BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Case 2 – Prediction of hotel cancellations

Group W

Ana Luísa Mestre, number: 20200599

Beatriz Pereira, number: 20200674

Mariana Domingues, number: 20201040

Nadine Aldesouky, number: 20200568

March, 2020

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

INDEX

1. INTRODUCTION

The Hotel chain C wants to predict the potential future booking cancellations in order to improve their policies and consequently their revenue.

Our team of data scientists has taken on this project to help Hotels C proactively identifying a possible cancellation to apply preventive measures to reduce them.

2. BUSINESS UNDERSTANDING

2.1. Determine Business Objectives

- **Background**: Hotels C is a chain of city hotels with resorts all over Portugal. It has severely suffered from high cancellation rates even after retaliating with restrictive cancellation and aggressive overbooking policies. City hotels H1 and H2 have endured almost 28% and 42% cancellations, respectively. This problem affects the Sales, Accounts, Food and Beverage Service, Kitchen, Housekeeping, Front Office, and HR departments. As a result, Michael the Revenue Manager Director has hired Group W to help predict future cancellations. Indeed, Michael aims to implement preventive measures to reduce the cancellation rate to 20%. Currently, the growth of deal-seeking customers and Online Travel Agencies has had a very negative effect on hotels.
- **Business Objective/Problem**: To forecast net demand based on existing bookings in order to predict future cancellations and reduce them to 20%.
- **Business Success Criteria**: Successful forecasting will result in very accurate predictions about which customers have a high likelihood of cancelling and which do not.

2.2. Assess the Situation

- **Inventory of Resources**: Currently we have a database of 79330 bookings which will be analyzed by a team of four business analysts known as Group W. To manage our problem we will use Python, Microsoft Word, Power Point, and Excel.
- **Requirements, Assumptions, and Concerns**: This report is for managers' use which means that it is for private use in Hotel chain C only. Additionally, this is not a technical audience meaning that all aspects must be comprehensible in business terms without technical jargon. For this project, we will be working with the *H2* dataset composed by a sample of 79330 booking records and 31 variables, including the target variable.
- **Risks and Constraints**: These bookings were extracted from H2 historical records based on the constraint that they took place between July 1st, 2015 and August 31st, 2017. Additionally, this report must be completed and presented to management within one week's time.

Risks	Contingencies
Useless features for the task	Ask for different variables or derive new features
Missing relevant information	Ask for more detailed explanation in meta data e.g. ADR
Lack of funding or available time	Request an extended deadline
Losing the data	Create a copy of all aspects of the project
Breakdown of system or data warehouse	Perform daily quality insurance checks on warehouse to verify that the data and the programs are updated and upload them correctly

Table 2.1.– Risks and Contingencies

- **Terminology**:

ADR: Average price per night (calculated by dividing the sum of all lodging transactions by the total number of staying nights)
Travel Agents: individuals selling a set of different holiday packages organized by tour operators
Tour Operators: individuals organizing holiday tours including hotel reservations and leisure activities
Special Requests: specific requests from the customer such as twin bed, high floor etc.
OTA: Online Travel Agency
Transient booking: last-minute booking or walk-in guests
Self-catering: no meals included
Half-board: meals include breakfast and another meal
Full-board: meals include breakfast, lunch and dinner
Bed and Breakfast: overnight accommodation and breakfast

- **Costs and Benefits**:

As this is an academic project, we have no information about the costs of the data collection neither the ones regarding the development and the implementation of our solution for the business problem.
The benefits include reducing cancelations, increasing bookings, increasing ROI, and improving customer satisfaction, leading to higher revenues and lower costs.

2.3. Determine Data Mining Goals

- **Data Mining Goals**: To create a predictive model to classify which bookings are more likely to be canceled in order to help the hotels reduce the cancellation rate in the future.
- **Data Mining Success Criteria**: Successful predictiveness of the model will result in a high precision, recall, accuracy, f1, and AUC scores. This means that the model should have generally correctly predict the customers who will cancel. Indeed, the model must also not overfit the data, so it has a higher capacity of generalizing and working on any dataset. Additionally, the model should be very insightful by providing specific actionable items.

2.4. Produce Project Plan

- **Steps**

1. General exploratory analysis, identifying clear problems in the data such as missing values, outliers, or incoherencies.
2. Deeper exploratory analysis.
3. Describe the insights obtained from the data exploration - explore; produce and interpret visualizations to understand the absolute frequencies of the variables, or the presence of outliers.
4. Prepare the data to the model input – select important features, apply features transformations, create new features, detect outliers and/or correcting incoherencies.

