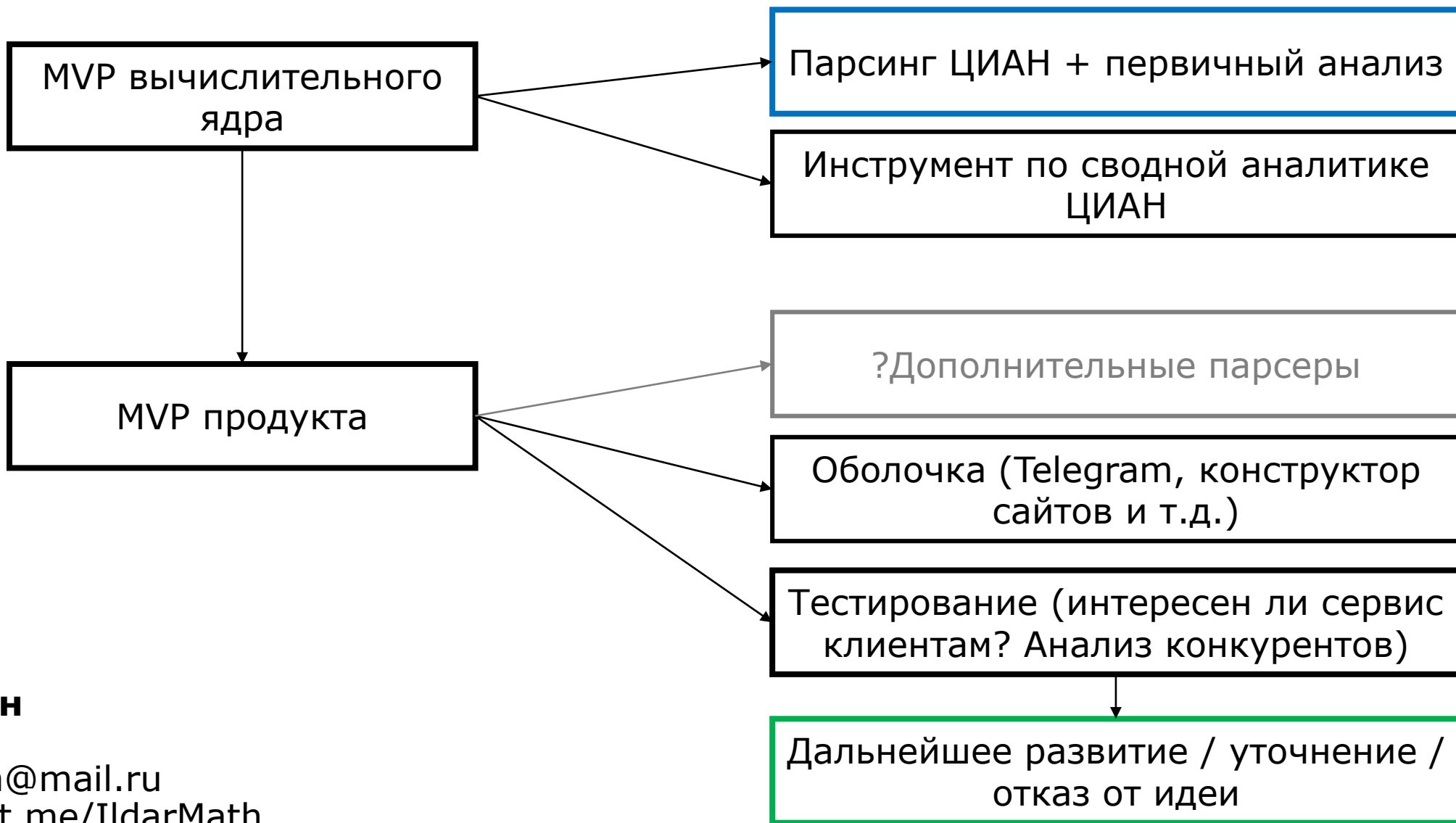


# Схема проекта

Сервис поиска интересных предложений для покупки недвижимости



**Ильдар Абдулин**

Контакты:

Email [nakiullovich@mail.ru](mailto:nakiullovich@mail.ru)

Telegram <https://t.me/IldarMath>

# Спринт №1: Парсинг ЦИАН + первичный анализ

**Цели:** выгрузить объявления по продаже жилой недвижимости третьего транспортного кольца Москвы, стоимость которых ниже рыночной оценки.

