|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **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)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Предсказание цен\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ на недвижимость\_\_\_\_\_\_\_\_­\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_в регионе Highland\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_ИУ5-63Б\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_М.Д. Иванченко\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_****Ю.Е. Гапанюк\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 г.*

**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)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_ИУ5\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_В.И. Терехов\_\_

(И.О.Фамилия)

«\_07\_» \_\_\_\_февраля\_\_\_\_\_\_\_ 2024 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_\_\_\_\_\_\_\_\_Предсказание цен на недвижимость в регионе Highland\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_ИУ5-63Б\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Иванченко Максим Дмитриевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ИССЛЕДОВАТЕЛЬСКАЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_КАФЕДРА\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_\_\_ нед., 50% к \_\_\_ нед., 75% к \_\_ нед., 100% к \_\_\_ нед.

Техническое задание \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_Исследовать методы машинного обучения для решения задачи регрессии\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на \_\_17\_\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания « 07 » февраля 2024 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_Ю.Е.Гапанюк\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_**М.Д. Иванченко**\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Содержание**

[**Введение** 4](#_Toc168431032)

[**Постановка задачи** 6](#_Toc168431033)

[**Выполнение работы** 7](#_Toc168431034)

[**Заключение** 16](#_Toc168431035)

[**Список использованной литературы** 17](#_Toc168431036)

# **Введение**

Проблема оценки недвижимости очень актуальна в наше время. Продавцы хотят получить больше, а покупатели не хотят переплачивать за недвижимость. Однако, можно примерно оценить реальную стоимость здания и облегчить выбор для покупателей.

В данной работе мы будем использовать данные, полученные из объявлений продаваемой недвижимости в районе Highland, Шотландия. Мы будем использовать алгоритмы регрессии для определения цены здания, учитывая район, улицу, количество этажей, жилую площадь, количество комнат, количество балконов, расстояние до метро, наличие лифта, количество фотографий в объявлении.

Целью данной работы является разработка эффективной модели, которая может помочь быстро определить вероятную стоимость недвижимость в регионе Highland.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи регрессии.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
6. Выбор метрик для последующей оценки качества моделей.
7. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

# **Постановка задачи**

Данная работа по машинному обучению направлена на решение задачи регрессии, а именно, предсказание цен на недвижимость.

Имеются данные о районе, улице, количестве этажей, жилой площади, количестве комнат, количестве балконов, расстоянии до метро, наличии лифта, количестве фотографий в объявлении каждого здания.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания цен последующих объявлений. Для этого мы будем использовать различные алгоритмы регрессии, такие как линейная регрессия, K-ближайших соседей, дерево решений, случайный лес и градиентный бустинг. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки ее точности и эффективности.

Результатом работы должна быть модель, которая сможет предсказывать цены на недвижимость.

# **Выполнение работы**

Для решения задачи регрессии был выбран набор данных, содержащий информацию объявлений о продаже недвижимости.

В наборе данных присутствуют следующие признаки:

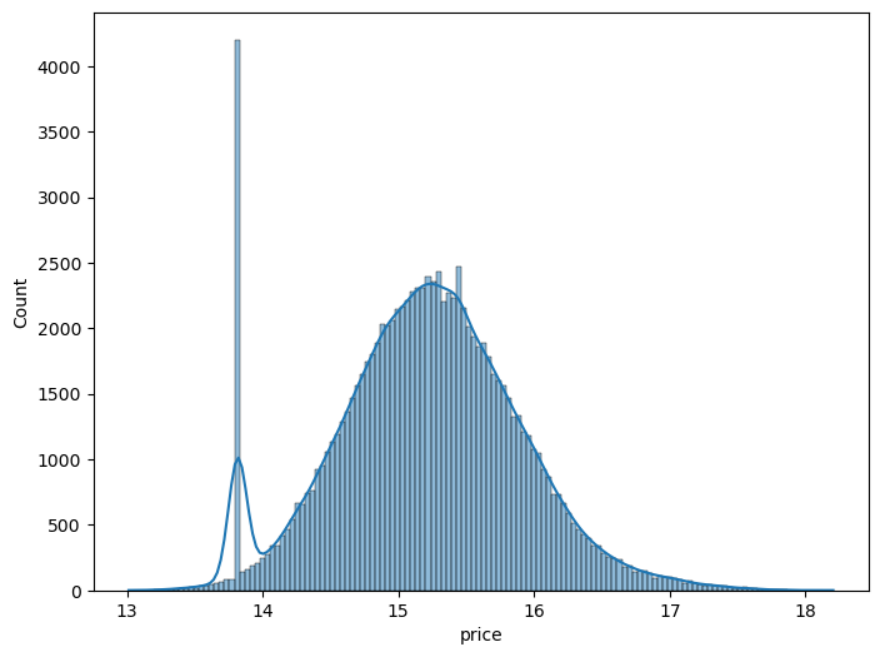
* date – Дата в формате ГОД-МЕСЯЦ
* street\_id – номер улицы
* build\_tech – сторительная методика
* floor – количество этажей
* area – жилая площадь
* rooms – количество комнат
* balcon – количество балконов
* metro\_dist – расстояние до метро
* g\_lift – наличие лифта
* n\_photos – количество фотографий в объявлении
* kw1-kw13 - район Highland

Данный датасет использован для решения задачи регрессии – предсказания цен на недвижимость.

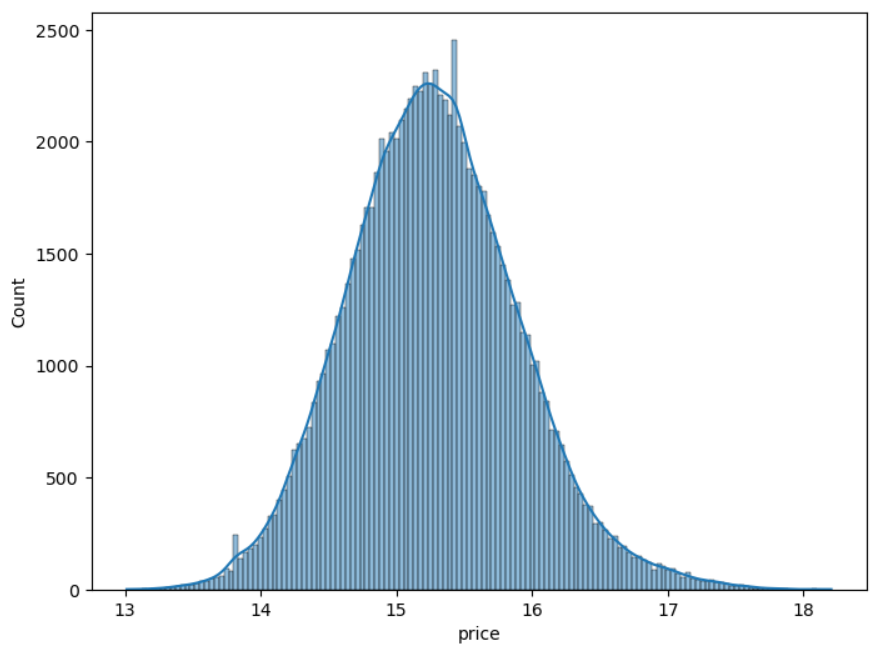
Загружаем данные, получаем общую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 100000 строк и 24 столбца, из которых 20 типа int64, 3 типа float64 и 1 типа object.

Были обнаружены пропуски в трех признаках. Заполняем их модой и средним, группируя по номеру улицы.

Строим график histplot для визуализации количества объявлений для всего диапазона цен, которые были предварительно прологарифмированы:

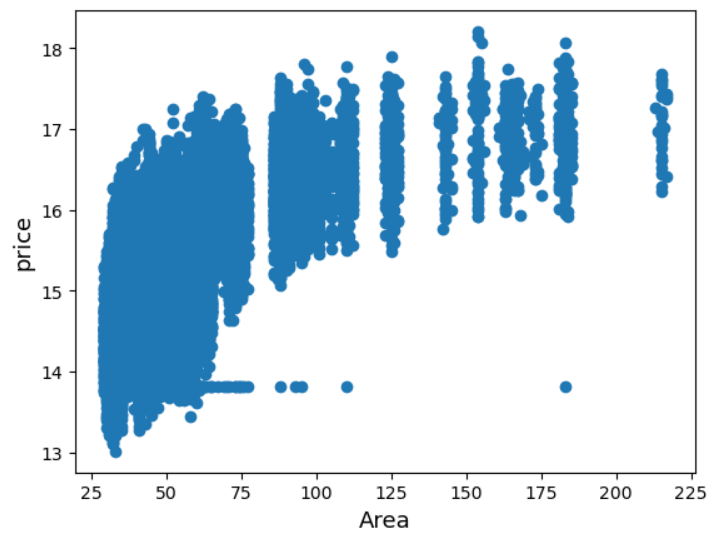


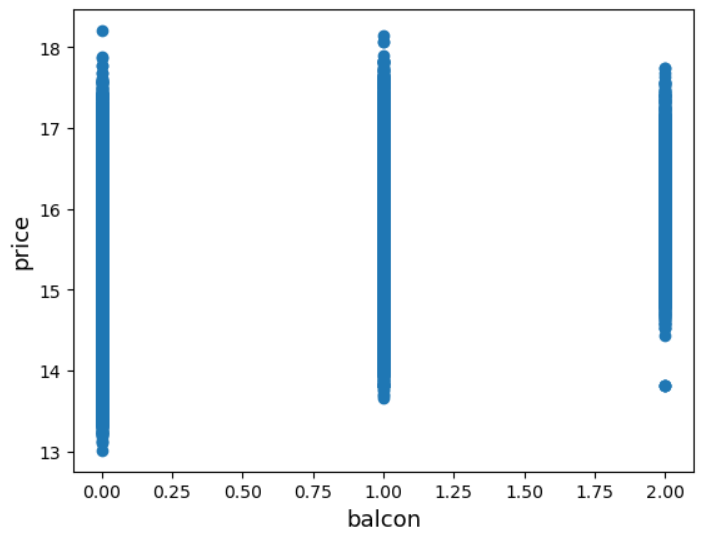
Эмпирическим путем было выяснено, что выделяющемуся столбцу соответствует цена в 1000000 долларов. Их количество было сокращено путем рандомного выбора нескольких семплов. В итоге диаграмма приобрела вид:



Теперь распределение выглядит хорошо, поэтому целевой признак был прологарифмирован.

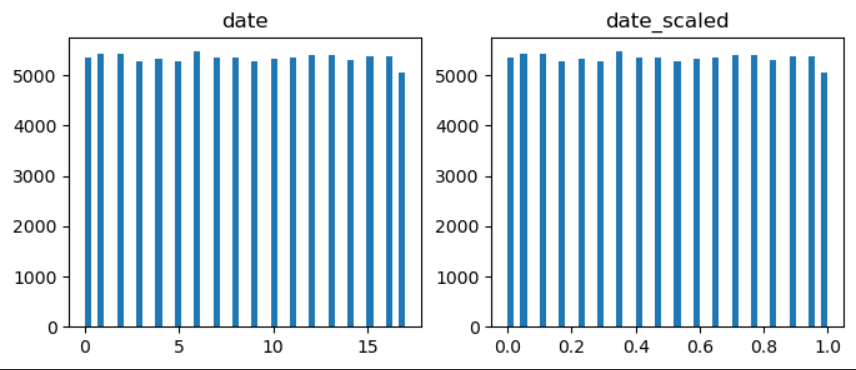
Далее были построены несколько scatter-диаграмм, отражающие зависимость некоторых признаков с целевым:

