

Two Sigma Connect: Rental Listing Inquiries



renthop



TWO SIGMA

Хинензон Евгений

Чуркин Никита

План

1. Описание и особенности задачи.
2. Обзор данных
3. Feature engineering
4. Кросс-валидация
5. Модели
6. Лик в данных
7. Работа с изображениями
8. Результаты

Описание и особенности задачи

1. Имеются данные объявлений по аренде недвижимости в США с сайта [renthop.com](https://www.renthop.com). Каждому объявлению сопоставлен ответ – «уровень интереса»: low, medium или high.
2. Необходимо предсказать уровень интереса для тестового множества объявлений. Метрика качества – multi-class loss.
3. Данные очень разнообразны и разнородны: содержат текст, числовые переменные, изображения и т.д.

Мотивировка

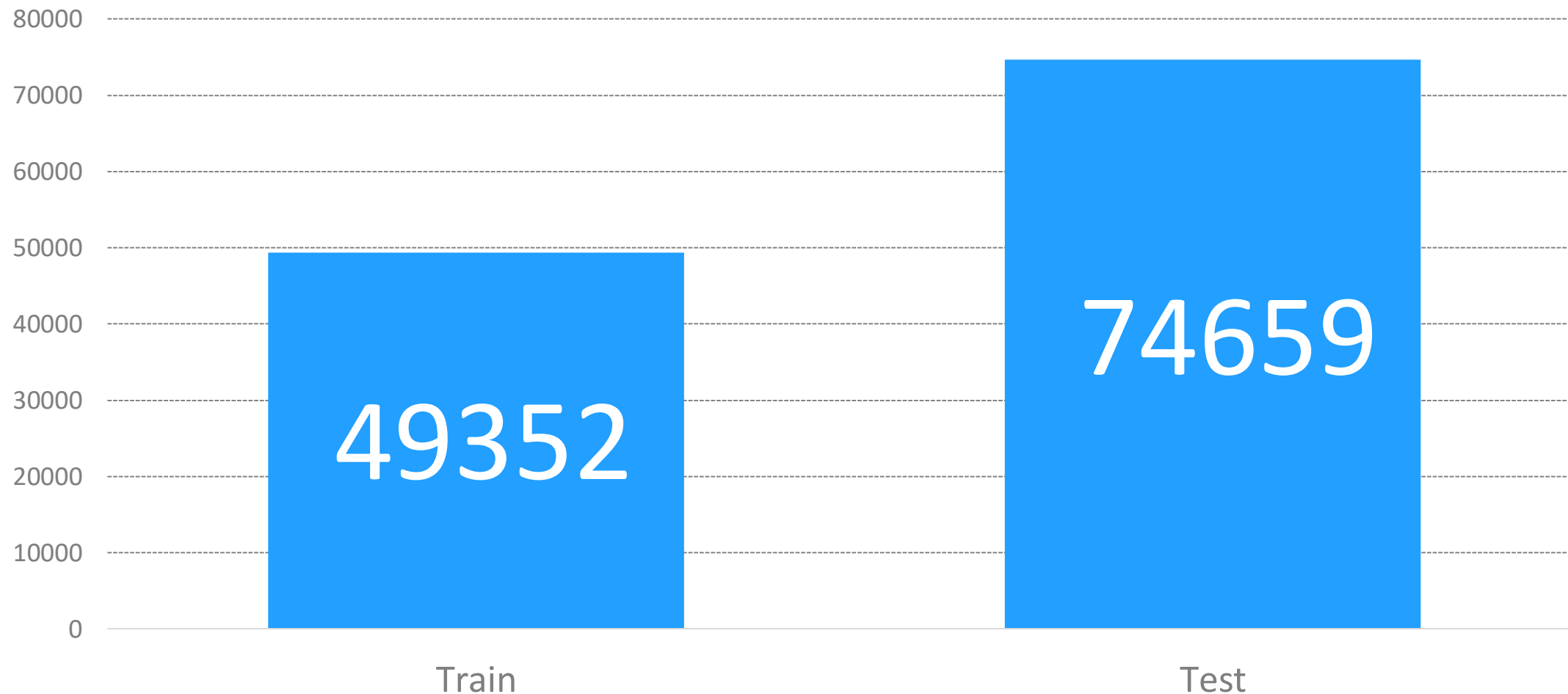
- Можно ли автоматизировать для покупателя процесс выбора объекта недвижимости?
- Можно ли исключить менеджера-посредника из следующей цепочки (пример: турагентства в России постепенно замещаются)



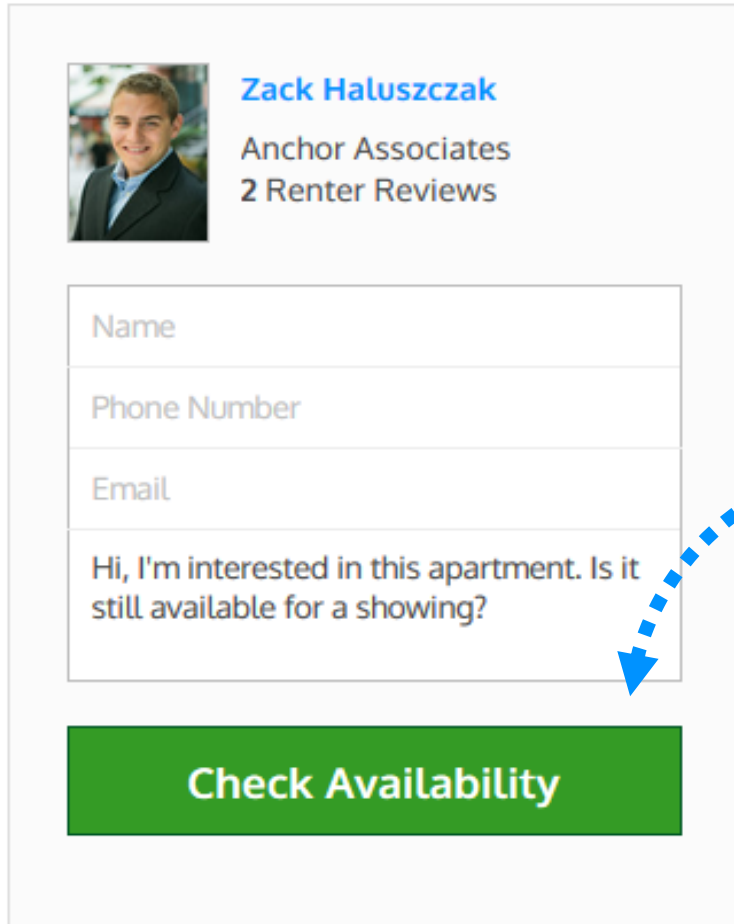
Обзор данных


- Числовые: количество санузлов, количество спален, цена
- Гео-данные: широта и долгота, полный адрес (display address), адрес без номера дома – адрес улицы (street address)
- Текстовые: список из ключевых слов (features), блок описания объекта (description)
- Id-шники: manager_id (id менеджера-агента), building_id (id здания), listing_id (id объявления, уникальный номер объявления).
- Изображения: фотографии объектов, ссылки на изображения
- Дата создания объявления

Количество объектов в train и test



Что такое «interest level»?



 **Zack Haluszczak**
Anchor Associates
2 Renter Reviews

Name

Phone Number

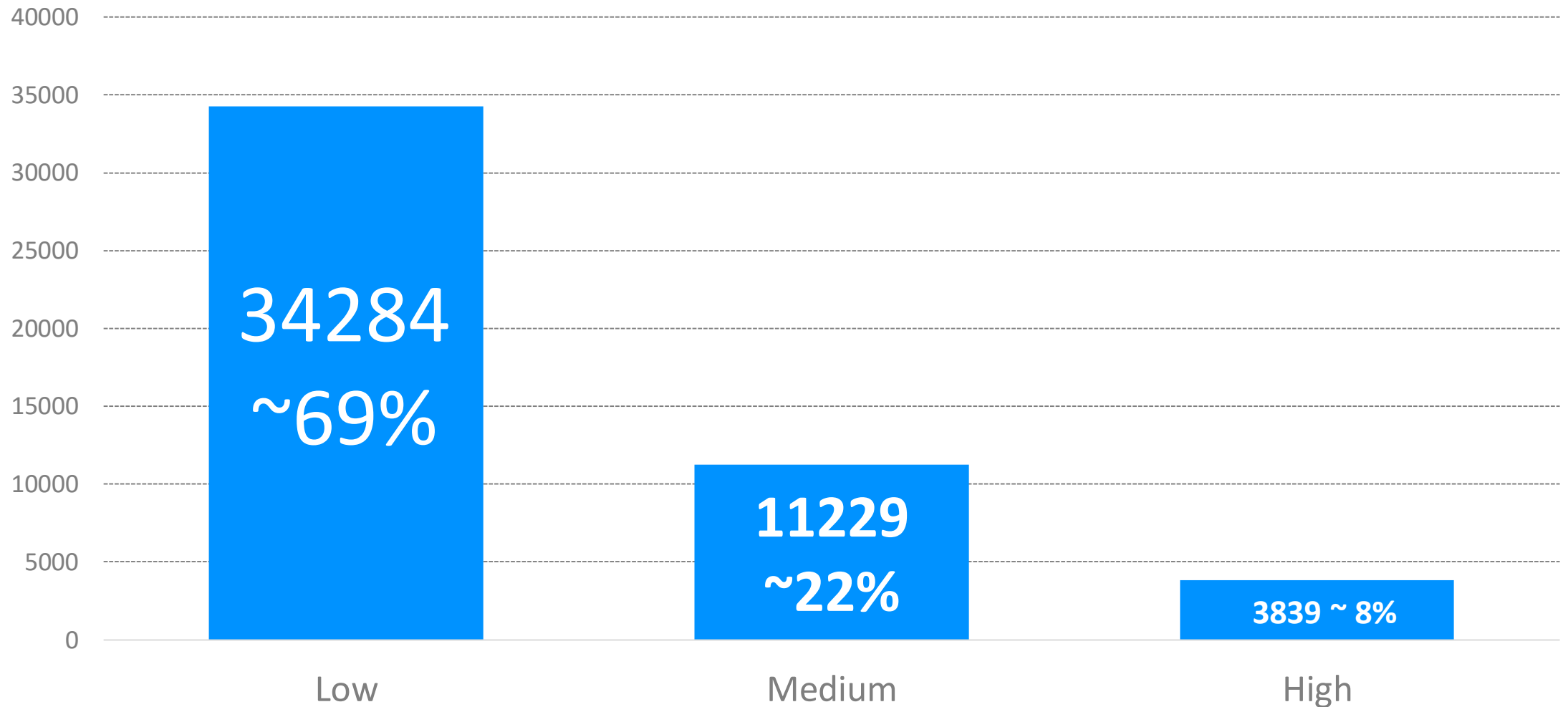
Email

Hi, I'm interested in this apartment. Is it still available for a showing?


Check Availability

- Количество кликов?
- Количество обращений к менеджеру через форму «check availability»?
- «Kagglers will predict the number of inquiries a new listing receives based on the listing's creation date and other features» – из description конкурса.

Распределение interest level в train





 **Jason Burke**
Citi Habitats

Name

Phone Number

Email

Hi, I'm interested in this apartment. Is it still available for a showing?

Check Availability

[Report Listing](#)

Менеджер и форма отклика (в профиле менеджера указаны его объявления + отзывы)

Studio at Washington street
Financial District, Downtown Manhattan, Manhattan

Studio | 1 Bath | **Immediate** Move-In
Listing Posted 26 mins ago

\$2,631 Per Month
100 HopScore



Количество санузлов

Цена

Description

One month free and no broker fee for a limited time!
Will not last long at this price, contact me today to view!

The apartment features:

Описание объекта

Ключевые слова (features)

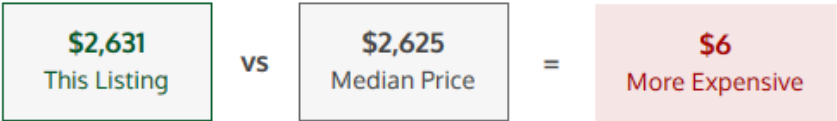
Features & Amenities

- | | | |
|--------------------|---|------------------------|
| ✓ No Fee | ✓ Featured | ✓ Cats Allowed |
| ✓ Dogs Allowed | ✓ Doorman | ✓ Elevator |
| ✓ Fitness Center | ✓ Laundry In Building | ✓ Common Outdoor Space |
| ✓ Storage Facility | ✓ One month rent free on a 13 month lease | ✓ no broker fee! |

Сравнение по цене →

Price Comparison

Comparing **this listing** against median prices for **Studio / 1BA apartments in Financial District with Doorman, Elevator**.



The price of this apartment is **\$6 more expensive** than the median price.

Плюсы и минусы
Объявления
(видимо, по ним вычисляется HopScore) →

HopScore Breakdown

This listing has a HopScore of **100** and was posted **26 mins ago**. The listing quality and manager score is **outstanding**. Some of the contributing factors to the HopScore are listed below.

| | |
|---------|--|
| GOOD | Listing is featured |
| GOOD | This manager has a VIP Account subscription |
| GOOD | Manager has registered with RentHop |
| GOOD | Manager is from a reputable firm |
| GOOD | Manager has great inventory and availability |
| NEUTRAL | Street number not provided |
| NEUTRAL | Manager does not yet have user reviews |

Цена, санузлы и спальни I

0. В данных есть ошибки: нулевые цены, гигантское кол-во спален в некоторых листингах и т.д.

«Плохих» строчек не очень много – спальни и санузлы исправляем руками по логике (22 спальни -> 2 спальни и т.д.), цену с помощью медианной цены.

1. Цену можно прологарифмировать для симметричности (асимметрия > 100 -> асимметрия ~1).

