ASML:Classification Assessment

Xiaoning Nie. (cshb26)

March 21, 2022

1 Executive Summary

Imagine you are a hotel manager and you want to know if your guests will cancel their orders after booking. Call each customer to ask if they want to cancel their order? That's not practical. It can actually be solved quite well. A computer can help you with this repetitive task, and given enough food - the datasets - it will do the job brilliantly.

This project is all about using a dataset from a hotel to analyse the likelihood of future guests making a booking withdrawal. The dataset contains 119,390 pieces of data, each containing 32 features such as hotel type, guest payment method, number of guests, guest wait time for confirmation, etc. Some of these features are related to cancellations and some are not.

What this project did was to filter out the relevant features and feed them to a computer, which generated a model that could be used to predict whether a guest would cancel in the future, and then optimise it to improve the accuracy of the computer's predictions. At the end of the project, the model achieved the desired accuracy rate!!!!——???, so it can be considered a successful model.

The use of computers to analyse data can be very challenging but also very interesting as when analysing a dataset, we can visualise the data to make our analysis more sufficient and efficient. For example, when first reading in the dataset, we can draw a few features of interest to see the weight of each attribute, as shown in Figure 1 for total_nights and adr, the average daily rate. It is clear from the plot that the vast majority of people choose to stay for less than 10 days and almost all hotels have an adr of no more than 300 per night.

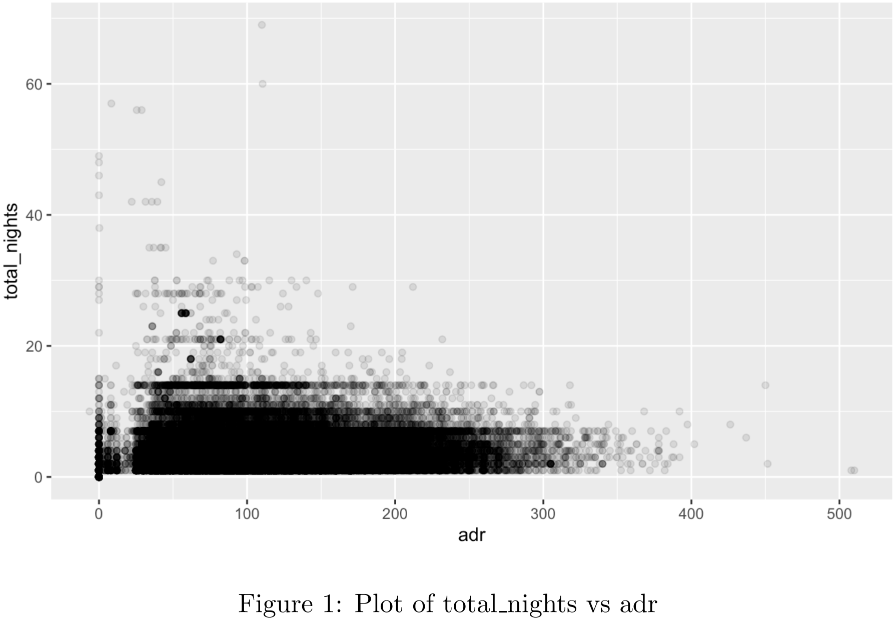

Figure 1: Plot of total_nights vs adr

2 Problem Description

The dataset is a hotel dataset containing 119,390 data and 32 features detailing each guest's booking and itinerary information, which is used to train the model. In detail, 13 of these 32 features are of type character, 18 of type numeric and 1 of type Date, with one of the features of type numeric, is_canceled, being intended to be predicted by modeling. Also, this dataset is of high quality with few missing values, which facilitates modeling. Figures 2, 3 and 4, 5 show the relationship between the data and predicted values for the numeric and character types respectively, and it is clear that certain features are not relevant to the predicted values and will be cleaned up later.

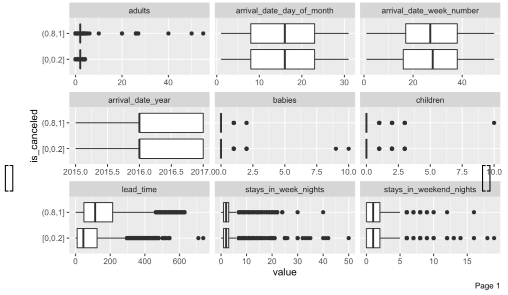

Figure 2: P1 of numeric data

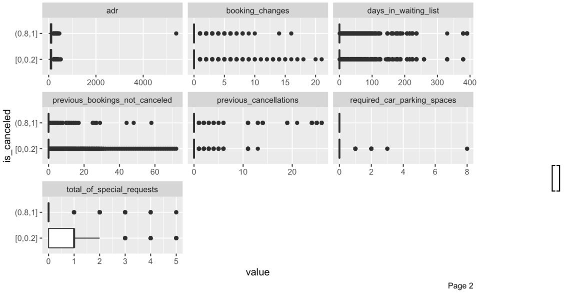

Figure 3: P2 of numeric data

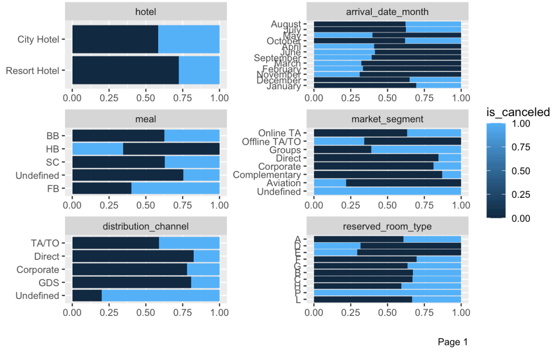

Figure 4: P1 of categorical data

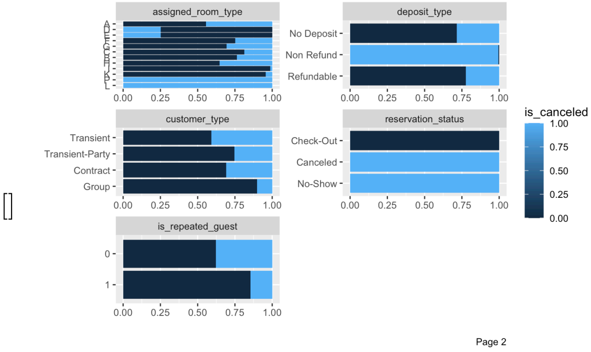

Figure 5: P2 of categorical data

3 Model Fitting

3.1 Feature engineering

3.1.1 Data cleaning

A look at the dataset reveals the following redundant features or non-relevant variables, so they are cleaned.

1. `reservation_status` and `reservation_status_date` are not meaningful to the predicted outcome and are therefore removed.

2. `stays_in_weekend_nights` and `stays_in_week_nights` have redundant information for the forecast, so `total_nights` is set to replace them.

3. `kids` is set to replace `children` and `babies`.

4. Set `room_type` to 1 if `reserved_room_type` and `assigned_room_type` are the same, 0 if they are not.

5. Create a new variable `parking` which is either `parking` or `none` depending on the value of `required_car_parking_spaces`.

6. Convert `arrival_date_month` to numeric type data.

7. `country`, `agent`, `company`, `date` and `week` categories are too many to be considered and are therefore removed.

3.1.2 Correlation analysis of the remaining variables

The results shown in Figure 6 and 7 can be obtained by correlation analysis of the remaining variables, and it was judged that ** had a low correlation with the predicted results, so they were removed.

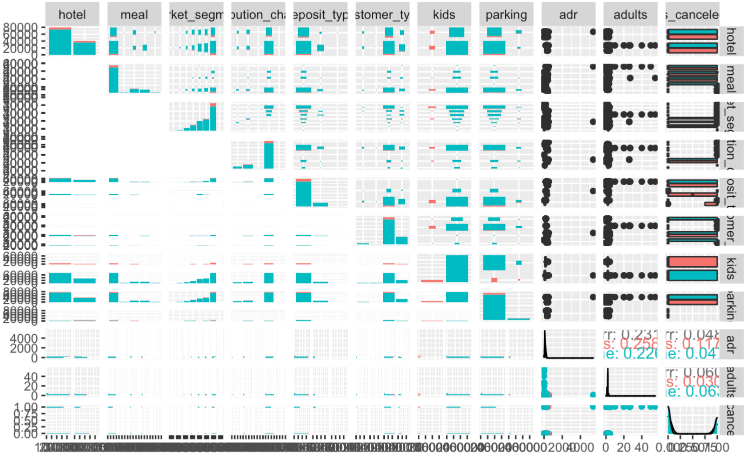

Figure 6: P1 of the analysis

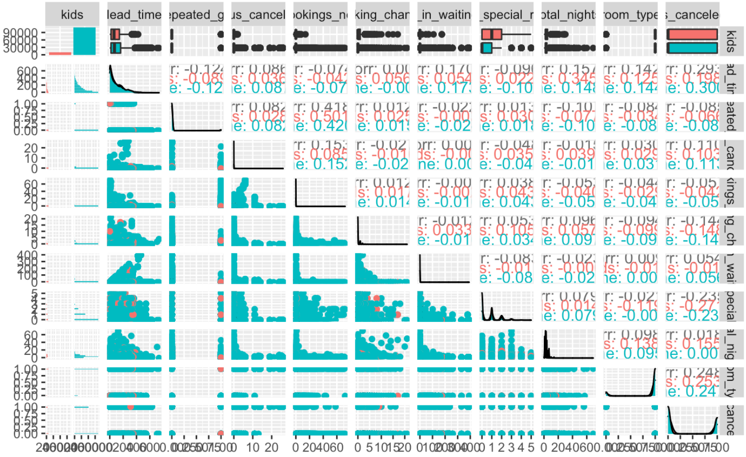

Figure 7: P2 of the analysis

3.1.3

2

1. 6days 99%0
2. 7lead load 4load1-4

3.2

4 Model improvements

5 Performance report

6 Reproducible code