BUSINESS CASES WITH DATA SCIENCE

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN DATA SCIENCE

Business Case 2 – HOTEL CHAIN C

Group H

Hiromi Nakashima, m20201025

Manuel A. F. Carreiras, m20200500

Luis F. R. Agottani, m20200621

Venâncio Munhangane, m20200579

March, 2021

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

TABLE OF CONTENTS

1. INTRODUCTION

In this project we were designated to analyze the Hotel chain C losses of revenues due to booking cancellations. With this problem the Hotel decided to adopt an overbooking policy which resulted in several issues.

At this moment, Hotel Chain C lost almost 28% of the bookings in Hotel H1 and 42%, in H2. The table below summarize the hotels cancelling behavior after the aggressive overbooking policy implemented.

Hotel	Metric	Not Canceled	Canceled
H1	Bookings	28,938 (72,2%)	11,122 (27,8%)
H1	Room Revenue	11,601,850€ (66,5%)	5,842,177€ (33,5%)
H2	Bookings	46,228 (58,3%)	33,102 (41,7%)
H2	Room Revenue	14,394,410€ (56,9%)	10,885,060€ (43,1%)

In order to reverse the hotel booking reality, were provided the H2 hotel's booking dataset from July 1, 2015 until August 31, 2017. The report was adjusted in four main parts based in CRISP-DM methodology (Pete Chapman, 1999).

The GitHub repository where all the present analysis is saved can be accessed through the following link: https://github.com/hnakashima96/Business-Case-.

2. BUSINESS UNDERSTANDING

At this stage we defined the essential business guidelines to grant a good result of the project and to develop the best solution for the Hotel Chain C, specifically in a city hotel (H2). As it can be seen from the table above the H2 hotel's loss of revenue were more than 10 Million Euros. To reverse this scenario the developing of predictive models can play a huge helping to address the problem of cancelations. The hotel will be able to reduce the losses by not confirming the bookings from customers who are predicted with high likelihood of canceling or making offers in advance to try to prevent cancelation. To sum up is important to have another model to predict the number of not canceled and canceled bookings per week giving insights for the management to know the weeks that they must prevent the cancelations or not confirming the bookings and accepting more reservation.

2.1. Business Objectives

The goals of Hotel Chain C, specifically in a city hotel (H2) are:

- Implement a predictive model to classify the reservations of the hotel.
- Identify the probability of booking with high likelihood of cancelling.
- Decrease in 20% the future booking cancel.

2.2. Business Success criteria

Based on the business objectives description, a main result was defined to guarantee the success of this project: defining the monthly goals to decrease 20% of booking cancellations over a year.

2.3. Determine Data Mining view

Based on the business goals we defined the Data Mining goals as shown in the table below (Table 1). Also, we defined the success criteria for these goals to support the confidence of our results.

Table 1. Data Mining Goal s and Success Criteria.

Data Mining Goals	Success Criteria
Apply a classification algorithm to identify the booking cancellations.	Evaluation of 90% of the classification model
Implement a forecast algorithm to predict the booking goals	Evaluation of 80% in forecasting model

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data understanding

The dataset provided with the information of the H2 hotel contains 31 variables describing 79,330 reservations. Those reservation are orders from clients that arrived and orders that were cancelled. In order do build the models the dataset was cleaned and modified.

3.1.1 Data Visualization

In order to understand the case we produced several data visualization, but we are going to present the ones that we considered most important than others can be consulted in the notebook.

a) Cancellations and non-cancellations bookings per month

The figure below shows that the average cancelation rate is above 30% every month. On another hand the not cancelled bookings rate is above 50%. The peak of the cancelation rate tends to occur in the second quarter (April until June) and low cancelation rates tend to occur in March and November.
For the H2 Hotel, for timeline observed in our database, we can see that:

- The Hotel has a loss per cancellation around 328,83 €.
- Each confirmed booking has a revenue around 311,38 €.

This numbers urges the management for the importance of reducing the rate of cancellation.

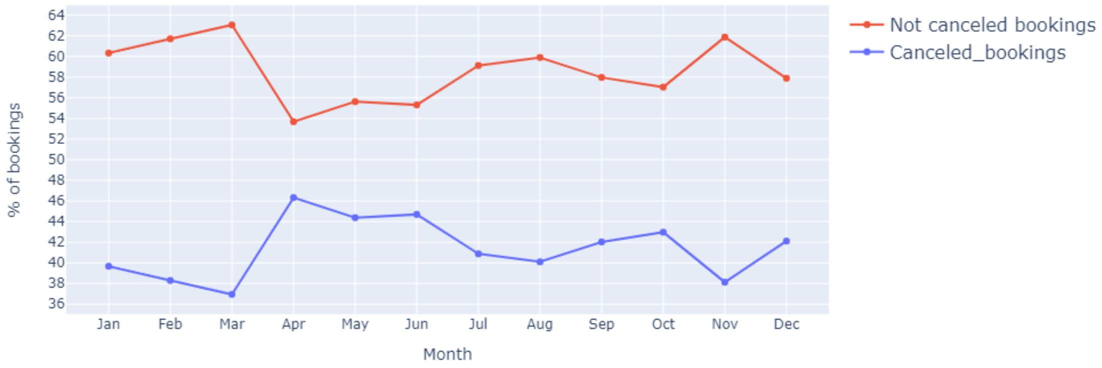

Figure 1. Cancellations and non-cancellations bookings per month (2005-2017)

a) Relation between Number of Nights and Number of Guests

From fig. 2 it is found that for the H2 hotel most customers that do not cancelled their bookings tend to make their reservations for the weekend and they also end up not staying for the night.
On the other end customers who booked in weeknights have preference to stay only two nights.

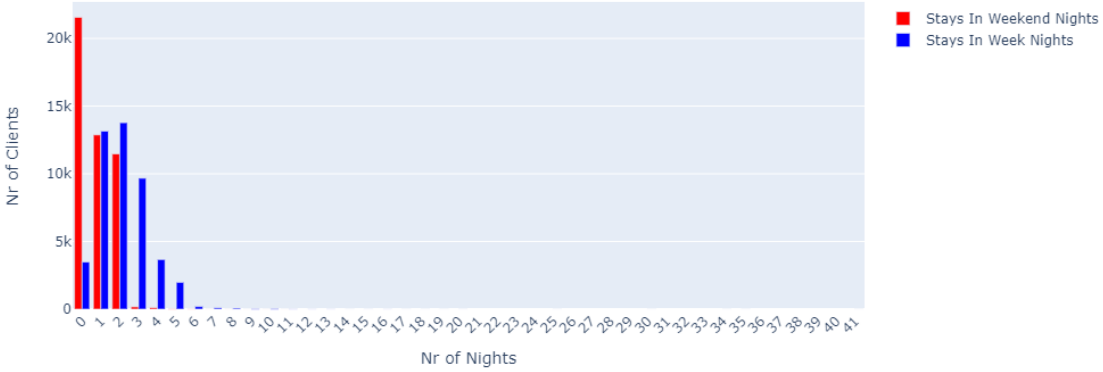

Figure 2. Clients per number of nights (2005-2017).

b) Relation between Lead Time and Cancellation

The figure below illustrates the Lead time difference between cancelled and not cancelled bookings, in days. It is clear that customers who cancelled their bookings tend to have bigger values for this variable. On the other side it is possible to observe that customers who have a short lead time rarely cancel.

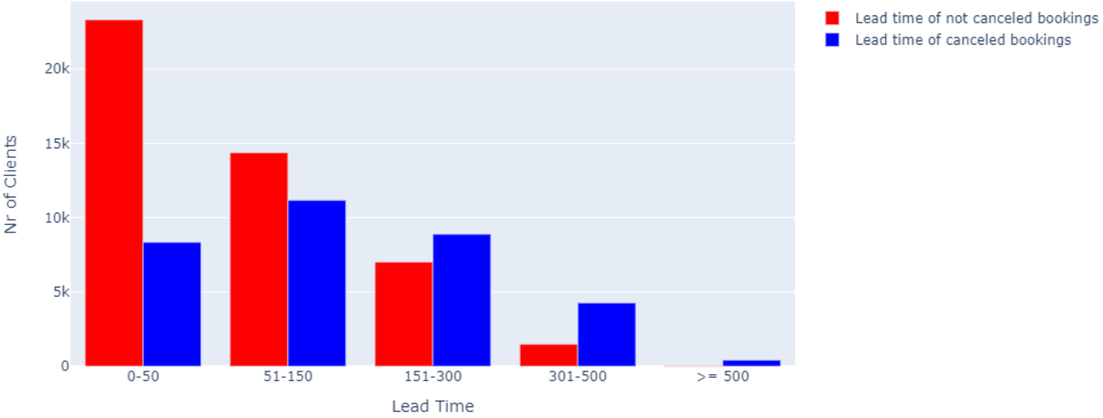

Figure 3. Lead Time by cancellation status (2005-2017)

3.2. DATA PREPARATION

Now we are going to describe the principal activities in order to obtain the final dataset used in the models.

3.2.1. Data preparation to predict booking cancellation

For this purpose, we dropped 9 features (“ADR”, “AssignedRoomType”, “ArrivelDateYear” “Agent”, “Company”, “DaysInWaitingList” “ReservationStatusDate”, “ReservationStatus”). They were discarded due to some provided information that supposedly is obtained after the Hotel confirms the booking and others are useless.

For categorical variables was applied the *one hot encode*. To detect and remove the outliers was used the Local Outlier Factor and 5% of the observations were considered outliers. All the metric features used in the model were rescaled on a range in [0,1]. For feature selection was used the embedded method which consists of training the model with all the variables and then train it again only with the variables considered more important.

During the analyses we found out that the dataset may be unbalanced. There are 41.7% of cancelled and 58.3% of non-cancelled bookings. Most of the classification models work well with a balanced dataset because, they have the assumption that the distribution in the target variable is balanced [6]. Because of that we decide to train the model with and without over-sampling. To over-sample the data we used Synthetic Minority Over-sampling Technique - Nominal Continuous (SMOTE-NC) which is a generalized SMOTE approach to handle with mixed datasets of continuous and nominal features [8].

3.2.2. Data preparation for Time Series

To apply the other predictive model for the number of not canceled and canceled bookings per week we created 2 times series. One with the number of cancelations and other with the non-cancelations from 2015 to 2017.

3.2.3. Missing Cases

In the H2 booking dataset there are two features that have missing data. Country, Children with twenty-four and four missing values, respectively. Country is a categorical variable representing customer home country. To fill the missing values, we used the mode for the Country, which in this case is ‘PRT’. For the metric variable (children) we used KNNImputer with three neighbors.

3.3. MODELING

In the following section is presented the models used (Random Forest, Gradient Boost descent and ARIMA (Autoregressive Integrated Moving Average)). In order to tune the models to predict the likelihood of cancelling the booking was used Randomized Parameter Optimization (RandomizedSearchCV).

3.3.1. Random Forest

Random forest is an ensemble classifier with bootstrap aggregation that consists of many decision trees and trends to performs better than a single tree when there is a huge number of predicators variable [12].

3.3.2. Gradient Boost

Gradient Boost is an ensemble method that can be used to is a technique that can be used to construct an ensemble of classifiers that uses a loss function to optimize the model.

4. EVALUATION

At this stage, the models are evaluated to ensure the best quality that can achieve the objectives of the H2 hotel. The evaluation of the models to predict the likelihood of cancelling the bookings was based on accuracy, recall, precision, and F1 values generated. For this we also used the hold-out test which was set randomly to 64% of the data to train the model, 16% for validation purpose and 20% for test.

The table below provides information about all the metrics used to evaluate each model to predict the likelihood of cancelling the bookings. The table compares the six models in terms of their performance of predictions before and after feature selection (1), with unbalanced and oversampled dataset(2) and to sum up with RandomizedSearch (3).

With unbalanced dataset the scores are lower in all scenarios. The Random Forest with the default parameters outperformed other predictive algorithms with an accuracy score of 0.94, F1 of 0.86 precision 0.86, recall 0.86.

Table 2.F1 score micro of the models before the over-sample.

Model	precision	recall	F1	accuracy
Random Forest + unbalanced dataset + default hyperparameter	0.87 (0): 0.86 (1): 0.88	0.86 (0): 0.92 (1): 0.79	0.86 (0): 0.89 (1): 0.83	0.93
Gradient Boosting + unbalanced dataset+ default hyperparameter	0.84	0.81	0.82	0.91
Random Forest + SMOTE-NC + Features Selection + default hyperparameter	0.86 (0): 0.87 (1): 0.86	0.85 (0): 0.91 (1): 0.80	0.86 (0) :0.89 (1): 0.83	0.93
Gradient Boosting + SMOTE-NC + Features Selection + default hyperparameter	0.82	0.82	0.82	0.90
Random Forest + SMOTE-NC + Features Selection + RandomizedSearch	0.86	0.84	0.85	0.93
Gradient Boosting + SMOTE-NC + Features Selection + RandomizedSearch	0.86	0.85	0.86	0.93

As can be seen the random forest (SMOTE-NC + Features Selection + default hyperparameter) had the best results for training so we used it for testing and got the following scores accuracy score of 0.93, F1 0.85, precision 0.86, recall 0.85.

With Random Forest, the 5 most important features were lead time, ArrivalDateWeekNumbers, DepositType when is not refund, TotalofSpecialrequests and country when the client is from Portugal.

5. FORECASTING MODEL

For the forecasting we applied the method ARIMA (Autoregressive Integrated Moving Average) for time-series where we could understand seasonality, trend, and noise from the data to predict future results for bookings cancelations according to date.

The figures 4 and 5 indicates how well our model can predict seasonality and trend from past behaviour with observed data and one step ahead forecast. The results are good since the forecast align with the observed data.

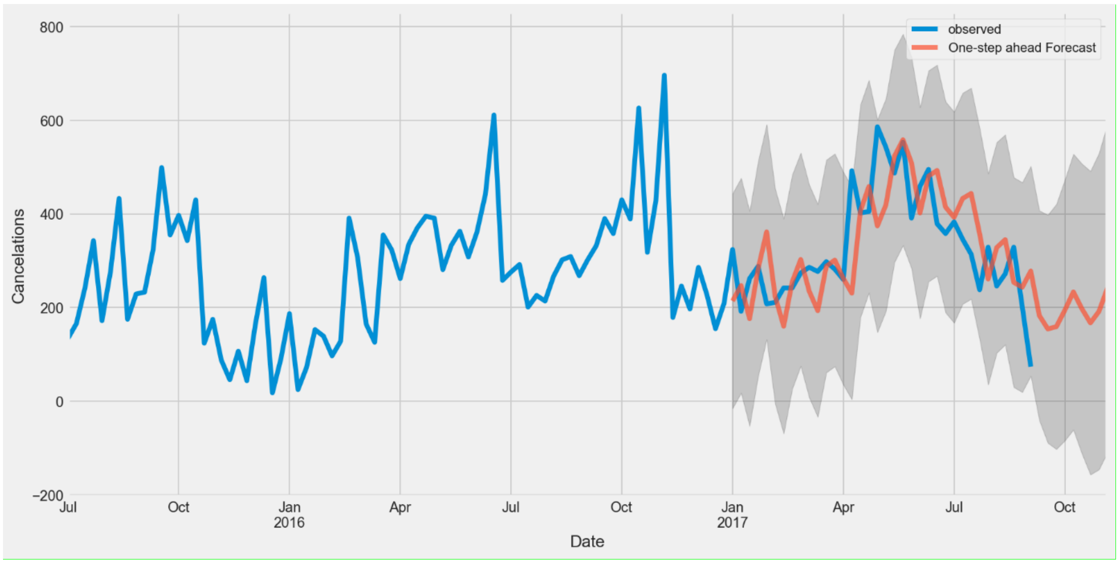

Figure 4. Validate forecasting and visualize future prediction for cancelled bookings.

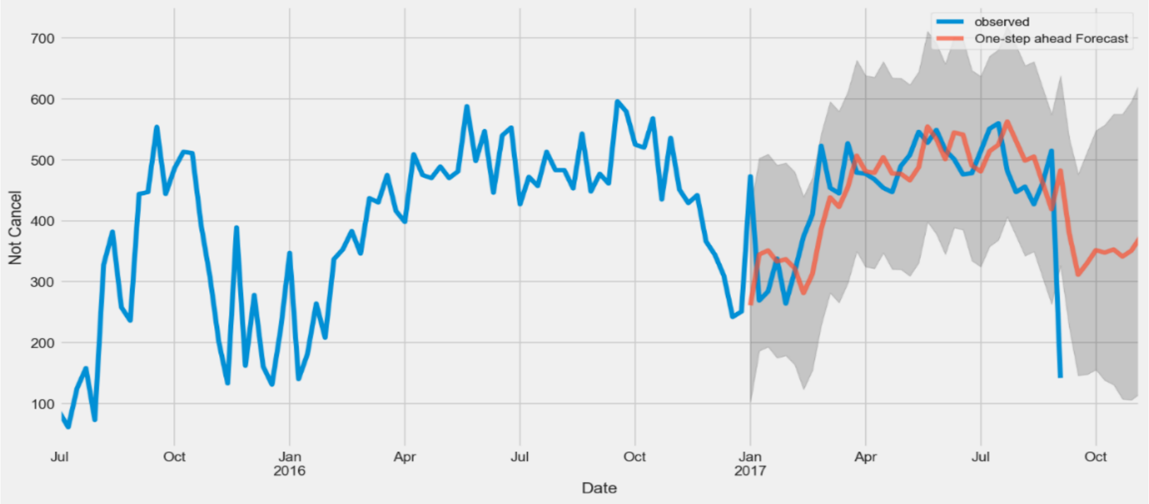

Figure 5. Validate forecasting and visualize future prediction for NOT cancelled bookings.

For the evaluation of our forecasting model, we used MSE and RMSE where the results indicates that our model was able to forecast the sum of weekly bookings status within 96 and 86 of the real observed data, which indicates good results when comparing to our data range as we can analyze in the table 3.

Table 3. Estimator quality with MSE and RMSE score.

Observations	MSE	RMSE	Range (Min – Max)
Cancelled bookings	9276.80	96.32	18 - 696
NOT Cancelled bookings	7517.83	86.71	61 – 597

With the model we can predict numbers for overbooking according to future canceled bookings to guarantee higher numbers of not canceled bookings without exceed the available numbers of rooms as we can see in figure 6.

