

Data Import and Preprocessing:

1. Preprocessing:

The original dataset is from

<https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis/code>, the data is modified to the dataset which contain 49673 rows and 10 columns:

1. CustomerID: Unique identifier for each customer.
2. Age: Age of the customer
3. Gender: Gender of the customer
4. Location: Geographic location of the customer.
5. MembershipLevel: Indicates the membership level (e.g., Bronze, Silver, Gold, Platinum). [this is created based on the 'TotalSpent']
6. TotalPurchases: Total number of purchases made by the customer.[this is created by total up the number of purchases for each customer)
7. TotalSpent: Total amount spent by the customer.
8. FavoriteCategory: The category in which the customer most frequently shops (e.g., Electronics, Clothing, Home Goods).
9. LastPurchaseDate: The date of the last purchase.
10. FrequencyOfWebVisit: Consider adding more attributes like customer's occupation, frequency of website visits, etc.
11. Churn: Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

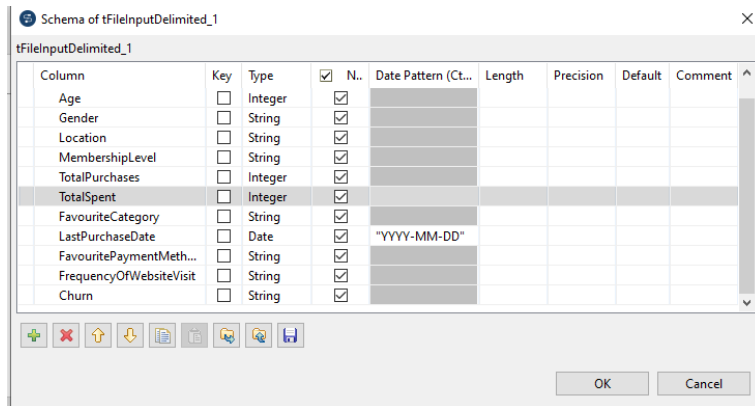
a. Using Talend Data Integration

-Import file into tFileInputDelimited

THEN TSZE YEN S2194020



-Set the Date format in “YYYY-MM-DD” so it can be accessed into SAS DataSource



-Get the output using tFileOutputDelimited_1

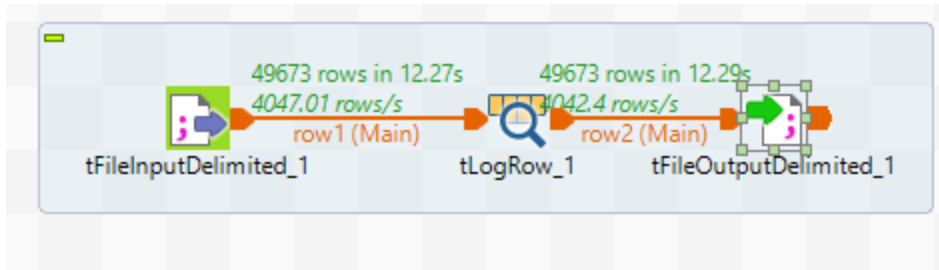
THEN TSZE YEN S2194020



- Check the columns

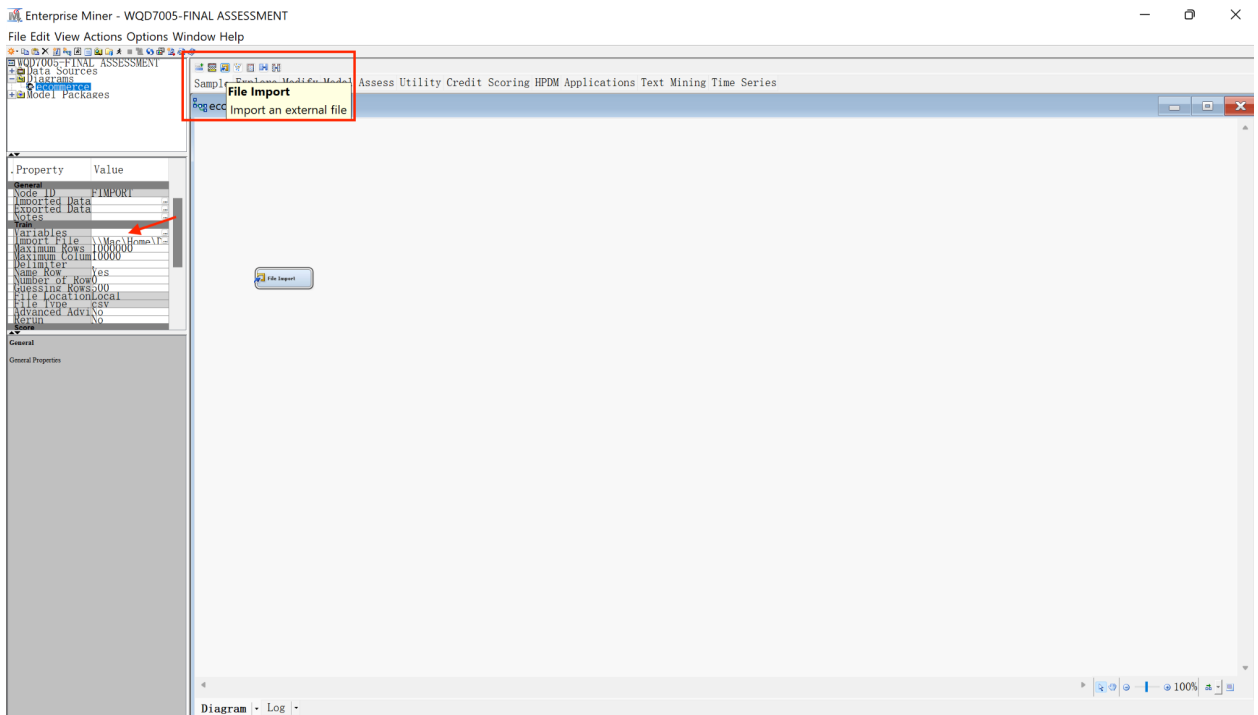
- Output in the form of csv file.

THEN TSZE YEN S2194020




Data Import

- Go to 'Sample', find 'File Import', then in Train panel, go to 'Import File' to import data.



Go to 'Data Source', create a data source.

THEN TSZE YEN S2194020

 Data Source Wizard -- Step 5 of 8 Column Metadata ✕

(none) ☐ not Equal to ☐ ... Apply Reset


Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Input	Interval	No		No	.	.
CustomerID	Input	Interval	No		No	.	.
FavouriteCat	Input	Nominal	No		No	.	.
FavouritePay	Input	Nominal	No		No	.	.
FrequencyOf	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchase	Time ID	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipLe	Input	Nominal	No		No	.	.
TotalPurchas	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

Show code Explore Compute Summary < Back Next > Cancel

2. Specify roles for the variables.

a. Assign roles

 Variables - FIMPORT

(none) ☐ not Equal to ☐ ... ☐ Mining ☐

Columns: ☐ Label

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Interval	No		No	.	.
Customer ID	Input	Interval	No		No	.	.
FavouriteCategory	Input	Nominal	No		No	.	.
FavouritePaymentMethod	Input	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
LastPurchaseDate	Time ID	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipLevel	Input	Nominal	No		No	.	.
Returns	Input	Interval	No		No	.	.
TotalPurchases	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

b. DataPartition: Split data into Training and Validation data.

-Training data: 70%

-Validation data: 30%

THEN TSZE YEN S2194020

Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning	Default
Random Seed	12345
Data Set Allocation	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	
Class Targets	
Status	
Create Time	1/7/24 7:04 AM

```
graph LR; DF_EDATA1 --> DataPartition
```

c. Impute missing values

Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Value	
Missing Value	0
Missing Value Surrogate	
Default Large None	
Default Large	
Default Large Median	
Default Large None	
Default Large	
Default Large	

```
graph LR; DF_EDATA1 --> DataPartition; DataPartition --> Impute
```

-There are missing values in ‘FrequencyOfWebsiteVisits

Imputation Summary

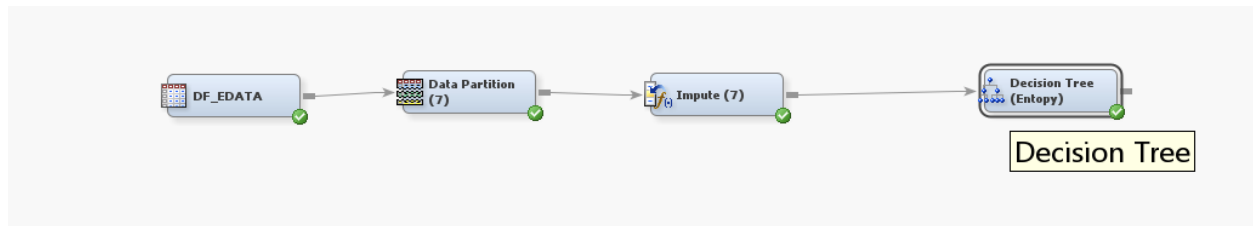
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
FrequencyOfWebsiteVisits	MEDIAN	IMP_FrequencyOfWebsiteVisits	93	INPUT	INTERVAL		1047

