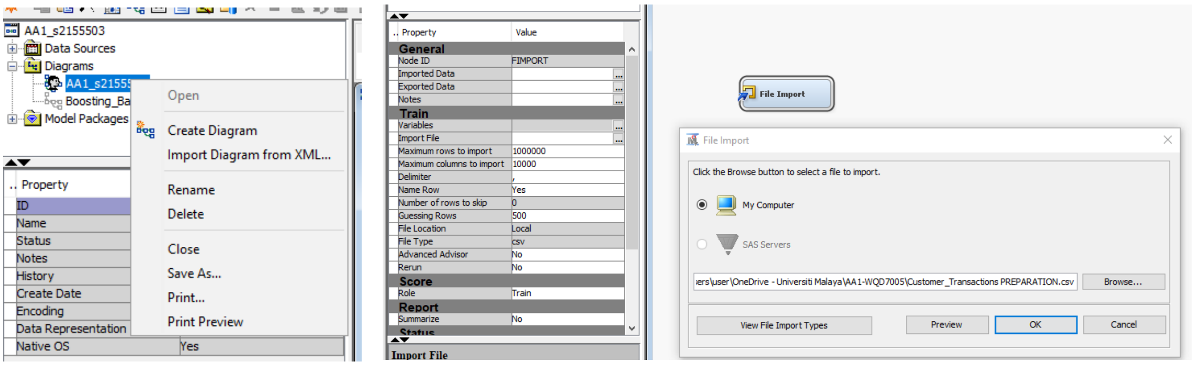
SAS Enterprise Miner

Data Import to SAS Enterprise Miner



Property	
ID	
Name	
Status	
Notes	
History	
Create Date	
Encoding	
Data Representation	
Native OS	Yes

Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	...
Maximum rows to import	1000000
Maximum columns to import	10000
Delimiter	,
Name Row	Yes
Number of rows to skip	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Rerun	No
Score	
Role	Train
Report	
Summarize	No
Status	

File Import

Click the Browse button to select a file to import.

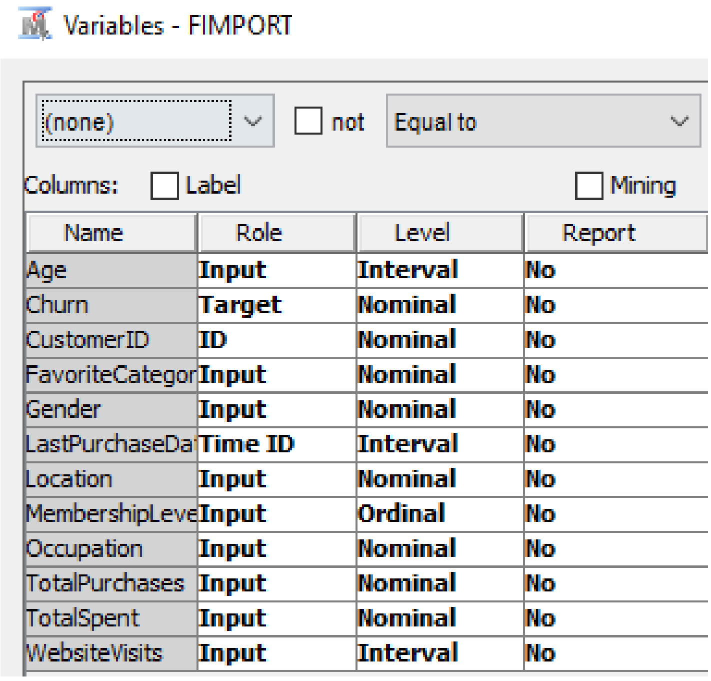
● My Computer

○ SAS Servers

ers\user\OneDrive - Universiti Malaya\AA1-WQD7005\Customer_Transactions PREPARATION.csv

In the initial phase, begin by creating a diagram in SAS Enterprise Miner. Then, choose the 'File Import' component from the sample options. Following this, navigate to the side tab, search for 'Import File', and select the path of the pre-processed CSV file already exported from Talend Data Preparation.

Specify Variable Roles



Variables - FIMPORT

(none) □ not Equal to

Columns: □ Label □ Mining

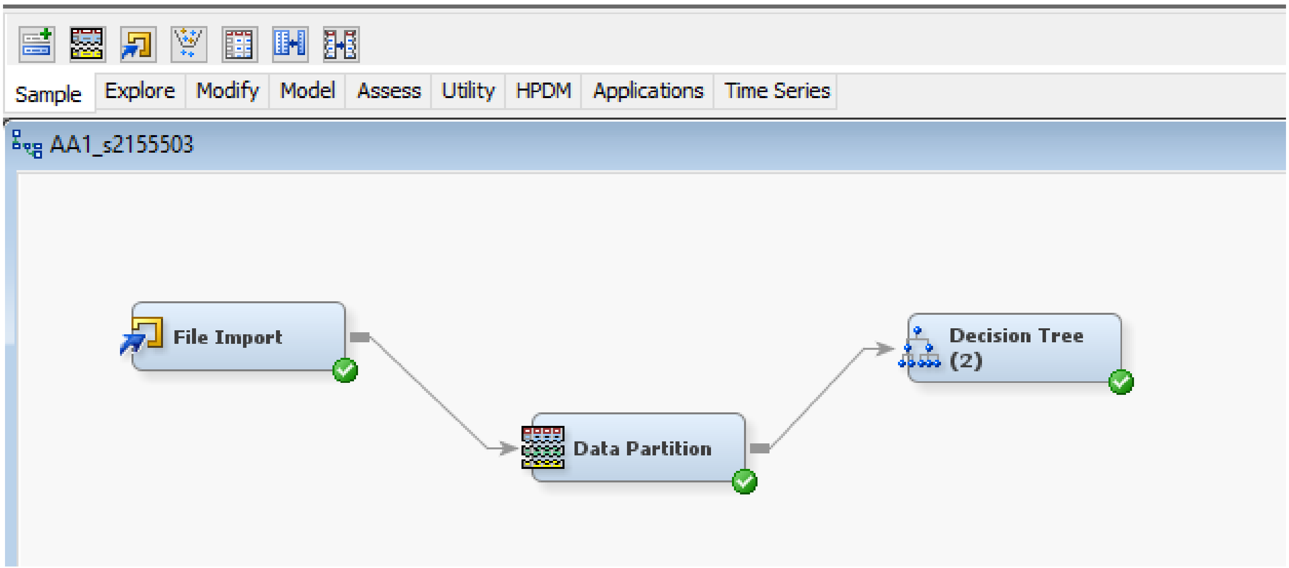
Name	Role	Level	Report
Age	Input	Interval	No
Churn	Target	Nominal	No
CustomerID	ID	Nominal	No
FavoriteCategor	Input	Nominal	No
Gender	Input	Nominal	No
LastPurchaseDa	Time ID	Interval	No
Location	Input	Nominal	No
MembershipLeve	Input	Ordinal	No
Occupation	Input	Nominal	No
TotalPurchases	Input	Nominal	No
TotalSpent	Input	Nominal	No
WebsiteVisits	Input	Interval	No

In the SAS Enterprise Miner process, specifying the variable roles is a key step to ensure the correct analysis of data. As shown in the variables list, each attribute has been assigned a specific role and level appropriate to its nature. For instance, 'Age', 'TotalSpent', and

'WebsiteVisits' are marked as 'Input' with an 'Interval' level, suitable for continuous data. 'Churn' is set as the 'Target' variable with a 'Nominal' level, as it is a categorical outcome we aim to predict. 'CustomerID' and 'LastPurchaseDate' are labeled as 'ID' and 'Time ID' respectively, recognizing their use as identifiers rather than variables to be analyzed. Other attributes like 'Gender', 'Location', and 'MembershipLevel' are categorized as 'Input' with a 'Nominal' or 'Ordinal' level, indicating their role in the model as categorical predictors. This designation ensures that each variable is treated correctly in the modeling process, facilitating accurate data analysis.

