Lessons from Kaggle: RentHop

2017-07-15, Eric Doi

Problem: Predict Listing Popularity

Favorites Find Roommates Post Rental Login

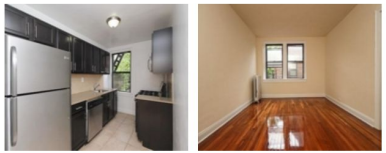
dio 1 2 3 4+ Lft Rm $ 1000 to $ 2000 Filters SEARCH

Updated - July 1, 2017

Upcoming Open Houses

Sat, Jul 1	9:00am - 1:00pm	$1,324	1BR, 1BA at 1541 Metropolitan Avenue
Sat, Jul 1	9:00am - 1:00pm	$1,289	1BR, 1BA at 1565 Odell Street
Sat, Jul 1	10:00am - 6:00pm	$1,500	1BR, 1BA at 82-17 Myrtle Avenue

« Back | Page 1 of 93 (1,849 Rentals) | Next » Sort: HopScore | Price



100 1BR, 1BA at Skillman Avenue

Sunnyside, Long Island City, Northwestern Queens, Que...

$1,995 Per Month By Matthew Viola

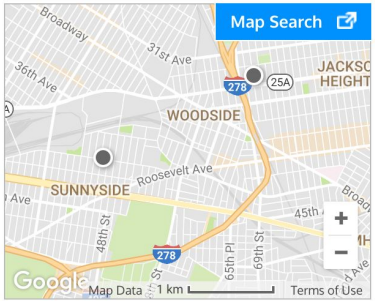
Check Availability

Posted 18 mins ago

100 1BR, 1BA at BRIGHTON 11TH

Queens

$1,850



New York City, NY

New York City is the world's greatest city. Of course, that is open to debate in many parts of the world, but among the residents long and new, there is simply no contest. Whether you seek gourmet dining, legendary Broadway shows, luxury brands, or more humble interests; you'll certainly find what you are looking for somewhere within the five boroughs of NYC.

Input / Target / Evaluation

Input:

- Main Data (< 150K rows)
  - *Primary Key*: listing_id
  - Price
  - # Bedrooms, Bathrooms, etc.
  - building_id
  - manager_id
  - Latitude / Longitude
  - Tags, Descriptions, etc.
- Photos (80 GB)
  - Multiple photos per listing

Target:

interest_level: ‘low’ / ‘medium’ / ‘high’

Evaluation:

Output: P(low), P(medium), P(high)

Metric: Multi-class Log Loss

Data Example

```
In [10]: raw_train_df.head(2).transpose()
```

Out[10]:

	10	10000
bathrooms	1.5	1
bedrooms	3	2
building_id	53a5b119ba8f7b61d4e010512e0dfc85	c5c8a357cba207596b04d1afd1e
created	2016-06-24 07:54:24	2016-06-12 12:19:27
description	A Brand New 3 Bedroom 1.5 bath ApartmentEnjoy ...	
display_address	Metropolitan Avenue	Columbus Avenue
features	[]	[Doorman, Elevator, Fitness Cent
interest_level	medium	low
latitude	40.7145	40.7947
listing_id	7211212	7150865
longitude	-73.9425	-73.9667
manager_id	5ba989232d0489da1b5f2c45f6688adc	7533621a882f71e25173b27e313

