

**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).

Currently, customers can book through Travel Agencies, Tour Operators, Corporate or Directly with the hotel. The advancement of the internet not only brought more exposure, but also 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:

* To implement better price and overbooking policies.
* To identify high cancellation likelihood bookings.
* To implement actions to prevent cancellation.
* To reduce 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.

On the D4B Consulting side, 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 H2 resort bookings, 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 | 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***

* Classification problem: In this type of problem, the objective in implementing machine learning techniques and algorithms is to predict a class of the target variable. On this specific problem, the variable *IsCanceled* is the target. Also, it is a supervised problem when the data provided already included desired outputs.
* Accuracy: A rate between the true outputs against the total. In other words, the proportion of correctly predicted cancellations or no cancelations against the total of bookings.
* Precision: measure how precise the model is out of those predicted positives. The proportion of correctly predicted cancelations against the proportion of total correctly predictions.

* Recall: how many actual positives, the model captures as being positive. The proportion of correctly predicted cancelations against the total of actual cancelations.

* F1 Score: Measure that represents a balance between recall and precision. This is measure of model quality is especially important for this analysis as it represents the balance between the false positive (precision) and false negative (recall). In this case, False positive is about to predict that a customer will cancel when, in reality, he will not cancel the booking. This kind of situation generates problems to the hotel such as: 1) overbooking and consequently extra expenses with rent a room in another hotel and reviews. 2) costs offering services to avoid a cancel that it would not happen. On the other hand, the false negative is about to predict that a booking will not be cancelled when, in reality, it will be cancelled. This kind of situation generates lower revenue to the hotel because the room will be empty

* AUC (Area under the curve):

## Determine Data Mining goals

The data mining goals states project objectives in technical terms:

1. Create a model that will be able to predict the probability of new bookings be cancelled or not.

*Success criteria*: High percentages of accuracy, precision, f1-score. Also, low percentage of false negatives.

1. Understand the main characteristics of the bookings.

*Success criteria*: Give some insights about the characteristics of the bookings helping to identify if it has or not a high probability to be cancelled.

## Project Plan

Figure 2.1 - 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 and Flask.

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 (predictive) using Random Forecast classifier algorithm. We opted for this model because it presented the best results compared with other algorithms. Further details will be presented in section XXX – Results. The model evaluation will be made using accuracy, precision and F1 score metrics. The accuracy metric is good to see the overall performance of the model. Precision to measure how good the model is predicting the cancellations. Finally, F1 to check the balance between recall and precision. It will also be presented a confusion matrix and an AUC curve.

# 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 missing entries) 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, it 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 with 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, guests those requiring parking space, booking changes and/or special requests are more willing to show up. Repeated guests also have low percentage of cancellations, but they represent less than 3% of bookings.

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.

## Data preparation

A classification algorithm can be used to analyze discrete data, however, the data should be prepared to ensure it meets the requirements for this purpose. The dataset must be as complete as possible, missing values must be treated (either removed or filled in) and outliers should also be checked for. To avoid underfitting (when the model cannot detect any relationships in the data), relevant variables must be included; on the other hand, to avoid overfitting, variables with no predictive power should be eliminated. Because the quality of the data can make a model perform well or not so well, this data preparation stage takes the majority of the time in the project.

On the first step of data preparation, we drop some observations with missing values (28 observations, representing 0,035% of the dataset) and one observation in which the ADR is too high and clearly noisy (ADR=5.400). We have also eliminated the bookings with zero adults, as we believe it is an error, once in those same case there were children and/or babies included (0,48% of the dataset). The next step was feature engineering which we built 4 new features, for data analysis purpose:

|  |  |
| --- | --- |
| **New variables** | **Description** |
| *Days\_before\_cancel* | *Number of days the booking is canceled before the entry date (=0 if not canceled)* |
| *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.

