Project: Predictive Analytics Capstone

by
Hari Sankaran Nair

Task 1: Determine Store Formats for Existing Stores

Summary of the Business Scenario

Your company currently 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 the 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.

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

   A K-centroids analysis was done using k-means clustering method (using k values as 3, 4, 5, 6) to determine the best number of each type for store format. The results are as shown below:

k = 3

K-Means Cluster Assessment Report

Summary Statistics

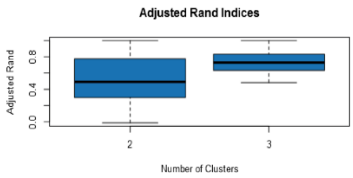
Adjusted Rand Indices:

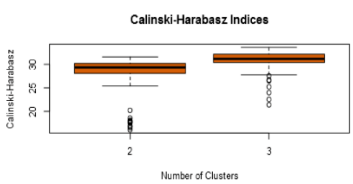
	2	3
Minimum	-0.01295	0.4832
1st Quartile	0.304	0.6334
Median	0.4924	0.7295
Mean	0.4822	0.7435
3rd Quartile	0.7759	0.8312
Maximum	1	1

Calinski-Harabasz Indices:

	2	3
Minimum	16.1	21.41
1st Quartile	28.06	30.36
Median	29.34	31.15
Mean	27.96	30.78
3rd Quartile	30.13	32.12
Maximum	31.58	33.57

Plots





k = 4

K-Means Cluster Assessment Report

Summary Statistics

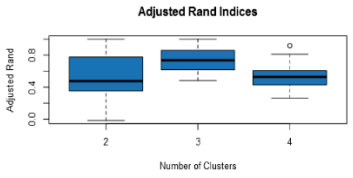
Adjusted Rand Indices:

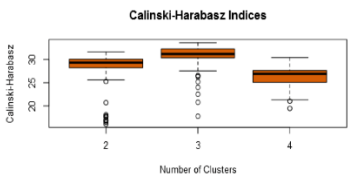
	2	3	4
Minimum	-0.0152	0.4826	0.2633
1st Quartile	0.3595	0.6224	0.4297
Median	0.4759	0.735	0.5289
Mean	0.5015	0.7367	0.5335
3rd Quartile	0.7555	0.8585	0.6055
Maximum	1	1	0.9184

Calinski-Harabasz Indices:

	2	3	4
Minimum	16.1	17.79	19.48
1st Quartile	28.24	30.32	25.09
Median	29.31	31.15	26.89
Mean	28.09	30.59	26.4
3rd Quartile	30.03	32.25	27.61
Maximum	31.58	33.57	30.37

Plots





k = 5

K-Means Cluster Assessment Report

Summary Statistics

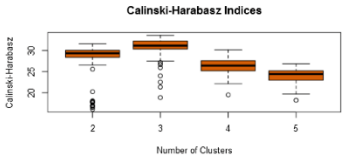
Adjusted Rand Indices:

	2	3	4	5
Minimum	-0.0152	0.3363	0.2633	0.2967
1st Quartile	0.3384	0.6596	0.4186	0.3877
Median	0.5092	0.7422	0.5244	0.4692
Mean	0.5014	0.7361	0.528	0.4896
3rd Quartile	0.7759	0.8185	0.6115	0.5851
Maximum	1	0.9586	0.7758	0.8518

Calinski-Harabasz Indices:

	2	3	4	5
Minimum	16.1	18.85	19.48	18.23
1st Quartile	28.4	30.32	25.22	22.99
Median	29.34	31.14	26.43	24.41
Mean	28.07	30.65	26.43	23.97
3rd Quartile	30.03	32.22	27.57	25.18
Maximum	31.58	33.57	30.14	26.85

Plots



k = 6

K-Means Cluster Assessment Report

Summary Statistics

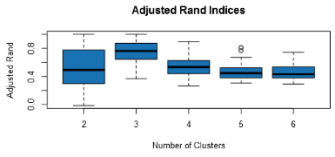
Adjusted Rand Indices:

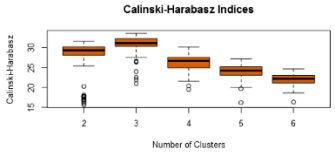
	2	3	4	5	6
Minimum	-0.0152	0.3682	0.2633	0.3041	0.2923
1st Quartile	0.304	0.6459	0.4432	0.3821	0.3802
Median	0.4925	0.7616	0.533	0.4475	0.4314
Mean	0.4816	0.7503	0.5423	0.4669	0.4496
3rd Quartile	0.7555	0.8694	0.6245	0.5231	0.5332
Maximum	1	1	0.8921	0.811	0.7427

Calinski-Harabasz Indices:

	2	3	4	5	6
Minimum	15.65	20.96	19.48	16.2	16.36
1st Quartile	28.06	30.35	24.93	23.07	21.12
Median	29.32	31.12	26.67	24.23	22.18
Mean	27.91	30.7	26.37	23.93	22.05
3rd Quartile	30.13	32.25	27.56	25.18	23.03
Maximum	31.58	33.57	30.14	27.19	24.62

Plots





From the k-means cluster analysis, cluster k = 3 seems to be the best one. Hence, I am using cluster k = 3 as the baseline to compare the other numbers of k terms. Comparing cluster k = 3 to all the other k terms, using k = 3 for the analysis offers the best results with cluster 3 having the tightest range and highest mean when k = 3.

