BUSINESS CASES WITH DATA SCIENCE

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

Wonderful Wines of the World

Customers' Segmentation

Group T

Ana Amaro, number: m20200598

Filipe Lourenço, number: r20170799

Gonçalo Almeida, number: m20200594

Guilherme Neves, number: r20170749

March, 2021

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

INDEX

1.Business Understanding

1.1. Business Objectives

Wonderful Wines of the World (W.W.W.) is a 7-year-old wine company that has the ambition to sell quality wines (and from time to time also some wine accessories). Their selling channels are composed of 10 small stores across the U.S., phone calls based on contacts with catalogues (updated frequently) and their website.

While up to this point their marketing efforts have been based on aggressive, mass-marketing and intuition-driven campaigns, W.W.W. has been collecting data about their customers for 4 years and now wants to explore it. With this goal in mind, W.W.W. contacted us asking for guidance in understanding what are the most important factors that differentiate their customers, how many different segments of customers they serve and how could they take advantage of this knowledge to drive a segmented based marketing strategy (to the existing and new customers). In a nutshell, it seems that W.W.W. is interested in identifying which groups of customers they serve and how to reach them based on their specificities.

Given the objective at hand, we will set two business success criteria, one more subjective – provide useful insight about the different types of customers – and the other more objective – improve the return of the marketing efforts.

1.2. Situation Assessment

W.W.W. provided us with an excel file containing 10 thousand customers, chosen randomly from the customers who have made any purchase in the last 18 months. Furthermore, this file contained 28 variables regarding the customers' wine taste preferences, some personal information, and the relation they have with the company. All these variables have associated metadata, that in general is clear regarding their meaning, with some exceptions (e.g., how the Lifetime Value of each customer was computed), that the management team was also not able to clarify. Despite this lack of knowledge about the data, the management team showed high availability to help in any business question we might face along the way.

We agreed with the timeframe purposed of one week to prepare a presentation of 5 minutes to the management team and to deliver a full report regarding the steps we followed and the recommendations we want to provide. Moreover, the main costs agreed are the working hours our team will devote to the project and that will be billed to W.W.W. in function of the benefits achieved that are closely related with the advantages mentioned in the business success criteria (knowledge about customers and a more efficient marketing strategy).

The main risks associated with this task are very generic and arise from a potential lack of representativeness or quality of the data we obtained, which can easily be mitigated by asking for more data if along the way we find reasons to suspect this happening.

Lastly, from a terminology point of view, we need to understand the concept of Marketing Mix, also known as the 4P's and that comprises: Product (satisfaction of customer needs); Price (based on customers' willingness to pay); Placement (the channel used to reach the customers); Promotion (deployment of strategies based on what was learned from the remaining P's that help us to target each customers' group).

1.3. Data Mining Goals

The data mining goal for this project is an effective categorization of customers in different groups according to their engagement with the company and their tastes. In order to measure success, we will focus on a subjective measure that will be validated by a majority vote of the team of consultants working on this project: Create clusters in a way that each one has enough customers to be worth to develop marketing campaigns for, while maintaining a clear separability between the characteristics of the "typical" customer from each cluster.

1.4. Project Plan

Concisely, our project will be based on the following steps: 1. Retrieve the dataset, gather some superficial understanding of the data and the variables and try to spot easy to detect problems, by using techniques like statistical analysis of the variables; 2. Deeper exploratory analysis and insight extraction about the data; 3. Clean the data by acting upon the insight found at step 2 (e.g., remove outliers, solve missing values and other problematic values); 4. A small review of the steps done until here, by confirming if all the problems encountered were solved; 5. Define the model to use and construct it, based on the data analysis done and on the business and data mining goals (for this step we already have as hypothesis that we will use a clustering algorithm build on two different perspectives - taste and engagement - and merged afterwards); 6. (Optional) In case the model constructed is based on different perspectives, we merge them (which may imply some restructuring), and, in case outliers have been removed, add them to the corresponding cluster; 7. Verify if the model meets the data mining success criteria, and interpret the results achieved in light of the business context; 8. Present the report to the management team and if the approval is met, deploy the plan.

2. Data Understanding

The data collection was straightforward, as W.W.W. provided us with a single excel file (that only needed some small adjustments), which also facilitated the process of data description, where all variable types were consistent with the metadata.

While looking more closely at the data, employing statistical summaries, boxplots, histograms, pair-plots, correlation matrixes (using the Phik correlation to allow for the use of both binary and metric features) and also taking advantage of the visualization techniques proposed by SOM, there were discovered some reasons for concern: 1. There seemed to exist some well-defined outliers, both in the univariate space and in the multivariate space; 2. The variable Recency had a highly skewed distribution with less than 5% customers engaging with the company more than 100 days ago (considering it has a maximum value of 549), which was clearly visible in the U-Matrix and component planes from SOM; 3. The variable Access also posed a problematic behaviour, presenting the value 0 for the vast majority of customers; 4. There were some prohibitively high values of correlation among some variables, while others did not show even a moderate correlation with the remaining.

Regarding the binary variables, we were also able to observe that a big proportion of customers had not bought accessories (as the metric variable 'Access' suggested), and no accessory was particularly more popular than the others, except for the Bucket that was, clearly, the least bought. Furthermore, only a very small proportion of customers has presented a complain, and there was a more or less balanced distribution of customers who have kids and/or teens at home.

From a quality perspective, the data presented neither null values nor duplicated customers. Nevertheless, while attempting to check for several inconsistencies (e.g., check if there was any customer that reported to have made their last purchase longer than for the amount of time that customer is with the company), we discovered several customers with less than 21 years old – the minimum age for drinking in the US – and in some cases, probably due to rounding, the sum of the different wine types purchased did not match precisely 100%.

3. Data Preparation

As mentioned above, some of the customers were younger than 21 years old, but after consulting W.W.W., we were informed that may be due to exportations to other countries so we will not act upon this. Nevertheless, there are some other significant aspects that we cannot ignore, and therefore we will need to select which data to keep in our project. Firstly, the variable Recency showed some problems mainly in its uneven distribution so we were considering that we might want to exclude the customers that have made a purchase longer ago than 100 days, but we also found out that this variable shows no particular correlation with any other, which suggests that people who purchased longer ago do not seem to have a different pattern, and therefore don't need to be excluded, while the variable itself should and was dropped. Based on the lack of correlation with other variables, we also decided to drop Complain and Dayswus (as none showed a higher than 0.35, in absolute value, correlation with any other variable). Additionally, and as discussed above, the variable Access also presented some problems and considering that this variable is completely conveyed in the sum of the binary variables Humid, Bucked, SMRack and LGRack, we decided to also drop it from the construction of the clusters and use the binary variables to, later on, interpret the clusters formed in replacement of this one. Lastly, we also decided we would need to filter the most correlated variables in our dataset, but while the decision to drop was taken in this step (complying thus with the CRISP-DM methodology), the actual process will be described in the construction of data step.

The main cleaning effort we undertook in the dataset was to try to find and exclude the outliers (to later add them back for clustering classification). Our team used a two-step confirmation for outlier detection (only considering a customer as an outlier if he was both an outlier in the univariate and multivariate space) based on the IQR method and with the Mahalanobis distance, that we had to set with some more flexible than usual parameters - a 2.5 multiplier and 3.5% of extremism, respectively - in order to not consider more than 6% of our dataset as outliers, as this amount already seemed to exclude the clearest outliers.

Regarding the data construction, our team considered it was a good opportunity to get rid of some of the prohibitive correlations found. Firstly, we noticed that the variables Income, Freq, Monetary, LTV and Age all shared a correlation above 0.83, so based on this and on the meaning the majority of these variables have in the business domain, we decided to create a new variable named Value Index that was formed from the first principal component of the previously mentioned variables. This technique allowed us to keep around 90% of the original variability summed in one variable that is highly correlated (> 0.9) with all variables used to construct it. On the same note, we used this same process to create a variable named Web, where we summed up in one principal component 94% of the variability of WebVisit and WebPurchase, which were also highly correlated, having this variable as meaning the Website "friendliness" of each customer. Note that in both cases the variables used needed to be standardized before proceeding to the PCA construction and that although the management team of W.W.W. may not understand directly how these variables were formatted (as they did not for the variable LTV), their meaning is very straightforward.

In this step, we also decided to normalize as a percentage, with the sum needing to be 100%, the sum of the different wine tastes purchased.

Lastly, and before proceeding to the modelling phase, we confirmed visually that the steps undertaken had improved the problems found in the Data Understanding phase of this report.

4. Modelling

4.1. Modelling Technique

As mentioned in the project plan (and confirmed along the data understanding and preparation), the idea to proceed with a clustering algorithm was to use a two-perspective technique (and subsequent merge) that would involve understanding separately what the personal preferences of the customers are and how they engage with the company.

Given the problem we are trying to solve, inserted in a business environment, we want to form clusters of spheric-like shape. On this note, the most known algorithms that we can use are K-Means, SOM and Gaussian. While the SOM algorithm allows us to have access to the visualization techniques that we used in the data understanding phase and the Gaussian provides us soft clustering (each customer will not only belong to a rigid cluster but will have a "percentage" belonging to each of the clusters), given that after developing the two perspectives we will merge them, these advantages become not usable for our final clustering solution. Based on this and how easy it is to explain the K-Means algorithm to non-technical people (it is already established that the W.W.W. lacks technical skills and in the short run, it will be important to have the support and comprehension of the managers, which becomes easier if they understand the logic behind), we will explore this approach, keeping in mind that if at the end the results are not satisfactory enough to fulfil our data mining success criteria, we can simply try different algorithms (with the disadvantage of not being as explainable).

Regarding K-Means, and being it based on the Euclidean distance between customers, it relies on the assumption that all the variables used are on the same scale and are all metric. Moreover, the model assumes spheric-like clusters (which we have already deemed appropriate for this context) and is very sensitive to outliers (that we have already carefully dealt with). Lastly, it requires us to select *à priori* the number of clusters to create, which we will address later on.

Another important technique we will use in our modelling process will be hierarchical clustering. We will use this algorithm to verify if when merging the two clustering perspectives some clusters seem sufficiently similar to also be merged. This algorithm's main assumptions (e.g., distances and the need for scaling) will already be dealt with before applying the K-Means.

We will also recur to a simple KNN-Classifier to attribute our previously separated customers-outliers to the clusters they more closely resemble, so we can also cover them in our marketing strategy (once more, the problematic assumptions of this model will already be solved when reaching this step, as they are also mainly related with the distance calculation and the exclusive use of metric features). Lastly, we will use a full-grown decision tree (to help us in the assessment of our clustering solution to discover which are the most relevant characteristics to classify customers as belonging to a specific cluster) and a t-SNE technique (to have more tools to make a reliable decision regarding the approval or not of the data mining success criteria). Note that we are only mentioning the use of Hierarchical Clustering, KNN-Classifier, Decision Tree and t-SNE here, as it seems to exist a sort of grey

area in the CRISP-DM methodology on if the models used to help in the construction and evaluation of the main model should be or not mentioned in the model selection phase.

4.2. Model Construction and Assessment

Given the nature of the method our team decided to use in this project, our modelling will not require the generation of a test design. Being this said and building on the last section of the present report, we decided to divide our variables into two groups: Wine Taste (with variables related to the preferences of the customers: Dryred, Sweetred, Drywh, Sweetwh, Dessert and Exotic) and Engagement (where the variables refer to the customers' interaction with the company and some personal information: Edu, Perdeal, Value_index and Web).

With these two groups formed, we applied the Inertia-Elbow technique and the Silhouette Score to attempt to determine what would be the number of clusters to form for each perspective. In the case of the Engagement group, the Inertia and the Silhouette Score both suggested clearly that 2 would be the correct number of clusters, while in the case of the Taste group, it was somehow more ambiguous (the Inertia plot indicated both 2 and 3 would be a good choice, and although the average Silhouette Score was higher for 2 clusters, the Silhouette plot showed that in that case we would end up with a very good cluster, but with a second one that barely met the average score), but we decided to keep 3 clusters.

After proceeding to the cluster formation using K-Means, we merged the two perspectives (resulting in 6 clusters) and applied a hierarchical clustering technique on top of the clusters formed. Then, by analyzing the dendrogram plot, our group opted to merge two additional pairs of clusters. Lastly, to conclude our model creation, we added back the outliers we had excluded in the data cleaning phase, using a KNN-Classifier between each outlier and the centroids of the clusters (we only used one neighbour, as we tried to "mimic" the last step of the K-Means process of considering the points closer to each centroid as belonging to the clusters centred in that centroid).

In order to assess the quality of our model and if it met our subjective data mining success criteria, we needed to recur to some visualization tools. Firstly, using a coordinate plot, we were able to observe the average behaviour of the "typical" customer of each cluster across metric features, and it was clearly verifiable that the clusters presented well-diversified and separable characteristics.

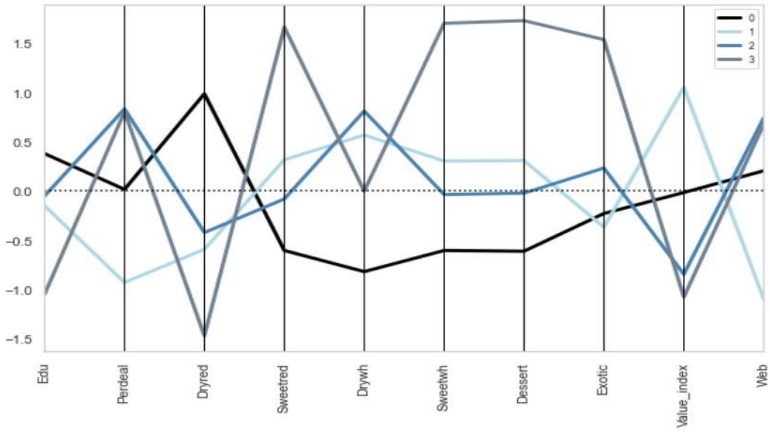

Figure 1: Representation of the behaviour of the average customer across clusters.

Building on this, when comparing how the proportions of the binary variables vary across the clusters, it was also found, in some cases, a substantial difference**.**

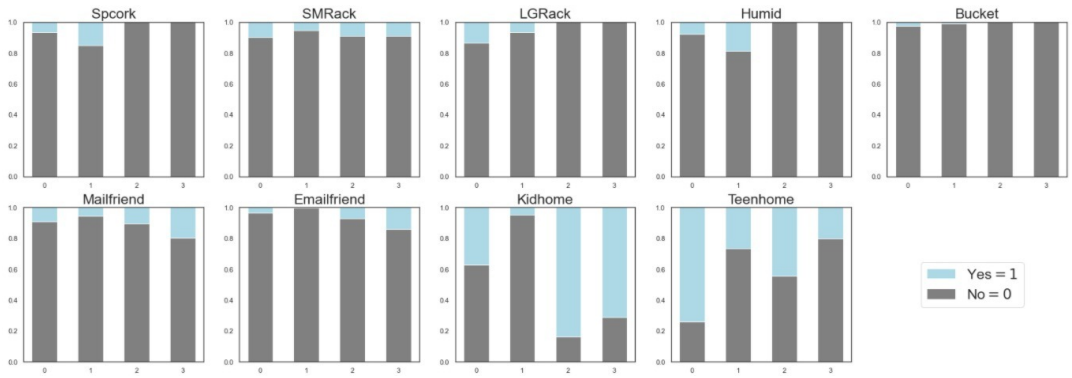

Figure 2: Difference in proportions (of the binary variables) across clusters.

Following the same reasoning, we decided to also use the t-SNE technique that allowed us to confirm in the 2D space that there was a good separability among the clusters formed and a Decision Tree to help us understand which were the variables that contributed the most for the separation of customers in different clusters. Additionally, no cluster was constituted by such a reduced number of donors that would not be worth to develop a specific strategy for them. Lastly, and although not being necessarily a good way to evaluate the model, our team decided to also verify the scores achieved in other more objective metrics: R^2 and Average Silhouette Score with 0.45 and 0.19, respectively.

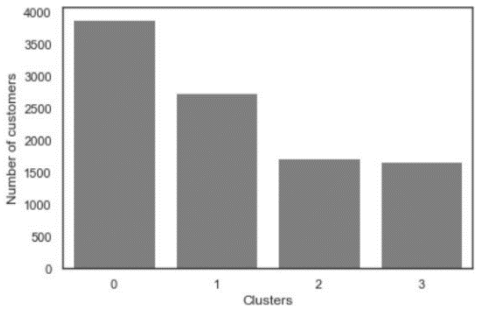

Figure 3: Distribution of customers per cluster.

Overall, our team decided with unanimity that the data mining success criteria was met and that there also seem to exist the right fundamentals to believe that the business goals will be met, more on this will be presented in the next section.

5 Evaluation

5.1 Evaluate results

The final solution obtained through the application of the K-Means algorithm successfully achieved one of the initial business success criteria (provide useful insight about the different types of customers), as the analysis above regarding the accomplishment of the data mining success criteria suggested, and gave us confidence in the possibility to, after deployment, also reach the other (improve the marketing return by adopting a segmented strategy). Building on this, it was possible to identify different characteristics for the different types of customers by looking at the parallel coordinates plot and the proportions of the binary variables presented in the last section, and that we will now explore a little deeper, providing context to the business side.

Cluster 0 is the largest one, being composed by almost 4000 customers, who are, on average, more educated and seem to be the group that likes the most the dry red wines and has the highest

distaste for sweet/semi-dry reds, dry white wines, sweet/semi-dry white wines and dessert wines. This cluster also represents the customers who are more likely to have teenagers under their care and on average are the ones more interested in buying large wine racks. Meanwhile, cluster 1 is composed by 2600 customers, who seem to dislike dry red and exotic wines, while, although never being the group that most strongly prefers any wine, still show a taste for dry white, sweet/semi-dry red, dessert and sweet/semi-dry white. Furthermore, this group seems to be especially important from a business perspective, as they have by far the highest value, according to the metric our team created to measure a customer's value (Value Index), while not being fond of buying wine at discount and through W.W.W.'s website. Given the importance of this cluster, it is important to note that they show a dislike for being contacted by either email or mail, but seem to be the ones who have enjoyed the most the sale of wine humidifiers and silver-plated cork extractors. On the opposite side, the average customer from cluster 2, compressing 1600 customers, is especially eager to buy products at discount and online, whilst not being particularly valuable for the company. Regarding the wines, they prefer dry white and exotic wines. The customers from this cluster are also not prone to buying accessories (the only event in which they even participated was the sale of the small wine rack) and are the ones more likely to have kids under 13 years old. Finally, cluster 3 has 1500 customers, being the smallest one. This group presents the lowest Value Index among all clusters, has the lowest education level and has a high aversion to dry red wine. Moreover, they purchase a significant number of wines on promotion and using the website, while being particularly keen on sweet/semi-dry reds and white wines, dessert wines and exotic wines. Lastly, the members from this cluster show the same behaviour as the ones from cluster 2 in respect to the accessories bought, but present an interesting characteristic that is being more prompt to receive mail and/or email communications.

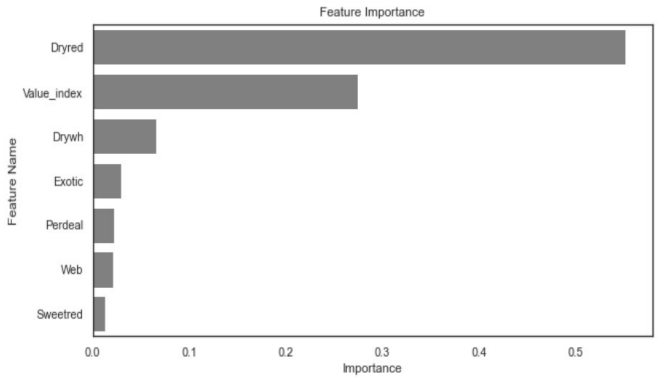

Figure 4: Difference of feature importance according to the reduction in Gini

Additionally, and constructing on one of the more generic requests that the W.W.W.'s manager made us (asking which were the most important characteristics that separated customers), we can now conclude that it is mainly the taste for dry red, followed by their value to the company (figure 4). It is also based on this, that we present a hint/direction for future exploration: a supervised learning technique to predict which may be the people more "worthy" of being targeted in the future, using as target the variable we created and that we found to be a good separator between groups of customers, Value Index. Note that for a more reliable use of this parameter, it would be good to know how the variable LTV was computed (as it is part of the construction of the variable we created), being that for this, W.W.W. should contact the team responsible for implementing it on its database.

