1/8/2019

Recommending New Store Format for a Grocery Store

Predictive Analytics for Business Capstone Project

Udacity Nanodegree

Tools: Alteryx, Tableau

Katerina Bosko, PhD
WWW.CROSS-VALIDATED.COM

The Business Problem

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 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. You've been asked to provide analytical support to make decisions about store formats and inventory planning.

Task 1: Determining Store Format

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. You've been asked to determine the optimal number of store formats based on sales data.

Task 2: Store Format for New Stores

The grocery store chain has 10 new stores opening up at the beginning of the year. The company wants to determine which store format each of the new stores should have. However, we don't have sales data for these new stores yet, so you'll have to determine the format using each of the new store's demographic data.

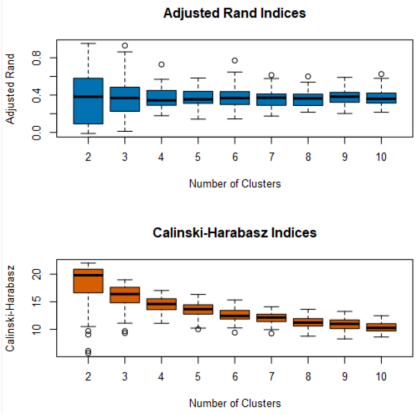
Task 3: Forecasting Monthly Sales

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast. You've been asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores.

Step 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. This was determined by comparing the ARI and CHI indices when using the K-Centroid Diagnostics tool in Alteryx.
ARI measures the stability of the clusters, while CHI – their distinctness. Judging from the whisker and box plot, the higher the median and smaller the variation, the better the results. From the picture below, we can conclude that 3 is the optimal number of clusters[1] because even though it has slightly less median on ARI than the 2 cluster solution (0.368697 vs 0.382353), it's whisker and boxer plot is more compact.



K-Means Clustering Model Validation Indices

2. **How many stores fall into each store format?**

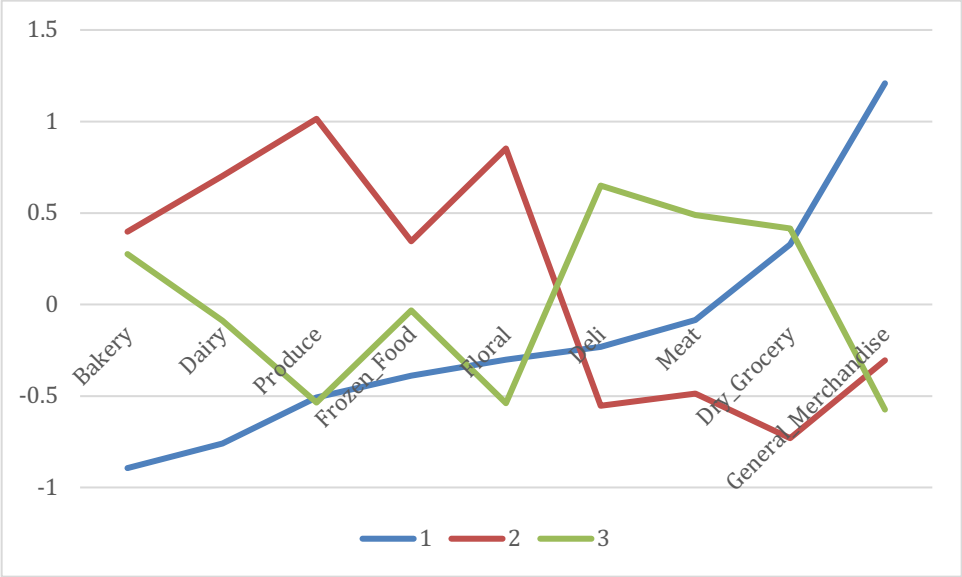
Segment 1 – 23
Segment 2 – 29
Segment 3 – 33

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

Having visualized the cluster centroids for each product category in Tableau (see picture below), it is easy to see the differences in product mixes upon which the three store segments were built. When the percentage sales per category of each segment are compared side by side[2], we can conclude that

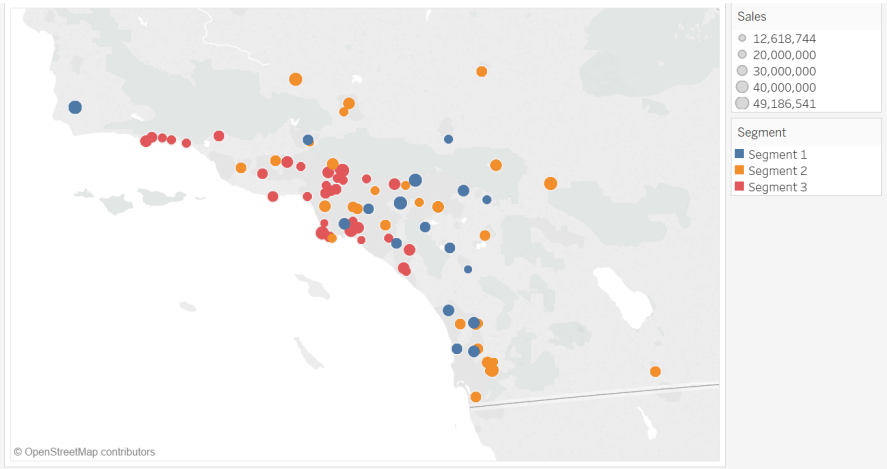
Segment 1 – stores in this segment sell on average more General Merchandise
Segment 2 – stores in this segment sell on average more Produce and Floral
Segment 3 – stores in this segment sell on average more Deli and Meat

[1] In the following, I will refer to clusters as segments.
[2] See my dashboard in Tableau Public.



Cluster Centroids for each Product Category in each Cluster

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.**



Map of existing stores divided in three segments

The visualization above is also accessible at https://public.tableau.com/profile/katerina.bosko#!/vizhome/AllocatingGroceryStorestoSegmentsandForecastingProduceSales/Dashboard3

Step 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.)**

To predict the best store format for new stores, I used three classification models (Decision Tree, Forest Model, Boosted Model) trained on the data we have for 85 existing stores. Based on the validation sample, both Forest Model and Boosted Model produce equally good results in terms of Accuracy - 0.8235 (see picture below), but Boosted Model has higher F1 score of 0.8889. Due to this, I used the Boosted Model in my further analysis.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	Accuracy_1	Accuracy_2	Accuracy_3
p8_2-Decision_Tree	0.7059	0.7685	0.7500	1.0000	0.5556
p8_2_ForestModel	0.8235	0.8426	0.7500	1.0000	0.7778
p8_2_BoostedModel	0.8235	0.8889	1.0000	1.0000	0.6667

Comparison of Different Classification Models based on the Holdout Sample

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

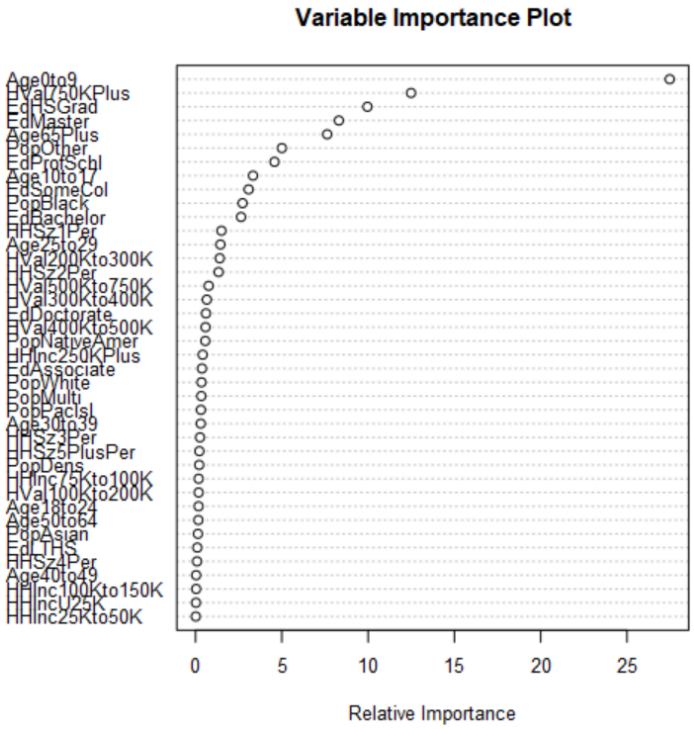
Having applied the chosen classification model on new data, I got the following composition of new stores:

Segment 1 – 3 stores
Segment 2 – 6 stores
Segment 3 – 1 stores

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

3. **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 demographic variables to differentiate between store formats seem to be presence of kids (Age0to9), the education variable EdHSGrad (probably graduation from high school) and the mysterious variable HVal750KPlus (see the picture below).



Variable Importance Plot for Boosted Model

Step 3: Predicting Produce Sales

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?

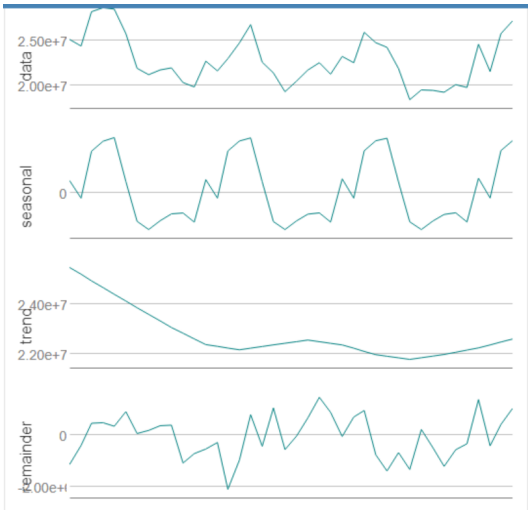
Type	Model
Existing stores	ETS (Auto)[3]
New Stores Segment 1	ETS (Auto)
New Stores Segment 2	ARIMA (1,0,0)(0,1,1)
New Stores Segment 3	ARIMA (1,0,0)(0,1,1)

[3] With specifications m,n,m.

The model for each time series was chosen based on the analysis of ACF and PACF plots and comparison of Accuracy Measures for different models using the TS Compare tool on the validation sample.
In each case, four time series models were checked – two ETS models and two ARIMA models (one with customized parameters based on own analysis and one with auto parameters suggested by Alteryx).
For instance, having analyzed the decomposition plot for existing stores (see picture below), I decided to check the following parameters:

Components:	**Parameter**	**Explanation**
Error term	Multiplicatively	the changes doesn't stay constant and vary across periods
Trend	No trend	the trend line changes direction twice, so it's neither linear nor exponential
Seasonal	Additive	the seasonal component seems to stay constant throughout the periods

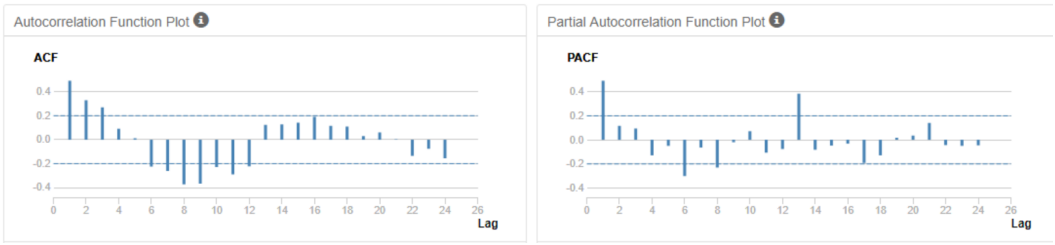
Hence, the model to test was specified as ETS (m,n,a).



Decomposition Plot for Time Series of Existing Stores

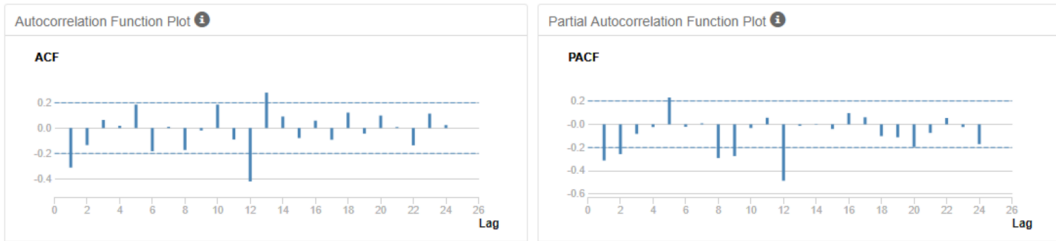
The auto mode in Alteryx, however, suggested to test ETS (m,n,m) model, which turned out to have better accuracy measures.

To customize ARIMA model, I analyzed the ACF and PACF plots.
For instance, for existing stores we had the plots as shown below.



The ACF and PACF plots of Seasonal Component for Existing Stores

From the plots above, we can conclude that the data needs to be differenced to become stationary. Since the first lag on the ACF and the PACF plots is positive, the ACF plot cuts decreases gradually and the PACF cuts off abruptly, we could specify the parameters of the Seasonal Component as ARIMA (1,0,0).



The ACF and PACF plots of the Seasonal First Difference for Existing Stores

The second batch of the ACF and PACF plots shows data after the first differencing (hence the integrated component = 1). From these plots we see that the first lag on both ACF and PACF plots is negative. At the same time, the ACF cuts off sharply after a few lags, while the PACF decreases more gradually. All this indicates that we should use MA terms.
Hence, I specified the final model as ARIMA (1,0,0)(0,1,1).
The auto mode in Alteryx, again suggested a different model – ARIMA Auto with specifications (1,0,0)(1,1,0).

In the next step, I compared the accuracy measures of four models. For existing stores, the accuracy measures of the four models - ETS(m,n,a), ETS (Auto), ARIMA (1,0,0)(0,1,1) and ARIMA (Auto) – for the holdout sample were as follows:

Accuracy Measures:

Model	ME	RMSE	MAE	MPE	MAPE	MASE
ARIMA_1_0_0_0_1_1	-335734.3	791161.4	686064.1	-1.4218	3.0188	0.4037
ARIMA_Auto	-604232.3	1050239.2	928412	-2.6156	4.0942	0.5463
ETS_Auto	210494.4	760267.3	649540.8	1.0288	2.9678	0.3822
ETS_MNA	-947831.4	1103108.1	970225.1	-4.3472	4.4493	0.5709

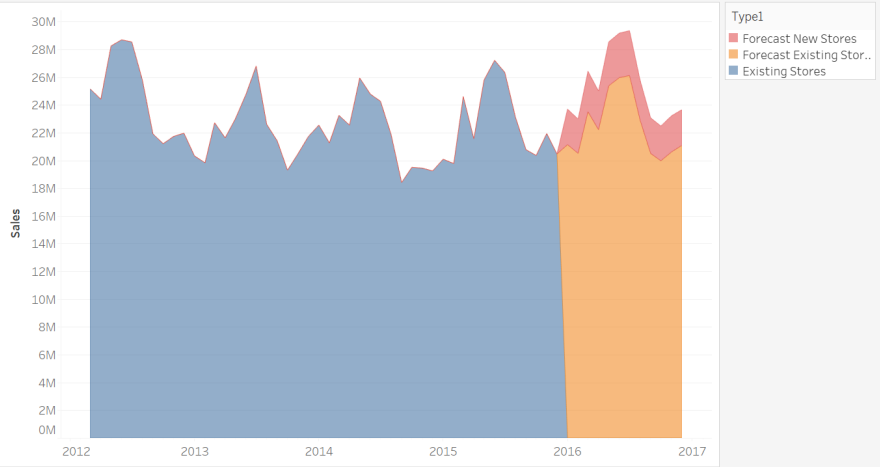
From the table above, we can conclude that ETS (Auto) model performs the best as it has the smallest error terms in comparison to other three models (e.g. MASE = 0.38 and RMSE = 760267). Thus, ETS (Auto) model was then used to forecast the produce sales for existing stores in 2016.
Above I described the analytical process using the example on existing stores. The same

analytical steps were then repeated for time series forecasts of the new stores.

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.

Based on time series analysis, the produce sales in all 95 stores in 2016 were predicted to make **$303.37 ± $42.47 millions**, with the confidence interval of 95%.

Month	New Stores	Existing Stores
2016-1	$2,560,502.42	$21,136,208.14
2016-2	$2,476,071.18	$20,506,604.69
2016-3	$2,921,765.24	$23,506,131.46
2016-4	$2,789,558.69	$22,207,971.24
2016-5	$3,168,359.45	$25,376,698.32
2016-6	$3,210,563.51	$25,963,559.45
2016-7	$3,233,213.72	$26,113,357.20
2016-8	$2,887,369.53	$22,904,671.92
2016-9	$2,569,187.10	$20,499,151.00
2016-10	$2,508,080.86	$19,970,808.95
2016-11	$2,607,788.06	$20,602,232.30
2016-12	$2,575,770.43	$21,072,786.92
	$33,508,230.20	**$269,860,181.58**



Produce Sales for Existing Stores in 2012-2015 and Forecast for 2016, including New Stores

The visualization above is also accessible at
https://public.tableau.com/profile/katerina.bosko#!/vizhome/AllocatingGroceryStorestoSegmentsandForecastingProduceSales/Dashboard3