Marketplace Machine Learning Data Science Take-Home Prompt

Background: Knowing whether a listing will be booked for a future date could provide huge business value to Airbnb. Therefore, we would like to predict this event. The dataset provided here contains a random sample of listings from three markets, San Francisco, Paris and Los Angeles.

- Each row in the dataset is a combination of a listing and a calendar night for which we try to predict if it will be booked, based on what we know 30 days prior to the calendar night. The calendar night is denoted by ds_night and ranges between 2015-01-01 and 2015-12-31.
- All the data in the dataset, except for column dim_is_requested, is current for 30 days before ds_night, denoted by ds (i.e., ds + 30 days = ds_night).
- `dim_is_requested` refers to whether or not the listing was ultimately requested a booking for the ds_night
- Listing, ds_night combinations that were already booked or are otherwise unavailable 30 days prior will not appear in the dataset

Features: The features can be categorized into the following groups:

- **Price**: price per night set on the calendar.
- **Listing attributes**: each home on Airbnb is unique on multiple dimensions, such as exact location, size, reviews, decor, etc. Users' willingness to pay may vary with these attributes.
- **Occupancy and availability** of the listing: these features are calculated looking at the status of the calendar.
- **Demand for a listing**: some listings are more popular from search results and generally attracts more views. This set of features are collected from the upper funnels of the booking process and represent interest from guests for a listing.
- **Demand and supply within the market**: aggregated features like number of searches and the number of contacts can depict the general demand for a market. The number of available listings represents supply.
- **Demand and supply within a KDT-Room type cluster**: we use a machine learning algorithm to cluster the listings that are close geographically, as an automated way to identify neighborhoods. Each cluster identified is called a KDT (k-dimensional tree) node, which contains 100 listings. Similarly as the market features, demand and supply are measured on this KDT node level which contains 100 listings for the same type of listings.

Your assignment: You have 48 hours to play with the data and tackle the problem using machine learning. The requirements are:

- Build a model to predict whether a listing will receive a booking request for a calendar night.
- Start with a baseline model that is more than a random guess and see how much you can improve from there.
- Show how you evaluate and improve your model performance. Explain your choice of evaluation technique.
- Using the provided dataset, derive additional features to demonstrate your data sense and creativity.
- Identify how you would use your model and findings to improve Airbnb's marketplace by writing out specific recommendations.
- What other prediction problems can be solved using this dataset? Suggest future work that could leverage this data.
- Please submit one document and provide code and a writeup (e.g. in R Markdown or iPython Notebook).
- In order to minimize unconscious bias in our review process, please don't include your name or any identifying personal details in your submission.