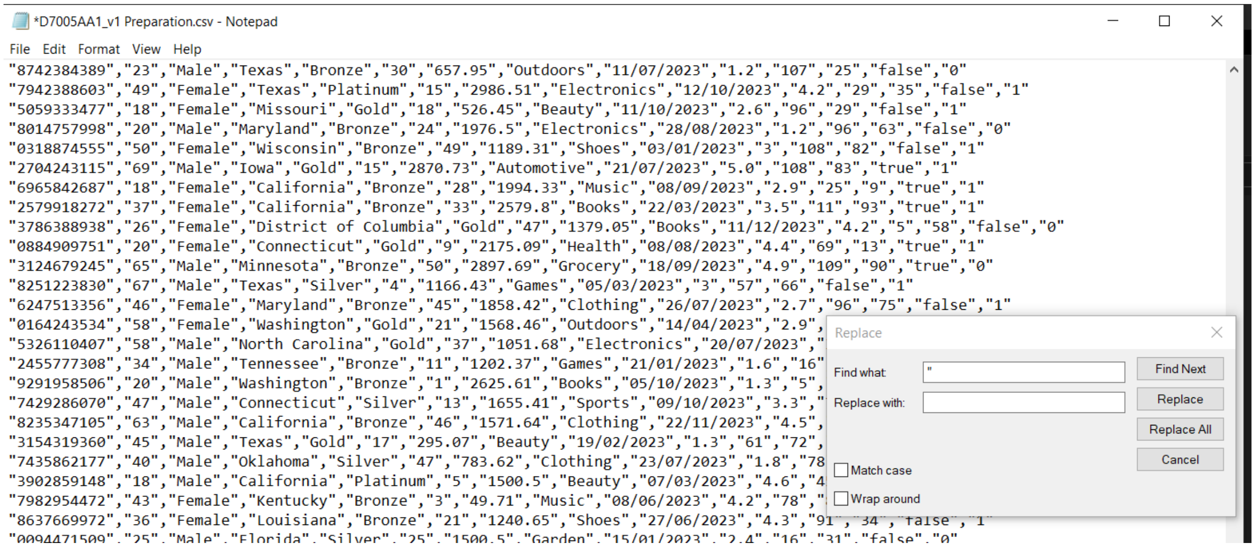
Part 4: Modelling using SAS Enterprise Miner

After a new diagram in a new project is created, the File Import node is placed into the diagram. This node is to import the processed dataset exported from the Talend Data Preparation.

Before that, open the CSV file with Notepad, notice that all values are covered with quotation marks. Hence, we can remove the quotation marks by using the Replace function as shown in Figure 17 below. This step is to ensure the data can be identified as Interval level when setting the schema of the dataset later in Figure 18.