5. Review the data again, after cleaning it (e.g.: see if the insights are the same or if they changed slightly).
6. Define and apply different techniques and classification methods to create a supervised predictive model with the binary variable IsCanceled as the target variable to classify bookings and outliers.
7. Fine Tuning (making small adjustments to parameters to achieve the desired output of performance) of the model, and selection of the one that provides the best solution for the classification of the bookings based on the higher scores.
8. Assess and interpret the results from the chosen model.
9. Provide a deployment plan.

- **Initial Assessment of Tools and Techniques**

Technique	Pros	Cons
Gradient Boosting	Flexible No need for pre-processing or imputation	Prone to overfitting Computationally expensive
AdaBoost	Easy to implement Flexible	Slow Sensitive to outliers
Random Forrest	Not sensitive to outliers Efficient Lower risk of overfitting	Costly in terms of time Biased towards variables with more values

Table 2.2.– Tools and Techniques

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data Understanding

The dataset was provided as a csv file by the Revenue Manager Director of Hotel Chain C, Michael, directly from the company. This dataset has 79330 rows and 31 columns, from which 13 are non-metric variables and 18 are metric variables. The data needs some preprocessing as it has some missing values and duplicate observations.

When exploring the dataset, we noticed that there are:

- 33076 canceled bookings, representing almost 42% of the data
- 390 bookings with no adults, but with children, which can refer to separate rooms reserved for the parents and/or children, and 167 bookings with no adults, children or babies involved in the booking, which refer to people who only used the hotel services, such as restaurant or gym
- 7192 bookings where the reserved room was not the assigned room
- 27172 bookings that were canceled and where the customer did not have made any cancellation before
- 916 bookings that were canceled and where the customers did not show up
- 4633 customers who were never hotel clients because they canceled their previous bookings

- 159 customers who were never hotel clients because they didn't show in their previous bookings
- 72492 customers who were never hotel clients because they have never made any booking in the hotel before

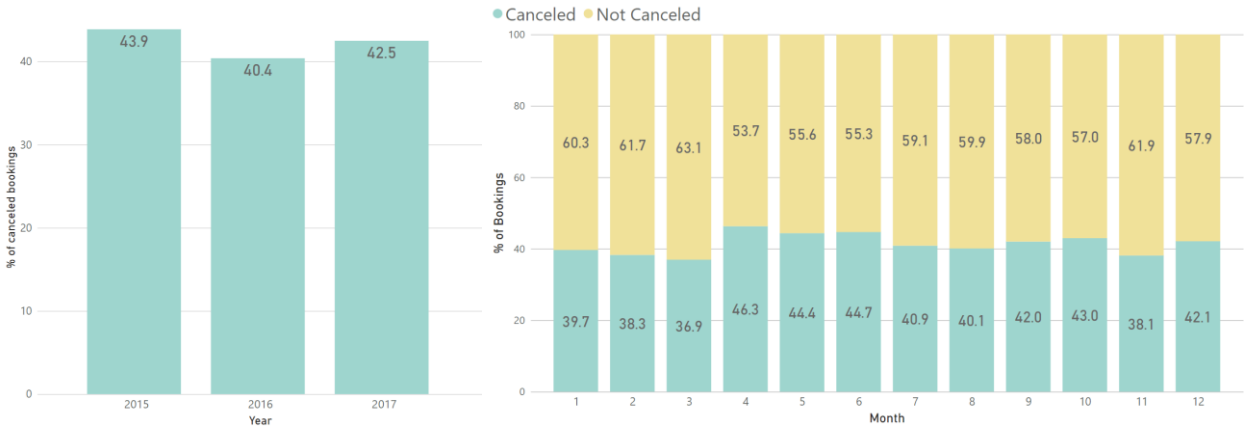

Figure 3.1.– Cancellations by arrival time

The year of 2015 has the highest percentage of canceled bookings – 43.9%. Grouping the cancellations by month of arrival, we could conclude that the months more common to have cancellations are April, May and June.

Figure 3.2.– Cancellations by type of guest

Besides this, between being a repeated guest or a new guest, the second one has a higher percentage of cancellations given us to conclude that clients who already know the hotel like to be there.

