ALTERNATIVE ASSESSMENT 1

Case Study: E-Commerce Customer Behaviour Analysis

Background:

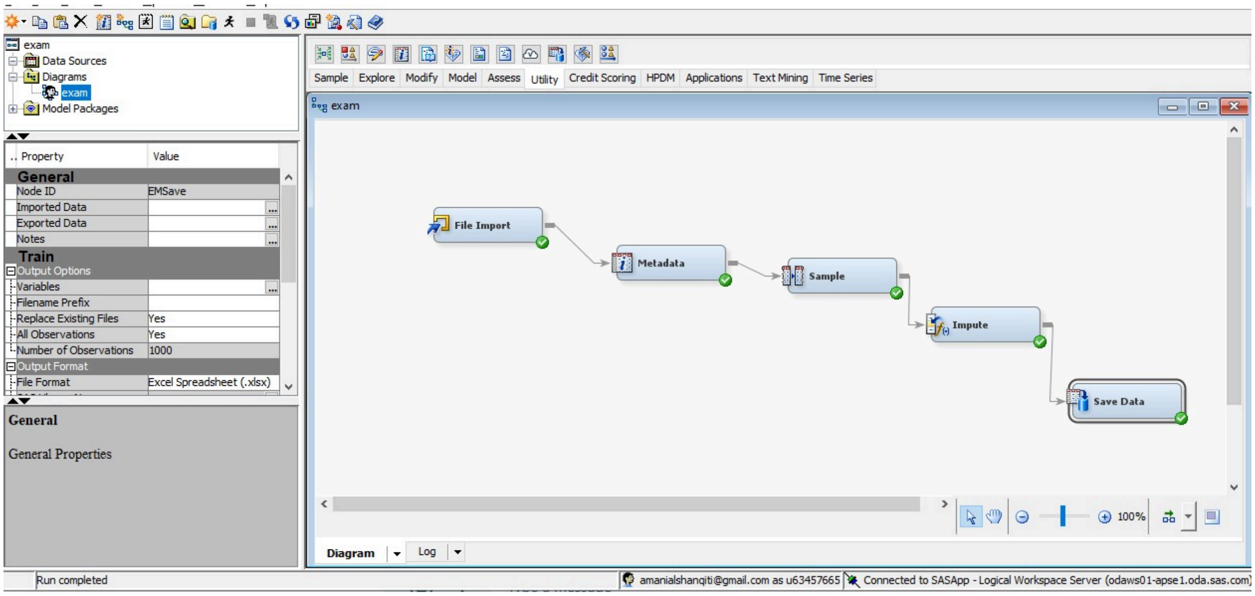
The foundation of this dataset was a pre-existing collection of customer transactions, which provided a robust starting point. To enhance its analytical value and, we have generated additional attributes, ensuring that our dataset more accurately reflects the nature of customer behavior in the e-commerce domain.

Dataset Structure:

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 (e.g., Electronics, Clothing, Home Goods). LastPurchaseDate: The date of the last purchase.
Churn: Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

In addition to these, these attributes were introduced to dataset:
TotalVisits: The total number of visits made by the customer.
PaymentMethod: Preferred payment method of the customer (e.g., Credit Card, PayPal), offering insights into payment preferences and potential trust levels with different payment modes.
DeviceUsedForShopping: The primary device used by the customer for shopping (e.g., Mobile, Desktop, Tablet), helping us understand shopping behaviors across different devices.

TotalVisits: provides insights into how engaged customers are with the e-commerce platform which can be helpful for the analysis.
Geographic data (Location): Helps in regional analysis and understanding location-based trends.
Device and Payment Method: Provides a deeper understanding of the technological and financial preferences of the customers, which are critical in today's digital shopping era.



- Metadata

After importing the data this step is crucial to ensure that for the analysis of the data correctly. Variables such as Churn also being assign as the target variable for predictive modeling purposes.

```
Variable Summary

          Measurement    Frequency
Role         Level         Count

ID          INTERVAL         1
INPUT       INTERVAL         4
INPUT       NOMINAL          7
TARGET      INTERVAL         1
```

