

**BUSINESS CASES WITH DATA SCIENCE**

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

**Business Case #2 – Hotel Booking Cancellations**



Data4Business Consulting

Débora Santos, number: 20200748

Diana Furtado, number: 20200590

Pedro Henrique Medeiros, number: 20200742

Rebeca Pinheiro, number: 20201096

March, 2021

INDEX

[1. INTRODUCTION 1](#_Toc65525049)

[2. BUSINESS UNDERSTANDING 1](#_Toc65525050)

[2.1. Background 1](#_Toc65525051)

[2.2. Business Objectives 2](#_Toc65525052)

[2.3. Business Success criteria 2](#_Toc65525053)

[2.4. Situation assessment 2](#_Toc65525054)

[2.4.1. Inventory of resources 2](#_Toc65525055)

[2.4.2. Requirements, assumptions and constraints 2](#_Toc65525056)

[2.4.3. Risks and contingencies 2](#_Toc65525057)

[2.5. Determine Data Mining goals 3](#_Toc65525058)

[2.6. Project Plan 4](#_Toc65525059)

[3. CLUSTERING ANALYSIS 4](#_Toc65525060)

[3.1. Data understanding 4](#_Toc65525061)

[3.2. Data preparation 5](#_Toc65525062)

[3.3. Clustering 6](#_Toc65525063)

[3.4. Evaluation 8](#_Toc65525064)

[4. RESULTS EVALUATION 8](#_Toc65525065)

[5. DEPLOYMENT AND MAINTENANCE PLANS 9](#_Toc65525066)

[5.1. Next steps 9](#_Toc65525067)

[5.2. Application 10](#_Toc65525068)

[6. CONCLUSIONS 10](#_Toc65525069)

[7. REFERENCES 10](#_Toc65525070)

# 

# INTRODUCTION

Thank you for choosing **Data4Business Consulting (D4B)** to help you with the challenge of better understanding your customers characteristics. Our main objective is helping Hotel Chain C to identify high cancellation likelihood bookings and consequently increase your gains and customers’ satisfaction.

The world is experiencing a great technological and digital revolution where understanding business data, customers 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 C.
* Jupyter Notebook with the code of the entire process.
* The files to run application developed by D4B to C.

All files can be accessed in Github:

https://github.com/Debs86/Business\_Cases\_Projects/tree/main/BC2

We are excited to take part of this challenge.

# BUSINESS UNDERSTANDING

## Background

**Hotel chain C (C)** has 2 hotels in Portugal: One is a resort located at the region of Algarve (H1) and the other one is a city hotel located at the city of Lisbon (H2).

The current customers have some ways of booking a room: Travel Agency, Tour Operators, Corporate and Direct. The advancement of the internet brought more exposure, but on the other hand it brought more competitiveness. With the appearance of the online travel agencies (OTAs), the number of the “deal-seeking” customers have grown immensely. “Deal-seeking” customers tend to make multiple bookings for the same trip to find the best deal. Consequently, it increases cancellations in the hotels.

C was severely impacted by cancellations, representing almost 28% of the bookings in H1 and nearly 42% in H2. The Revenue Manager Director of C, Michel, has already implemented several approaches to reduce cancellations, with no significant improvement.

Michael wants to implement prediction models to allow the chain’s hotels to forecast net demand based on reservations on-the-books, more specifically in H2, therefore he has reached D4B.

## Business Objectives

The customer’s primary objective is to implement prediction models to forecast net demand based on reservations on-the-books. With these models, the customer expects:

* Implement better price and overbooking policies.
* Identifying high cancellation likelihood bookings.
* Implement actions to prevent cancellation.
* Reduce the cancellations to a rate of 20%.

## Business Success criteria

The expected outcome will be the development of a predictive model to forecast the net demand based on the bookings. The success of the proposed task will be evaluated by C revenue manager director and, if needed, we will go back to the model until we get an outcome that matches with his expectation.

## Situation assessment

### 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 C’s Management and team.

From D4B Consulting, this project will be conducted by a team of 4 Data Scientists and Business Analysts.

We have been provided by the C team with a dataset of bookings made in H2, of customers due to arrive between July 1, 2015 and August 31, 2017. Along with this dataset, we were also provided with its metadata file.

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.

### Requirements, assumptions, and constraints

The completion date of the present phase of the project is March 15, 2021, but we expect to continue giving support and helping C to achieve the next goals for the growth of the business.

Even if the booking has as *Reservation Status*, *No-Show* and the *Deposit Type* is *No deposit*, the customers will be charged on their credit card.

Some bookings provided seems to be duplicated, but we are going to disregard that.

### 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 |
| Bookings with very similar characteristics | Split the data (train and test) avoiding bias. |
| Model overfitting[[1]](#footnote-1) | Ask for more observations (bookings) |

Table 2.1 - Risks and contingency.

#### Terminology

***Business glossary***

* “Deal-seeking” customers: tend to make multiple bookings for the same trip to find the best deal.
* Net demand is defined as demand minus cancellations.

***Data mining glossary***

* Prediction problems: It is a machine learning technique. The technique consists in build a model based on a previous real dataset which will predict with a high accuracy the target. In this case, the model built will predict if a booking will be cancelled or not.
* Normalization: The major predictive algorithms 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.

## Determine Data Mining goals

The data mining goals states project objectives in technical terms:

1. Training a model based on a real booking dataset that will be able to predict if the future booking will be cancelled or not. The dataset will be splitted in train and test, which the first one will be used to build the predictive model and the second one to test and check if the model built had a right outcome.

*Success criteria*: Right target prediction. The model should be able to predict if the next online booking will be cancelled or not.

1. Evaluate the predictive model.

*Success criteria*: Acuracia

1. Suggest marketing strategies and business applications.

*Success criteria*: Reduce to 20% the booking cancellations and give overbooking suggestions based on the evaluation of the combine variables cancellation and no show.

## 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. To present the results, we expect to use Word for the report and Power point for the presentation. Finally, to provide a user-friendly visualization of the results, we plan to build an application using Python.

The performance of the model 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 a supervised model (preditcion) using ??? 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 model evaluation will be made using R squared and some visualizations to check the good distribution between clusters.

# predictive 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.

## 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 ( 31 features, from which 13 categorical and 18 numerical as shown on Table 3.1) and observations (79.330 customers), how the variables are distributed (there are some skewed variables) , if there is noise, if there are missing values (28) 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.

|  |  |
| --- | --- |
| **Numeric** | **Categorical** |
| *ADR*, *Adults,Babies*, *BookingChanges,Children, DaysInWaitingList, LeadTime, PreviousBookingsNotCanceled, PreviousCancellations RequiredCardParkingSpaces, StaysInWeekendNights, StaysInWeekNights, TotalOfSpecialRequests* | *Agent, ArrivalDateDayOfMonth, ArrivalDateMonth, ArrivalDateWeekNumber, ArrivalDateYear, AssignedRoomType, Company, Country, CustomerType , DepositType, DistributionChannel, IsCanceled, IsRepeatedGuest, MarketSegment, Meal, ReservationStatus, ReservationStatusDate, ReservedRoomType,* |

Table 3.1 - Numerical and categorical features.

As we are dealing with a classification problem is important to check the distribution of the data according to the target variable (*IsCanceled*). Almost 42% of the total bookings are cancelled. From those cancelled bookings, 2.8% are no-show (meaning the customer got charged).

Going into more details, April to June are the months where there is a higher proportion of cancellations. Most of the bookings that have no children and/or babies have a higher percentage of cancellations. Also, Bed & Breakfast bookings tend to cancel more than average. On the other hand, repeated guests and those requiring parking space, booking changes and/or special requests are more willing to show up.

Bookings through Travel Agents or Tour Operators (representing nearly 87% of the total bookings) tend to cancel more often, while groups reservations tend to cancel less than others.

- Customers booking in advance are more willing to cancel the reservation

- In average, a non-cancelling customer, books 80 days in advance of the expected arrival date

- Regarding the cancelling customers, those book, in average, 150 days in advance and cancel the booking 90 days before the expected arrival date

## Data preparation

This is the stage when the input data for modelling is prepared, so we have ensured the data meets the requirements for this purpose. INCLUIR AQUI COMO DEVE FICAR AS VARIAVEIS PARA UM MODELO PREDITIVO.

On the first step of data preparation, we drop some observations with missing values (28, which represents 0,035%) and one observation which ADR is equal 5400 because it is clearly noisy if compared to the others. We also eliminated the bookings with zero adults, because we believe it’s an error, once in those same case there were children and/or babies included (0,48% of observations). The next step was feature engineering where we built 3 features:

|  |  |
| --- | --- |
| Variables | Meaning |
| Days\_before\_cancel | Number of days the booking is canceled before the entry date (=0 if not canceled) |
| DateDifference | Difference between the Arrival date and the last status update date |
| Days\_until\_cancel | Number of days between the reservation is made until it's canceled (=0 if not canceled) |
| RoomType\_change | Binary variable showing if the customer will get what he/she reserved (1 if ReservedRoomType=AssignedRoomType; 0 otherwise) |

Table 3.2 – New variables.

A Pearson correlation matrix was prepared to look to the correlations. From the analysis of full Pearson correlation matrix, we couldn’t identify any correlation between features. The next step, we prepared a Phik correlation, as we are dealing with a dataset with many categorical features. From this analysis, we could identify some strong correlations, as we can see in figure xxx.



Reservation date  
reservation status

Agent

Arrival date month

Market segment

Reservedroomtype

Assignedroomtype

INCLUIR AQUI AS ALTAS CORRELAÇÕES E EXPLICAR QUAIS DELETAMOS E PORQUE

After dropping these 07 features, we have checked once again the Phik correlation matrix, which can be found in the notebook provided with the report.

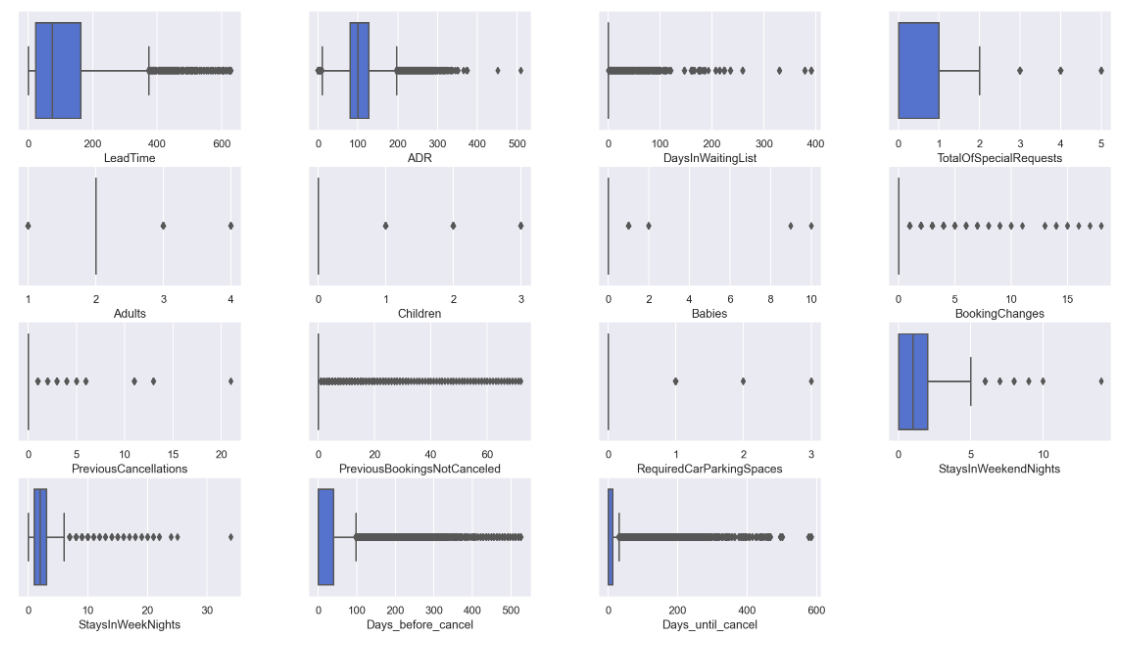
To check for the presence of outliers on the numeric variables we looked at the boxplots for each numeric feature and concluded that features shown in Figure 3.2 seem to have outliers.

Figure 3.2 – Boxplot for numeric feature – outliers.

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 3.3.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 3.3 - Number of outliers excluded with different approaches.

After testing removing based on 6 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 hand made removal outliers only on the six features mentioned above (*Babies, ADR, DaysInWaitingList, StaysInWeekendNights, StaysInWeekNights and Days\_until\_cancel*). 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 feature selection, we have one-hot-encoded categorical variables (to transform in binary values) and applied the StandardScaler so the dataset will have a mean value of 0 and a standard deviation of 1.

INCLUIR AQUI O FEATURE SELECTION.

We have finished the data preparation with the following variables to apply the model: *Dayswus, Edu, Freq, Recency, Perdeal, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebVisit, Access and Avg\_ticket*.

## modelling ( não atualizado).

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.

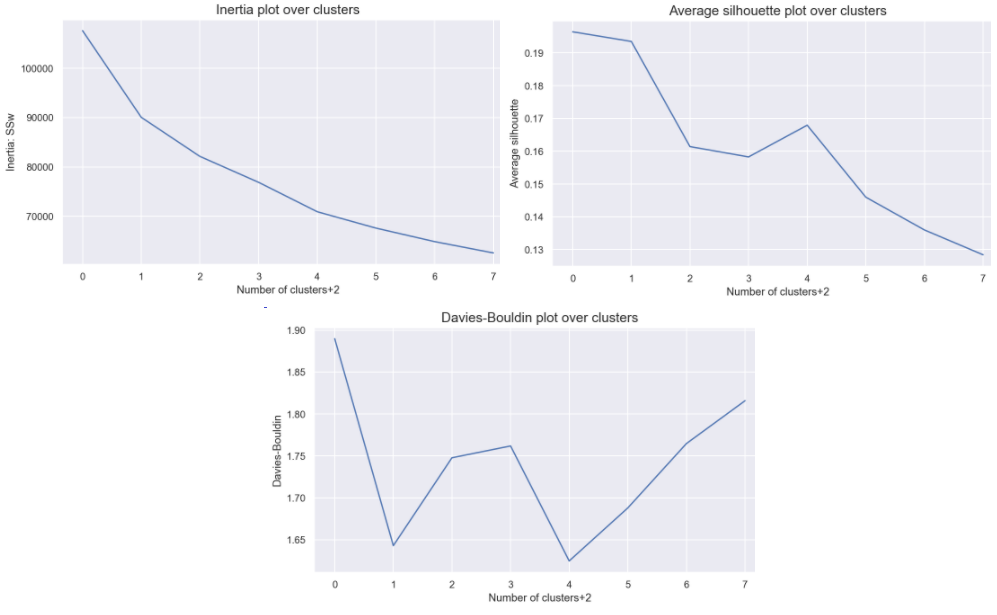


Figure 3.3 – Inertia, Average silhouette and David-Bouldin plots.

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

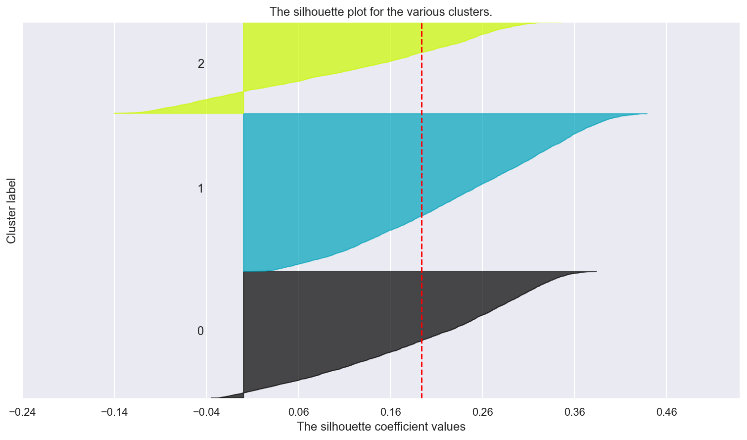


Figure 3.4 - Silhouette plot.

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.

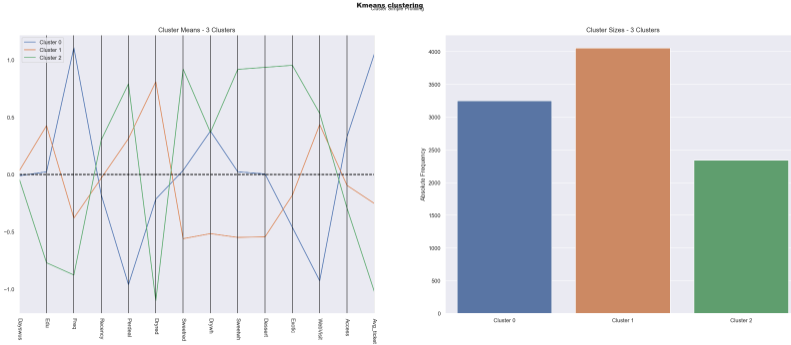


Figure 3.5 - Clusters profile.

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.

## 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.

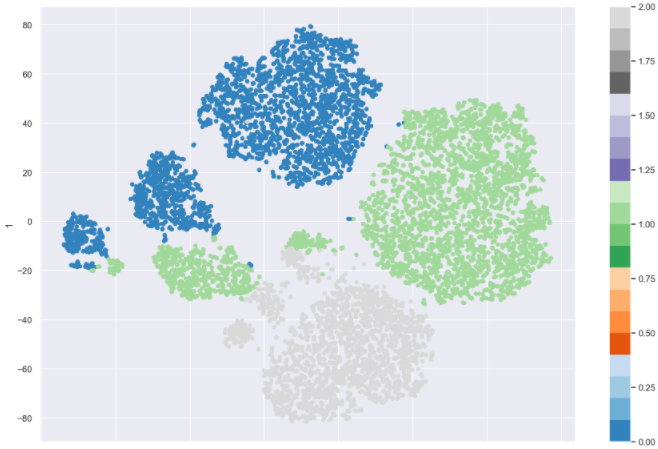
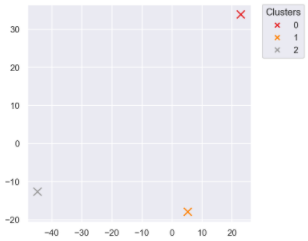
 

Figure 3.6 - t-SNE plots.

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.

# 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.

# DEPLOYMENT AND MAINTENANCE PLANS

## 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.

## 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”.

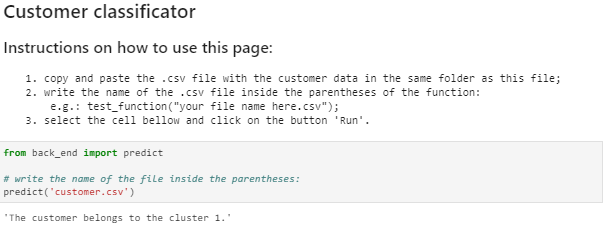


Figure 5.1 - Customer classifier application.

# 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.

# 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>>

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. [↑](#footnote-ref-1)