Two Sigma RentHop Competition

Matthew Emery (@lstmemery)

June 1st, 2017

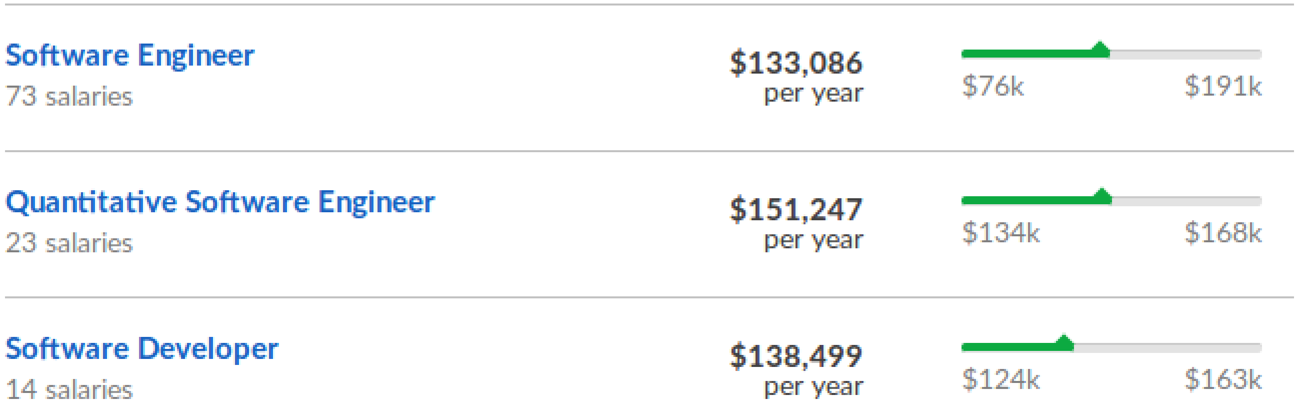
Winning Kaggle Competitions by KazAnova

1. Understand the Data
2. Understand the Metric
3. Cross-Validate Early!
4. Hyperparameter Tuning

Source

Who are Two Sigma and RentHop?

- Two Sigma: AI Heavy New York Hedge Fund
- RentHop: Smart Apartment Search (New York Only)
- Reward: Recruitment to Two Sigma

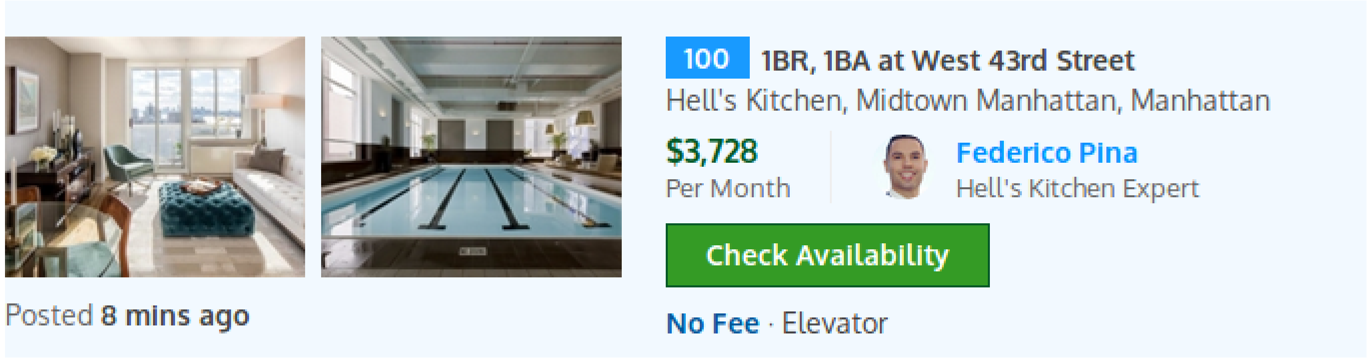


Software Engineer 73 salaries	$133,086 per year	$76k	$191k
Quantitative Software Engineer 23 salaries	$151,247 per year	$134k	$168k
Software Developer 14 salaries	$138,499 per year	$124k	$163k

Source

The Goal

- Predict how interested people will be in this:



100 1BR, 1BA at West 43rd Street

Hell's Kitchen, Midtown Manhattan, Manhattan

$3,728
Per Month

Federico Pina
Hell's Kitchen Expert

Posted 8 mins ago

No Fee · Elevator

Understanding the Data

Training: 49352 Rows

Test: 74659 Rows

- Location Data
- Natural Language Data
- Image Data (78.5 Gb compressed)
- ...and everything you would else you would expect (price, bedrooms etc.)

renthop

2σ TWO SIGMA

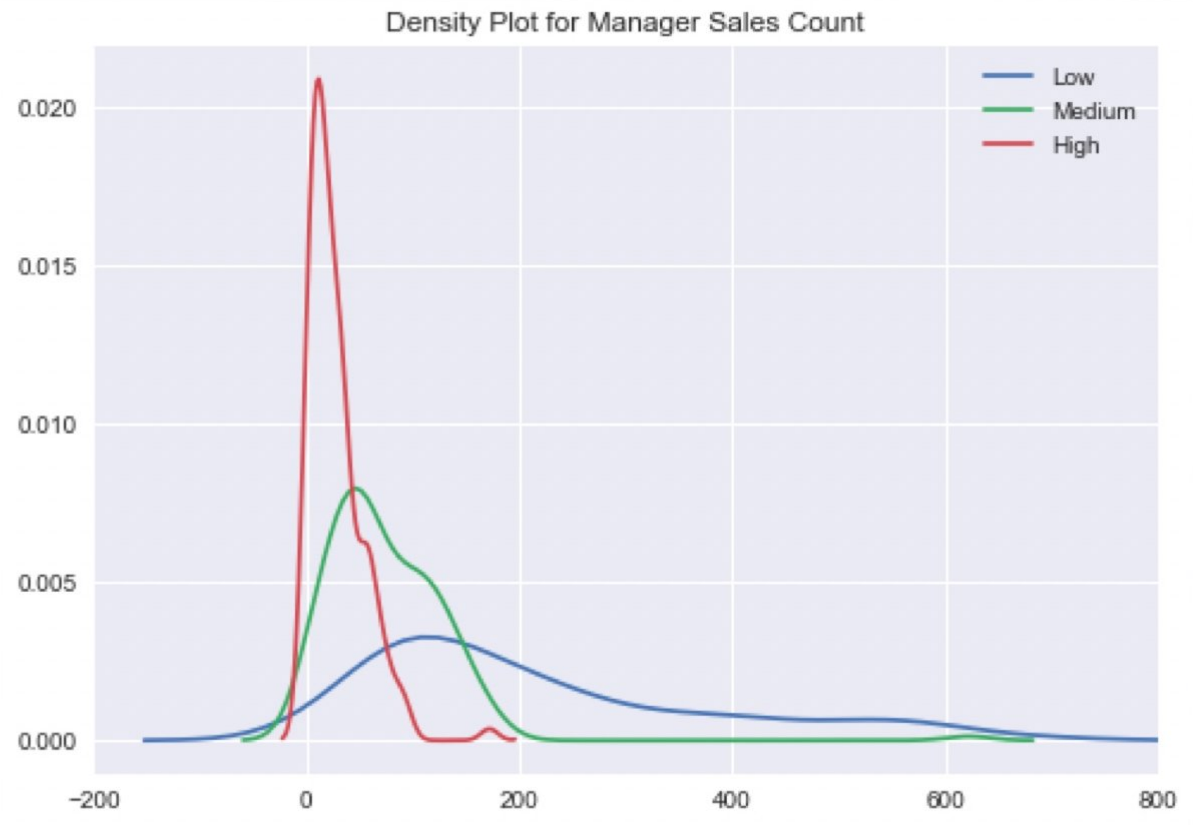
Understand the Metric

Multiclass Log Loss (Low, Medium, High Interest)

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

- Note: This isn't ordinal

Manager ID Count



Someone just used different transformations of Manager ID Count and scored in the top 15%

Source

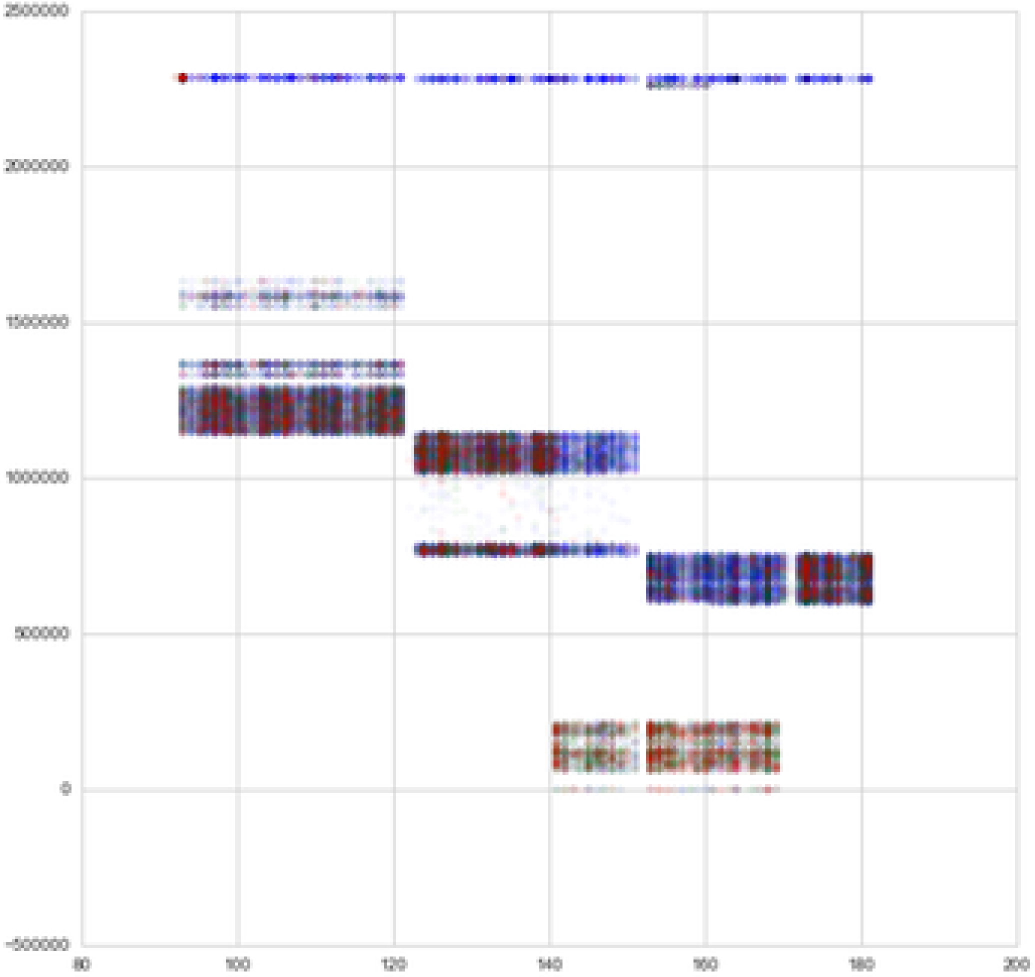
Listing ID

- This pattern hinted at a possible data leak...

Source

Data Leak

The creation time of the image folders were correlated with interest.



- X-Axis: Day
- Y-Axis: Seconds
- Blue=Low
- Green=Medium
- Red=High

Explanation

Feature Engineering

A few interesting ones:

- Grouping by categorical features and finding count/median/mean/standard deviation of numerical ones. (3rd Place)
- Inferring Points of Interest from text descriptions (Supermarket, Subway, etc.) (2nd Place)
- Leveraging duplicate data (Leads and lags on pricing) (11th Place)
- Exclamation marks in description
- Reverse GeoCoding New York Neighbourhoods

Second Place Solution

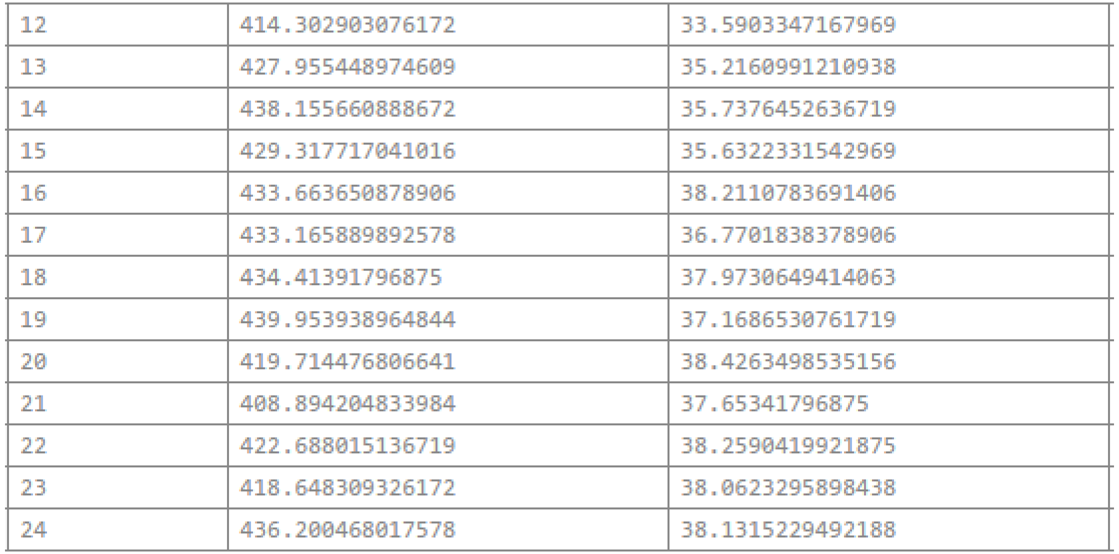
@Faron

```
- 32 LightGBM models
- 9 Extreme Tree models (sklearn)
- 7 RF models (sklearn)
- 5 Keras models
- 3 XGBoost models
- @KazAnova's StackNet example base-level predictions
```

