

**BUSINESS CASES WITH DATA SCIENCE**

**MASTER’S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

**Predict Hotel Booking Cancellations**

Suggestions on Overbooking and Discount policies

Group Y

João César (20200669)

João Henriques (20200670)

Pedro Sancho Vivas de Castro (20200132)

Vilmar Bussolaro (20200268)

March 15th, 2021

INDEX

[1. INTRODUCTION 1](#_Toc66181052)

[2. BUSINESS UNDERSTANDING 2](#_Toc66181053)

[2.1. Background 2](#_Toc66181054)

[2.2. Business Objectives 2](#_Toc66181055)

[2.3. Business Success criteria 2](#_Toc66181056)

[2.4. Situation assessment 3](#_Toc66181057)

[2.5. Determine Data Mining Goals 3](#_Toc66181058)

[2.6. Project Plan 3](#_Toc66181059)

[3. PREDICTIVE MODEL FOR HOTEL CANCELLATIONS 4](#_Toc66181060)

[3.1. Data understanding 4](#_Toc66181061)

[3.2. Data preparation 4](#_Toc66181062)

[3.3. Model implementation 4](#_Toc66181063)

[3.4. Model Evaluation 5](#_Toc66181064)

[3.5. Model Selection 5](#_Toc66181065)

[3.6. Prediction of no shows 5](#_Toc66181066)

[4. RESULTS EVALUATION 6](#_Toc66181067)

[4.1. Prediction of cancellations 6](#_Toc66181068)

[4.2. Prediction of no shows 7](#_Toc66181069)

[4.3. Overbooking and Discount policies 8](#_Toc66181070)

[5. DEPLOYMENT AND MAINTENANCE PLANS 9](#_Toc66181071)

[5.1. Plan deployment 9](#_Toc66181072)

[5.2. maintenance 9](#_Toc66181073)

[6. CONCLUSION 10](#_Toc66181074)

# 

# INTRODUCTION

In this project we are asked to make a Customer Segmentation for Wonderful Wines of the World (WWW), a 7-year-old enterprise that seeks out small, unique wineries around the world and brings their wines to its customers. The goal is for the company to have a better understanding of their customers’ behaviors and identify the different segments of clients within their database.

The database we are studying corresponds to a random sample of 10 000 of the company’s active customers. These customers have made at least one order in the last 18 months and were selected from a test promotion made for an accessory: a silver-plated cork extractor.

We are following the CRISP-DM methodology to achieve the final product, which is a cluster solution that segments the customers in 5 distinct groups.

Business recommendations are presented following the characterization of each group.

# BUSINESS UNDERSTANDING

## Background

Wonderful Wines of the World (WWW) is a 7-year old company with the mission to impress their customers with high quality wines.

They have 3 distribution channels: a web site, ten small stores in major cities around the USA and catalogs (telephone). Several hundred selections are available in each new catalog, sent every 6 weeks.

WWW now has around 350,000 customers in their databases, which were mainly acquired through aggressive promotions in wine and food magazines. Those customers are highly involved in wine, entertain frequently, and have sufficient money to indulge their passion for wine.

Their current market strategy is mass marketed. All customers get the catalog, and there are no loyalty programs or attempts to identify target markets for cross-selling opportunities. They also have simple market reports, feedback from salespeople and intuition.

## Business Objectives

This analysis has as objective understanding the characteristics that discriminate the customers. We want to know how many different customer segments there are and the respective behavior, preferences and social economic differences so we can better understand the aspects with more room for improvement.

Knowing who the customers are (social demographics) as well as what are their buying patterns and preferences will help boost the marketing efforts from WWW.

We want to expand the purchases from the customers that eventually order less from the company, as well as keeping and motivating the best customers so they keep buying from the company and defining priorities for which segment should get more attention.

## Business Success criteria

In order to be successful, the project needs to distinguish between different groups of customers and provide enough insight to drive decision making in a better way that is done today.

Using an A/B testing approach, by assessing the variation (delta) in sales and revenue between a subset of customers that will use the new marketing approaches and those that will not, it will be possible to assess the impact the customer segmentation has on the company. This application will be considered a success if the result (yield) is positive. Similarly, it will be a failure if it brings no results and no impact on revenue compared to the benchmark (status quo).

By extrapolating that impact, it´s also possible to calculate the ROI of this specific Data Science project, which will surely pay itself multiple times.

## Situation assessment

For this project we have the WonderfulWinesoftheWorld.xlsx dataset with a sample of 10,000 customers from its active database. These are all customers who have purchased something from WWW in the past 18 months (after 18 months with no purchase, a person is eliminated from the active database). It was these 10,000 randomly selected people who were sent the test promotion for the silver-plated cork extractor.

Our team is composed of 4 Data Scientists, which had a week to prepare a 5 minute presentation to the C-Suite, as well as a 10-Page Report (which you are reading) and the accompanying code.

## Determine Data Mining Goals

To segment customers based on their characteristics, the best approach is to use Unsupervised Learning. By using a clustering method, we can select an appropriate algorithm to identify which sets of objects are more similar than others, therefore forming groups: Those groups (or clusters) are easier to comprehend and act upon.

But before bringing information to the surface, we need to do the basics: preprocess data. On feature selection the goal is to choose a set of variables that is both broad and relevant but that do not generate the curse of dimensionality. Highly correlated features should be dropped.

By using Data Mining knowledge from previously learned algorithms and techniques to determine the optimal way of making our clustering, we expect to create a K number of clusters that make sense to the business (between 2 and 8), that will have a R2 measure of at least 35% in the implementation.

The number of clusters need to be assessed by elbow methods, validated by the results as a viable number when achieving clear differentiation between groups of customers.

The goal of the Predictive Model for Customer Segment for New Customers after their First Purchase is deemed acceptable above 55%.

## Project Plan

1. Data understanding  
2. Data preparation  
2.1. Outlier removal  
2.2. Scaling  
3. Cluster analysis and selection  
4. Cluster visualization  
5. Predictive Model for Customer Segment for New Customers after their First Purchase

# PREDICTIVE MODEL FOR HOTEL CANCELLATIONS

## Data understanding

After taking a first look to our data, we found out the given dataset has 30 different variables. Two of them were irrelevant for our analysis: ‘Custid’, which was an individual ID for each of the 10 000 customers and ‘Rand’, a randomly generated value.

From these variables we framed the variable in three different types of values:

- Metric values, that included the variables 'Dayswus', 'Age', 'Edu', 'Income', 'Freq', 'Recency', 'Monetary', 'LTV', ‘Access’ and 'WebVisit';

- Binary values, composed by 'Kidhome', 'Teenhome', 'SMRack', 'LGRack', 'Humid', 'Spcork', and 'Bucket';

- Ratio values: 'Perdeal', 'Dryred', 'Sweetred', 'Drywh', 'Sweetwh', 'Dessert', 'Exotic' and 'WebPurchase'.

The data is very consistent, with no null values and at a first look, without much processing it, looks like there is not a big number of major outliers.

QQplots between variables showed they don´t follow a normal distribution.

## Data preparation

We deleted the last row since it was the averages of all columns.

When talking about outlier detection we used 6 diferent methods, and for a point to be considered an outlier it needed to be labeled as so in at least 4 of them. The methods were 5 standart deviations from the median, IQR (but with 6 instead of 1.5 since none of the variables followed a normal distribution), IsolationForest, Local outlier factor, yeo johnson transformation and quantile transformation. This method ended up deleting 144 rows (1,44% of the dataset), we consider this number to be an acceptable amount keeping the integrity of the dataset.

Values of Recency had a lot of values classified as outliers according the IQR method, and because of the long tail of the distribution the clusters were concentrated in a single focal point around the mean. We did a simple Data Engineering modification, capping the Recency value at 100.

## Model implementation

We tried two different approaches to clustering:

**3.3.1 Approach One - 2 K-means for customer profile perspective and taste profile perspective + Hierarchical Clustering for merging the two approaches.**

We performed three different methods to estimate the best number of K clusters, using Distortion, Silhouette and Calinski-Harabasz.

For the first cluster, using the variables selected for *Customer profiling* (cust\_prof = ['Dayswus', 'Age', 'Edu', 'Recency', 'LTV']).

That got us a k-means clustering with R2 of 0.53 and K=4..

Then we performed another k-means with *Products-Taste profiling* (taste\_prof = ['Perdeal', 'Dryred', 'Sweetred', 'Drywh', 'Sweetwh', 'Dessert', 'Exotic', 'WebPurchase']).

That got us a clustering solution with R2 of 0.56 and K=4.

Then we built an hierarchical clustering on top of that, to merge both perspectives. Using agglomerative clustering, the R2 of the cluster was 0.43.

