МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Дальневосточный федеральный университет»

(ДВФУ)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

(ШКОЛА)

Департамент информационных и компьютерных систем

ОТЧЕТ

по дисциплине «системы искусственного интеллекта»

Выполнил студенты группы Б9122-09.03.03пикд

Зверев Р. И.

Проверил преподаватель

Бочарова В. В.

зачтено/не зачтено

г. Владивосток

2025 г

Оглавление

Цель работы

Целью работы является прогнозирование цен на недвижимость с помощью алгоритма Linear Regression на примере набора данных, который содержит данные о продажах индивидуальных домов в период с мая 2014 года по май 2015 года в округе Кинг, штат Вашингтон, США.

Постановка задачи

В данной работе рассматриваются вопросы поэтапной обработки и анализа набора данных kc_house_data.csv, содержащего информацию о характеристики домов и сопутствующих факторах.

Необходимо реализовать следующие этапы и функции:

- Анализ и удаление выбросов;
- Анализ и восстановление существующих пропусков;
- Стандартизацию данных;
- Обработать существующие признаки с возможностью генерации новых;
- Оценить качество регрессии по коэффициенту детерминации и RMSE (root mean squared error);
- Выделить значимые и незначимые коэффициенты регресии.

Предметной областью является анализ стоимости домов и предсказание цены. Инструментами для реализации задач выбраны библиотеки языка Python: sklearn (scikit-learn), pandas, seaborn, matplotlib, numpy.

Введение

В данной лабораторной работе рассматривается задача прогнозирования цен на жилье с использованием метода линейной регрессии на примере набора данных **kc_house_data.csv**, содержащего информацию о продажах индивидуальных домов в округе Кинг (штат Вашингтон, США) за период с мая 2014 по май 2015 года. Практическая цель — получить устойчивую модель, которая способна объяснить и предсказывать зависимость цены от характеристик домов и внешних факторов.

Работа включает полный цикл предобработки и анализа данных: обнаружение и удаление выбросов, анализ и восстановление пропусков, стандартизация признаков и их преобразование/генерация новых признаков при необходимости. После построения модели будет выполнена ее оценка по коэффициенту детерминации (R^2) и RMSE, а также исследована значимость коэффициентов регрессии для выделения ключевых факторов, влияющих на цену. В качестве инструментов используются библиотеки Python: pandas, numpy, scikit-learn, seaborn и matplotlib.

Анализ предметной области

Описание набора данных kc_house_data.csv.

Для проведения анализа был выбран набор данных kc_house_data.csv, который содержит информацию о стоимости домов и ряде сопутствующих характеристик. Датасет включает 21 тыс. записей (домов) и 21 признак (столбцов).

Основные столбцы, представленные в датасете, и их описание:

- ID: уникальный идентификационный номер проданного дома;
- Date: дата продажи дома;
- Price: цена проданного дома;
- Bedrooms: количество спален;
- Bathrooms: количество ванных комнат (где 0.25 обозначает, что комната с туалетом, 0.5 - комната с туалетом и раковиной);
- Sqft_living: общая площадь дома;
- Sqft_lot: площадь прилегающей территории;
- Floors: количество этажей;
- Waterfront: бинарный атрибут, указывающий на то, есть ли вид на реку или нет;
- View: оценка внешнего вида дома (от 0 до 4);
- Condition: оценка состояния дома (от 0 до 5);
- Grade: оценка качества строительства и дизайна (от 1 до 13);
- Sqft_above: общая площадь наземной части дома;
- Sqft_basement: общая площадь подземной части дома;
- Yr_built: год строительства дома;
- Yr_renovated: год последнего ремонта или реконструкции дома;
- Zipcode: почтовый индекс дома;
- Lat: широта;
- Long: долгота;
- Sqft_living15: средняя общая площадь 15 ближайших домов;

- Sqft_lot15: средняя площадь прилегающей территории 15 ближайших домов.

В рамках данного исследования основной акцент будет сделан на анализе влияния различных представленных факторов на числовую стоимость, выраженную в столбце Price. Понимание взаимосвязей между этой переменной и остальными характеристиками домов позволит выявить ключевые детерминанты стоимости дома.

Предобработка данных

Для построения линейной регрессии было решено провести предобработку признаков: убрать выбросы, пропуски, создать новые и так далее.

Исходный набор данных kc_house_data.csv содержит разнообразную информацию о домах, включая их географические характеристики, количество комнат, площади ближайших домов. Для эффективного анализа такого многомерного датасета и выявления неявных закономерностей требуется последовательный подход к обработке данных.

Сперва нужно проверить существуют ли пустые значения в данных и, если есть, то удалить или заполнить их.

```
df_copy.isna().sum()
```

	0
date	0
price	0
bedrooms	0
bathrooms	0
sqft_living	0
sqft_lot	0
floors	0
waterfront	0
view	0
condition	0
grade	0
sqft_above	0
sqft_basement	0
yr_built	0
yr_renovated	0
zipcode	0
lat	0
long	0
sqft_living15	0
sqft_lot15	0

Рисунок 1. Анализ пропусков в данных

Видно, что пропусков в данных нет, значит можно двигаться дальше.

Рассмотрим распределение домов по году ремонта (столбец «yr_renovated»):

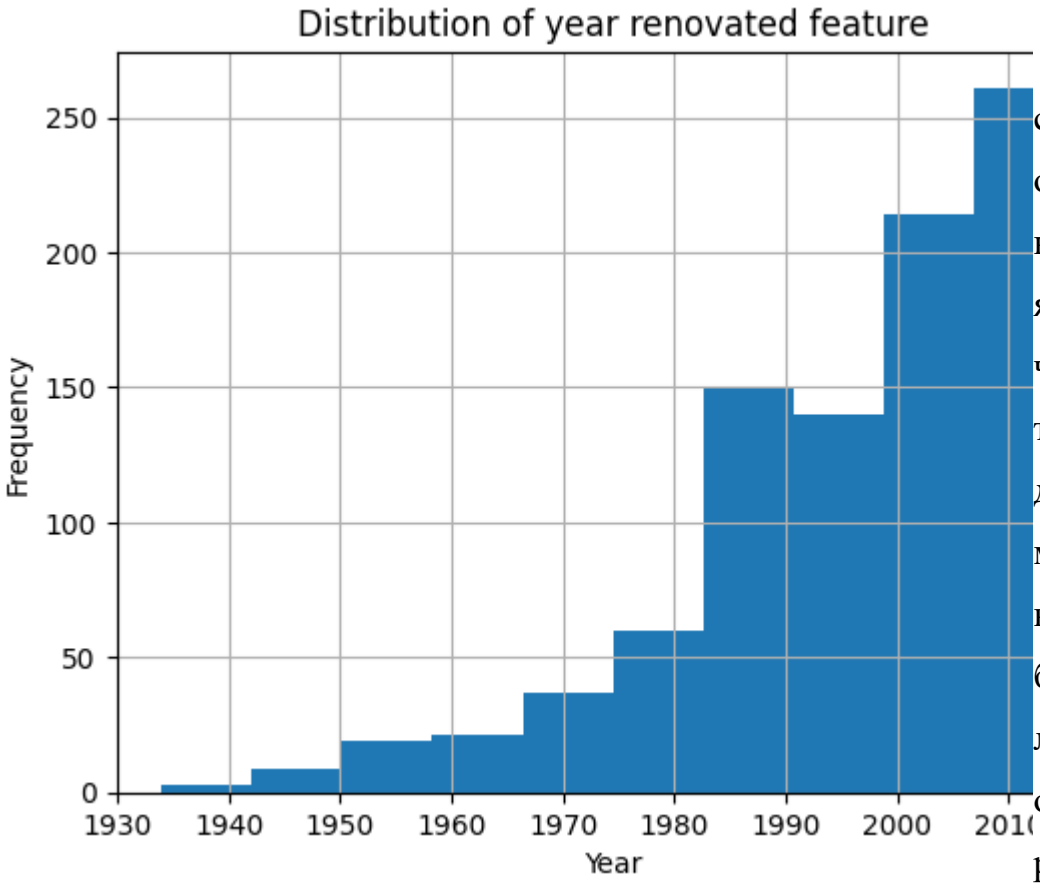


Рисунок 2. Распределение домов по году ремонта

Основная часть домов была отремонтирована после 1990 года, поэтому создаем новый бинарный признак «was_renovated_post90», у которого значение 1, если дом был отремонтирован после 1990 года и значение 0, если дом был отремонтирован раньше. Так же, удаляем признак «yr_renovated».

Рассмотрим признак «yr_built» - год строительства дома и как дома распределяются по нему:

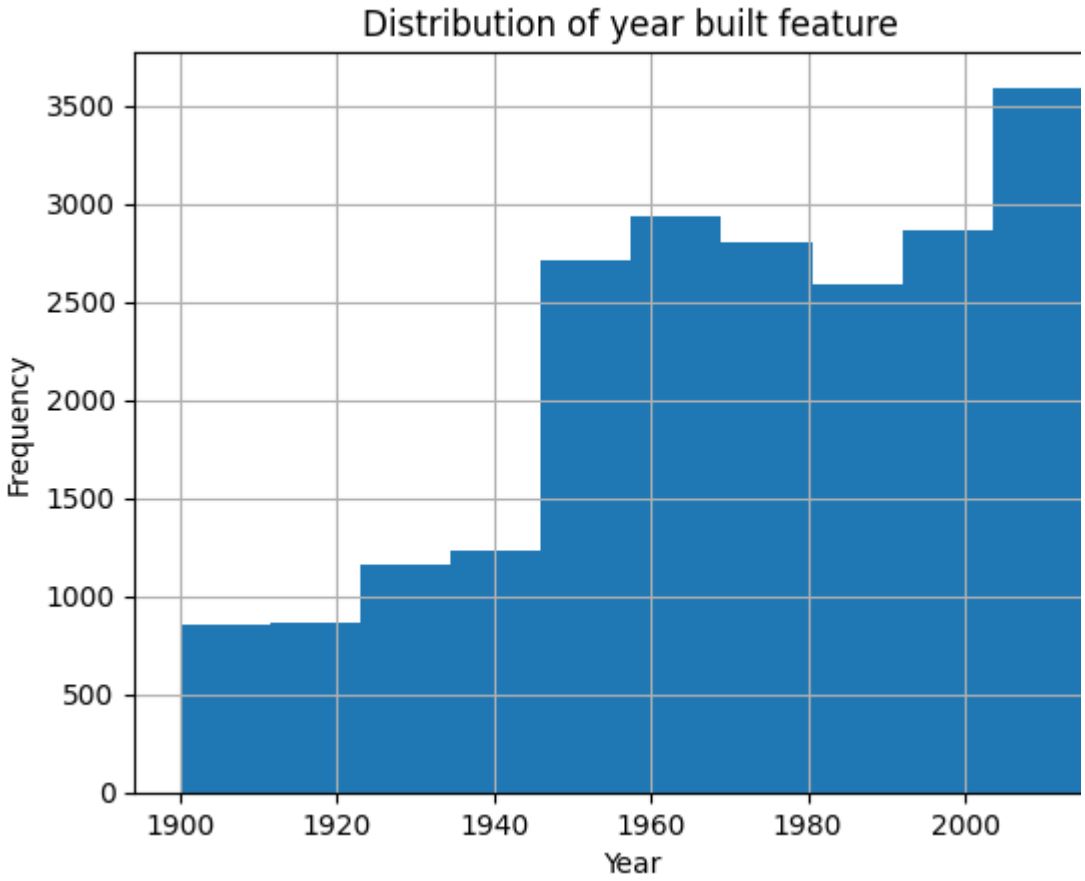


Рисунок 3. Распределение домов по году строительства

Данный признак можно сделать категориальным, разбив его на группы: до 1950 года, с 1950 до 1975 года, с 1975 до 1997 года, с 1997 до 2015 года. Поскольку этот признак категориальный, то применим One Hot Encoding метод для перевода категориального признака в числовой. Данный метод формирует n новых бинарных признаков, где n - число уникальных значений категориального признака, и каждая категория преобразуется в отдельный бинарный признак (0/1), показывающий принадлежность объекта к этой категории.

В исходных данных в столбцах «sqft_basement» и «view» преобладает значение 0. Можно создать на их основе новые бинарные признаки «has_basement» и «viewed», где 1 - любое значение, не равное 0 в исходном признаке.

Также, округлим до целого числа значения в столбце «bathrooms».

Рассмотрим численные признаки через гистограммы:

Р в

Виден некоторых разброс в данных. Один из примеров в столбце «bedrooms», где значение оси X уходит до 30, что показывает, что есть записи, где написано 30 спален. Проверим гипотезу:

```
df_copy['bedrooms'].value_counts()
```

bedrooms	count
3	9824
4	6882
2	2760
5	1601
6	272
1	199
7	38
0	13
8	13
9	6
10	3
11	1
33	1

Рисунок 5. Уникальные значения столбца и кол-во записей для каждого значения

Сделаем предположение, что опечатка и у записи 15870 поменяем количество спален с 33 на 3.

Также, уберем записи, которые отличаются в 3 раза от стандартного значения в колонках «bedrooms», «bathrooms», «sqft_living», «sqft_lot», «sqft_above», «lat», «long», «sqft_living15», «sqft_lot15», как выбросы и в итоге вместо 21613 записей осталось 20058 записей.

Рассмотрим распределение цен по квантилям:

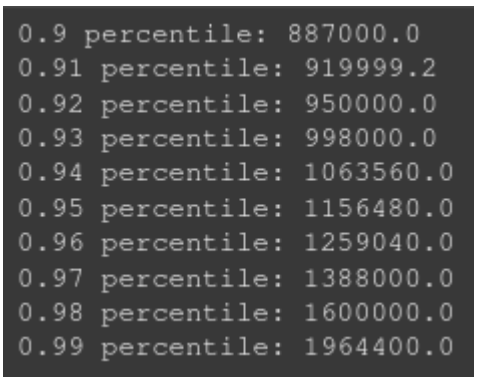


Рисунок 6. Распределение цен по квантилям

Удалим дома стоимостью чуть выше, чем у последнего квантиля, так как это какие-то уникальные дома и можно посчитать за выброс.

Также, вместо даты в исходном виде сформируем признак «sale_month», который показывает месяц продажи дома.

Рассмотрим матрицу корреляций признаков, чтобы определить какие признаки сильно коррелируют и удалить их:

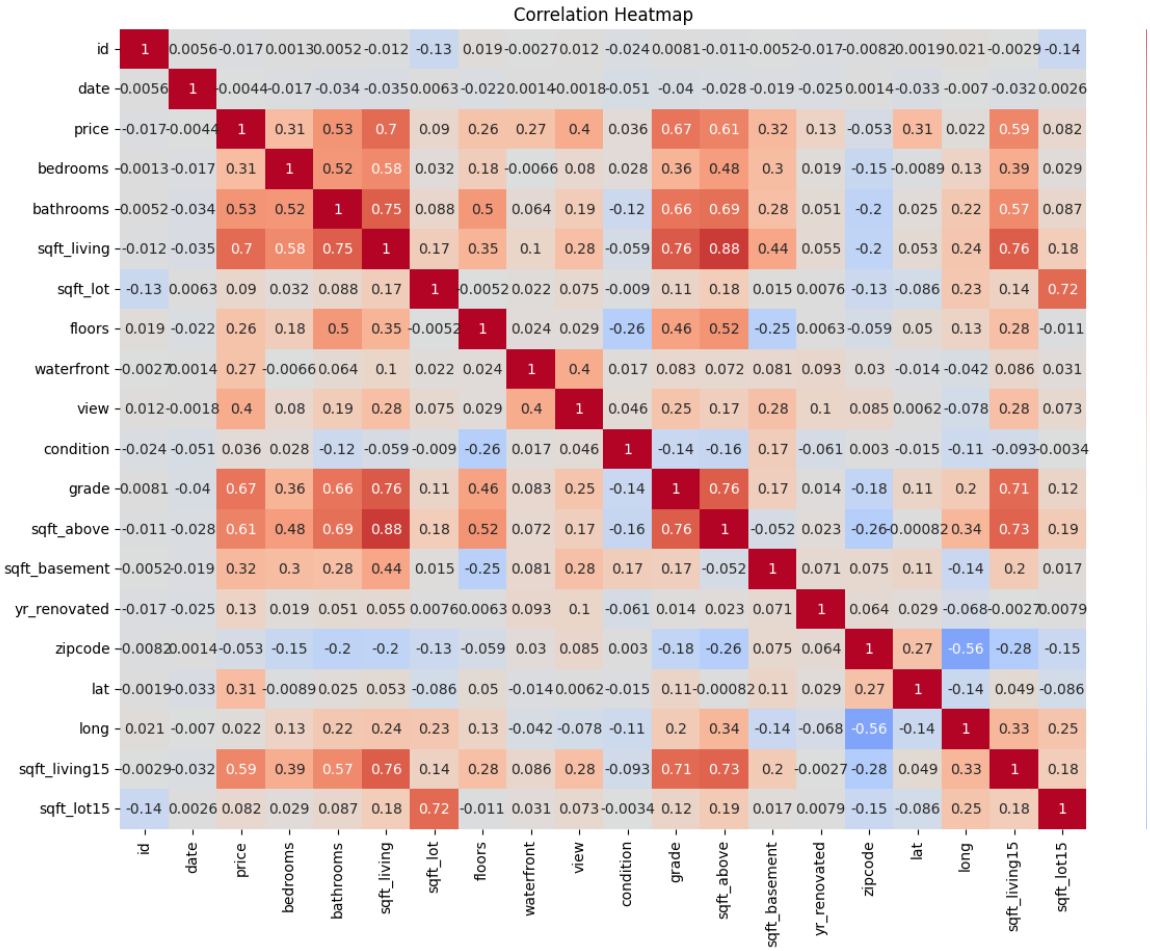
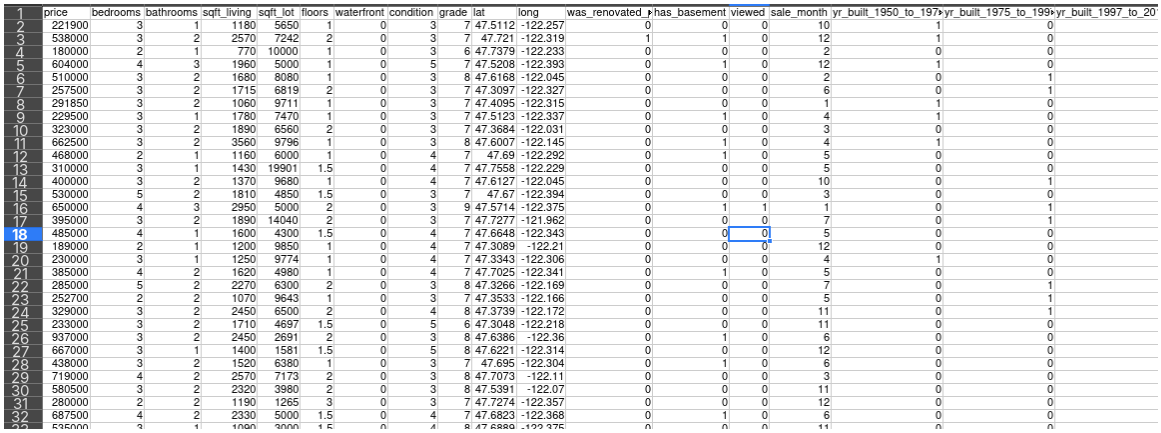


Рисунок 7. Матрица корреляций признаков

Видно, что признаки «sqft_above», «sqft_living15», «sqft_lot15» сильно кореелируют и лучше их удалить.

Наконец, уберем ненужные данные для обучения: «id», «date», «zipcode».



	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	condition	grade	lat	long	was_renovated	has_basement	viewed	sale_month	yr_built_1950_to_197	yr_built_1975_to_199	yr_built_1997_to_20
2	221900	3	1	1180	5650	1	0	3	7	47.5112	-122.257	0	0	0	10	1	0	
3	538000	3	2	2570	7242	2	0	3	7	47.721	-122.319	1	1	0	12	1	0	
4	180000	2	1	770	10000	1	0	3	6	47.7379	-122.233	0	0	0	2	0	0	
5	604000	4	3	1960	5000	1	0	5	7	47.5208	-122.393	0	1	0	12	1	0	
6	510000	3	2	1680	8080	1	0	3	8	47.6168	-122.045	0	0	0	2	0	1	
7	257500	3	2	1715	6819	2	0	3	7	47.3097	-122.327	0	0	0	6	0	1	
8	291850	3	2	1060	9711	1	0	3	7	47.4095	-122.315	0	0	0	1	1	0	
9	229500	3	1	1780	7470	1	0	3	7	47.5123	-122.337	0	1	0	4	1	0	
10	323000	3	2	1890	6560	2	0	3	7	47.3684	-122.031	0	0	0	3	0	0	
11	662500	3	2	3560	9796	1	0	3	8	47.6007	-122.145	0	1	0	4	1	0	
12	468000	2	1	1160	6000	1	0	4	7	47.69	-122.292	0	1	0	5	0	0	
13	310000	3	1	1430	19901	1.5	0	4	7	47.7558	-122.229	0	0	0	5	0	0	
14	400000	3	2	1370	9680	1	0	4	7	47.6127	-122.045	0	0	0	10	0	1	
15	530000	5	2	1810	4850	1.5	0	3	7	47.67	-122.394	0	0	0	3	0	0	
16	650000	4	3	2950	5000	2	0	3	9	47.5714	-122.375	0	1	1	1	0	1	
17	395000	3	2	1890	14040	2	0	3	7	47.7277	-121.962	0	0	0	7	0	1	
18	485000	4	1	1600	4300	1.5	0	4	7	47.6648	-122.343	0	0	0	5	0	0	
19	189000	2	1	1200	9850	1	0	4	7	47.3089	-122.21	0	0	0	12	0	0	
20	230000	3	1	1250	9774	1	0	4	7	47.3343	-122.306	0	0	0	4	1	0	
21	385000	4	2	1620	4980	1	0	4	7	47.7025	-122.341	0	1	0	5	0	0	
22	285000	5	2	2270	6300	2	0	3	8	47.3266	-122.169	0	0	0	7	0	1	
23	252700	2	2	1070	9643	1	0	3	7	47.3533	-122.166	0	0	0	5	0	1	
24	329000	3	2	2450	6500	2	0	4	8	47.3739	-122.172	0	0	0	11	0	1	
25	233000	3	2	1710	4697	1.5	0	5	6	47.3048	-122.218	0	0	0	11	0	0	
26	937000	3	2	2450	2691	2	0	3	8	47.6386	-122.36	0	1	0	6	0	0	
27	667000	3	1	1400	1581	1.5	0	5	8	47.6221	-122.314	0	0	0	12	0	0	
28	438000	3	2	1520	6380	1	0	3	7	47.695	-122.304	0	1	0	6	0	0	
29	719000	4	2	2570	7173	2	0	3	8	47.7073	-122.11	0	0	0	3	0	0	
30	580500	3	2	2320	3980	2	0	3	8	47.5391	-122.07	0	0	0	11	0	0	
31	280000	2	2	1190	1265	3	0	3	7	47.7274	-122.357	0	0	0	12	0	0	
32	687500	4	2	2330	5000	1.5	0	4	7	47.6823	-122.368	0	1	0	6	0	0	