Best Model: LightGBM (CV: 0.50135/ Test: 0.50557)

Meta-modeled with a 2-layer neural network.

An Aside on LightGBM



12	414.302903076172	33.5903347167969
13	427.955448974609	35.2160991210938
14	438.155660888672	35.7376452636719
15	429.317717041016	35.6322331542969
16	433.663650878906	38.2110783691406
17	433.165889892578	36.7701838378906
18	434.41391796875	37.9730649414063
19	439.953938964844	37.1686530761719
20	419.714476806641	38.4263498535156
21	408.894204833984	37.65341796875
22	422.688015136719	38.2590419921875
23	418.648309326172	38.0623295898438
24	436.200468017578	38.1315229492188

- Faster than XGBoost
- Requires more hyperparameter optimization

Second Place Solution

Grid-Search Bagging

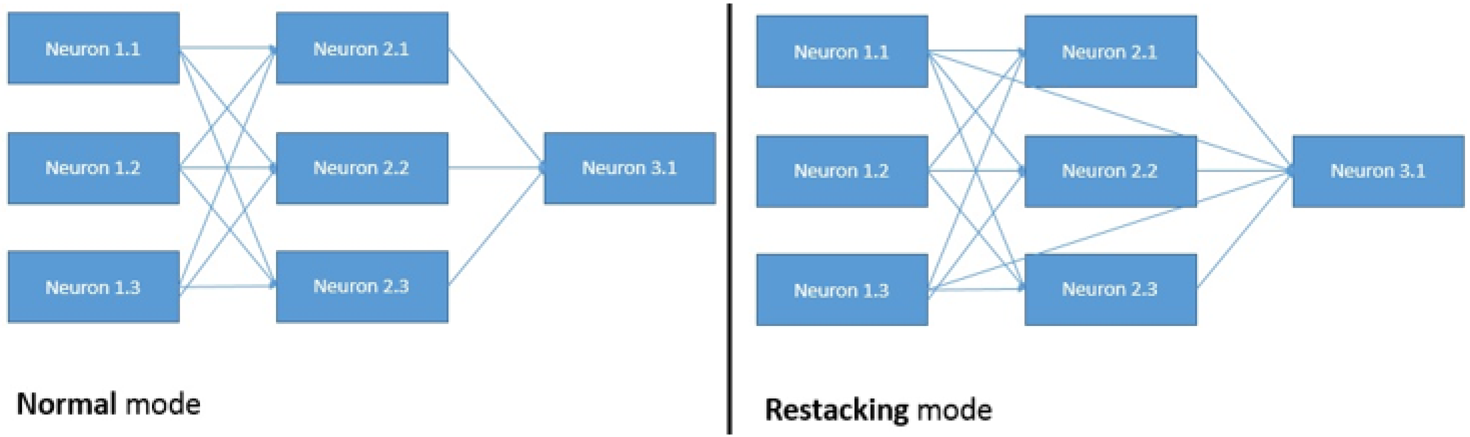
Grid Search: Check cross-validation scores for each hyperparameter in regular intervals. e.g. Check maximum depth of XGBoost from 1 to 10.

Bagging (Bootstrap AGGregating): Sample the data many times, with replacement

For each of 12 bags: Grid search hyperparameters If the new hyperparameters is better, blend it into the model

StackNet

Written by Marios Michailidis (kazAnova) for his PhD A Java-based, flexible meta-modelling network



Source

References

2nd Place Solution