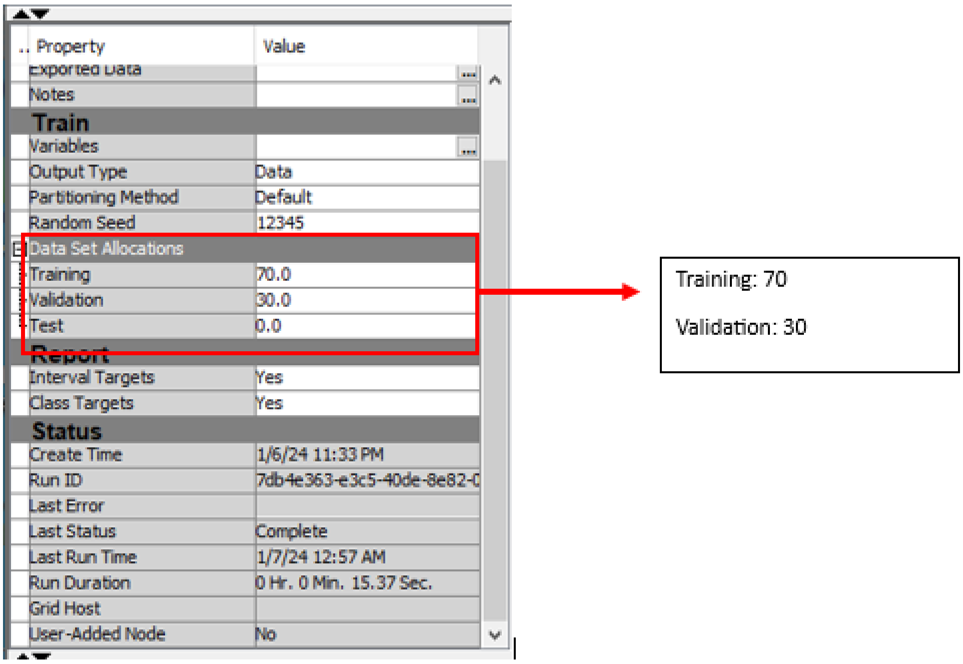
1.0 Tasks 2

Decision Tree Analysis: *Create a decision tree model in SAS Enterprise Miner to analyse customer behaviour.*



As shown in the diagram, establish a decision tree model by linking three components which are file import, data partitioning, and the decision tree. The specific purpose of the component and property configuration data partitioning and Decision Tree will be discussed in the next section.

Data Partitioning



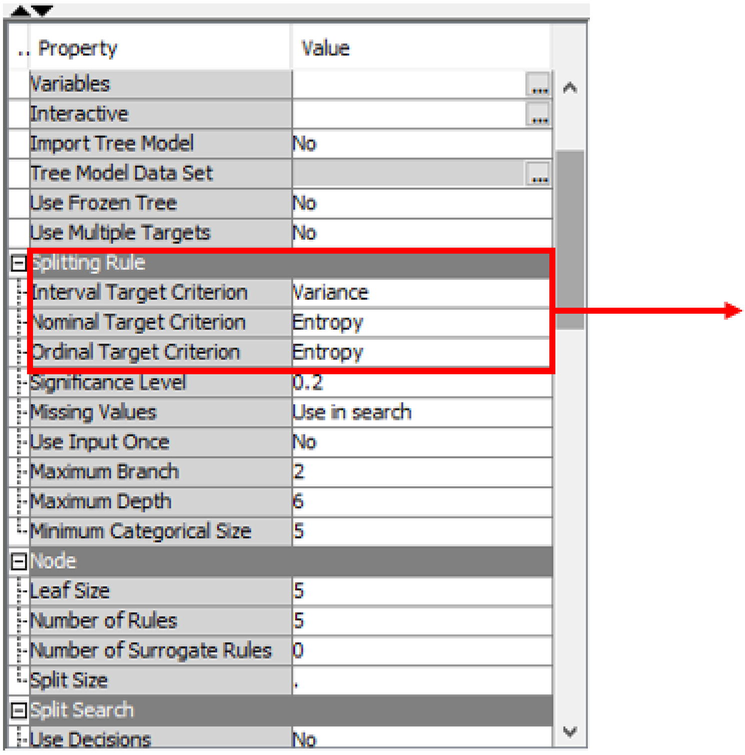
Property	Value
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/6/24 11:33 PM
Run ID	7db4e363-e3c5-40de-8e82-
Last Error	
Last Status	Complete
Last Run Time	1/7/24 12:57 AM
Run Duration	0 Hr. 0 Min. 15.37 Sec.
Grid Host	
User-Added Node	No

Training: 70

Validation: 30

In SAS Enterprise Miner, data partitioning is used to split a dataset into separate subsets for model building and validation, allowing for unbiased assessment of model performance. For this model, I use a 70% portion for training and a 30% portion for validation.

Decision Tree



Property	Value
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	Variance
Nominal Target Criterion	Entropy
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No

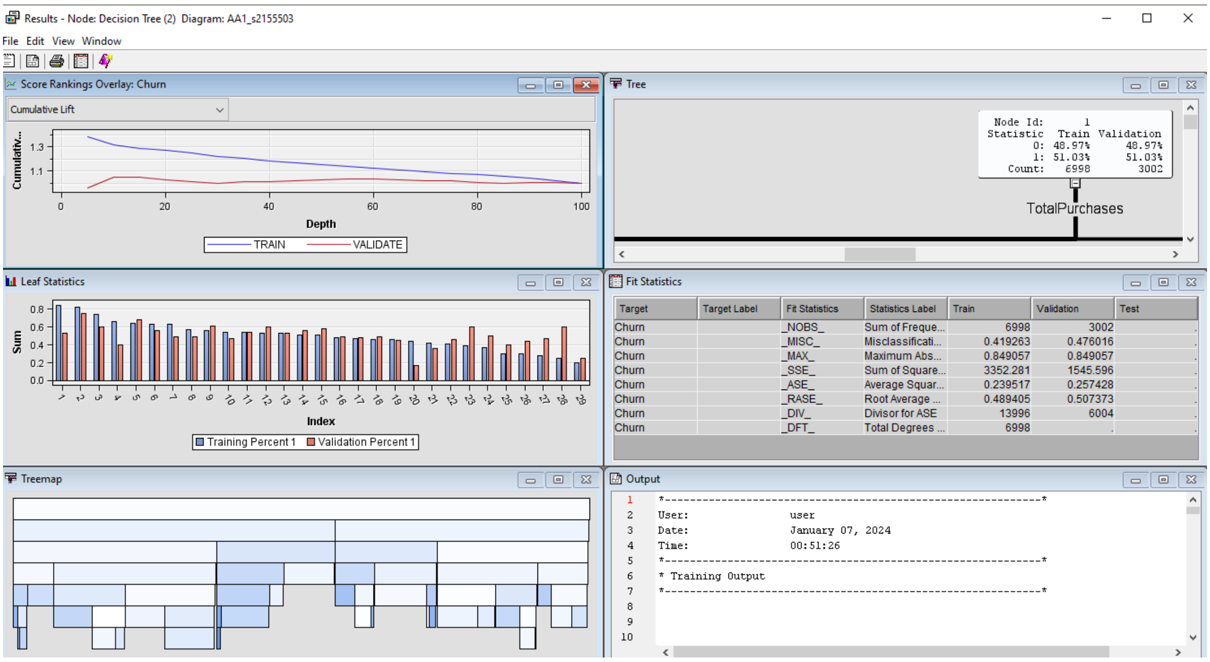
Change the Splitting Rule

Interval Target Criterion: Variance

Nominal Target Criterion: Entropy

Ordinal Target Criterion: Entropy

When building a decision tree model, the splitting rule based on the nature of our target variable. For interval target variables, the criterion often used is Variance, which minimizes the variance within each split. In cases with nominal target variables, the criterion of choice is typically Entropy, which measures the disorder or impurity of data at each split. Similarly, for ordinal target variables, the Entropy criterion is also favoured. By adapting the splitting rule to your specific target variable, you can optimize the performance and interpretability of your decision tree model.



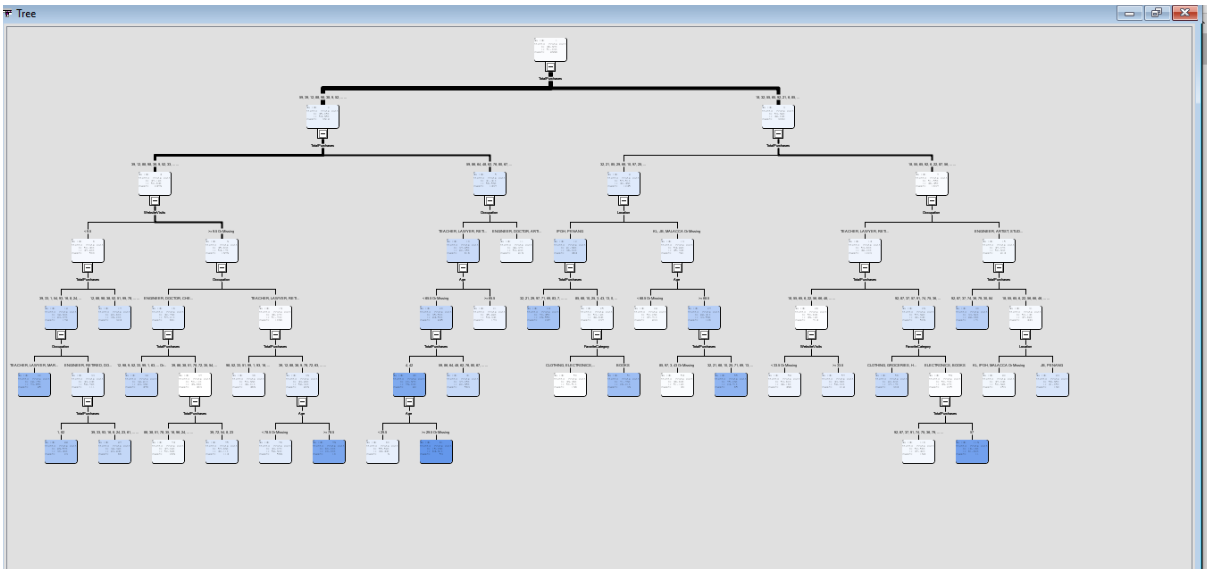
```
Node Id:        1
Statistic  Train  Validation
       0:  48.97%     48.97%
       1:  51.03%     51.03%
   Count:   6998        3002
```

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Freque...	6998	3002	.
Churn		_MISC_	Misclassificati...	0.419263	0.476016	.
Churn		_MAX_	Maximum Abs...	0.849057	0.849057	.
Churn		_SSE_	Sum of Square...	3352.281	1545.596	.
Churn		_ASE_	Average Squar...	0.239517	0.257428	.
Churn		_RASE_	Root Average ...	0.489405	0.507373	.
Churn		_DIV_	Divisor for ASE	13996	6004	.
Churn		_DFT_	Total Degrees ...	6998	.	.

```
*------------------------------------------------------------*
User:                     user
Date:                     January 07, 2024
Time:                     00:51:26
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*
```

The figure provides an overall result generated by this model. It encompasses various visual elements, including a score ranking overlay, a tree diagram, fit statistics, a treemap, and more.

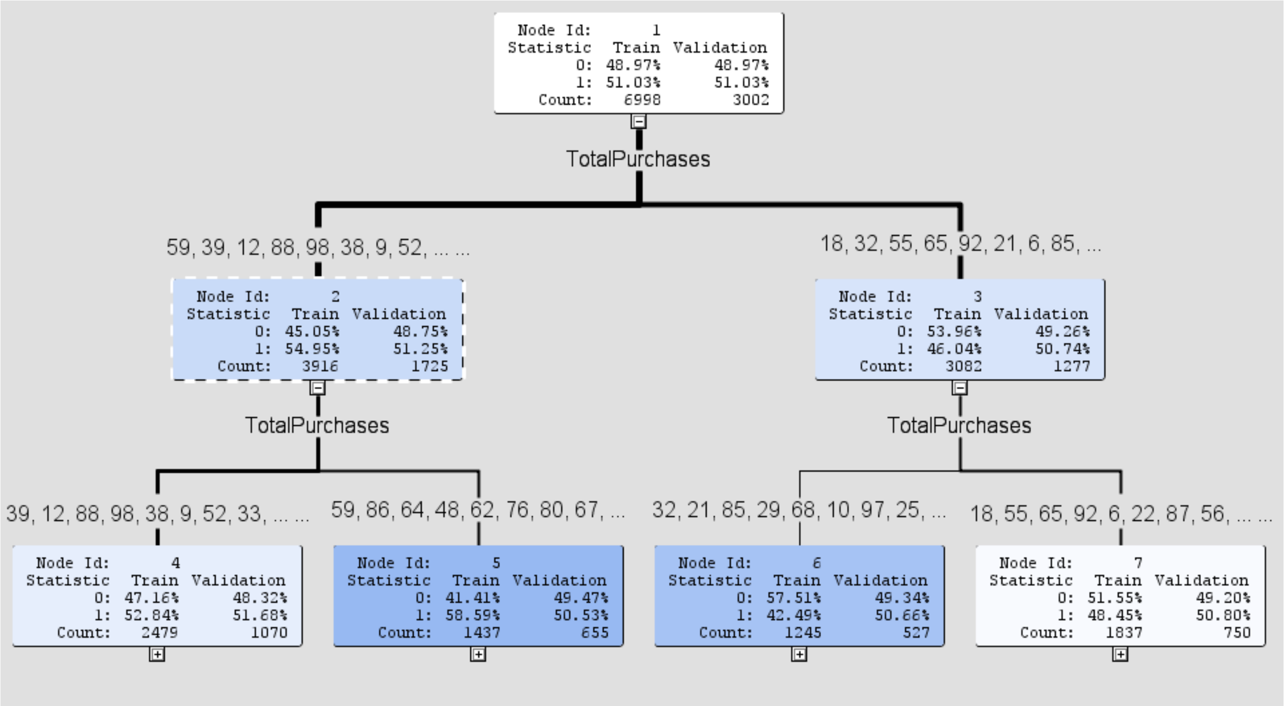
Here's the complete tree diagram for the model predicting observed target values as shown in below: -



Tree diagram with 14 splitting rules for the input column “TotalPurchases”.

```
Variable Importance

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

TotalPurchases                          14           1.0000        1.0000        1.0000
Occupation                               4           0.3809        0.4956        1.3014
Age                                      4           0.3516        0.6659        1.8936
FavoriteCategory                         2           0.2623        0.0000        0.0000
WebsiteVisits                            2           0.2532        0.4962        1.9596
Location                                 2           0.2510        0.5514        2.1966
```

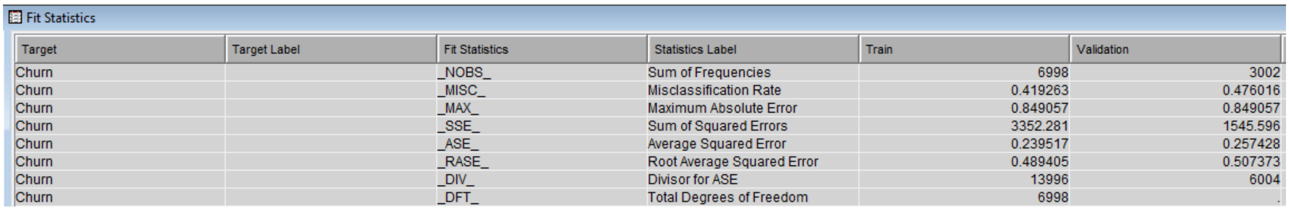
According to the Variable Importance table the most important predictors is TotalPurchases.



The decision tree analysis, focusing on 'TotalPurchases', reveals distinct customer segments based on their purchasing behavior. Initial splits in the tree indicate that the number of purchases is a significant predictor of the behavior under study, likely customer churn. The detailed node statistics suggest that as the total number of purchases varies, so does the likelihood of a customer falling into one of two categories, which could signify churned versus active customers.

The further Nodes 4 to 7 suggests more nuanced thresholds of purchase behavior that are influential in predicting outcomes. For instance, certain nodes with higher purchase counts may correlate with a greater likelihood of customer retention, while others with fewer purchases might indicate a risk of churning.

Overall, the decision tree provides actionable insights, suggesting that the frequency of purchases is a key behavioral indicator. This information can be leveraged to tailor customer engagement strategies, such as targeted marketing to increase purchase frequency among customers showing signs of potential churn.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Churn		_NOBS_	Sum of Frequencies	6998	3002
Churn		_MISC_	Misclassification Rate	0.419263	0.476016
Churn		_MAX_	Maximum Absolute Error	0.849057	0.849057
Churn		_SSE_	Sum of Squared Errors	3352.281	1545.596
Churn		_ASE_	Average Squared Error	0.239517	0.257428
Churn		_RASE_	Root Average Squared Error	0.489405	0.507373
Churn		_DIV_	Divisor for ASE	13996	6004
Churn		_DFT_	Total Degrees of Freedom	6998	.

From this Fit Statistics Table, it seems that the misclassification rates maximum error are in acceptable range. The ASE and RASE also suggest that the model's predictions are close to the actual values on average.

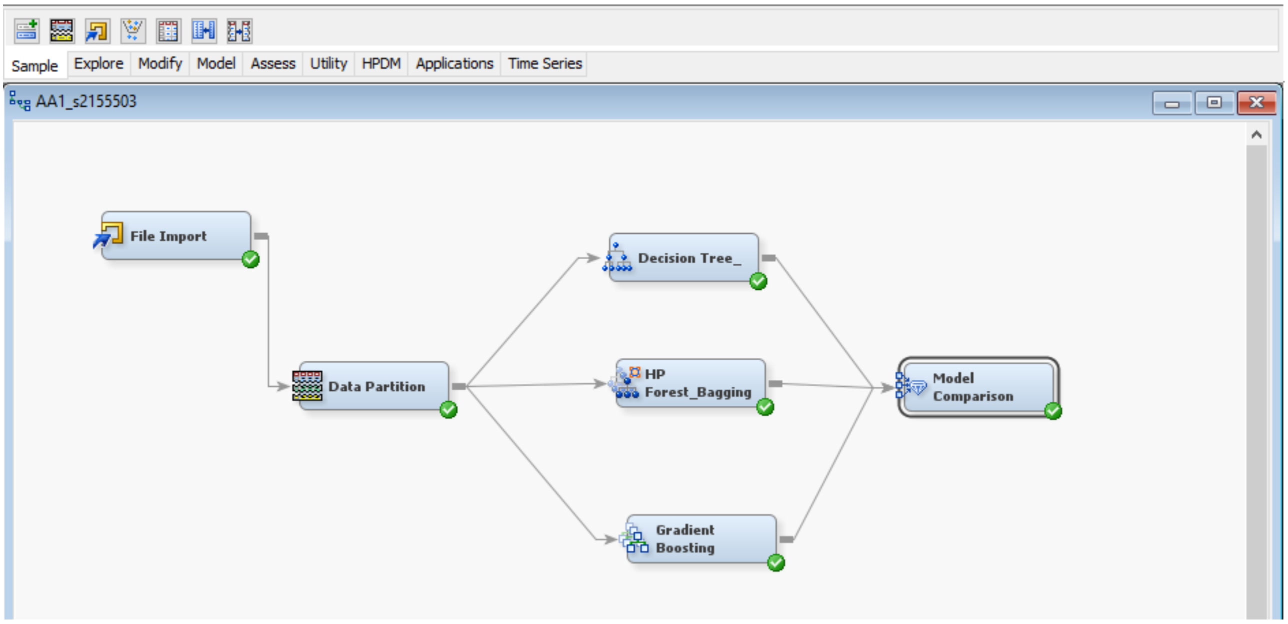
The customer behavior analysis, using decision tree modeling, indicates that the frequency of purchases ('Total Purchases') is the most critical factor in determining customer churn, with other variables such as 'Occupation' and 'Age' also playing significant roles. 'Total Purchases' is the principal variable used to split decisions in the predictive tree, suggesting thresholds in purchasing behavior are key indicators of churn risk. The variable importance scores reveal 'Age' as a more significant predictor in the validation set than in training, suggesting demographic factors may influence churn differently across datasets.

The fit statistics exhibit a moderate misclassification rate, with the model being more accurate on training data compared to validation data, hinting at potential overfitting. The Average Squared Error (ASE) and Root Average Squared Error (RASE) point to a moderate error level in predictions, with these errors slightly inflated in the validation set, indicating room for improvement in the model's predictive accuracy.

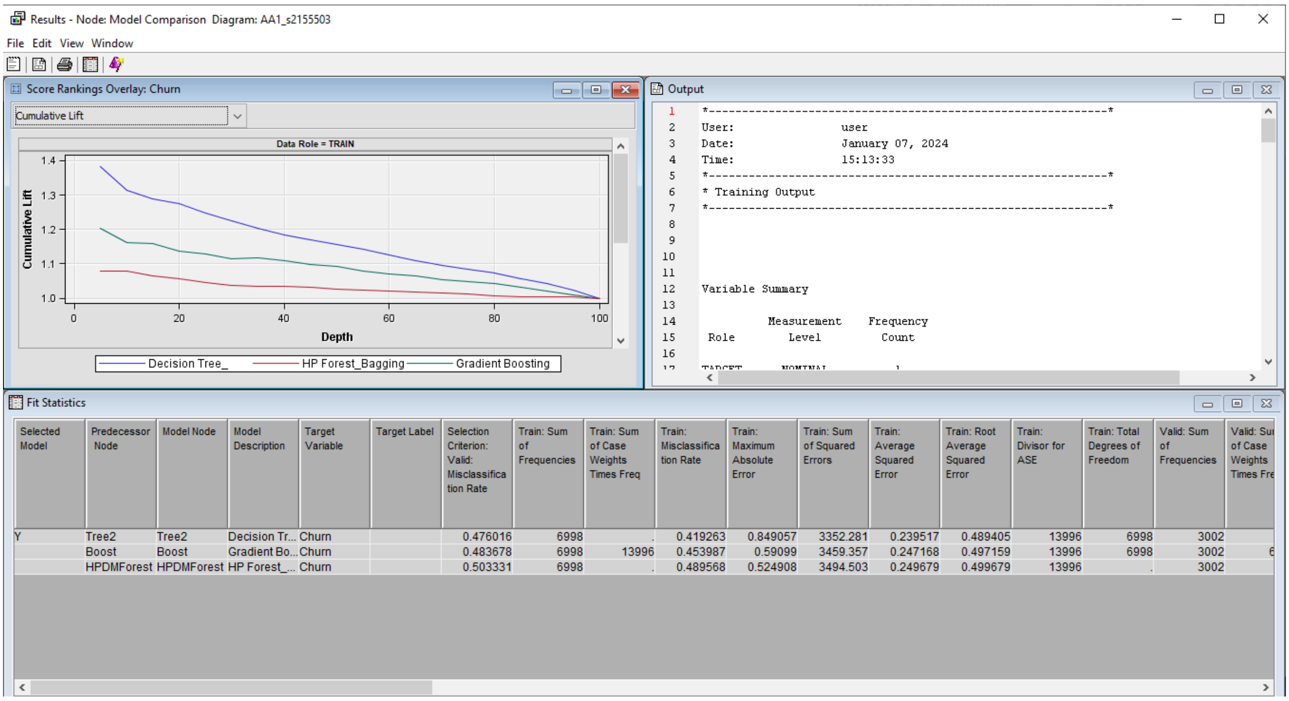
Overall, the analysis underscores the need to consider how various factors like purchasing frequency, occupation, and age interplay in influencing customer retention. There is also a suggestion of the need to address overfitting and enhance the model's generalization capabilities to better predict churn across different customer segments.

2.0 Tasks 3

Ensemble Methods: ***Apply Bagging and Boosting, using the Random Forest algorithm as a Bagging example.***



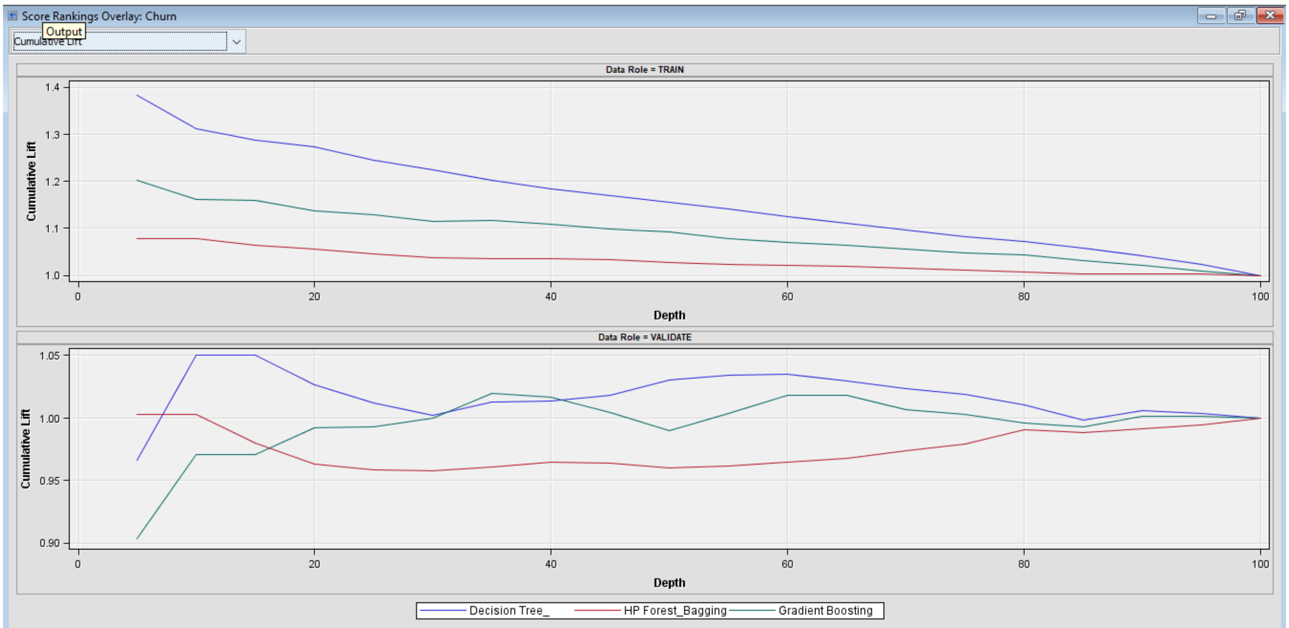
In this SAS Enterprise Miner workflow, I've constructed an ensemble methods approach to enhance predictive performance, building upon the foundation laid by the decision tree model established in the earlier task. The ensemble methodology is incorporating two components which are the HP Forest and Gradient Boosting. The HP Forest is for the bagging technique using the Random Forest algorithm, which aggregates multiple decision trees to reduce variance and improve stability. On the other hand, the Gradient Boosting component implements boosting, a method that sequentially builds models with each one focusing on the errors of the previous model to reduce bias. To systematically assess the efficacy of these models, the Model Comparison component is employed to leverages the collective strengths of bagging and boosting to potentially outperform the individual predictive capabilities of the models involved.



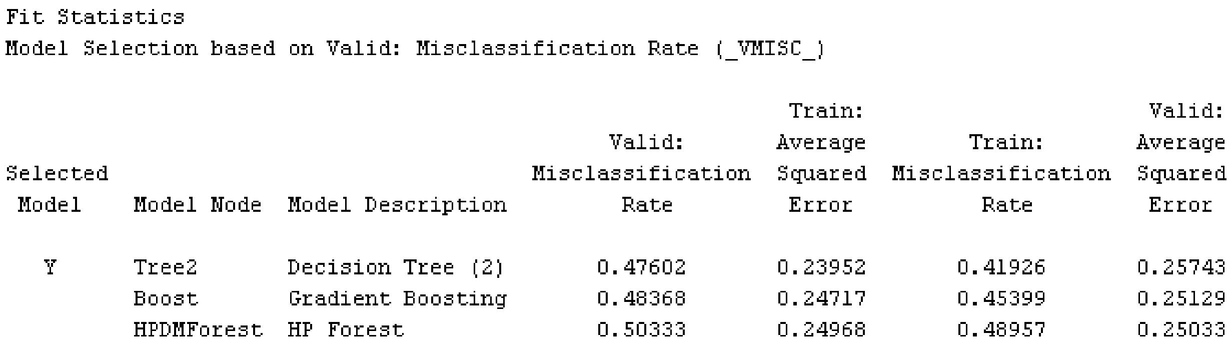
```
*------------------------------------------------------------*
User:                 user
Date:                 January 07, 2024
Time:                 15:13:33
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

           Measurement    Frequency
 Role         Level         Count

TARGET      NOMINAL           1
```

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Sum of Case Weights Times Freq	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	Valid: Sum of Frequencies	Valid: Sum of Case Weights Times Fre
Y	Tree2	Tree2	Decision Tr...	Churn		0.476016	6998	.	0.419263	0.849057	3352.281	0.239517	0.489405	13996	6998	3002	
	Boost	Boost	Gradient Bo...	Churn		0.483678	6998	13996	0.453987	0.59099	3459.357	0.247168	0.497159	13996	6998	3002	6
	HPDMForest	HPDMForest	HP Forest_...	Churn		0.503331	6998	.	0.489568	0.524908	3494.503	0.249679	0.499679	13996	.	3002	

The figure provides an overall result generated by the model comparison component. It encompasses of a score ranking overlay, fit statistics, and output file.



The chart displays the performance of three models which are Decision Tree, HP Forest (Bagging), and Gradient Boosting across different complexities, as indicated by depth. For the training data, Gradient Boosting consistently outperforms the other models, maintaining a higher cumulative lift across all depths. In contrast, the Decision Tree and HP Forest show a decreasing lift with increasing depth. On the validation set, the Decision Tree's performance drops as complexity grows, while the HP Forest and Gradient Boosting display more stable trends. Overall, Gradient Boosting seems to offer the best performance, particularly at higher depths, indicating its effectiveness in both training and validation scenarios.

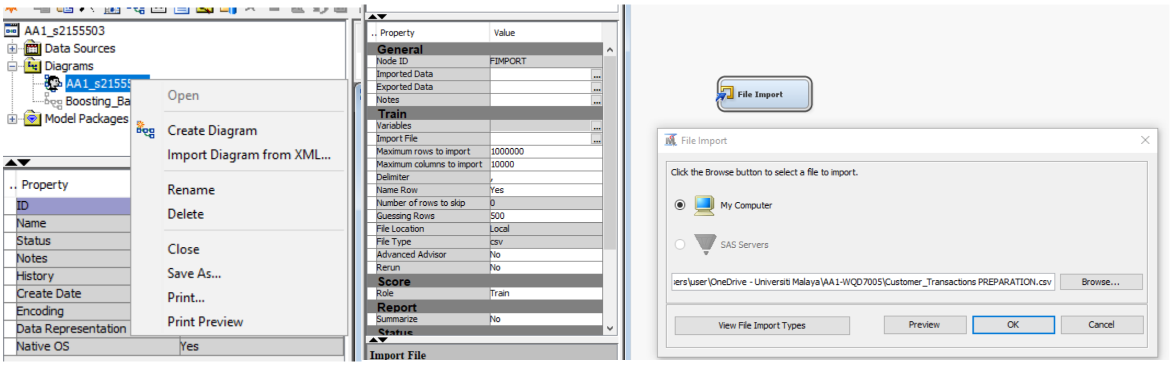


```
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                                                             Train:                         Valid:
                                              Valid:         Average        Train:          Average
Selected                                      Misclassification  Squared  Misclassification  Squared
 Model     Model Node  Model Description         Rate         Error           Rate          Error

   Y       Tree2       Decision Tree (2)       0.47602       0.23952        0.41926        0.25743
           Boost       Gradient Boosting       0.48368       0.24717        0.45399        0.25129
           HPDMForest  HP Forest               0.50333       0.24968        0.48957        0.25033
```

For the training set, the Decision Tree model has a misclassification rate of 0.41926 and an average squared error of 0.23952. The Gradient Boosting model shows a similar misclassification rate of 0.45399 but has a slightly better average squared error of 0.2417. The HP Forest model presents the highest misclassification rate among the three at 0.48957 with an average squared error of 0.24968. These statistics suggest that while the Decision Tree and Gradient Boosting models have a closer performance in terms of error, the HP Forest model is less accurate on the training data.

APPENDIX

Task 1



.. Property	
ID	
Name	
Status	
Notes	
History	
Create Date	
Encoding	
Data Representation	
Native OS	Yes

.. Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	...
Maximum rows to import	1000000
Maximum columns to import	10000
Delimiter	,
Name Row	Yes
Number of rows to skip	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Rerun	No
Score	
Role	Train
Report	
Summarize	No
Status	

Import File

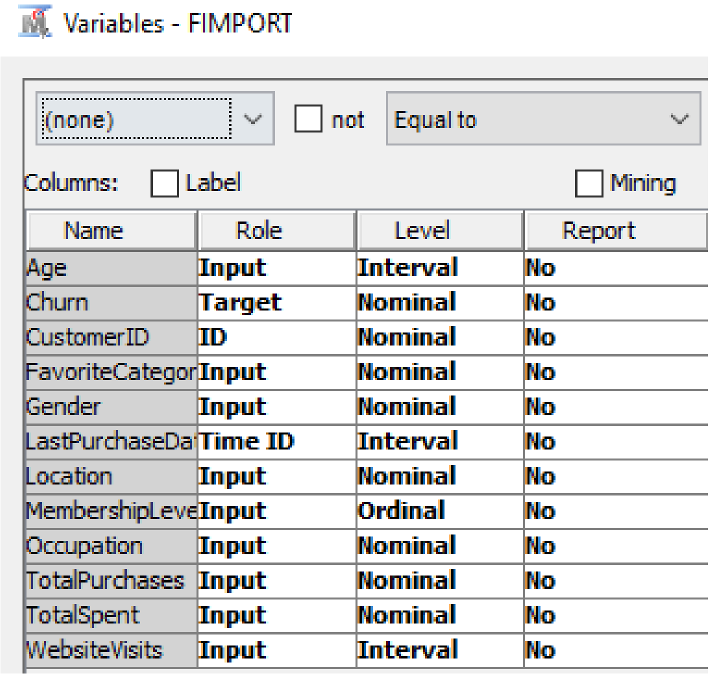
File Import

Click the Browse button to select a file to import.

● My Computer

○ SAS Servers

ers\user\OneDrive - Universiti Malaya\AA1-WQD7005\Customer_Transactions PREPARATION.csv

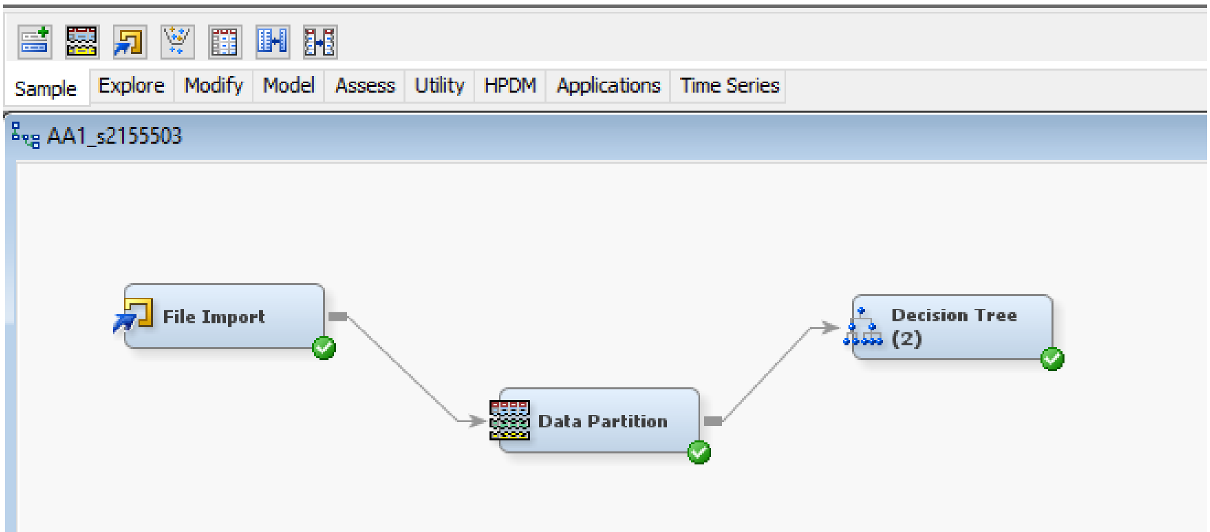

Variables - FIMPORT

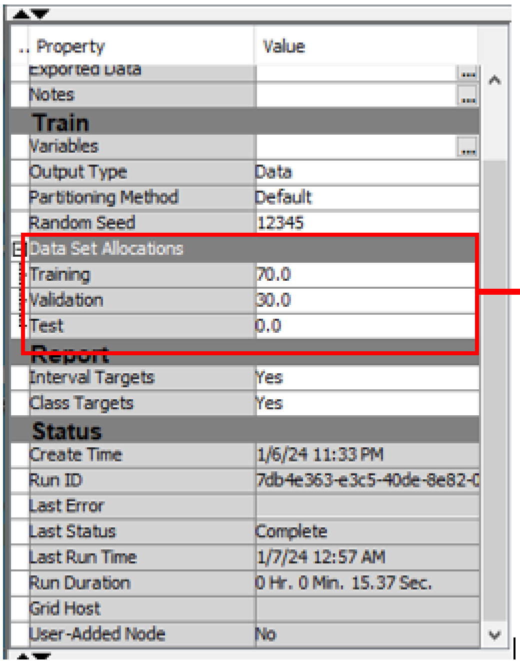
(none) □ not Equal to

Columns: □ Label □ Mining

Name	Role	Level	Report
Age	Input	Interval	No
Churn	Target	Nominal	No
CustomerID	ID	Nominal	No
FavoriteCategor	Input	Nominal	No
Gender	Input	Nominal	No
LastPurchaseDa	Time ID	Interval	No
Location	Input	Nominal	No
MembershipLeve	Input	Ordinal	No
Occupation	Input	Nominal	No
TotalPurchases	Input	Nominal	No
TotalSpent	Input	Nominal	No
WebsiteVisits	Input	Interval	No

Task 2

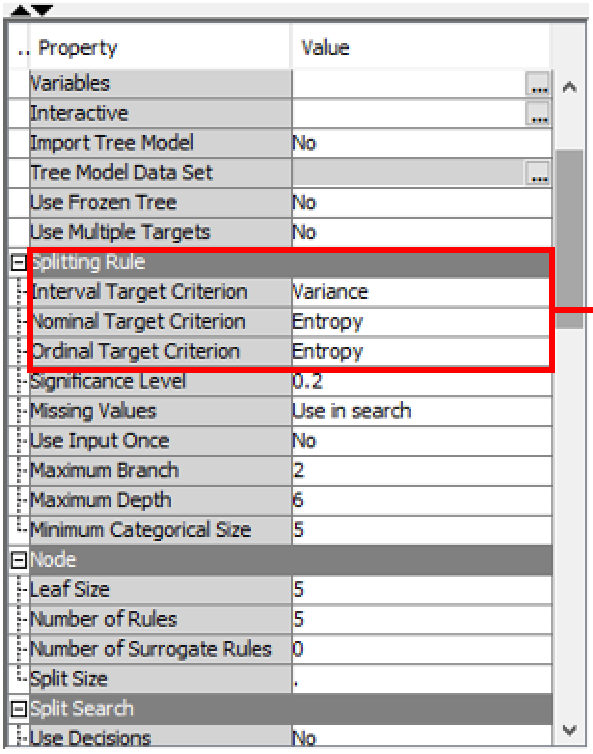




Property	Value
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/6/24 11:33 PM
Run ID	7db4e363-e3c5-40de-8e82-
Last Error	
Last Status	Complete
Last Run Time	1/7/24 12:57 AM
Run Duration	0 Hr. 0 Min. 15.37 Sec.
Grid Host	
User-Added Node	No

Training: 70

Validation: 30



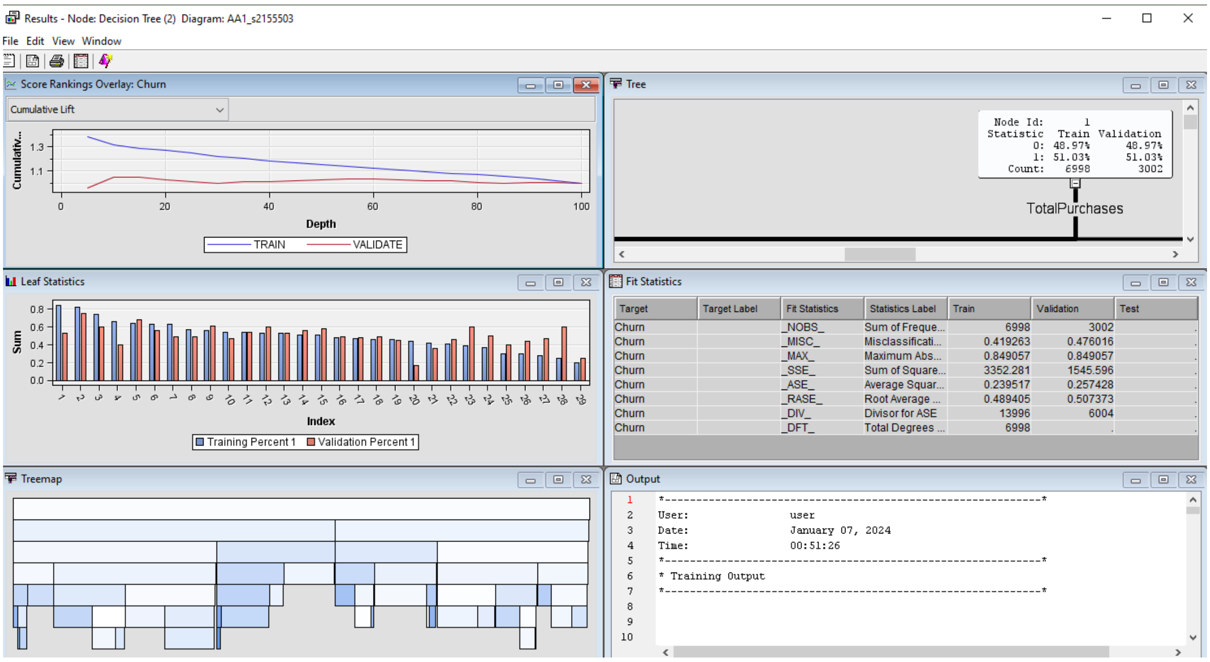
Property	Value
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	Variance
Nominal Target Criterion	Entropy
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No

