**Постановка задачи**

Цель настоящей работы – проанализировать и отобрать признаки, влияющие на оценку недвижимости. Также проанализировать возможность определения рыночной стоимости объектов недвижимости на основе технологий машинного обучения и оценить результаты оценки стоимости недвижимости

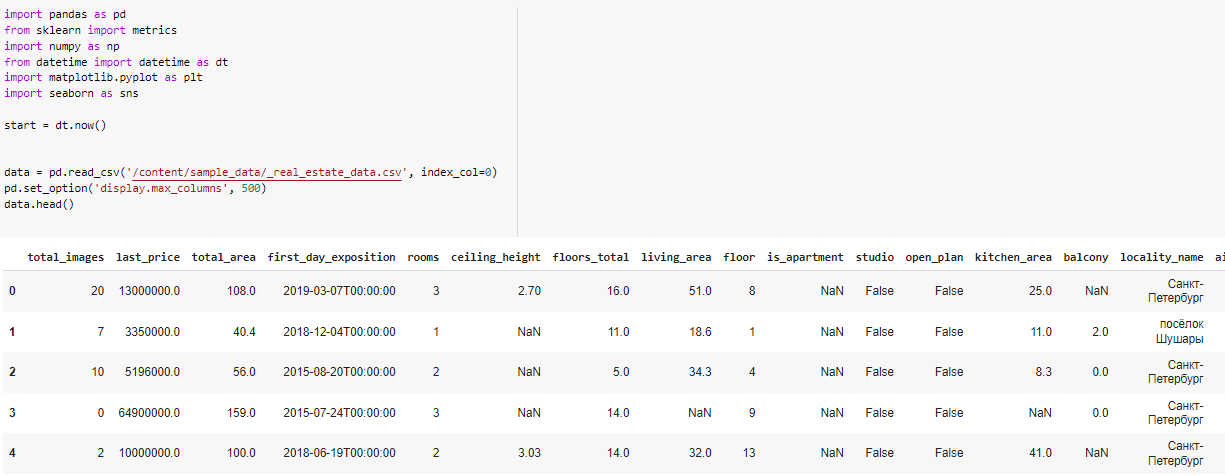
Данная работа имеет 2 основные задачи: отбор информативных признаков и прогнозирование целевой переменной.

# **Данные. Подготовка данных.**

**Загрузка данных**

Для практической части работы используется файл с выборкой данных с названием \_real\_estate\_data.csv, найденный на площадке kaggle.com (система организации конкурсов по исследованию данных). В нем хранятся собранные данные архивов объявлений о продаже квартир в городе Санкт-Петербург и рядом расположенных населенных пунктов за 2015-2019 годы.

Данный файл открывается в сервисе colab через модуль импортируемой библиотеки pandas, выводятся первые 5 строк из файла.



Исходные данные

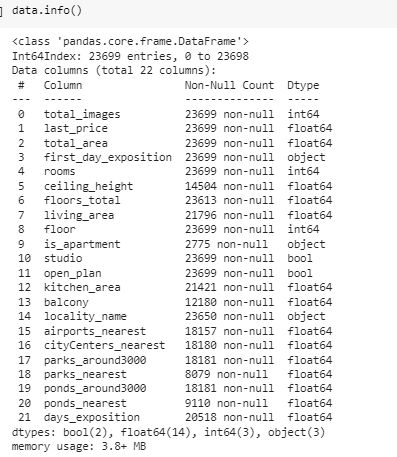
С помощью функции shape выводится размер данных. Данные состоят из 22 признаков и 23699 строк.

Признаки:

* airports\_nearest — расстояние до ближайшего аэропорта (м)
* balcony — количество балконов
* ceiling\_height — высота потолков (м)
* cityCenters\_nearest — расстояние до центра города (м)
* days\_exposition — сколько дней было размещено объявление
* first\_day\_exposition — дата публикации
* floor — этаж
* floors\_total — количество этажей в доме
* is\_apartment — апартаменты (True, False)
* kitchen\_area — площадь кухни (м²)
* last\_price — цена
* living\_area — жилая площадь (м²)
* locality\_name — название населённого пункта
* open\_plan — свободная планировка (True, False)
* parks\_around3000 — количество парков в радиусе 3 км
* parks\_nearest — расстояние до ближайшего парка (м)
* ponds\_around3000 — количество водоёмов в радиусе 3 км
* ponds\_nearest — расстояние до ближайшего водоёма (м)
* rooms — число комнат
* studio — квартира-студия (True, False)
* total\_area — площадь квартиры (м²)
* total\_images — количество фотографий квартиры в объявлении

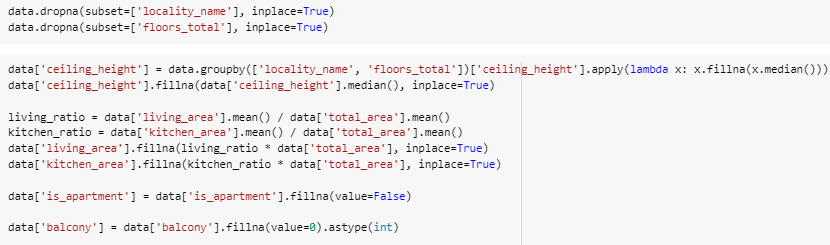
**Предварительная обработка данных**

При первичном анализе данных можно сделать вывод о большом количестве пропущенных значений и неверных типов данных. Для корректной работы модели данные необходимо перевести в числовой формат.



Информация о данных

Удаляются строки, где отсутствует локация и этажность. В данных столбцах отсутствует не более 1% значений, поэтому удаление будет не критично. Выполняется замена некоторых типов данных и заполнение пустых значений медианами.



Обработка данных

Рассчитывается и добавляется в датасет столбец со стоимостью за 1 м2 путем деления стоимости на жилую площадь квартиры.



Добавление столбца «Цена за 1 м2»

Необходимо сгруппировать по району и заполнить медианой и средним значением пропуски в столбцах цена за м2 и расстояние до ближайшего аэропорта, значения которых могут зависеть от района. Таким же образом заполняются пустые значения в столбце расстояние до центра по среднему значению.



Обработка данных

В столбцах апартаменты, студия, свободная планировка данные необходимо перевести в булево значение типа int (целое число) 0 или 1.

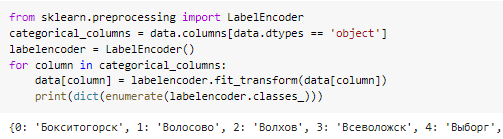
Заполнение пустых значений в столбцах с количеством парков в радиусе 3000км, расстоянием до ближайшего парка, и количеством водоемов в радиусе 3000км происходит путем подстановки средних значений. Пустые значения столбца с количеством дней размещенного объявления заполняются медианой.

Далее необходимо избавиться от даты в столбце для корректной работы модели. Для этого столбец с датой можно разделить на три столбца с годом, месяцем и днем недели. Теперь от столбца с датой публикации можно избавиться.



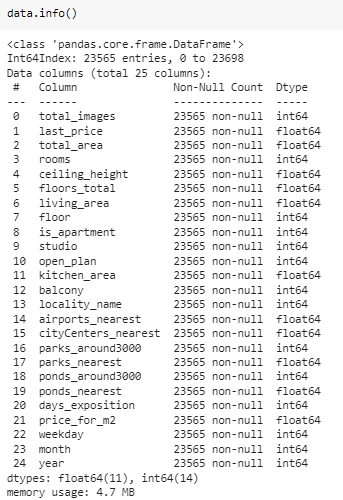
Обработка данных

Затем столбцы с категориальными признаками заменяются на числа.



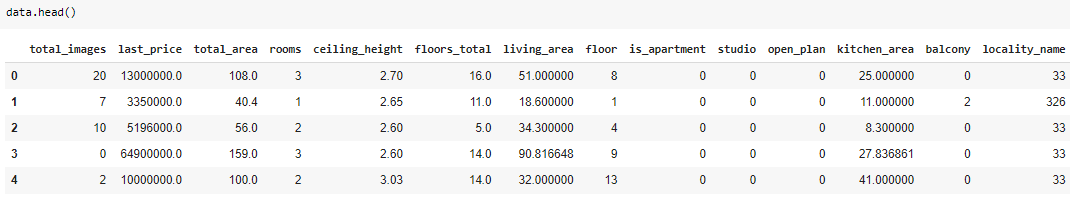
Обработка категориальных данных

На данном этапе имеем 25 признаков и 23565 строк. Теперь в данных нет пустых значений и типы данных изменены:



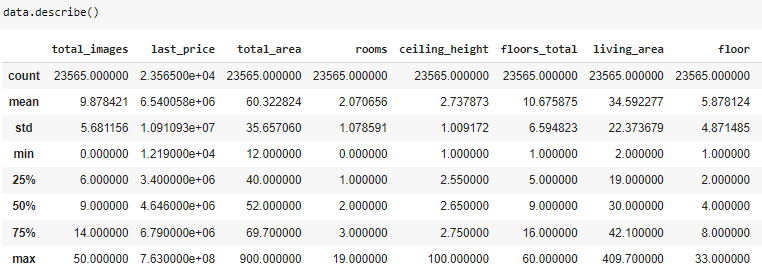
Информация об обработанных данных

Получаем следующий вид данных:



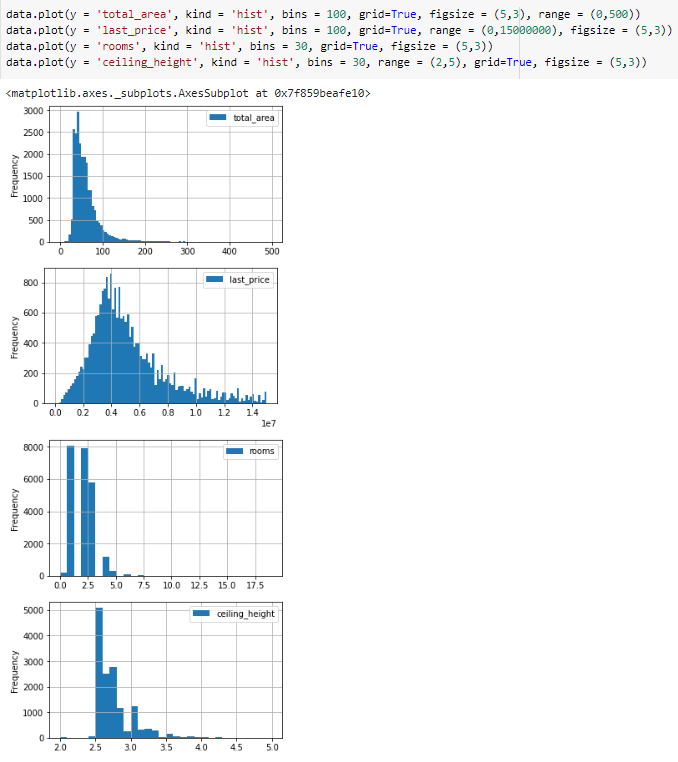
Вид данных

Следует изучить, есть ли выбивающиеся значения с помощью основных статистических показателей. В частности, мы видим количество наблюдений (count), среднее арифметическое (mean), среднее квадратическое отклонение (std), минимальное (min) и максимальное (max) значения, а также первый (25%), второй (50%) и третий (75%) квартиль каждой количественной переменной. Здесь можно заметить, что по ряду признаков есть аномальные значения.



Статистические показатели

Для наглядности построены гистограммы некоторых признаков, а именно, общая площадь квартиры, цены, количество комнат и высота потолков.



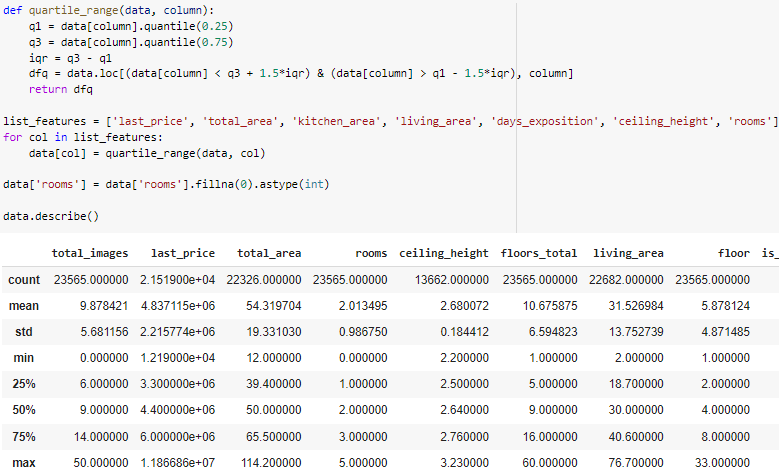
Построение гистограмм

Исходя из выведенных статистических показателей и гистограмм можно сделать вывод о том, что:

* минимальная площадь квартир составляет 12 м2, максимальная - 900 м2, а среднее значение площади квартир равно 60 м2;
* средняя стоимость квартир составляет 6540000 руб.;
* в основном количество комнат продаваемых квартир лежит в границах от 1 до 4;
* средняя высота потолков составляет 2,7 м.

В данных существуют некоторые аномальные значения, например, максимальная высота потолков в выборке, равная 100 м. От таких значений следует избавиться, используя перцентили.

Для обнаружения выбросов в данных воспользовались перцентилем и отбросили значения на основе 1.5 размаха, после чего значения стали лучше и адекватнее.



Обнаружение выбросов

На крайнем этапе подготовки данных получаем выборку без пустых строк и выбросов, состоящая из 25 столбцов и 18131 строк.

Статистические показатели:

* минимальная площадь квартир составляет 12 м2, максимальная - 114 м2, а среднее значение площади квартир равно 54 м2;
* средняя стоимость квартир составляет 4837115 руб.;
* среднее количество комнат равно 2;
* средняя высота потолков составляет 2,68 м.

По таким «расчищенным» данным можно строить модели оценки стоимости недвижимости.

# **Отбор признаков с помощью фильтров. Оценка стоимости недвижимости с применением линейной регрессии**

Теперь следует отобрать значимые признаки с помощью метода фильтрации, а именно корреляции. С помощь метода corr() вычисляется попарная корреляция столбцов. Здесь необходимо рассмотреть корреляцию к целевой переменной - last\_price (цена на жилье). Чем ближе значение к 1, тем сильнее корреляция.



Корреляция

Из таблицы видно, что заметная корреляция присутствует между общей площадью и стоимостью квартиры (коэффициент корреляции 0,67), между жилой площадью и стоимостью (коэффициент корреляции 0,55). Так же умеренную корреляцию от 0,3 до 0,5 можно наблюдать у переменных: цена за квадратный метр, площадь кухни, количество комнат и высота потолков.

Соответственно, исходя из шкалы Чеддока вышеперечисленные признаки являются значимыми для построения модели прогнозирования стоимости недвижимости.

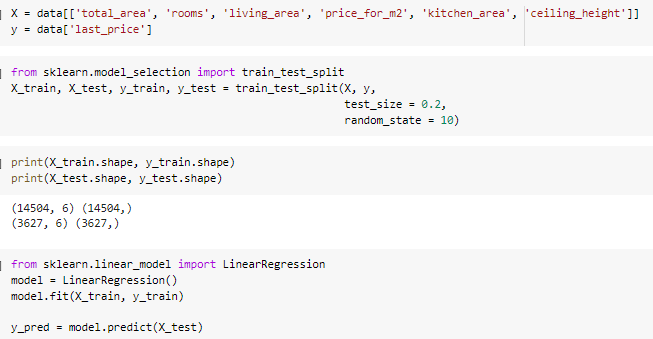
Так же существуют признаки, имеющие слабую корреляцию, менее 0,3: количество этажей в доме, название населенного пункта, расстояние до центра города, количество парков и водоемов на расстоянии 3000км. Такие признаки не значимы для целевого параметра, их не стоит вносить в модель.

**Обучение и оценка качества модели**

Теперь после обработки данных и отбора признаков для оценки стоимости недвижимости необходимо обучить модель.

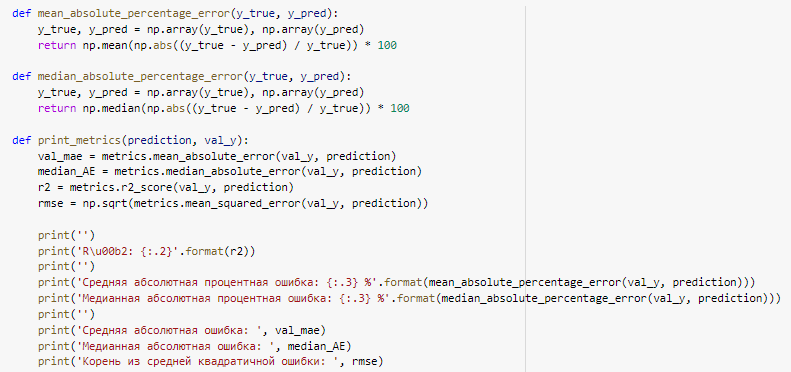
Поскольку вышеперечисленные переменные имеют наиболее высокую корреляцию с целевой переменной, поместим эти признаки в переменную Х, а в Y – стоимость квартиры.

Для этого разобьем данные на тренировочные (80%) и тестовые (20%). С помощью линейной регрессии необходимо обучить модель и предсказать цены.



Обучение. Линейная регрессия

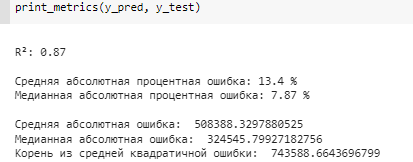
Для оценивания результатов работы алгоритма необходимо импортировать библиотеку с метриками (sklearn.metrics) и создать функции вычисления коэффициента детерминации, средней, медианной абсолютных процентных ошибок и корень из средней квадратичной ошибки:



Функции вычисления метрик

Далее необходимо вызвать данные функции и передать в них реальные тестовые и предсказанные значения.

Коэффициент детерминации R2 модели, оценивающий качество модели достаточно хорош и равен 87%. В данном случае мы видим, что 87% изменчивости цены объясняется независимыми переменными, которые были выбраны для модели.



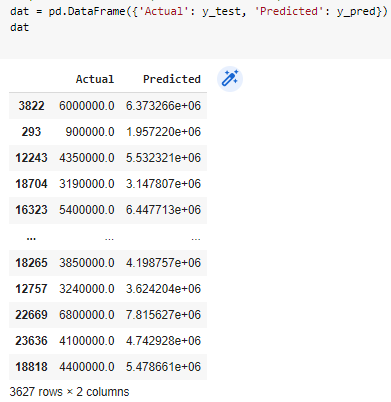
Метрики для построенной модели линейной регрессии

Средняя абсолютная процентная ошибка равна 13,4%, а медианная абсолютная процентная ошибка равна 7,87%. Средняя абсолютная ошибка равна 508388 руб. (11% от среднего значения стоимости квартиры 4618170 руб.), а медианная абсолютная ошибка равна 324545 руб. (7% от среднего значения).

Можно заметить, что корень из среднеквадратичной ошибки составляет 743588 руб. (16% от среднего значения). Это означает, что алгоритм проделал хорошую работу.

Теперь можно вывести актуальные (тестовые) и предсказанные цены.

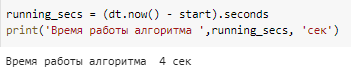
Хоть и модель не очень точна, прогнозируемые оценки близки к фактическим.



Сравнение реальных и предсказанных цен

Таким образом, создана неплохая модель линейной регрессии для оценки стоимости жилой недвижимости. Глядя на предыдущий рисунок можно сравнить актуальные (тестовые) и предсказанные цены на квартиры. Например, для квартиры в строке 3822, имеющая актуальную цену 6000000 руб., модель предсказала цену равную 6373266 руб. Таким образом, модель ошиблась на 373266 руб., а это на 6% больше актуальной стоимости квартиры. Хоть и модель не очень точна, прогнозируемые оценки стоимости жилой недвижимости близки к актуальным.

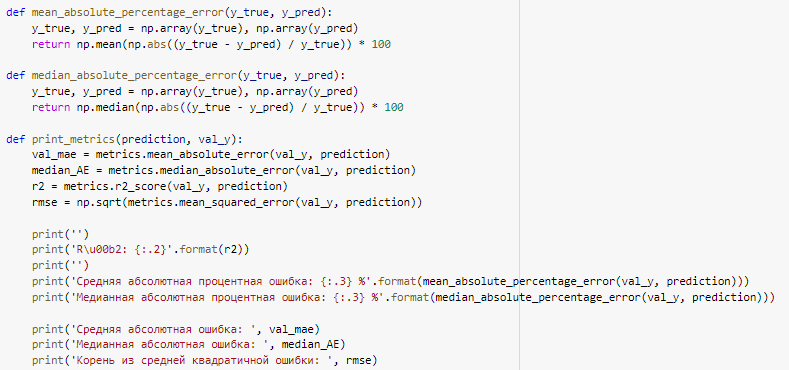
Время работы данного алгоритма составила 4 секунды.



5.18. Время работы алгоритма

# **Отбор признаков и оценка стоимости недвижимости с применением Random Forest**

Созданы функции, вычисляющие среднюю, медианную абсолютную ошибку, корень из средней квадратичной ошибки и коэффициент детерминации R2.



Функции для вычисления метрик

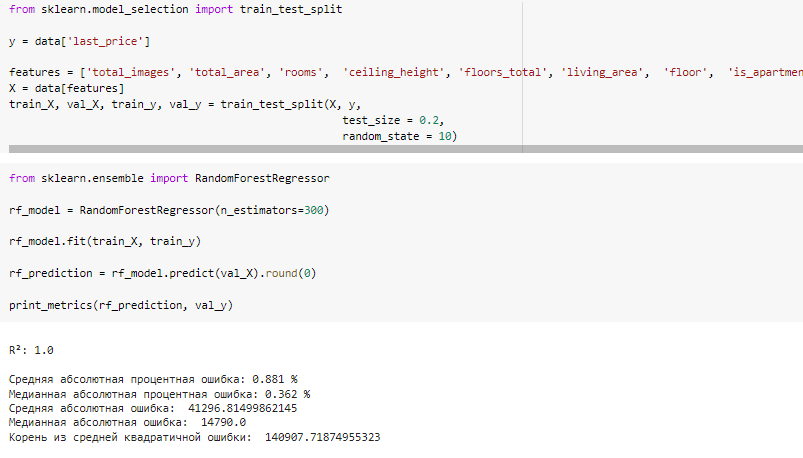
Затем необходимо разбить данные на тренировочные (80%) и тестовые (20%).

Переменная Y – стоимость квартиры.

Переменная X – все признаки:

1. total\_images — количество фотографий квартиры в объявлении
2. total\_area — площадь квартиры (м²)
3. rooms — число комнат
4. ceiling\_height — высота потолков (м)
5. floors\_total — количество этажей в доме
6. living\_area — жилая площадь (м²)
7. floor — этаж
8. is\_apartment — апартаменты (True, False)
9. studio — квартира-студия (True, False)
10. open\_plan — свободная планировка (True, False)
11. kitchen\_area — площадь кухни (м²)
12. balcony — количество балконов
13. locality\_name — название населённого пункта
14. airports\_nearest — расстояние до ближайшего аэропорта (м)
15. cityCenters\_nearest — расстояние до центра города (м)
16. parks\_around3000 — количество парков в радиусе 3 км
17. parks\_nearest — расстояние до ближайшего парка (м)
18. ponds\_around3000 — количество водоёмов в радиусе 3 км
19. ponds\_nearest — расстояние до ближайшего водоёма (м)
20. days\_exposition — сколько дней было размещено объявление
21. price\_for\_m2 — цена за м2
22. weekday — день недели
23. month — месяц
24. year — год

Далее создан объект RandomForestRegressor (ранее импортированный из библиотеки Scikit-learn) и задана ему настройка - количество деревьев: 300. В это время строятся и оптимизируются деревья. Когда модель обучится (за это отвечает метод fit), метод predict предскажет цены для тестовой выборки квартир (val\_X). А созданная нами функция print\_metrics() получит на вход предсказанные и известные нам заранее актуальные цены, и напечатает значения метрик.



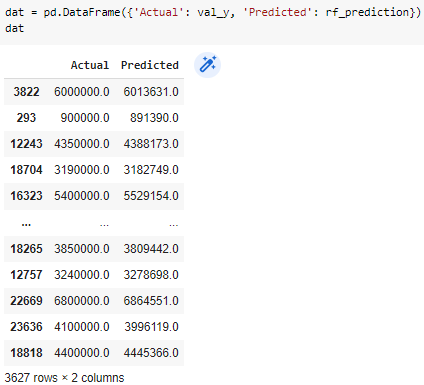
Обучение и вывод метрик

Получен R2, равный 1, а значит, 100% изменчивости цены объясняется с помощью построенной модели. Это отличный результат.

Так же получены значения других метрик:

* средняя абсолютная процентная ошибка: 0,88%;
* медианная абсолютная процентная ошибка: 0,36%;
* средняя абсолютная ошибка: 41296 руб. (0.8% от среднего значения);
* медианная абсолютная ошибка: 14790 руб. (0,3%);
* корень из средней квадратичной ошибки: 140907.72 руб. (3% от среднего).

Показатели данных метрик ошибок алгоритма являются очень низкими, менее 3%, что говорит о высокой точности предсказания.



Сравнение реальных и предсказанных цен

Глядя на рисунок со сравнением реальных и предсказанных цен можно убедиться в высокой точности предсказания, например, для квартиры в строке 3822, имеющая актуальную цену 6000000 руб., предсказана цена – 6013631 руб. А значит, модель ошиблась на 13631 руб., а это 0,2% от актуальной цены.

Так же в процессе обучения модель отобрала важные для построения признаки:

. 

Построение диаграммы «Важность признаков»

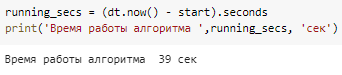
Можно увидеть, какие признаки наиболее значимы для модели на диаграмме, где по оси х – номер признака, а по оси y – значимость:



Диаграмма значимости признаков

А значит, признаки, играющие важную роль: цена за квадратный метр (45%), общая площадь (41%), жилая площадь квартиры (12%). Для модели случайного леса этих признаков достаточно, чтобы верно предсказывать стоимость жилой недвижимости.

Работа данного алгоритма составила 39 секунд:



Время работы алгоритма

# **Результаты эксперимента**

Отбор признаков методом фильтрации, а именно, корреляции показал, что для целевого показателя стоимости жилой недвижимости значимы признаки: общая и жилая площадь квартиры, цена за квадратный метр, площадь кухни, количество комнат и высота потолков. Эти признаки были внесены в модель обучения, после чего с помощью линейной регрессии были предсказаны цены квартир. Были вычислены величины ошибок при сравнении известных цен квартир из тестовой выборки с предсказанными моделью. Средняя и медианная процентные ошибки – менее 14%. Средняя абсолютная ошибка составила 11% от среднего значения стоимостей, средняя медианная ошибка равна 7%, а квадрат из среднеквадратичной ошибки - 743588 руб. (16% от среднего целевого показателя). Коэффициент детерминации R2 показал качество модели на 87%, значит 87% вариации зависимой переменной определяется вариацией независимой переменной. Эти показатели работы алгоритма говорят о его достойной работе.

Алгоритм случайного леса отобрав как значимые признаки цену за квадратный метр, общую и жилую площадь показал превосходную работу. Были спрогнозированы цены, максимально близкие к актуальным. Средняя и медианная процентные ошибки – менее 1%. Средняя и медианная абсолютная ошибки составили менее 1% от среднего значения, квадрат из среднеквадратичной ошибки равен 140907 руб., что равно 3% от среднего значения стоимостей квартир. Величина коэффициента детерминации R2 составила 1, т.е. 100%, что говорит об идеально полученной модели, в которой все точки наблюдений лежат точно на линии регрессии.

Таким образом, алгоритм случайного леса оказался лучше линейной регрессии и проделал идеальную работу, спрогнозировав стоимости на квартиры с высокой точностью, но затратив на работу алгоритма на 35 секунд больше, по сравнению с линейной регрессией.