

**Business Case 3: Predict Hotel Booking Cancellation**

Palm & Company

**P**edro Santos (M20190420)

**A**na Cláudia Alferes (M20190932)

**L**ennart Dangers (M20190251)

**M**ichael Machatschek (M20190054)

April 2020

**BUSINESS CASES WITH DATA SCIENCE**

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

INDEX

[1. INTRODUCTION 3](#_Toc37918484)

[2. BUSINESS UNDERSTANDING 3](#_Toc37918485)

[2.1. BACKGROUND AND CURRENT SITUATION 3](#_Toc37918486)

[2.2. BUSINESS OBJECTIVES & SUCCESS CRITERIA 3](#_Toc37918487)

[2.3. SITUATION ASSESMENT 4](#_Toc37918488)

[2.4. DETERMINE DATA MINING GOALS 4](#_Toc37918489)

[2.5. PROJECT PLAN 4](#_Toc37918490)

[3. PREDICTIVE ANALYTICS PROCESS 5](#_Toc37918491)

[3.1. DATA UNDERSTANDING 5](#_Toc37918492)

[3.1.1. Data Collection and Description Report 5](#_Toc37918493)

[3.1.2. Data Exploration 5](#_Toc37918494)

[3.2. DATA PREPARATION 7](#_Toc37918495)

[3.3. MODELLING 8](#_Toc37918496)

[4. EVALUATION CRITERIA 13](#_Toc37918497)

[5. DEPLOYMENT AND MAINTENANCE PLANS 13](#_Toc37918498)

[6. FUTURE IMPROVEMENTS 13](#_Toc37918499)

[7. CONCLUSIONS 14](#_Toc37918500)

[8. REFERENCES 14](#_Toc37918501)

# 1. INTRODUCTION

Nowadays, tourism and especially hospitality is a demand-driven and highly competitive market. Due to this competition, Hotels are forced to provide free bookings cancellation and are confronted with a significant increase in cancellations. Depending on the hotel, a common benchmark is a cancellation rate of 30 – 45 %. To overcome this issue, hotels often do so called overbooking. This involves risk: mainly reallocation costs and a damage in reputation. The following report illustrates the creation of a model to predict the net demand of hotel H2 following the data mining process CRISP-DM.

# 2. BUSINESS UNDERSTANDING

## 2.1. BACKGROUND AND CURRENT SITUATION

Even though the hospitality sector has different distribution channels, nowadays online travel agencies (OTA) are the most common distribution channels. The examined data of Hotel H2 illustrate that around 50 % of the bookings are made by OTA’s. Operating with OTA’s comes with merits and demerits. On the one hand, it is more likely to improve market exposure and the corresponding revenue. Having said that, OTA’s charge commissions (between 15 to 30 %) and force a hotel to provide price parity among all distribution channels. Moreover, to stay competitive, the hotels are constrained to offer the customers cancellation without any fees.

In fact, dealing with cancellation illustrates a serious business issue within hospitality. Hotel H2 has a cancellation rate of 41.7 %, which represents more than 40 % of their room revenue. One way to balance this problem is to implement an overbooking strategy or adjust the cancellation policies in a more restrictive way. Both contain pros and cons. Firstly, overbooking compensates cancellations and reduces the number of empty rooms but if the actual check-ins are higher than the number of rooms, overbooking can cause reallocation costs, social reputation damage as well as the loss of immediate and future revenue. More restrictive cancellation policies can cure these issues. However, it can cause a decrease in revenue because non-refundable rooms are usually discounted. Besides, due to the restrictions in flexibility, less customer will book a room.

An accurate model to predict the net demand of Hotel H2 can solve the above-described issues. The net demand is computed by the bookings subtracted by the cancellations.

## 2.2. BUSINESS OBJECTIVES & SUCCESS CRITERIA

Business Objectives

The main business objectives are divided into the forecast of the net demand and in the generation of valuable insights about customer patterns. A good working prediction of the actual demand can reduce the number of unused rooms due to spontaneous cancellations and avoid negative ratings and costs if a customer must choose another hotel. In other words, the cost of social reputation damage, as well as the reallocation cost, can be lessened. Also, H2 can maintain customer relationship management and do not lose long term customers. Due to more precise knowledge of the real demand, an appropriate pricing and overbooking strategy can be implemented. Eventually, knowing the correct occupancy rate can help to handle the right number of employees and therefore be more efficient.

As a side effect, a predictive analysis should deliver customer patterns, which are valuable for the whole business. Here it is essential to deal with questions, such as “who are the customers?”, “who is likely to cancel bookings?” and “why is the customer likely to cancel?”. These insights could be processed by the Marketing department to develop strategies to make offers with the goal to avoid cancellations.

Business Success Criteria

The future key success criteria are to reduce the cancellation rate to only 20 %. Besides, a reduction of the empty inventory rate depending on the previous rate is a further goal. Regarding the marketing campaigns to avoid cancellations, a potential objective can be to create proper offers with a success rate of 20 %. Finally, it is vital that the created model can be integrated within the current business processes.

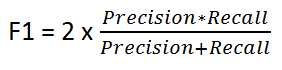
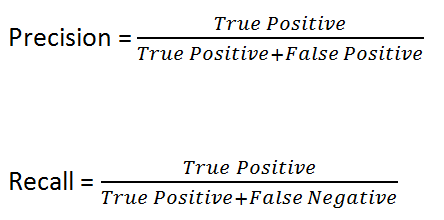
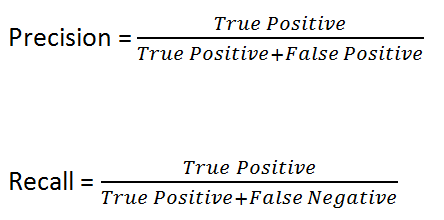
## 2.3. SITUATION ASSESMENT

In general, to obtain an accurate prediction is usually challenging, because it is just based on data which may have some hidden quality issues. Identically a prediction model still has risks left and should not have an accuracy of 100%, due to risks of overfitting the data. Nevertheless, the goal of the model is to reduce the risk of overbooking and maximize the revenue.

## 2.4. DETERMINE DATA MINING GOALS

Since the classes of the given dataset are somewhat balanced our goal metrics is accuracy with a rate higher than 80 %. (Provest, Fawcett, 2013). However, within our evaluation we will also consider further evaluation metrices, which are based on a so-called confusion matrix. The confusion matrix compares the prediction with its corresponding actual output and displays the sum of how many predictions were correctly classified.

|  |  |  |
| --- | --- | --- |
| **Prediction Actual Output** | **Is cancelled** | **Is not cancelled** |
| **Is cancelled** | True positives | False positives |
| **Is not cancelled** | False negatives | True negatives |

As mentioned above, the main evaluation measure is accuracy. Accuracy computes the ratio of all correct predictions (true positives and true negatives) and all samples. In addition, we will take the following metrices into account:

Source: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>

To obtain the desired goal we implement more than one algorithm to compare each result and choose the best working model based on the metrics. Furthermore, these comparisons should be based on different feature sets.As a side product of the predictive analysis, the receipt of valuable business insights through data exploratory analysis is also an essential data mining goal.

## 2.5. PROJECT PLAN

As mentioned in the introduction, this project follows the CRISP-DM process. The most important that stands out within this process is, that it is not hierarchical, and one can move one step back. At the beginning, the actual business must be understood by the project team (see chapter 2). After discussing the business background and effects, we started to prepare the available data. We began by an exploratory data analysis to get insights for the data preparation and data engineering phase. After selecting the most relevant features, we started the modelling phase. During all these phases we had several interactions between the phases to improve our decisions in previous steps.

# 3. PREDICTIVE ANALYTICS PROCESS

## 3.1. DATA UNDERSTANDING

The goal of data understanding is to explore and explain the data based on statistical methods and the obtained knowledge in business understanding. Following the CRISP-DM process, this part contains a data collection and description report as well as a data exploration.

### 3.1.1. Data Collection and Description Report

Data Collection Report

The examined dataset contains data about hotel clients of the years 2015, 2016 and 2017, of a hotel based in Lisbon (Portugal), which is owned by the company C. The variables display information about the number of visitors, quantity of nights in the hotel, the arrival and reservation dates, if it was booked by an agency or company, and the status from the reservation. In this hotel the cancelation rate is 42 % and the goal is to reduce cancelations to a rate of 20 %. In order to do that, we performed data exploration to get insights about the customers and their probability of cancelation.

Data Description Report

The raw dataset has 31 columns, which represent different features. The number of rows is 79.330 and each row represents one reservation made by a customer. Some of the features are categorical variables (14) containing strings or are dummy variables. In general, more information, such as the city of the customer, the type of the customer, the assigned and reserved room, the type of meal picked by the customer, the agency or company that made the reservation and how they have done it.

### 3.1.2. Data Exploration

Within the data exploration, most of the analysis requires aggregated data on the reservation level. Furthermore, a comparison between cancelled and not cancelled reservations is one of the approaches to obtain more insights. The exploratory analysis is separated into the subject areas of time, agency responsible for the reservation, room type of the reservation, family size, booking changes, country of the customers, customer type, deposit type, distribution channel and market segment, repeated customers, lead time, type of meal picked in the reservation, previous cancellations, number of nights in the hotel and special requests. We will be focusing on the ones deemed as more important.

Time exploration

The time analysis is divided into the years and month regarding both cancelled and not cancelled. Since we do not have the full year for 2015 and 2017, we cannot compare between years itself, but it is possible to compare the ratio of cancelations and see that in the year of 2016 the ratio its smaller. In the visualization of the months, it is possible to see useful patterns for further analysis. Since we can observe which are the months that have more reservations, we can also look into the one that have more cancellations and take precaution, as in those months it is useful to have a different approach to the customers.

Agency

This analysis examines the most common agency to make the reservation and their ratio of cancelations. In both, cancelled and not cancelled reservations, we can see that the agency “9” and “1” are the ones that contact the more the hotel but with a big difference in the cancellation’s ratio 41,5 % and 73,2 % respectively. With this information we can target agencies with different approaches since we know which one are likely to cancel their reservations.

Family Size

We decided to add up the number of adults, children and babies corresponding to a booking into one new feature. Once we have that feature computed we could see the most frequent stays in the hotel corresponds to groups of two people. Even though it is not very visible a difference between cancelled and not cancelled reservations, we could see a decreased of cancellations ratio in families that have children.

Booking Changes

In this visualization we decided to just check the customers that have made at least one book change, so that we could see if the customers who made changes are more likely to cancel their booking or not. The data tells us that the probability of cancellation of a customer who made at least one booking change corresponds to 16.4 % *vs* 83.6 % for those who did not not cancel.

Customer Type

In this analysis it is possible to see the customer types that made more bookings in an overall view but we can also check the ones that have a higher rate of cancellations, with the “Contract” and “Transient” customers being the ones that are more likely to cancel.

Deposit Type

Looking into this feature we could see that 99.8 % of the deposit type corresponding to “Non Refund” customers end up cancelling the booking, which translates to around 12,000 customers who cancelled the booking although it is non-refundable. In contrast, just 24 customers booked in the “Non Refund” deposit option and did not cancel a trip. This insight is not consistent with business logic and must be considered if it can be discarded for modelling.

Distribution Channel and Market Segment

In these two visualizations it is possible to see the most frequent distribution channels and market segments. For the distribution channel we did not find a useful pattern to get insights. In a further analysis, we decided to look in the market segment instead and check the rate of cancellation higher than 25 % in the different types of market segments.

Lead Time

This feature is particularly important in the analysis because we can see a clear increase in the rate of booking cancellations with the increasing of the lead time. It is also possible to observe when the cancelled booking start outgrowing the not cancelled ones, which is around the 160-179 days in lead time.

Previous Cancelations

In this feature it was possible to check that 84.5 % of customer who have made previous cancellation have cancelled the booking in this hotel.

Special Requests

Here, one can observe that the customers that usually do not make any kind of requests are the ones who cancel the most.

Nationality

Even though this information is important in a business point of view, we found out that, for most of the cases, it is only possible to confirm the nationality of a customer on the check-in moment. Therefore, cancelled bookings do not have check-in and, in consequence, this feature was dropped.

In order to develop the analysis above, visualizations were produced, and they are available for consultation in the Jupyter Notebook.

## 3.2. DATA PREPARATION

Data Cleaning

In the exploratory analysis features were discovered which contain conspicuous values. At first, we dropped bookings that have no numbers in adults, children, and babies (100 bookings). Furthermore, bookings with no count on nights were dropped equally (260 bookings). Additionally, a minimum (20 Euro per night) and a maximum (520 Euro per night) threshold are defined and consequently discard all bookings that are not in this range (1285 bookings). Bookings that required more than one car and have a family size above 10 are dropped as well. Since these records count only 10 bookings, this reduction will not affect the final modelling result. Eventually, within the step of data cleaning, 1652 bookings were dropped. This means a quota of 2.1 %.

Feature engineering

Instead of creating new feature such as ratios, our focus is on binning and transforming some of the features.

To obtain a more balanced distribution, we binned some of the features. “Binning methods smooth a sorted data value by consulting its “neighbourhood,” that is, the values around it. The sorted values are distributed into a number of “buckets,” or bins.” (Han, Kamber, Pei,2012). Furthermore, binning is a useful technique to reduce cardinality when categorical features have too many categories. For the bin sizes of the average daily rates, we decided to use the ranges of booking.com: 0-50, 51-100, 101-150, 151-200, and 200+. All other binned features are following neither the equal-width nor equal-frequency approach. Instead, we created bins, which improve the distribution and follow a business perspective supported by the cancellation rate. Further steps of transformation are to apply logarithmic and square root to some features to make right skewed distributions look more like a normal distribution. The logarithmic transformation was applied, for instance, to Lead Time.

Before the modelling starts, all categorical features are converted into numerical features, which is called encoding. It is vital because most of the algorithms cannot handle categorical data. Encoding splits a categorical feature into the number of given category names. Whether one category occurs, it counts 1 whereas all other columns contain zeros. (Provost, Fawcett, 2013). In this project two different techniques of encoding are used: OneHotEncoder for nominal and OrdinalEncoder for ordinal data. Both are supported within the scikit learn library.

Feature selection

The feature selection is based on a business perspective and the ability to split each feature into the two classed whether if it is cancelled or not cancelled. The distribution of each feature, which can be found in the exploratory analysis, is the key factor to measure this ability.

## 3.3. MODELLING

Our modelling process consist of three iterations. In an iteration we test the same algorithms with different parameters and different feature sets. In the first iteration we started with two initial feature sets and tested them with different algorithms and different parameter combinations. In the next iteration we focused on the best performing models of iteration one and tried new feature combinations and transformations. In the third and final iteration, we fine-tuned the best performing model of iteration two.

These are the base algorithms which we tested in the first iteration: Logistic Regression, Random Forest, Gradient Boosting, Multi-Layer Perceptron and XGBoost.

The parameters were chosen in a way that we have control over the speed of the algorithm and the possibility to balance over- and underfitting. The full list of features used in the iterations can be found in the Jupyter notebook.

In the next part of this section we give a short theoretical introduction about the algorithms we applied. The following parts will explain our approach in each of the three iterations. Eventually, we present the final model and explain the importance of the features.

***Logistic Regression***

Logistic Regression is a linear model that is used solely for classification tasks. It fits a sigmoid function to the training data by applying maximum likelihood estimation. This method is quite efficient; however, it is not possible to solve nonlinear problems with it. The output of logistic regression is a probability that y is equal to 1, given an input x.

***Random Forest***

A Random Forest is part of the ensemble family with a classification tree as the base model. A classification tree is built by splitting the dataset in a way that the purity of the new datasets is maximized, where the purity can be measured by the Gini coefficient or the entropy. This process is repeated for each new dataset until a stopping criterion is reached.

A random forest consists of several classification trees, which are built with bootstrapped datasets. For the final prediction we run the input through all trees and take the option that received the most votes.

***Neural Network – Multi-Layer Perceptron***

A Multi-Layer Perceptron is a neural network that consists of an input layer, one or more hidden layers and one output layer. In each layer we have several nodes - in the input layer we have a node for each feature, in the hidden layer we have as many nodes as we define and in the output layer, we have a node for each class. The algorithm consists of the feed forward part, where the inputs are passed through the network to get a prediction and the back propagation part, where the error of the prediction is calculated and some network characteristics (e.g. the weights) are updated according to the learning rate.

***Gradient Tree Boosting***

Gradient tree boosting is an ensemble method with a classification tree as a base model. Its architecture can be described as a chain of trees, where each tree is built based on the error of the previous tree and each tree has a weighted impact to the final prediction.

***XGBoost***

XGBoost is also an ensemble method with a classification tree as a base model. Its architecture is similar to Gradient tree boosting. The difference is in the tree itself. In XGBoost a special tree is used that is built differently than a classical classification tree. Additionally, this algorithm has different methods for tree pruning.

First Iteration

In the first iteration we ran two feature sets through the algorithms and made a first evaluation in the end. Both feature sets contain features that we deem to be important from a business perspective and that are able to separate the two classes based on the distribution which we analysed in the exploratory analysis. The difference between the two features sets is, that in the first set most of the continuous variables were binned and encoded with one-hot-encoding, whereas the second set contains mostly the original continuous variables. We did this split to test whether the implicit order of continuous variables bears relevant information for the prediction or not.

As you can see in the results csv files and in the Jupyter notebook, the Random Forest is with 84.12 % the best performing algorithm and XGBoost with 84.03 % the second best. Also, the precision and recall are in both algorithms satisfactory. Looking at the parameter combinations, we can see that in Random Forest and XGBoost not all initial parameter values made it into the top five models. For the next iteration we will therefore focus only on those values, that were present in any of the top five combinations. In five of the six algorithms the not-binned feature set performed better. That is why we use this feature set for the next iteration.

Second Iteration

The starting point in the second iteration were the two best performing algorithms (Random Forest and XGBoost), their best performing parameter values and the not-binned feature set of iteration one. In this iteration we then focused on transforming features and omitting less important features. Our test environment were 12 variants of Random Forest and 8 variants of XGBoost.

Firstly, we tried logarithmic and square root transformation for those features, which tend to have a right skewed distribution. After several combinations we came to the result that the feature set with a logarithmic transformation of *LeadTime* was the best performing option.

Secondly, we looked into the feature\_importance of the Random Forest model and excluded the five least important features from the feature set. Eventually, it turned out that the set with all features performed best.

Based on the results, which also can be seen in the Jupyter notebook, we choose XGBoost as our final algorithm that should be fine-tuned in the last iteration. Even though, the Random Forest had a slightly better accuracy, we decided us for XGBoost as this algorithm allows us more flexibility for hyperparameter tuning.

Third Iteration and Model Optimization

The third and last iteration consists in the optimization of the XGBoost algorithm while using the best feature set of the second iteration. Therefore, we used the feature set containing all features, including some binnings and transformations. Finally, we used Optuna, a hyperparameter optimization framework, which returned us the best working version of the model, where we aimed to get the highest accuracy. More detail about Optuna can be found on the Jupyter Notebook as well as the final results of the model. Nevertheless, it is important to note that the accuracy metric is 85,1% and the remaining metrics are also in accordance with our data mining goals.

Model Explainability and details

It is important to have a good model with good results, but it is also important to understand where the results come from and if they have business value. In this chapter, we aim to demonstrate how our model satisfies both conditions.

Firstly, we investigated the feature importance of the model. Here, a feature with a high importance shows how significant it is for the model to reproduce the current results. PreviousCancellations tops this list, with a feature importance of 33,2 %, followed by RequiredCarParkingSpaces (16,6 %) and transientAndContract (15,0 %). Note that, again, these values just represent how important the feature are to build the model and not how they affect the predictability of each customer.

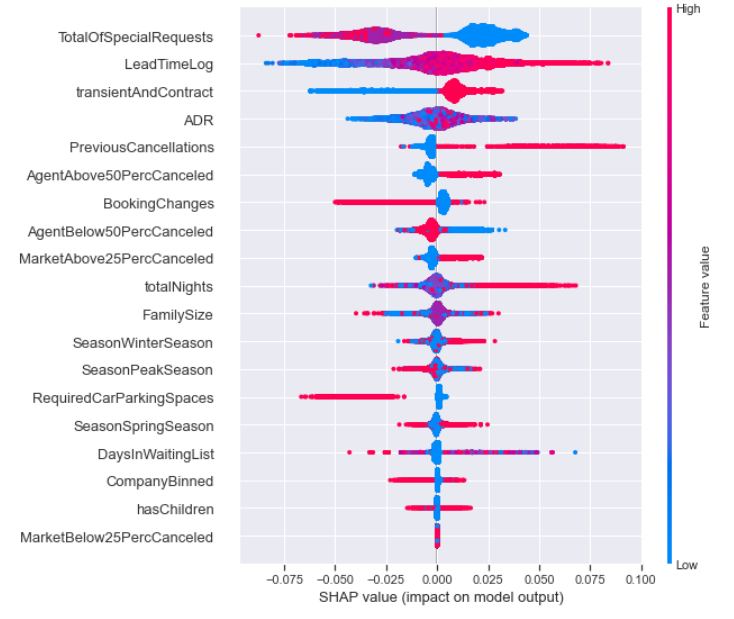
Secondly, we resorted to three visualizations from the SHAP Python package that, through the use of Shapley values explain how each feature affects the model. We will be focusing on the second visual, as It is the one that brings more value to the business. A short description of the other visuals as well as the second one and more information about the Shapley values are available on the Jupyter Notebook.

The plot below, which can also be found on Jupyter Notebook, shows with more detail how each variable impacts the model and how. The colour scale represents the feature value, where red are high values and blue low values, the X-axis is how much it impacted the model and the height of each bar is the distribution of the feature.

The first feature on the list is TotalOfSpecialRequests, a discrete variable. One can see that low values for Special Requests will increase the output of the model and, therefore, increase the likelihood of a cancellation, whereas a high number of requests as the opposite effect. This effect is expected and is in accordance with historical data, as it is described in the Data Understanding chapter.

LeadTimeLog shows that customers with longer leads times have a higher likelihood of cancelling, whereas the opposite is also true. Again, this goes in accordance with business understanding, as customers who book in advance always face more uncertainty than those who book later. transientAndContract also shows that customers who booked through this method have a higher chance of cancellation. This is also seen on the historical data and the Data Understanding chapter. PreviousCancellations has an expected behaviour, where people who cancelled before have a higher probability of cancelling again. This also goes in accordance with the business understanding. RequiredCarParkingSpaces shows that requiring a parking slot decreases the probability of cancellation.

The rest of the features and how they drive the model output can be observed on the visualization below, as we described mostly those features that paint a clearer picture on the results and what drives cancellations.



By having a better control of these factors, the Hotel will be able to decrease their cancellation rate or include a better way to monetize customers who have a higher likelihood to cancel (ie, apply more efficiently non-refundable fees).

# 4. EVALUATION CRITERIA

In terms of data mining objectives, one can say that the overall goal is fulfilled. The best model reaches an accuracy of 85.11 % and a precision of 85.94 %. Even recall and F1-Score are relatively high (76.51 % and 80.95 %, respectively). Due to the high number of different combinations of feature sets, algorithms, and their related hyperparameters, the chosen model proves to be suitable. Besides, the exploratory analysis as well as the feature importance within the modelling part provides insight, which can be used for further purposes such as marketing campaigns.

Concerning the business objectives, it is not yet possible to obtain a final evaluation. Since the model must be fully implemented in the business, it needs a certain amount of time to evaluate if the business objective is reached.

# 5. DEPLOYMENT AND MAINTENANCE PLANS

Regarding a potential deployment of the current model, it is essential to connect it to the current enterprise resource planning system to avoid cumbersome manual workload. One of the main goals here is to integrate a classification in the business process without changing the workflow. Whenever a new booking is recognized, it will be classified through the new model either as a potential cancellation or an actual booking. Additionally, whenever a change on this booking is made after the booking came in, the booking will be classified again. Based on this decision the occupation rate will be updated and helps the revenue manager to gain knowledge about the actual net demand. Furthermore, the classification of booking should be made available to the CRM system, so that the marketing department is able to proactively take countermeasures for potential cancellations. To maintain the model, it is recommended to do another training based on the new data. For instance, it is conceivable to retrain after each season. Furthermore, a frequent validation cycle in the first deployment phase is highly advisable. Since this model is not proofed by new data, this step is necessary to avoid possible inaccuracy even though the model has high accuracy, precision, and recall.

# 6. FUTURE IMPROVEMENTS

In general, modelling has usually some constraints and our model selection could be more advanced in the future. Because of time constraints within this project, not all possible combinations could be tested. It is not likely that this approach will change the metrics fundamentally, however, a slight improvement in accuracy is still thinkable. Secondly, additional data, such as data from the entire hotel chain, weather data or even competitor data, such as their prices and available room, may have a positive impact on the model. Lastly, the data quality of the given dataset is improvable, especially the feature of the deposit type should be reconsidered and readjusted.

# 7. CONCLUSIONS

To summarize, we followed a CRISP-DM approach to calculate the net demand, requested by the Hotel. After a thorough data analysis, the goal was to build an accurate model which could satisfy the hotel requests, by predicting cancellations while reaching our 80 % accuracy target. Several models and iterations were considered and tested against each other.

The final model, XGBoost, consists of an ensemble method with a special type classification tree. This model satisfies both our goals, with an accuracy and precision of over 85 %. We have also dived into the explainability of the model, with the inclusion of visuals which show how each feature impacts the model. This allows the hotel to take new insights and create new measures to avoid cancellations or to compensate with over-booking.

Finally, for future reference, it is necessary to deploy the model with new data, to further increase the confidence in the model. Nevertheless, if more time were available, other models, new features and feature transformations would be considered to improve the model.

# 8. REFERENCES

Provost, F., & Fawcett, T. (2013). Data Science - What You Need to Know About Data Mining and Data-Analytic Thinking. *Journal of Chemical Information and Modelling*. <https://doi.org/10.1017/CBO9781107415324.004>

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in Brief*. https://doi.org/10.1016/j.dib.2018.11.126