Therefore, my recommendation is that the optimal number of store formats is 3.

2. How many stores fall into each store format?

Store Format	Number of stores
1	23
2	29
3	33

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

Please find below, the summary of cluster analysis for k = 3:

Summary Report of the K-Means Clustering Solution Custer_ Analysis

Solution Summary

Call:

stepFlexclust(scale(model.matrix(~-1 + X._Dry_Grocery + X._Dairy + X._Frozen_Food + X._Meat + X._Produce + X._Floral + X._Deli + X._Bakery + X._General_Merchandise, the.data)), k = 3, nrep = 10, FUN = kcca, family = kccaFamily("kmeans"))

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

Convergence after 12 iterations.

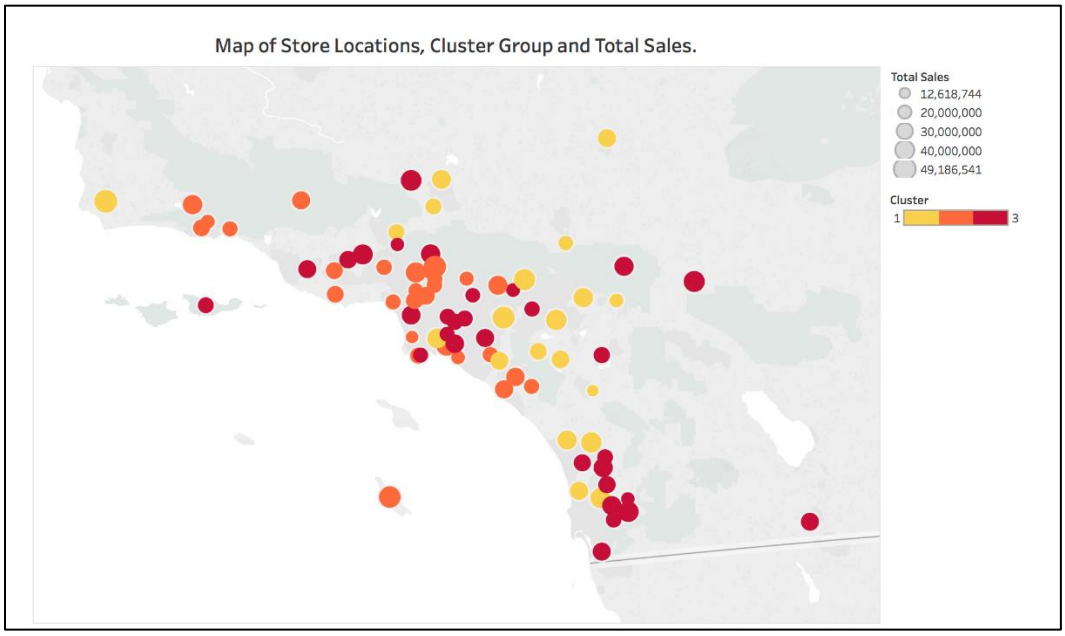
Sum of within cluster distances: 196.83135.

	X._Dry_Grocery	X._Dairy	X._Frozen_Food	X._Meat	X._Produce	X._Floral	X._Deli
1	0.327833	-0.761016	-0.389209	-0.086176	-0.509185	-0.301524	-0.23259
2	-0.730732	0.702609	0.345898	-0.485804	1.014507	0.851718	-0.554641
3	0.413669	-0.087039	-0.032704	0.48698	-0.53665	-0.538327	0.64952

	X._Bakery	X._General_Merchandise
1	-0.894261	1.208516
2	0.396923	-0.304862
3	0.274462	-0.574389

The analysis above shows that Cluster 1 is most positive for percentage of general merchandise sales versus Cluster 3 which is the most negative while Cluster 2 is slightly less negative. This indicates that these clusters differ in terms of the variable for percentage of general merchandise sales.

4. **Please provide a Tableau visualization (saved as a Tableau Public file) that shows the location of the stores, uses color to show cluster, and size to show total sales.**



Task 2: Formats for New Stores

Summary of the Business Scenario

The grocery store chain has 10 new stores opening at the beginning of the year. The company wants to determine which store format each of the new stores should have. However, sales data is not available for these new stores yet, so the format will have to be determined using each of the new store's demographic data.

1. **What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)**

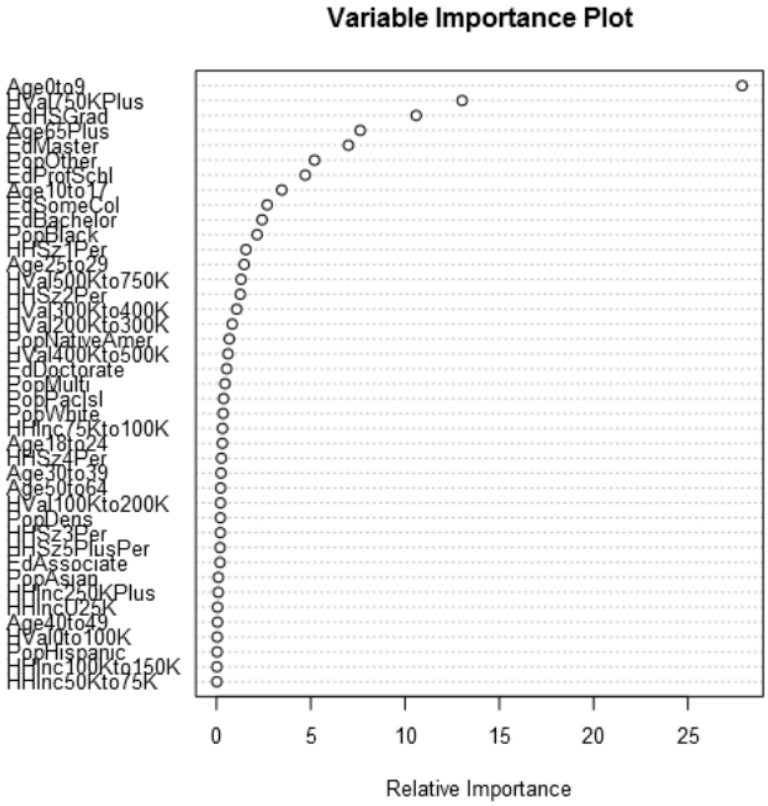
   In order to predict the new stores' formats, the demographic data available in the StoreDemographicData.csv was used. All the variables were used as predictor variables and they were run separately through a decision tree model, a random forest model and a boosted model. 80% of the data was used for training the models and 20% for validating the models.

   A comparison of the 3 models is provided below:

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
Boosted	0.8235	0.8543	0.8000	0.6667	1.0000
DT	0.7059	0.7327	0.6000	0.6667	0.8333
Forest	0.8235	0.8251	0.7500	0.8000	0.8750

   Based on the above comparison, the Boosted and Forest models have the same accuracy score, but the Boosted model has a higher F1 score. Based on this, it was decided to use the Boosted Model for the final prediction.



The Variable Importance Plot from the Boosted Model indicates that the 3 most important variables for the Boosted Model are Age0to9, HVal750KPlus and EdHSGrad.

2. What format do each of the 10 new stores fall into? Please fill in the table below.

The Boosted Model made the following predictions for the 10 new stores:

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

Task 3: Predicting Produce Sales

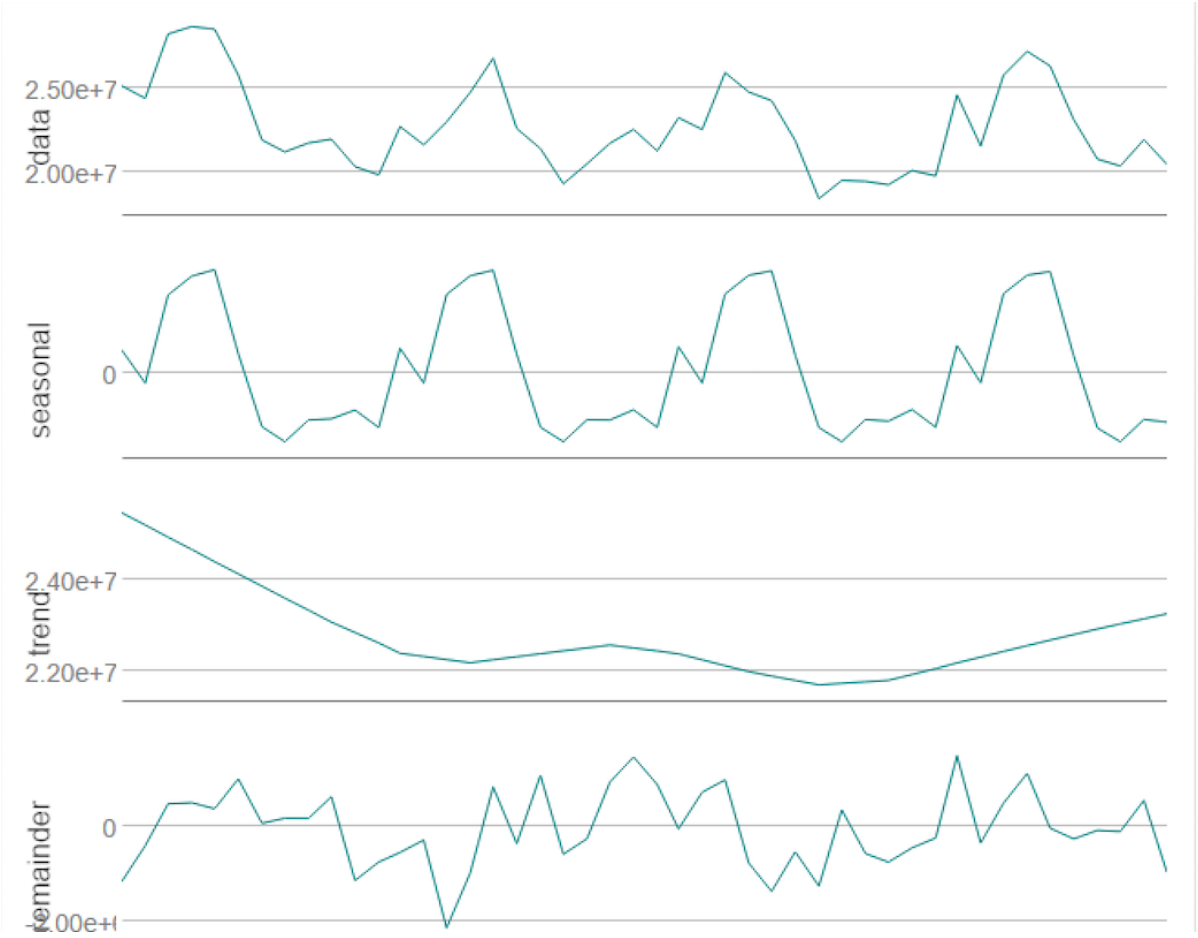
Summary of the Business Scenario

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast. A monthly forecast is to be prepared for produce sales for the full year of 2016 for both existing and new stores.

1. **What type of ETS or ARIMA model did you use for each forecast? Use ETS(a,m,n) or ARIMA(ar, i, ma) notation. How did you come to that decision?**

   Both the ETS and ARIMA models were run for comparison. Analysis of the initial time series decomposition plots as shown below allowed for further analysis of the model parameters to be established.

   The data used in this case was the sales for produce only per month for all the stores aggregated together.



From the decomposition plots shown above, the Error element is increasing, the Trend element is non-existent, and the Seasonal element is also increasing, therefore an ETS(M,N,M) will be used. For the ARIMA model, the model has been set to calculate the elements automatically.

For comparison purposes, a holdout time-period of 12 periods was used to validate the ETS and ARIMA model.

Below is the ETS Accuracy Measures:

ME	RMSE	MAE	MPE	MAPE	MASE
210494.4	760267.3	649540.8	1.0288	2.9678	0.3822

And here are the ARIMA model Accuracy Measures:

ME	RMSE	MAE	MPE	MAPE	MASE
-604232.3	1050239	928412	-2.6156	4.0942	0.5463

ETS model has higher accuracy when compared to the ARIMA model. This is indicated by the lower values of RMSE accuracy and MASE accuracy, which indicate a better fit.

The ETS model's RMSE accuracy is 760267.3 versus the ARIMA model's RMSE accuracy of 1050239.
The ETS model's MASE accuracy is 0.3822 versus the ARIMA model's MASE accuracy of 0.5463.

The ETS model will be used for forecasting since because of its higher accuracy and hence, better fit.

2. **Please provide a table of your forecasts for existing and new stores. Also, provide visualization of your forecasts that includes historical data, existing stores forecasts, and new stores forecasts.**

Month	New Store Sales Forecast	Existing Store Sales Forecast
Jan-16	2,600,354.85	21,381,830.22
Feb-16	2,505,198.46	21,081,311.62
Mar-16	2,889,940.32	24,502,171.96
Apr-16	2,743,927.30	22,352,993.13
May-16	3,110,813.81	25,331,350.65
Jun-16	3,191,154.55	26,330,255.79
Jul-16	3,219,369.78	25,715,514.09
Aug-16	2,852,751.79	23,458,933.07
Sep-16	2,543,602.66	21,801,458.48
Oct-16	2,477,331.44	21,509,922.65
Nov-16	2,569,169.56	22,619,212.99
Dec-16	2,535,481.94	21,582,321.09

A Tableau plot is provided below for the sales forecasts for existing, new and the combined stores' produce sales: