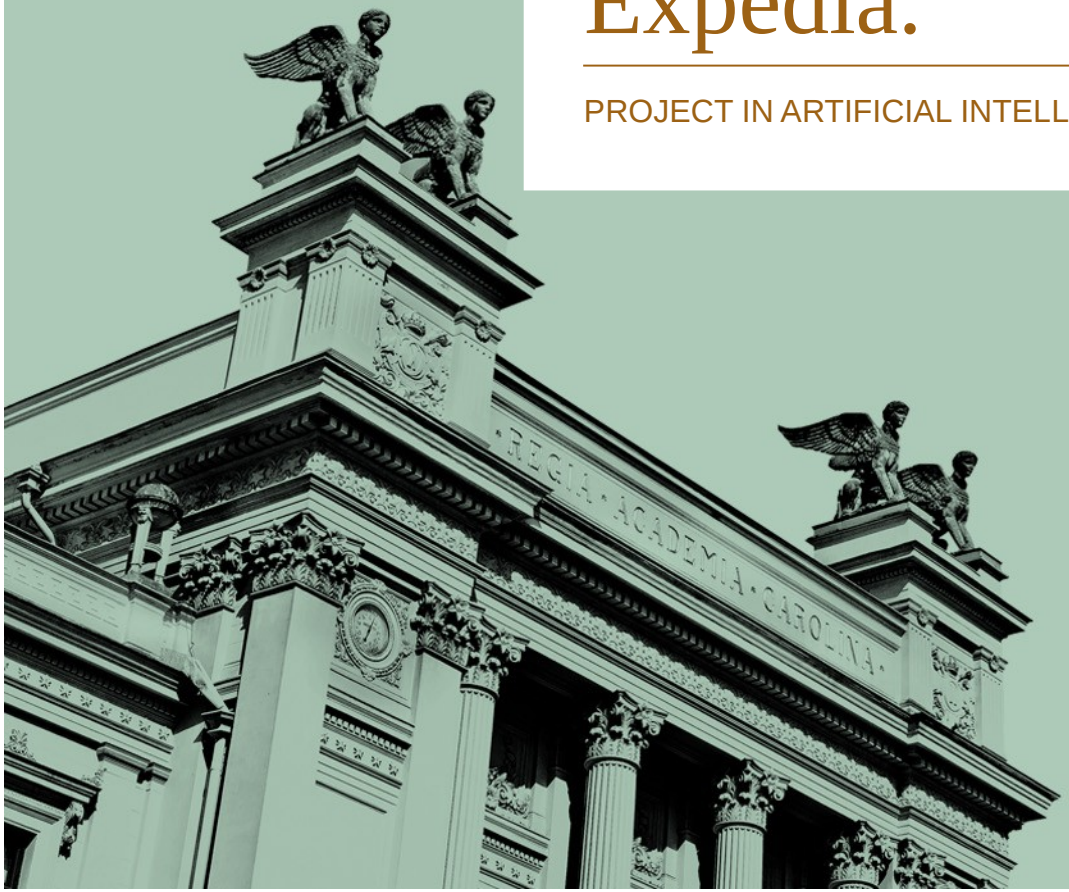
LUND UNIVERSITY

Classification of hotels for Expedia.

PROJECT IN ARTIFICIAL INTELLIGENCE - EDAN70



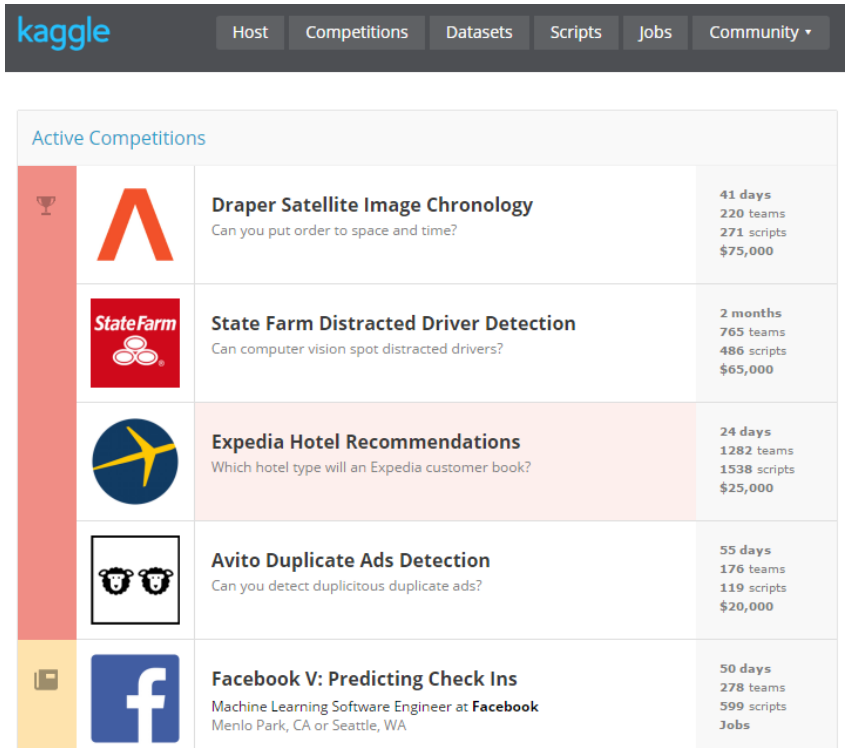
Introduction

- Who we are.
- Kaggle.com
- Our main problem. Expedia
- Random Forest Classifier.
- Expedia and workflow.
- Conclusions.

LUND UNIVERSITY

Kaggle.com

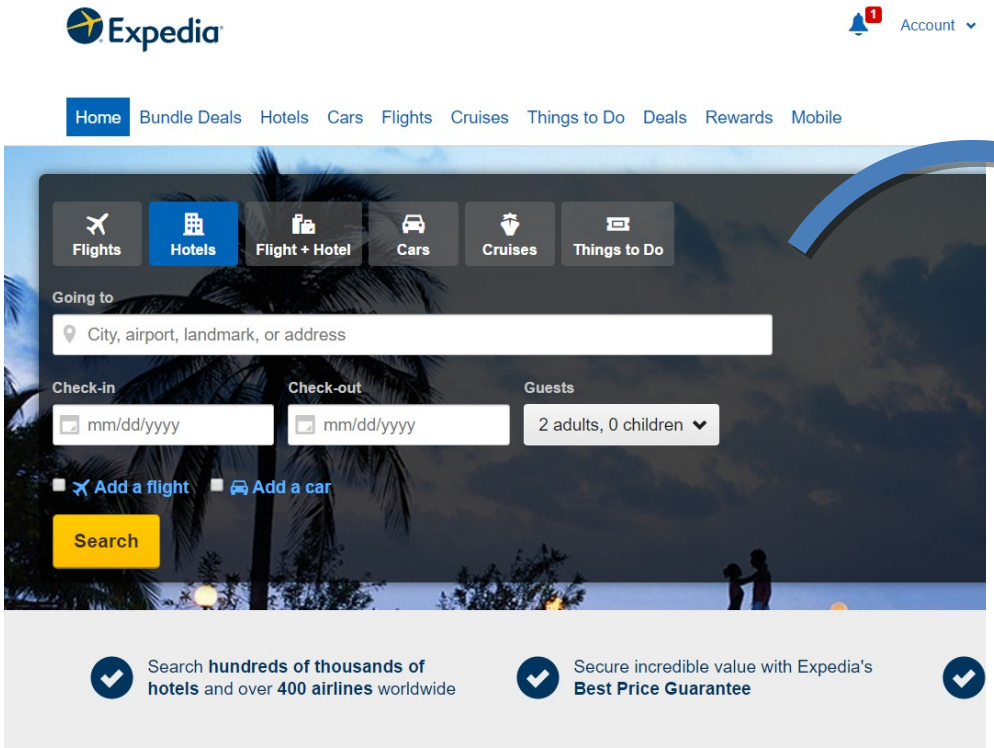
- Users from all over the world compete to produce the best machine learning models.
- Submissions, Scripts, Leaderboards.

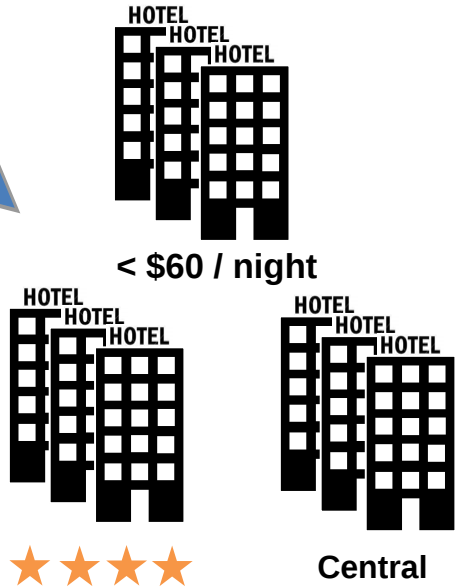


LUND UNIVERSITY

Expedia

- The problem – Expedia.





LUND UNIVERSITY

Ho Chi Minh City: 530 hotels

Sort By: Price | Guest Rating | Vacation Rentals | Hotel Name | Hotel Class | Recommended | More

45% booked

Ho Chi Minh City is a popular location on your dates.
Hint: you can get a lower price on other dates
Try searching one week later

Search by hotel name

Find what you want faster, filter by price or neighborhood.
Got it.

Hotel avg	3 star avg	4 star avg	5 star avg
$81	$69	$131	$307

Filter hotels by

- NEW! Free Cancellati

Hotel Class

- ★★★★★ 5 Stars (18)
- ★★★★ 4 Stars (52)
- ★★★ 3 Stars (215)
- ★★ 2 Stars (222)
- ★ 1 Star (23)

Price Per Night

- Less than $75 (95)
- $75 to $124 (20)
- $125 to $199 (5)
- $200 to $299 (2)
- Greater than $300 (1)

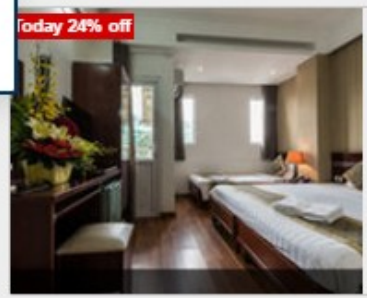
Search Nearby

- Opera House
- War Remnants Museum
- Saigon Square
- Diamond Plaza
- Reunification Palace

Show more

Neighborhood

- Ho Chi Minh City (and vicinity)
- Ben Thanh Market
- Bui Thi Xuan
- Consulate Area
- District 3



Saigon Odyssey 1 Hotel ★★
Pham Ngu Lao Map
1-866-264-5744 • Expedia Rate ✔ Free Cancellation

Good! 3.6/5
(28 reviews)
We have 5 left at
~~$46~~ $35
avg/night
Sale!
Hurry! Offer ends in 06:12:47
Earn 2,248 points

Saigon Domaine Luxury Residences ★★★½
#5 Guest Rated
Ho Chi Minh City Map
1-866-267-9053 • Expedia Rate

Wonderful! 4.5/5
(152 reviews)
~~$198~~ $149
avg/night
Sale!
Earn 9,620 points

Get an extra 10% or more off select hotels with Member Pricing!
Email | Sign up | Already a member? Sign In
You agree to receive deals and offers from Expedia, and may unsubscribe at any time

Lan Lan Hotel 2 ★★★
#73 Guest Rated
Ben Thanh Market Map
1-866-272-4856 • Expedia Rate ✔ Free Cancellation

Good! 3.8/5
(166 reviews)
In high demand!
We have 2 left at
~~$75~~ $52
avg/night
Sale!
Earn 6,130 points

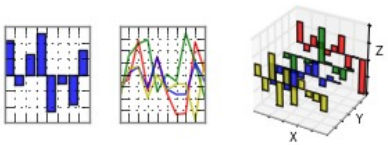
Tools

- Python 64-bit
- Alot of RAM

- Pandas – parsing data into data structures
- NumPy – scientific computing package
- Scikit learn – Machine Learning library, built on SciPy, NumPy and matplotlib

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



scikit learn

NumPy

LUND UNIVERSITY

Expedia - dataset

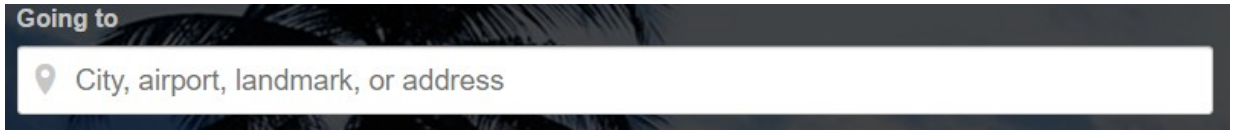
- 24 column in training
- 22 column in testing (no is_booking, no hotel_cluster)
- Most of the columns are integers or floats
- Output hotel cluster ID integer range from 1-99

	date_time	site_name	posa_continent	user_location_country	user_location_region	user_location_city	...
0	2014-08-11 07:46:59	2	3	66	348	48862	...
1	2014-08-11 08:22:12	2	3	66	348	48862	...
2	2014-08-11 08:24:33	2	3	66	348	48862	...
3	2014-08-09 18:05:16	2	3	66	442	35390	...
4	2014-08-09 18:08:18	2	3	66	442	35390	...

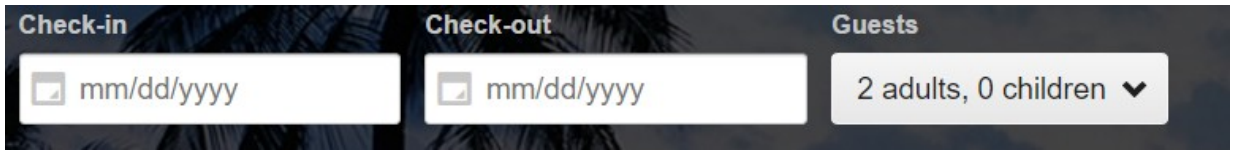
LUND UNIVERSITY

Expedia - workflow

- Understanding dataset



srch_destination_type_id, hotel_continent, hotel_country, and hotel_market



srch_ci srch_co are filled with dates
srch_adults_cnt, srch_children_cnt, and srch_rm_cnt is number of guests and rooms

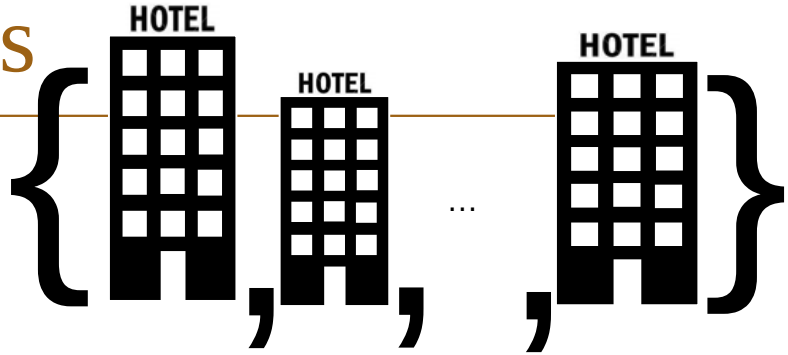
Add a flight maps to the is_package field

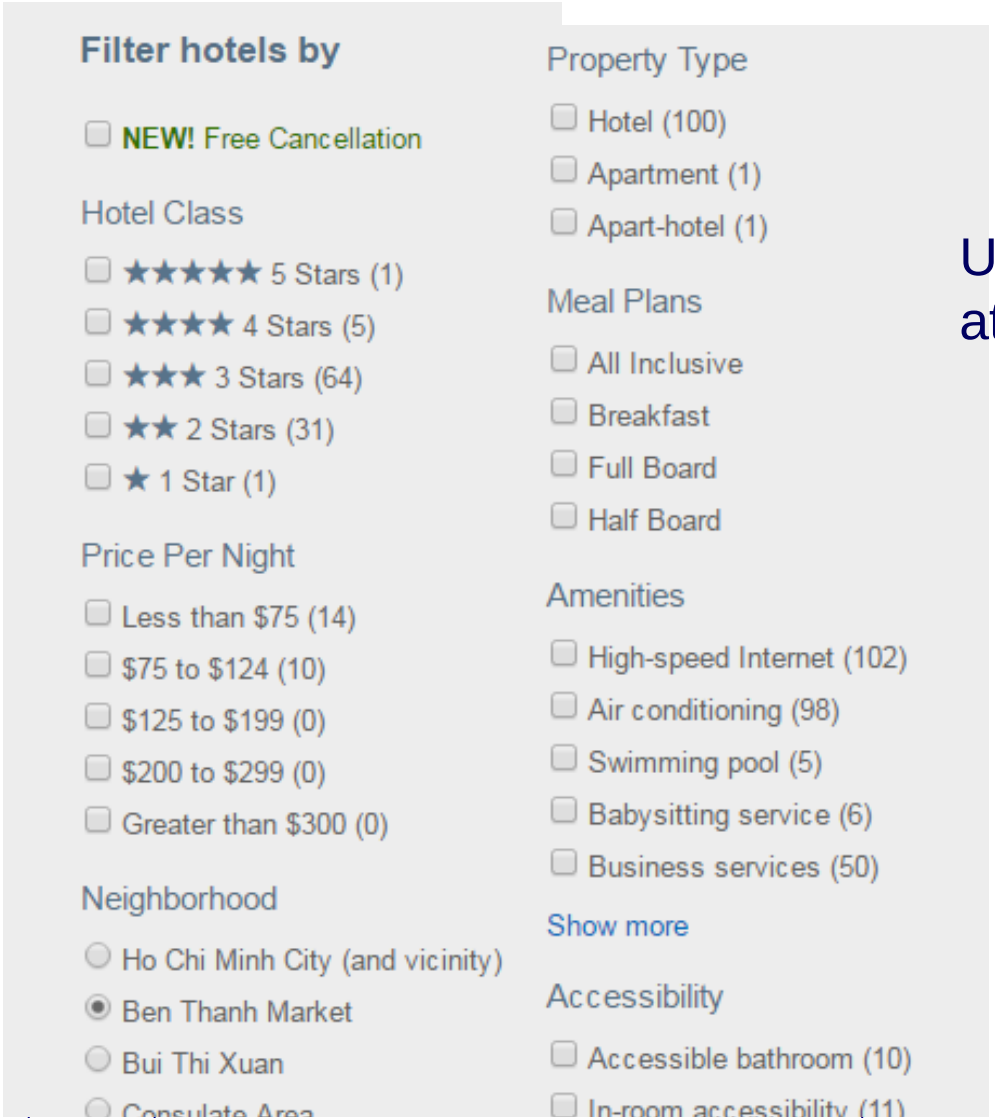
site_name – Expedia point of sale (Expedia.com, Expedia.se, ...)

posa_continent – ID of continent associated with site_name

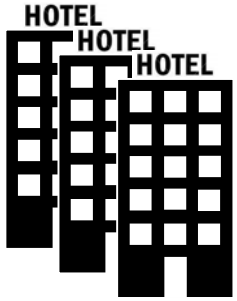
LUND UNIVERSITY

Expedia – Hotel Clusters

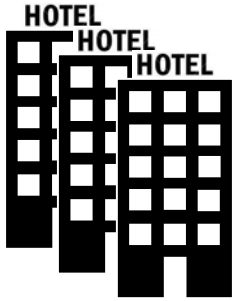




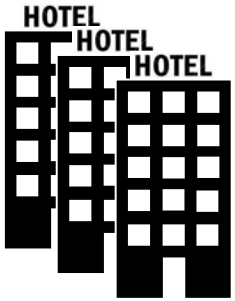
Useful! Expedia can much quickier at an earlier stage filter the hotels



< $60 / night



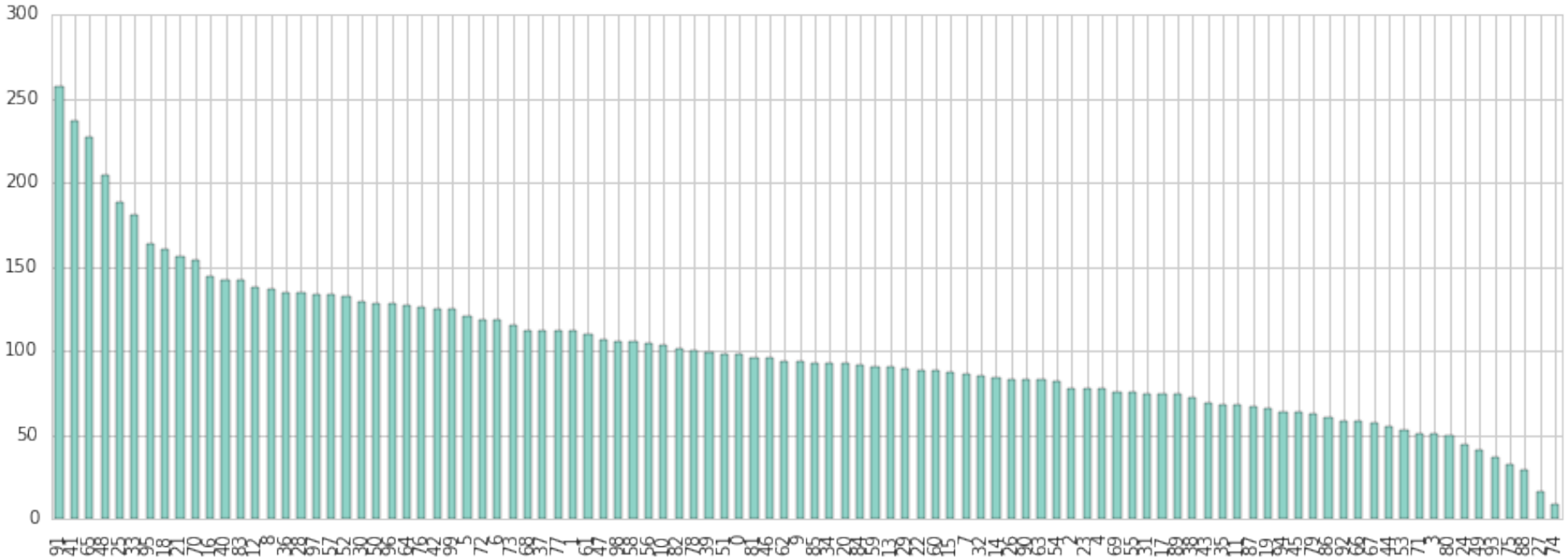
★★★★



Central

LUND UNIVERSITY

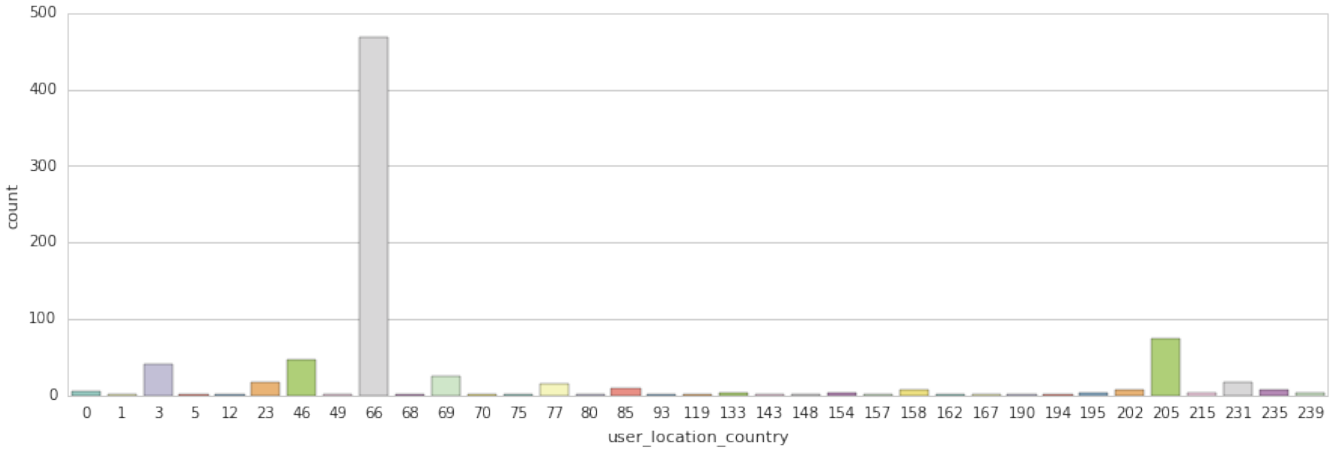
Expedia – most frequent hotel clusters

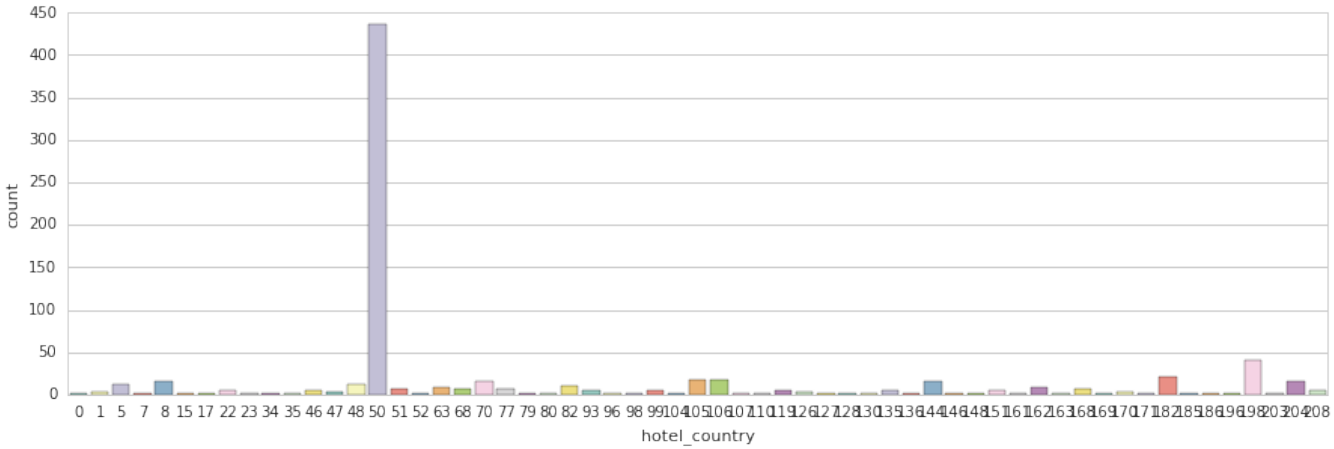


LUND UNIVERSITY

Expedia – examining features

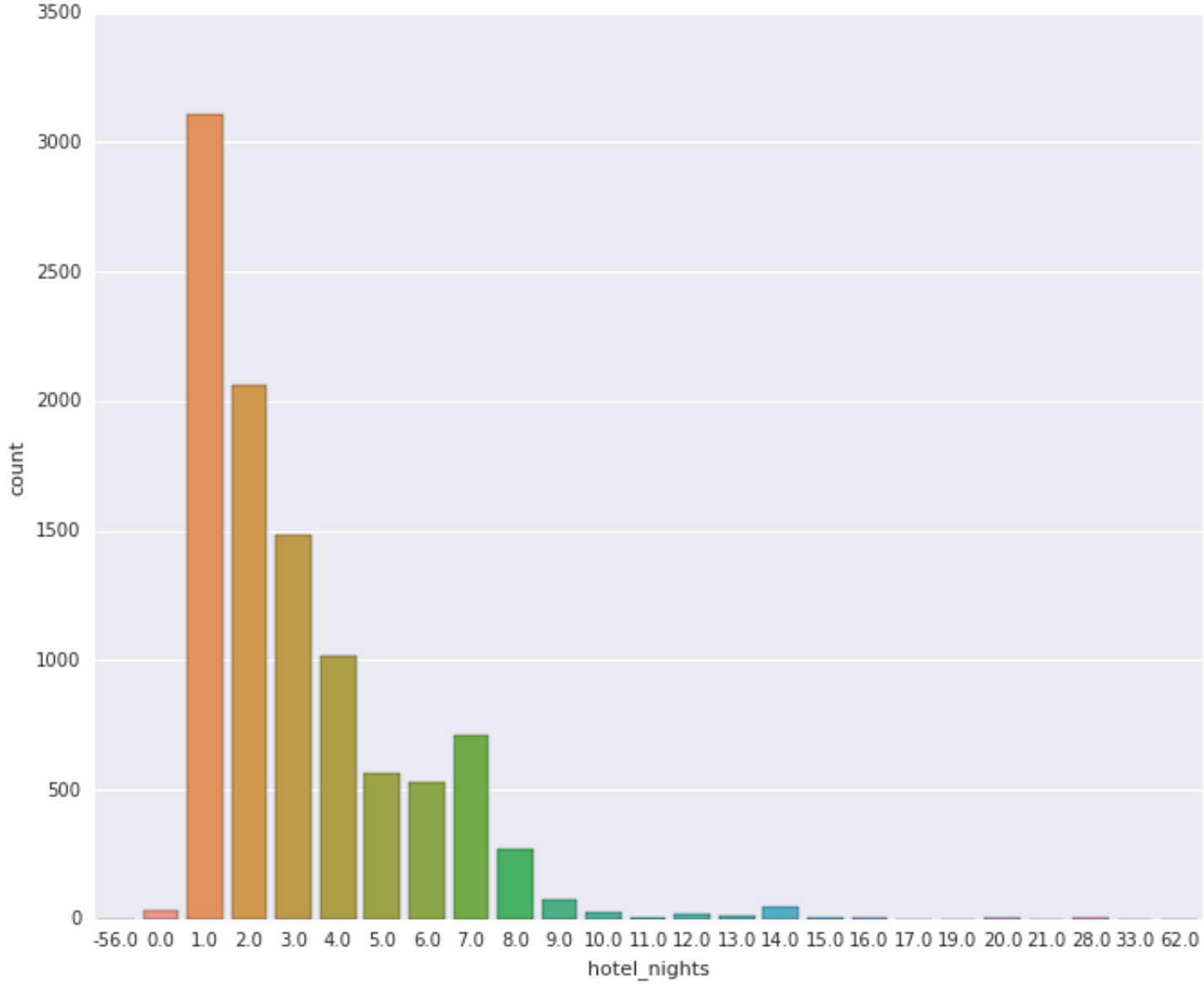
- What are the most countries the customer travel from/to?





LUND UNIVERSITY

- Nights of stay



LUND UNIVERSITY

Random Forest Classifier

- Supervised learning classifier – Uses bagging methods.
- Random sub-samples.
- Generates decision trees on each sub-sample.

feature A of the 1st sample

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$

feature B of the Nth sample

Create random subsets

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{B15} & f_{C15} & C_{15} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} \quad S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{B6} & f_{C6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

Decision tree 1

$$S_M = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & C_4 \\ f_{A9} & f_{B9} & f_{C9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

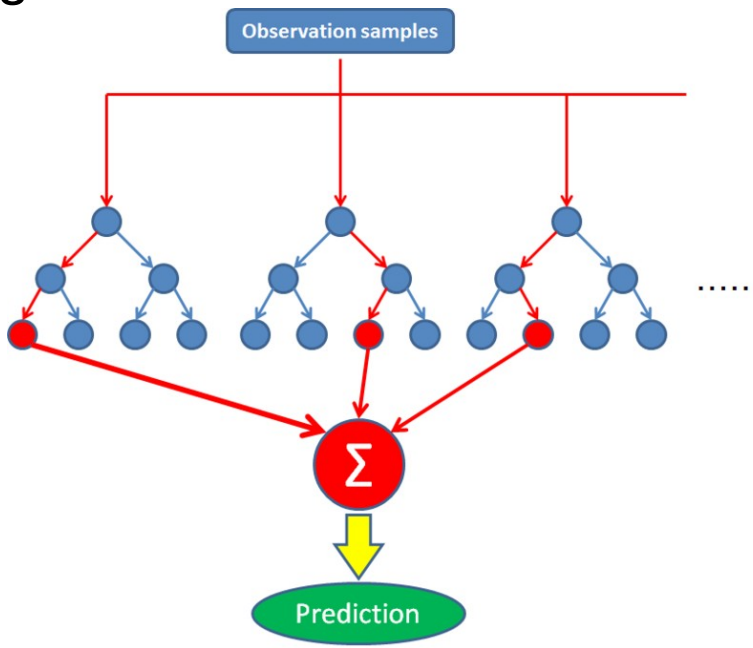
Decision tree 2

Decision tree M

LUND UNIVERSITY

Random Forest Classifier

- Sum all the decision trees.
- Mistakes are taken care of.
- The classifier corrects decision trees habit of overfitting to their training set.

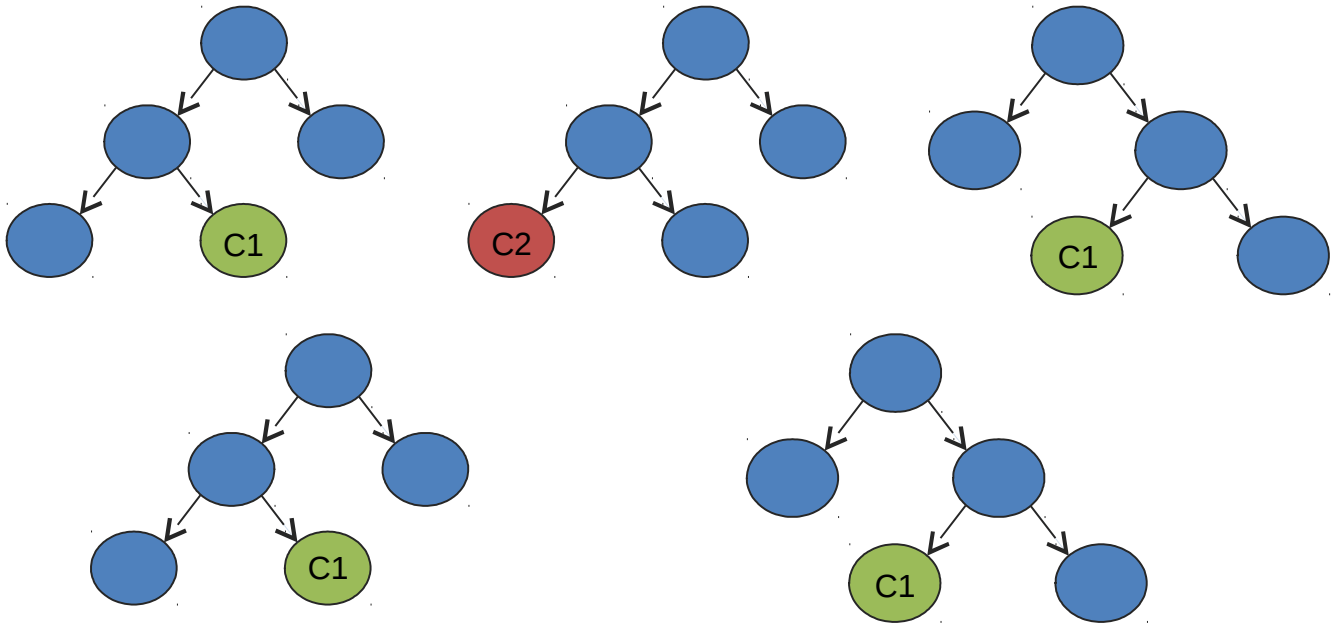


LUND UNIVERSITY

Random Forest Classifier

- Why does Random Forest work?

