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 - 82%, 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.

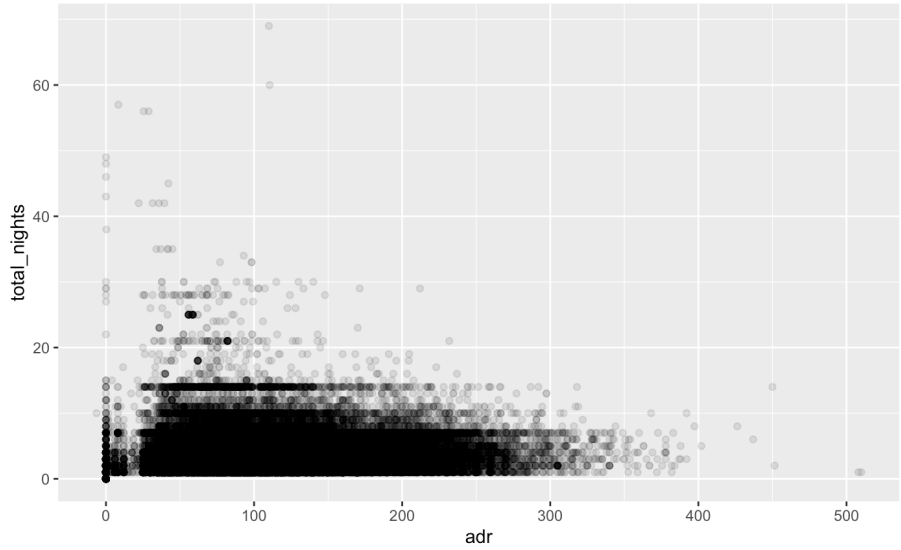

Figure 1: Plot of total_nights vs adr

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.

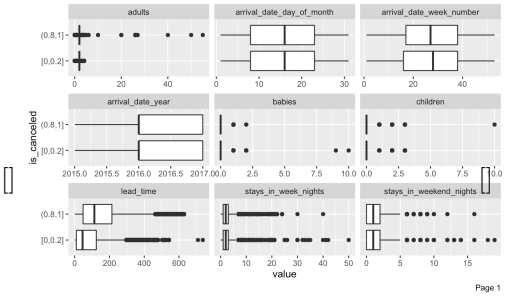

Figure 2: P1 of numeric data

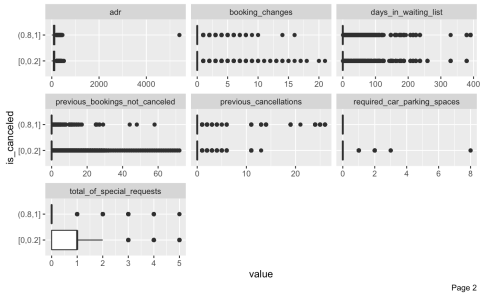

Figure 3: P2 of numeric data

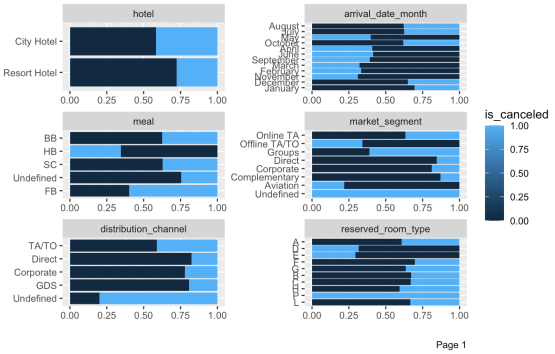

Figure 4: P1 of categorical data

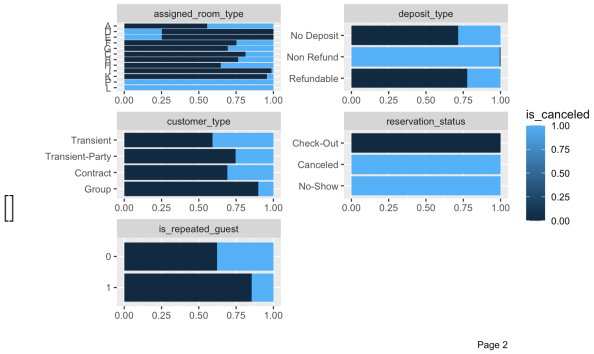

Figure 5: P2 of categorical data

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 some features had a low correlation with the predicted results, so they were removed.

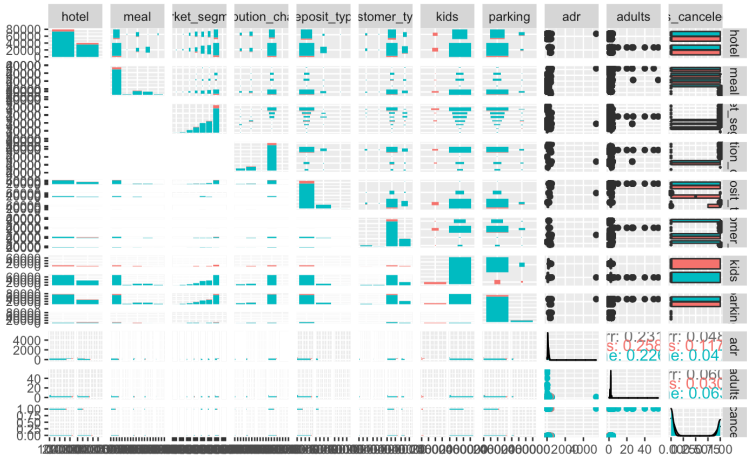

Figure 6: P1 of the analysis

Figure 7: P2 of the analysis

The above images show that there are several features such as `hotel`, `meal` and `adr` that have low correlation with the predicted value, so they are removed here.

3.1.3 Outlier handling

The analysis yielded the following 2 outliers.

1. We can see from the dataset that 99% of `days_in_waiting_list` are 0, which is not representative, so is deleted.

2. As shown in Figure 8, there are many values in the `lead_time` category, so it was discretised and binned, i.e., the values of `lead_time` were divided into 4 parts with equal frequency, and the values were replaced by categories 1-4 in order.

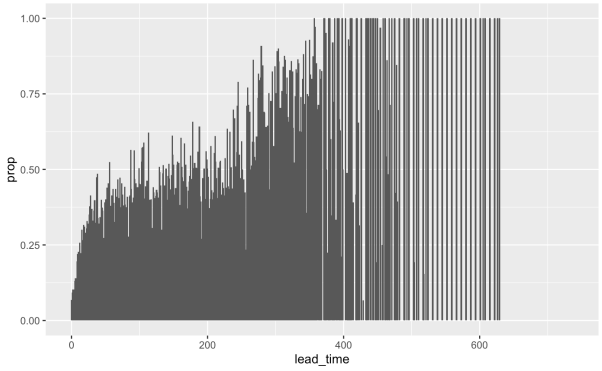

Figure 8: Data of category lead_time

3.2 Modeling

Logistic regression was used to solve this problem based on the choice of the number of target predicted values, and the results are shown in Figure 10.

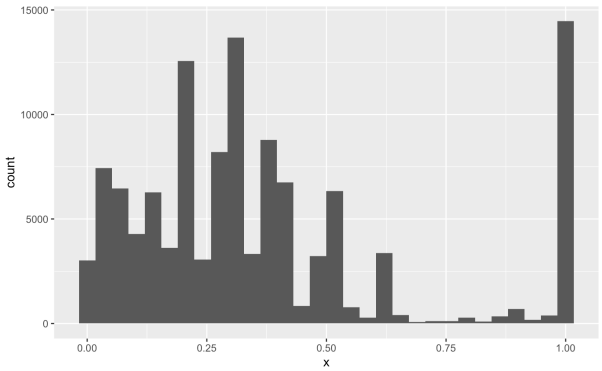

Figure 9: Simple logistic regression result

It can be seen that the model gives the probability of the full data and by setting the threshold to 0.5 the model prediction can be obtained.

The TPR of the model can be calculated as 93.58 and the FNR as 52.06, with an overall correct rate of 82%, which seems good but could be improved.

4 Model Improvements

The addition of the hyperparameter epsilon=1e-5 to the logistic regression was tried, however the results did not change. Therefore, cross-validation was used to improve the accuracy and the following cross-validation diagram was obtained.

A number of different models from the `mlr3` library were tried to test their accuracy, and a super learner was defined to find the best performing model. Figure 10 and 11 show the cross-validation and super learner diagrams respectively.

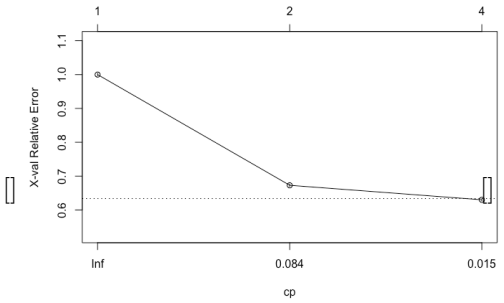

Figure 10: Number

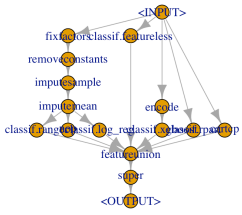

Figure 11: Proportion

5 Performance Report

5.1 Model selection

For a simple logistic regression model, it is good enough to get 82% correct, but the results improve again when cross-validation is used.

For problems with small datasets in general, using simple logistic regression will lead to an increase in efficiency because it is faster and gives a relatively good result; but for larger datasets, using other methods such as cross-validation can improve the correctness rate.

5.2 Post-model analysis

The number and proportion of TPR, FPR, TNR and FNR obtained by simple logistic regression are shown in Figure 12 and 13.

canceled	predict cancel FALSE	TRUE
0	70337	4829
1	21199	23025

Figure 12: Number

canceled	predict cancel FALSE	TRUE
0	93.575553	6.424447
1	47.935510	52.064490

Figure 13: Proportion

It is clear from the results that in this model, TPR=52.06, FPR=47.94, TNR=6.42 and FNR=93.58.

For false negatives and false positives, I would be more concerned about false positives in this problem. One reason is that false negatives are correctly predicted and more accurate, the other reason is false positives mean that the guest did not want to withdraw, but the system calculated that the guest would, which would cause the hotel to lose credibility.

6 Reproducible Code