Another important insight that we could extract from the most correlated variables with the target variable – IsCanceled, are:

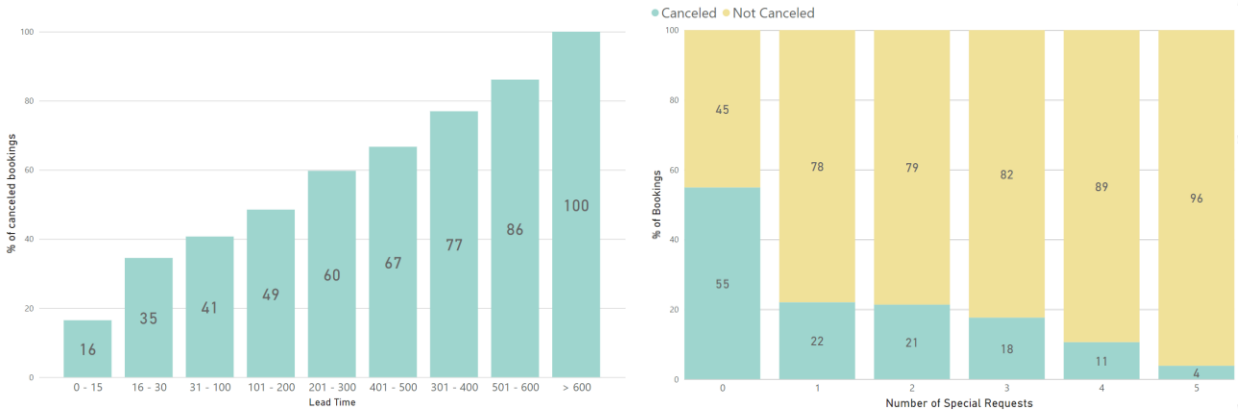

Figure 3.3.– Cancellations by correlated variables

- as the number of special requests increases, the percentage of canceled bookings decreases
- as the lead time increases, the percentage of bookings canceled also increases

3.2. Data Preparation

In this phase, we will apply some data cleaning, select the features that can be used for the predictive model and do several feature engineering.

3.2.1 Missing data

In the dataset there are only a few features with missing or undefined data. The missing data are only from two features – Country and Children, correspondent to 26 observations in total. To avoid any imputation errors and given the small quantity of observations we decided to drop them. Regarding the undefined data, we changed the NULL value of the variables Agent and Company to “not applicable” to represent cases where the booking was made in other circumstances.

3.2.2 Data Selection

Regarding outliers, we applied two different methods of outliers’ detection – Local Outlier Factor and Isolation Forest. However, the number of eliminated observations was very high in both cases and we wanted to prevent the representative sample of all booking's population, so we discarded this option. We also tried to look for some incoherencies manually where we discovered two rooms with 9 and 10 babies and ended up deleting these two observations. Analyzing the LeadTime variable we saw that it has bookings made with a large anticipation from the arrival date, almost until two years. We thought about creating a threshold, but not knowing what is the hotel’s policy we ended up assuming that the hotel allows these extreme early bookings.

It is important to understand that if we collected every variables' values after the arrival date, then we would be leaking information and biasing the results from the predictive model. To avoid this, most of the variables are only updated until the prior day to the arrival date, but there are variables, such as AssignedRoomType, DaysInWaitingList, ReservationStatus and ReservationStatusDate that are not supposed to be included in the performance of the predictive model. We want to perform the model

before the hotel gives a confirmation to the client, so these variables will not be included in the modeling section.
The variable Country is also not used because of a similar problem, is not common to know the client nationality before check-in and by default the hotel fills the variable with the country of origin – Portugal. One more time, the distribution of the variable will not be the same for people that checked in and for the ones who didn't.
There were one more variable not used for modeling – DepositType, since we got the information that it was extracted in a wrong way and its quality was compromised.

3.2.3 Data treatment and engineering

First, we treated the high cardinality from the variables Agent and Company, the two with more than 200 distinct values. For the variable Agent we kept the seven most frequent agents (including the "no agent" category) and grouped the rest of the values in a category – "Other". For the variable Company, since 95.4% of the data does not book by a company, we kept only the category "not applicable" and grouped the rest in a category – "Other".
Regarding feature engineering, we combined some features to obtain different insights. We created new features and applied some transformations that were chosen to be useful or not looking at their correlation with the target variable. This analysis can be seen in the table 3.2.
As shown in table 3.1, by summing some variables we could create a Total variable and calculate the percentage of each component.

Combined Variables	Total (Summing)	Percentages (Dividing variable by its total)
Adults	Persons	PercentageAdults
Children		PercentageChildren
Babies		PercentageBabies
StaysInWeekNights	StaysInNights	PercentageWeekNights
StaysInWeekendNights		PercentageWeekendNights
PreviousCancellations	PreviousBookings	PercentageCancellations
PreviousBookingsNotCanceled		PercentageBookingsNotCanceled

Table 3.1. – New variables created

For categorical features, we applied the one hot encoding technique. For the two features that are ordinal – Meal and ReservedRoomType, we created a second option of encoding, namely Meal2 and ReservedRoomType2, transforming them into integers from 1 to number of categories (following the quality rate).

Finally, to choose the better option for the same variable, we calculated their linear correlation with the target variable. The selection, represented in bolt in the next table, was made by the higher correlation between the options.

Option 1	Option 2
Dummy Meal	Meal2
Dummy ReservedRoomType	**ReservedRoomType2**
PreviousCancellations	**PercentageCancellations**
PreviousBookingsNotCanceled	**PercentageBookingsNotCanceled**
Adults	PercentageAdults
Children	PercentageChildren
Babies	PercentageBabies
StaysInWeekNights	**PercentageWeekNights**
StaysInWeekendNights	**PercentageWeekendNights**

Table 3.2. – Options considered in the selection of some variables

In the end, we got 58 variables and 79300 observations to proceed for the next step, where we applied a more cautious feature selection using a wrapper method to evaluate the importance of the features on a specific model.

3.3. MODELING

3.3.1 Select Classification Technique

In this step, we tried to change some parameters and the data splitting criterion. The validation of the models was made with the given data, using the hold-out method to split the data into a training and test set (80/20). However, as we have a good number of duplicates, we did not do a simple hold-out method. For our model not to repetitively use the same observations during the training and testing phase, we decided to force the duplicates to be in the same proportion in the train and test sets. The same thing was applied to the target variable so we can have the same proportion of cancellations in both sets, and do not bias our final model.
For the feature selection, we used RFE (Recursive Feature Elimination) for each model in order to select the number and the most important variables to be used. After fitting the model, we checked the confusion matrix as well as the scores of the different models to evaluate their performance and select the best one. The following table shows each model and its associated scores.

Model	Number of variables	Accuracy	Recall	Precision	F1 Score	AUC
Gradient Boosting	31	Train 0.823 Test 0.819	Train 0.706 Test 0.694	Train 0.844 Test 0.843	Train 0.769 Test 0.761	0.891
AdaBoost	25	Train 0.802 Test 0.800	Train 0.694 Test 0.682	Train 0.804 Test 0.807	Train 0.745 Test 0.739	0.861
Random Forest	17	Train 0.913 Test 0.860	Train 0.859 Test 0.762	Train 0.928 Test 0.885	Train 0.892 Test 0.819	0.929

Table 3.3. – Models' scores

Assumptions per modelling technique:

- Random Forests and decision trees make no prior assumptions about the dataset as they are non-parametric and scale invariant models.

To classify bookings into Group 0 (those who have a lower probability of cancelling) and Group 1 (those who have a high probability of cancelling), we used a random forest with a maximum depth of 18.
On the figure 3.4 you can see the 17 features used order by their importance (calculated with gini index). As we could expect, the LeadTime came on top as well as the arrival time.

