

Chapter 3 Story Hotel Booking Cancellations: eXplainable predictions for booking cancellation

Authors: Miłosz Michta (University of Warsaw), Kazimierz Wojciechowski (Warsaw University of Technology)

Mentors: Maciej Andrzejak (McKinsey & Company), Alicja Joško (McKinsey & Company)

Apr–Jun 2020

3.1 Introduction

Imagine you are the owner of a portuguese hotel. You notice that some of the reservations are cancelled on a regular basis. To maximize revenue you plan to practice room overbooking strategy. How exactly would you approach this strategy? How many rooms to oversell? Which customers are most likely to cancel their reservation? What is the probability of cancellation? And most importantly: **what is the reason behind the prediction?** We will make an attempt to answer this question using explainable machine learning.

From the standpoint of a machine learning engineer one might want to find a model that not only is high-performing, but is also based on intuitive insights about the data.

A persuasive manager of the facility might want to attempt to offer the booking on slightly different conditions to customers that are most prone to cancel their reservation e.g. by offering a non-refundable booking for customers from Portugal and Germany. It will be of crucial importance to determine which features to tweak to make the booking as secure as possible while trying not to make the customer feel too uncomfortable by insisting on too many changes. To solve this interesting problem we propose instance-specific explainable machine learning techniques.

3.2 Problem specification

In this chapter, we will use data from Hotel booking demand (Mostipak 2020) The data contains several information about when booking was made, what is the date of an arrival, how long visitors will stay, where are they come from, how many of them will come, etc. In terms of hotel, we know what type of the hotel is, what is the ADR, the deposit type, agent and company that made the booking or are responsible for paying booking, and much more.

Our machine learning task will be classification whether client will cancell their bookings or not. From business perspective, model like this might not be explicite useful, since predicted category should not affect decision of cancellation from the side of company, agent or hotel itself. On the other hand, use of explainable machine learning might give us the most important factors of cancellation process and use them to reduce rate of cancelled reservations.

To obtain most reliable results, we have used three different models: LightGBM (Guolin Ke 2017), Naive Bayes and Logistic regression. Each of them has distinct learning procedure and structure, thus our conclusions should not be biased by the choosed model.

3.3 Target leak detection

At the beginning, we test how baseline models work without any feature engineering. Suprisingly, all models get 100% accuracy score both on training and validation score. This could mean three things:

- We are awesome!
- We are lucky!
- We have a target leakage in the dataset!

Firstly, let's see partial dependence profiles for each feature. If there is some leakage, then average prediction profile should distinguish itself amoung the others. At first glance, the `reservation status` feature has strange plot. It looks like we ommit the feature represents almost the same as target.

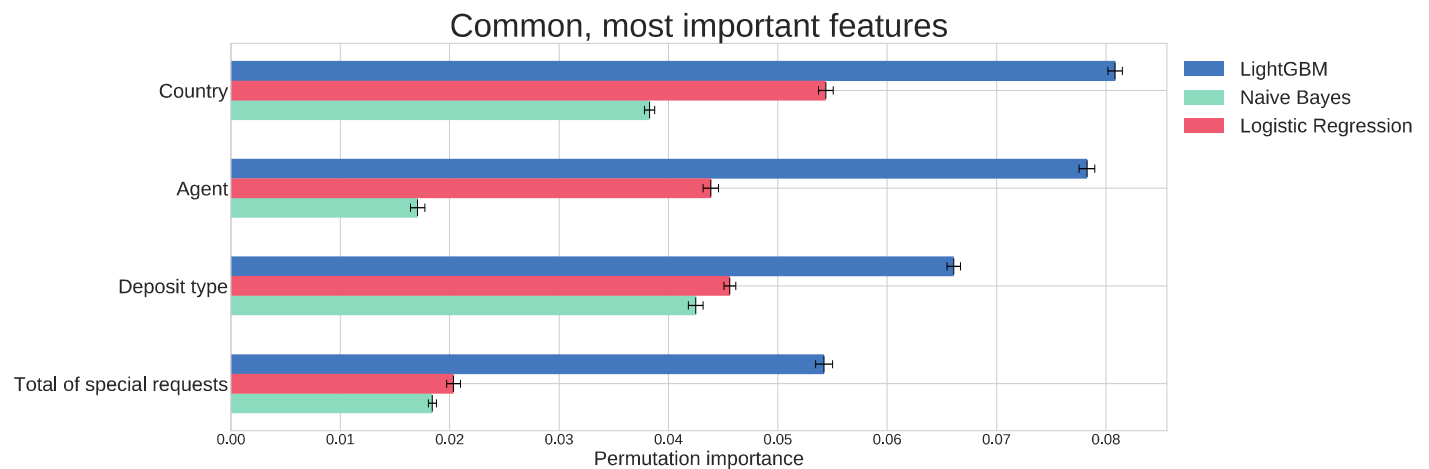


Partial dependence profiles for reservation_status feature using three different models: LightGBM, Naive Bayes and Logistic regression.

We had to remove this feature from our dataset even if we loose high performance score, because of laking interpretation power.

3.4 Bias correction

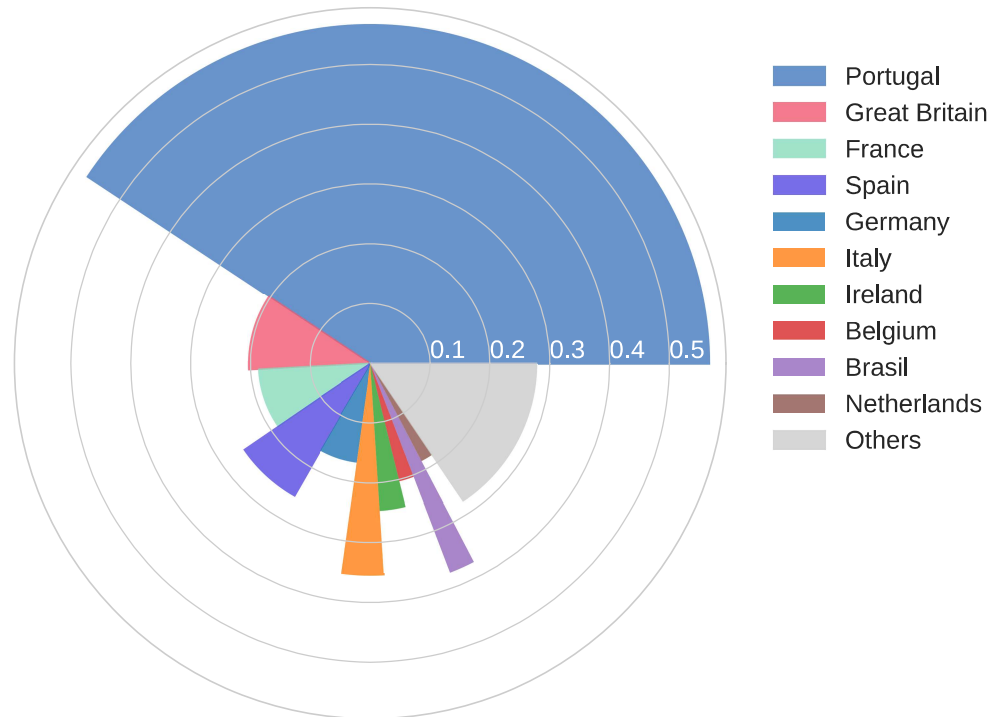
Since we have already remove leakage feature, lets figure out which variables are the most important.



The most important features for LightGBM, Naive Bayes and Logistic regression in according to Permutation Importance.

It looks like, the country of origin is the most important and the most impactful for 2 out of 3 models. The others variables looks reasonable, but logically country should not be the most important feature for predicting booking cancellation. Lets figure out what what is the cause if results like that.

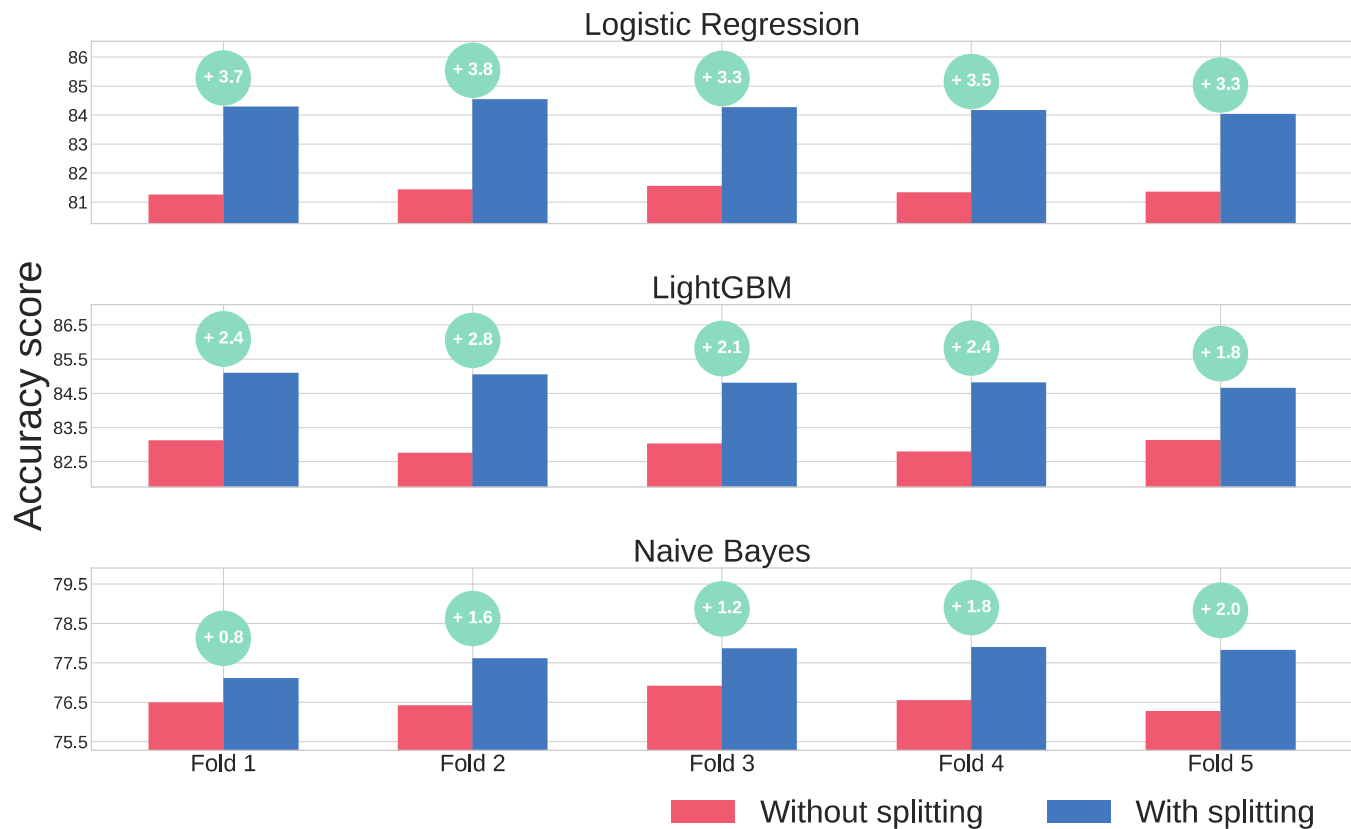
Number of observations and average target per country



Dependency between number of observations (angle) and average prediction (radius) for each country.

The Portugal seems to be reason of the bias in the dataset. Around 40% of the data comes from Portugal and their average prediction equals 56%, where the average is around 38%. In this case, we split model on 2 parts:

- One model for portugals
- One model for other countries.



Accuracy score uplift after splitting data for portugals and other countires.

After splitting the dataset in each fold and training two distinct models, we get significant uplift in model accuracy for each model and fold.

3.5 Offering different conditions

Here we will attempt to use a instance-specific explainable machine learning techniques in order to provide a similar offer that will reduce the probability of cancellation. We will determine which methods, if any, might be applicable for real-time negotiation.

3.6 Conclusions

As you probably noticed, explainable machine learning gives a lot opportunities to validate predictive models, find most insightful facts about data and set new directions to improve our results.

Explainable machine learning methods can be compared to the pointing fingers on every model weakness and to the compass guiding where we should go to improve final outcome.

Thank to XAI accompnied by strong analysis and visualtion, we was able to achieve better results of our models.

References

Guolin Ke, Thomas Finley, Qi Meng. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>.

Mostipak, Jesse. 2020. "Hotel Booking Demand." <https://www.kaggle.com/jessemostipak/hotel-booking-demand>.