BUSINESS CASES WITH DATA SCIENCE

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

PoS Appliance's Retail – Real World Data Science Project

Group F

Lorenzo Pigozzi	m20200745
Nguyen Huy Phuc	m20200566
Ema Mandura	m20200647
Xavier Gonçalves	m20201090

June 2nd, 2021

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

CONTENTS

1. INTRODUCTION

As a fundamental strategy for any retail company, it is important to analyze the existing information on transactions and points of sale, as well as preferences in terms of products. For this, the time dimension is relevant as well and can lead to very different and valuable insights. By studying points of sale, products, consumer behavior in terms of purchasing patterns and co-occurrences, retail companies can be one step ahead in terms of attracting new customers through marketing and strategy, while engagement of current customers to increase their purchases in the future, and even in cost containment and smarter and more efficient logistics management.

Moreover, insight captured form the analysis will be used as crucial information to effectively create the best forecasting models for future sales and revenue. Thus, enhancing the capability of decision making in terms of not only business planning and budgeting but also processes and resources optimization.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Point of Sale Appliance's Retail consists of a dataset for transactions carried out at defined points of sale and spread across Australia. The transactions are related to the sale of household appliances at these points of sale. In total, the transactions refer to several thousand different products, of more than one thousand and fifth brands and almost two hundred categories of home appliances, which are sold in almost half a thousand points of sale spread across the continent.

This data is particularly relevant, and its analysis will create greater effectiveness in the future sales strategy and in the customer experience at the points of sale.

2.2. BUSINESS OBJECTIVES

The main objective is the analysis and transformation of existing data on points of sale over almost four years. The data will be processed to obtain a greater understanding of the characteristics of each point of sale, on a quarterly basis, relative to the range of best-selling products, family and product category, brands, and consumer preferences. Thus, an analysis will also be made in terms of market share. The co-occurrences of products are also very relevant in this analysis and will consider the existing patterns and relationships of purchases with the objective of increasing the chances that more items will be placed in the cart, if all the logistics and advertising dynamics in the points of sale is effective.

Moreover, clustering by point of sale will also be carried out, considering the value and preferences of the products, to which will be added the useful forecasts in terms of product sales and by point of sale.

In short, the objectives in terms of business will involve better knowing the points of sale and the customers, increasing the effectiveness of future campaigns and logistics, avoiding waste and shortages on the shelves, and improving the customer experience.

2.3. Business Success criteria

The success of the project can be measured by the feedback from the Point-of-Sale manager and Financial Director on operational efficiency improved from the insight generated from the dataset. The planning efficiency and accuracy also should be monitored to see the improvement in planning activities after the forecasting models are deployed.

Although the result of the project is instantly, due to the management characteristic of this type of project, long-term improvements are outcomes to be expected after the project has been successfully delivered. In this context, some examples of efficiency improvement are a more intelligent logistics organization, with more concrete and reliable forecasts in relation to stocks, as well as an analytical view on various indicators and products, as well as specific points of sale.

2.4. Situation assessment

The dataset contains data about retail transactions, such as the product family, category, brand, name, SKU, point of sale where it was purchased, quantity, total sales, and time reference, among other data. Inside, there are 21 families of products, 178 categories of products, 1535 brands, 8660 different SKUs, sold in a total of 410 points of sales throughout Australia in almost 4 years, between January 1st. 2016 and November 1st. 2019.

Although the amount of data is sufficient to carry out meaningful analysis, the size of the data could be a significant challenge for any data science process. Thus, it is necessary to preprocess and reduce the size of the original dataset in order to improve the efficiency in the later data mining processes. Dealing with a large dataset, well developed Business Intelligent tool such as Power BI will be used to carry out quick analysis and dashboard to understand the pattern and insight from the dataset. Next, for clustering and forecasting, specific aggregated data will be generated to support developing the clusters analysis and forecasting model creation.

2.5. Determine Data Mining goals

In this context, the strategy used in terms of data mining will be a set of different analyzes of information about products and points of sale, which will evaluate the characteristics of each point of sale, by quarter, the set of best-selling products, but also the co-occurrences of products, which products are bought together. Customer preferences and families and product categories will also be scrutinized to try to understand reality as effectively as possible with the existing data.

Clustering by point of sale will also be carried out, considering the value and preference of products.

In addition, two types of forecasts will be borne by the team, relating to the time interval between the end of the available data and six weeks past that day. One of the forecasts will aim to anticipate sales of products and the other these sales referring to points of sale.

By conducting this comprehensive and extensive analysis, many insights can be drawn to improve and fulfill business objectives, creating value.

3. DATA MINING PROCESS

3.1. DATA UNDERSTANDING AND PREPARATION

First, the first problem that we had to face was the size of the dataset. Indeed, it's almost 20 giga-bytes, and due to our computational resources, we had to find a way to preprocess it and at the same time reduce the size, without losing any kind of meaningful information.

The csv file contains 182342304 rows and 9 columns.

ProductFamily_ID	ProductCategory_ID	ProductBrand_ID	ProductName_ID	ProductPackSKU_ID	Point-of-Sale_ID	Date	Measures	Value
Family_16	Category_11	ProductBrand_306	ProductName_649	ProductSKU_1970	POS_1	2017-03-04	Sell-out units	2.0
Family_16	Category_11	ProductBrand_306	ProductName_649	ProductSKU_1970	POS_1	2017-03-04	Sell-out values	1540.0
Family_16	Category_11	ProductBrand_306	ProductName_649	ProductSKU_1970	POS_1	2016-05-02	Sell-out units	4.0
Family_16	Category_11	ProductBrand_306	ProductName_649	ProductSKU_1970	POS_1	2016-05-02	Sell-out values	3080.0
Family_16	Category_11	ProductBrand_306	ProductName_649	ProductSKU_1970	POS_1	2016-10-24	Sell-out units	2.0

Figure 1 – Original Dataset Preview

The first 5 variables (ProductFamily_ID, ProductCategory_ID, ProductBrand_ID, ProductName_ID and ProductPackSKU_ID) contain the encoded characteristics of each product purchased.

Point-of-Sale_ID is the identification number of each store of the company. Moreover, the variable "Measures" explains if each record stores the "Value" regarding the Quantity or the Monetary Value: indeed, if the Measures variable is "Sell-out units" the "Value" variable contains the quantity of the product purchased, if the "Measures" is "Sell-out values" the variable "Value" contains the amount of Monetary Value.

Furthermore, it's important to know that the dataset was originated from a transactional document, and the one we're analyzing is basically the same structure but with transaction aggregated by Point of Sale, Date and Product SKU, that is the most granular level of the hierarchy of the products.

In order to preprocess the dataset and to reduce the size, first our team started dropping all the useless part of the text that all the records contain for the categorical variables using a regular expression, and keeping just the actual number corresponding to the ID. Then, we also changed the datatype of the variables, in order to get them even more manageable and easy to read. Furthermore, we noticed that for each aggregation of transaction, there are 2 different records stored: one for the quantity purchased for that product and one for the monetary value. Thus, for each pair of records the values of all the other variables are repeated except the "Value" variable, and so we decided to split the dataset in 2 based on "units" or "value" stored in the "Measure" variable and subsequently adding a new column for only one of the records.

As the dataset provided is similar to a transactional system, the records stored are ordered, thus both the 2 partial datasets created, df_units and df_values have the same order, and the "Value" variable of df_units for the record 'i' retrieves the quantity of the record 'i' in df_values.

Thus, even though the best practice to do this merging of the 2 datasets would be a join operation based on "ProductPackSKU_ID", "Point-of-Sale_ID" and "Date", due to computational issues of the operation we can just store df_units["Value"] as df_values["Quantity"], based on the index reset.

3.2. EDA and Quarter Analysis

While performing exploratory data analysis, top 10 products across all the points of sale were detected, as shown in Figure 2. For those products, we looked at the sales over the years and quarters. Two distinct trends can be distinguished, as seen on the line graph in Figure 3. The products either have a continuous growth over the years or they fluctuate over quarters. The latter trend has a pattern of distinct increases in sales in the first and last quarter and decreases in the second and third quarter.

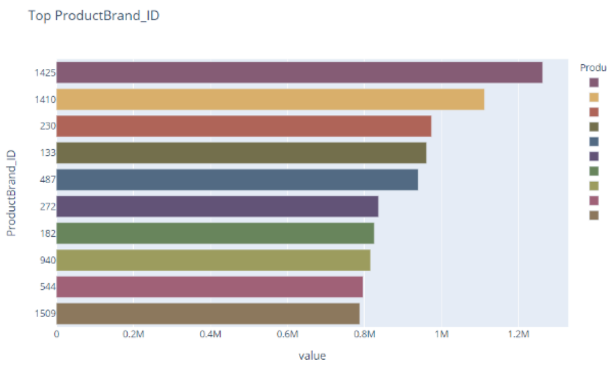

Figure 2 – Top 10 products

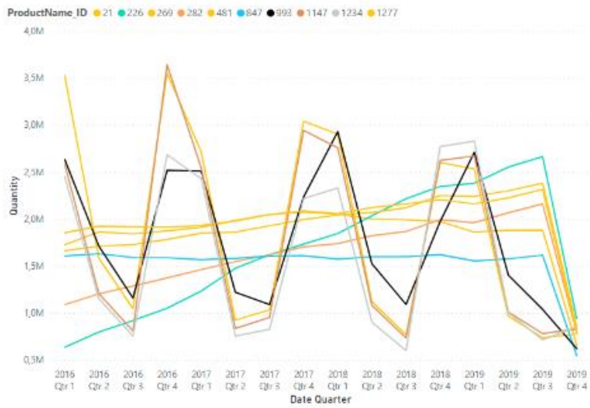

Figure 3 – Top 10 products over time

For point-of-sale analysis, a Power BI report was created that allows the client to visualize every point of sale by both year and quarter. The screenshot in Figure 4 shows an example page of the report. The user is prompted to choose a point-of-sale, year, and quarter. For each combination of these filters, the page shows the top ten products by both quantity and total sales value, as well as the market shares of product categories and product families. Additionally, the user can observe the sales of the top 5 products by quantity over the months. The information provided by these visualizations can help the management see the patterns and preferences specific to a place and time and consequentially make the right marketing and supply-vs-demand decisions.

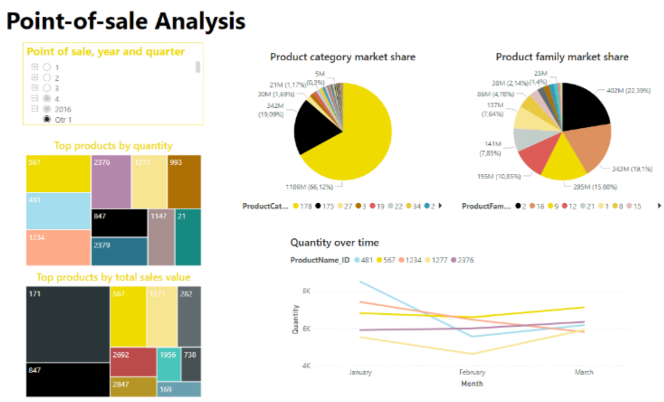

Figure 4 – PBI page for Point-of-Sale 4, year 2016, quarter 1

4. RESULTS EVALUATION

After the quarter analysis that was requested, we also performed two other analyses, that were more challenging in terms of data science approach. The results got in the data exploration and the seasonality check were useful in order to proceed and go deeply into the project, and to get more solid results.

4.1. CLUSTERING POINT OF SALE

The first analysis we caried on was the clustering challenge of the Point of Sales that the company requested. We considered two different perspectives in order to analyze the stores, by Value and by Product Preference.

4.1.1. Clusters by Value

For the clustering by Value, we generated a new dataset that was aggregated by Point of Sales, and 3 variables were considered: the Quantity, the Monetary Value and the Average of Monetary Value, that was just calculated by (Monetary Value / Quantity). The values stored are based on the mean of the aggregation of the records that belong to each of the Point of Sale.

After that, we scaled the data using a Standard Scaler and then we decided to use the K-means algorithm in order to cluster. Analyzing the inertia plot, we decide to choose 3 clusters.

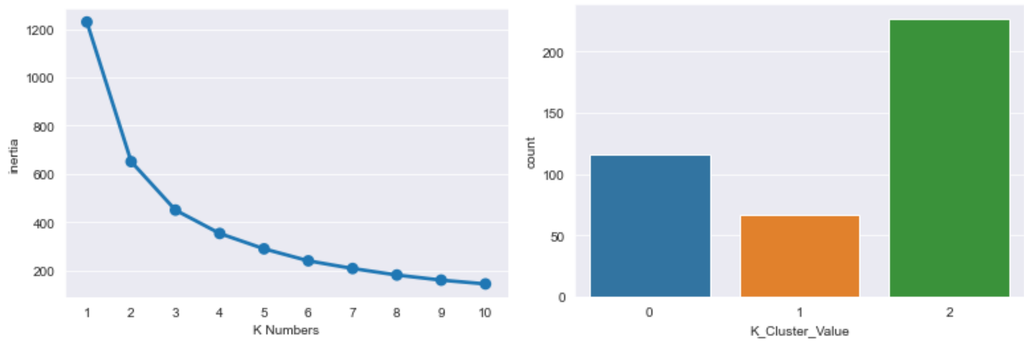

Figure 5 and 6 – Inertia Plot and Cluster's distribution

For this solution, we got a successful R-squared score of almost 70%.

Due to the lack of other variables rather than the categorical ones that define the product category, the profiling phase of the clusters was challenging, nevertheless we were able to identify the main characteristics of the 3 clusters based on the same variables that we used to cluster.

Cluster 0: "The Boutique"

Characteristics:

- Lowest Quantity of products purchased.
- Lowest Amount of Monetary Value.
- Highest Average of Monetary Value.

Cluster 1: "The standard Store"

- Highest Quantity.
- Highest total amount of Monetary Value.
- Regular Average of Monetary Value.

Cluster 2: “The Discount”

- Regular Quantity of products purchased.
- Regular Amount of Monetary Value.
- Lowest Average of Monetary Value.

In the plots below it is possible to see the differences that were detected for the profiling of the clusters.

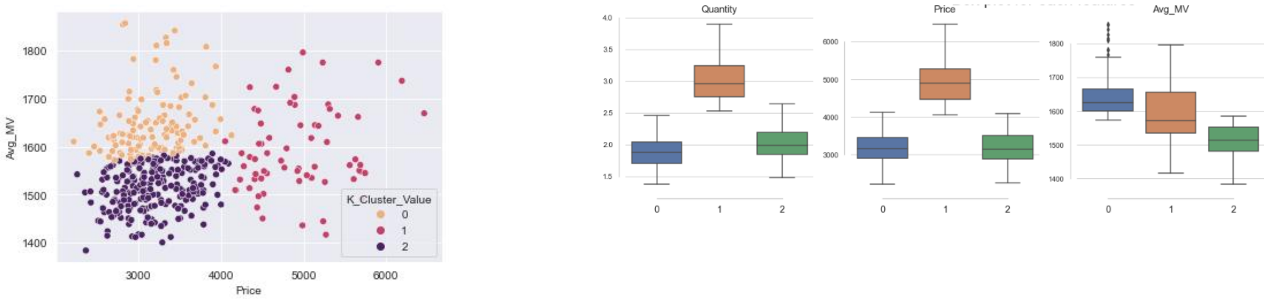

Figure 7 – Scatterplot and boxplots of the Cluster Solution of the PoS

4.1.2. Clusters by Product Preferences

In order to complete an analysis based on the product preferences for each of the point of sale, first we performed a groupby of the data based on Point of Sale and ProductFamily_ID. Then, on top of the data structured in this way we compute a pivot table, which gave us a result in which the index of the dataset is the ID of the Point of sales, and the columns are the 21 families of the product. The values stored in the table are the counts of the quantity, and basically retrieves the number of aggregated transactions that each store has for each of the product family. To complete the preparation of the dataset for this purpose, we also standardize the values.

ProductFamily_ID	1	2	3	4	5	6	7	8	9	10	...	12	
Point-of-Sale_ID													
1	0.170430	0.055578	0.014502	0.001244	0.007939	0.012418	0.005455	0.019979	0.224637	0.012917	...	0.193548	0.(
2	0.186305	0.059973	0.013296	0.000489	0.007232	0.011279	0.006276	0.020150	0.233292	0.008366	...	0.211235	0.(
3	0.189802	0.056752	0.013711	0.001008	0.006277	0.014679	0.006666	0.018486	0.241554	0.007252	...	0.199386	0.(
4	0.170438	0.042863	0.009649	0.000582	0.004488	0.012526	0.004153	0.013903	0.223651	0.010925	...	0.224961	0.(
5	0.161121	0.057175	0.013409	0.000751	0.007078	0.012736	0.005511	0.016340	0.258256	0.009598	...	0.213251	0.(

5 rows × 21 columns

Figure 8 – Preview of the Pivot Table used for clustering.

The choice for the solution we got is also that we decided to compute a Principal Components Analysis on top of the Product families. In this way, we reduced the number of variables considering for clustering and at the same time we can use the components to summarize the important patterns along the count of quantities for the different product families. We consider only 5 principal components, representing the 21 product families, because they already explained the 95% of the variability.

Also in this case, we performed the K-Means algorithm, and based on the inertia plot we got a result of 3 clusters. The R-squared score for the cluster solution in this case was 0.65%.

We profiled the clusters based on the combined interpretation of the clusters themselves and of the Principal Components. Based on it, we were able to identify different groups of product families that are mainly purchased in one or more of the clusters of the Point of Sale.

Cluster 0

- High percentage of Product-Families: 9, 2, 18, 3
- Low percentage of Product-Families: 21, 11, 10, 20

Cluster 1

- High percentage of Product-Families: 15
- Low percentage of Product-Families: 12

Cluster 2

- High percentage of Product-Families: 21, 11, 10, 20
- Low percentage of Product-Families: 9, 2, 18, 3

The table below shows the interpretation of the principal components that we used to profile the clusters.

		PC0	PC1	PC2	PC3	PC4	PC5
	ProductFamily_ID						
Quantity	1	-0.493000	0.308954	0.153471	-0.795657	-0.041922	-0.019788
	2	-0.727867	0.228528	0.295782	0.360345	-0.118803	-0.408422
	3	-0.533627	0.157969	-0.041832	0.195632	0.097227	-0.311371
	4	0.008231	0.106238	0.296871	-0.079438	-0.172893	0.170633
	5	-0.520571	0.206640	0.102412	0.222667	-0.049782	0.199694
	6	-0.279098	-0.161363	-0.298297	-0.228884	0.100456	-0.043682
	7	-0.321117	0.110296	0.042442	0.214823	-0.284239	0.208518
	8	-0.443859	0.239074	0.267360	0.454259	-0.303889	0.150357
	9	-0.845480	0.161837	-0.507303	0.035288	0.002126	0.007267
	10	0.581810	-0.206502	0.228598	-0.094543	-0.062004	0.319597
	11	0.604537	0.047805	0.327736	-0.003749	-0.114538	0.144709
	12	-0.526718	-0.847031	-0.033954	-0.054918	0.008292	-0.009621
	13	0.374740	0.211058	0.454074	0.082347	-0.028800	0.241625
	14	-0.069483	0.308372	0.558306	0.501219	-0.066223	0.466539
	15	-0.300187	0.381321	0.342885	0.162670	0.757559	-0.057372
	16	0.291299	0.153023	0.475949	0.004987	0.377968	-0.023924
	17	0.299958	0.163534	0.152348	-0.003688	0.047387	-0.073469
	18	-0.598348	0.273891	0.392612	0.309255	-0.328921	0.017780
	19	0.251169	-0.161216	0.009657	-0.004789	-0.136532	0.177963
	20	0.577980	0.023705	0.284443	0.018068	-0.253738	0.224906
	21	0.996131	-0.017611	-0.084003	-0.010644	-0.003683	-0.010491

Figure 9 – PCA interpretation

Furthermore, as a result we also tried to perform a hierarchical clustering merging the two previous results, by Value and by Product Preference. Nevertheless, data shows that merging the two previous cluster's solutions is quite difficult and not optimal.

For this reason, we will provide as final solution the original clusters created without merging them, and it will be possible to consider one perspective (clusters by value) or the other one (clusters by product preference) based on the business needs.

4.2. FORECASTING

4.2.1. Model evaluation

There are several robust timeseries forecasting methods including both classical models and modern machine learning models. To objectively evaluate the best model for the data of this project, we compared a classical method ARIMA (SARIMA) and a modern deep learning model RNN (LSTM) and evaluated the forecasting results.

The two models were evaluated using the metrics of Rooted Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) for the Out-of-Sample (20% of the whole dataset) time series forecasting with actual data.

Models	LSTM			SARIMA		
Time-step	T+1	T+2	T+3	T+1	T+2	T+3
RMSE	4565.48	4607.78	4997.17	2743.15	2955.68	2792.47
MAPE	0.0432	0.0480	0.0573	0.0285	0.0334	0.0316

The forecasts of 2 models were performed on the total sales quantity of product ID 2609 across all Point-of-sale, with the time period from 01/2016 to 10/2019. The result showed that the classical SARIMA model outperformed the modern one. Thus, the SARIMA will be used in the next processes.

4.2.2. Sales quantity forecasting

During the data exploration process, we discovered that there are two main sales patterns in the top products sold. The first one is the weekly fluctuated patterns and the second one is the seasonal sales patterns. Thus, it was decided to build and optimize two different SARIMA model for each pattern.

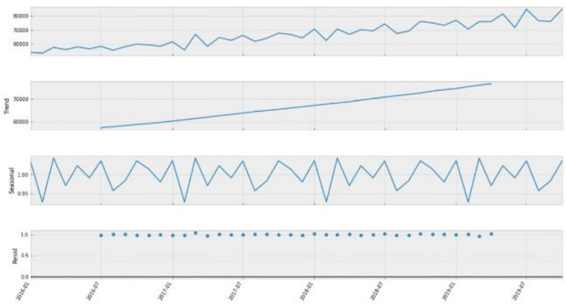

Figure 10. Pattern 1 - Weekly fluctuated (Product ID 2609)

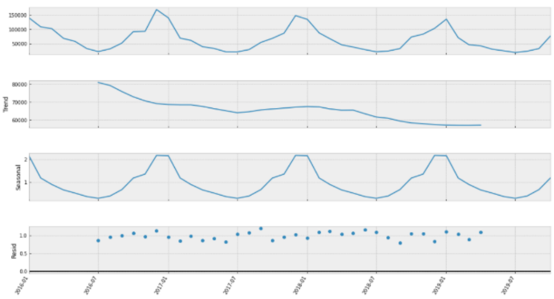

Figure 11. Pattern 2 – Seasonality (Product ID 481)

In order to find the optimal parameters of the SARIMA models, we combined the analysis of the auto correlation and partial correlation plot and the grid search method. The best SARIMA model for forecasting on both monthly basis and weekly basis for the top product sold across all Point of Sale will be saved for final production.

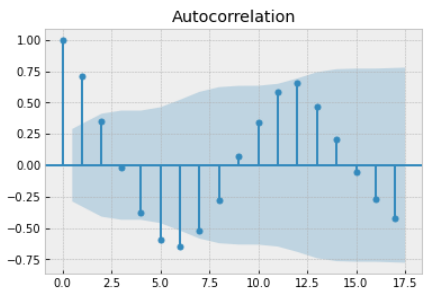

Figure 12 - Example of ACF plot

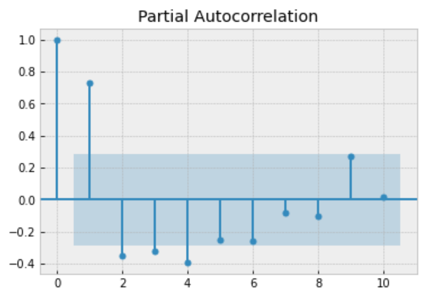

Figure 13 - Example of PACF plot

In order to comprehensively evaluate the models, we calculate the metrics of RMSE and MAPE for both in-sample forecasts (1 step ahead) and out-of-sample forecasts (3 steps for monthly modes and 6 steps ahead for weekly models) for each model on monthly basis and weekly basis. The test set cover 20% of the dataset will be used for evaluation the forecasted values.