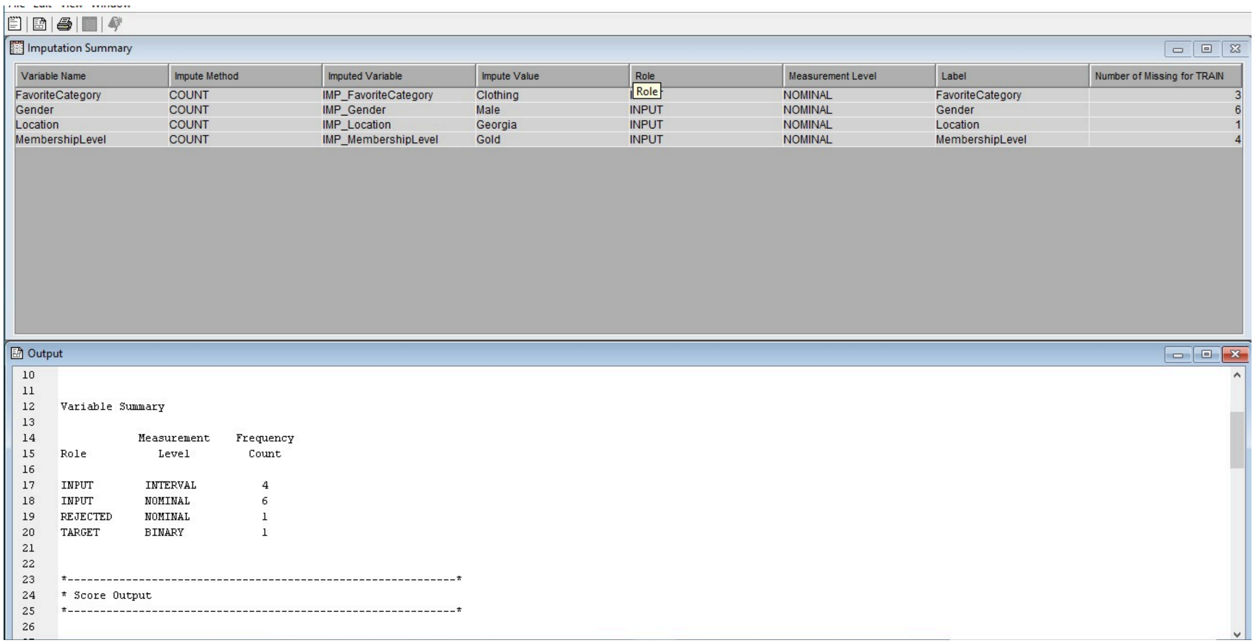
```
Sampling Summary

                                  Number of
 Type          Data Set          Observations

DATA       EMWS1.Meta_TRAIN         49673
SAMPLE     EMWS1.Smpl_DATA           4967
```

-sampling

The sampling technique chosen for this study is probability sampling or to be precise, stratified sampling. This sampling technique is the process where a sample is selected, and each stratum of the population is represented. 10% of data were sampled to 4967 rows.



Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
FavoriteCategory	COUNT	IMP_FavoriteCategory	Clothing	Role	NOMINAL	FavoriteCategory	3
Gender	COUNT	IMP_Gender	Male	INPUT	NOMINAL	Gender	6
Location	COUNT	IMP_Location	Georgia	INPUT	NOMINAL	Location	1
MembershipLevel	COUNT	IMP_MembershipLevel	Gold	INPUT	NOMINAL	MembershipLevel	4

Output

```
Variable Summary

             Measurement    Frequency
Role            Level         Count

INPUT         INTERVAL          4
INPUT         NOMINAL           6
REJECTED      NOMINAL           1
TARGET        BINARY            1


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

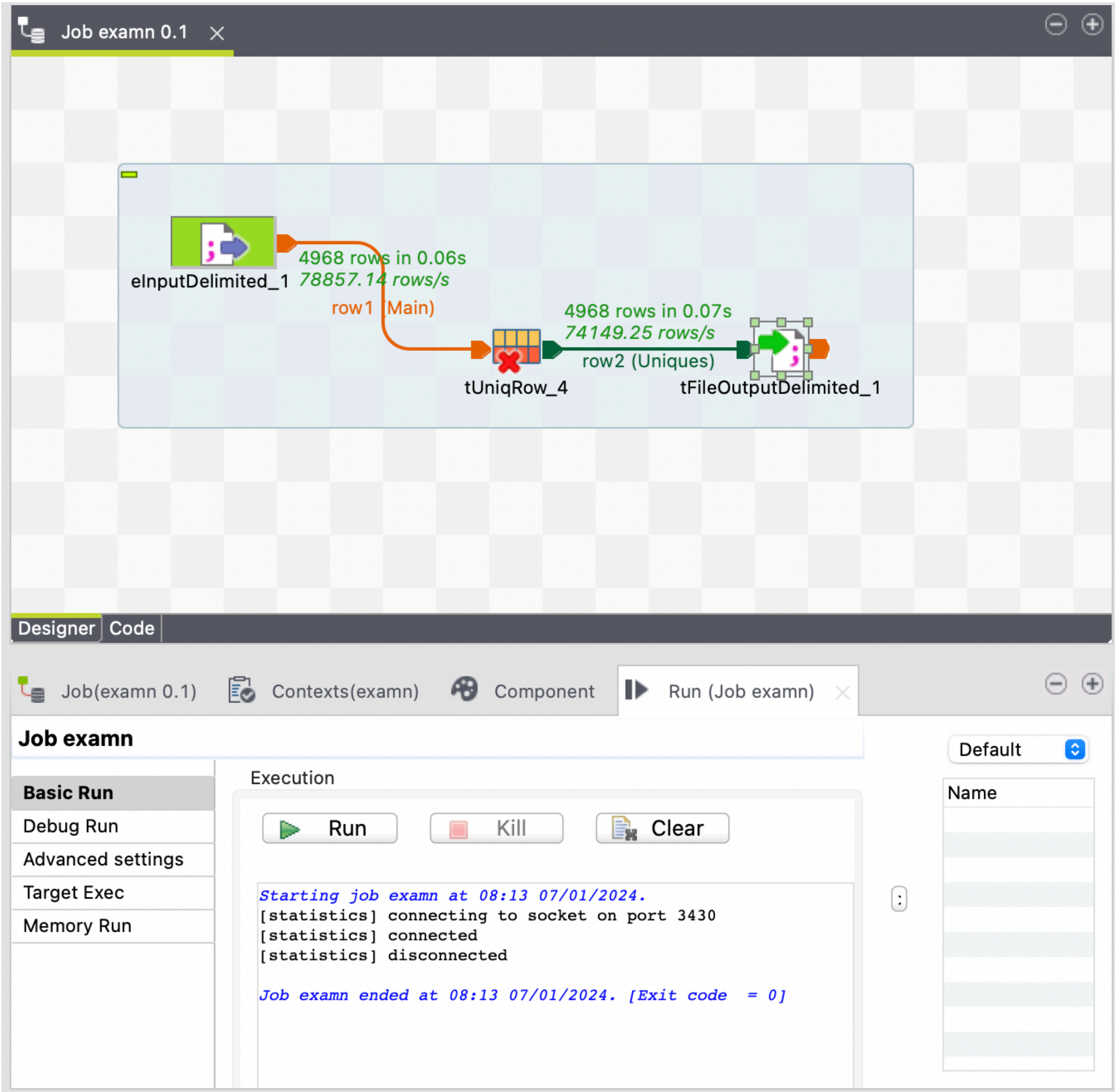
- Handling Missing Values

Missing values within the dataset were addressed using two distinct methods depending on the nature of the data:

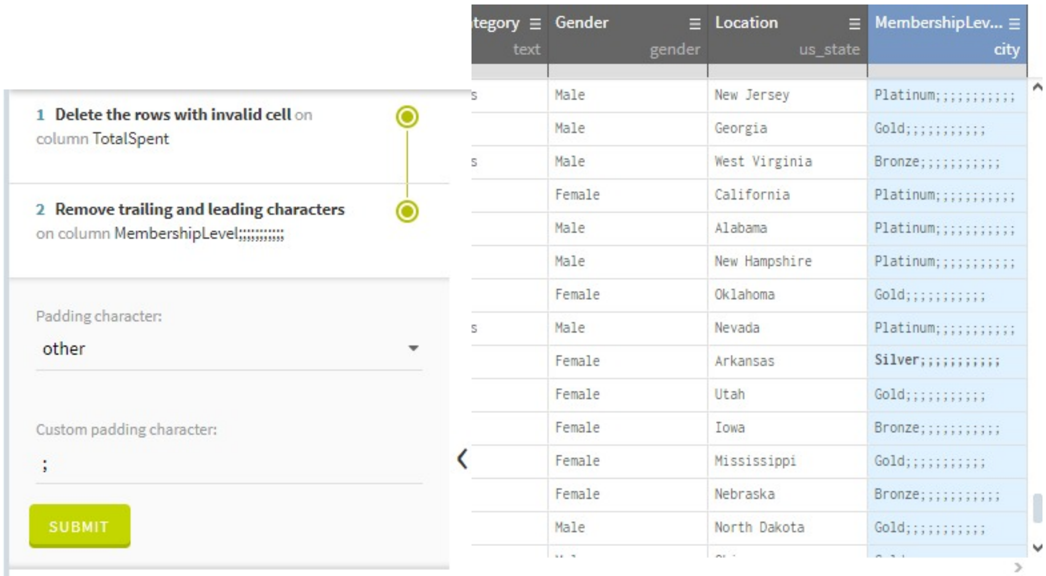
- Categorical Data: Missing values were imputed using the mode, which is the most frequent category within the data.
