NOVA IMS Information Management School

MARCH 2022

BC2: PREDICT HOTEL BOOKING CANCELLATIONS

ALICE VALE R20181074
EVA FERRER R20181110
RAFAEL SEQUEIRA R20181128
RAQUEL SOUSA R20181102

BUSINESS CASES FOR DATA SCIENCE

This report was written under the CRISP-DM guidelines.

Business Understanding

Hotel Chain C is an accommodation chain with numerous resorts and city hotels in Portugal. Following the rise of the cancellation trend in Europe, *chain C* was negatively impacted by the phenomenon registering cancellations of nearly 42% in some of its hotels. To counteract, several approaches were put into effect by the revenue manager director, however, without attaining success. They now intend to employ a predictive model solution to forecast cancellations, specifically in a city hotel located in Lisbon, Hotel 2. The main goal is to decrease cancellations down to a rate of 20%. Such outcome could lead to speedier pricing strategies and cancellation policies that would position the company ahead of its competitors and the market trends.

Data Understanding

The dataset provided by the Second Hotel comprised 79330 observations, with information regarding 3 distinct years (2015,2016 & 2017) and 31 variables respecting to the stay. The binary target variable to be predicted is whether a reservation was cancelled or not, assuming respectively the values 1 or 0. As a preliminary data exploration analysis, visualization techniques were applied to understand the raw behavior of the given data: the distribution of the Target, Revenue and Booking per month/year were intensely accessed. Consequently, it was decided to divide the further steps into two segments, based on seasonality: the high season comprising data related to the months April-September; and the low season to the months October-March.

Data Preparation

The preparation process began with duplicate checking. A total of 25902 observations were identified as being replicates, which is believed to be due to the fact that the given dataset might be a merging of multiple datasets from different data warehouses or sources. This could have led to the same records being brought together into the same dataset. Since duplicates are redundant and a threat to the performance of predictive models, all observations under this category were removed, remaining in the dataset the first example of each incident.

Secondly, missing value inspection was employed. For the *high* season split, the variable *Children* was missing four observations, which were attributed the value zero, since this was the most frequent instance of the variable's range of values. For the *low* season split, only the variable *Country* was missing two instances, which were imputed, following the method applied before, the *mode*. It is important to mention that *MarketSegment* and *DistributionChannel* had cases of *NULL* or *Undefined,* however, upon further inspection it was determined that these occurrences are valid and do not signify missing values or inaccuracies. For both season splits, *high* and *low*, the variables' histograms and boxplots were investigated so to assess the existence of skewness or outliers. Subsequently, and based on the previous evaluation, manual removal clauses were created to proceed with the elimination of said values. Lastly, less than 1% of data was removed, although outliers were identified more often in the *low* season split.

Regarding incoherencies, a booked room must have at least one individual assigned to it, that is, *Adults*, *Children* and *Babies* cannot all be equal to zero. Following this assumption, a booked room should not be occupied merely by babies due to safety and legal reasons.

It was thought that only this category of individuals should be integrated in this incoherence since, in the hotel industry, children above a particular age are allowed to have their own hotel room. Finally, a repeated guest should have at least one previous booking in the hotel to be recognized as such. A total of 388 observation fell under these three hypotheses and were removed from the dataset to secure data quality.

A total of 15 new variables were created with the goal of increasing the predictive power of the model. The new dimensions were generated based on a thorough assessment of the already existing variables, resulting in either combination of variables or information extracted from them. Concerning categorical features, a *One-Hot Encoding* strategy was applied to prevent multicollinearity and ensure their ability to offer predictive information. All features were scaled using the *Standard* scaler, to allow for variable comparability and consistency of distributions.

Variable	***Description***
Infants	Total number of both children and babies, per booking
ArrivalDate	Complete date of arrival to the hotel
ReservationDate	Complete date of reservation
DaysBet	Number of days between the *ReservationStatusDate* and *ReservationDate*
GroupSize	Total number of individuals per booking
TotalDays	Length of stay, in days, in the hotel
TotalValue	Total revenue of a booking
ValuePerson	Revenue of a booking per individual
ADR_ADJ	Measures the relationship between *ADR* per distinct combination of year, week distribution channel and room type and the corresponding value of the 3rd quantile
LT_ADJ	Measures the relationship between *LeadTime* per distinct combination of year, week distribution channel and room type and the corresponding value of the mean
Year_Month	Year and month of arrival to the hotel
Year_Week	Year and week of arrival to the hotel
AMonth	Number of the arrival month to the hotel
AYear	Year of arrival to the hotel
Day	Day of arrival to the hotel

Great efforts were made to prevent Data Leakage when performing data engineering. In fact, the variables which originated from Lead Time were inspired by the provided Article[1], so not to manipulate future conclusions.

To select the features with the highest potential of leading to a successful predictive model, both *Lasso* and *RFE* approaches were used for each season split. For both splits, *DaysBet* and *LeadTime* showed the greatest predictive significance. *(As seen in the following page)* After assessing the results, the correlation matrix was examined to evaluate the relationship between features and ensure that redundancy was not present among the selected variables. Finally, one subset of variables was designed for each split.

[1] ANTONIO, Nuno, ALMEIDA, Ana de and NUNES, Luis. Hotel booking demand datsets. Data in Brief. February 2019. Vol. 22, p. 41–49. DOI 10.1016/j.dib.2018.11.126

High	*IsRepeatedGuest, LT_ADJ, Adults, PreviousCancellations, PreviousBookingsNotCanceled, BookingChanges, ADR, RequiredCarParkingSpaces, TotalOfSpecialRequests, DaysBet, GroupSize, TotalDays, ADR_ADJ*
Low	*LT_ADJ, ADR_ADJ, StaysInWeekendNights, PreviousCancellations, PreviousBookingsNotCanceled, BookingChanges, ADR, RequiredCarParkingSpaces, TotalOfSpecialRequests, DaysBet, TotalValue*

Modeling

Initially, the dataset was split into train and test datasets, accordingly to the year the data referred to, taking advantage of a natural differentiation among observations. The decision was to train and validate the model with 2015,2016 observations, reserving 70% of the data for training purposes and the remaining for validation; Thus, 2017 data was used for testing the model.

Several classifier algorithms were tested, namely: *Random Forrest, Light Gradient Boosting, eXtreme Gradient Boosting, C-Support Vector, Neutral Network - Multi-layer Perceptron – applied on a Sequential Model - Keras, Logistic Regression* and *Voting*. These models' parameters were tuned to improve the test's recall metric value, when possible. The choice of primarily accessing recall instead of precision was the fact that the initial focus was on decreasing the ratio of False Negatives (FN). In addition, for each model iteration, a precision versus recall curve was plotted, to determine the best threshold of the Target for the prediction. For instance, a threshold of 0.4 means that predictions from 0.4 up to 1 are ultimately considered as Cancelled.

Finally, all the models were compared among each other, using a *ROC Curve*. However, the high and low seasons have unique practical implementations in the real business world, therefore the approached followed to compare and evaluate algorithms was shaped to each case. In fact, it was to minimize the ratio of FN in the low season: these predictions are translated into empty rooms (loss of revenue), because it is wrongly predicted that the customer will not cancel. While in the high season, minimizing the ratio of False Positives (FP) was the main concern, since it is an epoch in which overbooking can lead to higher expenses in reallocation, because the Hotel may already be booked to its max capacity. To do that so, 2 distinct metrics were accessed to compare the most promising models: the Cancellation Rate and the Overbooking Risk. It is important to highlight that the *Voting Classifier* was disregarded due to possible future technical difficulties in its implementation by the Hotel, due to its required computational and time effort, hence a simpler model was chosen.

High Season	Cancelation Rate	Overbooking Risk
XGB	0.087	0.062

Low Season	Cancelation Rate	Overbooking Risk
LBG	0.077	0.102

Legend: Final Solutions for each Season. Interpretation of Overbooking Risk for High Season: About 6% of the model predictions, result in wrongly assessing that the customer will cancel. Interpretation of Cancelation Rate for Low Season: About 8% of the model predictions, result in wrongly assessing that the customer will not cancel.

After merging the two Seasons, the 2 latter metrics were re-accessed.

Final Prediction	Cancelation Rate	Overbooking Risk
	0.084	0.074

Evaluation

The features importance was re-accessed, now with the support of the final chosen models for each Season. The results found are concordant with our preliminary analysis: *DaysBet* and *LT_ADJ* can be considered the most impactful on the hotel booking behavior, however in the high season special requests and the number of previous cancelations are more characterizing of the cancelation behavior, whereas in the low season the rate of the room booked has a higher significance.

While performing an annual analysis for 2017, it can be seen that the cancelation rate is higher in the months of peak season, contrasting with the low season, where overbooking is more predominant. This falls in line with the rationale applied when choosing the best model, since it is less risky to overbook off peak as there are more rooms available, in case of double booking. Nonetheless, the maximum value achieved by any of these rates does not surpass the 20% sought by hotel management. The deployment of the models created, drastically reduced the original cancelation rate observed for 2017, keeping it steadily around 10%.

A study performed regarding booking's value led to the conclusion that around 30% of the bookings are responsible for 50% of the total income registered by the hotel. These are not only characterized by stays with a higher average rate but also with average expenditures per individual roughly double the regular reservation. The true cancelation rate in this segment is one percentual point lower.

A leverage analysis was then performed, segmenting reservations based on their percentual monetary value in three different bins with approximately the same size. This again demonstrates the significant difference in value between reservations in the dataset, supporting the immense turnout this segment produces where a booking is twice as valuable.

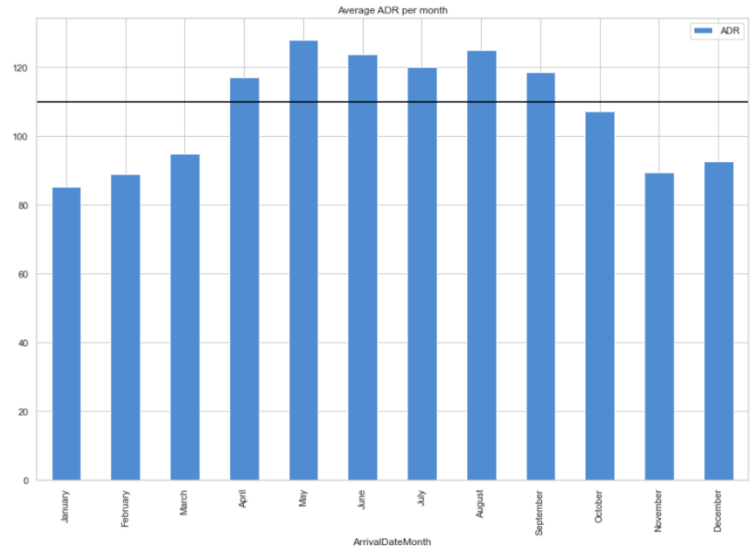
Deployment

With the knowledge provided by the results of this implementation, the hotel will be capable of designing a more certain and definite policy regarding overbooking, which will surely culminate in higher profit margins.

Should the hotel decide to use this model to predict future cancelations, and due to the lack of an internal data science department, we at *Datalin* offer our services to help implement and also maintain its correct functioning. It is important to note that, due to the data available coverage of years between 2015 and 2017, which does not include Covid-19 fluctuations, the practical applicability of the model can be affected when deployed for prediction of current/future years. To tackle that mishap, a new model could be constructed. Our advice would be to improve the accuracy of the developed model by testing it for the current year of 2022, using data from 2019 (pre-Covid) and 2021 (last year), avoiding the year 2020 for its obvious atypical behaviors.

ANNEXES

Analyzing the average ADR per month, combining data from 2015, 2016 & 2017: Explaining the Seasonality Choices



Analyzing the Final Metrics results, per each Month of 2017