Variable Distribution Training Data

Obs	Number of Missing for TRAIN	Number of Variables	Percent of Variables
1	1047	1	100

Decision Tree Analysis:

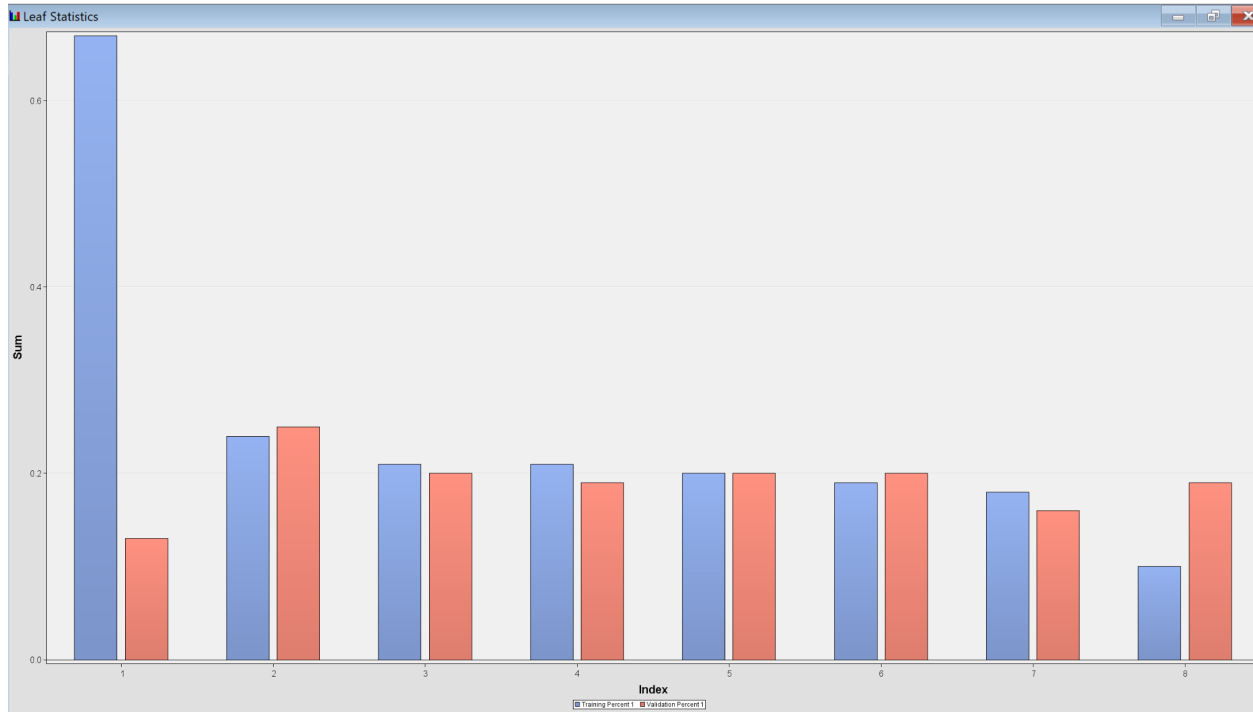


Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid:	Train:	Train:	Valid:
			Misclassification Rate	Average Squared Error	Misclassification Rate	Average Squared Error
Y	Tree10	Decision Tree (Entropy)	0.20023	0.16007	0.20012	0.16014

The misclassification rate for the decision tree is 0.20012 which considered as low.

