BUSINESS CASES WITH DATA SCIENCE

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

Mind Over Data's Retail Challenge

Sales Forecast

Group T

Ana Amaro, number: m20200598

Filipe Lourenço, number: r20170799

Gonçalo Almeida, number: m20200594

Guilherme Neves, number: r20170749

June, 2021

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

INDEX

1.Business Understanding

1.1. Introduction and Presentation of the Business Objectives

Mind Over Data is a consulting company that outsourced to our team a case involving a physical retailer. Due to confidentiality, Mind Over Data was not able to provide us with many specificities regarding the end customer and the dataset we have available to work with. Being this said, we are dealing with an Australian retailer that sells appliances across 410 points-of-sale. Their main objectives are to gather insight about the stores, at a quarterly level (e.g., top products sold per store, market share per family/category and products' categories and families frequently bought "together"), to leverage segmentation techniques to extract even more knowledge than the quarterly analysis firstly uncovered and to forecast the next 6 weeks of sales, both at the aggregate level and by point-of-sale. Considering the requests presented, we believe that success will be achieved if we are able to: develop the tools that will allow the company to efficiently understand their stores; create an effective segmentation of the points-of-sale; and create an appropriate forecast for the next 6 weeks sales. Note that we were also asked to do a proper data enegineiring and analysis, which we don't consider as a goal in itself, but more as a first step towards achieving the remaining goals.

1.2. Situation Assessment

To complete this task, we were provided with a single CSV file, containing data regarding the purchases made, daily, at each point-of-sale between January of 2016 and November of 2019 (although we were told there was an "anonymous" component to the dates, meaning there was a temporal shift), split at the SKU level, and including an encoded version of the Product Name, Brand, Category and Family. In this regard, the main limitation we will face is the encoding of data that does not allow for a full business understanding of the problem. Moreover, to develop this project, we agreed with the deadline suggested, involving the preparation of a 5 minutes presentation and the delivery of a full report regarding the steps followed and the recommendations we want to provide. Additionally, the main costs agreed are the working hours our team will devote to the project and that will be billed to Mind Over Data in function of the benefits achieved. Finally, from the terminology point of view, the language used does not arise any questions, mainly as we have no direct contact with the company culture and environment as we are not the direct client of the Appliance Chain we are working to.

1.3. Data Mining Goals

From the data mining perspective, and regarding the business need to understand, at a quarterly level, the points-of-sale, our team believes that, given the lack of context and the high number of stores and products, the main goal is to be able to develop a dashboard that contains the information required in an easy to access and understand way so that members of the Appliance Chain can make more informed decisions regarding their stores. For this, our team believes that success will be achieved if a dashboard with the aforementioned characteristics is constructed. Note that given the limitations mentioned –and the fact we lack some important context, like the number of visitors per store, the location of the store, which business questions are the company particularly interested in, among others-, it would not make sense to try to derive more appealing business conclusions than just organize the data in such a way that someone in the company can understand it and use it to answer

business questions they may find relevant. Regarding the segmentation of stores, it implies the need to use a proper clustering technique, that will capture the nature of the (temporal) data we were provided with, creating relevant groups of stores. Overall, our team believes that if we can meaningfully segment the points-of-sale, we will have achieved success from this goal's data mining perspective. Lastly, as mentioned, in a less technical way, while presenting the business objectives, we will have to construct a model that can predict, not only for the overall company, but also for each point-of-sale, the amounts of products that will be purchased in the next 6 weeks. In this regard, we will try different approaches and will evaluate the models constructed mainly based on the metric Root Mean Squared Error (RMSE) -although the MAPE being widely used to access the forecast's quality, its limitation is well addressed in the literature, presenting significant bias, while the RMSE presents the precise behavior we are interested in, this is, to penalize disproportionally higher the large errors (when our decisions lead to tremendous amounts of spare inventory or unfulfilled demand) than the small ones-, being also this metric that we will use to define success. It shall be achieved if our team believes the forecasting error is low enough, in line with the business domain our team has gathered up to that point.

As a final note, and referring to all the data mining success criteria, due to their subjective nature, our team will conduct a meeting at the appropriate time to decide by majority vote if success was achieved or not.

1.4. Project Plan

Concisely, our project will be based on the following steps: 1. Retrieve the dataset and try to compact its size to a more manageable one; 2. Gather some superficial understanding of the data; 3. Deeper exploratory analysis, insight extraction about the data and plan how to act upon it; 4. Organize the data in such a way it can be used to construct a dashboard to create a quarterly analysis; 5. Define the models to use, create the conditions to apply it (e.g., split train and test set) and construct it, based on the data analysis done and on the business and data mining goals (in this case we will rely on both supervised and unsupervised techniques); 6. Attempt to improve the model(s) performance by tuning the parameters, when possible; 7. Verify if the data mining success criteria were achieved, and contextualize those in light of the business goals defined; 8. Present the report to the management team and if the approval is met, deploy the plan.

2. Data Analysis: Understanding and Preparation

The data collection was fairly straightforward, despite the file size: to process the large file, instead of using Pandas, we relied on Vaex, which allows for a significant increase in efficiency while working with large datasets. After the importation, we started to work on the dataset to make it smaller, for instance by removing strings and converting those to integers (only keeping the numeric part of the string). Moreover, we grouped our data at the Product Name level, instead of the SKU one. Employing a pivot table, we also joined the rows regarding the same transaction's day, which meant dealing with cases that for some reasons had for the same day more than one transaction (after a careful check, we realized those were not duplicated entries, but probably more like simple introduction errors mistakes, meaning the data was not inserted all at once, most likely due to some delay in getting the full day of operations of all stores) and with the fact that we had always one row for the value of the transaction and other for quantity, which we combined into two columns.

In this specific project, due to the size of the dataset and the multiple amounts of hierarchies to explore, we decided to rely more on statistical data than on visualizations, as that allowed for insight to be more quickly gathered. In this regard, we started to notice some problems with the data associated with the value of the transaction, particularly some had negative values and others had values that seemed suspiciously low. To deal with this, we decided it would be better to drop transactions with negative values, and although we were warned that the value/price of products varied highly across time and stores, some of the values were too low to be significant for our analysis, which implied further analysis from our side. The main reason was that even if the product price was legit and it was properly registered, that does not allow to infer characteristics about any specific point-of-sale or help to forecast demand, because, with a basic understanding of microeconomy, it is easy to understand that if prices are low enough (in relation to its value), there will always be customers buying the product, independently of the store that makes that promotion. Concerning this, we analysed every transaction that had a price per unit lower than 50 AUD (which seemed a low threshold for appliances and allowed us to not have to explore the full dataset which would be too computationally intensive) and decided that if the price was discounted more than 90% in comparison with its mean price, we would remove that transaction, by considering it a non-desirable outlier (either by a heavy discount that does not reflect the capability of a store or because of imputation errors). Another interesting point that we found was the reduction of sales in 2019 that seems clearly due to the Covid-19 situation (once more, we were informed there was a shift in dates to "anonymize" also this variable, but the patterns seem to suggest it was a backward shift, being that the actual dates already happened partially in a Covid-19 scenario), and in that regard, we will have to be extra careful when devising our validation environment for the forecasting model, to capture this dynamic both in the training phase, but also in the evaluation, as the next 6 weeks that we will want to forecast will still be on a Covid-19 scenario.

Being this done, and with no significant exploration still demanding data at the daily level, we decided to group our daily data at the quarterly and weekly level, which allowed us to have a significantly smaller dataset to work with. With the data at the weekly level, we started to try to uncover further problems with our data, mainly with potential discontinued products, brands, categories, families and even closed stores. In this field, we were able to detect a store (96) that did not sell any products after April of 2019, which was enough for us to consider as a closed store, and to remove it from analysis, as any further exploration of an already closed store did not seem worthy, especially when it seems that was a store that closed due to Covid-19 and did not reopen for business (which may be due to cases like being on a mall that entered in insolvency or another external factor not related directly to the performance of the store). Moreover, there were 13 categories, 312 brands and 558 products that were not sold in the last 9 months, and that therefore we considered as "discontinued". Note that we considered the need for 9 months to consider a product, brand and/or category discontinued, as some appliances may have a very seasonal component and by setting this threshold, we believe to have given enough room for this situation to have been accommodated. Furthermore, those were not removed, yet, from our analysis, as they may still be interesting to understand the performance of the points-of-sale at the quarterly level (e.g., a store did not stop being a top seller in the Q1 of 2016, because their best-seller product was discontinued in Q2 of 2016) and when performing the clusters, being only non-relevant for the forecast phase.

Considering the exploration done this far, we decided to start constructing the visualizations for the understanding of the points-of-sale at the quarterly level. In this regard, we only left in the

Jupyter Notebook a simple proof-of-concept, as the main output of this part of our analysis will be delivered in Power BI. There, the Appliance Chain can explore, per quarter, the most sold product, brand, category and family, not only per point-of-sale (as requested) but also in overall terms (tab Top Products). Moreover, and focusing on the dynamics of the points-of-sale, we have also created a ranking with the top stores by quarter in total quantity sold and monetary amount (tab Top Stores). Additionally, to transmit the idea of market share per point-of-sale, at the family and category level, we have constructed a tree-map, with the total quantity of products sold for the intended hierarchy (tab Market Share). With the combination of all these visualizations, we believe to have developed an easy-to-use tool for the Appliance Chain to understand their points-of-sale and the preferences referring to the most diverse hierarchies in analysis. Being this said, some insights caught our immediate attention, as for instance the store 292 being the top seller in every quarter, while the Family 9 and the Category 178 are also always top performers (even in most scenarios, when analysing by store, and not just aggregated terms), with more variability in preferences in the other hierarchical levels, Brand and Product Name.

3. Modelling

Regarding the modelling techniques needed, we will have three different problems: market basket analysis, clustering and forecast. Regarding the first two, we do not need to construct any specific design to validate our hypothesis, while for the latter we will perform a temporal train and test split, using the last 6 weeks for testing (to match the period we want to forecast), and using all the remaining for training (which also captures the stores' behaviour in a Covid-19 scenario).

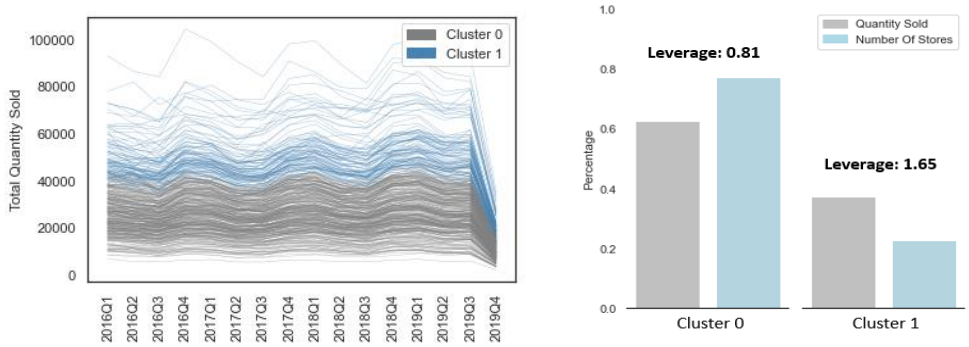
Starting with the Market Basket Analysis (which was only proposed as an extra challenge), as requested it was only done in 4 quarters (Q3 and Q4 of 2018 and Q1 and Q2 of 2019), and at the family and category level. In both situations we applied the Apriori algorithm, analysing the lift, confidence and support as seen fit, to understand the data. At the family level, the co-occurrences were very dense, with most families appearing in all quarters for all stores, which lead to results like no substitute families in any quarter (when considering the value of 0.975 as the lift threshold). On this same note, we were also not able to find complimentary products, having all families lifts below 1.05. In this regard, we found what was already expected, a significant number of rules with 1 of confidence (families always purchased together).

Moving to the product category, we had to set higher values for the rules' lift (1.5), support (0.15) and confidence (0.75) in the pursuit of complementary products, to narrow down the number of rules found. Even with these thresholds, we found for each quarter more than 1750 rules (being the one with the highest number of rules the Q2 of 2018 with 3175 rules). One interesting point is the appearance of category 88 (alone) as one of the most frequent categories on the antecedent side for all quarters except for the last one, where it stops being so frequently positioned as an antecedent. Moreover, at this level, we kept finding categories that were always purchased together, although some not being captured in the rules constructed due to the lift and/or support thresholds. For the substitute analysis, we also attempted more restrictive thresholds than at the family level (with lift and confidence at most being 0.9 and 0.2, respectively), which still led to a fairly high number of rules (93, 117, 53 and 67, for each quarter, correspondingly), with the particularity that from quarter to quarter we tend to have different categories appearing as the most "quoted" category in rules of substitution.

A more detailed exploration is available at our Power BI dashboard (as well as written in code in the Jupyter Notebook), where it is possible to explore the rules found (tab Market Basket Analysis).

Concerning the clustering problem, we were asked to segment both considering the value (in this case in total quantities, as Mind Over Data warned us that the value in monetary amounts should not be trusted) and the preferences of the points-of-sale. Our first decision was to respect the temporal nature of our data and approach this problem using an adaptation of the K-Means algorithm specially designed for time-series clustering, which uses as metric the Dynamic Time Wrapping (DTW). Additionally, we decided to use data at the quarterly level, as we believe this will allow us to filter better the noise from the signal for this particular application, and will also help to bridge the results with some of the outcomes of the quarterly analysis of the points-of-sale.

Starting with the value perspective, and applying the elbow method for the K-Means with DTW, we discovered that the ideal number of clusters seemed to be two. In this regard, we were able to construct a clear separation between the stores with a stronger "value" performance and the ones with a lower one, being the distinction clear, as visible below. The smaller group of stores (the one containing the "most valuable" stores) is composed of 93 stores, while the largest has 316 stores. Being this said, the typical store from the "most valuable stores" cluster sells on average twice the volume (close to 800000 units) than a typical store from the "lowest value stores" cluster (around 400000).



Figures 1, 2 – On the left, the time-series representing total value per quarter by store, allowing to see the clear distinction between clusters. On the right the leverage of each cluster.

Regarding the preferences perspective, we had to devise a more original approach. Firstly, we decided that our analysis should be done at the category level, being the main reasoning that looking at the product level would be impractical and lead to too much noise, while analysing at the family one would already imply a too high loss of information. Moreover, as we wanted to analyse preferences, and not just absolute values, we divided the quantity sold per point-of-sale per category per quarter by the total amount of quantity sold per quarter, which allowed us to get the percentage of quantity sold of each category per quarter per point-of-sale (in this way, we are truly capturing preferences as we are not punishing or rewarding stores based on aspects like the location, which can imply more/fewer visitants to the store and higher/lower quantities sold of every category, without implying preference). Then, having 178 categories for each store, we created 178 clustering problems (each category led to the need to cluster the time-series across point-of-sales). Being this said, the first step towards solving those problems was to plot 178 inertia plots and define to each category the ideal number of clusters to be formed. With this information, we ended up creating for each point-of-sale 178 different labels of belonging (the cases of stores that did not sell any unit of a specific category were joined together in an extra cluster for each category where this happened, to differentiate those from the remaining). As a final step in our clustering problem, we needed to find a way to compile all

those clustering labels into just one. For this, we recurred to K-Modes, and used the labels of belonging to different clusters as categorical variables, and that according to the inertia/cost plot obtained (using matching dissimilarity to assort it) led to the creation of 3 different clusters. Bellow, we have presented a sneak-peak in our results and the different preferences found, which reveal that even though the clustering was performed at a category level, there seems to exist clear distinctions between the top preferences for each cluster at the product level (which can also be observed at the most diverse levels of hierarchy. Still on this regard, cluster 0 is composed of 248 stores, while 1 represents 40 stores and the last one, 121 stores. Lastly, given the constrained space in this report, and how we believe that there is more business value in providing the right tools in an easy to interpret way for the people that know the business, instead of making simplistic analysis based on encoded data and with no specific business problem in mind, we will not elaborate much on the results achieved (having us already shown the good separability achieved between clusters in both perspectives and the most interesting findings, namely the difference in leverage between value clusters and the differentiated value brought by stores depending on the cluster they are inserted in, and a clear distinction in the top of products preferences, between the preferences clusters, being that the main intention, this is, to differentiate stores at the product level by preference), instead, we will provide in our Power BI, the tools for this cluster-effort exploration, allowing to understand better the typical behaviour of the clusters formed in both perspectives (tabs POS Clustering by Value and by Preference).

Cluster 0

Product Name	Average Units
1277	9012.04
2609	8810.11
481	8439.63
993	7992.53
1147	7867.49

Cluster 1

Product Name	Average Units
1147	9106.05
481	8525.98
993	7914.25
567	7519.25
2609	7177.6

Cluster 2

Product Name	Average Units
1277	5616.98
2609	4913.54
2802	4834.48
481	4514.67
847	4484.31

Figures 3 - Top 5 products most sold, for the “typical” store (on average) of each cluster.

Moving now to the supervised learning problem, the first step we took was to remove the “discontinued” products, categories and brands. Then, we decided to apply three different algorithms (at the weekly level), being two of them particularly developed to work with time series, Prophet and SARIMAX (which includes the classic ARIMA model), while the other although being developed to deal with general regressions, tend to be highly used with success for forecasts, XGBOOST.

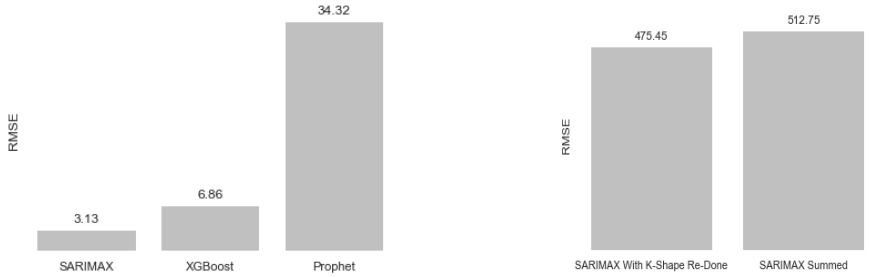
While the Prophet was fairly ready to apply, the XGBOOST and SARIMAX needed some data pre-processing. Regarding the XGBOOST, we had to convert the forecast problem into a regression one. For this, the first step was to construct a train set, that contained for each store and date all the combinations of products sold by that store in any particular period of the train set, filling with 0 the entries where a specific product was not sold, this is, there was the need to create a way for the model to learn the “concept” of 0, and how frequently it happens. Moreover, we needed to decompose the date into different variables: week of the year, month and year. The same process was followed for the test set, while in this case, we just created empty entries for the combinations of the dates to forecast and the products sold for each specific store in the training set. Concerning the SARIMAX model and the number of possible variations of this model, it would be impractical (considering the constraints we had mainly in terms of time and computational power) to assess for each time-series their ideal SARIMAX (doing it manually would imply significantly more hours than the ones billed at the beginning of the project while using automatic detection of the correct SARIMAX would make the predictions too slow to be deployable). Due to this, we decided to recur to a time-series clustering algorithm named K-Shape that segments time-series based on the shape of the same. Note that this

clustering model has a different intention than when we used the K-Means DTW, as here we are just interested in the shape of the time-series and not in the actual proximity of the time series values. This allowed us to group all the time-series into just 4 groups, which we now can explore to understand the best model parametrization to apply to combinations of stores and products belonging to each cluster, by analysing the centroid time-series shape. Although this method leads to the loss of some specificity, it is expected to capture the general "behaviour" of the time-series, being that if extreme variations exist, we expect the Prophet algorithm to "shine", as it is particularly good in those type of scenarios.

Overall, we created a model that will run efficiently, XGBOOST, one that is known to work well with seasonality and noise, Prophet, and a last one, SARIMAX, where we will attempt a more flexible approach to mix statistical rigour with the efficiency (introduced by the K-Shape algorithm) of not having to define the parametrization for each single time-series.

Note that for all algorithms, we only tried to forecast products already sold for each particular store, which we also took into consideration when evaluating the model, not considering new products introduced on stores or products never sold by a particular store (considering that each store has purchased on average around 1000 less different products than the total amount of different products available, if we had included all those combinations, the models would most likely correctly forecast those as 0, which would artificially largely inflate the results achieved), while avoiding data leakage, by not using just the combinations of stores and products that we knew would be on the true test set (using all combinations of products sold in any temporal period from the train set by each store). It is also worth mentioning that in cases that result in forecasts below 0, we will convert those to 0, and all values will be rounded to the closest unit.

For the Prophet Algorithm -and considering that parameter tunning would not be an option, as we have only run it once and it took around 110 hours to forecast all combinations of products and stores, with a fully dedicated CPU, working in multi-thread- we allowed the model to run with the default automatic attempt to find seasonality at the year and week level, while also asking the model to look for quarterly seasonality. Regarding the XGBOOST, we attempted a very small parameter tunning phase, being the best results achieved when setting the learning rate to 0.1 and the maximum tree depth to 15. Lastly, for the SARIMAX, our initial K-Shape inertia plot suggested the best number of clusters to be between 2 and 3, being that we selected 3 to accommodate a higher diversity. With those, and looking at the centroid time-series shape we found the best SARIMAX models and applied to each member of the cluster the corresponding model given by the centroid, taking only a few hours to achieve results, contrary to the days/weeks that would have been required if done manually (or with automatic detection) for each single time-series.



Figures 4, 5 – On the left, the RMSE scores achieved by the 3 different models attempted to forecast sales by combination of product and store. On the right, the RMSE scores achieved with the two attempts involving SARIMAX models to forecast sales by product (at the aggregate level).

Comparing the results achieved by the models, SARIMAX was the one achieving the best score, by a significant margin, as visible above. The results suggest the robustness of the combination of K-

Shape with the SARIMAX models, which significantly decreases the labour hours needed to manually identify thousands of time-series and/or the computation power required to detect automatically for each the best model. To better understand this model, we also computed the average forecast error, which gives us the notion if the model tends to forecast in excess or defect, being that in this case, the results was -0.04, suggesting on average a little under-forecast. Finally, to fulfil the desire of having a total quantity forecast per product, we decided, to be consistent, to only attempt either the SARIMAX model, repeating the same process as before (with the K-Shape also suggesting 3 clusters, but with different parametrization), or to just sum all the quantities predicted per product at the store level, achieved previously. Once more, we present the results above with the two processes attempted, being that developing the full process from scratch achieved the best results, this time with an average forecast error of -0.87, once more under-forecasting on average "what the real quantity is", but again not significantly.

At this point, the only modelling activity left to perform was to make an unsplit between train and test, and repeat the process followed previously using the full dataset to train the model allowing it to predict the quantities of products sold per store and in overall terms for the next 6 weeks. Here we had to repeat the clustering process, with now the best number of clusters being 4 (both for the forecast of products per store and for the total products' quantity forecast). Then, using the same logic as before, we found for each centroid shape the best SARIMAX's fit, and applied it accordingly to the members of each cluster. The results for this predictive output were also present in our Power BI tool, where the team from the Appliance Chain can quickly understand the products' expected demand for each store and for the global chain.

Overall, and considering the data mining goals defined, our team approved with unanimity that the success criteria were achieved. Although -due to the cardinality of stores and products in analysis and the lack of clear guidelines of aspects of interest- it was hard to provide specific feedback and interpretation for the quarterly analysis of the points-of-sale and for the clusters formed, all the information and insight is presented in the Power BI report (tabs Forecasting's), allowing the Appliance Chain to extract knowledge as needed according to the future business requirements that may appear. Furthermore, the forecasting tool created seems to have done a proper job, being able to achieve an RMSE of 3.13, at the point-of-sale level and 475.45, when at the total level.

4. Evaluation

Looking at the work developed, there seems that the business goals originally proposed were achieved, with a proper analysis conducted in Power BI (accessible and self-explanatory to our clients explore at will), of the points-of-sale at the quarterly level, which as challenged includes a simple 4-quarters Market Basket Analysis of the families and categories of products. Moreover, by using appropriate techniques, we were able to segment the Appliance Chain customers at two levels (value and preferences), which we believe offers a significant insight and will help in future analysis, that can now be performed at the cluster level when appropriated. In this regard, we have also shown (figures 1,2 and 3) that the clusters/segments formed seem well differentiated with distinct tastes and group-values being clearly identifiable. Finally, the model created to forecast sales, not only achieved good results, but was also constructed in such a way that requires minimum effort to be updated and used, while also not demanding significant computational power and time to run.

Discussing now future efforts, we believe that more data could and should be collected and provided to the team analysing the data (starting with the signature of a non-disclosure agreement that would allow knowing to which company we are working), for instance, the number of visits per store (which may help us to relativize the good and bad performance of some stores, that may just be due to their locations and not the quality of service itself).

In a nutshell, our project followed, in general terms, the plan defined in step 1.4., and we believe to have complied with the CRISP-DM methodology to the possible extent given the specificities of the project. In this sense, the different steps of the CRISP-DM needed to be presented in broader divisions, which gave some flexibility to deal with some problems in the applicability of the framework. Being this said, there were still some problems that we had to deal without respecting the CRISP-DM, as removing some entries (e.g., products discontinued) only in a specific phase of the modelling process, instead of being in the data selection phase. Regarding the next steps, and considering the limited resources available, we believe that delaying the deployment by attempting new models, new parameters or new approaches altogether, would lead to huge opportunity costs not only for our team but also for our client.

5.Deployment

5.1. Deployment Plan

Firstly, it is not possible to have a good machine learning system in place if the data is not properly stored and accessible, and, in this particular case, delivering data to a consulting firm in a large CSV file, seems a poor choice. In this regard, it would be important to clarify if the company does not have a proper storage system or if they just do not feel comfortable letting our team access their data warehouse. We believe the case to be the latter, and therefore we will not elaborate on how to build a proper data storage system, focusing more on the implementation of the machine learning system that we believe should be developed. The analysis conducted on Power BI, at the quarter level should be optimized to be constantly updated, quarter after quarter, which in our understanding will help the Appliance Chain to have an efficient summarization of their stores' quarter results, while being able to compare them with the past periods.

Regarding the forecasting efforts, we believe that the next steps should involve the decomposition of the phases performed up to the application of the model into different scripts, with a direct connection to the database, to achieve automatic updates of the forecasts estimated. This will be possible as our model is fairly quick to run, which allows for weekly updates of forecasts instead of relying on the ones made 6 weeks before. On this note, we believe that the forecasting tool should be complemented with access to the inventory system of each store, to automatically signal which products should be ordered and in which amounts, while considering the inventory at the store (Amount to Order = Amount Forecasted – Amount in Stock), making the process significantly more efficient. Note that this would imply an automatization of the registering process of discontinued products, as there would be no point in keeping forecasting those.

5.2. Plan Monitoring and Maintenance

As mentioned before, we expect our model to be able to perform forecasts on weekly basis. Considering this, if possible, the best-case scenario would involve having a weekly phase of clustering

using the K-Modes, being that this would require some supervision when selecting the number of clusters and then applying the corresponding models to each cluster. If this scenario is not possible to be conducted, we still believe the model can be used to achieve good results for a reasonable period, without needing to be updated. The need for this update could be accessed employing a trigger that should be activated once the RMSE score of the model was inferior to a certain threshold, being that the point where we would have to re-do the clustering and apply the proper SARIMAX to each cluster, which should take minimal effort and time.

On this same topic, and considering the clustering of the points-of-sale, we do not believe there will be significant changes at a very alarming rate, which would imply the need to re-cluster frequently, but this should be validated at a quarter level, being probably the input of the marketing team a good "human trigger" on when the clusters start to become too "mixed", implying a lost in segmentation power.

5.3. Conclusion and Brief Review of the Project

Although the review of the project should be done after the deployment and in close collaboration with all stakeholders involved, we believe to be already in a good position to present some conclusions regarding the project developed this far.

Despite the lack of labelling (all data anonymised), that would help to understand the context of the analysis we were asked to perform, our team believes to have been able to deliver a tool that will significantly improve the decision-making process of the Appliance Chain, by using a data-driven approach. In this regard, the dashboard created will help to understand the different points-of-sale, what do they bring to the table and how are their performance is varying over quarters.

Moreover, and as challenged, we created a Market Basket Analysis per quarter, that at the family level did not allow to find complementary or substitute products (which in itself is still a valid result, and that may ask for, in future efforts with more computational power, this analysis to be conducted at the weekly level, to have a less dense co-occurrence matrix), but that at the category level raised some interesting rules. Concerning this, we once more noticed that the lack of space in this report did not allow us to develop further our findings, but we strongly believe that producing a self-explanatory dashboard is significantly more important than any light-weighted approach we would perform here, as we have very little context on the problem.

Regarding the segmentation, we have used an interesting technique (K-Means DTW), that allowed for what looks like a powerful segmentation among groups, with some noticeable differences both at the value and preference level (that can be explored further with some goal in sight), while the forecasting tool created, although innovative (combining supervised and unsupervised learning on the same problem), provided the best results, being very efficient and easy to use and implement.

Lastly, we have provided a proper deployment plan, that although short (due to space limitations), will help define a proper strategy on how to leverage the insight gathered in this project and how to make the benefits not limited to this endeavour, but compounding over time.