Цена, санузлы и спальни II

2. Разные осмысленные комбинации bedrooms, bathrooms, price, полученные с помощью арифметических операций +, -, *, /,

Пример: $\text{price} / (\text{bathrooms} + \text{bedrooms})$ – «цена за комнату»

3. Есть дробные санузлы, видимо, это означает наличие туалета без ванной или душа ($\text{bathrooms} == 2.5$) -> (есть 2 полных санузла и 1 отдельный туалет).

Получаем фичу: является ли bathrooms целым числом?

Цена, санузлы и спальни III

4. Восстанавливаем фичу:

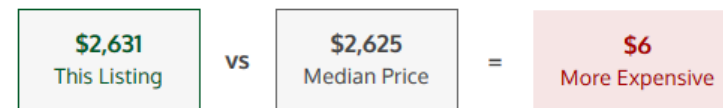
Price comparison.

Используем KNNRegressor

для предсказания цены в листинге на основе только широты или долготы. Истинную цену делим на получившуюся предсказанную цену – получаем «переоценку» данного объявления. Это и будет аналогом «price comparison».

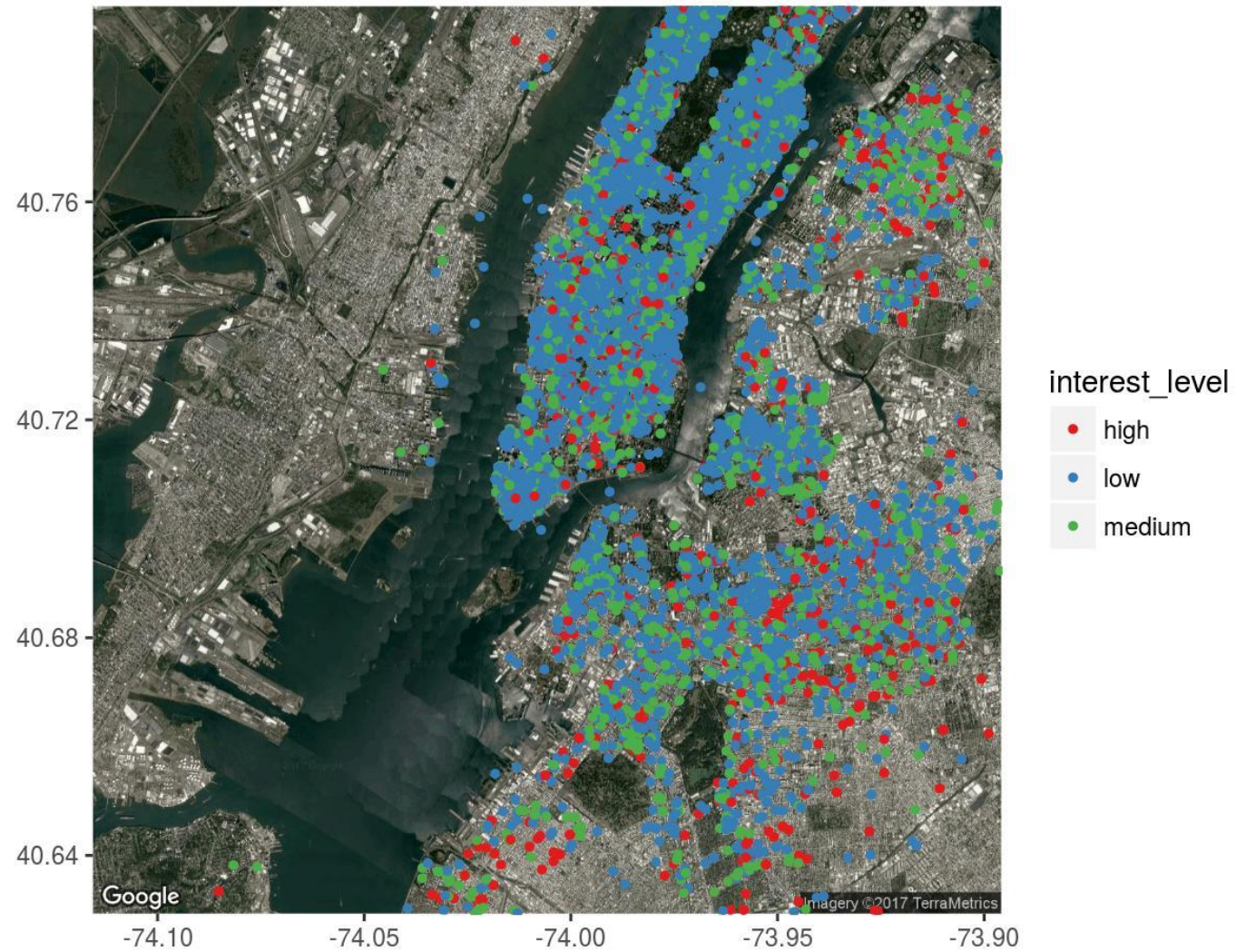
Price Comparison

Comparing **this listing** against median prices for **Studio / 1BA apartments in Financial District with Doorman, Elevator**.



The price of this apartment is **\$6 more expensive** than the median price.

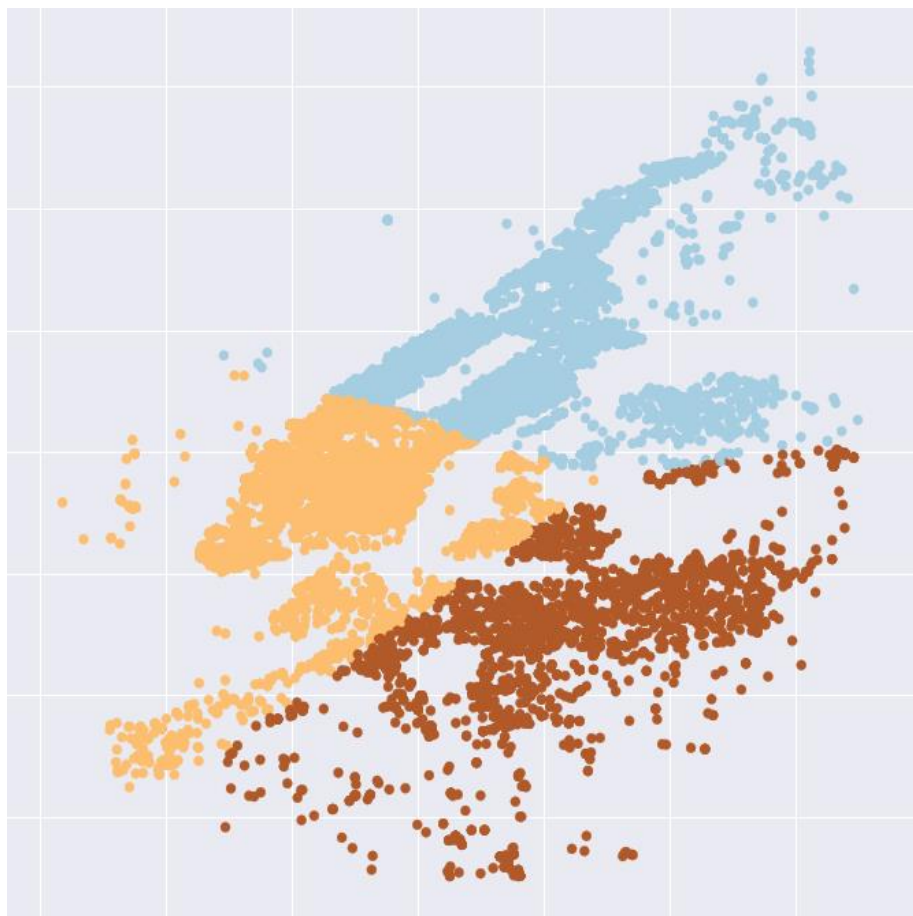
Гео-данные



Гео-данные I

- 0. В координатах ошибки: порядка 39 некорректных. Эти координаты легко восстановить по адресу.
- 1. В условии соревнования сказано, что все объекты в данных из Нью-Йорка. Однако, есть (и в train, и в test) объекты из других городов.
- 2. Посчитаем расстояние от здания объявления до медианных координат.
- 3. Расстояние от здания объявления до Центрального Парка.

