BUSINESS CASES WITH DATA SCIENCE

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

Business Case #1 - Wine store

Data4Business Consulting

Débora Santos, number: 20200748

Diana Furtado, number: 20200590

Pedro Henrique Medeiros, number: 20200742

Rebeca Pinheiro, number: 20201096

March, 2021

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

INDEX

1. INTRODUCTION

Thank you for choosing **Data4Business Consulting (D4B)** to help you with the challenge of better understanding your customers characteristics and segmentation. Our main objective is helping your business to improve its set-up, increase your gains, bring new customers and retain the current ones.

The world is experiencing a great technological and digital revolution where understanding business data, customers segmentation and their needs is essential for the business success. The exponential technological advances, such as data mining techniques, artificial intelligence, internet of things, can help taking the business to the next level.

Through innovative technological programs, well-referenced data mining methods and insights of digital marketing, the present report intends to provide an overview of the process behind the analysis, presents the results and insights you need to be successful in this new era.

In addition to the present report, the following deliverables will be submitted:

- Outcomes presentation to WWW: https://prezi.com/view/0etl5qTn6cfrrZbQH8XP/
- Jupyter Notebook with the code of the entire process.
- The files to run application developed by D4B to WWW.

All files can be accessed in Github: https://github.com/Debs86/Business_Cases_Projects

We are excited to take part of this challenge.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Wonderful Wines of the World (WWW) has been present in the wine market for 7 years. The company aims to provide customers with a premium selection of wine and wine accessories. WWW has been using a marketing strategy based on the previous experience acquired, on how to maximize the sales: send a catalog (which is renewed every 6 weeks) to the 350,000 customers on the database (from the past 4 years) and expect the customer to approach the company to buy wine and accessories. Also, from previous analysis, the company already knows that most of its customers are wine lovers who have no financial constrains to get good quality wine.

The current customers have three different ways of purchasing wine and accessories from WWW: in person (through one out of the ten stores WWW has in major cities around the USA), by telephone (through the catalog) or online (on WWW's web site).

The key persons in this business are the owner (Fernando Bação) and the managers (João Fonseca and David Silva). Fernando is interested on increasing wine and accessories selling. João and David are looking at the actions needed to get the outcome the owner expects.

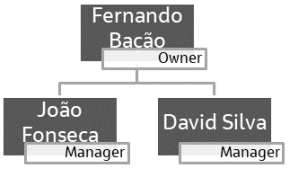

Figure 2.1 - Organizational chart.

The main problem WWW is facing is how to improve the wine and accessories sales (not only to existing, but also to new customers) using the knowledge on the current customers and this is why they have reached D4B.

2.2. Business Objectives

At this point, WWW has no specific knowledge about the customers. All marketing actions are based on market reports, information from sales team and intuition.

The customer's primary objective is to increase wine and accessories sales by understanding the following:

- The key characteristics that best distinguish the customers.
- Which and how many customer segments there are in the provided database.
- How the business can reach new and existing customers from each segment and which segments should be prioritized.
- Improve the interaction with the customers by creating new marketing strategies.

2.3. Business Success criteria

The expected outcome will be well defined customers' segments which can make possible to build a customized marketing strategy and maximize the return of investment. Another expected outcome of this report is suggestions of marketing strategies and business applications for the findings.

The success of the proposed task will be evaluated by WWW's owner and managers and, if needed, we will go back to the model until we get an outcome that matches with the board's expectation.

2.4. Situation assessment

2.4.1. Inventory of resources

This project was made following the CRISP-DM reference model (Cross Industry Standard Process for Data Mining). CRISP-DM is a standard process built in the end of 90's and it was built by more than 200 members lead by a consortium of big companies. *CRISP-DM succeeds because it is soundly based on the practical, real-world experience of how people conduct data mining projects.*[1]

This project has the support of WWW's Management and IT team.

From D4B Consulting, this project will be conducted by the following team: Débora Santos (Executive sponsor), Diana Furtado (Project leader), Pedro Medeiros (Data miner) and Rebeca Pinheiro (Data expert).

We have been provided by the WWW's IT team with a database of the customers who purchased in the last 18 months. It was also provided a metadata file of this dataset.

The main technology used to achieve the objectives of this report was Python. Python is one of most important and commonly used program languages in data science projects.

