BUSINESS CASES WITH DATA SCIENCE

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

Wonderful Wines Of the World

Group_N

Bruno FIlipe Prazeres Ines Soares, number : 20200658

Xavier Golaio Gonçalves, number: 20201090

Li-Lou Dang-Thai, number: 20200743

March, 2021

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

1. Introduction 3
    1. General Context 3
    2. Methodology 3
2. Business Understanding 4
    1. Background 4
    2. Business Objectives 4
    3. Business Success Criteria 4
    4. Situation Assessment 4
    5. Determine Data Mining Goals 5
3. Data Mining Process 5
    1. Data Understanding 5
        1. Initial Data Collection 5
        2. Data Description Report 5
        3. Data Exploration Report 5
    2. Data Preparation 5
        1. Feature Selection 5
        2. Feature Engineering 7
    3. Modeling 7
        1. Customer Segmentation 7
        2. Wine Segmentation 8
        3. Cluster Merging 8
    4. Evaluation 9
4. Results Evaluation 9
Marketing Strategy 10
5. Deployment and Maintenance Plans 11
6. Conclusions 11
Consideration for model improvement 12
7. References 12

1. Introduction

1. General Context

Finding new customers is vital in every industry, this process begins by learning as much as possible from the existing customers. By understanding current customers, organizations are able to identify groups of customers that have different product interests, different market participation or different responses to marketing efforts.

Market segmentation, the process of identifying customers' groups, makes use of geographic, demographic, psychographic and behavioral characteristics of customers. By understanding the differences between the different segments, organizations can make better strategic choices about opportunities, product definition, positioning, promotions, pricing and target marketing.

2. Methodology

The analysis followed a simplified version of the CRISP-DM framework recommended by Kellcher, Mac Namee, & D'arcy (2015, p.53).[1] as well as the information available at CRISP-DM Overview at IBM Knowledge Center.

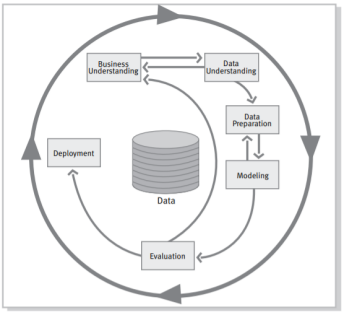


Figure 1 – CRISP-DM2

This framework of analysis is a proven way to guide data mining efforts and tools. Including descriptions of the phases lead to the guidance of a project, it explains each one and the relationships between tasks. Representing the process model, it presents an overview of the life cycle of data mining. In this context, the life cycle is composed of six phases which are the most relevant, as well as their dependencies, represented with arrows. The model presents itself as a flexible representation that can be easily customized. CRISP-DM allows analysts to create models that fit our particular needs (IBM Knowledge Center). Since this is an iterative process where advances and insights acquired along the steps can influence the prior decisions, the steps are not entirely followed in a strictly linear or straightforward logic.

According to the framework, data understanding and preparation phases are more significant than modeling, evaluation and deployment phases nevertheless these latter are important to consider for future data mining goals, in the context of eventual questions raised. Moreover, certain phases in CRISP-DM are more closely linked together than others. For example, Data Preparation and Modeling phases are closely linked, and analytics projects often spend some time iterating between these two phases.[2]

2. Business Understanding

1. Background

Wonderful Wine of the World is a seven year old American company that is looking for unique wineries across the world, in order to sell the best wines to its customers. WWW's purpose is to find the most rare, tasteful, and exotic wines to charm and satisfy its clientele.

The company owns multiple stores around the United States. Their products are also available on their website, or it can be purchased directly by telephone (after looking at the catalog). Several hundred selections are available in each new catalog, sent every 6 weeks.

2. Business Objectives

WWW wishes to better understand the profile of their customers. They want to know their behavior, and their preferences, in order to be able to define new marketing approaches to better provide each customer. The goal is to put in place a marketing strategy (marketing mix : product, price, place, promotion) for each segmentation (clusters) that the data science team will uncover.

3. Business Success Criteria

Through aggressive promotion in wine and food magazines, WWW now has 350,000 customers in its database. Most customers are highly involved in wine, and have sufficient money to indulge their passion for wine. WWW sometimes offers wine accessories as well – wine racks, cork extractors, etc.

WWW is trying to make use of the database it started 4 years ago. So far, it has simply mass-marketed everything with no differentiation between different customers. The entire clientele get the catalog, and there are no loyalty programs or attempts to identify target markets for cross-selling opportunities. Now, WWW wants to "get smart" about its database, and start differentiating customers, and developing more focused programs which can greatly improve the results of their marketing strategies.

WWW has provided a sample of 10.000 customers from its active database. These are all customers who have purchased something from WWW in the past 18 months (after 18 months with no purchase, a person is eliminated from the active database). It was these 10.000 randomly-selected people who were sent the test promotion for the silver-plated cork extractor

Since WWW usually does not differentiate their marketing strategies towards the different client segments, finding these different types of clients and creating relevant strategies towards their different characteristics is the main criteria to allow the modeling to be useful for WWW in business terms.

4. Situation Assessment

Our team was given an Excel file, which in itself indicates the level of digital maturity WWW processes, with a sample of the much larger database WWW keeps of its clients. We will have the possibility of speaking directly with the CEO of WWW, allowing us the continual improvement of our work process with possible feedback from the administrative board. In terms of costs, our model will require man-hours to build but, from the relatively small database size, will be able to run on economical hardware while generating considerable returns when the model is applied.

As requirements, assumptions and constraints, our team have access to a limited dataset made up of 10.000 clients which contain information on WWWs' consumer's behavior. The company sends catalogs to the clients every 6 weeks with advertising and promotions. In order to apply business

and marketing strategies defined after the data mining procedures apart from these catalogs, WWW is also using its website as an online sales platform. The types of available wines are dry red, sweet or semi-dry reds, white wines, sweet or semi-dry whites, dessert wines which compose the total array of products sold, of which some are considered exotic. The dataset also has personal information about the clients as well as mail and e-mail friendly clients, and complaints in the last 18 months.

5. Determine Data Mining Goals

As part of the used framework, our team must fulfill business requirements, as well as Data Mining success criteria. On one hand, customer segments must be identified according to the particularities of the customers, in order to support any business insights and applications based on the conclusion, in light of business and marketing strategies to maximize potential utility for the company. On the other hand, Data Mining goals will be established in the context of data exploration, identifying variables that should segment the customers, number of clusters, dimensionality reduction among others. All these objectives shall be readable from a non-technical point of view and data visualization might be an important instrument. Wonderful Wines of the World pretends to use its database in a smarter way so that more focused and customized programs can be provided to differentiated clients.

3. Data Mining Process

1. Data Understanding

1. Initial Data Collection

The dataset was provided by the IT team of Wonderful Wines of the World, it is in a single excel file. We only encountered one problem with the excel file, the last line and column of the data was equal to 'Rand' which we decided to remove before importing it to our Jupyter Notebook.

2. Data Description Report

Our dataset is composed of 10 000 customers and 29 features. All the customers from the data have made at least one purchase in the past 18 months. We identified four categories of variables: the customer's characteristics, their consumer behavior, the type of wines and the accessories they buy. (Table with the division of features into the four categories can be seen in the notebook).
After analyzing the data through Pandas Profiling and the DTale library, we found that there weren't any missing values or duplicated rows.

3. Data Exploration Report

The data is composed of metric features, and binary features, which we divided into two different groups. The data was coherent, having plausible values even for some of the most extreme cases of some variables.

2. Data Preparation

1. Feature Selection

The data preparation process is key to ensure the accuracy of the data our group has prior to any analysis to be able to gather accurate insights. We started by analyzing the number of missing values present in each feature to verify if any data imputation was needed (Table with each features description can be seen in the notebook).

All the rows had values for each feature so our team proceeded to study each feature in terms of its relevance to the business problem in question. From the four main categories of variables we concluded that all 28 variables (excluding the customer ID) would be useful for our segmentation. From these variables there are two noticeable ones: ACCESS and LTV that appear to be calculated from other variables available in the dataset, our team decided to use them for the clustering process as they could still provide some insight for the marketing strategies for each cluster. Besides the calculated variables mentioned, we tried to find correlations between the variables by using Pearson Correlation to take into account the existence of binary variables, and to find any redundancies in the variables we could use.

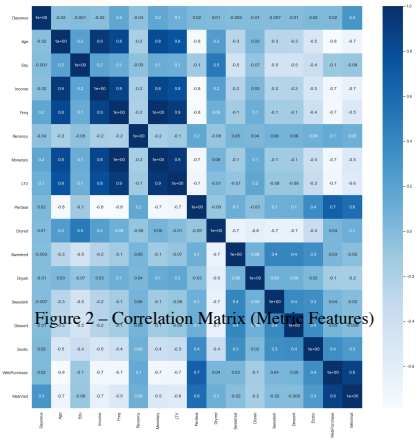


Figure 2 – Correlation Matrix (Metric Features)

Even though there were some high correlations between some variables, for instance Income with Monetary, Age, and LTV, or WebPurchase with WebVisit, our team found that all these variables were meaningful for the understanding and differentiation of the clusters so we decided to keep all of them for the segmentation.

Another important step in data preparation is the study of outliers, after checking the Box-plots of the metric features, we applied the LOF (Local Outlier factor) to evaluate them, and chose to remove the outlier customers because they only represented 2,6% of the whole sample.

Finally we studied the discriminatory ability of the variables by studying the variance and the distribution of the values of each feature.

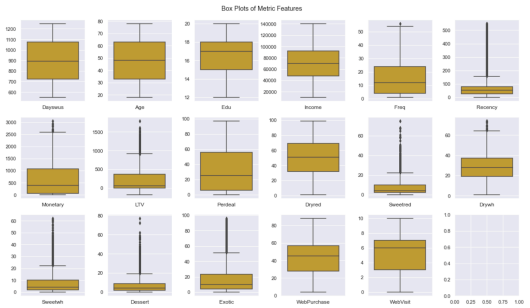
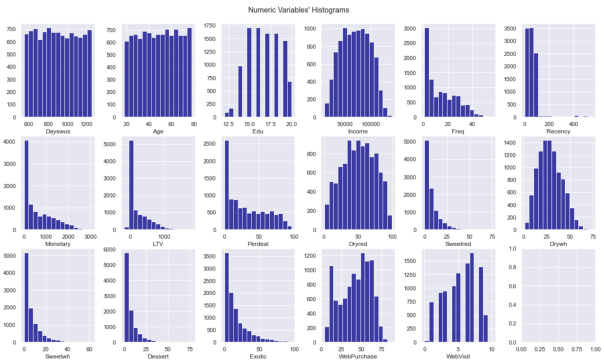


Figure 3 – Box Plot of Metric Features



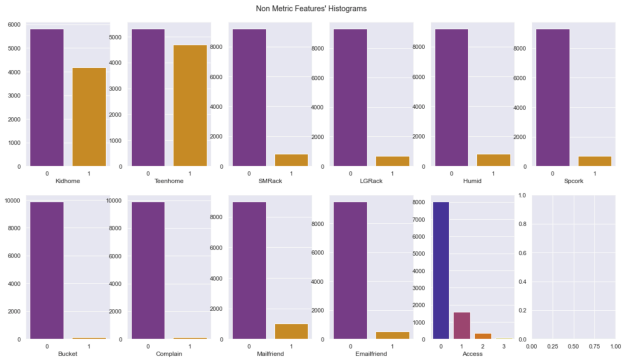


Figure 4 – Histogram of the Numeric Variables and Non-Metric Variables

2. Feature Engineering

Besides the removal of the outliers for the modeling, we noticed the presence of different scaled variables in our data set which would bias the result of the segmentation. For this reason we decided to standardize all variables of the dataset, using the MinMax Scaler.

3. Modeling

For the clustering, we decided to use the K-Means on the numerical features and K-Modes on the binary variables.

Regarding the binary variables, after clustering them with KModes, we realized that only « Kidhome » and « Teenhome » had significant discriminatory ability as the other variables had over 98% of the same value. Finally we concluded that they weren't that relevant as they had high correlation with age variable, and decided to focus more on the numerical variables.

The metric data is divided in two: customer and product features. To support our decisions on the number of clusters k, we used the inertia K-Elbow plot and the silhouettes. From the clustering results we arrived at, we labeled and described each one according to their characteristics taking special consideration on the main inter-cluster differences.

1. Customer Segmentation

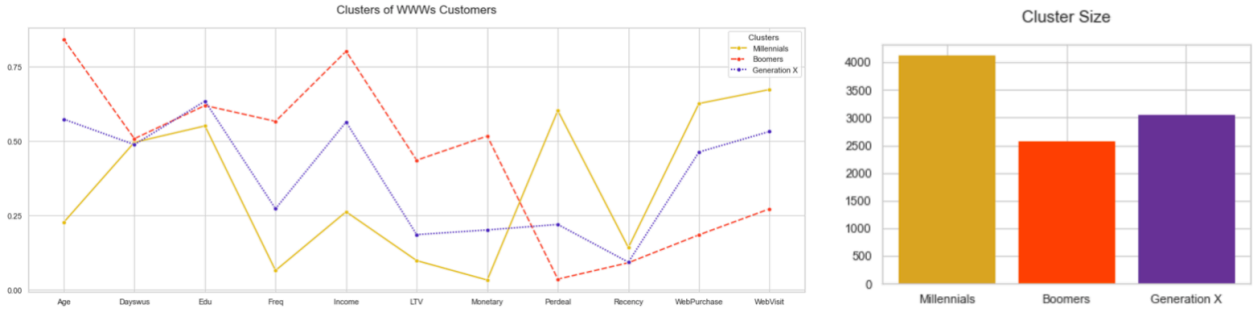


Figure 5 – Customer Clustering

Cluster 0 - Millennials : The principal characteristic of this cluster is that they are the younger customers of WWW. They represent the biggest cluster with almost half (around 4000) of our dataset. Millennials mostly purchase on the website, and compare to the other clusters, they have the highest percentage of wine bought on discount. In the past 18 months, cluster 0 has the lowest number of purchases, and total sales amount, however their last purchase was the most recent.
Cluster 1 - Boomers : This cluster shows the older and the wealthier customers of WWW. They spend the most money, and made the most purchases in the last 18 months. Cluster 1 rarely ever buys a wine on discount, and will usually favor buying directly in a store or by telephone. It is the smaller cluster with approximately 2500 customers.
Cluster 2 - Generation X : Generation X is the "in between" cluster, they are located in the middle of Millennials and Boomers. Customers from cluster 2 have a "medium high income", but they sometimes purchase wine on discount and they often buy it online. Generation X has a significant higher income, but compared to Millennials, its Life-Time-Value is not as high as we could expect.

2. Wine Segmentation

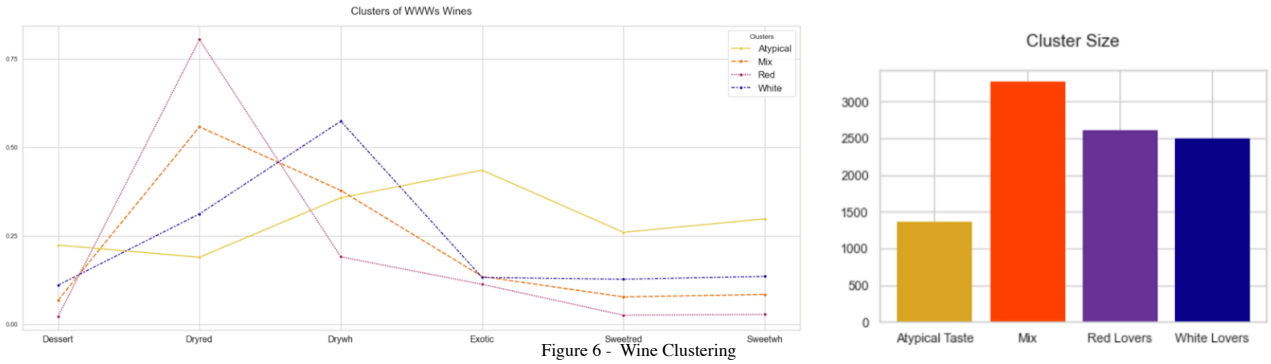


Figure 6 - Wine Clustering

Cluster 0 - Atypical Taste : The primary feature of this cluster is its high interest in exotic - unusual - wine. It follows an opposite pattern in comparison of the three other clusters. Customers with Atypical Taste will mostly enjoy sweet and semi-dry wine. In general, they also prefer white to red wine. It can be implied that this segment of consumers appreciate original wine and go for bolder choices. Cluster 0 is composed of approximately 1450 customers, and is the smaller cluster.

Cluster 1 - Mix : It is our bigger product's cluster with almost 3500 customers. Their taste in wine is variate, even though they frequently buy more dry red and white wine. Compared to the other clusters, its distribution lies in the middle.

Cluster 2 - Red Lovers : This cluster represents the Red Wine Lovers, more than ¾ of their purchase is composed of dry red wine. Customers from cluster 2 rarely purchase sweet wine, and the rest of their purchase is lower than 25% of the total. Red Lovers has around 2500 customers.

Cluster 3 - White Lovers : On the contrary to cluster 2, White Lovers particularly appreciate white dry wine. They represent almost the same amount as the Red Lovers, with 2500 consumers. Compared to Mix and Red Lovers, cluster 3 purchases around 12,5% of each exotic, sweet, semi-dry wines.

3. Cluster Merging

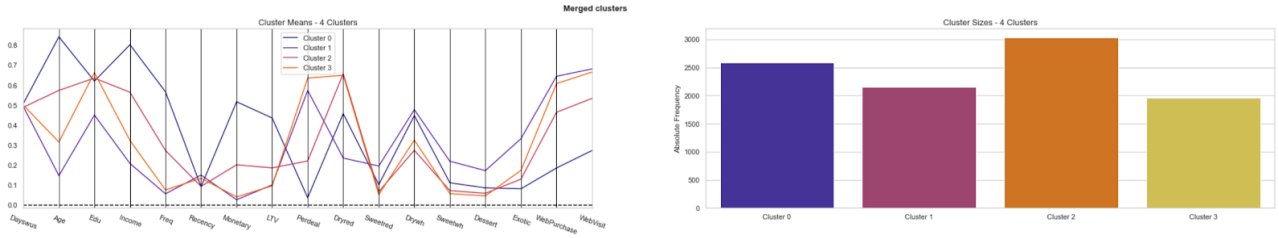


Figure 7 Merge of Clusters

Cluster 0 - Boomers with mixed taste : Our first cluster is composed of the "Boomers" clusters, they are the older customers with a variation of different tastes in wines. It represents the customers with the highest income, and their total amount of sales in the past 18 months is greater than the three other clusters combined. In contrast with the rest, cluster 0 rarely buys on the website. They must prioritize in-store sales, or by telephone with the catalog.

Cluster 1 - Millennials with a "sweet tooth" : Our second cluster shows that the younger generation is more interested in tasting and purchasing atypical wines. They tend to consume more sweet wines and are not as interested as all the other clusters in dry red wine. More than half of their purchases were bought on discount, and they have the higher percentage of buying directly online.

Cluster 2 - Generation X "The casuals": It is the bigger merged cluster with a little more than 3000 customers. ⅜ of their purchase is dry red wine, and they don't buy much of the other different

wines. The total amount of their sales in the last 18 months is higher than cluster 1 and cluster 3 but it is less than half of cluster 0.

Cluster 3 - Young educated with dry red wine : The last cluster is the smaller one with almost 2000 customers, a little less than cluster 1. Their age group stands in between Millennials and Generation X while having an education considerably higher than Millennials. It concerns mainly individuals who didn't purchase a lot in the last 18 months, with a low total amount of sales, and they are the higher cluster to buy their wine on discount. They're mostly Red Lovers and prefer to buy on the website

4. Evaluation

Since clustering is an unsupervised learning method, we don't have any outputs to compare it to. However we can evaluate our model with the number k clusters we chose. To determine the optimal number of clusters we used the K-elbow method and the silhouette that can be seen in the notebook. To assess the model and feature importance, we used a decision tree classifier, and got the following results.

```
It is estimated that in average, we are able to predict 90.25% of the customers correctly
```

4. Results Evaluation

Modeling the metric variables gave us main merged clusters. In terms of binary variables, we tried to use K-Modes to model them but found that most binary variables to not have enough discriminatory ability to use as most had the value 0, for example, accessories bought from most customers in the sampled data set. We decided to study these binary variables *a posteriori*, from the already established clusters to then better direct our marketing suggestions to which one.

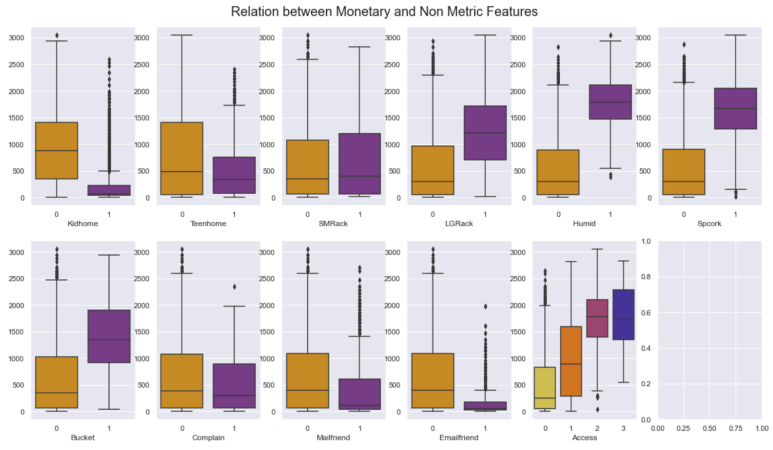


Figure 8. Box Plot of Monetary and Non Metric Features

The final segmentation results had quite interesting results in terms of the Business Problem, as the different clusters showed enough significant differences between them to allow for the creation of different marketing strategies for each one. With the Business criteria in mind we labeled and defined each cluster and found the most adequate marketing strategies for each segment of customers of WWW.

Marketing Strategy

Cluster 0 - Boomers with mixed taste :

Boomers with mixed taste, the cluster containing the older and the wealthier customers of WWW, spend the most money, buy most frequently, rarely buy on discount and do not buy online. Their preferences fall from dry white to red wine. For this cluster we should focus less on promoting discounts towards them, as they won't be scared to put higher amounts for a good bottle of wine. The company should use direct marketing efforts towards phone calls, and store catalog contacts. Information to this segment of the market must be clear, customer service should be always available, extraordinarily attentive and there must be given information directed to improving their lifestyle. When doing campaigns, our products should always be associated with activities related with this segment tastes such as enjoying the wine while spending time with their families.

Additionally, WWW might offer a wider variety of high end wines while also having special interest in avoiding losing these clients as they represent an important segment of the clients. Another alternative would be to send Boomers, catalogs addressed with a premium list of exclusive expensive top quality wine.

Cluster 1 - Millennials with a "sweet tooth" :

Millennials, the cluster representing the youngest customers in the sampled data set, have the lowest amount of income and money spent with our products. At the same time they are the ones who mostly buy online (stores might not be so important to this segment) while also purchasing on discount the most. There also seems to have a preference for sweeter, and fruity wines, while also enjoying tasting atypical wines from all over the world.

Because of these characteristics WWW should market to the elements of this segment by sending more promotions including discounted wines, with special focus on displaying them online. Taking into consideration their lower purchasing power, besides the discounted wines they can be marketed less expensive wines. A useful strategy for this segment would be staying socially connected on the internet using both social networks and channels and also engage them with customer comments and opinions, create brand evangelists to represent WWW to create contests and content about our products in their platforms. Communication should happen using web presence or by email contact and catalogs might not be so important.

The Exotic product line shouldn't be available in-store, since millennials are the main buyers of this type of wine, and mostly buy online, the products would just sit in the back-store and expire.

Cluster 2 - Generation X "The Casuals":

Being the "in between" cluster, this cluster can be further segmented and studied to try to understand underlying characteristics of these customers. This generation's favorite product is dry red wine.

Their ratio of frequency and monetary is lower than expected considering their high income. They enjoy dry red wine but they are not spending the expected amount of money on it. Almost half of their purchases are made on the internet and the other half is divided between phone purchases and in real stores. Generation X are occasional customers, they enjoy drinking wine at the dinner table, and mostly buy directly in-store. For that reason, an interesting campaign should be through web advertisement. Most of the people in this generation prefer more personal and authentic media, and this marketing efforts could be made through social networks and email marketing and even through the website.

Cluster 3 - Young educated with dry red wine:
Cluster 3 is mostly composed in young educated people, and their preference falls to dry red wine. On some occasions, white wine is also chosen, but all the remaining types are deprecated. In this cluster, the Life Time Value is very low, which translates into very low value for the company. This factor is also explained by a low amount of sales.
In order to increase sales in this segment, one possible path is to increase discounts and promotions related to dry red wine. Since this is their preference, we will likely enlarge our current customer sales and penetrate into this sector, gathering new future clients. Similar to what happens with the millennials, these young educated people mostly buy online, but since it is a common product, it should also be available in store. Additionally, a good way to engage clients from this segment would be sending discounts through email (and less through catalogs).

5. Deployment and Maintenance Plans

To allow the operationalization of the segmentation model there could be a dashboard to better understand better the current customers to easily explore the different client segments. The deployment of the model could be applied to the full database of customers, allowing the possibility of better understanding the total population, so that the marketing strategy to follow is as accurate as possible.

After the deployment of the model the data sources should be validated and updated frequently as the data concerns a time span of 18 months and customers behavior can change over time. With the current data set of around 10.000 clients (of a total of 350 thousand) that were sent the test promotion for the silver-plated cork extractor there is no problem of performance with the model in terms of hardware and software performance. In the eventual case of a bigger sampled population there may be the need to get better hardware performance to be able to produce the same results in useful time.

6. Conclusions

In this project, our purpose was to understand WWW customers behavior and preferences and try to find the most adequate best marketing solutions based on their characteristics and needs. This process was accomplished by grouping the customers with similar behaviors and characteristics by using the segmentation model and then retrieving intelligence from the shared characteristics of each cluster/segment. After this segmentation we used the rest of the relevant variables in the data set to target specific interests of each cluster to allow for a more accurate and effective target marketing.
The results showed us significant varying behavior between the different customer segments we arrived to which, in consequence, gave us the possibility of creating marketing strategies much more relevant for each customer. This targeted marketing can greatly decrease costs in marketing efforts while also maintaining a healthier customer base as any promotions offered to each customers will be more relevant to their interests. In conclusion, not only increasing the customer's profit to the company but also keeping them more engaged as they will not receive irrelevant promotions to their interests. All these efforts consist of a great advantage to the company's efforts in keeping the current clients, understanding them better in terms of their relation with WWW and

can even be used for future strategies to capture new clients. In general the use of this model will always be valuable for the company as they are more informed on what business decisions to make with the help of data driven information.

Consideration for model improvement

The model used to segmentate the customers could be improved if more data was collected, not only new features, but also with the information we already have, such as how was the response to mail or e-mail contacts, time series of purchases, more detail on which type of wines and accessories are purchased and not just a general category, relation between discounted purchases and customer behavioral change. Examples of new features would be location, complaints details, working from home, number of children in the household, neighborhood religion census (as different cultures represent different perspectives about alcohol consumption), car ownership and its use to commute, smoking, job sector and liquor accessibility. Another option for possible model improvement is to create a trigger after the period of 18 months to try to recapture the inactive clients.

7. References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000) CRISP-DM 1.0: Step- by-Step Data Mining Guide. SPSS, Copenhagen.

Kellcher, Mac Namee, & D'arcy (2015, p.53).

The data mining life cycle. Image from IBM Knowledge Center, https://www.ibm.com/support/knowledgecenter/it/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html