BUSINESS CASES WITH DATA SCIENCE

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

Prediction of customers cancellations

Group O

Eleonora Sbrissa, M20200628

Luis Reis, M20200636

Pedro Godeiro, M20200396

Sara Michetti, M20200626

March, 2021

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

INDEX

INTRODUCTION

1. BUSINESS UNDERSTANDING

Hotel bookings cancellations are the biggest nightmare for hotels: having a low occupancy rate is not sustainable for any hotel manager as rooms are their products. In this sense, predicting cancellations can help them improve their revenue.

Nowadays more and more customers are deal-seekers and tend to make multiple reservations and keep the one with the best deal, also with the help of online travel agencies (OTAs). There are mainly two approaches to fight cancellations: Overbooking and restrictive cancellation policies, both coming with pros and cons. This team of data scientists was hired to build a model that predicts cancellations in order to understand which approach is better to use and why.

1.1. BACKGROUND

The dataset was provided by a Revenue Manager of hotel chain C, a chain with resort and city hotels in Portugal. The dataset contains information related to bookings made to the hotel H2, which were due to arrive between 1st July 2015 and 31st August 2017. On that period, they experienced a 42% cancellation rate.

1.2. BUSINESS OBJECTIVES

The hotel would like to implement a model that is able to predict cancellations done by the customers, hence forecast the net demand based on reservations on-the-books. Among the business objectives the hotel manager would like to set how many overbooking the hotel can do together with better pricing policies.

1.3. BUSINESS SUCCESS CRITERIA

The business success criteria is the reduction of cancellations to a rate of 20% taking into account that the model should have the lowest False Positive rate. For the hotel it is expensive to predict a cancellation that at the end does not happen, as it may induce to overbooking, additional costs for the company as well as disappointed customers.

1.4. SITUATION ASSESSMENT

The Revenue Manager Director of Hotel Chain 2 gave access to a dataset containing all info related to Hotel H2 bookings from 1st July 2015 to 31st August 2017. The dataset, contained in an excel file, is in a table format in which each row relates to a different booking. In the dataset there are 41.7% of bookings cancelled, 43% related to 2015, 40.45% related to 2016, 42.55% related to 2017.

1.5. DETERMINE DATA MINING GOALS

To build a successful machine learning algorithm, the first thing to do is to understand the dataset, clean data and select the features that are most important to predict the binary target variable, “Is Canceled”. Categorical variables will be transformed into useful information through one hot encoding.

To assess the accuracy of the final algorithm both F1 scores and confusion matrix are going to be used. Finally, to predict the net demand, Monte Carlo simulations will be run.

2. PREDICTIVE ANALYTICS PROCESS

2.1. DATA UNDERSTANDING

2.1.1. What do we know about the dataset?

The initial dataset contains 79330 rows and 31 variables related to bookings information. Each row represents a booking, but as the customer ID is missing it is difficult to track customers behaviour.

From the 31 variables, 10 were numerical, 20 categorical, 2 boolean. The target variable is “isCancelled”, with 1 representing the reservation cancelled.

A first analysis was conducted using Pandas Profiling in order to check data distribution, data consistency and to detect missing data. There is some duplicated information that were not considered as duplicates, but they were considered as different bookings by the business.

Some observations:

- There are some rooms with no adults, but only with children and babies
- Most number of bookings occurred, on average, in the months of May and June. While December and January are the ones with the least booking. The average ADR also follows this trend
- There are some agents that cancel more than the average of 42%
- It was noticed that July 2015 and July 2017 have the highest lead time, and highest cancellation %. This is probably related to the fact that July is considered high season and many customers book in advance due to the flexible cancellation policy, then cancel when they find a better deal.

To understand better the dataset an analysis has been conducted visually and it is reported below.

Regarding the Lead time, it is clear from Figure 1, that it is not linear thorough the months, and the same can be applied to the rate of cancellations for every month. The peak has been reached in July 2015 when the average lead time was 180 days and the cancellation rate resulted in 67%. The month in which most cancellations occur is April with a rate of 46% (data available just for 2016 and 2017), followed by June with 45%. The month with the least cancellations is March with 37% of cancellation rate (Figure 2).