Trying AWS

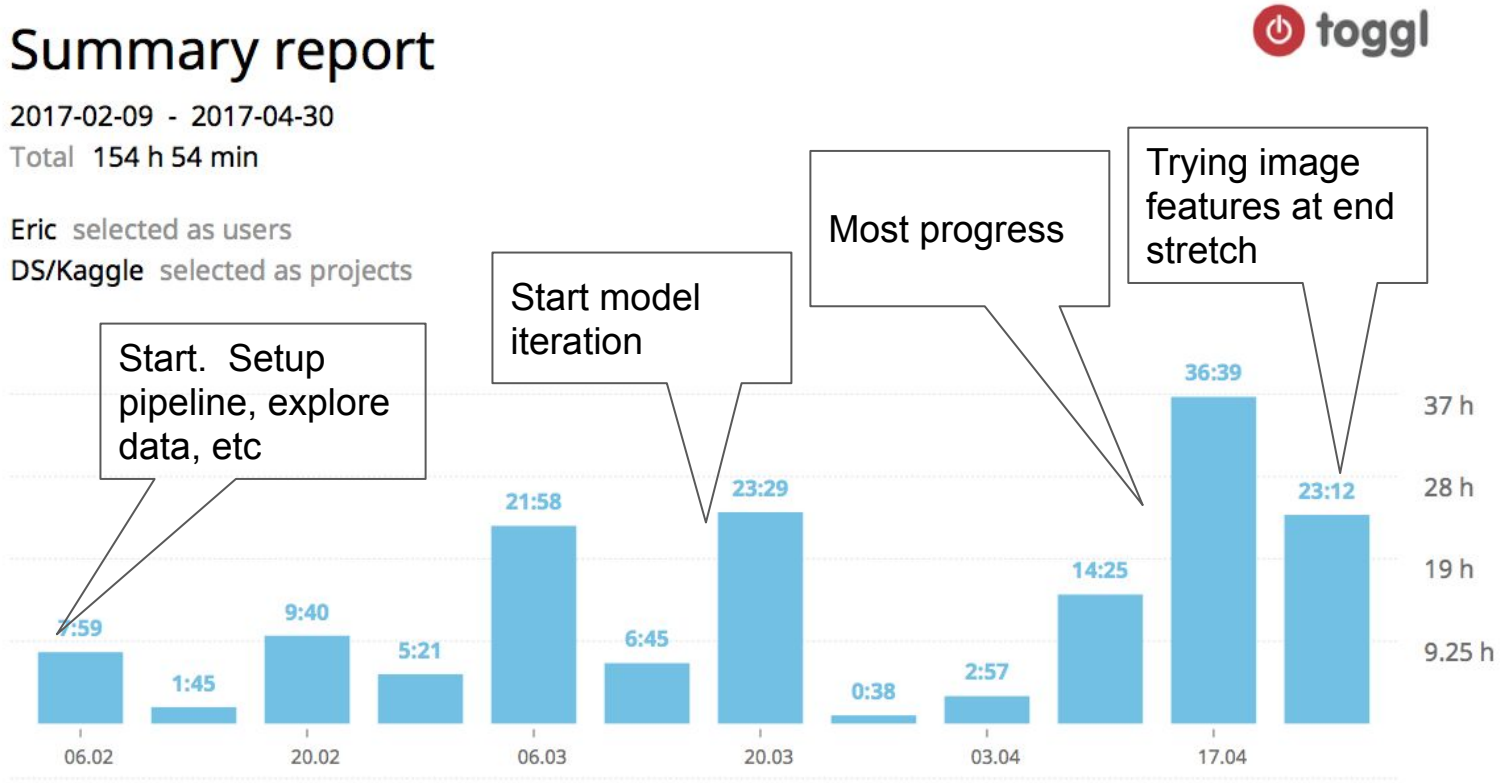
+ **Cheap**! Only **$3** for c4.2xlarge (8-core, 15GB) * 27 hours
- **Troublesome**. Install libraries, check out repo, mount storage drive, *torrent* the 80GB of images...

- Tip: Using Ubuntu image instead of Amazon Linux is much simpler for package installation

Keeping a Log

- **Fixed f4b…**
  - Try shallow train:
    - xgb_10000_3_0.3_f4b: CV 0.5364, n_best 318: (LB: 0.54362, overfitting ~0.006 still)
  - Change concats back to outer…??? Nope, same.
  - Try revert to bad-order commit: c9ffc7e027725f9d5ba1a18be4a36b7e8404a09e
    - OMG, that was it. Back to 0.5328. Why??
  - **Was it because the re-ordering increased variation? Try removing random seed…**
    - Yes! xgb_10000_3_0.3_f4b: CV 0.5331, n_best 288 (LB 0.54143)

Where Did the Time Go?



Solution Study: Plantsgo

Plantsgo: 1st Place Winner

- Best ensemble: 0.492
  - Compared to my best ensemble: 0.512
- Best single xgb: 0.503 (top ~50)
  - Compared to my best single xgb: 0.52

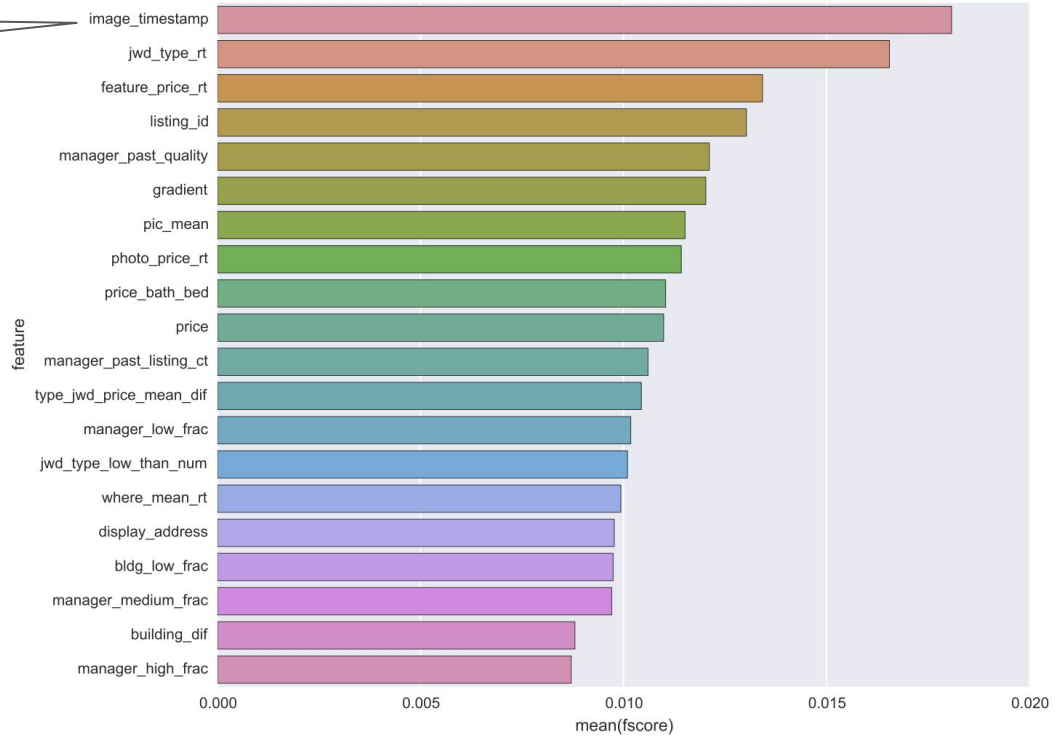
Feature Engineering++

Added over 100+ hand-crafted features

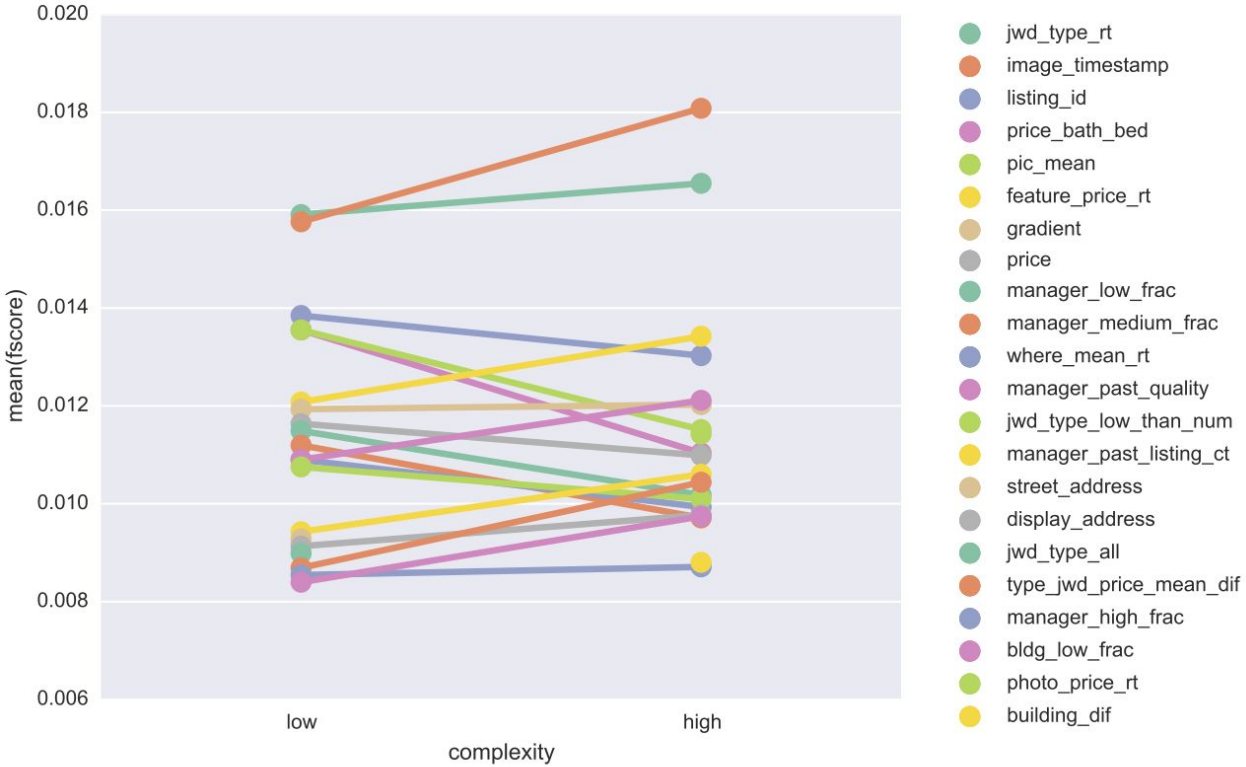
- “jwd”: Local-area features
  - E.g., jwd_type_low_than_num: The number of competing listings of the same type (lat/long/bath/bed/price) with lower price
- “Gradient”: position-in-listing-ids / position-in-days. Increases when lots of listings are added in a short time & v.v.
- Manager-related:
  - E.g. manager_pay_jwd: avg of fangxing_mean_rt_jwd for this manager (~this manager is in pricey areas on average~)
- 1 Image-related feature, pic_mean: avg pixel ct of images
- Building-related
- Combinations

Top 20 Features -- Plantsgo

“Magic” leak feature



Feature Changes with Model Complexity -- Plantsgo



Github Repo

https://github.com/ericdoi/kaggle-renthop-2017