**Входные данные:** данные о продаже недвижимости в Москве с сайта ЦИАН.

**Проблемы:** наличие фейков.

## Задачи:

1. Разработать парсер датасета с сайта ЦИАН (минимальный набор факторов: широта и долгота расположения дома, общая площадь, количество комнат, материал стен, этаж квартиры и этажность здания, адрес, округ, тип продажи);
2. Провести обработку пропусков и поиск аномалий;
3. Первичный анализ факторов, фича-инжиниринг исходя из бизнес логики (пример – расстояние до центра);
4. Провести вывод объявлений в .xlsx документ со стоимостью ниже выборочного среднего значения похожих объявлений.

**Методы и библиотеки:** Jupyter Notebook, Python 3, pandas, scikit-learn, matplotlib, geopy.

# Резюме результата спринта

**Цель спринта:** выгрузить объявления по продаже коммерческой недвижимости третьего транспортного кольца Москвы, стоимость которых ниже рыночной оценки.

**Результат спринта:** проведен обзор ограничений парсинга ЦИАН, разработан скрипт для выгрузки объявлений раз в 1-7 дней, разработана грубая модель оценки стоимости.

**Шаг 1.** Парсинг коммерческих объявлений ЦИАН.

**Результат:** .xlsx датафрейм (253 объекта).

По парсеру: можно начать запускать каждый день с выгрузкой полного кода HTML страницы, а сам постобработчик для формирования .xlsx совершенствовать по мере потребности данных в модель (постобработчик есть в грубом формате для 200+ факторов и в аккуратном формате для 10 факторов). Это убережет нас от ситуаций с изменением конфигураций исходного кода страницы.

**Шаг 2.** Фильтрация данных для обучения модели .

**Результат:** Подвыборка А – наблюдения со «справедливой» стоимостью (143 объекта).

По стоим.>15 млн – 185 объектов (много продаж бизнеса, а не недвижимости);

Не включаются наблюдения ниже 20% относительно средней стоимости – 143 объекта.

**Шаг 3.** Обучение модели на подвыборке А.

**Результат:** обученная модель на подвыборке А.

По оценке средней стоимости: медиана ошибки - 5%, добавив расстояния до красной площади, Арбат, Тверского района и т.д. Модель пока что переобучена.

**Шаг 4.** Вывод списка недвижимости со стоимостью ниже рыночной оценки вне выборки А.

**Результат:** Топ 32 (вне А) наблюдений по отклонению фактической стоимости от оценочной.

Минусы: в датасете по прежнему много продаж бизнеса и пока что быстрого решения по их отлову нет. Из-за этого при выводе Топы по отклонению фактической от реальной на первых местах идут аренды бизнеса.

# Обзор парсинга ЦИАН: открытые инструменты

URL: <https://github.com/lenarsaitov/cianparser>

## Возможности:

- Конфигурация входных данных по 7 параметрам: аренда/продажа, город, вид жилья – квартира/таунхаус/и т.д., кол-во комнат, страница начала сбора данных, страница конца сбора данных, скорость сбора, собственник;
- Конфигурация выходных данных по 2 параметрам: кодировка, выгрузка в формате .csv.

## Выгружаемые данные:

Признаки, получаемые в ходе сбора данных

- **district** - район
- **underground** - метро
- **street** - улица
- **floor** - этаж
- **floors\_count** - общее количество этажей
- **total\_meters** - общая площадь
- **living\_meters** - жилая площади
- **kitchen\_meters** - площадь кухни

- **rooms\_count** - количество комнат
- **year\_construction** - год постройки здания
- **price** - стоимость
- **price\_per\_m2** - стоимость на квадратный метр
- **author** - автор объявления
- **author\_type** - тип автора
- **phone** - номер телефона в объявлении
- **link** - ссылка на объявление

Возможные значения поля **author\_type**: **real\_estate\_agent** - агентство недвижимости, **homeowner** - собственник, **realtor** - риелтор, **official\_representative** - ук оф.представитель, **representative\_developer** - представитель застройщика, **developer** - застройщик, **unknown** - без указанного типа.

# Обзор парсинга ЦИАН: открытые инструменты

## Тестирование:

[https://cian.ru/cat.php?engine\\_version=2&p=1&region=1&offer\\_type=flat&deal\\_type=rent&room2=1&room3=1&with\\_neigh\\_bors=0&type=4](https://cian.ru/cat.php?engine_version=2&p=1&region=1&offer_type=flat&deal_type=rent&room2=1&room3=1&with_neigh_bors=0&type=4) (14.04.23)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	author	author_ty	link	city	deal_type	accommo	floor	floors_co	rooms_co	total_me	price_per	price_per	commissi	year_of_c	living_me	kitchen_n	phone	district	street	underground		
2	Excellent real estate	real_esta	https://w	Москва	rent	flat	11	15	2	45	1444	65000	30	2019	30	10	7.96E+10	Даниловс	Автозаво	ЗИЛ		
3	ЦАН	real_esta	https://w	Москва	rent	flat	4	9	2	48	875	42000	50	1963	35	6	7.97E+10	Марфино	Академи	Фонвизинская		
4	ЦАН	real_esta	https://w	Москва	rent	flat	11	15	2	41	1585	65000	50	2019	24	12	7.96E+10	Даниловс	Автозаво	ЗИЛ		
5	Excellent real estate	real_esta	https://w	Москва	rent	flat	4	9	2	45	933	42000	30	1963	30	10	7.97E+10	Марфино	Академи	Фонвизинская		
6	ID 18835544	official_r	https://w	Москва	rent	flat	10	12	3	121	7024	850000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
7	ID 47020496	homeowr	https://w	Москва	rent	flat	26	44	2	50	2140	107000	0	2018	33.3	-1	7.5E+10	Пресненс	Ходынская	Улица 1905 года		
8	ID 18835544	official_r	https://w	Москва	rent	flat	5	12	2	80	6875	550000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
9	ID 24309666	homeowr	https://w	Москва	rent	flat	9	16	2	40	2125	85000	0	2018	22	9	7.91E+10	Филевски	Большая	Фили		
0	ID 18835544	official_r	https://w	Москва	rent	flat	7	12	2	112	8928	1000000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
1	ID 18835544	official_r	https://w	Москва	rent	flat	10	12	2	88	9090	800000	0	2015	-1	-1	7.96E+10	Тверской	Охотный ф	Охотный ряд		
2	Лиедел Инвестментс Л	official_r	https://w	Москва	rent	flat	42	75	2	98	3469	340000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
3	Лиедел Инвестментс Л	official_r	https://w	Москва	rent	flat	52	75	2	100	3300	330000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
4	Лиедел Инвестментс Л	official_r	https://w	Москва	rent	flat	52	75	2	101	3168	320000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
5	Лиедел Инвестментс Л	official_r	https://w	Москва	rent	flat	56	75	2	90	3666	330000	0	2013	-1	-1	7.92E+10	Пресненс	1-й Красн	Деловой центр		
6	ID 32373776	homeowr	https://w	Москва	rent	flat	2	9	2	56	803	45000	0	1973	46	8	7.5E+10	Ивановск	Молостое	Новогиреево		
7	ЦАН	real_esta	https://w	Москва	rent	flat	3	10	2	75	1733	130000	0	2020	55	10	7.97E+10	Пресненс	Мантулин	Выставочная		

Структура кода самого парсера перегружена ненужными для нас элементами (около 920 строк кода, нет комментариев).

Библиотека <https://github.com/lenarsaitov/cianparser> не работает с коммерческими объявлениями, а также объявлениями на продажу.

Аналогично с <https://www.youtube.com/watch?v=NR4lAlaQ1u4&t=2698s>, инструкции уже не актуальны.

**Вывод:** открытых работающих инструментов нет, поэтому пишем с нуля.

# Обзор парсинга ЦИАН: сложности

```
1 #Коммерческая
2 #url = 'https://www.cian.ru/sale/commercial/277193588/'
3 #Краткосрочная аренда
4 url = 'https://www.cian.ru/rent/flat/285274355/'
5
6 #Загрузка всего HTML кода страницы
7 session = cloudscraper.create_scraper()
8 res = session.get(url=url)
9 res.raise_for_status()
10 html = res.text
11
12 #Показать первые 200 символов исходного кода страницы
13 soup = BeautifulSoup(html, 'lxml')
14 print(str(soup.get_text())[0:200])
```

**Краткосрочная аренда  
жилой недвижимости**



Сдам двухкомнатные апартаменты 171м² ул. Охотный Ряд, 2, Москва, ЦАО, р-н Тверской м. Охотный ряд - база ЦИАН, объявление 285274355

**Коммерческая  
недвижимость**



**HTTPError: 403 Client Error: Forbidden for url:**  
8/

**Вывод:** ЦИАН защищает (капча) страницы объявлений от парсинга (в отличие от страниц поискового запроса). Судя по алгоритму cianparser данной проблемы не было 2 месяца назад.

# Обзор парсинга ЦИАН: сложности

## НО:

(исходный код страницы поискового запроса)

год постройки и материал стен (не отображается на странице поискового запроса)

```
84 "heatingType":null,"classType":null,"accessType":null,"materialType":null,"buildYear":1967,"  
85  
86
```

координаты объекта

```
83  
84 , "boundedBy": {"lowerCorner":{"lng":38.684018,"lat":43.760146}, "upperCorner":  
85  
86
```

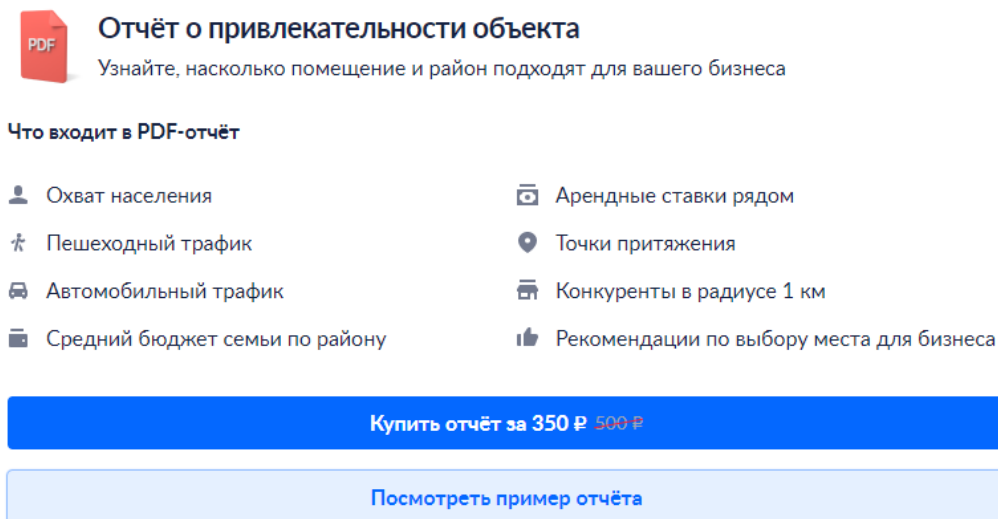
**Вывод:** однако вся информация (факторы) об объявлениях находится на странице поискового запроса.

# Обзор парсинга ЦИАН: сложности

## ЦИАН против парсинга:

- мгновенный сброс парсера на страницах объявления при использовании популярных библиотек парсинга (капча) – **фактически запрещает отправлять напрямую объявление в разрабатываемое приложение;**
- **нужно регулярно обновлять парсер:** [парсер который работал год назад](#) – не работает, парсер который работал 2 месяца работает частично (ЦИАН проводит регулярное обновление названий переменных в коде).

Возможная причина – ЦИАН сам является нашим конкурентом.



**Отчёт о привлекательности объекта**  
Узнайте, насколько помещение и район подходят для вашего бизнеса

Что входит в PDF-отчёт

Охват населения	Арендные ставки рядом
Пешеходный трафик	Точки притяжения
Автомобильный трафик	Конкуренты в радиусе 1 км
Средний бюджет семьи по району	Рекомендации по выбору места для бизнеса

Купить отчёт за 350 Р 500 Р

Посмотреть пример отчёта

Некоторых элементов PDF-отчета нет даже в примере (возможно сервис не развивается)  
[https://files.cian.ru/files/commercial/geo-analytics/report\\_example.pdf](https://files.cian.ru/files/commercial/geo-analytics/report_example.pdf)

**Выводы:** проблемы с парсингом будут всегда, на крайний случай есть сервисы для обхода тяжелых защит (например), имеет смысл изучить вопрос легальности парсинга ЦИАН.