1. Most trees provide correct predicition for the most part of the data.
2. Trees make mistake at different place.



LUND UNIVERSITY

Expedia – How good is the classifer?

- We predict 5 hotel clusters for each sample in test.csv
- The evaluation function is Mean Average Precision @ 5

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

where |U| is the number of user events, P(k) is the precision at cutoff k, n is the number of predicted hotel clusters.

Test0 : Truth is 1, Predicted [1,2,3,4,5] => Average precision = 1

Test1 : Truth is 2, Predicted [1,2,3,4,5] => Average precision = $\frac{1}{2}$

Test2 : Truth is 5, Predicted [1,2,3,4,5] => Average precision = $\frac{1}{5}$

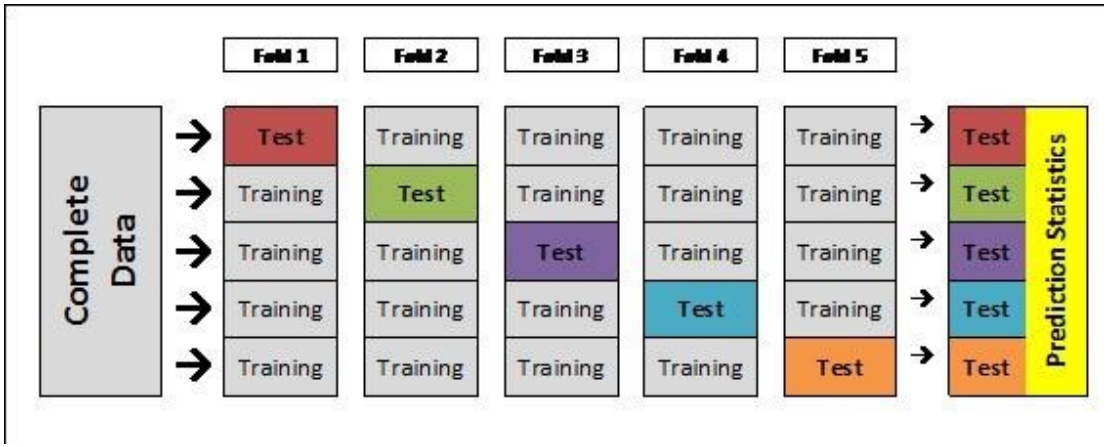
Test3 : Truth is 6, Predicted [1,2,3,4,5] => Average precision = 0

Mean average precision = 0.425

LUND UNIVERSITY

Expedia – How good is the classifer?

- k-fold cross-validation for model tuning



- We could more easily tune the model with a Grid Search for the best parameters

LUND UNIVERSITY

Expedia - Results

- Results with Random Forest classifier:

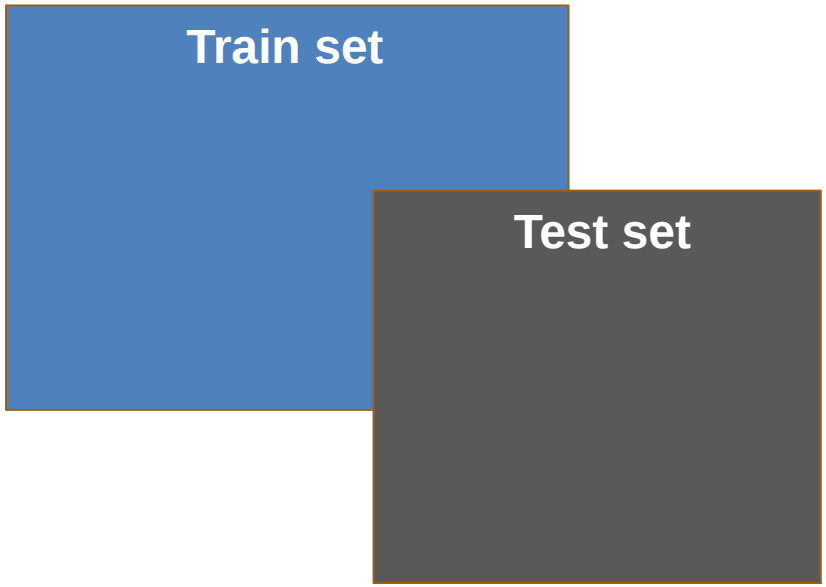
0.18584

- Results with most popular local hotels:

0.30090

LUND UNIVERSITY

Leakage



- user_location_country, user_location_region, user_location_city, hotel_market and orig_destination_distance

LUND UNIVERSITY

Leakage - Results

- Using a more advanced approach with most popular hotels and leakage we got:

0.50050

LUND UNIVERSITY

Expedia - Conclusion

- Machine learning can be used in real-life situations to optimize a product or service
- It is very important to not leak training examples into the test set because the model will overfit
- Here the best model will have to find the leak (1/3) and train itself to catch the rest of the holdout data (2/3)

LUND UNIVERSITY

LUND
UNIVERSITY