# BUSINESS CASES WITH DATA SCIENCE

**MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

## Predict Hotel Booking Cancellations

Suggestions on Overbooking and Discount policies


Group Y

João César (20200669)

João Henriques (20200670)

Pedro Sancho Vivas de Castro (20200132)

Vilmar Bussolaro (20200268)


March 15th, 2021


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

# INDEX

# 1. INTRODUCTION

In this project we are asked to make a predictive model for the cancellations of a hotel of the Hotel Chain C. This Chain has currently two hotels: H1 and H2. The hotel we will be analyzing and working on is H2, a city hotel.

As every other hotel, H2 is having more and more cancellations throughout time, a good prediction is needed in order to know how many rooms to make available to minimize the loss occurred from those cancellations.

For us to make a good predictive model, H2 provided us a dataset with every booking for the dates between July 2015 and August 2017.

We will apply our Machine Learning skills in order to find the best model to help the hotel chain to better predict the cancellations for H2.

Our main notebook, pre-processed data, assessment on predicting no shows and presentation will also be present in this GitHub repository: https://github.com/PedroSancho/BC2_Predicting_Hotel_Cancellations

## 2. BUSINESS UNDERSTANDING

### 2.1. BACKGROUND

Digital transformation happened in all areas of business, the hospitality industry included. The rise of Online Travel Agencies (OTAs) introduced to the market a new ideology, free booking cancelations. This cancellation option puts the hotels, that must honor the bookings, at risk, since at the same time they must support the opportunity costs of having empty rooms when someone cancels.

All around the globe cancellations are becoming a lot more common, in Europe for example, the cancellation rate by reservation value, from 2014 to 2018, rose from 33% to 40%. Cancellations can occur for understandable reasons such as business meeting changes, vacations rescheduling, illness, or adverse weather conditions. However, cancellations are now occurring due to not so understandable reasons, such as finding a better deal.

Hotel chain C, a chain with resort and city hotels in Portugal, was severely impacted by cancellations, representing almost 28% in H1 and almost 42% in H2.

### 2.2. BUSINESS OBJECTIVES

Hotel chain C's main objective is to reduce the uncertainty about demand, allowing the hotel to ensure less risk and therefor mitigating monetary losses on cancelations. The Hotel chains director expects to implement better pricing and overbooking strategies through getting a prediction process that could point out which bookings would be cancellations.

### 2.3. BUSINESS SUCCESS CRITERIA

In order to be successful, the project needs to help Hotel chain C predict costumer's intentions with enough accuracy to reduce cancellations of hotel H2 to a rate of at most 20%.

### 2.4. SITUATION ASSESSMENT

Before this project Hotel chain C already tried to implement different policies with hopes of losing less money, they have tried to have a limited number of rooms restrictive cancellation policies and have also tried to have a move aggressive overbooking policy. However, both of this failed, with the last one ending up generating more losses. Their last attempt was to have a less aggressive overbooking policy that resulted on the hotel having empty rooms due to cancelations.

As for the dataset the hotel provided data from bookings made at hotel 2, which were due to arrive between July 1, 2015, and August 31, 2017.

Our team is composed by 4 Data Scientists, which had a week to prepare a 5-minute presentation to the C-Suite, as well as a 10-Page Report (which you are reading) and the accompanying code for the assessment of dealing with hotel cancellations.

## 2.5. DETERMINE MACHINE LEARNING GOALS

To have a good prediction on the cancellations, the best approach is to use Supervised Learning. By building a predictive model, we can better know the variables that have a greater impact on the outcome (if the reservations will or will not be cancelled) and therefore predict with more accuracy the true number of real cancellations, for the company to have the best conditions upon decision making.

But before bringing information to the surface, we need to do the basics: preprocess data. On feature selection the goal is to choose a set of variables that is both broad and relevant but that do not generate the curse of dimensionality. Highly correlated features should be dropped.

We will focus on the positive results of the variable 'IsCanceled', which indicates if the reservation from the database was cancelled or not and identify the variables that can better help us predict that outcome.

We will test various algorithms in terms of recall and precision values, pursuing the one that best suits our analysis. Particularly we are seeking the highest recall value possible, maintaining our precision levels over 60%.

## 2.6. PROJECT PLAN

1. Data understanding
2. Data preparation
   2.1. Outlier pre-processing (wrong input assessment)
   2.2. Encode categorical features
3. Compare and apply different predictive models
4. Deployment and maintenance

# 3. PREDICTIVE MODEL FOR HOTEL CANCELLATIONS

## 3.1. DATA UNDERSTANDING

Having our first look at the data provided by H2, we found that this dataset has 31 different variables and 79330 reservation records on the period from July 2015 to August 2017.

We framed these variables into four different types of values: Metric, Binary, 'OHE' (which will be later subject to a One Hot Encoder) and Other. The variables are distributed as the following:

- Metric, that included the variables: *'LeadTime', 'ArrivalDateYear', 'Adults', 'ArrivalDateWeekNumber', 'ArrivalDateDayOfMonth', 'StaysInWeekendNights', 'StaysInWeekNights', 'Children', 'Babies', 'ADR', 'RequiredCarParkingSpaces' and 'TotalOfSpecialRequests';*

- Binary, composed by: 'IsRepeatedGuest', 'PreviousCancellations', 'PreviousBookingsNotCanceled', 'BookingChanges' and 'Different_Room';

- The variables that will be subject to a One Hot Encoder: 'Meal', 'MarketSegment', 'DistributionChannel', 'ReservedRoomType';

- Two other variables - 'Agent' and 'ReservationStatusDate'.

This dataset has been previously pre-processed, so the data is pretty much consistent, with a small number of null values: 4 on 'Children' and 24 on 'Country' and at a first glimpse, we may need some more pre-processing to optimize our Machine Learning concepts implemented.

## 3.2. DATA PREPARATION

Our first preprocessing step was the dealing with outliers. As the dataset was already semi-prepared, we decided to just remove or change some rare cases of outliers by applying some thresholds to some feature distributions. In the case of the variable 'Babies', we had two values that were absurd, so both were set as the value 2. In two other variables, we had just one case on each where we also changed the value to the maximum value (without the outliers). In the case of 'PreviousCancellations' it was 13 and in 'ADR' 510. Also, we set all the values in the feature 'BookingChanges' over 10 to be 10.

The managers told us the variable 'DepositType' was extracted in a wrong way and its quality was compromisedso we decided to drop the variable 'DepositType'. Also, we removed the variable 'ReservationStatus' to prevent data leakage.

We then encoded some variables. Converted the feature 'Children' into an integer and 'ReservationStatusDate' into the date time dtype.

After that, we did some minor but needed changes in some features. We removed the excess spaces in the variables 'Meal', 'Agent', 'Company', 'ReservedRoomType', 'AssignedRoomType', we spotted that in 'Country' there was the case of China being in ISO A2, when all the other countries were in ISO A3, so we made the correspondent change, in the variable 'Company' we replaced the string "NULL" with "NA" and we did set all the undefined values as 'Nans' for the features 'MarketSegment' and 'DistributionChannel'.

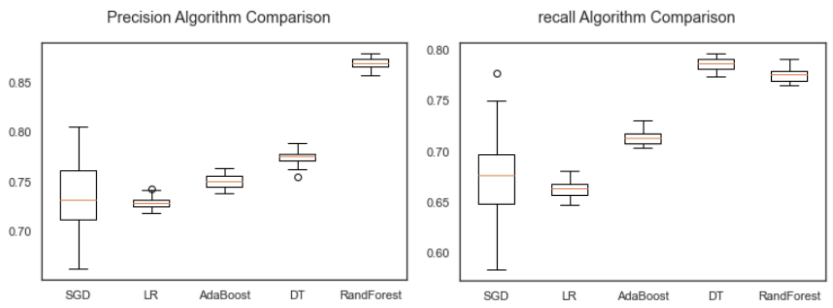
In terms of missing values, we spotted that all combined in every variable summed up to be 0.0003% of the dataset, so, due to the insignificant value we dropped those values.

Finally, we did some feature engineering. We created a new binary variable called ‘Different_Room’, which indicated if the client did or did not change room and we applied a one hot encoding to the variables 'Meal', 'MarketSegment', 'DistributionChannel', 'ReservedRoomType', so we could treat every possible value as a binary for a better analysis.

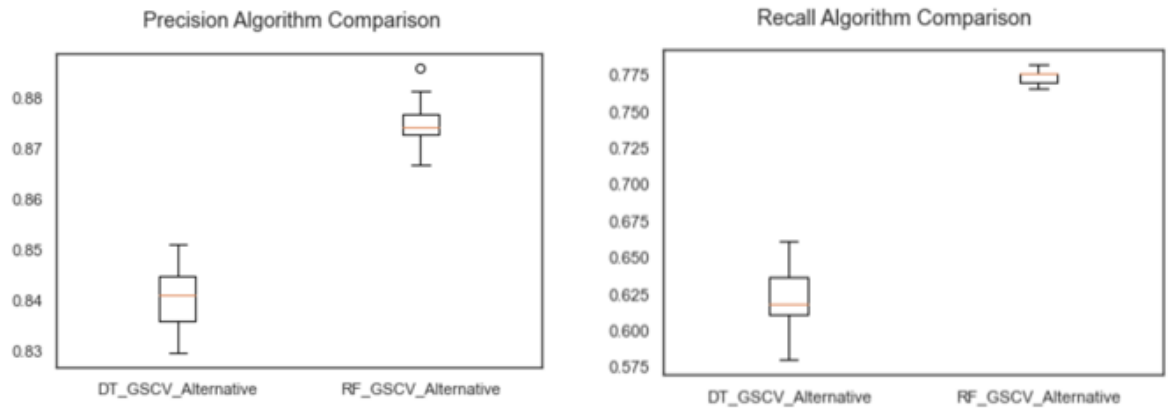
### 3.3. MODEL SELECTION

In order to reach Michael's goal of lowering cancellation rate from 40% to 20%, a precision of 50% would do the trick on achieving that goal of lowering cancellation rate. Additionally, we understand that it is advisable to have at least 68% in precision in order to deploy a model and monitor precision rate with a margin that will let the model be in production without the need to be fine tunning constantly. Although, recall is the main focus on the predictive model we will select. We are looking for a model that provides at least an 79% recall.

We proceeded to perform a repeated stratified 10-fold cross-validation assessment of different "vanilla" algorithms to select the best ones.



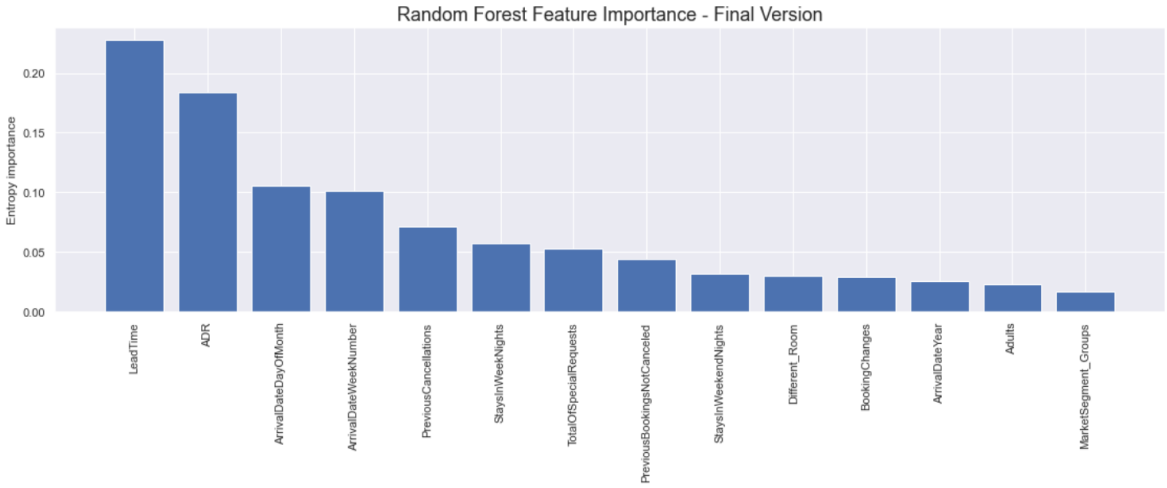

We will proceed with a deeper analysis on the Decision Tree and Random Forest models. Due to their overall better metrics both in their own terms, one in the more white-box spectrum of models and the other one on the black-box realm.




Between these two models we can see that Random Forest Classifier Algorithm has both a better precision and better recall. Our main focus is on precision and despite not being as interpretable of a model as Decision Tree we will stick with the Random Forest.

RF_GSCV_Alt had the best results for what we are seeking. RF stans for Random Forest, GSCV stands for GridSearchCV and Alt stands for alternative arguments used constructed upon the result of GridSearchCV procedure maximizing F1-Score. Our final model was Random Forest Classifier Model with the following criteria: class_weight = {0:20, 1:2}, criterion = 'entropy', random_state = 1, max_depth = None, max_features = None, min_samples_split = 4, n_estimators = 100.

### 3.4. FEATURE SELECTION




After selecting the Random Forest Classifier Algorithm, we decided to assess if the same algorithm could yield similar results with a more trimmed feature set. We ended up with 14 features with the most important being the lead time of each reservation, the total number of requests in each reservation and whether the room was changed or not.

Through an assessment using *RepeatedStratifiedKfold* for this trimmed feature set we encountered very similar precision and recall and therefore we proceed to select this implementation with the trimmed feature set over the more extensive one. Thus, by selecting this specification described by this feature importance graph we may be able to put in production a less computationally daunting predictive model.

### 3.5. POSSIBILITY OF PREDICTING NO SHOWS

We investigated the possibility of predicting No shows using a similar approach we did for predicting cancelled bookings, it did not show any progress for a variety of different vanilla algorithms. Furthermore, selecting the best algorithm from those we tested in the vanilla specification, Random Forest Classifier, did not return a good enough result even on the *GridSearchCV* F1-Score specification and therefore we did not proceed to recommend implementing such predictive model.

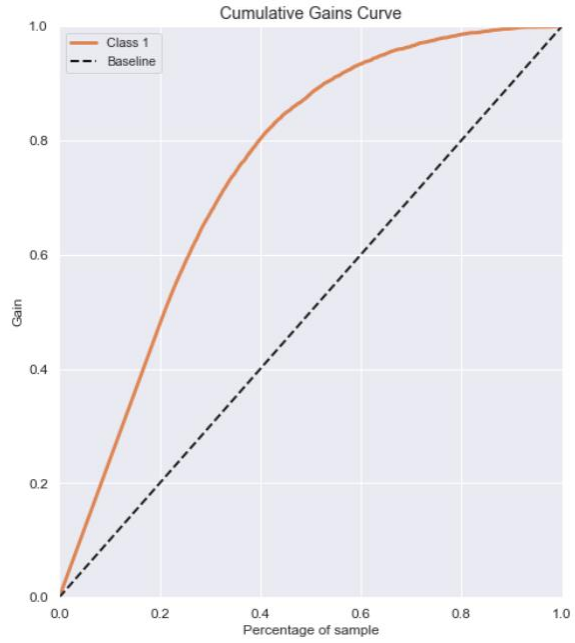
Different specifications that aimed to maximize recall generate extremely poor precision and that lead us to leave aside the possibility of generating intelligence with predicting No Show for the hotel industry

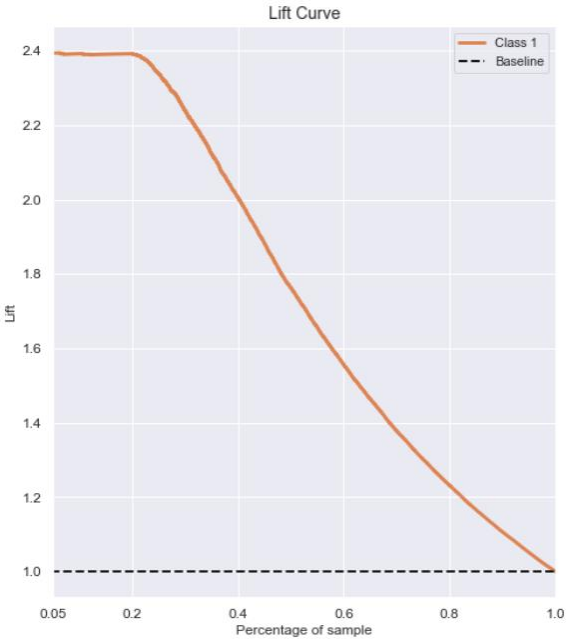
# 4. RESULTS EVALUATION

## 4.1. PREDICTION OF CANCELLATIONS

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.92 | 0.88 | 9240 |
| 1 | 0.88 | 0.77 | 0.82 | 6615 |
| accuracy | | | 0.86 | 15855 |
| macro avg | 0.86 | 0.85 | 0.85 | 15855 |
| weighted avg | 0.86 | 0.86 | 0.86 | 15855 |

After selecting our algorithm and trimming our variable set, we found that with this model we managed to get a precision score of 88% and a recall of 78% for the cancelation predictions, which is inside our desired standards to meet business requirements.




| Predicted | 0 | 1 |
|---|---|---|
| Ground Truth | | |
| 0 | 8506 | 734 |
| 1 | 1581 | 5034 |

In the table above we can see in another perspective the results we got from this model, which presented indistinguishable metrics compared to the more robust method, namely *RepeatedStratifiedKfold*.

We clearly wanted to minimize the amount of booking cancellations that we did not predict as cancellations, but the ones that are more crucial to minimize are predicting a cancellation and ending up with an overbooked customer in your hotel lobby without having a room to provide them. This only happened in 734 of the cases, 5% of total bookings and 12% of the predicted cancellations, and we can call that a positive result for our model.

True negatives and positives are bookings that we predicted correctly, contrarily to false negatives and positives that are bookings that we failed to predict correctly, therefore a false negative is someone that we predicted to not cancel, and ended up cancelling (this is restriction goal, getting it below 20% cancellation rate), a false positive is someone that was predicted to cancel but ended up not cancelling (the error that we aimed to focus).

Precision is the count of true positives divided by the sum of true positives and false positives. This fraction gives us the proportion of accurately predicted cancelations. Our model obtained a precision of not cancelled bookings of 0.84 and 0.87 precision for cancelled bookings.

Recall is the count of true positives divided by the sum of true positives and false negatives. Our model obtained a recall of not cancelled bookings of 0.92 and 0.76 recall for cancelled bookings.

Despite not aiming for improving accuracy, the ratio between true predictions and the total count of predictions, accounted for 0.85.

Our main goal with this project is to reduce the false negatives to below 20%, without generating a new problem, like the predicting cancellation with low ability to get it right (something that is measure through recall metric). Therefore, this was the main reason we wanted to get at least 60% recall and a high precision, at least 80%.

## 4.2. OVERBOOKING AND DISCOUNT POLICIES

The problem now relies on the number of false positives, since we cannot simply overbook every room we predict to be cancelled since these predictions have a precision of 88%. Overbooking client's rooms that end up not cancelling brings high costs to the hotel since if they do not have any other room available, they will be forced to put them on a different hotel.

Therefore, we understand that by overbooking 4 predicted cancellation bookings for every 5 predicted cancellation bookings will generate enough cushion to avoid a new problem appearing, such as overbooked customers in the hotel lobby without a room to check-in and will also account for lowering cancellation rate up to 14%.

Additionally, we recommend that management may study the possibility of applying already pre-made go-to discount and upgrade policies for customers since there is high confidence that 14% of bookings will still be cancelled despite being predicted as non-cancelled bookings. Thus, the goal is to avoid having idle rooms or having emergency discount policies that would generate negative margin streamlines.

# 5. DEPLOYMENT AND MAINTENANCE PLANS

## 5.1. PLAN DEPLOYMENT

In terms of deployment, we suggest collecting more data (more recent than August 2017). As the data may differ, the model can become less precise, as well as refining the model throughout time alongside with the new data that the hotel may have in the future.

Also, an application should be created for the bookings manager to have the predictions for example 7 days prior to the arrival time. Along with that, creating a live dashboard for the manager to have all the information provided by the predictive model is the best way to interpret the data and be able to better take decisions of possible promotions to eventually persuade the client not to cancel.

This could be done through a Dash implementation and should have a time series visualization of the overall impact of the predictions the model is recommending and how the results are matching the recommendations.

Overall, we generated a prediction that could have accounted for 5% increase in Revenues in the 2015-2017 timeframe. At last, to make this type of effect on Revenue happen, the company should continuously invest in Machine Learning solutions for this model to stay fine-tuned and avoid too much degradation of the model metrics since this would lead the cancellation rate to converge back to 40%.

## 5.2. MAINTENANCE

By having this type of a predictive model, we may yield to H2 the ability to hamper the cancellation rate with good confidence of getting the results it wants for a good timeframe, without the need to take the model out of production and re-tune it back. Thus, there is enough space for this metric to degrade as time goes by without endangering the goal of lowering cancellation rate to 20%.

Additionally, by having a precision of 88%, we have high confidence that by deciding on overbooking on 4 predicted cancellation once every 5 predicted cancellations, there won't be a problem of getting overbooked customers on the lobby without having rooms for them and a high confidence that it will take a good timeframe in order for the model to degrade downwards of 81% precision.

Lastly, it is important to look for these two metrics in order to guarantee that the model has not degraded to a point that it will make the overbooking and discount policies out-of-date and therefore hamper H2 business. Therefore, once recall reaches 70% or Precision reaches 81% it is important to fine-tune the model back into metrics that will allow for the overbooking policies to be effective in generating more revenue and dealing with cancelled bookings.

# 6. CONCLUSION

The hotel H2 had a high cancellation rate of 40% and this was proved to not be effective. The hotel was losing money with the amount of rooms that were empty and a better model to predict the cancellations was needed for the Hotel to operate more efficiently.

After doing some preprocessing and testing various predictive model algorithms, we were able to get two different algorithms matching our initial requirements. From those two algorithms we picked the one that better suited our needs for the predictive model, and then we were able to play inside the final algorithm we chose, the Random Forest, in order to have the best model possible. For that model we were able to select the features we found that have the greatest impact on the final result.

With the model we built we managed to get very good results from what were our initial goals. Our main goal was to reduce cancellations to a rate of, at most 20%. For that, we needed the precision of our model to be over 79%, together with a recall of 68%. Despite that, we ended up getting 88% precision and 78% recall, which is more than enough to produce the desired results and therefore the model should be deployed.

In conclusion, we believe that this model will help the hotel H2 when predicting the number of cancellations, allowing the hotel to more precisely predict the rooms that will be able to be available for further late reservations. This will reflect in better results for the future, as well as a better reputation for the company.