MDSAA

Master Degree Program in Data Science and Advanced Analytics - Major in Business Analytics

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

Customer Segmention

Group C

Catarina Moreira, number: 20201034

Luisa Barral, number: 20201045

Madalena Valério, number: 20200657

Yu Song, number: 20200572

February, 2021

INDEX

1. INTRODUCTION

An enterprise by the name of Wonderful Wines of the World is looking to modernize its approach to marketing and would like to make use of a database they started four years ago and that has accumulated around 350 000 customers. With this project and with the help of the CRISP-DM process we intend to advise the company on how to target the right people with the right messaging about its products, which will in turn increase the success of its marketing campaigns. Through the CRISP-DM framework we are able to structure our project in six different phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. In short, we will start by stating the business objectives, define the Data Mining problem, describe and explore the data, perform some data cleansing, build the clustering models, evaluate results and, finally, plan for deployment and maintenance.

All the deliverables are present in the following github repository: **https://github.com/MadalenaValerio/Business-Cases-for-Data-Science**

2. BUSINESS UNDERSTANDING

2.1. Background

Wonderful Wines of the World (WWW) is a company that specializes in selling exclusive wines from all over the world through catalogs, a website and ten small stores located in the USA. Its client base consists of individuals that have enough money to indulge in such unique products, which wouldn't otherwise be available outside their countries of origin. Besides understanding that its customers are wealthy and have a real passion for wine, not much else is known, consequently all decisions relating to promotions and pricing are solely based on intuition. All clients receive a catalog every 6 weeks and there are no loyalty programs or marketing strategies for different types of client, everything is mass marketed and there is no effort to understand what customers are interested in and what their buying behavior is.

2.2. Business Objectives

The goal of this project is to further understand the customers of WWW based on their demographic and behavioral characteristics. We want to be able to identify different subgroups of clients and analyze them so the company can make more strategic choices about what products they promote to certain clients, at what price the products should be sold to each group and where they should sell their products to make them more easily available. From a business perspective, we need to figure out how many subgroups of clients we have, who are they, what distinguishes them most, how should the company approach them from a marketing perspective and which are the most profitable ones WWW should prioritize.

2.3. Business Success criteria

The purpose of this project is to give insights into the behaviors of WWW's clients and what strategies can be used to interact with these clients, in a more effective way, given the information we have about their demographic features, the purchases they've made, what items they tend to buy and how they buy them. A useful outcome of this study would be to separate clients into understandable segments that could be analyzed, to know what the company's target market is and how they can reach new and existing customers in an informed way.

2.4. Situation assessment

The company has made available a dataset with 10 000 customers and 29 features. These are all customers that have made a purchase in the last eighteen months and were sent a test promotion for the silver-plated cork extractor. In addition, we have some metadata about the dataset.

2.5. Determine Data Mining goals

In Data Mining terms, we intend to perform a segmentation of the buyers of WWW through clustering techniques, in order to have defined subgroups of clients that share similar characteristics within their own group. This ensures the company makes use of its available database to plan what types of customers should be selected for certain promotions and how to target possible new consumers. The assessment of the models success will be dependent on how well the model differentiates between each cluster and how much insight it provides for marketing purposes.

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data understanding

First of all, we start to see the data that was given and to understand the meaning of each variable. After a brief analysis, we passed the given Excel file to a pandas DataFrame and, using some methods of pandas, we checked the data type of each variable and the columns present in the dataframe. As we can see, we only have float and int data.

Each line of our dataframe represented a customer. In each client we had information about their age, the number of days as a client, their income, the number of years of education, among others. These variables were very important in determining the different interests of the different customers. And, in this way, find new customers to purchase the product.

After analyzing both the features and the observations, we came across variables that had no meaning in solving this problem and, therefore, we eliminated them. We also eliminated Nan values and created new variables that we believe are important for the realization of this problem. In the next step, we will explain in detail the changes made so that we could have a good use in solving the problem.

3.2. Data preparation

As we are dealing with a huge amount of data (both features and observations), it's really important to prepare correctly the data.

Initially, we have in this dataset 29 variables that were given. During preparation, we eliminated some variables as well as removing missing values and outliers. We also created additional variables that we thought were important for solving the problem.

Feature Engineering

First of all, we removed the last row of our dataset because it was not relevant for our problem because it represented the average of every value of each column.
After analyzing the different features presented, we found that both the 'Access' and 'Rand' variables were not determinants, they were not significant to analyze and achieve a solution for the acquisition of customers. In this way, we have eliminated them from our dataset.

After eliminating these variables, we checked if the dataset had duplicate values that were not important. However, as we can see, it did not have any duplicated values.

We also created some variables that would help us to get to know the dataset better. We started by creating 'buyer' which is a binary variable that tells us whether or not the customer is a buyer.
After checking the high-correlation between 'Monetary' and 'Freq', we've created an 'AveragePurchase' variable that tells us the average purchase. In this way, we can have a better understanding of the taste of customers.

Correlation

Before proceeding to correlate the variables, we separated the metric features from the non metric features.

The purpose of making the phik matrix was to be able to detect the high-correlated variables.
If two variables have a high correlation, this means that they are variables that give us redundant information and, in this way, we can eliminate one of them. In this category, we found that the 'Monetary' and 'Freq' variables are high-correlated. However, instead of eliminating one of them, we found it more beneficial to create a new variable (explained in *Feature Engineering*).

As we can see, in **Figure 1**, 'Custid' and 'Dayswus' variables have a low relationship with all other metric features.

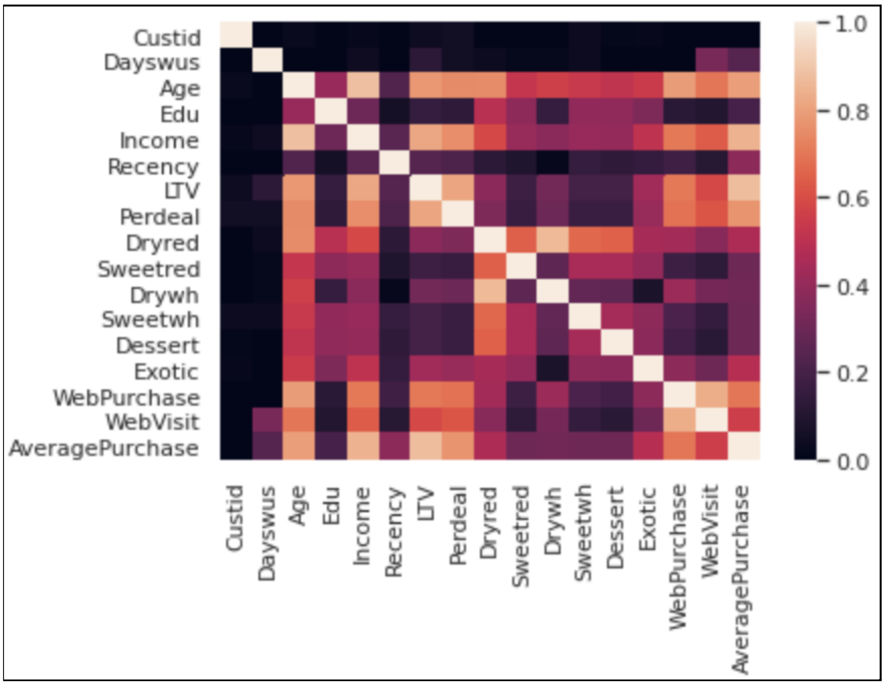
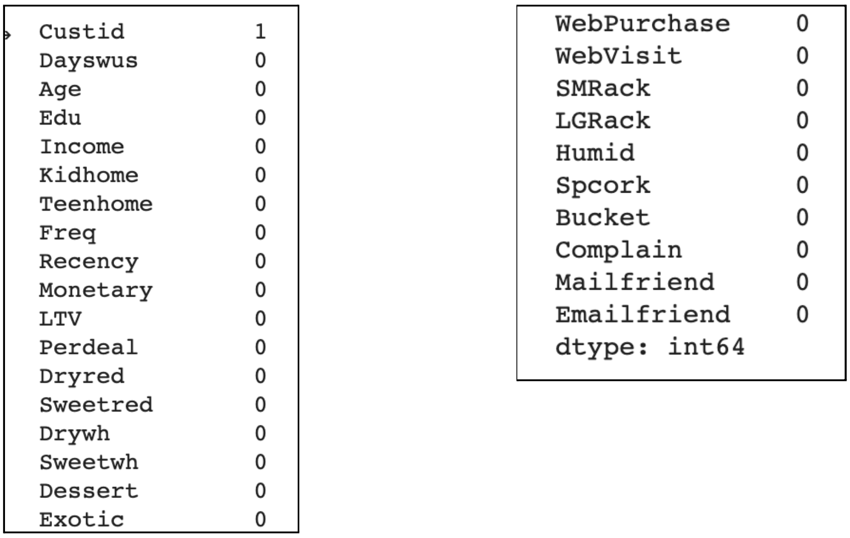


***Figure 1* -** Phik Matrix

Missing Values

We found that the variable 'Custid' had a Nan value. In this way, we eliminated that line as it is an insignificant number of Nan values. (Compared to the value of customers presented).



```
Custid          1
Dayswus         0
Age             0
Edu             0
Income          0
Kidhome         0
Teenhome        0
Freq            0
Recency         0
Monetary        0
LTV             0
Perdeal         0
Dryred          0
Sweetred        0
Drywh           0
Sweetwh         0
Dessert         0
Exotic          0
```

```
WebPurchase     0
WebVisit        0
SMRack          0
LGRack          0
Humid           0
Spcork          0
Bucket          0
Complain        0
Mailfriend      0
Emailfriend     0
dtype: int64
```

***Figure 2* -** Missing Values of the features

Data Cleaning - Outliers

At this stage of the project, we tried to solve the problems related to the quality of the data.
In this step, we started by drawing the boxplot for the metric features.
We realized, then, that there were values that were significantly different from the other observations - outliers. This type of value can cause some problems in the distribution of the data. Therefore, we tried to apply the iqr method to eliminate the outliers of the variables. However, after applying this method, we found that we eliminated more than 3% of the total data. Therefore, this solution was no longer suitable because we would lose many values and, thus, compromise our project.

We then did a manual filtering. Looking at the boxplot, we remove the most outliers. After doing this procedure, we found that we deleted only 3% of the data.

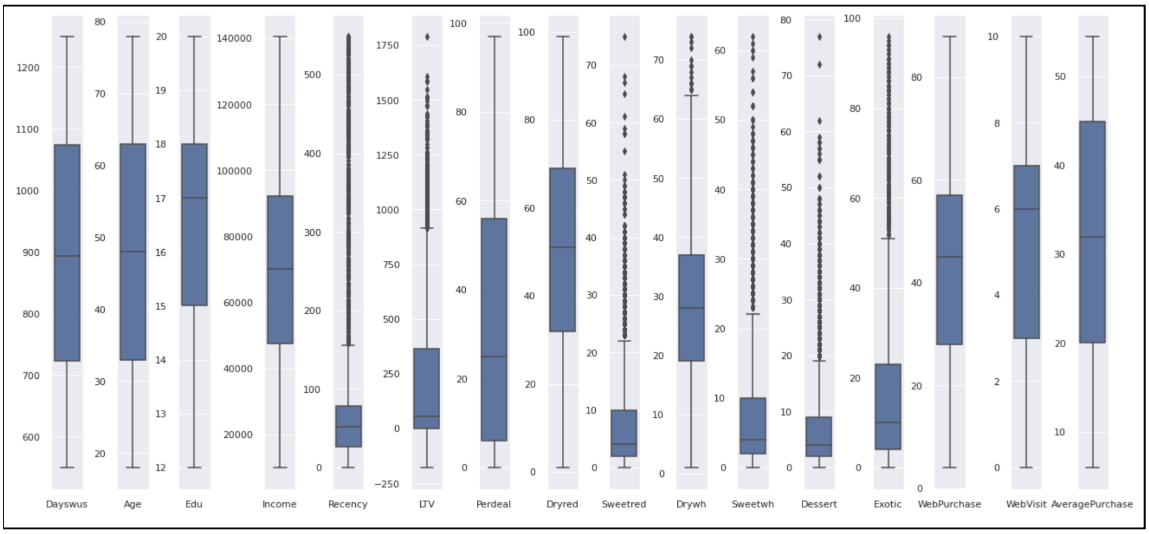


Figure 3 - Boxplot of the different metric features

3.3. Modeling

We took several approaches to find a solution to the problem.

It should be noted that, before applying each of the cluster algorithms, we normalized the data.

We started by using the K-means method. K-means is a clustering algorithm that tries to group in K groups the data points that are similar. The choice of K was made with the aid of the elbow plot and, therefore, we found that a good value to assign to K would be 4.

Posteriorly, we used the Gaussian Mixture model. This model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters[1].

Then, we implemented the hierarchical cluster model with K-means. Through the presented Dendrogram - ***Figure 4*** - we took the number of clusters to use in the K-means algorithm.

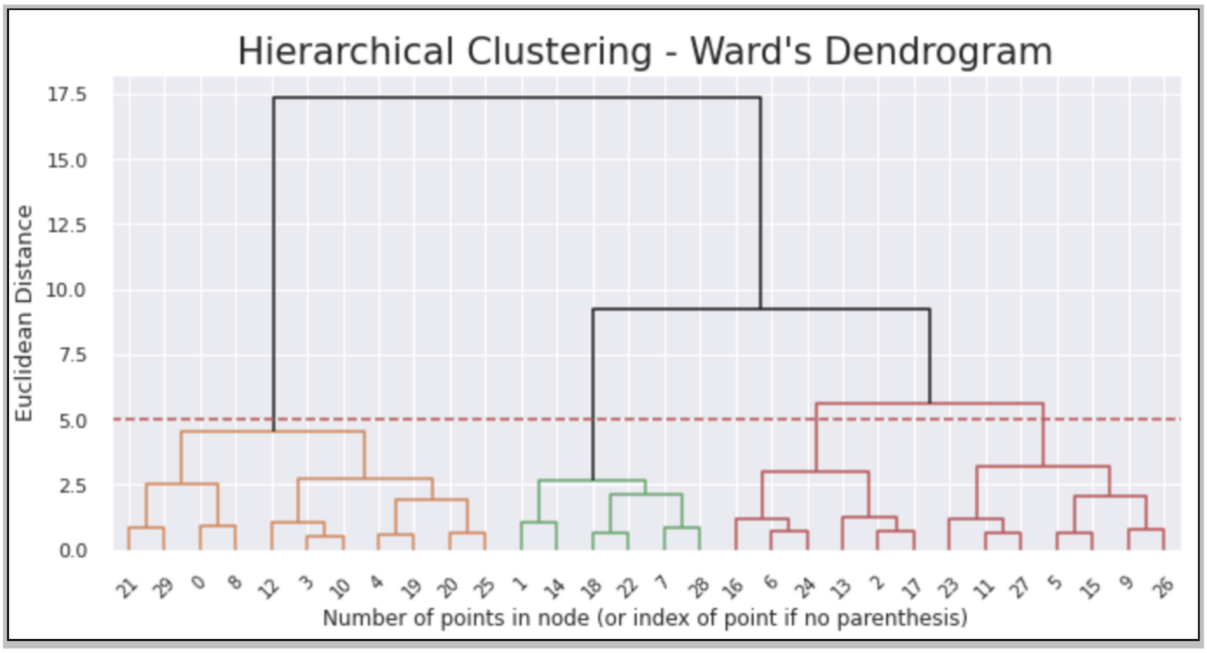


Figure 4 -Hierarchical Clustering - Ward's Dendrogram

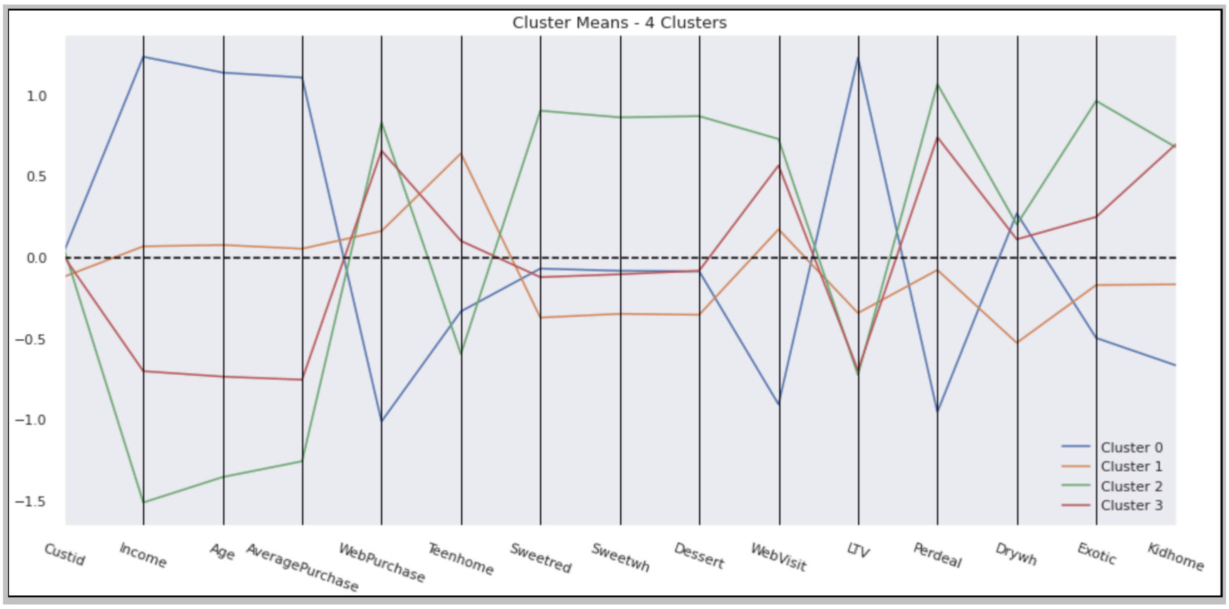


Figure 5 - K-means - Cluster means

Finally, we used the MiniBatch K-means. Its main idea is to use small random batches of examples of a fixed size so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence. As the number of iterations increases, the effect of new examples is reduced, so convergence can be detected when no changes in the clusters occur in several consecutive iterations [3].

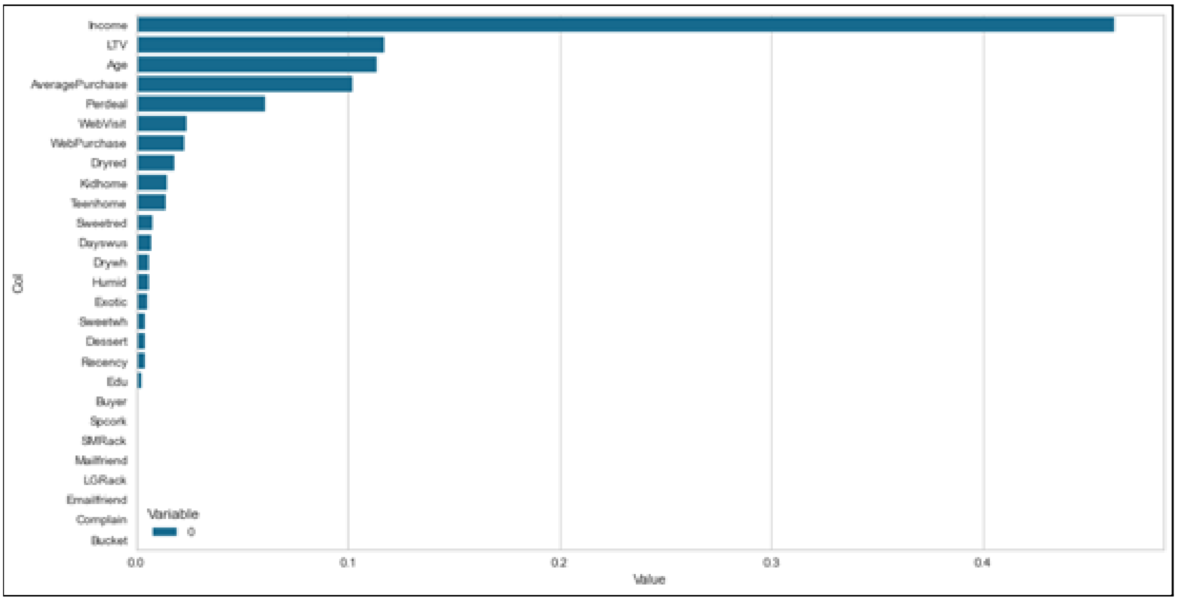
To conclude, we use the Hierarchical algorithm and K-means to solve our problem.

3.4. Evaluation

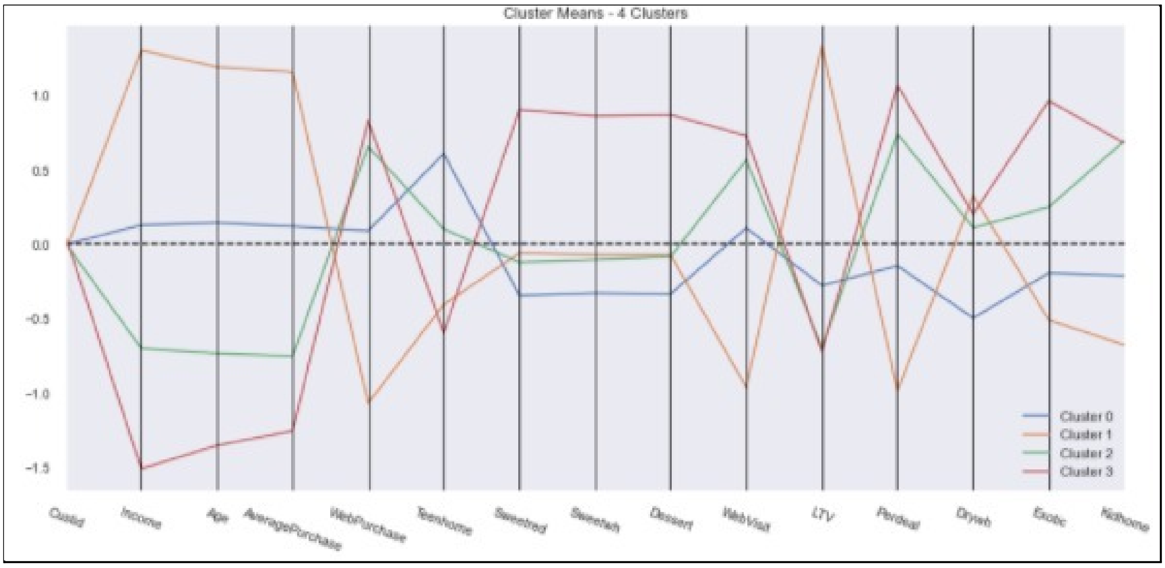
Wonderful Wines of the World is another experience of our group applying unsupervised learning algorithms. The dataset has 10,000 entities and 30 columns, we changed most data types, deleted some columns and converted to meaningful ones. For example, we create 'AveragePurchase' from 'Monetary' and 'Freq' to show the average purchase money of customers.

There are no missing values in the dataset. Because of the special meaning of some features, outliers don't mean the values are abnormal such as 'Monetary', 'Recency' and 'LTV'. So we did the filtering based on the boxplots.

Feature importance: the top 5 are 'Income', 'LTV', 'Age', 'AveragePurchase' and 'Perdeal', among which 'Income' is 3 times higher than the second important feature.



After trying different algorithms, Hierarchical clustering was the optimal clustering solution for Wonderful Wines of the World case, because its results were easy to understand and self-explanatory. The final clusters represent well segmented groups for re-engagement marketing purposes such as:



4. RESULTS EVALUATION

Cluster	Marketing Strategy
Cluster 0: Middle age clients with above average income, purchasing lower quantities of most wine.	Pay more attention to our middle age clients and cultivate them as our high value clients in the future.
Cluster 1: Older clients with highest income not sensitive with discounts and normally less shop online. Highest percentage of dry white wine consuming and average other wine purchasing.	Keep maintaining our most lifetime value clients through supporting various good quality wines and encouraging them to visit our web to purchase online.
Cluster 2: Young adults with lower income with teenager at home, purchasing more during discount on the internet with lower price	Every month or so provide them with discounts to be used online in a selected range of dry and exotic wines as well as notifications of promotions.

Cluster 3: Teenagers with lowest income and kids at home, purchasing more during discounts on the internet with lower average prices are more likely to buy all kinds of wine especially exotic wines.	These clients spend more time visiting our webs, post more ads on emails with various kinds of wine and attractive discounts.

DEPLOYMENT AND MAINTENANCE PLANS

During the execution of the project we detected some problems in our data which can be solved, in future data collections and cleaning. The first thing we noticed was that the last row in our dataset was not compatible with the rest of our data, it represents the average of every value on that column. Additionally an average purchase could be created and provided to allow customers to be aggregated based on their average purchase value, this would be useful for a better understanding of the clients as well as design a specific marketing strategy to each group. Some changes were also made to the data types of some variables like “Custid”, “Dayswus”, “Age”, “Edu”, “Income”. The group was not able to extract any valuable information from some variables like “SMRack” and “LGRack”, probably because they represent a really small part of the customers, 8% to be more exact. We believe that an investment should be made on the data collection process, related to these variables. Other than that, we felt that most of the variables that we had available for modeling were useful and adequate.

After deployment we should monitor the data, since it could change and therefore diverge from our original model, as well as our model performance, if it is still valid and accurate. For the monitorization of the data we should review some descriptive statistics, data types, missing values and compare the new data with the original data. Similarly to the data monitorization, we can also monitor our model, where we will compare the distribution of the labels of our original set versus the new data.

Since this monitorization requires a significant amount of time for analyzing models in production, we suggest the use of a monitoring system such as the Domino Model Monitor (DMM) where Data scientists can focus on value-added projects and receive alerts when production models degrade. This platform provides an easy-to-understand dashboard we can use to assess the business impact and proactive actions.

5. CONCLUSIONS

By following the CRISP-DM methodology we ensured that our plan to solve WWW’s issues was well structured and could be easily implemented by the company. After understanding the problem that needed to be solved, setting the goals we wanted to accomplish with this project and performing the necessary feature engineering and data cleaning, we implemented various clustering techniques, and ended up choosing the one that best met our business objectives, which was a combination of the Hierarchical algorithm with K-means. With this model we ended up with four clusters that were easily distinguishable between each other. The features used to describe each segment were a mixture of demographic such as Income, Age, whether they have teens or kids living at home, and behavioral characteristics such as what their average purchase is, how much they use WWW's website, what types of wines they purchase, if they tend to buy products on sale and their lifetime value to the company. With the identification of different customer segments and marketing strategies to approach every client group in a more personalized way, we expect WWW to increase its user engagement, thus creating more value for the enterprise.

5.1. Considerations for model improvement

We found some risks to this deployment that should be documented and mitigated.

In order to improve our model, the group suggests an investment on the data collection process which will gather substantial benefits for this application. This investment will provide us with more details about our customers and therefore construct better marketing strategies to each cluster of customers. To support continuous improvement, the group suggests the use of Excel spreadsheets to track improvement projects, PowerPoint files to document the strategy, and a document management system to capture and store training content and best practices.

6. REFERENCES

[1] Scikit-learn.org. 2021. *2.1. Gaussian mixture models — scikit-learn 0.24.1 documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/mixture.html> [Accessed 28 February 2021].

[2] Scikit-learn.org. 2021. *sklearn.cluster.MiniBatchKMeans — scikit-learn 0.24.1 documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html> [Accessed 28 February 2021].

[3] Upcommons.upc.edu. 2021. [online] Available at: <https://upcommons.upc.edu/bitstream/handle/2117/23414/R13-8.pdf> [Accessed 28 February 2021].