*D7005AA1_v1 Preparation.csv - Notepad
File Edit Format View Help

```
"8742384389","23","Male","Texas","Bronze","30","657.95","Outdoors","11/07/2023","1.2","107","25","false","0"
"7942388603","49","Female","Texas","Platinum","15","2986.51","Electronics","12/10/2023","4.2","29","35","false","1"
"5059333477","18","Female","Missouri","Gold","18","526.45","Beauty","11/10/2023","2.6","96","29","false","1"
"8014757998","20","Male","Maryland","Bronze","24","1976.5","Electronics","28/08/2023","1.2","96","63","false","0"
"0318874555","50","Female","Wisconsin","Bronze","49","1189.31","Shoes","03/01/2023","3","108","82","false","1"
"2704243115","69","Male","Iowa","Gold","15","2870.73","Automotive","21/07/2023","5.0","108","83","true","1"
"6965842687","18","Female","California","Bronze","28","1994.33","Music","08/09/2023","2.9","25","9","true","1"
"2579918272","37","Female","California","Bronze","33","2579.8","Books","22/03/2023","3.5","11","93","true","1"
"3786388938","26","Female","District of Columbia","Gold","47","1379.05","Books","11/12/2023","4.2","5","58","false","0"
"0884909751","20","Female","Connecticut","Gold","9","2175.09","Health","08/08/2023","4.4","69","13","true","1"
"3124679245","65","Male","Minnesota","Bronze","50","2897.69","Grocery","18/09/2023","4.9","109","90","true","0"
"8251223830","67","Male","Texas","Silver","4","1166.43","Games","05/03/2023","3","57","66","false","1"
"6247513356","46","Female","Maryland","Bronze","45","1858.42","Clothing","26/07/2023","2.7","96","75","false","1"
"0164243534","58","Female","Washington","Gold","21","1568.46","Outdoors","14/04/2023","2.9",
"5326110407","58","Male","North Carolina","Gold","37","1051.68","Electronics","20/07/2023","
"2455777308","34","Male","Tennessee","Bronze","11","1202.37","Games","21/01/2023","1.6","16"
"9291958506","20","Male","Washington","Bronze","1","2625.61","Books","05/10/2023","1.3","5",
"7429286070","47","Male","Connecticut","Silver","13","1655.41","Sports","09/10/2023","3.3","
"8235347105","63","Male","California","Bronze","46","1571.64","Clothing","22/11/2023","4.5",
"3154319360","45","Male","Texas","Gold","17","295.07","Beauty","19/02/2023","1.3","61","72",
"7435862177","40","Male","Oklahoma","Silver","47","783.62","Clothing","23/07/2023","1.8","78
"3902859148","18","Male","California","Platinum","5","1500.5","Beauty","07/03/2023","4.6","4
"7982954472","43","Female","Kentucky","Bronze","3","49.71","Music","08/06/2023","4.2","78","
"8637669972","36","Female","Louisiana","Bronze","21","1240.65","Shoes","27/06/2023","4.3","91","34","false","1"
"0094471509","25","Male","Florida","Silver","25","1500.5","Garden","15/01/2023","2.4","16","31","false","0"
```

Replace

Find what: "

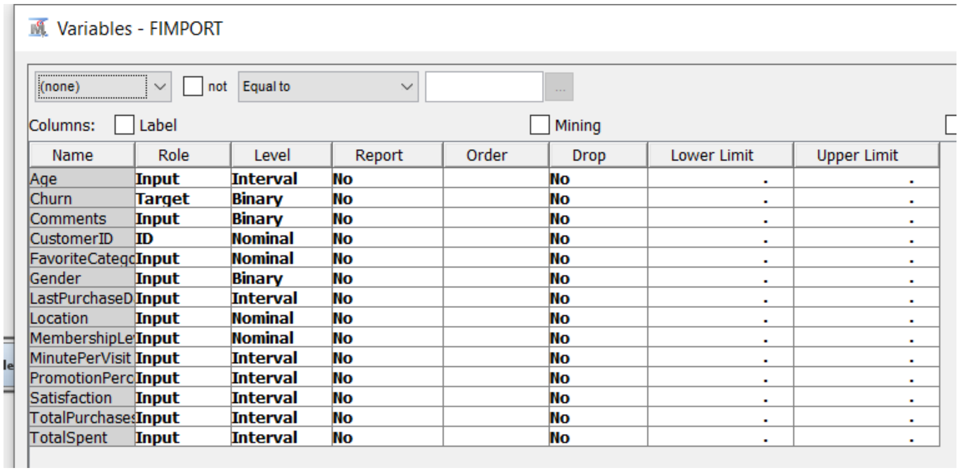
Replace with:

□ Match case

□ Wrap around

Figure 17: Removing quotation marks in NotePad.

The roles of each variable can be assigned during or after data import by selecting “Edit Variables” on the File Import Node. The setting for each variable is shown in Figure 18.



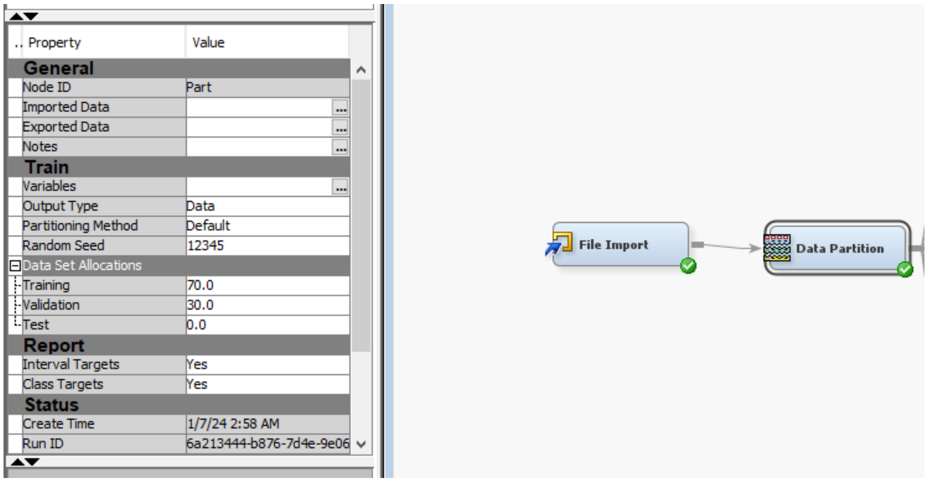
(none) □ not Equal to

Columns: □ Label □ Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Binary	No		No	.	.
Comments	Input	Binary	No		No	.	.
CustomerID	ID	Nominal	No		No	.	.
FavoriteCateg	Input	Nominal	No		No	.	.
Gender	Input	Binary	No		No	.	.
LastPurchaseD	Input	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipLe	Input	Nominal	No		No	.	.
MinutePerVisit	Input	Interval	No		No	.	.
PromotionPerc	Input	Interval	No		No	.	.
Satisfaction	Input	Interval	No		No	.	.
TotalPurchase	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

Figure 18: Data Schema settings.

The most important setting in Figure 18 is to set “Churn” as the target variable and make “CustomerID” as ID type to avoid the inclusion of this column during modelling. Next is data partitioning. A “Data Partition” node can added to the diagram and linked with “File Import”. In this study, the partition of data is set as 70% testing and 30% visualization.



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
⊟Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/7/24 2:58 AM
Run ID	6a213444-b876-7d4e-9e06

Figure 19: Setting up data partition.

Now, the data is ready to undergo modelling. In this study, decision tree-related models are used to model customer behaviour. Various decision tree models are constructed based on step-by-step tuning. The first two models are constructed as below:

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT1	Entropy	2	6	5	0.475	0.352
DT2	Gini	2	6	5	0.449	0.336

The misclassification results can be view via construction and running of Model Comparison node with the node linked to the model node. DT1 and DT2 are using fully default parameters except the nominal and ordinal target criterion, which is a part of parameter tuning. Observing the misclassification rate for validation, both models seem to have a large room for improvement available. As there are 51 locations and multiple options in FavouriteCategory available, the maximum branch is now increased to 3 as did for models DT3 and DT4.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332

Compared to model DT1, DT3 with extra maximum branch brings a huge improvement in reducing misclassification rate for both validation and training set. For the Gini tune, DT4 does improve in classification rate although it is not as large as the Entropy tune. Now, the maximum allowed branches is increasing to 4 to observe what will happen.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT5	Entropy	4	6	5	0.445	0.360
DT6	Gini	4	6	5	0.425	0.299

DT5 suffering a huge drop in performance of validation set classification tasks when compared to DT3 which has the same Entropy tune. The Gini-tuned DT6 is better than DT5 but not DT4. This situation indicates that is overfitting. Next, the maximum depth is adjusted based on models DT3 and DT4. The maximum depth is increasing by 1.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332
DT7	Entropy	3	7	5	0.385	0.250
DT8	Gini	3	7	5	0.418	0.332

Increasing maximum depth do not improve the classification rate for both tuned criterion. That means increasing the maximum depth to 7 is causing overfitting. The DT3 and DT4 are still selected as the best models for each tune. Now, the maximum depth is decreased to 5 to observe whether this action can reduce the misclassification rate for validation set.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332

DT9	Entropy	3	5	5	0.408	0.316
DT10	Gini	3	5	5	0.425	0.346

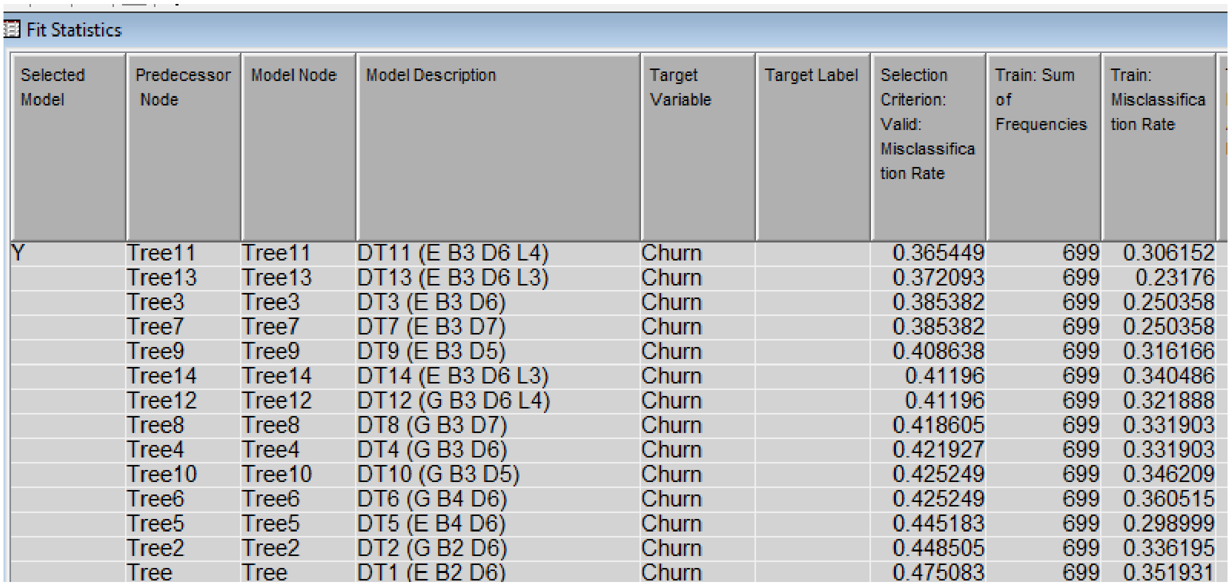
Both reduced maximum depth models (DT9 and DT10) are performing worse than before with increasing of misclassification rate for both validation and training set, indicating occurrence of underfitting. Hence the maximum branch of six is the best depth for this model. The next tuning is based on minimum leaf size.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332
DT11	Entropy	3	6	4	0.365	0.306
DT12	Gini	3	6	4	0.412	0.322

Both reduced minimum leaf size models (DT11 and DT12) are improving the performance in reducing misclassification rate for validation rate. For now, DT11 is the best Entropy-tuned model while DT12 is the best Gini-tuned model. Next, the minimum leaf size is further reduced to 3.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT11	Entropy	3	6	4	0.365	0.306
DT12	Gini	3	6	4	0.412	0.322
DT13	Entropy	3	6	3	0.372	0.231
DT14	Gini	3	6	3	0.412	0.340

Both DT13 and DT14 does not further improve the classification result of validation set. For DT13, the increasing difference between the validation and training set misclassification rate indicates the model is overfitting at this stage. There is no further decision tree tuning so the model DT11 with Entropy-tuned, maximum 3 branches, maximum 6 layer depth and minimum leaf size of 4 is the best decision tree model for this study. The overall performance ranking is tabulated in the output of Figure 20.



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate
Y	Tree11	Tree11	DT11 (E B3 D6 L4)	Churn		0.365449	699	0.306152
	Tree13	Tree13	DT13 (E B3 D6 L3)	Churn		0.372093	699	0.23176
	Tree3	Tree3	DT3 (E B3 D6)	Churn		0.385382	699	0.250358
	Tree7	Tree7	DT7 (E B3 D7)	Churn		0.385382	699	0.250358
	Tree9	Tree9	DT9 (E B3 D5)	Churn		0.408638	699	0.316166
	Tree14	Tree14	DT14 (E B3 D6 L3)	Churn		0.41196	699	0.340486
	Tree12	Tree12	DT12 (G B3 D6 L4)	Churn		0.41196	699	0.321888
	Tree8	Tree8	DT8 (G B3 D7)	Churn		0.418605	699	0.331903
	Tree4	Tree4	DT4 (G B3 D6)	Churn		0.421927	699	0.331903
	Tree10	Tree10	DT10 (G B3 D5)	Churn		0.425249	699	0.346209
	Tree6	Tree6	DT6 (G B4 D6)	Churn		0.425249	699	0.360515
	Tree5	Tree5	DT5 (E B4 D6)	Churn		0.445183	699	0.298999
	Tree2	Tree2	DT2 (G B2 D6)	Churn		0.448505	699	0.336195
	Tree	Tree	DT1 (E B2 D6)	Churn		0.475083	699	0.351931

Figure 20: Overall decision tree model ranking.

Notice that the best Gini-tuned model (DT12) is only can be ranked six out of fourteen models, indicating that Gini-tuning is not quite suitable in this study.

Several models are included in the modelling such as Bagging Ensemble Method (Random Forest), Gradient Boosting HP Tree and HP Forest. The Ensemble model 1 is the ensemble of the Top 2 best decision tree models (DT11 and DT13) while the Ensemble model 2 is the ensemble of the best Entrophy-tuned model (DT11) and the best Gini-tuned model (DT12). Both HP environment tree and forest and the Gradient Boosting use the default parameters. The overall workflow is shown in Figure 21 and the overall results are in the output of Figure 22.

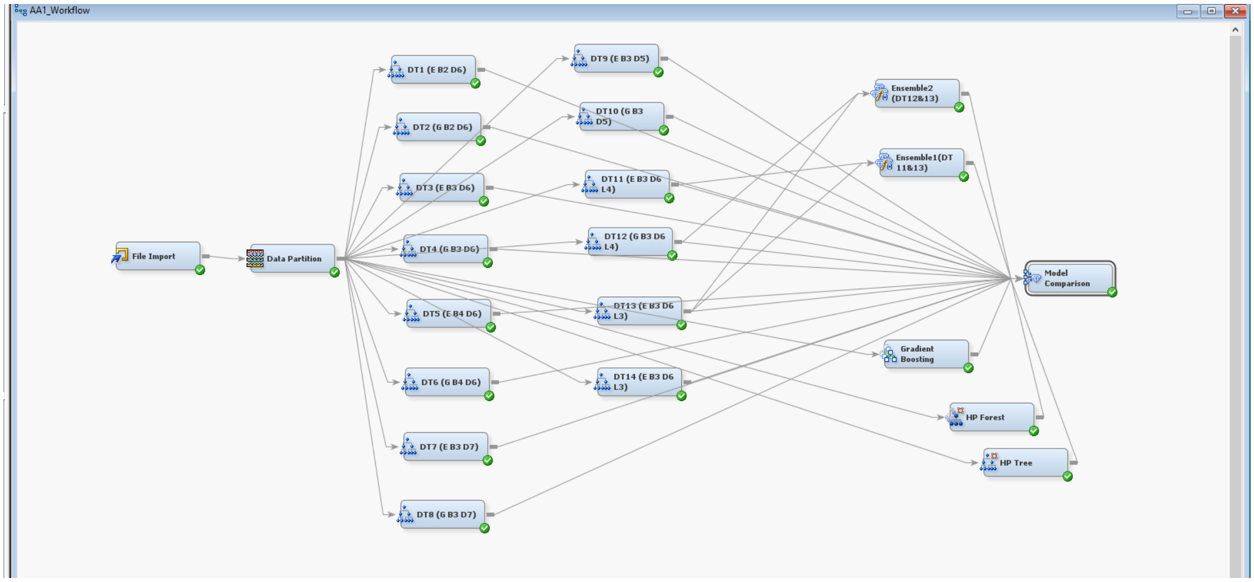
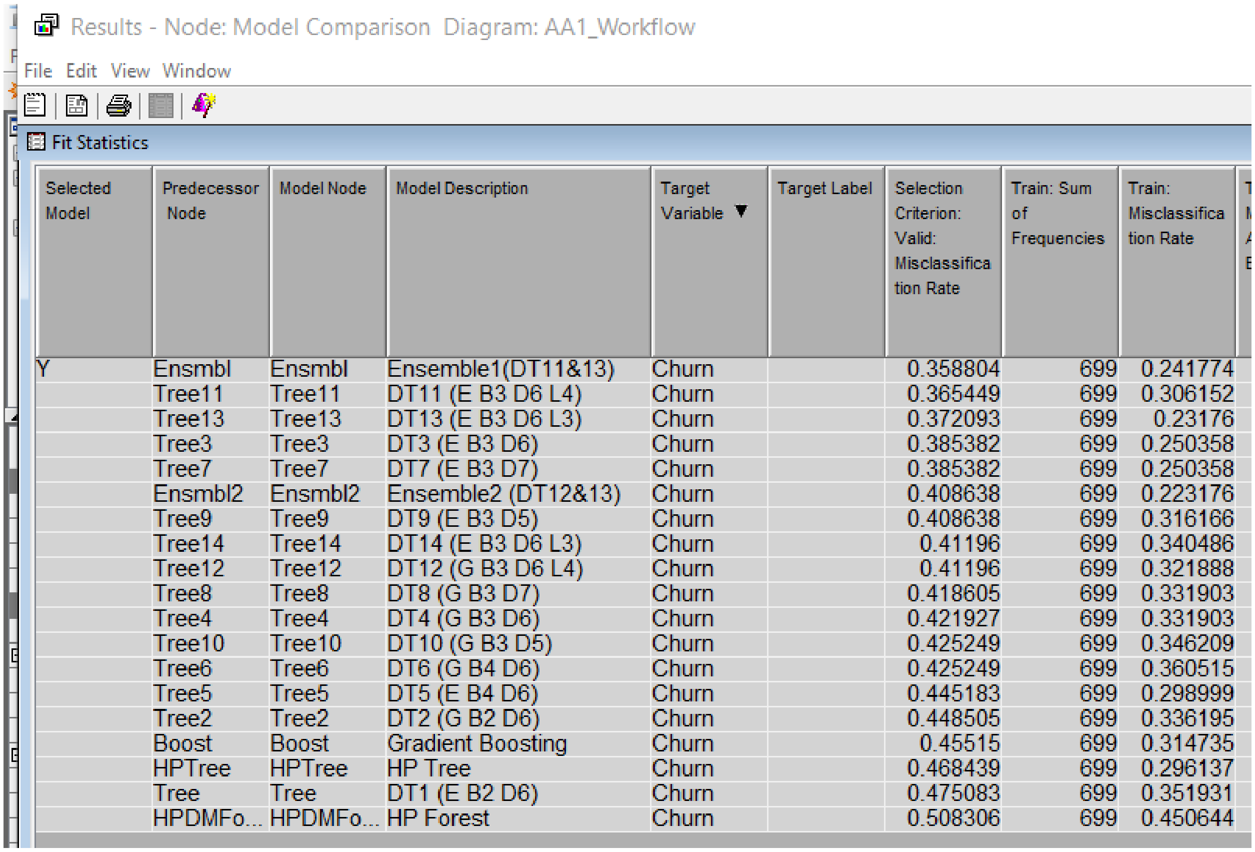


