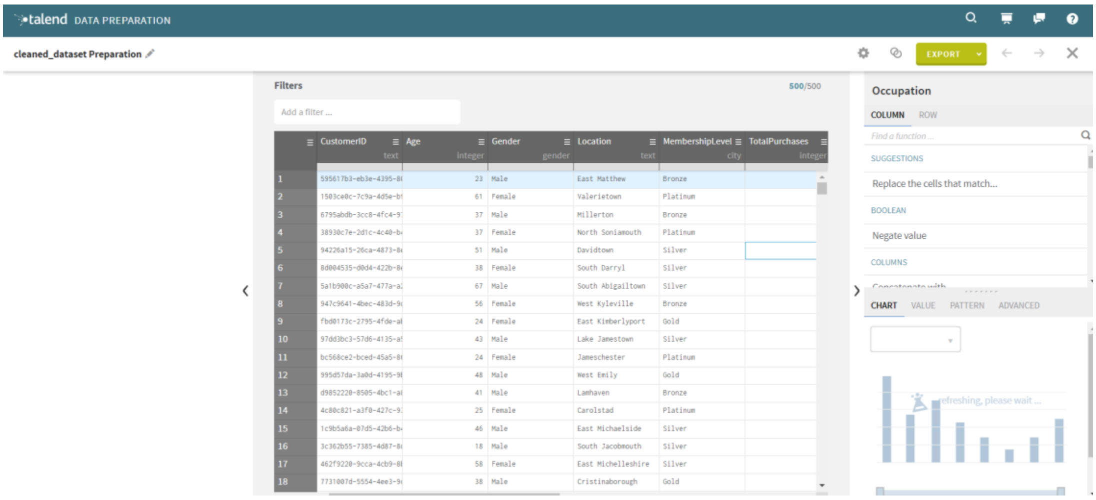
b. Check duplicate data – No duplicate data is found.
   Advantages: Can output the dataset into remained dataset and rejected dataset.
   Challenges: Need to take attention which column should be checked for the unique data.



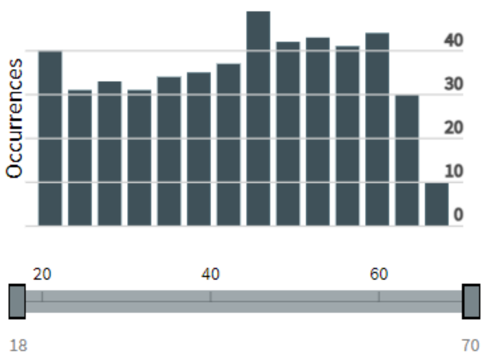
3. Export cleaned data as CSV.



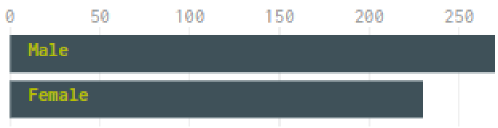
4. Extract cleaned data into Talend Data Preparation.



5. Carry out data processing in Talend Data Preparation.
   a. Data profiling – General statistics summary and pattern of each variable are shown in the table below.
      Advantages: Can output the summary statistics without any coding.
      Challenges: Some statistics is not useful such as duplicate count for group data.

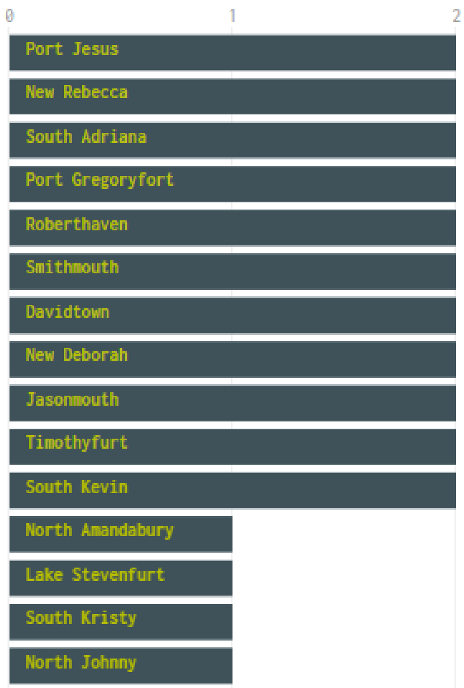
Variable	Value
CustomerID	Count: **500** Avg length: **36** Distinct: **500** Duplicate: **0** Min length: **36** Valid: **500** Empty: **0** Max length: **36** Invalid: **0**
Age 	Count: **500** Min: **18** Max: **70** Distinct: **53** Mean: **44.94** Duplicate: **447** Variance: **225.67** Valid: **500** Median: **46** Empty: **0** Lower quantile: **32** Invalid: **0** Upper quantile: **57.75**

Gender



Count: **500**
Avg length: **4.92**
Distinct: **2**
Duplicate: **498**
Min length: **4**
Valid: **500**
Empty: **0**
Max length: **6**
Invalid: **0**