Change the Splitting Rule

Interval Target Criterion: Variance

Nominal Target Criterion: Entropy

Ordinal Target Criterion: Entropy

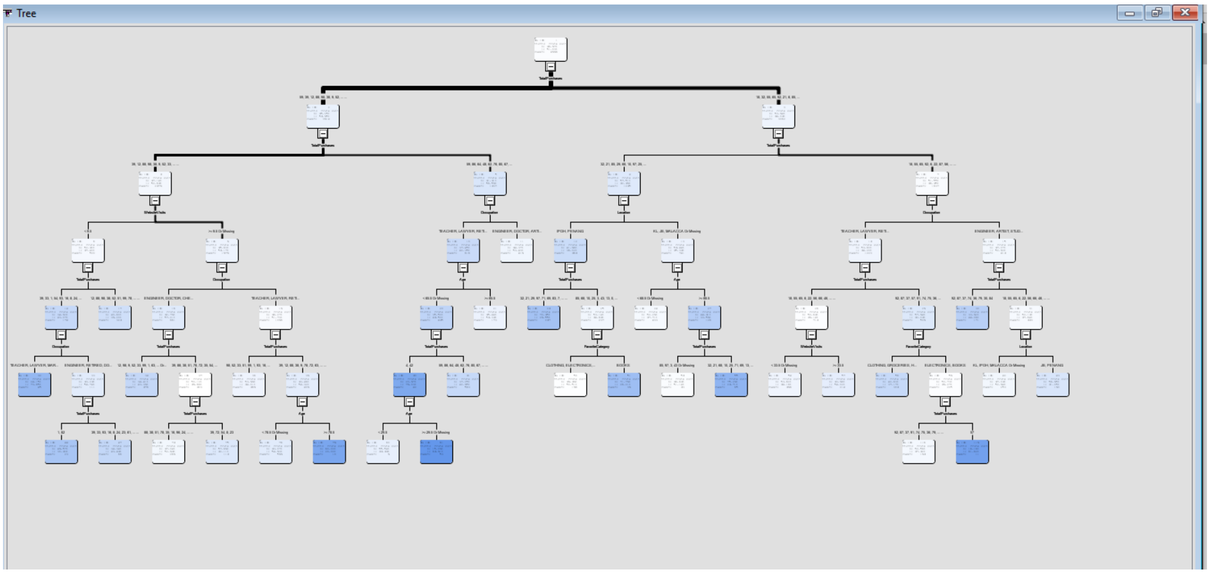


```
 Node Id:         1
Statistic   Train  Validation
       0:  48.97%      48.97%
       1:  51.03%      51.03%
   Count:    6998        3002
```

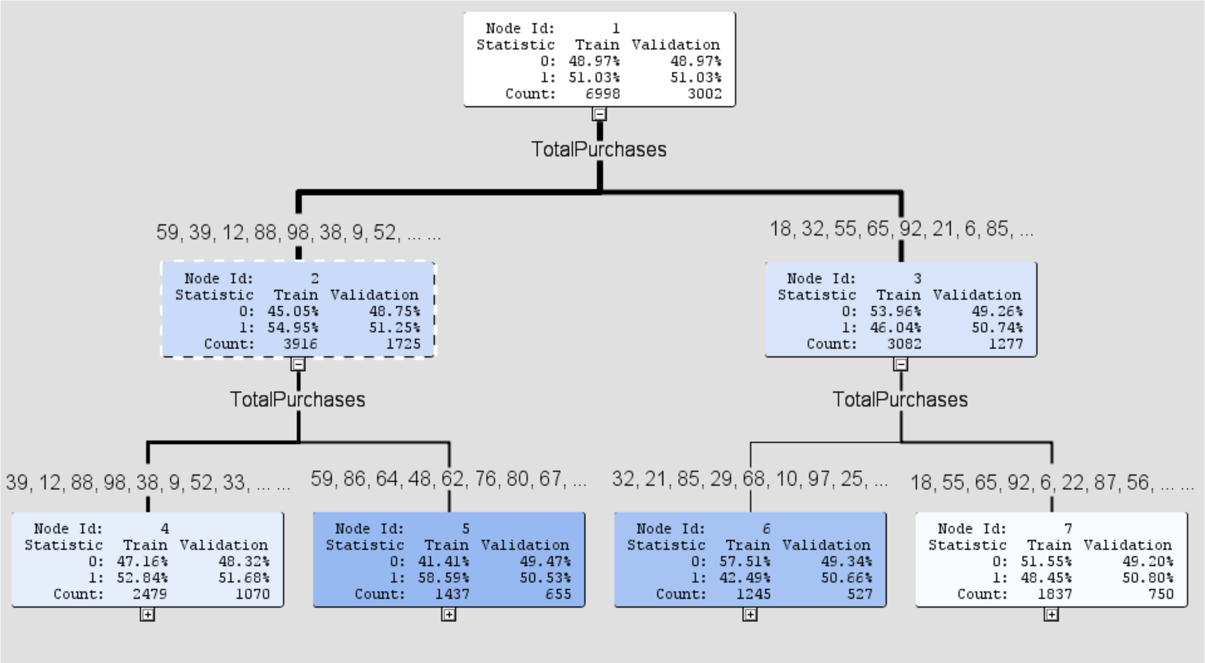
TotalPurchases

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Churn		_NOBS_	Sum of Freque...	6998	3002	.
Churn		_MISC_	Misclassificati...	0.419263	0.476016	.
Churn		_MAX_	Maximum Abs...	0.849057	0.849057	.
Churn		_SSE_	Sum of Square...	3352.281	1545.596	.
Churn		_ASE_	Average Squar...	0.239517	0.257428	.
Churn		_RASE_	Root Average ...	0.489405	0.507373	.
Churn		_DIV_	Divisor for ASE	13996	6004	.
Churn		_DFT_	Total Degrees ...	6998	.	.

```
*------------------------------------------------------------*
User:                user
Date:                January 07, 2024
Time:                00:51:26
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



```



```
Variable Importance

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

TotalPurchases                    14         1.0000        1.0000        1.0000
Occupation                         4         0.3809        0.4956        1.3014
Age                                4         0.3516        0.6659        1.8936
FavoriteCategory                   2         0.2623        0.0000        0.0000
WebsiteVisits                      2         0.2532        0.4962        1.9596
Location                           2         0.2510        0.5514        2.1966
```



```
Node Id:        1
Statistic   Train  Validation
       0:  48.97%      48.97%
       1:  51.03%      51.03%
   Count:   6998        3002
```

```
Node Id:        2
Statistic   Train  Validation
       0:  45.05%      48.75%
       1:  54.95%      51.25%
   Count:   3916        1725
```

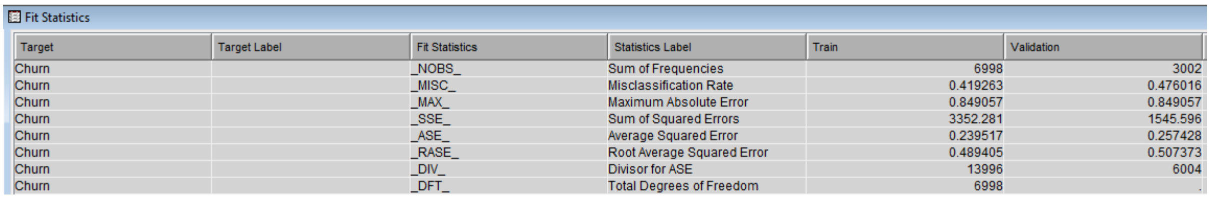
```
Node Id:        3
Statistic   Train  Validation
       0:  53.96%      49.26%
       1:  46.04%      50.74%
   Count:   3082        1277
```

```
Node Id:        4
Statistic   Train  Validation
       0:  47.16%      48.32%
       1:  52.84%      51.68%
   Count:   2479        1070
```

```
Node Id:        5
Statistic   Train  Validation
       0:  41.41%      49.47%
       1:  58.59%      50.53%
   Count:   1437         655
```

```
Node Id:        6
Statistic   Train  Validation
       0:  57.51%      49.34%
       1:  42.49%      50.66%
   Count:   1245         527
```

