# Predictive Analytics

*Alteryx and functions*

## The Business Problem

The company where I currently work has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. I have been asked to provide analytical support to make decisions about store formats and inventory planning.

## Details

In this project I am supposed to perform a few tasks in which my last and the final task is to predict the sales but prior to that I have to analyze 85 groceries stores which already exist.

For this project I have given three data sets Store information.csv, storesalesdata.csv and storesdemographicdata.csv for the first step I am supposed to work with **storessaledata.csv** and **Storeinformation.csv.**

The very first step in this I need to classify the 85 stores into segments so that we can supply products accordingly based on the segments to the stores. This will help us to maintain the storage and the product according to the demand and the usage of the product in the particular segment Store.

After classifying the stores I am going to to predict the same segments for the newly

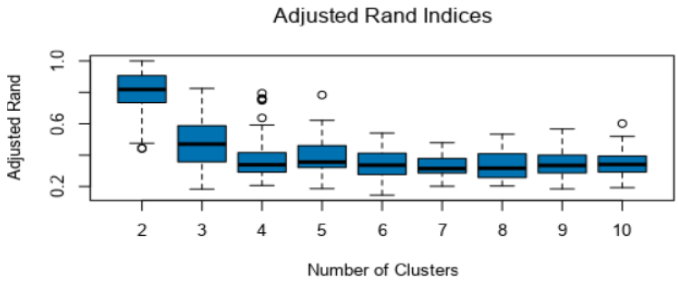
launched 10 stores so that I can classify those New stores into the same segments for that I am going to use **classification model** for predicting of classifying the stores in this I am going to use 3 models name **boosted model,  decision tree and random forest**.

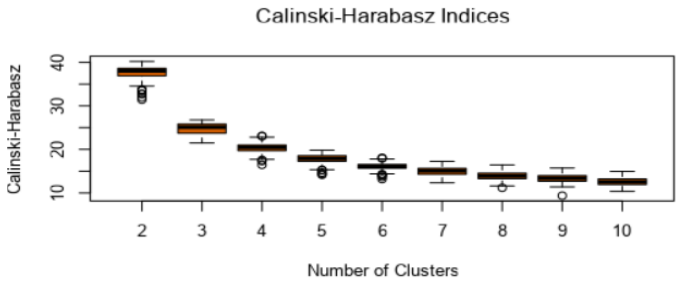
After successfully predicting 10 new stores into the segments I am going to predict the monthly sales for the  2016 year.

## Steps performed

For clustering the stores the very first step I did was importing the data through input data tool

- After importing the data I used the **auto field tool** to change the data types of all the variables in the data set and after performing basic preparation with the data I started using a **predictive grouping tool palette** for clustering purposes.
- My first step was using the k centroid diagnostic tool to know the maximum number of clusters I can get from the data.




- 
- From above plots scenes that are clustered to all the number of clutcher 3 provides the **maximum stability** to the clusters as they are higher than the other increasing number of clusters.

- Then performing cluster analysis through the **k-centroid analysis** tool I distributed the stores into the segments.
- And later I also used the **append cluster tool** to put all the stores into their respective clusters.

Cluster Information:

| Cluster | Size | Ave Distance | Max Distance | Separation |
|---|---|---|---|---|
| 1 | 28 | 2.171857 | 4.988594 | 1.906056 |
| 2 | 35 | 2.453516 | 4.414411 | 1.857263 |
| 3 | 22 | 2.251618 | 3.777476 | 1.819576 |

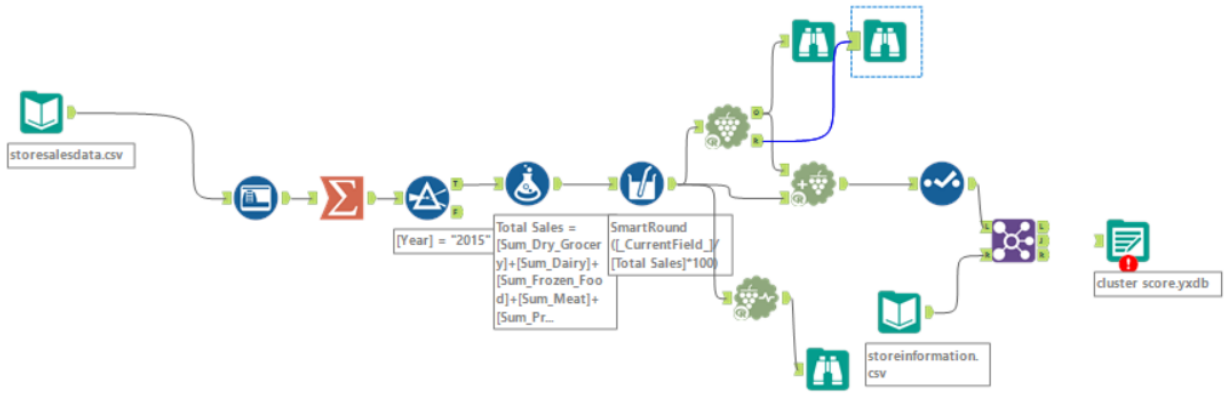
Convergence after 8 iterations.
Sum of within cluster distances: 196.22067.

| | X._Sum_Dry_Grocery | X._Sum_Dairy | X._Sum_Frozen_Food | X._Sum_Meat | X._Sum_Produce | X._Sum_Floral | X._Sum_Deli |
|---|---|---|---|---|---|---|---|
| 1 | 0.654896 | -0.281371 | -0.037878 | 0.500887 | -0.643036 | -0.707666 | 0.66778 |
| 2 | -0.619835 | 0.672634 | 0.301029 | -0.375935 | 0.847051 | 0.742157 | -0.422162 |
| 3 | 0.152597 | -0.711992 | -0.430701 | -0.039415 | -0.529171 | -0.280039 | -0.178279 |

| | X._Sum_Bakery | X._Sum_General_Merchandise |
|---|---|---|
| 1 | 0.221245 | -0.631946 |
| 2 | 0.346546 | -0.325416 |
| 3 | -0.832908 | 1.322002 |

- From the above picture we can observe the differences between the clusters and products needed for the affection towards the clusters.
- We can clearly see from the picture that there are **28 stores in cluster 1** , **35 stores in cluster 2** and **22 stores in cluster 3.**

**Here is the picture of the workflow.**




In the next step I am going to use my output of the first task which is the clusters data

and **storedemographicdata.csv** to use classification models to classify 10 new stores into the Store format for the clusters which we have created in the previous task.

For that first I need to join both the data and perform **association analysis tool** to know which variable will be significant or important to use classification prediction model and which variables I can skip so that I can escape my model from over and under fitting of variables.

*Focused Analysis on Field Cluster1*

| | Association Measure | p-value |
|---|---|---|
| HVal400Kto500K | -0.345028 | 0.0012215 ** |
| PopMulti | 0.330032 | 0.0020382 ** |
| Age0to9 | 0.321871 | 0.0026651 ** |
| Age65Plus | -0.290501 | 0.0069956 ** |
| PopPacIsl | 0.260891 | 0.0158822 * |
| EdSomeCol | 0.248887 | 0.0216235 * |
| Age50to64 | -0.224812 | 0.0385903 * |
| HVal500Kto750K | -0.218836 | 0.0442033 * |
| Total.Sales | 0.216569 | 0.0465029 * |
| Age10to17 | 0.205541 | 0.0591415 . |
| HVal100Kto200K | 0.204673 | 0.0602449 . |
| EdMaster | 0.200723 | 0.0654800 . |
| EdDoctorate | 0.195853 | 0.0724346 . |
| PopAsian | 0.185563 | 0.0890783 . |
| Age30to39 | 0.171436 | 0.1166949 |
| Age40to49 | 0.168857 | 0.1223834 |
| HHInc50Kto75K | -0.166982 | 0.1266495 |
| PopOther | 0.164833 | 0.1316792 |
| HHInc150Kto250K | 0.156704 | 0.1520835 |
| EdLTHS | -0.153873 | 0.1597162 |
| HVal300Kto400K | -0.147223 | 0.1787623 |
| EdHSGrad | -0.143124 | 0.1913024 |
| HHSz4Per | 0.138283 | 0.2069113 |

From a lot of variables now I can pick my significant and important variables which will definitely help the model to predict better and perform better with the final data.

Now I know the variables which I am going to use now I can create sample from my data I am going to use **80% and 20%** rule using **create sample tool** rule for the data samples to perform classification models on it and then I'll compare the models using **model comparison tool** to know which model predicts the better results with lower **ACCURACY** and then I'll use that model with **score tool** to give pass the 10 new stores to the respective segments.

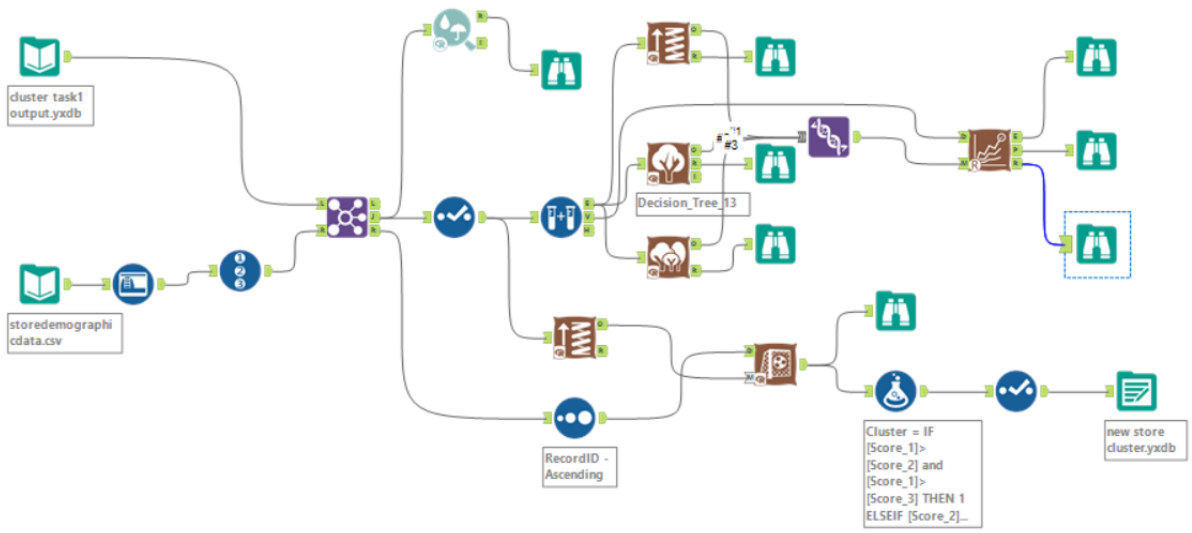
**Fit and error measures**

| Model | Accuracy | F1 | Accuracy_1 | Accuracy_2 | Accuracy_3 |
|---|---|---|---|---|---|
| boosted | 0.7647 | 0.7833 | 0.6000 | 0.7500 | 1.0000 |
| Decision_Tree_13 | 0.4118 | 0.4250 | 0.4000 | 0.3750 | 0.5000 |
| forest | 0.5882 | 0.5917 | 0.4000 | 0.6250 | 0.7500 |

This is the report of the model comparison tool in which we can see that the boosted model has provided the accuracy with 76% which is very high compared to other two models so now we can make a decision that we can use the Boosted  model on the whole data set.

| RecordID | Store | Cluster |
|---|---|---|
| 86 | S0086 | 3 |
| 87 | S0087 | 2 |
| 88 | S0088 | 3 |
| 89 | S0089 | 2 |
| 90 | S0090 | 2 |
| 91 | S0091 | 3 |
| 92 | S0092 | 2 |
| 93 | S0093 | 3 |
| 94 | S0094 | 2 |
| 95 | S0095 | 2 |

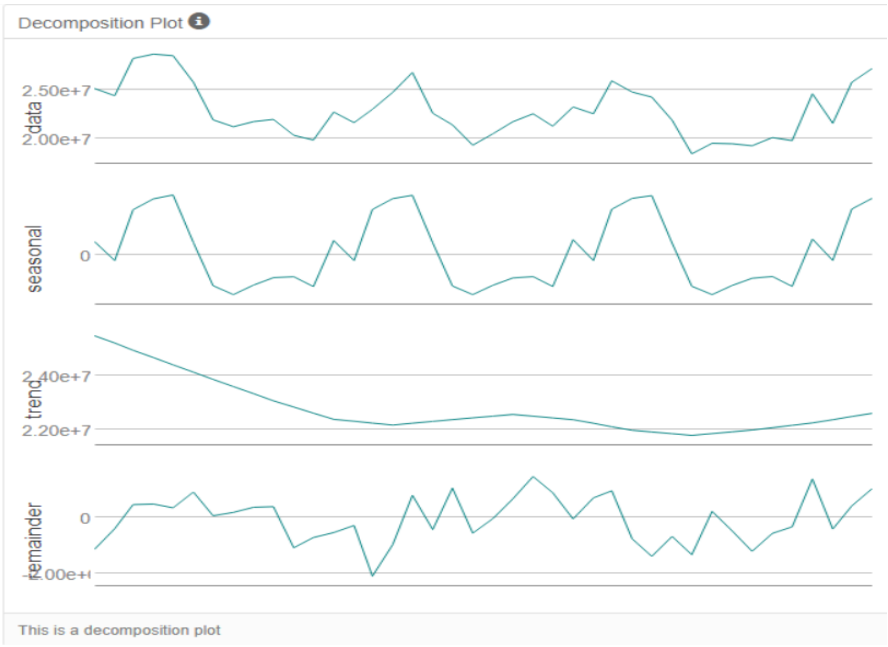
From the score tool we have concluded that 6 stores false under cluster 2 and 4 stores fall under cluster 3.




Here is the picture of the workflow.

In the last step of forecasting the sales for the stores were going to use storesalesdata.csv

To determine the parameters to apply on ETS model building I began visualizing a time series decomposition plot. I used the Summarize tool grouping Store Sales data by Year and by Month and summing Produce. After I attached the TS Plot tool and customized it for monthly target field frequency and plot type as time series decomposition plot. Based on plot it was possible to determine trend, seasonal and error components.



In the composition plot of the data we can observe that there is no trend in the data but we can see some **seasonality** in the data there is a spike in the middle and then down and then spike this is a continuous process and we can say by observing that there is seasonality in the data but we can't see any trend in the data  for that we need to opt function in the configuration window of time series forecasting tools for that I would be used multiplicative technique to make the data stationery.

After using forecasting models which is ETS and ARIMA for time series and comparing the results or the report of the data we will pick the model with the lowest RMSE

Accuracy Measures:

| Model | ME | RMSE | MAE | MPE | MAPE | MASE |
|---|---|---|---|---|---|---|
| ETS | -21581.13 | 663707.2 | 553511.5 | -0.0437 | 2.5135 | 0.3257 |
| ARIMA | -604232.29 | 1050239.2 | 928412 | -2.6156 | 4.0942 | 0.5463 |

In the report we can see **ETS** has given the **lowest RMSE**  which is **663707.2** compared to **ARIMA**. so, we can use **ETS** for our further prediction.

Below is the picture of work flow.