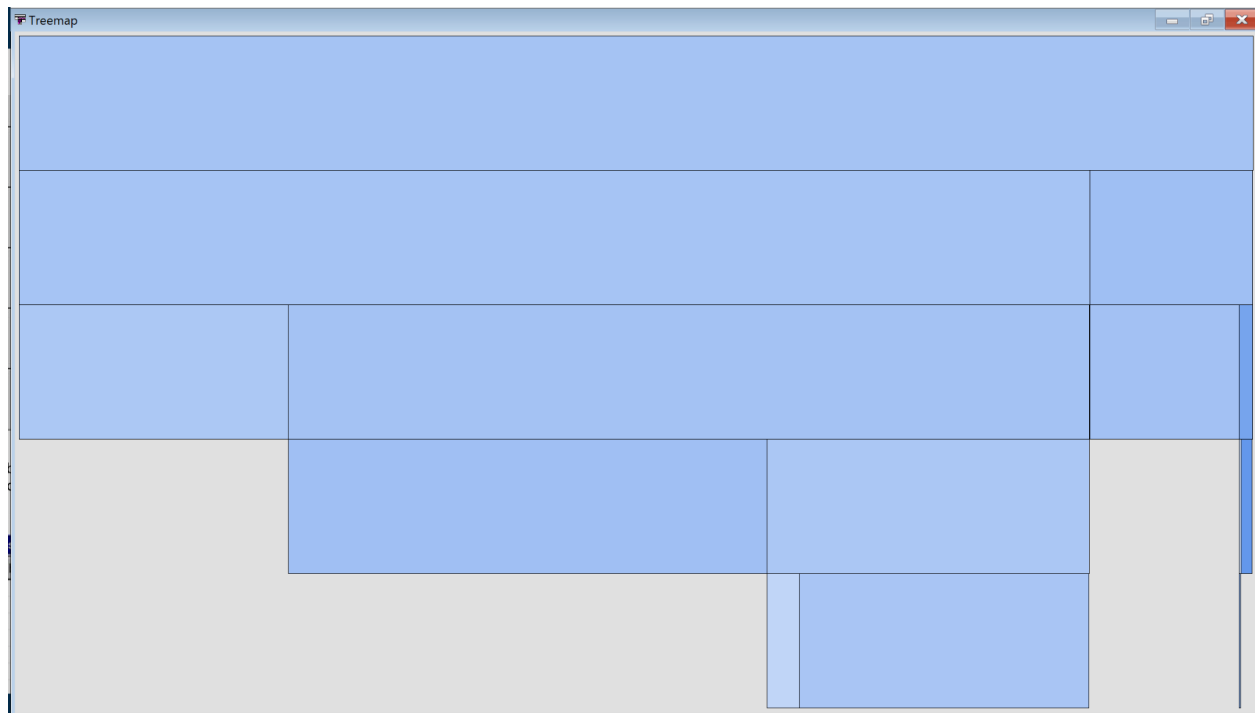


The Leaf Statistics shows two bars for each leaf index, one blue and one red. The blue is for "Training Percent" while the red bars represent the "Validation Percent." This interprets which data is split into training and validation set.

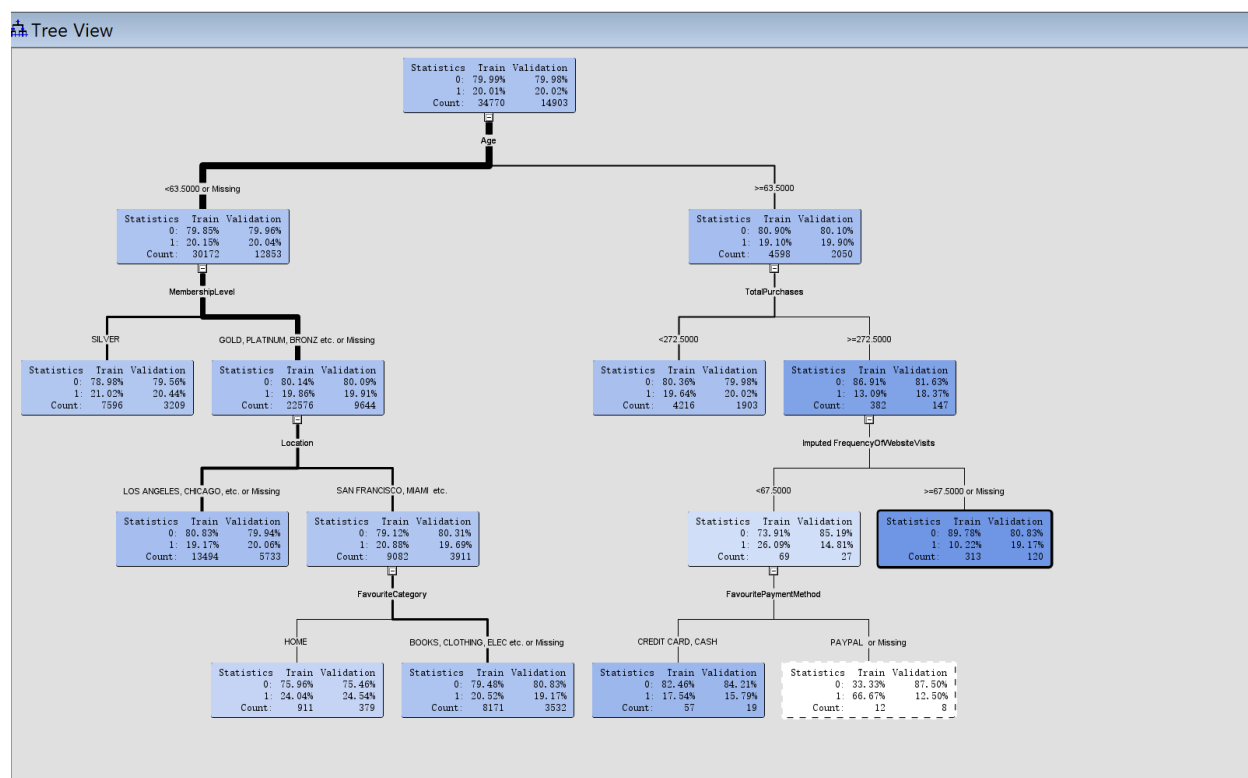
The leaf indices along the x-axis (ranging from 1 to 8) correspond to individual leaves of the decision tree or individual trees in the forest. The y-axis represents the sum of the metric being measured. The specific metric is not labeled, but it could be error, accuracy, or another performance measure.

