Project: Predictive Analytics

By: Anthony Nguyen

Project Overview

Task 1: Determine Store Format for Existing Stores

A grocery store currently has 85 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. The project deliverables are to provide analytical support to make decisions about store formats and inventory planning.

To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms formats and segments will be used interchangeably throughout this project.

Task 2: Store Format for New Stores

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, we do not have sales data for these new stores yet, so we'll have to determine the format using each of the new store's demographic data.

Task 3: Forecasting

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

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.
   - Information based off the K-Means Cluster Assessment Report from the K-Centroids Diagnostics tool shows that at 3 store formats, the AR (0.708) & CH (31.02) Indices median level is the highest compared to all other formats. The AR & CH plots reflect that the 3 store format may not have the tightest box & whisker plot, but the high value median suggest that the 3 store format is optimal.

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

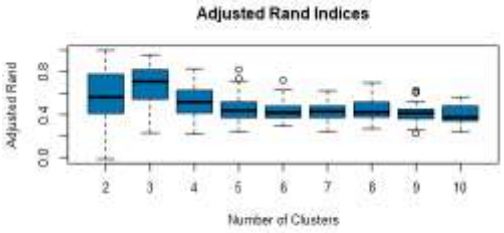
	2	3	4	5	6	7	8
Minimum	-0.0152	0.2276	0.2198	0.2392	0.2903	0.2399	0.2674
1st Quartile	0.4196	0.5498	0.4171	0.3733	0.3714	0.3754	0.3784
Median	0.562	0.7083	0.5162	0.4366	0.4184	0.4288	0.4228
Mean	0.533	0.678	0.5246	0.4563	0.4341	0.4254	0.4398
3rd Quartile	0.7656	0.8173	0.6249	0.5156	0.4768	0.4774	0.5136
Maximum	1	0.9583	0.8277	0.8215	0.7202	0.6221	0.6977

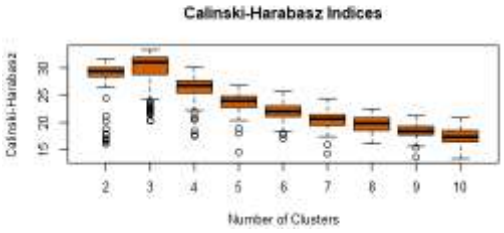
	9	10
Minimum	0.2232	0.2398
1st Quartile	0.3626	0.3412
Median	0.4117	0.3743
Mean	0.41	0.3984
3rd Quartile	0.4501	0.4736
Maximum	0.6294	0.5636

Calinski-Harabasz Indices:

	2	3	4	5	6	7	8
Minimum	16.1	20.27	17.55	14.58	17.03	14.3	16.11
1st Quartile	28.42	28.82	25.46	22.79	21.1	19.5	18.57
Median	29.39	31.02	26.77	23.95	22.02	20.66	19.85
Mean	28.21	29.65	26.21	23.67	21.98	20.48	19.72
3rd Quartile	30.07	31.96	27.68	24.8	23.01	21.45	20.89
Maximum	31.58	33.31	30.09	26.78	25.65	24.37	22.5

	9	10
Minimum	13.64	13.4
1st Quartile	17.76	16.43
Median	18.39	17.36
Mean	18.47	17.46
3rd Quartile	19.31	18.55
Maximum	21.24	20.87