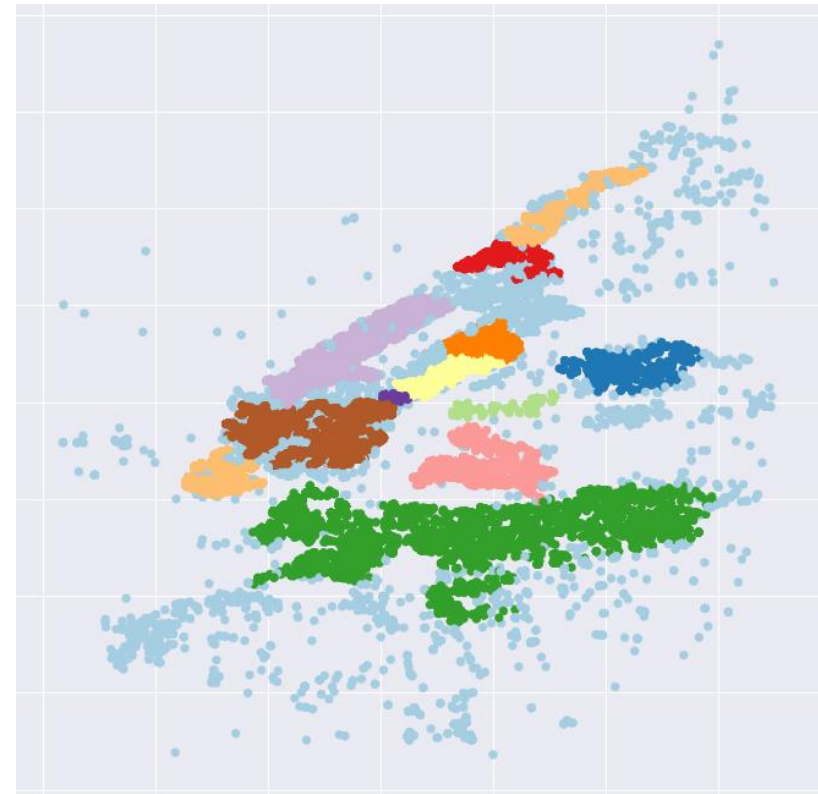
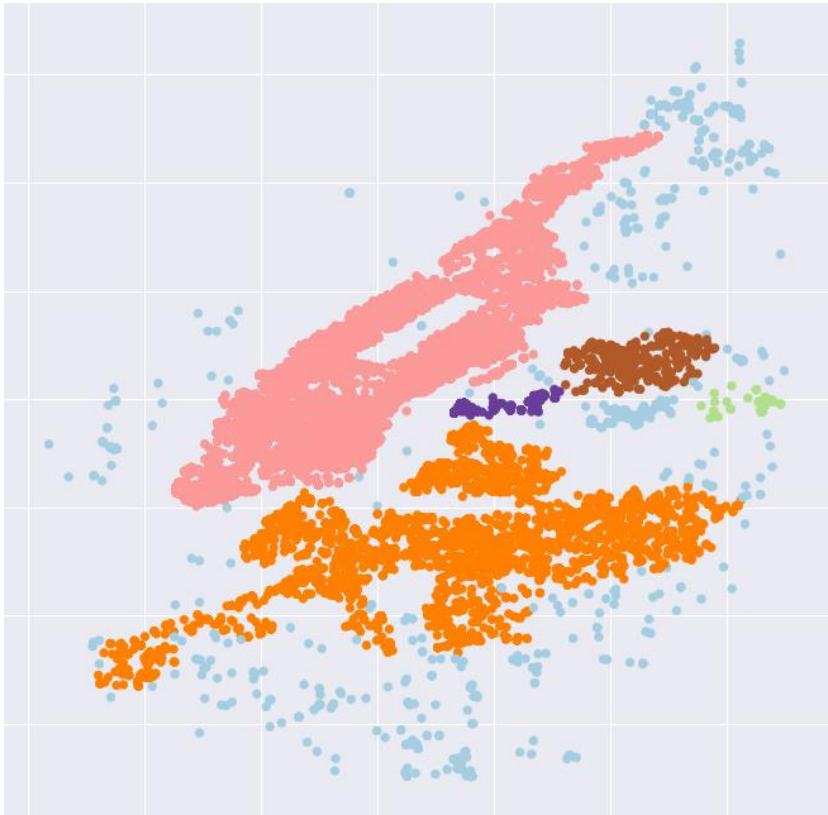
3Means и 50Means



Гео-данные II

4. Метки полученных кластеров – новый признак.
5. Посчитаем по каждому кластеру всевозможные характеристики:
медианную цену внутри кластера, частотность кластера в данных, количество уникальных зданий в кластере, количество менеджеров, работающих в данном кластере, среднее значение целевой переменной в кластере.
6. Хочется кластеризовать локации получше и проделать пункты 4-5 с новой кластеризацией.

HDBSCAN (разные min_cluster_size)



<http://hdbscan.readthedocs.io/en/latest/index.html>

Гео-данные III

7. В данных есть два вида адреса: с номером дома и без. Вычислим расстояние Левенштейна (количество исправлений, требующихся для приведения одной строки к другой) между этими двумя адресами.

Для большинства объектов это расстояние лежит в интервале 4-6, есть объекты с расстоянием 120 и более: обычно в одном из адресов в этом случае находится не адрес, а реклама или посторонние данные.

Гео-данные IV

8. Адреса без номеров дома приводим к одному виду:

приводим слова к нижнему регистру, удаляем пунктуацию, трансформируем слова следующим образом:

Str -> st, street -> st, avenue -> av, boulevard -> blvd и т.п.

9. Делаем one-hot encoding по улице.

10. Вычисляем медианную цену по улице, кол-во зданий и т.д.

Текстовые признаки I

0. Считаем количество ключевых слов, т.е. слов в features.

1. Слова из features приводим к нижнему регистру и удаляем похожие словосочетания, «дубликаты», например:

Wifi_in_building -> wifi, wifi -> wifi_available -> wifi и т.д.

2. Считаем ключевые слова с помощью CountVectorizer (берем 75 самых частотных слов).

Features & Amenities

- | | | |
|---------------------------------|----------------------------|--|
| ✓ No Fee | ✓ Floorplans Available | ✓ Featured |
| ✓ Exclusive | ✓ Doorman | ✓ Elevator |
| ✓ Fitness Center | ✓ Common Areas | ✓ Wifi In Building |
| ✓ Outdoor Areas | ✓ Storage | ✓ Valet |
| ✓ Swimming Pool | ✓ Dishwasher | ✓ Wheelchair Accessible |
| ✓ Sun Deck | ✓ Floor To Ceiling Windows | ✓ Credit Card Payment Accepted (fee Applies) |
| ✓ Electronic Rent Payment (ach) | ✓ Concierge Services | ✓ On-site Management |
| ✓ Cats Allowed | ✓ Dogs Allowed | |

Текстовые признаки II

0. В блоке `description` удаляем пунктуацию, приводим слова к нижнему регистру.

1. Вычисляем длину `description` и количество слов в `description`.

2. Считаем кол-во слов, написанных CAPSom.

3. Определяем, если в описании телефон или e-mail менеджера.

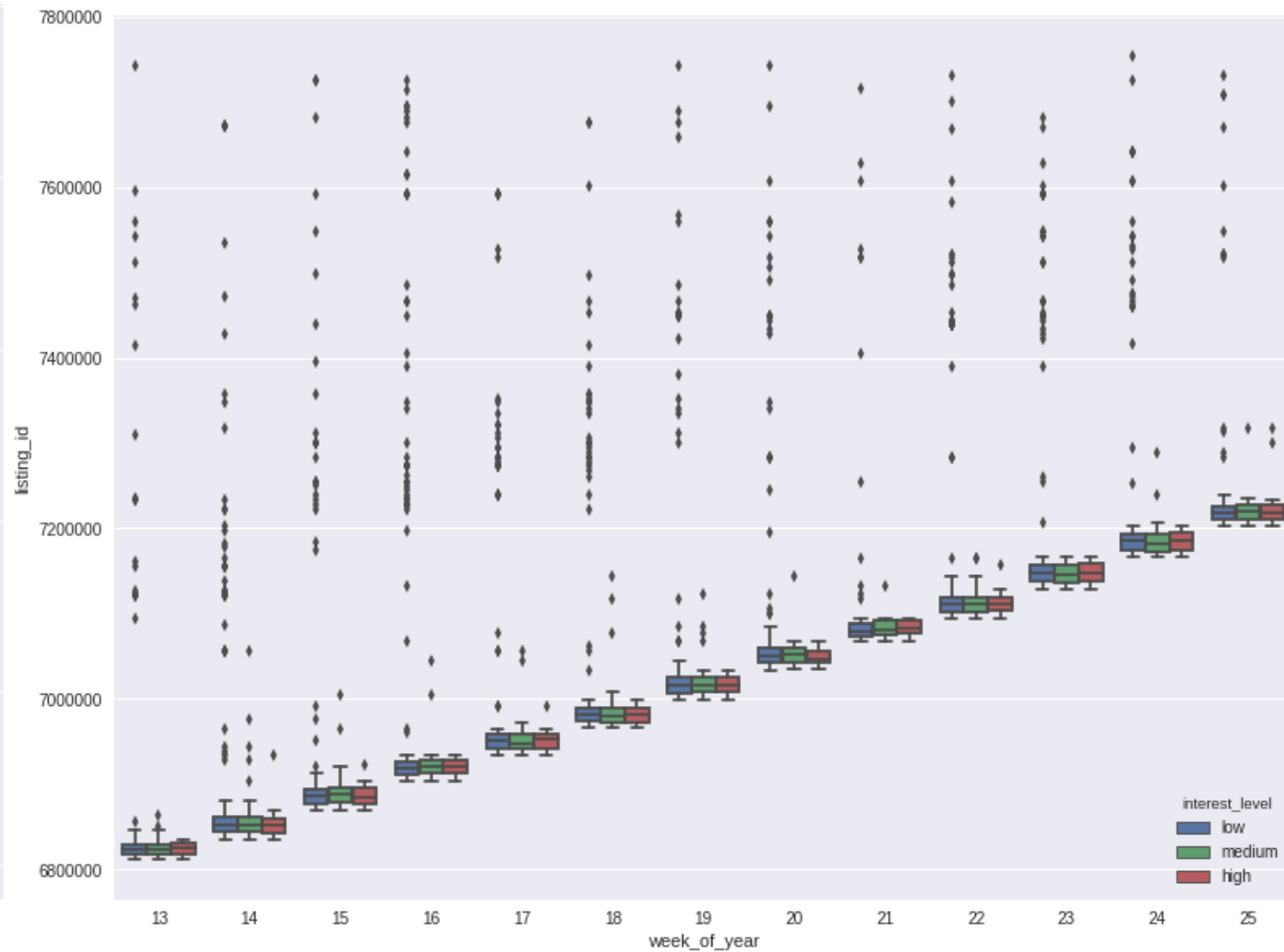
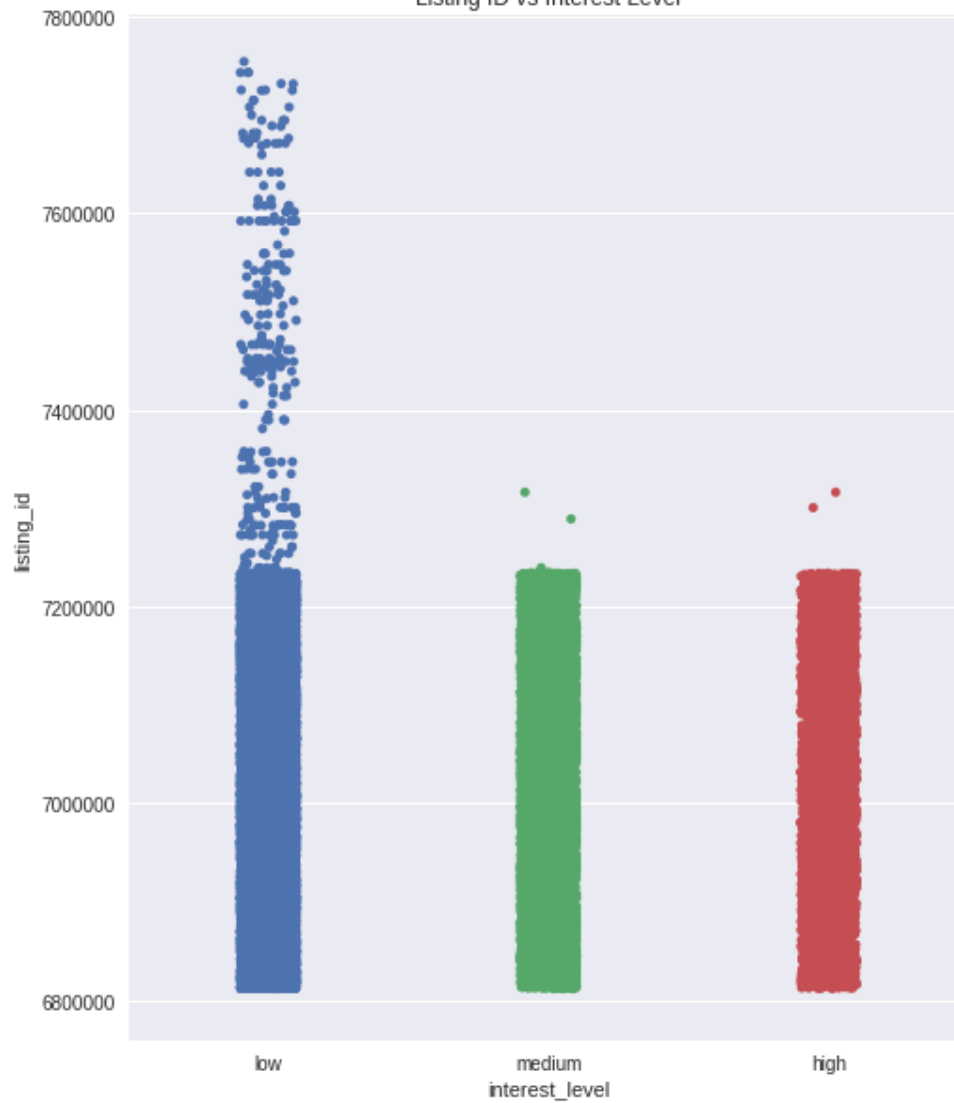
4. Вручную формируем несколько тематических словарей, например:

`transportation = ['subway', 'bus', 'taxi', 'train', 'railroad', 'railway', 'transport']` и считаем, сколько слов в описании из каждого словаря.

Дата и время

0. Получаем год/месяц/день недели/час появления публикации на сайте и т.д.
1. Считаем количество прошедших секунд с самой ранней даты из train/test до текущей даты объявления.
2. listing_id оказался полезной фичей, т.к. имелась связь между датой и listing_id

Listing ID vs Interest Level



Building_id и manager_id I

0. Building_id и manager_id – числа в шестнадцатеричной системе счисления.
1. manager_id – уникальные идентификаторы, а в building_id есть id 0, который встречается у разных зданий.
2. Для каждого менеджера/здания посчитаем число зданий/менеджеров, которые он обслуживает.
3. Медианная цена объявлений данного менеджера/здания.
4. Средняя длина описания и среднее количество ключевых слов у листингов данного менеджера.

Building_id и manager_id II

5. Для каждого менеджера/здания определяем частотность его появлений.
6. Вычисляем, сколько секунд прошло с предыдущего листинга с данным менеджером/зданием.
7. То же самое, что в п.6, но смотрятся листинги лишь с ответом high.
8. Уровень интереса предыдущего листинга данного менеджера/здания.
9. Кодлируем id зданий/менеджеров средним таргетом.
10. «Скорость» менеджера $\log(1 + \text{Сумма «преодоленного пути»} / \text{интервал активности})$.
11. One-hot encodi'м менеджеров и здания.

Кластеризация менеджеров

«Listings were all created by either an independent landlord, a professional real estate marketing company, a New York licensed real estate agent, or a NYC brokerage firm» - из description соревнования.

Согласно описанию, есть 4 типа посредников.

Возьмем данные по менеджерам, рассказанные на предыдущих слайдах и используем 4Means для кластеризации менеджеров на 4 группы.

Модели

Использовались следующие модели: KNN, DT, ET, LR, SVC, RF, XGB, LGBM.

Лучшая модель – lgbm (!) - 0.504 CV, 0.509 LB

Схема решения

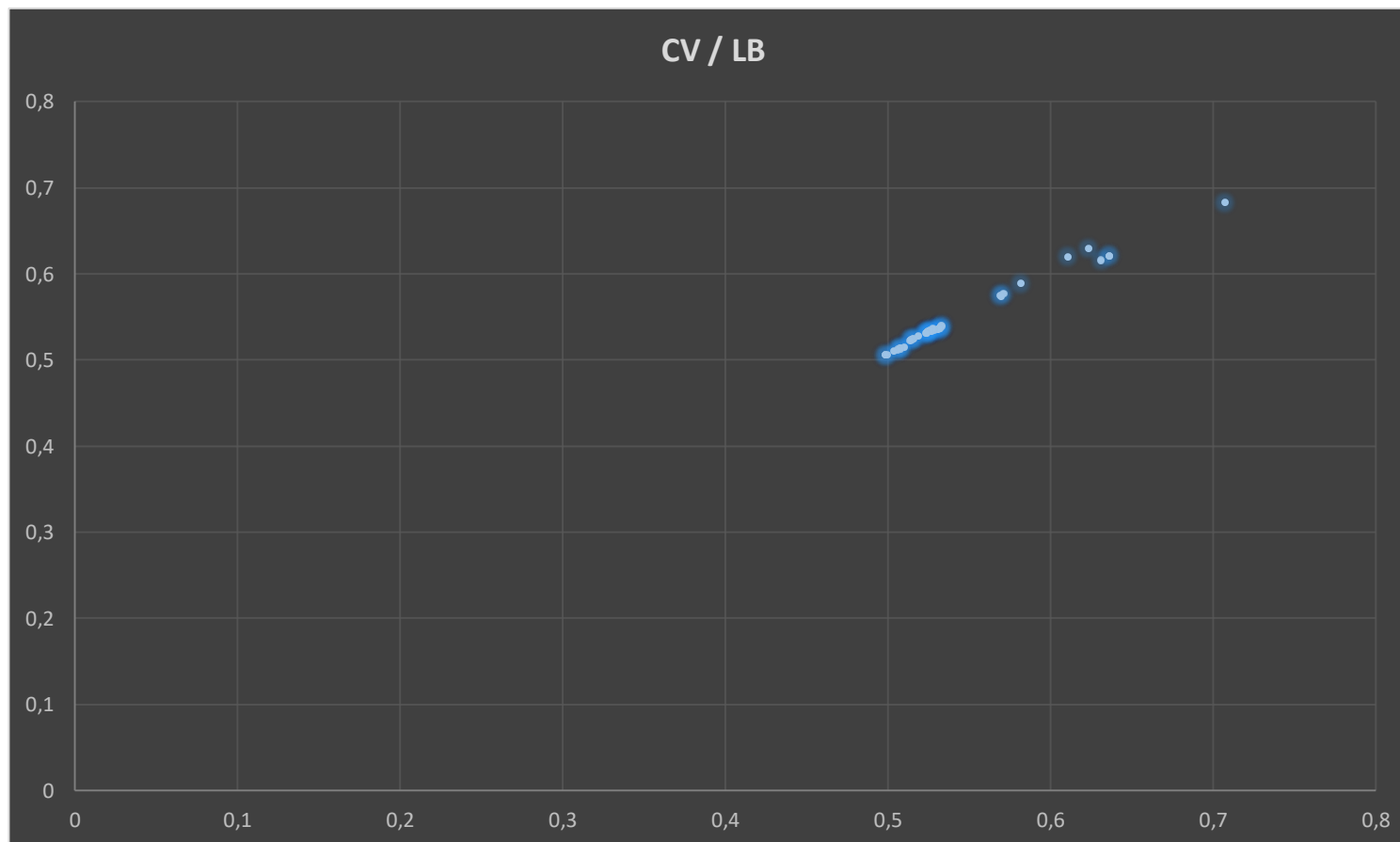
0. Всего было около 30 моделей первого уровня, модели обучались на разных признаковых пространствах.
1. На втором уровне выбирались несколько самых лучших признаков и все модели первого уровня, рассчитывались стандартные характеристики по матрицам.
2. Используем LGBM для финальной модели.

Выбор гиперпараметров

Все модели оптимизируем с помощью BayesianOpt:

<https://github.com/fmfn/BayesianOptimization>

Связь CV и LB



Изображения



78.5GB данных

В общем датасете представлены ссылками на каждый listing_id

Изображения II

0. Была замечена проблема, в некоторых папках содержались дубликаты фотографий.

Для нахождения количества уникальных фотографий по listing_id искалось среднее значение серого цвета для каждой фотографии (с точностью до 10 знака) – некое уникальное значение для картинки.

Фотографии с одинаковым значением удалялись

- Фича – количество уникальных фотографий по каждому listing_id

Изображения III

1. В некоторых листингах с одним building_id было замечено много повторяющихся изображений. Могли содержаться одна или две уникальные фотографии.

Снова использовалось среднее значение серого для нахождения уникальных.

- Фича – количество уникальных фотографий по building_id



Изображения IV

2. В немногих листингах содержались планировки квартир. Для нахождения считалось количество белых пикселей, если это значение было больше 60%, то данная картинка считалась планировкой.

- Фича : наличие планировки



Изображения V

3. Было видно, что качество фотографий сильно отличалось. Хотелось получить более подробную информацию о изображении и камере, например производитель камеры, модель, дата и время съёмки и тд. Проверил все изображения на наличие Exif данных. Их оказалось мало порядка 6 000. Поэтому взял только наличие/отсутствие метайнформации.

- Фича: Наличие exif данных

Изображения VI

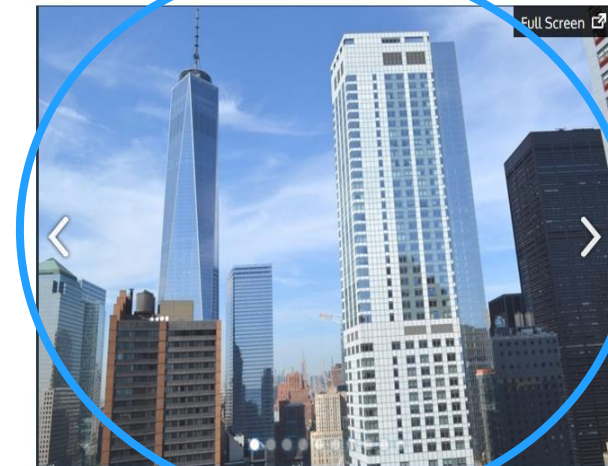
4. Были получены некоторые ключевые характеристики изображений – средняя яркость, средний тон, средняя насыщенность. Предполагалось, что яркие и светлые фотографии более нравятся больше. Для этого изображение переводилось в цветовую модель HSV и находилось среднее по каждому каналу.

Фичи: средняя яркость, средний тон, средняя насыщенность.

Изображения V

Требовалось выбрать, по какому из изображений находить предыдущие фичи. Было сделано предположение, что самая ранняя фотография из листинга закреплялась первой на превью на сайте и по ней судили об объявлении.

Для нахождения времени создания фотографий использовали архиватор 7zip. Он позволяет получить список файлов с датой создания.



Studio at Washington street
Financial District, Downtown Manhattan, Manhattan

\$2,631
Per Month

100
HopScore

Studio 1 Bath Immediate Move-In
Listing Posted 26 mins ago



Description

One month free and no broker fee for a limited time!

Will not last long at this price, contact me today to view!

The apartment features:



Jason Burke
Citi Habitats

Name
Phone Number
Email
Hi, I'm interested in this apartment. Is it still available for a showing?

Check Availability

[Report Listing](#)

Заметки из решения с 0,5 LB

Использовалась 3-х уровневая модель: первый уровень – 20 моделей (микс из xgb, LightGBM, keras, RF и тд), второй уровень 4 модели и третий уровень - среднее арифметическое с сабмитов второго уровня

Каждая модель из первого уровня использовала примерно 200 фичей. (Не использовалась конвертация категориальных переменных в числовые для избегания лика).

По мнению Силограма, отличием между его решением и решениями в районе 0,51 был создание признакового пространства. И в частности, он говорит о какой-то «магической» фиче, которая еще не была описана на форуме. Силограм, счел нечестным раскрыть эту фичу в тот момент.

Основные выводы:

Методы GBM работают лучше, когда данные в соревновании не однородные

LightGBM – очень хорошо, точность близкая к XGB, но намного быстрее работает

Для стекинга, очень важно использовать модели результаты которых сильно отличаются, даже слабые модели могут пригодиться.

15

—

Silogram



0.49786

71



1mo

Magic или Leak



- Казанова(занявший 11 место), был не согласен с Силограмом, и выложил на форуме «магическую» фичу. Он считал это ликом.
- Лик заключался в известном времени создания каждой папки изображений в миллисекундах, то есть легко было получить зависимость от времени всех данных.
- Фича дала +0.015 к LB
- Фича была получена архиватором 7zip, функцией вывода файлов архива с датой их создания.
- Стоит отметить, что фичу можно было получить только скачав изображения с торрента, а не напрямую с сайта.

Результаты на LB – Топ 2% из 2488

- Public LB

| | | | | | | |
|----|---|------------------|---|---------|----|-----|
| 54 | — | Evgeny Khinenzon |  ●●●● | 0.50366 | 33 | 1mo |
| 61 | — | Nikita Churkin |  ●●●● | 0.50400 | 71 | 1mo |

- Private LB

| | | | | | | |
|----|---|------------------|---|---------|----|-----|
| 46 | — | Evgeny Khinenzon |  ●●●● | 0.50260 | 33 | 1mo |
| 49 | — | Nikita Churkin |  ●●●● | 0.50298 | 71 | 1mo |

Другие решения и что можно было сделать

0. Решения многих топовых участников очень похожи.

1. Можно было превратить задачу классификации в задачу регрессии и увеличить количество моделей для второго уровня в 2 раза.

2. Попробовать StackNet (<https://github.com/kaz-Anova/StackNet>)

3. Лучше поработать с объектами не из Нью-Йорка.

4. При target-кодировании попробовать комбинации признаков.

5. Можно было восстановить нулевые building_id по координатам, адресам и т.д.