The next stage was to check the correlation matrices to identify redundant variables. We could not identify any strong correlation on the Spearman’s matrix, but we identified strong correlations on Phi-k matrix, as we are dealing with a dataset with many categorical features. The following features were dropped, based on redundancy and their correlation between each other: *MarketSegment*, *RoomType\_change*, *Agent*, *Company*, *ReservationStatusDate*, *AssignedRoomType*, *ReservationStatus*, *Days\_before\_cancel, Days\_until\_cancel*. We have also dropped 8 variables considered non-relevant for the model (*ArrivalDateYear*, *ArrivalDateWeekNumber*, *ArrivalDateDayOfMonth*, *ArrivalDate*, *Country*, *BookingChanges*, *RequiredCarParkingSpaces*, *TotalOfSpecialRequests*).

We have then reproduced the Phi-k correlation matrix to confirm it looked reasonable, once more. For further details, the correlations are available on the notebook.

Looking at the boxplots for each of the numeric variables, we have identified 5 variables (*ADR*, *DaysInWaitingList*, *StaysInWeekNights*, *Babies* and *StaysInWeekendNights*) with outliers (Figure 3.1).

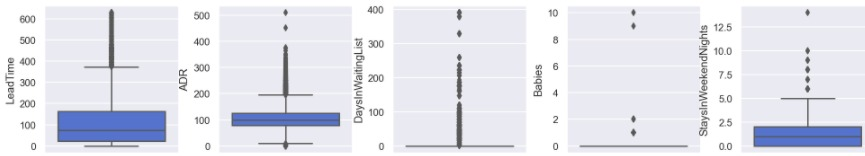


Figure 3.1 – Boxplot for numeric feature before outliers removal.

To remove these outliers, we tested applying five different methods on the entire dataset and using them individually or in combination. However, after testing these different approaches, and realizing that, outliers wise, it did not seem to make a big difference on the box and whiskers plots, we have opted for manually removing the outliers (considering it was feasible to do it for the six features where outliers were found). This way we have removed 2.723 observations, representing 3,45% of the dataset.

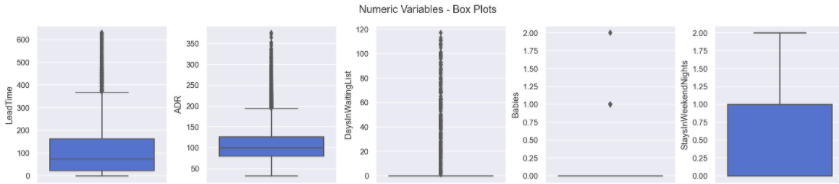


Figure 3.2 - Boxplot for numeric features after outliers removal.

In order to prepare the data for modelling, we have one-hot-encoded categorical variables (to transform in binary values). The next step was split the data set in train (80%) and test (20%) set. As we stated on section Risk and contingencies, the data set has many “duplicated” rows. So, we stratified the data keeping the proportion of duplicate rows similar in both train and test set.

## modelling and evaluation

The first stage of this process was to split the dataset into train and test, 80% and 20% of the dataset, respectively. This dataset has entries that are exactly the same and, even though they are real observations, to the model they can be considered as duplicates. Because of this, we had to be especially careful when splitting the dataset, to ensure the same pattern was followed both in the train and test datasets, so we split the repeated values equally between both datasets.

We start this process creating a pipeline to perform a baseline modelling. The pipeline includes two steps: excluding the features with variance low than 10% and running 4 different algorithms (Decision Tree, Random Forest, AdaBoost and Gradient Boosting) with the default parameters. We chose these algorithms based on their ability to predict the positive results (i.e. cancelations) and its ease of interpreting the results. We did not scale the data, because the algorithms chosen don’t need it.

To measure the performance of each algorithm, we used a 10-fold cross validation only on the train set and calculated the average score for each fold. The metrics used were Accuracy, Precision and F1-Score. The results are presented on Table 3.3.