- Numerical Data: Missing values were imputed using the mean of the available values, providing a central tendency measure for the imputation.

There were no Numerical missing values.

-Cleaning Data



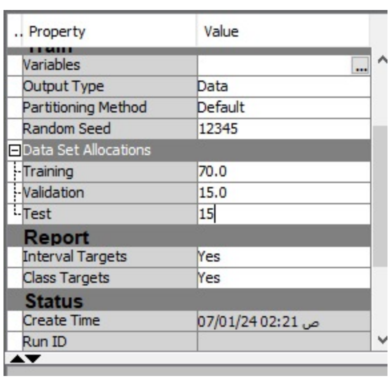
- Making sure that there are no duplicate rows in the data using Talend and saving the result.



- Cleaning Error in MembershipLevel data using Talend data preparation.

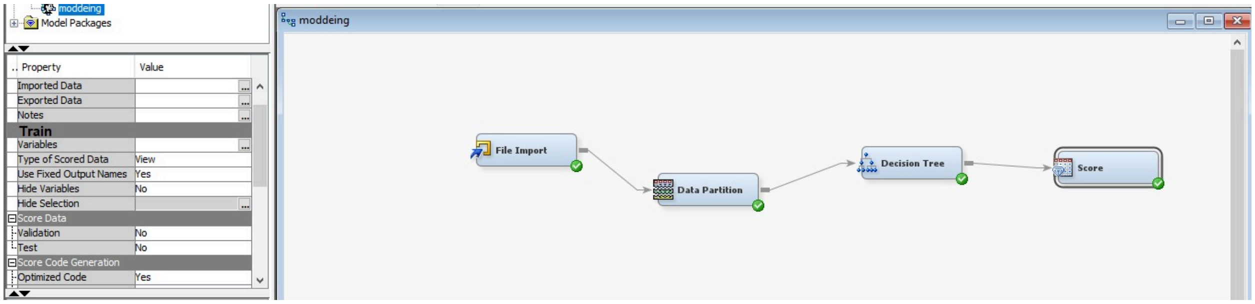
Data modeling:

Splitting data for modeling using 15% validation 15% testing and 70% training

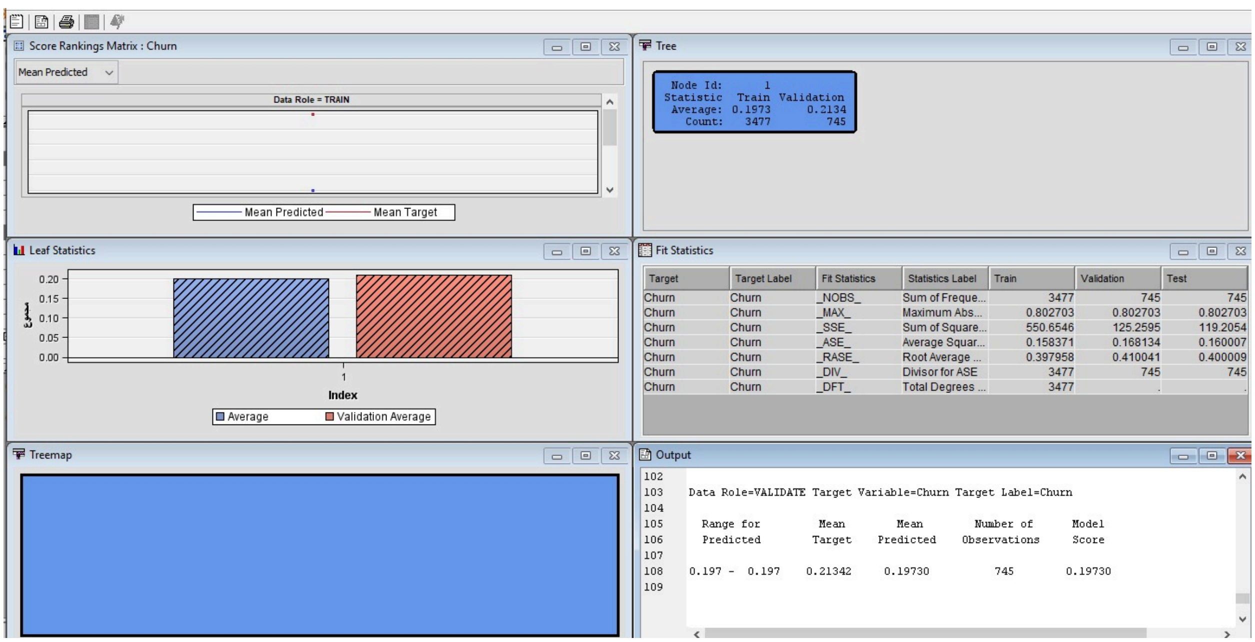


Property	Value
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	15.0
Test	15
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	07/01/24 02:21 ص
