Expedia Hotel Recommendations

Team 5

Michael, Hamza, Christian

Berlin, 15th November 2025

Context

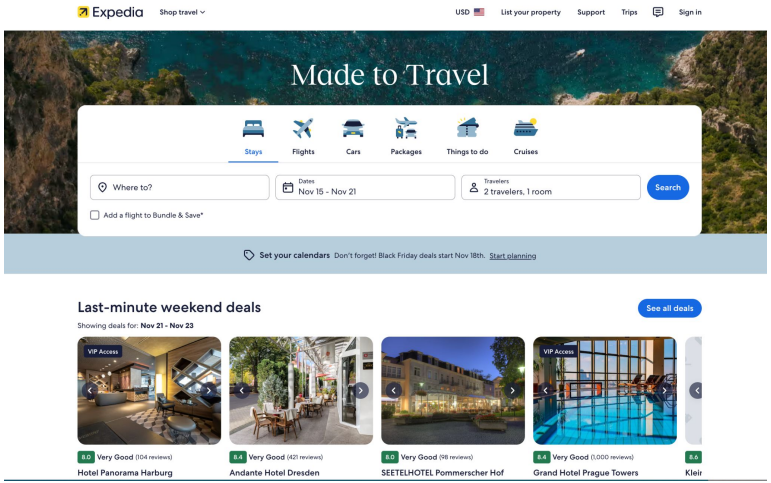
This Kaggle competition challenged us with predicting what hotel a user will book based on some attributes about the search the user is conducting on Expedia.
Expedia is interested in predicting which hotel group a user is going to book.

Dataset:
Train data: 37.670.293 observations, 2013-2014
Test data: 2.528.243 observations, 2015
Sample Submissions
Destinations