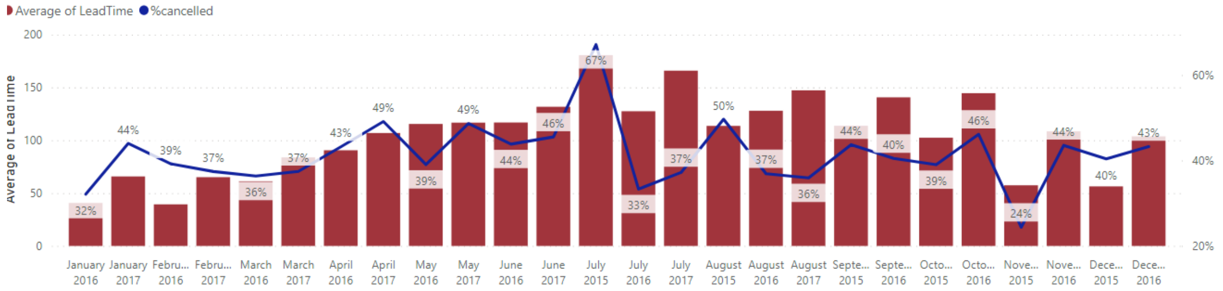


Figure 1: Lead time vs % of cancelled bookings by month-year

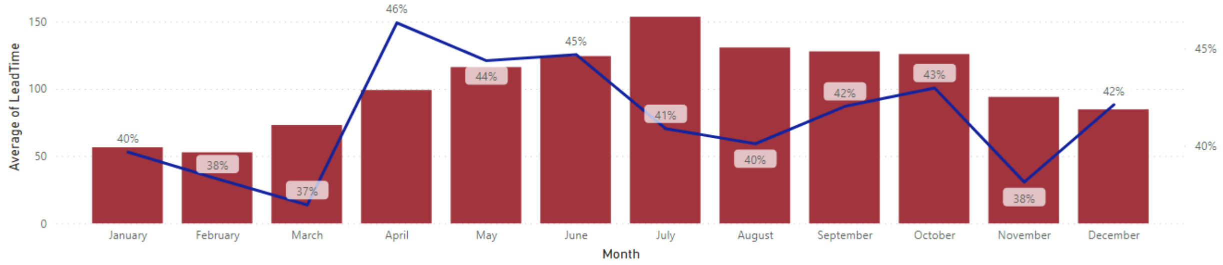


Figure 2: Lead time vs % of cancelled bookings by month

In Figure 3 instead it is noticeable that the increase of special requests by the customer significantly decreases the number of bookings that ended up being cancelled.

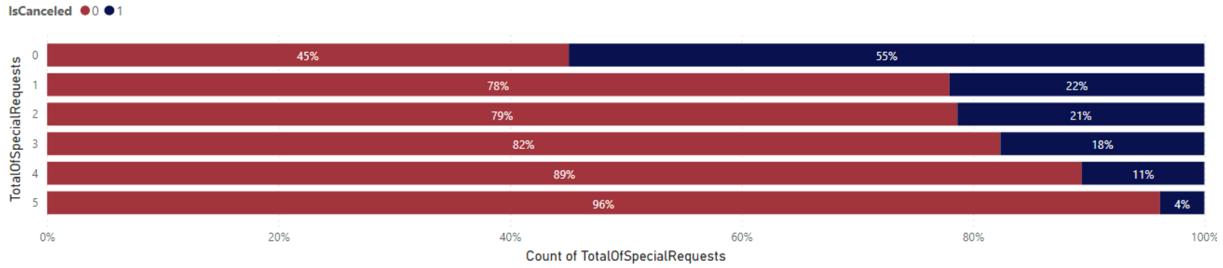


Figure 3: Special requests vs cancellations

2.2. Data preparation

2.2.1. Data Cleaning

In table 1 it is possible to see the variables that were cleaned from null values

Features	***Preprocess***
Agent	NULL values have been replaced with 0 and considered reservation with no agent
Children	Null values have been replaced with 0 children per reservation
Adults & Children	Rows with 0 adults and 0 children were deleted (considered as error) - 167 rows
Company	Null values have been replaced with 0 as there is no company for those observations

Table 1: Features cleaning

Country was dropped from the dataset, as this information is unclear and can be biased: if the hotel doesn't know the Country of Origin of the guest, they assume it is from Portugal (39% of the bookings are associated with Portugal). Also, **DepositType** was considered as mistake and dropped because it was seen that all the people that reserved with "No-Refund" then cancelled.

ArrivalDateYear, **ArrivalDayOfTheMonth** is also dropped as it is related to the check-in date, and it is not useful to predict future cancellations.

ReservationStatus is redundant as it is related to the cancellation, and **ReservationStatusDate** is related to the reservations that happen. Both of them were dropped as well.

2.2.2. Feature engineering

To add value to our model, some new features were created. The first one created was "**Number of Previous Bookings**". Since we have information about the previous bookings who were cancelled and not cancelled (in "PreviousBookingsNotCanceled and PreviousCancellations"), doing a sum of both of them makes it possible to see how many bookings were made in the past.

Another feature created was the exact **check-in date**, combining the day, month and year. This feature was not used in the model itself but was used to create another feature, which is called **DayOfTheYear**, which is the exact day of the year from 1 to 366. Since it is not possible to use this categorical feature in the model without encoding and the interval is a really similar number to how many degrees there is in a circle, the variable as the Cosine of the day of the year was encoded, having then a continuous interval between -1 and 1.

The last feature created was **RatioPreviousCancelled**, which accounts for the percentage of cancelled bookings on the total amount of bookings.

2.2.3. One Hot Encoding

One hot encoding was used to transform categorical variables so they could be used in the machine learning algorithm.

2.2.4. Split train/validation/test

To evaluate the model, it was decided to split the data into three partitions: Train, Validation and Test. First, data were split into Train (80%) and Test (20%). After that, from the training set 20% was used as a Validation dataset, which is going to be used to tune the hyperparameters of the models. The Test set is going to be used then to assess the quality of the model after all hyperparameters have been tuned.

2.2.5. Outliers

Looking at lead time, it was decided to exclude rows that contained more than 550 days before the booking. The percentage of cancellation for those customers is 100%.

Also, for the variable "StaysinWeekendNights" all the customers that stayed more than 10 weekend nights were deleted. It was assumed that those customers had special deals and did not cancel the reservation.

As can be seen in Figure 4, the ADR variable presents one extreme outlier at 5400. This has been deleted to obtain a better distribution of it.

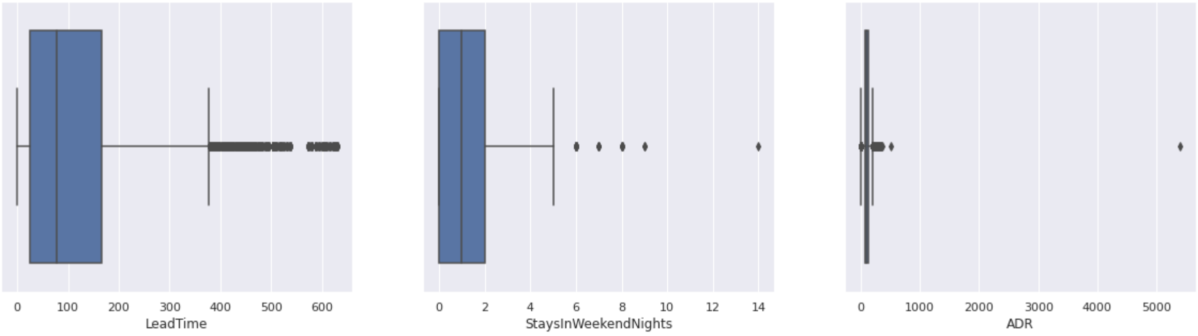


Figure 4: Boxplot before outlier removal

2.2.6. Feature Selection

To select the most important features for the model a Decision Tree Classifier was used. The parameters applied were 12 as maximum depth and entropy as criterion.

Starting from 518 variables it was decided to keep only the 107 variables that impacted the prediction. In other words, the ones with feature importance 0 were excluded. In Figure 5 there is a list of the top 10 features.

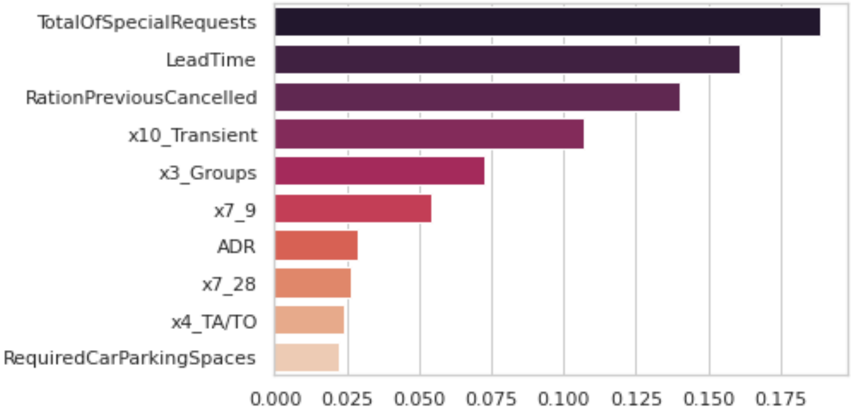


Figure 5: Feature importance - top 10 variables

2.3. MODELING

In order to reach the final algorithm various were tested: Adaboost, Random Forest, Gradient Boosting, Neural Networks and XGBoost. They were all evaluated using F1 score. The best one in terms of performance was Random Forest, then used as final model.

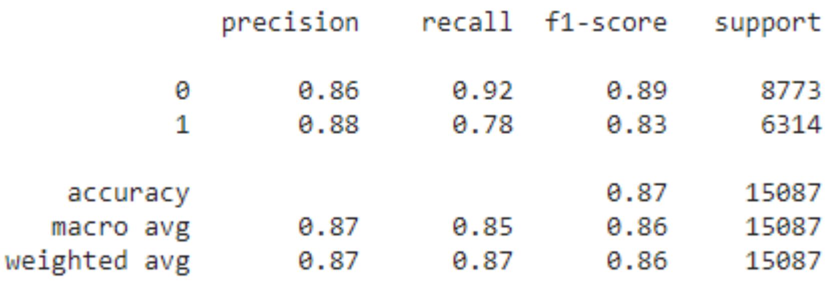
2.3.1. Random Forest

Random Forest belongs to the family of ensemble methods. The main concept is to build different trees, each one considering a randomly selected subset of variables. It is a good way of modeling since with each random tree we can get different aspects of the given problem.

For the training of the model, 48123 bookings were used along with 114 variables. To tune the hyperparameters of the model a gridsearch was run combined with cross-validation to make sure the model was using the best set of parameters. The criterion used was gini.

2.4. Evaluation

To Evaluate the model a sample of 15087 clients was used. The team was relying on the F1 score metric which was 83% and the model had an accuracy of 87%. Here follows the classification report.



```
              precision    recall  f1-score   support

           0       0.86      0.92      0.89      8773
           1       0.88      0.78      0.83      6314

    accuracy                           0.87     15087
   macro avg       0.87      0.85      0.86     15087
weighted avg       0.87      0.87      0.86     15087
```

Figure 6: Classification Report

The confusion matrix for the test set is shown in Figure 7.

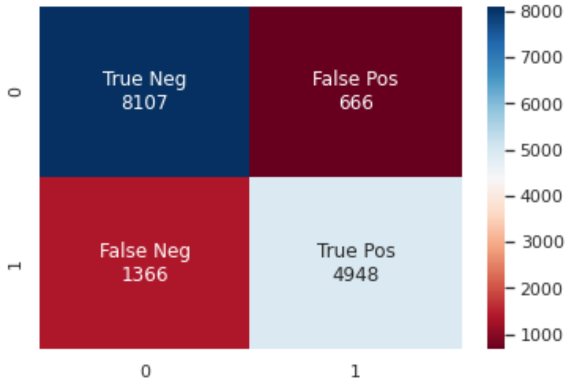


Figure 7: Confusion matrix

We can see that the model predicted 85.6% correct the people that wouldn’t cancel and did not, as for the people that the model said they would cancel it got correct 88.1% of the clients.

The top five most important variables to predict the cancellations of the clients are listed in the Figure 8.

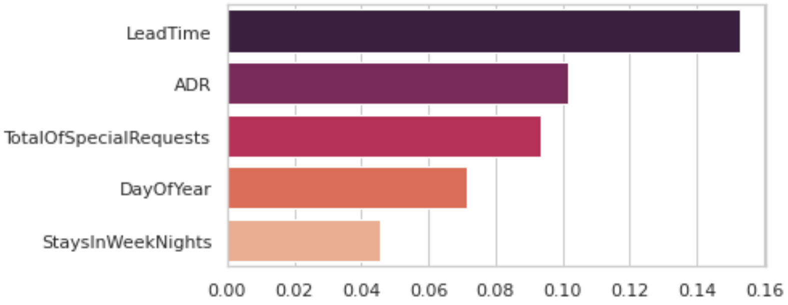


Figure 8: Model's top 5 features

2.5. Monte-Carlo simulation

To be able to give more valuable insights to the client, it was decided to create a probability distribution on the proportion of cancellations of any given dataset. In order to do that, since the probability distribution of this proportion is not known, a Monte Carlo Simulation is needed to construct this distribution.

Monte Carlo Simulation consists of repeating a process many times and compute the results of every iteration into a probability distribution. In order to do that, it was decided to retrieve from the Test Dataset a set of random observations, do the predictions on them and check how much was cancelled on percentage. Repeating that process many times helps to store the values and constructs a distribution, shown on Figure 9.

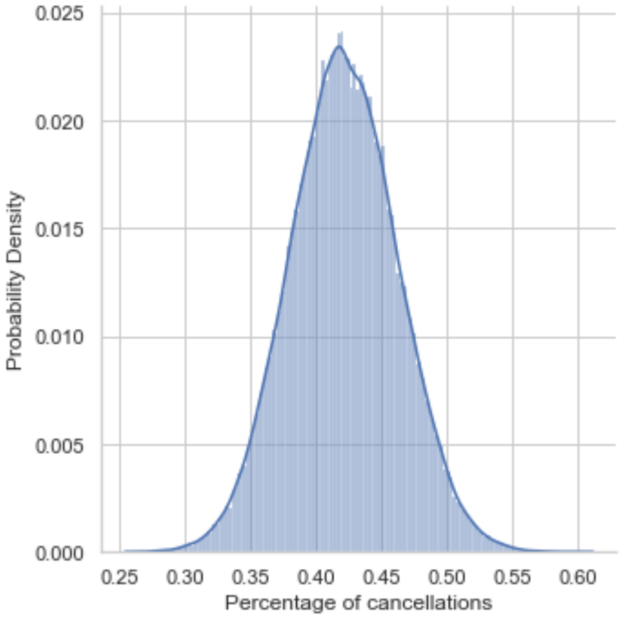


Figure 9: Distribution of Percentage of Cancellations

Exploring the probability distribution in Figure 9, valuable insights can be captured. If the client wants to have, for example, a 95% probability of having at least a certain percentage of cancellations he can overbook by this percentage, we can use the 5th quantile, which is 35,4% of cancellations. Using this distribution, the client can set any probability to define how many overbookings they can do.

3. RESULTS EVALUATION

With our high-rate prediction model and our Monte-Carlo simulation, the hotel management can utilize both, to obtain a prediction about whether their customers are going to cancel their reservations or not and choose how many overbookings should be done considering all the information. This can help reducing the percentage of cancellations to a rate of 20% or less like it was asked by the CEO.

4. DEPLOYMENT AND MAINTENANCE PLANS

4.1. DEPLOYMENT

To deploy the model, cancellations need to be included in the booking system so that the manager knows by how much is possible to overbook and also understands which are the bookings at risk of cancellation.

4.2. MAINTENANCE

As per the maintenance aspect, it is key to run the model daily to update the status of the cancellation possibility and the net demand. The variables used change continuously, both because they can vary based on customer behavior and because of the OTAs.

5. CONCLUSIONS

In order to increase the revenue of the Hotel H2, it is needed to have an overbooking strategy based on data. To do that, the Hotel needs to keep using our model for future bookings, always analyzing for every cohort of bookings by how many percent they can overbook, balancing the possible financial gains for every booking with the risk of leaving someone without a room.

Not only the model itself gives us insights on how to act, but the explanation of each rule gives us something, since we are dealing with a white box algorithm. For example, insights on seasonality are really important so it is possible to create different strategies on different times of the year. Other examples are knowing how to act with people with different lead times, average daily rate and etc.

5.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

In the first data analysis it was found out that in July 2015 there was a peak of cancellations, even if on average it is not the month with the highest number of bookings. We suggest having a deeper look into this in order to find the root cause and include also this kind of information in the model to improve the prediction.

To have a holistic approach also at customer level, and not only at booking level, we advise to collect more demographic data to also have a different point of view. This is to improve the general knowledge about your customer and eventually include this information in the model.

6. REFERENCES

https://www.redalyc.org/pdf/3887/388751309003.pdf

https://medium.com/tech4she/investigating-factors-affecting-hotel-booking-cancelations-9ec9bf81b0a8