Location

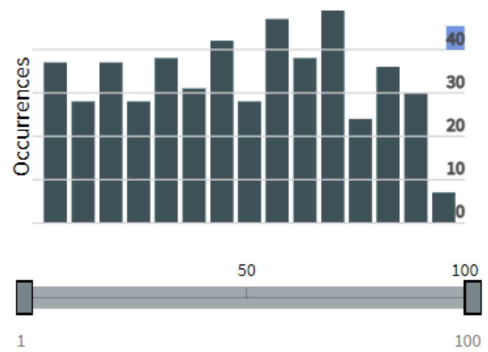


Count: **500**
Avg length: **12.21**
Distinct: **489**
Duplicate: **11**
Min length: **7**
Valid: **500**
Empty: **0**
Max length: **20**
Invalid: **0**

MembershipLevel

Count: **500**
Avg length: **6**
Distinct: **4**
Duplicate: **496**
Min length: **4**
Valid: **500**
Empty: **0**
Max length: **8**
Invalid: **0**

TotalPurchases



Count: **500**

Distinct: **99**

Duplicate: **401**

Valid: **500**

Empty: **0**

Invalid: **0**

Min: **1**

Max: **100**

Mean: **50.58**

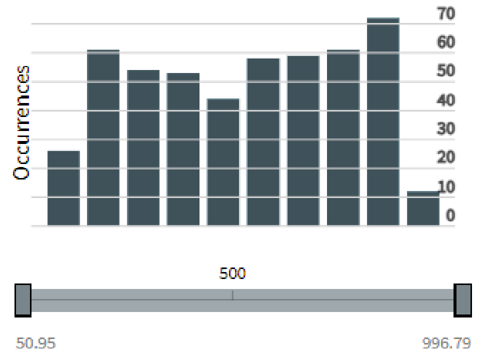
Variance: **786.63**

Median: **52**

Lower quantile: **26.25**

Upper quantile: **73**

TotalSpent



Count: **500**

Distinct: **500**

Duplicate: **0**

Valid: **500**

Empty: **0**

Invalid: **0**

Min: **50.95**

Max: **996.79**

Mean: **546.05**

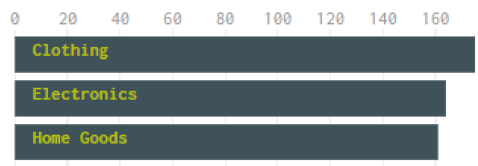
Variance: **79305.64**

Median: **567.57**

Lower quantile: **288.34**

Upper quantile: **804.79**

FavoriteCategory



Count: **500**

Distinct: **3**

Duplicate: **497**

Valid: **500**

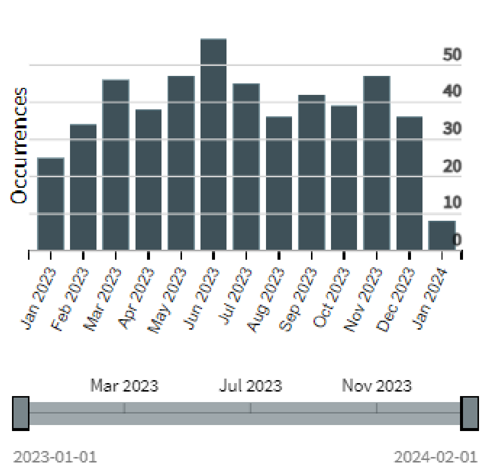
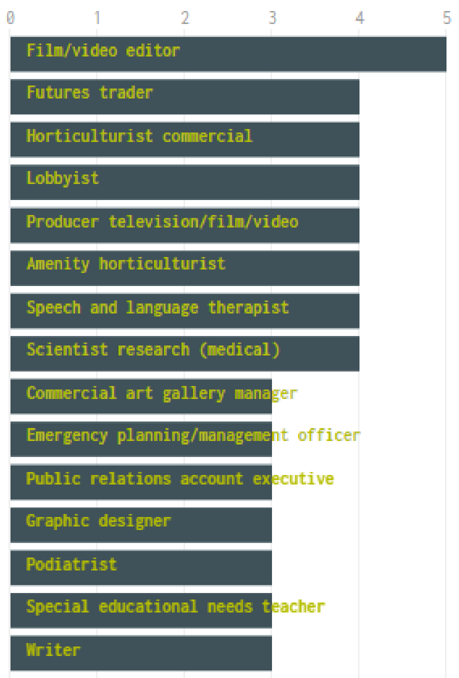
Empty: **0**

Invalid: **0**

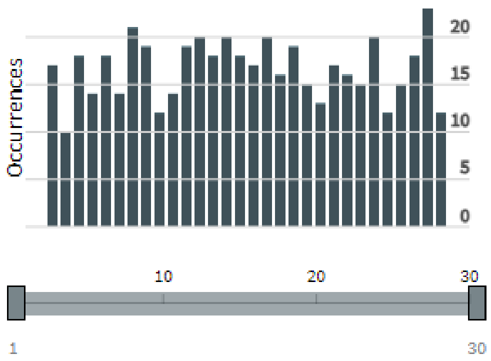
Avg length: **9.63**

Min length: **8**

Max length: **11**

LastPurchaseDate	
	Count: **500** Distinct: **273** Duplicate: **227** Valid: **500** Empty: **0** Invalid: **0**
Occupation	
	Count: **500** Avg length: **20.69** Distinct: **340** Duplicate: **160** Min length: **4** Valid: **500** Empty: **0** Max length: **43** Invalid: **0**

FrequencyOfVisits



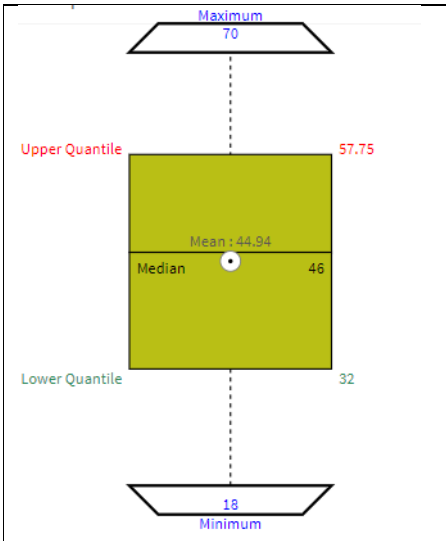
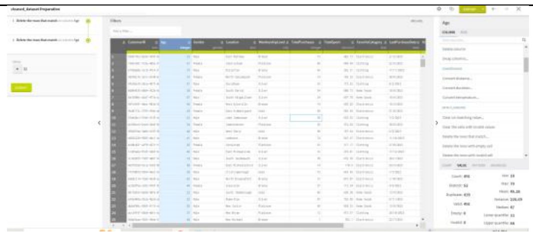
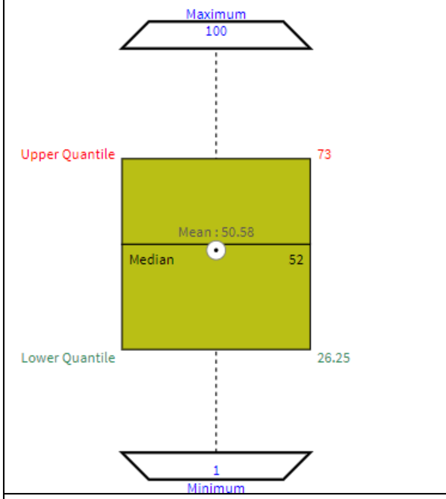
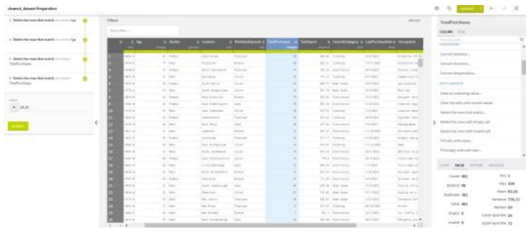
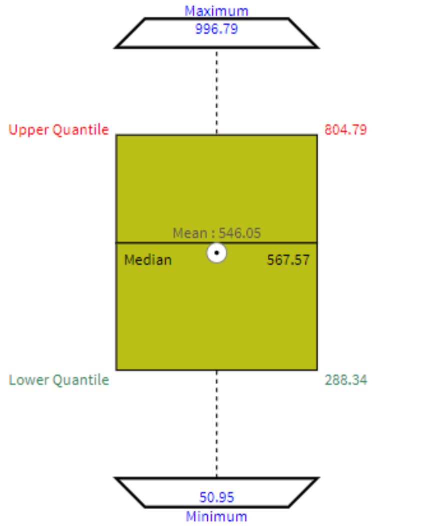
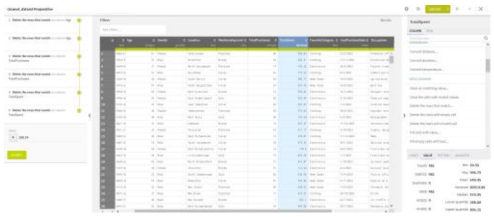
Count: 500	Min: 1
Distinct: 30	Max: 30
Duplicate: 470	Mean: 15.57
Valid: 500	Variance: 72.21
Empty: 0	Median: 15
Invalid: 0	Lower quantile: 8
	Upper quantile: 23

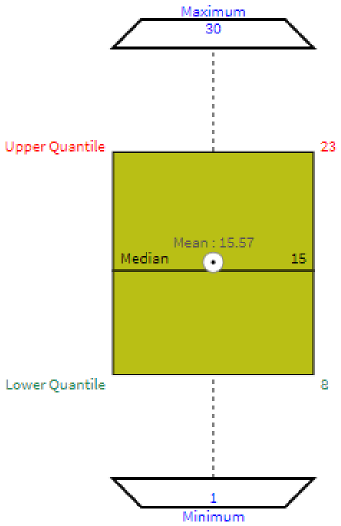
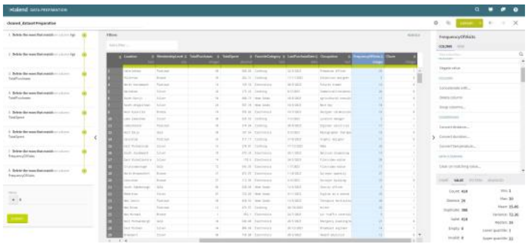
Churn

Count: 500	Min: 0
Distinct: 2	Max: 1
Duplicate: 498	Mean: 0.48
Valid: 500	Variance: 0.25
Empty: 0	Median: 0
Invalid: 0	Lower quantile: 0
	Upper quantile: 1

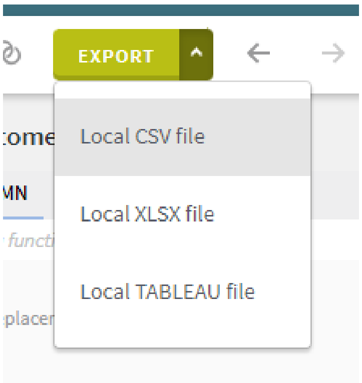
b. Outlier detection – Only applied to numeric data. Outliers detected are removed.
Advantages: Can plot the box plot without any coding.
Challenges: Outliers are not clearly indicated in the box plot.

