

Question/Need:

Cancellation of hotel bookings is a problem that has a substantial impact on nearly every online management system. These cancellations limit the production of accurate forecasts, reduce profits, and harm the hotel. Therefore, to reduce the negative effects of cancellation, it is significantly useful for hotels to identify which bookings are likely to get canceled.

The goal of this project is to build a classification model to predict whether a new booking will be canceled or not. It will allow hotel managers to plan the hotel operations accordingly, decide more assertive cancellation policies, and define better overbooking tactics.

Data Description:

I am planning to use the Hotel booking demand dataset available in <https://www.kaggle.com/jessemostipak/hotel-booking-demand>.

The dataset has 32 features and more than 11,000 observations.

The model will predict the status of the booking by labeling the target variable is _canceled as Canceled (1) or Not Canceled (0).

The features I will probably use (not sure yet) as indicators of whether a booking will be canceled or not are:

- lead_time: number of days that elapsed between the entering date of the booking into the PMS and the arrival date
- previous_cancellations: number of previous bookings that were cancelled by the customer prior to the current booking
- deposit_type: indication on if the customer made a deposit to guarantee the booking this variable assumes three categories: No Deposit / Non Refund / Refundable.

Tools:

- NumPy
- Pandas
- Scikit-learn
- Matplotlib
- Seaborn