BUSINESS CASES WITH DATA SCIENCE

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

Wonderful Wines of the World

Group F

Lorenzo Pigozzi	m20200745
Nguyen Huy Phuc	m20200566
Ema Mandura	m20200647

March, 2021

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

INDEX

1. INTRODUCTION

Every business's goal is to increase income and decrease expenses. To do that, broadening the customer network is a crucial step of the process. In order to survive in the today's market, companies are forced to adapt to modern technologies. Data mining techniques are increasing in popularity when it comes to optimizing marketing strategies.

Since advertising means tend to be costly, it is critical to do it in the right way and aim it at the right audience. When looking for new customers, some level of certainty needs to be present before money is invested in advertisement that might not result in conversion. Therefore, to find new customers, the company's best approach is to first understand their existing customers well.

For this project, the team will carry out customer segmentation, in order to best understand different interests, responses and participation levels present in the current customer database. The goal of this project is to track the behavior of existing customers and propose appropriate marketing strategies for potential customers, based on which customer segment they seem to fit best.

The data is available here.

Github repository: https://github.com/LorenzoPigozzi/Business_Cases/tree/main/Case%201

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Wonderful Wines of the World (WWW) is a company based in the USA focused on selling wines and wine-related accessories. All sales are conducted either over the phone, on the official website, or in one of the ten physical stores located around the USA. The company is 7 years old, and its main effort is discovering wines from across the world and delivering their unique taste to the consumer.

So far, the marketing approach of the company is mass-marketing through catalogs. The catalogs are sent out once in 6 weeks and they each offer a selection of hundreds of products. However, the company has orderly kept a customer database for 4 years. The database contains about 350,000 wine enthusiasts, which are defined by their unique data, but not yet grouped in any way.

WWW has decided to utilize its collected data and has made available a subset of 10,000 customers from their database. All those customers are active – given that active means that they have made a purchase at WWW in the past 18 months. Providing this data, WWW expects to receive a business solution involving customer segmentation and a suggested marketing approach for each of the segments.

2.2. BUSINESS OBJECTIVES

2.2.1. Increasing revenue

One of the man objectives of the project is to increase revenue. Since WWW's marketing was based on sending out physical catalogs, money was spent on design, production, and the delivery cost. Given that not all customers always responded to the offer, WWW was losing money per customer. By using strategic marketing, these expenses can be cut.

2.2.2. Customer satisfaction

Another side-effect of mass-marketing is the lack of connection to the customer. Customers appreciate feeling understood, and they respond well to offers that correspond well to their interests and preferences. Using customer segmentation, a level of personalization can be added to the advertisement.

2.2.3. Sustainable growth

Future growth can be planned better by exploring past data. It is determined by the best possible use of resources. By getting a better understanding of customer purchasing trends, better predictions about future demand can be made, and financial resources can be organized optimally.

2.3. BUSINESS SUCCESS CRITERIA

The usefulness of the clustering analysis project can be addressed by 6 key characteristics (Gavett, 2014):

1) Identifiable: The customers in each cluster should be well distinguished by their key features, like demographics (Age, Income, Education) or behavior (wine-type purchases).

2) Substantial: It's usually not cost-effective to target small clusters — a cluster should be large enough to be potentially profitable.

3) Accessible: The defined clusters should be feasible to connected through their favorite communication and distribution channels.

4) Stable. In order for a marketing effort to be successful, a cluster should be stable enough for a long enough period of time to be marketed strategically. Thus, the features used for clustering should be consistent and concise by definition through time.

5) Differentiable. The customers in a cluster should have similar characteristics that are clearly different from the ones of other people in other clusters.

6) Actionable. The defined clusters should be able to be applied in business processes through an application that help business stakeholders easily access to the insights and action plan gained from the clustering analysis

2.4. SITUATION ASSESSMENT

For this project, the most used and valuable resource is the provided dataset. The data provided meets most of the quality criteria.

The data is complete – there is no missing values on attribute level, and no imputation needs to be done. However, some imputation was done on the dataset before.

The data is accurate – the data accurately reflects real-world situations.

The data is reliable – there is no inconsistencies, and WWW is trusted as a source, as it is in their interest to give consistent and accurate data

The data is relevant – most of the attributes are relevant, but due to some high correlation, some of them can be considered redundant.

The data is timely – it is up to date.

WWW also provided their full support during the project, as their IT team was available for questions and assistance during the full length of the project.

2.5. DETERMINE DATA MINING GOALS

The main data mining goal is performing customer segmentation through the data mining process. The aim is to identify variables that best differentiate customers from each other and apply a clustering algorithm to them. It is important to find a number of clusters that naturally occurs in the data. This number needs to be small enough for it to be feasible to make that many different marketing campaigns. The final clusters should be easy to describe by their unique characteristics.

3. PREDICTIVE ANALYTICS PROCESS

In order to complete this analysis, the IT department of the W.W.W. provides us a dataset containing the information of the customers who have completed a purchase in the last 18 months.

3.1. DATA UNDERSTANDING

The dataset contains 10001 observations and 30 variables.

Opening the file in excel, we quickly discovered that the last column and the last row are generated by a previous analysis conducted probably by the IT department itself.

The first column, *Custid*, can easily be identified as the index of the dataset, because contains exactly 10000 unique values.

For the other columns, we can identify 3 different main groups of variables.

Characteristics of the customers		Product's interest by each customer (in %)		Further infos about the purchases (binary features)	
DAYSWUS	number of days as a customer	PERDEAL	% purchases bought on discount	SMRACK	1=bought the small wine rack $50
AGE	customer's age or imputed age	DRYRED	% of wines that were dry red wines	LGRACK	1=bought the large wine rack $100
EDUC	years of education (may be imputed)	SWEETRED	% sweet or semi-dry reds	HUMID	1=bought wine cellar humidifier $75
INCOME	household income (may be imputed)	DRYWH	% dry white wines	SPCORK	1=silver-plated cork extractor $60
KIDHOME	1=child under 13 lives at home	SWEETWH	% sweet or semi-dry white wines	BUCKET	1=bought silver wine bucket $150
TEENHOME	1=child 13-19 years lives at home	DESSERT	% dessert wines (port, sherry, etc.)	ACCESS	number of accessories (not SPCORK)
FREQ	number of purchases in past 18 mo.			COMPLAIN	1=made a complaint in last 18 mo.
RECENCY	number of days since last purchase			MAILFRND	1=appears on a purchased list of "mail friendly" customers
MONETARY	total sales to this person in 18 mo.			EMAILFRD	1=appears on a purchased list of "e-mail friendly" customers
LTV	Lifetime value of the customer				
WEBVISIT	average # visits to website per month				

3.2. DATA PREPARATION

Looking at the data, it seems that the IT department completed a very good work, as the data doesn't present particular issues.

There aren't missing values and duplicated rows; the features are all numerical, and we simply split them in metric and non-metric features. This split is necessary because we need to analyze them in different ways, that we will see afterwards.

First of all, we begin the data preparation checking the correlation among the variables. For this purpose, we use 2 different metric measures for the metric and the non-metric features; for the first group, we use the Spearman correlation, because we know that some features are left-skewed.

For the non-metric features, instead, we use a more appropriate measures for the binary or descriptive variables, called the Phi_k correlation and introduced by Karl Pearson, however, for the binary features we don't have a very high correlated variables, and so we decided to keep all of them.

Instead, for the numerical variables we detected a really high correlation among 2 pair of variables, *LTV*, *Monetary* and *Freq.* Thus, we keep only the *LTV*, dropping *Monetary* and *Freq*. However, we will also generate a new feature called *Average purchase* which is the division of *Monetary* and *Frequency*.

The second step of the data preparation that we realize is the transformation of the distribution. To solve the problem of the non-normality, we compute the Cox-Box transformation to the metric features, and this transformation seems to be really relevant and effective.

As last step, thus, we scale the variables using the Standard Scaler function of scikit-learn.

3.3. MODELING

In order to segment the customers, different techniques were been considered. For this phase of the Data Mining process, we focus only on the metric features.

We start the analysis computing the DBSCAN algorithm. The main purpose for the use of this algorithm is that we want to leverage the automatic detection of the outliers, and in this case, they seem to be not really relevant: the total number of outliers detected is 27. We decide to keep them in the data

The second algorithm that we apply is the Self Organizing Maps. The reason also in this case is simple: we want to use the functionality of the plots this algorithm provides to better understand the feature-space.

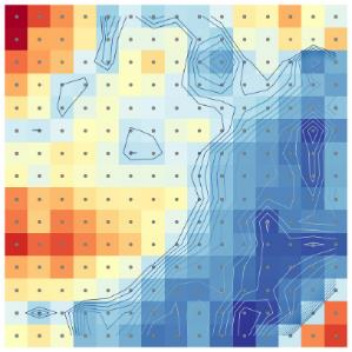

Figure 3. U-matrix

The final clustering-algorithm that we use is K-means. Using the elbow method on the inertia plot, 3 different clusters are clearly detected.

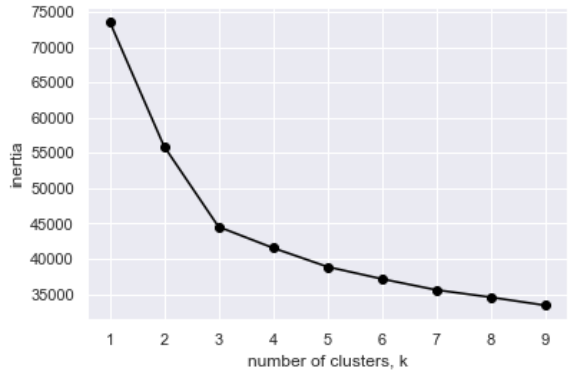

Figure 4. Inertia plot

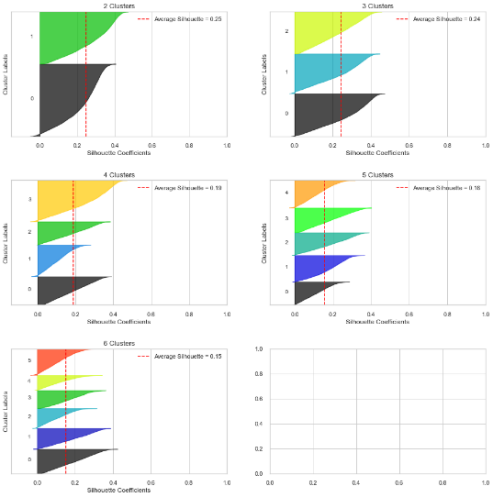

Figure 5. Sihouette score plot

Thus, finally we run the K-means algorithm with the parameter number of clusters equal to 3, and we obtain the cluster labels of the observations.

3.4. EVALUATION

The result obtained with the last algorithm described above can be measured using the R-squared. The value for this metric in our case is equal to 0.45. Furthermore, another positive conclusion of the results that we want to take into consideration is that the 3 clusters look really balanced, and this fact can be important for the marketing managers of the Wonderful World of Wine.

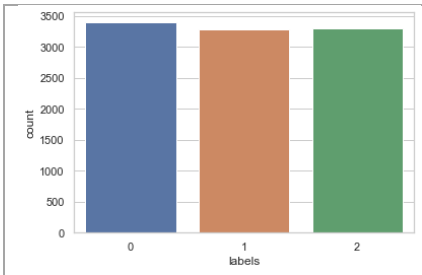

Figure 6. Size of the clusters

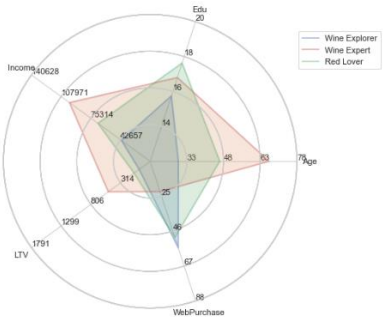

Figure 7. Key features

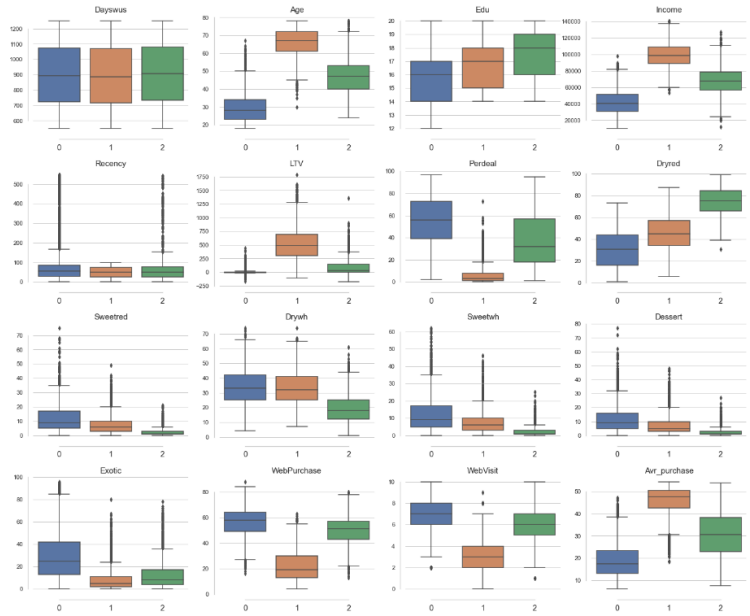

Figure 8. Distribution of each clusters by numeric features

Table: Analysis of numeric features

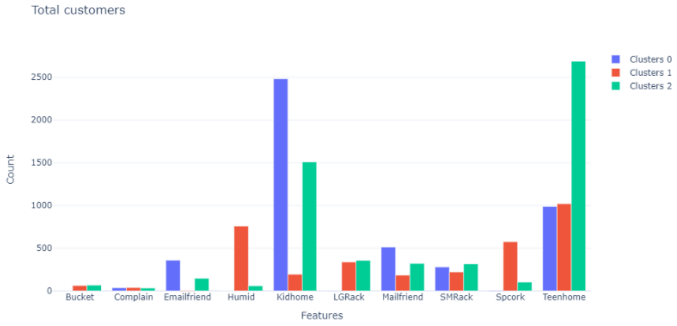

Figure 9: Analysis of binary features

For better understanding the clusters, a decision tree model is created to re-classify the dataset. The prediction result on test set (30% of the dataset) is 91% which is a very good result. Features importance are measured and the two features showing huge discriminative ability is `Dryed` and `Age`, follow by less significant features such as `Perdeal`, `LTV` and `Avr_purchase`.

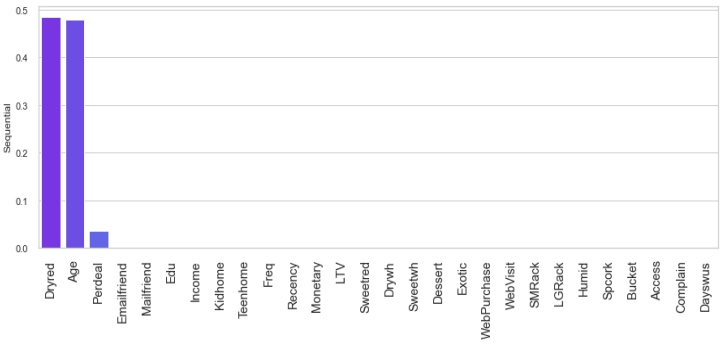

Figure 10: Features importance

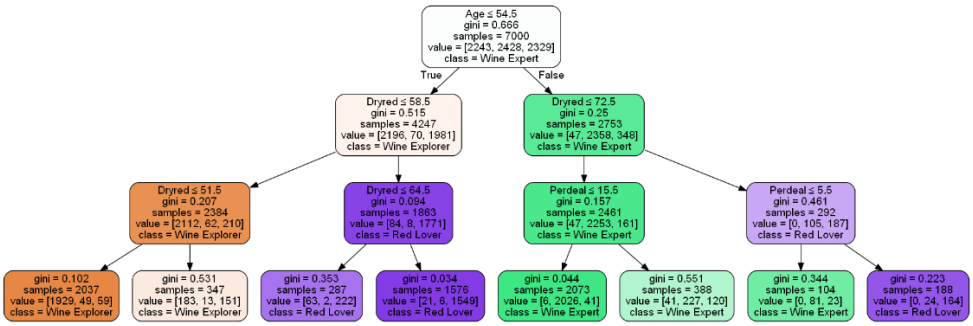

Figure 11: Decision tree classification model

Besides the classification of the outliers, the decision tree classifier provides a set of simple rules that allow future customer classification into each cluster by performing some simple queries. Below, it is illustrated what are the questions necessary to classify a customer into each cluster:

- **Wine Explorer**: If a customer has `Age` ≤ 54.5 and `Dryred` ≤ 51.5 then the customer belongs to this cluster with a **94.7% probability**;
- **Wine Expert**: If a customer has `Age` > 54.5, `Dryred` ≤ 72.5 and `Perdeal` ≤ 15.5 then the customer belongs to this cluster with a **97.68% probability**;
- **Red Lover**: If a customer has `Age` ≤ 54.5 and `Dryred` > 58.5 then the customer belongs to this cluster with a **94.81% probability**;

The top 3 important features `Age`, `Dryred` and `Perdeal` are all the business need to classify new customer with accuracy up to 91%. This show that the classification model is well generalized

To be more practical in the deployment of the model, we also build an API to test the model prediction on new customer to know which cluster they belong to. The API will be addressed in the deployment section

4. RESULTS EVALUATION

Interpretation of the clusters and marketing ideas

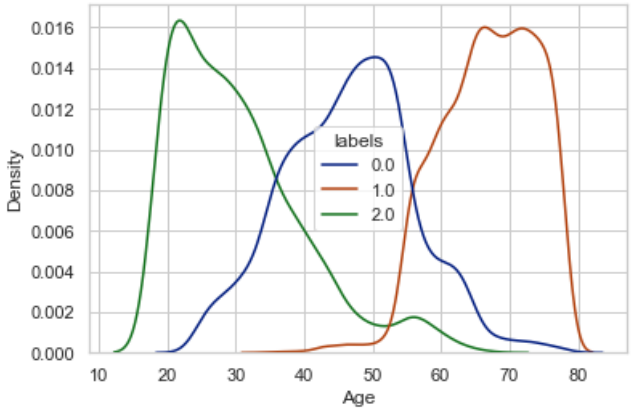

Figure 6. Distribution of the age by clusters

The Wine-Expert

The characteristics of this customer allow us to suppose that he is a very wine-addicted.

He represents the oldest group of customers of the company, and he spends a considerable amount of time, effort and money about wine-products. Indeed, he is the customer with the highest income, and also the most active considering the frequency of his purchases for the company products. Furthermore, he is the client who bought more recently and isn't interested to the discount opportunities that the WWW offer to his customers. The reason is simple: when the Wine-Expert wants a wine, he buys it. No matter the price. He's not looking for offers or cheap wines, instead he

wants to discover the different typologies and tastes that the company offers. Indeed, he is also the customer with the highest level of variety, considering the categories of wine and different flavors. He enjoys the sweet red wine, but also the white ones and after eating he usually likes to continue also with the dessert wine. In short: a real expert on the field.

For the marketing purpose, we suggest to reach him by mail or other no digital ways. He is not internet friendly. Furthermore, an idea could be to propose a loyalty card that can be sent at home after a defined good number of purchases completed. On the card is possible to accumulate points, and with those points it will be possible to achieve a particular accessory for free, or even better to participate to some wine-event as a VIP-member, for free as well. The number of points necessary for the gifts must be balanced with the costs of the accessories or event that the company has to pay for. In particular, in our opinion the wine expert could be really interested in wine-tasting event, maybe in the wineries themselves; he is also really curious about wine, and this is a fact to be considered: he would love to discover more insights about what he is drinking and the origin themselves of the products he enjoys the most.

The Red Lover

The key characteristic of this customer segment which helps us easily distinguish from other clusters is the dominant amount of red wine purchases

This segment has the customers in the middle range of age and income of the whole business customers. However, this segment group has the highest education level which can tell that they are the combination of sophisticated individuals that love to drink wine. About 80% of their purchases from the store are for dry red wines, the rest 20% is mostly dry white wine which indicates that this segment is very consistent in their taste. They are also actively looking for wine from the web. Moreover, good deals of wine are good way of gain attraction from them. Another good information about their personal life is that they are successful couples currently living with children from pre-teenage or teenagers

There could be several marketing approaches for this group. Assuming that this customer wants to have dinner with red wine together with their family, we can have a giveaway of promo-codes from partner restaurant for family dining or take-home orders. Since red wine is usually go well with cheese, another opportunity is to collaborate with web-food-companies (for example cheese or beef steak) and send out a workshop for wine tasting in collaboration with these food companies. Include rack cross-sale in web-purchase

The Wine-Explorer

The characteristics of a typical customer from this segment allow us to describe them as a beginner level wine enthusiast.

A customer from this group represents the youngest generation in the dataset. Most members also have low education, but this might be the consequence of young age – they might be active students. A typical customer is also defined by lower income, which can also be explained by age, as they are not yet well experienced. They are also web-oriented, which makes sense for people who grew up in the age of technology. The average customer in the group is more likely to have a young child living at home than a person in any other group. When it comes to shopping habits, they are the customers with the lowest frequency which in combination with their preference for discounted items makes their lifetime value also low. This pattern can be seen as a result of their low income. However, the customers in this group have good potential, given that they are curious about different types of wines, but mostly in the range of sweet and unusual wines. This group is not very interested in dry wines.

Concerning the marketing the approach, the suggested advertising medium for this group is social media. Given the low lifetime value of its members, it is important to keep the advertising expenses low, as acquiring new customers from this group will not increase the revenue much, but costly advertising methods will very likely lose money for WWW. Therefore, instead of using actual social

media ads, the suggested method is to simply use official WWW social media accounts for promoting products. One possible course of action is making an Instagram Giveaway, in which participants need to share a WWW post on their account, as well as follow the page and tag other users. The winner can, for example, get a discount coupon to use on the web shop. This will animate existing customers but also expand the reach to potential new ones. Also, for this group, physical catalogs would work better if changed to e-mail newsletters. Since those customers respond well to discounts, a special offer can be made and sent by e-mail to users that have not made a purchase in a specified time period. For new customers, offer a discount on the first purchase if they subscribe to the newsletter on the web store.

5. DEPLOYMENT AND MAINTENANCE PLANS

Deployment is a crucial part of every business project. A machine learning model will not be able to drive values if it is not effectively deployed in the real-world context.

The dataset collected by IT department is in a very good condition as no missing value and data error was found during the exploratory process. However, to ensure the quality of the predictive model in future use as well as scalability of the clustering model if there are more data fed in, any newly collected data should have the same quality as the current one.

Current objective achieved:

- Cluster analysis: a cluster analysis with insightful information about each customer segments to support on marketing activities. The clusters are well defined with detailed profiling carried out to understand the key features and characteristics that distinguish each cluster from the others

- Predictive model and testing environment: to deliver a predictive model that able to classify the segment of any current customer or predict that of any future customer with high accuracy. To test the deployment, we also create an API to predict the belonging cluster of any new customer. The API is currently under testing platform but the aim of it is to carry out a practical approach for any business stakeholder to easily understand the result of the project.

Future steps to successfully deploy the model:

- Develop fully functioned application: the final solution that able to be scale up to all the organization. The application should be able to allow continuous data fed in, continuous training and evaluation. The application will accept new data for prediction and validated results of new data as additional training data.

- Periodical maintenance plan should be able to continuous evaluate key metrics of the cluster analysis objectives which is, for example, the successful of marketing plan suggested by the analysis or new customer achieved through the findings.

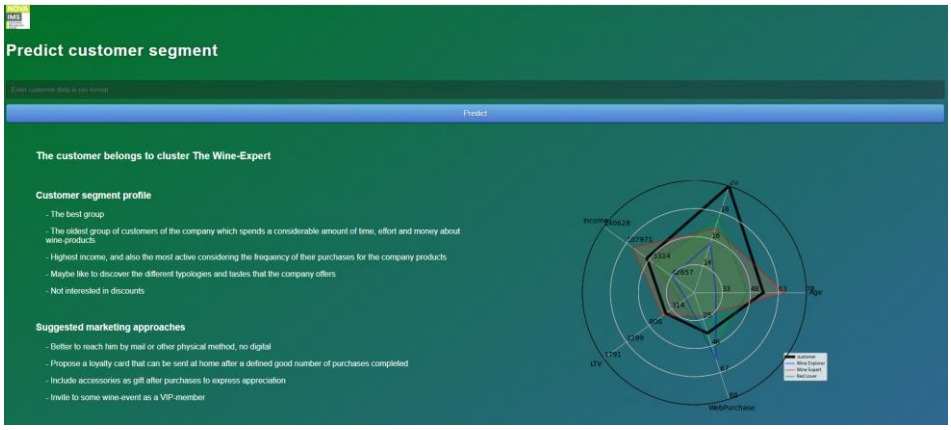

Figure 8: API

6. CONCLUSIONS

In order to support the marketing team of the WWW company, in this project we analyzed the customer profiles of the previous 17 marketing campaigns. The overall objective for the company was really ambitious: creating a new marketing approach based on data driven decisions.

We received from the IT department a dataset containing the information of 10000 customers.

After a brief preparation and pre-processing of the dataset, in which we cleaned and transform some variables, we started the customer segmentation analysis. Different clustering algorithms were used for this purpose: in particular, DBSCAN was used for the detection of the outliers, SOM was used for the understanding of the dimensional space of the variables and finally the K-means algorithm identified 3 segments of customers.

The 3 clusters detected were pretty well defined by a few numerical variables, nevertheless the binary features were useful for the profiling phase as well. Specifically, the 3 groups of clusters are characterized by the age and LTV variables and by the history of the products purchased, expressed in percentages.

We are confident that the results obtained and the marketing ideas suggested can be very useful for the company, in order to increase the sales and maybe also decrease the costs, throughout a marketing campaign more focused on the 3 different customer profiles.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

In order to improve this analysis and the results obtained, first of all we can suggest to use more observations: more observations available means always better analytical results.

Then, we recognize that the IT department is doing a good job in terms of storing data, even though we can suggest to insert more variables. For instance, a possible suggestion could be to insert in the dataset the last 3 products purchased by each customer, and not only the percentages. This information could be useful to understand if a customer uses the WWW service to purchase many times exactly the same product or if he likes to differentiate, or also to understand if there are some products that are usually purchased at the same time, establishing in this way some association rules among the products.

7. REFERENCES

1. Gavett, G., 2014. What You Need to Know About Segmentation. [online] Harvard Business Review. Available at: https:/hbr.org/2014/07/what-you-need-to-know-about-segmentation

2 https://en.wikipedia.org/wiki/Phi_coefficient

3 https://phik.readthedocs.io/en/latest/