COLOCAR TABELA DA DEBS COM OS RESULTADOS FINAIS

Table 3.3 - Metrics results for the different algorithms.

The next step was to find the optimal number of features and get their importance; this analysis was made applying the Mutual information coefficient, as most of the variables have non-linear relationships. Features *DepositType\_Non Refund, DepositType\_No Deposit* and *ADR* stand out as the most relevant features to the target (altogether weighting 57%). The ranking with the features which weight>1% is presented in XXXX.

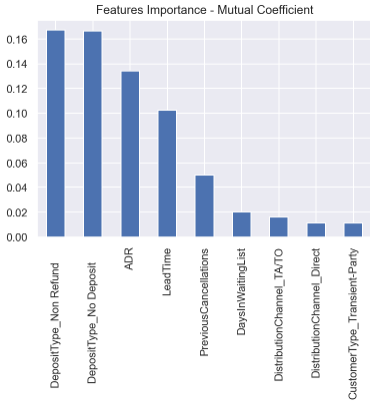
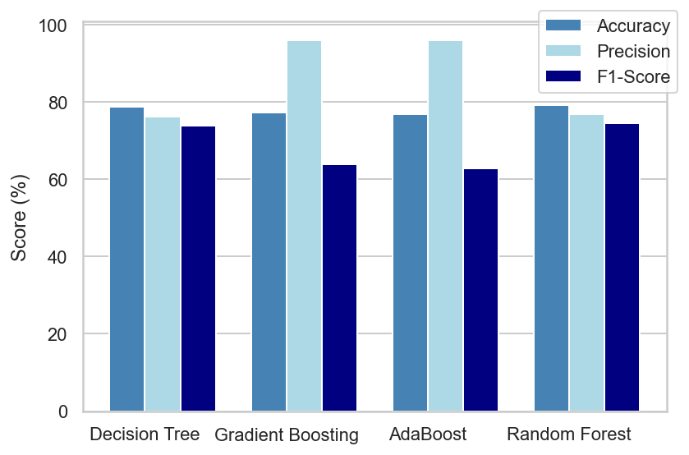
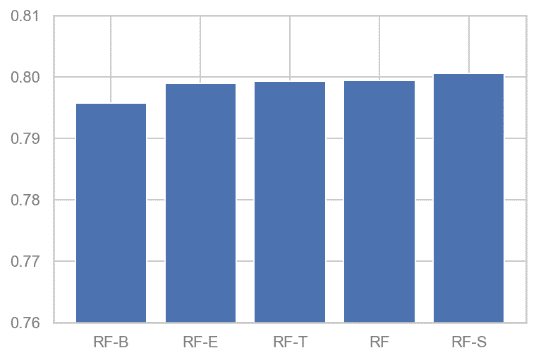


Figure 3.3 - Features importance from Mutual Information Coefficient.

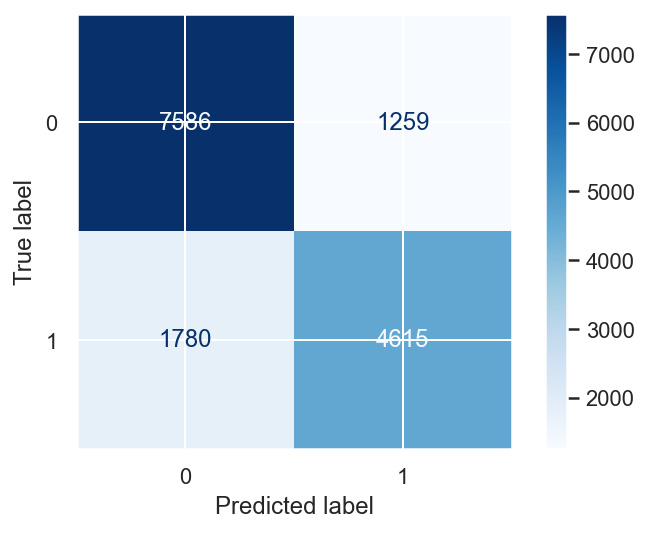
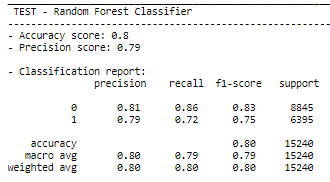
We repeated the process of applying some algorithms with the default parameters and cross validation on the train set, but now only with the top 9 features selected by mutual coefficient. The results are presented on the table XXX.



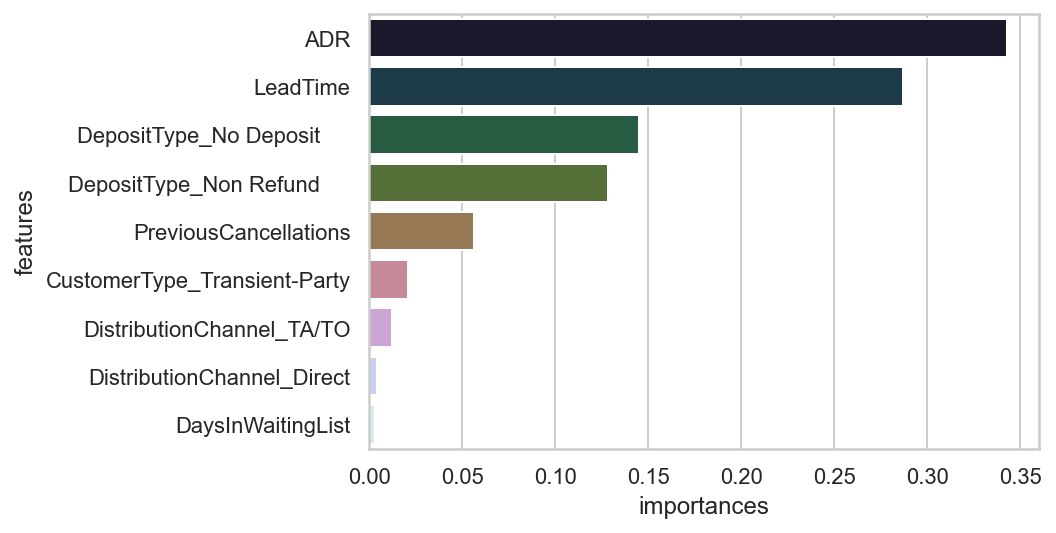
Analyzing the results, we chose as our final algorithm the Random Forest Classifier. Even the Gradient Boosting and Adaboost had presented very good scores for precision metric, as they presented low score for F1 metric, we decided to not use them. Compared with Decision trees, Random Forest presented better accuracy and precision.

 The next stage was trying to improve the scores of Random Forest tunning the parameters. At this time, we applied the model in the test set. We tested 6 different types of parameters. The table with the parameters of each model and the graphic with the results are presented below:

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameter changed** | **Value** |
| modelRF | random\_state | 5 |
| modelRF\_ent | Criterion | Entropy |
| modelRF\_300trees | n\_estimators | 300 |
| modelRF\_b | bootstrap | False |
| modelRF\_s | max\_samples | 0.7 |



The model with the best performance was the one with the parameter *max\_samples* *= 0.7*.



# 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”.

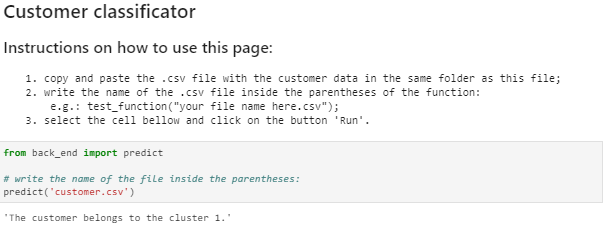


Figure 5.1 - Customer classifier application.

# 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>>