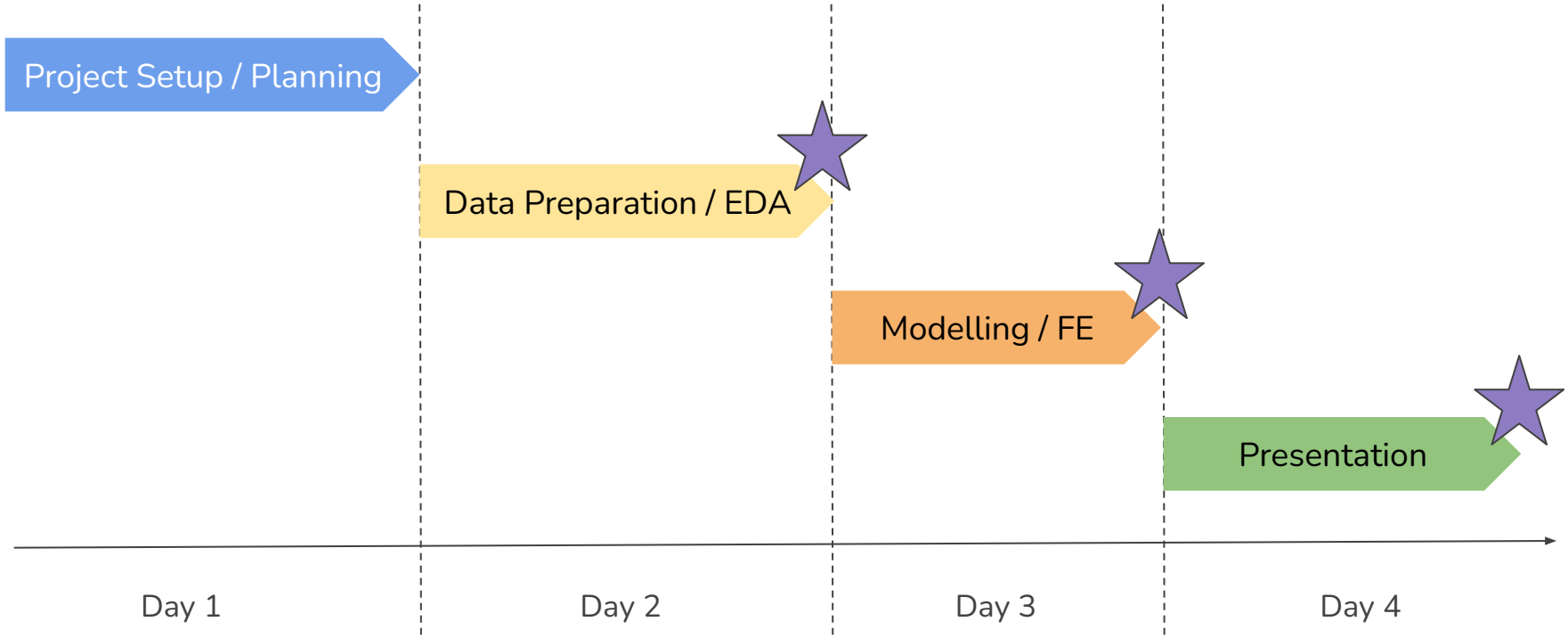
Overall data: >4.50 GB

Target for prediction:
hotel_cluster

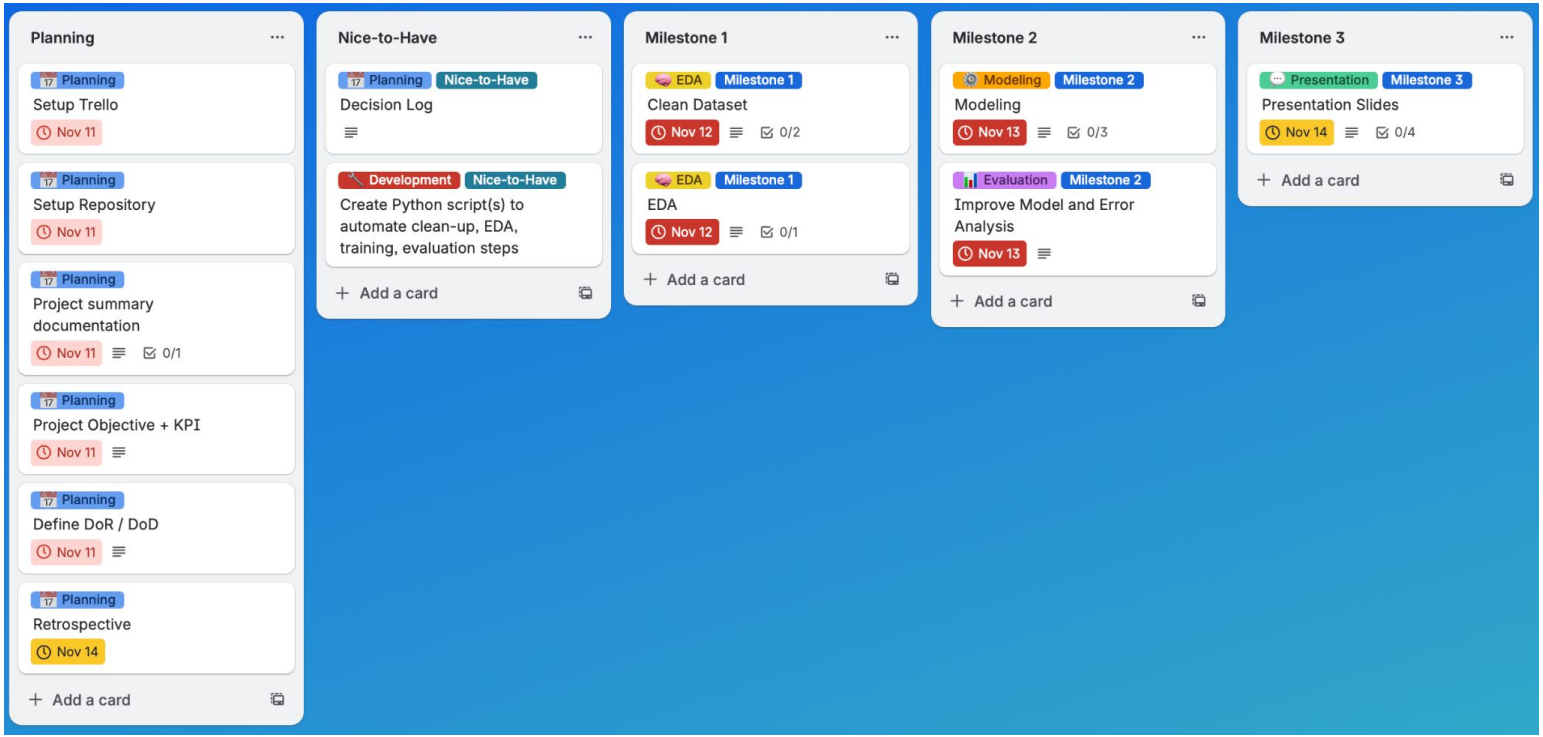
Submissions are evaluated according to the Mean Average Precision @ 5 (MAP@5)



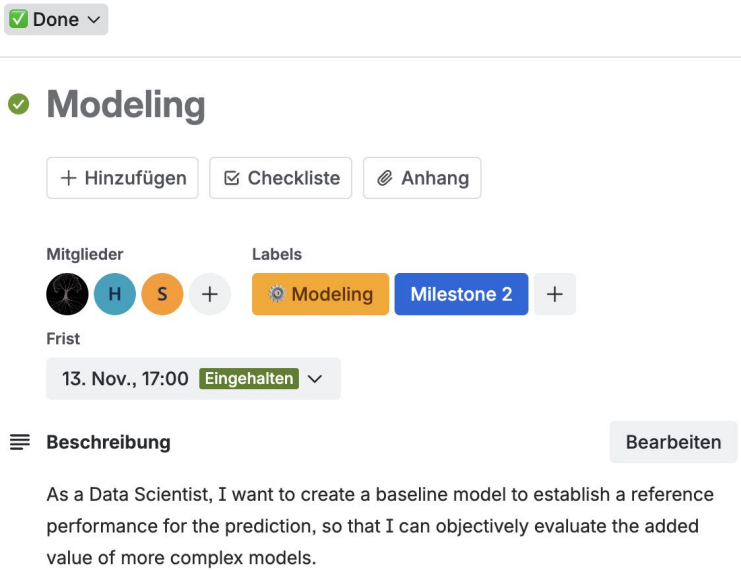
Roadmap


Project Setup / Planning
Data Preparation / EDA
Modelling / FE
Presentation
Day 1
Day 2
Day 3
Day 4

Methodology 1/2



Methodology 2/2



Acceptance criteria:

1. A simple baseline model (e.g., predicting the likelihood of a user booking a hotel cluster) is implemented.
2. The baseline model runs successfully on the prepared dataset without errors.
3. Evaluation metrics (e.g., accuracy score, confusion matrix, classification report, mean squared error, MAP@5 — depending on the task) are calculated and documented.
4. The model's performance is clearly reported as a benchmark for future models.
5. Code and results are reproducible (fixed random seeds, local development environment).
6. Assumptions, limitations, and chosen metric rationale are clearly explained.
7. Results are stored or visualized in a way that enables easy comparison with later models.

Methodology 2/2

DoR / DoD

DoR:

- The Card is well defined (Title, Description, Sub-tasks)
- Acceptance criteria are clearly defined and agreed upon
- The Card can be completed within one day
- Dependencies are identified and managed

DoD:

- All acceptance criteria are met
- Code has been peer-reviewed and merged into the main branch
- The team has reviewed and accepted the story

Objectives & KPIs

Objective
Predict the likelihood of a user booking a hotel cluster.

KPI
Submissions are evaluated according to the Mean Average Precision @ 5:

$$MAP@5 = \frac{1}{|U|}\sum_{u=1}^{|U|}\sum_{k=1}^{min(5,n)} P(k)$$

Why MAP@5 instead of Accuracy?

Because multiple hotel clusters can be plausible for each event. MAP@5 measures ranking quality rather than only top-1 correctness.

Team/Roles

1. Christian - Superman
2. Michael - Ironman
3. Hamza - Batman

Scope

Create a machine learning model to predict the probability that a user will book a specil

Objective

Predict the likelihood of a user booking a hotel cluster.

KPI

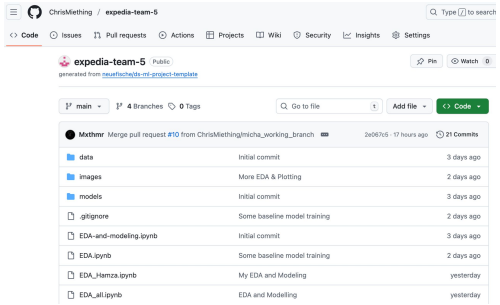
MAP@5 (Mean Average Precision @ 5):

$$MAP\ @\ 5 = \frac{1}{|U|}\sum_{u=1}^{|U|}\sum_{k=1}^{\min(5,n)} P(k)$$

1. Number of User Events |U|
2. Precision at Cutoff P(k)
3. Number of Predicted hotel Clusters (n)

Deliverables

- up to 5 predictions for each user event
- Presentation



Cleaning & EDA 1/2

Understanding the data

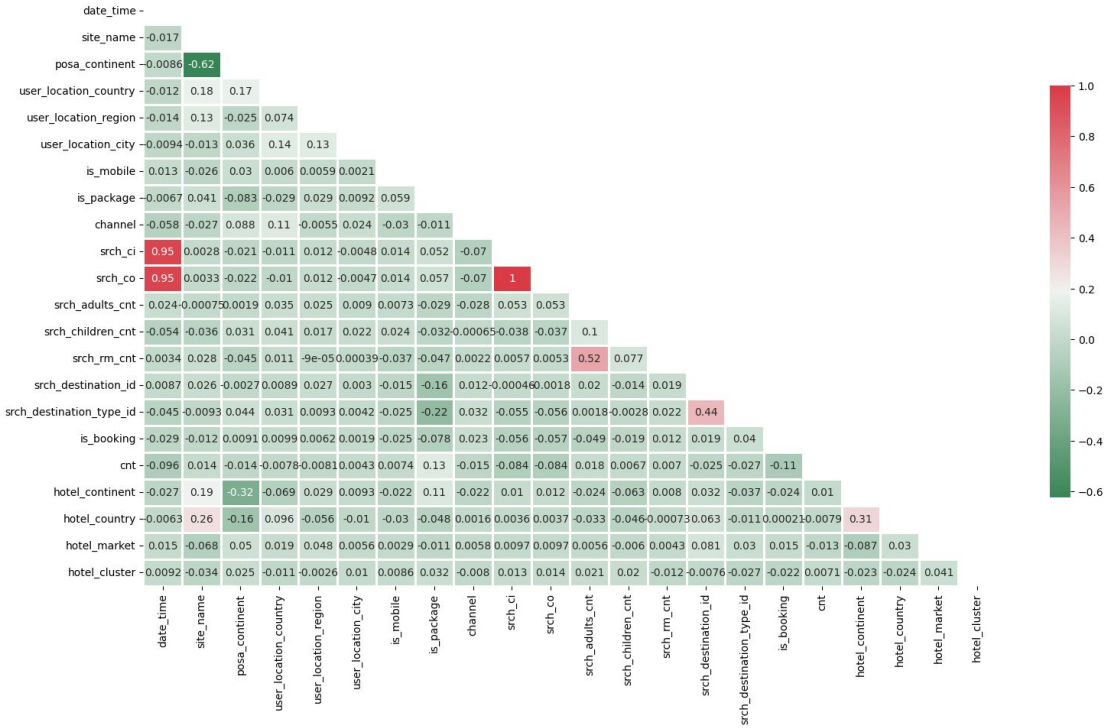
- Basic descriptions
- Categorical variables with a large number of distinct values
- Most columns are integers or floats -> feature engineering is limited

First steps working with the data

- Chunking the data
- No dropping of columns (except orig_destination_distance)
- Handling datetime columns
- Dropping rows with missing values
- User ids in test.csv are a subset of the user ids in train.csv
- No columns correlate linearly with hotel cluster

Cleaning & EDA 2/2

Heatmap for train data



Modeling (Regression) - Baseline

Features: Using all Columns

Baseline Model: Logistic Regression

Data Size: 20000

Accuracy Score	0.0296
MSE	2271.3013
MAP@5	0.0585

Modeling (Classification) - Baseline

Features: Using all Columns

Baseline Model: LGBM Classifier

Data Size: 20000

Accuracy Score	0.0248
MSE	1595.2278
MAP@5	0.0317

Feature Engineering to Improve Model

Separate Categorical Columns

Date Conversion from String to DateTimeStamp

Dropping rows with NaN

Calculation on Date Columns to add new Features

Used hotel_cluster as the target

Models (Regression)

Models Used: Linear Regression, Random Forest, XGBoost Regressor

	MSE	Accuracy	MAP@5
Linear Regression	819.615482	0.01600	0.0282
RandomForest	740.459459	0.02600	0.0470
XGB Regressor	715.417368	0.01625	0.0335

Since the goal is to calculate MAP@5 to evaluate the best model, regression is not the best way to get the top 5 recommendations.

Regression is the wrong modeling approach for this classification and ranking problem.

