BUSINESS CASES WITH DATA SCIENCE

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

Business Case 5:
Mind Over Data Retail Challenge

Group AA

Ahmadov, Emil (m20201004)

Macean, Doris (m20200609)

Shin, Doyun (m20200565)

Tagiltseva, Anastasiia (m20200041)

June, 2021

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

INDEX

1. INTRODUCTION 1
2. METHODOLOGY 1
3. BUSINESS UNDERSTANDING 1
   3.1. Background 1
   3.2. Business Objectives 1
   3.3. Business Success criteria 1
   3.4. Situation assessment 1
   3.5. Determine Data Mining goals 2
4. PREDICTIVE ANALYTICS PROCESS 2
   4.1. Data understanding 2
   4.2. Data preparation 3
   4.3. Exploratory Data Analysis – Quarterly Analysis per PoS 3
   4.4. Product co-occurrences 4
   4.5. Clustering 4
   4.6. Forecast 5
5. RESULTS EVALUATION 6
6. DEPLOYMENT AND MAINTENANCE PLANS 7
7. CONCLUSIONS 7
   7.1. Conclusion of Findings 7
   7.2. Limitations 7
   7.3. Future Research 8
8. REFERENCES 8

1. INTRODUCTION

Our team was tasked with understanding the purchasing patterns and behaviors of the client's 410 stores across Australia. Sales patterns vary drastically and being able to forecast sales is essential for supply chain planning and ensuring low out of stocks, as well as low idle inventory. Our analysis includes answers for questions related to product popularity, complementary products, grouping stores with similar sales behavior and forecasting sales for the subsequent 6 weeks.

2. METHODOLOGY

Cross-Industry Standard Process for Data Mining (CRISP-DM) [1] was the applied methodology in the execution of this research. CRISP-DM [2] provides a six-step process: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The following sections offer a succinct description of the execution of each of these steps.

Our team began with an exploratory data analysis of the dataset. Once our data exploration was complete, we proceeded with Quarterly analysis of top products sold, market share and product co-occurrences. The Point-of-Sales were clustered to obtain similar behaviors across stores. We then proceeded to develop forecasts of sales for the following six weeks.

3. BUSINESS UNDERSTANDING

3.1. BACKGROUND

Our client is an Australia-based retailer for a wide range of products. Within the given data, there were 21 Families of Products in 178 Categories of Products from 1 535 Brands Products and 8 660 SKUs located in 410 different stores.

3.2. BUSINESS OBJECTIVES

Our client wishes to understand each Point-of-Sale characteristics and achieve the following goals:

- quarterly analysis of top products sold, market share by product family and category,
- understanding of product co-occurrences,
- insight into the preference of customers at each point of sale to capitalize on improved customer relations by points of sales clustering,
- forecasting sales for the next 6 weeks by product and point of sale

3.3. BUSINESS SUCCESS CRITERIA

The solution for the above business objectives would be considered successful if the difference between products supplied to a point of sale and the following product sales decreases after the deployment. This will lead to increased revenue due to potentially more sales, as well as limiting costs and write-offs associated with mismatch between product supply and demand. Finally, the appropriate supply chain management also leads to customer satisfaction and brand loyalty.

3.4. SITUATION ASSESSMENT

For this project, we utilized a mid-range PC of our own and 182 342 304 entries of sales history data from our client (daily sales from 2016-01-01 to 2019-11-01 for each of 410 points of sales by 8660

sku products with respect to quantity and sales amount). Addex Consulting consists of four data scientists and python as the primary coding language. Various python modules and Power BI were used to analyze the data and develop models. The specific tools are addressed in our enclosed notebook.

We also had an access to one virtual machine C2 Compute-Optimized instance on Google Cloud Computing. At the time of initializing the virtual machine, it was utilizing Intel Xeon Scalable Processor (Cascade Lake), an octacore processor with base frequency of 3.1Ghz, up to 3.8 Ghz (all-core max) or 3.9 Ghz (single-core max). We attempted to maximize the utilization of its parrarrel processing capabilities (especially since the CPU is hyper-threading capable) through dask library, but due to a number problems throughout the project, we opted out of it. The compute engine also had 32Gb RAM, although the exact frequency is unknown.

3.5. Determine Data Mining goals

The business objectives can be translated to the following data mining goals:

- Proper data cleaning and data engineering
- Developing dashboard: Power BI it's a data visualization tool that helps represent huge chunks of data in a simpler form that is easy to understand.
- Clustering: In this, Data Mining organizes data into meaningful subgroups (clusters) such that points within the group are similar to each other and as different as possible from the points in the other groups.
- Association rules: The main goal of association is to establish the relationship between items which exist in the market. The typical examples of association modeling are Market basket Analysis and cross selling programs [3]. The tools used for association rule mining are Apriori Algorithm
- Predictive model to forecast demand: there are many types of technical methods of data mining, which mainly include: classification algorithm, clustering algorithm and time-series mining algorithm [4]

4. PREDICTIVE ANALYTICS PROCESS

4.1. Data understanding

The sheer size of the data required us to apply several transformations to minimize the computational burden, especially the memory usage. Unlike the common data mining process, where we first understand the data as a whole then move on to data preparation, the size of the data at hand required us to merge these two steps together. In other words, the steps were as follows:

1. Identify a specific pattern or anomaly in the whole data.
2. Propose potential options to clean the data.
3. Conducted analyses to check the applicability of these options.
4. Transform the data, repeat from step 1.

This approach helped us to reduce the computational burden for each subsequent cycle, as each cycle reduced the memory usage and usually decreased the amount of the data points. The details are presented in the following section.

4.2. DATA PREPARATION

The application of this section can be found in the attached notebooks: "BC5 1_Data Understanding and Cleaning.ipynb" and "BC5_2 Product Group Level Relationships.ipynb," where further details can be found. As previously mentioned, we paid special attention to limiting the memory usage. Our first step was to hardcode the datatype when loading the original data. By checking the min-max values and the presence of decimal numbers, we were able to convert each variable to data types with minimum memory usage to correctly contain the data. Note, although the product group details are categorical, we stored them as integers (instead of object/categorical) to minimize the memory requirement.

We checked for the nature of the relationships among Product Name, Category, Family, and Brand. If they are 1-to-1, we can easily create a dictionary of these variables connected to the ProductName_ID, drop them from the data, compute analysis, then recall the group level details (i.e., category, family, etc.) after the computation. However, as found in "BC5_2 Product Group Level Relationships.ipynb," this was found to be untrue, deterring our approach.

Our initial analysis indicated the presence of duplicate data. Upon further analysis, these were found to be incomplete aggregations. Before aggregating again, we conducted a coherence check on the 'Value' variable. Whereas there were no anomalies in sell-out units, we noticed sell-out prices that are equal to zero or negative. Even from accounting perspective it was hard to argue the negative price entry was made for sales correction. Zero-price entries were more justifiable (i.e., promotions, sales quantity adjustment etc). As such, we dropped the negative price entry and its paired quantity entry, while keeping zero price entries within our data. The anomalies found from the coherence check appear to be rather insignificant, and the 'Value' variable's data integrity appeared to be largely intact.

After the above pre-processing, we checked whether the number of unit rows and price rows were equal. Then, only the sell-out unit (quantity) data was extracted, as our client requested the analysis only on the quantity element. Pickle was used to reduce the file size. Note, the price analysis was still conducted, largely following the same approach taken for quantity analysis in separate notebook. After the separation, the final data type conversion was done. The final data type and memory usage is summarized in Table 1.

```
RangeIndex: 77529406 entries, 0 to 77529405
Data columns (total 7 columns):
 #   Column              Dtype
---  ------              -----
 0   ProductFamily_ID    uint8
 1   ProductCategory_ID  uint8
 2   ProductBrand_ID     uint16
 3   ProductName_ID      uint16
 4   Date                datetime64[ns]
 5   Point-of-Sale_ID    uint16
 6   Quantity            uint16
dtypes: datetime64[ns](1), uint16(4), uint8(2)
memory usage: 1.3 GB
```

Table 1 Final Data Types and Memory Usage summary

4.3. EXPLORATORY DATA ANALYSIS – QUARTERLY ANALYSIS PER POS

Our client requested to share the exploration/understanding of the data in a form of quarterly analysis, per point-of-sale. All analysis was done exclusively on Python and results are saved as csv files. In addition, we built a simple dashboard to visualize the outputs in more interactive and easy-to-understand manner. At the current state, the dashboard only contains information specifically requested by our client. Additional analyses (for example, top product families for each PoS per year per quarter), were excluded from our dashboard, but provided in csv form for our client to access. Of course, such an information can be added to the dashboard during the deployment stage depending on the company's needs.

To highlight a few of the findings, most of the stores appear to be the busiest during Q1 and Q4. This can be attributed to the drastic sales increase towards the end of December, followed by a rapid decline soon after.

For Top Products analysis, we ranked the products by quantity for each PoS. This ranking drastically differs from one PoS to another. We also looked at top families and categories, however, as expected, the due to their relatively high aggregation level, the rankings were not as different among different PoS. For the most popular families and categories identified, we produced a breakdown by brand. To be specific, if Family 1 was one of the top 5 families in PoS1 in 2019 Q2, the breakdown of Family 1 by brand for that quarter for the PoS was provided. Meaning, the brand breakdown and top family/category results should be studied together if our client wishes to make decision based on our findings.

Market share analysis for family and category were also done. Our outputs include all family and all category market shares, then we checked for top 5 for each PoS per quarter. FamilyID 9 and 12 were consistently the most popular products across all PoS. These families showed a steady increase in quarter-to-quarter sales. The sales was frequently highest on Q3, however, the fluctuation throughout the year was minimal. In terms of the category, Category ID 178 was by far the most dominant factor, occupying more than 70% of the total sales units in most of the shops.

4.4. PRODUCT CO-OCCURRENCES

Our initial plan was to provide an elaborate market-basket analysis via apriori algorithm. We hoped to loop through a list of dataframes divided by each PoS. As such an approach would require a minimum of 155Gb memory, even with forced memory overcommitment it was not possible to run the function. Instead, we created a simpler function with other features stripped down, which takes the PoS ID, year, and quarter as the input. Ideally, this would be deployed as an application, so the end-users can intuitively utilize the function.

Market basket analysis utilizes order/invoice level data, checking items within different orders, to which we do not have access. The excessive amount of data required us to decide between

1. Using daily aggregation as a proxy with high minimum support value for apriori or,
2. Using weekly aggregation as a proxy with significantly lower minimum support value.

We believed weekly aggregation would cause very serious information loss, which, in turn, the result would be highly unreliable. In contrast, option 1 would produce more reliable, yet limited insight, hence Option 1 was adopted. There was no way of validating this, as the products were only represented in ID, hence the nature of items and the associations among them could not be analyzed.

4.5. CLUSTERING

One of the approaches to clustering involves evaluating sales performance by category. This approach leads to a more precise understanding of variations in the level of demand across stores

This way, assortment breadth, and depth can be treated differently for each category. In this case, a store does not necessarily end up in the ‘top’ segment for all different product categories, because the store’s sales performance for each category is evaluated separately [6].

To achieve this goal, we used unit sales per product to do point of sale clustering. We have noticed that Point of Sale with ID = 96 has no sales record after the first quarter of 2019 and we decided to exclude it from the analysis.

We first needed to define the optimal number of clusters for each section. We plotted the elbow plots seen in the figure below (inertia as metric was used in this study.). As the elbow is at around 2-4 cluster for the data, we decided to apply the k-means algorithm with 3 clusters.

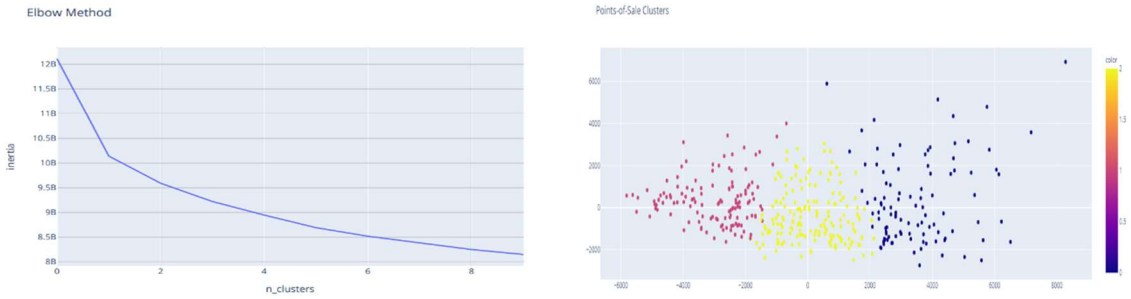

Figure 1: Clustering

In the table below you can find the top-5 Product Name for each cluster, cluster size, and the average sales of each cluster.

Clusters	size	Avg sales	1	2	3	4	5
1	174	460 688	356	993	2609	1277	481
2	107	617 506	356	253	993	1369	2609
3	128	403 497	356	993	2609	1277	481

Table 2: Cluster's overview

4.6. FORECAST

In order for the client to stock their stores appropriately, they have asked for a forecast of product sales for the following 6 weeks. As we are interested in future sales, discontinued items were excluded from the analysis. Based on the advice of the client, items that were not sold in the last 6 weeks were considered as discontinued.

After excluding these products, we still had a lot of products at hand which we needed to analyze. The models require a lot of computing power and lots of additional data to analyze each product separately. Regarding constraints in time and computing power, our team decided to come up with an automated approach. This automated approach takes data for each product as input and firstly separates it into two parts: training data and test data. This division is made according to the 80/20 rule. After dividing the data into two parts, different forecasting methods are fit to the training data. Some of these forecasting methods are ARIMA, SARIMA, Exponential Smoothing and Holt's Method. These forecasting methods try to capture characteristics of the given time series. After fitting the models to the training data, forecasts are done for the test period for each model. The program then compares these forecasted values to the real values and calculates R-square metrics for each model. As a result of this procedure, our code automatically finds the best performing model and uses this model to forecast the following periods.

We first used this approach to forecast for the aggregated number of products sold across all point of sales, and then for forecasting per product for each point of sale.

5. RESULTS EVALUATION

Before deploying the models, it is important to validate and review the way the model was built in order to assure that a faulty model that does not meet the business criteria is not deployed.

To find the overall performance of our approach, it was not feasible to forecast for all the products and find final R-square value because of computing power and time constraints. Instead, our team developed a forecast for the top 5 products and the average R-square score is found to be around 0.40. When we look to a specific case, we can see that product number 1277, which is the most sold product, has R-square value of 0.73.

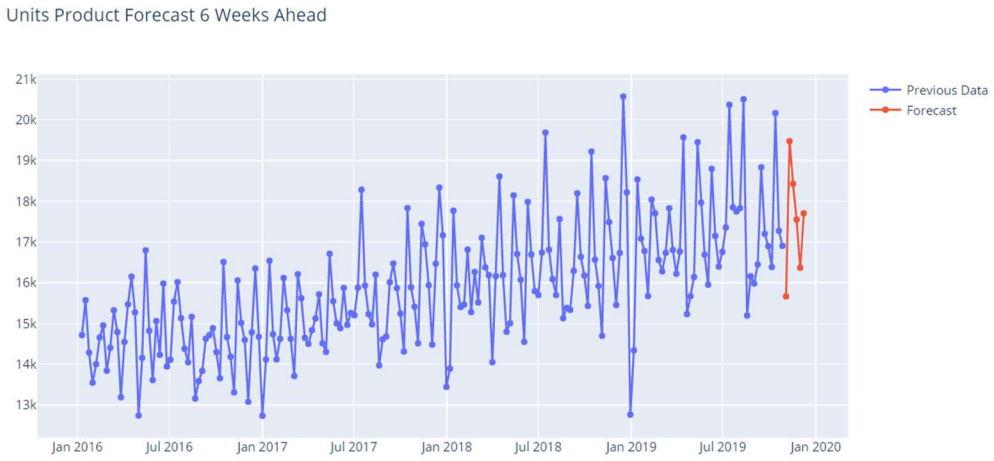

Figure 2: Units Product forecast

For the forecasts of each product at a specific point of sale, average R-square value for the top 5 products at location 292 was found to be 0.50. To look at the specific case of product number 481, which is also one of the most sold products, the R-square score is found to be around 0.80.

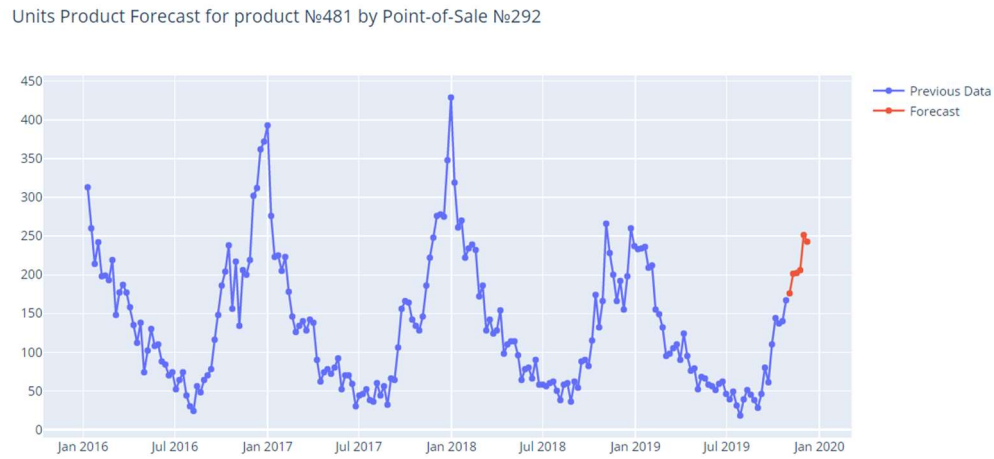

Figure 3: Units Product forecast for Product by PoS

6. DEPLOYMENT AND MAINTENANCE PLANS

We recommend a multi-stage deployment plan to ensure the effective implementation of our analysis and forecasting model. The plan begins with a pilot program designed to assess the accuracy of the projected sales. Our team has developed the design of an application that will output point of sale specific information. In the first stage of the program, we will implement the proposed design into a prototype. This application will provide internal users who are responsible for the allocation and replenishment of products with sales forecasts and inventory recommendations based on past sales behavior. At this point, the application can be adapted according to the feedback obtained from users.

During full deployment, this forecasting system will need to be updated to match the changing customer behavior and market dynamics. This system will need to be continuously evaluated every 6 months to provide better performance. As old products are discontinued and new products emerge on the market, new associations will arise. For implementing these changes, we will need to work with the user interface engineers from the company. We will work together with them to connect back-end of the association rules and forecasting system to the front-end.

7. CONCLUSIONS

7.1. CONCLUSION OF FINDINGS

GroupAA studied retail transaction data aggregated at daily level. An effort was put into making the data more manageable and providing our client with information that can be used to make business decisions. Quarterly analysis for each PoS was done on Product family, category, brand and name level to find the top sellers and the respective market shares. Product co-occurrence was also studied, albeit its findings are rather limited. Instead, we provide our client with functions that, hopefully, can be utilized with their much more powerful machine. Different PoS clusters were also found, and we have addressed the most popular products within each cluster. The final analysis was a 6 weeks demand forecast, with reasonable R^2 performance measures of 0.73 (across the company) and 0.8 (on a selected PoS). Of course, we believe more improvement can be done via further optimization and calibration process. We finished our project with a deployment plan proposal. However, as always, such plan would require a close communication with the client, and should be carefully implemented to avoid any unexpected interruption to the company's operation.

7.2. LIMITATIONS

The most significant limitation was the size of the data and lack of computational power to handle such data. The rigorous memory usage reduction process allowed us to load the data, however throughout the project we had to be very careful not to produce unnecessary objects to save memory usage. Nonetheless, the memory reduction does not reduce any data points, unless the method is via aggregation for a specific analysis. This, in turn, limited the number of experiments we could apply to different problems to find the best methods as each experiment would require significant amount of time commitment. Furthermore, some tasks could only be done on the virtual machine, the output of which was shared back to different team members to continue the analysis, incurring significant delays throughout the project.

For the same reason, many of the techniques and tools were rendered useless to handle such data. For smaller datasets, loop could have been used to iterate the analysis through different PoS, which,

in this case, was deterred by the limited computing power. Some libraries, such as pandas, has limitation on the number of data points (often related to the number of rows) it can handle, hence we were challenged to seek alternative methods to conduct certain tasks.

Lastly, the utility of our solutions also suffered. Product co-occurrence being an example, minimum support level of 0.7 is very high value which seriously limits the usage to only highly popular (sets of) items. Balancing between limitation and quality loss was, in fact, a common problem throughout the project. However, all these limitations made us to be very careful in handling of data; understand the in-depth workings of different tools at the source-code level and their limitations; and to commit a lot more time to design out study.

7.3. FUTURE RESEARCH

In the future, the prototype can be expanded for a more granular assessment. Considering individual orders at the customer level can allow for even more advanced recommendation and forecasting systems. Obtaining some additional meta data from the sales at each point-of-sale will allow for enhanced forecasting predictions.

Lastly, we have committed significant effort into creating a dataset that is much more manageable than it was originally provided, and develop a framework to work with such data. We hope our effort found a sound basis for the future research on this data.

8. REFERENCES

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Retrieved September 10, 2015, from https://themodeling-agency.com/crisp-dm.pdf

[2] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.

[3] Manpreet Kaura, Shivani Kang. Market Basket Analysis: Identify the changing trends of market data using association rule mining. Bhai Gurdas Institute of Engineering and Technology, Sangrur 148001, India. 2016
https://core.ac.uk/download/pdf/82094674.pdf

[4] V. Dhote, S. Mishra, J.P. Shukla, S.K. Pandey. Runoff prediction using Big Data analytics based on ARIMA model. November 2018. Indian Journal of Geo-Marine Sciences 47(11):2163-2170
http://nopr.niscair.res.in/bitstream/123456789/45311/1/IJMS%2047(11)%202163-2170.pdf

[5] Sohaib Zafar Ansari. Market basket analysis: trend analysis of association rules in different time periods. February 2019
https://run.unl.pt/bitstream/10362/80955/1/TEGI0458.pdf

[6] Approaches to Retail Store Clustering. Solvoyo Website. 2020
https://www.solvoyo.com/resources/approaches-to-retail-store-clustering/