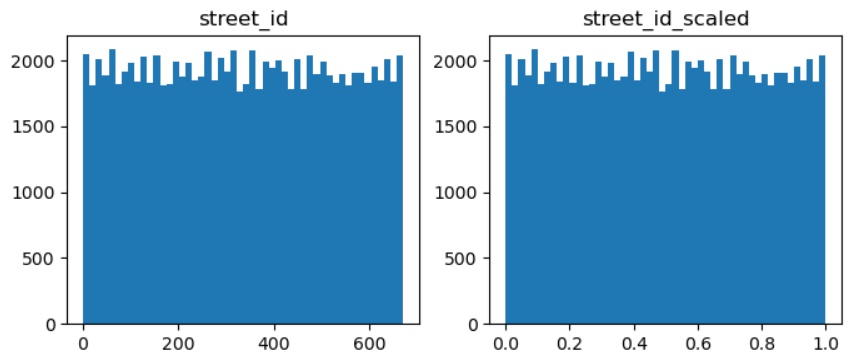


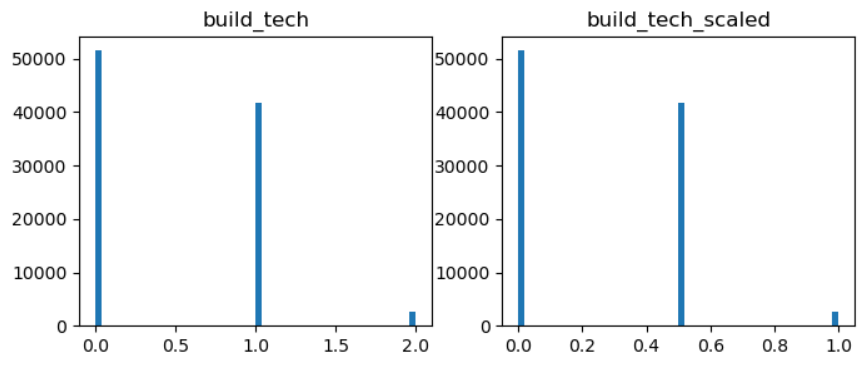


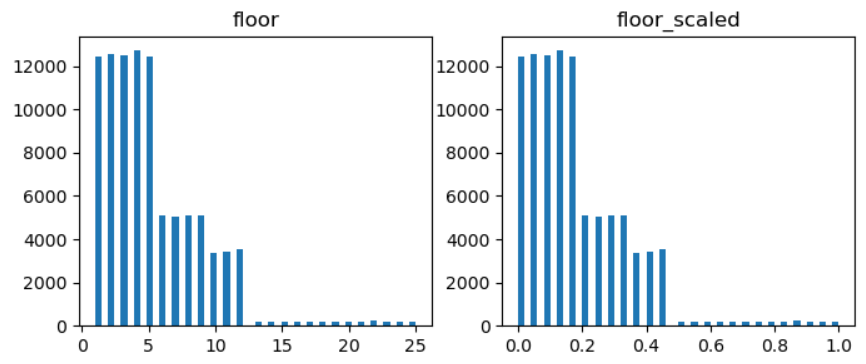
Заметные выборы были удалены из выборки.

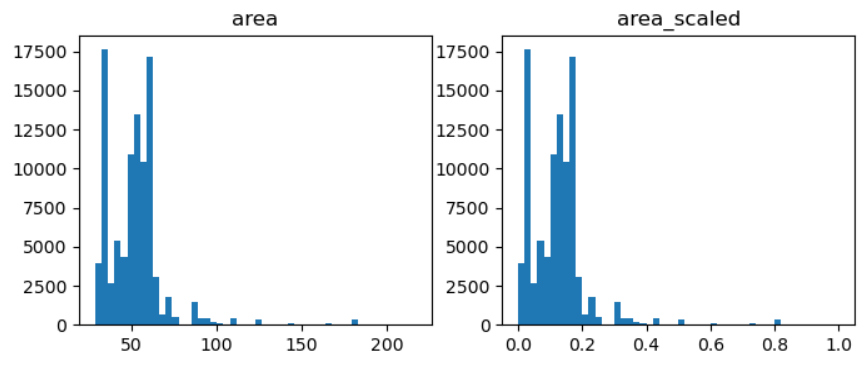
Далее приведем данные к нужному формату. Сначала закодируем даты с помощью LabelEncoder, а затем масштабируем численные признаки методом MinMax Scaler, который преобразует каждый признак таким образом, чтобы он имел среднее значение равное 0 и стандартное отклонение равное 1. Посмотрим на распределения колонок до и после масштабирования.

