**Demand Transference – Modelling Approach**

17nth July 2020

****

Contents

[Overall Scope Definition: 3](#_Toc45896744)

[Solution Framework: 3](#_Toc45896745)

[Initial Approach: 3](#_Toc45896746)

[Improved Approach: 4](#_Toc45896747)

[Data used: 4](#_Toc45896748)

[Level of model fit: 5](#_Toc45896749)

[Data preparation: 5](#_Toc45896750)

[Feature Engineering: 6](#_Toc45896751)

[Modelling: 7](#_Toc45896752)

[Results: 8](#_Toc45896753)

[Codes: 9](#_Toc45896754)

# Overall Scope Definition:

* To design a modelling framework to capture the demand transference for a set of products given each product has an adequate sales history
* Create a POC by using Australia Nielsen sales data present for three years across two retailers

# Solution Framework:

## Initial Approach:

The initial approach was a two-step model, where firstly we use the group of discontinued SKUs and few similar products to model the relationship between ids & volume sale. Then use the sales estimates to calculate lost sales (for discontinued SKUs) and lift (for similar SKUs). Finally, we can calculate the retention factor.

The second step was then to model these retention factors against the SKU attributes to predict whether a SKU has to be retained or not. For more details please refer the following document



We did few iterations with the explained approach; our inferences are given below:

Advantages:

* One of main advantage is that we don’t need any competitor sales data for this approach
* As SKU interactions are considered, model is fit on a very limited number of SKUs and data till discontinuation is used in case of discontinued SKUs. This results in a good model fit thus reinforcing our estimates

Disadvantages:

* There is an underlying assumption that the demand is transferred only within few fixed similar mars SKUs which are selected based on similarity score
* Also, there is an assumption that sales estimate will be higher than the actual sales, which holds true for discontinued SKUs after discontinuation but cannot be generalized to normal SKUs

## Improved Approach:

To improve upon the shortcomings of the previous approach we added a few more features and made slight modifications to the proposed approach. Firstly, we wanted an approach which can be used to calculate the retention for all SKUs at a given point in time. The new approach also utilizes one model to arrive at the numbers unlike the two-step approach proposed before. To reduce the sales fluctuations, we use the month level rolled up data instead of retailer-week level data.

### Data used:

There is one primary data source

* Import\_LDESC: Retailer-Week level data with all the information corresponding to the SKUs
* Cleaning Rules:
  + Removed the records that contains the null values in APN
  + Removed the records corresponding to this particular LDESC 'Double D Sugar Fre Choc&Jely Ronds 1x70g'. Since this LDESC is wrongly mapped

Data Pre-Processing:

* Aus\_SKUMonthYr\_data\_v1:
* Rolled up the data to APN-Month-Year level
* Seasonality Index (SI) is calculated at the Segment Level
* BrandSales is calculated by rolling up the monthly level data to Brand-Month
* CategorySales is calculated by rolling up the monthly level data to Category-Month
* price\_per\_vol is calculated by dividing the total\_ValueSales with the total\_VolSales.
* Description Mapping File:
* Created the mapping file by taking all the required SKU level attributes and stored it for further usage.

### Level of model fit:

Unlike the proposed iteration where we chose the discontinued SKUs along with some similar SKUs, in this approach we fit a model at a segment level thus reducing the number of models to be build. Please find the segment wise SKU counts in the summary sheet attached in the results section.

### Data preparation:

We used the data till December 2019 to train the model. Below given are the preliminary filters that are applied to clean the data.

* Removing few points at the start of the SKU where the sales are particularly low, if these points are to be included than a separate flag indicator can also be used
* Removing SKUs with less than a year of data
* Date filter for training the model, we used data till December 2019 (Change depending on the available data)

Next, we have to find few of the discontinued SKUs present in the data. To flag these SKUs, we look for 70% drop in sales when compared to last four-month average (this logic can be tweaked to make it more strict or lenient). If the drop in sales is consistent for next 3 months then we flag the SKU to be discontinued. Among the flag discontinued SKUs check how many data points are present before the discontinuation period. To make the model understand the difference between pre and post discontinuation we have put a filter of at least 11months of data before discontinuation,

For the rest of the SKUs which are normal we used the same filters that are being used in TPO thread to select the modellable SKUs.



Filter out the data for only selected discontinued SKUs and the normal ones for the next step.

The list of final discontinued SKUs can be found in the (Segment)\_modelresults\_(General)\_re1.xlsx (“DiscontinuedSKUs” sheet) embedded in Results section

### Feature Engineering:

Now that we have all the SKUs required for modelling, we can create the required features to model the relationship between our IDVs and DV (log of volume sales). Apart from the SKU attributes (like flavor, pack size, product type etc.) and sales features, we have also tried creating few features which can capture the effect of assortment change. One of the main reasons behind including the discontinued SKUs in the train mix is to see what effect of assortment change on the sales of the similar SKUs. So, we need a particular feature which can indicate this change in assortment and also differentiates between the similar and dissimilar SKUs.

In the final results we used the score based on the Walmart paper but the TAE score has also been tried out for one of the segments. Below given are the explanation for each feature.

#### Similarity Score (WM):

This similarity score loosely based on the Walmart similarity score explained in the below given paper. This score tries to capture the effect of change in assortment and the expectation is that the direction of change is different for similar SKUs when compared to non-similar ones.



Below given is a sample calculation on how the score works. The score calculated using the presence of SKU attributes within the month/week. Let’s take brand for example, if a particular brand is widely present in a particular timeframe then the SKUs which belong to the brand should have a higher score indication presence of potential replacements.

When we remove a SKU from the assortment (as given in the sheet), we can see that the score reduced for the SKUs which are more similar to the one which is discontinued thus capturing the required effect.



Since by design the variable decreases for the similar SKUs, we already know the direction (-ve) of the coefficients we would want the variable to have if the effect is visible in the data. Apart from the similarity score we also add in a flag to differentiate between the prep and post period for the discontinued SKUs. This done to ensure the model learns the difference in sales for these SKUs. We have also clustered the SKUs together based on the Sales History to help model differentiate between inherently low and high selling SKUs.

Final variable list can be found in the (Segment)\_modelresults\_(General)\_re1.xlsx (“Variables” sheet) embedded in Results section

#### TAE (Total Assortment Effect) Score:

This similarity score loosely based on the similarity score explained in the below given paper from oracle.



Below given is a sample file with the TAE calculation illustration. This variable can be used as a proxy or along with the Walmart based similarity score as well.



### Modelling:

Once we have the features ready, we fit a lasso at a segment level. We can either use the complete set of variables or do a variable selection before fitting if the datapoints are too low. For our iteration we have used the RF variable importance to choose the final IDVs.

Performance metrics can be found in the (Segment)\_modelresults\_(General)\_re1.xlsx (“Performance” sheet) embedded in Results section. Similarly, the coefficients can also be found in the sheet “Coeff”.

Retention calculation and evaluation:

To calculate retentions, we selected the month of Jan 2020. Now in this month we prepare the data in the required model format for all the SKUs present. We cannot prepare the data for the SKUs which were introduced in that month as there is no history present.

For example, let’s say there were 11 SKUs present in the month of Jan 2020, out of which only 10 had history. Now we can loop through these 10 SKUs and calculate the retention. Below are the steps used to arrive at the numbers.

In every loop we are trying to simulate a scenario where the SKU under consideration is discontinued. Given this change in assortment, our goal is to find what percent of the sales would be retained by others in the mix?

Let’s say we are trying to calculate retention for SKU\_1 (SKU under consideration) among the mix of 10 SKUs

Calculate the sales lost: loss in sales assuming the SKU\_1 is not discontinued

To get this number we create two estimates, one with trans\_flag (flag that indicates pre and post discontinuation period) as zero (Prediction (Old WithZeroFlag)) and another with flag as one (Prediction (Old WithOneFlag)).

The difference between the estimated will give the amount of sales lost if the SKU under consideration is not discontinued.

Sales Lost= Prediction (Old WithZeroFlag)- Prediction (Old WithOneFlag)

Next, we try to simulate the scenario where SKU\_1 is discontinued. For this we remove SKU\_1 from the mix and recalculate the similarity score and number of SKUs per month variable.

We can now predict the estimates (Prediction (New WithZeroFlag)) for the remaining 9 SKUs using the recalculated scores.

To calculate the lift generated from these SKUs we filter out the ones with incremented sales estimate when compared to earlier estimate with zero flag (Prediction (Old WithZeroFlag))

Apart from incremental sales we should also take into consideration the similarity between SKUs. So, the next step is finding some of the similar SKUs to the SKU\_1.

As explained in the feature engineering section that there will be a reduction in the sim score value for the similar SKUs after the assortment change

Hence, we create a score reduction flag checking for reduction in the similarity score before and after reduction

The final SKUs selected from the 9 are the ones which have incremental sales estimate along with a reduction in similarity score

Let’s say we get 3 similar SKUs, then we can calculate the lift generated from these 3 SKUs using the below equation

Lift=sum (Prediction (New WithZeroFlag)- Prediction (Old WithZeroFlag))

Retention%= (Lift/Sales Lost) \*100

Final retention metrics can be found in the (Segment)\_modelresults\_(General)\_re1.xlsx (“Retention” sheet) embedded in Results section.

### Results:

|  |  |
| --- | --- |
| **Category** | **Sheet** |
| Summary SKU Counts |  |
| Comparison Gum WM Vs TAE |  |
| WM Bitesize |  |
| WM Gum |  |
| WM Self Consumption |  |
| WM Share pack |  |
| TAE Gum |  |

### Codes:

|  |  |
| --- | --- |
| **Category** | **Code** |
| Data Clean/Pre-Processing |  |
| WM Bitesize |  |
| WM Gum |  |
| WM Self Consumption |  |
| WM Share pack |  |
| TAE Gum |  |