# Обзор парсинга ЦИАН: выводы

## В ходе обзора найдено:

- Блокировка (проверка браузера, капча и т.д.) страниц самих объявлений. Загружаются только у объявлений с краткосрочной аренды (на момент 24.04), на остальных стоит блокировка. Страницы поискового запроса загружаются для всех.
- Обновление конфигурации исходного кода – меняются названия переменных раз в некоторый период.
- Готовые инструменты из открытого доступа не работают.
- ЦИАН сам предоставляет услуги по аналитике недвижимости, поэтому проблемы с парсингом будут всегда.

## Критерии качества:

- Минимальная конфигурации исходных данных;
- Минимальный объем кода для удобства поправок парсера;
- Упор на читабельность кода для дальнейшей передачи командам разработки парсеров;
- Без оптимизации по скорости;
- Разрабатывать отдельным модулем.

# Парсинг ЦИАН: описание скрипта

**Задача:** разработать парсер для выгрузки объявлений по фильтрам алгоритма экспертной оценки коммерческой недвижимости.

## Описание результата:

- Используемые библиотеки: BeautifulSoup, cloudscraper, библиотеку регулярных выражений re.
- Выгрузка в 2 форматах. Формат №1: 10 факторов с обработкой данных: \* Ссылка; \* Id продавца; \* Полная стоимость; \* Полная площадь; \* Номер этажа; \* Дата создания объявления; \* Широта, долгота; \* Район; \* Год построения здания. Формат №2: 202 фактора без обработки данных - перед загрузкой в модель данные нужно еще обрабатывать.).
- Фильтры поискового запроса: Коммерческая недвижимость: офис, торговая площадь, склад, свободное назначение, производство; стоимость До 35млн, площадь от 200 м2; Этаж>1; Прямая продажа.
- Парсер учитывает ранее найденные проблемы: 1. с быстрой переконфигурацией в случае изменения структуры исходного кода страницы (загрузка html и парсинг отдельно, использует библиотеку регулярных выражений); 2. нет привязки к конкретным названиям переменных для формата №2.

# Парсинг ЦИАН: результат

10 факторов с обработкой данных:

	Id Promotion	Id User	Full Price	Total Area	Floor Number	Creation Date	Lat	Lng	District	Build Year
0	www.cian.ru/sale/commercial/279443521	49659440	1450000	200	1	2022-10-25	55.8116457	37.65242095	"р-н Алексеевский"	
1	www.cian.ru/sale/commercial/279642085	49659440	1450000	420	2	2022-11-01	55.73822235	37.53170005	"р-н Дорогомилово"	
2	www.cian.ru/sale/commercial/287027078	28717	1500000	261	1	2023-05-04	55.663751	37.705329	"р-н Печатники"	2011
3	www.cian.ru/sale/commercial/286823156	78028008	1890000	240	3	2023-04-28	55.703773	37.640066	"р-н Даниловский"	
4	www.cian.ru/sale/commercial/284247750	49659440	2000000	220	1	2023-02-28	55.61815525	37.60375495	"р-н Чертаново Центральное"	
5	www.cian.ru/sale/commercial/274265152	279413	2000000	480	2	2022-06-02	55.782182	37.599911	"р-н Тверской"	
6	www.cian.ru/sale/commercial/285018884	49659440	2100000	300	1	2023-03-20	55.69886839999999	37.7762746500	"р-н Кузьминки"	
7	www.cian.ru/sale/commercial/286877301	29646	2300000	900	1	2023-05-01	55.677829	37.717034	"р-н Печатники"	
8	www.cian.ru/sale/commercial/284484685	203808	2500000	200	1	2023-03-06	55.852137	37.354582	"р-н Митино"	
9	www.cian.ru/sale/commercial/277358165	49659440	2900000	211	1	2022-08-22	55.65456895	37.7372002000	"р-н Маюино"	

203 фактора без обработки данных:

FeatureId	Warning	Marking	Offer	Unized	Time	ription	Min	ssport	Verif	FeatureLa	enger	LiftsC	"commercialOwnership":	mmmercialC	Moderatio	icePerMet	atPrice	Total	dedFrom	ifiedDocu	trackingEn	totalPer
	false	null	"3 недели	[68823851	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	false	true	null		
	false	null	"2 недели	[16688262	false	[]		null		{"agentAccountType":null	false		{"isUserIc	null	null	null	false	true	true	null		
	false	null	"вчера"	[28608300	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	true	true	null		
	false	null	"3 недели	[71836029	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	false	false	null		
	false	null	"2 недели	[8051077	true	[]		null		{"ownerType":null	true		{"isUserIc	null	null	null	false	true	true	null		
	false	null	"3 недели	[3402320	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	false	true	null		
	false	null	"вчера"	[39820564	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	true	true	null		
	false	null	"3 недели	[60535334	false	[]		null		{"agentAccountType":"age	false		{"isUserIc	null	null	null	false	false	true	null		
	false	null	"3 недели	[40592067	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	false	true	null		
	false	null	"3 недели	[41808060	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	false	true	null		
	false	null	"3 недели	[67199344	false	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	false	true	null		
	false	null	"неделю"	[42412721	true	[]		null		{"ownerType":null	false		{"isUserIc	null	null	null	false	true	true	null		

# **Модель:** описание, обучение модели на подвыборке А

## **Датасет - результат парсинга от 05.05:**

Убраны наблюдения со стоимостью менее 15 млн. руб. – много продаж бизнеса;

Подвыборка А: 143 наблюдений

Тренировочная выборка - 107 наблюдений.

Тестовая выборка - 36 наблюдений.

## **Целевая переменная:**

Стоимость объекта за квадратный метр.

## **Факторы:**

Полная площадь,

Номер этажа,

Год постройки.

## **Фича инжиниринг:**

Расстояния в метрах до географических центров районов Красной площади, Арбат,

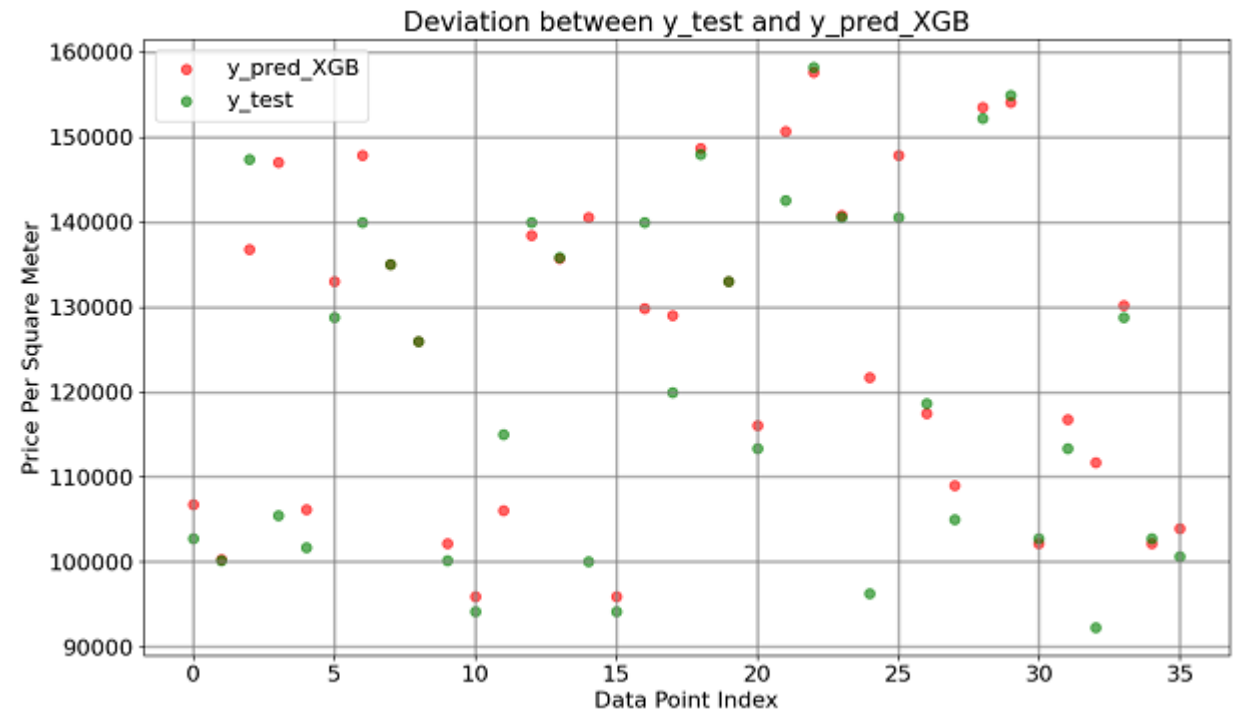
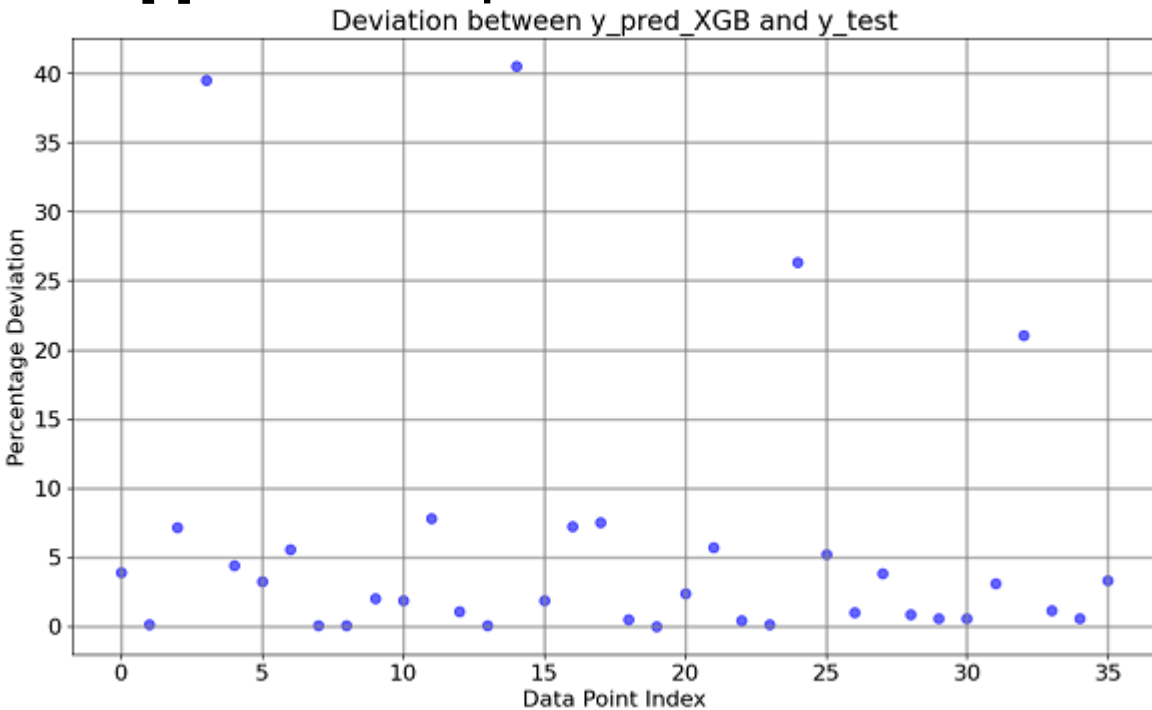
Тверского района, Пресненского района, Хамовники, Басманный район, район

