



**NOVA**

**IMS**

Information  
Management  
School

# BUSINESS CASES WITH DATA SCIENCE

---

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

## **Chain C, H2 - Predictive Model on Cancellations**

Group X

Beatriz Chumbinho, number: R20170867

Inês Costa number: R20170775

M<sup>a</sup> Leonor Morgado, number: R20170871

Rodrigo Matias, number: R20170880

**March, 2021**

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

## INDEX:

1. INTRODUCTION.....	3
2. BUSINESS UNDERSTANDING.....	3
2.1. Background.....	3
2.2. Business Objectives.....	3
2.3. Business Success criteria.....	3
2.4. Situation assessment.....	3
2.5. Costs and Benefits.....	4
2.6. Risk & Contingency.....	4
2.7. Determine Data Mining goals.....	4
3. PREDICTIVE ANALYTICS PROCESS.....	4
3.1. Data understanding.....	5
3.2. Data preparation.....	6
3.3. Modeling.....	8
3.4. Evaluation.....	9
4. RESULTS EVALUATION.....	9
5. DEPLOYMENT AND MAINTENANCE PLANS.....	10
6. FURTHER ACTIONS.....	11
7. CONSLUSION.....	11
8. REFERENCES.....	11

## 1. INTRODUCTION

Hospitality companies are always vulnerable to a specific risk: Cancellations. A lot of factors can culminate in this event, namely, situations like changes in business meetings, vacations rescheduling, health factors or any other incident. Nowadays, Hotels are even more exposed to the cancellation risk due to the emergence of the Online Travel Agencies - OTA - which promote "Deal-seeking" behaviors. Thus, the competition is getting stronger and hotels are adopting a defense mechanism - Overbooking- that has its basis on allowing more reservations than the real capacity, believing that some customers will cancel their bookings. In fact, the supposed defensive mechanism also constitutes a threat for the company since it may compromise the reputation as well as the revenue which includes the customers reallocation costs. On the other hand, the cost of avoiding the overbooking strategy is also high since it has the consequence of having empty rooms. Motivated by this problem, Chain C hired our team in order to provide a predictive model that may be capable of predicting who are the customers who will cancel and, as well, who are the ones that will not show up.

## 2. BUSINESS UNDERSTANDING

### 2.1. Background

Hotel chain C, a chain with resort and city hotels in Portugal, was severely impacted by cancellations, representing almost 42% in H2. For this reason, Revenue Manager Director of Hotel chain C, Michael decided to limit the number of rooms sold with restrictive cancellation policies. To balance that decision, Michael implemented a more sophisticated overbooking policy that started generating costs. To mitigate those costs, Michael decided to soften the overbooking policy, which in turn proved to be inadequate. The less aggressive overbooking policy resulted in the hotel having inventory not sold, even on high demand dates.

---

### 2.2. Business Objectives

Implementation of a prediction model in order to identify in advance clients that might cancel their reservations and the ones that don't show with no justification. By identifying those two types of clients Hotel Chain C, would be able to make an offer to the ones with high likelihood of cancellation, and be free to be more aggressive with the overbooking policy when a customer with high likelihood of no show books a room.

---

### 2.3. Business Success criteria

Our team's goal in the present project is to provide to Chain C the following, shaped as a report and a presentation:

- Reduce cancellations to a rate of 20%
  - Increase the net revenue with overbooking by identifying the 'No shows'
  - Deliver it by the 15th of March.
- 

### 2.4. Situation assessment

To develop this project the "Chain C" hired four data scientists for the time of seven days. To the data scientists, a dataset from one of the hotels of the company, H2, was made available in order to perform exploration and develop a model which is capable of predicting the different commitment behaviors of the customers in a matter of booking cancellations. The platform available to perform the data analysis / machine learning task in order to accomplish the objective is Python Jupyter Notebook.

## 2.5. Costs and Benefits:

Component	Description	Benefit	Assigned Cost
Labour	Estimated cost for the human resources needed to execute project activities  Rates usually include Overheads	Vast data scientist team  Will accomplish the company goals	= Junior days * rate
Materials	Hardware, Software	High quality technology	Purchased cost
Contingencies	Risk provision	Continuous of the project in case of constraints	Only if needed (to be defined)

table 1

## 2.6. Risk & Contingency:

Risk	Preparation	Response	Probability
A large number of employees call in sick	Develop an incentive plan for taking unscheduled shifts  Create routine processes to operate a shift with fewer workers	Immediately communicate to employees to request that they come for an unscheduled shift	High
Network or system outage	All networks and systems need to be prepared with quality backups	Switch to backup and escalate to IT	High
A machine breaks down	Keep parts and components in stock for quick maintenance	Address the problem to machine suppliers  Have a maintenance team available	High

table 2

## 2.7. Determine Machine Learning goals

The main goal of this project is to predict future behavior of Hotel Chain C customers in a matter of cancellations. The machine learning methodologies will enable our team to provide Hotel Chain C a pattern recognition program, able to predict who are the customers that will not show up as well as the ones who intend to cancel and also the ones who are committed with their reservation. In order to do so, we must select the best indicators and most appropriate algorithm to forecast booking cancellations of those clients. In a nutshell:

- Predict who are the customers that will not show up
- Predict who are the customers that intend to cancel
- Predict who are the customers that will comply with their booking
- This goal will be successfully fulfilled with a model accuracy above 80%
- Having a precision above 80% regarding the cancellations
- Having a recall above 90% regarding the cancellations

## 3. PREDICTIVE ANALYTICS PROCESS

In order to develop the present cancellations predictive model, several steps were performed. Our team started by understanding the data we had in hands, taking into consideration the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology - a standard process model that describes common approaches used to conduct data mining studies. CRISP-DM methodology can be understood by looking at the image on the right.

This way, we started by understanding your business, the project objectives, the requirements and the data itself that Chain C provided us in a circular, iterative and interactive perspective. The data was prepared, in the way presented lately in the present section. Then, the data was the input for our model - using Random Forest.

This model enabled us to create a predictive model about customer's behavior in a matter of cancellations. Having done all these steps, our team evaluated the achieved results comparing them with your business needs.

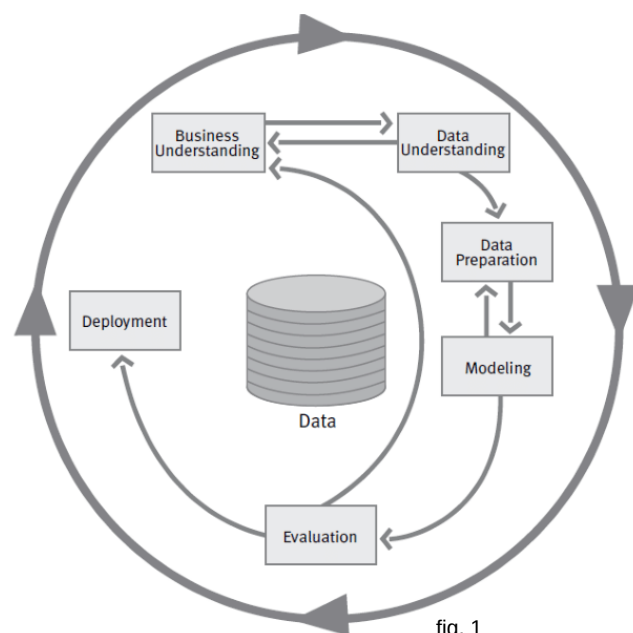


fig. 1

## 3.1. Data understanding

### 3.1.1. Variables:

The dataset provided relates to the bookings made in a specific hotel, H2, of the Hotel chain C for a period of two years and a month. It contains 31 indicators and 79330 reservations from which only 58% were enjoyed, since 42% were cancelled or the customer did not show.

Our team worked on understanding each variable taking into account their meaning in the process of reaching both business and data mining goals the best we can. All variables refer to the bookings that were made and how the customer experience affects the hotel. For example, number of days staying, requests, meals and how the customers made their reservations. The meaning of each variable - Metadata - can be found in the appendices section.

From those 31 variables our team only made use of 28. The 'country' variable was useless to our analysis since that information is provided only in the check-in moment. The variable 'IsCanceled' is redundant since it is already embedded in the 'ReservationStatus' categories. Due to corrupcy in the data, 'DepositType' variable was also discarded. During the exploratory process our team noticed that some of the variables would benefit from some engineering, in a further stage of this document that thematic will be approached.

### 3.1.2. Descriptive Statistics:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
LeadTime	79330	NaN	NaN	NaN	109.736	110.949	0	23	74	163	629
StaysInWeekendNights	79330	NaN	NaN	NaN	0.795185	0.885026	0	0	1	2	16
StaysInWeekNights	79330	NaN	NaN	NaN	2.18296	1.45642	0	1	2	3	41
Adults	79330	NaN	NaN	NaN	1.85098	0.509292	0	2	2	2	4
Children	79330	NaN	NaN	NaN	0.0913652	0.372168	0	0	0	0	3
Babies	79330	NaN	NaN	NaN	0.00494138	0.0843233	0	0	0	0	10
Meal	79330	4	BB	62305	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MarketSegment	79330	8	Online TA	38748	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DistributionChannel	79330	5	TA/TO	68945	NaN	NaN	NaN	NaN	NaN	NaN	NaN
IsRepeatedGuest	79330	NaN	NaN	NaN	0.0256145	0.157983	0	0	0	0	1
PreviousCancellations	79330	NaN	NaN	NaN	0.0797428	0.415472	0	0	0	0	21
PreviousBookingsNotCanceled	79330	NaN	NaN	NaN	0.132371	1.69341	0	0	0	0	72
ReservedRoomType	79330	8	A	62595	NaN	NaN	NaN	NaN	NaN	NaN	NaN
AssignedRoomType	79330	9	A	57007	NaN	NaN	NaN	NaN	NaN	NaN	NaN
BookingChanges	79330	NaN	NaN	NaN	0.187369	0.60862	0	0	0	0	21
Agent	79330	224	9	31955	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Company	79330	208	NULL	75641	NaN	NaN	NaN	NaN	NaN	NaN	NaN
DaysInWaitingList	79330	NaN	NaN	NaN	3.22677	20.8709	0	0	0	0	391
CustomerType	79330	4	Transient	59404	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ADR	79330	NaN	NaN	NaN	105.304	43.603	0	79.2	99.9	126	5400
RequiredCarParkingSpaces	79330	NaN	NaN	NaN	0.0243666	0.154919	0	0	0	0	3
TotalOfSpecialRequests	79330	NaN	NaN	NaN	0.546918	0.780776	0	0	0	1	5
ReservationStatus	79330	NaN	NaN	NaN	0.605824	0.51176	0	0	1	1	2
ReservationStatusDate	79330	864	2015-10-21	1416	NaN	NaN	NaN	NaN	NaN	NaN	NaN

table 3

### 3.1.3. Clean Noise:

After analyzing the dataset we noticed that the variable 'Children' was the only that contained null values. Since this variable is numeric, we imputed it with the median.

### 3.1.4. Balance Checking:

In order to keep aware about the purpose of the study, check the unbalanced learning was important to understand what was the most common classification of the target attribute - 0, 1 or 2. The most common classification is 1 - that corresponds to normal situations - followed by the 0 - which corresponds to cancelations - and the less common classification is the 2 - corresponding to 'no show'. However, the number of canceled reservations, as we already know, is just slightly lower than the normal situations which is a problem.

In order to face this problem we oversampled the train dataset with a Synthetic Minority Oversampling Technique (SMOTE), this procedure was used to create synthetic examples for the 'no show' class, those examples are plausible, and relatively close in the feature space to existing examples of the minority class. After evaluating the results and accuracy, we realized that SMOTE procedure was not a benefit for our predictive model. we ended up by do not using that process. The results are presented in the table 4.

Label	Original X_train	Oversampled X_train
0	21685	21685
1	31551	31551
2	630	9000

table 4

## 3.2. Data preparation

### 3.2.1. Splitting the data:

In order to perform a prediction it is necessary to split the dataset into training - X\_train - and testing - X\_test. We allocated 70% of the data for the train and the remaining 30% for testing. To make the pre-processing tasks simpler we created the following dataset: combine [X\_train,X\_test] which enables us to perform the tasks simultaneously for both of them.

It is important to note that for both X\_train and X\_test there is the corresponding Y\_train and Y\_test. Being the datasets with 'X' the ones which contain all the variables but the target and the ones with 'Y' containing only the target.

### 3.2.2. Outliers:

Regarding the outliers, our team started by analysing each variable. Our final decision consists in the usage of the LOF method - Local Outlier Factor - this algorithm calculates the local density of a point and compares it with its neighbors. If a point has quite lower density that its neighbor is considered an outlier. In order to comply with the best practices we removed at total 3% of the data, applying a contamination of 0.03.

Note: Only non-categorical data was removed. The anomalies present in the categorical data were treated in the feature engineering section, explained in a further stage of this document.

### 3.2.3. Normalization:

We ended up by choosing the Random Forest which performs worse with normalized data so we did not standardized our dataset.

### 3.2.4. Encoding:

The result of encoding the categorical variables with One Hot Encoding was the creation of a column for each different category in every categorical variable. These new columns are binary, having the values 0 or 1 depending on its category in that variable. We removed one category in each variable to avoid multicollinearity.



### 3.2.3. Feature Engineering:

Our goal was to produce tools to retrieve the maximum information from the dataset. To reach our objective we created new variables that are transformations of variables already existing in the dataset and we also adapted existing variables. The table at the right (table 5) explains the feature engineering results.

Variable Name	Description	Creation
<i>Months</i> (Already existing)	Indicates the respective month.	We used the 'ArrivalDateMonth' variable. We associate the respective number of each month.
<i>Meal</i> (Already existing)	Represents the type of meal booked.	We used the variable 'Meal', we decided to join the categories 'FB' and 'HB'.
<i>MarketSegment</i> (Already existing)	Represents the Market segment designation.	We used the variable 'MarketSegment' to create a new category called 'Other' where we put the categories 'Undefined', 'Aviation' and 'Complementary' that already existed.
<i>DistributionChannel</i> (Already existing)	Represents the booking distribution channel.	We used the variable 'DistributionChannel' to create a new category called 'Other' where we put the categories 'Undefined' and 'GDS' that already existed.
<i>Agent</i> (Already existing)	Represents if it was an agent or not that made the booking.	We mapped each character to 0 or 1.
<i>Company</i> (Already existing)	Represents if it was a company or not that made the booking or responsible for paying the booking.	We mapped each character to 0 or 1.
<i>Nr People</i> (New variable)	Represents the number of people per booking.	We sum the variables 'Adults', 'Children' and 'Babies'.
<i>Totalnights</i> (New variable)	Represents the number of nights per booking.	We sum the variables 'StaysInWeekNights' and 'StaysInWeekendNights'.

table 5



fig. 2

### 2.3.3. Feature selection:

The correlation matrix available in the figure below, fig.2, shows that do not exist correlations above 0.54 between metric features. This way, we did not removed any.

### 3.3.Modeling

#### 3.3.1. Feature Selection:

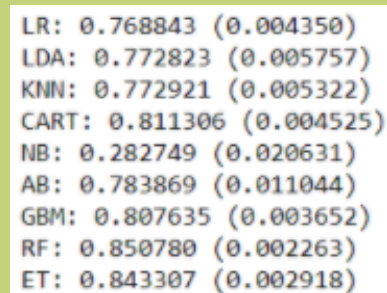
Due to the importance of this step, to achieve the best predictive model, our team decided to apply to the preprocessed dataset a variety of different methods to guide our choice. With this, we started by checking the correlations using the Spearman Correlation but we ended up not removing any variable from the model because the highest correlation that existed was 0.54%. Then, we performed the three techniques: RFE , Lasso regression and Ridge regression but we did not get really good results so we decided to do it by trial and error where we got the best results.

The final subset of features was: 'LeadTime', 'StaysInWeekendNights', 'StaysInWeekNights', 'Adults', 'Children', 'Babies', 'PreviousCancellations', 'PreviousBookingsNotCanceled', 'BookingChanges', 'DaysInWaitingList', 'ADR', 'RequiredCarParkingSpaces', 'TotalOfSpecialRequests', 'Nr People', 'Totalnights', 'Meal\_HB/FB', 'Meal\_SC', 'MarketSegment\_Direct', 'MarketSegment\_Groups', 'MarketSegment\_Offline TA/TO', 'MarketSegment\_Online TA', 'MarketSegment\_Other', 'DistributionChannel\_Direct', 'DistributionChannel\_Other', 'DistributionChannel\_TA/TO', 'IsRepeatedGuest\_1','Agent\_1', 'Company\_1', 'CustomerType\_Group', 'CustomerType\_Transient', 'CustomerType\_Transient-Party'.

#### 3.3.2. Classifier choice:

A great variety of classifiers was tested, like MLP, neural networks (NN), K-Nearest Neighbors (KNN), Support Vector Machine (SVC), random forest (RF), decision tree (Dtree), Extra Trees (ET), Naive Bayes (NB), Logistic Regression (LR), Latent Dirichlet allocation (LDA), and also some which combine more than one method, ensemble methods such as, gradient boosting (GBM), adaboost (AB), bagging. In the figure it is possible to verify the respective score of each classifier, which is the mean accuracy on the X\_test dataset.

After examining all the mentioned classifiers and their respective performance, we concluded that Random Forest (RF) was most appropriate for our data and taking into account this project objectives.



Classifier	Mean Accuracy	Standard Deviation
LR	0.768843	(0.004350)
LDA	0.772823	(0.005757)
KNN	0.772921	(0.005322)
CART	0.811306	(0.004525)
NB	0.282749	(0.020631)
AB	0.783869	(0.011044)
GBM	0.807635	(0.003652)
RF	0.850780	(0.002263)
ET	0.843307	(0.002918)

fig. 3

#### 3.3.3. Model Fine Tuning - Random Forest Parameters:

Using a grid search algorithm our team selected the best parameters to tune the elected model - Random Forest (RF). This way, it was possible to increase the accuracy and, consequently, to better predict the customer's behaviors.

The best parameters to fulfill the prediction objective are: criterion='entropy', max\_depth=200, max\_features=None, min\_samples\_split=0.005.



### 3.4. Evaluation:

After applying the Random Forest classifier to the selected features, previously cleaned, analyzed and worked on, and with the most appropriate parameters, we are able to predict customers cancelation ahead of time.

The resulting model has an overall accuracy of 82%, predicting the cancellations with a precision of 85%. For the No Show category, the model is considering the majority of it as Checked-out and some as Cancelled, this way the precision of the model to predict No Shows is zero.

This is happening because the No show situation in the Hotel Chain C is insignificant since it only happened a very small number of times, unlike the cancellations and normal situations. Also, in trying to predict the No shows better we applied some oversampling to the dataset but realized this would worsen not only the accuracy of the overall model but also the cancellations. Given this and considering that the No shows do not bring any expense to the hotel we decided to prioritize the cancellations predictions.

Also, we believe that it is better to be mistaken and consider someone as it is not going to cancel and they end up cancelling than considering that it cancelling and the person ends up not doing it. If the second append and their room is not available anymore, for example, the hotel in question would gain a very bad reputation instead of just losing a booking.

Furthermore, this final model is predicting the normal situations of check-out almost perfectly, with a precision of 81% as is shown in the fig.4.

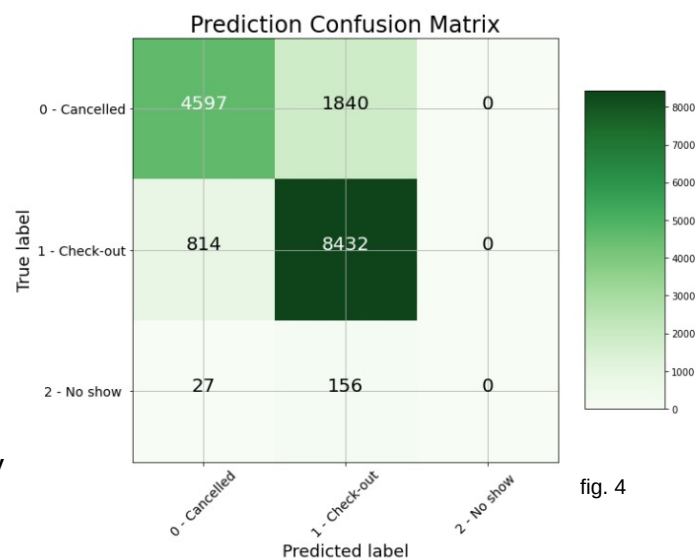


fig. 4

## 4. RESULTS EVALUATION

To meet the business objectives the cancelations predictive model program was developed with success with an high accuracy on predicting who intended to cancel, the main business objective. The data scientists team had complied with the programmed schedule - 15th March.

On the other hand, the business objective referent to increasing the revenue through overbooking was not possible to fulfil in 100% with the data we had access. This fact is justified once the proportion of 'No-Show' situations was too low and thus the model is not able to create a pattern to recognize those cases. This way, the Hotel Chain C can use the knowledge about cancelations to perform overbooking.

By carefully analyzing the most important variables to predict booking cancellations, such as 'TotalOfSpecialRequests', 'LeadTime' or 'PreviousCancellations' our team realized that:

- People who have 1 or more special requests are very likely to not cancel their bookings;
- The further the advance in booking the stay, bigger is the probability of someone cancelling it;
- If a person has canceled once, is more likely to cancel again;
- 'Group' bookings tend to be more canceled than others.

Business Success Criteria	Machine Learning Results
Reduce cancellations to a rate of 20%	Predictive model was developed with na high accuracy, namely, on the calcelation's prediction
Increase the net revenue with overbooking by identifying the 'No shows'	Lack of data diversity to properly analyse the situation

table 6

## 5. DEPLOYMENT AND MAINTENANCE PLANS

We found some problems in the data that can be solved in future data collection and cleaning efforts, namely a reservation id. Additionally, we found that there were some rows with only childrens which can be conflicting when training the model. We also could not be able to understand the meaning of the 'undefined' value in some variables. Our team felt that few useful and adequate variables were available to use for modelling since none of them was correlated with the target variable.

We defend that this project should be deployed using a flask application where every-time a reservation is made in the Hotel Chain C system an automatic prediction of the customer behavior would take shape. This way, our team guarantee that it is important to consider this system as tool embedded in the quotidian activity of the hotel. For this reason, it is imperative to make the employees familiar with the system.

We believe in the value of an iterative and interactive deployment. This way, we suggest you to keep with us in order to frequently evaluate the system and business needs in this field.

As an addition, we would like to provide you, in the near future, the necessary functionalities to predict 'No show' situations, having the right resources for that.

In order to retrieve the best advantage of the present analysis, if a new reservation is predicted to be cancelled in the near future automatically Hotel Chain C should proceed in order to retain this customer. A way to captivate those customers is, per example, offering them free parking spots or free meals.

There needs to be a continuous experimentation/exploration of new ideas (e.g. feature engineering, model architecture and hyperparameters).

To finalize, we suggest Hotel Chain C to apply Continuous Integration and Continuous Delivery system. This way, you will be able to keep always on an innovation journey.

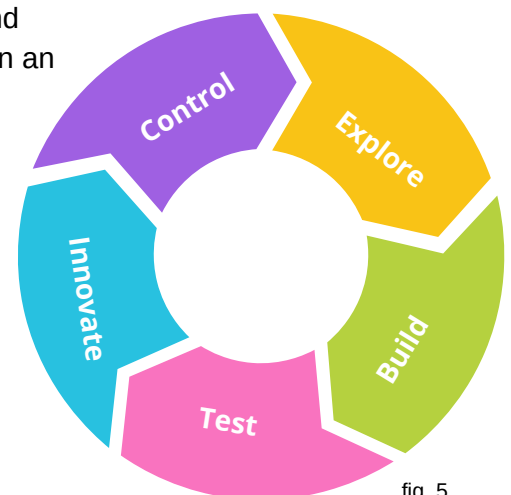


fig. 5

## 6. FURTHER ACTIONS

Further Actions	Pros	Cons
Analyse the remain hotels from Chain C	Verify the previous conclusions / avoid bias	Might provide redundant information
Collect more data about people who has 'No Show' bahavior	Better predict those cases	Might do not add value in terms of information gain
Study the suggested marketing initiatives	Estimate costs and benefits	Might spend resources bad allocated

table 7

## 7. CONCLUSION

After this study, Michael will have the right tools to achieve his main goal: reduce cancellations to a rate of 20%. Despite not being able to predict no show customers due to lack of those situations on the hotel given has an example, through the model developed is possible to detect efficiently the majority of bookings cancellations ahead of time as well as the normal situations, considered check-out as states the objective of this project. The Hotel Chain C will be able to implement better pricing and overbooking policies and not least, identify bookings with high likelihood of canceling.

## 8. REFERENCES

- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) (2000). "Crips-DM 1.0". Step-by-step data mining guide.
- Lucas Bação, Fernando - Business Cases with Data Science- March 13, 2021.
- N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer(2011). SMOTE: Synthetic Minority Over-sampling Technique