Hotel Bookings

Master Degree Program in Data Science and Advanced Analytics

Business Cases with Data Science

Github repository:

https://github.com/andremforte/BC2_GroupV

Group V:

Anis Tmar (m20211157)

André Forte (m20210590)

Opeyemi Mary Akande (m20211320)

Rafael Nunes (m20210832)

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Index

1. Business Understanding

The tourism and travel industries demand are managed by advanced booking. Hotel chain C, a chain with resort and city hotels in Portugal, has been severely impacted by cancellations. To surpass that, the hotel manager decided to employ an overbooking policy to control the cancellations. However, it created more problems.

In order to manage this situation, the manager evaluated the possibility of developing predictive models to forecast the net demand for their hotels by hiring our Consulting Company. A dataset with information from H2, one of their city hotels, were provided to us, comprehending bookings between the 1st of July of 2015 and the 31st of August 2017.

The company's objective is to reduce the number of cancellations in the total of bookings. The manager's purpose is to get the cancellation rate down to 20%.

The data mining's goal is to build and optimize a classification algorithm that can predict if a booking is cancelled or not, based on customers' behavior and bookings' characteristics.

2. Data Understanding

The initial dataset contains 79330 rows and 31 variables (see Annex 1), and after analysing the distribution of the target variable, we understood that we were facing an imbalanced dataset. In order to have more information without any transformation, we continue our exploratory data analysis. These are some patterns we extracted from all of the process:

- There are some rows without any guest (zero adults, zero children and zero babies), which means that they don't represent reservations.
- Most of the customers don't require car parking spaces (it could be a fast stay indicator).
- Most bookings don't register any change during its process (zero booking changes).
- Most of the customers are new, so they don't register any previous cancellations.
- The main booking distribution channels are Travel Agents (TA) and Tour Operators (TO).
- The majority of the customers tends to demand between 0 and 2 special requests.

Looking at the distribution of bookings per Arrival Date, we can see some evidence of seasonality (high number of cancellations during summer and low number of cancellations during winter) (Annex 2). Analysing the Cancellation Ratio (computed by the number of cancellations divided by the total of bookings per month), it's possible to see some pattern after July 2016 - cancellation ratio is higher during spring and summer months. (Annex 3).

We extracted more important information for Data Preparation step. We saw a lot of duplicated entries (25902 rows) and there are missing values in two variables ('Country' (24) and 'Children' (4)). From boxplots, we can see some extreme values in 'Babies' and 'ADR' variables that could be outliers and there are high cardinality variables – 'Country', 'Company' and 'Agent'. Finally,

based on linear correlation from pairwise plots, we chose Pearson's method to visualize the variables' redundancy and concluded that there were two variables highly correlated.

3. Data Preparation

3.1. Cleaning Data

We started to transform the dataset by deleting the duplicated entries. We also removed the rows (154 entries) which were not bookings i.e., zero adults, zero children and zero babies. After that, we imputed the missing values, using the most common value (mode) for the categorical variable, 'Country', and the median value for the numeric one, 'Children'.

We used the manual filtering approach to remove the outliers, deleting 3 entries. In Babies', we considered any row with above 4 babies and less than or equal to two adults as outliers (2 rows). In 'ADR', there was a single row with an extreme value, so it was considered as an outlier. We also removed "ReservationStatusDate" and "ReservationStatus" because they can influence the model's performance. If we include them, we were giving a variable to our algorithm that have information about the target, so our analysis would be biased.

3.2. Feature Engineering

Next, we decided to create new features that would bring value to the models:

- "Kids" - numeric variable that represents the sum of the "Babies" and "Children".
- "RoomChange" - categorical variable that compares the "ReservedRoomType" to the "AssignedRoomType", attributing "0" when there is no change in the rooms, "1" when the rooms are not the same.

To reduce the cardinality in categorical variables, we selected the dominant categories and aggregated the ones with fewer entries. These are the variables we transformed with their final categories: Country (top 10 countries and "Other"), Agent ('9', 'Other' and 'NULL'), Company ('NULL' and 'Company'), MarketSegment ('OnlineTA', 'Offline TA/TO', 'Direct', 'Corporate' 'Groups' and "Other"), DistributionChannel ('TA/TO', 'Direct' and 'Other').

3.3. Data Normalization

In this dataset, we have categorical and numeric variables. To improve the efficiency of our modeling algorithm, we needed to transform the features. For non-metric variables, we encoded them using OneHotEncoder. For metric variables, we decided to use MinMax Scaler, changing the values to a common scale between 0 and 1.

4. Modeling

4.1. Algorithms and evaluation metrics

In the Modeling phase, we decided to apply threen algorithms. Considering our goal, we decided to choose DecisionTreeClassifier, RandomForestClassifier and CatBoostClassifier[1], since all of them can deal with numeric, but also categorical variables, which is important in our case. Besides, the algorithms belong to the group of ensemble models, so they can select the best features during their induction process. For that reason, and since we don't have a dataset with a high number of variables and we don't have IDs that could influence the predictions, we let the algorithms selecting their most meaningful features. Another advantage of these models is that we can understand which variables are the most important to implement business campaigns and to avoid future cancellations.

We will select the final algorithm based on the comparison of three metrics: precision, recall and f1 score. The metrics' choice reflects the business goal of cancellations' reduction. For that, we need to make sure that we have a considerable proportion of predicted cancellations that correspond to actual cancellations (precision), but also to be confident that our algorithm can predict a considerable proportion of all the actual cancellations (recall). After that, we selected the final model considering the f1 score, since it's a weighted average of the two metrics above and it's suitable for imbalanced datasets.

4.2. Modeling process

Regarding modeling process, we split the data into two groups: train and test with 80% and 20% of the main dataset, respectively. To make sure that we have the same proportion of the target categories in each subgroup, we used random stratified sampling. We chose Stratified 10-Fold Cross Validation, which is an iterative method that allows us to evaluate the model 10 times, using 10 partitions with the same proportion of the target categories, that we can use to select the best parameters – 9 subgroups to train and 1 to test changing consecutively. We used GridSearch to optimize the models. One of the parameters, 'class_weights', was used in both algorithms to surpass the different target proportion.

The final metrics' results can be seen in Table 1. The results of the confusion matrix of each algorithm can be seen in Table 2. Analysing DecisionTreeClassifier, we notice that 65% of our predicted cancellations were truly cancelled (precision = 0.65). We can predict 73% of the actual cancellations (recall = 0.73). Looking at RandomForestClassifier's results, 56% of the expected

[1] CatBoost is an algorithm for gradient boosting on decision trees.

cancellations were well predicted (precision = 0.56). Our model is able to predict 82% of the actual cancellations (recall = 0.82). From CatBoostClassifier, 64% of our predicted cancellations were predicted correctly. (precision = 0.64). We can predict 81% of the actual cancellations (recall = 0.81).

	Labels	Precision	Recall	F1 Score
Decision Tree	**0**	0.88	0.83	0.85
	1	0.65	0.73	0.69
Random Forest	**0**	0.9	0.73	0.81
	1	0.56	0.82	0.67
CatBoost	**0**	0.91	0.81	0.85
	1	0.64	0.81	0.71

Table 1. Classification Report: final metric results

Algorithm	TP	TN	FP	FN
Decision Tree	6159	2354	853	1289
Random Forest	5415	2627	580	2033
CatBoost	6012	2583	624	1436

Table 2. Confusion Matrices' results

Analysing the results from Table 2, it's possible to see a high value of well predicted cancellations in Random Forest Classifier (2627 records). However, the number of not well predicted cancellations is also high (2033 records), which could be a problem when we are assessing the performance of our algorithm. The Decision Tree has the highest value of False Positives (853 records) and the number of False Negatives is the lowest (1289 records). The CatBoost' results are more balanced than the other algorithms.

Our goal is to build a model that guarantees us the best results for the defined evaluation metrics, but also the best performance considering our goals. Taking into consideration the conclusions above, the results provided in the Confusion Matrix and the F1 score values for each model, we selected the CatBoostClassifier as our final solution.

5. Business Implications

Applying this algorithm to the business could provide us useful information. It is possible to understand which are the seven most important features that are responsible for cancellations: CountryPRT, Agent9, LeadTime, TotalofSpecialRequest, RequiredCarParkingSpaces, PreviousCancellations and ADR (Annex 4).

We can select services/products that the hotel could offer to avoid cancellations. In this case, we can see that the ones who have a booking with a high number of special requests register a low probability to cancel it. Another example is with the car parking spaces: if the customers request car parking spaces, they have a low probability to cancel their bookings. Since these two services could affect the customers' decision, the hotel should consider providing them with discounts to prevent cancellations. It's possible to see that bookings that are coming from Travel Agencies have a high probability to cancel, especially from Agent 9. We can notice that transient bookings associated with other transient bookings (Transient-party) have a low probability to cancel.

Looking at the violin plot[2] (Annex 5), Portuguese bookings have a high probability to cancel. However, this could not be 100% true because, considering hotels' procedures, when they don't know the nationality of the customers, they assume the default one (in this case, the hotel is in Portugal, so they assume Portuguese) until guests' check-in. Another important aspect is that reservations with a high number of days between the booking and the arrival dates (high LeadTime) have a high probability to cancel.

Moreover, the False Negatives predictions of the model can be associated to overbooking. Nevertheless, H2 can compensate this situation by offering them special services based on bookings' value. This might reduce the profit margin of the hotel in a short term. However, this approach is a way to create important loyalty relationships with the guests (since we can see in this dataset that this is the first time of the most of them) that could be profitable in a long term. Finally, this model could be used to improve user experience, customize/individualize services, understand customer behavior, predict future cancellations and, eventually, increase the hotel's revenue.

6. Suggestions for Deployment

To have some insights from the model's application, it's important that H2 has a proper structure to make use of it. For that reason, our suggestion is to create a framework that could link different departments in the business process. This could be useful because when the model predicts a cancellation, the information could be immediately passed to the responsible team. Then, the department contacts the customers.

It is important not to forget that this algorithm was created based on past data i.e., behavior of previous bookings. However, due to the occurrence of different events, changes in the behavior of the customers could arise (COVID pandemic, economic crisis are some examples of events). For that reason, we need to make sure that the model will be revised over time to guarantee its best performance.

[2] The violin plot is presenting the SHAP (Shapley Additive Explanations) values, which is a useful technique for machine learning algorithms' interpretation. It's important to know that this plot is just showing the importance of each variable regarding predictions, not evaluating the quality of the algorithm's performance.

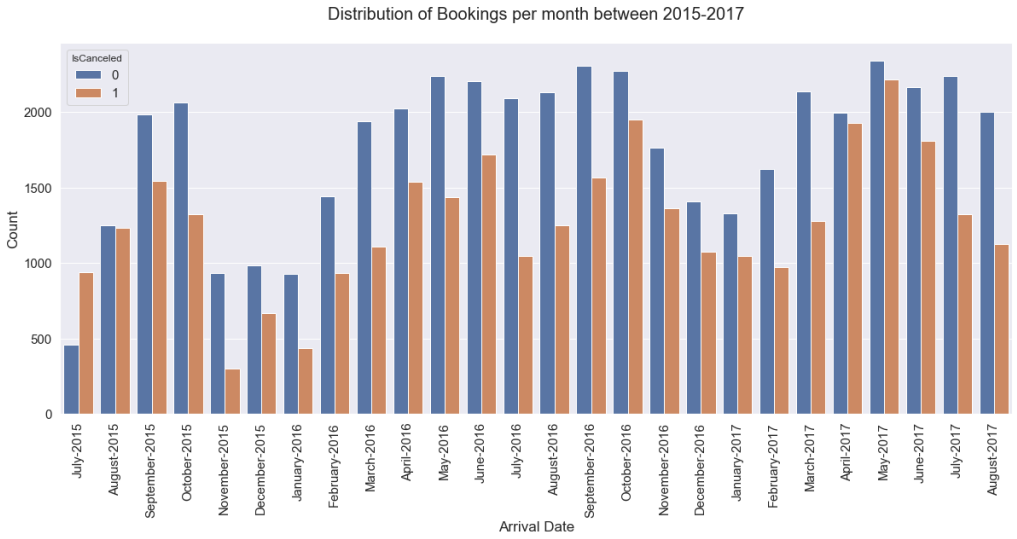
Annexes

Metric Variables	Mean	Minimum Value	Maximum Value
ArrivalDateWeekNumber	27.18	1.0	53.0
ArrivalDateYear	-	2015	2017
ArrivalDateMonth	-	-	-
ArrivalDateDayOfMonth	-	1.0	31.0
LeadTime	109.74	0.0	629.0
StaysInWeekendNights	0.80	0.0	16.0
StaysInWeekNights	2.18	0.0	41.0
Adults	1.85	0.0	4.0
Children	0.09	0.0	3.0
Babies	0.01	0.0	10.0
PreviousCancellations	0.08	0.0	21.0
PreviousBookingsNotCanceled	0.13	0.0	72.0
BookingChanges	0.19	0.0	21.0
DaysInWaitingList	3.23	0.0	391.0
ADR	105.30	0.0	5400.0
RequiredCarParkingSpaces	0.02	0.0	3.0
TotalOfSpecialRequests	0.55	0.0	5.0

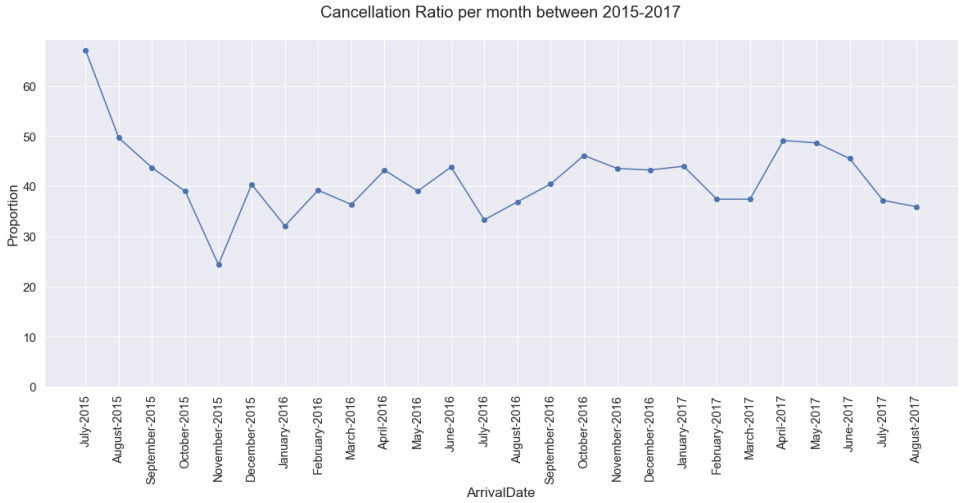
Annex 1a. *Metric Variables/Features*

Non-Metric Variables	Mode
IsCanceled	0
Country	PRT
MarketSegment	Online TA
DistributionChannel	TA/TO
Meal	BB
IsRepeatedGuest	0
ReservedRoomType	A
AssignedRoomType	A
DepositType	No Deposit
Agent	9
Company	NULL
CustomerType	Transient
ReservationStatus	Check-Out
ReservationStatusDate	-

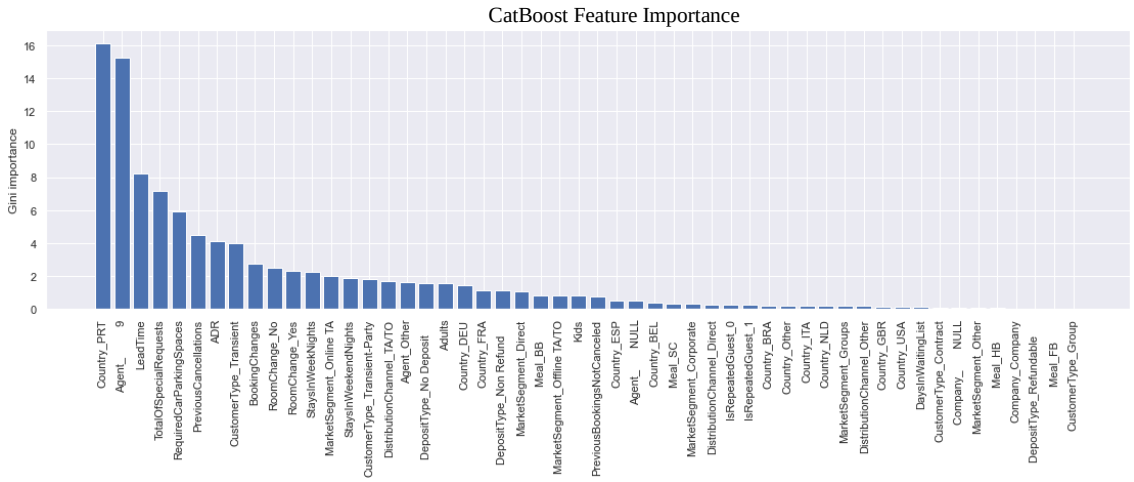
Annex 1b. *Non- Metric Variables/Features*



***Annex 1**. Distribution of Bookings per month (2015-2017)*



***Annex 3.** Cancellation Ratio per month (2015-2017)*



***Annex 4.** Feature Importance (CatBoost Algorithm)*

***Annex 5**. Violin Plot (CatBoost)*