BUSINESS CASES WITH DATA SCIENCE

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

Wonderful Wines Of the World

Group_N

Bruno FIlipe Prazeres Ines Soares, number : 20200658

Xavier Golaio Gonçalves, number: 20201090

Li-Lou Dang-Thai, number: 20200743

March, 2021

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

1. Introduction

1. General Context

In the hotel industry, as in many other travel-related industries, demand is managed through advanced bookings. Bookings (also known as reservations) are a forward contract between the hotel and the customer that gives the customer the right to use the service in the future at a settled price, but often with an option to cancel. This cancellation option puts the risk on hotels who have to honor the bookings that they have on-the-books, but, at the same time, have to support the opportunity costs of having vacant rooms, when someone cancels, and there is no time to try to sell the room or sell it at a discounted price. In Europe, the cancellation rate by reservation value, from 2014 to 2018, rose from 33% to 40%. Cancellations occur for understandable reasons such as business meeting changes, vacations rescheduling, illness, or adverse weather conditions. However, cancellations also occur for not so understandable reasons, such as finding a better deal. “Deal-seeking” customers, or customers that tend to make multiple bookings for the same trip or make one booking, but continue to search for better deals (e.g., looking for hotels with better social reputation, better price, or better location). The number of “deal-seeking” customers has grown immensely with the appearance of Online Travel Agencies (OTAs) in 1996.

2. Methodology

The analysis followed a simplified version of the CRISP-DM framework recommended by Kellcher, Mac Namee, & D’arcy (2015, p.53).[1] as well as the information available at CRISP-DM Overview at IBM Knowledge Center.

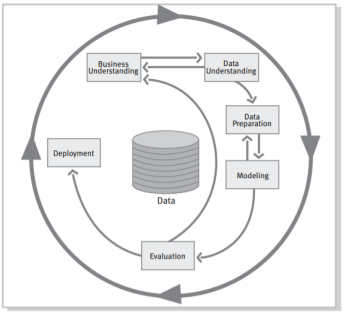

Figure 1 – CRISP-DM2

This framework of analysis is a proven way to guide data mining efforts and tools. Including descriptions of the phases lead to the guidance of a project, it explains each one and the relationships between tasks. Representing the process model, it presents an overview of the life cycle of data mining. In this context, the life cycle is composed of six phases which are the most relevant, as well as their dependencies, represented with arrows. The model presents itself as a flexible representation that can be easily customized. CRISP-DM allows analysts to create models that fit our particular needs (IBM Knowledge Center). Since this is an iterative process where advances and insights acquired along the steps can influence the prior decisions, the steps are not entirely followed in a strictly linear or straightforward logic.

According to the framework, data understanding and preparation phases are more significant than modeling, evaluation and deployment phases nevertheless these latter are important to consider for future data mining goals, in the context of eventual questions raised. Moreover, certain phases in CRISP-DM are more closely linked together than others. For example, Data Preparation and Modeling phases are closely linked, and analytics projects often spend some time iterating between these two phases.[2]

2. Business Understanding

1. Background

Hotel chain C, a chain with resort and city hotels in Portugal, isn't any different than other independent and non-independent hotel chains. Hotel chain C was severely impacted by cancellations, representing almost 28% in H1 and almost 42% in H2, as shown in the table below. For this reason, Michael, Revenue Manager Director of hotel chain C, decided to limit the number of rooms sold with restrictive cancellation policies. To balance that decision, Michael implemented a more aggressive overbooking policy. However, the latter started to generate costs. To counterbalance those costs, Michael softened the overbooking policy, which in turn also revealed to be not good. The less aggressive overbooking policy resulted in the hotel having inventory not sold, even on high demand dates.

2. Business Objectives

Concerned about the increasingly negative impact caused by cancellations, Michael hired a consultant to evaluate the possibility of developing predictive models to predict the net demand for their hotels, specifically in a city hotel (H2). The hotel provided the consultant a dataset with the bookings made in that hotel, which were due to arrive between July 1, 2015, and August 31, 2017.

3. Business Success Criteria

To reduce the uncertainty about demand, Michael wants to implement prediction models to allow the chain's hotels to forecast net demand based on reservations on-the-books. With these models' estimations, Michael expects to implement better pricing and overbooking policies and identify bookings with high likelihood of canceling. Identifying those bookings could allow the hotels to try to contact those bookings' customers and make offers to try to prevent cancellation (e.g., dinner, car parking, spa treatments, discounts, or other perks). Michael's goal is to reduce cancellations to a rate of 20%.

4. Situation Assessment

The team was given an excel file containing information about the city hotel H2 and its bookings, with some detailed information and, most relevant, if it was cancelled or not. Our main goal is to forecast the real demand taking into consideration the cancellation rate. We will have the possibility of speaking directly with the Revenue Manager Director of the hotel Chain, allowing us the continual improvement of our work process with possible feedback from the administrative board. Customer types may vary from contract to group, transient or transient-party, while deposits can be refundable, non refundable or no deposit at all. Meal categories may be undefined (no meal package, only room), bed & breakfast, half-board and full board. Moreover, there are three possible status for a booking reservation status, and there are: canceled, check-out (when the client has checked in but already left), and no-show, when the client booked but did not show up. In terms of costs, our model will require man-hours to build but, from the database and will be able to run on economical hardware while generating considerable returns when the model is applied, while solving for the uncertainty of cancelations and overbooking problem.

5. Determine Data Mining Goals

Our team is expected to create a model that will be able to classify hotel bookings, using the hotel dataset provided and the cancelation binary feature as the target variable. Since the main objective is to find the best predictive model according to the highest test accuracy, several models shall be applied, such as random forest classifier, decision tree, neural network, stacking classifier and logistic regression. Despite the main goal of the model, Data Mining goals will be established in the context of data exploration by identifying the variables that are more relevant for the prediction analysis, with the help of data visualization throughout the research.

3. Data Mining Process

1. Data Understanding

1. Initial Data Collection

The dataset was provided by Michael, the Revenue Manager Director of Hotel Chain C, it is in a single csv file.

2. Data Description Report

Our dataset is composed of 79930 rows (bookings) and 31 features. All the booking orders were made in the period from the 1st of July 2015 to the 29th of August, 2017.
After analyzing the data through Pandas Profiling and the DTale library, we found that there were just 24 missing values in “Country”, 4 in “Children” and 25902 duplicated rows. As there is not an identification number to separate every booking, it is a possibility that similar bookings were made on the same day, and thus cannot be deleted from our data.

3. Data Exploration Report

The data is composed of numerical, and categorical features, which we divided into two different groups. The target of our model is the feature “IsCancelled”. In the next figure, we can see the total number of clients who cancelled against the ones who didn’t.

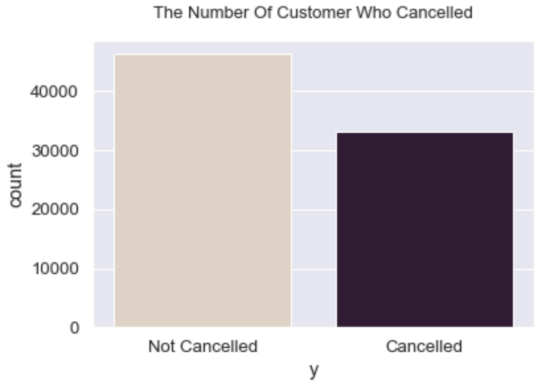

Figure 2 – Countplot of our target ‘Is Cancelled’

In order to better understand our dataset, we used the Phik Matrix to see the correlation of our features to our target. The first figure below shows that the numerical features have a low correlation with “y”. Compared to the other variables, only “LeadTime” and “TotalOfSpecialRequest” have a higher correlation.

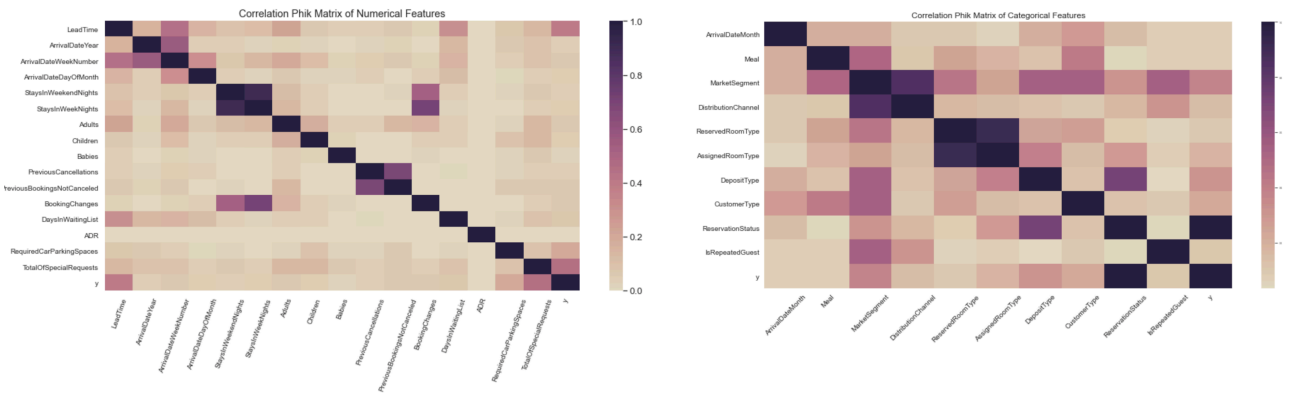

Figure 3 – Phik Matrix of the Numerical and Categorical Variables with Y

In the second Phik Matrix above, we can see that categorical features have a higher correlation with the target than our metric features.. As “ReservationStatus” is highly correlated with “IsCancelled” and is giving the same amount of information, it risks to alter negatively our model, therefore it was removed from our data.

2. Data Preparation

1. Data Cleaning

Since the feature “Deposit Type” was said to be extracted wrongfully and its quality to be compromised, it was removed. As a result of data exploration, we decided to drop multiple variables such as : “ReservationStatus” due to its high correlation with our target ‘IsCancelled” and “Company” which had 95% of NULL values and could still be related to the variable “Agent”.

Regarding missing values, only 2 features had them : “Country” and “Children”, they had respectively, 24 and 4 missing values. For “Country”, the 24 NaN were mostly rows with target = 1 meaning they had cancelled their booking. Since the country with most cancelled booking is Portugal which is also the mode of the feature, we decided to impute the missing values with “PRT”. The 4 Nans from “Children” were also filled with the mode which is 0.

The feature “Agent” has 8131 NULL values, we interpret it as a customer that did not book through an Agent, and replace NULL with a 0.

As “Agent”, “Company” had 75641 NULL values, which we put as 0, meaning they did not book with a company.

Concerning the outliers, we only dealt with the feature “Babies” which had values : 9 and 10 which we considered being typing errors, and deleted those 2 rows. Other features showed some potential outliers such as “LeadTime”, and “Stays WeekNights” but we considered them as not being true outliers, and important values for our model. Thus we decided to not treat them as such.

2. Feature Engineering

After the data exploration, we discovered that some of the values appeared to be redundant or open to misinterpretation, and decided to make several modifications. As the feature “Agent” is the ID of the Agent and thus couldn’t be used in our future model, we decided to turn it into binary variables, such as : 0 being a customer that did not book through an Agent, and 1 being a customer book through an Agent. The same change was made to the feature “Company” that worked as “Agent”.

The feature “ArrivalDateMonth” was changed as number-month : January = 1, February = 2, March = 3, April = 4, May = 5, June = 6, July = 7, August = 8, September = 9, October = 10, November = 11 and December = 12.
Following those steps, we used the function get_dummies from the pandas library to One Hot Encode our categorical features.

3. Feature Selection

In order to see the feature importance of our different variables, we used Chi-Square, Mutual Information and RFE.

Using the Chi-Square test it is possible to evaluate the independence of two events, or how one observed event deviates from the expected other event. This way, we can determine the relationship between the category feature, or predictor, that is independent and the dependent feature as a response. When the features are independent, we will have a smaller Chi-Square

value. As opposite, a high value for the test means that the hypothesis of independence is incorrect. For model training, we would choose the features that have higher Chi-Square as the higher this one is, the more dependent on the response. With our current variables, the test only gave us one feature : LeatTime with the highest value of 666403.

Subsequently, as we have categorical features, we decided to use Mutual Information from the sklearn library. This function measures the relation between variables. A result close to 0 shows that the variables are independent, while a higher value demonstrates a strong correlation between the features. Mutual Information relies on the entropy estimation from KNN distances. In the figure below, the feature importance from the Mutual Information shows that as Chi-Square it also identified LeadTime as the most dependent and important feature. On the contrary to Chi-Square, MI classified more features as “important” : ADR - with a score almost as high as LeadTime - PreviousCancellations, TotalOfSpecialRequests, MarketSegment_group, etc.

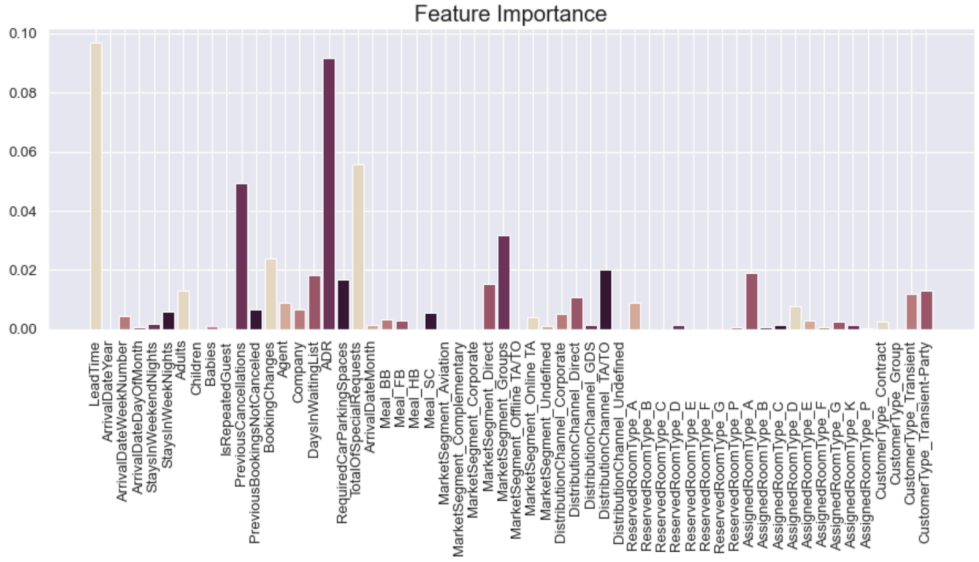

Figure 4 – Feature Importance of Mutual Information

For our last feature importance test, we used RFE (Recursive Feature Elimination) with the Logistic Regression Model. In the next figure, the ranking of those features is made from 1 till 39, where one is the most important.

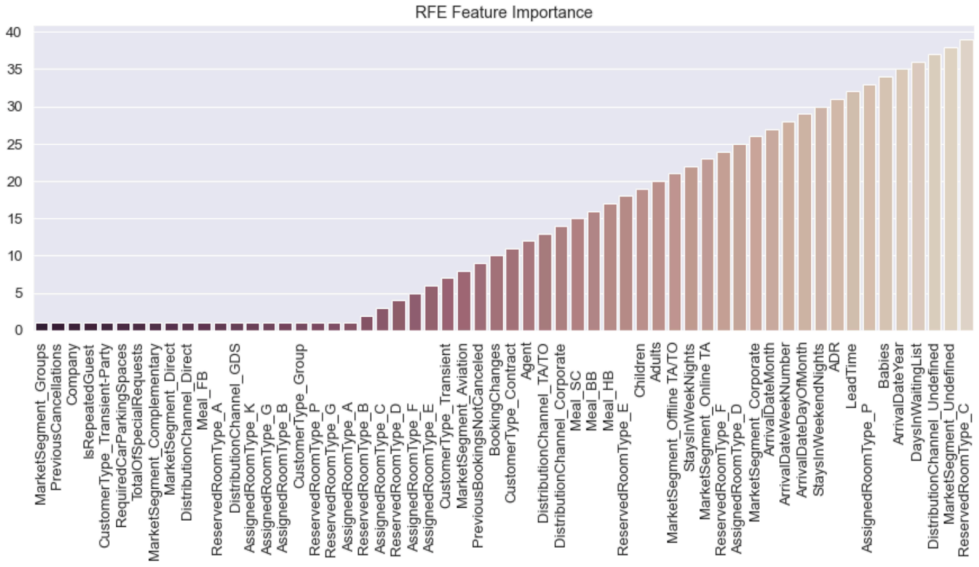

Figure 5 – Feature Importance Ranking of RFE

Following the feature selection, and the different try outs of our model, we decided to keep the entire set of features, as it gave us a better score both for training and the test set.

3. Modeling

Before attempting to find a suitable model, we standardize our dataset with the StandardScaler. To ensure that we create an accurate model capable of predicting the cancellations' booking, we tested different algorithm models : Random Forest Classifier, Decision Tree, Neural Network, (Stacking Classifier), and Logistic Regression.

1. Random Forest Classifier

We started with the Random Forest Classifier, and began by using the GridSearchCV in order to find the best parameters. The algorithm gave us the following parameters :

```
{'n_estimators': 700, 'max_features': 'auto', 'max_depth': 100}
```

After applying our new parameters, we had F1 Score of 0.876.

2. Decision Tree Classifier

The GridSearchCV gave us the following parameters :

```
tree_cv.best_params_
```

```
{'class_weight': None,
 'criterion': 'gini',
 'max_depth': 9,
 'max_features': None,
 'min_samples_split': 0.005,
 'splitter': 'best'}
```

After multiple tries, we decided to use the Decision Tree Classifier with its default parameters as it gave us a better score : 0.83 against 0.80 with the GridSearch parameters.

3. Neural Network

Regarding the Neural Network, we used the default parameter and got a f1 score of 83.8.

4. Stacking Classifier

We tried to use an ensemble learning technique and combine : Logistics Regression, K Nearest Neighbors, Decision Tree Classifier, SVC, and GaussianNB. We used Logistic Regression as a meta-classifier. As the stacking classifier model takes a long time to run, we didn't use the GridSearchCV and preferred to apply the default parameters.
The model gave us our second best F1-Score : 0.85

5. Logistic Regression

With the default parameters of the Logistic Regression Model, we manage to get a score of 0.790 which is our lowest. Then, we applied the GridSearch to improve our model, here are the parameters given : { C = 100, penalty = "l2", solver = "liblinear"}
With the new parameters we reach a score of 0.792.

6. Classifier Selection

The following bar plot shows the F1 Score of the five models : Random Forest Classifier gave us a higher score of .876.