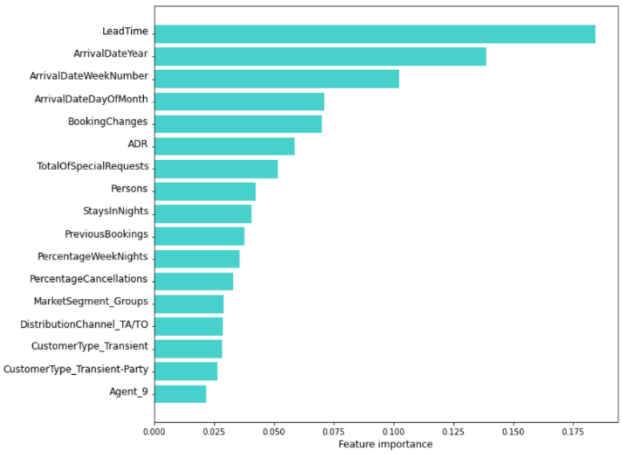

Figure 3.4. – Feature importance of the variables used

3.3 EVALUATION

3.3.1 Evaluate Classification Models' Results

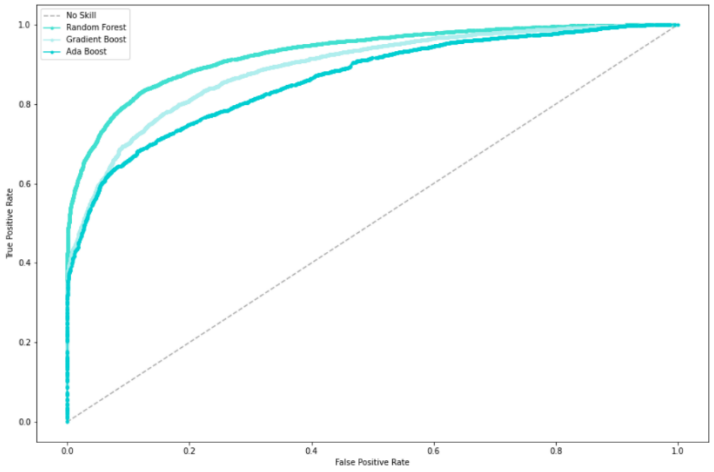

Figure 3.5. – ROC Curve

This ROC curve uses the recall and specificity metrics (with macro average). Looking at the curve, the Random Forest model has the most ideal curve which means it has the best performance. Choosing a threshold of 0.5, the model reached an accuracy of 86% on the test dataset using only 17 variables.
From the remaining 14% badly classified, 4.1% were classified as cancellations when the customer did not cancel the booking (FP), and 9.9% were classified as non-cancellations when the customer ended up cancelling the booking (FN). The following table represents the confusion matrix for the test set.

		Predicted	
Actual		TN = 8592	FP = 653
		FN = 1573	TP = 5042

Table 3.4. – Confusion Matrix

This matrix helps us understanding the scores presented in table 3.3. From the actual cancelled bookings, the model was able to identify 76.2% of them (recall metric). From the predicted cancelled bookings, the model could identify correctly 88.5% (precision metric).

3.3.3 Review Process

During this project, we had some problems regarding data leakage as we were unsure about the time to perform the model and until when do we have updated information in our dataset. Of course, we should predict a possible cancellation with as many days in advance as possible, but the manager was not clear defining this limit. The only certain we had was that the model should be applied after giving any confirmation to the customer, whenever the answer is given (either on the same day as the booking or one week later). Also, both the data and metadata were missing important information for us to understand the context and the meaning. Besides this, the time constraint was quite restricting. It would have been helpful if we had additional time to play with different parameters of the models and provide the best solution possible. Applying a bigger grid search for different parameters and subsets of the data can take days running in a common machine.

3.3.4 Determine Next Steps

- Move to the deployment step.
- Fine tuning of the classification model.
- Improve the data preparation step – outliers, feature selection, etc.
- Improve the predictive model - tuning the parameters or trying new methods (e.g. Logistic Regression, XGBoost, Neural Networks).
- Training and testing the predictive model with different data splits - using K-fold cross-validation, for example.
- Classify more than two groups (cancelled and not cancelled) creating more intervals from the output (likelihood of cancelling – 0 to 1) and be able to act differently for each type of booking.

4. DEPLOYMENT

4.1. Deployment Plan

After having classified the bookings into two groups, the next step is to determine how to proactively reach customers who are more prone to cancelling and convince them to not convert. This could be done

by providing more price sensitive deals, or packages which include activities and meals as these tend to be the Deal-seeking non-loyal customers.

Furthermore, if Hotel Chain C has more budget in the future, they can create a more automated solution like an app which uses our predictive model. Being able to input values into a list of features for the model which then outputs the likelihood of the customer cancelling in the future as a result would be sufficient.

4.2. Plan Monitoring and Maintenance

- Review the model when its predictiveness decreases a lot
- Following the pandemic situation, the model cannot be applied now because the data used to train it was collected before; so, in order to predict well the bookings more likely to be canceled, we must have more recent data (from 2020) that matches the current context.
- If the Hotel moves to a more digital business model, it can benefit of an app/program that alerts when the data from the booking is updated if it is probable to cancel.
- If the booking base increases by 25%, new competitors enter the market, or the landscape of the market changes e.g. Travel agencies shutdown; this project must be reviewed because this could give rise to new customer profiles and behaviours.
- If current conditions do not change, the analysis must be reviewed within one quarter's time at most in order to ensure its viability and relevance since touristic and hotel data is quite seasonal.

4.3. Review Project

To sum up, although the dataset was of good quality, we believe that additional data would have been very beneficial for the project. Supplementary information about the hotel i.e. its location or policies would have been beneficial in order to analyze the Lead time variable to choose a certain threshold for cutting off outliers. Moreover, the date at which the special requests and booking changes took place would also be helpful to better understand and predict the booking behaviour of the customers. Consequently, the generalization capability of the model could have been more accurate with more data for training and testing to ensure their credibility.