**Исследование в области применения алгоритмов ML для расчета стоимости квартир**

**на примере города Кирова**

С использованием написанного на Python парсера, с авито было собрано 2 692 предложения о продаже квартир в городе Кирове. В процессе анализа собранных данных были удалены предложения о продаже, не относящиеся к городу Кирову, ДДУ, ЖСК, жилые дома, двухуровневые квартиры (это отдельный рынок), квартиры в дома, которые еще не введены в эксплуатацию, а также дубликаты. После исключения осталось 2 405 предложений о продаже квартир.

Основные предикторы, которые имеют достаточный уровень описания на сайте авито, приведены в таблице ниже.

Описание основных предикторов (авито):

|  |  |  |
| --- | --- | --- |
| **Наименование** | **Описание** | **Заполняемость** |
| Общая площадь, кв. м | Основной ценообразующий фактор, хорошо коррелирует с ценой квартиры | Отличная (100%) |
| Количество комнат | Коррелирует с общей площадью, приведет к мультиколлинеарности. Категориальный признак. | Отличная (100%) |
| Площадь кухни, кв. м | Коррелирует с общей площадью, приведет к мультиколлинеарности. По данным наблюдается гетероскедастичность. | Хорошая (89%) |
| Жилая площадь, кв. м | Очень хорошо коррелирует с общей площадью, приведет к мультиколлинеарности. | Хорошая (88%) |
| Этаж | Можно разделить на 3 категории: первый, средний, последний. | Отличная (100%) |
| Этажность | С 1 по 27 | Отличная (100%) |
| Балкон/Лоджия | Сомнительный фактор, так как нет информации об отсутствии лоджии/балкона у квартиры. | Хорошая (76%) |
| Тип комнат | Изолированные, смежные | Средняя (61%) |
| Тип санузла | Совмещенный, раздельный | Отличная (100%) |
| Направление окон | Во двор, на улицу | Хорошая (71%) |
| Материал стен | Кирпич, панель, монолит, блок, дерево | Отличная (100%) |
| Год постройки | Можно сделать категориальную переменную | Средняя (67%) |
| Хронологический возраст | Получен из года постройки (созданный предиктор) | Средняя (67%) |
| Высота потолков | – | Низкая (44%) |
| Адрес | Необходимо перевести в координаты. Координаты в модели учтут полностью все характеристики местоположения | Отличная (100%) |
| Координаты | Получены из адреса (созданный предиктор). Координаты получены с использованием Yandex геокодера. Проблема в том, что бесплатно Yandex позволяет перевести не более 1000 адресов с одного ключа. | Отличная (100%) |
| Цена, руб. | Зависимая переменная, которую необходимо предсказать. | Отличная (100%) |

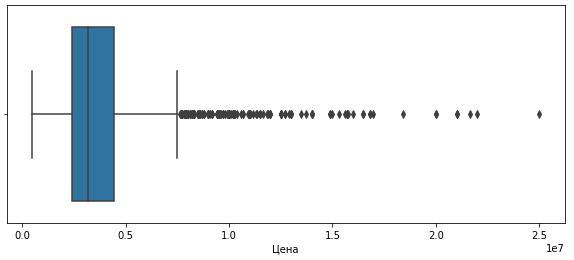
Абсолютное количество пропусков по данным приведено ниже:

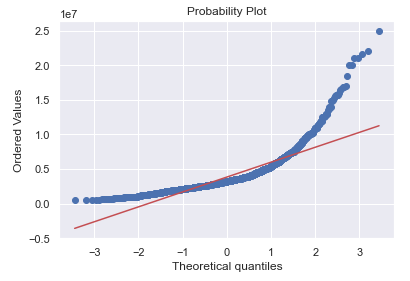
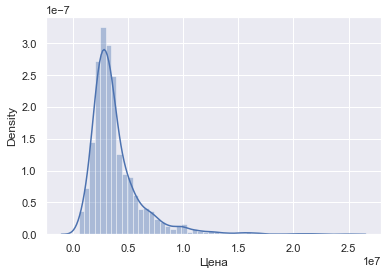


**Анализ данных (зависимой и независимых переменных):**

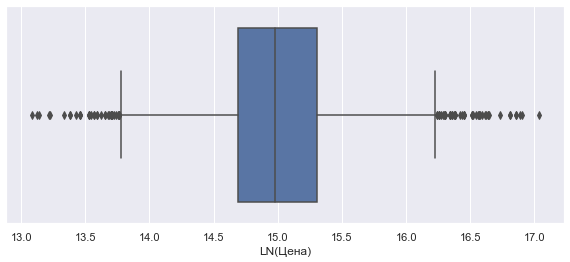
**1. Анализ цены квартиры:**

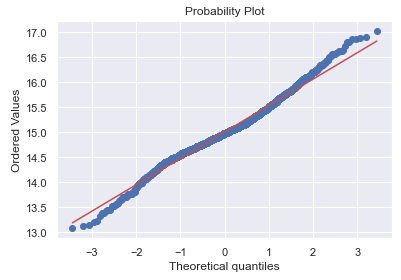
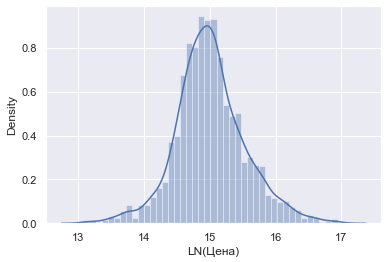
Распределение цены до преобразования выглядит следующим образом:





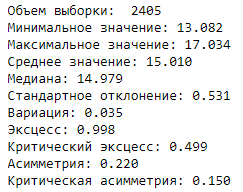
Наблюдается правосторонняя асимметричность и большое количество выбросов справа. Кроме этого, по графику Q-Q plot видно, что точки существенно отличны от прямой, это говорит об отсутствии нормального закона распределения в цене квартиры. Было принято решение посмотреть на распределение логарифмированной цены квартиры (см. рисунок ниже).





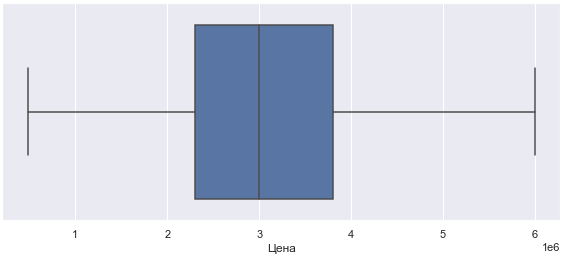
По графику «ящик с усами» видно равномерное распределение выбросов с двух сторон. Гистограмма и график Q-Q plot показывают, что логарифм цены квартиры имеет распределение, приближенное к нормальному.

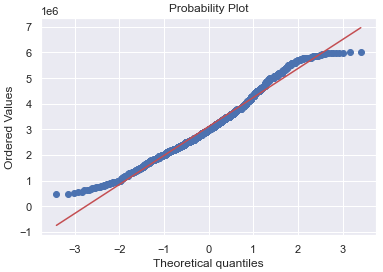
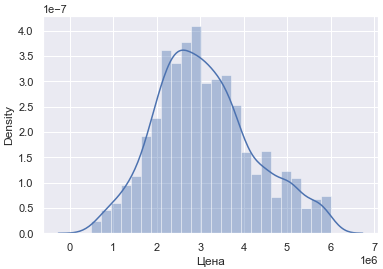
Описательная статистика:



Среднее значение и медиана логарифмической цены сопоставимы, вариация несущественная, по эксцессу наблюдается небольшая остроконечность (значение эксцесса выше его критического значения), что отличает распределение цены от нормального, но, учитывая, что у нас хороший объем выборки, распределение можно считать близким у нормальному.

Границы усов «ящика с усами» имеют следующие значения: 13,76 (левая) и 16,23 (правая). Посмотрим на количество (объем) данных, которые попали в выбросы. В нижнюю границу попали квартиры (35 квартир) преимущественно маленькой площади (до 37 кв. м), расположенные в старых кирпичных/деревянных домах (от 35 лет и старше) этажностью до 5 этажей, на первых и последних этажах. В верхнюю границу попали квартиры (48 квартир) преимущественно большой площади (от 80 кв. м, комнатность от 3 и более), расположенные в относительно новых кирпичных/монолитных домах (до 26 лет и младше) этажностью от 5 и выше, на средних этажах, в т.ч. жилье бизнес-класса. Всего квартир, попавших в выбросы, – 83 квартиры. В целом количество для исключения некритичное. Если цену не логарифмировать, то левая граница «ящика с усами» составит примерно 6 000 000 руб., таким образом, под исключение в качестве выбросов попадет 298 квартир, что в разы больше. Исключение выбросов позволяет приблизить цены квартир к нормальному закону распределения.



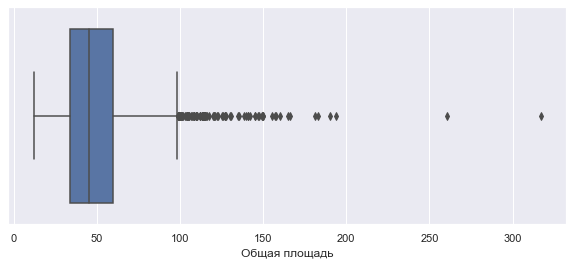


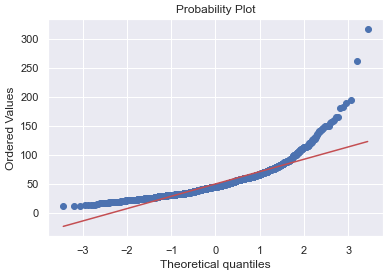
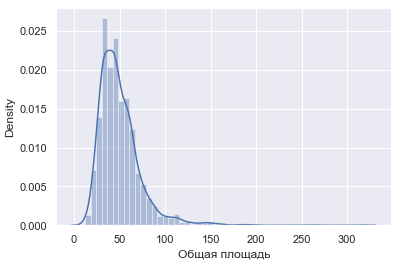
Обращаю внимание, что при построении регрессии выбросы необходимо исключать, также необходимо исключить наличие мультиколлинеарности и гетероскедастичности, а остатки должны быть распределены нормально. Иные алгоритмы ML устойчивы к данным показателям.

Для дальнейшего исследования мы сделаем несколько различных выборок, чтобы проверить как это повлияет на качество моделей. Первая выборка (df\_1) будет включать в себя данные без выбросов, вторая (df\_2) без выбросов логарифмированной цены и третья/четвертая – выбросы не будут исключаться из выборок.

**2. Анализ общей площади квартиры:**

Распределение общей площади квартир до преобразования выглядит следующим образом:

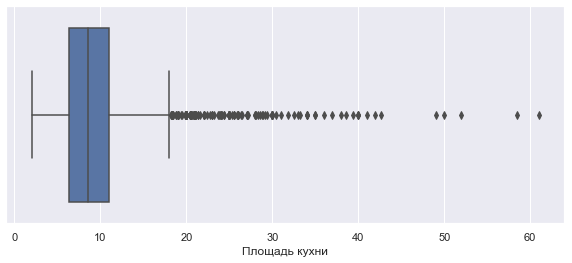


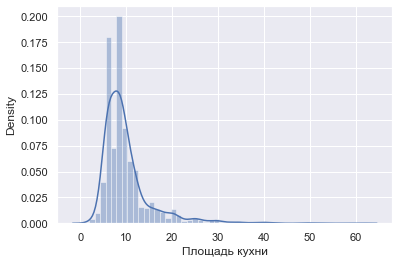


По графику «ящик с усами» видно большое количество выбросов справой стороны, в выбросы попало 93 квартиры. Это квартиры стандартного класса и бизнес-класса, расположенные на средних в домах с годом постройки преимущественно до 25 лет и материалом стен – кирпич/монолит. Гистограмма и график Q-Q plot показывают, что распределение общей площади квартир отличается от нормального. Исключение выбросов позволит приблизить распределение общей площади квартир к нормальному закону распределения. Квартиры с общей площадью более 250 кв. м необходимо исключить в обязательном порядке, так как они существенно отдалены от основной массы квартир.