Run ID	

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

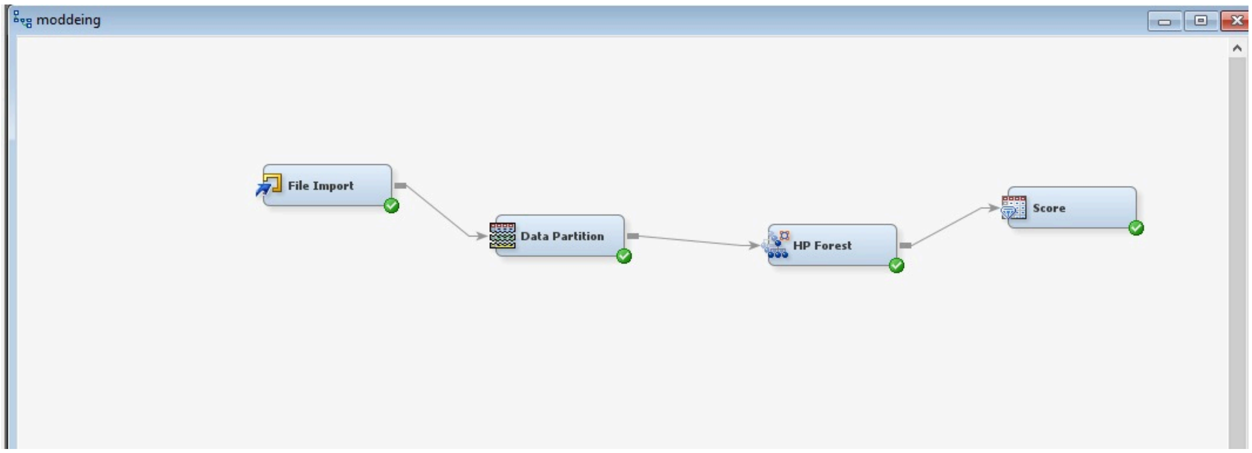


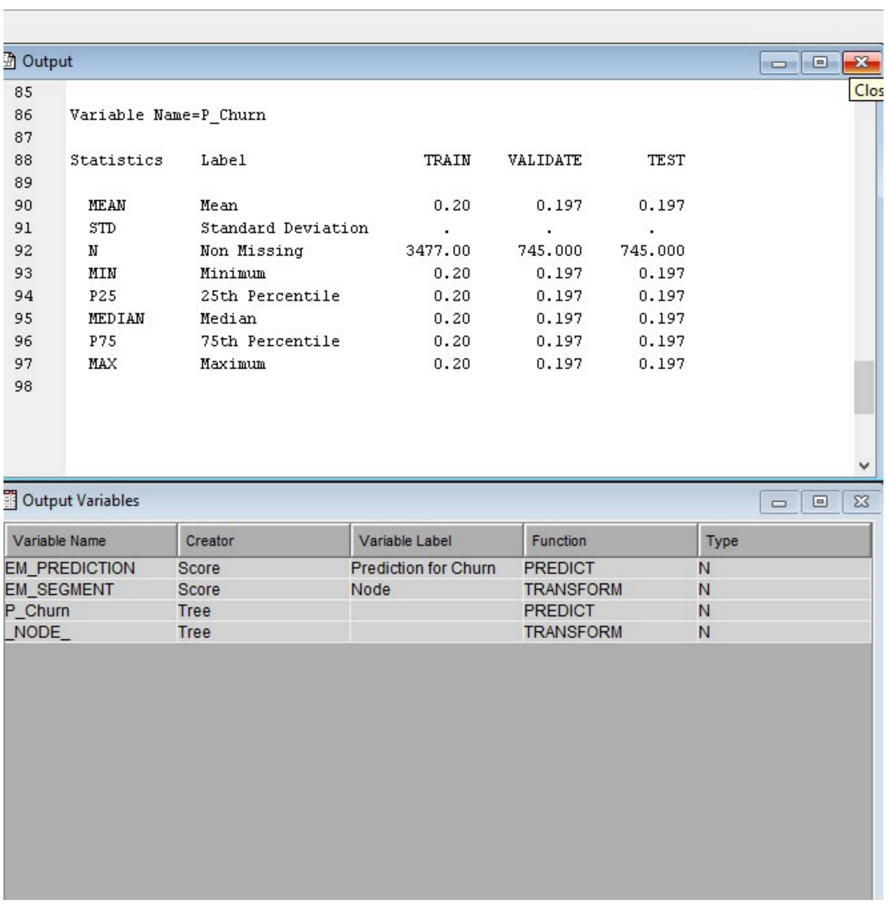
- The mean target is higher in the validation set (0.21342) than in the training set (0.19730)a higher churn rate in the validation data.
- The Average Squared Error (ASE) is slightly lower for the validation (0.168134) compared to the training (0.158371), which is promising as it indicates the model is not overfitting.
- The Root Average Squared Error (RASE) is consistent between training (0.397858) and validation (0.410041), which again suggests the model is generalizing well.



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

