МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science Pro»

Слушатель Кропотова Александра Сергеевна

Москва, 2024

Содержание:

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

1.Постановка задачи

Актуальность темы

Прогнозирование стоимости жилья – актуальный вопрос как в России, так и в других странах мира. Несмотря на общеизвестные методы предсказания стоимости цены на квартиры, задача не теряет актуальности, поскольку вызывает интерес как покупателей квартир, так и компаний-застройщиков, государства и банков.

Сбор и подготовка данных

Для данной работы был использован датасет, данные которого были собраны с сайта для поиска недвижимости для покупки, продажи и аренды – Циан.ру (https://www.cian.ru), данные были собраны с помощью библиотеки ‘Beautiful Soup’. При сборе информации были использованы только данные о квартирах, данные о продавцах/агентствах недвижимости и иные сведения, не являющиеся характеристиками квартир, не использовались.

В целях формирования scv-файла в формате таблицы с признаками-характеристиками квартир для сбора данных были выбраны следующие признаки:

Название признака	Описание признака
price	Целевая переменная. Стоимость квартиры.
min_to_metro	Количество минут, затраченное на передвижение от Объекта недвижимости до ближайщей станции метро пешком
region_of_moscow	Округ Москвы: ● ЮАО – административный округ

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

	• СВАО – Северо-Восточный административный округ • ЗАО Западный административный округ • ЦАО Центральный-Восточный административный округ • ЮЗАО Юго-Западный административный округ • ВАО Восточный административный округ • СЗАО Северо-Западный административный округ • САО Северный административный округ • НАО Новомосковский административный округ • ЗелАО Зеленоградский округ города Москвы
total_area	Общая площадь Объекта недвижимости
living_area	Жилая площадь Объекта
floor	Этаж расположения квартиры
number_of_floors	Общее количество этажей в доме
construction_year	Год постройки (сдачи) здания, в котором находится квартира
is_new	Бинарный признак. 1 – новостройка, 0 – вторичное жилье
is_apartments	Бинарный признак. 1 – апартаменты, 0 – не апартаменты
ceiling_height	Высота потолка квартиры
number_of_rooms	Количество комнат в квартире

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Сайт Циан.ру имеет ограничение – 28 объявлений на 54 страницах , для просмотра большего количества объявлений нужно нажать на кнопку «показать больше». Так как используемая библиотека ‘Beautiful Soup’ не обладает данным функционалом, парсинг данных проводился в 5 этапов:

1. Получение ссылок с 54 страниц и дальнейший парсинг данных объявлений квартир с параметрами:

- студии и 1-комн. квартиры, новостройки, не апартаменты
- 2-комн. квартиры, новостройки, не апартаменты
- 3-комн. квартиры, новостройки, не апартаменты

2. Получение ссылок с 54 страниц и дальнейший парсинг данных объявлений квартир с параметрами:

- студии и 1-комн. квартиры, новостройки, апартаменты
- 2-комн. квартиры, новостройки, апартаменты
- 3-комн. квартиры, новостройки, апартаменты

3. Получение ссылок с 54 страниц и дальнейший парсинг данных объявлений квартир с параметрами:

- студии и 1-комн. квартиры, вторичное жилье, не апартаменты
- 2-комн. квартиры, вторичное жилье, не апартаменты
- 3-комн. квартиры, вторичное жилье, не апартаменты

4. Получение ссылок с 54 страниц и дальнейший парсинг данных объявлений квартир с параметрами:

- студии и 1-комн. квартиры, вторичное жилье, апартаменты
- 2-комн. квартиры, вторичное жилье, апартаменты
- 3-комн. квартиры, вторичное жилье, апартаменты

5. Объединение получившихся четырех мини-датасетов в единую таблицу. Итоговый размер датасета: (1937 строк, 12 столбцов).

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Благодаря тому, что данные были собраны парсером, полученный датасет не содержит пропусков в целевой переменной, то есть удалять строки нет необходимости.

Количество пропущенных значений в переменных:

min_to_metro: 29

region_of_moscow: 18

total_area: 18

living_area: 673

floor: 33

number_of_floors: 289

construction_year: 692

is_new: 0

is_apartments: 0

ceiling_height: 1306

number_of_rooms: 0

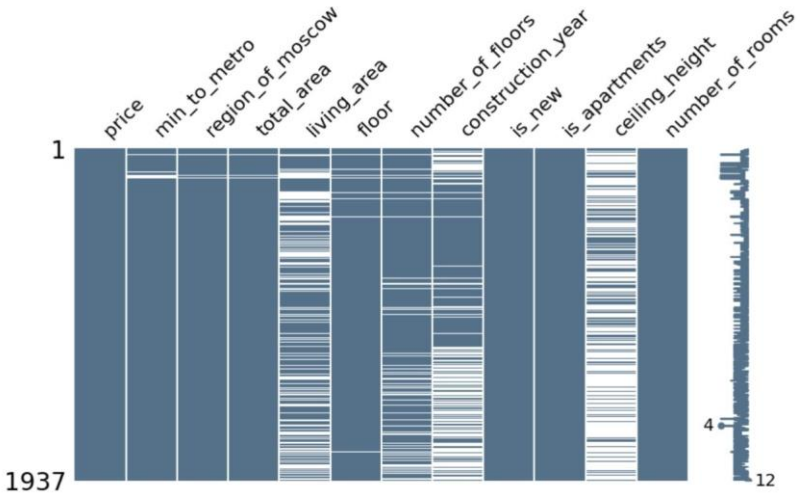


Рисунок 1 – Визуализация пропусков в датасете с рисенением библиотеки missingno

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

2.Характеристики переменных в датасете

	price	min_to_metro	region_of_moscow	total_area	living_area	floor	number_of_floors	construction_year	is_new	is_apartments	ceiling_height	number_of_rooms
price	1.000000	-0.266555	-0.240513	0.782794	0.692339	0.173664	0.088058	0.078912	-0.097024	0.080837	0.363075	0.458756
min_to_metro	-0.266555	1.000000	0.173402	-0.271201	-0.215160	-0.141498	-0.130139	0.103726	0.127824	-0.110511	-0.147725	-0.129680
region_of_moscow	-0.240513	0.173402	1.000000	-0.203291	-0.176903	-0.128194	-0.134269	0.066288	0.118369	-0.139442	-0.166696	-0.124049
total_area	0.782794	-0.271201	-0.203291	1.000000	0.900997	0.288062	0.231385	0.037386	-0.149324	0.062057	0.358493	0.708081
living_area	0.692339	-0.215160	-0.176903	0.900997	1.000000	0.260766	0.245794	-0.038494	-0.236518	0.062045	0.287757	0.696982
floor	0.173664	-0.141498	-0.128194	0.288062	0.260766	1.000000	0.775255	0.237217	-0.029638	0.134029	0.197921	0.175736
number_of_floors	0.088058	-0.130139	-0.134269	0.231385	0.245794	0.775255	1.000000	0.320282	0.064298	0.081422	0.168462	0.094108
construction_year	0.078912	0.103726	0.066288	0.037386	-0.038494	0.237217	0.320282	1.000000	0.513752	0.266663	0.245801	-0.050676
is_new	-0.097024	0.127824	0.118369	-0.149324	-0.236518	-0.029638	0.064298	0.513752	1.000000	-0.066651	0.244369	-0.219900
is_apartments	0.080837	-0.110511	-0.139442	0.062057	0.062045	0.134029	0.081422	0.266663	-0.066651	1.000000	0.286919	0.065411
ceiling_height	0.363075	-0.147725	-0.166696	0.358493	0.287757	0.197921	0.168462	0.245801	0.244369	0.286919	1.000000	0.149365
number_of_rooms	0.458756	-0.129680	-0.124049	0.708081	0.696982	0.175736	0.094108	-0.050676	-0.219900	0.065411	0.149365	1.000000

Рисунок 2 – Матрица корреляций признаков

min_to_metro

Более половины квартир в датасете находятся в 5-15 минутах от метро пешком.

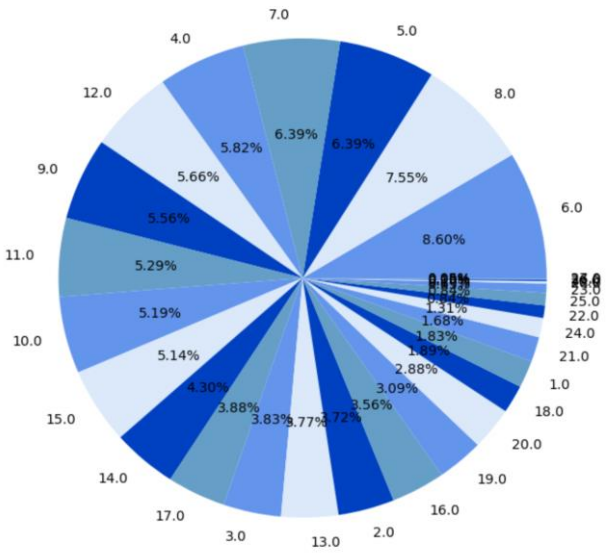


Рисунок 3 - Визуализация количества квартир по дальности от метро

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

region_of_moscow

Самый часто представленный район Москвы в датасете—ЦАО (27% квартир), на втором месте—ЗАО (13%), на третьем-САО(10%).

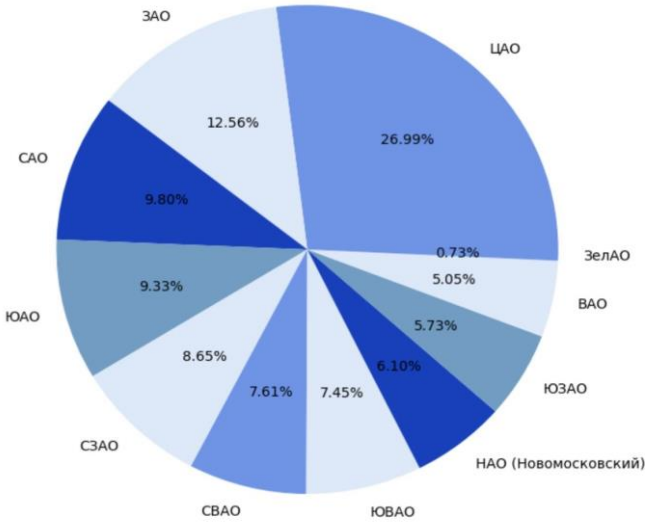
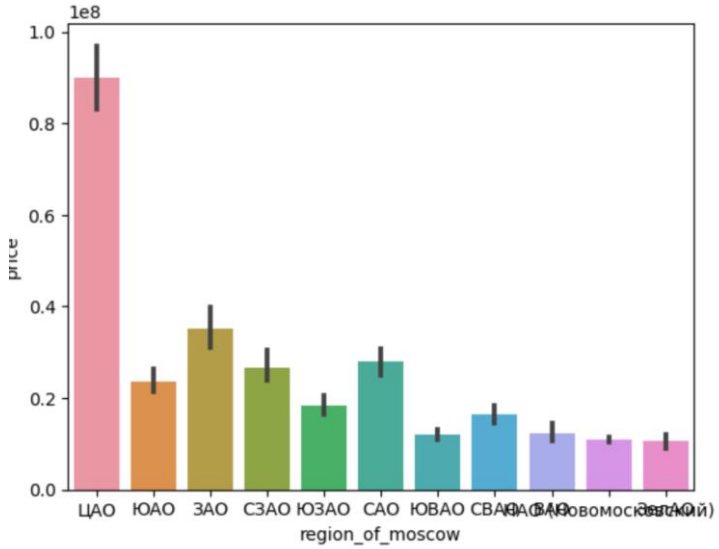


Рисунок 4 - Визуализация количества квартир по району

При этом стоит обратить внимание на то, что в датасете ЦАО является районом с самыми дорогими квартирами.



Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Рисунок 5 - Визуализация стоимости квартир в зависимости от района расположения

total_area

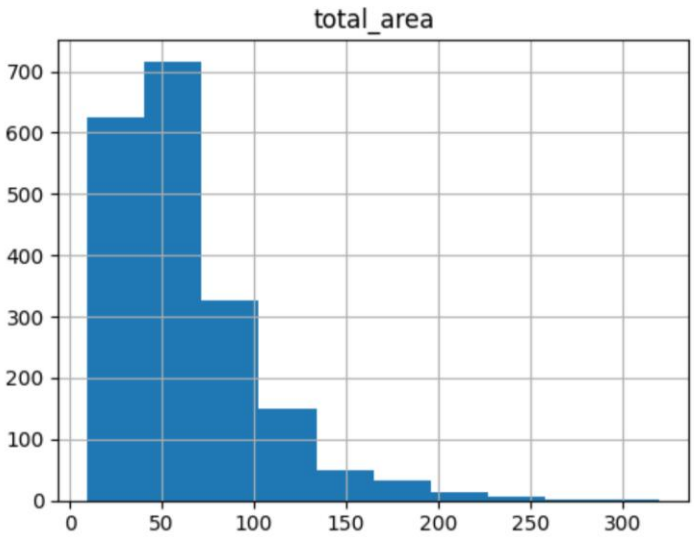


Рисунок 6 - Распределение признака общей площади квартиры

Данную переменную можно назвать главной для целей предсказания стоимости Объекта недвижимости, поскольку данный признак имеет самую высокую корреляцию с целевой переменной и четко прослеживающуюся линейну связь.

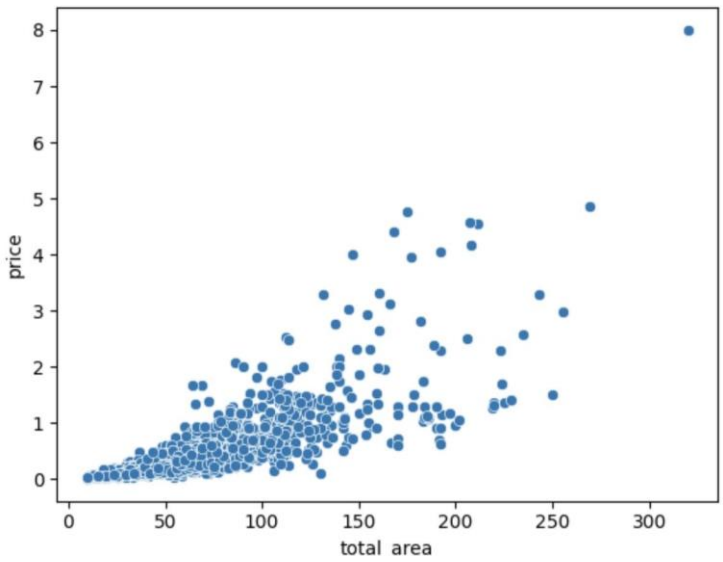


Рисунок 7 - Зависимость цены от общей площади

Необходимо обратить внимание на что, что признак общей площади на графика «ящик с усами» показывает «выбросные» значения - присутствуют квартиры с большим количеством квадратных метров, но и стоимости этих квартир соответствующая.

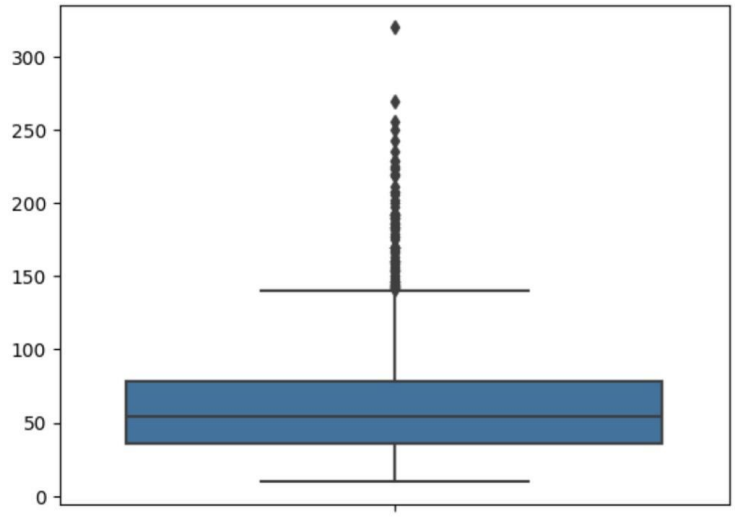


Рисунок 8 - Ящик с усами для признака общей площади living_area

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

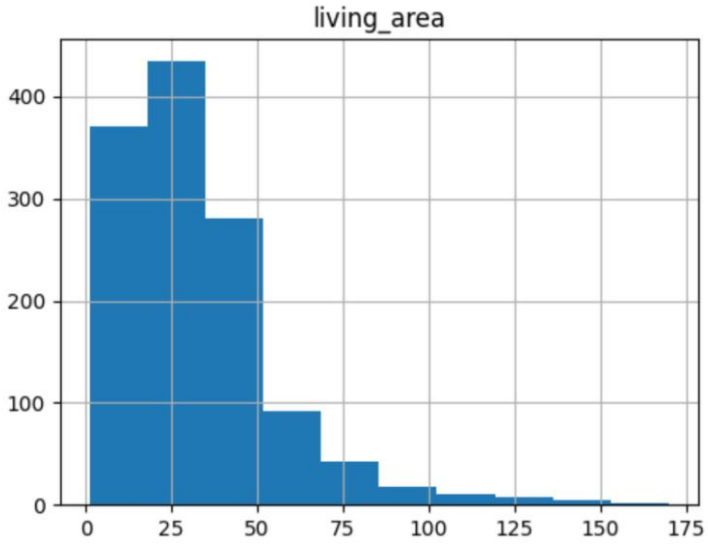


Рисунок 9 - Распределение признака жилой площади квартиры

Признак схож с предыдущим. Помимо того, что признаки имеют схожие графики распределения, имеют высокую корреляцию друг с другом (0.9), два эти признака по сути своей являются одним и тем же. Жилая площадь— усеченная общая площадь(площадь именно жилых комнат).

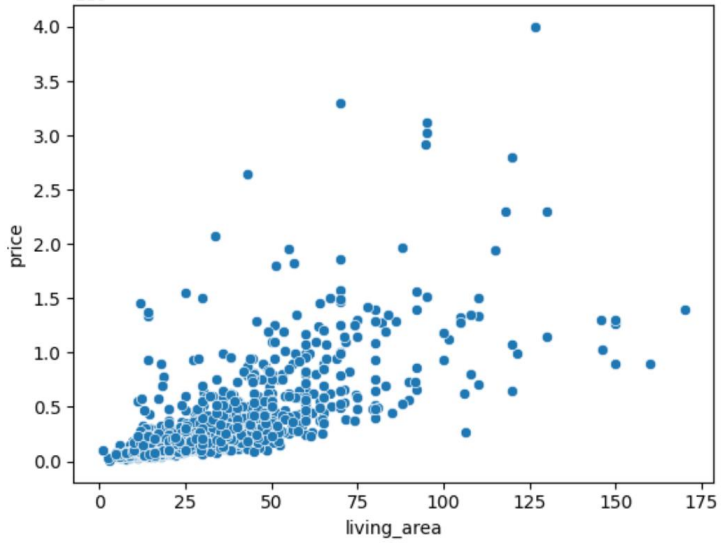


Рисунок 10 - Зависимость цены от жилой площади

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

В данном случае верным решением для подачи данных в модель будет оставить одну из этих переменных, так как для модели эти две модели практически идентичные.

Визуализация выборочных значений для жилой площади квартиры - ящик с усами демонстрирует наличие авбросныз значений в признаке.

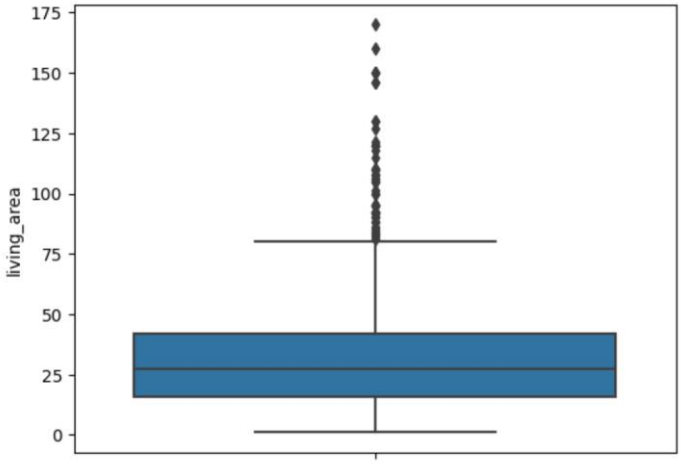
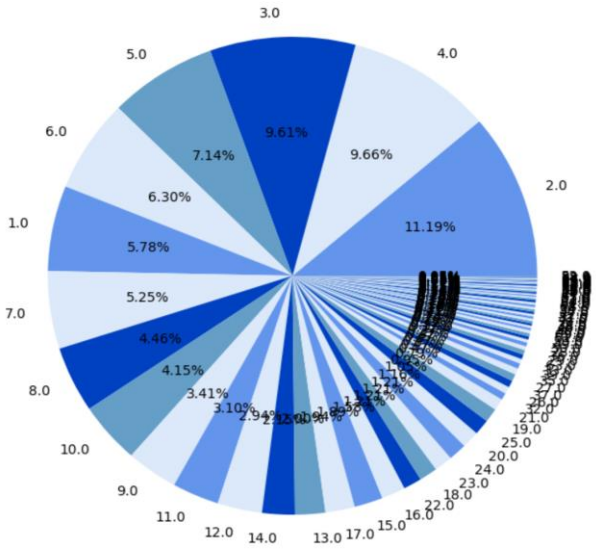


Рисунок 11 - Ящик с усами для признака жилой площади

floor

Большинство квартир в датасете располагаются на 1-10 этажах. Квартиры, расположенные на высоких этажах,встречаются реже.



Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Рисунок 12 - Визуализация количества квартир по этажу

Наличие выбросов в признаке «этаж» подтверждается графиком «ящик с усами».

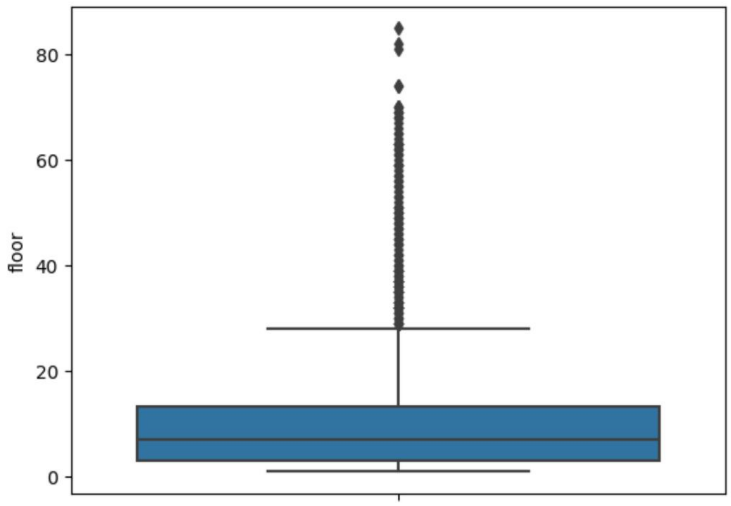


Рисунок 13 - Ящик с усами для признака этажа квартиры

number_of_floors

В датасете чаще всего встречаются квартиры в 9-23 этажных домах.

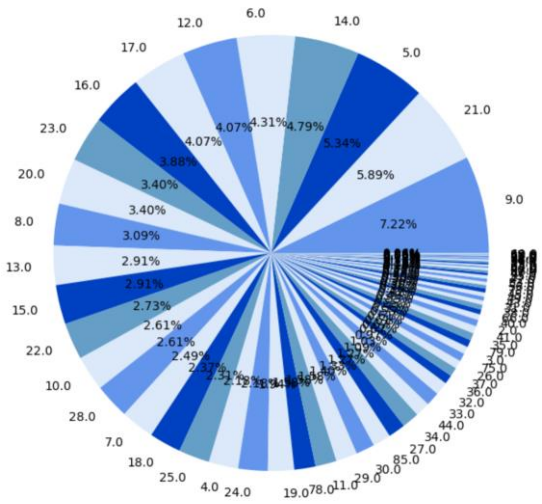


Рисунок 14 - Визуализация количества квартир по этажности дома

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Проверка на наличие выбросов с помощью графика «ящик с усами» показала наличие выбросных значений в этажности дома.

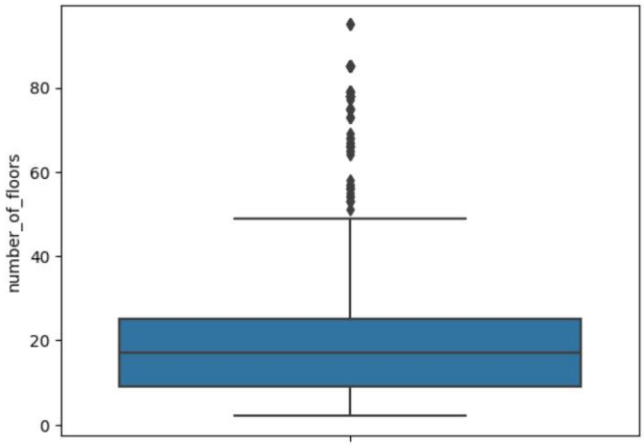


Рисунок 15 - Ящик с усами для признака этажности дома, в котором располагается квартира

construction_year

Подавляющее число квартир в датасете находится в домах-новостройка, сдача которых состоялась/запланирована в 2020-2026 г.

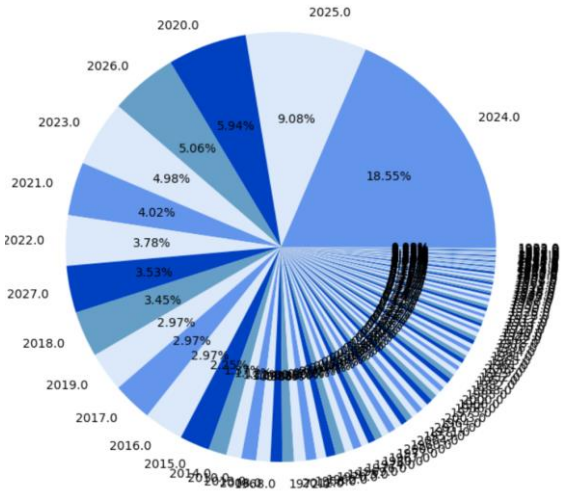


Рисунок 16 - Визуализация количества квартир по году сдачи дом

is_new

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Большинство квартир в датасете является вторичным жильем.

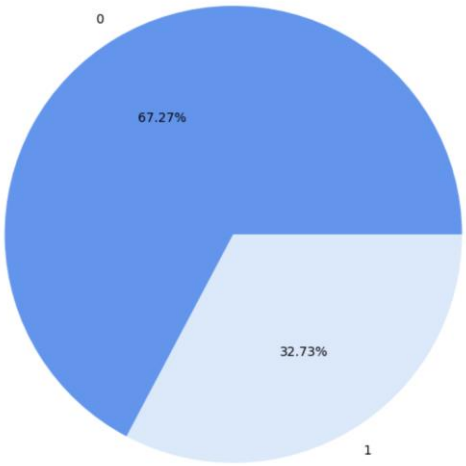


Рисунок 17 - Визуализация количества квартир в зависимости от того, является квартира новостройкой (1) или вторичным жильем (0)

is_apartments

Большинство куартир в датасете является апартаментами

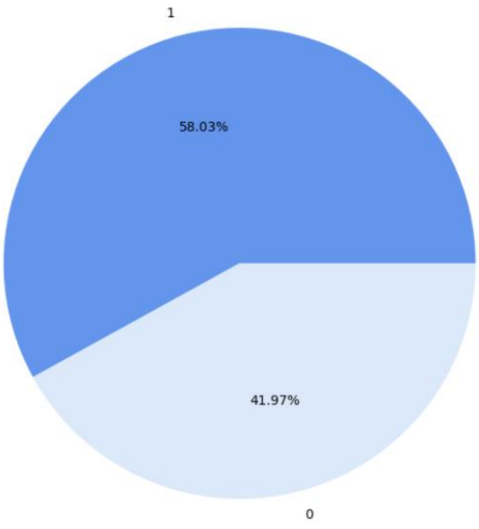


Рисунок 18 - Визуализация количества квартир в зависимости от того, является квартира апартаментами (1) или нет (0)

ceiling_height

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Большинство квартир в датасете со стандартной высотой потолков (около 3 метров).

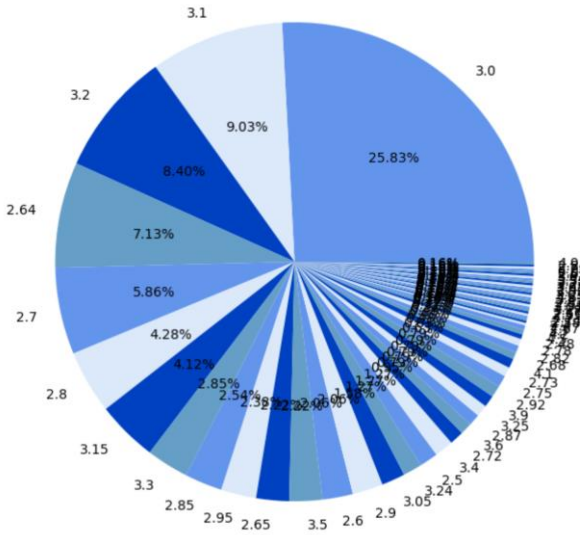


Рисунок 19 - Визуализация количества квартир в зависимости от высоты потолка

Стоит обратить внимание на наличие выбросов в признаке, согласно графику «ящик с усами».

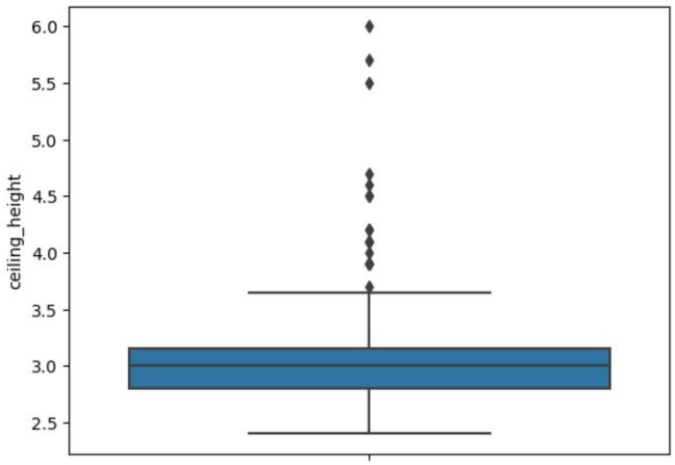


Рисунок 20 - Ящик с усами для признака высоты потолка квартир

number_of_rooms

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

В датасете примерно одинаковое количество 1 (включая студии), 2 и 3-комплектных квартир

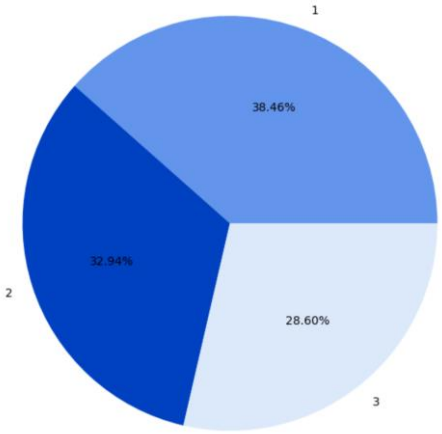


Рисунок 21 - Визуализация количества квартир в зависимости от количества комнат квартиры

Разведочный анализ данных

С помощью простых методов языка программирования Python describe() и info() проведено первичный простой анализ данных, с помощью которого определено несколько тезисов:

1. Средняя стоимость квартиры в датасете составляет около 40 000 000, что поможет в дальнейшем ориентироваться в оценке качества моделей машинного обучения.
2. В датасете присутствуют примеры квартир с очень большим количеством квадратных метров, что можно было бы счесть выбросами, одна стоит помнить, что общая площадь—самая сильно влияющая на таргет переменная, и стандартные методы борьбы с выбросами, такие как замены выбросов на среднее, медиану или моду в данном случае не подойдут. Удаление таких примеров так же не очень хорошо скажется на

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

предсказательной способности модели, поэтому переменная «общая площадь» не будет обработана от выбросов.

3. До 75% квантиля стоимость квартир варьировалась от 0,95 до 47 млн руб. А с 75% квантиля до 100% значное стотмости варьировалась от 47 до 800 млн руб., что говорит нам о нетипичности таких дорогих и больших квартир.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

3.Предобработка данных

Работа с выбросами

Поиск и замена на NaN выбросов в переменных floor, number_of_florr, ceiling_height. Поиск выбросов осуществлен с помощью способа трех сигм. Суть предобработки заключается в нахождении стандартного отклонения признака и оставлении признака в пределах трех сигм вправо и трех сигм влево. Важно, что значения, которые не попадают в данный промежуток значений, не будут удалены, а буду заменены в дальнейшем при обработке пропусков в датасете.

Признак «общая площадь» намеренно не был обработан от выбросов, поскольку значения, которые попадают под понятие «выбросов» важны для модели и формирования цены в датасете.

Замена пропущенных значений

Пропущенные значения в числовых признаках заменены методом KNNImputer. Замена данных методом KNNImputer, или k ближайших соседей, является методом машинного обучения для решения задач классификации и основан на оценивании сходства объектов. Используется функция расстояния Евклида:

$$d(x_i, x_j) = \| x_i - x_j \| = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

где x_{ik} и x_{jk} — к-е элементы векторов x_{i} и x_{j} соответственно

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Рисунок 22 - Функция расстояния Евклида

Несмотря на то, что данный метод используется в задачах классификации, его также целесообразно использовать для восстановления (или импутации) пропущенных значений в датасете. Модель оценивает расстояние от пропущенного значения до k-точек, которые больше всего похожи на рассматриваемое значение, и уже на их основании выбирается значение для пропущенного значения.

$$y_k = \frac{\sum_{i-1}^{n} k_i d(x, a_i)^2}{\sum_{i-2}^{n} d(x, a_i)^2}$$

где a_{i} — i-ый объект, попавший в область, k_{i} — значение атрибута k у заданного объекта a_{i}, x — новый объект, x_{k} — ый атрибут нового объекта.

Рисунок 23- Вычисление дистанции от попавших в область объектов и соответствующих значений этого же атрибута у объектов

Пропуски в категориальной признаке (регион Москвы) заполнены с помощью припенерич метода SimpleImputer-пустые значения заменены на моду, стратегия заполнения пустых значений—наиболее часто встречающееся.

Преобразование категориальных признаков

Для подачи данных в модель необходимо преобразовать категориальные признаки в числовые.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Существует два популярных метода преобразования категориальных признаков:

1.OneHotEncoder

Присвоение каждому значению в категориальном признаке отдельной колонки и проставление единиц в строках, к которым относится данное значение, и нулей, в случаях, если строка не относится к данному значению категориального признака.

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Рисунок 24 - Принцип работы OneHotEncoder

Достоинством метода являются более точные прогнозы при использовании данного метода, а недостатком - громоздкость датасета, иногда неудобство в использовании.

2. LabelEncoder

Более простой в использовании метод. Присвоение значению в категориальном признаке определённой цифры, замена всех значений на цифры.

Достоинством метода является простота использования, а недостатком является склонность моделей к поиску закономерностей в зависимости от числа, иными словами есть шанс, что модель решит, что чем число выше, тем

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

больше будет цена, хотя значения в категориалтных признаках могут быть не лучше, и не хуже один другого.

Для получения решений и финальной модели будем использовать OneHotEncoder, а для создания приложения для простоты будем использовать LabelEncoder для категориального признака region_of_moscow.

Масштабирование данных

В датасете присутсвуют признаки с разными шкалами измерения: годы, миллионы рублей, и комнаты имеют разный масштаб. Чтобы модель воспринимала данные не в таком разрозненном виде, необходимо провести масштабирование данных. Переменные, которые измеряются в разных масштабах, не вносят одинаковый вклад в функцию модели, и изучения модели и могут в конечном итоге вызвать смещение.

Существует два самых популярных способа масштабирования данных: StandartScaler и MinMaxScaler. StandartScaler проводит стандартизацию и значений и приводить все данные к виду, близкому к нормальному распределению-мат.ожилание = 0, дисперсия =1. MinMaxScaler приводит данные к единому масштабу (0,1). В данной работе использован StandartScaler; чтобы привести распределения к признакам к распределениям, напоминающим нормальное распределение.

В финале предварительной обработки данных из датасета необходимо удалить переменную living_area, поскольку переменная сильно коррелирована по отношению к другой переменной датасета-total_area, жилая площадь по своей сути является вырожденной переменной из общей площади.

Выбор и обучение моделей машинного обучения

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

1. Линейная регрессия

Для предсказания стоимости недвижимости самая простая и интуитивно понятная модель—модель линейной регрессии. Она и будет первой моделью в данной работе.

Модель линейной регрессии используется для моделирования зависимости между целевой переменной и предикторами. Этот метод позволяет предсказывать значения целевой переменной на основе значений предикторов. Со сути своей линейная регрессия является прямой, описывающей зависимости переменных от таргера, данная прямая строится и подстраивает веса, наклон и коэффициенты с целью минимизации среднеквадратичной ошибки-то есть разницы между значением, предсказанный прямой регрессии и истинным значением.

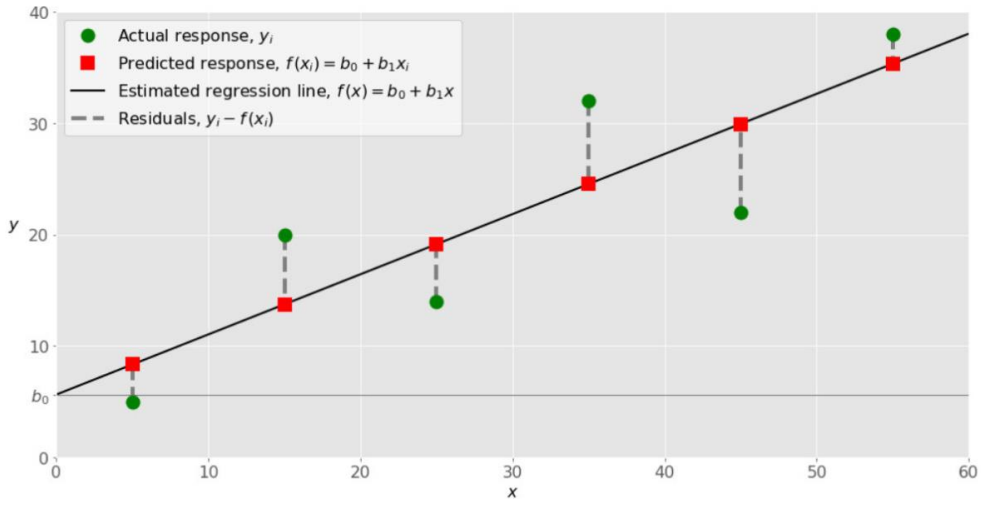


Рисунок 25 - Принцип построения линейной регрессии

2.Дерево решений

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Данная модель использует древовидную структуру, ветви представляют собой результаты узла, которые могут быть как следующими условиями (узлами принятия решений), так и результатом (конечным узлом).

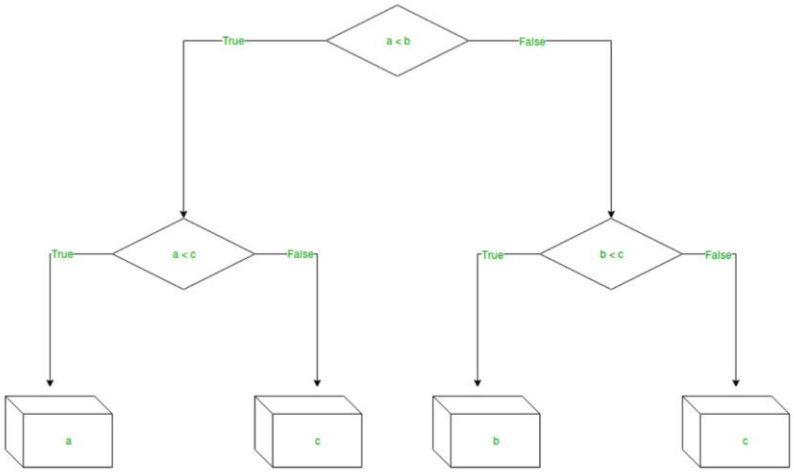


Рисунок 26 - Принцип работы модели дерева решений

3. Бэггинг

С помощью бустрэпа формируются выборки из исходной выборки, подаваемой в модель. На каждой новой маленькой выборке модель обучается, в зависимости от выбранного базового алгоритма. Итоговый результат будет являться усреднением результатов обучения на каждой маленькой выборке. В качестве базовой модели в работе было выбрано дерево решений.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

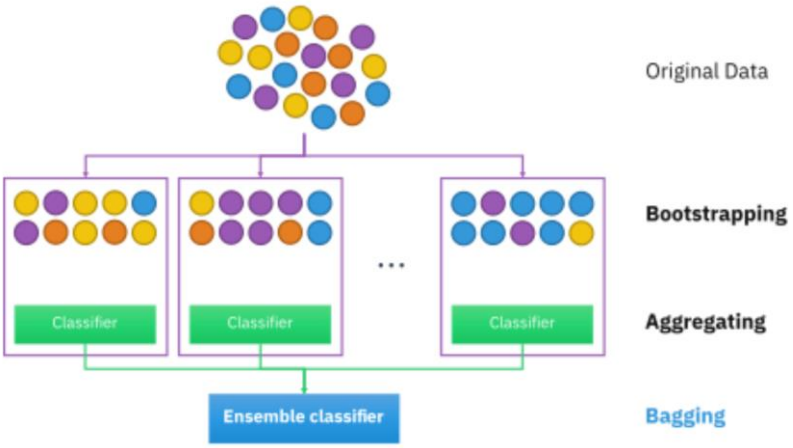


Рисунок 27 - Принцип работы модели бэгинга

4.Случайный лес

С помощью бустрэпа формируется выборка из исходной. На сформированной выборке строится дерево решений. По заданному критерию (по умолчанию—среднеквадратическая ошибка, в работе используется критерий по умолчанию) выбирается лучший признак, на котором далее делается разбиение, после снова выбирается лучший признак, алгоритм продолжается того момента, когда в листе будет минимальное количество объектов.

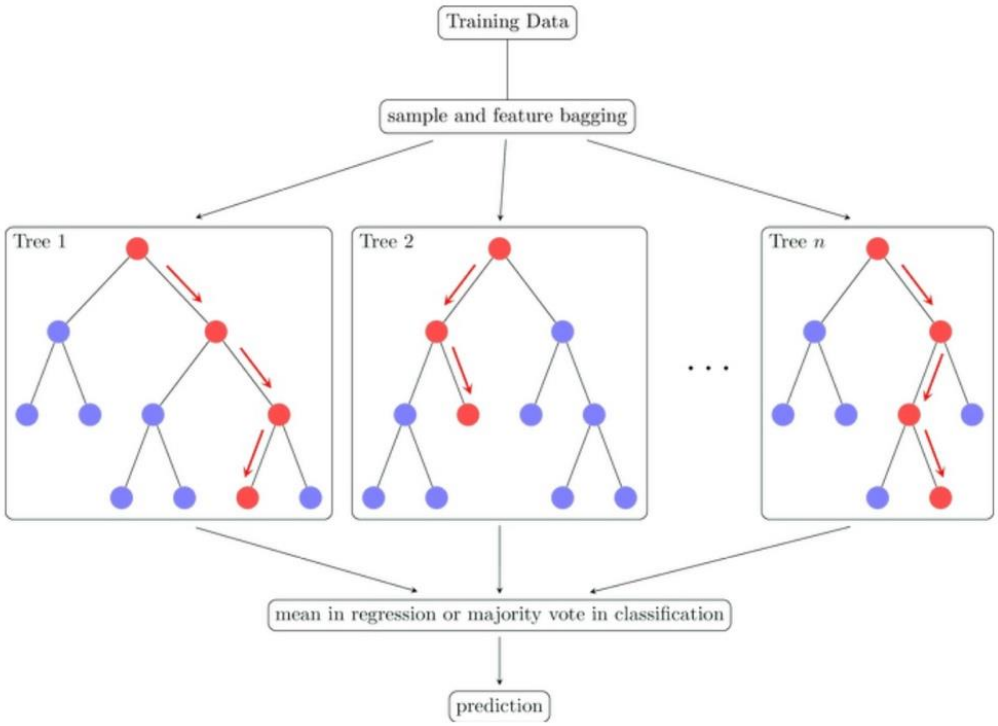


Рисунок 28 - Принцип работы модели случайного леса

5. Воутинг (ансамбль моделей)

По сути не является отдельной моделью, а скорее моделью принятия решений путем голосования различных моделей. В данной работе в качестве голосующих моделей использованы случайный лес, бэгинг, решающее дерево. Ансамбль решений за счет использования разных моделей часто помогает избежать переобучения, что достаточно распространено для деревьев решений (особенно для случайного леса).

6.Стэкинг

Данная модель также является комбинацией из предсказаний моделей различных моделей (как в случае Воутинга-предыдущей модели), однако главное отличие от Воутинга заключается в том, что в Стэкинге выбирается финальна ямодель для принятия решения, которая на финальном голосовании моделей будет вносить наибольший вклад в финальное предсказание.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

7.Градиентный бустинг

Данная модель позволяет сформировать решение в поэтапном режиме, минимизируя ошибки предыдущего решения. По умолчанию— среднеквадратическая ошибка, в работе используется критерий по умолчанию.

4.Оценка качества обученных моделей

Для оценки качества моделей в работе использованы 2 метрики качества:

1. RMSE

Корень из среднеквадратичной ошибки. Метрика показывает, на сколько в среднем ошибается прогноз модели в отличие от реального значения.

2. R2

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Коэффициент детерминации. Измеряет долю дисперсии, объясняемую моделью, в общей дисперсии целевой переменной. Чем ближе R2 к 1, тем лучше модель объясняет данные, чем R2 ближе к 0, тем меньшая доля предсказаний объясняется моделью, т.е. прогноз модели можно считать случайностью при R2 = 0.

Результатом обучения выбранных моделей являются следующие результаты:

	model	RMSE_train	RMSE_test	R2_train	R2_test
0	LinearRegression	31 969 695	26 853 582	0.69	0.60
1	DecisionTreeRegressor	17 253 224	21 576 890	0.90	0.74
2	RandomForestRegressor	10 958 846	23 908 002	0.96	0.68
3	VotingRegressor	11 170 754	21 825 841	0.96	0.74
4	StackingRegressor	17 206 619	26 618 958	0.90	0.60
5	GradientBoostingRegressor	15 397 894	25 984 686	0.93	0.63

Рисунок 29 - Метрики качества обученных моделей

В результате обучения моделей один из лучших результатов показала модель DecisionTreeRegressor, лучше этой модели показала результаты на обучающей выборке VotingRegressor, однако слишком большой разрыв ошибки на обучающей и тестовой выборке говорит о том, что модель, скорее всего склонна к переобучению, т.е. запомнила все данные и закономерности на обучающей выборке, а на тесте справляется значительно хуже, такая модель будет делать неточные прогнозы на новых данных.

Поиск лучших параметров

Получив лучшие результаты на модели решающего дерева, можем осуществить поиск лучших параметров с помощью метода GridSearch - модель для нахождения оптимальных параметров модели. Кроме того, существует

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

второй метод поиска лучших параметров - RandomizedSearchCV, который случайным образом передает набор гиперпараметров, рассчитывает оценку и выдает лучший набор гиперпараметров, который на выходе дает лучший результат. В работе представлен поиск лучших параметров в признанной лучшей моделью DecisionTreeRegressor_GS с помощью GridSearch, а также поиск лучших параметров модели GradientBoostingRegressor с помощью RandomizedSearchCV.

	model	RMSE_train	RMSE_test	R2_train	R2_test
0	LinearRegression	31 969 695	26 853 582	0.69	0.6
1	DecisionTreeRegressor	17 253 224	21 576 890	0.9	0.74
2	RandomForestRegressor	10 958 846	23 908 002	0.96	0.68
3	VotingRegressor	11 170 754	21 825 841	0.96	0.74
4	StackingRegressor	17 206 619	26 618 958	0.9	0.6
5	GradientBoostingRegressor	15 397 894	25 984 686	0.93	0.63
6	DecisionTreeRegressor_GS	22 945 650	27 655 807	0.84	0.58
7	GradientBoostingRegressor_RS	12 059 829	25 847 816	0.96	0.63

Отбор признаков

Помимо поиска лучших параметров модели, был проведен анализ значимости признаков. Из имеющихся в датасете признаков нет ни одно, который был бы не важен для предсказания.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

	feature_name	importance
2	floor	0.766339
5	is_new	0.072733
0	min_to_metro	0.044178
4	construction_year	0.038531
1	total_area	0.030422
8	number_of_rooms	0.022665
3	number_of_floors	0.013958
7	ceiling_height	0.008030
9	region_of_moscow_BAO	0.001891
6	is_apartments	0.001253

Рисунок 30 - Значимость признаков для предсказания стоимости цены квартиры в финальной модели (таблица)

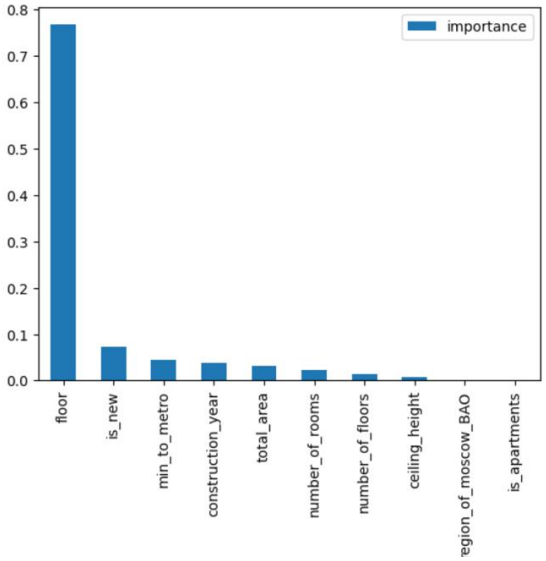


Рисунок 31 - Значимость признаков для предсказания стоимости цены квартиры в финальной модели (график)

Выбор лучшей модели

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Несмотря на подбор лучших параметров модели, лучшей моделью все же осталась DecisionTreeRegression с изначально заданными параметрами: max_depth=8, min_samples_leaf=3

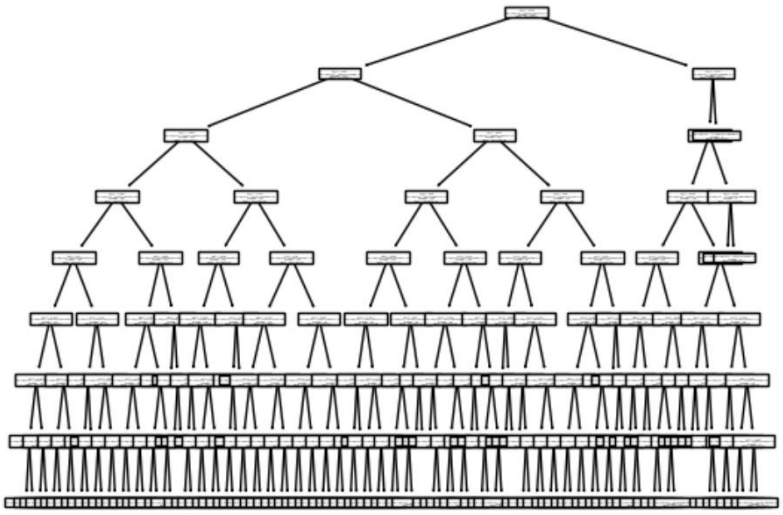


Рисунок 32 - Визуализация финальной модели дерева решений

Разработка приложения

Приложение для предсказания стоимости квартиры по заданным параметрам представляет собой форму для заполнения данных с желаемыми параметрами квартиры.

После заполнения всех критериев формы, пользователь может получить прогноз относительно стоимости квартиры по заданным параметрам.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Рисунок 33 - Приложение-форма для предсказания стоимости квартиры по заданным параметрам

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Заключение

В результате работы были проведены следующие мероприятия:

1) Сбор данных с сайта поиска недвижимости (циан.ру) с помощью бибиотеки BeautifulSoup;
2) Проведен первичный и визуальных анализ данных;
3) Проведена предобработка данных, в том числе работа с выбросными значениями, заполнение отсутствующих значений, преобразование категориальных признаков,масштабирование данных;
4) Обучены различные модели машинного обучения: LinearRegression, DecisionTreeRegressor, BaggingRegressor, RandomForestRegressor, VotingRegressor, StackingRegressor, GradientBoostingRegressor;
5) С помощью метрик оценено качество обученных моделей, выбрана лучшая;
6) Проведен отбор признаков и подбор лучших параметров модели;
7) Создано приложение для предсказания стоимости цены квартиры в Москве по заданным пользователем параметрам.

Москва, ул. 2-я Бауманская, д.5, стр.1
+7 (495) 182-83-85, do@bmstu.ru
do.bmstu.ru

Библиографический список:

1. Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 480 с.: ил. — (Серия «Бестселлеры O’Reilly»).
2. Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow:концепции, инструменты и техники для создания интеллектуальных систем.Пер. с англ. - СпБ.: ООО "Альфа-книга': 2018. - 688 с.: ил
3. Брюс, П. Б89. Практическая статистика для специалистов Data Science: Пер. с англ. /. П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.: ил.
4. Документация по библиотеке scikit-learn. Режим доступа: https://scikit-learn.ru/category/supervised_learning/