From the chart, it can be observed that Leaf 1 has the highest sum for both training and validation, while other leaves show varying sums. The training and validation sums are relatively close for each index, which may indicate that the model generalizes well from the training data to unseen validation data. However, without more context or a defined metric, we cannot draw specific conclusions about the model's performance.

The purpose of such a chart could be to identify which leaves or trees are contributing the most to the model's performance and to check for overfitting or underfitting. Overfitting would be indicated if the training bars were significantly higher than the validation bars, suggesting the model performs well on training data but not on validation data. This chart appears to show a relatively balanced performance between the two datasets.



The color variation in a treemap usually represents a category or a range of values, but since all visible rectangles here are the same shade of blue which with one rectangle unfilled, this indicates that there is no additional categorical data indicated in this visualization.



At the root of the tree, we see a node split by the "Age" feature, dividing the data into two branches: one for data points with an "Age" less than 63.50 (or missing this data) and the other for those with an "Age" greater than or equal to 63.50. Each branch shows the proportions of data in the train and validation sets, followed by the count of data points in each.

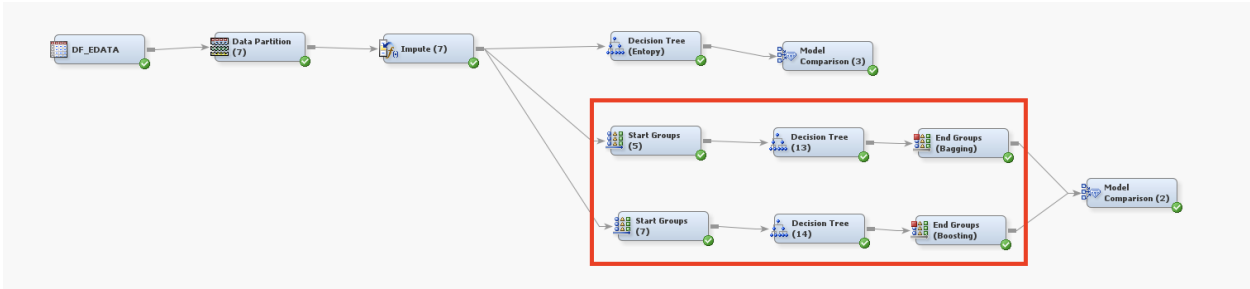
From these initial branches, the tree further splits based on other features such as "MembershipLevel," "TotalPurchases," "Location," "ImputedFrequencyOfWebsiteVisits," "FavouriteCategory," and "FavouritePaymentMethod." For example, under "MembershipLevel," the data is divided into categories such as "SILVER" and "GOLD, PLATINUM, BRONZE etc. or Missing." In "Silver", there are only 21.02% of the customer to be churned. For "GOLD, PLATINUM, BRONZE etc. or Missing," there 80.14% of customers are active.

Further down the tree, there are splits based on "Location" which distinguish between places like "LOS ANGELES, CHICAGO, etc. or Missing" and "SAN FRANCISCO, MIAMI etc." This indicates that location is a significant feature that affects the decision-making process in this model. There are 79.12% of the customers are active which often buying things only. What is more, the "FavouriteCategory" the customer preferred is "HOME" which computes of 75.96% only itself, whereas for "BOOKS, CLOTHING, ELEC,etc or Missing" computes 79.48% of customer from total all.

For "TotalPurchases", there are more than 275 items are bought by 382 customers with 80.36% of them are active customers. For "Imputed FrequencyOfWebsiteVisit", there are more than 67.5 times of the customer visit the store, 86.91% out of them are active customers. For the customer visit the store not that frequent which less than 67.5 time, their "FavouritePaymentMethod" is by "CREDIT CARD", and "CASH".

