BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Appliances Retail:
Real World Data Science project

Group O

Eleonora Sbrissa, M20200628

Luis Reis, M20200636

Pedro Godeiro, M20200396

Sara Michetti, M20200626

May, 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

For more information related to the python code please check the Github repository available here.

1. BUSINESS UNDERSTANDING

1.1. BACKGROUND

This project aims to apply the group's knowledge over a real-world problem. Through this challenge, the group aims to learn how to deal with 27 Gb of data, how to transform and use them to complete the business objectives. The dataset provided contains information related to an appliances retail chain, that contains 410 points of sales spread in Australia. The products sold are over 8k, grouped in 21 families of product.

1.2. BUSINESS OBJECTIVES

The Business would like to understand each Point-of-sale characteristic through a quarterly analysis of the top products sold, market share preferences and product co-occurrences. They want then to group clusters according to similar characteristics and have an analysis considering value and product preferences. Lastly, they want to have a forecast of the next 6 weeks, related to units sold and point-of-sale.

1.3. SITUATION ASSESSMENT

The initial dataset contains the daily sales of each item inside each store of the chain, containing information related to both units sold and sales. The records of sales go from Jan 2016 till Nov 2019. Products information are built in a hierarchy, starting from family until SKU level. All the names are identified through numbers for privacy of the company.

1.4. DETERMINE DATA MINING GOALS

The goals of this project are:

- Understand each point-of-sale characteristics: The team built a dashboard in Power BI to show the quarterly analysis of each point of sale. The analysis aims to show the top products sold and the market share preferences at Family and Category level. The team implemented also a table with products co-occurrences to understand which products are mostly bought together.
- Point-of-sale clustering: to group different types of point-of-sale, products were first clustered at category level and then point-of-sales clusters were built using the previous results using k-means and hierarchical methods.
- Forecast: two forecasts have been built, one focused on products sales and the other one focused on which products will be sold in each point of sale. Both forecasts focus on the 6 weeks after the end of the sales records.

1.5. DATA PREPARATION

The raw dataset contained 19GB of data, with 9 different columns and over 180M rows. 5 columns were related to the product hierarchy and the others were about point of sale, date of sale, units sold, and value. As it is hard to work with a dataset as big as this one, the first thing to do was to separate it by chunks and remove in each column the unnecessary text type characters to lighten the memory.

Some strings and titles were also shortened up so they would not occupy space but have the same meaning. After making those changes, the chunks were merged into a new dataset with a good size but still not ideal. Next, the team pivoted the column 'Measures' into 2 columns, 'Value' and 'Units' to have half of the rows. Lastly, the duplicates have been removed. As a result, the current dataset has a weight of 4Gb: the number of columns is the same that as the original dataset and rows quantity is halved. Multiple Jupyter notebooks are present on the Github folder for the steps here mentioned.

2. PREDICTIVE ANALYTICS PROCESS

2.1. DATA UNDERSTANDING

Data is related to daily sales of 410 different point of sales, belonging to the same chain.

The different columns of the dataset build the product hierarchy, available in the next figure:

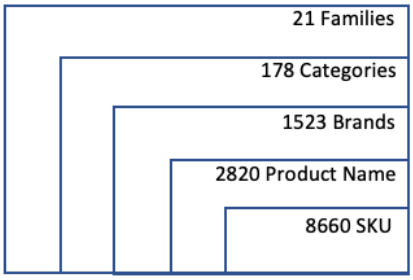


Figure 1: Products Hierarchy

There is no information related to the product type, all the different names were covered for privacy issues. All the names are identified through numbers.

Looking at years 2016-2018, the trend of units sold and total price is equal towards the different years. Here below an analysis that joins the 3 years (2019 was excluded as there is no info related to Nov and Dec):

Figure 2: Average of units sold and average of total price. Years 2016-2018

There is a peak on average of units sold in January and December, as well as a peak on average total price. There is another peak on the average price in October and November which means that during these months the products sold are more expensive. In the months of April and August even though the average of total price decreases, the number of units stays on average: this means the stores in those months sold products with discounts.

Regarding 2019, there is no information regarding the months of November and December, but the trend related to the other months follows the same as the previous years.

Among the 21 families available in the market, here below an analysis based on units sold across the 4 years:

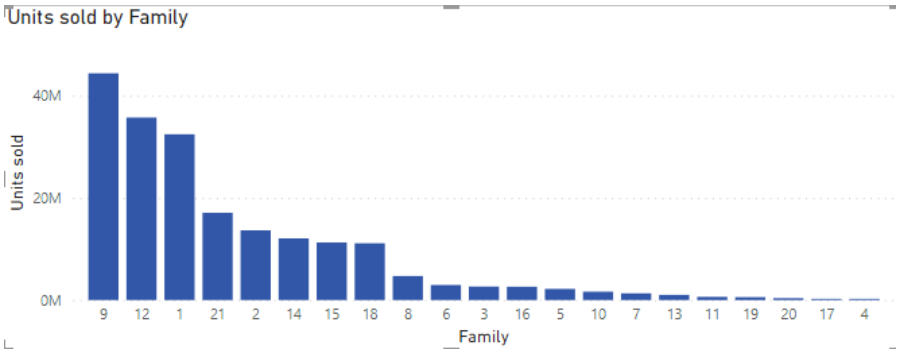


Figure 3: Total units sold by Family

The most sold family is number 9, followed by number 12.

The biggest store in terms of units sold and revenue is store number 292, which takes the lead across all the years.
The most sold category is 178, which accounts for 70% of the total units sold across the stores. Looking at the following graph we can see the number of units sold related to this category distributed across the months, and the total price divided by units sold.

Figure 4: Category 178 - Units sold and total price divided by units sold

The positive trend is a sign that for this category, even though the units sold stay the same, the price increases throughout the year, with a clear increase in the months of June and October.

The report will be now divided into 4 parts, according to the different challenges the Company wants to address. The first part will be related to Point -of-sale quarterly analysis, the second one on point-of-store product co-occurrences, the third one on the clustering, the fourth one will focus instead on the products forecast.

2.2. PoS Quarterly Analysis of Top Products Sold & Market Share

In the notebook called point-of-store analysis the team did an analysis related to the top 5 products sold per store per quarter, throughout the 4 years. Four tables were built to show for each store the

top 5 products and the units sold for each one. As the table is yearly, there is the possibility to check how the top 5 products change across the year. In the same notebook there is also an analysis related to the market share per family and category: the team built yearly pivot tables in which you can check the distribution of families and categories in terms of units sold throughout the quarters for each store. An example was made available for store 292.

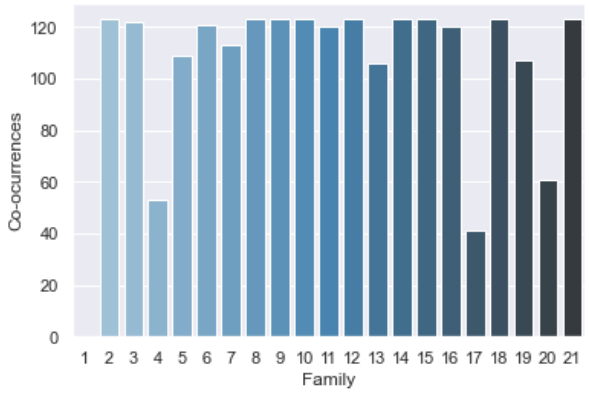
For a clearer understanding of the analysis made on python, the team built a dashboard using Power BI accessible at this link. Each viewer can select the store and have a general overview of the quarterly units sold as well as the yearly trend related to total price divided by units sold. By then selecting a specific year and quarter, first thing to show up is a KPI based on last year same quarter units sold, in order to understand the trend of the selected quarter based on previous year data. There is also a card showing the total sales for that quarter. Next, a table shows the top 10 products in terms of units sold for the selected quarter. The hierarchy goes down to product level. Finally, there are two graphs explaining the market share division for both family and category in percentages for that quarter.

2.3. PRODUCT CO-OCCURRENCES

The client asked for a quarterly analysis of product co-occurrences, so the team decided to analyze only 2019 since throughout the years there are multiple changes in price of products, discontinuity, Promotions.

To better understand the co-occurrences, a function was created where the user can indicate the point-of-sale and which quarter of the year he wants to check, and it provides a visualization of the family of products co-occurrences on that specific store for that specific quarter.

It was created an example with the store that has the most sales throughout the years which is 292, for the co-occurrences on the 1st quarter for the product family 1.




```
(Family
1      0.0
2    123.0
3    122.0
4     53.0
5    109.0
6    121.0
7    113.0
8    123.0
9    123.0
10   123.0
11   120.0
12   123.0
13   106.0
14   123.0
15   123.0
16   120.0
17    41.0
18   123.0
19   107.0
20    61.0
21   123.0
Name: 1, dty
```

Figure 5: Product co-occurrences for Family 1 during quarter 1 year 2019

The team also created association rules to check correlations between the product families, also to see if some families of products are more sold with others and which products are more likely to be bought together.

For the first quarter of 2019 we have the following rules. Meaning that the product families on the consequents are very likely to be bought if the antecedents are bought.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(3, 8, 9, 12, 15, 16, 21)	(1, 2, 5, 14, 18)	0.818695	0.740560	0.669296	0.817516	1.103915
(1, 2, 5, 14, 18)	(3, 8, 9, 12, 15, 16, 21)	0.740560	0.818695	0.669296	0.903770	1.103915
(3, 8, 9, 12, 16, 21)	(1, 2, 5, 14, 15, 18)	0.818903	0.740373	0.669296	0.817308	1.103914
(1, 3, 8, 9, 12, 16, 21)	(2, 5, 14, 15, 18)	0.818903	0.740373	0.669296	0.817308	1.103914
(3, 8, 9, 12, 16, 21)	(2, 5, 14, 15, 18)	0.818903	0.740373	0.669296	0.817308	1.103914
(2, 5, 14, 15, 18)	(1, 3, 8, 9, 12, 16, 21)	0.740373	0.818903	0.669296	0.903998	1.103914
(1, 2, 5, 14, 15, 18)	(3, 8, 9, 12, 16, 21)	0.740373	0.818903	0.669296	0.903998	1.103914
(2, 5, 14, 15, 18)	(3, 8, 9, 12, 16, 21)	0.740373	0.818903	0.669296	0.903998	1.103914
(1, 3, 8, 9, 12, 15, 16, 21)	(2, 18, 5, 14)	0.818695	0.740581	0.669296	0.817516	1.103884
(2, 18, 5, 14)	(3, 8, 9, 12, 15, 16, 21)	0.740581	0.818695	0.669296	0.903745	1.103884

Table 1 : Quarter 1 Year 2019 Rules

2.4. CLUSTERING

Two types of clusters have been run for the client to understand and generalize better their business:

- Store perspective: taking in consideration only units and value dimensions for different years;
- Product perspective: taking in consideration how different categories are sold across store clusters

Every cluster is in its own Jupyter notebook for an easier analysis. You can find them on GitHub.

2.4.1. Store Value Clustering

For the first clustering, the data have been cleaned and turned into a table of the following variables: units sold in 2019, revenues in 2019, the growth in both units and value from 2018 to 2019 and from 2017 to 2018 and the average price of the products by store.

After scaling the data and calculated the correlation, both units 2019 and unit growths have been dropped due to the high correlation with their counterparts in the value section.

A hierarchical clustering with 100 clusters and the inertia plot were run to check which is the appropriate number of clusters for the k-means algorithm. All the plots and tables can be found in the notebook “6.Clusters_Store” on GitHub. After choosing 5 as the number of groups, the following graphs show the results:

Figure 6: Mean distribution of units (left) and value (right) across clusters

As noticeable, cluster 2 is the one that counts most of the units and revenues, but the highest growth can be seen from the stores belonging to cluster 0. Cluster 1 3 and 4 are the clusters with the lowest unit and value on average; having cluster 1 the same value growth as cluster 2. The dimension of the clusters can be seen inside the table on the left graph. Cluster 3 is the most popular, on the other hand cluster 2 is the least popular.

The R^2 for this clustering is 0.61.

2.4.2. Product Preference (Category) Clustering

To get to the product preference at category level of stores, two clustering needed to be calculated.

First, as the business has 178 categories, it was needed to group the categories discriminating between the units sold in 2019, the growth between 2018 and 2019, the store count for every category (in sales) and its growth compared to 2018, the average price per category across the years and the percentage split of the category sales on quarters.

After scaling and dropping the mostly correlated variables, it was decided to keep only the categories which were continuously sold throughout the past two years, 2018 and 2019. As a result, 16 categories have been dropped.

With hierarchical clustering on k=100 and the inertia plot, 4 clusters of categories have been found.

Figure 7: Clusters characteristics

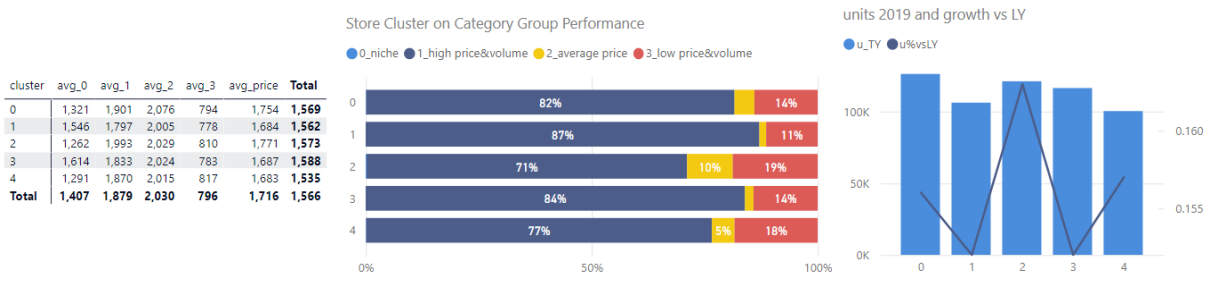
On the left graphs is possible to see the distribution of the units throughout the quarters. For example, the categories in cluster three are mostly sold during the first and the fourth quarter, while the categories belonging to cluster two prefer the central two quarter.

The count per cluster can be seen in the table in the centre: the least common cluster in the number 3, while the first one is the most popular.

On the right side, the average price per cluster for 2019 and how many stores sell the category out of the 410 stores in the business. The bottom right plot instead shows the units sold on average by every store in 2019 and the growth vs the previous year. Cluster one is the one selling the most, while Cluster 2 is the one growing the most compared to 2018. The R^2 for this clustering is 0.89.

2.4.3. Store and Product Preference Clustering

Last, stores performances have been checked for these categories' clusters, and the following clusters of stores have been found:



cluster	avg_0	avg_1	avg_2	avg_3	avg_price	Total
0	1,321	1,901	2,076	794	1,754	**1,569**
1	1,546	1,797	2,005	778	1,684	**1,562**
2	1,262	1,993	2,029	810	1,771	**1,573**
3	1,614	1,833	2,024	783	1,687	**1,588**
4	1,291	1,870	2,015	817	1,683	**1,535**
Total	**1,407**	**1,879**	**2,030**	**796**	**1,716**	**1,566**

Figure 8: Product preferences per store cluster

Cluster 2, that is the smallest group, is the one that sells the most "non-standard" categories. In the stacked bar plot is noticeable that they do sell the group 2 and 3 of categories in a higher portion compared to the rest of the clusters. It also has the highest average price, except for the categories belonging in group 0, that is, as seen before, the group that sells the least. Regarding the group 0 of categories, cluster 2 is the one selling the most of it (0.1% of the total sales) but they are also the ones with the lowest average price for these categories. Store in the cluster one on the other side are the ones selling the most the standard products (87%) and have the lowest growth of units vs 2018. The R^2 for this clustering is 0.78.

2.5. FORECASTING

In order to make predictions on units sold by each product by store, it is needed to train a forecasting model for each product name and for each store. Considering there is 2820 product names and 410 stores, it would not be feasible, in terms of computing power, to perform complex Machine Learning algorithms for this task.

Considering that, it was decided to use simple univariate statistical models to perform the predictions. The first idea on mind was to use ARIMA models to perform the predictions, since they are vastly use to make forecasting on time series. Although, ARIMA models require parameter settings to fit the model, and one way to avoid that is using Auto ARIMA, a function that set the parameters automatically, but in the training each model would take more than 1 minute to train.

Considering it is not possible to use the ARIMA model, other alternatives would be using other well used statistical models to train our time series. Looking for some references, we realized that the Theta Model and ETS Model have good performances on forecasting, so it was decided that both would be used to forecast since they are fast and good performance algorithms.