Variable	Value
CustomerID	N/A
Age	491 rows left after removed outliers.

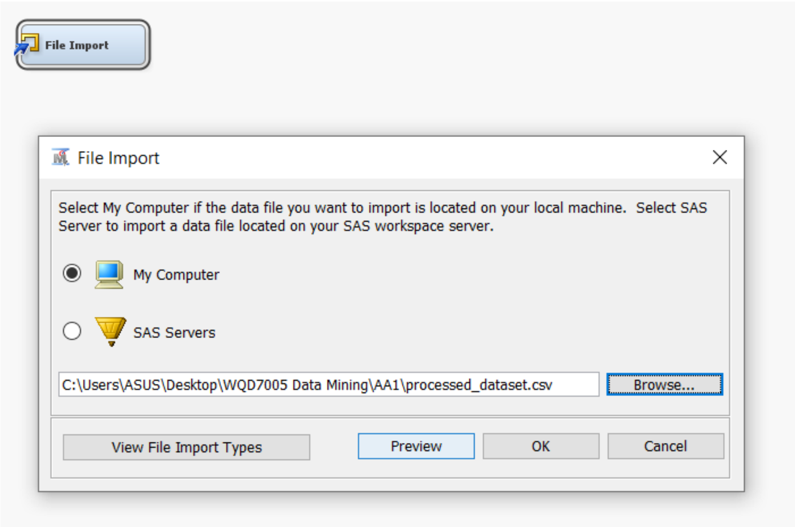
<table>
<tr><td>
</td><td></td></tr>
<tr><td>Gender</td><td>N/A</td></tr>
<tr><td>Location</td><td>N/A</td></tr>
<tr><td>MembershipLevel</td><td>N/A</td></tr>
<tr><td>TotalPurchases

</td><td>481 rows left after removed outliers.
</td></tr>
<tr><td>TotalSpent

</td><td>481 rows left after removed outliers.
</td></tr>
<tr><td>FavoriteCategory</td><td>N/A</td></tr>
</table>

LastPurchaseDate	N/A
Occupation	N/A
FrequencyOfVisits  	414 rows left after removed outliers. 
Churn	N/A

6. Export processed data as CSV.

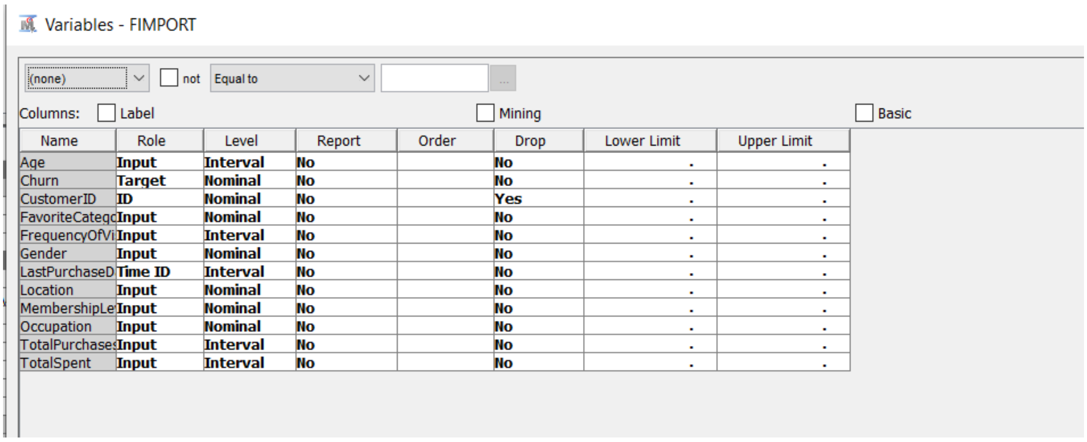


7. Import processed data into SAS Client Miner.



8. Specify variable roles.
   Advantages: Can determine the role and data level of each variable.

Challenges: Some useful information is not provided such as lower and upper limit for numeric data.



Variables - FIMPORT

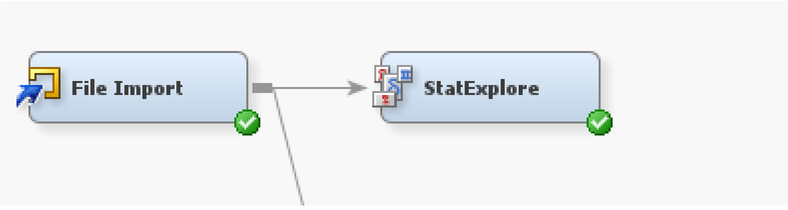
(none) not Equal to

Columns: Label Mining Basic

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

Decision Tree Analysis:

1. StatExplore



Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=FavoriteCategory

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Churn	0	3	0	Home Goods	36.62	Electronics	34.74
Churn	1	3	0	Clothing	41.38	Electronics	33.00
OVERALL		3	0	Clothing	34.86	Electronics	33.89

