**Muhammad Syahirul Khaliq bin Mohamed Aidi Shahriz**

**22075208**

**Github link:** https://github.com/Syahirulkhaliq/Khaliq_22075208/

## Introduction to dataset and objective

Dataset consists of 1020 rows of unique customers and 13 attributes which are Customer ID, Gender, Age, City, Membership Type, Total Spend, Items Purchased, FavoriteCategory, Average Rating, Discount Applied, Days Since Last Purchase, Satisfaction Level and Churn. Below are the details of the attributes.

| Column | Type | Description |
|---|---|---|
| Customer_ID | Numeric | A unique identifier assigned to each customer, ensuring distinction across the dataset |
| Gender | Categorical | Specifies the gender of the customer, allowing for gender-based analytics |
| Age | Numeric | Represents the age of the customer, enabling age-group-specific insights. |
| City | Categorical | Indicates the city of residence for each customer, providing geographic insights |
| Membership_Type | Categorical | Identifies the type of membership held by the customer, influencing perks and benefits |
| Total_Spend | Numeric | Records the total monetary expenditure by the customer on the e-commerce platform |
| Items_Purchased | Numeric | Quantifies the total number of items purchased by the customer |
| FavoriteCategory | Categorical | Records most-explored category of a user, represents customer biasness towards products |
| Average_Rating | Numeric | Represents the average rating given by the customer for purchased items, gauging satisfaction |
| Discount_Applied | Boolean | Indicates whether a discount was applied to the customer's purchase, influencing buying behavior |
| Days_Since_Last_Purchase | Numeric | Reflects the number of days elapsed since the customer's most recent purchase, aiding in retention analysis |
| Satisfaction_Level | Categorical | Captures the overall satisfaction level of the customer, providing a subjective measure of their experience |
| Churn | Numeric | A binary column indicating whether the customer has churned (0 for retained, 1 for churned), indicating customer retention |

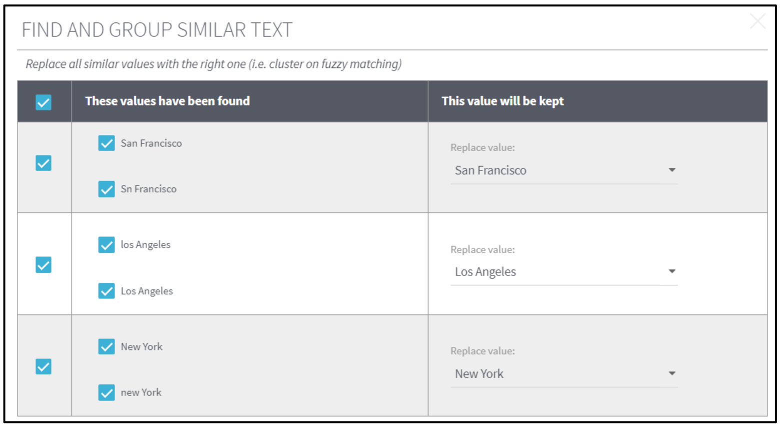
Based on the dataset, "Churn" variable is the most suitable target variable. This is because it directly relates to customer retention, a key aspect of understanding and predicting customer behaviour. By analysing churn, you can identify patterns and factors that indicate whether a customer is likely to stop purchasing (churn) or continue being active. Analyzing churn can provide valuable insights into customer loyalty, satisfaction, and overall engagement with the e-commerce platform. It can also help in developing strategies to improve customer retention and targeting interventions to reduce the churn rate. Thus, objective of this study to

- Analyse the customer behaviour dataset and derived meaningful insights from the model analysis
- Assess and compare the performance of models which can emphasize the reliability of the insights develop

In this study, SAS Enterprise Miner will be mainly utilised to extract information for the following objectives with the inclusion of other tools such as Talend Data Prep.

## Dataset Import and Preprocessing

Prior from using SAS Enterprise Miner, Talend Data Preparation was utilized to provide comprehensive details on dataset, ensure data consistency of categorical data values such as spelling errors and alphabet casing, as well as checking missing values. There were some data inconsistencies that can be group using “Find and group similar text” function in city column as below:



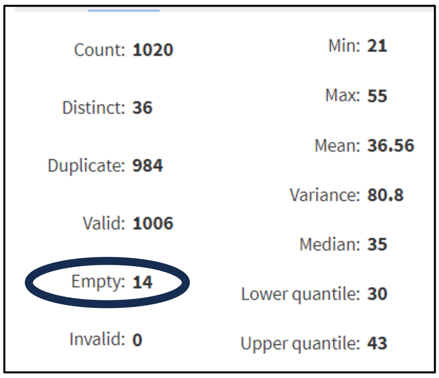
FIND AND GROUP SIMILAR TEXT

*Replace all similar values with the right one (i.e. cluster on fuzzy matching)*

| ☑ | These values have been found | This value will be kept |
|---|---|---|
| ☑ | ☑ San Francisco<br>☑ Sn Francisco | Replace value:<br>San Francisco |
| ☑ | ☑ los Angeles<br>☑ Los Angeles | Replace value:<br>Los Angeles |
| ☑ | ☑ New York<br>☑ new York | Replace value:<br>New York |

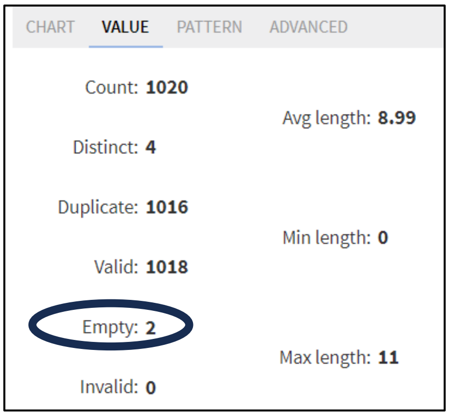
Such approach would be easier to conduct in Talend Data Preparation compared to SAS. In addition, missing values were detected in ‘Age’ and ‘Satisfaction Level’ column as shown below:

‘Age’ Column



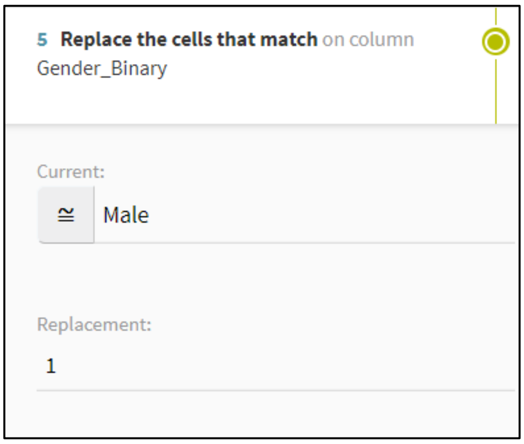
| | |
|---|---|
| Count: **1020** | Min: **21** |
| Distinct: **36** | Max: **55** |
| Duplicate: **984** | Mean: **36.56** |
| Valid: **1006** | Variance: **80.8** |
| Empty: **14** | Median: **35** |
| Invalid: **0** | Lower quantile: **30** |
| | Upper quantile: **43** |

‘Satisfaction Level’ column



Count: 1020
Avg length: 8.99
Distinct: 4
Duplicate: 1016
Min length: 0
Valid: 1018
Empty: 2
Max length: 11
Invalid: 0

Considering the amount of dataset rows to be quite limited, imputation method is preferable in comparison with deleting rows to prevent the loss of data. Such method will be implied in SAS Enterprise Miner. In addition, columns containing 2 unique values like 'Gender' and 'Discount Applied' was duplicated, and the duplicated columns were replace with binary values using 'replace the cells that match' function with below as an example:

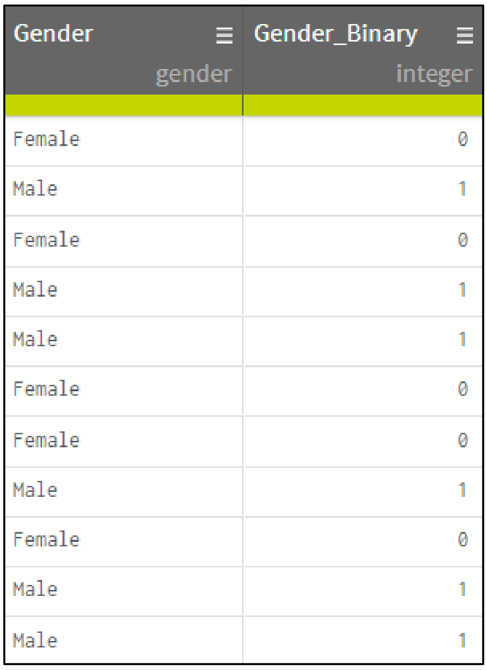


5 **Replace the cells that match** on column Gender_Binary

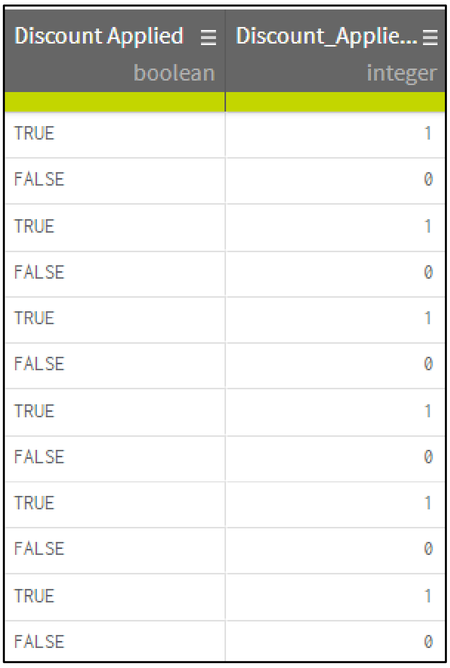
Current:
≅ Male

Replacement:
1

For 'Gender' column, 'Male' value represents as 1 while 'Female' value represents as 0. While for 'Discount Applied' column, 'TRUE' value represents as 1 while 'FALSE' value represents as 0. The result of columns are as below:

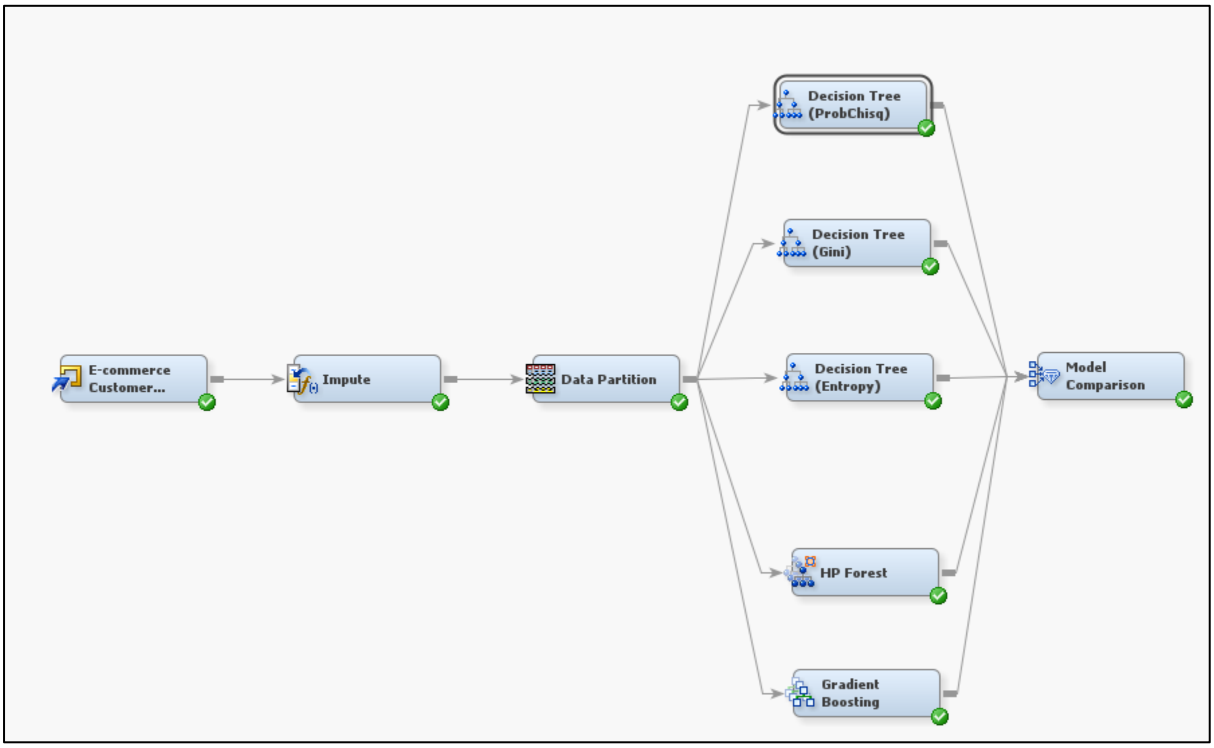


| Gender<br>gender | Gender_Binary<br>integer |
|---|---|
| Female | 0 |
| Male | 1 |
| Female | 0 |
| Male | 1 |
| Male | 1 |
| Female | 0 |
| Female | 0 |
| Male | 1 |
| Female | 0 |
| Male | 1 |
| Male | 1 |

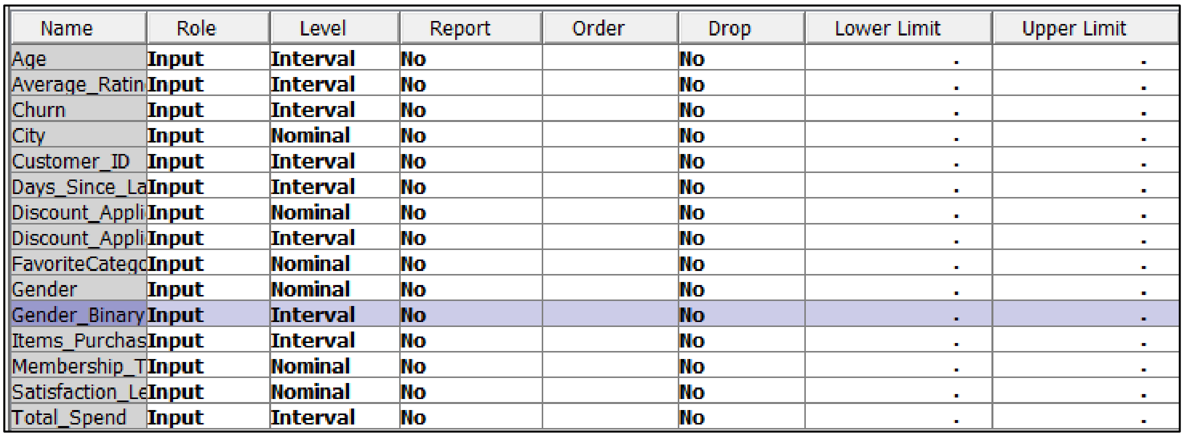


| Discount Applied<br>boolean | Discount_Applie...<br>integer |
|---|---|
| TRUE | 1 |
| FALSE | 0 |
| TRUE | 1 |
| FALSE | 0 |
| TRUE | 1 |
| FALSE | 0 |
| TRUE | 1 |
| FALSE | 0 |
| TRUE | 1 |
| FALSE | 0 |
| TRUE | 1 |
| FALSE | 0 |

The dataset was then exported to be imported in SAS for further preprocessing. Below

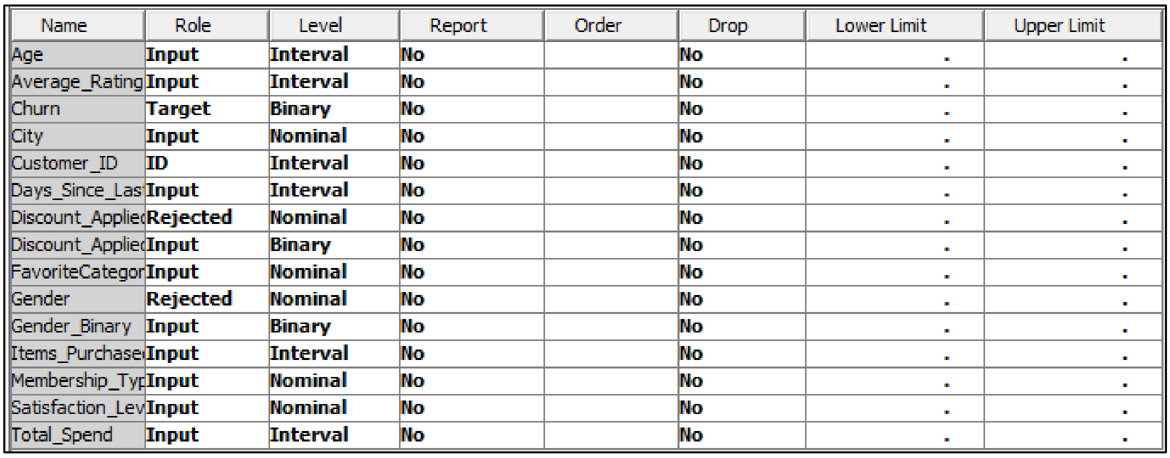


The file was imported via ‘File import’ node. Considering the dataset consists of only 1020 rows, no sampling was needed as dataset itself will be considered as representation. Details of initial roles were as below:



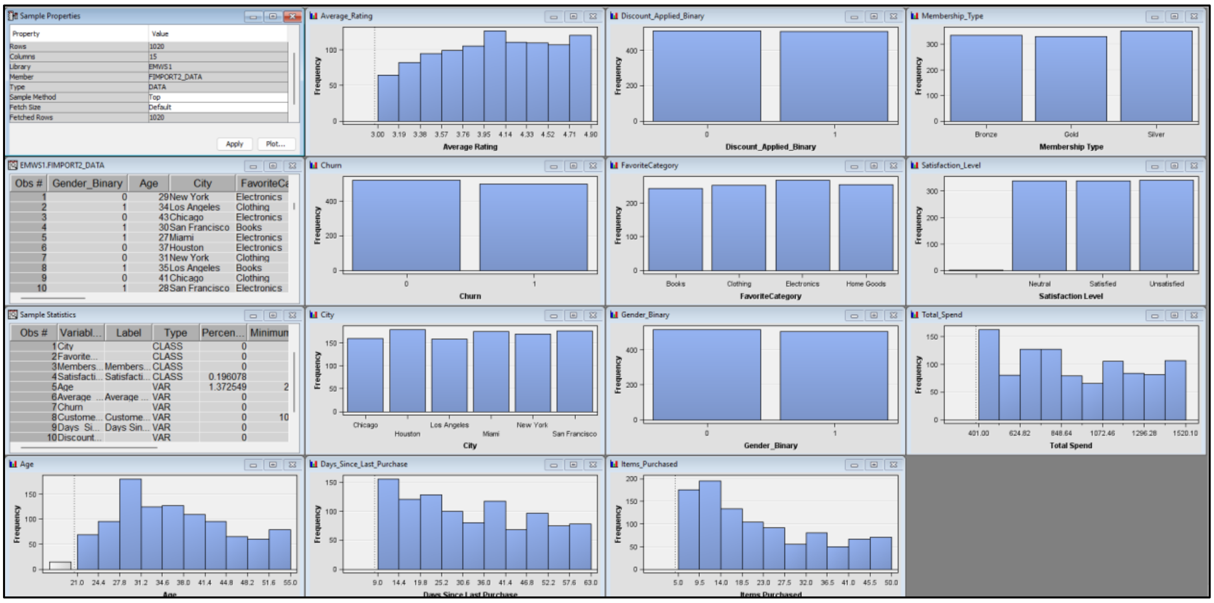
| Name | Role | Level | Report | Order | Drop | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|---|
| Age | Input | Interval | No | | No | . | . |
| Average_Ratin | Input | Interval | No | | No | . | . |
| Churn | Input | Interval | No | | No | . | . |
| City | Input | Nominal | No | | No | . | . |
| Customer_ID | Input | Interval | No | | No | . | . |
| Days_Since_La | Input | Interval | No | | No | . | . |
| Discount_Appli | Input | Nominal | No | | No | . | . |
| Discount_Appli | Input | Interval | No | | No | . | . |
| FavoriteCatego | Input | Nominal | No | | No | . | . |
| Gender | Input | Nominal | No | | No | . | . |
| Gender_Binary | Input | Interval | No | | No | . | . |
| Items_Purchas | Input | Interval | No | | No | . | . |
| Membership_T | Input | Nominal | No | | No | . | . |
| Satisfaction_Le | Input | Nominal | No | | No | . | . |
| Total_Spend | Input | Interval | No | | No | . | . |

In the ‘edit variable’ section of file import. Minor changes were conducted in terms of role and level. Churn was selected as the target variable for customer behaviour analysis while Customer_ID role was changed to ID. The role of nominal level columns of Discount_Applied and Gender were set to rejected to prevent redundancy with the modified columns of these 2 columns. In addition, the level of Discount_Applied_Binary, Gender_Binary and Churn were set to Binary due to its binary value. Below are the overall changes made:



| Name | Role | Level | Report | Order | Drop | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|---|
| Age | Input | Interval | No | | No | . | . |
| Average_Rating | Input | Interval | No | | No | . | . |
| Churn | Target | Binary | No | | No | . | . |
| City | Input | Nominal | No | | No | . | . |
| Customer_ID | ID | Interval | No | | No | . | . |
| Days_Since_Las | Input | Interval | No | | No | . | . |
| Discount_Applied | Rejected | Nominal | No | | No | . | . |
| Discount_Applied | Input | Binary | No | | No | . | . |
| FavoriteCategor | Input | Nominal | No | | No | . | . |
| Gender | Rejected | Nominal | No | | No | . | . |
| Gender_Binary | Input | Binary | No | | No | . | . |
| Items_Purchase | Input | Interval | No | | No | . | . |
| Membership_Typ | Input | Nominal | No | | No | . | . |
| Satisfaction_Lev | Input | Nominal | No | | No | . | . |
| Total_Spend | Input | Interval | No | | No | . | . |

A quick exploration on attributes was conducted via choosing every related attributes in ‘edit variable’ section and clicking ‘explore’. Below are some visualizations of the exploration:



Sample Properties

| Property | Value |
|---|---|
| Rows | 1020 |
| Columns | 15 |
| Library | EMWS1 |
| Member | FIMPORT2_DATA |
| Type | DATA |
| Sample Method | Top |
| Fetch Size | Default |
| Fetched Rows | 1020 |

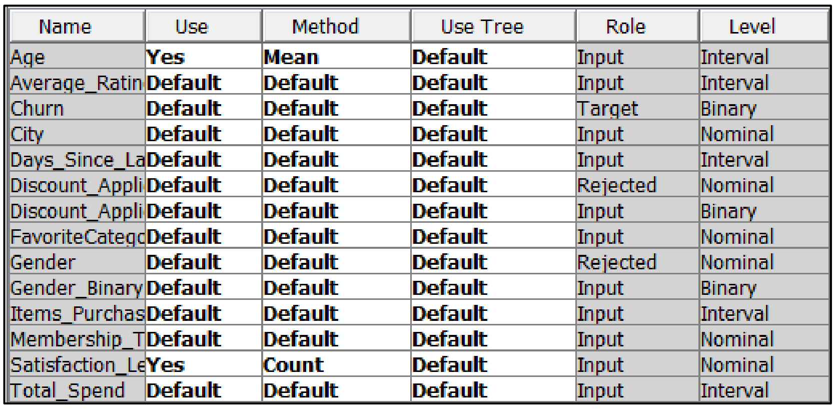
EMWS1.FIMPORT2_DATA

| Obs # | Gender_Binary | Age | City | FavoriteCa |
|---|---|---|---|---|
| 1 | 0 | 29 | New York | Electronics |
| 2 | 1 | 34 | Los Angeles | Clothing |
| 3 | 0 | 43 | Chicago | Electronics |
| 4 | 1 | 30 | San Francisco | Books |
| 5 | 1 | 27 | Miami | Electronics |
| 6 | 0 | 37 | Houston | Electronics |
| 7 | 0 | 31 | New York | Clothing |
| 8 | 1 | 35 | Los Angeles | Books |
| 9 | 0 | 41 | Chicago | Clothing |
| 10 | 1 | 28 | San Francisco | Electronics |

Sample Statistics

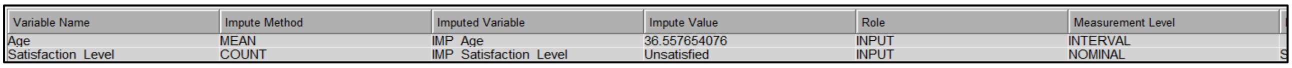
| Obs # | Variabl... | Label | Type | Percen... | Minimun |
|---|---|---|---|---|---|
| 1 | City | | CLASS | 0 | |
| 2 | Favorite... | | CLASS | 0 | |
| 3 | Members... | Members... | CLASS | 0 | |
| 4 | Satisfacti... | Satisfacti... | CLASS | 0.196078 | |
| 5 | Age | | VAR | 1.372549 | 2 |
| 6 | Average ... | Average ... | VAR | 0 | |
| 7 | Churn | | VAR | 0 | |
| 8 | Custome... | Custome... | VAR | 0 | 10 |
| 9 | Days Si... | Days Sin... | VAR | 0 | |
| 10 | Discount... | | VAR | 0 | |

Based on the graphs, there is opportunity to perform log transformation for skewed distribution, particularly 'Age' and 'Items_Purchased' attribute. Log transformation smoothens the data for better distribution. In handling the missing values, imputation method was selected by using the impute node as in SAS workflow diagram. In the edit variable section, the 'Use' column were changed to 'Yes' for 'Age' and 'Satisfaction_Level'. Considering the missing value type is missing completely at random (MCAR) as well as having low count (only 14 for 'Age' and 2 for 'Satisfaction_Level'), the method of imputation was set to Mean and Count respectively. Below are the changes made:



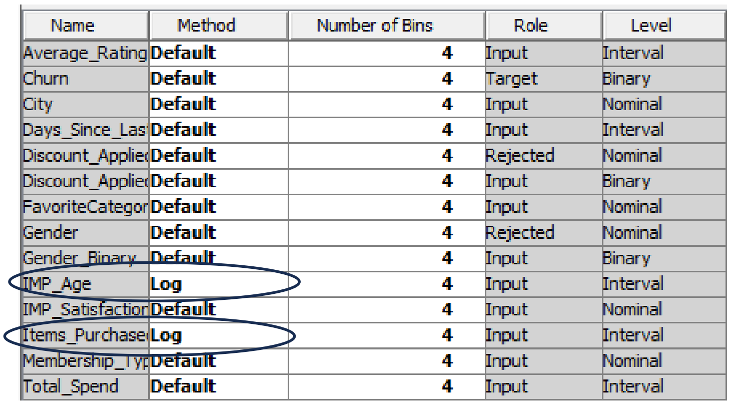
| Name | Use | Method | Use Tree | Role | Level |
|---|---|---|---|---|---|
| Age | **Yes** | **Mean** | **Default** | Input | Interval |
| Average_Ratin | **Default** | **Default** | **Default** | Input | Interval |
| Churn | **Default** | **Default** | **Default** | Target | Binary |
| City | **Default** | **Default** | **Default** | Input | Nominal |
| Days_Since_La | **Default** | **Default** | **Default** | Input | Interval |
| Discount_Appli | **Default** | **Default** | **Default** | Rejected | Nominal |
| Discount_Appli | **Default** | **Default** | **Default** | Input | Binary |
| FavoriteCatego | **Default** | **Default** | **Default** | Input | Nominal |
| Gender | **Default** | **Default** | **Default** | Rejected | Nominal |
| Gender_Binary | **Default** | **Default** | **Default** | Input | Binary |
| Items_Purchas | **Default** | **Default** | **Default** | Input | Interval |
| Membership_T | **Default** | **Default** | **Default** | Input | Nominal |
| Satisfaction_Le | **Yes** | **Count** | **Default** | Input | Nominal |
| Total_Spend | **Default** | **Default** | **Default** | Input | Interval |

Below is the output after running impute node:



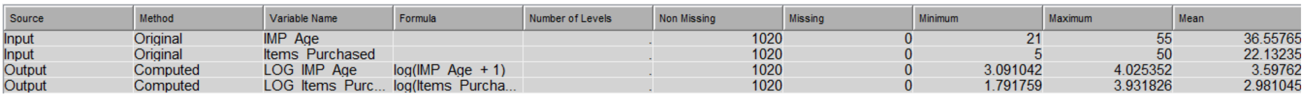
| Variable Name | Impute Method | Imputed Variable | Impute Value | Role | Measurement Level |
|---|---|---|---|---|---|
| Age | MEAN | IMP_Age | 36.557654076 | INPUT | INTERVAL |
| Satisfaction_Level | COUNT | IMP_Satisfaction_Level | Unsatisfied | INPUT | NOMINAL |

After that, 'Transform Variable' node was connected to normalize data via performing log transformation. This can be done on edit variable section of Transform variable node such as below:

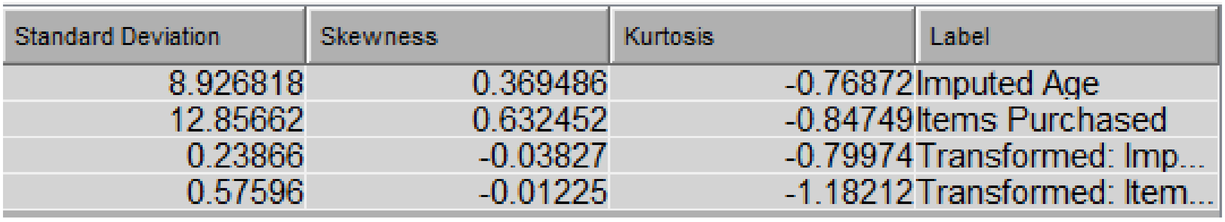


| Name | Method | Number of Bins | Role | Level |
|---|---|---|---|---|
| Average_Rating | **Default** | 4 | Input | Interval |
| Churn | **Default** | 4 | Target | Binary |
| City | **Default** | 4 | Input | Nominal |
| Days_Since_Las | **Default** | 4 | Input | Interval |
| Discount_Applie | **Default** | 4 | Rejected | Nominal |
| Discount_Applie | **Default** | 4 | Input | Binary |
| FavoriteCategor | **Default** | 4 | Input | Nominal |
| Gender | **Default** | 4 | Rejected | Nominal |
| Gender_Binary | **Default** | 4 | Input | Binary |
| IMP_Age | **Log** | 4 | Input | Interval |
| IMP_Satisfaction | **Default** | 4 | Input | Nominal |
| Items_Purchase | **Log** | 4 | Input | Interval |
| Membership_Typ | **Default** | 4 | Input | Nominal |
| Total_Spend | **Default** | 4 | Input | Interval |

The result of transformation is indicated as below:



| Source | Method | Variable Name | Formula | Number of Levels | Non Missing | Missing | Minimum | Maximum | Mean |
|---|---|---|---|---|---|---|---|---|---|
| Input | Original | IMP Age | | . | 1020 | 0 | 21 | 55 | 36.55765 |
| Input | Original | Items Purchased | | . | 1020 | 0 | 5 | 50 | 22.13235 |
| Output | Computed | LOG IMP Age | log(IMP Age + 1) | . | 1020 | 0 | 3.091042 | 4.025352 | 3.59762 |
| Output | Computed | LOG Items Purc... | log(Items Purcha... | . | 1020 | 0 | 1.791759 | 3.931826 | 2.981045 |

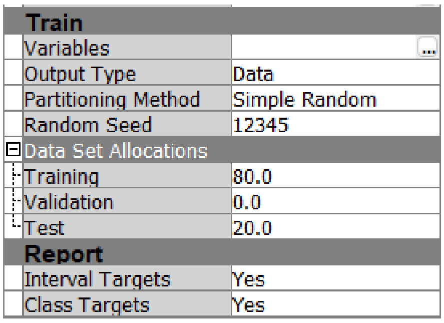


| Standard Deviation | Skewness | Kurtosis | Label |
|---|---|---|---|
| 8.926818 | 0.369486 | -0.76872 | Imputed Age |
| 12.85662 | 0.632452 | -0.84749 | Items Purchased |
| 0.23866 | -0.03827 | -0.79974 | Transformed: Imp... |
| 0.57596 | -0.01225 | -1.18212 | Transformed: Item... |

The reduction in standard deviation and skewness of indicates a better distribution of the transformed attributes. The higher negative value of Kurtosis for transformed 'Item_Purchased' attributes indicates fewer outliers and less extreme values in dataset.

## Data Partition process

Dataset must undergo partition into Training, validation and test set. Data partition node was dragged and connected to impute node from previous data preprocessing. In this case, dataset was only divide into training (80%) and test set (20%) due to limited dataset amount. Partitioning method was set to simple random method as every data points have equal chance to be selected, subsequently reducing biasness. Below are details of data partition process:

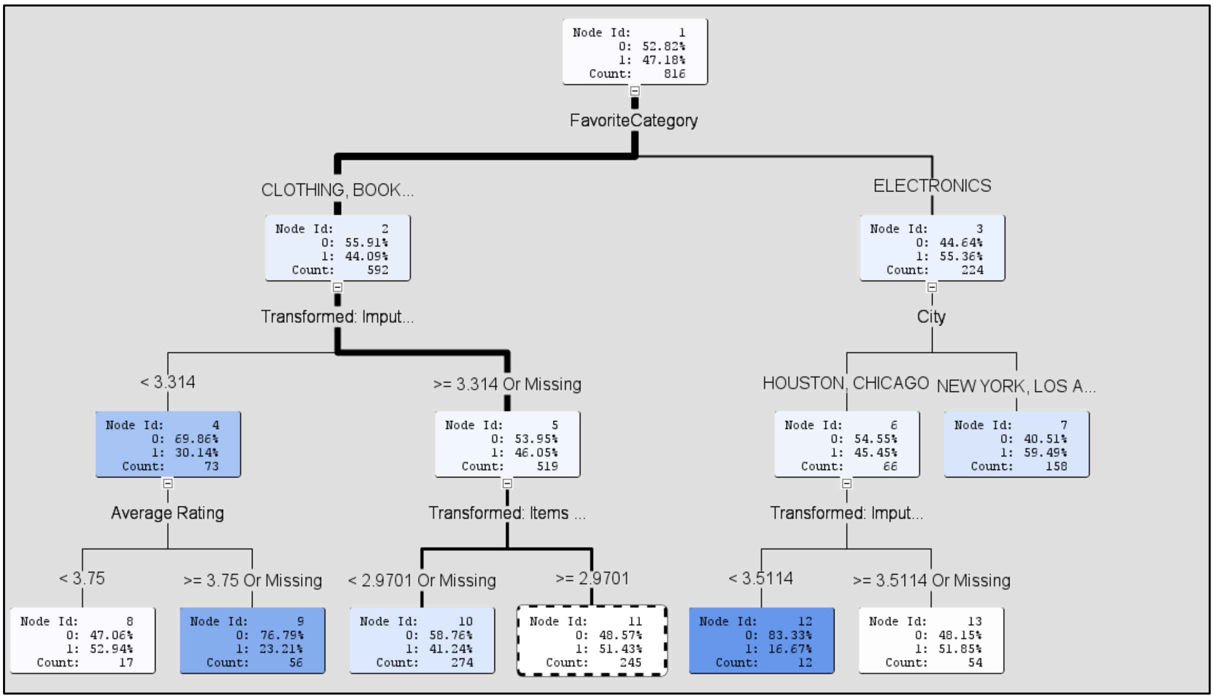


| | |
|---|---|
| **Train** | |
| Variables | ... |
| Output Type | Data |
| Partitioning Method | Simple Random |
| Random Seed | 12345 |
| ⊟ Data Set Allocations | |
| Training | 80.0 |
| Validation | 0.0 |
| Test | 20.0 |
| **Report** | |
| Interval Targets | Yes |
| Class Targets | Yes |

## Data Modelling and Analysis

## Decision Tree Model

Since the target variable is binary, Gini is used as target criterion as it is effective in binary classification as it is fast to be computed. The maximum depth was limited to 3 to provide main overview of the important details of customer behaviour. Below are the decision tree details:

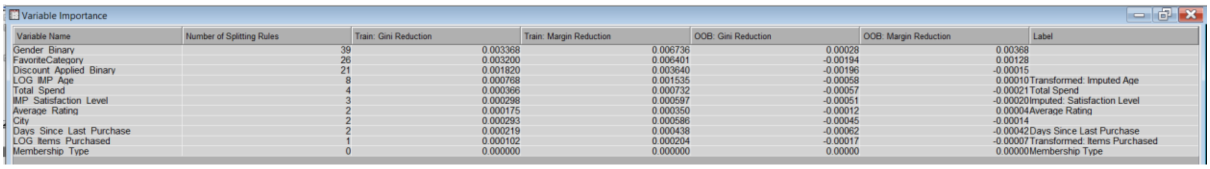


From the decision tree formed, this indicates that favorite category affects the most outcome of Churn. While electronics from category are mainly impacted by cities bought, other category mainly being impacted by age. While age is either impacted by average rating if log age <3.314, or number of items purchased if it is more than or equal to 3.314. From the cities that impacted electronics, Houston and Chicago city is dependent on age as well.

There a few strategies that can be developed from this model:

1. Target customers in specific cities like Houston and Chicago with electronics, as city location appears to be a significant factor for this category
2. For other categories, age seems to be the most significant factor. Marketing and product recommendations can be age-specific to cater to different preferences.
3. Lower age groups can focus on higher rating products as it is heavily influenced from it. For higher age groups, items purchased seems to be a significant factor. Strategies like bundle offers can be introduced for higher volume purchasing
4. Since electronics are more influenced based on cities, company can adjust stock inventory to these locations

## Bagging using High Performance Random Forest (HP Random Forest)



Variable Importance

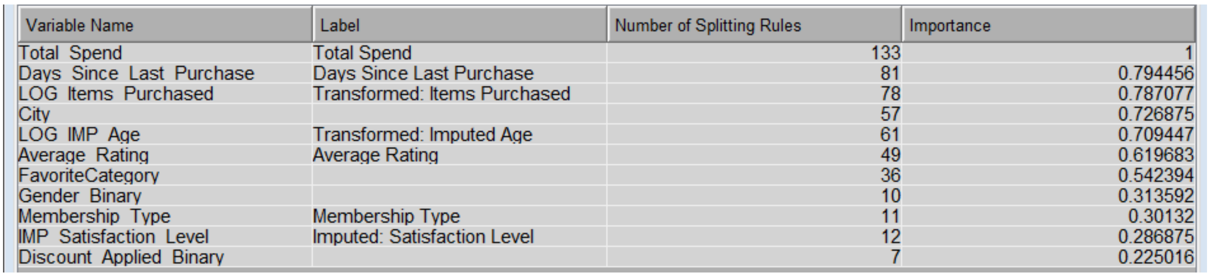
| Variable Name | Number of Splitting Rules | Train: Gini Reduction | Train: Margin Reduction | OOB: Gini Reduction | OOB: Margin Reduction | Label |
|---|---|---|---|---|---|---|
| Gender Binary | 39 | 0.003368 | 0.006736 | 0.00028 | 0.00368 | |
| FavoriteCategory | 26 | 0.003200 | 0.006401 | -0.00194 | 0.00128 | |
| Discount Applied Binary | 21 | 0.001820 | 0.003640 | -0.00196 | -0.00015 | |
| LOG IMP Age | 8 | 0.000768 | 0.001535 | -0.00058 | 0.00010 | Transformed: Imputed Age |
| Total Spend | 4 | 0.000366 | 0.000732 | -0.00057 | -0.00021 | Total Spend |
| IMP Satisfaction Level | 3 | 0.000298 | 0.000597 | -0.00051 | -0.00020 | Imputed: Satisfaction Level |
| Average Rating | 2 | 0.000175 | 0.000350 | -0.00012 | 0.00004 | Average Rating |
| City | 2 | 0.000293 | 0.000586 | -0.00045 | -0.00014 | |
| Days Since Last Purchase | 2 | 0.000219 | 0.000438 | -0.00062 | -0.00042 | Days Since Last Purchase |
| LOG Items Purchased | 1 | 0.000102 | 0.000204 | -0.00017 | -0.00007 | Transformed: Items Purchased |
| Membership Type | 0 | 0.000000 | 0.000000 | 0.00000 | 0.00000 | Membership Type |

The variable importance indicated that Gender_Binary, FavoriteCategory, Discount_Applied_Binary and Age are the top 4 attributes of the importance. These provides details of insights for strategic business such as:

1. Strong influence of gender bias indicates necessity of creating gender-based marketing campaigns
2. Promoting products that are categorical based on the customers/users' favourite. These can increase engagement of users and prevent churning
3. Analyse effectiveness of discounts and considering personal ones as well

## Boosting using Gradient Boosting Decision Trees

Gradient boosting tree is a great approach as it corrects the error of previous tree decision under a number of iterations until it reaches its minimum loss function. Different iterations were implied and tested to find its minimal globalization, which is around 189 iterations. Thus, it is used in case study. The variable indicator based on result running is as below:



| Variable Name | Label | Number of Splitting Rules | Importance |
|---|---|---|---|
| Total Spend | Total Spend | 133 | 1 |
| Days Since Last Purchase | Days Since Last Purchase | 81 | 0.794456 |
| LOG Items Purchased | Transformed: Items Purchased | 78 | 0.787077 |
| City | | 57 | 0.726875 |
| LOG IMP Age | Transformed: Imputed Age | 61 | 0.709447 |
| Average Rating | Average Rating | 49 | 0.619683 |
| FavoriteCategory | | 36 | 0.542394 |
| Gender Binary | | 10 | 0.313592 |
| Membership Type | Membership Type | 11 | 0.30132 |
| IMP Satisfaction Level | Imputed: Satisfaction Level | 12 | 0.286875 |
| Discount Applied Binary | | 7 | 0.225016 |

As indicated, the top 5 main factors influencing such result would be 'total spend', 'Day Since Last Purchase', 'Items Purchased', 'City' and 'Age'. Thus, such strategies that can be implied in business strategy would be:

1. Target customers who have spent more but have not purchased recently with personalized offers or reminders
2. Develop loyalty programs to reward repeat purchases, thereby reducing the likelihood of churn.
3. Develop strategies to local tastes and purchasing habits, possibly reflecting regional differences and different age

## Models' comparison based on performance metrics

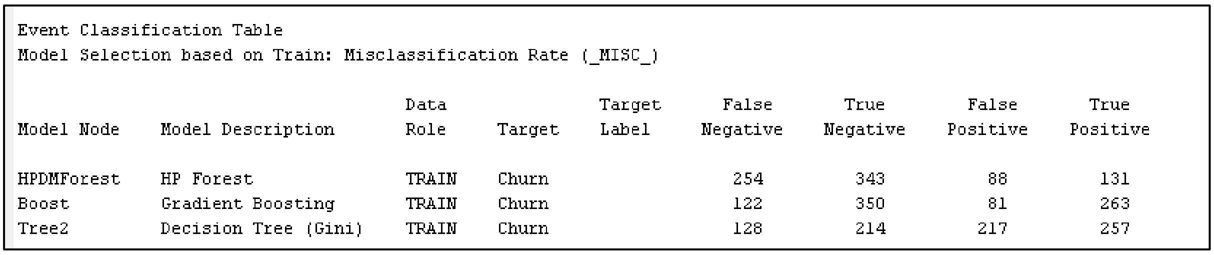
The model comparison node was utilized the compare the performance of tree models utilized

## Classification table

Classification table involves 4 labels, which are True Positive, True Negative, False Positive, False Negative. These are the description of these 4 labels in terms of churn analysis:

- True Positive (TP): Indicates the customers who were predicted to churn and did churn. It helps to understand the effectiveness of retention strategies targeted at at-risk customers.
- True Negative (TN): Represents customers who were predicted to stay and did stay. It shows the accuracy of the model in identifying loyal customers.
- False Positive (FP): Customers who were predicted to churn but did not churn. This could lead to unnecessary spending on retention efforts or incentives for customers who were not at risk
- False Negative (FN): Customers who were not identified as at-risk but churned. This is a missed opportunity for intervention to retain the customer

FN is the most important label in this aspect as it describes the failure of capturing potential customers that will churn. While other labels are important as well, FN is the main focus on this evaluation.



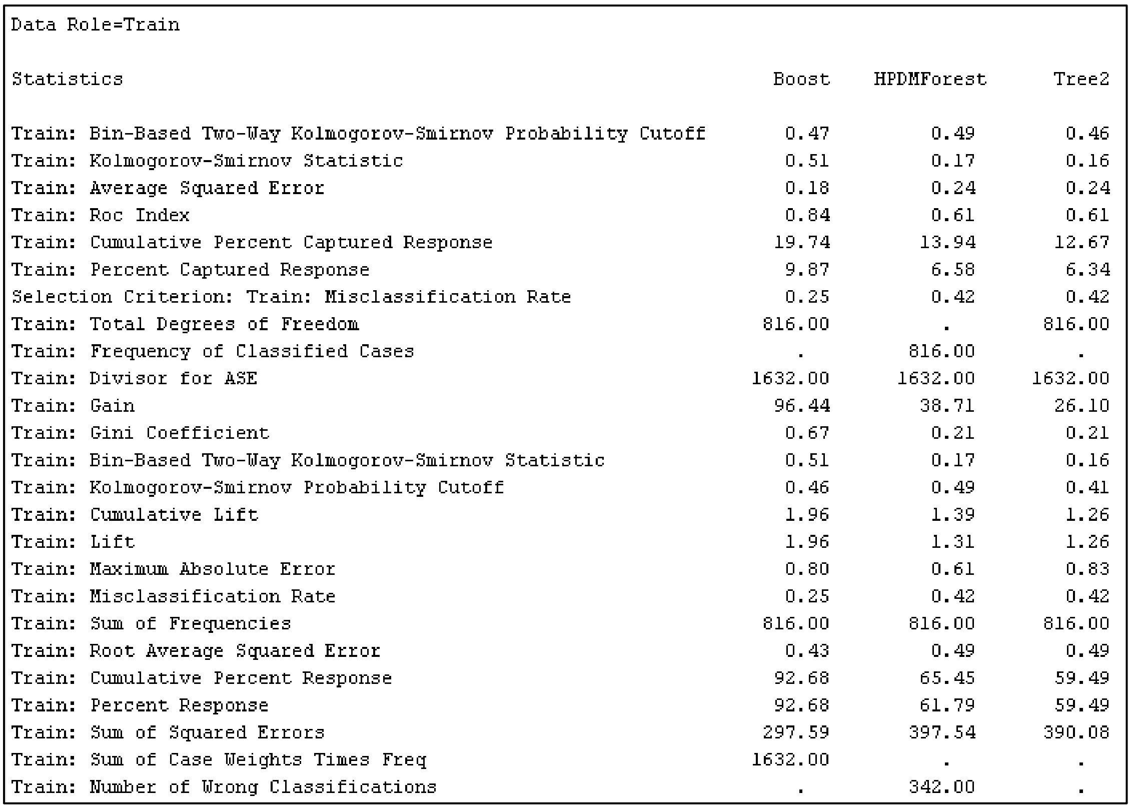
```
Event Classification Table
Model Selection based on Train: Misclassification Rate (_MISC_)

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

HPDMForest    HP Forest            TRAIN    Churn                254        343         88        131
Boost         Gradient Boosting    TRAIN    Churn                122        350         81        263
Tree2         Decision Tree (Gini) TRAIN    Churn                128        214        217        257
```

Based on training dataset classification table , HP Forest model is the least suitable model in predicting churn risk as it generates FN more than 2 times the amount of both Gradient Boosting and Decision Tree respectively. In addition, Gradient boosting is the most suitable model in prediction of churn risk with the lowest FN of 122 followed by Decision Tree with 128

## Evaluation Metrices

This will involve comparison based on few metrices based on model in both trained and test data

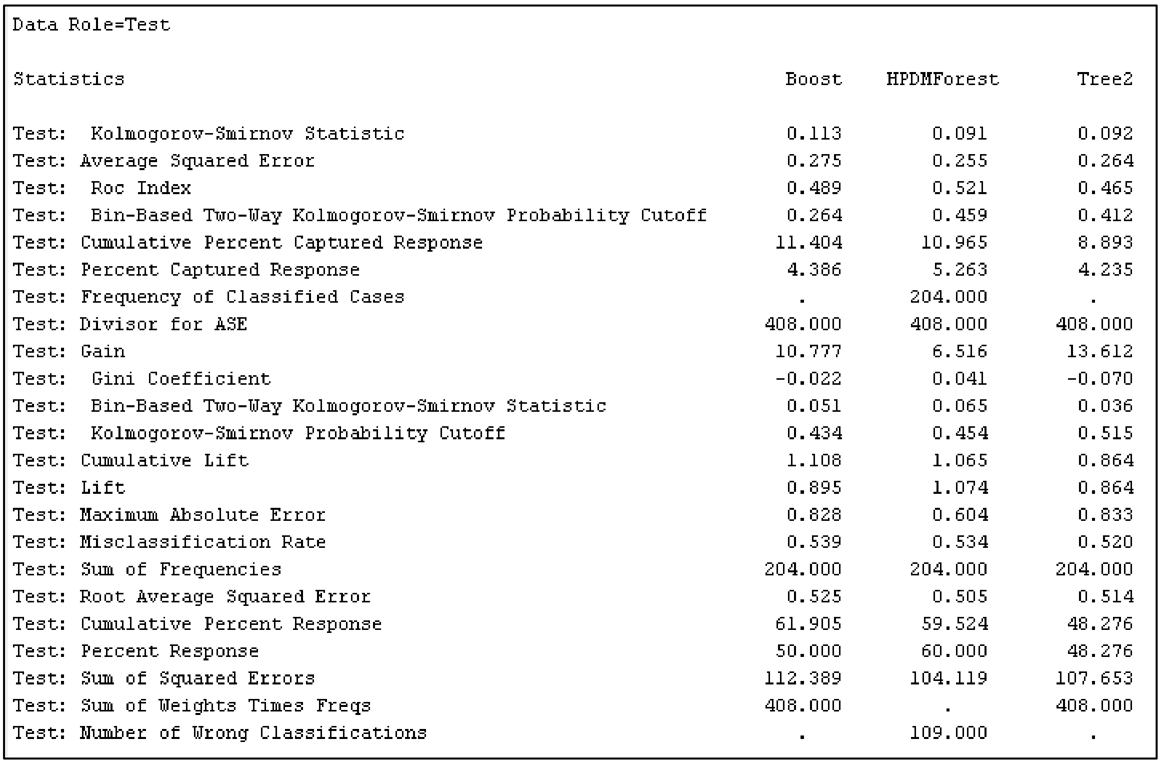
<u>Trained data</u>



```
Data Role=Train

Statistics                                                         Boost    HPDMForest       Tree2

Train: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff      0.47          0.49        0.46
Train: Kolmogorov-Smirnov Statistic                                 0.51          0.17        0.16
Train: Average Squared Error                                        0.18          0.24        0.24
Train: Roc Index                                                    0.84          0.61        0.61
Train: Cumulative Percent Captured Response                        19.74         13.94       12.67
Train: Percent Captured Response                                    9.87          6.58        6.34
Selection Criterion: Train: Misclassification Rate                  0.25          0.42        0.42
Train: Total Degrees of Freedom                                   816.00           .        816.00
Train: Frequency of Classified Cases                                 .          816.00         .
Train: Divisor for ASE                                           1632.00       1632.00     1632.00
Train: Gain                                                        96.44         38.71       26.10
Train: Gini Coefficient                                             0.67          0.21        0.21
Train: Bin-Based Two-Way Kolmogorov-Smirnov Statistic               0.51          0.17        0.16
Train: Kolmogorov-Smirnov Probability Cutoff                        0.46          0.49        0.41
Train: Cumulative Lift                                              1.96          1.39        1.26
Train: Lift                                                         1.96          1.31        1.26
Train: Maximum Absolute Error                                       0.80          0.61        0.83
Train: Misclassification Rate                                       0.25          0.42        0.42
Train: Sum of Frequencies                                         816.00        816.00      816.00
Train: Root Average Squared Error                                   0.43          0.49        0.49
Train: Cumulative Percent Response                                 92.68         65.45       59.49
Train: Percent Response                                            92.68         61.79       59.49
Train: Sum of Squared Errors                                      297.59        397.54      390.08
Train: Sum of Case Weights Times Freq                            1632.00           .           .
Train: Number of Wrong Classifications                               .          342.00         .
```

Overall, Gradient Boosting decision tree is significantly the better model compared to decision tree and HP random forest as it has a relatively lower misclassification rate, sum of squared error, as well as higher gini coefficient in comparison to HP Random Forest and Decision Tree model. This indicates that gradient boosting likely better to handle noise in data, overfitting issues, and provide better prediction accuracy.

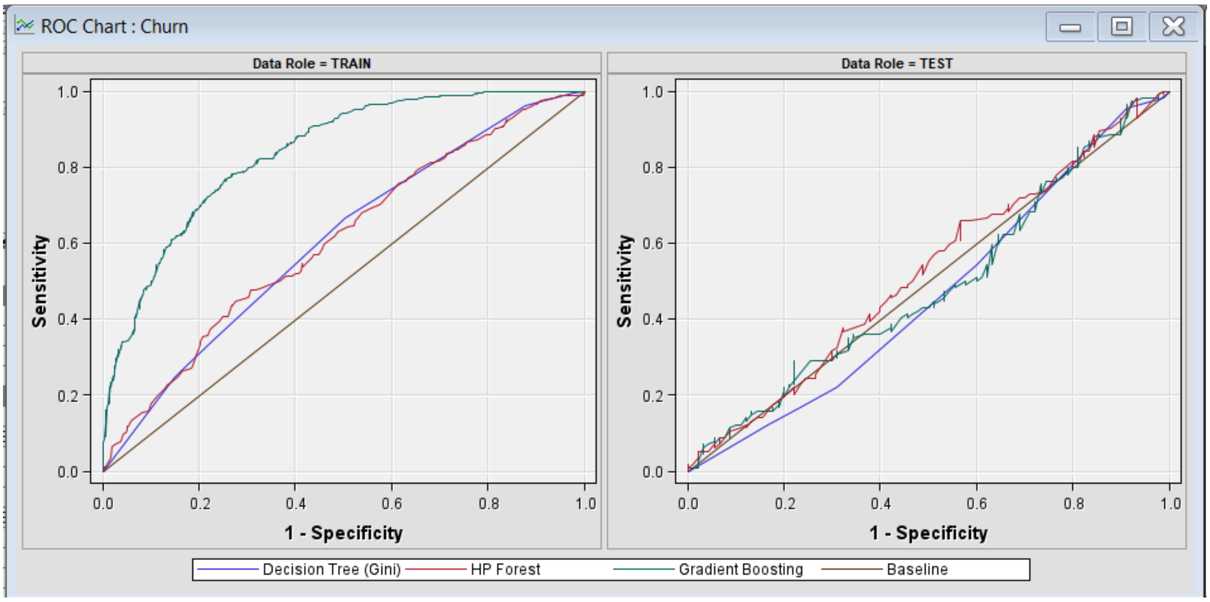
Test data



```
Data Role=Test

Statistics                                                     Boost    HPDMForest       Tree2

Test:  Kolmogorov-Smirnov Statistic                            0.113         0.091       0.092
Test: Average Squared Error                                    0.275         0.255       0.264
Test:  Roc Index                                               0.489         0.521       0.465
Test:  Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff 0.264         0.459       0.412
Test: Cumulative Percent Captured Response                    11.404        10.965       8.893
Test: Percent Captured Response                                4.386         5.263       4.235
Test: Frequency of Classified Cases                              .         204.000         .
Test: Divisor for ASE                                        408.000       408.000     408.000
Test: Gain                                                    10.777         6.516      13.612
Test:  Gini Coefficient                                       -0.022         0.041      -0.070
Test:  Bin-Based Two-Way Kolmogorov-Smirnov Statistic          0.051         0.065       0.036
Test:  Kolmogorov-Smirnov Probability Cutoff                   0.434         0.454       0.515
Test: Cumulative Lift                                          1.108         1.065       0.864
Test: Lift                                                     0.895         1.074       0.864
Test: Maximum Absolute Error                                   0.828         0.604       0.833
Test: Misclassification Rate                                   0.539         0.534       0.520
Test: Sum of Frequencies                                     204.000       204.000     204.000
Test: Root Average Squared Error                               0.525         0.505       0.514
Test: Cumulative Percent Response                             61.905        59.524      48.276
Test: Percent Response                                        50.000        60.000      48.276
Test: Sum of Squared Errors                                  112.389       104.119     107.653
Test: Sum of Weights Times Freqs                             408.000          .        408.000
Test: Number of Wrong Classifications                            .         109.000         .
```

However, for test dataset, Gradient Boost have a higher misclassification rate, sum of squared errors with lower gini coefficient compared to the other two models. This might indicate the model being overfitting. While Decision Tree has the lowest misclassification rate providing a better prediction accuracy, HP Random Forest produces the highest gini coefficient and lowest sum of squared errors.

## ROC Chart



ROC curve provides indication on models predictive ability with higher area under curve (AUC) indicates better ability of predicting. In train dataset, it is depicted that all three models AUC is above the baseline, indicating its reliability on developing a reliable prediction model. However, different trend was seen in test dataset, with Gradient boosting and Decision Tree model covers below the baseline with only HP random forest barely covering above the baseline. This indicates poor performance of prediction test dataset. This indicates high possibility of outliers and noise as models learned the data too well, and unable to generalize to new and unseen data. Generalizability issues are related a few aspects such as:

1. Limited amount of dataset to capture overall representation of the analysis
2. High complexity of the model, which needs to be reduced
3. Lack of more generalized training methods such as utilization of cross validation technique

Thus, these limitations can be address for further studies

# Conclusion

Based on the data preprocessing and modelling, certain insights can be gained from the customer behaviour analysis which includes:

- Localized and Demographic Targeting: Develip marketing strategies for electronics in cities like Houston and Chicago and align product recommendations with age-specific preferences for other categories.
- Personalized Engagement: Implement gender-specific campaigns and promote items based on individual customer's favourite categories to enhance user engagement.
- Incentive Programs: Conduct analysis on the significance of discounts and introduce personalized offers, reminders, and loyalty programs to encourage repeat purchases and higher spend.
- Inventory and Marketing Optimization: Adjust inventory such as electronics in specific cities and use bundle offers for higher age groups to incentivize volume purchases.

Based on the performance evaluation, Gradient boosting decision tree provides the best model on training dataset. However, due to possibility of outliers and noise, as well as limited amount of dataset, the model provided is overfitted. This indicates further study with dataset with better representation, different models of different complexities, as well as different training approach such as cross validation technique. Limitation of this studies include time-constraint of implying pre-processing techniques such as sequence analysis and associate rule mining, as well as multiple sources of dataset to increase generalizability of model developed.