The whole idea of the forecasting is to create a pipeline in which for each product name the Theta and ETS models are trained (dataset from 2017 to June 2019) and both evaluated on the test dataset (dataset from July 2019 to October 2019). If one of the models have a R2 score higher than 0.13, that represents a medium effect size according to Cohen (1992), it can be said that this model can be used to forecast units sold for November and December 2019. If both models are above the threshold, the one of the highest score is selected.

The problem with using this logic is that since several products do not have a consistent sales pattern, we will have a lot of products that did not manage to train a good enough model, leaving them without predictions. This happens because the models couldn't find the seasonality, in which for this case was identified to be 52 cycles, considering we work with weekly values. To assert this problem, when a product name did not manage to reach a good forecasting model, its category will be trained in the same pipeline and use the proportion of sales from the product in the category to make the prediction. If the category from the product still fails to train a good model, the family of the product will be used with the same logic. In the table below we can check the performance of the models.

	ETS Model	Theta Model	No model found
Product Name	16%	5%	79%
Product Category	19%	18%	63%
Product Family	53%	38%	9%

Table 2: Performance of the diferente models

To make predictions at store level, it was decided to use the product predictions and spread them across the point of sales based on the contribution of each product name on the store. Specifically, two contributions were taken and averaged out: the last four weeks of 2019 (weeks 45, 46, 47, 48) and the next four weeks for the prior year 2018 (weeks 45, 46, 47, 48). Both these were averaged out, understanding how much the specific product takes up, percentage wise, on the store sales for both before and after the week in which the forecast is starting. Like this, the product prediction could tackle on the stores which actually sell the product. Considering that for some product names it was not possible to forecast, not even at family level, 400 store-name combinations could not be forecasted.

This solution just offered here was deployed for this case because it was going to take too much computer power to do predictions of time series for each store-product combination.

3. DEPLOYMENT AND MAINTENANCE PLANS

3.1. DEPLOYMENT

In order to apply the models created in forecasting the products being sold for each store in the future, it is interesting that the models being used ate updated with time. This means that if one month is by, we could use the previous month to forecast the next months. It is important that we keep updating the model as the time go by, because we could see that years too much in the past were not helping with training the model.

Not only that, but it is really important that the predictions should be used to guide strategy from each store, understanding that these predictions could be key to prepare stock for each product in the future.

3.2. MAINTENANCE

Both Product Co-occurrences and clustering analysis can be performed quarterly, when there is the biggest shift of the performances per season. Eventually both clustering can converge to one final solution in the future in which both the category groups and the store clusters can be calculated at the same moment.

4. CONCLUSIONS

As for the conclusions, working with this set of data was challenging enough. Most of the time spent by the team was in first place, making sure that the data was not large enough to be used in Python with the pandas package; but secondly, preparing the data for any type of analysis. Each small analysis or table had to be pivoted, cleaned, merged and reviewed by the team multiple times. The fact that 2019 did not include data for November and December also created a problem in seeing the differences in quarters, which the team solved averaging out the quarter 4 for the other years. Each small set of problem had to be thoughtfully taken in consideration.

Interpretation was also a big challenge for this business case. Not being able to understand which products and categories we were talking about put a limitation on the interpretability of any result of the analysis that are reported here. Also, not being able to know where the point of sales were located, put an extra layer of difficulty to the case.

In any case, all the analysis has been performed, and the team is confident that they can be used by your business to improve the way you see your products. Clustering the categories and seeing them reflected of groups of stores should help understanding where to invest more both from a product and point of sales perspective, taking also in consideration the market share analysis and dashboard created. Both association rules and product co-occurrences can make you better understand at store level which kind of customers you have, and which are their needs. Finally, the forecast, to always have on store the right number of products, making sure to replenish them on time and have satisfied customers.

4.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

More qualitative data is needed to improve what have been done until now. A market basket analysis can be performed if also transactional data is available and if both promotions, discounts, and inventory data are present.

Regarding what the team did until now, anything that it could not be performed because of machine power can be improved. Specifically worth mentioning the association rules, forecast at store-product combination.

5. REFERENCES

https://www.statsmodels.org/stable/examples/notebooks/generated/theta-model.html

https://www.statsmodels.org/devel/examples/notebooks/generated/ets.html

https://machinelearningmastery.com/findings-comparing-classical-and-machine-learning-methods-for-time-series-forecasting/

https://link.springer.com/content/pdf/10.3758/BF03207705.pdf