Data Role=TRAIN Variable Name=Gender

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Churn	0	2	0	Male	54.93	Female	45.07
Churn	1	2	0	Male	56.65	Female	43.35
OVERALL		2	0	Male	55.77	Female	44.23

Data Role=TRAIN Variable Name=MembershipLevel

Target	Target Level	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
Churn	0	4	0	Platinum	26.29	Bronze	25.35
Churn	1	4	0	Bronze	25.12	Platinum	25.12
OVERALL		4	0	Platinum	25.72	Bronze	25.24

There are three group data, FavouriteCategory, Gender, and MembershipLevel compared to Churn. For active purchasing customers, most of them prefer to buy home goods (36.62%) online, followed by electronic (34.74%). While most of the churn are prefer for clothing category (41.38%) and followed by electronic (33.00%). For gender, since majority of the observations are male, hence both churn and non-churn data comes from male, which are 56.65% and 54.93% respectively. For membership level, most of the Platinum members (26.29%) active purchasing goods online followed by Bronze members (25.35%). While those stop practice e-commerce are Bronze members (25.12%) followed by Platinum members (25.12%).

Interval Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable=Age

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Churn	0	45	0	213	18	70	44.76056	14.65335	-0.0721	-1.12154	INPUT	Age
Churn	1	47	0	203	18	70	45.26108	15.00349	-0.20431	-1.0907	INPUT	Age
OVERALL		46	0	416	18	70	45.00481	14.80947	-0.1372	-1.11008	INPUT	Age

Data Role=TRAIN Variable=FrequencyOfVisits

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Churn	0	14	0	213	1	30	15.10329	8.398598	0.079448	-1.09471	INPUT	FrequencyOfVisits
Churn	1	16	0	203	1	30	15.49261	8.727089	0.02628	-1.20224	INPUT	FrequencyOfVisits
OVERALL		15	0	416	1	30	15.29327	8.552351	0.054405	-1.15011	INPUT	FrequencyOfVisits

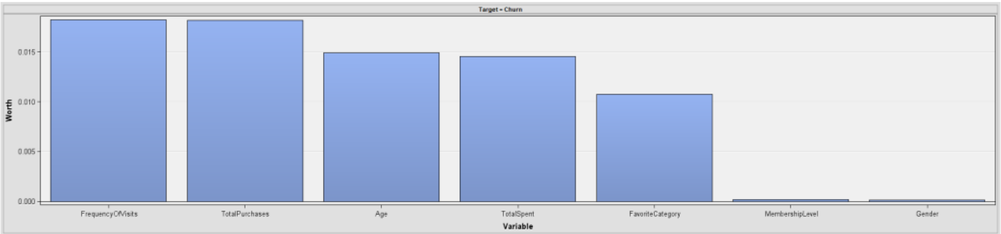
Data Role=TRAIN Variable=TotalPurchases

Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Churn	0	49	0	213	1	100	49.77465	27.69027	0.017955	-1.12148	INPUT	TotalPurchases
Churn	1	55	0	203	1	100	51.04433	28.41285	-0.13416	-1.15851	INPUT	TotalPurchases
OVERALL		52	0	416	1	100	50.39423	28.01855	-0.05719	-1.14596	INPUT	TotalPurchases

Data Role=TRAIN Variable=TotalSpent

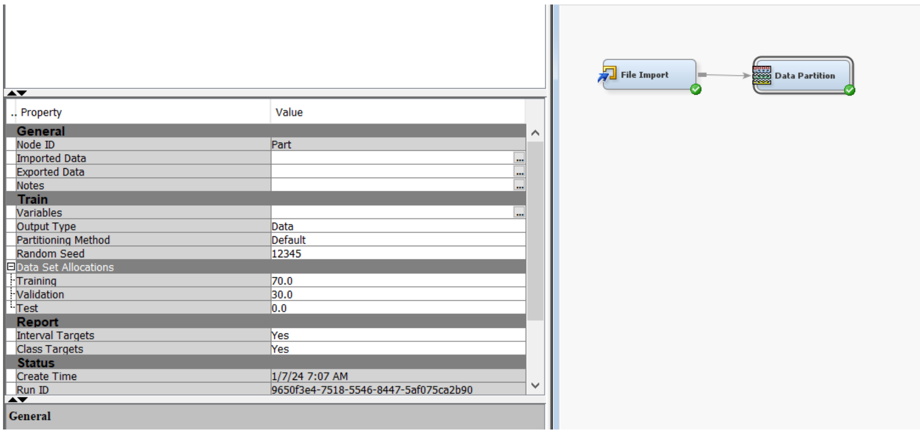
Target	Target Level	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
Churn	0	520.05	0	213	50.95	996.79	524.6474	287.8086	0.016174	-1.32264	INPUT	TotalSpent
Churn	1	597.78	0	203	68.89	994.61	569.2042	278.576	-0.19585	-1.23418	INPUT	TotalSpent
OVERALL		568.73	0	416	50.95	996.79	546.3903	283.8769	-0.08872	-1.29553	INPUT	TotalSpent

For interval variables in this dataset are Age, FrequencyOfVisits, TotalPurchases, and TotalSpent. The mean for the customers continues purchasing based on these four variables are 44.76053 year. 15.10329 times monthly, 49.77645 items and USD 524.6474 respectively. The mean for churn customers is 45.26108 year, 15.49261 times, 51.04433 items and USD 569.2042 respectively.



