NOVA IMS Information Management School

D | A | T | A | L | I | N

Bringing Intelligence

BC1: WONDERFUL WINES OF THE WORLD

Business Cases for Data Science

March 2022

Alice Vale R20181074
Eva Ferrer R20181110
Rafael Sequeira R20181128
Raquel Sousa R20181102

This report was written under the CRISP-DM guidelines.

Business Understanding

Wonderful Wines of the World (WWW) is a well-establish winery enterprise which currently sells in store, through catalog and by e-commerce, that has been on the market for 7 years. For the past 4 years, the company has kept a transactional database; However, *WWW* has not extracted any relevant intel regarding their customers' buying patterns or the monetary value they generate. Therefore, missing on a great opportunity to conceive targeted marketing campaigns and loosing revenue on poor decision-making approaches. *WWW* aims to change their ongoing *modus operandi* by focusing on performing customer profiling, resulting in clearly identifying clusters of clients based on engagement and behavioral similarities.

Data Understanding

The data provided by *WWW* to conduct the costumer segmentation analysis comprised information regarding the behavior and the demographical characteristics of 10 thousand customers of its active database, over the past 18 months. It is crucial to mention that the dataset was already well-conditioned, however, the usual steps of data exploration and transformation were still employed.

Data Preparation

The first stage was to evaluate the basic statistics and variables' histograms to get a comprehensive understanding of the data's raw comportment. No unexpected patterns were unveiled by this assessment. A duplicated *customer ID* was found and one of the records removed, so to guarantee data truthfulness. Following, no missing values were detected in any of the variables. To inspect for outliers, the boxplots of the variables were explored, though no observations were removed as it was concluded it was more beneficial to the study to represent all the original behaviors rather than excluding potentially relevant information.

Concerning incoherencies, the summary of all the wine buying percentages should be between 99 and 101. This interval was considered to account for possible case decimal rounding, that might lead to events where this sum is not precisely equal to 100. Moreover, according to the dataset's information, no customer should be below or above the age of 18 or 78, respectively. In addition, there should not be any clients in the dataset that have not made a purchase in the last 18 months and the last purchase made by a customer cannot be older than the acquisition of such customer. Finally, a person cannot be a customer longer than when the database was set up, as there was no client registration before that time. There were no observations under these five potential incoherencies.

In what concerns to feature engineering, three new variables were created. *Mont_income* represents the average monthly income of each customer. *Neg_LTV* is a binary variable defining whether a client has a negative *LTV,* signalized by 1, or a positive *LTV*, signalized by 0. *Avg_Purchase* illustrates the average amount spent by each customer when they make a purchase. All the variables were scaled using the *Min-Max* scaler, to account for any eventual data distribution skewness. The relationship between *Neg_LTV* and other variables was further investigated so to analyze possible customer behaviour patterns.

To create the different customer segments two variable perspectives were created: *Value* and *Buying Behavior*. The earliest gathers the variables that symbolize the value of the client to the enterprise, for instance *Recency*, *LTV* and *Dayswus*. The latter contains the ones that translate the clients' purchasing behavior, such as *Perdeal* and the various wine buying percentages. The correlations were assessed so to guarantee that no redundant information was included in the perspectives, which could affect the design of the clusters.

Modeling

To proceed with the segmentation of the customers based on the data available, two different clustering techniques were assessed so to reach the optimal solution. For each perspective both *hierarchical* and *k-means* clustering techniques were used since data is numeric and these are distance-based methods. Each application will be described in more detail below.

The critical decisions when using hierarchical clustering techniques are to first decide the linkage method to be applied and then the optimal number of clusters to construct. This is done through the analysis of the R^2 and dendrogram plots respectively. For both perspectives, it was extremely clear that the linkage method to use was *Ward's*, since it detained the higher values of the coefficient of determination, with 4 clusters to be constructed, conclusion derived from the analysis of the dendrogram.

To use k-means, the only requirement is to provide the number of clusters a priori which is done through the analysis of the inertia and silhouette plots, attempting to reach the lowest error possible. The conclusion was not straightforward, so the options of 3 and 4 clusters were applied to both perspectives. In the end, the *value* segment thrived with 3 clusters and the *behavior* with 4.

In order to compose a final decision about the method to be applied to each perspective, four metrics were evaluated, so to support on the assessment of the quality of the clustering solutions obtained. These were the R^2, the *Calinski-Harabasz*, the *Silhoutte* and the *Davies-bouldin* scores. In a reliable clustering solution, the first three should have higher values whilst the latter lower values. From this analysis, it became clear that the best solutions were the use of k-means for the *behavior* perspective and the use of hierarchical clustering for the *value* perspective, both with 4 final clusters.

METRICS	***Value perspective***		***Behavior Perspective***	
	Hierarchical	**K-means**	**Hierarchical**	**K-means**
R^2	0.6078	0.5415	0.5349	0.5828
Davies-bouldin	0.4	0.4351	0.5457	0.5435
Silhoutte score	0.7263	0.704	0.6435	0.684
Calinski-Harabasz	63419.1346	47390.5012	35115.1413	47175.515

Value perspective		
Clusters	One_Timers	Customers that have not visited the store in the longest time and have the lowest life-time values and income. Can be seen as individuals who visit the store occasionally.
	Senior_Customers	The customers who have been registered with the store for the longest, however register low life-time value and income.
	Prospect_Customers	The most recent clients but that reveal extreme potential since show both high income and life-time values.
	High_Valuable	The most valuable and loyal customers, registering small recency values and the highest life-time value and income, along with seniority within the company.

Behavior perspective		
Clusters	DryRed	The clients who purchase the most of Dry Red wine and lowest of all other categories.
	Open_Minded	Customers who show a very versatile behavior, buying a lot in any category except for Dry Red wine. To note that these are the ones who buy the most in the Exotic category and have a significant interest in discounts.
	DryWhiteWine	Gathers clients who show a preference for Dry White wine and do not engross in promotional deals.
	Discount_Red	Customers that show an interest in dry wines, being the most engaging with deals.

The merging of perspectives was imperative to get a clearer picture of the overall behavior of the customers. From that initial merge, one cluster remained empty, resulting in 15 different clusters. Our rule of thumb was to then merge, based on proximity, all clusters which registered less than 500 observations, so to avoid niche behaviors that are not representative of a common behavior. In the end of this step, the result were 10 clusters. After closely analyzing their construction, some appeared to be particularly similar and prone to grouping. To ensure the validity of this conclusion, hierarchical clustering was performed. The dendrogram legitimated the merging of 2 pairs of clusters, allowing then for the final solution to be composed of 8 segments.

Evaluation

The final solutions were given updated names that considered both initial segmentation perspectives, creating a single label which identifies each cluster. Thus, the available variables' means were compared for each cluster to assess distinctive conducts, resulting in the following eight groups of customers and their preliminary interpretations:

Premium – Customers highly monetarily profitable for the company as well as the most frequent ones, naturally showing the highest value for LTV. Demographically characterized as the eldest, they have the highest Income and the lowest percentage of purchases bought on discount. It is worth mentioning their apparent, yet moderate interest for Dry wines.

Gold – These customers show great room for improvement, since much likely the Premium customers, they have high values for Monetary, Frequency and Income, resulting in the second highest mean value for LTV. In terms of their preferences, there is a clear interest for White Dry Wines.

Strong Dry Red – The most latent characteristic of these customers is their preference for Dry Red wines. In addition, they stand out for their highest educational level and for being recent client acquisitions. They are labeled as "Strong" due to the great potential they hold, since it is believed that with the right marketing campaigns, they can increase their sales significantly.

Weak Dry Red – These customers have the strongest preference for Dry Red wines. Nonetheless, their engagement indicators, including the LTV value, show considerably less potential than the previous cluster. It is worth mentioning that these cluster has the lowest consumption levels of Dry White, and any sweet wines.

Loyal – Customers that have been purchasing from us for the longest. Since their Recency values are fairly low, that indicates that such clients are still faithful to our brand. Consequently, these clients can easily increase their sales, and perhaps venture into trying new wines. They are not uniquely characterized by a strong wine preference, although they tend to mainly buy Dry Red wines.

Sporadic – These customers are the most challenging ones in what concerns to identifying patterns and designing successful marketing campaigns. There has been a while since they last purchased, they show low values for Monetary, yet still average to low values of Frequency. Thus, they were labeled as casual clients, that are not necessarily loyal to the company, but act accordingly to convenience. Their main purchases are White Dry wines, and, in comparison to the remaining clusters, they show a slightly preference for Dessert and Sweet White wines.

Weak Versatile – This cluster is characterized by the youngest demographic which explains the lowest values for education and Income. Their value for the business is currently extremely low, due to their absent frequency and poor value purchases, resulting in the second lowest LTV. On the other hand, they show the highest preferences for Sweet Red and White, Dessert and Exotic wines. Hence, being considered versatile since they do not strict their purchases to a single wine category. It is important to highlight that they also show a great interest for items on sale, possibly explaining their unfortunate values for LTV. On a different note, they are the cluster which buys mainly from the Web, that is plausible due to their age group.

Least Profitable – In terms of LTV, this cluster has its lowest value, while having the highest value for percentage of items bough on discount. Their online presence is significant and worth considering, since, following the previous cluster, they are the second set of customers with the highest website visits and purchases. These customers prefer Dry Red and White wines.

Cluster	***Number of Individuals***
Premium	1456
Gold	1399
Strong Dry Red	999
Weak Dry Red	761
Loyal	1120
Sporadic	471
Weak Versatile	1553
Least Profitable	2241

Followed by this analysis, several visualization techniques were applied to capture the similarities and dissimilarities between clusters, further understanding the profile of the finding results. Those can be consulted in the annexes of this report.

To conclude the Evaluation stage, a Leverage Analysis was performed. This analysis consists in calculating the value to the company of one individual in a given cluster, when compared to other individuals of any other clusters. Taking into account not only the distribution of the total value, in terms of monetary revenue, per cluster; but also, the percentage of individuals in each cluster. Therefore, this proved our previous conclusions based on the percentage of customers with Negative LTV. In fact, the clusters with the highest leverage were the *Premium, Gold* and *Strong Dry Red*; While the worst leverage results are attributed to *Weak Versatile, Least Profitable* and *Sporadic*.

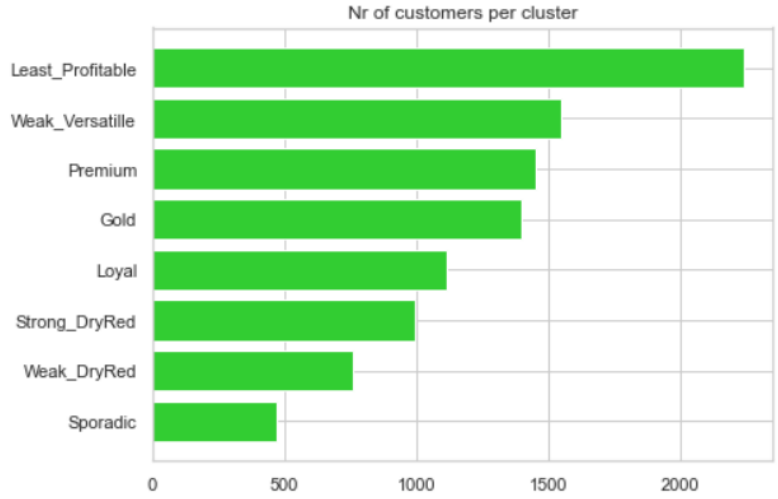
Deployment

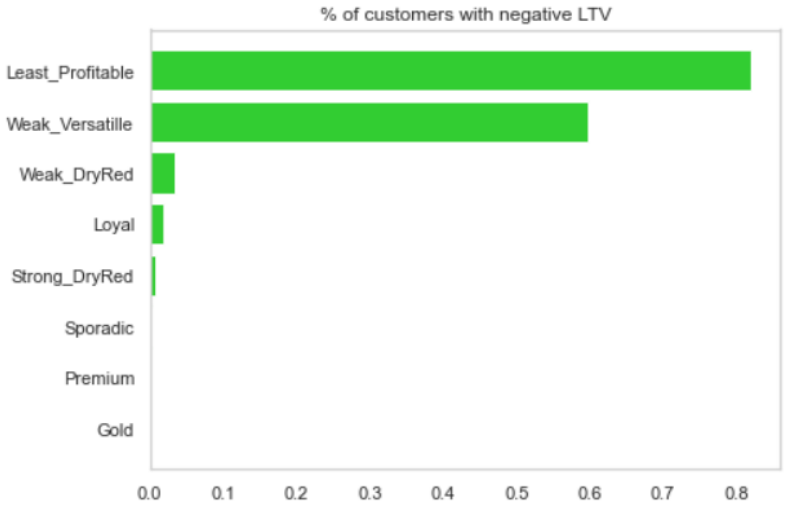
The marketing campaigns were design to be applied specifically and exclusively to its respective segment of customers. Nevertheless, we are aware of the mismatch of theorical models and its practical implementation in the real world. For that reason, in an initial stage of testing the applicability and success of these business-related strategies, *Datalin* suggests beginning the implementation with the *Strong Dry Red* cluster. For two motives: it is a comparatively smaller cluster in the current clients' asset; along with being a potential segment very receptive to a targeted marketing campaign, with room for growing into frequent and loyal customers of *WWW*.

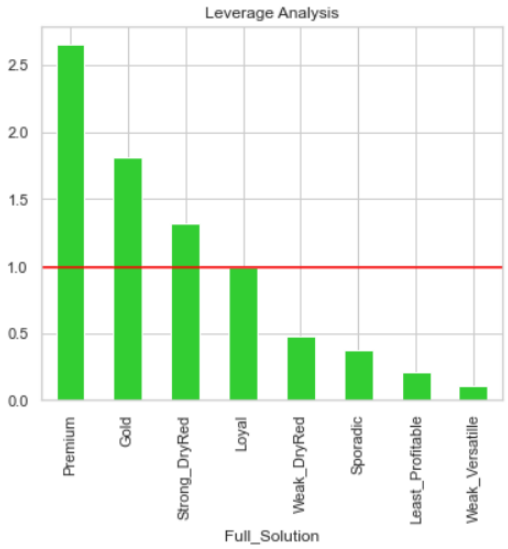
Cluster	***Marketing Strategy***
Premium	An exclusive in paper magazine, promoting high quality wines, send every two months. To encourage trying new items, every month there is a featured selected wine that these customers can savor free of charge in store while enjoying some delicacies, such as tapas. In each store, there is an annual gathering for wine tasting and socializing, with a specialist: such as an oenologist or a sommelier.
Gold	For any 100 $ spent, they receive a collectable premium glass of wine. The collection can be expanded, namely for seasonal celebrations, like Christmas.
Strong Dry Red	If they enroll in the loyalty program, they will be offered a monthly discount in the total of 20$ when purchasing at least 100$ worth of items.
Weak Dry Red	For every purchase worth at least 60$, the customer can choose from a list of wine accessories, featuring a wide range of items in different shapes and colors.
Loyal	For every purchase worth at least 60$, the customer gets a 30% discount in Exotic wines.
Sporadic	If they enroll in the loyalty program, they will get a one-time 50$ credit to spent on purchases worth 100$ or higher. Subsequently, they get access to a seasonal discount of 10%-30% in specific categories of wines to be defined.
Weak Versatile	If they subscribe to the newsletter, they get a monthly 15% coupon to spend in any category of wines.
Least Profitable	If they subscribe to the newsletter, they get a monthly 5% coupon to spend in any category of wines.

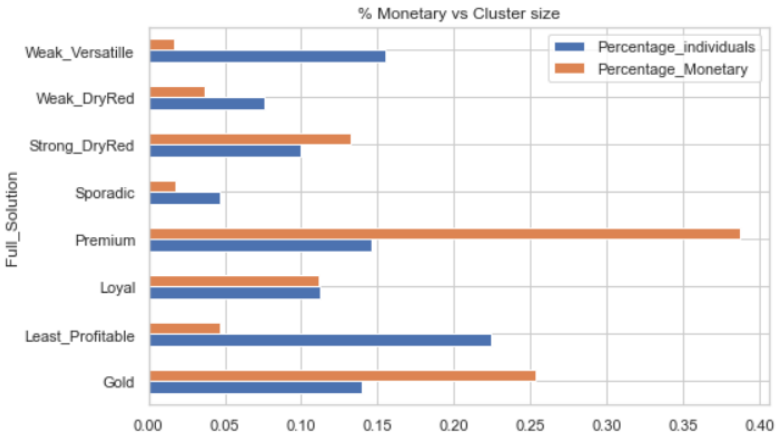
ANNEXES

Data Visualizations

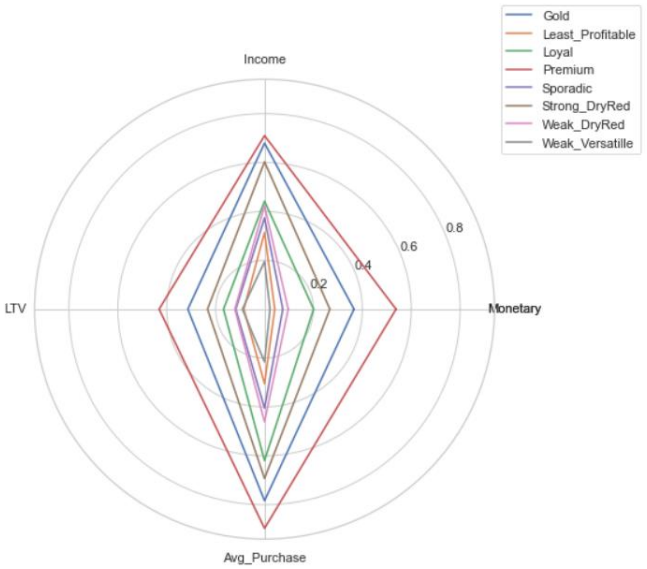


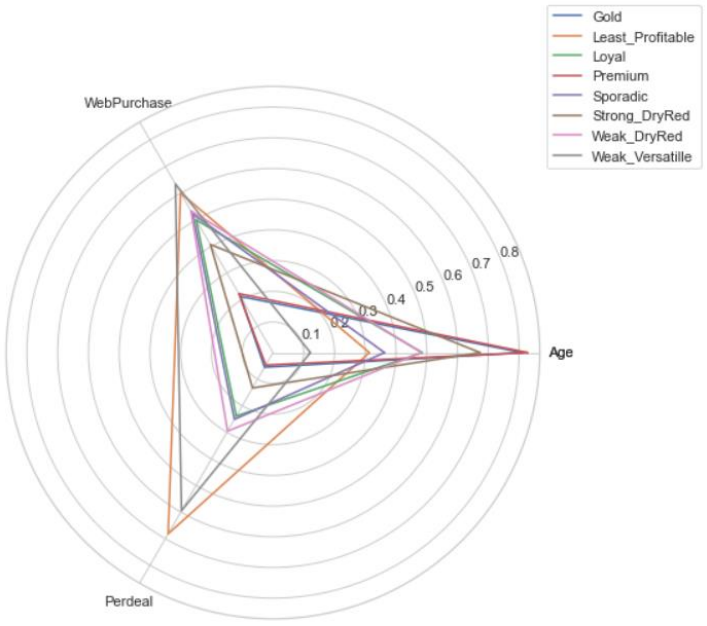


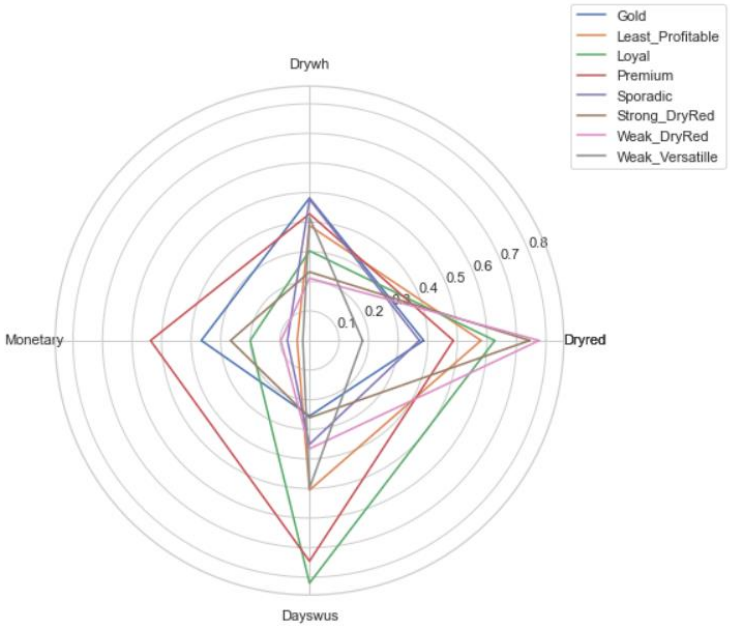


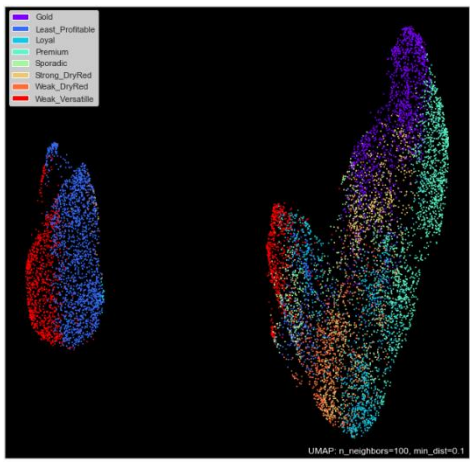


Radar Charts to evaluate the relationship between important discriminatory variables and the finding clusters









UMAP Map – Multidimensional Visualization