BUSINESS CASES WITH DATA SCIENCE

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

Business Case #5 – Appliance Retail Analysis

Group D

Pedro Henrique Medeiros, number: 20200748

Débora Santos, number: 20200748

Rebeca Pinheiro, number: 20201096

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

May, 2021

INDEX

1. INTRODUCTION

Thank you for choosing **Data4Business Consulting (D4B)** to help you with the challenge of better understanding your customer's preferences. Our main objective is to help **Mind Over Data** improve its setup, increase your gains, bring new customers, and retain the current ones.

The world is experiencing a great technological and digital revolution where understanding business data, customers' segmentation, and their needs are essential for business success. The exponential technological advances, such as data mining techniques, artificial intelligence, internet of things, can help to take the business to the next level.

Through innovative technological programs, well-referenced data mining methods, and insights into digital marketing, the present report intends to provide an overview of the process behind the analysis, presents the results and insights you need to be successful.

In addition to the present report, the following deliverables will be submitted:

- Outcome's presentation to Mind Over Data;
- Jupyter Notebook with the code of the entire process.;

All files can be accessed in GitHub:

- *https://github.com/Debs86/Business_Cases_Projects/tree/main/BC5*

We are excited to take part in this challenge.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Mind Over Data is an appliance retailer with 410 points of sales most of them located in Australia. The company works with 1535 brands products, which sells in total 8660 different products spread in the points of sales. The products are also divided into 178 product categories and 21 product families.

Given the number of possible product choices available, it is important to have an understanding not only about the point of sales itself but also about the products marketed, understanding their profitability over the years. Another important approach is to comprehend customer's behavior in specific periods, so that way, the points of sales can be prepared for different scenarios as a boost sales scenario and contacting suppliers or decrease sales scenario and then making appropriate warehouse management. These analyses have become a very important part of the retail industries, by providing an assertive overview of products or services which hopefully increase sales and reduces damages and this is why they have reached D4B.

2.2. BUSINESS OBJECTIVES

The customer's primary objective is to build models to answer the following problems:

- Understand each Point-of-Sale characteristics (top products sold, market share preference);
- Points of sales clustering divided by value and product preference;
- Units' products forecast 6 weeks: one split by product and by point-of-sale and another one total by-product.

2.3. BUSINESS SUCCESS CRITERIA

The main expected outcome will be providing analysis about the business in general but in a different aspect. Regarding the Points of Sales, reports will be given to answers questions mentioned about their particularities, considering the products marketed in it.

Another expected outcome will be well-defined point-of-sale clusters which can make it possible to build a customized marketing strategy and maximize the return of investment, once providing this segmentation is possible to have a better understanding of customer behavior.

And at least, a 6-week forecast demand ahead making possible to the point-of-sales be well prepared considering the sales history provided.

2.4. SITUATION ASSESSMENT

2.4.1. Inventory of Resources

This project was made following the CRISP-DM reference model (Cross Industry Standard Process for Data Mining). CRISP-DM is a standard process built at the end of the ’90s and it was built by more than 200 members lead by a consortium of big companies.

This project has the support of Mind Over Data Management and staff.

On the D4B Consulting side, this project will be conducted by a team of 3 Data Scientists and Business Analysts.

We have been provided by the Mind Over Data team with a dataset with transactions occurring between January 1, 2016, until November 1, 2019.

The main technologies used to achieve the objectives of this report were Python, with many of its data analysis libraries, and Power BI. Python is one of the most important and commonly used programming languages for data science projects.

Power BI is a powerful business analytics service by Microsoft that provides interactive data analysis and business intelligence capabilities.

2.4.2. Requirements, Assumptions, and Constraints

The completion date of the present phase of the project is June 02, 2021, but we expect to continue giving support and helping Mind Over Data to achieve the next goals for the growth of the business.

Demand forecasting is a field of predictive analytics that is composed of techniques used for the estimation of probable demand for a service or product in the future to optimize supply decisions. It is an essential business process around which a company's operational and strategic plans are developed. Based on the demand forecast, a company can formulate strategic and long-range plans like capacity planning, budgeting, financial planning, sales, and marketing plan,d risk assessment.

2.4.3. Risks and Contingencies

Table 2.1 identifies the list of risks and contingencies proposed.

Risk	Contingency
The Dataset extensive size	Move data to a cloud-based platform
Transactions grouped by day	opening the data for a shorter period

Table 2.1 - Risks and contingency.

2.4.4. Terminology

Business glossary

Point of sales appliances retail usually have as a principal business glossary:

- SKU: Related to the unique stock code for each product marketed. Each product is assigned a different SKU for warehouse management according to the product attributes. The code is also crucial for sales tracking.
- Point-of-sale: Where transactions can be completed and the customer can pick up the product desired. To operate a point-of-sale, many activities must be done as administrative, management, marketing, maintenance, stock, and so on. Considering that each business is very singular, most point of sale has their system assuring that the transactions can be done and also connecting supplier-business-customer.

Data Mining glossary

- Clustering: It is a data mining technique. The technique consists of applying algorithms that classify the observations (customers) according to the similarity of their attributes.
- Normalization: The majority of clustering algorithms need the data to be scaled to a standard range. The process of applying specific transformations in the data to have it in the same range is called normalization.
- Forecast demand: Related for predicting the future demand in the retail products. It is anticipating the product's desire considering both controllable and uncontrollable effects. Improving forecast demand accuracy the business will be able to make assertive decisions regarding inventory, marketing, warehouse, and others.

2.5. DATA MINING GOALS

The data mining goals states project objectives in technical terms:

- Create a model that will be able to segment points of sales according to the customers' product preferences.
- Success criteria: The t-NSE (visualization technique to visualize the distribution of the clusters) must have a good distribution of the clusters; Also compare the R^2 metric of some cluster techniques applied and chose the one with the highest $R^{2.}$ High percentages of accuracy, precision, F1-score, and AUC.
- Create a model that will predict forecast demand.
- Success criteria: For forecasting a low value for the RMSE (Root Mean Square Error) and R^2

2.6. PROJECT PLAN

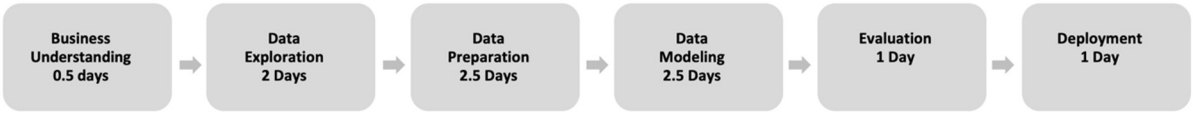

Figure 2.1 – Project Timeline.

Resources-wise, for the business understanding it was planned to use all the information provided in the kickoff meeting's presentation. For the core stages of the project, we plan to use Python to work the data provided. To present the results, we expect to use Word for the report and Powerpoint for the presentation.

The performance of the model will be directly connected with the quality of the input data. For this reason, we identify the Modelling stage as dependent on the Data preparation stage. During the project, we must go and back between Data preparation and Modelling many times, repeat this iteratively until we get the desired outcome.

For the Modelling stage, we aim to build an unsupervised model (clustering) using the K-means algorithm. Due to the timescales, we opted for using this algorithm as it is fast and efficient in terms of computational cost, simple to implement and the interpretation of clustering results is straightforward. The clustering quality evaluation will be made using R^2 and some visualizations to check the good distribution between clusters.

For the Forecast Analysis, we used the LGBMRegressor model from the LightGBM Python package, a gradient boosting framework that uses a tree-based learning algorithm. This model can handle the large size of the dataset, takes lower memory to run, and also focuses on the accuracy of results. The quality of the model's output was evaluated based on the R^2 and RMSE results. Furthermore, to measure the accuracy of the model, we used WAPE(Weighted Average Percentage Error)

3. PREDICTIVE ANALYSIS

3.1. DATA UNDERSTANDING

At this stage, we analyzed the dataset to understand its potential and limitations. This is an essential step to extract an initial understanding of what variables are in the dataset, what is their meaning to the problem at hand, how many they are, what are their distributions, and which of these features are relevant for the final goal.

The dataset is composed of 182.342.304 records dating from January 1, 2016, to November 1, 2019, with 21 different Families of Products, 178 Categories of Products, 1535 Brands Products, 2820 Product Names, 8660 SKUs, and 410 Points of Sales.

As can be seen in Table 3.1, the dataset contains 9 features (8 categorical and 1 numerical): Date of transaction, Point of Sale, 5 attributes that make up the product hierarchy, one attribute that identifies if that row is the quantity sold or its respective value and the last one with the values.

3.2. DATA PREPARATION

At this stage, the dataset should be prepared to ensure a good quality of the data for the algorithms to achieve good results.

We started preparing the data by filtering the column Measure, we created two new Data Frames. The first Data Frame was created with the records that have "Sell-out units" in the column Measure, and the second one with the records with "Sell-out values". Doing so, it was possible to extract the Quantity and Sales Values that were previously stacked in the same column. These two new columns were created, and the new Data Frames were combined. The columns ProductFamily_ID, ProductBrand_ID, ProductCategory_ID were then removed.

As a next step, the resulting DataFrame was filtered to contain only the records with positive Sales Values. A new column containing the average price was created from the division of the columns Sales Values by Quantity.

To deal with outliers, a new temporary DataFrame was created from grouping by the column ProductPackSKU_ID and calculating the mean from the recently created Avg_Price Column. This new

DataFrame contained the column Mean_Avg_price with the average prices for each distinct product, and the data from this column was merged to the Dataframe we were working on.

Following next, a new column named Perc_Variation was created from calculating the percentage in the price variations from the columns Avg price and Mean_Avg_Price. The transactions with sales values less than 10, and the price percentage variation less than -360% were removed. The records with Perc_Variation less than -500% were also removed, as well as the records with Perc_Variation above 50 and Avg_Price higher than 20.000. Additionally, as we considered the sales records in which the discount is over 50% and the quantity is higher than 150 aren't propper observations to cluster analysis and sales forecasting, these records were also excluded.

In Figure 3.1, It is possible to see the final distributions of the numeric features.

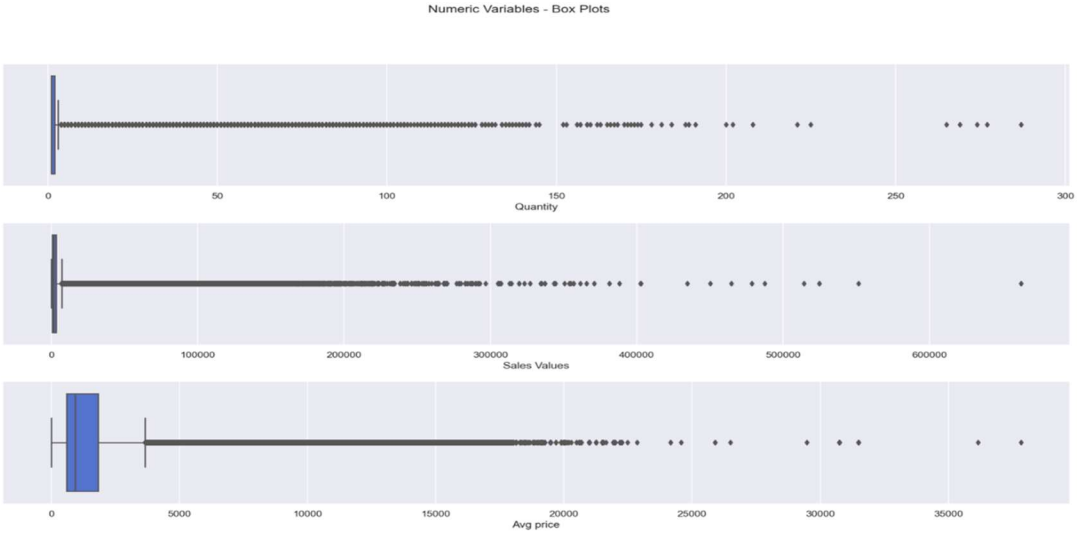

Figure 3.1 – Numeric Features Box Plots

Subsequently, we extracted the dates of the last sale of each product and store. This data was used to filter out the records in which the products haven't been sold for more the 180 days. We removed transactions regarding 633 products. Additionally, stores without sales in the last 45 days were also removed. Only Pos 96 was removed.

3.3. MODELING AND EVALUATION

3.3.1. Clustering

For the clustering, we have opted for using K-means clustering as it is the most reliable and efficient method. During the project, we performed several clustering tests and ended up choosing to cluster the data based on values, as it presented better results in terms of metrics and visualizations when compared to the clusters based on categories. The results presented here refer only to the clusters based on values.

During this process, we have created a set of functions that can be consulted in the notebook. From these we highlight the following outcomes from these functions on the cluster:

- Inertia plot, average silhouette plot, and Davies-Bouldin plot – showing the dispersion of the points within the cluster, how well each object lies within the cluster and a ratio between distances within the cluster and distances between clusters, for the different numbers of clusters to facilitate the decision on the optimum number of clusters.

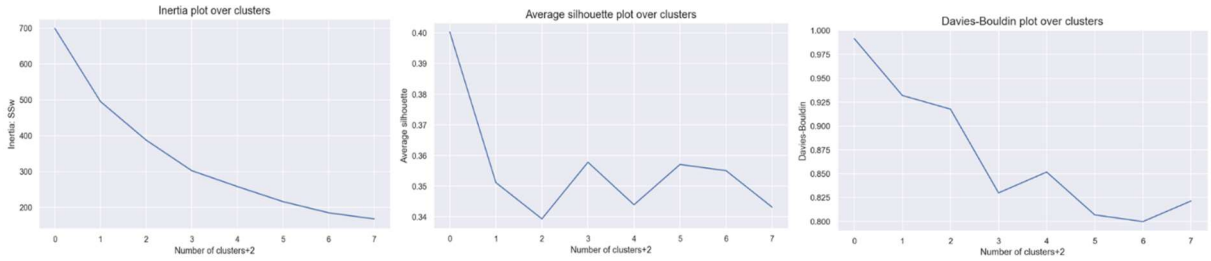

Figure 3.2 – Inertia, Average Silhouette, and Davies-Bouldin plots

- Silhouette plot – displays a coefficient for the clusters' quality depending on the number of clusters chosen

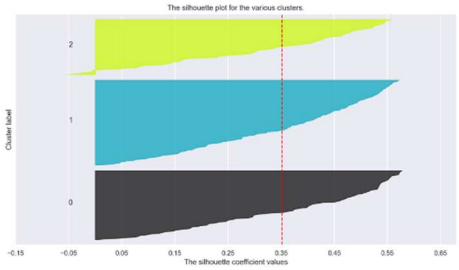

Figure 3.3 – Silhouette plots

From the analysis of the plots above, we have chosen to proceed with 3 clusters and produced their profiles, showing the cluster's means for each feature and the clusters' absolute frequency, which enabled us to confirm the clusters are well balanced.

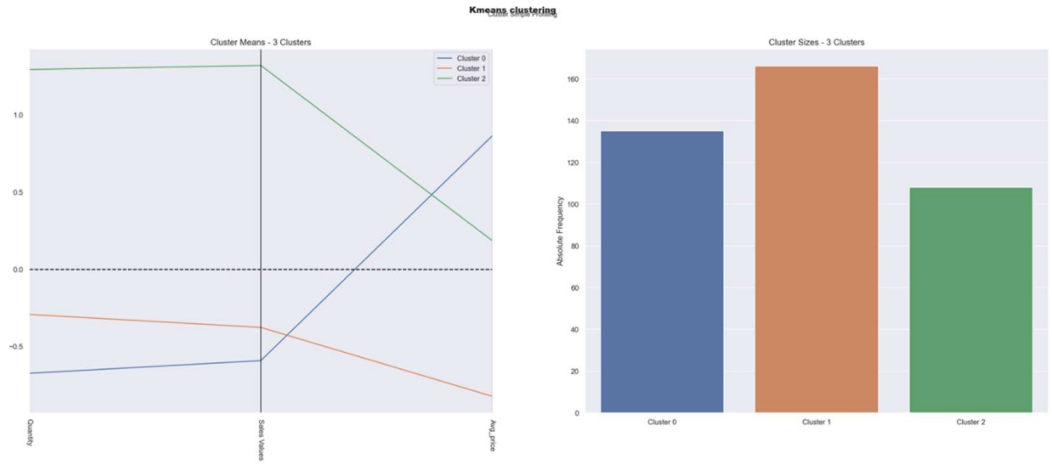

Figure 3.4 – Clusters profile

To evaluate the quality of our clusters, we have confirmed visually, in two dimensions, that the clusters are well defined (Figure 3.6, on the left) and the position of its centroids (Figure 3.6, on the right) through the T-Distributed Stochastic Neighbor Embedding (t-SNE) plots.

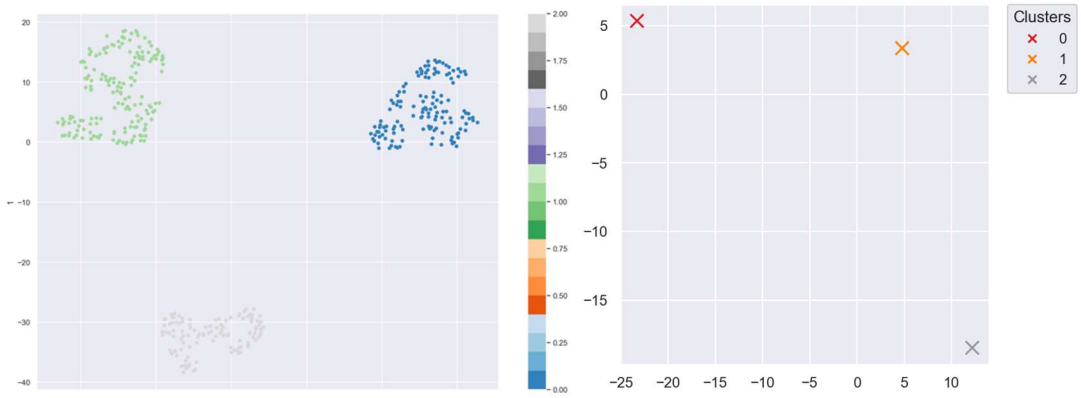

Figure 3.5 – t-SNE plots

3.3.2. Forecast

For the forecast demand, a feature engineering were done to create some variables to improve the performance of the model. First, we created a new column with each week of each year. After we group the quantity sold by store, product and week. The next step were create a new DataFrame of the Cartesian Product of the unique stores and unique products for each week. In the resulting Dataframe, a new column with the number of each product sold each week was added.

The DataFrame was then divided into training and testing sets based on the values of the column "Year_Week". The the testing set was with the records related to the last 6 weeks of 2019. After we split, we create again in the training set the data related to the last 6 columns but the quantities were replaced by 0.

The next steps were create some new columns:

- Mean quantity sold by product
- Mean quantity sold by store
- Mean quantity sold by store and product
- Lags 1, 2 and 3 : Quantity sold on the week before related to the same product and store
- Create a mean of the 3 lags
- Create 2 lags gradiente dividing lag by lag 2 and lag 2 by lag3.

After all this feature engineering the dataset has 21 columns, we split the data into train, validation and test. As stated before, the testing set were split being the last six weeks of 2019. The validation set was splited being the last 6 weeks that they are before the testing set. And the other weeks remaining in the training set.

Two models were applied: a Decision tree regressor and LightGBM.

The Decision tree parameters modified were max depth = 3 and min samples = 0.5. The other parameters were kept as default parameters.

In LGBM, the parameters are : 'num_leaves': 500, 'min_data_in_leaf':10, 'feature_fraction':0.7, 'learning_rate': 0.01, 'early_stopping_rounds':10, 'seed': 1.

Both models have almost the same performance in the training and the validation sets.

Lgbm achieves 0.845 for the adjusted R^2 score, and RMSE score of 1.88.

Decision tree achieves 0.8957 for the adjusted R^2 score, and RMSE score of 1.2680.

We applied another metric to measure the performance of the predictions. WAPE was calculate for

each store per product. The overall wape for lgbm is and for DT is

But both models when applied to the testing dataset, don't have a good performance.

The table below shows the the wape by train, test and validation by each store.

Store_ID	Test	Train	Validation	Total
292	0,39	0,20	0,13	**0,24**
72	0,39	0,19	0,13	**0,24**
73	0,36	0,16	0,12	**0,21**
333	0,33	0,18	0,10	**0,20**
42	0,32	0,17	0,11	**0,20**
282	0,37	0,14	0,11	**0,20**
78	0,33	0,16	0,11	**0,20**
92	0,35	0,15	0,11	**0,20**
280	0,36	0,14	0,10	**0,20**
330	0,31	0,17	0,10	**0,19**
383	0,35	0,13	0,11	**0,19**
323	0,31	0,17	0,10	**0,19**
389	0,34	0,13	0,11	**0,19**
93	0,30	0,17	0,10	**0,19**
71	0,31	0,16	0,10	**0,19**
329	0,33	0,13	0,10	**0,19**
356	0,30	0,15	0,11	**0,18**
391	0,34	0,12	0,09	**0,18**
22	0,30	0,16	0,08	**0,18**
48	0,30	0,15	0,10	**0,18**
382	0,29	0,15	0,10	**0,18**
103	0,33	0,12	0,09	**0,18**
Total	**0,24**	**0,10**	**0,07**	**0,13**

4. RESULTS EVALUATION

The clustering model developed by Data4Business Consulting reached good values on many important measures which will certainly help Mind Over Data to improve its setup, increase gains, better understand its business and make better decisions in the future. Please refer to the interactive dashboards developed in PowerBI by our team for further details on the values and insights extracted and provided in our analysis.

During the Sales Analysis, we extracted the following highlights:

1. Quarters 1 and 4 have the highest volumes in sales. Sunday is the day with the lowest sales volume. The year with the most sales was 2018.
2. The Product family with the highest percentage in sales is product family 2 (19% in total which correspond to 61B in sales)
3. POS_292 is the most representative in sales - 2B in sales (12% of the total)
4. ProductName_ID is 2802 has the most purchases - 19B on sales (18,96% in GT). ProductCategory_ID 178 is by far the most frequent - 239B in sales (82% in total)

As for the clustering, it was possible to identify three distinct groups:

1. Valuable: stores with higher sales volume mostly selling products with an average price between the ones from Bulk and Premium
2. Bulk: stores with average sales volume and predominantly selling products with lower prices
3. Premium: stores with the lowest sales volumes that sell the products at the highest prices

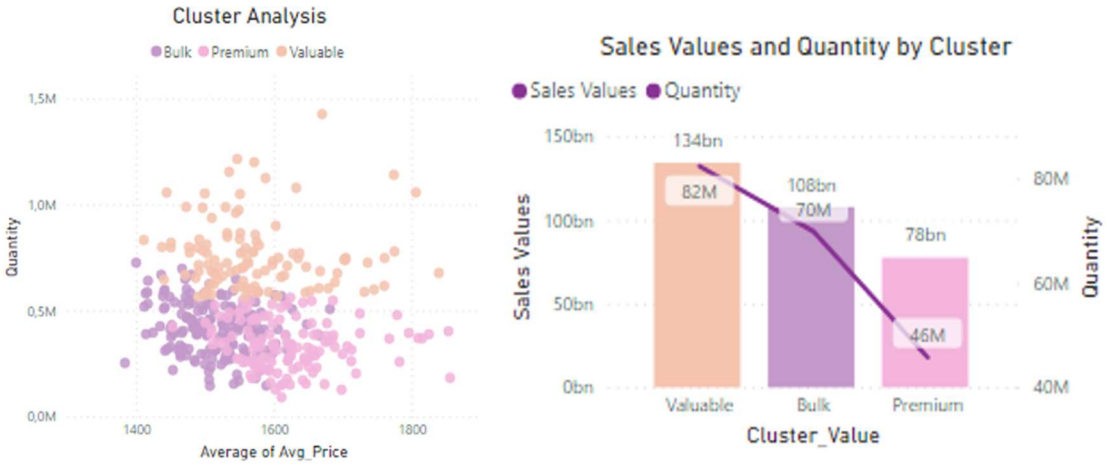

Figure 4.1 – Cluster Analysis

In the Demand Forecast analysis, our algorithm achieved good results in the training and validation sets, but it doesn't achieve good results when applied to test data.

5. DEPLOYMENT AND MAINTENANCE PLANS

We understand that the process of implementing the model analysis is very critical to your business. Appliance retails shops market is very dynamic and every day, there are new purchases on the system. So, we understand that it would be very important to implement the analysis provided integrated with the appliance retail store system. Integrating both, the results will be generated in a daily basis and the Mind Over Data could react faster and implement promotion actions to the customers according with their preferences, specially in the weekends which are the period with less sales according to the analysis provided.

Another important point is the data is that we have some problems of quality in the data. We advise to improve the quality of variable Sales Values, which there was found values below zero for a purchase.

We are going to monitor the performance of the model after it is implemented for some months and make some adjustments if necessary. We also suggest updating the model from time to time as the behavior of the customers, and even the market trends, can change, reducing the model performance.

6. CONCLUSION

As state in the section 2.2 – Business objectives, the main objective of this project was to provide Point of Sales analysis, considering top products sold, market share preference and product co-occurrences, providing point of sales clusters and also implement a forecast demand model based on units' products.

With the Point of Sales analysis it was possible to have a better understanding not only about customers preferences but also covering the what is the main products for each point of sales and which period of the year are considered sales critical in order to propose solutions.

The cluster aspects detected 3 well defined clusters (Valuable, Bulk and Premium). The main charachteristcs for grouping the stores were considering the sales volumes and selling products. It is important to add that clusters outputs meets the point of sales analysis, which confirm that the data processing was accurate.

Finally, for the forecast demand, the model developed by B4C reached good values on training and validation data but it need to be improved.

We hope Mind Over Data is satisfied with D4B work and we can continue working together

7. REFERENCES

[1] Chapman, P, Clinton, J, Kerber, R., Khabaza, T., Reinartz, T, Shearer, C. & Wirth, R. (2000). CRISPDM 1.0, CRISP-DM Consortium

[2] Riva, M 2021, Understanding Forecast Accuracy: MAPE, WAPE, WMAPE, viewed 28 May 2021, <https://www.baeldung.com/cs/mape-vs-wape-vs-wmape>

[3] Samir, S 2020, Machine Learning for Retail Demand Forecasting – Comparative Study of Demand Forecasting for Retail Store (XGBoost Model vs Rolling Mean), viewed 24 May 2021, <https://towardsdatascience.com/machine-learning-for-store-demand-forecasting-and-inventory-optimization-part-1-xgboost-vs-9952d8303b48>

[4] Schchur, Andri 2015, Demand Forecast with different data science approaches – Data science in demand forecast, viewed 25 May 2021, <https://towardsdatascience.com/demand-forecast-with-different-data-science-approaches-ba3703a0afb6 >

[5] Varesvik, V, Matijevic M, 2020, Demand Forecast using Machine Learning with Python, viewed 25 May 2021, <https://medium.com/vm-programming/demand-forecast-using-machine-learning-with-python-e8a4dca5aa0a>