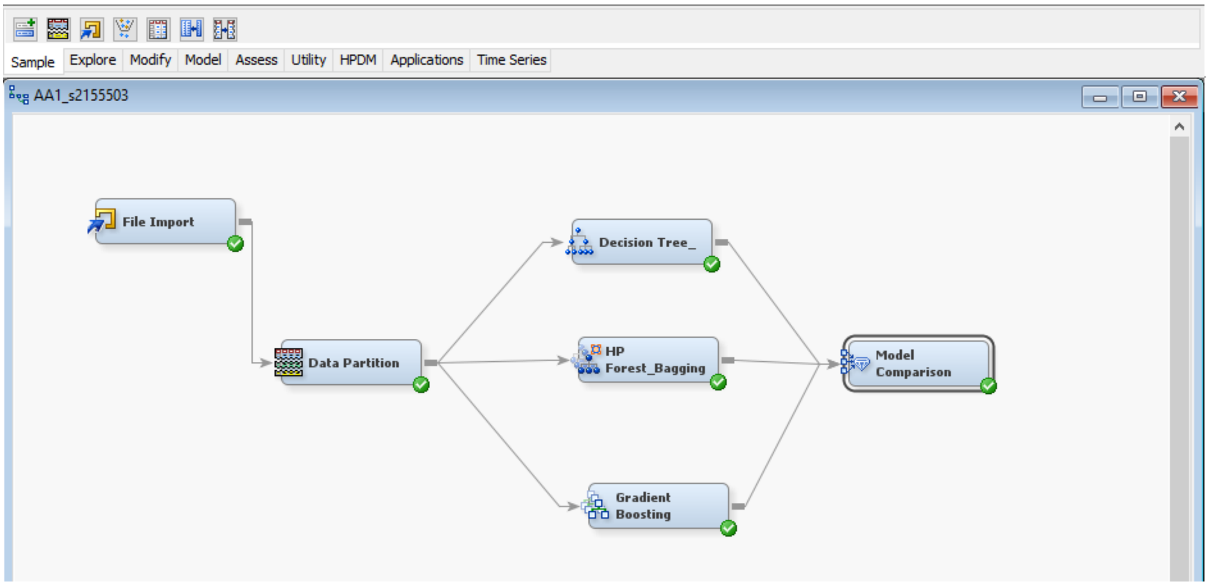
```
Node Id:        7
Statistic   Train  Validation
       0:  51.55%      49.20%
       1:  48.45%      50.80%
   Count:   1837         750
```

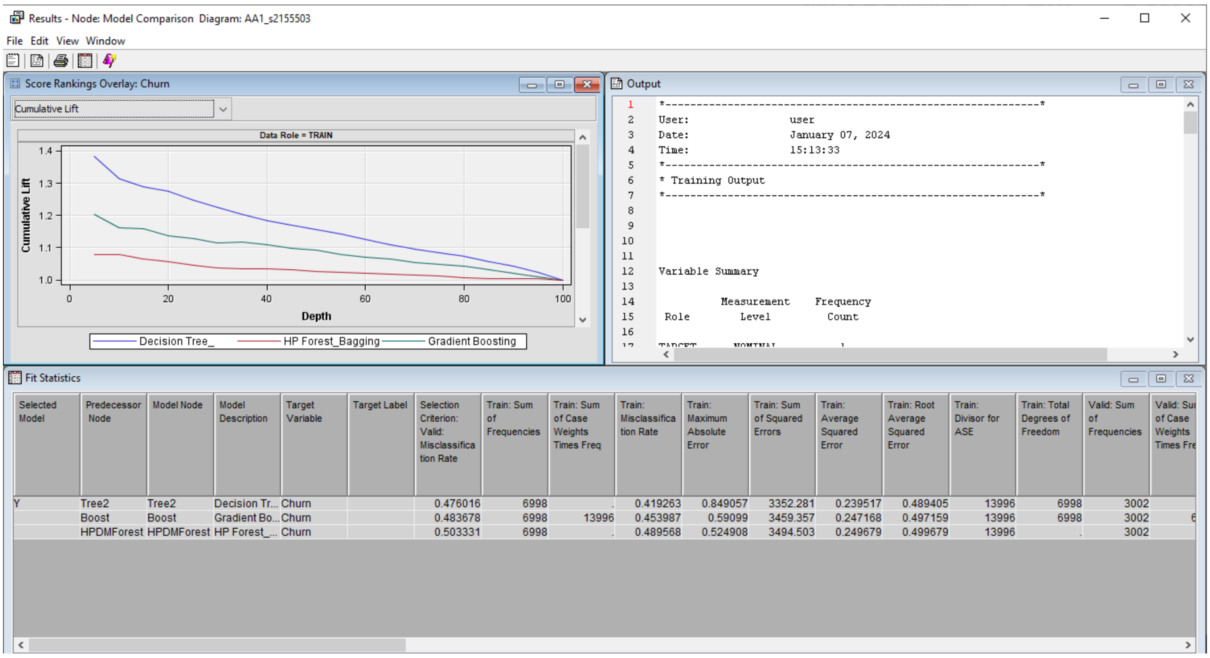


Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Churn		_NOBS_	Sum of Frequencies	6998	3002
Churn		_MISC_	Misclassification Rate	0.419263	0.476016
Churn		_MAX_	Maximum Absolute Error	0.849057	0.849057
Churn		_SSE_	Sum of Squared Errors	3352.281	1545.596
Churn		_ASE_	Average Squared Error	0.239517	0.257428
Churn		_RASE_	Root Average Squared Error	0.489405	0.507373
Churn		_DIV_	Divisor for ASE	13996	6004
Churn		_DFT_	Total Degrees of Freedom	6998	.

Task 3

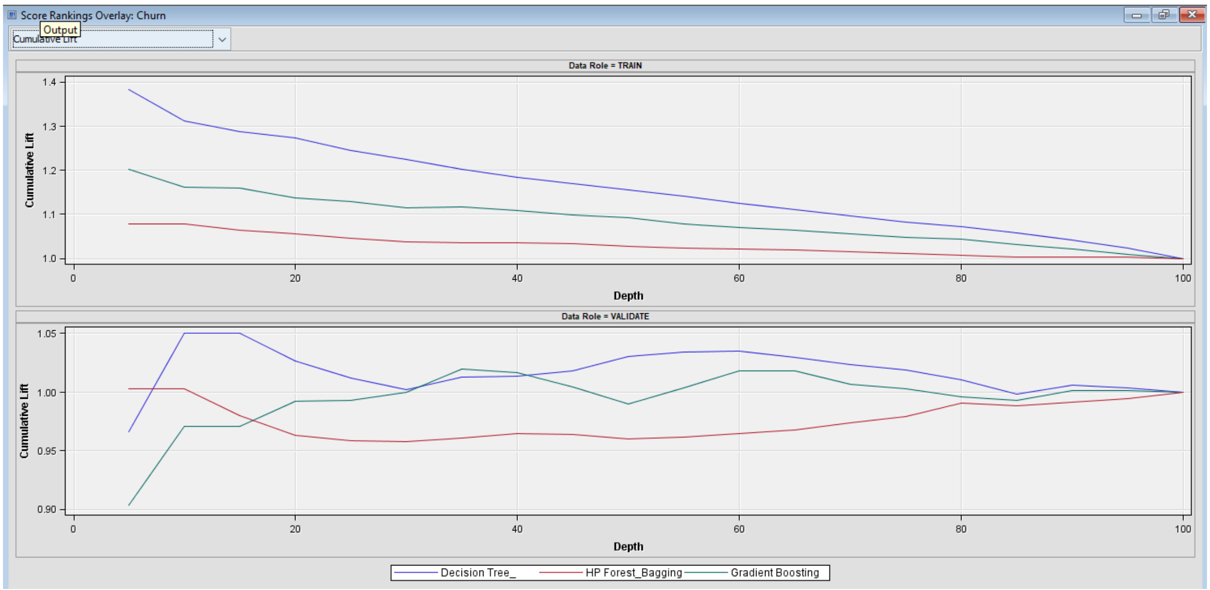




```
*-----------------------------------------------------------*
User:                user
Date:                January 07, 2024
Time:                15:13:33
*-----------------------------------------------------------*
* Training Output
*-----------------------------------------------------------*




Variable Summary

          Measurement    Frequency
 Role         Level        Count

TARGET      NOMINAL          1
```

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Sum of Case Weights Times Freq	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	Valid: Sum of Frequencies	Valid: Sum of Case Weights Times Fre
Y	Tree2	Tree2	Decision Tr...	Churn		0.476016	6998	.	0.419263	0.849057	3352.281	0.239517	0.489405	13996	6998	3002	
	Boost	Boost	Gradient Bo...	Churn		0.483678	6998	13996	0.453987	0.59099	3459.357	0.247168	0.497159	13996	6998	3002	6
	HPDMForest	HPDMForest	HP Forest_...	Churn		0.503331	6998	.	0.489568	0.524908	3494.503	0.249679	0.499679	13996	.	3002	



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

                                                            Train:                  Valid:
                                              Valid:        Average      Train:     Average
Selected                                 Misclassification  Squared Misclassification Squared
 Model    Model Node   Model Description       Rate          Error         Rate        Error

   Y      Tree2        Decision Tree (2)     0.47602        0.23952      0.41926     0.25743
          Boost        Gradient Boosting     0.48368        0.24717      0.45399     0.25129
          HPDMForest   HP Forest             0.50333        0.24968      0.48957     0.25033
```