Project 2

STA 695-234 group

University of Kentucky

03/08/2018

Data Exploration

- Training set: 37670293
  - Missing value:
    orig_destination_distance 13525001
    srch_ci 47083
    srch_co 47084
- Test set: 2528243
  - Missing value:
    orig_destination_distance 847461
    srch_ci 21
    srch_co 17

Data Exploration Cont.

Is imputation needed?

- Test set only contains is_booking=1
- Training set contains is_booking=1 (3000693, 7.96%) but is_booking=1 AND non-NA (1985514, 0.05%)
- Also, test set contains a lot of missing values too!

Clean Data

- Remove wrong dates: date_time, srch_ci, srch_co; dates beyond 2050
- Remove is_booking=0
- Fill in srch_ci and srch_co
  - Fill in srch_ci: date_time + avg_diff, where avg_diff is the average difference time between date_time and srch_ci (2 ways: user-specific and overall)
  - Fill in srch_co: srch_ci + avg_stay, where avg_stay is the average stay time (2 ways: user-specific and overall)

Clean Data Cont.

- Fill in orig_destination_distance
  - 4 keys: user_location_country, user_location_region, user_location_city, srch_destination_id
  - Average distance by group: some missing in a group
  - Overall average distance: all missing in a group

Leakage Problem

- Problem: All user_id in the test set can be found in the training set
- Exact match:
  - Use the 6 keys, user_location_country, user_location_region, user_location_city, hotel_country, hotel_market, orig_destination_distance, to match the training and test sets.
  - Predict using the hotel cluster in the training set.
  - Reuslts:
    - 29% of the test set can be exactly matched in the training set.
    - The score for the exact match method is only 0.23. (some hotels might belong to different clusters in different seasons)

Leakage Problem Cont.

- Multiple match: (including exact match)
  - Use the same 6 keys mentioned above
  - Predict with multiple hotel clusters (closer date will have higher weights)
  - 33% of the test set can be multiple matched in the training set.
- Popular match:
  - user_id most popular
  - user_location_city most popular
  - srch_destination_id most popular
  - etc.

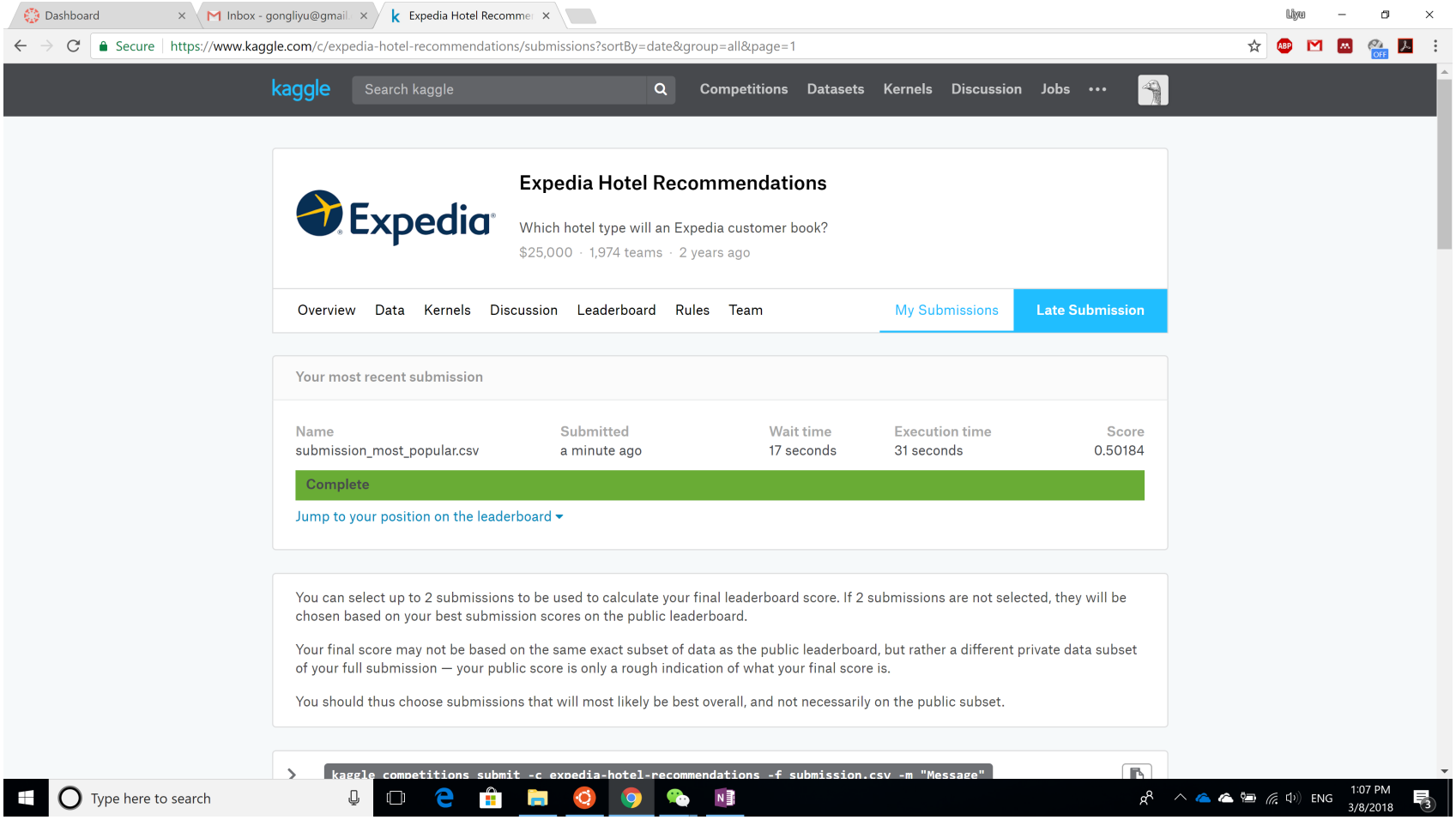
Machine Learning

- Random Forest and Xgboost
  - Feature engineering:
    New features: days of stay, days to checkin, checkin month, checkin day, book month, book day, book hour, weighted booking, destination
- SGD and Naive Bayes
  - Feature engineering
    Features: user_id, user_location_city, srch_destination_id, srch_destination_type_id, hotel_continent, hotel_country, hotel_market, is_mobile, is_package, season of destination, days to checkin for destination

Results

Table 1: Results

Method	Score
Exact Match	0.23082
Exact + Multiple Match	0.29690
Exact + Most Popular	0.50184
Random Forest	0.30528
Xgboost	0.29930
SGD	0.26065
Naive bayes	0.17375
Blended	0.30385
Combined	0.49820


Dashboard
Inbox - gongliyu@gmail
Expedia Hotel Recomme
Secure
https://www.kaggle.com/c/expedia-hotel-recommendations/submissions?sortBy=date&group=all&page=1
kaggle
Search kaggle
Competitions
Datasets
Kernels
Discussion
Jobs
Expedia Hotel Recommendations
Which hotel type will an Expedia customer book?
$25,000 · 1,974 teams · 2 years ago
Overview
Data
Kernels
Discussion
Leaderboard
Rules
Team
My Submissions
Late Submission
Your most recent submission
Name
submission_most_popular.csv
Submitted
a minute ago
Wait time
17 seconds
Execution time
31 seconds
Score
0.50184
Complete
Jump to your position on the leaderboard
You can select up to 2 submissions to be used to calculate your final leaderboard score. If 2 submissions are not selected, they will be chosen based on your best submission scores on the public leaderboard.
Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.
You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.
kaggle competitions submit -c expedia-hotel-recommendations -f submission.csv -m "Message"
Type here to search
ENG
1:07 PM
3/8/2018

Future Work

What can we do better?

- Use Field-aware Factorization Machine to do feature engineering.
- Impute distance by mapping location to sphere surface.

Thank you !