Project: Determining the Format for the New Stores and Forecasting Produce Sales

Task 2: Formats for New Stores

To predict the best store format for the new stores, I used classification Model as we have a categorical target variable. And since our categorical target variable has more than 2 categories I used the multinomial models.

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Decision_Tree_Capstone	0.6471	0.6667	0.5000	1.0000	0.5000
Random_Forest_Capstone	0.7647	0.7917	0.6250	1.0000	0.7500
Boosted_Model_Capstone	0.7647	0.8333	0.5000	1.0000	1.0000

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_Model_Capstone

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	0
Predicted_2	2	5	0
Predicted_3	2	0	4

Confusion matrix of Decision_Tree_Capstone

	Actual_1	Actual_2	Actual_3
Predicted_1	4	0	2
Predicted_2	3	5	0
Predicted_3	1	0	2

Confusion matrix of Random_Forest_Capstone

	Actual_1	Actual_2	Actual_3
Predicted_1	5	0	1
Predicted_2	1	5	0
Predicted_3	2	0	3

Model comparison reports above indicates that the two highest accuracy scores belong to **Boosted Model** and **Random Forest Model** both being **76.47%.** I also looked at the **F1 score** which indicates the model precision. This score ranges from 0 to 1 and higher values indicates higher precision. As a result, I chose the boosted model as the best, among others.

Then I applied the **Score tool** to get the scores for the new 10 stores and then I determined the clusters for the new stores using the values I obtained from the Score tool's analysis. The results below indicate what format each of the 10 new stores falls into.

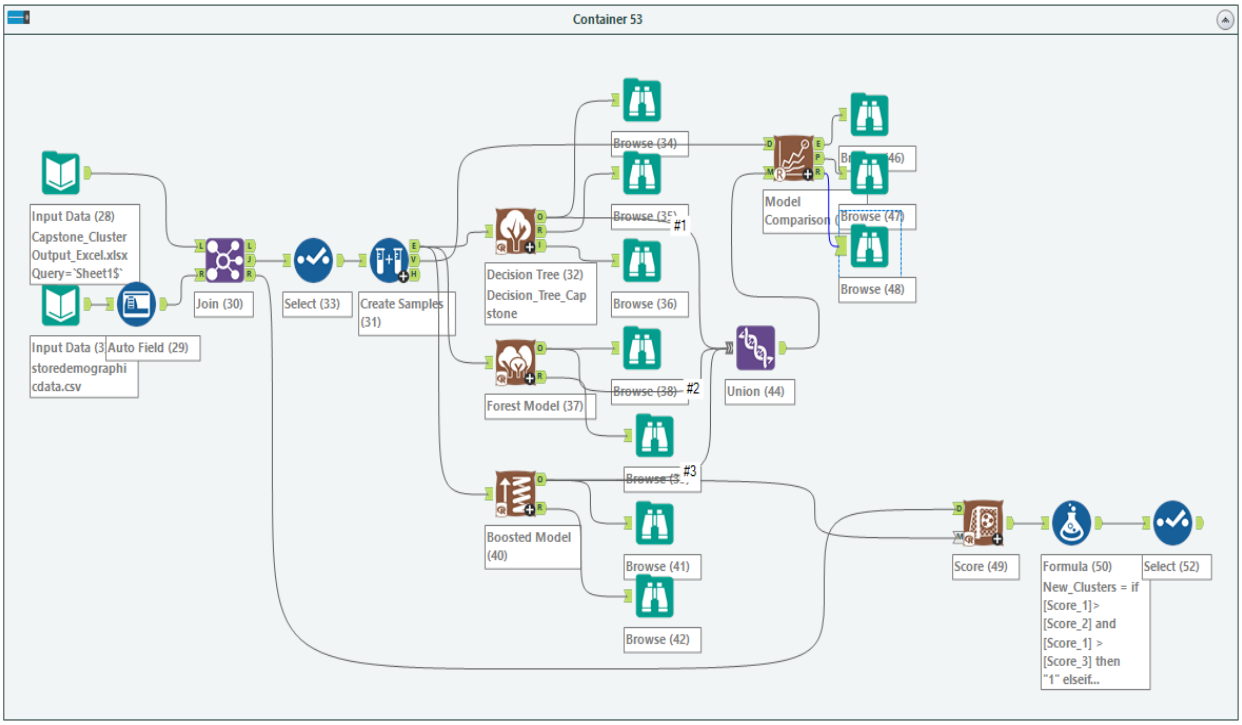
Results - Select (52) - Output

2 of 2 Fields ▾ | Cell Viewer ▾ | 10 records displayed

Record	Store	New_Clusters
1	S0086	1
2	S0087	2
3	S0088	3
4	S0089	2
5	S0090	2
6	S0091	3
7	S0092	2
8	S0093	3
9	S0094	2
10	S0095	2

Store Number	Segment
S0086	1
S0087	2
S0088	3
S0089	2
S0090	2
S0091	3
S0092	2
S0093	3
S0094	2
S0095	2

Below is the Alteryx workflow for Task 2.



To identify the most important variables that help explain the relationship between demographic indicators and store formats I evaluated the **Variable Importance Plot below**. Accordingly, the most important three variables are "*Age 0to9*", "*HVal750KPlus*", "*Age65Plus*"

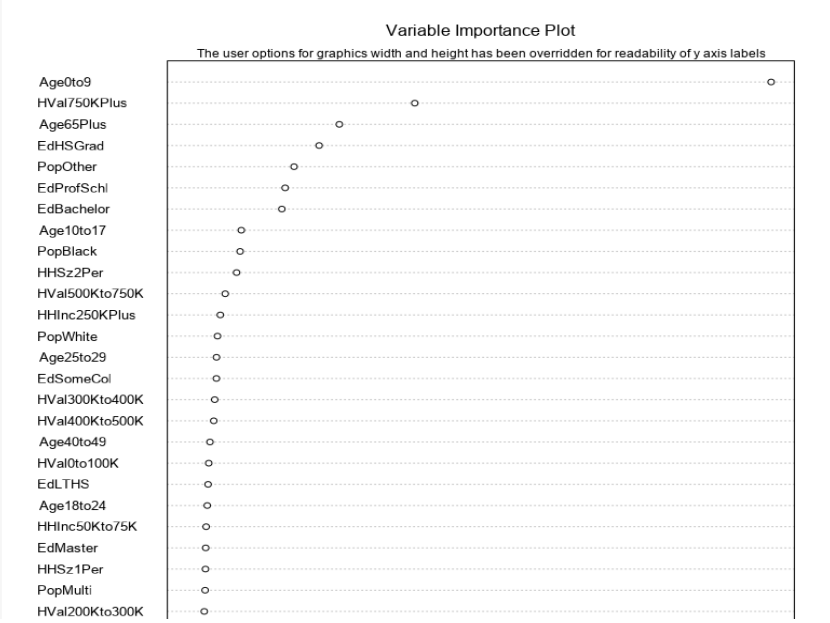
Report

Report for Boosted Model Boosted_Model_Capsto

Basic Summary:

Loss function distribution: Multinomial
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1829

Plots:



Kadir AKYUZ, Ph.D.