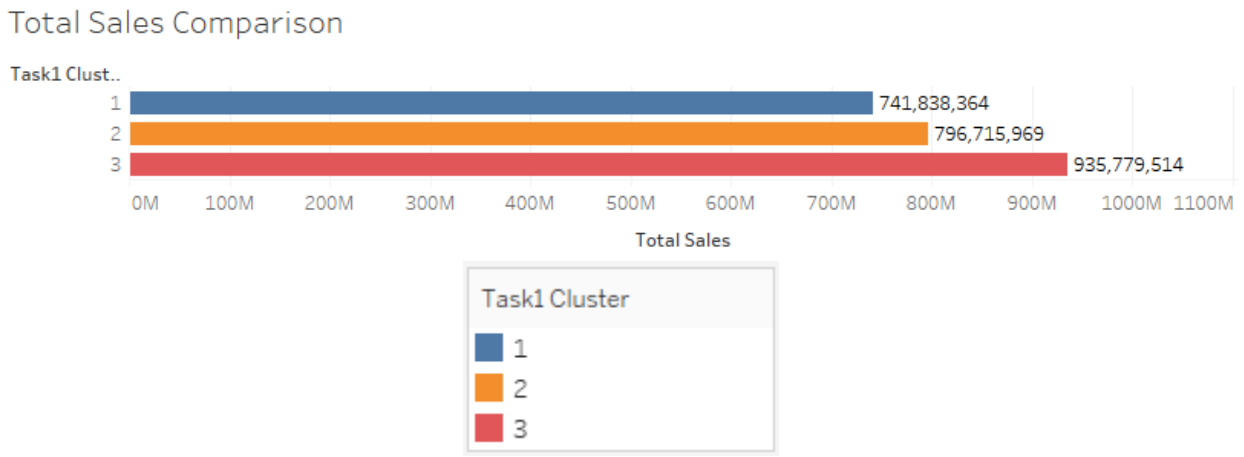


2. How many stores fall into each store format?
   - The number of stores that fall into each store format are as follows:
     - Cluster 1 = 23
     - Cluster 2 = 29

- Cluster 3 = 33

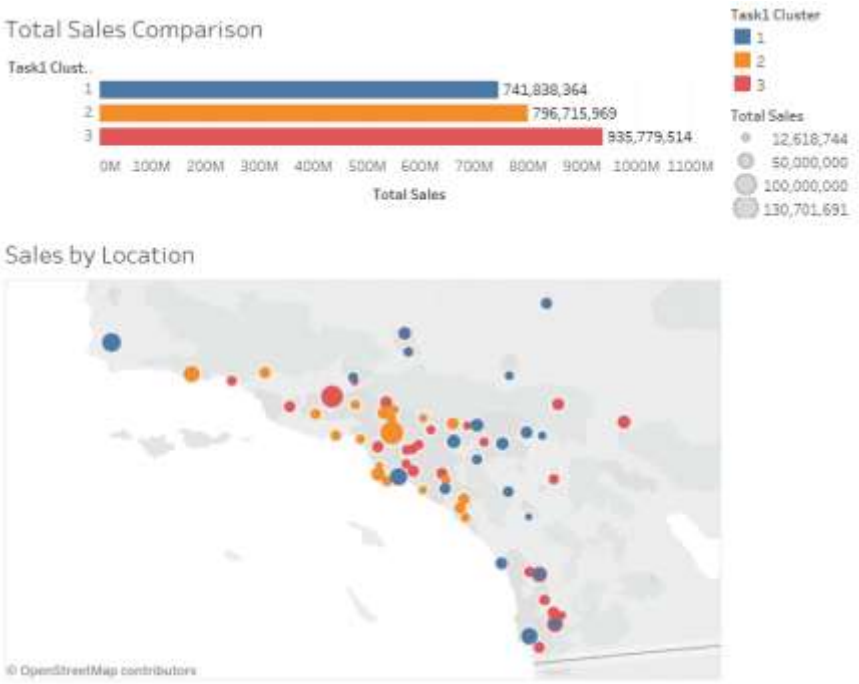
Cluster	Size	Ave Distance	Max Distance	Separation
1	23	2.320539	3.55145	1.874243
2	29	2.540086	4.475132	2.118708
3	33	2.115045	4.9262	1.702843

3. Based on the results of the clustering model, what is one way that the clusters differ from one another?
   - Based on the results of the clustering model, the one way that the clusters differ from each other is the amount of total sales. Cluster 3 has the highest sum of total sales compared to Cluster 1 & Cluster 2.



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.
   - Below is a visualization that shows the location of the stores with distinctions in color, size of total sales, and location.

https://public.tableau.com/views/Task1Dashboard_0/Dashboard1?:embed=y&:display_count=yes&publish=yes



Task 2: Formats for New Stores

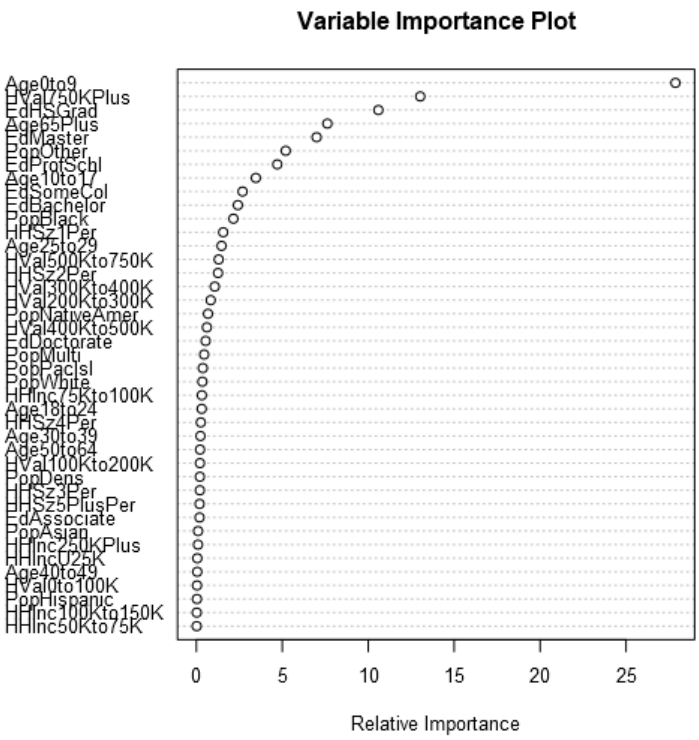
1. What methodology did you use to predict the best store format for the new stores? Why did you choose that methodology? (Remember to Use a 20% validation sample with Random Seed = 3 to test differences in models.)
   - With having the Boosted Model, Decision Tree, and Forest Model compared to each other through the Model Comparison Tool, the Boosted Model methodology was chosen to predict the best store format.
   - Both the Boosted Model & Forest Model had the same accuracy. The Boosted Model F1 score was higher than the Forest Model. This was the justification to select the Boosted Model as best to predict the store format.