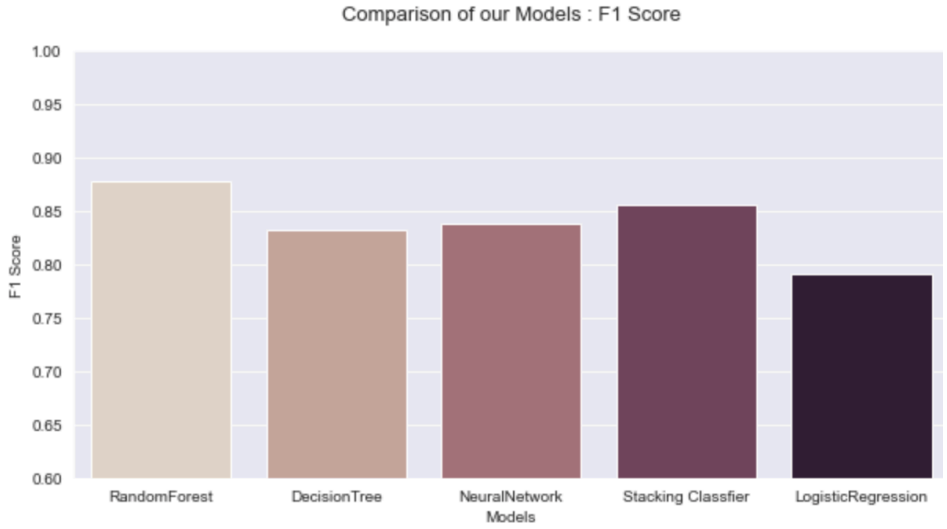

Figure 6 – F1 Score of the Models

4. Evaluation

To evaluate our final model, we used a confusion matrix, and compute the precision, accuracy and recall.

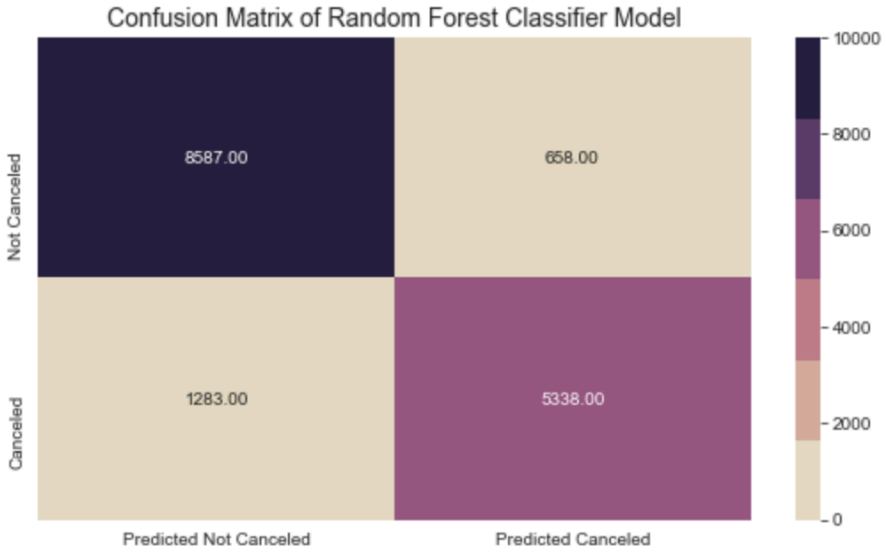

Figure 7 – Confusion Matrix of the Random Forest Classifier

With the confusion matrix and the TP, TN, FP, and FN we can compute the evaluation metrics. Here are the different conclusion made :

- 87.7% of booking from the testing set were properly classified,
- 92.8% of not cancelled bookings were properly classified,
- 80.6% of cancelled bookings were properly classified,
- 87% of bookings predicted not cancelled which were actually not cancelled,
- and 89.0% of bookings predicted cancelled which were actually cancelled.

In the notebook, you will find the ROC AUC curve, as well as the lift curve.

4. Results Evaluation

Looking at the matrix, there are 658 cases that were predicted as cancelled when in fact it wasn't. In such cases it would mean that the Hotel would have an overbooking situation in their hands, and would be required to reallocate 4% of their customers to different available hotels.

There are 1283 situations where the model predicted as not cancelled when it was indeed cancelled. 8% of customers will cancelled, and not be predicted by the model, which imply that the Hotel will lose money with empty rooms.

On the good side the Random Forest Classifier Model has an accuracy score of 87.7%, and is able to determine 89% of cases that are cancelled.

With this model, Michael will be able to forecast a more accurate demand based on the booking reservations, knowing that 89% of cancelled customers will not come. It would be possible to improve the overbooking policies, and find the right balance to reduce the possibility of being in a situation where the Hotel would have to reallocate clients, and lower the social reputation damages.

5. Deployment and Maintenance Plans

For the operationalization of the predictive model, it should be correctly implemented in the system available for reservations in the hotel management. If this happens, there must be synchronization between the application of the model and the hotel's reservation system, as well as the most relevant variables, such as BookingChanges and LeadTime. These variables can be changed over time and are important tools in the implementation of the predictive model, and can be constantly updated.

This way there could be a dashboard to better understand customers to easily access relevant information in terms of bookings. When predicting cancellation, the hotel may deliver efforts in the attempt to avoid this cancellation, offering advantages in the experience, such as discounts, car parking, spa treatments, or even local show tickets. The deployment of the model could be applied using the full database of customers, allowing the possibility of predicting the behavior of future clients, so that the marketing and management strategies to follow are as accurate as possible.

6. Conclusion

Using all the information in the data as well as data mining tools in the areas of data visualization and machine learning, our team was able to address the goals of the analysis. We were able to identify the most relevant features that can more easily help prediction of cancellation in the booking. All built models have considerable accuracy values, between 79.01% for the Logistic Regression, being the lowest, to 87.6% of the Random Forest, the highest F1-score achieved. This has shown that the Random forest algorithm presents itself as a good machine learning approach to build a booking cancellations prediction model.

The model created will facilitate the prediction of hotel bookings and promote the reduction of revenue loss that is generated from the cancellations. Also, the lightning of the restrictive cancellation policies will increase or sustain revenues, as well as the fact that most potential clients do not like this type of policies. Moreover, the problem with overbooking will be alleviated and this fact will lead to less uncertainty, while also saving money. On one hand, the hotel will not have to pay such a big amount for relocation costs in other hotels if it is overbooked at any moment. On the other, much less customers will feel the burn of overbooking and consequently bad rates and damage to the social reputation will considerably fall.

Consideration for model improvement

The model used to predict the booking cancelations could be improved if more data was collected. We recommend the « Deposit Type » feature to be extracted properly, because we believe that it could be an asset for the future models. Adding new data may lead to more accurate models, although it is hard to quantify and specify. In order to understand more the topic of booking cancelations in this hotel, different data in the property management system may be incorporated. New features may be added such as price level, deposit policies, weather information, current reputation of the hotel, exchange rates between currencies of the local and the most important nationalities in terms of number of clients are important features that can be added to improve the model.g

7. References

- Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

-Antonio, Nuno; De Almeida, Ana; Nunes, Luis (2018). Hotel booking demand datasets. Data in brief, 22, 41-49

-Antonio, Nuno; De Almeida, Ana; Nunes, Luis (2017). Tourism & Management Studies: Predicting hotel booking cancellations to decrease uncertainty and increase revenue, vol. 13, núm. 2, pp. 25-39.

- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000) CRISP-DM 1.0: Step- by-Step Data Mining Guide. SPSS, Copenhagen.
- [4]Kellcher, Mac Namee, & D'arcy (2015, p.53).
- [2]The data mining life cycle. Image from IBM Knowledge Center, https://www.ibm.com/support/knowledgecenter/it/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

Mutual Information, scikit-learn.org.

Rohit Madan, DecisionTree Classifier — Working on Moons Dataset using GridSearchCV to find best hyperparameters, https://medium.com/, 18/11/19