Факультет информатики, математики
и компьютерных наук

Программная инженерия

Нижний Новгород
2023

АНАЛИЗ СТОИМОСТИ НЕДВИЖИМОСТИ НА PYTHON С ПРИВЯЗКОЙ К ГЕОДАННЫМ И ВИЗУАЛИЗАЦИЕЙ В ВИДЕ ИНТЕРАКТИВНЫХ КАРТ LEAFLET JAVASCRIPT ПОСРЕДСТВОМ БИБЛИОТЕКИ FOLIUM

Выпускная квалификационная работа

Руководитель:

старший преподаватель
кафедры информационных систем
и технологий

Маслова Екатерина Александровна

Рецензент:

старший преподаватель
кафедры прикладной математики
и информатики

Шадрина Елена Викторовна

Выполнила:

студентка группы ВПИ 20

Бондарева Татьяна Павловна

Объект исследования

Рынок жилой недвижимости Нижегородской области

Предмет исследования

Использование современных инструментов анализа и визуализации данных с целью последующего прогнозирования стоимости объектов недвижимости с применением методов машинного обучения

Задачи исследования:

1. рассмотреть современные инструменты визуализации и анализа данных;
2. подготовить данные для проведения анализа и обучения модели;
3. выполнить одномерный и двумерный анализ данных с использованием инструментов визуализации;
4. рассмотреть методы машинного обучения, подходящие для прогнозирования стоимости недвижимости;
5. создать и обучить модель на полученных данных, интерпретировать результаты.

Основная часть работы

1 Описание и предобработка данных для проведения анализа

- 1.1 Описание исходных данных
- 1.2 Используемые инструменты
- 1.3 Предварительная обработка исходных данных

2 Проведение анализа данных. Выявление взаимосвязей и закономерностей в данных

- 2.1 Одномерный анализ данных
- 2.2 Двумерный анализ данных

3 Прогнозирование стоимости объектов недвижимости с использованием алгоритма xgboost

- 3.1 Обзор методов машинного обучения для определения стоимости объектов недвижимости
- 3.2 Описание алгоритма XGBoost
- 3.3 Выбор метрики качества модели
- 3.4 Построение модели. Интерпретация результатов

категориальные признаки

- building_type
- object_type

геопространственные признаки

- geo_lon
- geo_lat

временные признаки

- date
- time

количественные признаки

- level
- levels
- rooms
- area
- kitchen_area
- price
- region

Исходный набор данных содержит 5 477 006 объектов недвижимости, расположенных в 84 регионах Российской Федерации. Каждый объект характеризуется 13 признаками.

python

Язык программирования, широко применяемый для анализа данных

ANACONDA

Дистрибутив Python, содержащий необходимые для анализа и обработки данных инструменты и библиотеки

SciPy

Библиотека для выполнения научных и инженерных расчетов

pandas

Библиотека для анализа и обработки структурированных данных

jupyter

Интерактивная среда для разработки на Python

NumPy

Библиотека для работы с многомерными массивами, включающая набор математических функций

matplotlib

Библиотека для визуализации данных. Применяется для построения и отображения диаграмм размаха, круговых диаграмм, столбчатых диаграмм, гистограмм

seaborn

Библиотека для создания статических графиков. Применяется для построения диаграмм размаха, круговых диаграмм, столбчатых диаграмм, гистограмм

kepler.gl

Инструмент для визуализации больших наборов геопространственных данных и создания интерактивных карт. Применяется для отображения анализируемых объектов на карте с использованием цветового кодирования

Folium

Библиотека визуализации геопространственных данных. Применяется для создания интерактивной карты с отображением информации об анализируемых объектах

scikit learn

Библиотека машинного обучения. Применяется для разделения набора данных на обучающую и тестовую выборки, а также для оценки качества модели

dmlc XGBoost

Библиотека, используемая в машинном обучении и предоставляющая функциональность для решения задач на основе алгоритма градиентного бустинга. Применяется для создания модели прогнозирования стоимости

поиск объектов недвижимости, расположенных в Нижегородской области	выделено 101 086 объектов недвижимости Нижегородской области
поиск пропущенных значений	пропущенные значения отсутствуют
поиск повторяющихся записей	найдено и удалено 11 дубликатов
поиск неверных значений	откорректированы отрицательные значения признаков price, rooms
поиск выбросов	удалены выбросы из признаков price, area и kitchen_area
удаление ненужных признаков	удалены признаки region, date и time, на основе признака date сформирован признак year

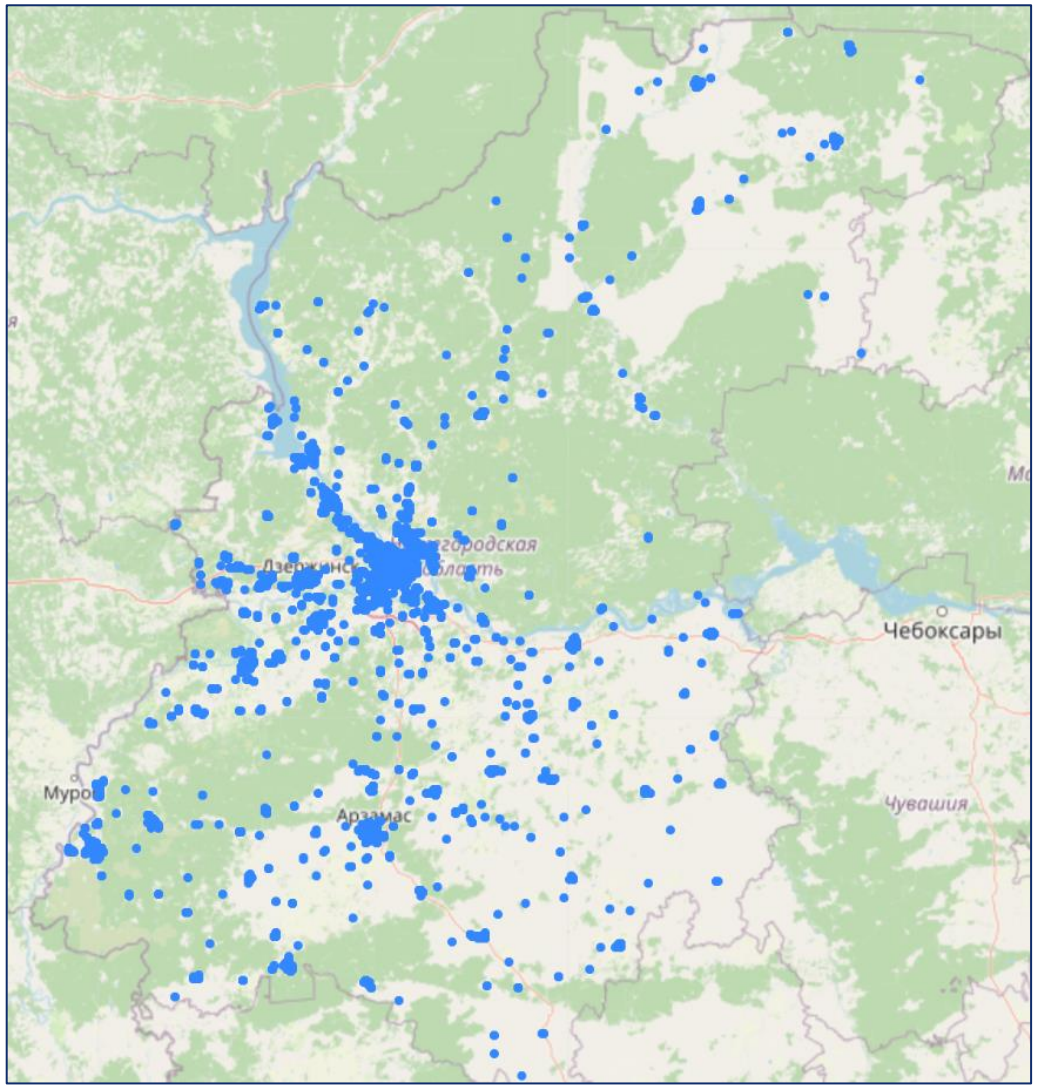
Исходный набор данных - 5 477 006 объектов, 13 признаков

	price	date	time	geo_lat	geo_lon	region	building_type	level	levels	rooms	area	kitchen_area	object_type
0	6050000	2018-02-19	20:00:21	59.805808	30.376141	2661	1	8	10	3	82.6	10.8	1
1	8650000	2018-02-27	12:04:54	55.683807	37.297405	81	3	5	24	2	69.1	12.0	1
2	4000000	2018-02-28	15:44:00	56.295250	44.061637	2871	1	5	9	3	66.0	10.0	1
3	1850000	2018-03-01	11:24:52	44.996132	39.074783	2843	4	12	16	2	38.0	5.0	11
4	5450000	2018-03-01	17:42:43	55.918767	37.984642	81	3	13	14	2	60.0	10.0	1

Обработанный набор данных – 79 506 объектов, 11 признаков

	price	geo_lat	geo_lon	building_type	level	levels	rooms	area	kitchen_area	object_type	year
2	4000000	56.295250	44.061637	1	5	9	3	66.0	10.0	1	2018
22	3843000	56.346027	43.871648	2	16	25	2	61.0	11.0	11	2018
23	2697200	56.346027	43.871648	2	6	25	1	44.0	20.0	11	2018
24	4214700	56.346027	43.871648	2	16	25	2	67.0	20.0	11	2018
25	3773700	56.346027	43.871648	2	12	25	2	60.0	12.0	11	2018

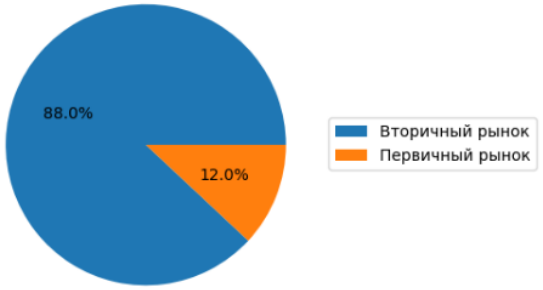
Наибольшее количество продаваемых объектов сосредоточено в административных центрах Нижегородской области: Нижнем Новгороде, Дзержинске, Кстово, Балахне, Павлове, Богородске, Городце, Арзамасе, Выксе и на Бору



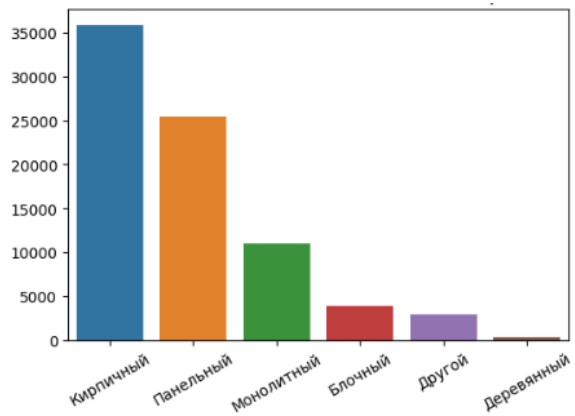
Факультет информатики, математики и компьютерных наук

Анализ стоимости недвижимости на Python с привязкой к геоданным и визуализацией в виде интерактивных карт

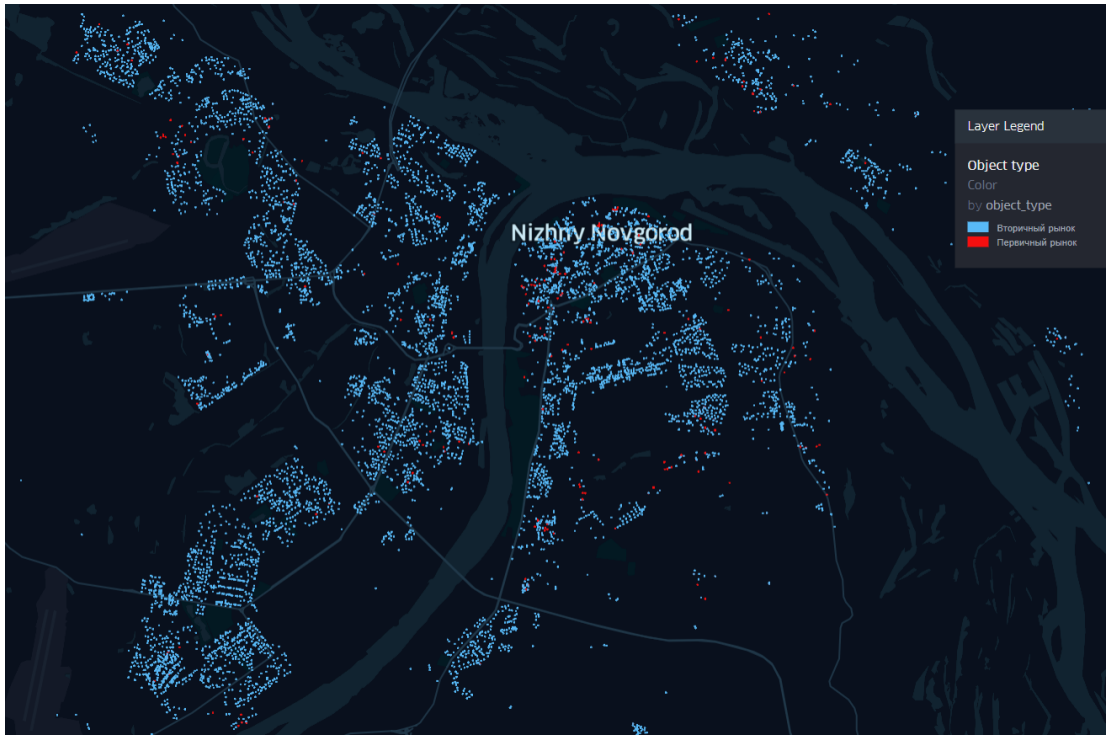
Проведение анализа данных. Выявление взаимосвязей и закономерностей в данных. Одномерный анализ данных

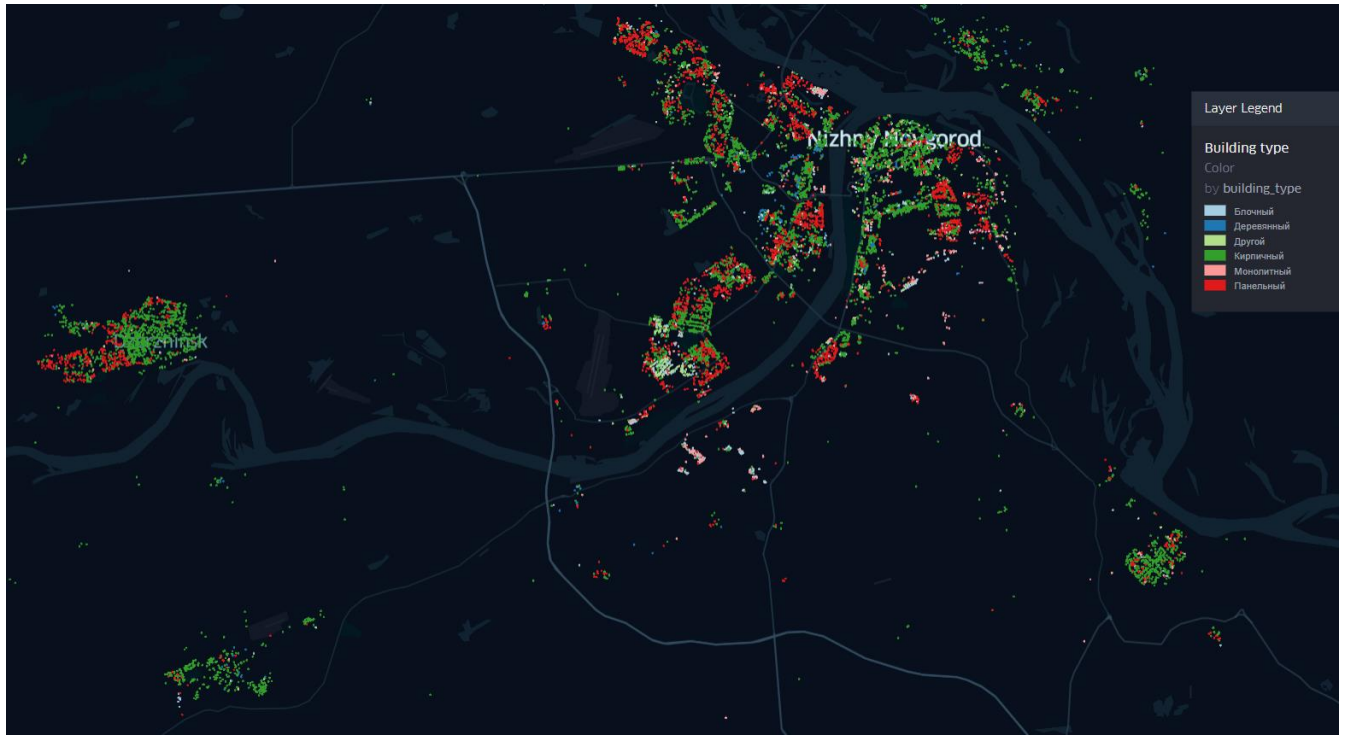


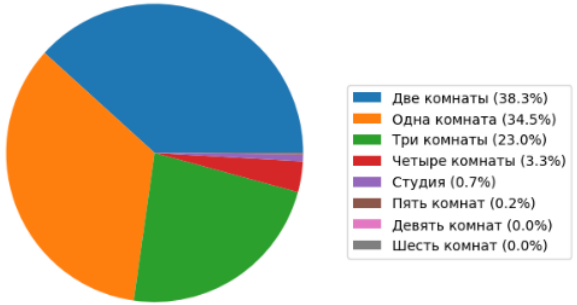
Наибольшая доля продаваемых объектов относится к рынку вторичного жилья. Их количество составляет 70 005 из 79 506



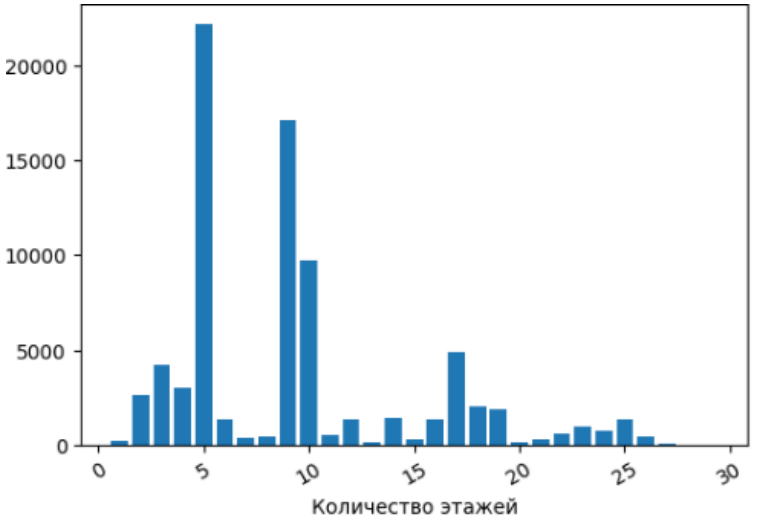
Наибольшее количество выставленных на продажу объектов – это кирпичные здания. Их число составляет 35 903 из 79 506

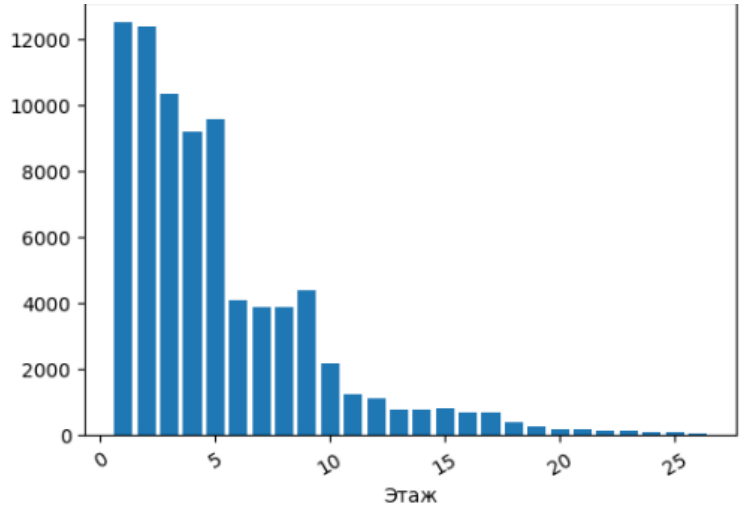




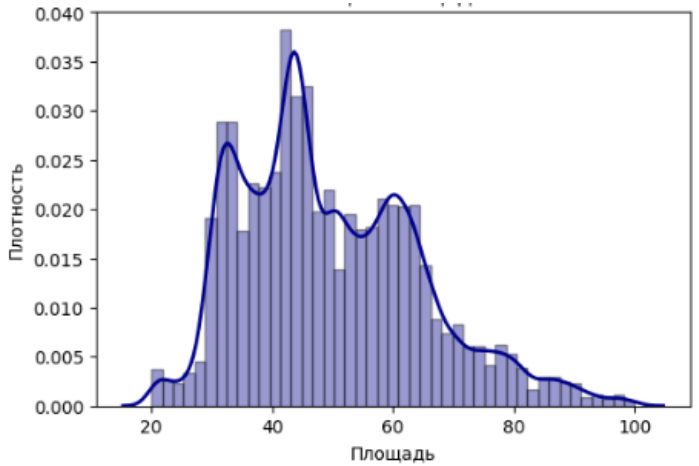


Наибольшее количество объявлений подано о продаже двух- и однокомнатных квартир. Их число составляет 30 443 и 27 404 соответственно

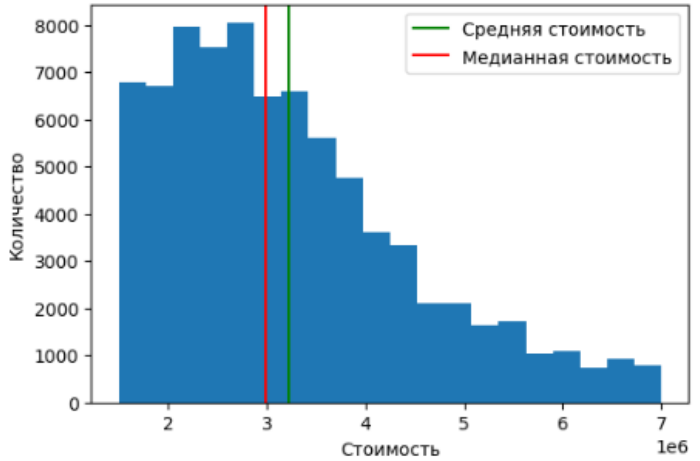




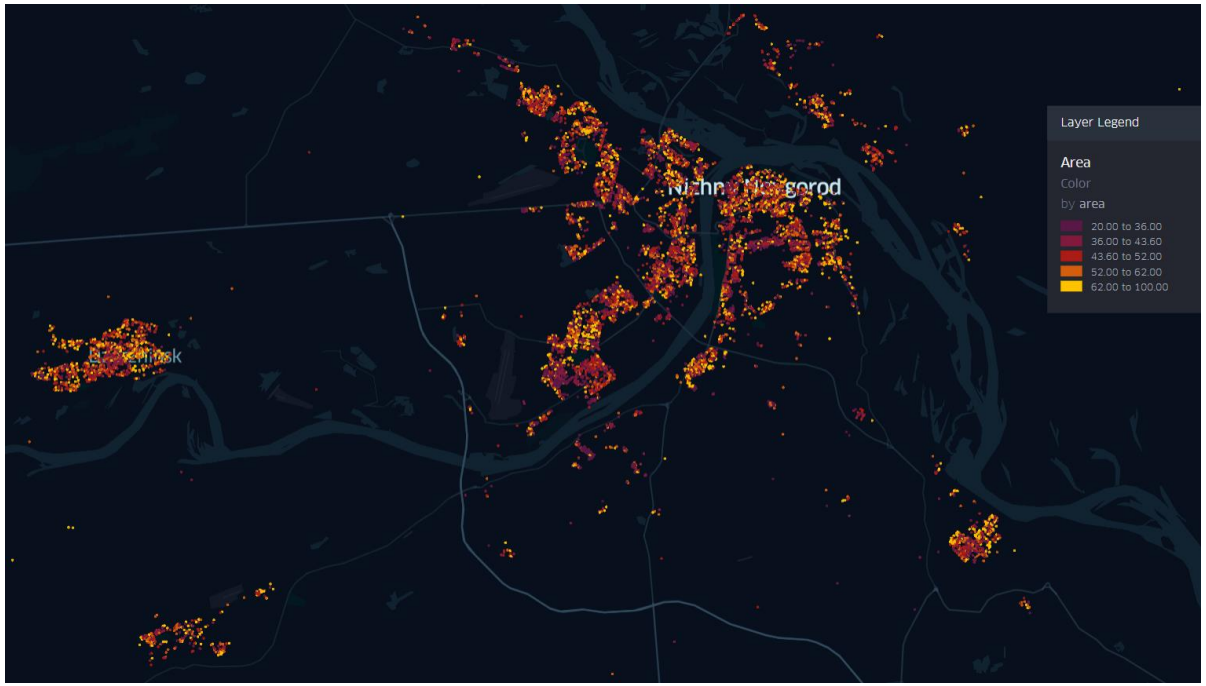
Большинство продаваемых квартир находятся в пятиэтажных зданиях и расположены преимущественно на первых двух этажах

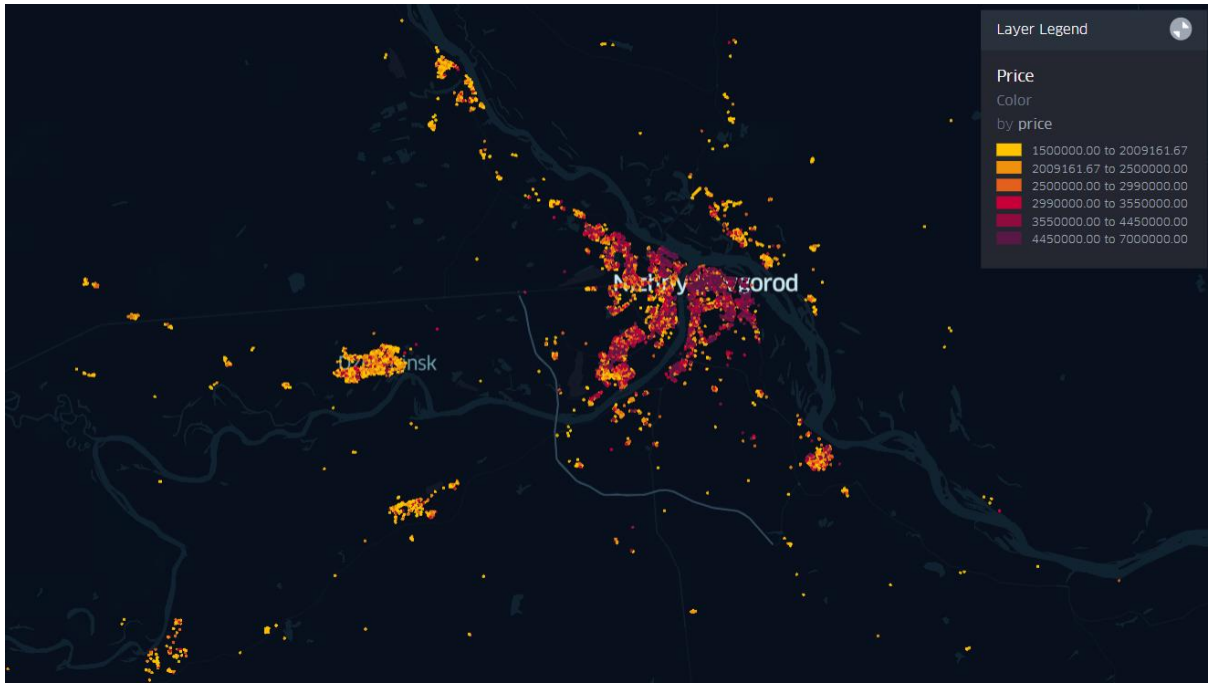


Средняя площадь продаваемых объектов составляет ориентировочно 47 м^2 с разбросом значений примерно 15 м^2, при этом 25% от общего количества имеют площадь ниже 38 м^2, а 75% - ниже 60 м^2



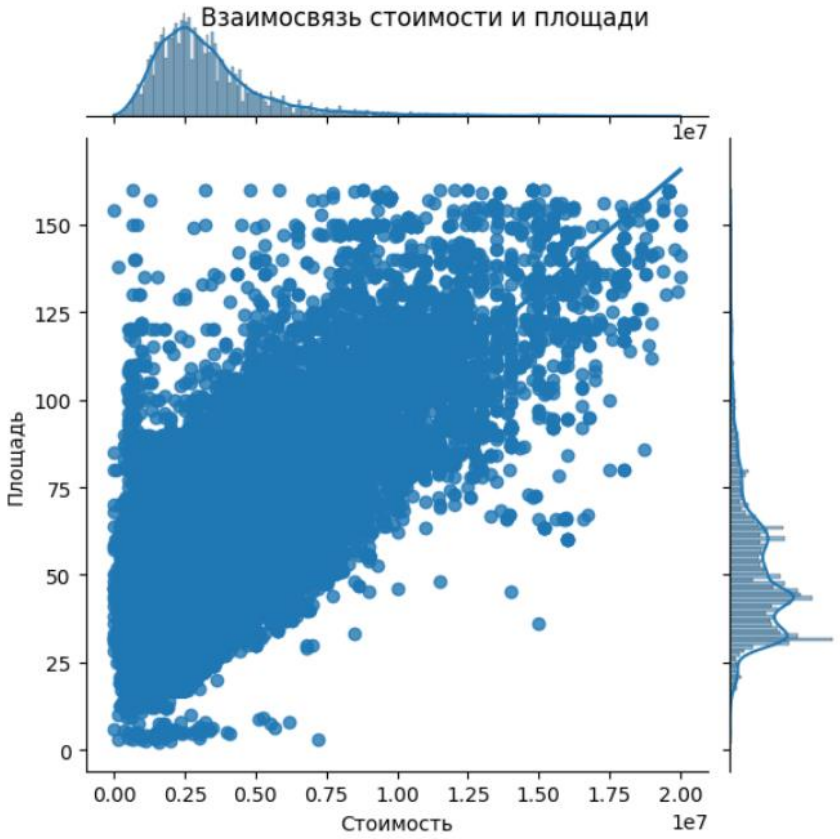
Средняя стоимость продаваемых объектов составляет ориентировочно 3,0 млн руб. с разбросом значений примерно 1,3 млн руб., при этом 25% от общего количества имеют стоимость ниже 2,3 млн руб., а 75% - ниже 3,9 млн руб



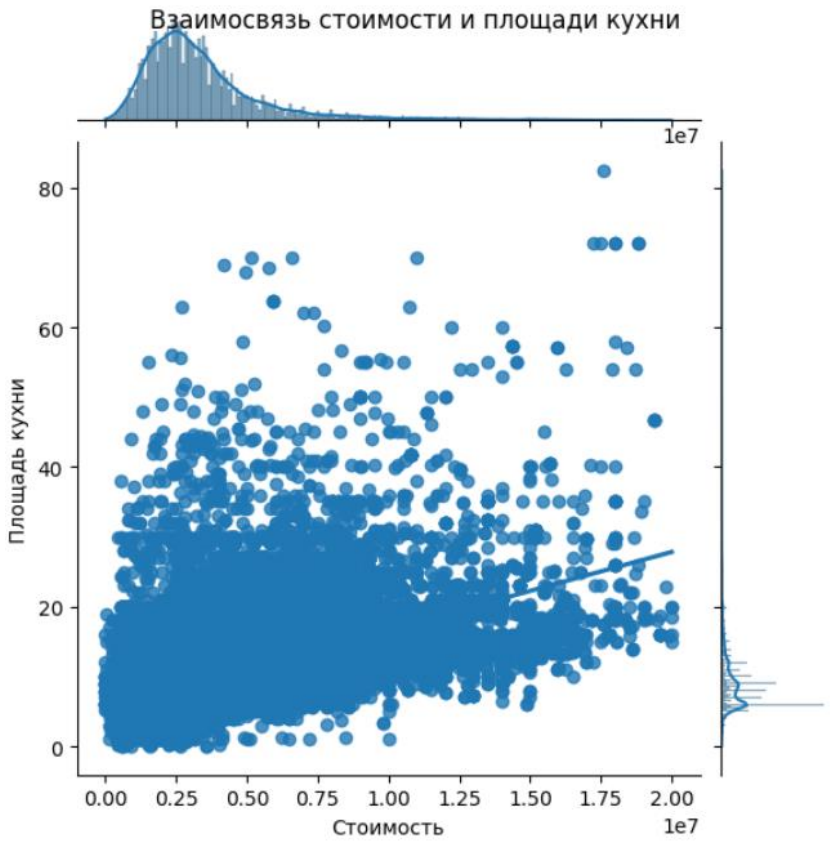


В результате проведения одномерного анализа данных можно сделать вывод, что средняя квартира, выставленная на продажу в Нижегородской области в период с 2018 года по 2021 год, имеет следующие характеристики:

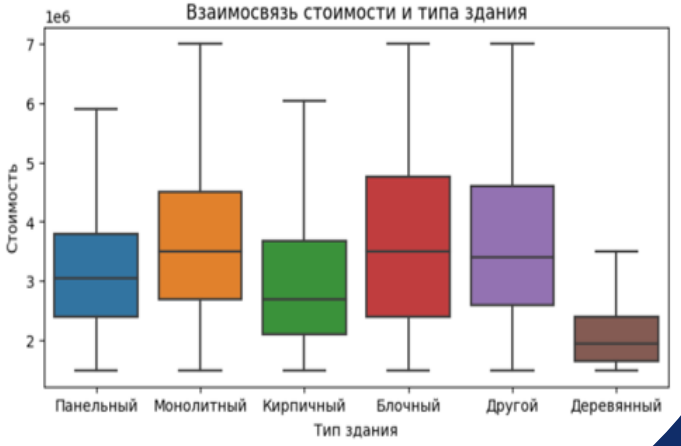
- **общая площадь около 47 м^2;**
- **двухкомнатная;**
- **расположена в кирпичном пятиэтажном доме;**
- **преимущественно на первом этаже;**
- **относится к рынку вторичного жилья;**
- **имеет стоимость около 3 млн руб.**



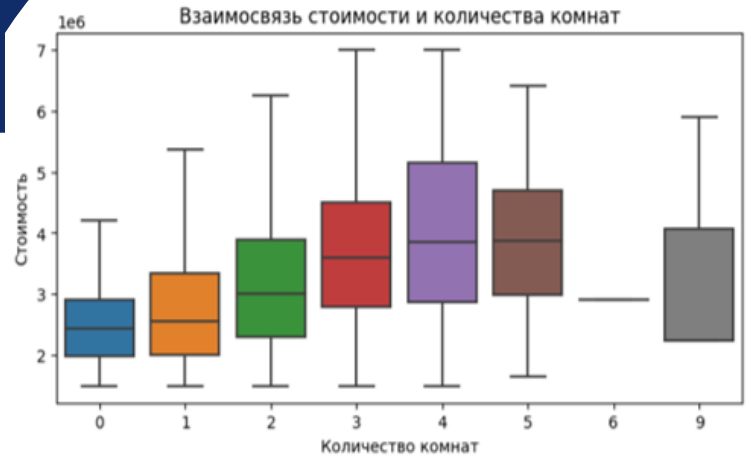
Коэффициент корреляции между признаками price и area составляет 0,75. Это говорит о наличии сильной линейной зависимости между ними – чем больше общая площадь квартиры, тем выше ее стоимость



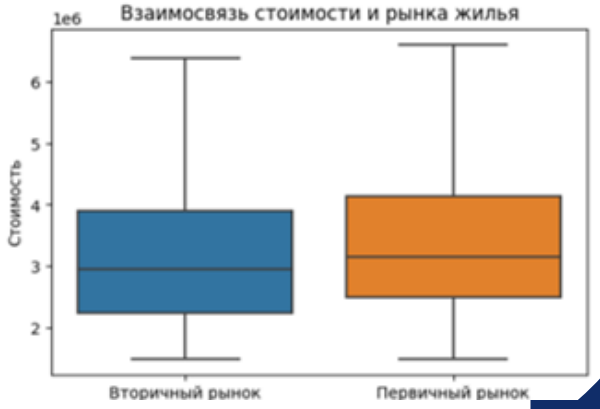
Коэффициент корреляции между признаками price и kitchen_area составляет 0,54. Это говорит о наличии средней линейной зависимости между ними – стоимость квартиры растет с увеличением площади кухни



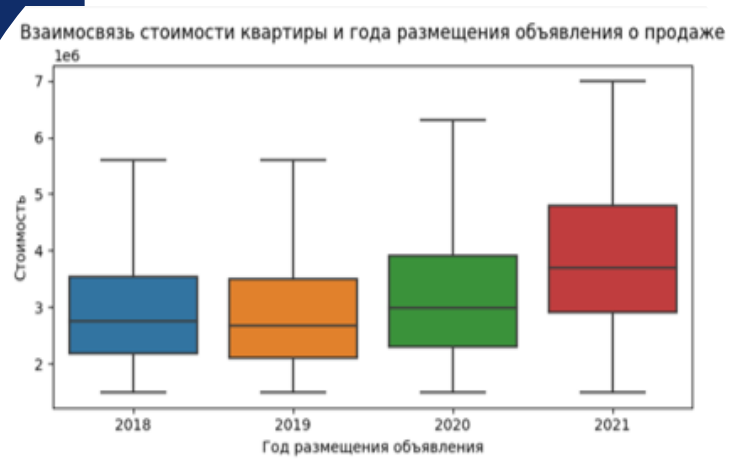
С увеличением количества комнат в квартире ее средняя стоимость увеличивается. Это справедливо, пока число комнат не превышает четырех. Средняя стоимость квартир с бóльшим количеством комнат соизмерима со стоимостью четырехкомнатной квартиры, либо ниже ее



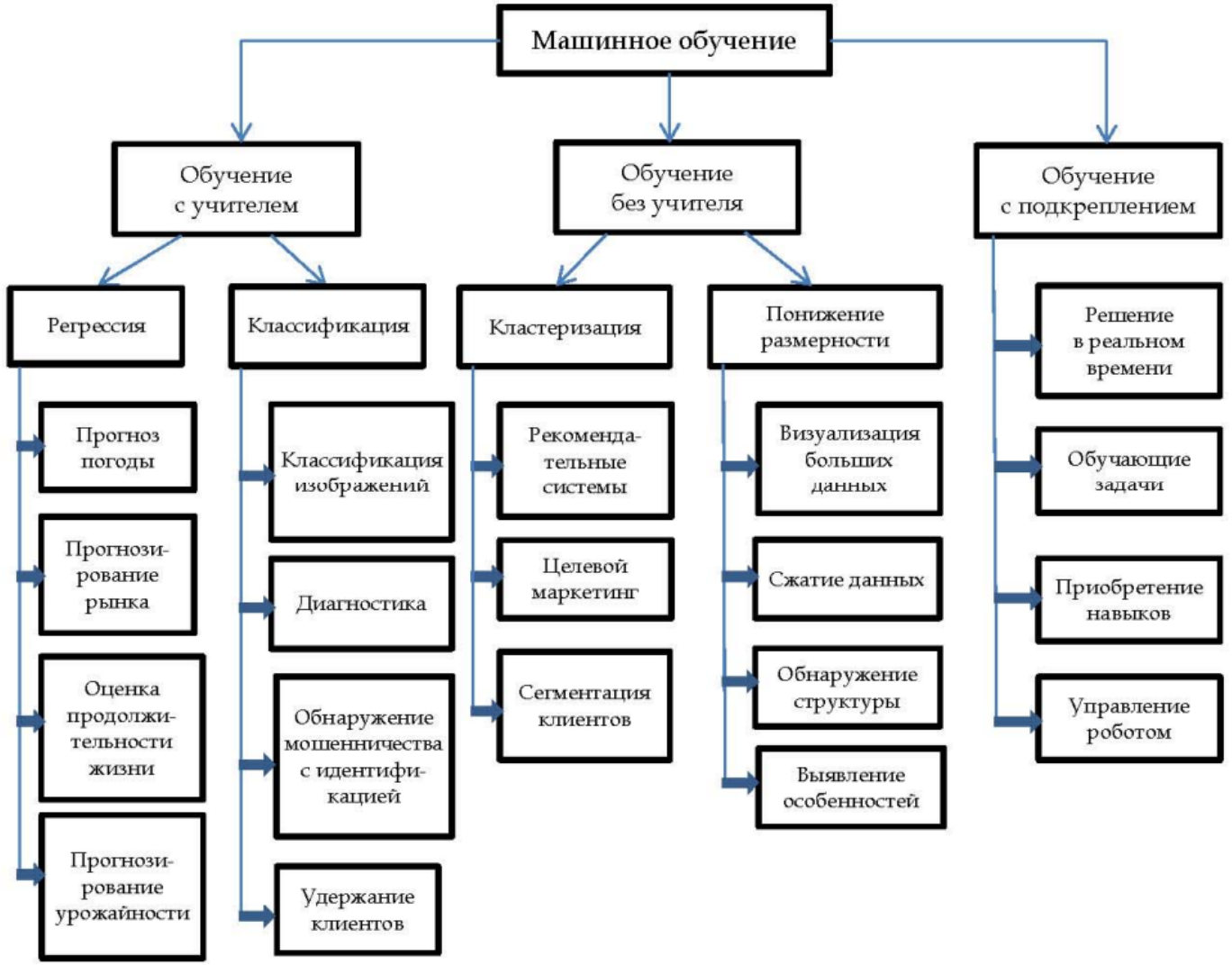
Средняя стоимость квартир в монолитных и блочных домах выше, чем в панельных и кирпичных. Наименьшую стоимость имеют деревянные дома



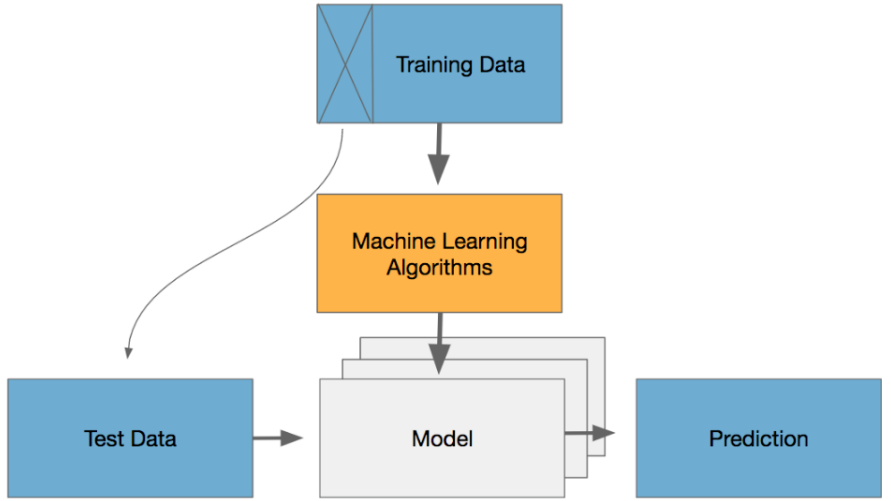
В 2018 и 2019 годах средняя стоимость квартир находилась примерно на одном уровне, в 2020 году отмечался небольшой рост и более резкий подъем в 2021 году

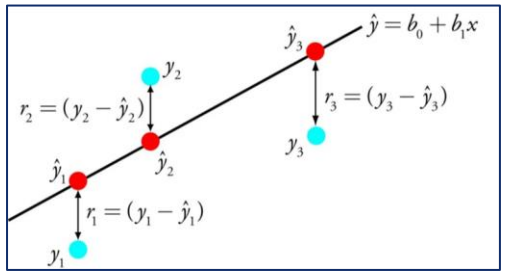


Средняя стоимость квартир на первичном рынке несущественно выше, чем на вторичном

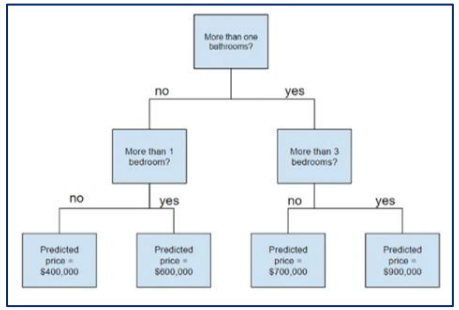


Для прогнозирования стоимости недвижимости необходимо решить задачу регрессии, которая относится к типу машинного обучения - обучение с учителем. Обучение с учителем предполагает наличие тренировочных данных, на основе которых создается модель, делающая прогнозы на новых ранее не встречавшихся данных.

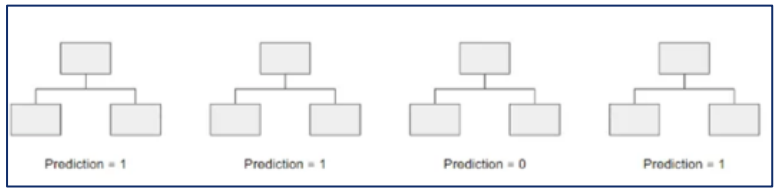




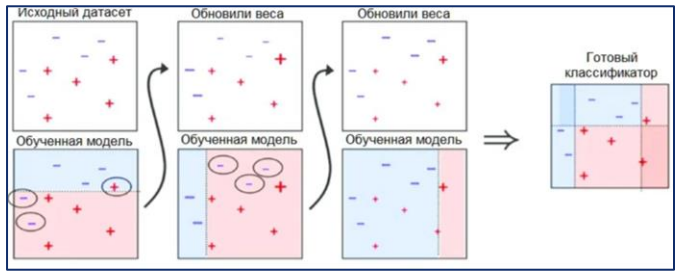
Линейная регрессия



Деревья принятия решений



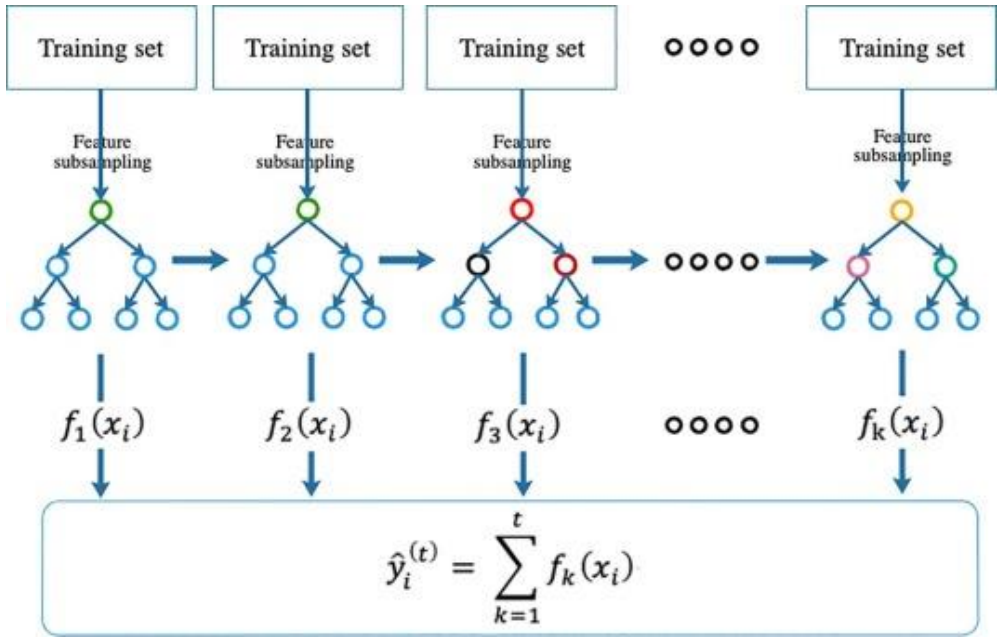
Случайный лес

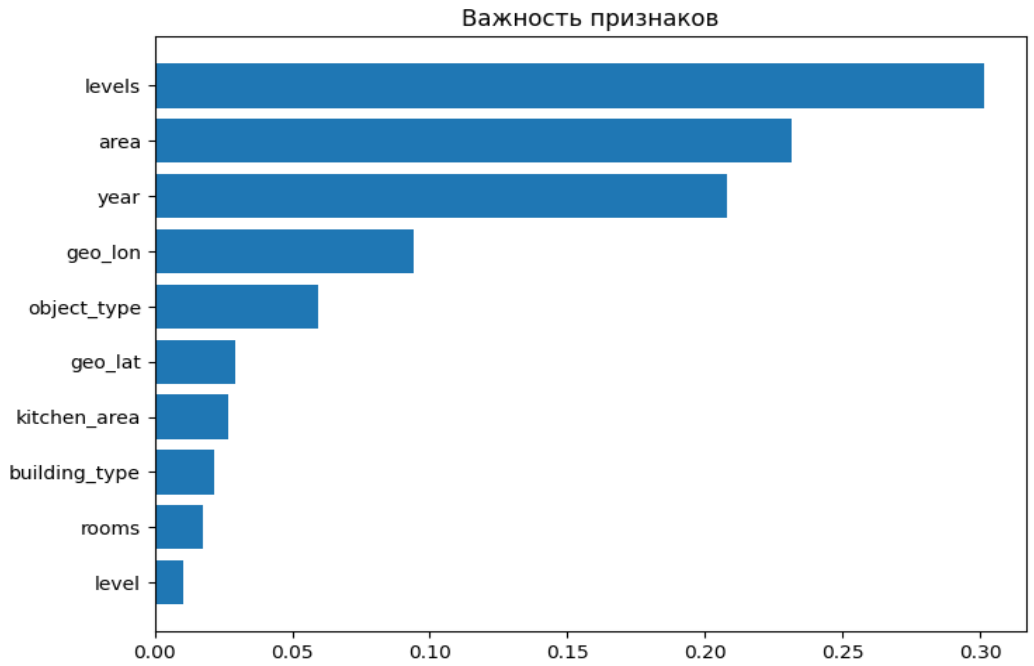


Градиентный бустинг

1. **Лейфер Л.А., Чёрная Е.В. Массовая оценка объектов недвижимости на основе технологий машинного обучения. Анализ точности различных методов на примере определения рыночной стоимости квартир. Имущественные отношения в РФ № 3 (222), 2020**
2. **Алексеева Ю. А., Гусев К. А. Использование методов машинного обучения для определения стоимости объектов недвижимости. СибГУ науки и технологий имени академика М. Ф. Решетнёва. Решетневские чтения, 2022**
3. **Q. Truong, M. Nguyen, H. Dang, B. Mei Housing price prediction via improved machine learning techniques. Procedia Computer Science, vol. 174, pp. 433–442, 2020**

XGBoost (extreme gradient boosting)





Полученное значение коэффициента детерминации составляет $R^2 = 0{,}92$, что указывает на высокую степень соответствия модели данным

Предварительная обработка данных	Одномерный анализ данных	Интерактивные карты	Двумерный анализ данных	Машинное обучение
подготовка набора данных для проведения анализа	описание средней квартиры, продаваемой в Нижегородской области с 2018 г. по 2021 г.	где сосредоточено наибольшее количество анализируемых объектов как меняется значение той или иной характеристики объекта в зависимости от его местоположения	выявление зависимости стоимости объекта от его характеристик	прогнозирование стоимости объекта недвижимости

Репозиторий
https://github.com/Tatiana0611/PYTHON-REAL-ESTATE-VALUE-ANALYSIS


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •