BUSINESS CASES WITH DATA SCIENCE

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

Business Case 1 – Customer Segmentation

Group AA

- Emil Ahmadov (m20201004)
- Doris Macean (m20200609)
- Doyun Shin (m20200565)
- Anastasiia Tagiltseva (m20200041)

March 1, 2021

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

INDEX

1. INTRODUCTION

TESTWonderful Wines of the World (WWW) is a company that aims to introduce its customers to well-made, unique, and interesting wines from all over the world – wines that would not normally travel far beyond their points of origin. The company sell wines to its 350,000 customers through catalogs (telephone), a web site, and ten small stores in major cities around the USA.

Our team of data scientists has been asked to perform an analysis on 10,000 customers that have made a purchase in the past 18 months to help the company better understand its cutomer base – how many customer segments exist in the database and what characteristics distinguish these segments. Primarily the company would like to know how they can reach new and existing customers from each segment, as well as which ones to prioritize.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Prior analysis by the company has led to the understanding that its customers are highly involved in wine, entertain frequently, and have sufficient money to indulge their passion for wine. Past marketing has been mass marketed and based on simple market reports, feedback from salespeople and intuition. Catalogues with several hundred selections are typically sent to all existing customers once every 6 weeks. The company will also occasionally offer wine accessories – wine racks, cork extractors, etc.

2.2. BUSINESS OBJECTIVES

Our client wishes to start differentiating customers, and developing more focused marketing programs and **achieve these goals**:

- understand current customers to identify groups of customers with shared patterns/characteristics.
- understand the differences between the different segments to make strategic choices in various campaigns.
- gain an insight into the behavior of their customers and to capitalize on improved customer relations.

2.3. BUSINESS SUCCESS CRITERIA

The 10,000 customers included in the dataset were sent the test promotion for the silver-plated cork extractor. Of these individuals, only 6.82% ended up purchasing this accessory. This conversion rate was obtained by random mass-marketing. Our proposal can be considered effective if we are able to obtain a higher conversion rate using our customer segmentation and tailored marketing approach. That is, the goal of our analysis is to obtain one or several segments that are more likely to respond to this promotion, essentially cutting promotion costs and improving the company’s profitability.

While the silver-plated cork extractor is simply one example of the benefits of an effective customer segmentation, the same logic applies to all products sold by WWW.

2.4. SITUATION ASSESSMENT

Our team had access to 10,000 randomly sampled customer data from WWW. Group AA consists of four data scientists. We used python and various modules for coding /analysis to study the given data.

Much of the information regarding WWW's history and data processing steps taken by the previous data scientist team was unavailable. In such uncertain cases, we avoided using definitive statements. or provided contingency plans for different assumptions.

The biggest limitation for analysis was the amount of data regarding WWW's customers. Currently, WWW only has data on different wine flavor preferences. As we hoped to translate the customers' preference to monetary values, we noticed there was no detailed data regarding which wines were bought on promotion and which were purchased on premium. The data may indicate that an individual purchased an equal amount of dry red and sweet white wine, but their actual commitment and interest for them are considerably different. Furthermore, there are many non-exotic wines that exceed the price of exotic wines. Hence, the wine preference data provided is not sufficient to measure the utility to consumers.

In this report, we use the term wine "flavor" and "type" interchangeably. These terms refer to the color and sweetness of wine, not whether the wine is an exotic type.

2.5. DATA MINING GOALS

In order to reach the company's business goals:

- Use multiple clustering methods to come up with different and well-defined clusters or segments.
- Form cluster profiling to find main characteristics of the clusters.
- Come up with beneficial models to predict customer behavior.

3. CUSTOMER SEGMENTATION PROCESS

3.1. DATA UNDERSTANDING

The dataset contains 10,000 rows and 30 columns. There are 21 columns with data type of float and the rest of the columns are stored as integer type. After checking data types and meanings of each column, we concluded that some of the columns have an incorrect data type. For example, some of the binary and integer variables are stored as floating numbers.

We found that there were no missing values in the dataset, as well as no duplicate rows. We then plotted the correlation matrix and found a strong correlation between some variables (i.e., income and age). These were considered in the data preparation.

Histograms of metric features were created, and it was found that some variables are uniformly distributed while others have skewed distributions. We also developed frequency plots for binary features to see how each category is distributed within the dataset. To check for outliers in data, we plotted boxplots of metric features.

3.2. Data preparation

We prepared the data by converting percentage values to monetary values. Percentage values for wine types, except exotic wines, and web purchasing were multiplied by monetary column to achieve this goal. In some cases, the percentage value can be high, but it may not translate to a large monetary value. For example, let's say that a specific customer uses website for purchases for 90 percent of time and this will generate 200 dollars. On the other hand, there might be a customer which uses website for 30 percent of his/her purchases, however this generates 300 dollars of financial value. This approach was used to eliminate this kind of misconception.

Secondly, we changed all the recency values which are higher than hundred, to the value of one hundred. We did that because all the customers having recency values greater than hundred can be considered as lost customers, people which are no longer buying wines from WWW. For example, this variable might have big importance for distinguishing customers when they have smaller values of recencies and might not have big importance as recency value gets larger.

Thirdly, we checked for the outlier values in each variable. By visually analyzing these graphs, we saw that there are a lot of values lying on the outside of the whiskers. We decided that these outliers are in fact special types of customers and kept all these values in our dataset.

3.3. Modeling

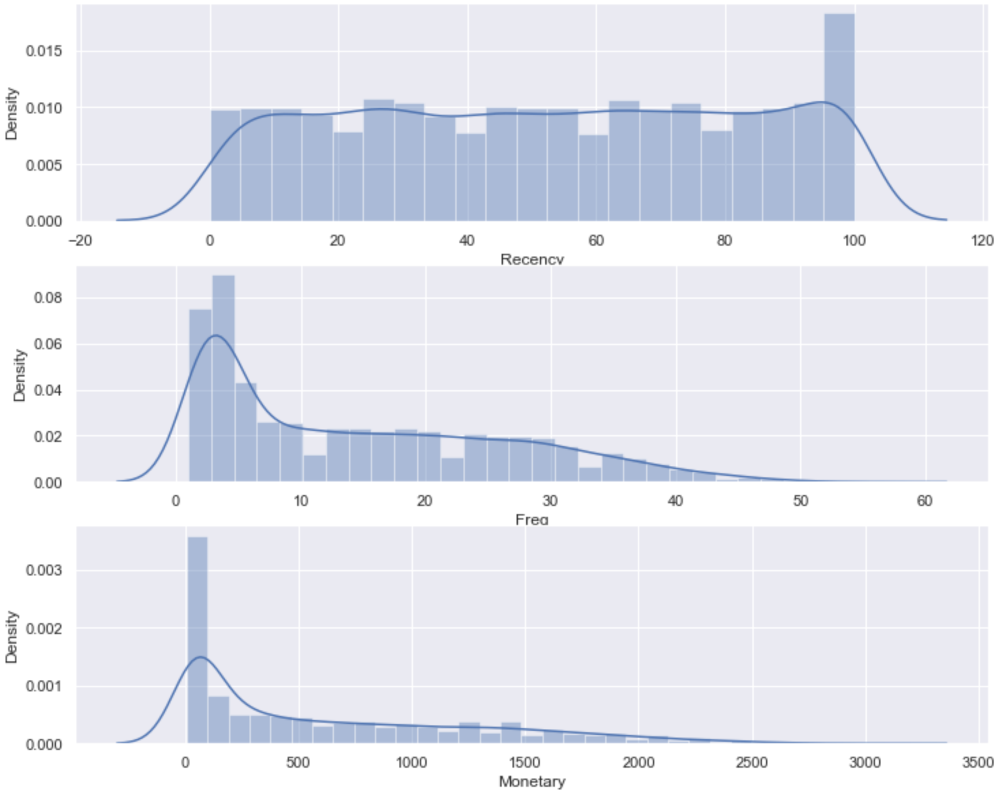
3.3.1. RFM Analysis

To have general insight of our current relationship with customers we used Recency, Frequency and Monetary (RFM) analysis. RFM analysis is a behavior based technique used to segment customers by examining their transaction history. RFM helps to identify customers who are more likely to respond to promotions by segmenting them into various categories[1].

The typical workflow for RFM analysis can be broadly divided into the following:

1. Collect transaction data.
2. Generate an RFM table from the raw data available.
3. Generate scores for recency, frequency, and monetary value, and use them to create the RFM score for each customer.
4. Use the recency, frequency, and monetary scores to define customer segments and design customized campaigns, promotions, offers and discounts to retain and reactivate customers.

We can see the distribution of our Recency, Frequency, and Monetary features in the plot below.



This plot provides us with some very interesting insights regarding how skewed our data is. The important thing to take note here is that we will be grouping these values into quantiles.

As shown in the workflow, the third step in RFM analysis is to generate the individual score for each metric and generate the RFM score.

We follow the below steps to create the score:

- Use quantiles to generate cut off points
- Create intervals based on the cut off points
- Use the intervals to assign score

Finally, we can group our customers by their RFM level.

RFM_Level	Description	Recency	Freq	Monetary	Count	
		mean	mean	mean		
Can't Loose Them	Bought recently, buy often and spend the most	42	27.2	1297.8	3720	37%
Champions	Made big purchases and often	42.5	13.4	492.2	1199	12%
Loyal	Spend good money, responsive to promotions	56.3	11.4	401.7	1209	12%
Needs Attention	Made some initial purchase but have not seen them since	72	3	47.2	654	7%
Potential	Bought more recently but not often	40.2	5.1	121.5	1246	12%
Promising	Above average recency, below average frequency and monetary	61.4	4.7	105.3	1113	11%
Require Activation	Lowest recency, frequency and monetary values	94.8	1.8	26.5	859	9%

We can see that a large percentage (about 60%) of our customers are in the top tier RFM levels. For the remaining 40%, we must take appropriate action to improve our relationship by understanding the client's needs. To achieve this goal, we will do clustering by perspectives

3.3.2. Clustering by Perspectives

We decided to segment our data set into two fields that may explain the behavior of our WWW's customers. We reviewed the demographics of our customer base, as demographics are likely to have some correlation to your ability and willingness to purchase wine. Simultaneously, we also explored the preferences of our customers, as we considered past behaviour to have the highest contributing factor for describing a customer's future purchases. Understanding the interests of our customers and segmenting them accordingly could help to ensure that WWW takes a targeted approach to marketing.

We first needed to define the optimal number of clusters for each section. We plotted the SSE for each section and obtained the elbow plots seen in the figure below. As the elbow is at around 2-4 cluster for both sections of the data, we decided to proceed with 3 and 4 clusters for each of the demographics section and the preferences section, respectively.

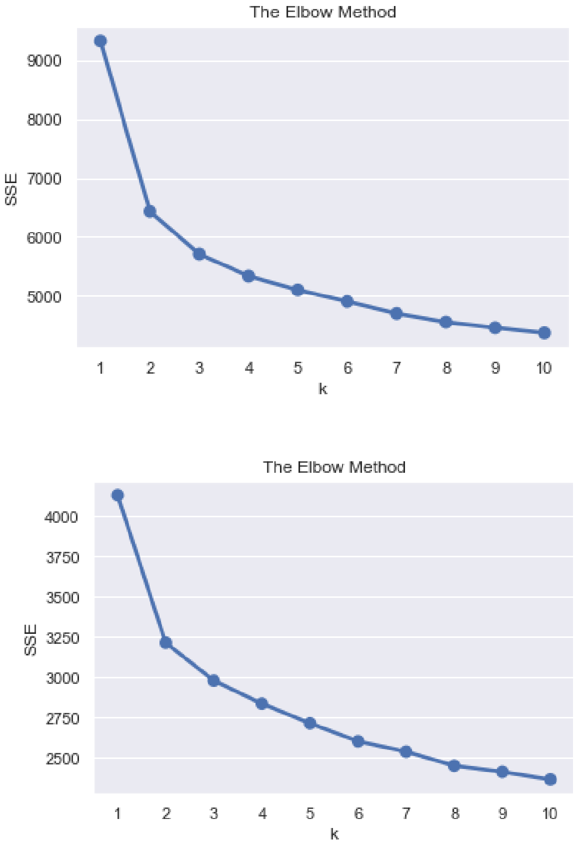


Figure 1: Plots of the sum of squared errors for both the customer demographics and preferences

The features we selected to include in our clustering took the form of numerical as well as binary. As such, we proceeded with a k prototype clustering for mixed data types on each of the two sections using the number of clusters chosen. We then merged the two clustering solutions and did a hierarchical clustering on top. Subsequently, we assessed the dendrogram and decided to proceed with a combined 4 clusters. The dendrogram can be seen in figure 2.

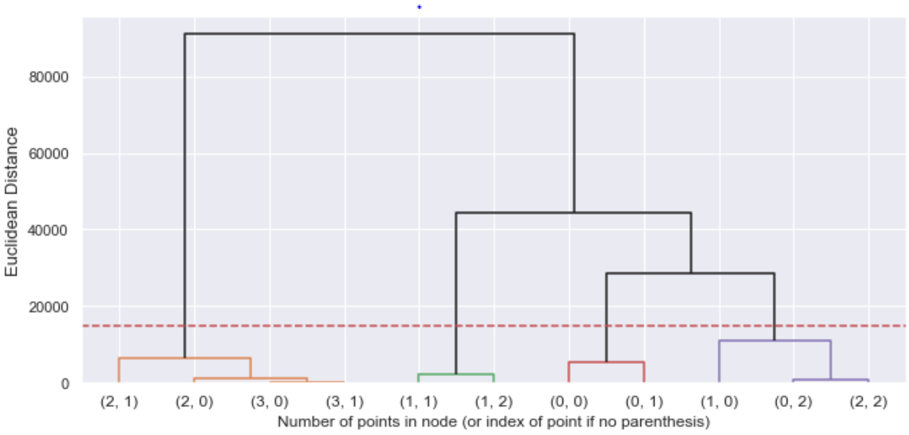


Figure 2: Hierarchical clustering

3.3.3. Leverage Analysis

In order to understand the behavior and relative value of our clusters, we conducted a leverage analysis. This essentially indicates the proportion of monetary spend of the clusters relative to the size of the clusters. A value greater than 1 indicates valuable and loyal customers and a value less than 1 represents a potential area for growth. As can be seen, clusters 1 and 3 would appear to be valuable customers as they have higher leverage, whereas clusters 0 and 2 are potential customers. The company could benefit from a targeted marketing approach to these clusters – namely cluster zero that shows a low spend relative to the amount of customers.

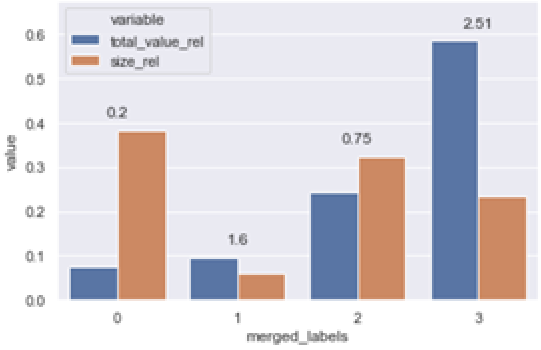


Figure 3: Leverage analysis

3.4. CLUSTER PROFILING

We proceeded to evaluate the 4 clusters for the purposes of understanding the company's major cus tomer segments. With regards to the demographic variables, we noticed that the recency and education of the customers did not have much differentiating power.

The remaining variable seemed to vary somewhat and can help us to describe our clusters. Within th e personal preference/past behaviour features, we noticed several glaring differences among our clu sters. Please see the figures below accordingly.

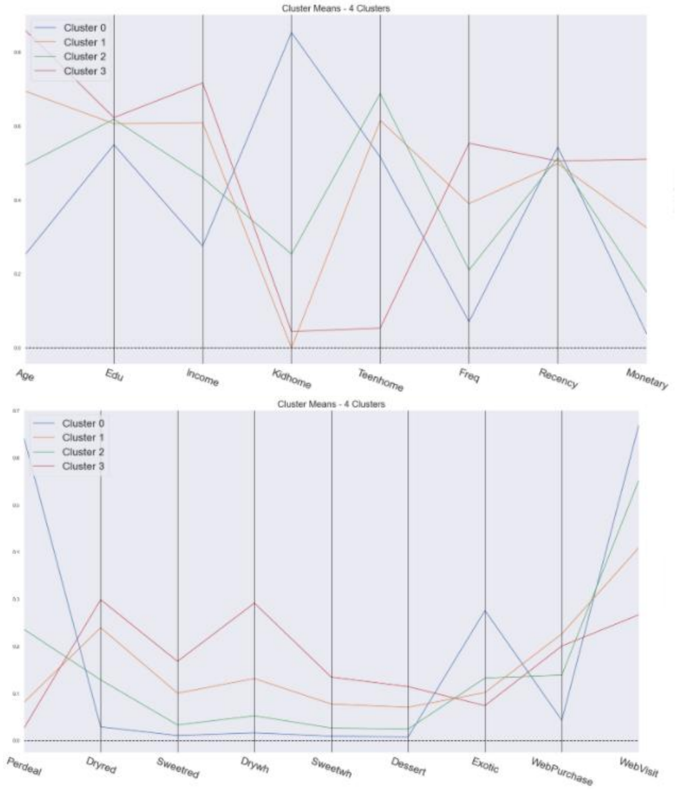


Figure 4: Cluster profiling

Clusters	Age	Income	Kid	Teen	Freq	Money	%deal	WebP	WebV
0	33	46K	0.85	0.52	4.88	122.88	62.46	64.24	6.69
1	60	90K	0	0.62	22.46	996.64	7.87	323.01	4.09
2	48	70K	0.25	0.69	12.57	467.32	22.95	199.58	5.52
3	70	104K	0	0.05	31.44	1560.65	2.54	286.93	2.67

Table 1: Cluster overview

Cluster 0 is the youngest group with an average age of 33. This cluster also has the lowest income and is the group most likely to have young children in the home. The individuals in this group are infrequent customers and have a low monetary spend. It seems that the cluster often takes advantage of discounts provided by WWW. In addition, it appears that they have a particular interest in exotic wines. This cluster visits the website often; however, has a relatively low spend online. This group has a high interest in wines and is also quite sensitive to promotions. This could be an area where a targeted marketing approach would be beneficial.

Cluster 2 contains individuals that are middle aged. They have a somewhat higher income than the previous cluster discussed. These individuals are the most likely to have a teen in the home.

According to our leverage analysis, the above two clusters are areas for potential growth. The subsequent two clusters are more consistent, with more predictable spending patterns.

Cluster 1 seems to be those individuals that are approaching retirement. This cluster has a significantly higher monetary spend than cluster 2. This could be attributed to the fact that they have older children and likely less responsibilities. This cluster also has the highest amount of online purchases.

Cluster 3 represents the oldest segment of customers with an average age of 70. The demographics data of Cluster 3 indicates the cluster consists mostly of retirees. These individuals, on average, spend more than any other groups on the wine purchases from WWW. We believe this is due to the high-income level, and their lack of children provides them with money and time for leisure activities. Cluster 3 also shows the least interest in discounts. This indicates a high propensity to purchase wine regardless of the price. This cluster also shows the most interest in wine accessories.

Across the segments, dry wines were favored over sweet wines. This preference was more prominent as the mean age of the cluster increased. This may be because the older customers have had time to experiment to find their favorite wine type and have grown to a particular flavor. The younger generation, however, appears to be going through the experimental phase. Cluster 0’s particularly high interest in exotic wines, as opposed to the low interest of Cluster 3 also supports this idea.

3.5. Evaluation

To evaluate our clusterization we are going to treat the clusters as labels and building a classification model on top. If the clusters are of high quality, the classification model will be able to predict them with high accuracy. At the same time, the models should use a variety of features to ensure that the clusters are not too simplistic. Overall, we are checking the following attributes:

1. Distinctiveness of clusters by cross-validated F1 score
2. Informativeness of clusters by SHAP feature importances

We will use LightGBM as my classifier because it can use categorical features and we can easily get the SHAP values for the trained models.

3.5.1. Demographic Perspective

A CV score for K-Prototypes for demographic data with 3 clusters is 0.996 which means that the customers are grouped in meaningful and distinguishable clusters. To determine if the clusters are also distinct and informative, we need to look at SHAP values for this classifier.

To classify the K-Prototypes clusters, LightGBM needs 8 features and only 4 of them are quite important and all the others have marginal importance.

3.5.2. Behavior perspective

The CV score here is 0.972 which is a bit smaller than in the demographic perspective. This means that these clusters are harder to perfectly distinguish, yet the score is high enough to conclude that K-Prototypes clusters of behavior perspective are meaningful and distinguishable. To determine if the clusters are also distinct and informative, we need to look at SHAP values for this classifier.

Overall, classifiers for both perspectives have F1 score close to 1 which means K-prototypes have produced easily distinguishable clusters. To classify the K-Prototypes correctly, LightGBM uses more features in behavior perspective (7-8). This contrasts with the demographic perspective which could have been almost perfectly classified using just 4-5 features. This proves that the clusters produced by merging different perspectives are more informative.

4. RESULTS EVALUATION

The overall results of the clustering experience with data mining are easy to communicate from a business perspective: once we have classified a customer into a particular segment, we can take appropriate action to increase his/her lifetime value. The study produced what are hoped to be better marketing recommendations and the way how we can improve the relationship.

By harnessing the power of data mining tools, customers were partitioned into segments based on demographic and behavioral traits, that will influence the language used in multichannel marketing campaigns and resulting in higher CTR and conversion rates than we have for the test promotion for the silver-plated cork extractor: 6.82%.

5. DEPLOYMENT AND MAINTENANCE PLANS

5.1. MODEL DEPLOYMENT

This project identified 4 distinctive clusters of WWW's customers based on their demographic data, preferences and spending history. Group-AA believes these four clusters will help WWW to better understand its consumer base, particularly for the marketing purposes. Without a proper deployment strategy, unforeseen anomalies may occur. For the optimal integration we believe the following seven steps are required:

1. Starting with these 10,000 customer data, check if our findings in the consumer behavior pattern still holds when compared against the entire database. In case of a computational power limitation, we can randomly sample multiple times and test against them.

3. Implement a trigger based on our clustering into the database. This will avoid the need to develop an end-user application.
3. New customers will be placed into different segments depending on their demographic information and temporarily placed into one of the four clusters we've defined.
4. Over time, their behavioral pattern data can be collected, which will be used to adjust their membership in the cluster.
5. We recommend re-clustering of the customers every 6 to 12 months to evaluate whether the consumer behavior still follows our findings in this project.
6. In case of a significant divergence, customer relationship management strategy, as well as the marketing strategy will need to be reviewed.
7. Further training of employees, especially the ones responsible for customer relationship management. Importance of correct data entry must be highlighted.

5.2. Marketing Recommendation

Cluster 0 was found to be the lowest spenders, which has resulted in a negative average customer lifetime-value. We strongly argue against giving up on this segment. This segment represents the largest number of customers in the given sample, a dollar increase in the profitability of this segment would have the highest impact to WWW's bottom line. In general, law of diminishing return would dictate it will be easier to increase the profitability of this cluster compared to more profitable segments. However, without the information on WWW's historical investment in customer acquisition and the configuration of LTV calculation, Group AA is reluctant to suggest the potential value of these customers.

To increase the profitability of Cluster 0, we recommend offering "surprise collection." The collection would consist of various samples of wine, which in turn can be charged more per volume of wine. This would appeal to Cluster 0, as they have higher adventurous tendency. The customers would benefit as they can experience various wines with far less commitment than buying multiple full bottles, while WWW can increase the profitability per volume and ease the inventory handling. It is important to advertise this collection to include many exotic wines and highlight its economic benefits. Depending on WWW's logistic capabilities and its partnership with delivery, subscription-based model for this collection would be explored as well, to stabilize the cashflow. Furthermore, this many individuals in Cluster 0 has limited free time due to their career and young children. These small portion wine could appeal to their busy lifestyle. We also suggest offering a small discount if the collection buyer likes one of the samples and wishes to order a full bottle. This strategy will create a sense of "smart-spending" among them, even if the discount amount is less than what WWW offered in its previous discounts.

Cluster 1 is the smallest cluster - they do not have children at home, they recently made their last purchase and are actively using online. They are quite interested in accessories and therefore, as a possible communication option, they are recommended to increase the average order receipt - offering a free accessory for purchase starting from a certain online price.

Cluster 2 is interested in promotions. Showing promising signs with quantity and value of their purchase but it has been a while since they last bought from you. We can target them with their wish list items and a limited time offer discount.

We also recommend pushing accessories to **Cluster 3**. Our analysis indicates Cluster 3 is most likely to invest in these items.

6. CONCLUSIONS

We have conducted RFM analysis and leverage analysis to better understand WWW's customer base. We have identified four distinct segments, showing wide range of demographics and behavioral patterns. Cluster 0 and Cluster 3 lie on the opposite end of the spectrum, showing a significant difference in all the elements of demographics. Such a difference was also observed in the behavioral pattern, namely, usage of discounts, their adventurousness in experimenting with exotic wines, and the frequency of visits to WWW's web shop, where cluster 0 showed high values in all three areas. Cluster 1 and 2 mostly reside between these two ends of the spectrum.

We have offered varying marketing approaches for each cluster. We aimed at increasing the profitability of Cluster 0 by offering "surprise collection" box to tab into their tendency to experiment, as well as exploring developing this new offering into a subscription-based model depending on the company's logistical capabilities. For Cluster 1, we see benefits in offering small accessories for wine purchases. Cluster 2 could be best served with wish-list and limited-time offer discount. Lastly, more expensive accessories can be advertised to Cluster 3.

We believe through clustering and customer-centric marketing approach, WWW can further improve its products and services provided to its customers.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

WWW can expand its data collection efforts, namely, collecting more specific data for exotic wines. Currently, Exotic and Perdeal data provides a blanket information for all different flavors. If WWW can specify exactly how much discount was applied to which type of wine, as well as which type of exotic wines were purchased, it could serve as a starting point to measure individuals' monetary commitment for different types of wine (I.e., perceived utility of each wine type for each customer).

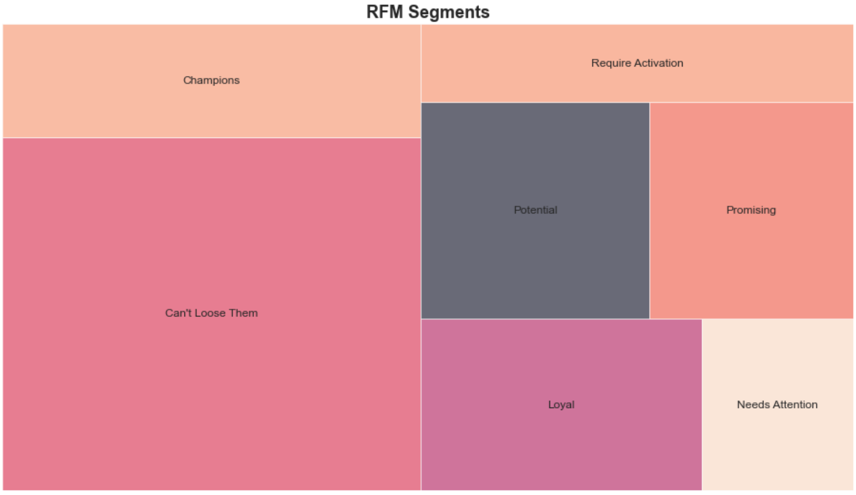
7. REFERENCES

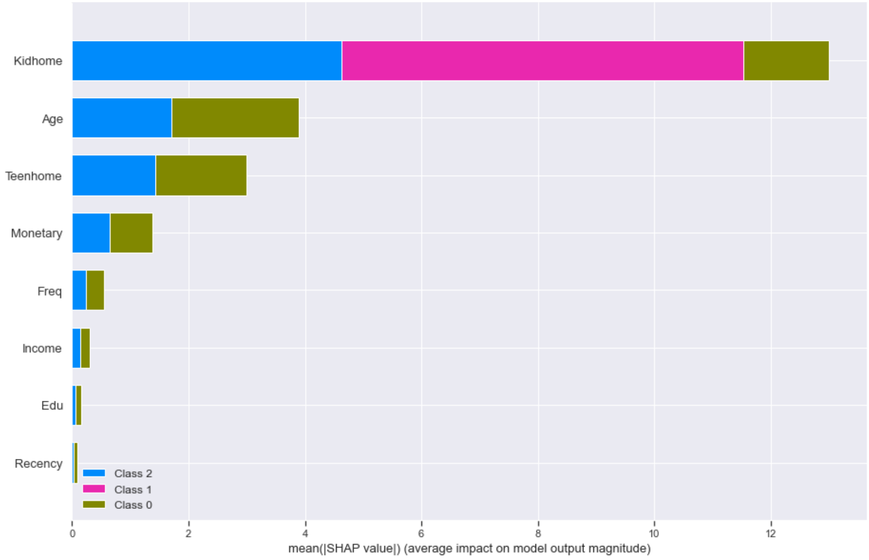
[1] Rsquared Academy (2019, July) Customer Segmentation using RFM Analysis https://blog.rsquaredacademy.com/customer-segmentation-using-rfm-analysis/

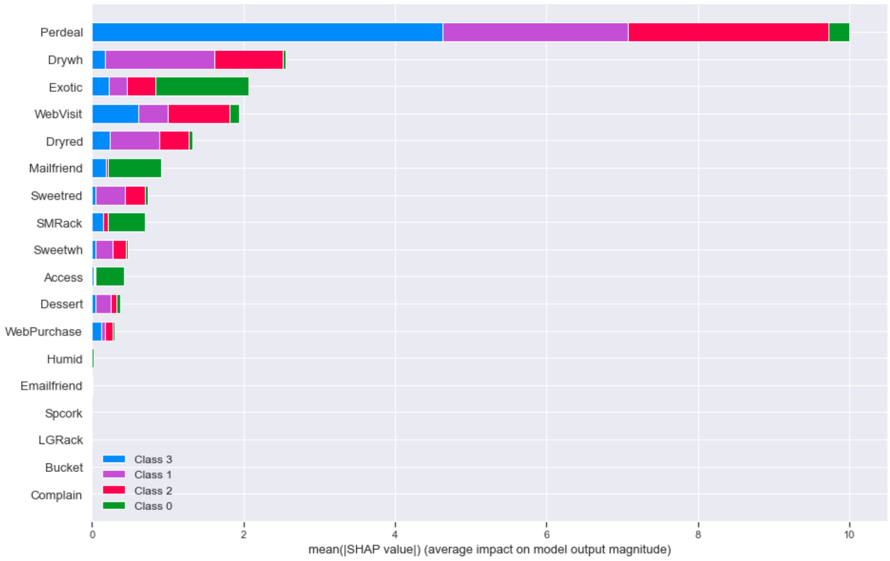
[2] Allison Kelly (2020, Feb) Customer Segmentation with K Means Clustering https://towardsdatascience.com/customer-segmentation-with-kmeans-e499f4ebbd3d

[3] Yexi Yuan (2019, Aug) Recency, frequency, monetary model with python — and how sephora uses it to optimize their google and facebook ads https://towardsdatascience.com/recency-frequency-monetary-model-with-python-and-how-sephora-uses-it-to-optimize-their-google-d6a0707c5f17

8. APPENDIX







	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
Product	Exotic	Dry red, dry white	Exotic, dry red	Dry read, dry white, sweet red
Price	Low price	High price	Avg price	High price
Place	Email, mail	Online	Online	Store
Promotion	Promotions and special offers	Free accessorize with the purchase more than avg sum	Wishlist items and a limited-time offer discount	Special events and individual subsription