**Project: Predictive Analytics**

## **Task 1: Determine Store Formats for Existing Stores**

1. What is the optimal number of store formats? How did you arrive at that number?

The optimal number of store formats is 3. I used the k-centroids cluster analysis model, and the model clusters the stores into 3 segments.

1. How many stores fall into each store format?

In cluster 1, there are 25 stores.

In cluster 2, there are 35 stores.

In cluster 3, there are 25 stores.

1. Based on the results of the clustering model, what is one way that the clusters differ from one another?

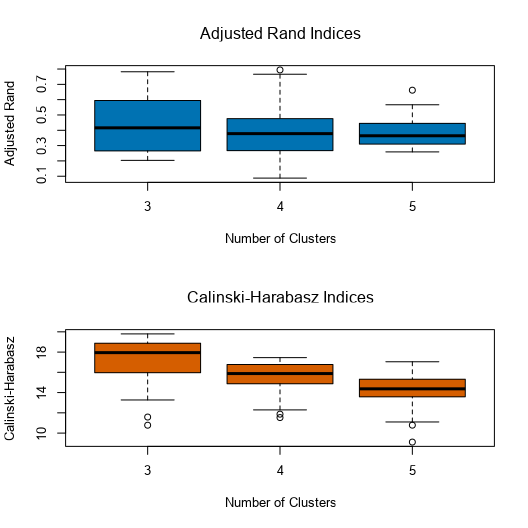
|  |
| --- |
|  |
| 5 | | Cluster | Size | Ave Distance | Max Distance | Separation | | --- | --- | --- | --- | --- | | 1 | 25 | 2.099985 | 4.823871 | 2.191566 | | 2 | 35 | 2.475018 | 4.412367 | 1.947298 | | 3 | 25 | 2.289004 | 3.585931 | 1.72574 | |
| 6 | Convergence after 8 iterations.  Sum of within cluster distances: 196.35034. |
| 7 | |  | Dry. | Dairy | Frozen.Food | Meat | Produce | Floral | Deli | | --- | --- | --- | --- | --- | --- | --- | --- | | 1 | 0.528249 | -0.215879 | -0.261597 | 0.614147 | -0.655028 | -0.663872 | 0.824834 | | 2 | -0.594802 | 0.655893 | 0.435129 | -0.384631 | 0.812883 | 0.71741 | -0.46168 | | 3 | 0.304474 | -0.702372 | -0.347583 | -0.075664 | -0.483009 | -0.340502 | -0.178482 | |  | Bakery | General.Merch |  |  |  |  |  | | 1 | 0.428226 | -0.674769 |  |  |  |  |  | | 2 | 0.312878 | -0.329045 |  |  |  |  |  | | 3 | -0.866255 | 1.135432 |  |  |  |  |  | |
|  |  |

The cluster differs from one another with their Rand indices and Calinski-harabasz. We can see that format 3 has a high median and low spread according to the AR and CH indices.

According to the AR indices, we can see that formats 4 and 5 have approximately the same AR indices. However, format 4 has a higher CH index than format 5.

Cluster 1 has higher sales for Dry, Meat, Deli and Bakery products overall compared to the other clusters. Cluster 2 has higher sales for Dairy, Frozen Food, Produce, Floral and Bakery products overall compared to the other clusters.

Finally, cluster 3 has higher sales compared to the other clusters for Dry and General Merchandise products.



1. Tableau visualization that shows the location of the stores, uses color to show cluster and size to show total sales.

Chart, scatter chart

Description automatically generated

## **Task 2: Formats for New Stores**

1. I used the Boost Model to predict the best store format for the new stores because the accuracy of the model is higher: 76.47 % compared to the Forest Model (70.59%) and Decision Tree Model (70.59%).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Fit and error measures**   | Model | Accuracy | F1 | Accuracy\_1 | Accuracy\_2 | Accuracy\_3 | | --- | --- | --- | --- | --- | --- | | DecisionTree\_CPT | 0.6471 | 0.6667 | 0.5000 | 1.0000 | 0.5000 | | ForestModel\_CPT | 0.7059 | 0.7500 | 0.5000 | 1.0000 | 0.7500 | | BoostedModel\_CPT | 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. | |

1. What format do each of the 10 new stores fall into?

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

## **Task 3: Predicting Produce Sales**

I chose the ETS(M, N,M) model. The ETS model has a better accuracy. I used a TS Compare Tool to compare the different model and the TS Compare tool is in favor of the ETS model. In fact, it forecast with better accuracy than the ARIMA Model. Using the decomposition plot, we can see that there is a seasonal pattern, and the peaks appear to be decreasing overtime. Therefore, the seasonality component should be added and hence should be multiplicatively(m).

Graphical user interface, chart, line chart

Description automatically generated

Concerning, my choice of ETS over the ARIMA model can be justified by this graph:

Chart, line chart

Description automatically generated

We can see that the ETS is better.

2. Table of your forecasts for existing and new stores.

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Months | Existing Stores | Forecast |
| 2016 | 1 | 21,174,989.40 | 2,590,567 |
| 2016 | 2 | 20,479,354.58 | 2,503,135 |
| 2016 | 3 | 23,580,340.68 | 2,910,154 |
| 2016 | 4 | 22,236,546.23 | 2,772,193 |
| 2016 | 5 | 25,427,255.46 | 3,142,262 |
| 2016 | 6 | 26,143,967.40 | 3,203,694 |
| 2016 | 7 | 26,399,993.27 | 3,233,436 |
| 2016 | 8 | 23,172,393.88 | 2,884,618 |
| 2016 | 9 | 20,544,268.64 | 2,562,089 |
| 2016 | 10 | 20,182,471.09 | 2,506,671 |
| 2016 | 11 | 20,966,876.35 | 2,598,151 |
| 2016 | 12 | 20,965,097.00 | 2,566,314 |

**Visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts**

Chart

Description automatically generated