**3.3.1 Approach Two – Single K-means**

By selecting both groups of variables, we built a comprehensive view of features, but taking into consideration the multicollinearity of some pairs to avoid the curse of dimensionality.

vc\_taste\_alt = ['Dayswus', 'Edu', 'Dryred', 'Sweetred', 'Drywh', 'Sweetwh', 'Dessert', 'Exotic']

vc\_custo\_alt = ['LTV', 'Recency', 'WebVisit', 'Perdeal']

The K-means cluster was implemented on vc\_taste\_alt + vc\_custo\_alt, a total set of 12 features, with 5 clusters and a R2 of 0.45. The number of K was also assessed using Distortion, Silhouette and Calinski-Harabasz.

## Model Evaluation

Regarding the Approach One, the total combined cluster R2 was 0.43. With Approach Two that value was 0.45. It had greatly defined clusters, and a higher R2. In alignment with the idea that the simple is always preferable over the complicated, we sticked to the Approach Two as the final solution.

## Model Selection

The final cluster solution was single K-Means of 12 variables mentioned on 3.3.1. The R2 achieved was 0.45.

## Prediction of no shows

The idea behind a Decision Tree is because it´s a white-box algorithm, that allows for an early assessment of who new customers might be.

By making its first purchase, a new customer leads to a small data collection, that fed into this algorithm will help the company on who to expect him to become, and how to target him accordingly. The decision tree used to classify new customers got an F1\_Score of 62%.

Different client acquisition campaigns paired with this prediction algorithm could diagnose if WWW´s current campaigns are acquiring the intended type of customer (A/B testing).

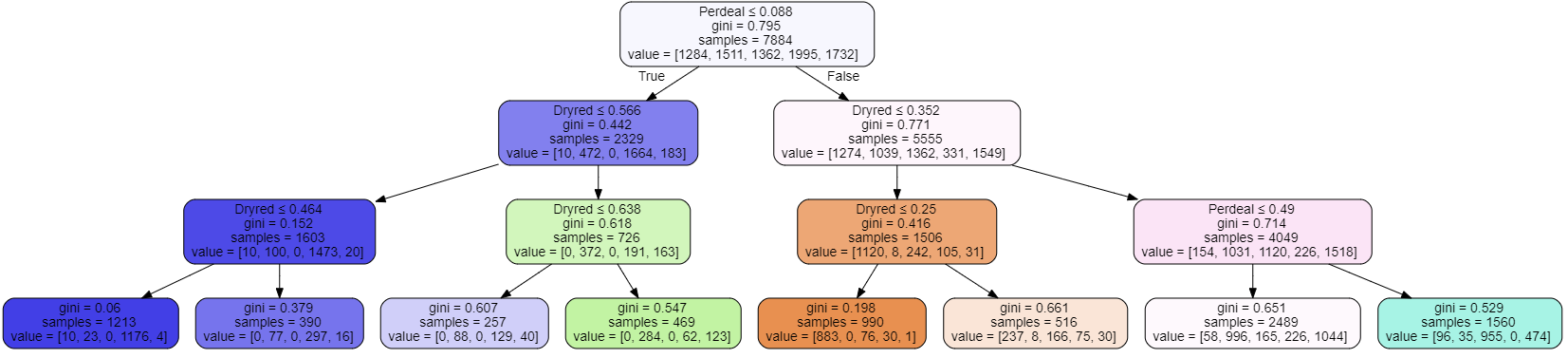


Figura 1: decision tree to predict costumer segment

# RESULTS EVALUATION

In order to evaluate in the most efficient way possible, for visualization purposes we will adapt and create some of the variables.

We joined the variables “Kidhome” and “Teenhome” into a new variable called “Childhome”. This variable is still binary and reflects if the customer has or not at least one child below 19 years old. Also, the same for “Mailfriend and “Emailfriend”, where we created the variable “Mailable” that returns the value 1 if the customer desires to be contacted either by mail or e-mail and 0 if else.

Also, we changed “Access” to include the variable “Spcork”, then created a variable named “bought\_access” that is 1 if Access different from 0, this variable tells us if costumer bought or not an accessory.

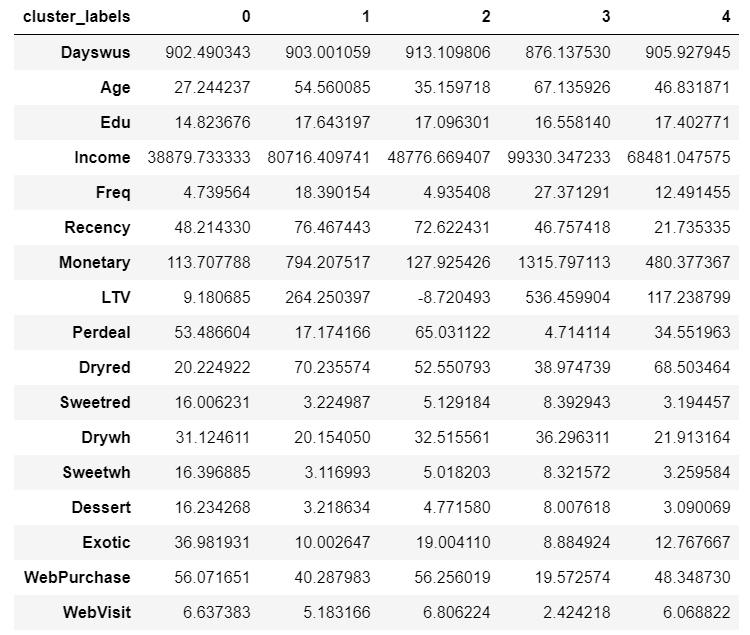


Figura 2: centroids dos clusters

## Prediction of cancellations

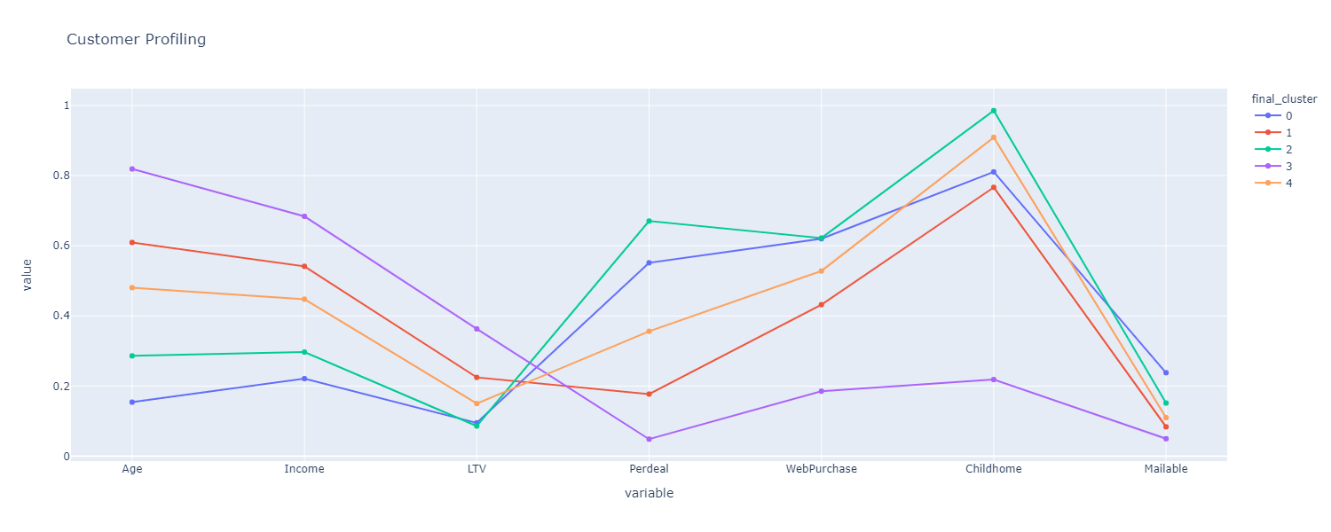


Figura 3: different variables side by side to create a customer profile

As we can see in the graph above, cluster 0 is the cluster with the lowest income and age, with an average of 27, therefore we called the customers of those cluster the “Young Crowd”. They have a very low life-time value, and we can also see that the group tends to prefer ordering online - more than 60% of the times - and often looks for the best deals with discounts over an eventual preferred wine.

In contrast, Cluster 1 is represented by customers with double the average, in terms of age and income, then the “Young Crowd”. This segment still makes a considerable amount of online orders, but takes less discount deals, which indicates that have more interest in wine and have more defined preferrable bottles of wine. We call this segment the “Silver” customer profile.

Cluster 2 is very similar to cluster 0, with a bit more maturity. Higher average age, income, but very similar online purchases. It is the segment with the highest ratio of deal sales and number of children, with almost all of the customers having at least one children under 19 years old. As they are the group that most searches for the best discounts, we will call them the “Bargain Hunter”.

The segment with the oldest people, with an average age of 67 years old, is cluster 3, with an average age of 67 years old. It is also the segment with the highest income and life-time value and for those reasons we called this segment the “Gold” customer profile, they are our best customers. This group is the more conservative and the most interested and experienced in wines, as they are the segment that buys more in the physical stores and less discount wines. Only 19,5% of the orders were made online.

At last, we have the cluster 4, called the “Average Joe” customer profile. This is the average people segment, that follows between the other four clusters and is the most represented segment, with approximately 25% of all the customers.

## Prediction of no shows

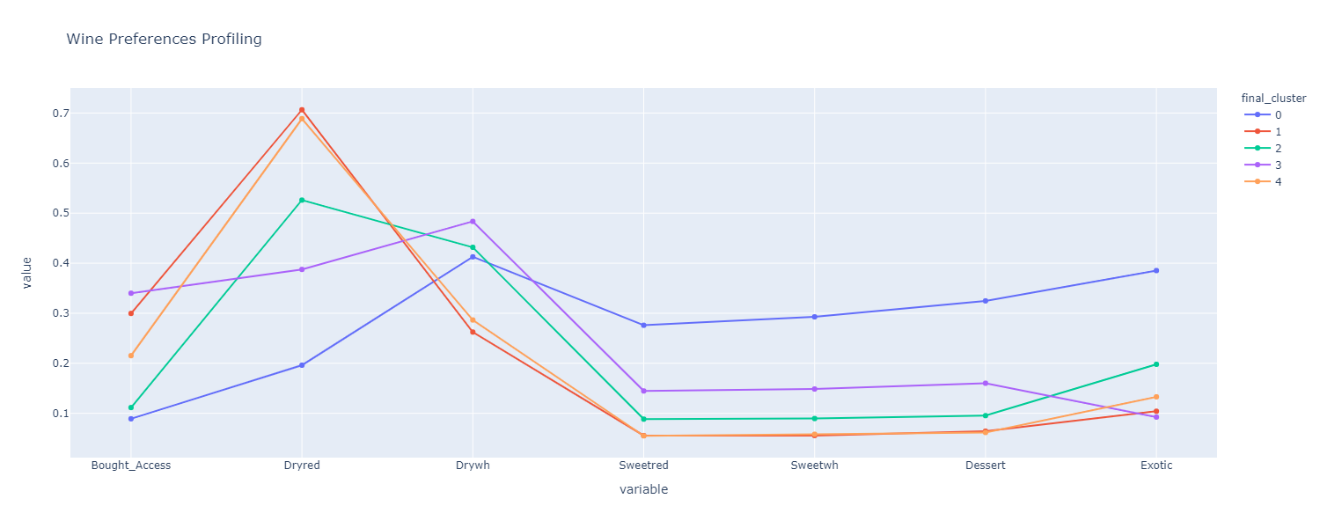


Figura 4: different variables side by side to create a taste profile

Regarding the wine preferences along our segments, we can observe that the variable “Bought\_Access” follows the same ranking as the variables “Age” and “Income”. The meaning of this is that the more the customer knows and is interest about wines, the more likely he is to own an accessory.

The dry wines are the favorite of the customers, being the red the clear winner. Around 70% of the wines bought from the “Silver” and the “Average Joe” customer profiles was red dry wine. In the other side, the white dry wine is the favorite among the “Young Crowd”, “Bargain Hunter” and the “Golden” customer segments, although, these values are much closer to the other two clusters than in dry red wine.

It is important to also notice that the “Young Crowd” have their preferences pretty much evenly distributed with, curiously, the red dry wine as the least favorite of the group, when it is the clear favorite among the whole sample. This may happen because, as the red dry wines have more demand, they may be less frequently in good deals and we know this segment prioritizes the good deals and the low prices when choosing the wine.

## Overbooking and Discount policies

The two least profitable segments, the “Young Crowd” and the “Bargain Hunter” are not our prioritized groups of customers, although, every customer is important, and we managed to develop a similar strategy for both. As they are more online-oriented and tend to look for best deals, we suggest a new tab, containing discounts, on the website so every time these customers want to acquire a bottle of wine their process is easier and more pleasant. Also, we suggest a creation of a loyalty program where discounts, for mostly red and white dry wines, are given throughout the purchases made online. We believe this will encourage the younger customers to order more times, in order to pursue better prices and bigger discounts in their preferred wines. And lastly, regarding these clusters, we feel it would be better, and clearly less expensive, to move the newsletter online. These target groups will be more exposed to our offers and it will certainly increase the demand for these clients, as little as this increase may be.

For the “Average Joe” customers, we believe the loyalty program above mentioned is a good suggestion. If we offer discounts in the most demanded wines and gifts, we will encourage these clients to stick to WWW every time they want to buy a wine. Also, the newsletter becoming digital would be good for this segment for better information of the loyalty program.

Next up, in terms of priority for us, is the “Silver” segment. We propose to offer these clients a package of wine and a discount on higher price tag products at the points of sale of Dry Red Wines. Also, we propose to create a high-value customers loyalty program that sends gifts and a personalized catalog of wines and high price tag products. It is important to make the clients feel special. And this is a crucial step to start turning these customers into “Gold”.

Finally, our priority, our most loyal and best customers, the “Gold” customer group. They represent almost 25% of our dataset and there for our goal here should be to retain them, with that in mind we need create strategies to praise them and make them feel like they are very important to our business. We suggest promoting the sell of premium high-priced wines offering invitations to exclusive wine tasting tours and events. And lastly, we should also include them in the high value customers loyalty program.

# DEPLOYMENT AND MAINTENANCE PLANS

## Plan deployment

Assess whether this 10k customers extracted from a test promotion made for an accessory (a silver-plated cork extractor) can eventually bias the sample and diminish the capacity to extrapolate the significance of the results for the whole 350k dataset. The assessment procedure must check: (I) if by assigning the closest centroid for each customer will change the nature of the differences in variables observed for the customer segmentation implemented; (II) if the R2 of the specification changes a lot when calculating it for the whole dataset after being clustered, compared with the R2 of having 10k customers.

We believe that in the future there can be made improvements in order to optimize future data collection and cleaning. A date time of the orders and the identification of some marketing campaigns would be extremely helpful to better understand where the customers come from and more specifically to measure the success and reachability of the campaigns.

We strongly suggest a considerable investment on the process of collecting data for better and more simple future analysis.

## Maintenance

Additionally, as time progresses to keep track of the R2 of the specification and the overall difference in the centroids at both scaled and unscaled data is also important. Thus, the need for tuning the specification again, by reselecting features that comprises the same intent from the selected specification from this work or even reassessing the best number of clusters is important if we observe detraction on the business objective measures, like conversion, retention, increase in any of the RFM features etc.

Lastly, it is important to notice that our model uses LTV (as it is), and since we received 10k of 350k customers with these feature set, we concluded that WWW have this LTV value for all the customers database, even if WWW does not know how it was constructed. The reason for selecting it was already explained earlier (correlation matrix assessment) but despite that, LTV can eventually be replaced in two ways, right now and a few months forward with a creation and tracking of a new KPI, respectively: (I) Substituting by Frequency or Monetary, since they are super highly correlated and convey a lot of this customer value aspect we are tying to bring for the segmentation solution; (II) Substituted by a new LTV measure that would be specified by this equation () specified for each and every of the 5 clusters we suggested.

# CONCLUSION

The initial marketing strategy of the company, to send the catalog to everyone, showed to not be effective, in some cases there were costumers that were giving loss instead of profits (negative LTV). This was obviously not the best way to approach their customers.

After prototyping with different algorithms and clustering techniques, we were able to successfully identify five different customer segments and study them deeply both in terms of demographics as well as product preferences. By giving information on who these customers are, we provided inputs on how the marketing strategy can be maximized.

The company has different target audiences, with the older demographic being the best clients, but we believe focusing only on them is not the best approach to satisfy the long-term strategy for WWW. It is important to keep and cherish the most valuable clients but also to stimulate the younger ones.

With the plan defined in this project we can say that the company will have a better approach for all the different segments. And we can say that both the company and the customers will most likely come out winning from the outcome of the deploying this pipeline of procedures and strategies.

This company was already promising and now its potential can reach new levels. By digitally transforming their marketing efforts using our customer segment and marketing strategy recommendations we may see a lot of value being created for the company.