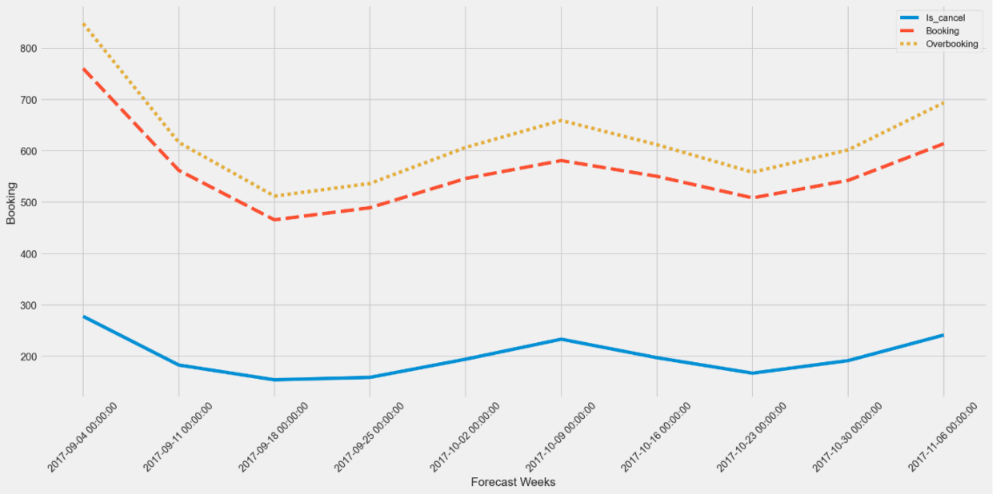

Figure 6. Validate forecasting and visualize future prediction for NOT cancelled bookings.

6. DEPLOYMENT AND MANTENANCE PLANS

At this stage we are going to describe the strategy to deploy the models.
The booking cancellation prediction model should be implemented with the forecast of not canceled and canceled bookings per week and other systems of the hotel. On the following points we described all the process.

1. Implement the predictive model for cancellation on the reservation system. These results should be passed to the manger department.
2. The manger department should combine these results with the cancellations forecast. This combination should help the manger department identifying the clients that could benefit from promotions.
3. Also, could help to define a possible strategy for the overbooking policy.

All the model and the forecast should be update frequently for better generalization of the results.

7. CONCLUSIONS

Based on the business objectives description, a main result was defined to guarantee the success of this project: defining the monthly goals to decrease 20% of booking cancellations for the next three months.

In order to finish, we relate our Business Goals with all the steps that we covered so far:

- Created a model to predict the likelihood of canceled booking with a precision of 86%, recall 85%, F1 score 85% and accuracy of 93%.
- From that point we developed a model to better understand the net demand forecast, this one with an MSE of 96,31 and 86,71.

Those results demonstrates that machine learning algorithms, in this case, the Random Forest algorithm, using datasets with the rightly identified features, is a good technique to build booking cancellations prediction models.

Understanding the net demand forecast we are able to implement in advance a strategy to fulfil the success criteria for this business. The Forecast model give us the necessary insights to develop a targeted overbooking policy for the ten upcoming weeks. The classification allows us to forecast the customer behaviour when it comes to cancellation. With this information, aligned with the timeline built for the other model, we could define a strategy to target possible future cancellations and try to drop this number by 20%.

7.1. FUTURE WORK

At this point we already have a mixed strategy to reduce cancellations and at the same time improve room revenue. We can achieve this by defining an overbooking policy and a targeted marketing to our clients with high likelihood of cancelling. But overbooking arises an issue: What if all the customers decide to show? This would lead us to loss of revenue through refunding.

For this reason, is important to define a strategy to control the overbooking percentage. First, we could try to reduce this number by trying to predict a third class in our model: Number of no-show ups. With this number we could redefine our overbooking policy making it less aggressive since we could use this rooms for possible customers that might need refunding or new accommodation.

To finish we could try to calculate the probability of overpast the capacity of the Hotel due to the overbooking and with that defining a metric of risk for this matter. This could be calculated with a Binomial distribution and more information about the capacity of the hotel.

8. REFERENCES

Michael J. A. Berry, G. S. (2004). *Data Mining Techniques - For Marketing, Sales and Customer Relationship Management.* Wiley Publishing.

Pete Chapman, J. C. (1999). CRISP-DM 1.0. *Step-by-step data mining guide.*

[5] Nguyen, Son & Olinsky, Alan & Quinn, John & Schumacher, Phyllis. (2018). *Predictive Modeling for Imbalanced Big Data in SAS Enterprise Miner and R. International Journal of Fog Computing*. 1. 83-108. 10.4018/IJFC.2018070103;

[8] Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953;