BUSINESS CASES WITH DATA SCIENCE

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

Business Case 1 – WWW Customer Segmentation

Group W

Ana Luísa Mestre, number: 20200599

Beatriz Pereira, number: 20200674

Mariana Domingues, number: 20201040

Nadine Aldesouky, number: 20200568

March, 2020

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

INDEX

1. INTRODUCTION

The company Wonderful Wines of the World wants to segment their customers and develop marketing strategies for each segment in which they are organized.

Our team of data scientists has taken on this project to help WWW make data-driven decisions to optimize their business processes.

2. BUSINESS UNDERSTANDING

2.1. Determine Business Objectives

- **Background**: Wonderful Wines of the World is a 7-year old company selling wines and wine accessories around the world. Wines are sold online, by phone and in-store. Currently, WWW has around 350000 customers and 10 US stores. Most of the customers are deeply involved in the world of wine and have a good financial condition. Unfortunately, the company has no knowledge about its customer database. This problem affects the sales, marketing, and IT department. Currently, all campaigns are mass-marketed and developed based on feedback from agents, basic marketing reports and intuition. Although this might have been working in the past, as WWW starts to grow and its customer base begins to diversify, it must be able to keep up by providing better targeted marketing efforts. This will improve its profitability and allow it to strategically reach new customers.
- **Business Objective/Problem**: To segment existing customers and develop marketing strategies for each segment in order to better target new customers.
- **Business Success Criteria**: Successful segmenting of customers will result in clearly defined customer *personas* with associated actionable marketing plans.

2.2. Assess the Situation

- **Inventory of Resources**: Currently we have a database of 10000 customers which will be analyzed by a team of four business analysts known as Group W. To manage our problem we will use Python, Microsoft Word, Power Point, and Excel.
- **Requirements, Assumptions, and Concerns**: This report is for managers' use which means that it is for private use in WWW company only. Additionally, this is not a technical audience meaning that all aspects must be comprehensible in business terms without technical jargon. For this project, we will be working with the *WonderfulWinesoftheWorld.xlsx* dataset, collected on the 14th of February 2021, and composed by a sample of 10000 customers and 29 variables.
- **Risks and Constraints**: These customers were randomly chosen based on the constraints that they must be above 18 years old and that they have purchased something from the company in the last 18 months. Additionally, this report must be completed and presented to management within one week's time.

Risks	Contingencies
Useless features for the task	Ask for different variables or derive new features
Not enough observations	Ask for more customer contacts
Missing relevant information	Ask for more detailed explanation in meta data e.g. Lifetime Value
Lack of funding or available time	Request an extended deadline
Losing the data	Create a copy of all aspects of the project
Competitors reaching potential customers first	Prioritizing customers who have a higher risk of converting
Breakdown of system or data warehouse	Perform daily quality insurance checks on warehouse to verify that the data and the programs are updated and upload them correctly

Table 2.1 – Risks and Contingencies

- **Terminology**:

Dessert wine: very sweet wines that are drunk after a meal with dessert.
Wine rack: a set of shelves to store wine bottles.
Humidifier: small device used to increase the humidity levels in a wine cellar to prevent the corks from shrinking and air to enter the bottle deteriorating the quality of the wine.

- **Costs and Benefits**:

As this is an academic project, we have no information about the costs of the data collection neither the ones regarding the development and the implementation of our solution for the business problem.
The benefits include reaching new customers, increasing ROI, and improving customer satisfaction, leading to higher revenues.

2.3. Determine Data Mining Goals

- **Data Mining Goals**: To segment existing customers using clustering techniques to determine better targeted marketing plans for each cluster. Also, create a supervised learning model to classify and target potential customers into one of the defined cluster groups.
- **Data Mining Success Criteria**: Successful segmenting of customers will result in a high performing model with a relatively high R2 score. This means that the model should produce clusters which are homogenous within and highly distinct from each other. The current situation is the benchmark which is no clustering or analysis of the existing customer database. Additionally, the model should be very insightful by providing specific actionable items. Finally, the predictive model which classifies new customers should have a relatively high accuracy.

2.4. Produce Project Plan

- **Steps**

1. General exploratory analysis, identifying clear problems in the data such as missing values, outliers, or incoherencies.
2. Deeper exploratory analysis.
3. Describe the insights obtained from the data exploration - explore; produce and interpret visualizations to understand the absolute frequencies of the variables, or the presence of outliers.

4. Clean the data, perform feature selection, outlier detection, and/or correcting incoherencies.
5. Review the data again, after cleaning it (e.g.: see if the insights are the same or if they changed slightly).
6. Define and apply different techniques and clustering methods.
7. Fine Tuning (making small adjustments to parameters to achieve the desired output of performance) of the model, and selection of the one that provides the best solution for the segmentation of the customers based on higher Calinski, higher R2 and lower Davies score (closer to 0).
8. Assess and interpret the results from the chosen model.
9. Create a supervised predictive model having the cluster labels as the target to classify new customers and the taken outliers.
10. Provide a deployment plan.

- **Initial Assessment of Tools and Techniques**

Technique	Pros	Cons	Selection Criteria
Mean shift	Can find arbitrarily shaped clusters Not dependent on initialization	Computationally expensive Selection of bandwidth beforehand	R2 score, Calinski Harabasz score, Davies Bouldin score
K-means	Good for metric features Good for spherical-shaped clusters Efficient and fast implementation	Sensitive to outliers and initialization method Need to set # of clusters beforehand	R2 score, Calinski Harabasz score, Davies Bouldin score
K-prototype	Good for mixed features (numerical and categorical)	Costly in terms of time and computing power	R2 score, Calinski Harabasz score, Davies Bouldin score

Table 2.2– Tools and Techniques

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data Understanding

The dataset was provided as an excel file by the manager directly from the company. This dataset has 10000 rows and 28 columns, from which 11 are non-metric variables and 17 are metric variables. The data seems clean, there are no missing values, no customers under 18 years old and no duplicate observations. Regarding the variables of the types of wines (Dryred, Sweetred, Drywh, Sweetwh, Dessert) - the sum of these variables should be 100% - we noticed that some records have a sum equal to 99% and others 101%. However, this is considered correct as it was probably caused by the rounding. The Exotic variable is not included in this 100% sum since it is related to very unusual wines – specific wines from unusual places or some special edition that can be included in the other categories.

In order to determine which variables were redundant and/or irrelevant we did a correlation analysis of the variables. To examine the pairwise relationship of the variables more precisely, we created a variety of heatmaps. The heatmaps were based on these correlation matrices: Pearson, Spearman and one of the Phi k (for both numeric and non-numeric variables). The threshold we defined was 80%, meaning that all

variables that had a correlation of more than 80% were up for examination in the next step. The variables selected here were Age, Income, Freq, Monetary, LTV, Webpurchase, Webvisit and Perdeal.

By analyzing these relationships, we can extract some useful information about our customers.

- The strong linear relationship of the number of purchases and all the money spent in the last 18 months (Freq vs. Monetary) allows us to conclude that, in general, our customers waste the same money by purchase in average (provided by the slope of this linear relation).
- The customers that have a bigger proportion of purchases bought on discount, in total, have a lower number of purchases made (Freq). The lowest frequency of purchases and highest number of purchases on discount represent the customers that only take advantages of the company discounts (buying mostly on promotions) and represent the younger people (relationship with Age).
- Regardless of not having the LTV formula, its high relationship with Freq and Monetary (and consequently with Age and Income) can lead us to conclude that the lifetime value includes their information already.

From plotting all the numeric variables' box plots and applying the Inter Quartile Range Method with the default multiplier, we noticed that the variables Freq, Recency, Monetary, LTV, Sweetred, Drywh, Sweetwh, Dessert, Exotic have potential outliers.

3.2. Data Preparation

Before uploading the data to Python, we noticed that the last row of the data didn't have a customer Id and the values were related to a function applied to each variable (specifically the average and the sum for different variables).
Moreover, we dropped the Rand column as it was not included in the metadata, so we have no information to explain or analyze it.

- Redundancy analysis:

Going through the list of highly correlated variables (made in the previous section) and using the metadata as well as the business objective in mind, we decided on the following:

Variable to Keep	Variables to Drop	Reason
SMRack, LGRack, Bucket, Humid	Access	Correlated, the specific accessory provides us with more info
Webpurchase	Webvisit	Linear correlation, the % of purchases made on website will give us more info than an average of visits per month
LTV	Perdeal, Freq, Monetary	Linear and nonlinear correlation (assuming all their info is included in the calculation of LTV)
Age	Income	Linear correlation (also very correlated with monetary and consequently with LTV)

Table 3.1 – Highly correlated variables

Regarding the outliers, from the variables we selected before, we used the Inter Quartile Range method with the multiplier equal to 5 for the variables Recency, Sweetred, Sweetwh and Dessert to define the range limits; for the variables Drywh and Exotic we decided to keep all the observations given that the limits provided by this method was out of the values range. Manually, we dropped only one observation from the variable LTV. In the end, 268 observations were dropped (2.68% of the dataset).

Regarding new data construction, no new attributes were constructed given the deadline of the project and since the existing features were assumed to be informative and sufficient for the goal. As for formatting, we used MinMax scaler to normalize all the metric variables in order to have the same scale and proportionate impact on the model.

3.3. MODELING

3.3.1 Select Clustering Technique

In this step, we tried to change the parameters regarding the number of clusters in each clustering method to analyse the different solutions that they gave us. We checked the distributions, the metrics, and the profiles of the different clusters. The following methods were applied by perspective (Table 3.2) and with all the features (Table 3.3).

	SOM+KMeans	**KMeans**	**KPrototypes**
Profile	**R2**= 0.426 **Calinski**= 3603.650 **Davies**= 0.854	**R2**=0.798 **Calinski**=9601.646 **Davies**=0.830	**R2**=0.468 **Calinski**=2141.652 **Davies**=2.113
Features used	Age, Edu	Age, Edu	Age, Edu, Kidhome, Teenhome
Clusters	3	5	5
Preferences	**R2**=0.154 **Calinski**=885.030 **Davies**=1.834	**R2**=0.624 **Calinski**=4031.922 **Davies**=1.492	**R2**=0.159 **Calinski**=460.528 **Davies**=4.385
Features used	Dryred, Sweetrd, Drywh, Sweetwh, Dessert, Exotic	Dryred, Sweetrd, Drywh, Sweetwh, Dessert, Exotic	Dryred, Sweetrd, Drywh, Sweetwh, Dessert, Exotic, SMRack, LGRack, Humid, Spcork, Bucket
Clusters	3	5	5
Engagement	**R2**=0.307 **Calinski**=2150.096 **Davies**=1.50	**R2**=0.670 **Calinski**=5655.195 **Davies**=1.072	**R2**= 0.289 **Calinski**=986.535 **Davies**=6.229
Features used	Dayswus, Recency, LTV, WebPurchase	Dayswus, Recency, LTV, WebPurchase	Dayswus, Recency, LTV, WebPurchase, Complain, Mailfriend, Emailfriend
Clusters	3	5	5

Table 3.2 – Clustering results by perspective

Metrics	SOM+KMeans	KMeans by perspective + Hierarchical merging	KMeans	KPrototypes
R2	1	0.421	0.430	0.465
Calinski	3654.106	3540.436	3666.389	4233.410
Davies	1.174	1.633	1.418	1.360
Features used	Dayswus, Edu, LTV, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebPurchase	Dayswus, Edu, LTV, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebPurchase	Dayswus, Edu, LTV, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebPurchase	Dayswus, Edu, Kidhome, Teenhome, LTV, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebPurchase, SMRack, LGRack, Humid, Spcork, Bucket, Complain, Mailfriend, Emailfriend
Clusters	3	3	3	3

Table 3.3 – Clustering results for all the features

Assumptions per modelling technique:

- SOM, Kmeans and Kprototypes are based on a distance metric so they assume all features are on the same scale which is why the data must be normalized prior.
- KMeans assumes that the cluster are of spherical shape and are of similar size. Additionally, it assumes that all the data used is numerical.
- KPrototypes[1][2] assumes that the relative distance between the centroids of each cluster is large. It also assumes that local neighbourhood around the centroid is denser than the neighbourhood of the surrounding points within the cluster. Finally, it assumes that the data is mixed which means it includes both numerical and categorical features.

To model our dataset, we chose to use KPrototypes given the criteria mentioned earlier.

3.3.2 Select Classification Technique

To classify new customers into our resulting clusters, we used a decision tree with a depth of 4. The validation of this model was made with the given data, using the hold-out method to split the data into a training and test set (80/20). We also could relocate the outliers that were taken in the processing part with the following distribution – Cluster 2: 191 customers, Cluster 1: 73 customers and Cluster 3: 4 customers.

[1] https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

[2] https://www.hindawi.com/journals/mpe/2020/5143797/

3.4 EVALUATION

3.4.1 Evaluate Segmentation Results

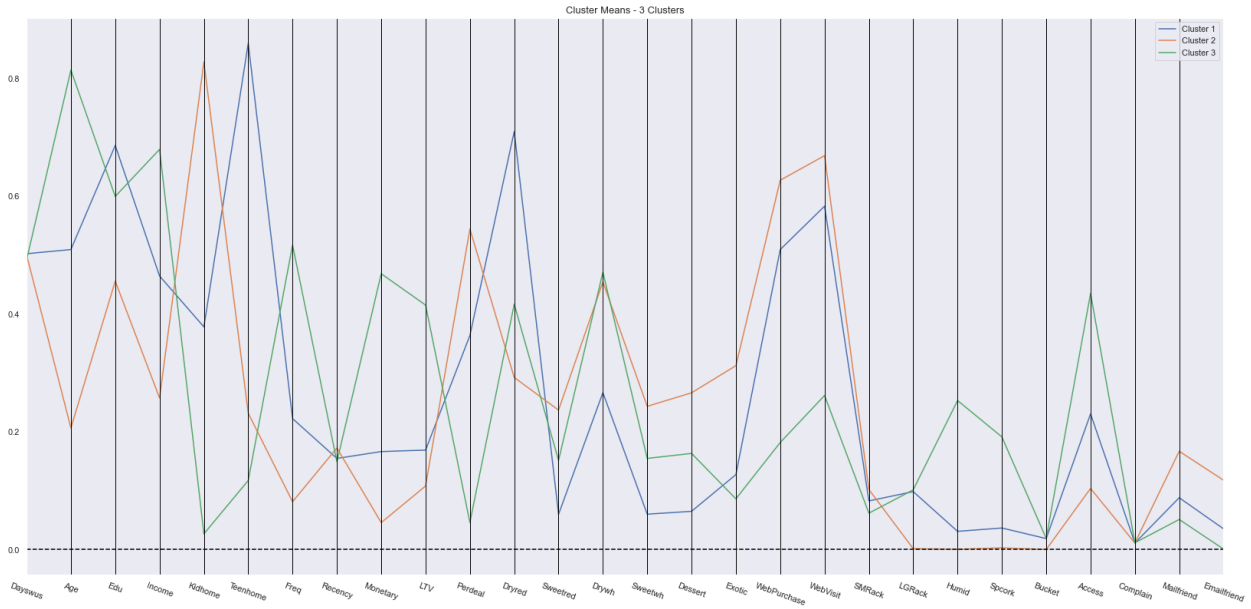

Figure 3.4 – Cluster Profiling

Cluster 1

This is the largest cluster, composed by a total of 4241 customers. This is the parents cluster as it has the highest number of customers with teens living at home. Their average age is 49, with an average income but also with the highest education. They use the website for online purchases almost half of the time. These customers consume the highest amount of dry red wine, they rarely buy sweet wines.

Cluster 2

This cluster comprises 2787 customers and they represent the least frequent customers. They are the youngest customers (average age is 30), and the ones who buy most by website and on discount– with the highest percentage of sales made on website and the highest number of visits per month. They are also the most engaged as they are the most mail and email friendly. Additionally, most of them have kids at home so they are young parents with minimal time and more expenses. Since they have the lowest income, they spend the least amount of money. They prefer exotic wines but have an even distribution across the different types of wines, meaning they are open to exploring different kinds of wines and taking the risk.

Cluster 3

This is the richest and the smallest cluster including 2704 customers. It is the grandparents cluster as they have an average age of 66 and very few of them have kids or teens at home. They are the customers who buy the most frequently and spend the largest amount of money making them the most valuable and with the highest Lifetime Value. They prefer dry red and dry white wine. They do not like exotic or sweet wines and do not interact with the website as much. They are also not motivated by discounts. They buy the most humidifiers and corks as they probably have wine cellars at home.

3.4.2 Evaluate Classification Results

For the test set we reached an accuracy of 90.91%.

3.4.3 Review Process

There were no major problems during the project process, however, the time constraint was quite restricting. It would have been helpful if we had additional time to experiment with more algorithms and provide the best solution possible.

3.4.4 Determine Next Steps

- Move to the deployment step.
- Fine tuning of the clustering model.
- Improve the data preparation step – outliers, feature selection, etc.
- Improve the predictive model - tuning the parameters or trying new methods (e.g. KNN).
- Training and testing the predictive model with different data splits - using K-fold cross-validation, for example.
- Build different marketing approaches for each segment, based on their characteristics/preferences.

4. DEPLOYMENT

4.1. Deployment Plan

After having the customers segmented in three different groups, and after building a model that predicts to which group each new customer belongs, the next step is to determine how to reach new and existing customers based on the four elements of the Marketing Mix. Additionally, we must identify which group should be prioritized to apply several approaches to reach them:

Cluster 1

Considering their preferences and their level of education, monthly video content can be created to provide some storytelling about the wine (i.e. history about the origin of the wine, as well as some curiosities/studies done regarding the topic) on the WWW website. Additionally, more variety of non-sweet dry red wine should be provided faster than the competitors (e.g. weekly new wine). Finally, it is possible to sell these wines at slightly discounted prices (e.g.: the wine of the month, bundles) or entice more website traffic by making online – only promotions.

Cluster 2

In this segment, WWW should offer more diversified wines and create an *app.* This can allow consumers to share and rate all the wines they tried, win points by providing reviews/ratings which then earn them discounts. Another option would be to create the nostalgic customer e-card with stamps to get a free wine after buying X number of bottles/reaching X stamps. These options would act as a loyalty program to maintain and attract new customers in this segment. The messaging can be done through mail flyers or online through targeted website banners for online orders. For this group, WWW must provide very low prices with a focus on discounts by quantity.

Cluster 3

Since this cluster represents our most valuable customers, WWW should try to maintain them by offering more variety of non-sweet luxurious dry wines and more accessories for wine enthusiasts who tend to

have cellars (i.e. wine glasses, tap barrels, reusable flask bottles etc.). Moreover, WWW can create wine tours, customized gift sets and blind tasting events. Since this group spends a lot of time in stores and have high income, they would have a willingness to pay for the tours or a wine sales agent. This agent would explain to them the history and ingredients of the wines, how they are to be enjoyed, smelled, or handled. Besides this, we can also provide higher prices or promote in store. WWW can reach out to them by telephone (given the pandemic situation) or in-person during better conditions.

4.2. Plan Monitoring and Maintenance

- We would have a program that would give us an alert when a certain percentage (threshold provided by WWW's sales/marketing manager) of customers do not fit into any of the presented clusters.
- If the customer base increases by 25% (to be reviewed by WWW's sales/marketing manager), this project must be reviewed because this increase could give rise to new customer profiles and segments.
- Following the end of the pandemic, these models and marketing strategies must be updated due to the drastic change in context associated.
- If WWW moves to a more digital business model, it will attract a different customer base which would make this analysis no longer relevant and effective.
- If current conditions do not change, the analysis must be reviewed within one year's time at most in order to ensure its viability and relevance.

4.3. Review Project

To sum up, although the dataset was of good quality, we believe that additional data would have been very beneficial for the project. This is because, in this case we were building two models and their generalization capability would have been more accurate if we had more data to train and test them to ensure their credibility. Additionally, a more detailed definition of the variables in the metadata would have also been helpful, i.e. LTV and Access.

5. APPENDIX

labels	1	2	3
Dayswus	901.534072	900.268030	894.100222
Age	48.558123	30.384284	66.844305
Edu	17.487621	15.645138	16.795488
Income	70656.980193	43607.389666	98750.658654
Freq	12.778354	5.301399	28.400888
Recency	52.592313	58.616792	50.680843
Monetary	493.711860	140.040545	1378.804734
LTV	123.424664	14.637962	563.252589
Perdeal	35.248526	52.784356	4.454142
Dryred	70.542089	29.626480	41.819527
Sweetred	3.049752	12.071762	7.716716
Drywh	20.444235	34.160029	35.373151
Sweetwh	3.000000	12.157158	7.737056
Dessert	2.920066	11.972372	7.342456
Exotic	12.206555	29.963760	8.263683
WebPurchase	46.776468	56.666667	19.265533
WebVisit	5.829050	6.686760	2.615385

Table 5.1 – Mean values of each metric variable per cluster

labels	1	2	3
Kidhome	0.377977	0.827772	0.026627
Teenhome	0.859467	0.231432	0.117234
SMRack	0.082528	0.101902	0.061760
LGRack	0.098326	0.001435	0.101331
Humid	0.030889	0.000359	0.252959
Spcork	0.036548	0.002870	0.191198
Bucket	0.018628	0.000000	0.018861
Access	0.230370	0.103696	0.434911
Complain	0.011082	0.011123	0.011464
Mailfriend	0.087951	0.166846	0.051036
Emailfriend	0.035133	0.117689	0.000740

Table 5.2 – Proportion of 1's of each binary variable per cluster