Algorithms Data Science Take-Home Prompt

Home Listing Recommender

Background: Knowing which home listings to recommend to a guest could provide huge business value to Airbnb. Therefore, we would like to train a recommender model that can predict which listings specific user is likely to book. The dataset provided here contains a random sample of our 7-day search log from two markets: Rio de Janeiro and Sao Paulo

Every time a user conducts a search they are shown certain number of listings that are available for the searched location, dates and search filters. Given the search results, the user can conduct one or more action on a specific listing: impress (no action), click (which takes user to listing page), contact host (to inquire about listing) and finally book.

Data Description
- Each row in the dataset is one of the listings that is a result of a search conducted by a user (identified with id_search)
- Each row has a label that tells us what is the ultimate action performed on the listing: impression, click, contact host or book. Keep in mind that we use the latest action as label. Therefore, if label is contact host, it means that before that user also did an impression and click. Or if the label is book, the user also did impression and click on the listing and may have contacted the host (if ds_contact is present) or may have just directly booked without contacting the host.
- Listings are uniquely defined using id_listing field in the dataset.
- Searcher (Booker) is uniquely defined using id_user field in the dataset

Features: The features can be categorized into the following groups:
- **Query Features**: (they start with query_* in the dataset) These are features related to user's search query (location, check-in & check-out dates and filters such as number of guests, etc)
- **Listing Features**: (they start with listing_* in the dataset) These are various attributes of the listings (such as price, review count, review rating, location, etc.)

Your assignment: You have 72 hours to play with the data and tackle the problem using machine learning to build a recommender system for a specific searcher. The requirements are:
- Build a model that will be able to recommend the most relevant (bookable) listings to users for the given search parameters.
- You can formulate the problem as a **ranking problem** or a **top-K recommendation problem** as long as you can justify your choice and test the recommendation model using applicable metics.
- 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. Since this is a recommendation/ranking problem that we are addressing, use at least one metric that tests how well you rank or recommend at top-K
- Using the provided dataset, derive additional features to demonstrate your data sense and creativity.
- Note that no user personalization features are provided in the dataset. Leverage id_user within the search data to derive at least one feature that captures users' historical preferences for use in your recommender.
- What consequences does your model have on new listings? Are they recommended enough? How would you change the recommendation model such that you optimize not only for bookings in general but for bookings of new listings as well> Demonstrate your approach and evaluate it using a metric of choice.
- Identify opportunities of using your model in Airbnb's marketplace. For what purposes could it be used?
- 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.

These candidate materials are confidential and candidates should not share them with anyone outside of Airbnb. The information, data and/or facts contained in these candidate materials are purely hypothetical and intended only for assessment purposes. All data included has been simulated for this exercise.

Explanation of features: Below find a table of original features provided to you. It is very useful to read the descriptions to understand the meaning behind each features. Reading the descriptions will also help you come up with derived features. Note that some of the features (such as dates) you can't directly use in modeling but you can use them to build very useful derived features.

Feature Name	Description
id_search	Unique ID of the search
label	Listing label (booked, contact host, clicked, impressed)
id_user	Unique ID of the user
id_listing	Unique ID of the listing
ds_search	date of the search
ds_book	date when listing was booked by user
ds_contact	date when host was contacted
query_market	market of user search (e.g. Sao Paulo)
query_checkin	searched checkin date
query_checkout	searched checkout date
query_num_guests	searched number of guest (filter)
query_num_children	searched number of children (filter)
query_num_infants	searched number of infants (filter)
query_radius	search query radius (map size)
query_price_max	maximum price search filter
query_price_min	minimum price search filter
query_center_lat	latitude of searched location center
query_center_lng	longitude of searched location center
listing_is_new	listing is new (has 0 reviews and bookings)
listing_total_price	total price of listing for selected dates
listing_instant_bookable	is listing instant bookable (possible to book without the need to first contact the host)
listing_lat	listing latitude
listing_lng	listing longitude
listing_review_rating	average review rating of listing given by guests (1 to 5)
listing_review_count	number of guest reviews
listing_property_type	property type id
listing_room_type	room type
listing_num_beds	number of beds
listing_num_bedrooms	number of bedrooms
listing_num_bathrooms	number of bathroom
listing_person_capacity	how many guests listing can host (set by host)
listing_has_pro_pictures	if listing has pro photos
listing_num_recent_reservations	number of recent reservation
listing_location_rating	average location rating (given by guests)
listing_cleanliness_rating	average cleanliness rating (given by guests)
listing_checkin_rating	average checkin rating (given by guests)
listing_value_rating	average value rating (given by guests)

listing_communication_rating	average communication rating (given by guests)
listing_accuracy_rating	average accuracy rating (given by guests)
listing_num_books_90day	number of bookings in last 90 days
listing_occupancy_rate	listing occupancy rate (what fraction of nights get booked)
listing_monthly_discount	if listing has monthly discount
listing_weekly_discount	if listing weekly discount provided by host
listing_cleaning_fee	cleaning fee
listing_monthly_price_factor	monthly discount price multiplier
listing_weekly_price_factor	weekly discount price multiplier
listing_minimum_nights	minimum nights allowed by host
listing_maximum_nights	maximum nights allowed by host