BUSINESS CASES WITH DATA SCIENCE

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

Business Case 2 – Hotel Chain C

Group J

Francisco Hermenegildo, number: 20200737

Gil Gonçalves, number: 20201066

Ikram Bouziri, number: 20200753

March 2021

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

INDEX

1. INTRODUCTION

Hotel chain C is a chain with resort and city hotels in Portugal. The company's revenue is being severely impacted by a big amount of bookings cancellations, especially in hotel 2 where almost 42% of total reservations are canceled. The management of the hotel decided to limit the negative impact of cancelations by applying restrictive cancelation rules along with aggressive overbooking policies which resulted in either non-sold rooms or additional costs.

This project will enable the manager of the hotel to predict net demand and identify bookings with a high likelihood of being canceled. This report contains an in-depth analysis of the dataset, extracting valuable information that the manager of the hotel chain C can use to make better decisions regarding overbooking policies and possible offers that could prevent cancellations to finally achieve the manager's goal to reduce cancellations to a rate of 20%.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Hotel chain C is a chain with resort and city hotels in Portugal that was severely impacted by cancellations, representing almost 28% in H1 and 42% in H2. To limit the number of rooms sold with restrictive cancelation policies, the revenue manager director of hotel chain C implemented an aggressive overbooking policy which generated costs. For this reason, the manager eased the overbooking policy but this resulted in the hotel having inventory not sold, even on high demand dates.

2.2. BUSINESS OBJECTIVES

In this project, we seek to solve the problem of the negative impact caused by cancelations.

This goal will be achieved by forecasting the net demand of the hotels based on reservations on-the-book, especially in hotel 2. The manager will be able to detect the reservations with a high likelihood of canceling and find solutions to prevent the cancelation. Getting insights into the customers' behavior will help the manager to implement better pricing and overbooking policies.

2.3. BUSINESS SUCCESS CRITERIA

At the end of this project, the hotel management will be able to predict the likelihood of a customer canceling his booking in order to efficiently implement its policies to reduce the impact of cancelations.

2.4. ASSESS SITUATION

The development of this project implied a team of three data mining students, all with the necessary computing and software skills. The company provided access to a sample of its customer's database through the excel H2.xslx. This dataset contains information related to all the bookings made in hotel H2, which were due to arrive between July 1, 2015, and August 31, 2017

Hotels chain C required this task to be completed and presented to the management team by the 15th of March 2021.

2.5. DETERMINE DATA MINING GOALS

The goal of this supervised learning activity is to develop a classifier that enables Hotel Chain C to forecast net demand (demand minus cancellations) based on past reservations, implementing at the same time better pricing and booking policies. To achieve this, the group set the following goals for this project:

- Determine which features contribute the best to the predictive model.
- Build new features from the original ones (if possible).
- Experiment with at least 3 different classification algorithms.
- Accuracy above 0.8 and below 0.9 (prevent overfitting).

3. PREDICTIVE ANALYTICS PROCESS

3.1. DATA UNDERSTANDING

(Data Collection) All the necessary data for the development of this project was provided by the Hotel Chain C management team, in the form of an excel file containing data about the bookings of Hotel2 which were due to arrive between July 1, 2015, and August 31, 2017. As complementary data, it was also provided the respective metadata and a PDF file with the business description and requirements for the project. The data was collected directly to a Jupyter notebook, which allowed all the necessary operations to develop a successful predictive model.

(Data Description) The excel file contained one single dataset with 79330 bookings (rows) and 31 attributes (columns), being characterized as following:

Numerical	Categorical	Binary	Dates
LeadTime	Meal	**IsCanceled**	ReservationStatusDate
ArrivalDateWeekNumber	MarketSegment	IsRepeatedGuest	ArrivalDateYear
StaysInWeekendNights	DistributionChannel		
StaysInWeekNights	ReservedRoomType		
Adults	ReservationStatus		
Children	ArrivalDateMonth		
Babies	AssignedRoomType		
PreviousCancelations	Country		
PreviousBookingsNotCanceleled	DepositType		
ADR	CustomerType		
RequiredCarParkingSpaces	Company		
TotalOfSpecialRequests	Agent		
LenghtOfStay			
ArrivalDateDayOfMonth			
BookingChanges			

Table 1 - Types of features present in H2 dataset

(Data Exploration) An extensive initial analysis of the dataset was performed using several data mining and visualization tools, each one for the respective purpose.

Tool	Use
Pandas-Profiling	Overview of the entire dataset and its characteristics
Boxplot	Identify possible univariate outliers
Histograms	Visualize variable frequencies
Scatterplot	Identify possible bivariate outliers

Table 2. Different data exploration tools used

(Data Quality) With the use of the different tools in the previous stage, its was stated the presence of outliers, missing values, and duplicated rows.

3.2. DATA PREPARATION

(Data Cleaning) Regarding missing values, only two features possessed them, which were "Children" and "Country". For these variables, the following solutions were implemented:

- For the variable "Children": the missing values were replaced by 0 (4 rows affected).
- For the variable "Country): the rows with missing values were dropped (24 rows affected).

The presence of outliers was immediately identified in various features through the visualization of the respective boxplots. The implemented approach was based on manual filtering while keeping the total amount of entries removed below 3%. To note that the IQR method was used as a first attempt, however, this approach wasn't by any means the appropriate one as it would remove more than 50% of the bookings.

The final set of filters implemented are present in Table 3 - Filters applied in outlier removal and translate into a total removal of ≈ 2.5% of records.

Variable	Filter (entries to remove)
LeadTime	> 500 days
LengthOfStay	> 10 nights
BookingChanges	> 3 changes
Babies	> 8 babies
DaysWaitingList	> 140 days
PreviousBookingsNotCanceled	> 7 bookings

Table 3 - Filters applied in outlier removal

Finally, it was decided to keep the duplicated rows since in the real world these are situations that can happen as a coincidence. (representing different bookings with the same characteristics).

(Data transformation)

After the exploration of the dataset, a few variables were transformed in order to possibly improve the training of the classifier and make it more robust.

- For the variables "RequiredCarParkingSpaces" and "PreviousCancellations", the conversion to binary was performed but it proved to not be useful at the end.

- The values of the variable 'ArrivalDateMonth' were transformed from the names of the months to the number of the months.
- Unnecessary spaces were identified and quickly removed of the features "Company", "Agent", "AssignedRoomType", and "ReservedRoomType".

(Feature engineering) During the initial data analysis, we found it critical to create new attributes from the original ones, adding mixed measures, enabling a better understanding of the data without as many variables.

New Attribute	Attributes
ArrivalDate	Agregation of "ArrivalDateYear", "ArrivalDateMonth" and "ArrivalDateDayOfMonth"
LenghtOfStay	"StaysInWeekendNights" + "StaysInWeekNights"
CancelLeadTimeDays	"ArrivalDate" – "ReservationStatusDate"
TotalCostOfStay	"ADR" x "LenghtOfStay"

Table 4. Features created and their rationale

(Data Formatting) Two data formatting techniques were necessary before pursuing the modeling phase. The fact that all metric features were in different scales meant that it was necessary normalizing them, and for this purpose, we used StandardScaler(). Contrarily, since all supervised learning algorithms require only numerical input, we used One Hot Encoding technique to convert the categorical features into numerical to enable them to be applied to the Classifier.

(Data Selection) The feature selection process was based upon the use of various techniques that evaluates the respective importance in terms of relevancy to the target variable. From this perspective, different correlation techniques were considered (present on Pandas Profiling report) since the dataset contained different types of features. Besides correlation between variables and the target variable, RFE, RIDGE, and LASSO were also applied to obtain complementary insights. In the end, the chosen subset of features with the purpose of training the model was the following:

- LeadTime
- PreviousCancelations
- PreviousBookingsNotCanceleled~
- ADR
- RequiredCarParkingSpaces
- TotalOfSpecialRequests
- LenghtOfStay
- BookingChanges
- x0_Online TA
- x0_Groups
- x2_Transient

3.3. MODELING:

3.3.1. Select Modeling technique

Having achieved the final subset, the selection process of the final model was iterative, experimenting with different algorithms and then selecting the best performing one, giving the best accuracy possible. Logistic Regression, Random Forest Classifier, GradientBoostingClassifier, and XGBoost were the Classifiers with the best initial results without any tunning of the respective parameters.

3.3.2. Build Model

Since we are dealing with a supervised learning problem with a binary dependent variable, some known models and also the not-so-common XGBoost were tested to see which one returned the best accuracy.

Parameters	F1-Score	1's Score	0's Score
Logistic Regression	0.77	0.69	0.81
Random Forest	**0.84**	**0.80**	**0.87**
GBC	0.79	0.72	0.83
XGBoost	0.82	0.76	0.85

Table 5. F1-Scores for each Classifier

As verified in the table above, the RandomForest Classifier yielded the best accuracy for both 1's and 0's. Naturally, the next step was to optimize its hyperparameters resorting to GridSearchCV and check for improvements. The optimized parameters are found in the table below.

Min_samples_split	N_estimators	random_state
7	800	42

Table 6. RandomForest optimized parameters

The improvement in the overall accuracy was marginal but nevertheless, it improved. The confusion matrix associated with the final run of the classifier is the following.

		Predicted Values	
		Yes	No
Real Values	Yes	8172 TP	856 FN
	No	1575 FP	4836 TN

Table 7. RandomForest Confusion Matrix

3.3.3. Model Assessment

To assess the consistency of the model's predictive capability, we introduced the Repeated K-Fold cross-validation. We defined k=10 as this is known to be a good value and the results are the following.

Table 8. Repeated K-Fold results. K=10

Accuracy	Standard deviation
0.848	0.005

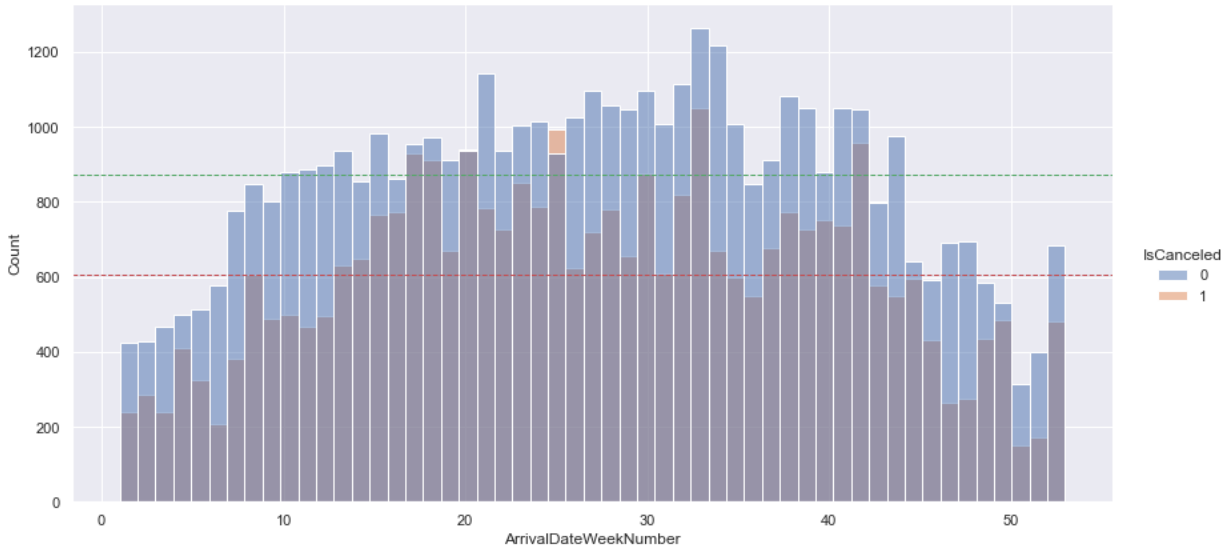
4. RESULTS EVALUATION

Putting all the pieces of the puzzle together we can finally understand our objective key concepts and put them into context.

The first key concept we are addressing is **Net Demand**.

Since Net Demand is all the Bookings that were not canceled, we decided to observe this Statistic by week of the year. This way, we believe it is easier to understand which weeks have the highest net demand and consequently, the ones that have the least.

The following distribution plot represents the distribution of all the reservations made from July 2015 to August 2017.



Interpretation of the distribution plot colors:

- Light-blue - indicates the number of not canceled reservations
- Dark-blue - indicates the number of canceled reservations
- Orange - indicates the number of canceled reservations exceeding not-canceled ones

We can verify that all weeks have a net demand higher than 50% of the total of reservations except for Week 25, where more than half of the total of reservations were canceled. We can also verify that some weeks the number of canceled reservations almost equals the number of successful reservations

On the plot, there are also two lines, both representing the average amount of reservations, – the green one with 872 successful reservations/week - and the number of canceled reservations – the red line with 605 cancellations/week.

Hence, the next key concept is **Overbooking.**

After an eye-level analysis of the distribution, it becomes clear that there are some weeks that, due to having almost the same number of reservations canceled and not canceled, have a bigger margin to allow overbooking.

To further complement this information, additional analysis on the canceling customers was required. The focus of the analysis went into reservations canceled within 14 days, 5 days, and on the day of Arrival.

Cancel LeadTime	0	5	14	**All**
Number of Reservations	1382 ~4%	3169 ~10%	6620 ~21%	**32.056**
Revenue "lost"	392.734€ ~4%	897.271€ ~9%	2.001.096€ ~20%	**10.229.455€**
ADR	100€	106€	110€	**104€**
Room Type Choice	A 77,6%	A 77,6%	A 76,8%	**A** **82,3%**
Meal Choice	BB 78,8%	BB 78,7%	BB 76,2%	**BB** **80,3%**

From the table above there are a few conclusions to be taken:

1. Customers tend to book Rooms type A;
2. Customers tend to include the Bed & Breakfast Meal type;
3. The impact on the revenue just from the canceling customers with 14 days to check-in, amounts to 20% of the total revenue lost from cancellations;
4. Customers who canceled on the day of the check-in made the Hotel lose 400k€.

This leads us to our last key concept, the **Restrictive CancelationPolicies.**

To avoid last-minute cancelations that prevent the hotel from filling the free room on such short notice, implementing some sort of Cancelling policies that discourage customers from canceling on the day of arrival could mean an increase of revenue of about 400k€ according to the research done from July 2015 to August 2017.

Also, the customers that cancel 5 days prior to the arrival date don't give much time for the hotel to generate an alternative customer for the vacant room, it is worth taking a closer look at this segment since they represent 10% of the total number of cancellations and of not earned revenue.

5. DEPLOYMENT AND MAINTENANCE PLANS

With the predictive model at the disposal of the hotel management, the decision-making process is now data-driven, enabling a bigger certainty regarding all the decisions made and also much more reliable expectations.

5.1. PLAN DEPLOYMENT

Our team suggests the implementation of the model that allows real-time feedback in order to make decisions such as the amount of Overbooking or defining the necessity of restricting cancellations.

5.2. PLAN MONITORING AND MAINTENANCE

After the deployment of the plan, it is advised a close follow-up of the generated results. If any inconsistencies are identified, it is suggested to schedule a meeting at any point in time with the team in order to assess and modify if needed, the marketing strategies provided in this report.

6. CONCLUSIONS

This second business case served to further improve the implementation of the CRISP-DM framework in a Data Science project.

Being the scope of this project the implementation of a predictive model, the insights we gained from the first one, where the aim was to segment a market, differ in a sense that when working with a predictive model, we are more tied down in terms of creativity, but in return, we got into a more factual, theoretical approach.

- The presence of bookings with the same characteristics (duplicated rows), although difficult to understand at first, can happen in the real world due to the large number of bookings that the hotel chain C has in specific timeframes.
- In the end, we managed to build a predictive model with very good accuracy, however, it is not certain that it is not overfitting, since the test set can have duplicates found on the training set.

Overall, it was a captivating and enriching experience to once again work on a real-life case.