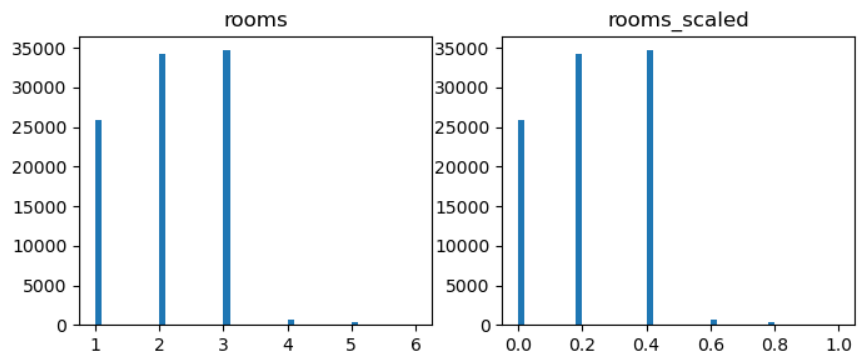


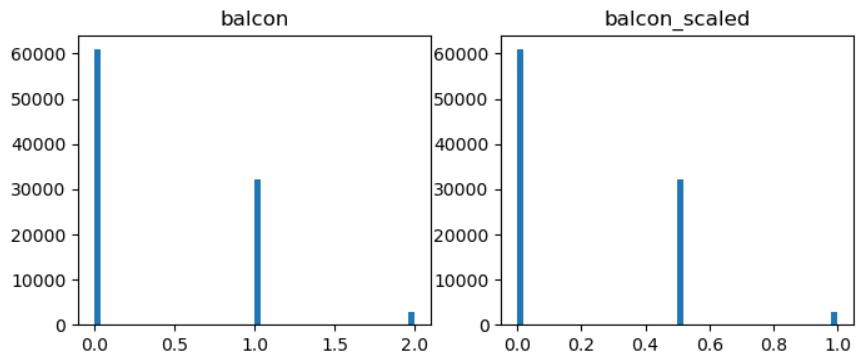


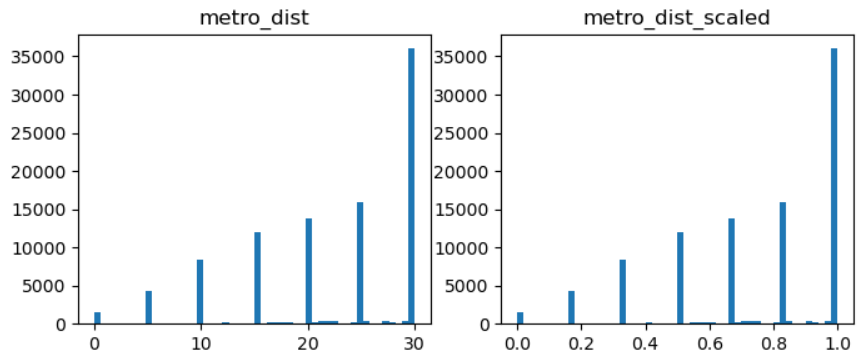


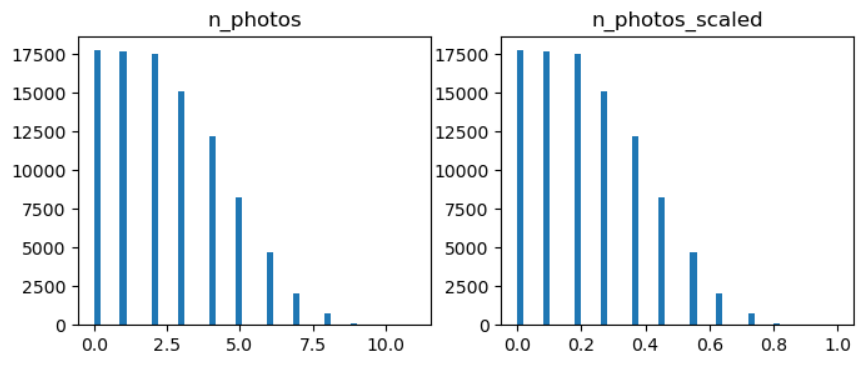






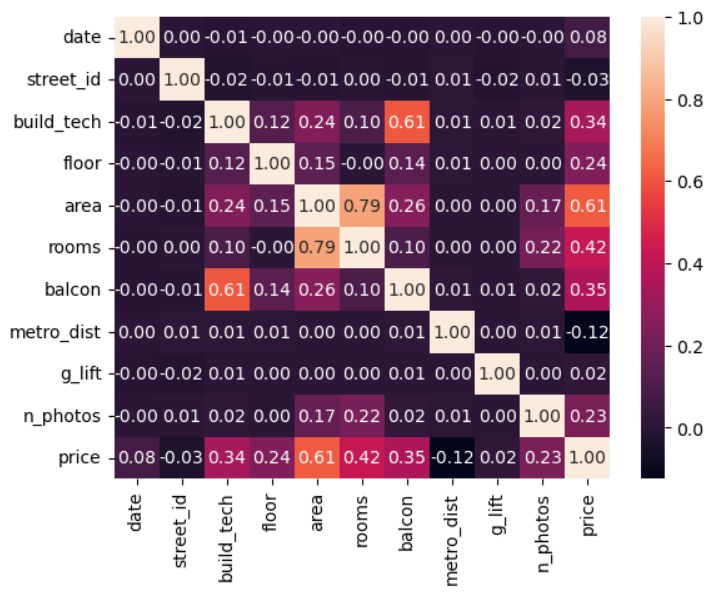






Распределение не изменилось.

Проводим корреляционный анализ данных. Строим тепловую карту корреляций.



Выводы:

* целевой признак price больше всего коррелирует с площадью (0.61);
* расстояние до метро и цена имеют отрицательный коэффициент корреляции (-0.12), но оставим для построения модели, т.к. он тоже оказывает влияние на целевой признак;
* столбцы с датой (0.08), улицей (-0.03) и наличием лифта (0.02) тоже оставим, т.к. выявлена хоть и небольшая, но возможность влияния на целевой признак;

Выберем метрики для оценки качества модели:

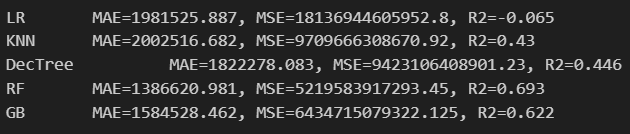
* Mean absolute error – средняя абсолютная ошибка.
* Mean squared error – средняя квадратичная ошибка.
* Метрика R2 или коэффициент детерминации.

Выберем модели для решения задачи классификации:

* KNN;
* Линейная регрессия;
* Дерево решений;
* Случайный лес;
* Градиентный бустинг.

Формируем обучающую и тестовую выборку в соотношении 80/20.

Строим базовое решения, выводим значения метрик:



Используем GridSearch для поиска оптимальных гиперпараметров для каждой модели.

KNeighboursRegressor:

Best hyperparameters: {'n\_neighbors': 20}

DecisionTreeRegressor:

Best hyperparameters: {'max\_depth': 10, 'max\_features': 1.0, 'min\_samples\_leaf': 5}

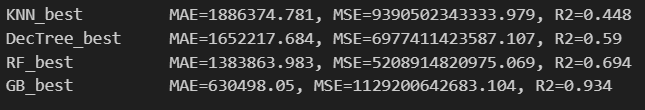
RandomForestRegressor:

Best hyperparameters: {'n\_estimators': 150}

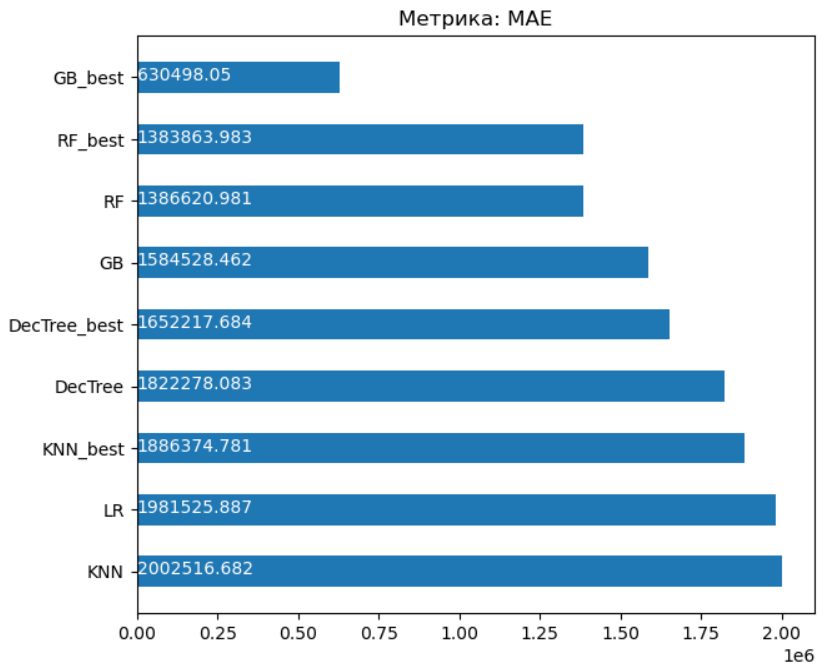
GradientBoostingRegressor:

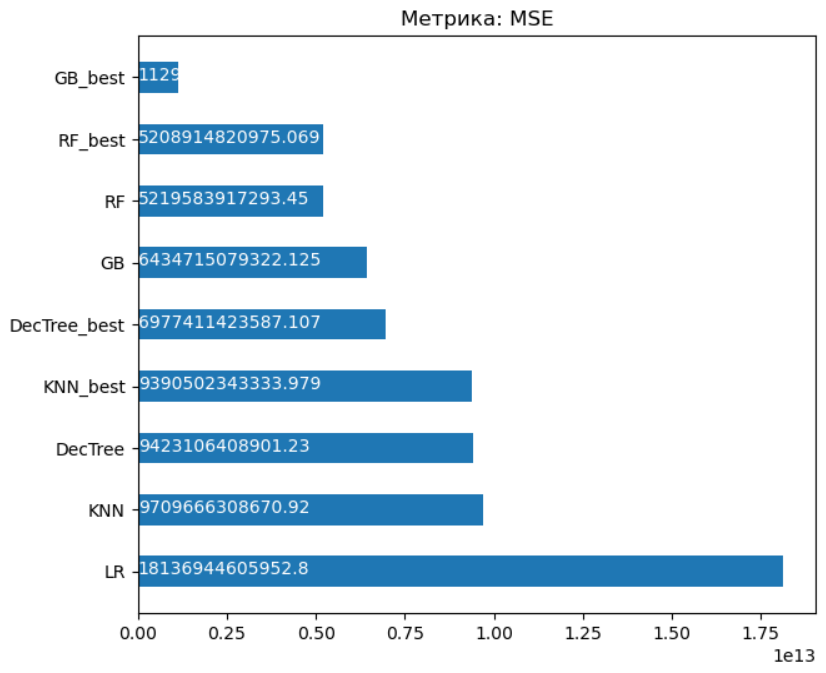
Best hyperparameters: {'learning\_rate': 0.8, 'max\_depth': 5, 'n\_estimators': 150}

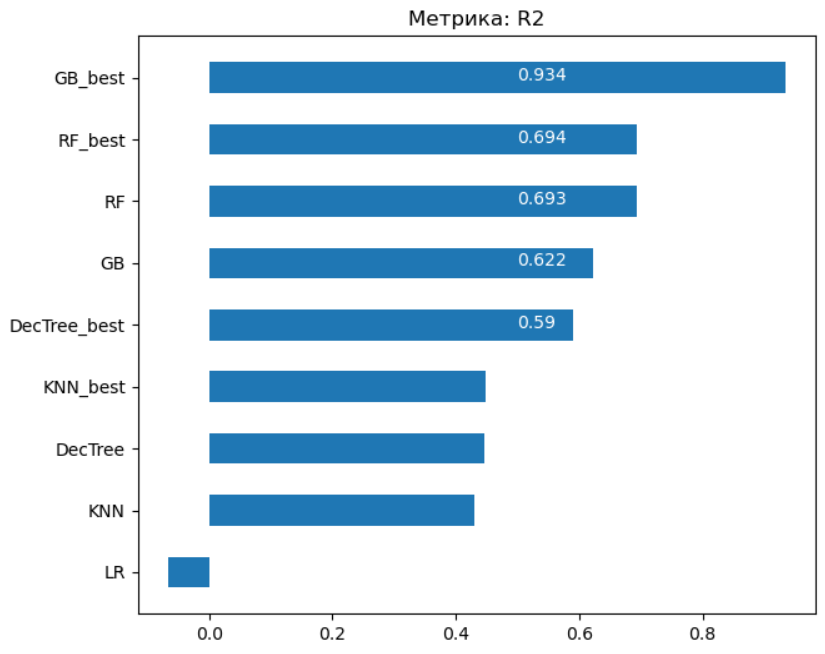
Обучаем и получаем новые значения метрик:



Сравним результаты моделей, построив диаграммы для каждой из метрик:







На основании трех метрик лучшей моделью в решении данной задачи регрессии оказался градиентный бустинг с подобранными гиперпараметрами.

# **Заключение**

Предсказание цен на недвижимость с помощью методов машинного обучения является актуальной и перспективной задачей в области продаж.

В рамках НИРС была разработана эффективная модель, которая может помочь покупателям быстро и с высокой точностью оценить, сколько разумно заплатить за конкретный объект недвижимости.

Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как метод ближайших соседей, линейная регрессия, дерево решений, случайный лес и градиентный бустинг.

В результате исследования было показано, что большинство использованных методов могут достичь хороших результатов, но самым точным оказался градиентный бустинг с подобранными гиперпараметрами.

# **Список использованной литературы**

1. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6
2. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://nbviewer.org/github/ugapanyuk/courses\_current/blob/main/notebooks/ml\_project\_example/project\_classification\_regression.ipynb
3. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses\_current/wiki/COURSE\_TMO\_SPRING\_2024/