Model Improvement (Classification) 1/3

Baseline Model: **LGBM Classifier**

Data Size: 25000

Accuracy Score	0.0126
MSE	1566.4619
MAP@5	**0.0254, 0.2673**

Model Improvement (Classification) 2/3

Baseline Model: **CatBoost Classifier**

Data Size: 10000

Accuracy Score	0.1325
MSE	1582.5790
MAP@5	**0.2126**

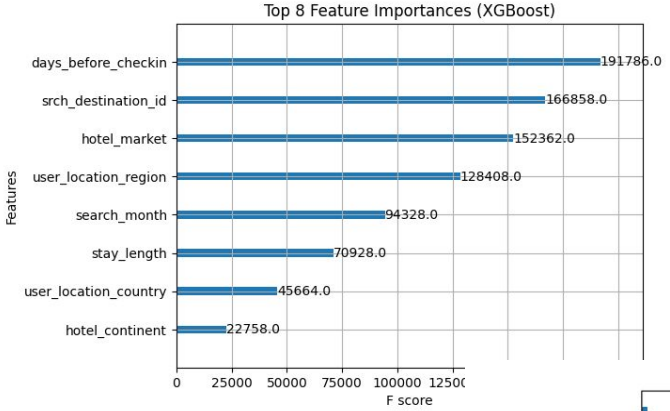
Model Improvement (Classification) 3/3

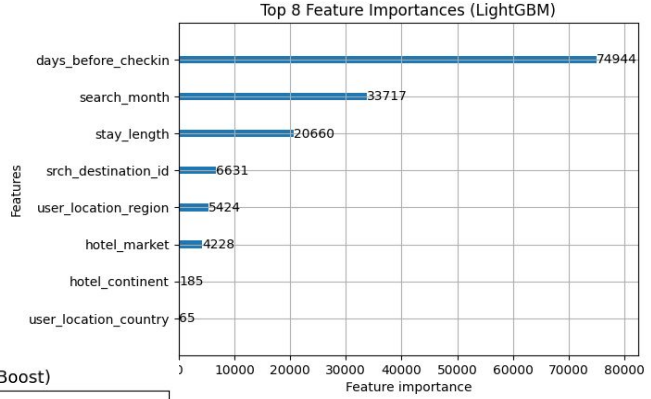
Baseline Model: **XGBoost Classifier**

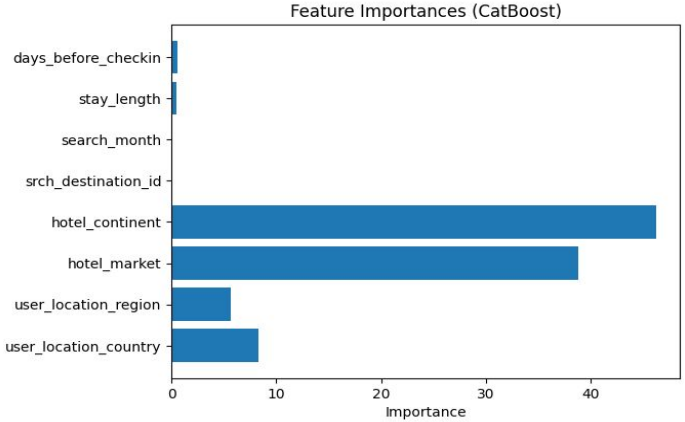
Data Size: 25000

Accuracy Score	0.3228
MSE	1119.1175
MAP@5	**0.4288**

Results


Top 8 Feature Importances (XGBoost)
days_before_checkin
srch_destination_id
hotel_market
user_location_region
search_month
stay_length
user_location_country
hotel_continent
191786.0
166858.0
152362.0
128408.0
94328.0
70928.0
45664.0
22758.0
Features
0
25000
50000
75000
100000
F score


Top 8 Feature Importances (LightGBM)
days_before_checkin
search_month
stay_length
srch_destination_id
user_location_region
hotel_market
hotel_continent
user_location_country
74944
33717
20660
6631
5424
4228
185
65
Features
10000
20000
30000
40000
50000
60000
70000
80000
Feature importance


Feature Importances (CatBoost)
days_before_checkin
stay_length
search_month
srch_destination_id
hotel_continent
hotel_market
user_location_region
user_location_country
0
10
20
30
40
Importance

Results

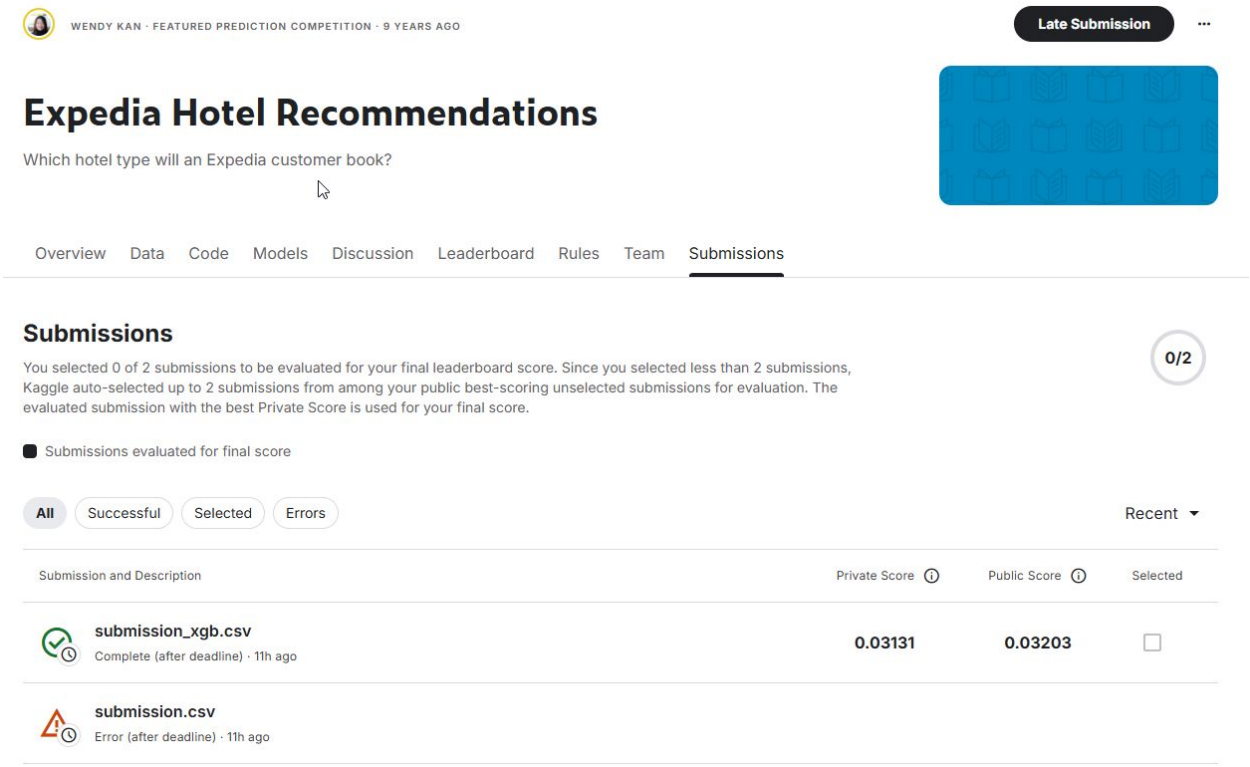
Best Model Based on the MAP@5 comparison: **XGBoost Classifier**

The top 5 scoring clusters per booking are submitted as the solution.

Submission:

id	hotel_cluster
1	1 51 24 45 49
4	80 0 26 17 96
5	80 0 26 17 96
8	41 70 98 21 10
10	41 70 98 21 10

Results


WENDY KAN · FEATURED PREDICTION COMPETITION · 9 YEARS AGO
Late Submission
Expedia Hotel Recommendations
Which hotel type will an Expedia customer book?
Overview
Data
Code
Models
Discussion
Leaderboard
Rules
Team
Submissions
Submissions
You selected 0 of 2 submissions to be evaluated for your final leaderboard score. Since you selected less than 2 submissions, Kaggle auto-selected up to 2 submissions from among your public best-scoring unselected submissions for evaluation. The evaluated submission with the best Private Score is used for your final score.
0/2
Submissions evaluated for final score
All
Successful
Selected
Errors
Recent
Submission and Description
Private Score
Public Score
Selected
submission_xgb.csv
Complete (after deadline) · 11h ago
0.03131
0.03203
submission.csv
Error (after deadline) · 11h ago

Thanks for listening, Questions?