NYC Apartments: Analyzing Rental Listings

Prashant Tatineni

Problem Statement

Understanding the Dataset

Feature Engineering

Classification Modeling Methods

Submitting a prediction in kaggle

Next steps

Can we predict the popularity of a rental listing?

- **Motivation and Assumptions**
  - Rental listings should receive more inquiries to be more profitable for rental owners.
  - More inquiries means that the listings are of higher quality.
  - Higher quality listings benefit the rental market as a whole.

- **Dataset from current Kaggle competition**
  - *"How much interest will a new rental listing on Renthop receive?"*
  - Actual listing data from April through June 2016
  - Interest level = High, Medium, or Low

kaggle renthop TWO SIGMA

Problem Statement

Understanding the Dataset

Feature Engineering

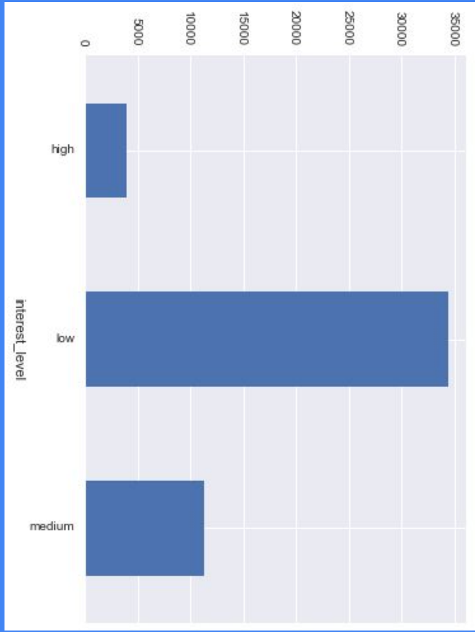
Classification Modeling Methods

Submitting a prediction in kaggle

Next steps

Understanding the dataset

- **One target variable:**
  - *interest_level = High, Medium, Low*



- **ID columns:**
  - *listing_id*
  - *manager_id*
  - *building_id*

- **Numerical:**
  - *price*
  - *bathrooms*
  - *bedrooms*
  - *created_date*

- **Text/Photos:**
  - *description*
  - *[features]*
  - *[photos]*

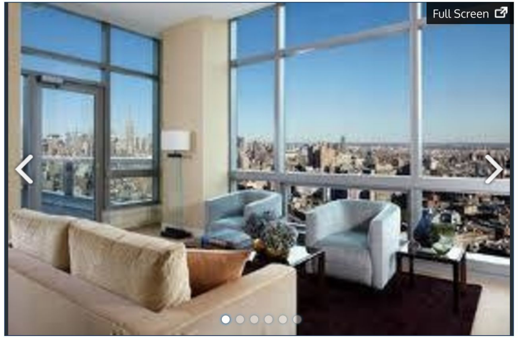
- **Location:**
  - *latitude*
  - *longitude*
  - *display_address*
  - *street_address*

Renthop Listings - Example

renthop | Search | Buildings | Resources | Need a Roommate?

« Back to Search

New to NYC? Check out our NYC Renter's Guide



Studio at W 42 St.

Hell's Kitchen, Midtown Manhattan, Manhattan

$2,570 Per Month

85.1 HopScore

Studio | **1** Bath | **Immediate** Move-In

Listing Posted 6 hours ago

shai gil

caliber associated

Description

The building located in prime midtown west location.The amenities include:Weekday Shuttle Service in theMornings & Evenings to & from Second Avenue. 24-hour white glove service.Concierge services.On-site parking garage.Hertz Connect for car rental,Off-street private drivewayQuarter-acre park,?s playground,Dog run..Health club.swimming pool.

brand new studio features the following:Floor to ceiling windows, City Skyline & the Hudson RiverHigh ceilings 9?-11?Window shades,Bosch Washer/Dryer.Streamlined open kitchensCabinets of Zebrano/Wenge wood grain laminate.Stainless steel appliancesWine coolers.shower stalls in addition to bath tubs, double sinksPorcelain tile floors and wallsCarrera marble countertopKohler bath fixtures.

Don't miss it!!!!!**NO FEE & 2 MONTH FREE***Net effective price*AVAILABLE 24/7!!

I SPECIALIZE IN THE RENTAL MARKET FOR THE PAST FEW YEARS , FEEL FREE TO CONTACT ME FOR FURTHER ASSISTANT. SHAI GIL 9543484066 sgil@calibernyc.com I'LL BE MORE THEN HAPPY TO ASSIST YOU WITH YOUR SEARCH AND MAKE SURE TO SAVE YOUR SEARCH AS STRESS FREE! All photos, amenities, and descriptions are a direct feed from the brokerage.

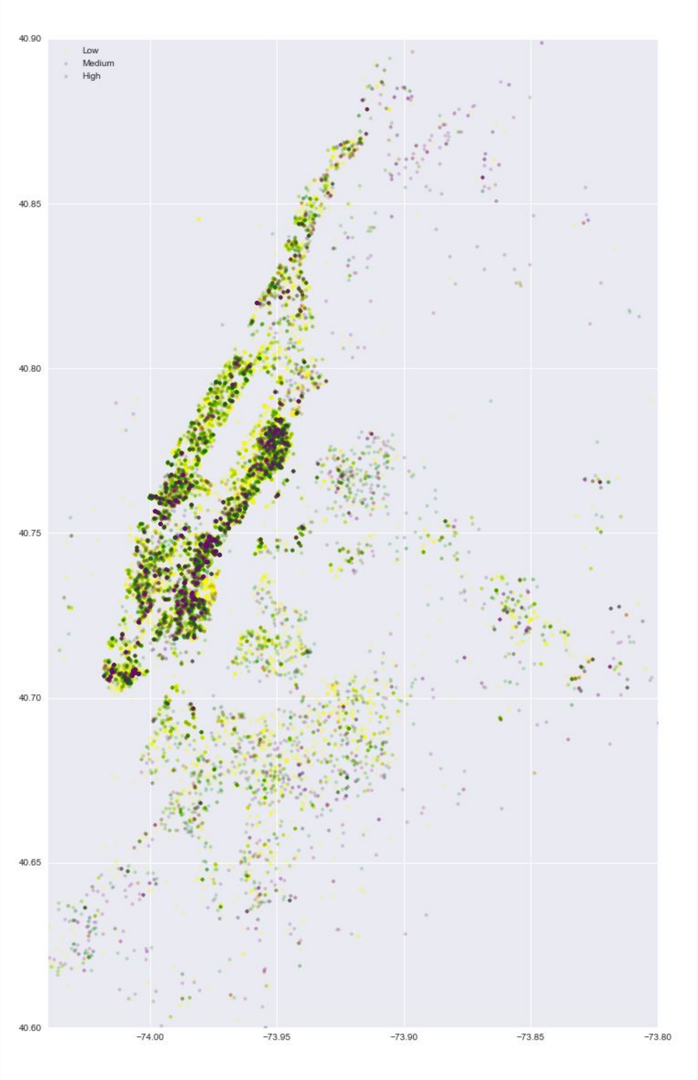
When calling refer to RentHop MX ID 928822

Features & Amenities

- **No Fee**
- Swimming Pool
- Roof Deck
- Dining Room
- Doorman
- Elevator
- Fitness Center
- Laundry in Building
- Laundry in Unit
- High Speed Internet
- Dishwasher
- Hardwood Floors
- Outdoor Space
- New Construction
- Dogs Allowed
- Cats Allowed

On the Map

Listing Data plotted (Lat/Long)



Understanding the dataset

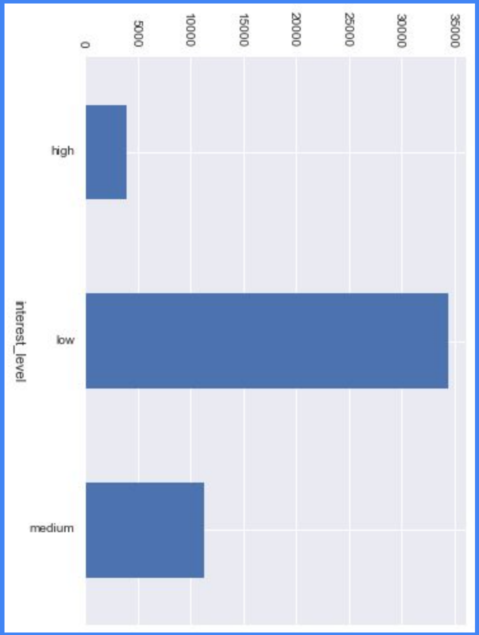
- **One target variable:**
  - *interest_level = High, Medium, Low*

- **ID columns:**
  - *listing_id*
  - *manager_id*
  - *building_id*

- **Numerical:**
  - *price*
  - *bathrooms*
  - *bedrooms*
  - *created_date*

- **Text/Photos:**
  - *description*
  - *[features]*
  - *[photos]*

- **Location:**
  - *latitude*
  - *longitude*
  - *display_address*
  - *street_address*



Problem Statement

Understanding the Dataset

Feature Engineering

Classification Modeling Methods

Submitting a prediction in kaggle

Next steps

Understanding the dataset

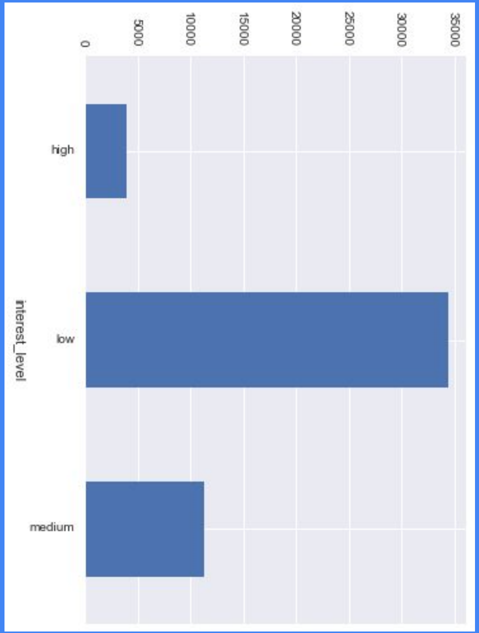
- **One target variable:**
  - *interest_level = High, Medium, Low*

- **ID columns:**
  - *listing_id*
  - *manager_id*
  - *building_id*

- **Numerical:**
  - *price*
  - *bathrooms*
  - *bedrooms*
  - *created_date*

- **Text/Photos:**
  - *description*
  - *[features]*
  - *[photos]*

- **Location:**
  - *latitude*
  - *longitude*
  - *display_address*
  - *street_address*



Using “manager performance” as a feature

- Some managers have many listings
- We can group by manager and assign weights:
  - High = 1
  - Medium = 0
  - Low = -1
- Use training set only to calculate manager performance

```
df_train.shape
```

```
(49352, 15)
```

```
df_train.groupby('manager_id')['listing_id'].count().shape
```

```
(3481,)
```

manager_count	manager_skill
34	-0.117647
29	0.413793
29	0.413793
10	-0.700000
78	-0.320513

Reducing the number of feature categories in the listings

Features & Amenities

- No Fee
- Dining Room
- Fitness Center
- High Speed Internet
- Outdoor Space
- Cats Allowed
- Swimming Pool
- Doorman
- Laundry in Building
- Dishwasher
- New Construction
- Roof Deck
- Elevator
- Laundry in Unit
- Hardwood Floors
- Dogs Allowed

```
(pd.DataFrame({'category' : cats.keys(),
 'count' : cats.values()})).sort_values('count',
                          ascending=False).head(10)
```

	category	count
505	elevator	26273
121	hardwood floors	23558
1293	cats allowed	23540
1232	dogs allowed	22035
902	doorman	20967
215	dishwasher	20806
82	laundry in building	18944
1031	no fee	18079
338	fitness center	13257
833	laundry in unit	9435

```
(pd.DataFrame({'category' : cats.keys(),
               'count' : cats.values()})).shape
```

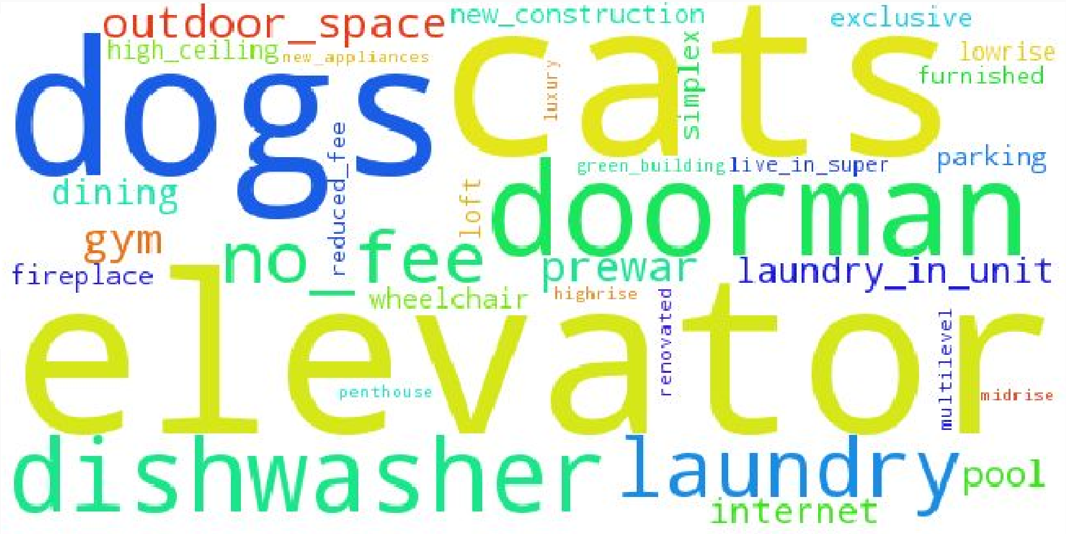
```
(1294, 2)
```

features
[Doorman, Elevator, Fitness Center, Cats Allow...
[Hardwood Floors, No Fee]

Reducing the number of feature categories in the listings

Features & Amenities

- No Fee
- Swimming Pool
- Roof Deck
- Dining Room
- Doorman
- Elevator
- Fitness Center
- Laundry in Building
- Laundry in Unit
- High Speed Internet
- Dishwasher
- Hardwood Floors
- Outdoor Space
- New Construction
- Dogs Allowed
- Cats Allowed



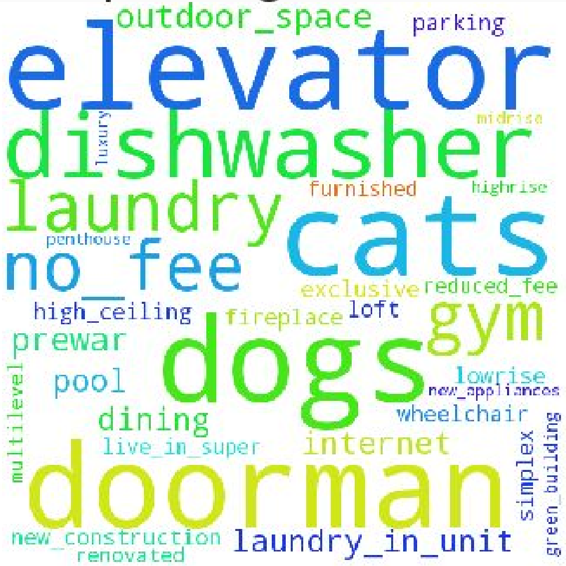

```
pd.get_dummies(df_train['categories']
.apply(pd.Series).stack()).sum(level=0)
```

Reducing the number of feature categories in the listings

Top Categories: High

Top Categories: Medium

Top Categories: Low



Problem Statement

Understanding the Dataset

Feature Engineering

Classification Modeling Methods

Submitting a prediction in kaggle

Next steps

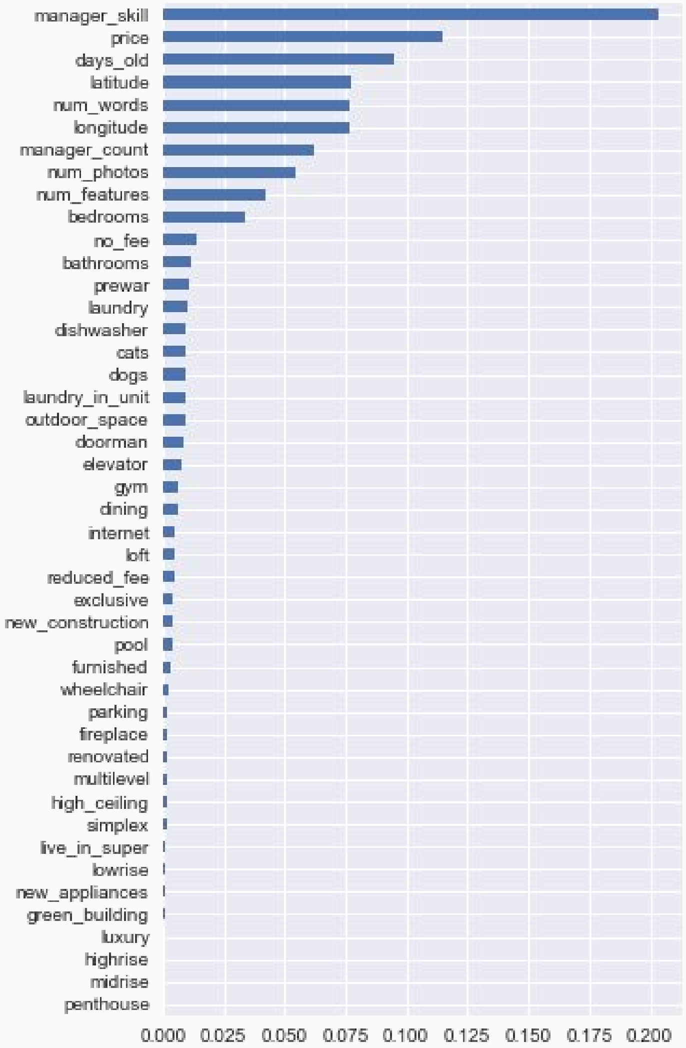
Classification Methods: Comparison

Initial Feature Set: 'Bathrooms','bedrooms','price','latitude','longitude', 'days_old','num_words','num_features','num_photos'	Logistic Loss **sklearn.metrics.log_loss()**
Logistic Regression (binomial)	0.71776
Logistic Regression (multinomial/newton-cg)	0.71441
KNN (n_neighbors=100)	0.75956
Bernoulli NaiveBayes	0.76274
MLP (hidden_layer_sizes=(100,50,10))	0.65063
Random Forest(n_estimators=1000)	0.62541

Classification Methods: Random Forest

Input Variable Set sklearn.RandomForestClassifier(n_estimators=1000)	Logistic Loss sklearn.metrics.log_loss()		
'Bathrooms','bedrooms','price','latitude','longitude', 'days_old','num_words','num_features','num_photos'	0.63817		
+ Feature categories	0.62023		
+ Manager count and manager skill	0.59985 / with 73.5% accuracy	Precision	Recall
	Class = High	50.2%	30.4%
	Class = Medium	48.6%	35.3%
	Class = Low	80.2%	91.0%

Classification Methods: Random Forest
- Feature Importances



Problem Statement

Understanding the Dataset

Feature Engineering

Classification Modeling Methods

Submitting a prediction in kaggle

Next steps

Submitting prediction in kaggle

Two Sigma CONNECT renthop

Two Sigma Connect: Rental Listing Inquiries

How much interest will a new rental listing on RentHop receive?

591 teams · 2 months to go

Overview | Data | Kernels | Discussion | Leaderboard | More | My Submissions | Submit Predictions

File descriptions

- **train.json** - the training set
- **test.json** - the test set
- **sample_submission.csv** - a sample submission file in the correct format
- **images_sample.zip** - listing images organized by listing_id (a sample of 100 listings)
- **Kaggle-renthop.7z** - (optional) listing images organized by listing_id. Total size: 78.5GB compressed. I

Submissions are evaluated using the multi-class logarithmic loss. Each listing has one true class. For each listing, you must submit a set of predicted probabilities (one for every listing). The formula is then,

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

Submission File

You must submit a csv file with the listing_id, and a probability for each class.

The order of the rows does not matter. The file must have a header and should look like the following:

```
listing_id,high,medium,low
7065104,0.07743170693194379,0.2300252644876046,0.6925430285804516
7089035,0.0, 1.0, 0.0
...
```

https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries

Submitting prediction in kaggle

Two Sigma Connect: Rental Listing Inquiries

How much interest will a new rental listing on RentHop receive?

591 teams · 2 months to go

Overview | Data | Kernels | Discussion | **Leaderboard** | More | My Submissions | **Submit Predictions**

Public Leaderboard | Private Leaderboard

This leaderboard is calculated with approximately 40% of the test data.
The final results will be based on the other 60%, so the final standings may be different.

Refresh

#	△1w	Team Name * in the money	Kernel	Team Members	Score	Entries	Last
1	—	* Silogram			0.51399	20	19h
2	—	Michael Jahrer			0.52067	27	4d
3	▲1	NxGTR			0.52810	55	4h
353	▼70	Pau Bellot Pujalte			0.60639	9	5d
354	▲110	Prashant			0.60653	8	4h
355	▼71	Matthijs Brouns			0.60668	11	3d

https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries

Problem Statement

Understanding the Dataset

Feature Engineering

Classification Modeling Methods

Submitting a prediction in kaggle

Next steps

Next steps

- **Opportunities remaining in the dataset:**
    - Incorporate image data
    - Further refine treatment of *manager_id*

- **Classification modeling:**
    - Apply text classification on the *description* feature
    - Incorporate cross-validation
    - Improve understanding of the effect of features

Questions?

Prashant Tatineni