Expedia Hotel Ranking

Kaggle: hiqbal2
Score: 0.33075
Github: https://github.com/hamzahiqb/expedia_challenge

Metric of Choice:

From the initial analysis, the dataset is very unbalanced with only 2.8% of the observations having a booking. In order to build an accurate model to rank the hotel, we will need to ensure that we have enough positive samples in the cross-validation.

Since the training accuracy can be as high as ~97%, it would be more appropriate to consider precision as the metric to optimize for. In my modelling, I assume that there is no cost of a false positive, hence, recall is not an issue. Given that every search ID in the dataset has an associated booking, I take a simplifying assumption that there is a high likelihood of a search/visitor leading to a booking. Hence, the cost of a false positive is low.

However, in reality, there is high cost associated with acquiring a customer. Considering F1-Score might be a good alternative in this case.

Model Choice

I started with a simple logistic model with base setting. Due to the unbalanced data, I chose to oversample the booking examples and run a cross validation over logistic and random forest regressors.

Due to the lack of time, I did not look deeply into the dataset itself and ran a brute force GV over the two models. While I aimed to build up a full pipeline, I am still not completely sure how to model one prediction per search ID.

Future Improvements:

Due to time limitations I was unable to build more than a basic grid search setup. With more time, I would prefer to build a better understanding of the dataset itself. Some additional work, I would like to look into:

- Effect of Check-in Month, Week
- Effect of long-weekends/holidays (append external data)
- Interactions between user and property features, e.g. domestic vs international search.
- Try controlling for recall. We can assume that showing a bad hotel on the top might lead to lower bookings. This will increase the cost of acquisition of a customer. Hence, a bad prediction can have a negative effect
- Improve the over sampling strategy
- Every search ID has a booking associated with it. The model shold aim to predict only one booking per search ID. (Run a separate model per property?)
- Derive within-data information on number of times the property was booked before the given search date. This might be an indicator of popularity.
- Build a dataset of country-region-continent mapping for misssing data.
- Better data imputation, especially for prices and room capacity.