Ensemble Method:

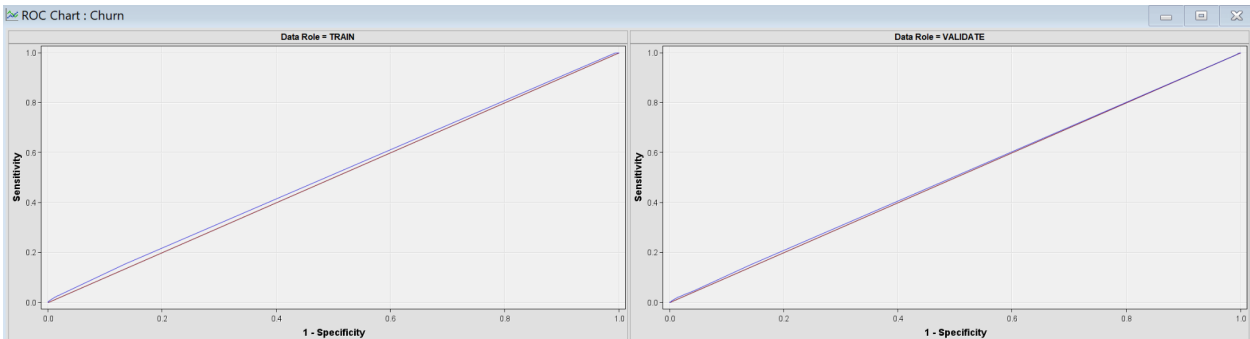
Diagram Flow of Bagging and Boosting Ensemble Models with Comparison



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

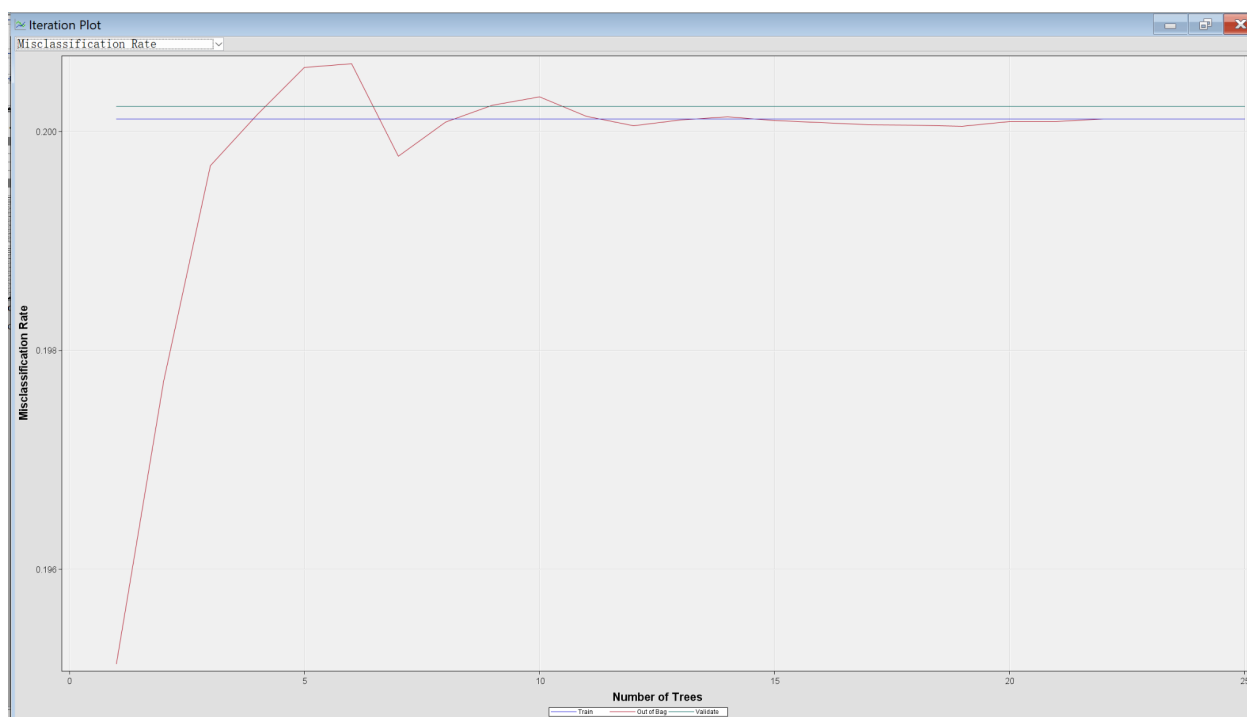
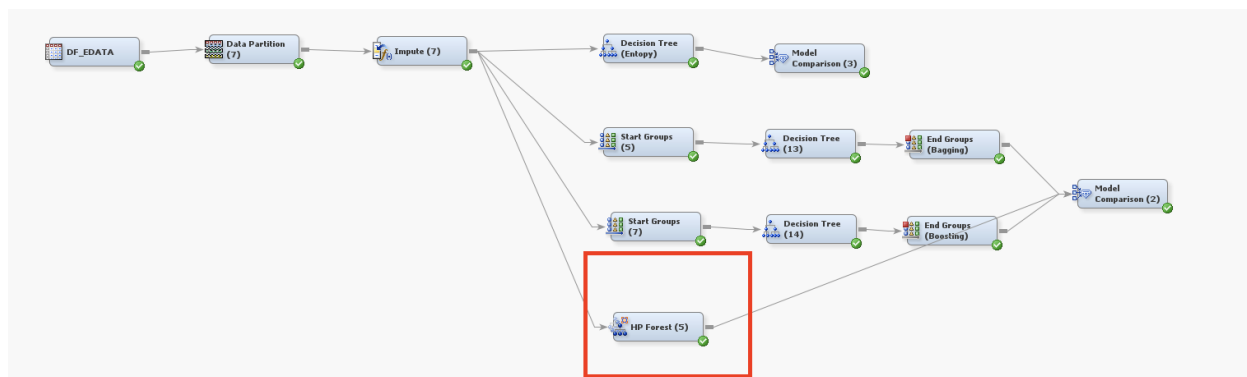
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	EndGrp6	End Groups (6)	0.20023	0.16008	0.20012	0.16014
	EndGrp8	End Groups (8)	0.20023	0.21987	0.20012	0.21998

The misclassification rate for End Groups (6) for Bagging Emsemble Method is 0.20012, and for End Groups (8) for Boosting is 0.20012.

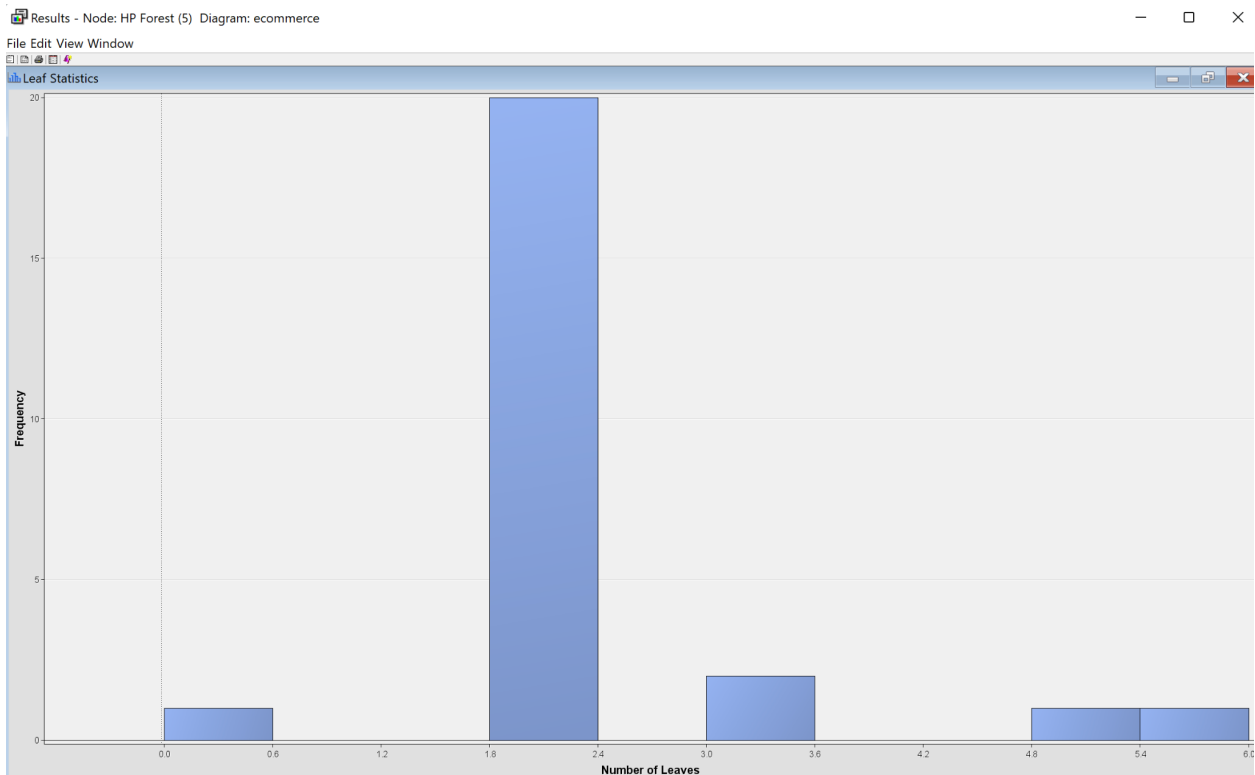


The ROC curve for this model is not desirable, as it is straightline for both Train and Validate which indicates with no predictive power.

Diagram of Random Forest



By looking at this iteration plot, the train data and validate data are both straight line. For Out of the bag, the misclassification rate starts relatively high and drops sharply after a few trees are added. Then the rate appears to stabilize. However, since the plot lines are very close together and relatively flat after the initial drop, this suggests that the model quickly reaches a point where adding more trees does not improve the performance significantly. It's also indicative that the model is not overfitting, as the lines for the training and validation errors remain close to each other without a large gap.



The histogram has bars at 2, 3, and 6, which shows the count of trees with that many leaves in the ensemble. The tallest bar is at 2 leaves, indicating that most of the trees in this ensemble have 2 leaves. There are fewer trees with 3 leaves and even fewer with 6 leaves. A decision tree with only 2 leaves is a very simple model, suggesting that the ensemble contains many simple trees. The presence of trees with more leaves indicates varying complexity within the ensemble.

Diagram Comparison of 4 models:

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected	Model	Model Node	Model Description	Valid:	Train:	Train:	Valid:
				Misclassification Rate	Average Squared Error	Misclassification Rate	Average Squared Error
Y	Tree10		Decision Tree (Entropy)	0.20023	0.16007	0.20012	0.16014
	HPDMForest5		HP Forest (5)	0.20023	0.15999	0.20012	0.16014
	EndGrp5		End Groups (Bagging)	0.20023	0.16008	0.20012	0.16014
	EndGrp7		End Groups (Boosting)	0.20023	0.21987	0.20012	0.21998

For comparison between Random Forests, Bagging Ensembles, Boosting Ensembles and Random Forest, it shows all the model have the same misclassification rate of 0.20012 value and for Average Squared Error, the Random Forest model has the lowest value.

Deliveries

In Data Preparation, the initial phase involved importing and preprocessing the data. Talend Data Integration was employed to set the correct date format and convert the data into a usable CSV format. The dataset was then split into a training set, comprising 70% of the data, and a validation set, containing the remaining 30%. This division was essential for training the model and subsequently validating its predictions. Variables were assigned specific roles, and any missing values, particularly in the 'FrequencyOfWebsiteVisits' field, were imputed based on the number of purchases done by the customer.

A decision tree model was constructed, yielding a misclassification rate of 0.20012. This low rate indicated the model's effectiveness in fitting the data. Leaf statistics provided insights into the model's complexity, suggesting an appropriate balance between model simplicity and predictive power, which is indicative of the model's ability to generalize beyond the training data.

Key insights emerged from the decision tree analysis. The 'Age' feature was identified as a primary node, highlighting its significance in predicting customer churn. Additional variables such as 'MembershipLevel', 'TotalPurchases', 'Location', 'ImputedFrequencyOfWebsiteVisits', 'FavouriteCategory', and 'FavouritePaymentMethod' also influenced the model, with 'Location' and 'MembershipLevel' emerging as particularly impactful. Notably, customers with a 'Silver' membership level exhibited a lower churn rate.

The project also explored ensemble methods, including bagging and boosting, to compare their efficacy. Both methods showed a misclassification rate mirroring that of the decision tree. However, the ROC curves for training and validation sets displayed no predictive power, which was a point of concern. The Random Forests method, in particular, demonstrated an iteration plot where the misclassification rate decreased sharply with the addition of a few trees before stabilizing, implying no overfitting.

This analysis faced challenges, particularly in handling missing data and avoiding model overfitting. Another significant challenge was interpreting the ROC curve effectively to enhance the model's predictive capabilities. The dataset is imbalanced and the output for the model is not desirable; this is due to the over-large dataset with 49673 rows, a sampling method should be done before the data partition. More an excessive bias in the dataset can affect the accuracy of the model overall, viewing the ROC curve, it is a straightline indicating no predictive power. Future suggestions of this study is reducing the data size.

Based on the findings, several strategic recommendations can be made. Focusing on 'Age' and 'MembershipLevel' could lead to more effective customer retention strategies. The lower churn

rate among 'Silver' members warrants further investigation to understand and leverage these insights across other membership tiers. Additionally, the regional differences suggested by the 'Location' data point to opportunities for localized marketing strategies. Lastly, it is crucial to keep the model up-to-date with new data, ensuring its ongoing relevance and adaptability to changing customer behaviors.