**3. Анализ жилой площади квартиры и площади кухни:**

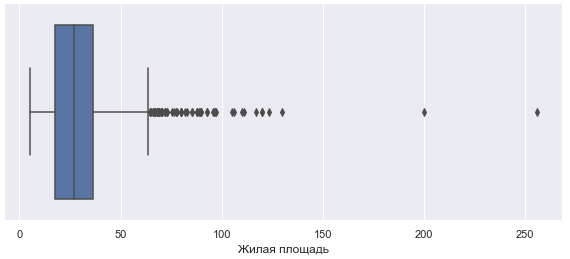
Распределение площади кухни до преобразования выглядит следующим образом:

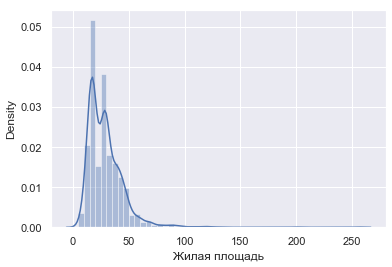




Количество квартир, попавших в выбросы: 161 квартира. Это квартиры преимущественно площадью от 70 кв. м, расположенные на средних этажах жилых домов (стандарт и бизнес-класс) годом постройки до 25 лет с материалом стен – кирпич/монолит.

Распределение жилой площади квартиры до преобразования выглядит следующим образом:





Количество квартир, попавших в выбросы: 74 квартиры. Это квартиры преимущественно площадью от 80 кв. м, расположенные на средних этажах жилых домов (стандарт и бизнес-класс) с материалом стен – кирпич/монолит.

Обращаю внимание, что общая площадь коррелирует с комнтаностью, площадью кухни и жилой площадью (см. рисунки ниже), что приведет к мультиколлинеарности.

**4. Анализ хронологического возраста:**

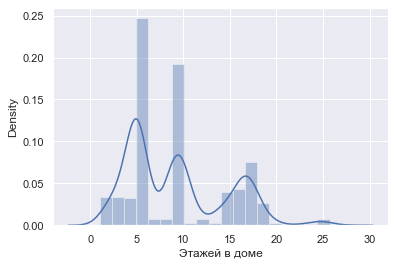
Распределение хронологического возраста до преобразования выглядит следующим образом:



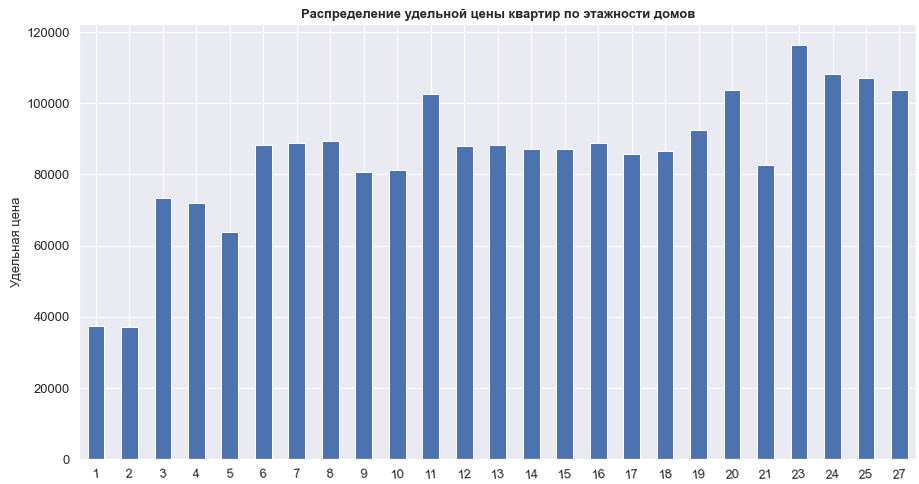
Согласно гистограмме 40% многоквартирных домов в г. Кирове, в которых выставлены на продажу квартиры, имеют возраст до 10 лет (средняя цена – 89 380 руб./кв. м), 16% – имеют возраст от 11 до 25 лет (средняя цена – 87 848 руб./кв. м), 25% – имеют возраст от 26 до 50 лет (средняя цена – 66 306 руб./кв. м) и 19% – имеют возраст свыше 50 лет (средняя цена – 57 604 руб./кв. м). По хронологическому возраст наблюдается большое количество пропусков – 729. Хронологический возраст влияет на стоимость квартиры – чем выше возраст дома, тем ниже удельная стоимость квартиры.

**5. Анализ этажности домов:**

Распределение этажности домов до преобразования выглядит следующим образом:

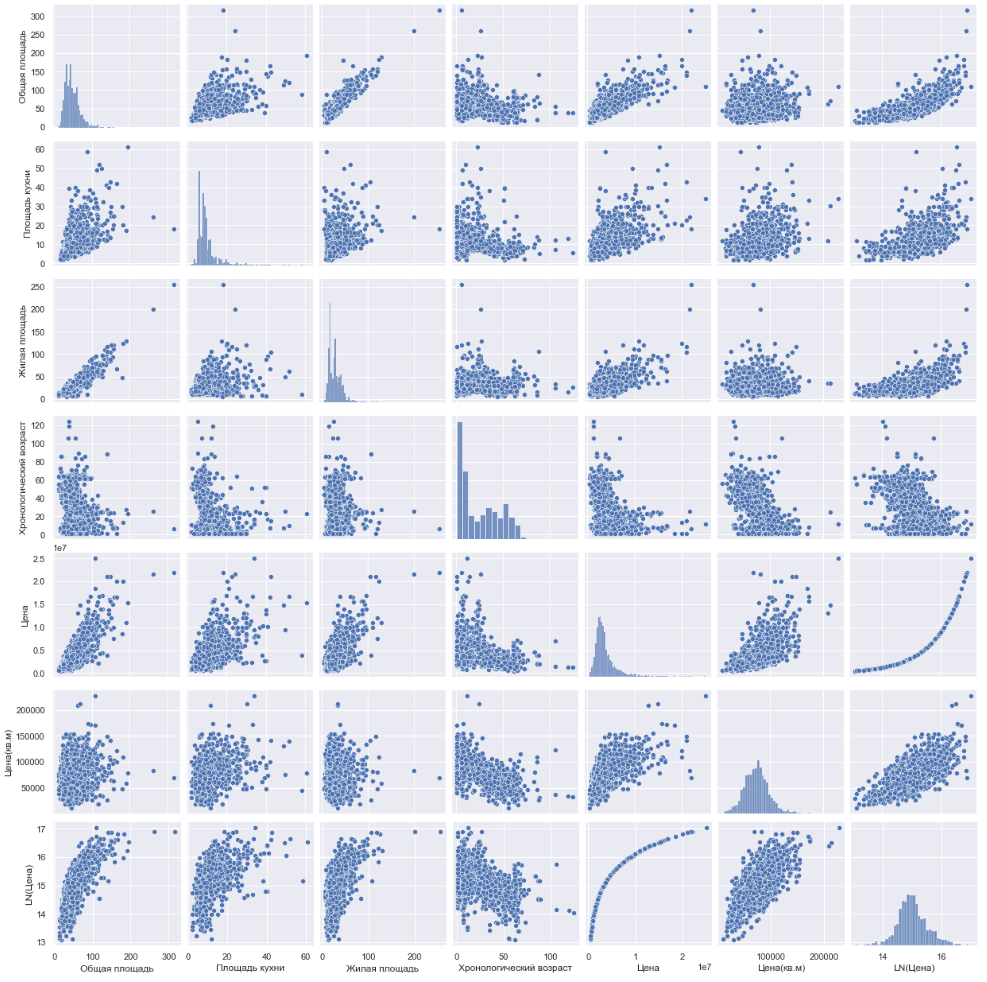


Основная доля рынка (1 999 квартир или 83%) приходится на квартиры, расположенные в домах этажностью – 2, 3, 4, 5, 9, 10, 16, 17. Зависимость удельной цены от этажности дома приведена на графике ниже:

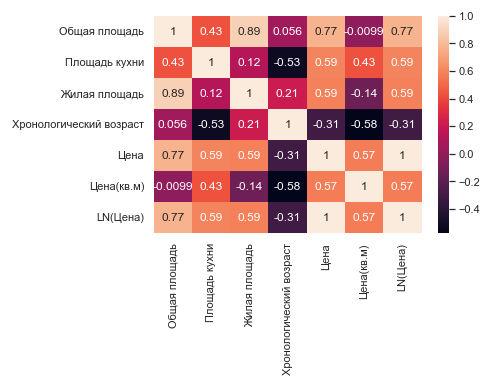


По графику видно, что самые дешевые квартиры расположены в домах этажностью до 2 (до 40 000 руб./кв. м), далее этажностью от 3 до 5 (в диапазоне 60 000 – 80 000 руб./кв. м), от 6 до 19 (в диапазоне 80 000 – 100 000 руб./кв. м) и от 20 до 27 (в диапазоне 100 000 – 1120 000 руб./кв. м). Этажность оказывает влияние на стоимость квартиры.

Построим диаграммы плотности распределения и рассеяния по следующим независимым переменным: общая площадь, площадь кухни, жилая площадь, хронологический возраст, цена, удельная цена, логарифмированная цена.



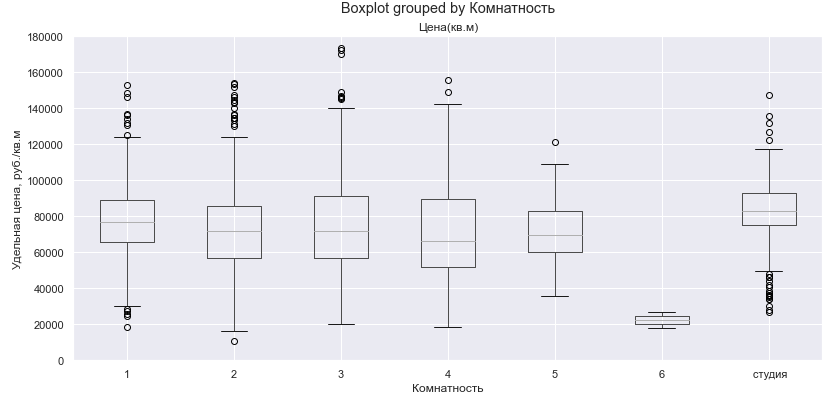
Согласно вышеприведенным диаграммам рассеивания видно, что имеется зависимость цены (в т.ч. логарифмированной цены) от общей площади, площади кухни и жилой площади – чем больше площадь, тем выше цена. Также видна обратная зависимость цены (в т.ч. логарифмированной цены и удельной цены) от хронологического возраста – чем выше хронологический возраст, тем ниже цена. Кроме этого, наблюдается прямая зависимость между общей площадью и площадью кухни, жилой площадью. Наблюдается обратная зависимость между площадью кухни и хронологическим возрастом – чем больше кухня, тем моложе дом (такая зависимость объяснима, так как в новых домах площадь кухни, как правило, больше, чем в старых домах). Все вышесказанное подтверждается корреляционной матрицей Спирмана (см. рисунок ниже). Обращаю внимание, что для восстановления пропусков по хронологическому возрасту можно попробовать, в том числе, использовать такие предикторы, как площадь кухни и удельная цена.