- **Bagging using Random Forest**





Output

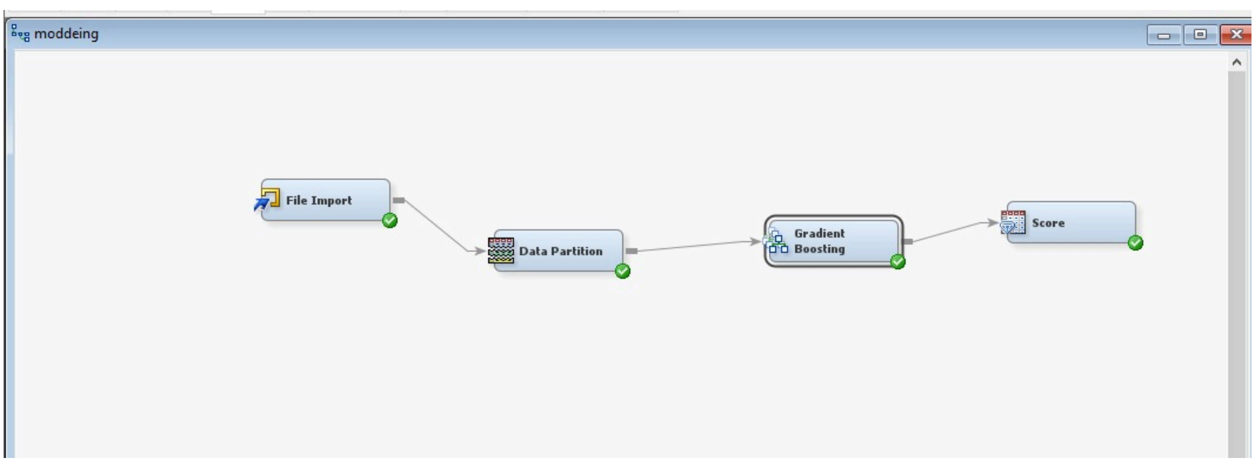
```
85
86  Variable Name=P_Churn
87
88  Statistics    Label                  TRAIN    VALIDATE      TEST
89
90    MEAN        Mean                    0.20       0.197     0.197
91    STD         Standard Deviation       .          .         .
92    N           Non Missing          3477.00     745.000   745.000
93    MIN         Minimum                 0.20       0.197     0.197
94    P25         25th Percentile         0.20       0.197     0.197
95    MEDIAN      Median                  0.20       0.197     0.197
96    P75         75th Percentile         0.20       0.197     0.197
97    MAX         Maximum                 0.20       0.197     0.197
98
```

Output Variables

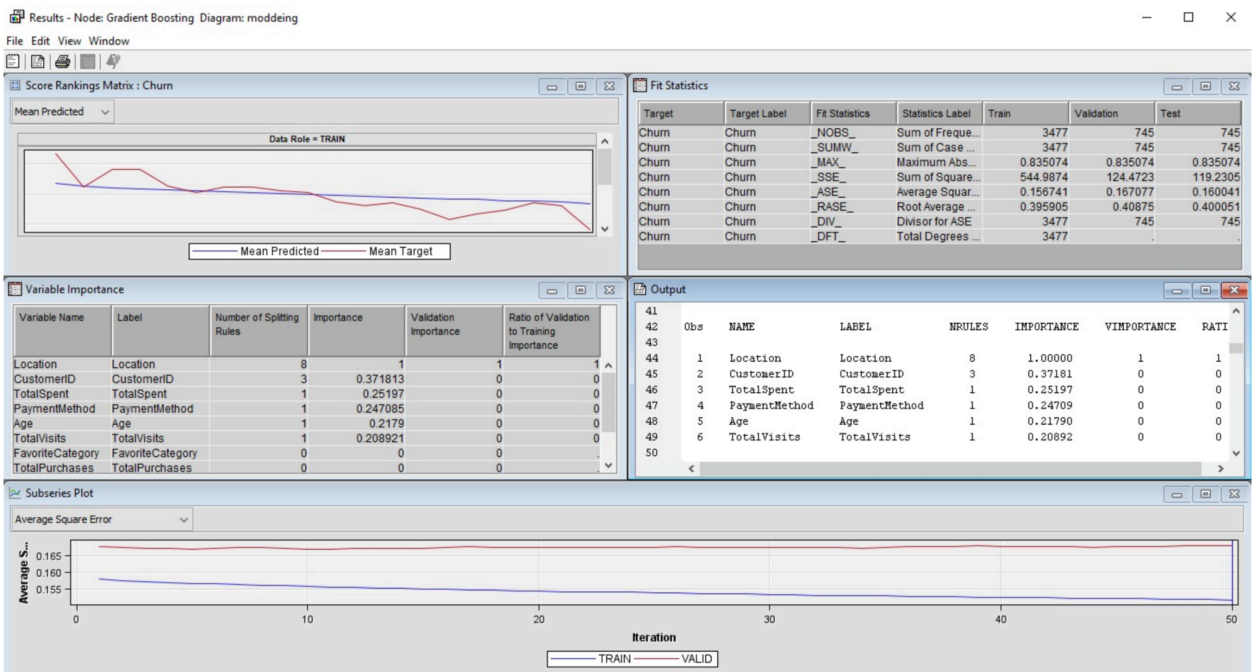
Variable Name	Creator	Variable Label	Function	Type
EM_PREDICTION	Score	Prediction for Churn	PREDICT	N
EM_SEGMENT	Score	Node	TRANSFORM	N
P_Churn	Tree		PREDICT	N
NODE	Tree		TRANSFORM	N

- The Average Square Error (ASE) is low at 0.158 for training and 0.168 for validation.
- The Root Average Squared Error (RASE) is 0.4 for training and 0.41 for validation.
- The model used 26 trees to achieve these statistics, with an inbag fraction of 0.6.

- **Boosting using Gradient boost model**

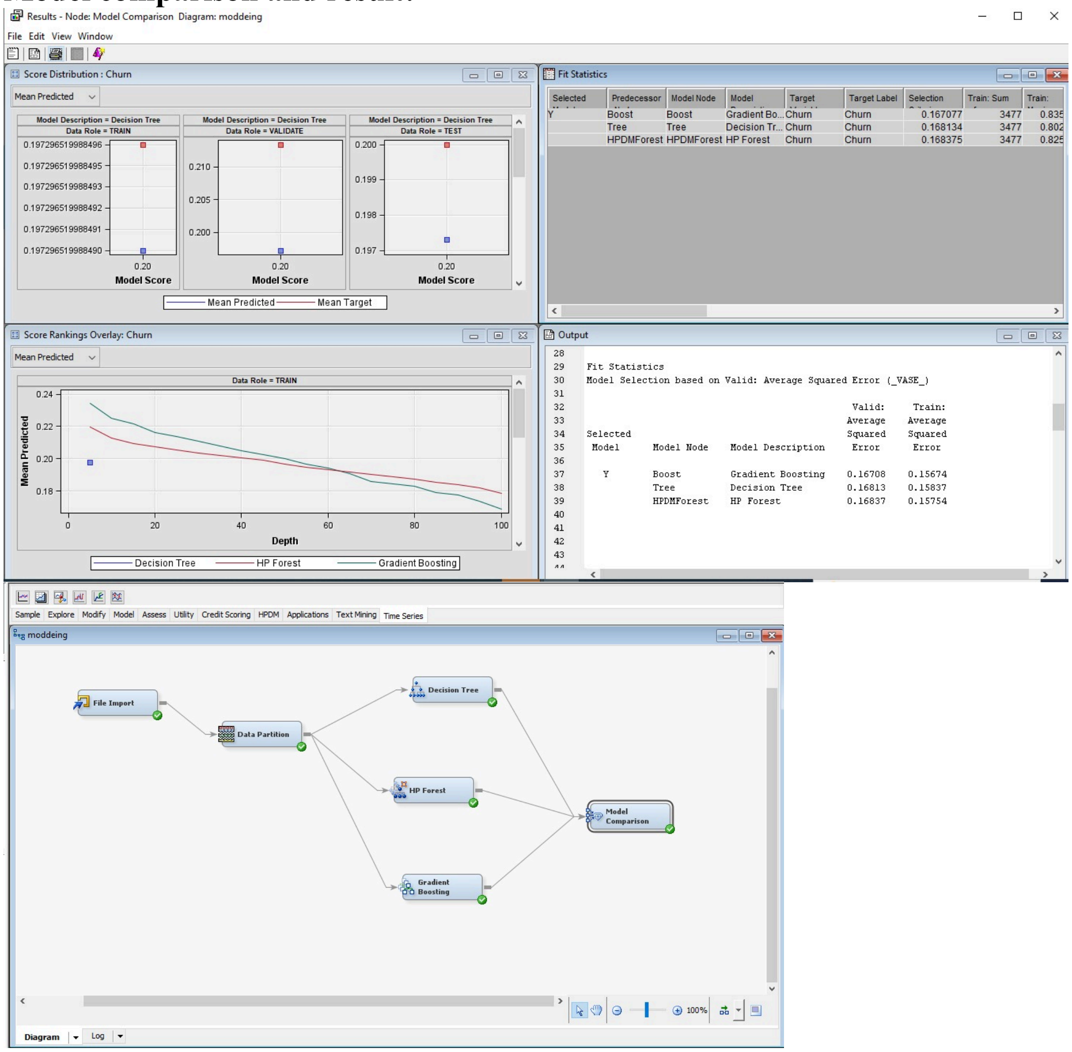


```
63     Fit
64  Statistics     Statistics Label                     Train
     Validation        Test
65
66     _NOBS_      Sum of Frequencies                 3477.00
      745.000       745.000
67     _SUMW_      Sum of Case Weights Times Freq     3477.00
      745.000       745.000
68     _MAX_       Maximum Absolute Error                0.84
        0.835         0.835
69     _SSE_       Sum of Squared Errors               544.99
      124.472       119.230
70     _ASE_       Average Squared Error                 0.16
        0.167         0.160
71     _RASE_      Root Average Squared Error            0.40
        0.409         0.400
72     _DIV_       Divisor for ASE                    3477.00
      745.000       745.000
73     _DFT_       Total Degrees of Freedom           3477.00
          .             .
74
75
76
```



- The Average Squared Error (ASE) is 0.16 for the training dataset and 0.167 for the validation dataset.
- The Root Average Squared Error (RASE) for the training dataset is 0.40 and for the validation dataset is 0.409.

Model comparison and result:



The comparative analysis of three predictive models—Gradient Boosting, Decision Tree, and HP Forest—reveals a closely matched performance in predicting customer churn. Each model shows a consistent Average Squared Error (ASE) of 0.16 in training. In validation, Gradient Boosting slightly outperforms with an ASE of 0.16708, followed by Decision Tree at 0.16813, and HP Forest at 0.16837. Testing ASE remains constant at 0.160 across all models. This near parity suggests that any of the three could be deployed for churn prediction with similar expected accuracy.

In this document, I conducted a thorough analysis of e-commerce customer behavior using a detailed dataset. My approach involved preprocessing the data, implementing statistical techniques, and deploying various predictive models like Gradient Boosting, Decision Tree, and HP Forest to analyze customer churn. The study highlighted the importance of handling missing values and employing probability sampling for more accurate insights. The models tested showed close performance in predicting customer churn, underscoring their efficacy in e-commerce customer behavior analysis. This work not only provided valuable insights into customer behavior but also enhanced my skills in data analysis and model implementation.