2.4.2. Requirements, assumptions and constraints

The completion date of the present phase of the project is March 1st, 2021, but we expect to continue giving support and helping WWW to achieve the next goals for the growth of the business.

Although the sale of alcoholic beverages to people under 21 is prohibited in the USA, on this project we considered all customers with age 18 or more that bought in the past 18 months.

The dataset provided has only 10.000 customers, even though the total database has 350.000 customers. One of the assumptions is that 10.000 customer will well represent the entire data.

2.4.3. Risks and contingencies

Table 2.1 identifies a list of risks and contingency proposed.

Risk	Contingency
Insufficient number of features	Work with remaining features or ask for different variables
Insufficient observations	Ask for more observations (customers)
Model overfitting[1]	Ask for observations

Table 2.1 - Risks and contingency.

2.4.3.1. Terminology

Business glossary

- Small wine rack, large wine rack, wine cellar humidifier, plated cork extractor and silver wine bucket are all wine accessories:
  1. Wine cellar humidifier is designed to increase humidity levels in any size commercial or residential wine room.
  2. Wine rack is a frame to store wine bottles.
  3. Plated cork extractor is used to open the wine bottle.
  4. Silver wine bucket is a premium accessory to keep the wine temperature during its consumption.
- Dry red wines, sweet or semi-dry reds wines, dry white wines, sweet or semi-dry white wines, dessert wines (port, sherry, etc.) and exotic wines are types of wine according to its ingredients, flavours, and other characteristics.

Data mining glossary

- Clustering: It is a data mining technique. The technique consists in apply some algorithms that will classify the observations (customers) into groups according to the similarity of their attributes.
- Normalization: The major algorithms of clustering need the data be scaled to a standard range. The process of applying some transformations in the data to have it in the same range is called normalization.

2.5. DETERMINE DATA MINING GOALS

The data mining goals states project objectives in technical terms:

1. Segment customers according to their willingness to purchase wine and accessories, considering their demographic and social information (age, years of education, presence or absence of children, income, etc.), the commercial information records for the last 18 months (purchases, complaints, websites visits, etc.).
   Success criteria: Visualize the cluster solution. Each cluster must have similar characteristics between the customers into it and distinct characteristics from the other clusters. We should be able to describe the characteristics best distinguish the clusters.
2. Evaluate the cluster quality.
   Success criteria: The TNSE (visualization technique to visualize the distribution of the clusters) must have a good distribution of the clusters; Also compare the R2 metric of some cluster techniques applied and chose the one with the highest R2.

[1] Model overfitting may happen as we are dealing only with only 3% of the entire dataset (10.000 out of 350.000 customers). This problem will be identified when the WWW starts applying the solutions proposed to the other customers. If it happens, our consultancy is up to work to improve the modelling as many times as needed.

3. Suggest marketing strategies and business applications.
   Success criteria: Present marketing strategies for each cluster and show business applications to help implement these strategies.

2.6. PROJECT PLAN

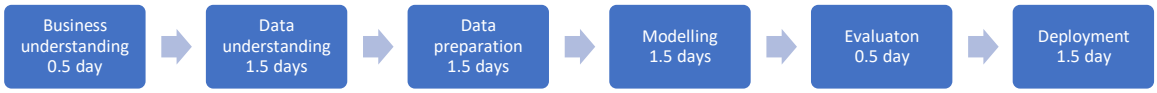

Figure 2.2 - Project's timeline.

Resources wise, for the business understanding we plan to use all the information provided in the kickoff meeting's presentation. For the core stages of the project, we plan to use Python to work the data provided by WWW's IT team. To present the results, we expect to use Word for the report and Prezi for the presentation. Finally, to provide a user-friendly visualization of the results, we plan to build an application using Python.

The quality of the clustering process will be directly connected with the quality of the input data. For this reason, we identify the Modelling stage as dependent of the Data preparation stage. During the project, we must go and back between Data preparation and Modelling many times, repeat this iteratively until we get the desired outcome.

For the Modelling stage we aim to build an unsupervised model (clustering) using K-means algorithm. Due to the timescales we opted for using this algorithm as it is fast and efficient in terms of computational cost, simple to implement and the interpretation of clustering results is straightforward. The clustering quality evaluation will be made using R squared and some visualizations to check the good distribution between clusters.

3. CLUSTERING ANALYSIS

In this section we go through the process of understanding and preparing the data for modelling, the modelling itself, the different algorithms used and, finally, the results evaluation.

3.1. DATA UNDERSTANDING

At this stage we analyzed the dataset to understand its potential and limitations. We have used the Pandas profiling to have an overview of the dataset: what variables are in the dataset, what they mean, number of variables (29 features, from which 11 categorical and 18 numerical as shown on Table 3.1) and observations (10.000 customers), how the variables are distributed (there are some skewed variables) , if there is noise, if there are missing and/or duplicated values (none) , which of these features are relevant for the final goal and which features are redundant.

We have also looked at the metadata file provided to understand the meaning of each feature to understand their relevancy in the project – in here we flagged feature *LTV* (Lifetime value of the customer) for posterior analysis since there is no clear information on how it was calculated and what does it exactly mean.

Numeric	Categorical
Dayswus, Age, Edu, Income, Freq, Recency, Monetary, LTV, Perdeal, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebPurchase, WebVisit, Access	*Custid, SMRack, LGRack, Humid, Spcork, Bucket, Complain, Mailfriend, Emailfriend, Kidhome, Teenhome*

Table 3.1 - Numerical and categorical features.

3.2. DATA PREPARATION

This is the stage when the input data for clustering is prepared, so we have ensured the data meets the requirements for this purpose [2]: Numerical variables only, data has no noise or outliers, data has symmetric distribution of variables, variables are on the same scale, there is no collinearity, few numbers of dimensions.

On the first step of data preparation, we set the variable *CustId* as index. The next step was feature engineering where we built a feature representing the average spent per purchase (*Avg_ticket*) which was calculated by dividing the feature *Monetary* by *Freq*.

As states on the first paragraph of the present section, we have focused on cleaning only the numeric variables sub-dataset, since those are the ones contributing to the model. A Pearson correlation matrix was prepared to look at these correlations. From the analysis of full Pearson correlation matrix, we have identified two groups of strongly correlated variables, as we can see in Figure 3.1:

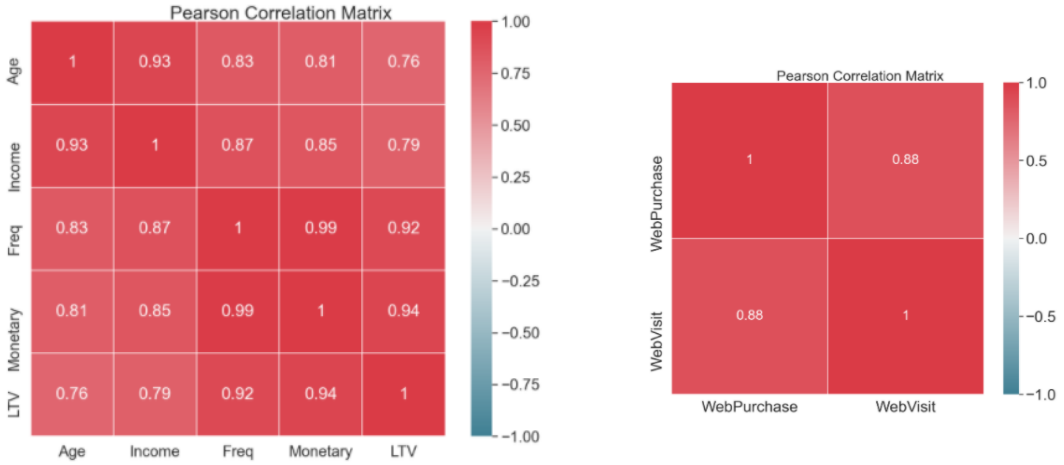

Figure 3.1 – Pearson correlation sub-matrices for numeric variables strongly correlated.

1. *Age*, *Income*, *Freq*, *Monetary* and *LTV* – to avoid redundancy and overfitting we kept only *Freq* feature, considering it is the most relevant for the analysis.
2. *WebPurchase* and *WebVisit* – for the same reason as stated in the previous point, we decided to drop *WebPurchase*.

After dropping these five features, we have checked once again the Person correlation matrix, which can be found in the notebook provided with the report.

We have then looked for missing and duplicated values and features which variance is lower than 10%:

- Missing and/or duplicated values: we concluded there were none in this dataset.
- Features which variance is lower than 10%: the outcome was only categorical variables, so we did not drop them as they will not contribute to the clustering.

To check for the presence of outliers on the numeric variables we looked at the boxplots for each numeric feature and concluded that features shown in Figure 3.2 seem to have outliers.

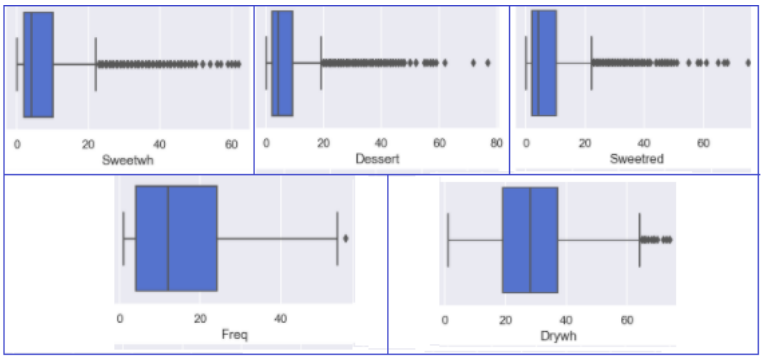

Figure 3.2 - Box and whiskers plot for features *Sweetwh*, *Dessert*, *Sweetred*, *Freq* and *Drywh*.

To remove these outliers, we tested applying six different methods for the entire dataset and using them individually or in combination. The number of observations removed by each method is summarized in Table 3.2.

Individual methods		**Combined methods**	
Z-score	1355 (14%)	1 method	1865 (19%)
Inter Quartile Range (IQR)	2848 (28%)	2 methods	628 (6%)
Local Outlier Factor (LOF)	634 (6%)	3 methods	506 (5%)
Isolation Forest	1663 (17%)	4 methods	412 (4%)
Support Vector Machines (SVM)	632 (6%)	5 methods	200 (2%)
Density based spatial clustering of applications with noise (DBSCAN)	293 (3%)	6 methods	23 (0,23%)

Table 3.2 - Number of outliers excluded with different approaches.

After testing removing based on 4 combined methods and realizing that this did not seem to make a big difference on the box and whiskers plot in terms of outliers, we have dropped this approach and decided to apply Inter quartile range (IQR) only on the five features mentioned above (*Sweetwh, Dessert, Sweetred, Freq* and *Drywh*). This way we have removed 354 observations, representing 3.5% of the dataset. The resulting, clean, box and whiskers plot for each feature can be found in the notebook.

In order to prepare the data for clustering, we have one-hot-encoded categorical variables (to transform in binary values) and applied the StandardScaler so that the dataset will have a mean value of 0 and a standard deviation of 1.

We have finished the data preparation with the following 14 numeric variables for clustering: *Dayswus, Edu, Freq, Recency, Perdeal, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebVisit, Access and Avg_ticket*.

3.3. CLUSTERING

As previously stated in section 2.6 - Project Plan we have opted for using K-means clustering as it is the most reliable and efficient method. In this process we have created a set of functions that can be consulted in the notebook. From these we highlight the following outcomes from the functions:

- Inertia plot, average silhouette plot and Davies-Bouldin plot – showing the dispersion of the points within the cluster, how well each object lies within the cluster and a ratio between distances within the cluster and distances between clusters, for the different numbers of clusters to facilitate the decision on the optimum number of clusters.

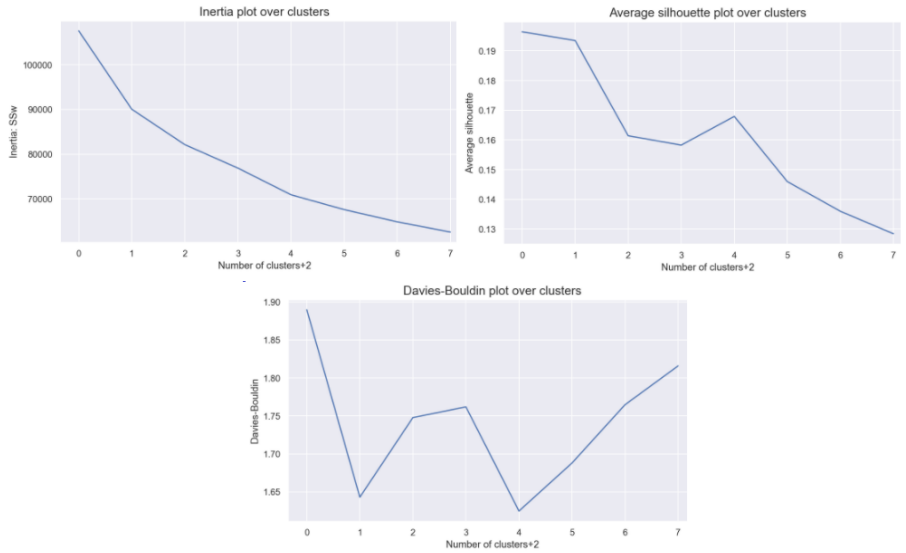

Figure 3.3 – Inertia, Average silhouette and David-Bouldin plots.

- Silhouette plot – shows a coefficient for the clusters' quality depending on the number of clusters chosen.

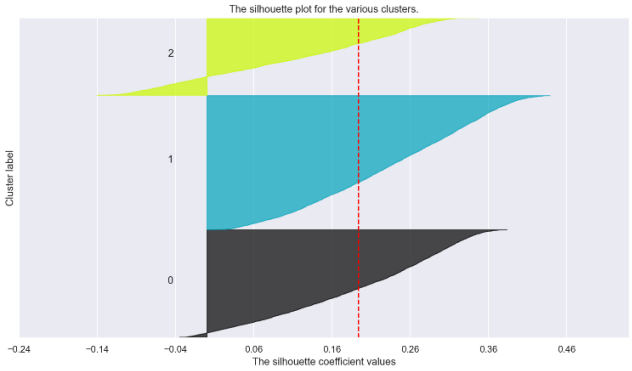

Figure 3.4 - Silhouette plot.

From the analysis of the plots above, we have chosen to proceed with 3 clusters and produced their profiles, showing the cluster's means for each feature and the clusters' absolute frequency, which enabled us to confirm the clusters are well balanced.

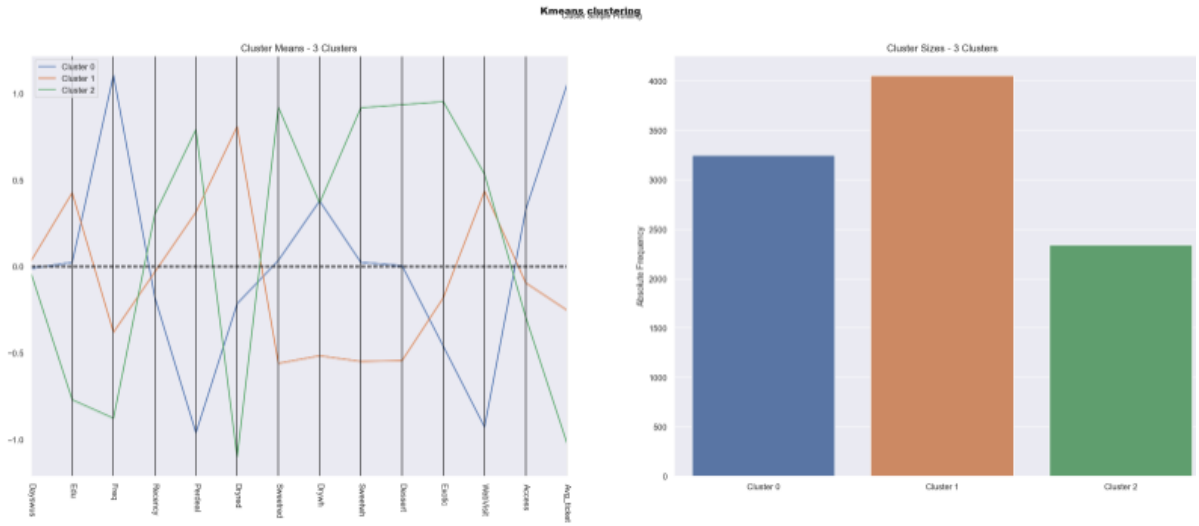

Figure 3.5 - Clusters profile.

It is worth to point that we also tested Hierarchical clustering whose outcome was reasonable, however, because K-means showed a higher R2, we have opted for proceeding the analysis with the latter method.

3.4. EVALUATION

To evaluate the quality of our clusters, we have confirmed visually, in two dimensions, that the clusters are well defined (Figure 3.6, on the left) and the position of its centroids (Figure 3.6, on the right) through the T-Distributed Stochastic Neighbor Embedding (t-SNE) plots.

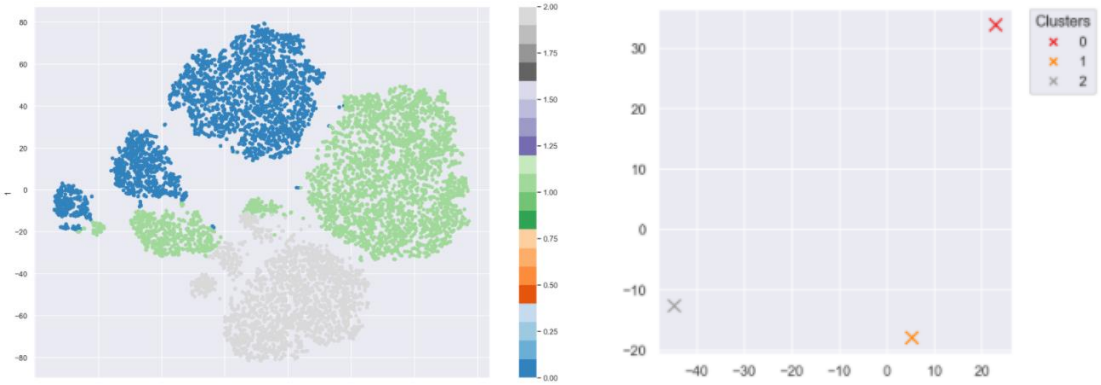

Figure 3.6 - t-SNE plots.

In addition, we calculated the R2 metric for K-means that presents a result of 0,36, higher than hierarchical cluster that presents a result of 0,31.

After splitting the data into training and test, we also applied a decision tree classifier to test our solution. It was able to predict 93,94% of the customer correctly. We also got the feature importance of each variable in predicting the cluster. In terms of feature importance, *Avg_ticket* and *Dryred* stand out as the most relevant features to the target. The full ranking can be found in the notebook provided.

4. RESULTS EVALUATION

With this model we were able to identify three different segments of customers who have different characteristics and purchase behavior.

Segment 1 (Cluster 0) is the group of WWW should prioritise since those are the ones that spend more and more often. They also tend to acquire more accessories when compared to the other groups however they do not purchase online or visit the website as often as the other groups. For those customers, the contact should be closer and personal, so we suggest approaching this segment via SMS or phone calls. This group prioritizes quality over price and is willing to pay more. Promotions and discounts were discarded in stores since this segment does not seem to react to items' discounts.

For this segment, we suggest the creation of small in-store events, such as wine tasting evenings and small workshops. We would like to explore more the customer experience as wine lovers in addition to expanding knowledge about the product consumed. To promote these events, we suggest implementing the following strategies:

- As a way of attracting new customers, the sales of tickets for the events would be processed in the following ways: a) the first customers to confirm would pay a lower price on the ticket or b) the customer would get 10% off the ticket price when inviting another potential customer.
- The idea with these events is to increase the working capital and encourage the sales of the least popular wines (*exotic*/*sweetred*/*dessert*).

- On the events days we suggest having some personalised accessories to be distributed, such as glass of wine or cork stoppers with the WWW branding.

To stimulate the frequency beyond the days when events don't occur, strategies such as buying a best seller and worst seller get an accessory would be adopted. Regarding the wine preferences of this segment, its customers prefer dry wines (*dryred* and *drywh).*

Segment 2 (Cluster 1) is characterised by highest level of education and a high number of accesses to the website and online purchases. The strategy adopted for this segment would be contacting the customers via e-mail as they use the internet as their main way of communication. Promotional emails would also include information about the product consumed (e.g. curiosities, composition, benefits, ideal consumption rate, recipes, etc.), the production chain and distribution. In that way, customers would have knowledge of the company purpose in addition to its concern regard quality and commitment to the consumer.

It was also noted that customers on segment 2 tend react to products' discount. Considering that the *dryred* product has the highest sales frequency in this segment, there is no need to promote it. The promotions are made with the second-best seller (*drywh*) combined with the product with lower sales alternately, i.e. *drywh* + *sweetred*, *drywh* + *sweetwh* or *drywh* + *dessert*.

Finally, customers on segment 3 (Cluster2) are the ones that purchase least frequently, have the highest percentage of online purchases, the highest volume of website visits and are also the youngest group, with the lowest income and money spent per purchase. They are also real lovers of exotic wines. Since this is the segment with the highest percentage of purchases of products on discount, we propose two different approaches:

- Send sporadic promotional links to the client's account on the websites and app, there would also be gift vouchers to be given on holidays and promotion packages including the second-best seller combined with the least bought products.
- Give the customers the opportunity to indicate new customers. For each new consumer referred, the current customer would accumulate points that to be converted into discounts on future purchases.

These strategies aim to increase the frequency on the websites and to attract new customers through referral.

For all segments, three categories could be created to classify customers according to the value of accumulated purchases (regular, gold and premium customers).

We suggest the creation of a loyalty program for gold and premium customers who would be sent 2 types of wine monthly, based on their preferences. In addition to this benefit, the premium customers would also have access to pre-sales and exclusive accessories. The transition from one category to another would occur as the value of purchases increases.

5. DEPLOYMENT AND MAINTENANCE PLANS

5.1. NEXT STEPS

- If it does not exist yet, create a customer account.
- Use the telephone (either call or SMS) as the main way of contacting customers on segment 1 and create in-store events for these customers as well.
- Create promotions for segments 2 and 3 when the customer buys the second largest sales product combined with the product with lower sales.
- Send sporadic promotional links for customers on segment 3, to their client account and app and give them gift vouchers on holidays.

- Create a points program for segment 3, which the customer that indicates new customers would accumulate points to be converted into discounts on future purchases.
- Create the three customers categories proposed in the section above (regular, gold and premium customers) and start the loyalty program proposed for gold and premium customers.
- Monitoring the model performance for the current and new customers.
- Start using digital marketing to reach new customers (e.g. Tweet Sentiment Visualization: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/).
- Implement the app provided by D4B to help the segment identification of current and new customers. It might be necessary to install some free programs to run the app.

5.2. APPLICATION

The application developed will enable WWW to simulate which of the presented three segments the customer belongs to, in a user-friendly environment. By simply updating the name of the CSV file containing the customers we want to simulate, we will get the output "The customer belongs to the cluster X".

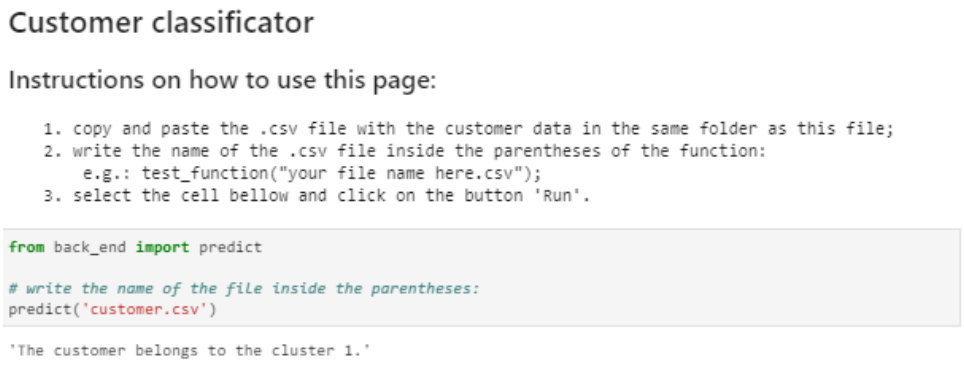

Figure 5.1 - Customer classifier application.

6. CONCLUSIONS

As state in the section 2.2 – Business objectives, two main objectives of this project were to identify the key characteristics that best distinguish the customers and understand which and how many customer segments there are in the provided database. Our final solution was able to detect 3 segments of customers. We were able to describe the key characteristics of customers in each segment in the section 4 – Results Evaluation. Also, one of the expected outcomes of this report was suggestions of marketing strategies and business applications for the findings. Marketing strategies were presented in section 4 – Results Evaluation and business applications were recommended in section 5 – Deployment and Maintenance Plans.

In addition, we set some risks on this project. One of these risks is the model performance, as we have been working with less than 3% of the entire data base. The model has improvement margin if we get additional datasets to test its performance.

We hope WWW will be satisfied with D4B work and we can continue working together.

7. REFERENCES

[1] Chapman, P, Clinton, J, Kerber, R., Khabaza, T., Reinartz, T, Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0*, CRISP-DM consortium

[2] Ryzhkov, E 2020, 5 Stages of Data Preprocessing for K-means clustering, viewed 27 February 2021, <https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932>