BUSINESS CASES WITH DATA SCIENCE

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

Business Case 1 - Wonderful Wines of the World

Group AC

Gabriel Cardoso, number: m20201027

João Lucas, number: m20200758

João Chaves, number: m20200627

Luís Almeida, number: m20200666

February 2021

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

INDEX

1. INTRODUCTION

Finding new customers is vital in every industry. The process for finding new customers begins by learning as much as possible from the existing customers. Understanding current customers allow organizations to identify groups of customers that have different product interests, different market participation, or different response to marketing efforts.

Market segmentation, the process of identifying customers' groups, makes use of geographic, demographic, psychographic, and behavioural 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. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Wonderful Wines of the World is a company which wants to delight their customers with an immersive wine experience. As of recently, it felt the need to better understand their customer base and provide unique promotions to the possible customer segments that they might possess.

As of now, WWW promotes their products across all customers in the same way, by providing them virtually the same channels to seek the company's products and has a strategy founded on giving the customer a catalog with all of the products, providing no tailored promotion.

As such, WWW aims at understanding the profile of their customers and devise promotions to target their market segments in a sharper manner so that it can provide the customer a more valuable experience.

To conclude, the main goal of this project is to provide meaningful information about the behaviors of WWW's customers and construct a customer profile so that it can deploy specific actionable strategies, which will also be explored, to target their different customer segments.

2.2. 2.2.BUSINESS OBJECTIVES

We would like to understand:

- Which characteristics best distinguish the customers;
- Which and how many relevant customer segments they have in the database;
- How to reach new and existing customers from each segment (marketing mix) and which ones should be prioritized;
- Increase the Life Time Value (LTV) of our clients, with special concern for the clients with low or even negative LTV.

2.3. BUSINESS SUCCESS CRITERIA

The main criterias to measure the success of our project will be set in accordance with the Business Objectives:

- Have well defined and separated clusters that can be easily distinguish and described;
- Be able withdraw insights from each cluster that allow to draw specific conclusions about marketing approaches to each cluster;
- Measure and Increase the overall LTV after applying the Marketing Mix;
- Eradicate clients with a negative LTV, and significantly increase the LTV of clients near to zero.

2.4. SITUATION ASSESSMENT

For this project we have a dataset containing information on all customers that made a purchase in the past 18 months (10000) and 29 features.

2.5. DETERMINE DATA MINING GOALS

These goals success will be determined by the quality of the clusters we're going to create, which can be accessed by a F1 score metric from an evaluation made by a predictive model. The score should be higher than .85 in order to trust the clusters created.

3. PREDICTIVE ANALYTICS PROCESS

Describe only the major steps involved in the process. Do not replicate what is already described in the Notebook. If necessary, reference the reader to the Notebook.

3.1. DATA UNDERSTANDING

3.1.1. Excel Exploration

In order to get useful insights into the problem ahead, first order of business demands that we do some exploratory analysis of the dataset provided.

As advised, we first opted to look at the excel file provided and realized that some columns, namely “Rand” and “Custid” didn’t provide any information to explain consumer behaviour, being the first one a column with random numbers which have no relation whatsoever to the remaining dataset and the later was an automatic integer assignment when the customer was added to the dataset (we use the “Custid” column as indices in the final solution for administrative purposes).

It is also noteworthy to point out that the last row in the dataset was an aggregative column, containing the mean of the values in the column. As such, and since this information and much more is easily obtained using python functions, we decided to clear this row and preserve data integrity.

As for the analysis of the excel file, what was afore described was the majority of the extent of our preliminary work.

3.1.2. Python Exploration

After transporting the data into a python script, we continued our analysis to gather even more detailed insights, for this we used common tools like the “.describe()”, “.info()” and “.dtypes()” to spot possible anomalies and trends in the data. We concluded that the dataset was fairly clean since there were no apparent missing values and the data types were pretty “on point” to what was expected out of the variables and their description.

Later we decided to use pandas profiling to get a more thorough analysis and spot possible problems to be corrected in the data pre-processing step. We came to the realization that only the variable “Recency” had very odd value distributions, which indicated that there were outliers that needed to be removed.

From the pandas profiling we also saw that there were quite a few variables with high correlations which could be addressed. Note: later on, we do a more comprehensive exploration of the correlations and explore various methods of removing features due to high correlation factor.