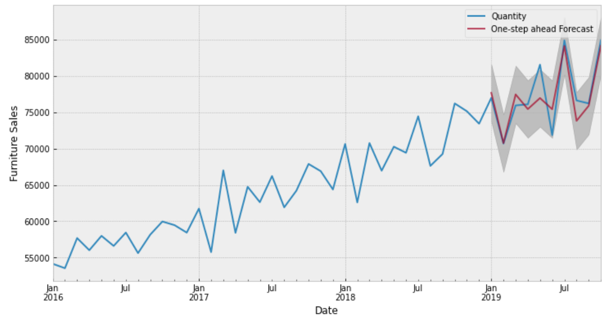

Figure 14 - Example of in-sample 1-step forecast monthly.

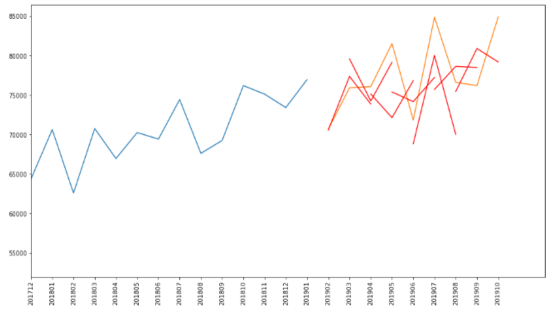

Figure 15 - Example of out-of-sample 3 steps forecast on monthly basis.

Evaluation result on best models found.

	Weekly fluctuated data (Product ID 2609)				Seasonality data (Product ID 481)			
Monthly forecast evaluation result	**SARIMA(1, 1, 1)x(0, 1, 1, 12)**		**RMSE**	**MAPE**	**SARIMA(4, 0, 2)x(0, 1, 2, 12)**		**RMSE**	**MAPE**
	In-Sample Forecast	T+1	2146.52	0.0203	In-Sample Forecast	T+1	5700	0.047
	Out-of-Sample Forecast	T+1	2743.15	0.0285	Out-of-Sample Forecast	T+1	7883.66	0.1370
		T+2	2955.68	0.0334		T+2	8228.25	0.1631
		T+3	2792.47	0.0316		T+3	5840.91	0.1108
Weekly forecast evaluation result	**SARIMA(0, 1, 1)x(0, 1, 1, 52)**		**RMSE**	**MAPE**	**SARIMA(2, 1, 0)x(0, 1, 1, 52)**		**RMSE**	**MAPE**
	In-Sample Forecast	T+1	1046.29	0.0451	In-Sample Forecast	T+1	1589.9	0.1218
	Out-of-Sample Forecast	T+1	831.93	0.0335	Out-of-Sample Forecast	T+1	1067.62	0.1216
		T+2	892.30	0.0350		T+2	1634.71	0.1760
		T+3	898.41	0.0355		T+3	1870.05	0.1783
		T+4	877.25	0.0326		T+4	1936.145	0.1741
		T+5	863.11	0.0317		T+5	2057.68	0.1834
		T+6	1177.92	0.0467		T+6	2133.52	0.1817

4.2.3. Sales Revenue forecasting for Point-of-Sale:

It is also noticed that it is important for the financial and budgeting planning to forecast the monthly revenue of each Point-of-sale. Using the insight from clustering analysis, that most of the POS from the same clusters have quite similar sales pattern which each other, it is decided that for each PoS cluster, one ARIMA model with optimized parameters can be used for monthly revenue forecasting.

Cluster 1

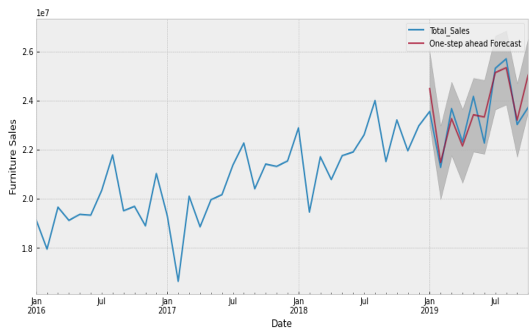

Figure 16 - PoS 1 Total sales timeseries

Cluster 2

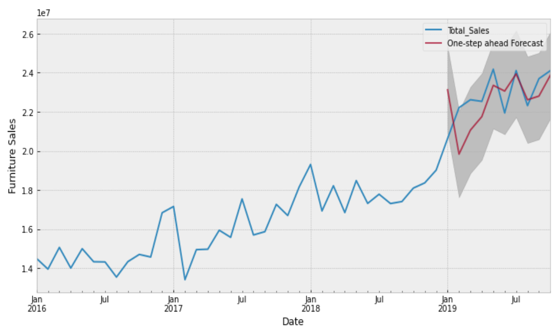

Figure 17 - PoS 19 Total sales timeseries

SARIMA(0, 1,3)x(0, 1, 1, 12)		RMSE	MAPE
In-Sample Forecast	T+1	686916.94	0.0237
Out-of-Sample Forecast	T+1	782189.66	0.0293
	T+2	696350.81	0.0235
	T+3	808669.95	0.0291

SARIMA(2, 1, 0)x(1, 1, 1, 12)		RMSE	MAPE
In-Sample Forecast	T+1	1334018	0.0484
Out-of-Sample Forecast	T+1	2468038	0.0858
	T+2	1950022	0.0665
	T+3	1738001	0.0525

The above results only showed for one representative Point of Sale of each cluster. It is required more validation and optimization for the models of each cluster to finally be good enough to apply for the activities of the business.

5. DEPLOYMENT AND MAINTENANCE PLANS

The deployment phase of every company is crucial for the business context. It provides a better idea about the effective usage of the data mining and exploratory data analysis models created, building a bridge between the business goals and the machine learning goals, applying the algorithm in a real-world context.

In order to establish this bridge and be visually attractive and user friendly, dealing with this set of insights can be done in a more accessible way by creating an interactive web dashboard. In this platform, several possible conclusions and insights are present to be easily figured out from the analyzed data. On the one hand, all the exploratory analysis is related to products, categories, product families, quantities and time dimension. On the other hand, it is equally important that this dashboard has information about the forecasts made for sales in the coming weeks and months, according to the analysis made by the team.

By entering this platform, the user who wants to analyze the indicators and forecasts will be able to select specific points of sale, year, quarter, and, in another dimension, various information related to the products, such as the most sold products, the products that generate the most revenue, the market share of product families and product categories, as well as a simple graphical analysis for the month of sales of some products or a specific product. All these indicators can be sliced and selected so that some of the remaining graphical representations can be related only to that(these) product(s).

Moreover, it is equally relevant that the dashboard also has information about the forecasts in terms of sales for the coming weeks and months regarding the number of units sold by products and by product and point of sale. This information is extremely useful in terms of logistics and organization of any company at any point of sale. Its availability grants the right to predict with a high degree of certainty the number of units sold, without relevant losses due to excess or defect, so that there is no shortage of units in stock at the points of sale and that there are not too many, which would be increasing the transportation and exhibition costs, making the commercialization process inefficient.

6. CONCLUSIONS

In order to support the company, our team analyzed a dataset containing aggregated transactions by Point of Sales, Date and Product SKU. Additionally, the most useful variables used for this project were the quantity purchased and the monetary value for each record. The business needs were multiple, and we were able to analyze the data from different perspectives, obtaining results that were successful despite the challenging size of the dataset and the few numerical variables which were provided.

During the project, the objectives that were achieved were a quarter analysis of the products, in which a seasonal pattern was detected for some of the products and other insights explained previously. Furthermore, we were challenged with the achievement of cluster solutions for the Point of Sales, that we analyzed based on two different perspectives, by Value and by Product Preference. Finally, our team carried out a forecast for the quantity of the next six weeks based on the seasonality detected and some of the models consider the forecast based on the clusters of the Point of Sales, providing a more specific prediction for each store of the company.

6.1. Considerations for project improvement

Even though the dataset provided was already preprocessed in some way, we suggest to the IT department of the company to improve the storing of the data. Indeed, the company is storing many transactions, and it would be better if some of the preprocessing that we have done were avoided simply improving the storing phase of the transactional system.

7. REFERENCES

1. Johann Robette, *The last will be first and the first last…*, https://towardsdatascience.com/tagged/forecasting
2. Tony You, *Understang SARIMA (More time series modelling)*, https://towardsdatascience.com/understanding-sarima-955fe217bc77
3. Jason Brownlee, A gentle introduction to SARIMA, https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/
4. Makridakis, S., Spiliotis, E. and Assimakopoulos, V., 2018. Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), p.e0194889.