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 J

Francisco Hermenegildo, number: 20200737

Gil Gonçalves, number: 20201066

Ikram Bouziri, number: 20200753

February 2021

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

INDEX

1. INTRODUCTION

WWW is a wine company that although having a big database of clients does not have a meaningful way of differentiating the types of customers present in it. This lack of analytical knowledge means that every month WWW loses money on campaigns for customers that don't care or that it is inappropriate to.

The analysis present in this report enables the company to better understand its customers, promote tailored approaches to each group, retain the most profitable customers and obviously drop the investment made on the worst ones.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Wonderful Wines of the World is a 7-year-old company that sells primarily wine from small and unique wineries around the world, although also offering related accessories such as wine racks and cork extractors to its customers. The customers can the products online, by visiting the 10 stores across the USA or through telephone, after consulting the catalogue that is sent every 6 weeks.

Four years ago, WWW decided to create a database of its customers, containing already 350.000 customers. Most customers are highly involved in wine, being great entertainers and with sufficient money to indulge their passion. The existing marketing strategy is not efficient since it is based on mass-marketing and one fit-all strategy for all the customers in the database.

2.2. BUSINESS OBJECTIVES

With this project, we seek to offer the company valuable information from the existing database and create more personalized marketing campaigns. This will be achieved mainly by differentiating the existing customers.

2.3. BUSINESS SUCCESS CRITERIA

At the end of this project, the company will be able to not only clearly segment its customers through shared characteristics but also implement targeted marketing approaches.

2.4. ASSESS SITUATION

The development of this project implied a team of three data mining students, all with the necessary computing and software skills. The company provided us access to a sample of its customer's database through the excel file WondefulWinesoftheWorld.xslx. This file contains information regarding 10000 active customers - a customer is considered active if he purchased in the past 18 months - each one characterized by 29 features.

WWW required this project to be completed and presented to the management team by the 1st of March 2021.

2.5. DETERMINE DATA MINING GOALS

The goal of a market segmentation activity is to ultimately define tailored approaches for each group and improve its revenue. To achieve this, we will apply the clustering algorithm that best fits the database, optimizing its segmentation, and study at the same time which features distinguishes groups the best. If needed, new features will be created to extract the most information possible, and finally, analyze the segments looking for clear traits and stimulus required. We will also determine how many groups we can segment the customers given the amount, frequency, and recency of their purchases over the last 18 months, their demographic information, and their preferences.

3. PREDICTIVE ANALYTICS PROCESS

3.1. DATA UNDERSTANDING

(Data Collection) All the necessary data for the development of this project was provided by the Wonderful Wines of the World management team, in the form of an excel file, the respective metadata, and a PDF file with the business description and requirements for the project.

(Data Description) Jupyter notebook was the selected software to load the excel file and perform all the necessary data mining stages. The original dataset was composed of 29 attributes and 10000 customers, covering both demographic and behavioral perspectives.

(Data Exploration) An extensive initial analysis of the dataset was performed using several data mining and visualization tools, each one for the respective purpose.

Tool	Use
Pandas-Profiling	Overview of the entire dataset and its characteristics
Boxplot	Identify possible univariate outliers
Histograms	Visualize variable frequencies
Scatterplot	Identify possible bivariate outliers

Table 1. Different data exploration tools used

(Data Quality) This analysis demonstrated the overall good quality of the provided data, with no missing values identified, no duplicated rows, or other types of errors.

3.2. DATA PREPARATION

(Data Selection) Since this project requires the implementation of an unsupervised learning model, only continuous features were considered for its development. Both redundancy and relevancy guidelines for feature selection were followed to achieve an ideal subset of variables

PCA was used not only with the purpose of dimensionality reduction, allowing the projection and subsequent visualization of the data in a 2-dimensional space, but to also obtain a better understanding of the most important variables, through the analysis of the variables most correlated with the 2 main principal components.

Contrarily, Pearson Correlation and phick matrix were implemented and considered to obtain a more restricted set of attributes. Note that, after a satisfactory final clustering solution was achieved, the remnant attributes, namely the binary ones were used for cluster profiling.

(Data Cleaning) The overall good quality of the dataset reduced in great measure the time spent in the data cleaning process.

It was immediately noticed the presence of one row containing the mean values across all the variables. Since this information can be easily extracted with python code it was decided to discard it.

Outliers were also identified, present in the variables “SWEETRED” and” SWEETWT” however, due to the marginal number of cases these records were kept.

(Feature Engineering) During the data analysis, we found it fundamental to create new attributes from the original ones, adding mixed measures, enabling a better understanding of the data without as many variables.

New Attribute	Attributes
CustomerImportance	LTV/Dayswus
AvgSpending	Monetary/Freq

Table 2. Features created and their rationale

(Final Feature Selection) Several different combinations of features were tested in SOM and KMeans to get the clearest segmentation possible. The variables used for the Clustering were ‘Age’, ‘CustomerImportance’, ‘AvgSpending’, ‘Income’, and ‘Perdeal’.

(Data Formatting) The fact that the various clustering models assign the same importance to the different distances or directions of the input space justifies the normalization of all the applicable features since they referenced very different scales.

3.3. MODELING:

3.3.1. Select Modeling technique

Having achieved a normalized final subset, the selection process of the final model was iterative, experimenting with different algorithms and then selecting the best performing one, respecting the following criteria:

- Better segmentation power of the clustering solution (clear clusters)
- Uniform size across all clusters
- Good assessment results

SOM and KMeans were the two models with the best result when applied to the dataset, giving both very similar results, although, in the end, KMeans was optimal.

3.3.2. Build Model

The first model to be considered was Self-Organizing Maps because it's an algorithm that, although complex to understand, powerful at finding patterns in a dataset.

mapsize	initialization	neighborhood	training	lattice	train_rough_len	train_finetune_len
15x15	random	gaussian	batch	hexa	200	200

Table 3. SOM parameters used

KMeans was then used to segment the database in the chosen amount.

n_clusters	init	n_init	random_state
4	k-means++	30	42

Table 4. KMeans parameters for segmenting

label	Age	Income	Perdeal	AvgSpending	CustomerImportance
0	1.212422	1.219848	-1.042812	1.205530	1.405156
1	-0.255251	-0.225308	0.229587	-0.298454	-0.639703
2	0.413487	0.396658	-0.580108	0.475343	-0.061595
3	-1.072826	-1.083648	1.057934	-1.083700	-0.726506

Figure 1. SOM clustering results

Kmeans was a model considered from the start due to its acknowledged use to segment customer databases and was in-fact the most competent at it. Although SOM results came close, KMeans was able to segment furtherly and behaved well either with many or just a few features.

n_clusters	init	n_init	random_state
4	k-means++	30	42

Table 5. KMeans clustering parameters

labels	Age	Income	Perdeal	AvgSpending	CustomerImportance
0	0.580544	0.547980	-0.686840	0.631938	0.121503
1	-1.097828	-1.130475	1.144120	-1.134664	-0.729855
2	1.288722	1.323958	-1.077510	1.284766	1.655225
3	-0.313683	-0.268822	0.182924	-0.316585	-0.622394

Figure 2. KMeans clustering results

The last model considered was the Gaussian Mixture of Models because of its prowess to separate different cluster masses accurately and as a comparison to our benchmark values of Kmeans and SOM. Ended up not segmenting the database well enough, performing poorly with many or just a few variables.

3.3.3. Model Assessment

One of the first questions raised was what is the optimal number of segments to describe accurately the customer's dataset. And although visualizing the output of each different clustering solution was enlightening, we decided to also sustain our decision with quality metrics. All the corresponding plots for each metric are present in the notebook in the Quality Check section.

All the models were assessed firstly considering not only the segmentation power of the dataset but also the relative size of the labeled clusters. Then, if the clustering solution was successful at an initial eye-level analysis, various tools would be applied to assess the quality of the solution in more technical-based testing.

Tool	Best nº of clusters
Inertia plot	Two visible elbows at 3 and 4 clusters
R2 plot	4 Clusters
Silhouette plot	4 and 7 clusters
Distortion	4 Clusters

Table 6. Different metrics used

4. RESULTS EVALUATION

In the table below we have a comparison of distinguishable features between each segment, and also to the entirety of the dataset.

Cluster	**0**	**1**	**2**	**3**	**Dataset**
Importance	$$$	$	$$$$	$$	Average
Size	2482	3052	2066	2400	-
Avg Age	58	29	70	42	47
Avg Income (in $)	85.035	38.691	106.460	62.482	69904
Avg Purchases last 18 months	19,7	3,3	32,5	8,4	14,6
Avg % of purchases made on promotion	13,2 %	64,3 %	2,3 %	37,5 %	32,4 %
Avg LTV (in $)	253,6	-6,3	687	25,7	209
Avg % of purchases made on the website	35,9 %	57,2 %	18,8 %	50,5 %	42,4
Avg Purchase Value (in $)	40,6	16,4	49,6	27,6	31,9
Avg CustomerImportance (in $)	0,28	-0,007	0,79	0,03	0.2
Total LTV of Segment (in $)	629.390	-19.218	1.418.873	61.670	Total 2.090.715

Table 7. Comparison between segments

Relevant characteristics - Segments ordered from most to least relevant:

- The customers from **Cluster 2** are considered the most important customers since they have the highest lifetime value of all, do not wait for promotions to make purchases, and spend much more per purchase than average.
- **Cluster 0** customers represent those customers who have a high income, although, for some reason, they do not spend accordingly. It is the segment that has the highest potential to increase revenue.
- **Cluster 3** is maybe the one with the highest potential for an increase in profitability since its customers are earning and spending just below the dataset average. Although this segment

does not create much revenue it's still a worthwhile investment to keep these customers as ours.

- Finally, **Cluster 1** is our worst group of customers. These customers are young, low paid, and our promotions are the main reason they buy. It is also the only segment that instead of making us profit, ends up costing us money.

Cluster	**0**	**1**	**2**	**3**
Dry Red	59,8	37,8	45,4	61
Sweet Red	5,6	10,3	6,7	4,7
Dry White	23,8	31,2	34,9	24,5
Sweet White	5,4	10,4	24,5	4,9
Dessert	5,4	10,2	4,9	4,9
Exotic	10,6	27,3	7,1	17,2

Table 8. Segment preferences

5. DEPLOYMENT AND MAINTENANCE PLANS

The approaches suggested focus on the reduction of the operational costs of Wonderful Wines of the World as a whole and ultimately reduce the identified segment cost to the company.

5.1. PLAN DEPLOYMENT

Cluster 2 - Most valuable customers:
The objective in this segment is to maintain and reward the customers since they are the most important customers to the company. For that purpose, it is suggested to:

- Reward the customers with chocolates and cheese for each 250$ they spend.
- Create and financially support a top-10 customer reward that promotes a tour to Portugal to taste some of the best wines of the last harvest.

Cluster 0 – Customers with high value:
Since these customers have a high percentage of teens at home, it is suggested to run targeted campaigns through social media making preference to sweet wines.

Run small loyalty programs offering:

- Offer cork extractor when a customer reaches an annual spending of more than 500$
- Offer a humidifier when a customer reaches an annual spending of more than 600$

Cluster 3 – Customers with potential:
For this segment, it is suggested to promote the Sweet wines as this is the least-buying type, by:

- Organizing free tasting.
- Offering tasting packages of sweet wine with each dry wine purchase.

Cluster 1 - Customers with least potential:
Since these are customers that cost us money to keep, we suggest:

- Make the catalog e-mail based and offer a discount of 3$ on the next purchase of each customer that subscribes (free) to the digital format (newsletter).
  In case the customer doesn't subscribe, stop spending on him.
- These customers visit the webpage a lot, most probably looking for a good deal, maybe WWW could hold each week a different promotion deal.
  Preferably hold an exotic week, each month, since more than 27% of their purchases are exotic wines.

5.2. PLAN MONITORING AND MAINTENANCE

After deployment of the plan, it is advised a close follow-up of the generated results. If any inconsistencies are identified, it is suggested to schedule a meeting at any point in time with the team in order to assess and modify if needed, the marketing strategies provided in this report.

6. CONCLUSIONS

This first business case served as a first practical introduction to all group members of the CRISP-DM framework, which ensures the correct planning, organization, and deployment of any Data Science project. Although some of the stages present in this framework were already familiar, others were more challenging to implement.

- It is fundamental to have not only an excellent understanding of the business but also the requirements for the project prior to any type of model creation.
- As with any data science project, a lot of iterative steps occur in order to achieve the best performing model.
- Being the scope of this project the implementation of an unsupervised learning model, it is important to achieve a good level of understanding of a great number of variables since the characterization of each segment depends on how familiarized you are with them.