Figure 21: Overall Modelling Diagram in SAS EM.



Results - Node: Model Comparison Diagram: AA1_Workflow

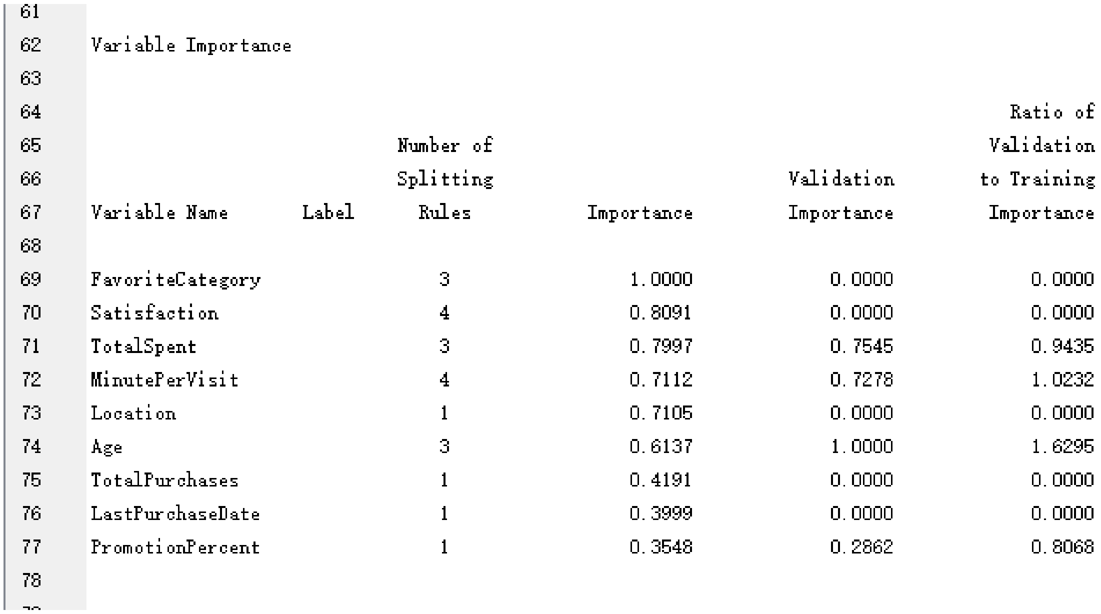
Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable ▼	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate
Y	Ensmbl	Ensmbl	Ensemble1(DT11&13)	Churn		0.358804	699	0.241774
	Tree11	Tree11	DT11 (E B3 D6 L4)	Churn		0.365449	699	0.306152
	Tree13	Tree13	DT13 (E B3 D6 L3)	Churn		0.372093	699	0.23176
	Tree3	Tree3	DT3 (E B3 D6)	Churn		0.385382	699	0.250358
	Tree7	Tree7	DT7 (E B3 D7)	Churn		0.385382	699	0.250358
	Ensmbl2	Ensmbl2	Ensemble2 (DT12&13)	Churn		0.408638	699	0.223176
	Tree9	Tree9	DT9 (E B3 D5)	Churn		0.408638	699	0.316166
	Tree14	Tree14	DT14 (E B3 D6 L3)	Churn		0.41196	699	0.340486
	Tree12	Tree12	DT12 (G B3 D6 L4)	Churn		0.41196	699	0.321888
	Tree8	Tree8	DT8 (G B3 D7)	Churn		0.418605	699	0.331903
	Tree4	Tree4	DT4 (G B3 D6)	Churn		0.421927	699	0.331903
	Tree10	Tree10	DT10 (G B3 D5)	Churn		0.425249	699	0.346209
	Tree6	Tree6	DT6 (G B4 D6)	Churn		0.425249	699	0.360515
	Tree5	Tree5	DT5 (E B4 D6)	Churn		0.445183	699	0.298999
	Tree2	Tree2	DT2 (G B2 D6)	Churn		0.448505	699	0.336195
	Boost	Boost	Gradient Boosting	Churn		0.45515	699	0.314735
	HPTree	HPTree	HP Tree	Churn		0.468439	699	0.296137
	Tree	Tree	DT1 (E B2 D6)	Churn		0.475083	699	0.351931
	HPDMFo...	HPDMFo...	HP Forest	Churn		0.508306	699	0.450644

Figure 22: Overall modelling result.

From the output in Figure 22, all three models with the default setting (Gradient Boosting, HP Tree and HP Forest) are among the weakest models in classification. Future tuning with better understanding is recommended so that the fined-tuned models have the opportunity to surpass the best decision tree model. The ensemble model of DT11 and DT13 is the best-performed model in validation set churn classification. Although the bagged ensemble model is the best model, for easier interpretation, the second best model which is DT11 is used to generate insight regarding customer behaviour.

Although the misclassification rate is 36% which is considered high, by observing the decision tree diagram, the Location and the favoured Content are the top two layers for every branch. However, by observing the variable importance in Figure 23, the most important variable for validation is age. Hence, I would like to suggest the e-commerce owner put more focus on location, age and favoured category as these factors are likely to influence the churn of the customer.



```
Variable Importance

                                                                                 Ratio of
                                Number of                                       Validation
                                Splitting                          Validation    to Training
Variable Name       Label         Rules          Importance        Importance     Importance

FavoriteCategory                    3              1.0000            0.0000         0.0000
Satisfaction                        4              0.8091            0.0000         0.0000
TotalSpent                          3              0.7997            0.7545         0.9435
MinutePerVisit                      4              0.7112            0.7278         1.0232
Location                            1              0.7105            0.0000         0.0000
Age                                 3              0.6137            1.0000         1.6295
TotalPurchases                      1              0.4191            0.0000         0.0000
LastPurchaseDate                    1              0.3999            0.0000         0.0000
PromotionPercent                    1              0.3548            0.2862         0.8068
```

Figure 23: Result output of model DT11.