# Table of contents:

[Table of contents: 1](#_Toc98685038)

[Introduction: 2](#_Toc98685039)

[1 – Business Understanding 2](#_Toc98685040)

[2 – Data Understanding 2](#_Toc98685041)

[3 – Data Preparation 3](#_Toc98685042)

[3.1 – Data Visualization and Preprocessing 3](#_Toc98685043)

[3.2 – Feature Selection 4](#_Toc98685044)

[4 - Modeling 5](#_Toc98685045)

[5 - Evaluation 5](#_Toc98685046)

[6 - Deployment 5](#_Toc98685047)

[Conclusion: 5](#_Toc98685048)

[Appendix: 5](#_Toc98685049)

# Introduction:

# 1 – Business Understanding

The hotel chain C, located in Portugal, is severely impacted by cancellations and our dataset (H2) represents 42% of cancellations among the bookings. Restrictive cancellation policies and aggressive overbooking policies are few of the actions done by the managers of the hotel to mitigate and avoid revenue loss. We were hired as consultants to help improve their cancellation numbers through a predictive model where we can determine if a customer is likely to cancel their booking or not. This approach may enable the hotel directors to act more accurately regarding pricing and overbooking policies and reduce cancellation rates to less than 20%.

We are expected to explore the data and build a model to predict cancellations, elaborate on business implications of employing the model and make suggestions on how to deploy and the impact on the business processes of the hotel chain.

We were presented with a dataset composed of 79330 bookings made between 2015 and 2017. The information is distributed through 30 different variables, regarding information about the customers, the booking, and the cancellation behaviors.

# 2 – Data Understanding

We started by performing some exploration of the data and understanding what each feature represented.

The database is composed of 79,330

observations and 30 features (date, numeric and categorical variables) related to:

* Booking basic information: date of arrival, stays in week and weekend nights, number of adults, children, and babies.
* Information about the customer: country of origin, if they have cancelled before, if they have booked before, if they require parking spaces or special requests.
* Additional booking information: agency and companies used to book, Monetary information about the rates of each room booked and type of rooms and meals.
* Whether or not the booking was cancelled.

We noticed that the top 20 countries account for 93% of the bookings, being Portugal the largest representative with 30.000 bookings (around 38% of the bookings).

Gráfico, Gráfico de barras, Histograma

Descrição gerada automaticamente

As mentioned before, the cancellation rate amongst the bookings currently represents 42% of the total bookings. However, at first glance, we noticed a significant number of duplicates, which will be treated in the next topic and the data will be skewed.

# 3 – Data Preparation

## 3.1 – Data Visualization and Preprocessing

As mentioned in the previous topic one of the first issues that we noticed on the dataset was the number of duplicates. Although there are no customer ID numbers for each booking, we understand that it would be highly unlikely to find two identical bookings with all the 30 variables equal, same amount of family members, booked at the same day in the same type of room for the same arrival date, with the same waiting time and lead time, same countries, market segments and all other features. For that reason, we understand that this duplicates could bring a significant amount of bias to our model and decided to drop all the duplicates.

After checking the countries and understanding that 93% of the data is represented by the top 20 countries, we decided to limit the Country feature to 20 different countries, and the observation is not one of the top 20 countries, it will be kept as “Other”.

After dropping the duplicates, we still kept around 50,000 observations and the target of our dataset was rebalanced for 30-70, with 30% being the number of cancellations.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

We also explored the annual cancellations/booking over the years and we did not find significant correlation of seasonality. We found that the average stay for non-business travelers (customers without a company booking) is around 2 days.

After understanding the basic features, the next step was to define which were the metric and the non-metric features. Also, to avoid having overfitting problems we decided to perform a data split in the beginning of the process in train and test datasets. Since the data is time-based, we performed not only a stratified split but also a shuffled split to mix the data from different years. We performed a manual outlier removal on 3 specific features that presented significant outliers: babies, ADR and days in waiting list (we had, for example, bookings with 8 and another one with 10 babies at the same time).

We also found missing values in the Children column and noticed that we have plenty of Countries, Companies and Agents filled with the string ‘NULL’. We decided to treat those as well. The NULL in “Agent” and “Company” indicated to us that those customers had booked by themselves, without agencies and not as businesses. In that case, we just substituted the “NULL” strings for zeroes. As for the country, we just replaced the nulls as “Unknown”, since those are probably customers who decided not to inform their country of origin. To keep the data coherent, we decided to change some datatypes (children, for example, was a float).

With the preprocessing steps done, we decided to create some new features that could be useful in the future.

* Metric features: logADR and logRevenue
* Non-metric features: Dummy (only 1 or 0 for yes or no) features for children, parking spaces, special requests and previous cancellations.

## 3.2 – Feature Selection

After performing the basic understanding and preprocessing steps, we started exploring the data itself and trying to select the features we would need to use in the predictive models.

Although useful for exploration, we dropped reservation status and reservation status date as they would cause information leakage on the modeling part. Also, removed all the date variables as we now have the ArrivalDate feature.

We also performed some pair plots to check correlation and collinearity between variables and the target (isCanceled) but nothing significant was found. For that reason, we decided to run a correlation matrix using the Spearman method.

We noticed that the only feature with at least 20% correlation with the target is the feature “LeadTime”, meaning the longer the wait time, the bigger the probability of cancellation. Customers that have previously cancelled also are slightly more likely to cancel the new bookings.

It’s worth mentioning that we found a slight inverse correlation for Special Requests and Booking Changes, so the more booking changes and special requests made by the customers, less likely they are to cancel their bookings. We also used RFE, Lasso, Decision Tree and Mutual Information for the numerical data and gathered all the results to decide which ones to keep or discard.

As for the categorical variables, we used the Chi-Square method, mutual information, and Decision Tree. We noticed that the models were presenting all the categorical variables are important to keep, so we decided to take a step further and split them into dummy variables to test them separately. After that step, we run some feature selection models again. The results for both numerical and categorical data are shown in the tables in the annex.

# 4 - Modeling

# 5 - Evaluation

# 6 - Deployment

# Conclusion:

# Appendix:

Numerical Data feature selection results:

Interface gráfica do usuário

Descrição gerada automaticamente

Categorical Data feature selection results:

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente