BUSINESS CASES WITH DATA SCIENCE

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

HOTEL BOOKING CANCELLATIONS

Group F

Lorenzo Pigozzi	m20200745
Nguyen Huy Phuc	m20200566
Ema Mandura	m20200647

March 15, 2021

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

Contents

1. INTRODUCTION 3
2. BUSINESS UNDERSTANDING 3
   2.1. Background 3
   2.2. Business Objectives 3
   2.3. Business Success criteria 3
   2.4. Situation assessment 4
   2.5. Determine Data Mining goals 4
3. PREDICTIVE ANALYTICS PROCESS 4
   3.1. Data understanding 4
   3.2. Data preparation 4
   3.3. Modeling 6
   3.4. Evaluation 6
4. RESULTS EVALUATION 7
5. DEPLOYMENT AND MAINTENANCE PLANS 8
6. CONCLUSIONS 9
   6.1. Considerations for model improvement 9
7. REFERENCES 10

1. INTRODUCTION

In the hotel industry, the relationship between a guest and the hotel is established through a booking. This contract is made with a specified cancellation policy, which can potentially be damaging to the hotel. When a cancellation is made close to the date of the reservation, the hotel might not be able to rent the room to a new guest or will be forced to rent it for a smaller price.

Sometimes, cancellations are made for unpredictable reasons, in situations where the guest simply no longer needs the accommodation. However, with the rise in popularity of online travel agencies, the trend of deal-seeking has also grown. Deal-seeking is the action of making several bookings for the same trip, with the intention to cancel all but one, which turns out to be the best deal.

In order to deal with those risks, hotels have the option of either introducing restrictive cancellation policies, which might result in loss of interest, or practice overbooking, which might damage the reputation. A better approach to the problem would be to implement predictive models.

With a predictive model, hotels can forecast net demand based on reservations on-the-books. By identifying customers that are likely to make a cancellation, the hotel can adjust prices and use overbooking accordingly.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Hotel chain C holds multiple resort and city hotels across Portugal. For their hotels H1 and H2, they suffered respectively 28% and 42% of their bookings worth in cancellations.

Their Revenue Manager Director, Michael, is in charge of coming up with a solution to reduce the number of cancellations. Since none of the previous attempts significantly improved the situation, a consultant was hired to try and implement a model to predict the net demand for the hotel H2.

2.2. BUSINESS OBJECTIVES

The main business objective for this case is predicting cancellations. The idea is to identify how likely a guest is to cancel a booking they made, in order for the hotel to be able to act preventively. The goal set by the revenue manager director is to reduce cancellations to a rate of 20%, decreasing in this way the net demand of the

2.3. BUSINESS SUCCESS CRITERIA

The success of the project can be measure considering the quality of the estimation that the model will provide, and subsequently considering more the long term another aspect to take into consideration will be the result that the predictive model developed with this project will provide. A possible measure based on business can be the difference between the objective of the manager, reduction of loss of revenue due to booking cancellations, with the effective monetary value that the model will allow to gain comparing with the previous years.

2.4. SITUATION ASSESSMENT

The hotel provided the consultant with a dataset of bookings for the period from July 2015 to August 2017. The data provided was mostly clean, with few insignificant missing values. It is mostly accurate and relevant.

The hotel also provided their full support during the project, as their revenue manager director was available for questions and assistance during the full length of the project.

2.5. DETERMINE DATA MINING GOALS

The main data mining goal is making a predictive model that forecasts whether a booking is likely to get cancelled or not. The aim is to identify variables that best differentiate bookings that get cancelled from bookings that get carried out and apply a predictive model to them. The model should be able to predict whether a booking will get cancelled with a good level of certainty.

3. PREDICTIVE ANALYTICS PROCESS

The hotel H2 provided us a dataset of the bookings between July 1 of 2015 and August 31 2017.

In this step it's presented a description of the process necessary to achieve the business goals, starting from the presentation of the data available and then describing the whole technical pipeline until the final prediction and assessment.

3.1. DATA UNDERSTANDING

The dataset contains 79330 observations and 31 variables, describing some of the characteristics of the H2 customers. The metadata lists the variables and explains their meanings, available here.

The variables are mainly classified in 3 different groups:

- History of the bookings (Ex: date of arrival, lead time)
- Characteristics of the customers (Ex: number of adults, children and babies, country, isrepeatedguest, customer type)
- Specific characteristics of the bookings (Ex: meal, parking, total of special requests)

3.2. DATA PREPARATION

A very brief preprocess of the dataset was already done, and considering the overall situation it doesn't present particular issues.

Working on it, the first thing that it's possible to notice is the presence of a relevant number of duplicated rows, that seems to be potential repeated guests. Nevertheless, the H2 hotel guarantees that it's only a coincidence, thus there is no need to drop customers. The problem is due to the absence of a unique variable that identifies the observations.