6. В данных были «почти» повторяющиеся строки, стоящие рядом во времени: они образовывали группы объявлений, отличавшихся лишь ценой или ключевыми словами. Можно было придумать множество признаков на основе этих групп.

sorted by group and time

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|-----|----------------------------------|----------------------------------|-----------------|--------------------|----------|-----------|-------|---------|---------|------------------|----------|-----------|----------|
| 1 | manager_id | building_id | display_address | street_address | latitude | longitude | price | Price+1 | price-1 | created | bedrooms | bathrooms | num_desc |
| 194 | 02e17b21a1814fb10336b2ee8ceb3e79 | 0c239739ea5fb30514f5968285259b58 | W 86 St. | 41 W 86 St. | 40.7863 | -73.971 | 2255 | 0 | 915 | 09/04/2016 03:14 | 0 | 1 | 86 |
| 195 | 02e17b21a1814fb10336b2ee8ceb3e79 | 0c239739ea5fb30514f5968285259b58 | W 86 St. | 41 W 86 St. | 40.7863 | -73.971 | 2255 | 1540 | 0 | 30/04/2016 04:04 | 0 | 1 | 86 |
| 196 | 02e17b21a1814fb10336b2ee8ceb3e79 | 1ade20f96090c7b55e32056fd0de1339 | W 110 St. | 501 W 110 St. | 40.8033 | -73.9641 | 3795 | -1300 | 1540 | 13/06/2016 02:42 | 3 | 1 | 142 |
| 197 | 02e17b21a1814fb10336b2ee8ceb3e79 | 1dbb2d7119b8365b5371e1e12690a032 | Amsterdam Ave. | 926 Amsterdam Ave. | 40.8002 | -73.9663 | 2495 | -95 | -1300 | 10/05/2016 03:13 | 1 | 1 | 143 |
| 198 | 02e17b21a1814fb10336b2ee8ceb3e79 | 1dbb2d7119b8365b5371e1e12690a032 | Amsterdam Ave. | 926 Amsterdam Ave. | 40.8002 | -73.9663 | 2400 | 100 | -95 | 20/05/2016 03:06 | 1 | 1 | 143 |
| 199 | 02e17b21a1814fb10336b2ee8ceb3e79 | 1dbb2d7119b8365b5371e1e12690a032 | Amsterdam Ave. | 926 Amsterdam Ave. | 40.8002 | -73.9663 | 2500 | 0 | 100 | 16/06/2016 02:36 | 1 | 1 | 143 |
| 200 | 02e17b21a1814fb10336b2ee8ceb3e79 | 1dbb2d7119b8365b5371e1e12690a032 | Amsterdam Ave. | 926 Amsterdam Ave. | 40.8002 | -73.9663 | 2500 | 700 | 0 | 23/06/2016 02:24 | 1 | 1 | 145 |
| 201 | 02e17b21a1814fb10336b2ee8ceb3e79 | 1e09d0cbacfe873404701d134f282dde | W 56 St. | 401 W 56 St. | 40.7676 | -73.9866 | 3200 | -200 | 700 | 24/06/2016 02:26 | 2 | 1 | 106 |
| 202 | 02e17b21a1814fb10336b2ee8ceb3e79 | 24f3dbb6ae6adc23a0b3220ce67d0a01 | W 58 St. | 330 W 58 St. | 40.7677 | -73.9836 | 3000 | 0 | -200 | 20/06/2016 18:28 | 2 | 1 | 103 |
| 203 | 02e17b21a1814fb10336b2ee8ceb3e79 | 251544637b2e1e915f5d287090d15130 | W 105 St. | 120 W 105 St. | 40.7988 | -73.9644 | 3000 | 0 | 0 | 09/04/2016 04:41 | 2 | 1 | 103 |
| 204 | 02e17b21a1814fb10336b2ee8ceb3e79 | 251544637b2e1e915f5d287090d15130 | W 105 St. | 120 W 105 St. | 40.7988 | -73.9644 | 3000 | 395 | 0 | 21/04/2016 03:31 | 2 | 1 | 103 |
| 205 | 02e17b21a1814fb10336b2ee8ceb3e79 | 251544637b2e1e915f5d287090d15130 | W 105 St. | 120 W 105 St. | 40.7988 | -73.9644 | 3395 | -295 | 395 | 06/05/2016 03:13 | 2 | 1 | 103 |
| 206 | 02e17b21a1814fb10336b2ee8ceb3e79 | 251544637b2e1e915f5d287090d15130 | W 105 St. | 120 W 105 St. | 40.7988 | -73.9644 | 3100 | 230 | -295 | 03/06/2016 03:42 | 2 | 1 | 103 |
| 207 | 02e17b21a1814fb10336b2ee8ceb3e79 | 263285dadae1eb73351df8edca92ff0a | Columbus Ave. | 792 Columbus Ave. | 40.7953 | -73.9667 | 3330 | 765 | 230 | 18/04/2016 02:57 | 2 | 1 | 123 |
| 208 | 02e17b21a1814fb10336b2ee8ceb3e79 | 263285dadae1eb73351df8edca92ff0a | Columbus Ave. | 792 Columbus Ave. | 40.7953 | -73.9667 | 4095 | -895 | 765 | 06/05/2016 02:28 | 2 | 1 | 123 |
| 209 | 02e17b21a1814fb10336b2ee8ceb3e79 | 35784800ca930a77ea64d2e67f001b33 | W 57 St. | 315 W 57 St. | 40.7674 | -73.9838 | 3200 | -350 | -895 | 13/06/2016 02:55 | 2 | 1 | 104 |
| 210 | 02e17b21a1814fb10336b2ee8ceb3e79 | 3a956bd42c50f06ac84cf072fc514f5f | W 42 St. | 650 W 42 St. | 40.761 | -74.0002 | 2850 | 150 | -350 | 20/05/2016 03:47 | 1 | 1 | 87 |
| 211 | 02e17b21a1814fb10336b2ee8ceb3e79 | 3a956bd42c50f06ac84cf072fc514f5f | W 42 St. | 650 W 42 St. | 40.761 | -74.0002 | 3000 | -605 | 150 | 27/05/2016 04:57 | 1 | 1 | 87 |
| 212 | 02e17b21a1814fb10336b2ee8ceb3e79 | 3cd0a4201a90325df0b5214db5a09051 | W 58 St. | 117 W 58 St. | 40.7655 | -73.9772 | 2395 | 0 | -605 | 03/05/2016 02:55 | 1 | 1 | 83 |
| 213 | 02e17b21a1814fb10336b2ee8ceb3e79 | 3cd0a4201a90325df0b5214db5a09051 | W 58 St. | 117 W 58 St. | 40.7655 | -73.9772 | 2395 | 805 | 0 | 13/06/2016 02:24 | 1 | 1 | 83 |
| 214 | 02e17b21a1814fb10336b2ee8ceb3e79 | 47eabd0f346864a6b77b630008efe56d | W 60 St. | 200 W 60 St. | 40.771 | -73.9876 | 3200 | 0 | 805 | 10/05/2016 02:59 | 1 | 1 | 100 |

$=2500-2400=100$
 lead lag
 $=2495-2400=-95$

Спасибо за внимание

jenia.khinenzon@gmail.com – Евгений Хинензон

nikita1994175@yandex.ru – Никита Чуркин