3.1.3. Metadata Exploration

After gathering insights, and with a clearer background of the problem, we decided to get another look at the metadata and correlate the behaviour in data with the description and knowledge that could be gained from the features at our disposal.

In particular, there was a feature that did not provide a clear answer to the problem at hand of defining customer segments and their behaviour. This column gave information, respectively, about customer complaints ("Complain" - this was a binary feature with a distribution of 1:100, being the majority class an absence of complaint) which was not useful since this had no relevancy in a pool of 10000 records.

Therefore, we opted to cut this column from the clustering for the reasons afore mentioned.

3.2. DATA PREPARATION

Powered by the initial analysis, we decided to look for typical possible anomalies such as missing data, wrong datatypes, inconsistencies in data, duplicated data, and others but there were no major problems on this department due to a fairly clean data.

To what concerns data engineering, we decided that it would be prudent to create some columns out of the existing ones to better summarize information, being useful since it is a way to decrease variables without compromising integrity. For example, we decided to merge "Kidhome" and "Teenhome" features into the column "Children" to gather information about customers with kids, not discriminating by age.

In terms of outlier detection, supported by box plotting in order to forecast the main steps of approach ahead and identify the most critical features to consider, we attempted many different methods of outlier removal to see the one that best preserved data integrity without removing too much data.

We finally settled on the Isolation Forest Method which removed 10% of our data. Despite it being a high percentage, it offered a better distinction between the values considered normal using the IQR method and the visual representation of the data.

To be more precise, the IQR method considered outliers values that were not really very far from the dispersion trend. As such, the Isolation Forest was a middle ground between these methods.

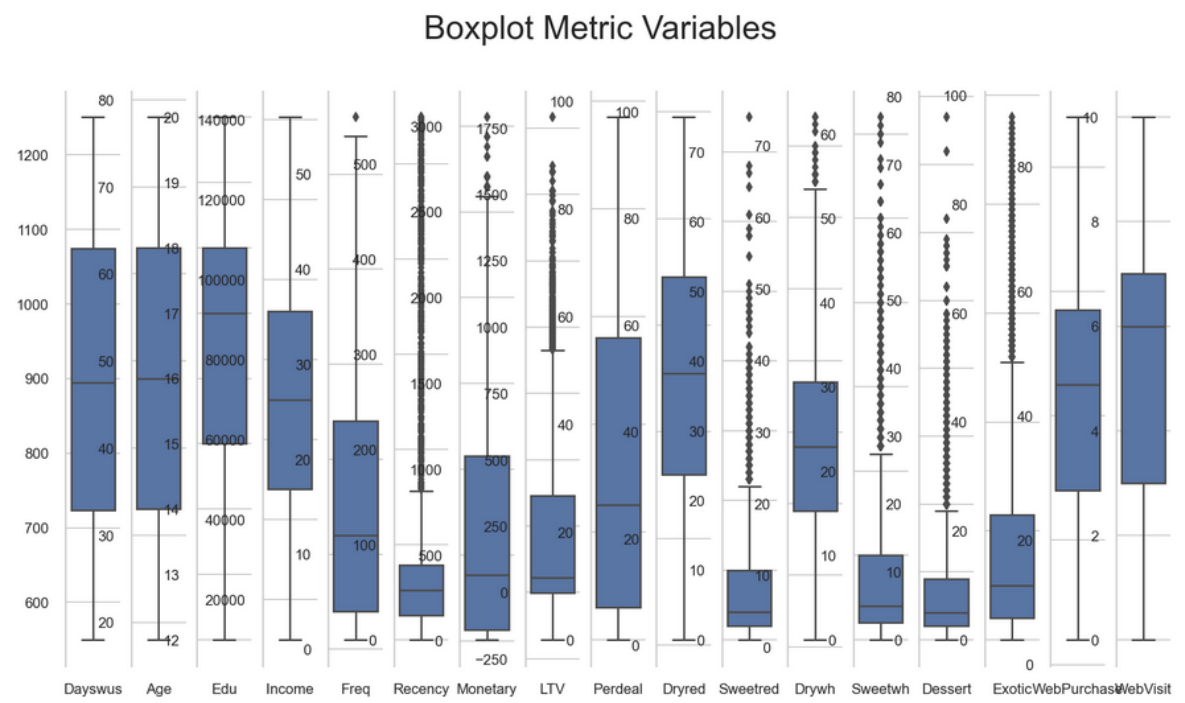

Fig.3.1. Boxplot of the data before outlier removal

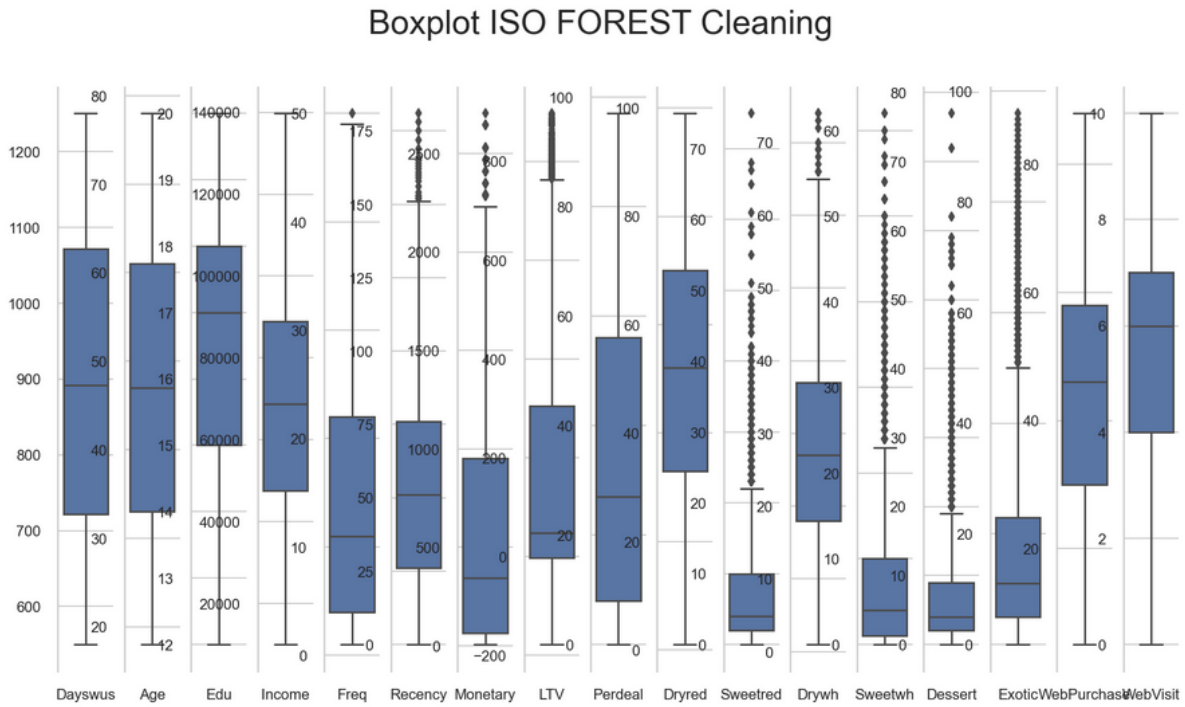

Fig.3.2 Boxplot of the data after outlier removal using Isolation Forest

In order to decide which features provided good information for later use in the model, we opted to identify high correlations which trimmed down even more our pool of features upon removal.

As the pandas profiling option already gave us a pretty good insight into possible correlations, we decided to explore the Phik even in more detail and decided to remove features which had a value close to 0.8 and above of correlation with others. In order to discern, which of the two to let go, our logic was to remove the one which had less diversity and provided less information gain to our clustering model. As such, we removed the feature with the highest absolute correlation between the two.

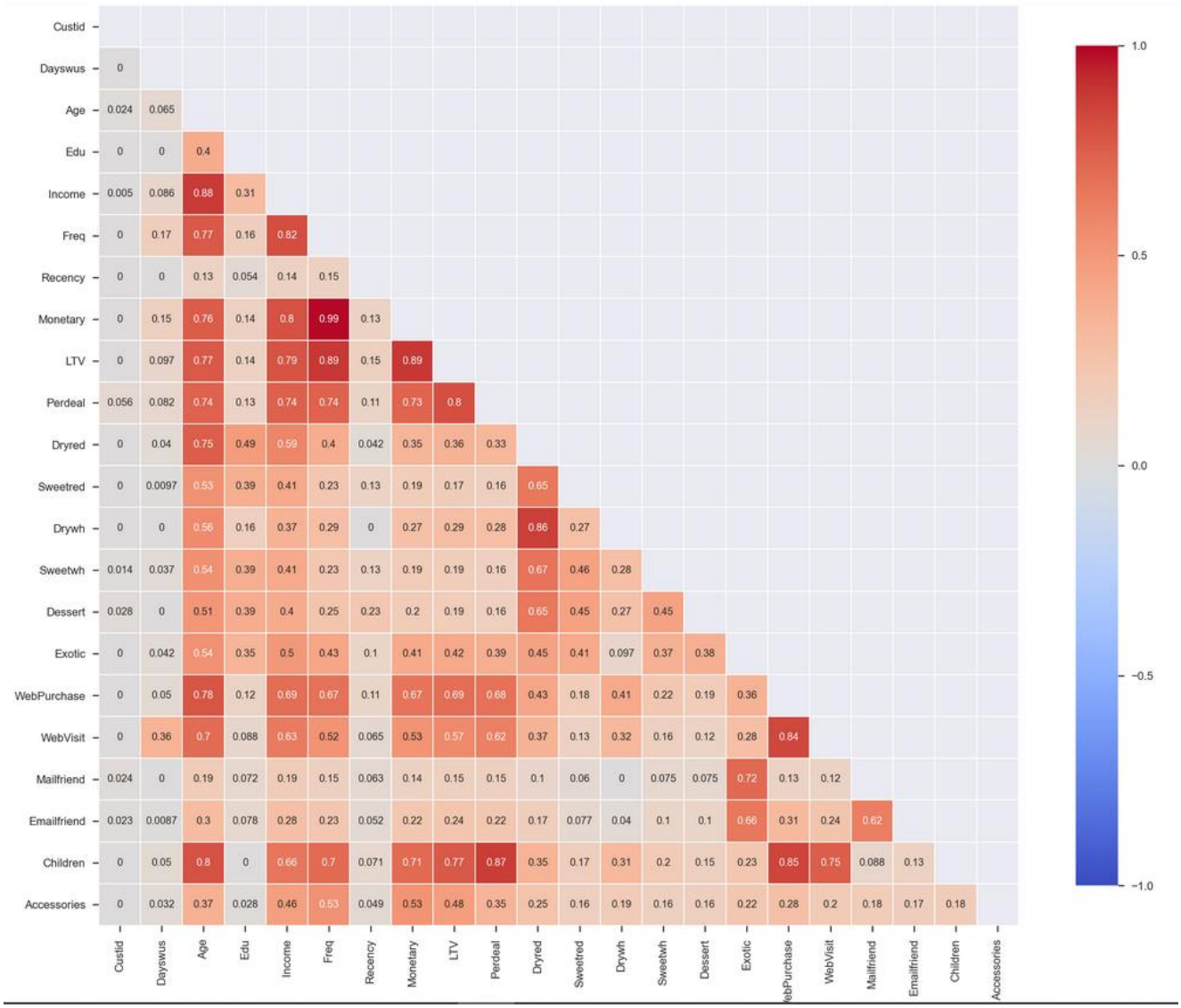

Fig.3.3. Phik correlation Matrix

3.3. MODELING

For the modeling part, the methodology chosen was to apply a clusterization in order to properly label each customer within a group where we can take a specific approach that will increase the value of the purchase per customer.

To apply a clusterization we selected two different models:

- K-Means, which is a simple, fast and extremely effective model but has the upside of only be appliable to numerical data and has the need of a preselection of the cluster number;
- K-Prototypes, virtually like K-Means with the plus of being a model that works on numerical and categorical data.

3.3.1. K-Elbow Plot

Firstly we had to determine the number of clusters to be used and for that purpose a K-Elbow plot of the inertia associated and thus select the best number of clusters. As result, it was observed that 4 clusters was the value to use.

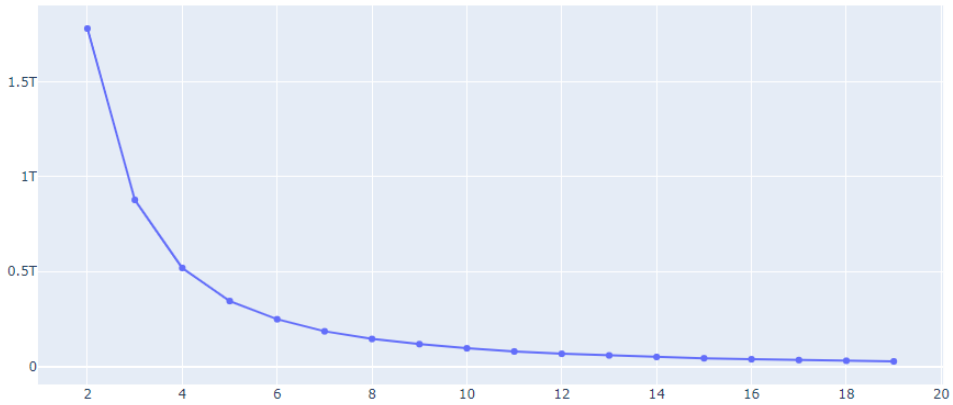

Fig. 3.4 K-Elbow Inertia Plot

3.3.2. K-Means

Next step was to apply the k-means method to the cleaned data, plot the results to visually check the dispersion and evaluate the final result.

To visualize the clusters, we've applied the UMAP embedding technique to the k-means data output. UMAP seeks to accurately represent local structure and better incorporate global structure. This technique has showed better performance to preserve both the local and global structure and it has proven to meet our needs.

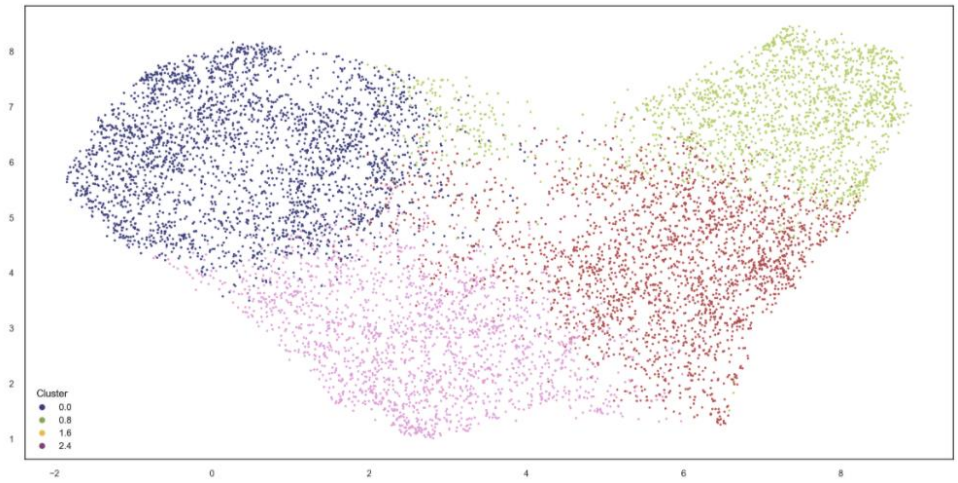

Fig.3.5. UMAP K-means cluster visualization

It can be observed that , although not perfectly, the clusters are very well defined and even without evaluate the results we can conclude that this segmentation was a success.

3.4. EVALUATION

Having our features segmented into the different clusters and labelled, we had to evaluate the model and consequent results. To do so we used 3 different methods and started by calculating the intercluster distance to understand if the centers of our clusters were properly separated.

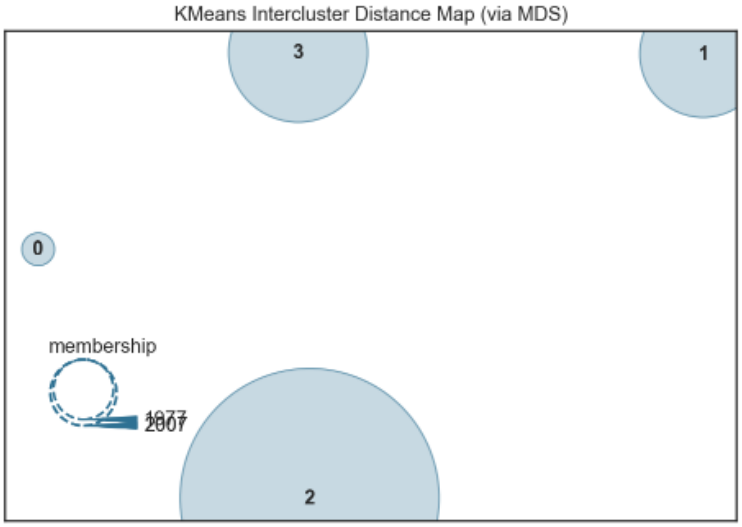

Fig.3.6. K-Means Intercluster Distance Map

As observable our cluster centers are perfectly separated, indicating a good data segmentation.

Then, to validate the measure our cluster's classification quality we applied LGBM (light gradient boosting machine) which treats the clusters as labels and builds a classification model. If the clusters are of high quality, the classification model will be able to predict them with high accuracy.

The F1 Score associated with our model was of .667, which is not very high, but since clusters are perfectly separated from each other and easily distinguishable, we don't discard K-means.

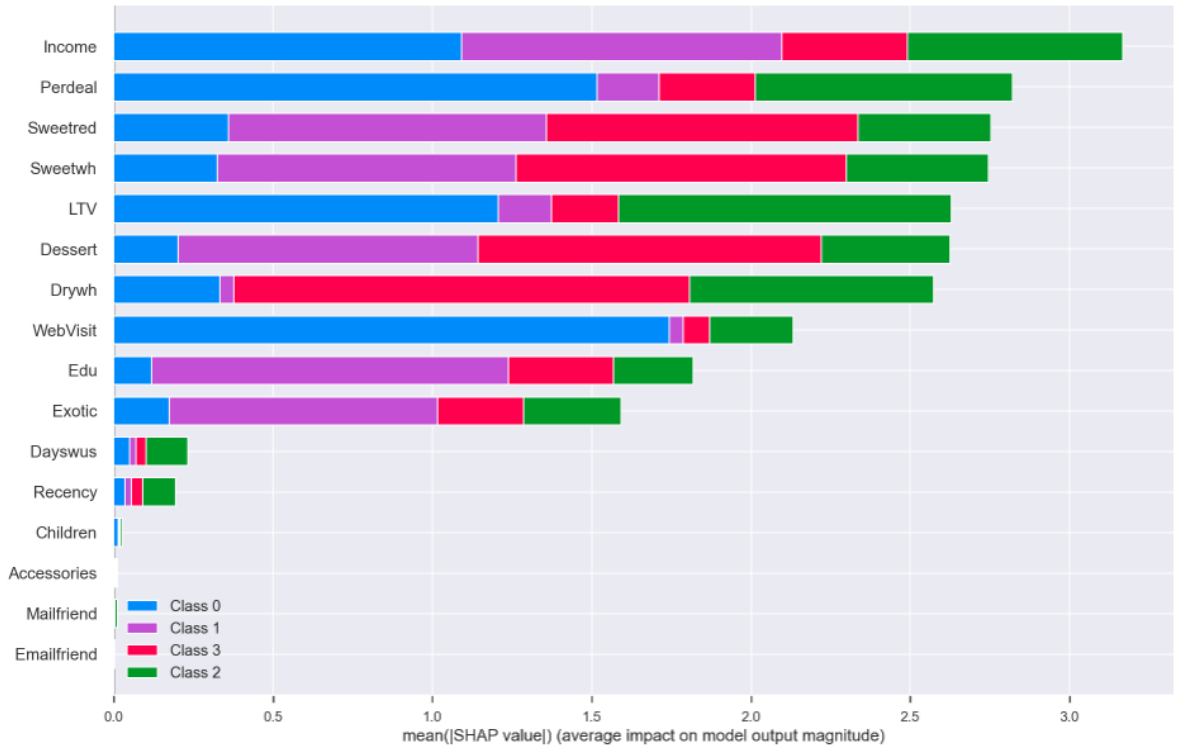

Fig.3.7. K-Means Summary plot

This summary plot lets us know how much of each variable's observations are distributed by the clusters and have a broader idea of the importance each variable has. The top 10 variables are the most important for this clusterization and are very balanced between each other.

Comparing k-means to k-prototypes, we confirm that the F1 score is close on both methods, which implies that k-means and k-prototypes have produced similar clusters that are easily distinguishable. Yet, the classification score on K-means is slightly superior and since k-means is a simpler and faster algorithm we choose it and proceed to the description and analysis of the clusters created.

4. RESULTS EVALUATION

Our K Means algorithm produced 4 distinct Clusters with very distinctive characteristics. We separated this characteristic in Demographic: Age, Number of Children living at home, Income, Life Time Value and percentage of purchases with discount; RFM: Recency, Frequency and Monetary; Chanel of Communication: Percentage of purchases online, Number of website visits, Mail or Email friendly (if responds to this kind of promotions); Accessories bought: Small wine rack, large wine rack, wine cellar humidifier, cork extractor, wine bucket; Type of Wines Consumed: Dry Red, Sweet red, Dry White, Sweet white, Dessert and Exotic.

In this chapter of the work, we not only describe each cluster individually and in comparison with each others, but we also provide suggestions for marketing approaches that work with the model and are specific for each of the produced clusters.

Cluster 0 – This cluster represents the oldest of our clients with an average of 64 years of age. They have the highest incomes at an average of 95542$ annually, and are also the ones with the highest Life Time Value for the company (410$ on average). Only 40% of them has children at home and the percentage of products bought at discount is very low and not very significant (7%).

On average they made 24 purchases and spent 1126$ in the last 18 months, making them both the ones who buy most frequently and spent the most. They are characterized by a small percentage of both web purchases and web visits, with only 24% of their purchases online and an average of 3 visits to the website per client, making them the cluster with the lowest online interaction; these customers are also quite unresponsive to email or mail communications. Their wine tastes aren't very diversified; their favourite wines are Dry Reds (45%) and Dry Whites (32%). Finally, the accessories these customers bought the most were the wine cellar humidifier (14%), and the silver plate cork (12%).

Useful Insights: Our most valuable group clients both on Lifetime Value as well as the money spent in the last 18 months. They are less prone to online services, promotions or to taste different and new kinds of wines. However, they have high incomes, make regular purchases, and like to buy expensive accessories.

Marketing Mix: The principal objective of this cluster is to keep the clients happy and satisfied since they are our most important clients, as so we suggest some marketing approaches: Product, even though they consume mostly dry reds and whites, 9% of them bought exotic wines, as so it would be a good approach to offer samples of more exotic wines to this group of costumers; Price is not a problem for this costumers, as so instead of Promotions or discounts a fidelity program based on money spent would be a good idea to keep them engaged and satisfied with the company; Placement, this clients are less prone to engage with the company via website, as so, on store experience as well as all kind of promotions or fidelity programs are key to keep them satisfied.

Cluster 1 – This group of customers are the youngest with an average age of 28 years. They have the lowest income at an average of 39580$ annually, and have a low Life Time Value at just 11$, most of them has children living with them and take advantage of promotions and discounts to make most of their purchases, 53% of their purchases were made on discount.

In the last 18 months, on average each client of this cluster did only 5 purchases and spent 117$, representing the clusters with the lowest RFM. They are characterized by making more than half of their purchases online (55%) and visiting the website more than 6 times in the last 18 months, they are also highly responsive to mail (53%) and email (36%) communications. Their wine tastes are very diversified and don't have a clear preference, however Exotics (35,5%) and Dry Whites (29,5%) are their favourites. Regarding accessories, they don't buy many, with the exception of the Small Wine Rack which 30% of them bought.

Useful Insights: Young group of clients, with low incomes as well as low Life Time Value. However, they are very prone to promotions, online communication and to experiment new wines.

Marketing Mix: This group of clients, however not representing a big income for the company now, are young people starting their life´s as wine lovers only now, as so, is important for us to keep them and develop the relation with the company to make them blossom on valuable and loyal customers. This is our marketing approach: Product, a diverse and big menu of different kinds of wines would be very good to make them satisfied and engaged; Price and Promotions are a key factor to keep this group of customers loyal, as so regular discounts specific to these customers, on sweet and exotic wines, are highly recommended. Finally, placement is also a key factor in this cluster, a constant communication via email or mail, in a format of newsletter or catalog for example, as well as a good website is essential to keep them engaged.

Cluster 2 – The average age of this group of customers is 37 years making them a young group of costumers. The income is higher than in cluster 1 at an average of 52489$ annually but have a lower Life Time Value at only 6$. They are the cluster with the highest percentage of children living at home, 96% of them have a kid or teen living with them and most of their purchases were also made on discount like in cluster 1.

In the last 18 months they made on average 6 purchases worth of 182$, representing a RFM just a bit higher than the one in cluster 1. They make more than half of their purchases online (54%) and visited the website more than 6 times in the last 18 months, they are not very responsive to mail (13%) and email (8%) promotions, although significantly more responsive than cluster 0 and 3. Their wine taste is very cantered in Dry Red (54%) and Dry White (33%). Their accessories purchases are almost inexistant.

Useful Insights: Very similar to cluster 1, however with a higher income and with a more defined wine taste.

Marketing Mix: This group of clients is similar to cluster 1, however they have higher incomes and higher power of purchase, as so is in our best interest to increase their Life Time Value, make them engage more and increase their purchase frequency. Our purposed marketing approach: Product, all promotions and communication should focus on Dry Reds and Dry Whites; an offer of low-Price Dry wines as well as regular Promotions directed to this group of costumers are important to increase their frequency of purchase; the Placement of this promotions, discounts and offers should be very well presented in the website since this is the main way of communication with this group of clusters.

Cluster 3 – This group of costumers is represented by customers in their 50's with high income and a high Life Time Value at an average of 146$, this value is however very inferior to the Life Time Value of cluster 0. This cluster also has a high percentage of children living at home (92%) and make a

reasonable number of purchases in discount (27%) however much smaller when compared to cluster 1 and 2.

In the last 18 months they made on average 14 purchases worth 567$. They make almost half of their purchases online (47%) and visited the website 6 time in the last 18 months. They are not very responsive to either Mail (8%) or Email (3%) promotions. They have a very well-defined wine taste, with Dry Red representing 81% of their purchases. Finally, they are not very prone to buy accessories, except for the Large Wine Rack which 17% of the clients of this cluster bought.

Useful Insights: This cluster of clients have a high income and a high Lifetime Value, however there is space to growth and to make them even more valuable. Are more prone to online communication and sales, and their wine of choice is without a doubt the Dry Red. Accessories sales are a good bet to increase their Life Time Value.

Marketing Mix: This group of clients is very important to the company, not only already provide a good income, but also present a good opportunity to increase their Life Time Value. The marketing Mix we purpose: Product, high incidence in Dry Reds as well as in Accessories; Price, since this group has a high income but also likes promotions, should be a mix between the fidelity program created to cluster 0 and discounts made to cluster 1 and 2; Promotions and discounts, should focus especially in the accessories with the objective to increase the sales of this products to this group of clients; finally the Placement, this cluster of clients divides their purchases between online and in store, as so is important to have a good communication in both places.

5. DEPLOYMENT AND MAINTENANCE PLANS

In order to put the plan in action, there are several steps to consider but they all should follow one major guideline which is have a small and dedicated team to assess new customers and position them according to the already employed model developed above.

We consider that in a company which is just taking its first steps into data driven decision making, a Kimball methodology of slowly working our way up data integration into the business would be best. As such, we need more and steady investment into gathering reliable and meaningful data.

Since our clusters provide us with very distinct customer profiles, we advise the strategy and planification to follow very closely the customer profiles for the short term. Although, we envision that the company will attain more customers with possible different characteristics, a more thorough and periodic approach to seek out new customer profiles should not be discarded as the market is in permanent change and new players might appear.

6. CONCLUSIONS

Usually finding new customers is more expensive than improving the value of already existing ones. This are great news for the company, since with this research we were able to understand that the already existing data set of customers has plenty of opportunities to increase the Life Time Value of our already existing clients. Although there is a big group of clients very engaged and valuable to the company, there is an even bigger group of clients that can become more profitable in a near future.

As so, the company should invest in the online presence and make specific marketing campaigns directed to the different clusters; since they all have special behaviors and tastes. Increasing the frequency of purchase is key to increase each clients Life Time Value for the company.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

For the next project, we should take more time trying to improve the model since in this project, although the modelling quality was acceptable, we didn't spend the time to perfect the accuracy of the labels created and associated to each customer. In this case we believe that the predicting power of our model wasn't as high as we wanted because of the features used and the data scaling, if we took the time to improve the data quality and try different feature combinations, we would obtain even better results.

Also, in the next project we could try some different model combinations, which again, we didn't do because the quality of our model was acceptable to perform the and deliver the results we needed.