Model Comparison Report

Fit and error measures

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

2. What are the three most important variables that help explain the relationship between demographic indicators and store formats? Please include a visualization.
   - The three most important variables that help explain the relationship between demographic indicators and store formats are:
     - Age0to9
     - HVal750K
     - EdHSGrad



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

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: Forecasting

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?

- The following ETS model was used: ETS (M,N,M) with no dampening.
- After evaluating the TS Plot report, the seasonality trend is increasing implying that multiplicatively should be applied (M).
- The trend line is not clear and shows no true sign of pattern, (N).
- The error shows variation and is irregular, multiplicatively should be applied (M).
- A holdout sample of 6 months was used.

Summary of Time Series Exponential Smoothing Model ETS

Method:
ETS(M,N,M)

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-62800.6583899	948208.5360249	762227.3150084	-0.4074302	3.3572352	0.4254973	0.1075188

Information criteria:

AIC	AICc	BIC
1275.8636	1292.6636	1299.5079

- The ARIMA (1,0,0)(1,1,0)[12] model was used.
- A holdout sample of 6 months was used.

Summary of ARIMA Model ARIMA

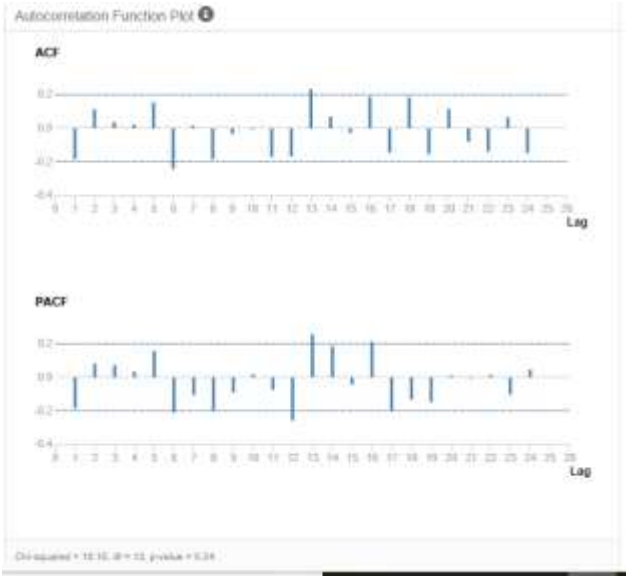
Method: ARIMA(1,0,0)(1,1,0)[12]

Information Criteria:

AIC	AICc	BIC
880.4445	881.4445	884.4411

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-102530.8278988	1042209.8520798	738087.5531171	-0.5465069	3.3006311	0.4120218	-0.1854462



- Comparing the ETS model to the ARIMA model, the ETS model accuracy is more favorable than the ARIMA model.
- ETS RMSE value (948208.54) is lower than the ARIMA RMSE value (1042209.85).
- The ETS (M,N,M) model was selected for forecasting.

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.

   - Below is a table of forecast sales for existing stores and new stores.

Month	Year	Existing Store Sales Forecast	New Store Sales Forecast
1	2016	21,298,345	2,600,355
2	2016	20,571,970	2,505,199
3	2016	23,883,412	2,889,940
4	2016	22,472,417	2,743,927
5	2016	25,650,806	3,110,814
6	2016	26,326,884	3,191,155
7	2016	26,541,048	3,219,370
8	2016	23,362,180	2,852,752
9	2016	20,677,598	2,543,602
10	2016	20,177,927	2,477,331
11	2016	20,984,723	2,569,170
12	2016	20,943,790	2,535,482

https://public.tableau.com/views/ProduceSalesByYear-HistoricalandForecastData/Dashboard1?:embed=y&:display_count=yes&publish=yes