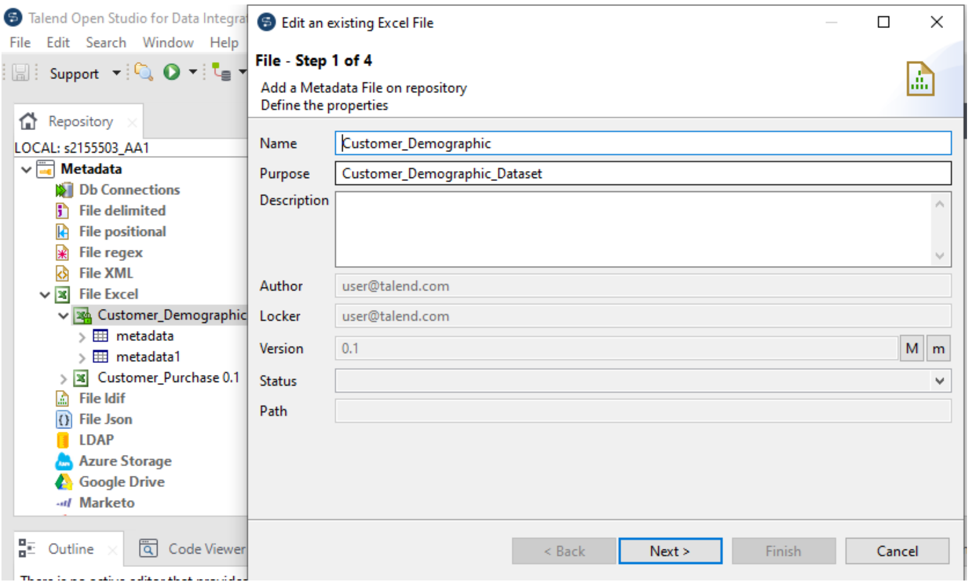
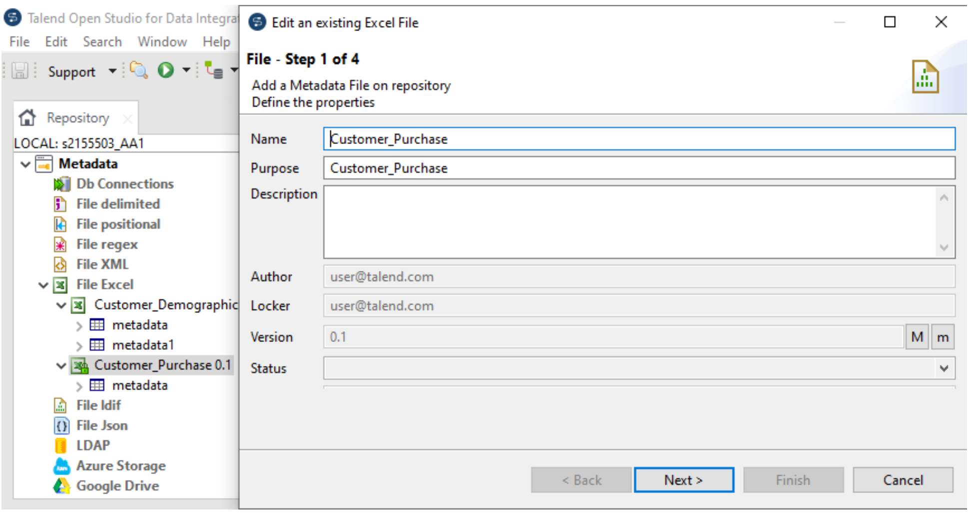
## RESULTS AND ANALYSIS

### Data Integration

I utilized the Talend data integration tool to carefully combine these datasets, ensuring the seamless merging of information while preserving accuracy and integrity.

Metadata creation for both Excel datasets: -




Edit Scheme and identify correct Data Type

Schema

Click Guess button to update the schema below according to your settings

Guess

Description of the Schema

| Column | Key | Type | ☑ N.. | Date Pattern (Ctrl+S... | Length | Precision | Default | Comment |
|---|---|---|---|---|---|---|---|---|
| CustomerID | ☐ | Integer | ☑ | | | 0 | | |
| Age | ☐ | Integer | ☑ | | | 0 | | |
| Gender | ☐ | String | ☑ | | 0 | 0 | | |
| Location | ☐ | String | ☑ | | 0 | 0 | | |
| MembershipLevel | ☐ | String | ☑ | | 0 | 0 | | |
| Occupation | ☐ | String | ☑ | | 0 | 0 | | |
| WebsiteVisitFrequency | ☐ | String | ☑ | | 0 | 0 | | |

Schema

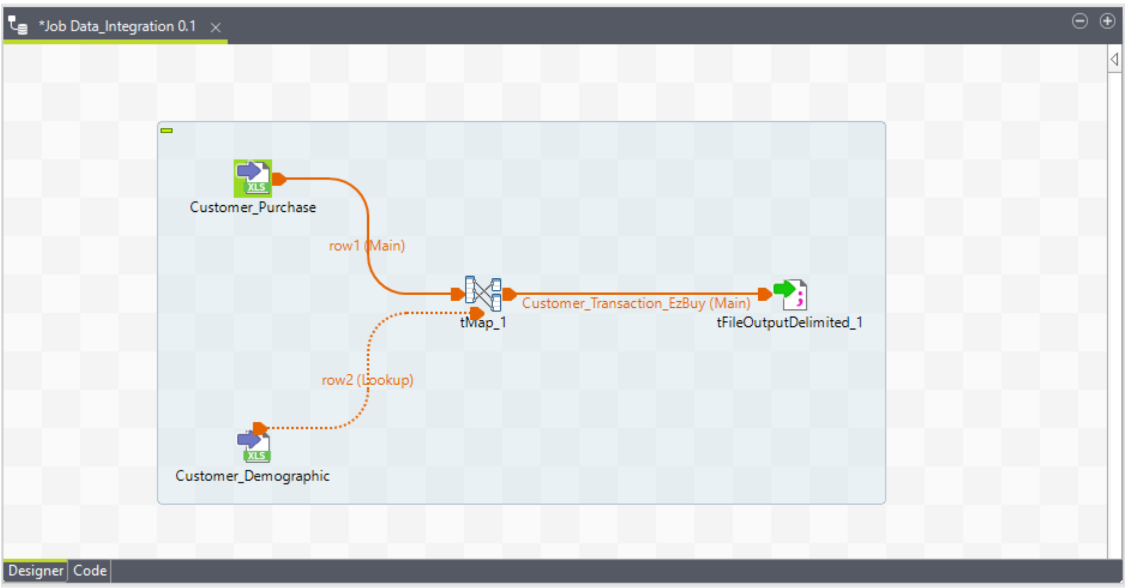
Click Guess button to update the schema below according to your settings

Guess

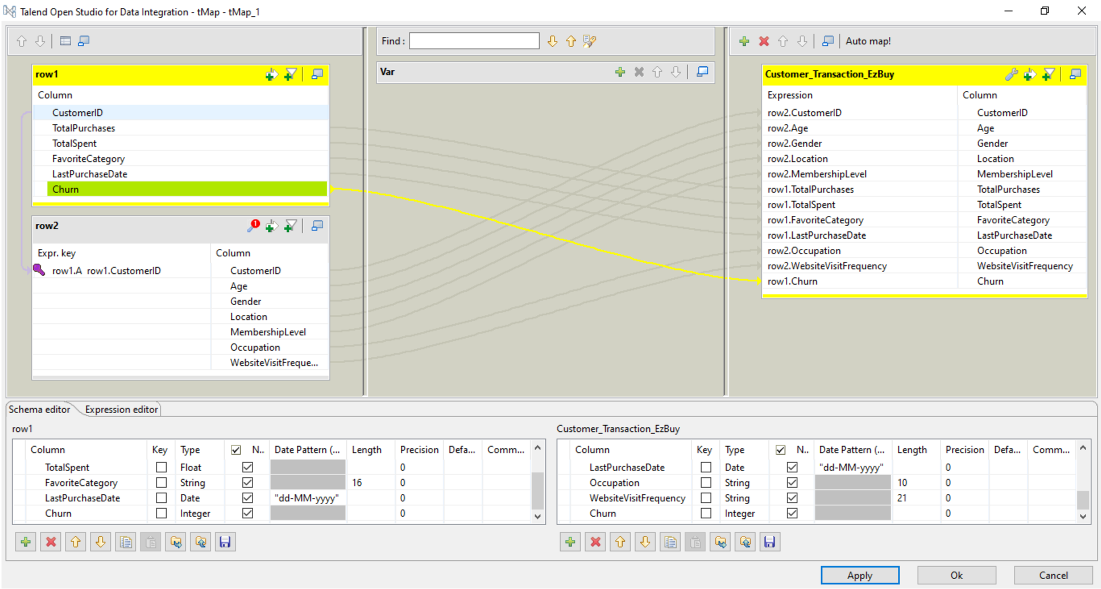
Description of the Schema

| Column | Key | Type | ☑ N.. | Date Pattern (Ctrl+S... | Length | Precision | Default | Comment |
|---|---|---|---|---|---|---|---|---|
| CustomerID | ☐ | Integer | ☑ | | | 0 | | |
| TotalPurchases | ☐ | Integer | ☑ | | | 0 | | |
| TotalSpent | ☐ | Float | ☑ | | | 0 | | |
| FavoriteCategory | ☐ | String | ☑ | | 16 | 0 | | |
| LastPurchaseDate | ☐ | Date | ☑ | "dd-MM-yyyy" | | 0 | | |
| Churn | ☐ | Integer | ☑ | | | 0 | | |

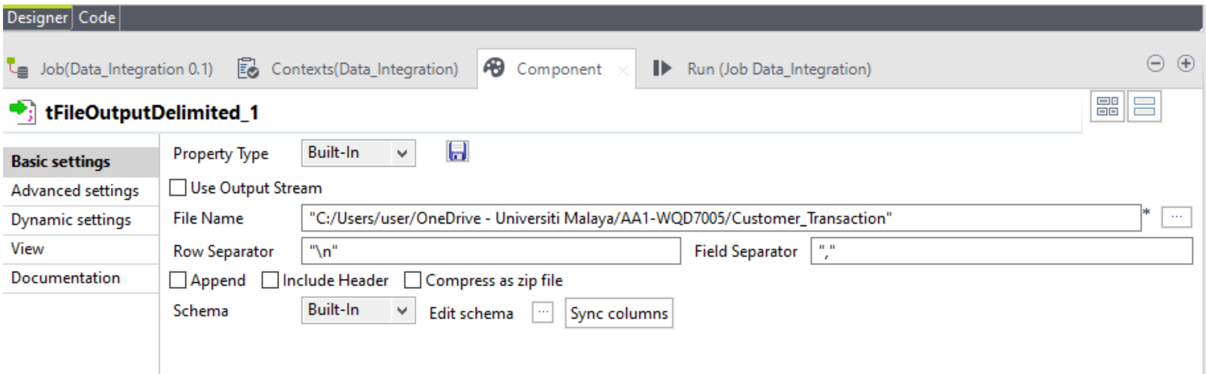
Set Up component's diagram.



In this process, I used four key components. First, source ExcelfileInput dragged from both metadata. Second component called TMap to connect the input data to the output data. Primarily serving the purpose of mapping input data to output data. This involves the transformation of one schema into another, ensuring seamless data integration. Lastly, TOutputDelimited facilitates the outputting of the processed data to a delimited file, adhering to the defined schema.

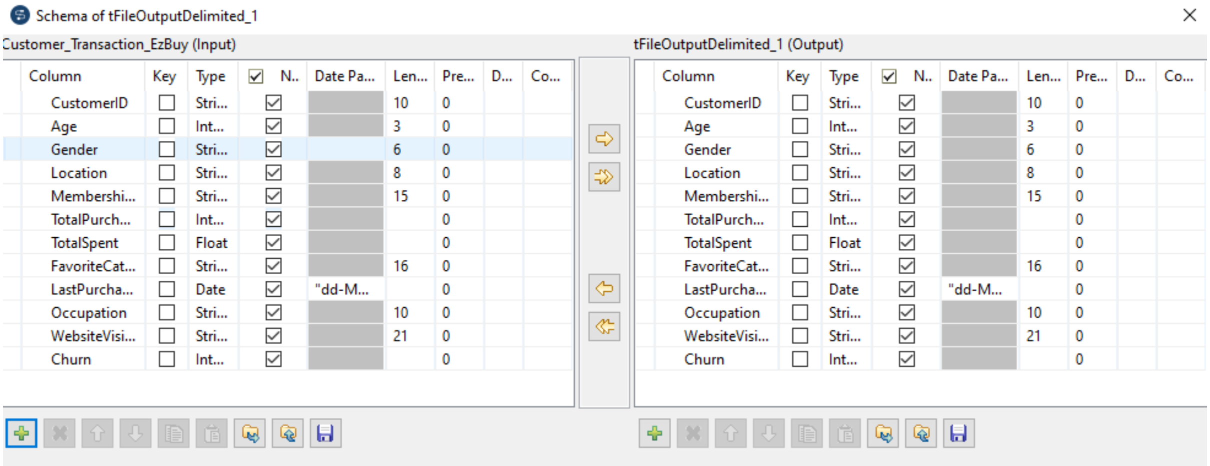



The component TMap figure above shows how was used to map both datasets, using a unique common key, the customerID from both datasets, to bring all the records together.




To ensure that TOutputDelimited functions as intended, it's important to follow a series of steps. First, we must verify the designated output path, making sure it's correctly specified.

Secondly, we should set the field separator to a comma (","), which defines how the data is separated in the output file.



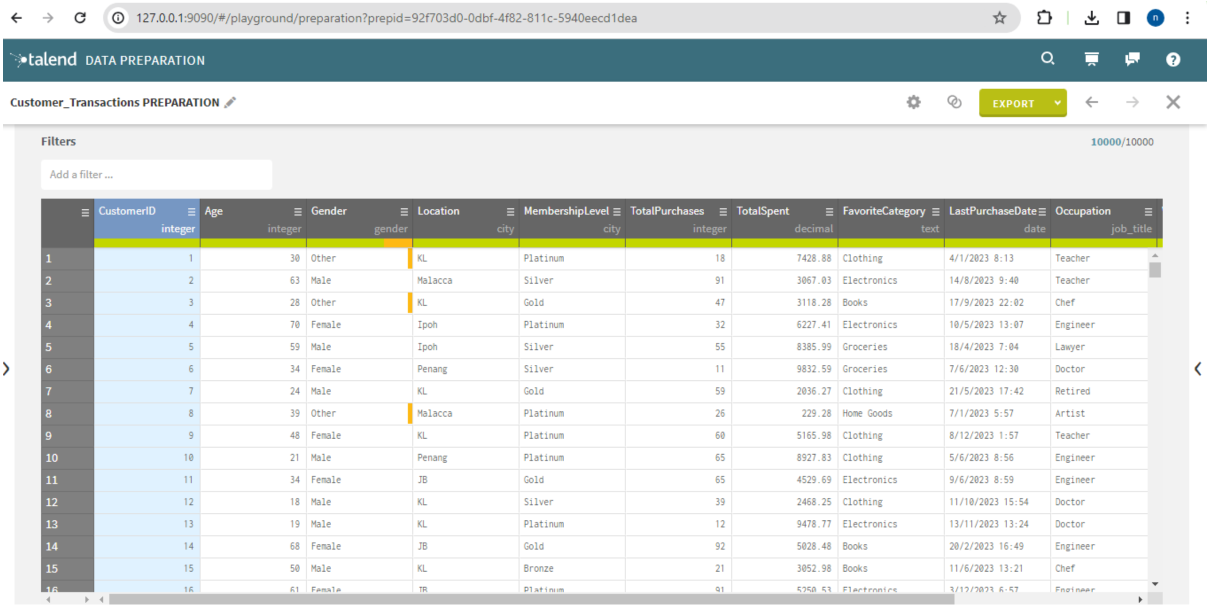

| Column | Key | Type | ☑ N.. | Date Pa... | Len... | Pre... | D... | Co... |
|---|---|---|---|---|---|---|---|---|
| CustomerID | ☐ | Stri... | ☑ | | 10 | 0 | | |
| Age | ☐ | Int... | ☑ | | 3 | 0 | | |
| Gender | ☐ | Stri... | ☑ | | 6 | 0 | | |
| Location | ☐ | Stri... | ☑ | | 8 | 0 | | |
| Membershi... | ☐ | Stri... | ☑ | | 15 | 0 | | |
| TotalPurch... | ☐ | Int... | ☑ | | | 0 | | |
| TotalSpent | ☐ | Float | ☑ | | | 0 | | |
| FavoriteCat... | ☐ | Stri... | ☑ | | 16 | 0 | | |
| LastPurcha... | ☐ | Date | ☑ | "dd-M... | | 0 | | |
| Occupation | ☐ | Stri... | ☑ | | 10 | 0 | | |
| WebsiteVisi... | ☐ | Stri... | ☑ | | 21 | 0 | | |
| Churn | ☐ | Int... | ☑ | | | 0 | | |

| Column | Key | Type | ☑ N.. | Date Pa... | Len... | Pre... | D... | Co... |
|---|---|---|---|---|---|---|---|---|
| CustomerID | ☐ | Stri... | ☑ | | 10 | 0 | | |
| Age | ☐ | Int... | ☑ | | 3 | 0 | | |
| Gender | ☐ | Stri... | ☑ | | 6 | 0 | | |
| Location | ☐ | Stri... | ☑ | | 8 | 0 | | |
| Membershi... | ☐ | Stri... | ☑ | | 15 | 0 | | |
| TotalPurch... | ☐ | Int... | ☑ | | | 0 | | |
| TotalSpent | ☐ | Float | ☑ | | | 0 | | |
| FavoriteCat... | ☐ | Stri... | ☑ | | 16 | 0 | | |
| LastPurcha... | ☐ | Date | ☑ | "dd-M... | | 0 | | |
| Occupation | ☐ | Stri... | ☑ | | 10 | 0 | | |
| WebsiteVisi... | ☐ | Stri... | ☑ | | 21 | 0 | | |
| Churn | ☐ | Int... | ☑ | | | 0 | | |

Lastly, to ensure that all columns from the input schema are correctly copied over to the output schema. This process guarantees that the data is formatted and exported accurately in accordance with the defined criteria.

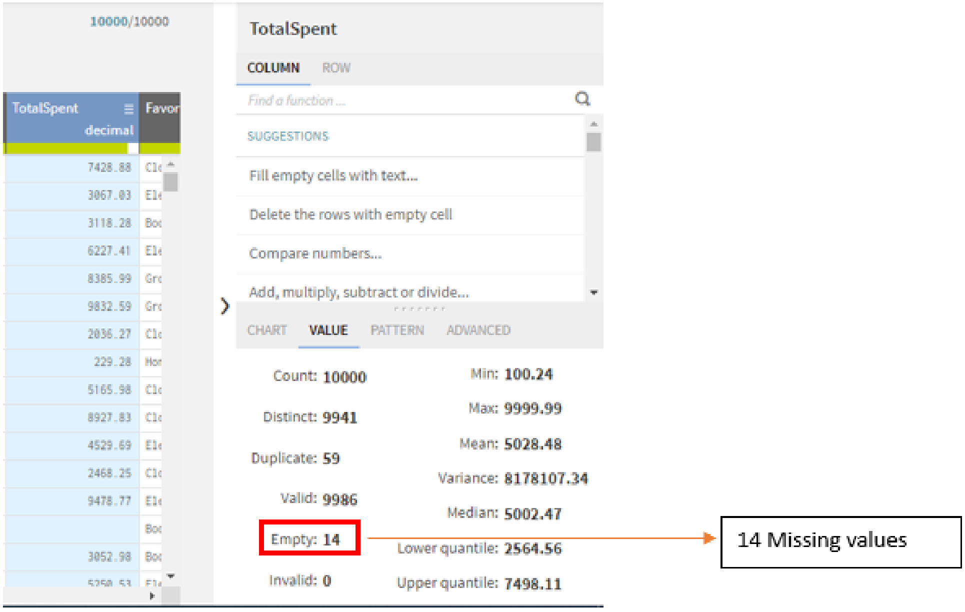
## Preprocessing using Talend Data Preparation

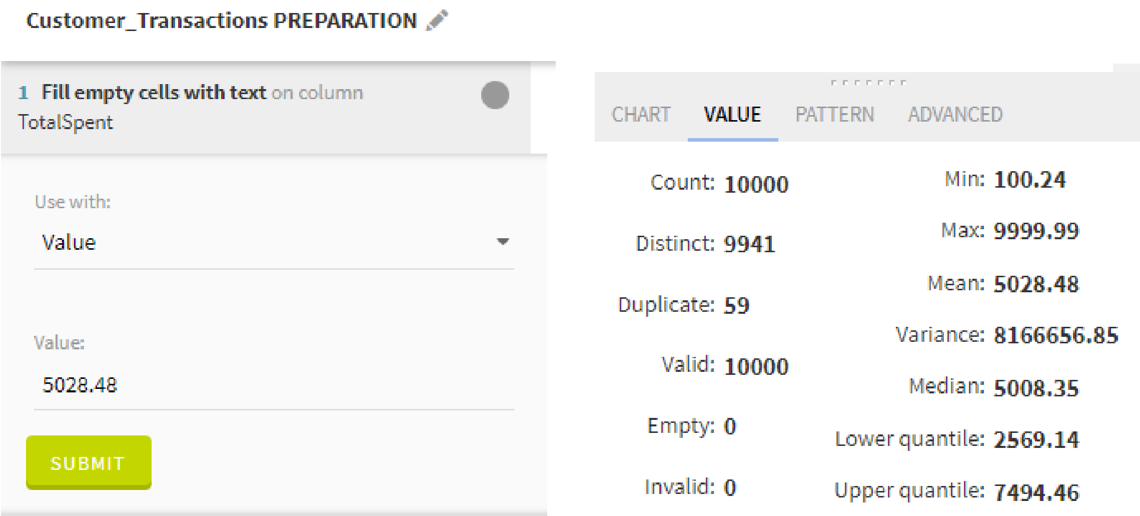
For the next preprocessing, I utilized Talend Data Preparation platform. I uploaded the integrated data from Talend Data Integration, which was produced in CSV delimiter format generated by the tFileOutputDelimited component. The figure below shows the uploaded CSV file in the talend data Preparation.



## Data Cleaning

The data cleaning process involves individually clicking on each column. During this, I noticed that the 'TotalSpend' column contains 14 missing values.

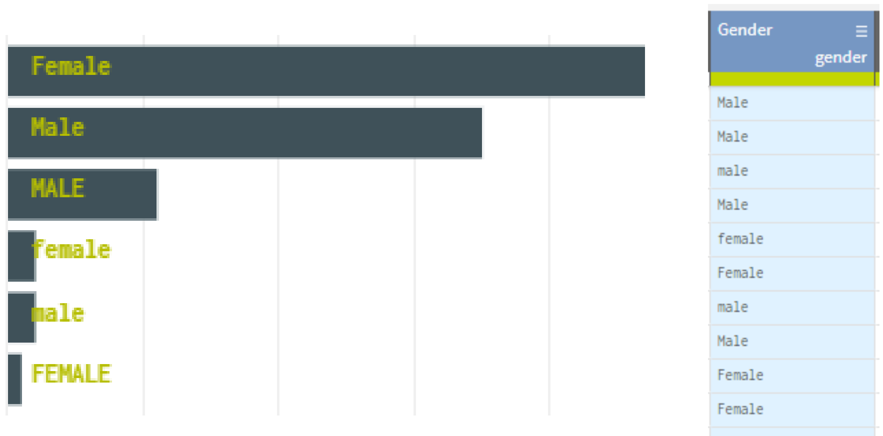


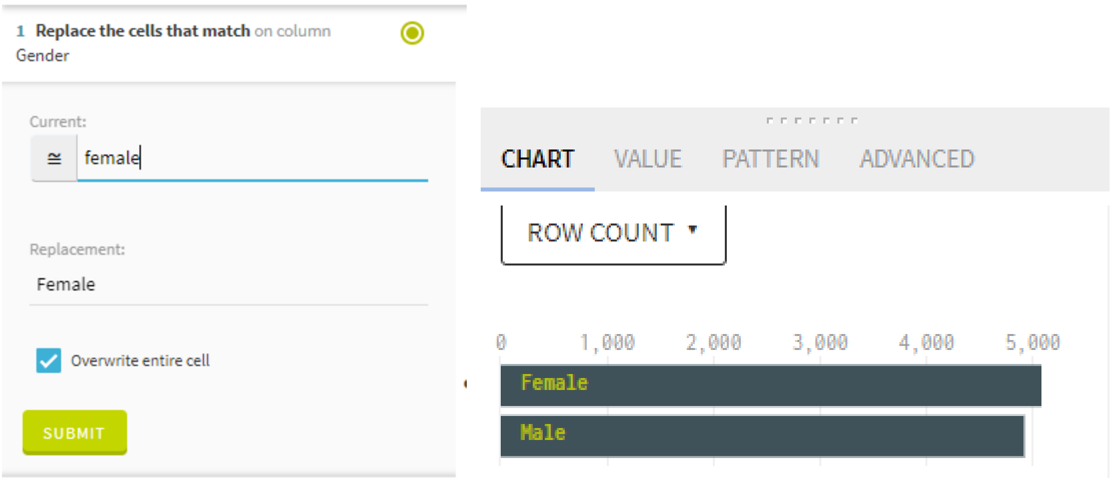

For the "TotalSpend" column, I identified 14 missing values. These were replaced with the mean value of 5028.48 to ensure that the column is free from any empty entries.

**Data Standardization**

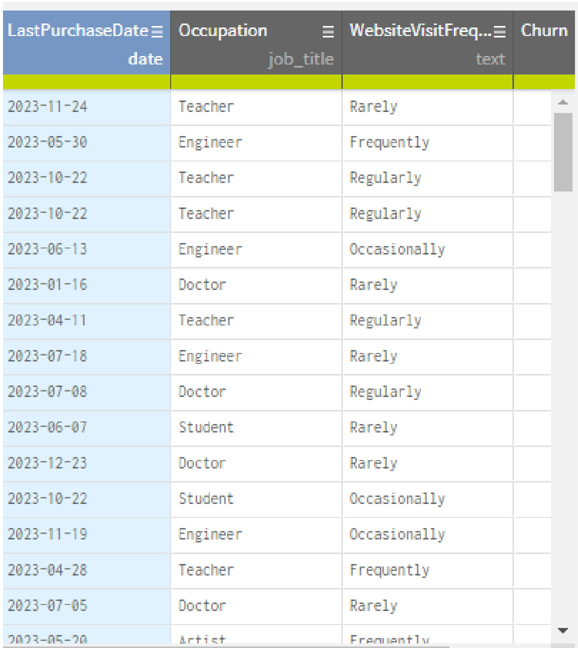
The data standardization process involves individually clicking on each column. During this, I noticed that the 'Gender' column contains varying letter cases, and the 'Date' column also lacks standardization.

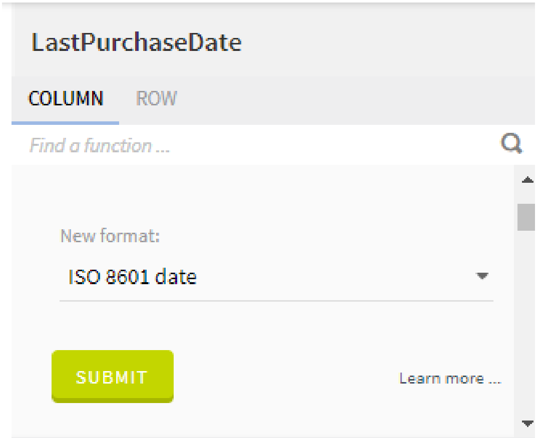



The figure above shows the gender contains varying letter cases.

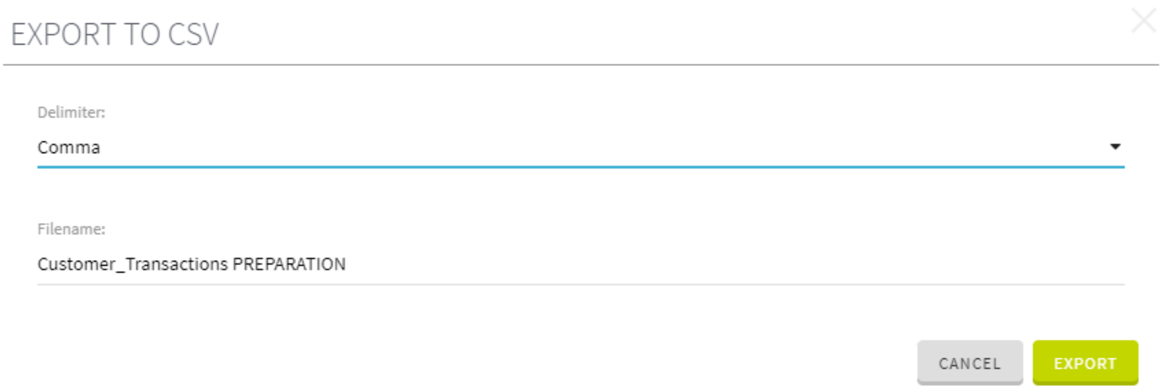



The figure above shows after standardized the gender.



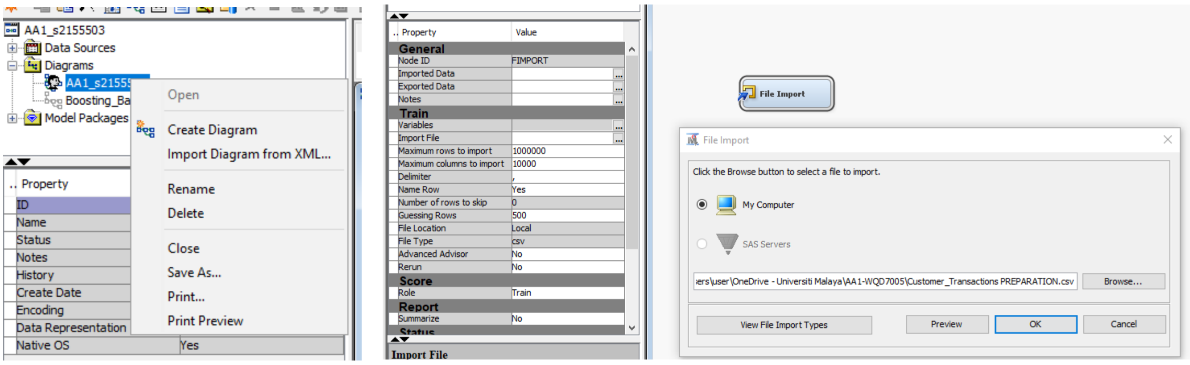




After identifying the date columns, a function within Talend was selected to standardize the dates. The ISO 8601 format was specified within this function to transform all date entries to this uniform standard. Following the transformation, the data was reviewed to ensure that the dates were correctly standardized, with any anomalies or errors requiring further attention being identified and possibly corrected.



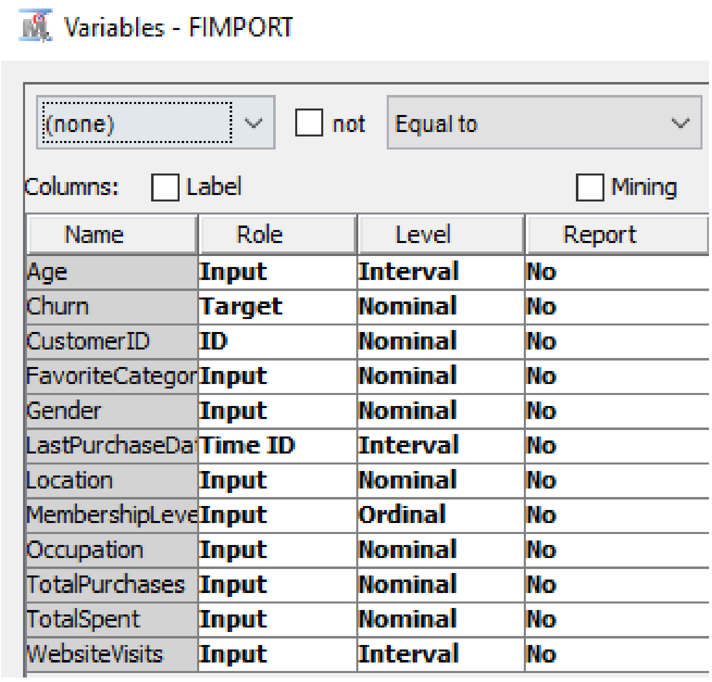

The final step involves exporting the file from Talend Data Preparation in CSV format, ensuring that the delimiter is set to a comma.

### Data Import to SAS Enterprise Miner




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

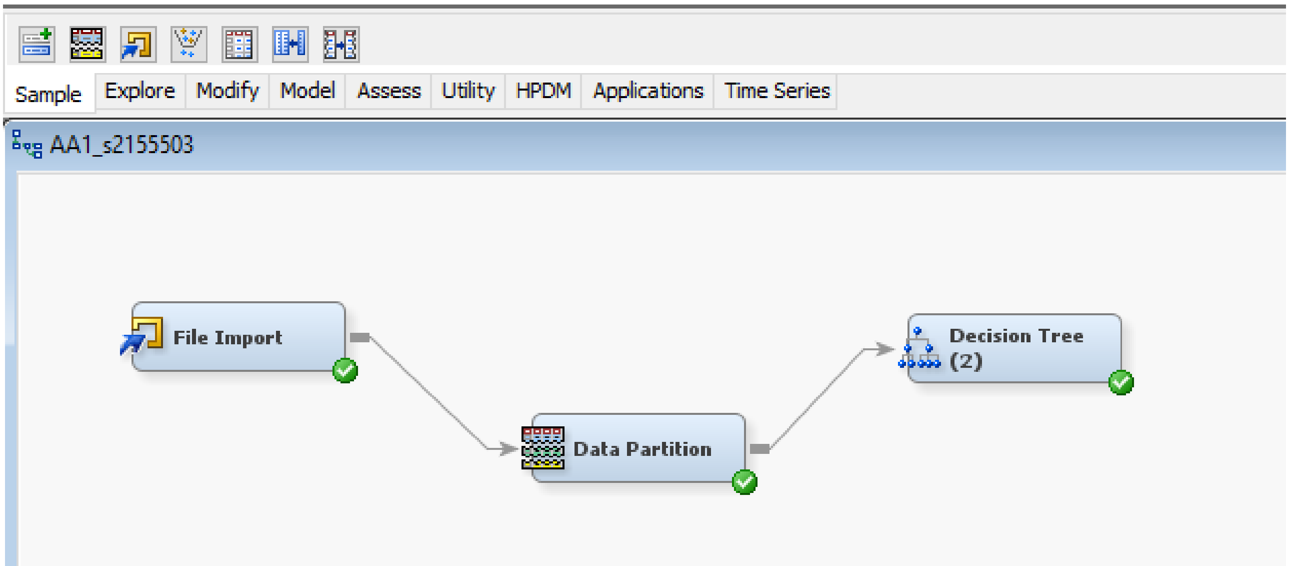
| Name | Role | Level | Report |
|---|---|---|---|
| Age | Input | Interval | No |
| Churn | Target | Nominal | No |
| CustomerID | ID | Nominal | No |
| FavoriteCategor | Input | Nominal | No |
| Gender | Input | Nominal | No |
| LastPurchaseDa | Time ID | Interval | No |
| Location | Input | Nominal | 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.

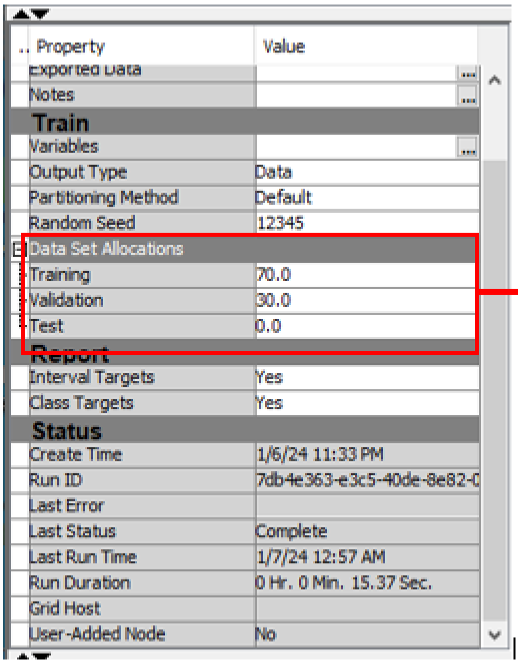
**1.0 Tasks 2**

***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**

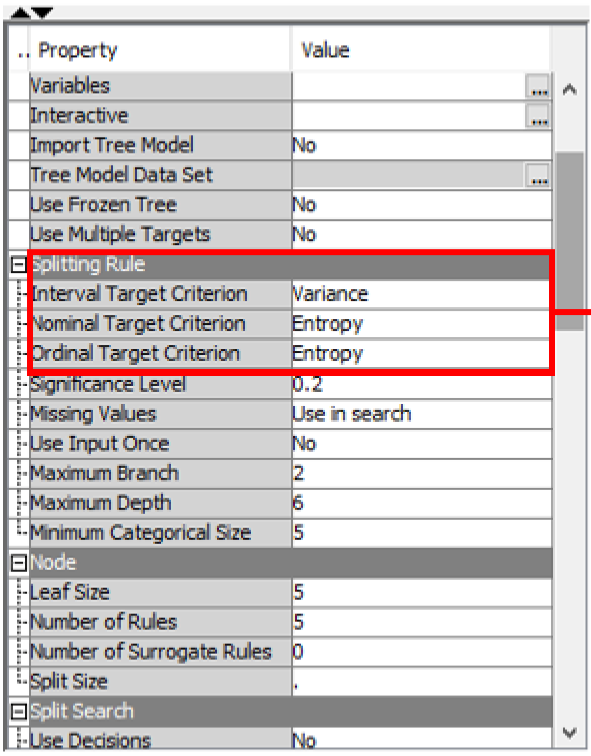



Training: 70

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**



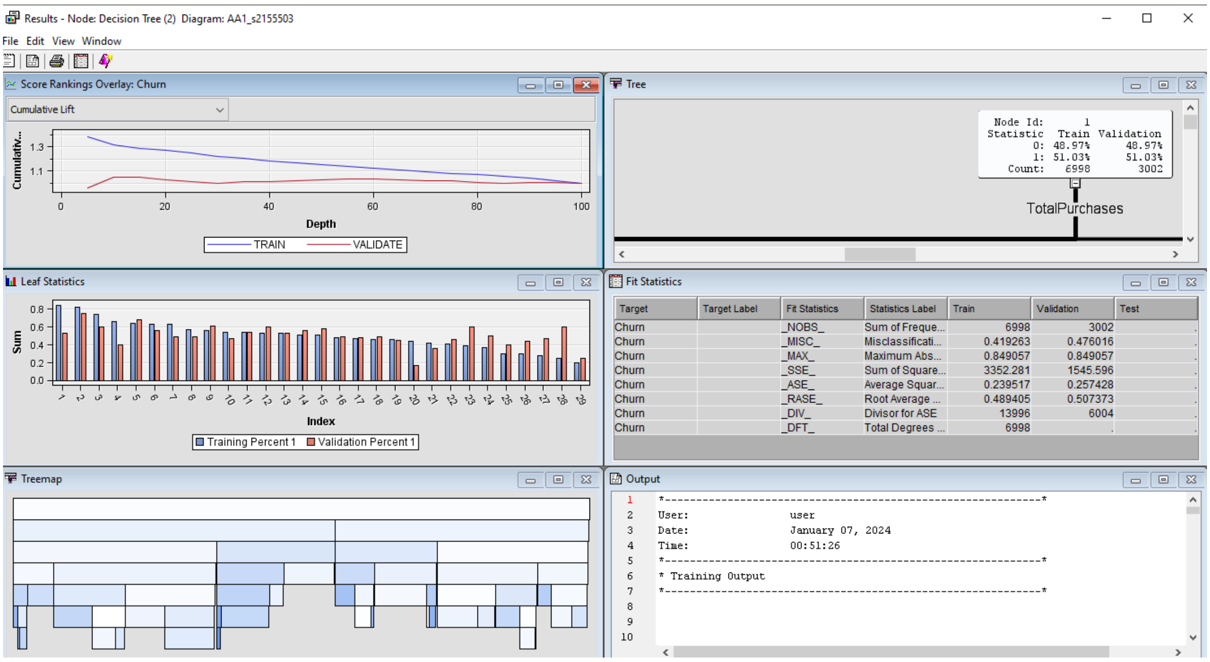

**Change the Splitting Rule**

Interval Target Criterion: Variance

Nominal Target Criterion: Entropy

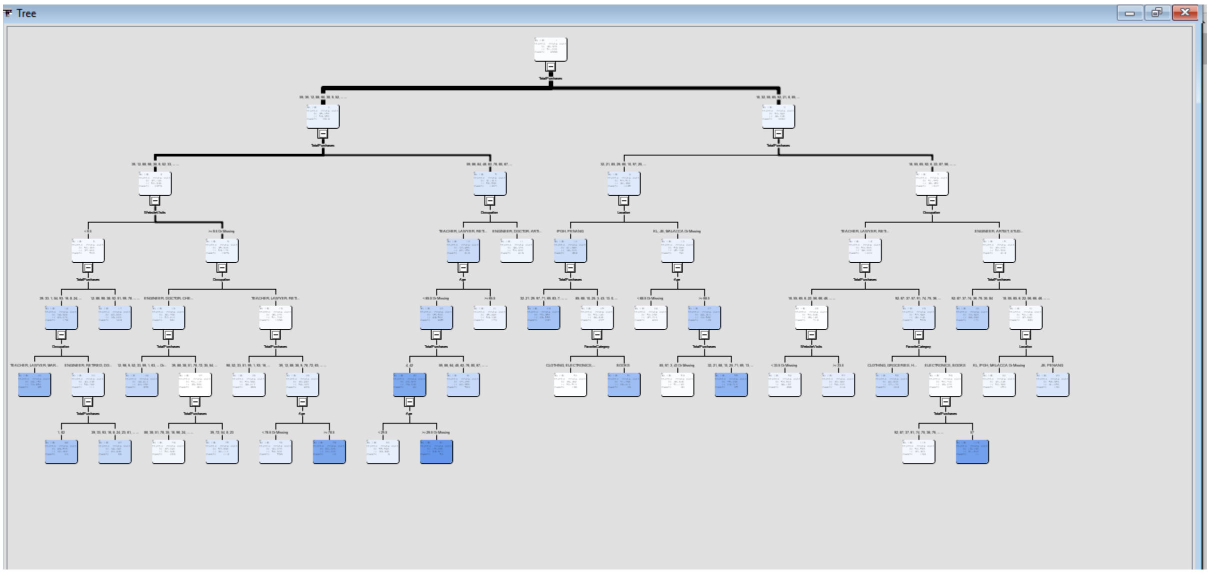
Ordinal Target Criterion: Entropy

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.



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: -

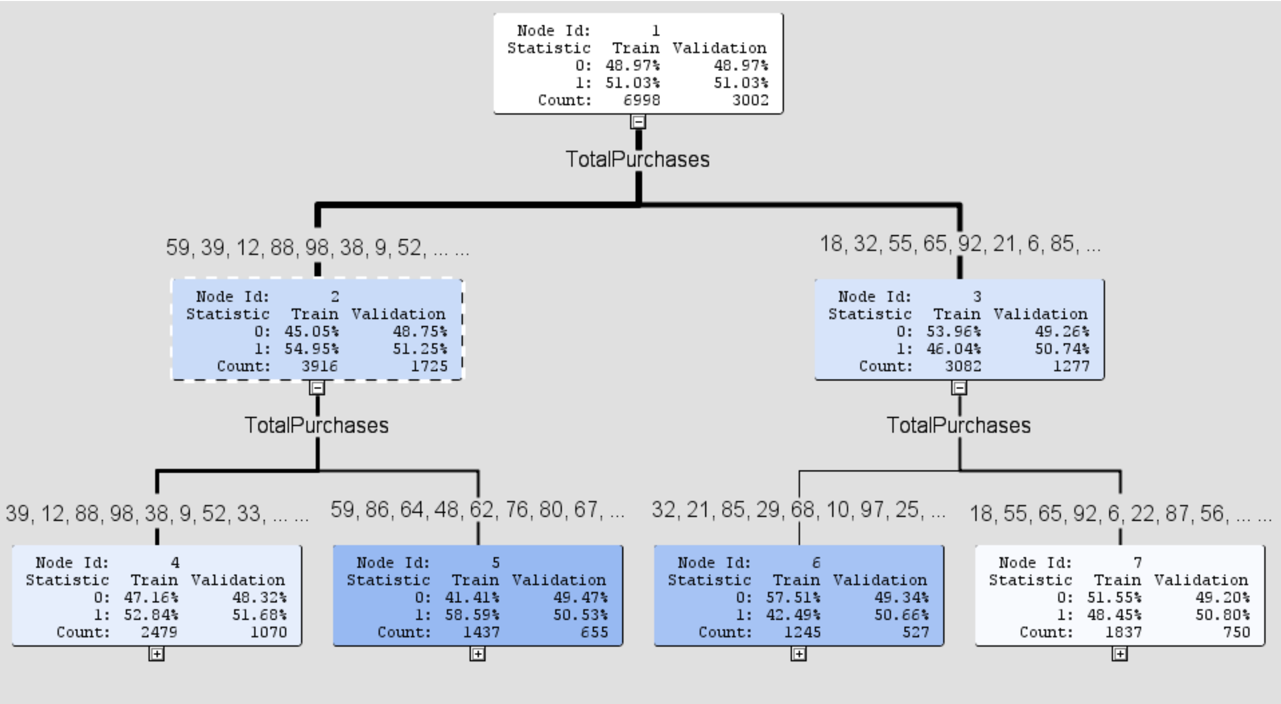



Tree diagram with 14 splitting rules for the input column “TotalPurchases”.

Variable Importance

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

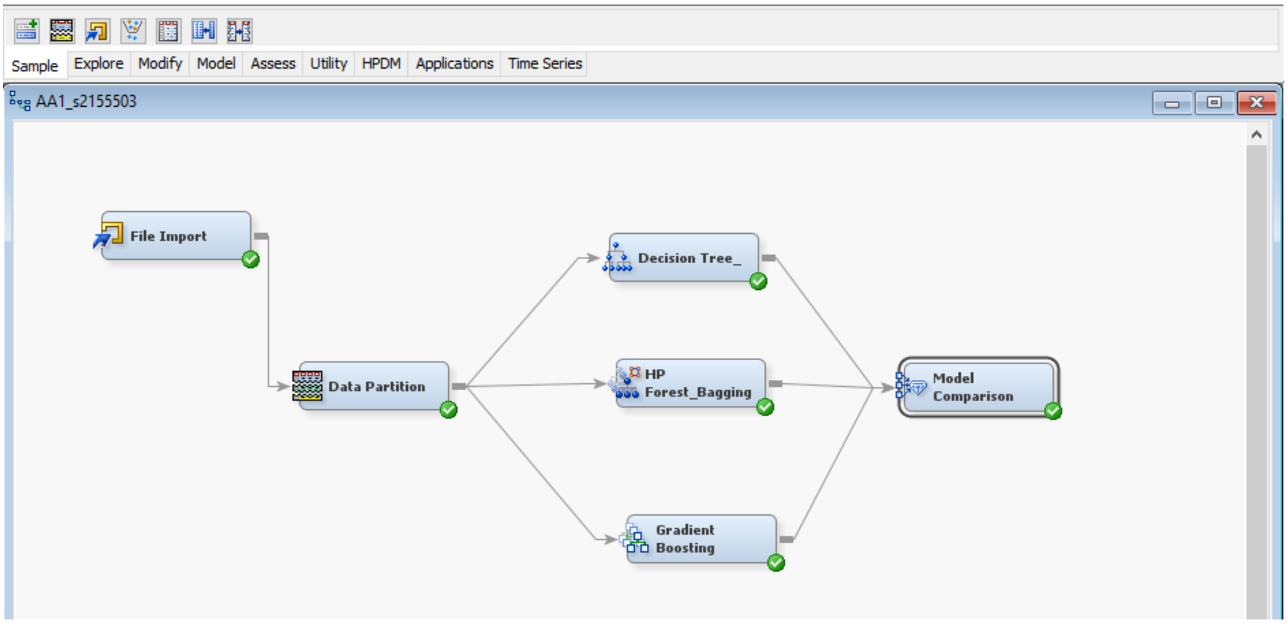
The customer behavior analysis, using decision tree modeling, indicates that the frequency of purchases ('Total Purchases') is the most critical factor in determining customer churn, with other variables such as 'Occupation' and 'Age' also playing significant roles. 'Total Purchases' is the principal 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.

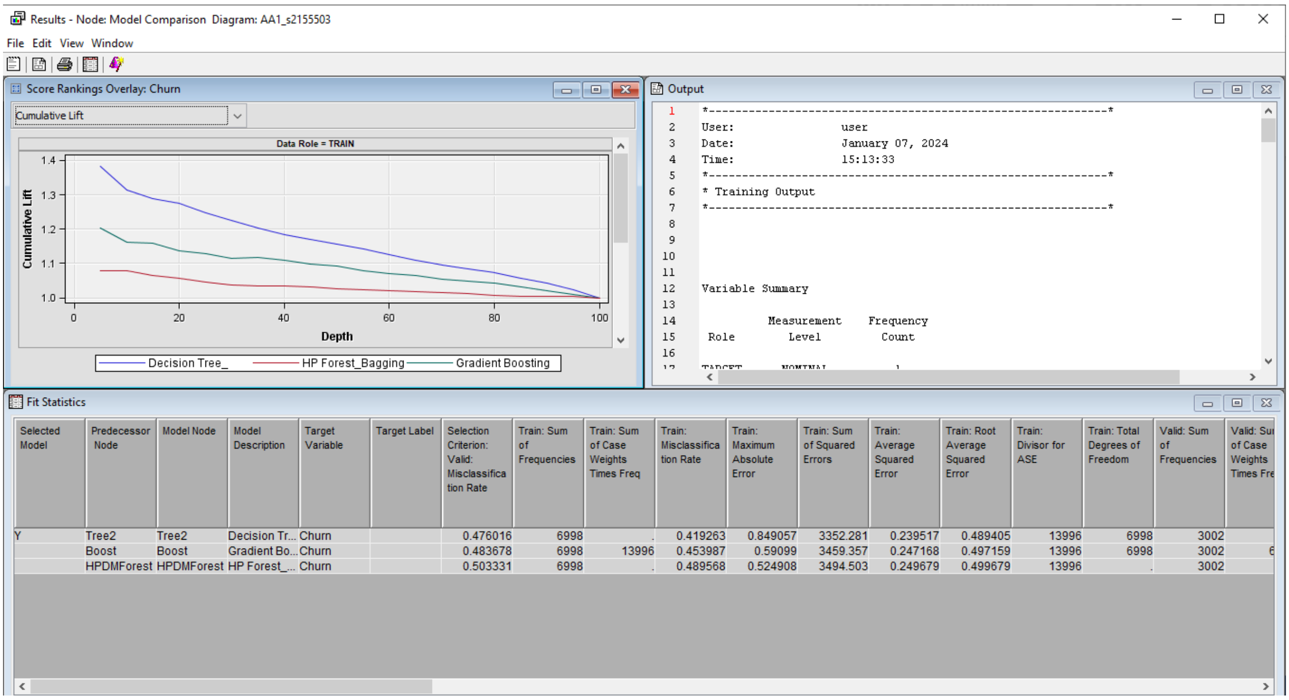
Overall, the analysis underscores the need to consider how various factors like purchasing frequency, occupation, and age interplay in influencing customer retention. There is also a suggestion of the need to address overfitting and enhance the model's generalization capabilities to better predict churn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

**2.0 Tasks 3**

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

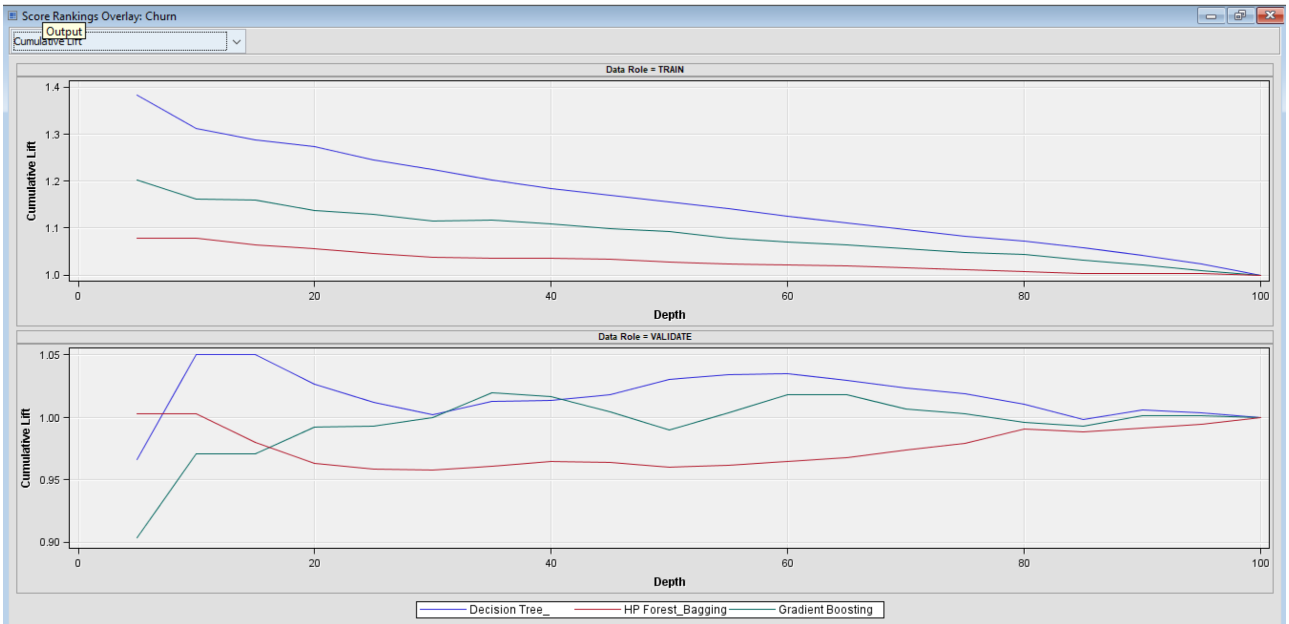



In this SAS Enterprise Miner workflow, I've constructed an ensemble methods approach to enhance predictive performance, building upon the foundation laid by the decision tree model established in the earlier task. The ensemble methodology is incorporating two components which are the HP Forest and Gradient 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.




| 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.

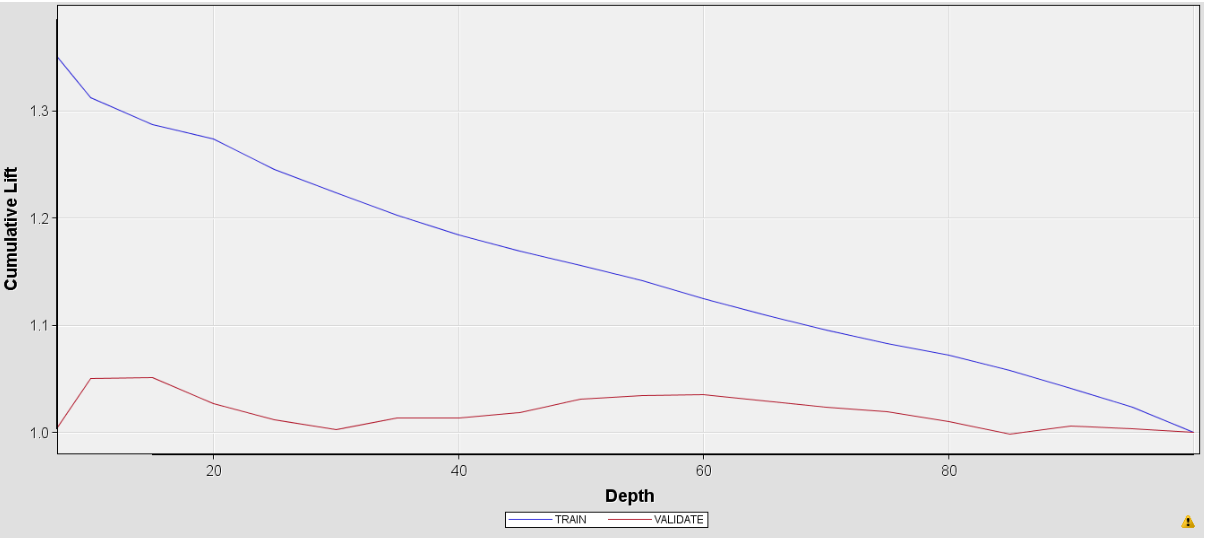
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

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

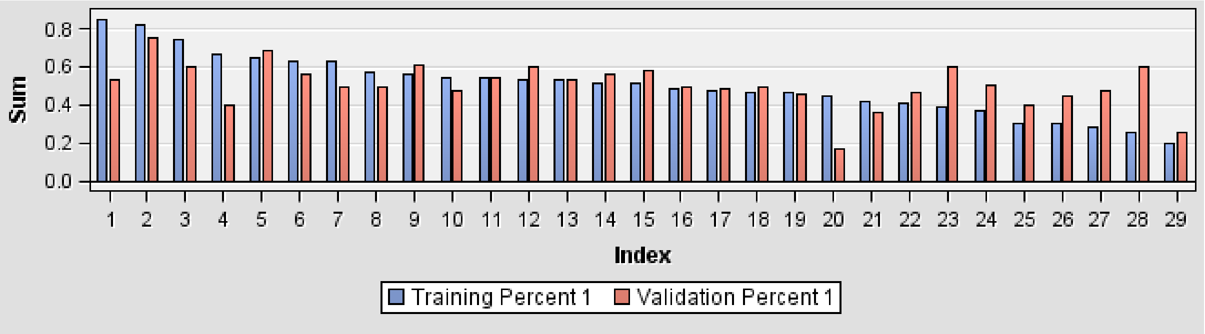
For the training set, the Decision Tree model has a misclassification rate of 0.41926 and an average squared error of 0.23952. The Gradient Boosting model shows a similar misclassification rate of 0.45399 but has a slightly better average squared error of 0.2417. The HP Forest model presents the highest 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**

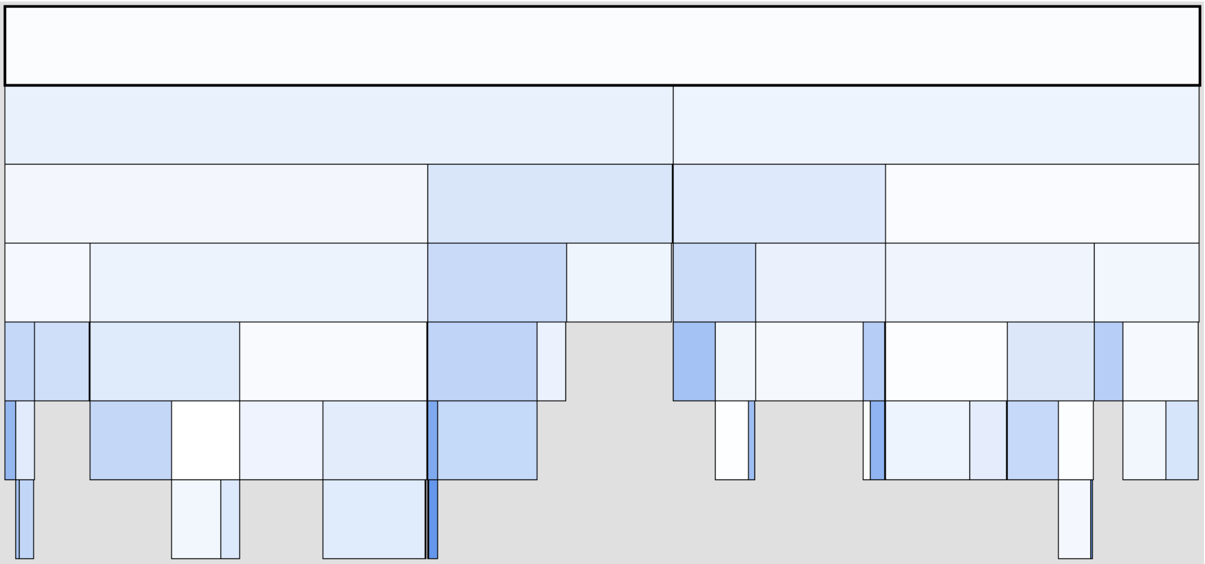
Score Rankings Overlays: Churn



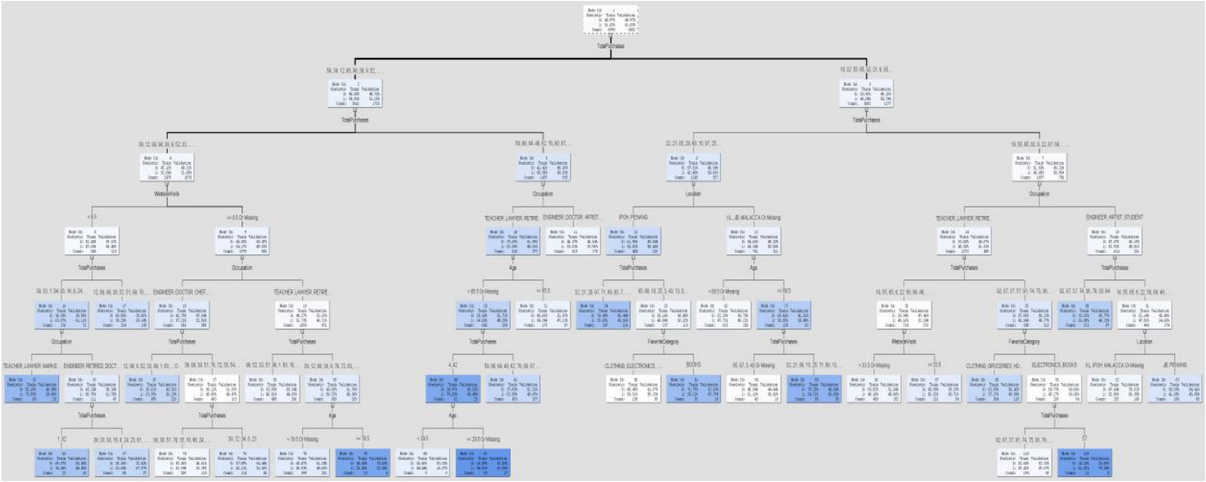

Leaf Statistics




TreeMap



Tree



Fit Statistics

| Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test |
|---|---|---|---|---|---|---|
| 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 | . | . |

```
*------------------------------------------------------------*
User:               user
Date:               January 07, 2024
Time:               00:57:51
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary


         Measurement    Frequency
Role        Level        Count


ID        INTERVAL         1
ID        NOMINAL          1
INPUT     INTERVAL         2
INPUT     NOMINAL          6
INPUT     ORDINAL          1
TARGET    NOMINAL          1
TIMEID    INTERVAL         1




Model Events
```

| Target | Event | Measurement Level | Number of Levels | Order | Label |
|---|---|---|---|---|---|
| Churn | 1 | NOMINAL | 2 | Descending | |

Predicted and decision variables

| Type | Variable | Label |
|---|---|---|
| TARGET | Churn | |
| PREDICTED | P_Churn1 | Predicted: Churn=1 |
| RESIDUAL | R_Churn1 | Residual: Churn=1 |
| PREDICTED | P_Churn0 | Predicted: Churn=0 |
| RESIDUAL | R_Churn0 | Residual: Churn=0 |
| FROM | F_Churn | From: Churn |
| INTO | I_Churn | Into: Churn |

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

*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*

Variable Importance

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

Tree Leaf Report

| Node Id | Depth | Training Observations | Training Percent 1 | Validation Observations | Validation Percent 1 |
|---|---|---|---|---|---|
| 26 | 4 | 633 | 0.48 | 258 | 0.49 |
| 11 | 3 | 619 | 0.54 | 278 | 0.54 |
| 78 | 6 | 599 | 0.57 | 261 | 0.49 |
| 41 | 5 | 583 | 0.63 | 267 | 0.49 |
| 56 | 5 | 498 | 0.46 | 192 | 0.49 |
| 38 | 5 | 489 | 0.46 | 206 | 0.45 |
| 36 | 5 | 478 | 0.63 | 226 | 0.56 |
| 17 | 4 | 324 | 0.39 | 138 | 0.60 |
| 58 | 5 | 304 | 0.37 | 119 | 0.50 |
| 74 | 6 | 289 | 0.53 | 118 | 0.53 |
| 62 | 5 | 253 | 0.53 | 106 | 0.60 |
| 24 | 4 | 247 | 0.30 | 101 | 0.44 |
| 57 | 5 | 216 | 0.56 | 84 | 0.61 |
| 50 | 5 | 198 | 0.51 | 96 | 0.58 |
| 118 | 6 | 194 | 0.47 | 86 | 0.48 |
| 63 | 5 | 190 | 0.41 | 68 | 0.46 |
| 21 | 4 | 173 | 0.54 | 87 | 0.47 |
| 30 | 4 | 171 | 0.66 | 87 | 0.40 |
| 75 | 6 | 114 | 0.42 | 45 | 0.36 |
| 67 | 6 | 88 | 0.64 | 37 | 0.68 |
| 55 | 5 | 85 | 0.25 | 35 | 0.60 |
| 32 | 5 | 65 | 0.74 | 25 | 0.60 |
| 81 | 6 | 53 | 0.85 | 17 | 0.53 |
| 54 | 5 | 43 | 0.51 | 18 | 0.56 |
| 51 | 5 | 39 | 0.28 | 19 | 0.47 |
| 66 | 6 | 23 | 0.30 | 10 | 0.40 |
| 119 | 6 | 11 | 0.82 | 8 | 0.75 |
| 79 | 6 | 10 | 0.20 | 4 | 0.25 |
| 80 | 6 | 9 | 0.44 | 6 | 0.17 |

Fit Statistics

Target=Churn Target Label=' '

| Fit Statistics | Statistics Label | Train | Validation |
|---|---|---|---|
| _NOBS_ | Sum of Frequencies | 6998.00 | 3002.00 |
| _MISC_ | Misclassification Rate | 0.42 | 0.48 |
| _MAX_ | Maximum Absolute Error | 0.85 | 0.85 |
| _SSE_ | Sum of Squared Errors | 3352.28 | 1545.60 |
| _ASE_ | Average Squared Error | 0.24 | 0.26 |
| _RASE_ | Root Average Squared Error | 0.49 | 0.51 |

| | | | |
|---|---|---|---|
| _DIV_ | Divisor for ASE | 13996.00 | 6004.00 |
| _DFT_ | Total Degrees of Freedom | 6998.00 | . |

Classification Table

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

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 57.8031 | 53.2828 | 1826 | 26.0932 |
| 1 | 0 | 42.1969 | 37.3285 | 1333 | 19.0483 |
| 0 | 1 | 41.7036 | 46.7172 | 1601 | 22.8780 |
| 1 | 1 | 58.2964 | 62.6715 | 2238 | 31.9806 |

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

| Target | Outcome | Target Percentage | Outcome Percentage | Frequency Count | Total Percentage |
|---|---|---|---|---|---|
| 0 | 0 | 51.5929 | 45.1701 | 664 | 22.1186 |
| 1 | 0 | 48.4071 | 40.6658 | 623 | 20.7528 |
| 0 | 1 | 46.9971 | 54.8299 | 806 | 26.8488 |
| 1 | 1 | 53.0029 | 59.3342 | 909 | 30.2798 |

Event Classification Table

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

| False Negative | True Negative | False Positive | True Positive |
|---|---|---|---|
| 1333 | 1826 | 1601 | 2238 |

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

| False Negative | True Negative | False Positive | True Positive |
|---|---|---|---|
| 623 | 664 | 806 | 909 |

Assessment Score Rankings

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

| Depth | Gain | Lift | Cumulative Lift | % Response | Cumulative % Response | Number of Observations | Mean Posterior Probability |
|---|---|---|---|---|---|---|---|
| 5 | 38.1953 | 1.38195 | 1.38195 | 70.5195 | 70.5195 | 350 | 0.70519 |
| 10 | 31.2350 | 1.24275 | 1.31235 | 63.4160 | 66.9677 | 350 | 0.63416 |
| 15 | 28.7466 | 1.23770 | 1.28747 | 63.1584 | 65.6979 | 350 | 0.63158 |
| 20 | 27.4005 | 1.23362 | 1.27400 | 62.9503 | 65.0110 | 350 | 0.62950 |
| 25 | 24.5630 | 1.13213 | 1.24563 | 57.7713 | 63.5631 | 350 | 0.57771 |
| 30 | 22.3518 | 1.11296 | 1.22352 | 56.7931 | 62.4347 | 350 | 0.56793 |
| 35 | 20.2979 | 1.07974 | 1.20298 | 55.0980 | 61.3866 | 350 | 0.55098 |
| 40 | 18.3990 | 1.05107 | 1.18399 | 53.6349 | 60.4177 | 350 | 0.53635 |
| 45 | 16.8829 | 1.04754 | 1.16883 | 53.4548 | 59.6440 | 350 | 0.53455 |
| 50 | 15.5748 | 1.03768 | 1.15575 | 52.9518 | 58.9765 | 349 | 0.52952 |
| 55 | 14.1882 | 1.00326 | 1.14188 | 51.1951 | 58.2689 | 350 | 0.51195 |
| 60 | 12.4633 | 0.93495 | 1.12463 | 47.7093 | 57.3887 | 350 | 0.47709 |
| 65 | 10.9944 | 0.93371 | 1.10994 | 47.6463 | 56.6392 | 350 | 0.47646 |
| 70 | 9.5886 | 0.91318 | 1.09589 | 46.5986 | 55.9218 | 350 | 0.46599 |
| 75 | 8.3108 | 0.90425 | 1.08311 | 46.1429 | 55.2698 | 350 | 0.46143 |
| 80 | 7.1767 | 0.90169 | 1.07177 | 46.0123 | 54.6911 | 350 | 0.46012 |
| 85 | 5.7421 | 0.82792 | 1.05742 | 42.2479 | 53.9590 | 350 | 0.42248 |
| 90 | 4.1389 | 0.76889 | 1.04139 | 39.2355 | 53.1409 | 350 | 0.39236 |
| 95 | 2.3740 | 0.70611 | 1.02374 | 36.0320 | 52.2403 | 350 | 0.36032 |
| 100 | 0.0000 | 0.54771 | 1.00000 | 27.9491 | 51.0289 | 349 | 0.27949 |

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

| Depth | Gain | Lift | Cumulative Lift | % Response | Cumulative % Response | Number of Observations | Mean Posterior Probability |
|---|---|---|---|---|---|---|---|
| 5 | 3.37381 | 0.96626 | 0.96626 | 49.3109 | 49.3109 | 151 | 0.70094 |
| 10 | 5.05116 | 1.13532 | 1.05051 | 57.9385 | 53.6104 | 150 | 0.63427 |
| 15 | 5.07229 | 1.05115 | 1.05072 | 53.6428 | 53.6212 | 150 | 0.63240 |
| 20 | 2.66019 | 0.95408 | 1.02660 | 48.6891 | 52.3902 | 150 | 0.62950 |
| 25 | 1.20505 | 0.95375 | 1.01205 | 48.6723 | 51.6476 | 150 | 0.59578 |
| 30 | 0.23008 | 0.95349 | 1.00230 | 48.6590 | 51.1501 | 150 | 0.56928 |
| 35 | 1.31000 | 1.07797 | 1.01310 | 55.0115 | 51.7012 | 150 | 0.55745 |
| 40 | 1.32711 | 1.01447 | 1.01327 | 51.7711 | 51.7099 | 150 | 0.53859 |
| 45 | 1.81597 | 1.05730 | 1.01816 | 53.9568 | 51.9594 | 150 | 0.53635 |
| 50 | 3.08227 | 1.14487 | 1.03082 | 58.4260 | 52.6056 | 150 | 0.53078 |
| 55 | 3.46247 | 1.07242 | 1.03462 | 54.7284 | 52.7996 | 151 | 0.52197 |
| 60 | 3.50337 | 1.03954 | 1.03503 | 53.0504 | 52.8205 | 150 | 0.48884 |
| 65 | 2.96194 | 0.96457 | 1.02962 | 49.2248 | 52.5442 | 150 | 0.47709 |
| 70 | 2.36280 | 0.94566 | 1.02363 | 48.2595 | 52.2384 | 150 | 0.47108 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 75 | 1.93094 | 0.95879 | 1.01931 | 48.9297 | 52.0181 | 150 | 0.46184 |
| 80 | 1.03057 | 0.87513 | 1.01031 | 44.6602 | 51.5586 | 150 | 0.46012 |
| 85 | 0.17616 | 0.80500 | 0.99824 | 41.0813 | 50.9427 | 150 | 0.43323 |
| 90 | 0.57151 | 1.13292 | 1.00572 | 57.8159 | 51.3243 | 150 | 0.39494 |
| 95 | 0.37581 | 0.96851 | 1.00376 | 49.4255 | 51.2244 | 150 | 0.36427 |
| 100 | 0.00000 | 0.92855 | 1.00000 | 47.3861 | 51.0326 | 150 | 0.27998 |

Assessment Score Distribution

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

| Posterior Probability Range | Number of Events | Number of Nonevents | Mean Posterior Probability | Percentage |
|---|---|---|---|---|
| 0.80-0.85 | 54 | 10 | 0.84375 | 0.9145 |
| 0.70-0.75 | 48 | 17 | 0.73846 | 0.9288 |
| 0.65-0.70 | 113 | 58 | 0.66082 | 2.4436 |
| 0.60-0.65 | 726 | 423 | 0.63185 | 16.4190 |
| 0.55-0.60 | 462 | 353 | 0.56687 | 11.6462 |
| 0.50-0.55 | 835 | 740 | 0.53016 | 22.5064 |
| 0.45-0.50 | 849 | 965 | 0.46803 | 25.9217 |
| 0.40-0.45 | 130 | 183 | 0.41534 | 4.4727 |
| 0.35-0.40 | 240 | 388 | 0.38217 | 8.9740 |
| 0.30-0.35 | 7 | 16 | 0.30435 | 0.3287 |
| 0.25-0.30 | 84 | 202 | 0.29371 | 4.0869 |
| 0.20-0.25 | 21 | 64 | 0.24706 | 1.2146 |
| 0.15-0.20 | 2 | 8 | 0.20000 | 0.1429 |

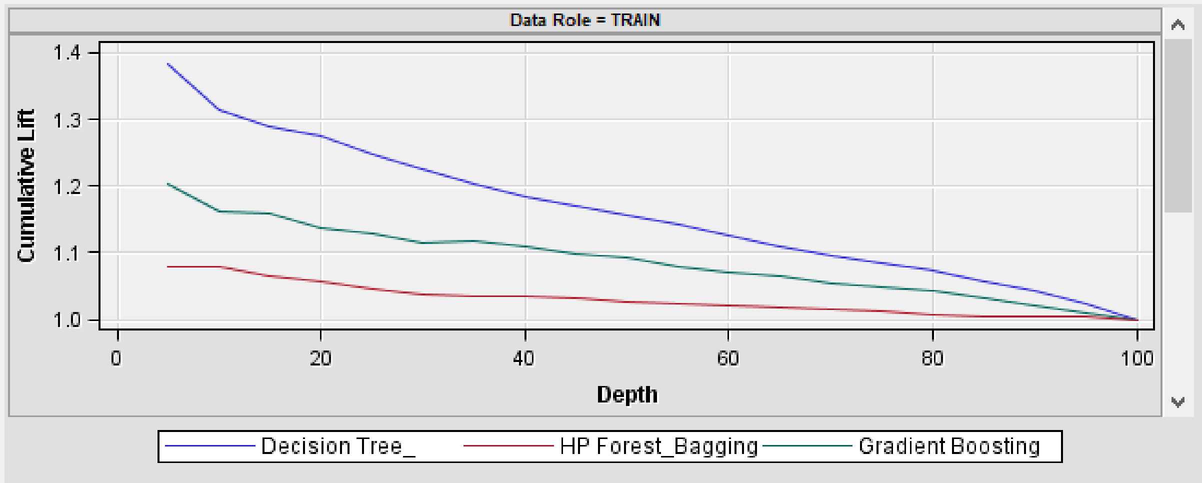
Data Role=VALIDATE Target Variable=Churn Target Label=' '

| Posterior Probability Range | Number of Events | Number of Nonevents | Mean Posterior Probability | Percentage |
|---|---|---|---|---|
| 0.80-0.85 | 15 | 10 | 0.83918 | 0.8328 |
| 0.70-0.75 | 15 | 10 | 0.73846 | 0.8328 |
| 0.65-0.70 | 35 | 52 | 0.66082 | 2.8981 |
| 0.60-0.65 | 282 | 248 | 0.63185 | 17.6549 |
| 0.55-0.60 | 178 | 167 | 0.56707 | 11.4923 |
| 0.50-0.55 | 384 | 319 | 0.53013 | 23.4177 |
| 0.45-0.50 | 354 | 388 | 0.46810 | 24.7169 |
| 0.40-0.45 | 48 | 71 | 0.41622 | 3.9640 |
| 0.35-0.40 | 142 | 115 | 0.38259 | 8.5610 |
| 0.30-0.35 | 4 | 6 | 0.30435 | 0.3331 |
| 0.25-0.30 | 53 | 67 | 0.29341 | 3.9973 |
| 0.20-0.25 | 21 | 14 | 0.24706 | 1.1659 |

| | | | | |
|---|---|---|---|---|
| 0.15-0.20 | 1 | 3 | 0.20000 | 0.1332 |

TASK 2

Score Ranking Overlays: Churn




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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |
| | HPDMForest | HPDMForest | HP Forest_... | Churn | | 0.503331 | 6998 | . | 0.489568 | 0.524908 | 3494.503 | 0.249679 | 0.499679 | 13996 | . | 3002 |

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

Variable Summary

| Role | Measurement Level | Frequency Count |
|---|---|---|
| TARGET | NOMINAL | 1 |

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 |
|---|---|---|---|---|---|---|
| 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 |

Fit Statistics Table
Target: Churn

Data Role=Train

| Statistics | Tree2 | Boost | HPDMForest |
|---|---|---|---|
| Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff | 0.50 | 0.51 | 0.51 |
| Train: Kolmogorov-Smirnov Statistic | 0.16 | 0.10 | 0.03 |
| Train: Average Squared Error | 0.24 | 0.25 | 0.25 |
| Train: Roc Index | 0.61 | 0.56 | 0.52 |
| Train: Cumulative Percent Captured Response | 13.13 | 11.61 | 10.78 |
| Train: Percent Captured Response | 6.22 | 5.60 | 5.39 |
| Selection Criterion: Valid: Misclassification Rate | 0.48 | 0.48 | 0.50 |
| Train: Total Degrees of Freedom | 6998.00 | 6998.00 | . |
| Train: Frequency of Classified Cases | . | . | 6998.00 |
| Train: Divisor for ASE | 13996.00 | 13996.00 | 13996.00 |
| Train: Gain | 31.23 | 16.07 | 7.77 |
| Train: Gini Coefficient | 0.23 | 0.13 | 0.04 |
| Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic | 0.16 | 0.10 | 0.03 |
| Train: Kolmogorov-Smirnov Probability Cutoff | 0.51 | 0.51 | 0.51 |
| Train: Cumulative Lift | 1.31 | 1.16 | 1.08 |
| Train: Lift | 1.24 | 1.12 | 1.08 |
| Train: Maximum Absolute Error | 0.85 | 0.59 | 0.52 |
| Train: Misclassification Rate | 0.42 | 0.45 | 0.49 |
| Train: Sum of Frequencies | 6998.00 | 6998.00 | 6998.00 |

| | | | |
|---|---|---|---|
| Train: Root Average Squared Error | 0.49 | 0.50 | 0.50 |
| Train: Cumulative Percent Response | 66.97 | 59.23 | 54.99 |
| Train: Percent Response | 63.42 | 57.09 | 54.99 |
| Train: Sum of Squared Errors | 3352.28 | 3459.36 | 3494.50 |
| Train: Sum of Case Weights Times Freq | . | 13996.00 | . |
| Train: Number of Wrong Classifications | . | . | 3426.00 |

Data Role=Valid

| Statistics | Tree2 | Boost | HPDMForest |
|---|---|---|---|
| Valid: Kolmogorov-Smirnov Statistic | 0.05 | 0.03 | 0.04 |
| Valid: Average Squared Error | 0.26 | 0.25 | 0.25 |
| Valid: Roc Index | 0.52 | 0.50 | 0.48 |
| Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff | 0.49 | 0.50 | 0.53 |
| Valid: Cumulative Percent Captured Response | 10.53 | 9.73 | 10.06 |
| Valid: Percent Captured Response | 5.67 | 5.19 | 5.01 |
| Valid: Frequency of Classified Cases | . | . | 3002.00 |
| Valid: Divisor for VASE | 6004.00 | 6004.00 | 6004.00 |
| Valid: Gain | 5.05 | 2.95 | 0.30 |
| Valid: Gini Coefficient | 0.03 | 0.01 | -0.05 |
| Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic | 0.04 | 0.02 | 0.00 |
| Valid: Kolmogorov-Smirnov Probability Cutoff | 0.48 | 0.50 | 0.51 |
| Valid: Cumulative Lift | 1.05 | 0.97 | 1.00 |
| Valid: Lift | 1.14 | 1.04 | 1.00 |
| Valid: Maximum Absolute Error | 0.85 | 0.59 | 0.52 |
| Valid: Misclassification Rate | 0.48 | 0.48 | 0.50 |
| Valid: Sum of Frequencies | 3002.00 | 3002.00 | 3002.00 |
| Valid: Root Average Squared Error | 0.51 | 0.50 | 0.50 |
| Valid: Cumulative Percent Response | 53.61 | 49.53 | 51.18 |
| Valid: Percent Response | 57.94 | 52.98 | 51.18 |
| Valid: Sum of Squared Errors | 1545.60 | 1508.76 | 1503.00 |
| Valid: Sum of Case Weights Times Freq | . | 6004.00 | . |
| Valid: Number of Wrong Classifications | . | . | 1511.00 |

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

| Model Node | Model Description | Data Role | Target | Target Label | False Negative | True Negative | False Positive | True Positive |
|---|---|---|---|---|---|---|---|---|

```
Boost       Gradient Boosting  TRAIN     Churn           1176      1426      2001      2395
Boost       Gradient Boosting  VALIDATE  Churn            535       553       917       997
Tree2       Decision Tree (2)  TRAIN     Churn           1333      1826      1601      2238
Tree2       Decision Tree (2)  VALIDATE  Churn            623       664       806       909
HPDMForest  HP Forest          TRAIN     Churn            510       511      2916      3061
HPDMForest  HP Forest          VALIDATE  Churn            258       217      1253      1274


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


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```