Among all variables, FrequencyOfVisits contributed highest percentage as churn customer, followed by TotalPurchases and Age.

2. Spilt dataset

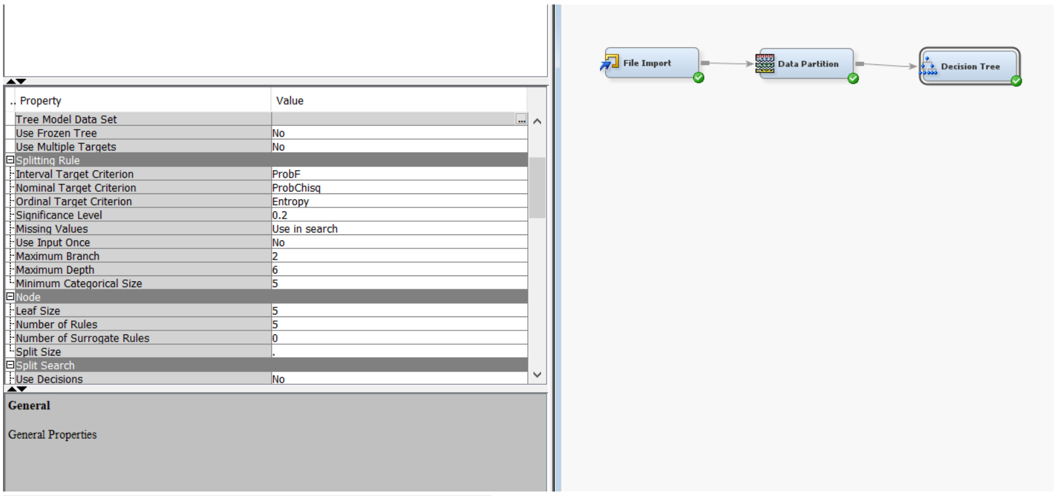


The dataset is spilt into 70% training data and 30% validation data.

```
Data=TRAIN

                Numeric     Formatted     Frequency
Variable         Value        Value         Count        Percent     Label

 Churn             0            0            149         51.3793
 Churn             1            1            141         48.6207


Data=VALIDATE

                Numeric     Formatted     Frequency
Variable         Value        Value         Count        Percent     Label

 Churn             0            0             64         50.7937
 Churn             1            1             62         49.2063
```

For training dataset, the data consists of 51.3793% churn customers and 48.6207% active customers. For validation dataset, the data consists of 50.7937% churn customers and 49.2063% active customers.

3. Model development




```
 Node Id:        1
Statistic  Train  Validation
       0: 52.30%      51.97%
       1: 47.70%      48.03%
   Count:    348         152
```

Classification Table

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	51.3793	100	149	51.3793
1	0	48.6207	100	141	48.6207

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

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	50.7937	100	64	50.7937
1	0	49.2063	100	62	49.2063

Event Classification Table

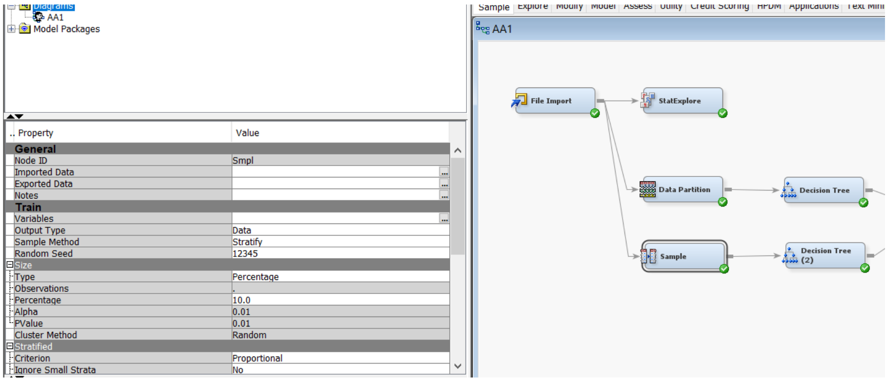
Data Role=TRAIN Target=Churn Target Label=' '

False Negative	True Negative	False Positive	True Positive
141	149	0	0

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

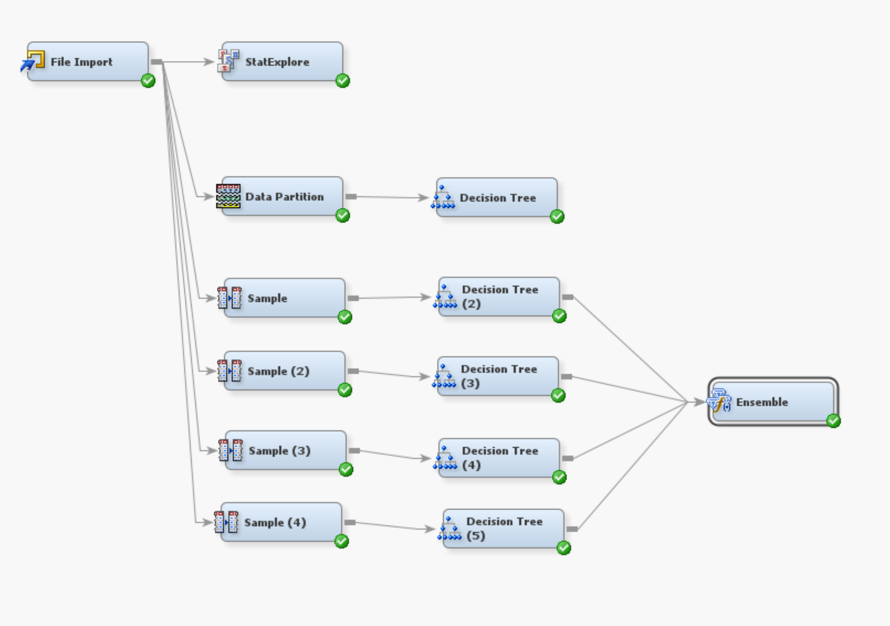
False Negative	True Negative	False Positive	True Positive
62	64	0	0

Since the first decision tree model is oversampling, all false positive and true positive are zero, the second model is generated by using stratified sample method.



```
Event Classification Table

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

  False       True        False       True
Negative    Negative    Positive    Positive

   20          22          0           0
```

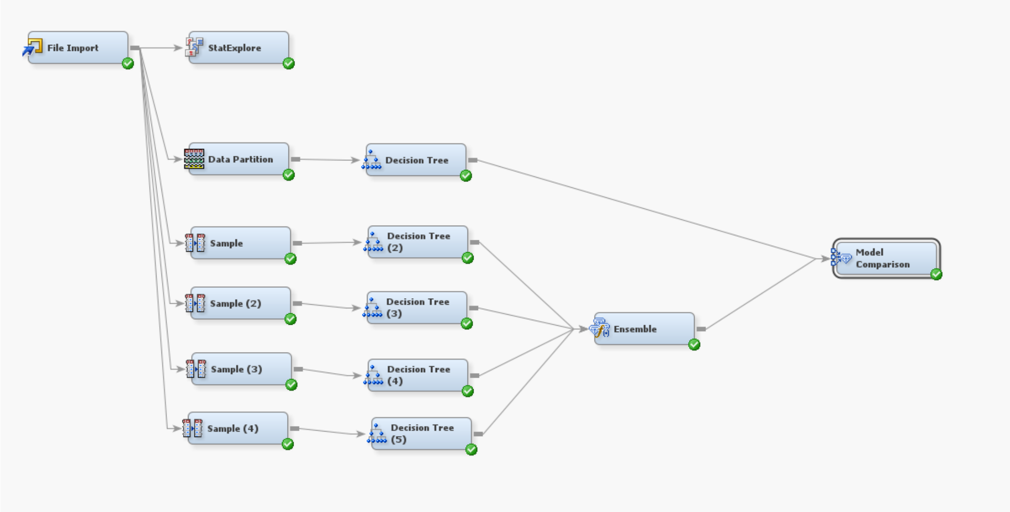
Since the results still showed as oversampling, ensemble method is used.




```
Event Classification Table

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

  False       True        False       True
Negative    Negative    Positive    Positive

   11          19          3           9
```

By using Bagging Ensemble model, the False Positive and True Positive results raised.

4. Model Evaluation



The results of first decision tree model and ensemble model are being compared.

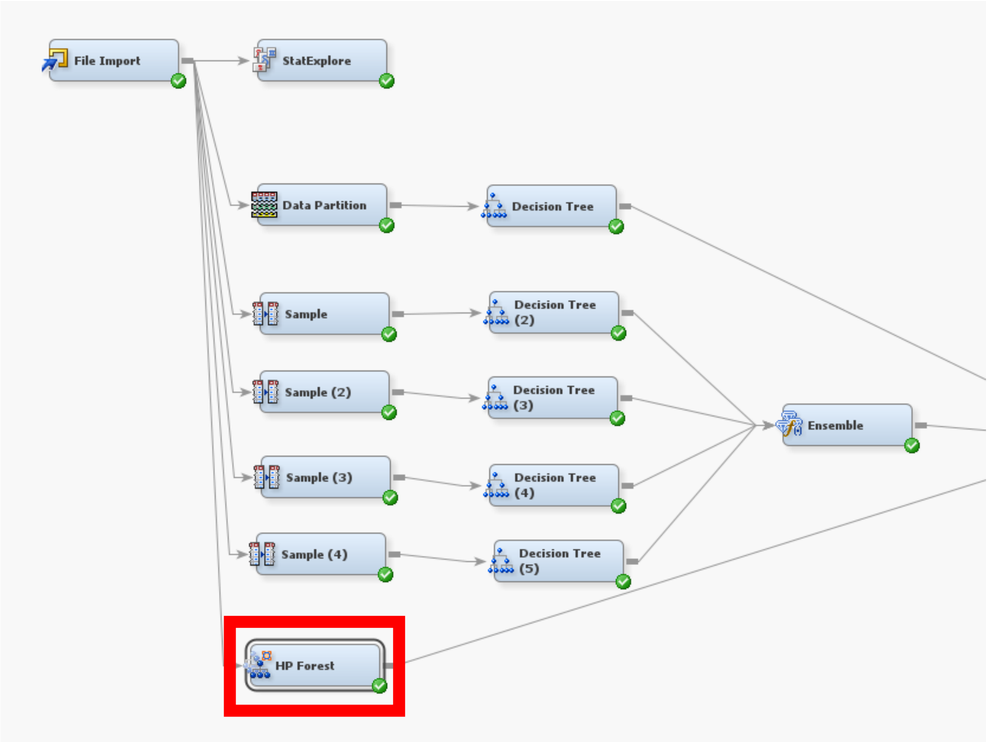
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
	Ensmbl	Ensemble	.	0.22132	0.33333	.
Y	Tree	Decision Tree	0.49206	0.24981	0.48621	0.24997

A lower misclassification rate indicates a lower ratio of the number of misclassified instances to the total number of instances while a lower average squared error indicates a lower average squared difference between the predicted and actual values. Both misclassification rate and average square error for Ensemble model (0.33333 and 0.22132) is lower than Decision Tree model (0.48621 and 0.24981), hence Ensemble model has a better predictive performance and accuracy.

Ensemble Methods:

1. HP Forest



The HPForest node creates predictive models by using a random forest ensemble methodology. It is similar to bagging in that it trains many decision trees by using different samples and then combines the predictions by averaging the posterior probabilities for interval targets or by voting for class targets (ensemble model used above).

From the iteration plot, the misclassification rate of out of bag fluctuates and lower than the misclassification rate of training data after around 27 trees at around 0.433.
Advantages: This model takes the equivalent amount of time to run a 31-tree forest as it takes to run a 4-iteration bagging or boosting.
Challenges: The model looks not so useful as the misclassification is unstable and may rise up again after 31 trees.

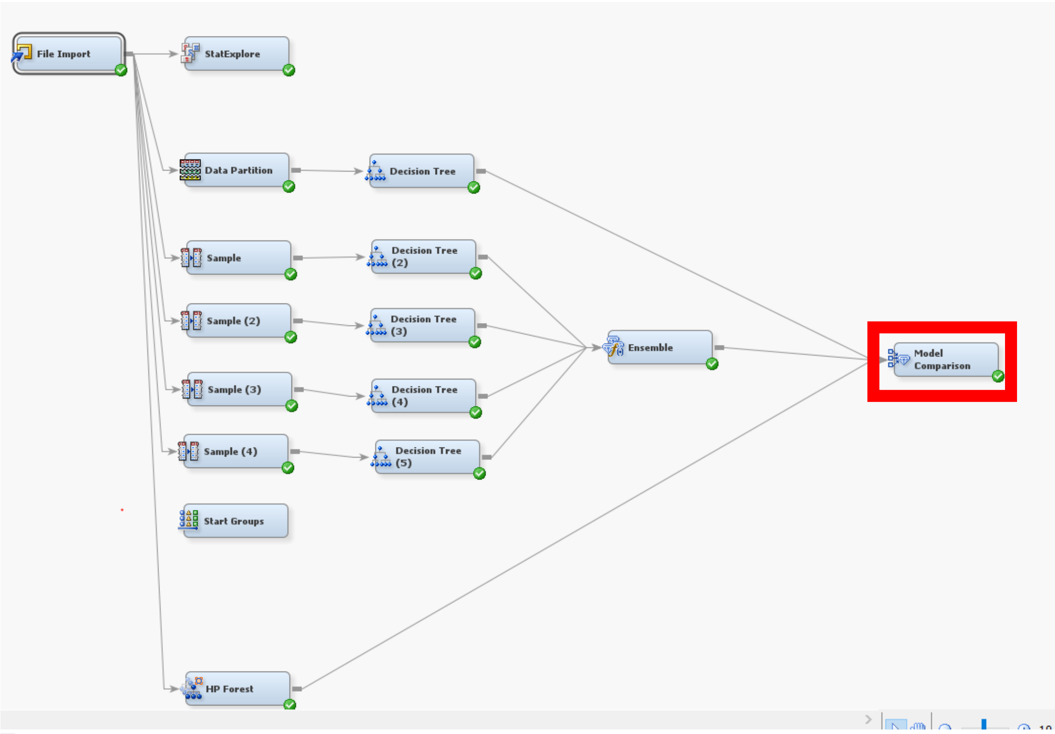
Event Classification Table

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

False Negative	True Negative	False Positive	True Positive
119	152	61	84

Besides, when we observed the classification table, there is no over-sampling occurred as in decision tree.

2. Model comparison



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
	Ensmbl	Ensemble	.	0.22132	0.33333	.
	HPDMForest	HP Forest	.	0.24341	0.43269	.
Y	Tree	Decision Tree	0.49206	0.24981	0.48621	0.24997

From the fit statistics, we can found that even HP Forest preform better than Decision Tree but Ensemble model perform the best among these 3 models. The misclassification rate and average square error for both HP Forest and Decision tree are almost the same whereas Ensemble model perform better with 0.33333 misclassification rate and 0.22132 average squared error.

Conclusion: Ensemble model performs better among these three models for this dataset.

Github link: https://github.com/s-s-yy/Soong-Sing-Ying
SAS file: https://drive.google.com/file/d/1FooKe1HRG7q7F6bGLaLvuuoiKH87ZYz1/view?usp=sharing