Lastly, we think it is worth to recommend, for future data mining endeavours, W.W.W. to start to gather some more information about its customers, as this would bring value for their profiling, as their location ('State' if it is from the USA or 'Country' if outside) and/or the physical stores from where they purchase the most, being that this would also help to understand the most profitable selling points (or the preferred selling points of our top customers), and thus allow to market those in a more aggressive way.

5.2. Review Process

Looking at the whole data mining project, our team did not spot any mistakes or failures made during its process. In every step, we attempted to be as careful as possible, which was sometimes unfruitful, but that assured us that everything was properly being done. Furthermore, and although some grey areas were found in the CRISP-DM methodology and how to proceed per it in more complex situations, our team followed it to the best of our capability, leaving everything well documented.

Lastly, it could be made the point in favour of attempting more clustering algorithms, such as SOM and Gaussian, in order to compare the results obtained with the ones provided by K-Means, but at the end of the day, our team decided, by the reasons expressed in the proper section, that doing so would be mainly to complete check-boxes and not because in practical terms it made sense for this particular case, which ends up saving man-hours/resources to both our consulting firm and our clients.

5.3. Determine Next Steps

Regarding the next steps in this project, we have several hypotheses of what to do if we decide to delay the deployment. Firstly, we could resort to different measures and algorithms to determine the outliers (as we have discussed the outliers are an important weakness of K-Means), nevertheless, we have already removed a significant part of the outliers and our model doesn't seem affected by non-detected outliers. We could also try more feature selection techniques, that would make the model more efficient, but we ended up with only 10 metric features, presenting them a well-differentiated pattern across clusters. Finally, another approach would be to try more clustering techniques, which would increase comparability, but again, it is harder to prove the value of what cannot be explained simply, mainly when the simple method seems to yield very good results.

Overall, considering all the delay options would increase the time and resources needed and that the business objectives seem to be on a good path for being achieved, we decided to proceed to the deployment stage, as no mistakes were spotted across the data mining project, the reasoning behind all of our decisions is well documented and the business objectives and data mining goals were met or are well directed to be met.

6. Deployment

6.1. Plan Deployment

As in most projects, the data mining phase is not the end of the project. In this particular case, the next step involves coming up with a deployment plan which must include a marketing campaign. To do this, we will follow a marketing mix approach with some of the steps we think W.W.W. should

undertake, as part of the deployment of this project, to reach new and existing customers from each segment presented above and know which ones should be prioritized.

Regarding Cluster 0, the product dimension for this cluster should be focused mainly on dry red wines (being this the only cluster interested in this type of wine), while the price level does not seem to be of a big relevance as this cluster shows to be indifferent to promotions. Furthermore, as they do not particularly like to purchase online or to receive mail and/or email catalogues, W.W.W. should focus mainly on the physical stores to reach them, particularly by having a large offering of different types of dry red wine (although these customers are the only ones that, on average, seem to like significantly dry red wine, we have to keep in mind they represent the largest share of our customers, and our second best segment in terms of value). From the promotion point of view, a good way to regain some latent customers from this segment may be to have a sale of large wine racks, and considering this group is very likely to have teenagers at home, it may be a good strategy to include "lucky draws" of prizes (e.g., laptops and videogames) only in the dry red wines (allowing to only target these customers and not incur in excessive spending on prizes).

When it comes to cluster 1, where our most valuable customers are, they do not have a particular clear favourite type of wine, so we should target this segment with all types of wine, except dry red and exotic wine. These customers have to be reached by physical stores only, given their avoidance of the other selling channels. Similarly to cluster 0 (but even at a higher extent), this group is not specifically eager to buy on promotion, and considering that these two clusters are the ones driven towards physical stores, it raises the question that making in-store discounts may be a waste of money and effort. Given the importance of this cluster, it is worth to spend a little more on advertising (also to capture new customers similar to this group), using mainly classic marketing channels (e.g., newspapers and TV) and even partnerships with golf clubs (as customers from this group tend to be wealthier and older) to reach them. Lastly, being these our top customers, it may be a good idea to offer a loyalty reward when they a reach a certain number of years with the company, for instance, a silver-plated cork extractor (as they seem to like this product).

To reach the clients from cluster 2, the efforts should be almost exclusively towards W.W.W.'s website, where they should have constant discounts (for both this cluster and cluster 3), but only if economically feasible, as this customer segment doesn't bring a lot of value. We also recommend some small budget social media campaigns towards ads advertising dry white wines, exotic wines and small wine racks. Furthermore, given the propensity this group presents to have kids at home, W.W.W. could try to establish partnerships with cinemas, zoos and thematic parks to offer family-pack tickets when a large number of bottles of wine is purchased (again, this should be focused only online to not give "free-meals" to customers from clusters 0 and 1).

Lastly, and keeping in mind that the customers from cluster 3 are the less valuable ones, and the smallest group, W.W.W. has to be very careful in the expenses incurred with campaigns. A low-cost option is to use the mail and email to send some online-vouchers and advertising of the website, showing mainly sweet/semi-dry reds and white wines, dessert wines and exotic wines. We have to stress that any effort towards this cluster should be proceeded by a careful profit analysis.

Being the main part of the deployment project presented (the marketing mix), our team decided to also develop an application where new customers' data can be inserted, returning the cluster where, on average, they would find their peers, and also the purposed marketing strategy. This will help to quickly categorize customers from outside this sample group, but that W.W.W. may find it would be worth it to be included in this pilot run of a segmented marketing strategy. Note that the application replicates all the needed transformations to the data and calculates the distances to each cluster, assigning the closest cluster to the inserted customer. Therefore, this data mining project can be applied to new customers, even by non-technical people.

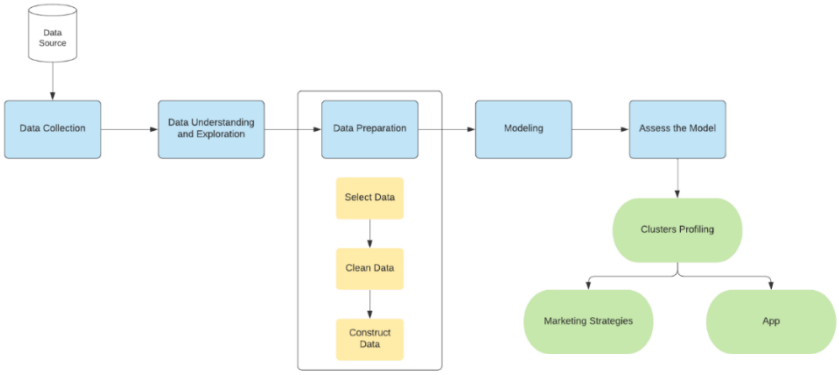

Figure 5: Final overview of the general steps taken according to the CRISP-DM methodology.

6.2. Plan Monitoring and Maintenance

Our team considers that will be critical to monitor and measure the impact of the marketing strategies adopted - by comparison of the marketing return before and after the implementation of these strategies -, to prove the value of a data-based decision-making process to the W.W.W.'s managers and to also verify if the remaining business success criteria were met. Moreover, given that some of the suggestions provided also include ways to capture new customers, W.W.W. could also implement a system of asking new customers how they have found out about the company, to keep track of the efficiency of the advertising channels. Furthermore, our team suggests that our application should not be used imprudently, meaning this, there is the need to keep a constant track of the characteristics of each centroid, and if at a certain point the characteristics deviate more than a certain threshold, the data mining process should be redone using the larger database. Being this said, we believe that if the customer segment we receive is representative, the difference between clusters is sufficiently high for the results to still be valid for a moderate number of new customers.

6.3. Review Project

Although the final project review should be done after the deployment in collaboration with the clients, we will provide some conclusions of the work undertaken. While applying the CRISP-DM, our team was able to assure all the steps were properly done and conduct a well-documented report that will help the W.W.W. to better understand our process and will facilitate future work. Not all the steps were straightforward, as methodologies have always to suffer some adaptation to the problem being faced, but overall, our team believes to have developed a well-constructed solution that met the requirements of our clients, and that will help them to launch a successful marketing strategy based on segmentation.