Якиманка, Мещанский район, Нижегородский район.

OneHotEncoder(Район).

**Модель:** XGBRegressor(n\_estimators=120, max\_depth= 13, subsample=0.8, colsample\_bytree=0.8, reg\_alpha=0.2, reg\_lambda=0.8, gamma=0.9, min\_child\_weight=3, learning\_rate=0.1)+StandardScaler+GridSearch(max\_depth, n\_estimators).

# Модель: оценка качества



## Метрики:

Результаты попаданий в тестовой выборке:

Наблюдений для тестирования всего - 35 шт.

Наблюдений с отклонением менее 10% - 88%.

Наблюдений с отклонением более 10% менее 20% - 12% (4).

## Тестовая выборка

MDAPE: 2.2092325739107332

Mean Absolute Error (MAE): 6388.935149470545

Median\_absolute\_error: 2369.3993142609834

R2: 0.6517132532455242

Explained\_variance\_score: 0.7003061224362462

## Тренировочная выборка

MDAPE: 0.07827727272727127

Mean Absolute Error (MAE): 313.72982166290336

Median\_absolute\_error: 112.20525568182347

R2: 0.999299564900331

Explained\_variance\_score: 0.9992996060826165

**Вывод:** По графикам видно, что модель обладает явными признаками переобучения, так как:

- имеются существенные различия между средним значением и медианой ошибки, что свидетельствуют о наличии выбросов в ошибках модели;
- значительные расхождения в метриках между тестовой и тренировочной выборкам.

Средний показатель R2 указывает на хорошую способность модели реагировать на изменения входных факторов. Низкий уровень MDAPE говорит о высокой точности модели в большинстве случаев. В результате, модель может быть использована для грубой оценки в рамках очень похожих на тестовую выборку объявлений.

**Результат:** Вывод списка  
недвижимости со стоимостью ниже  
рыночной оценки вне выборки А

Сравнение прогноза модели по 32  
наблюдениям, которые ниже 20%  
относительно средней стоимости  
(вне подвыборки А).

**Выводы:** Интересные наблюдения  
начинаются при разности в оценке  
стоимости фактической и  
прогнозной в 50000 рублей  
(остальные объявления чаще всего  
это продажа бизнеса).

Id Promotion		Difference
112	www.cian.ru/sale/commercial/286618914	96737.896205
92	www.cian.ru/sale/commercial/279677345	95911.458333
154	www.cian.ru/sale/commercial/285510386	90149.800703
79	www.cian.ru/sale/commercial/284067991	84989.648438
94	www.cian.ru/sale/commercial/285147210	81744.941231
80	www.cian.ru/sale/commercial/282906588	81351.077415
82	www.cian.ru/sale/commercial/287075294	80959.187601
83	www.cian.ru/sale/commercial/286023464	74377.864810
84	www.cian.ru/sale/commercial/286252219	71358.542570
86	www.cian.ru/sale/commercial/285074405	69947.137473
88	www.cian.ru/sale/commercial/282490665	66819.616575
128	www.cian.ru/sale/commercial/280701859	65697.994591
81	www.cian.ru/sale/commercial/271123117	62435.213068
137	www.cian.ru/sale/commercial/283139637	54572.257576
109	www.cian.ru/sale/commercial/284054423	54221.429688
89	www.cian.ru/sale/commercial/278629056	53467.437500
0	www.cian.ru/sale/commercial/243822357	51532.492188
85	www.cian.ru/sale/commercial/273448602	49489.669612
87	www.cian.ru/sale/commercial/286383572	47859.488006
91	www.cian.ru/sale/commercial/233051066	45456.136606
95	www.cian.ru/sale/commercial/286862620	44331.787741
90	www.cian.ru/sale/commercial/271680846	40712.085938
177	www.cian.ru/sale/commercial/286795670	36689.346354
93	www.cian.ru/sale/commercial/286600286	34966.516955
115	www.cian.ru/sale/commercial/271680778	34316.721073
129	www.cian.ru/sale/commercial/285858817	34229.055350
124	www.cian.ru/sale/commercial/286798983	32763.522289
157	www.cian.ru/sale/commercial/285637040	31813.812599
98	www.cian.ru/sale/commercial/286090600	30005.878348
179	www.cian.ru/sale/commercial/286716588	28345.364419
1	www.cian.ru/sale/commercial/286755992	27533.929688
97	www.cian.ru/sale/commercial/286275050	25345.639509
184	www.cian.ru/sale/commercial/285126100	24305.412666
96	www.cian.ru/sale/commercial/258363973	23988.588542
113	www.cian.ru/sale/commercial/286866017	20048.038800
102	www.cian.ru/sale/commercial/281138548	19095.027429
100	www.cian.ru/sale/commercial/280233762	19015.745739
130	www.cian.ru/sale/commercial/281645736	16259.541780
132	www.cian.ru/sale/commercial/281604494	14786.487092
140	www.cian.ru/sale/commercial/279550014	13883.841452
131	www.cian.ru/sale/commercial/281645237	13285.987092
143	www.cian.ru/sale/commercial/285740541	12757.848362

# Далее

## Возможные пути развития:

- Улучшение качества данных:
  1. Научиться отлавливать продажу бизнеса. Есть идея начать с поиска слова "аренд" из текста объявления, проблема - текст в HTML перепутан местами. А также отсеивание объявлений, для которых неизвестна стоимость за квадратный метр.
  2. Расширение списка наблюдений за счет ослабления фильтров поиска,
  3. Расширение списка наблюдений за парсинга других источников (Авито),
- Улучшение модели:

Использовать k ближайших соседей (повысит интерпретируемость), факторы материал стен и этажности здания. Добавление факторов по наличию определенных слов в тексте объявления.
- Разработка оболочки, первые наброски GUI - <http://researchmachine.pythonanywhere.com/>

**Ильдар Абдулин**

Контакты:

Email [nakiullovich@mail.ru](mailto:nakiullovich@mail.ru)

Telegram <https://t.me/IldarMath>