**6. Анализ комнатности:**

Распределение комнатности квартир и средний уровень цен выглядит следующим образом:

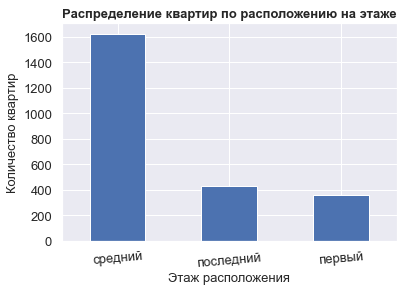


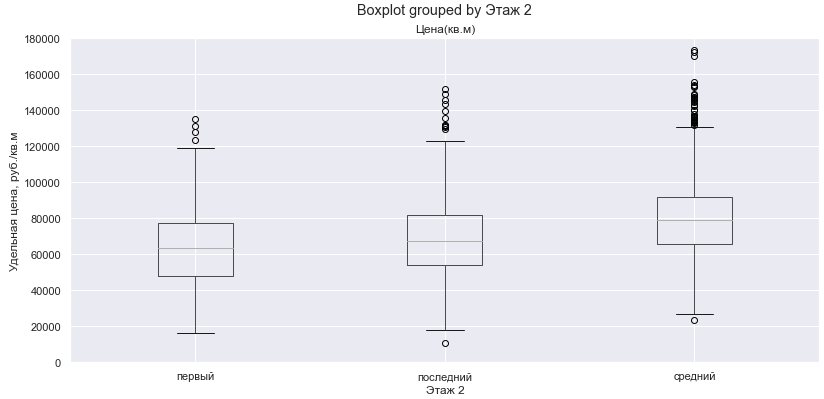


Основная доля рынка приходится на квартиры комнатностью 1, 2, 3 и студии. По графику видно, что самая высокая удельная цена у студий, затем у 1-к, 2-3-к и т.д. Наблюдается обратная зависимость удельной цены от комнатности – чем больше комнат у квартиры (количество комнат прямо зависит от площади), тем ниже удельная цена квартиры. В качестве независимой переменной в модель лучше взять площадь, которая прямопропорциональна комнатности. Комнатность является категориальной переменной.

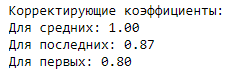
**7. Анализ расположения на этаже:**

Распределение расположения на этаже и средний уровень цен выглядит следующим образом:



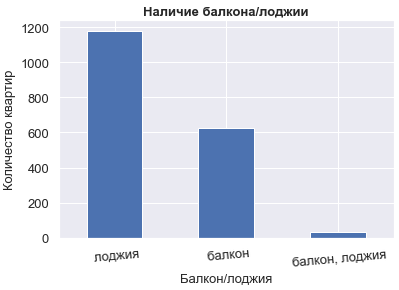


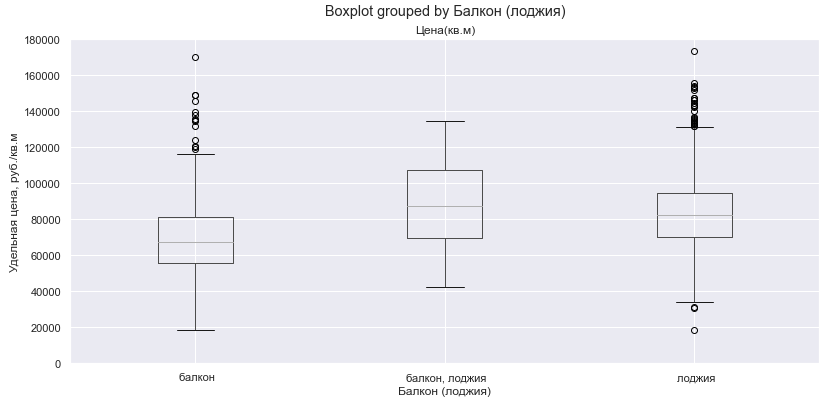
Основную долю рынка представляют квартиры (более 1 600 квартир), расположенные на среднем этаже. Согласно графику, самая высокая удельная цена у квартир, расположенных на среднем этаже. Квартиры, расположенные на последнем этаже, стоят чуть дороже квартир, расположенных на 1 этаже. Корректирующий коэффициенты (соотношения) представлены ниже:



**8. Анализ наличия балкона/лоджии:**

Распределение наличия балкона/лоджии и средний уровень цен выглядит следующим образом:

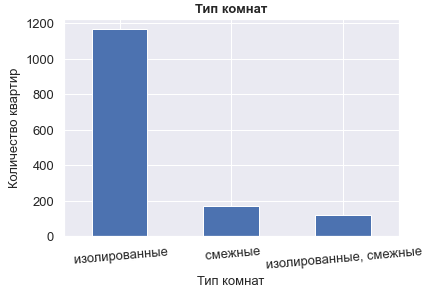


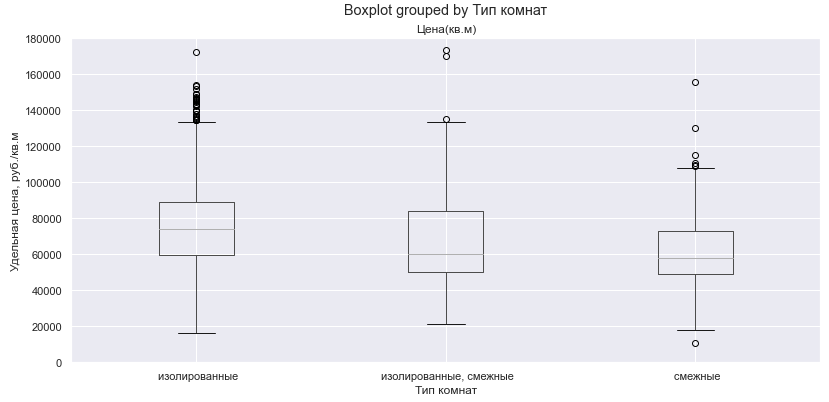


Основную долю рынка представляют квартиры (1 180 квартир), которые имеют лоджии. Квартиры, имеющие лоджию, стоят дороже квартир, имеющих балкон. Конкретного понимания разницы между понятиями «лоджия» и «балкон» в авито нет, гарантировать правильное заполнение со стороны продавца – невозможно. Информация о квартирах, которые не имеют балкон/лоджию – нет. В целом категорию «балкон, лоджия» можно прировнять к категории «лоджия», так как они сопоставимы по удельной цене и большая доля рынка представлена квартирами с лоджиями. Квартиры с категорией «балкон, лоджия» (35 квартир) представляют собой преимущественно квартиры площадью от 50 кв. м, расположенные в новых домах (возраст до 26 лет) на средних этажах. Использование данной независимой переменной в модели является сомнительным.

**9. Анализ типа комнат:**

Распределение типа комнат и средний уровень цен выглядит следующим образом:

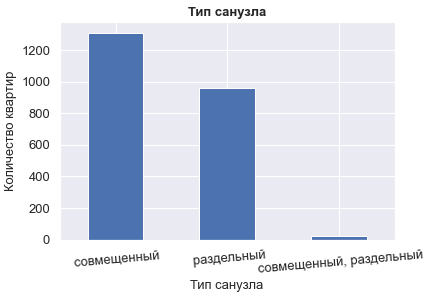




Большую долю рынка представляют квартиры с изолированными комнатами (более 1 100 квартир или 80%). Квартиры с изолированными комнатами стоят дороже, чем квартиры со смежными комнатами. По заполняемости параметра «тип комнат» наблюдает самое большое количество пропусков – 946. Пропуски можно заполнить изолированными квартирами, так как их частота составляет 80%. Категорию «изолированные, смежные» можно перевести в категорию «смежные» (понимается – проходные), так как их удельные цены сопоставимы.

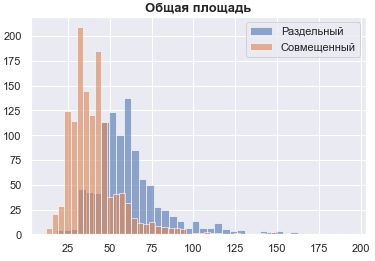
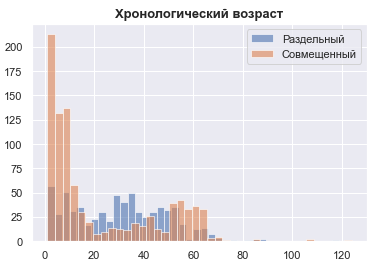
**10. Анализ типа санузла:**

Распределение типа санузла и средний уровень цен выглядит следующим образом:



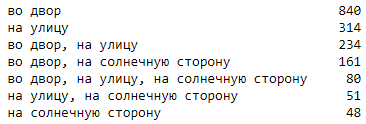


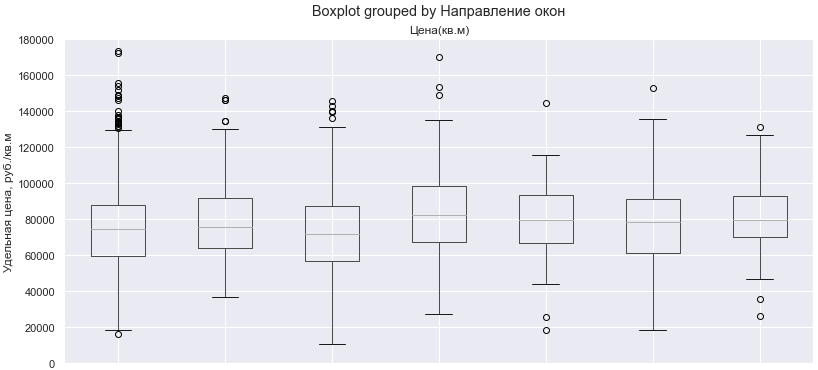
Большую долю на рынке представляют квартиры со совмещенным санузлом (более 1 300 квартир). Квартиры со совмещенным санузлом стоят дороже квартир с раздельным санузлом. Это обусловлен тем, что квартиры со совмещенными санузлами в большей степени находятся в домах возрастом до 25 лет (46%) и площадью до 50 кв. м, а квартиры с раздельными санузлами в домах возрастом от 25 лет и площадью более 50 кв. м. За счет возраста и эффекта масштаба удельная цена квартир с раздельным санузлом ниже удельной цены квартир со совмещенным санузлом (см. гистограммы ниже). Оставшаяся категория «совмещенный, раздельный» с 23 квартирами имеют самую высокую удельную цену. Это связано с тем, что в эту категорию попали преимущественно квартиры с 2-мя санузлами площадью более 80 кв. м, расположенные в домах возрастом до 25 лет на средних этажах.



**11. Анализ направления окон:**

Распределение направления окон и средний уровень цен выглядит следующим образом:

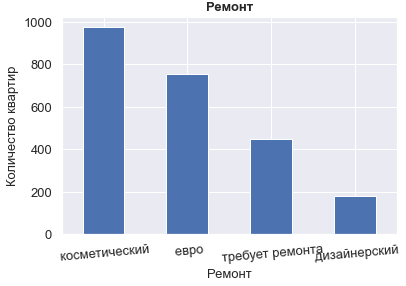


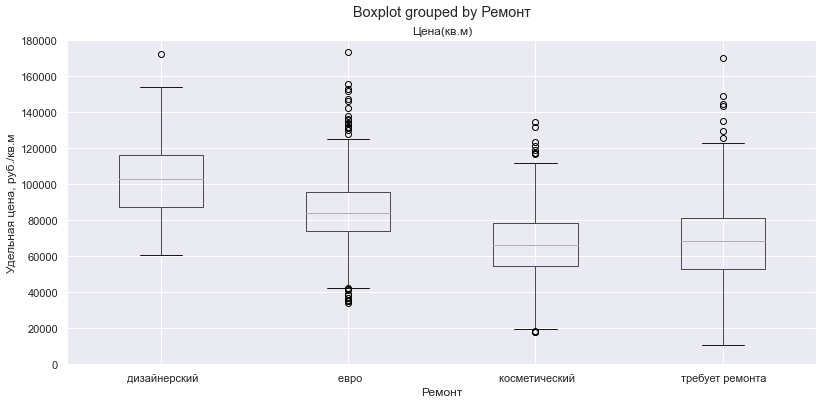


Существенной ценовой разницы не наблюдается.

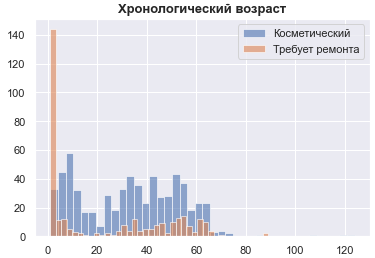
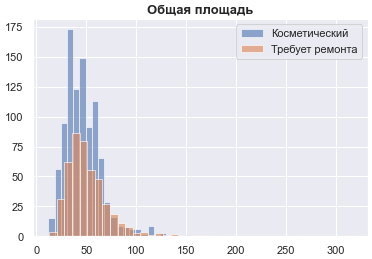
**12. Анализ состояния отделки:**

Распределение состояния отделки и средний уровень цен выглядит следующим образом:



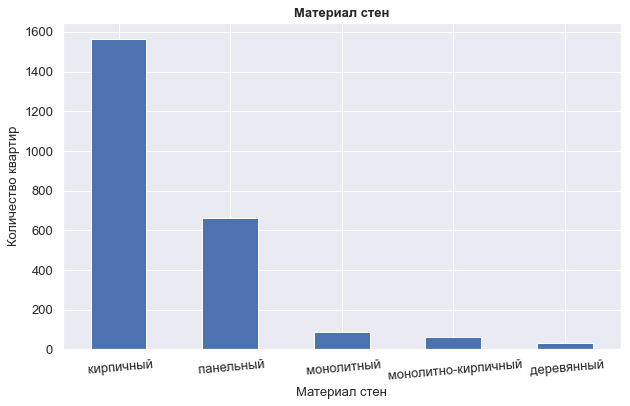


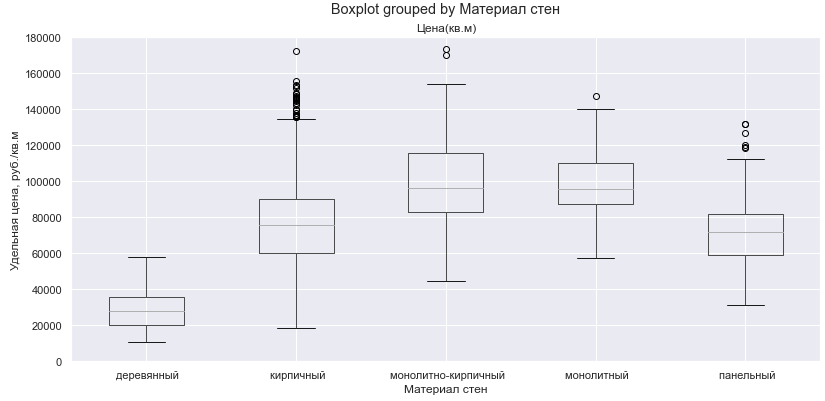
Большую долю на рынке представляют квартиры с косметическим (стандартным) ремонтом (более 900 квартир). Самые дорогие квартиры имеют дизайнерский ремонт, на втором месте квартиры с евроремонтом, на третьем месте – косметический ремонт и требует ремонта. Квартиры, которые требуют ремонта (без отделки) в среднем сопоставимы по цене с квартирами, которые имеют косметический (стандартный) ремонт. Это объясняется тем, что квартиры без отделки в большей степени расположены в домах новой застройки (см. гистограммы ниже).



**13. Анализ материала стен:**

Распределение материала стен и средний уровень цен выглядит следующим образом:

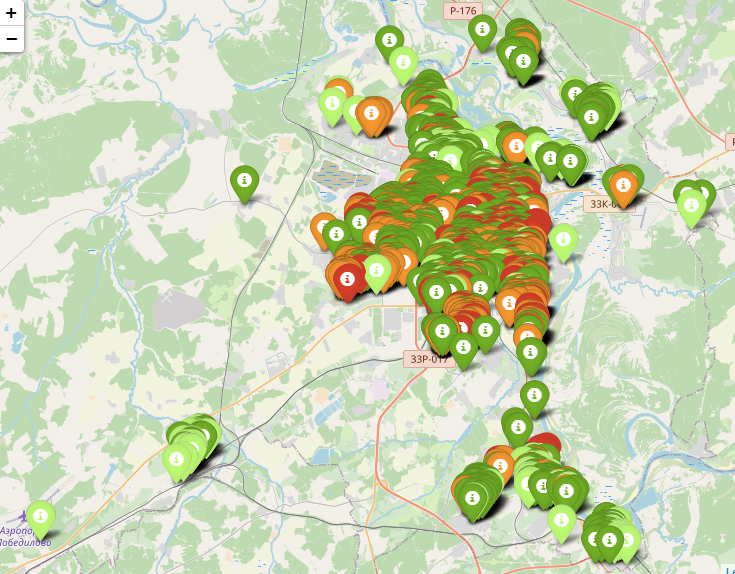




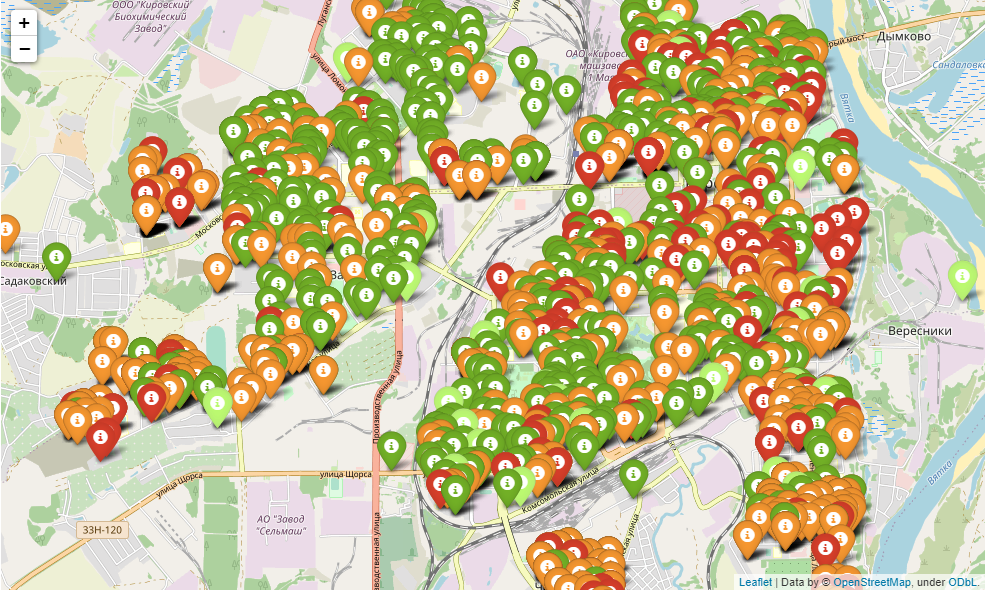
Большую долю на рынке представляют квартиры в кирпичных домах (более 1 500 квартир). Самые дорогие квартиры расположены в монолитных и монолитно-кирпичных домах, на втором месте квартиры в кирпичных домах, на третьем месте – дома в панельных домах и на последнем месте – квартиры в деревянных домах.

**14. Анализ местоположения (координат):**

Распределение по местоположению и средний уровень цен выглядит следующим образом:



На карте видно, что самые дешевые квартиры (светло зеленый и зеленый цвета, с удельной ценой до 75 000 руб./кв. м) в большей степени расположены в районах на окраине города. Квартиры удельной ценой свыше 75 000 руб./кв. м концентрируются в центральной части города и местах новой застройки.



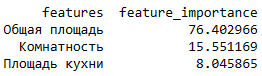
**Восстановление данных:**

**1. Восстановление данных по фактору «площадь кухни» (количество пропусков – 262):**

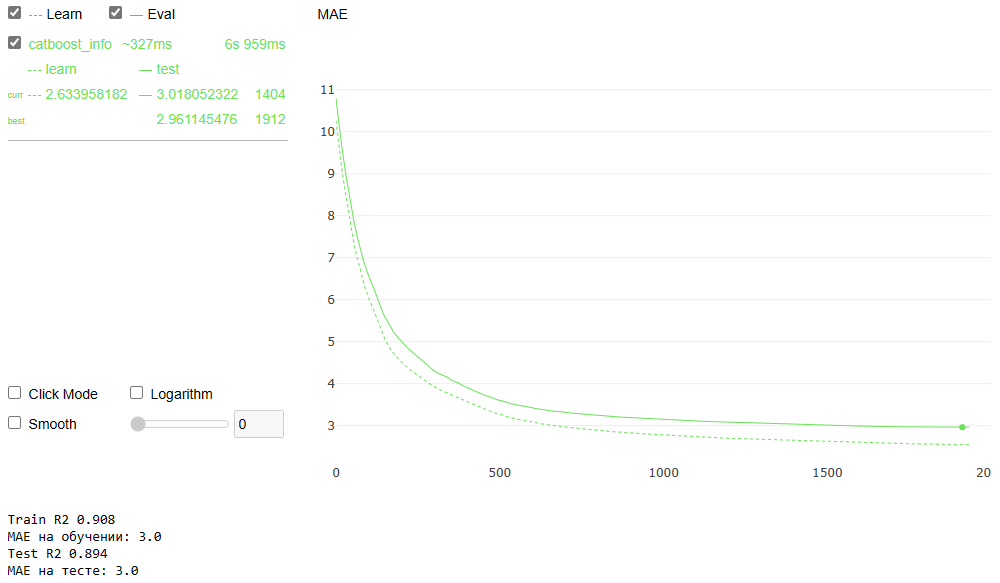
Пропуски по фактору «площадь кухни» заполнялись с использованием алгоритма KNNImputer (библиотека sklearn.impute). Такое решение было принято в связи с тем, что восстановить площадь кухни с использованием других факторов не представляется возможным, хорошая зависимость от других факторов не выявлена. Также, можно попробовать использовать алгоритм IterativeImputer. Класс IterativeImputer реализует многомерные алгоритмы восстановления пропущенных значений, оценивая другие значения в наборе данных. Данный класс моделирует каждый признак пропущенного значения как функцию от других признаков и использует оценку для замены значений. IterativeImputer фактически итеративно строит модель регрессии, используя подмножества столбцов для прогнозирования отсутствующих значений.

**2. Восстановление данных по фактору «жилая площадь» (количество пропусков – 290):**

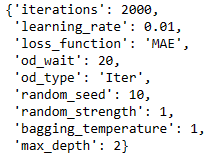
Для восстановления пропусков по фактору «жилая площадь» использовался алгоритм CatBoostRegressor. В процессе проработки алгоритма были подобраны такие гиперпараметры, которые позволяют получить модель с наилучшими метриками R2 и MAE. В качестве факторов, хорошо объясняющих размер жилой площади, использованы «общая площадь», «комнатность» и «площадь кухни». Степень важности факторов (в %) приведена в таблице ниже:



Результаты обучения алгоритма CatBoostRegressor:



Гиперпараметры:



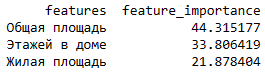
Модель, построенная на данных для обучения, приведена ниже:



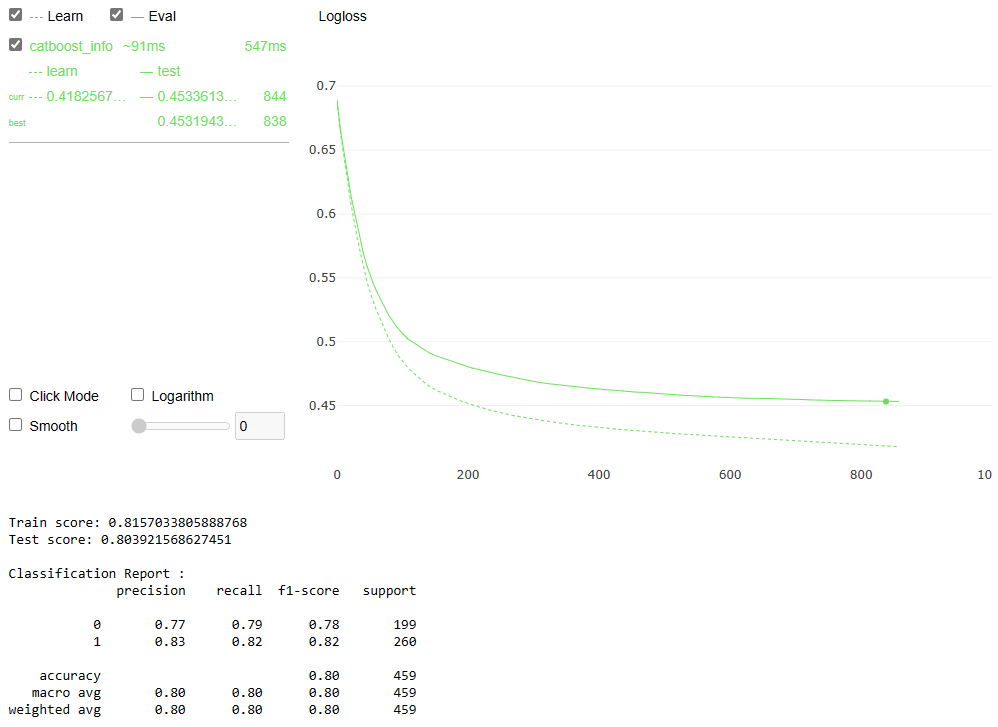
Коэффициент детерминации на обучении составил 0,908, на тесте – 0,894.

**3. Восстановление данных по фактору «тип санузла» (количество пропусков – 110):**

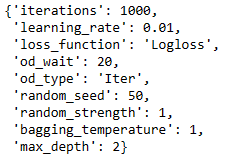
Для восстановления пропусков по фактору «тип санузла» использовался алгоритм CatBoostClassifier. В процессе проработки алгоритма были подобраны такие гиперпараметры, которые позволяют получить модель с наилучшими метриками R2 и Logloss. В качестве факторов, хорошо объясняющих тип санузла, использованы «общая площадь», «жилая площадь» и «этажность дома». Степень важности факторов (в %) приведена в таблице ниже:



Результаты обучения алгоритма CatBoostClassifier:

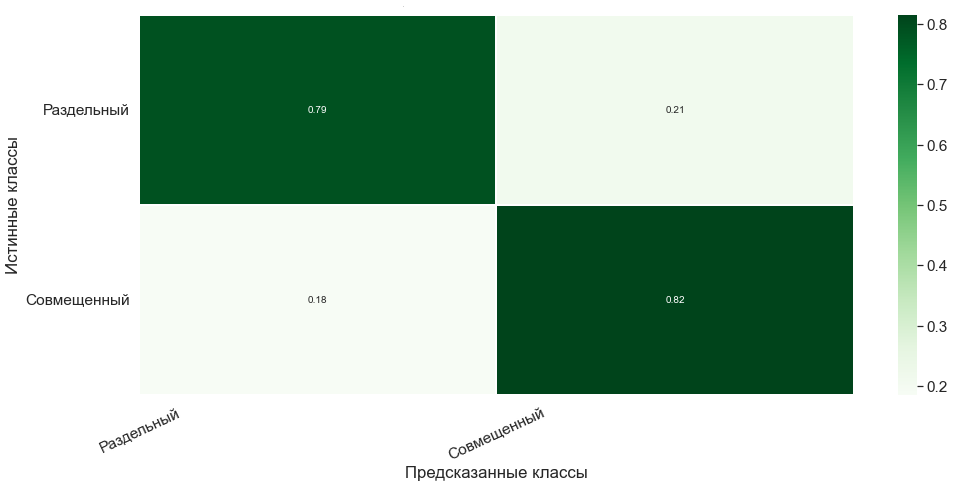


Гиперпараметры:



Accuracy на обучении составил 0,816, на тесте – 0,804, среднее значение accuracy на кросс-валидации – 0,810.

Confusion Matrix:





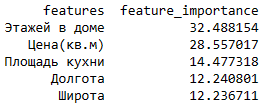
В части «раздельного» санузла модель ошиблась в 21% случаев, в части «совмещенного» санузла модель ошиблась в 18% случаев.

**4. Восстановление данных по фактору «ремонт (или состояние отделки)» (количество пропусков – 51):**

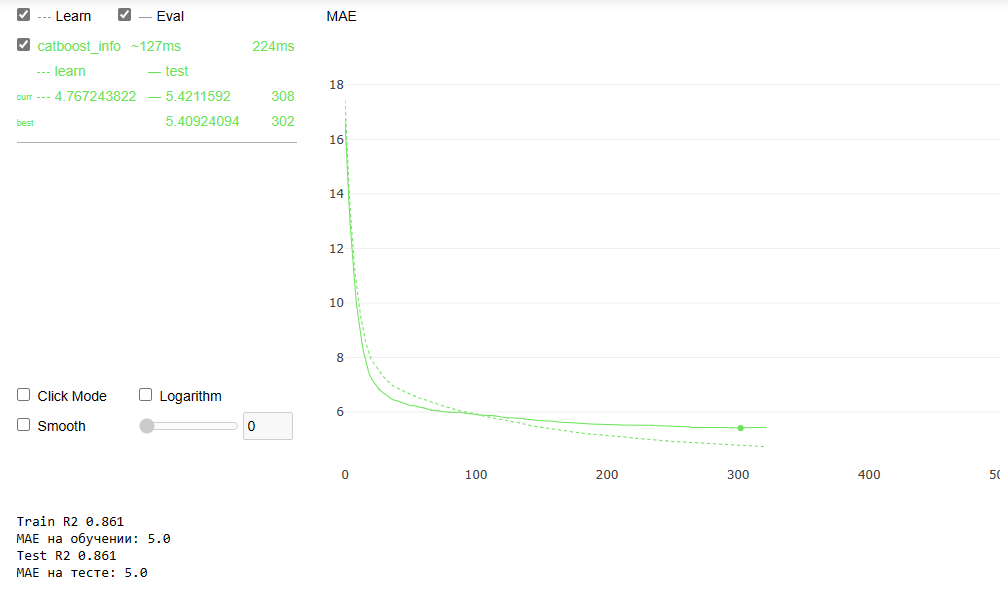
Пропуски по фактору «ремонт» заполнялись с использованием алгоритма SimpleImputer(strategy = 'most\_frequent') (библиотека sklearn.impute). Такое решение было принято в связи с тем, что пропусков очень мало и заполнить их можно модой (часто встречающимся значением, которое равно «косметический»).

**5. Восстановление данных по фактору «хронологический возраст» (количество пропусков – 729):**

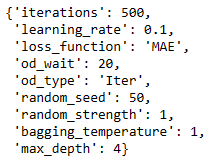
Для восстановления пропусков по фактору «хронологический возраст» использовался алгоритм CatBoostRegressor. В процессе проработки алгоритма были подобраны такие гиперпараметры, которые позволяют получить модель с наилучшими метриками R2 и MAE. В качестве факторов, хорошо объясняющих размер хронологического возраста, использованы «этажность», «площадь кухни», «местоположение (долгота и широта)» и «удельная цена». Степень важности факторов (в %) приведена в таблице ниже:



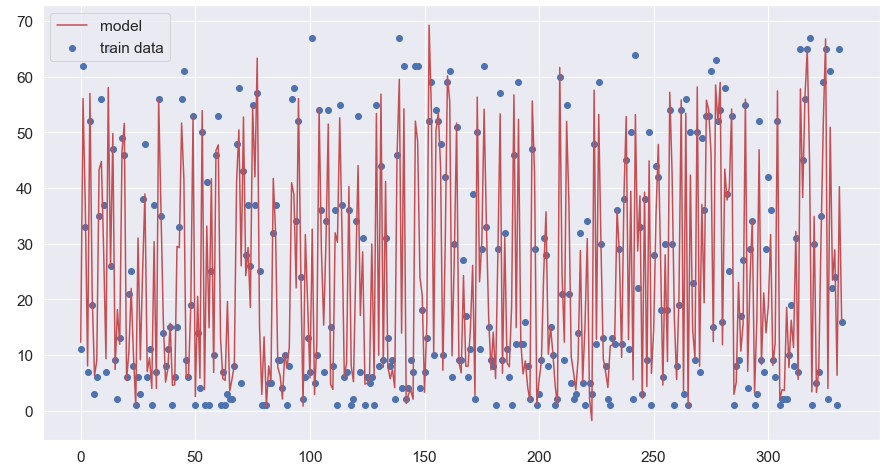
Результаты обучения алгоритма CatBoostRegressor:



Гиперпараметры:

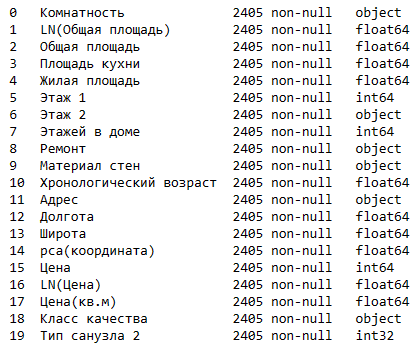


Модель, построенная на данных для обучения, приведена ниже:



Коэффициент детерминации на обучении составил 0,861, на тесте – 0, 861.

В результате использования алгоритмов все пропуски в данных были восстановлены:



**Восстановление данных (пропусков):**

– через алгоритмы классификации и регрессии.

– через специальные алгоритмы (SimpleImputer, KNNImputer, IterativeImputer).

– через описание в объявлениях.

– через специальные внешние источники данных (например, год постройки дома).

**Выбросы:**

– исключение выбросов из выборки (в т.ч. с использование алгоритмов).

– замена выбросов иным значением, полученным через алгоритм.

– исправление выброса.

**Построение регрессионной модели оценки стоимости квартир:**

В рамках регрессионного анализ было отработано 10 вариантов модели для достижения лучшего качества. Первоначально было сформировано 2 основные выборки:

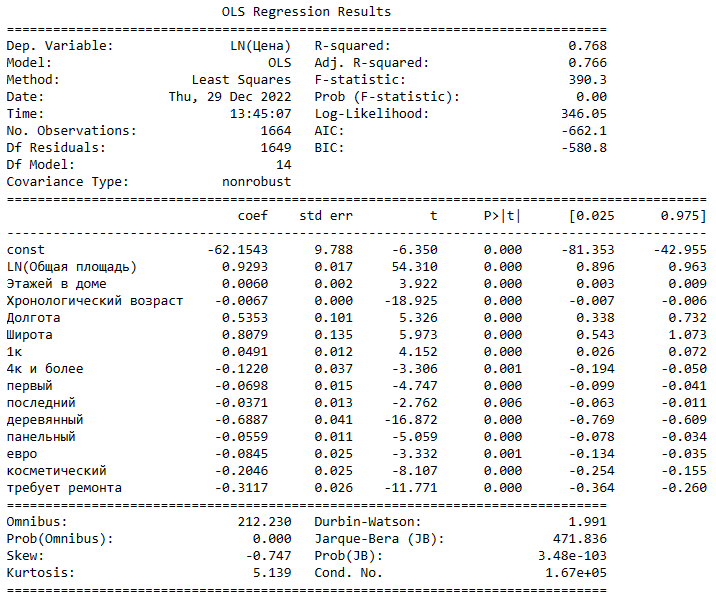
Выборка 1 – с учетом исключения выбросов в данных по общей площади квартиры, хронологическому возрасту и ценах предложения.

Выборка 2 – с учетом исключения выбросов в данных по логарифмированной общей площади квартиры, хронологическому возрасту и логарифмированным ценам предложения.

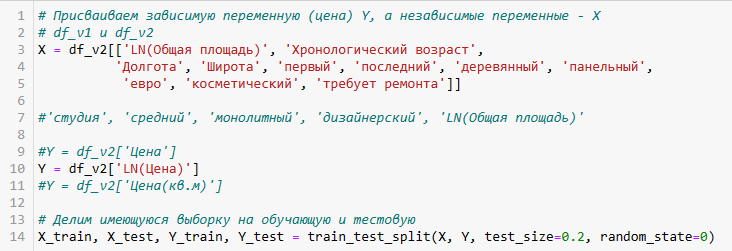
Различие в указанных выборках заключалось в том, что в выборке 1 содержалось намного больше выбросов, чем в выборке 2, что приводит к исключению большого количества информации в выборке 1 (боле 15%).

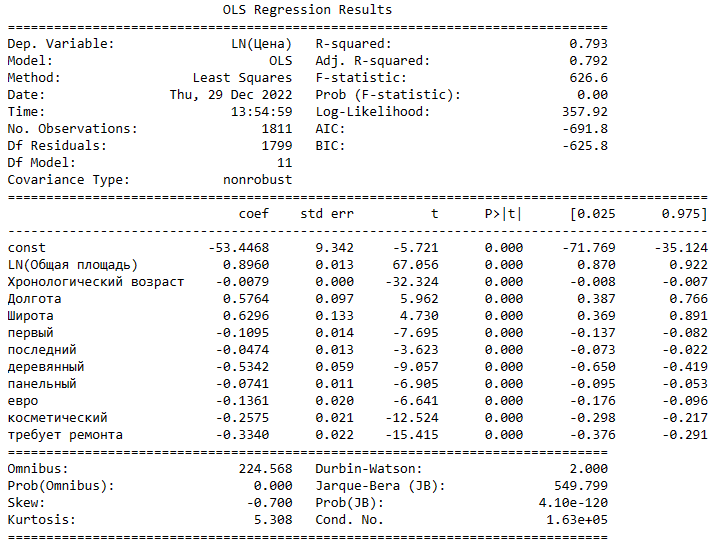
В части выборки 1 наилучшая метрика (R2) составила 77% (см. рисунки ниже).



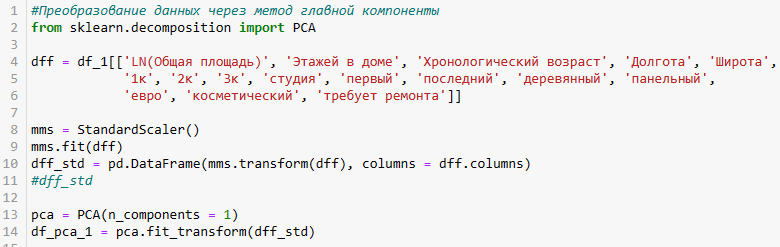


В части выборки 2 наилучшая метрика (R2) составила 79% (см. рисунки ниже).

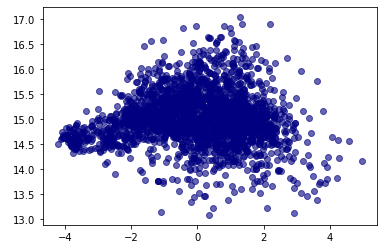




Далее, для исключения выбросов использовался алгоритм KNN. Так как в данных есть категориальные признаки, то все данные были преобразованы в одну переменную с использованием алгоритма PCA.



Зависимость логарифмированной цены от полученного признака приведена на рисунке ниже:

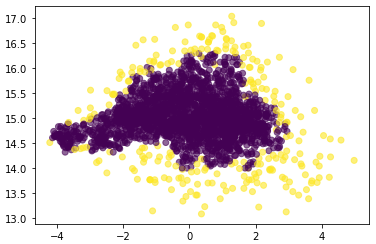


Далее, с использованием библиотеки pyod.models.knn была обучена модель KNN:

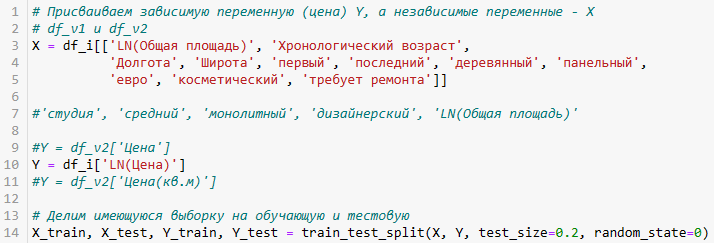




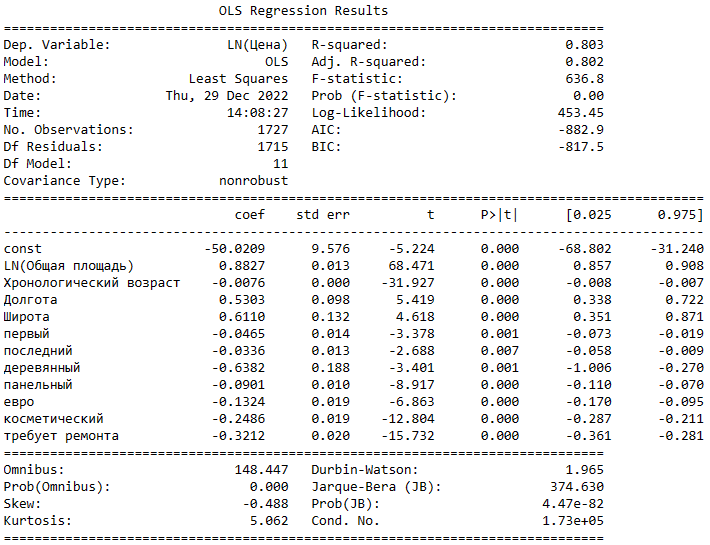
На рисунке ниже приведен графический результат алгоритма, выбросы отмечены желтым цветом:



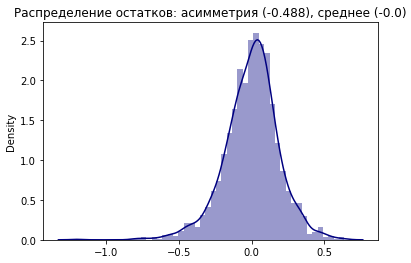
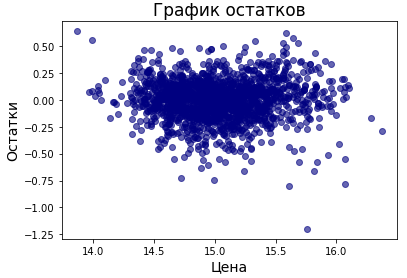
После исключения указанных выбросов была построена новая модель регрессии (также были исключены признаки – комнатность и этажность):



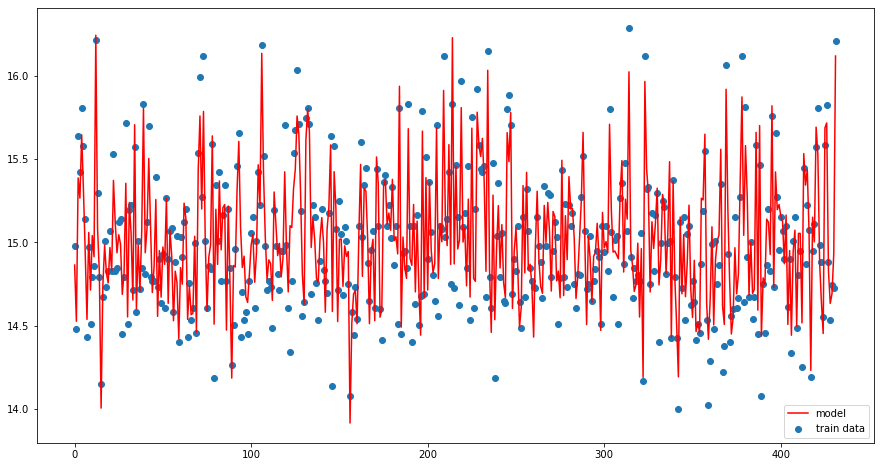
Результаты модели приведены ниже, метрика R2 стала лучше и составила 80%.



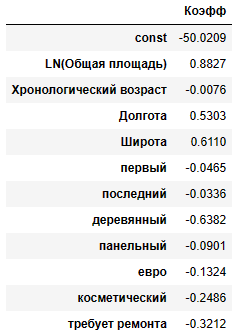
Результат на тестовой выборке составил 84%, результат по кросс-валидации – 79,8%. Согласно приведенным выше результатам, наблюдается мультиколлинеарность. Влияние мультиколлинеарности минимально, так как после исключения из модели факторов, коэффициенты других факторов существенно не изменяются. Остатки распределены ненормально, но гистограмма очень близка к нормальному распределению.



Модель, построенная на данных для обучения, приведена ниже:



Коэффициенты модели приведены ниже:



Нелинейная многофакторная регрессионная модель будет выглядеть следующим образом:

где:

1. LN(S) – логарифм площади квартиры.

2. Тхр – хронологический возраст дома.

3. Д – долгота (местоположение).

4. Ш – широта (местоположение).

5. Э1 – первый этаж.

6. ЭП – последний этаж.

7. Дер – материал стен - дерево.

8. Пан – материал стен - панель.

9. ЕВ – состояние отделки - евроремонт.

10. Кос – состояние отделки - косметический (не требует ремонта).

11. ТР – состояние отделки - требует ремонта.

Также учтено в модели:

Средний этаж – квартиры, расположенные на средних этажах (при этом Э1 и ЭП равны 0).

Дизайнерский ремонт – ремонт из дорогих материалов со сложными решениями (при этом евроремонт, косметический ремонт и требует ремонта равны 0).

Кирпич и монолит – материал стен (объединены в одну группу, при этом дерево и панель равны 0).

Отмечаю, что нормализация и стандартизация (MinMaxScaler и StandardScaler) на результат никакого влияния не оказывают.

**Построение модели оценки стоимости квартир с использованием алгоритма KNN:**

Для построения модели было рассмотрено 5 вариантов выборок:

1. Выборка 1 – без исключения выбросов (с учетом нормализации и стандартизации).

2. Выборка 2 – исключение экстремальных выбросов.

3. Выборка 3 – исключение выбросов по общей площади и цене.

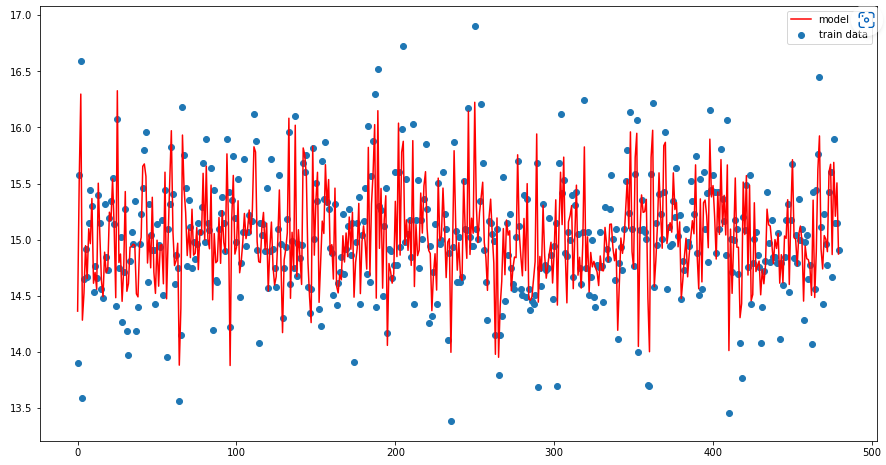
4. Выборка 4 – исключение выбросов по логарифмированным общей площади и цене.

5. Выборка 5 – исключение выбросов алгоритмом KNN.

Лучшее качество модели было получено на первоначальной выборке (Выборка 1) без исключения выбросов (с учетом логарифмирования цены и общей площади). Отмечаю, что лучшая модель была получена при использовании стандартизации данных (StandardScaler). Ниже приведена модель (учтены все факторы), качество модели и график.



R2 на обучении составил 82,3%, на тесте 79,2%, результат по кросс-валидации – 78,1%.



**Построение модели оценки стоимости квартир с использованием алгоритма RandomForest:**

Для построения модели было рассмотрено 5 вариантов выборок:

1. Выборка 1 – без исключения выбросов (с учетом нормализации и стандартизации).

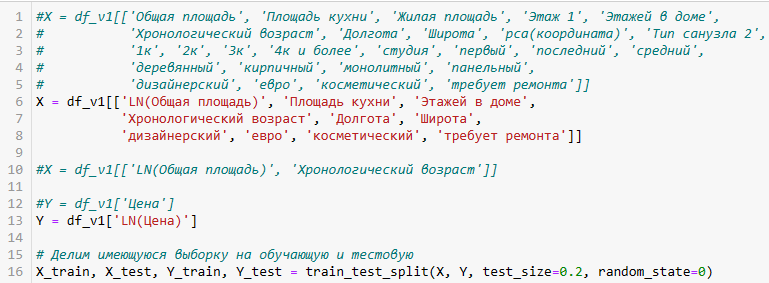
2. Выборка 2 – исключение экстремальных выбросов.

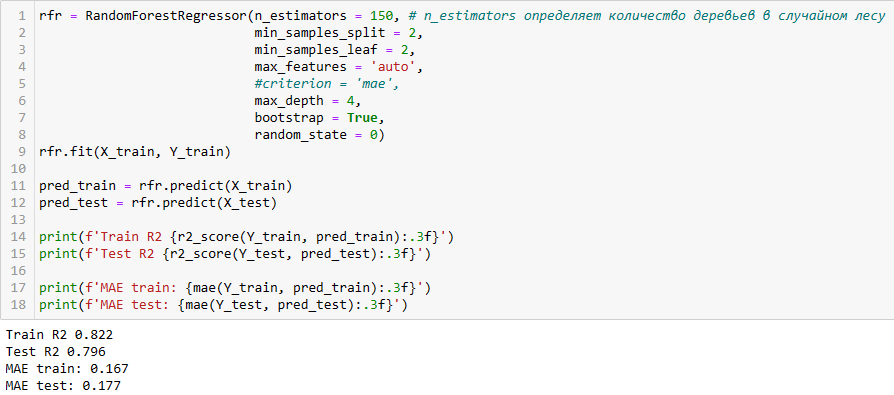
3. Выборка 3 – исключение выбросов по общей площади и цене.

4. Выборка 4 – исключение выбросов по логарифмированным общей площади и цене.

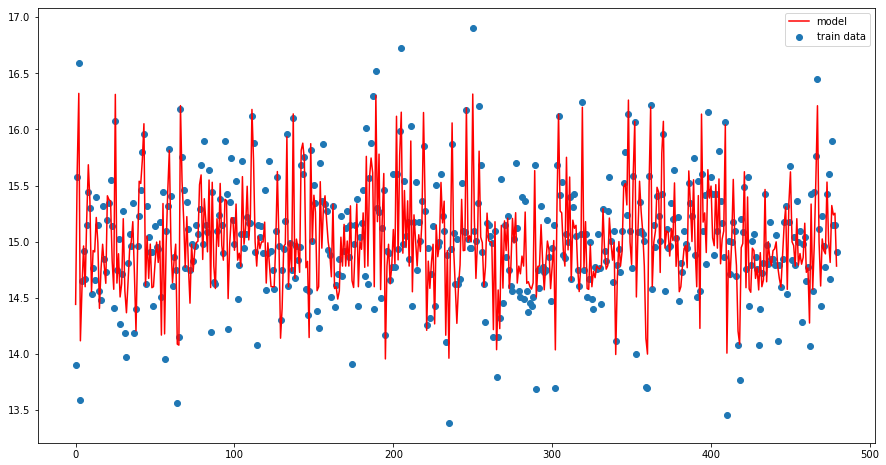
5. Выборка 5 – исключение выбросов алгоритмом KNN.

Лучшее качество модели было получено на первоначальной выборке (Выборка 1) без исключения выбросов (с учетом логарифмирования цены и общей площади). Нормализация и стандартизация не делают модель лучше. Ниже приведена модель, качество модели и график.





R2 на обучении составил 82,2%, на тесте 79,6%, результат по кросс-валидации – 79,3%.



**Построение модели оценки стоимости квартир с использованием алгоритма Bagging:**

Для построения модели было рассмотрено 5 вариантов выборок:

1. Выборка 1 – без исключения выбросов (с учетом нормализации и стандартизации).

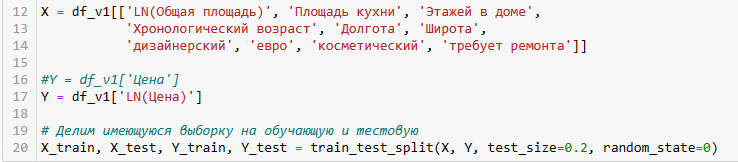
2. Выборка 2 – исключение экстремальных выбросов.

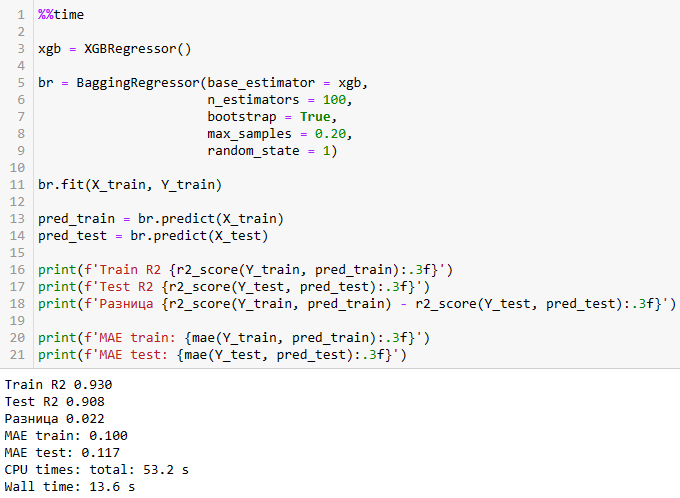
3. Выборка 3 – исключение выбросов по общей площади и цене.

4. Выборка 4 – исключение выбросов по логарифмированным общей площади и цене.

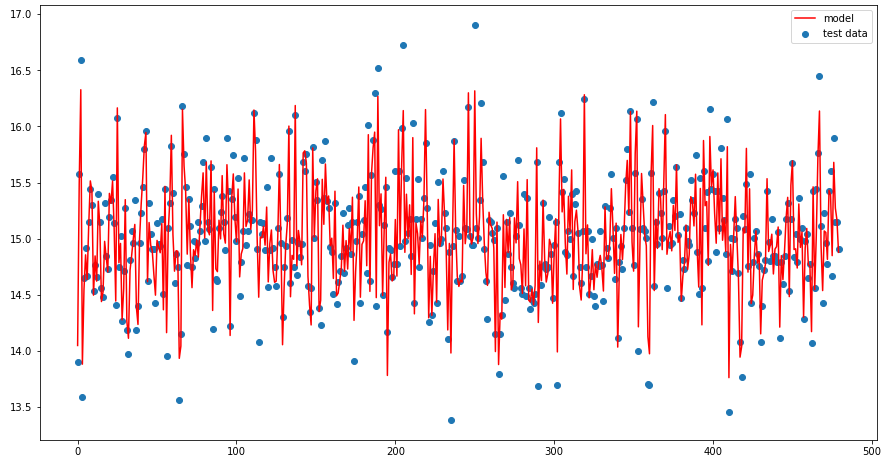
5. Выборка 5 – исключение выбросов алгоритмом KNN.

Лучшее качество модели было получено на первоначальной выборке (Выборка 1) без исключения выбросов (с учетом логарифмирования цены и общей площади), при этом, логарифмировать цену – необязательно. Нормализация и стандартизация не делают модель лучше (никак не сказываются на результате). Лучшее качество было получено при использовании алгоритма XGBRegressor. Ниже приведена модель, качество модели и график.





R2 на обучении составил 93%, на тесте 90,8%, результат по кросс-валидации – 88,8%. Также, хороший результат дает алгоритм CatBoostRegressor: на обучении 92,7%, на тесте 90,5%, кросс-валидация 82%. Нормализация и стандартизация на результате не сказываются.



**Построение модели оценки стоимости квартир с использованием алгоритма CatBoostRegressor:**

Для построения модели было рассмотрено 5 вариантов выборок:

1. Выборка 1 – без исключения выбросов.

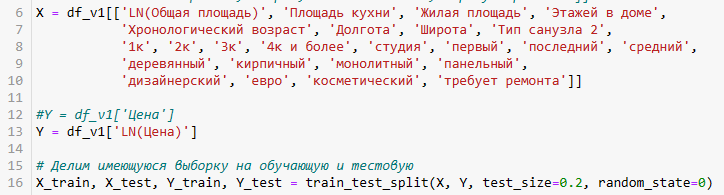
2. Выборка 2 – исключение экстремальных выбросов.

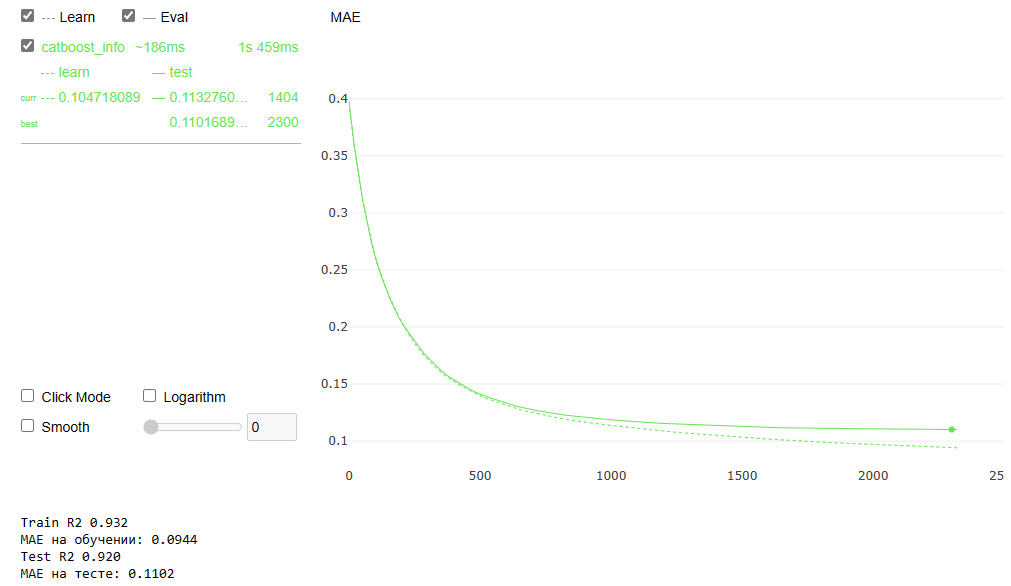
3. Выборка 3 – исключение выбросов по общей площади и цене.

4. Выборка 4 – исключение выбросов по логарифмированным общей площади и цене.

5. Выборка 5 – исключение выбросов алгоритмом KNN.

Лучшее качество модели было получено на первоначальной выборке (Выборка 1) без исключения выбросов (с учетом логарифмирования цены и общей площади). Нормализация и стандартизация не делают модель лучше (никак не сказываются на результате). Ниже приведена модель, качество модели и график.

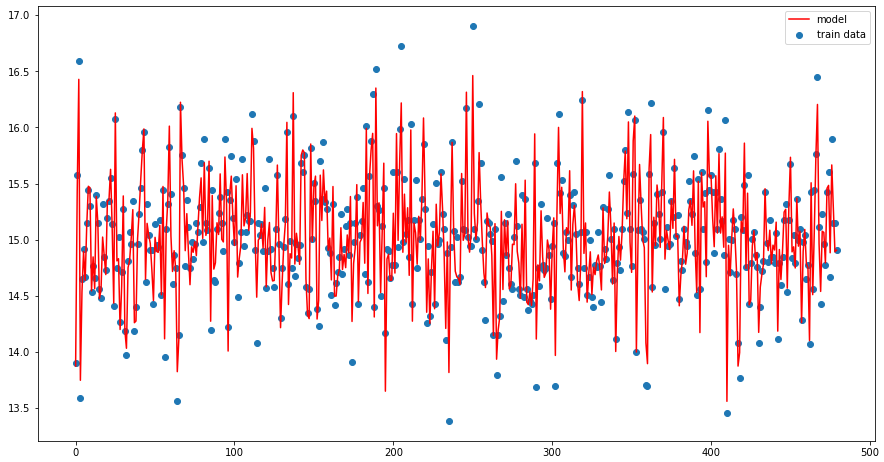




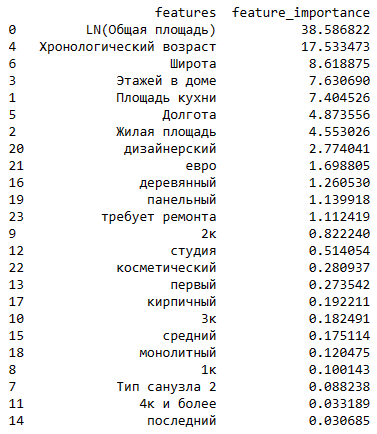
Лучшие параметры модели:



R2 на обучении составил 93,2%, на тесте 92%, результат по кросс-валидации – 90,59%. Нормализация и стандартизация на результате не сказываются.



Важность факторов:



**Построение модели оценки стоимости квартир с использованием Neural Network:**

Для построения модели было рассмотрено 5 вариантов выборок (с обязательной стандартизацией):

1. Выборка 1 – без исключения выбросов.

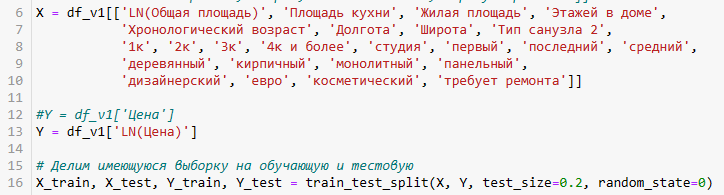
2. Выборка 2 – исключение экстремальных выбросов.

3. Выборка 3 – исключение выбросов по общей площади и цене.

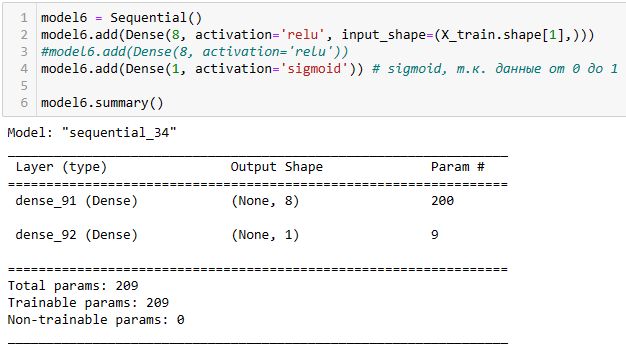
4. Выборка 4 – исключение выбросов по логарифмированным общей площади и цене.

5. Выборка 5 – исключение выбросов алгоритмом KNN.

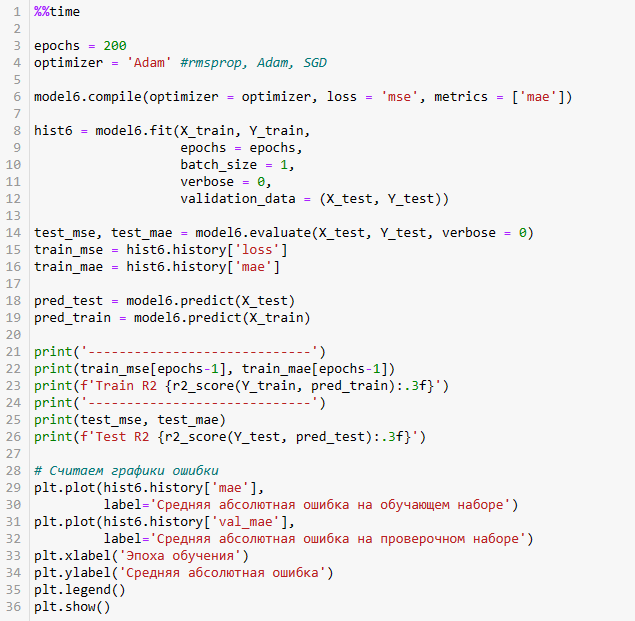
Лучшее качество модели было получено на первоначальной выборке (Выборка 1) без исключения выбросов (с учетом логарифмирования цены и общей площади). Данные по факторам были стандартизированы, данные по цене – нормализированы в диапазоне от 0 до 1. Ниже приведена модель, качество модели и графики.



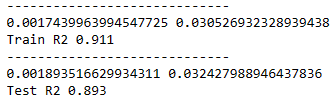
Нейронная сеть получилось простой по структуре, с одним слоем, на входе 8 нейронов, на выходе с функцией активации «sigmoid».



В качестве оптимизатора использовался «Adam», так как с его использованием получается наилучший результат. В качестве функции потерь принята MSE, в качестве метрики – MAE. Для получения хорошего результат были приняты: epochs = 200, batch\_size = 1.

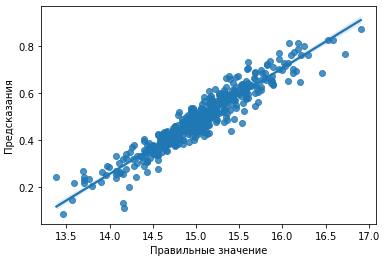
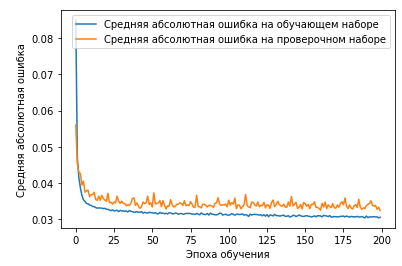


Результаты работы сети приведены ниже:



R2 на обучении составил 91,1%, на тесте 89,3%, кросс-валидация не делалась.

Графики метрики MAE и зависимости правильных значений цен от предсказанных приведены ниже.



**Построение модели оценки стоимости квартир с использованием MLPRegressor:**

Для построения модели было рассмотрено 5 вариантов выборок (с обязательной стандартизацией):

1. Выборка 1 – без исключения выбросов.

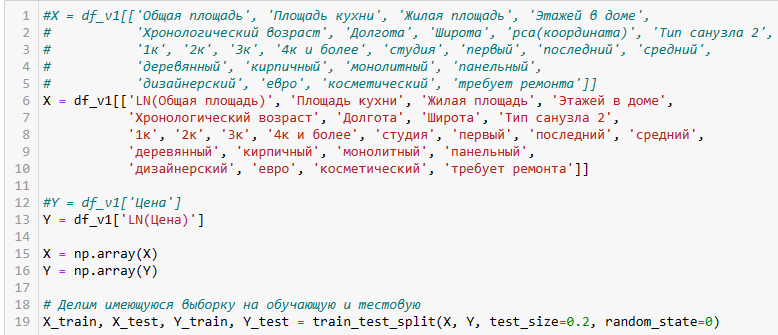
2. Выборка 2 – исключение экстремальных выбросов.

3. Выборка 3 – исключение выбросов по общей площади и цене.

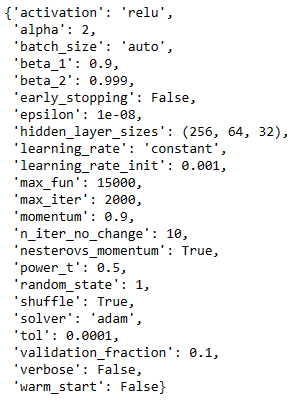
4. Выборка 4 – исключение выбросов по логарифмированным общей площади и цене.

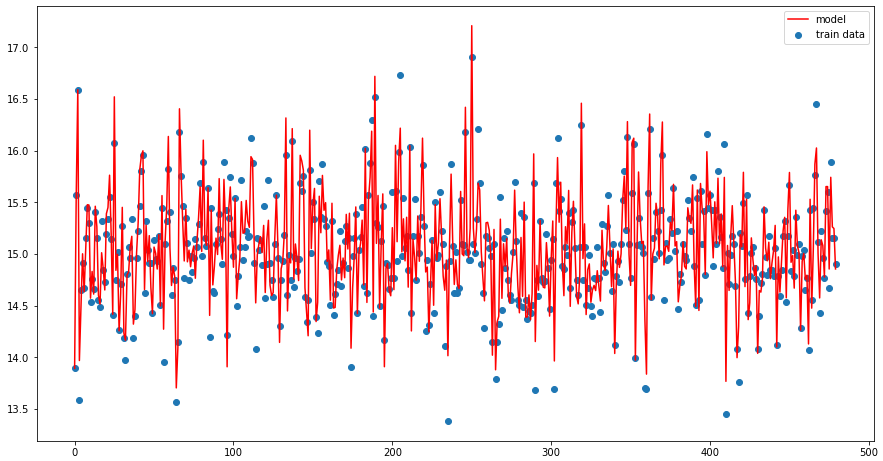
5. Выборка 5 – исключение выбросов алгоритмом KNN.

Лучшее качество модели было получено на первоначальной выборке (Выборка 1) без исключения выбросов (с учетом логарифмирования цены и общей площади). Данные по факторам были стандартизированы, данные по цене – нет. Ниже приведена модель, качество модели и графики.



Лучшие параметры модели:





R2 на обучении составил 90,8%, на тесте 89,3%, результат по кросс-валидации – 84,5%.

**Построение модели оценки стоимости квартир с использованием алгоритма Stacking:**

Для построения модели было рассмотрено 5 вариантов выборок:

1. Выборка 1 – без исключения выбросов (с учетом нормализации и стандартизации).

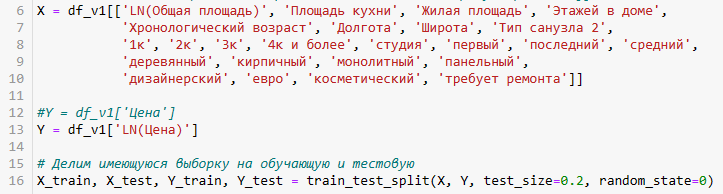
2. Выборка 2 – исключение экстремальных выбросов.

3. Выборка 3 – исключение выбросов по общей площади и цене.

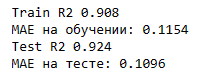
4. Выборка 4 – исключение выбросов по логарифмированным общей площади и цене.

5. Выборка 5 – исключение выбросов алгоритмом KNN.

Лучшее качество модели было получено на первоначальной выборке (Выборка 1) без исключения выбросов (с учетом логарифмирования цены и общей площади). Лучшее качество было получено при использовании алгоритма бустинговых алгоритмов, в качестве мета алгоритма использовалась линейная регрессия. Ниже приведена модель, качество модели и график.







R2 на обучении составил 90,8%, на тесте 92,4%.