Considering the missing values, there are only a few of them, 4 missing values for the variable Children and 24 for the variable Country. We removed these observations, as it's a very small percentage of customers.

Analyzing further the data through an Exploratory Data Analysis, we compute a no-sense observation check, and a problem raise up. Some records have no adults and neither no children registered, and actually those records do not represent bookings for the hotel, instead they represent other kind of expenses. Those observations are dropped.

Regarding the variables, first of all it's mandatory to identify the target variable, that in this case is 'IsCanceled'. It's a binary feature, in which the value 1 represents the cancellation of the booking. As already explained, the data mining goal is the prediction of target label of this variable.

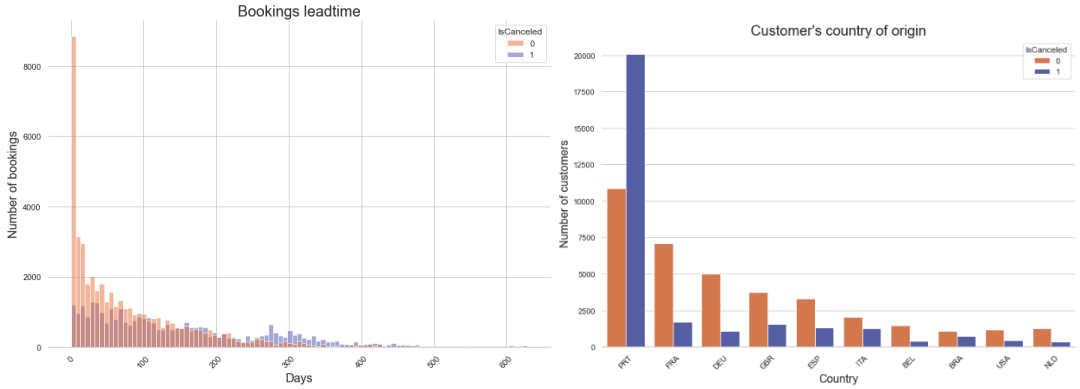


Figure 1. Variable distributions by target

Furthermore, in order to achieve the goals, it's necessary to focus only on some of the input features. Indeed, in the dataset a few variables can have a value only after the arrivals of the customers, and so those variables can't be useful for the prediction of a future customer cancellation: we need to focus only on the information we have before the date of the arrival of the customer.

Thus, interpreting the meaning and context and doing an analysis of the correlations among the variables, we arrived to this result.

Variables to drop:

- ArrivalDateYear: it's not useful for a prediction, because the goal is to predict the next customers
- ReservationStatus: it's possible to store this information only when the customer arrives
- ReservationStatusDate: other variables are more relevant and repeat info (ex. LeadTime)
- Babies: the majority of the customers has the same value (0 --> more than 95 %)
- Country: the value of the Country variable is only known for sure after the customer has made at least one check-in in the past, even then the quality of the variables isn't very reliable
- DepositType: this variable is incorrect

After that, the next step is to analyze and understand better the features considered: some of them are categorical, thus an encoding is necessary.

We decided to use the Target Encoder, a type of algorithm that takes into consideration the target variable in order to assign a value to each of the class presented for each of the variables. The encoding method is based on a calculation of a probability based on the relation among input and target variables.

The last step done before modeling was the standardization of the variables, due to the different scales they have. For this purpose, we used the Standard Scaler.

3.3. MODELING

Entering in this phase, the first task to complete is the split of the train (70 %) and test set (30 %) and the declaration of the target variable for both sets.

In order to tune and check the accuracy of the model, we also decided to use the cross validation, thus no further split of the train set is necessary.

The models we used and the results obtained are presented in the table below.

```
final_score
```

	Train acc	Test acc	Train Precision	Test Precision	Train Recall	Test Recall	Training time
Logistics Regression	0.802068	0.803195	0.799871	0.800432	0.699433	0.706679	0.184
LightGBM_baseline	0.853177	0.850105	0.885327	0.881310	0.743540	0.742379	0.423
LightGBM_tuned	0.969591	0.868768	0.968497	0.871556	0.958101	0.805656	6.033
Random Forest	0.991407	0.872005	0.992298	0.887720	0.987015	0.795227	6.231
Random Forest_tuned	0.925147	0.869945	0.938855	0.894743	0.877289	0.781689	1.541

Table 1. Model results

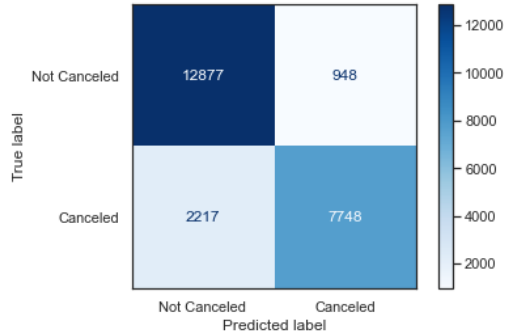
As we can see from the table, we obtained good results with different models. However, after the tuning phase of some hyper-parameters, for which we used a 5-folds cross validation, we decided to focus on the Random Forest with hyperparameters tunned in order to avoid overfitting. It's not the overall best result we obtained, but we chose this one because we have a good result for the recall measure on the test set, and also from the train we can see that the model is not overfitting: this means that we're gaining in generalization.

3.4. EVALUATION

The measure considered for the model selection in the last paragraph needs an explanation.

Considering the business context, the most damaging error for the prediction is the False Negative (FN), because a customer who will potentially cancel is a problem that really matters for the hotel in terms of revenue. The choice is made because based on the presentation of the Hotel Chain C, the marginal loss in terms of revenue due to a cancellation is worst situation the company faces up.

Considering the business needs, thus, it's advisable to focus on the reduction of the FN, as much as possible. In order to take it into consideration, the best choice of measure is the Recall, that consider the True Positive versus False Negatives, in particular focused on the prediction of class "IsCanceled" = 1. Getting a high recall means reducing the FN error.



For the evaluation of the model, so, we calculate the Recall for the test set, that we have not used so far in all the steps of model selection. A note that it's important to specify is that when we encoded using the

Target Encoder, obviously we have not used the information of the target variables of the test set, but only used the method *transform()* for the categorical features.

As showed previously, the recall obtained is 78 % on the class 1, which means that with the model created we are able to detect correctly that number of cancellations.

4. RESULTS EVALUATION

The feature important score generated from both Gradient Boosting and Random Forest models were evaluated to identify key features that affected the prediction of cancelation. The features that appeared in the top 5 of both models are *LeadTime, ADR, Agent, CustomerType, TotalSpecialRequest and ArrivalDateWeekNumber*. Having the features listed, we were able to profiling the reason behind canceled bookings to help business stakeholders have better understanding about the model result and knowledge of factors affect cancelation probability

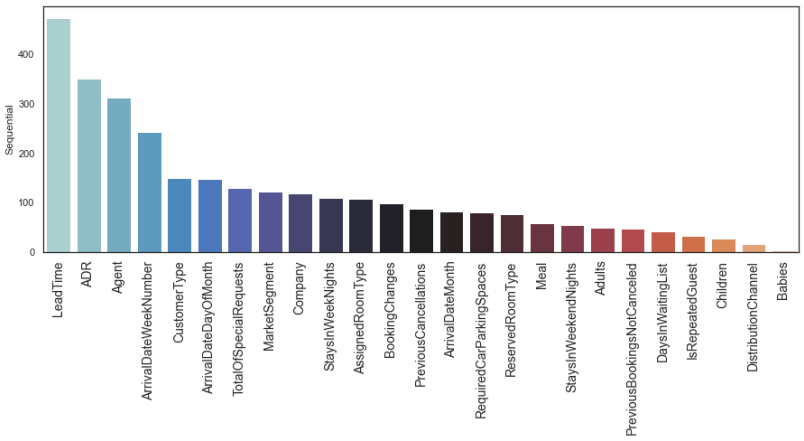


Figure 3 – Feature importances

In order to present the evaluation of the results based on the business context, the main focus we want to present is the interpretation of the probability of cancellation that the predictive model provides. Based on that, the company could take its conclusions and have a better idea of the business actions to avoid and prevent the problem of the booking cancellations, that really matters in terms of loss of revenue. For example, bookings with longer lead time, bookings from agent 1 and 3, or booking without special requests have higher probability of cancelation.

Then, having the predicted probability of each bookings, we created 3 different types of outputs based on intervals of the probability.

Probability of booking cancellations:

- 0 – 50 % of probability: based on data, this customer has a low probability to cancel the booking, thus no more action is required.
- 50 – 80 % of probability: this customer has a relevant percentage of probability to cancel. So, probably the revenue manager could be interested to make some business or marketing choices to avoid the cancellation of this customer, such as offering some kind of promotion to this customer in order to avoid the cancellation. Some possibilities that could interest the customer could be the free parking, discount for children, free meal or similar offers.

- 80 – 100 % of probability: this customer has a very high probability to cancel. The data will not provide more information for a better or accurate profile of the customer, but it's probable that having a really high probability of cancellations means that many factors can affect his decision. So, the manager has to take into consideration this probability in the business context. We can suggest to actually overbook the rooms of this customers or to apply some deposit policy, because they will very probably cancel the booking in the H2 hotel.

Obviously, these intervals we present above are just a suggestion, but other factors can be taken into considerations for the actions required for each probability of cancellation, such as the costs or the effective monetary loss of revenue for each cancellation.

5. DEPLOYMENT AND MAINTENANCE PLANS

The deployment phase of every company is really crucial for the business context. It provides a better idea about the effective usage of the machine learning algorithm created, designing a possible bridge between the business goals and the machine learning goals, applying the algorithm in a real-world context.

In this project we focused on the creation of a predictive model for the booking cancellations, thus it would be very useful for the manager of the Hotel Chain to have this prediction easily at handy.

In order to improve the accessibility of the results obtained for the company, our team have created an API that enables the directly use of the supervised algorithm developed.

The application retrieves a different kind of output based on the probability of the cancellation for each customer, as already explained in the previous paragraph. Basically, the application provides to the manager a directly insight and information about the typology of the customer based on the probability of cancellation, simply inserting the input data and pressing the predict button.

For further development, the predictive model can be used for several business application such as (1) an input for net demand estimating and resources planning, (2) a reference to build a classification model to classify different type of canceled bookings.

Regarding the maintenance plan, the dataset provided was in a good condition and was necessary just a relatively quick preprocessing phase, nevertheless in order to provide better results or at least to ensure the same level of quality of the predictions, the supervised model needs to be fed with the same amount of data, if available even more. As we created a predictive model, the more data are provided, the better will be the prediction. It's also very important to maintain the quality of the data and the structure of the variables as similar as possible. We can understand that this could not be possible always in a company, due to technical problems or maybe some improvements or changes on the storing phase of the data. For instance, the introduction of new records with different values can be supported by the model, but if the structure of the dataset will change, it's advisable to re-build and train another model from scratch, in order to have the best result.

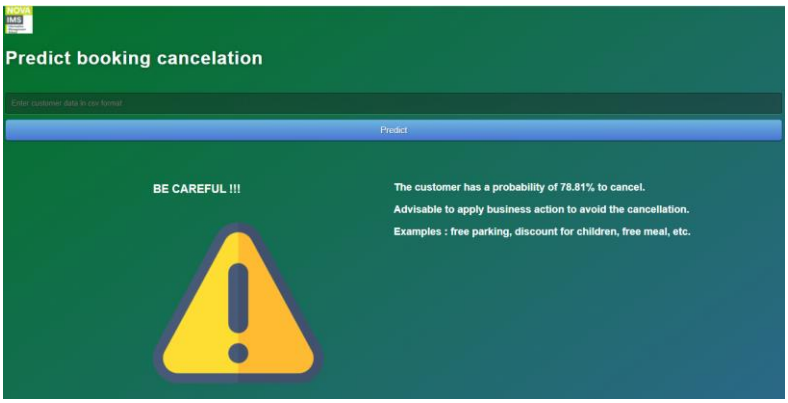


Figure 4. API

6. CONCLUSIONS

In order to support the Hotel Chain C company, in this project we analyzed a dataset of a urban Hotel B, placed in Lisbon. The overall objective of the company is to reduce the number of booking cancellations, that really affect the hotel H2 in terms of revenue. To complete and try to achieve this business objective, the data mining goal of the project was the creation of a model able to predict as well as possible the future cancellations: with this information, the manager could be able to contact the customers previously and offer them some promotions in order to avoid the cancellations of the bookings, which is the worst scenario considering the business of the hotel.

The dataset provided contains the informations of bookings of the Hotel H2, between the 1st of July 2015 and the 31th August 2017, with a total of 79330 records.

After a brief preparation and pre-processing of the dataset, in which we dropped some records and some variables based on the data mining goals, we started the modeling phase.

For the first prediction we used several models, such as Support Vector Machines, Decision Tree, Gradient Boosting and others, but then we focused on the best one: the Random Forest. The subsequent tuning phase for this model provides an even better overall result of the predictions.

Finally, we assess the built model using a set of unseen data, and it was able to provide a prediction with a recall on the cancellation class of 78 %.

We are confident that the result obtained could be very useful for the hotel, in order to have a prediction of the customers that have more probability of cancel the booking, and in this way getting the possibility to react in advance to decrease the number of effective cancellations.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Considering the condition of the dataset we started working with, the overall status was good and for sure the company is doing a good job of data-storing.

However, as we had to work on it, the main suggestion we can give is to provide a variable containing an id for each record, in order to be sure of the non-duplication of an observation; this could improve the quality of future analysis.

7. REFERENCES

1. Nishant Mohan, *All about target encoding for classification tasks*, https://towardsdatascience.com/all-about-target-encoding-d356c4e9e82