Рисунок 8. Подготовленный датасет

После подготовки данных датасет (19997 строк и 19 столбцов) выглядит так:

Обучение линейной регрессии

Рассмотрим 3 вида линейной регрессии: обычная (Linear Regression), гребневая с параметром α=3 (Ridge) и лассо с параметром α=120 (Lasso), сравним их на подготовленных данных и выделим коэффициенты признаков каждой регресии.

Датасет поделим на 2 части: тренировочную и тестовую в соотношении 80%:20%. Также, применим StandardScaler для стандартизации признаков.

Оценки качества регрессий и их коэффициенты признаков приведены в таблицах:

Таблица 1. Качества регрессий по R^2 и RMSE

	R^2	**RMSE**
Линейная регрессия	146542.995	0.68572962
Гребневая регрессия	146542.617	0.68573125
Лассо	146543.499	0.68572747

Таблица 2. Коэффициенты регрессий

Коэффициенты	**Линейная**	**Гребневая**	**Лассо**

	регрессия	**регрессия**	
bedrooms	-13595.8364205	-13577.0381500	-13187.3849808
bathrooms	15297.6431526	15303.3195651	15042.7784461
sqft_living	92248.0062107	92219.5011677	91973.6857992
sqft_lot	-4272.3421886	-4263.3558376	-4014.4149891
floors	7603.1950043	7608.5996268	7304.9944168
waterfront	26413.1897285	26409.2805075	26321.5313515
condition	22874.1924342	22872.4682724	22824.0025535
grade	108840.5535625	108816.9562686	108763.6061606
lat	76056.2627262	76047.9485564	76007.5300256
long	1275.0042280	1271.9736040	960.9669020
was_renovated_post90	12570.0171011	12572.1158276	12552.0033135
has_basement	-862.2968377	-854.0306843	-698.0702015
viewed	29079.9659699	29083.8267262	29048.6471429
sale_month	-8935.7638456	-8933.4699046	-8804.9986619
yr_built_1950_to_1975	2145.5715436	2140.2137699	0.0
yr_built_1975_to_1997	-23019.1724958	-23012.3811329	-24635.7864209
yr_built_1997_to_2015	-21221.2897152	-21209.6981002	-22569.4624442
yr_built_pre1950	41327.8961487	41315.3473755	39256.2819827

Как можно заметить из таблицы, для гребневой регрессии, фактически, неиспользуемым параметром оказался признак «yr_built_1950_to_1975». Все модели выделили в качестве не сильно значимых признаков «yr_built_1997_to_2015» и «beadrooms», а самыми вероятно значимыми являются «sqft_living», «grade», «lat», «viewed», «bathrooms», «condition», «waterfront».

Заключение

В ходе выполнения данной работы была успешно обучена модель линейной регрессии разных видов (гребневая и лассо). Целью работы являлось предсказание стоимости дома на основе набора признаков, что было достигнуто через последовательное решение ряда поставленных задач.

На начальном этапе была проведена характеристика исходного набора данных kc_house_data.csv. Были изучены структура датасета, что позволило сформировать первичное представление об исследуемых объектах.

Для наглядного представления данных и выявления начальных закономерностей были выбраны и, предположительно, применены методы визуализации, такие как гистограммы и графики.

Ключевым этапом стала предварительная обработка данных. В соответствии с поставленной задачей, было продемонстрировано конструирование производных признаков путем создания новых информативных полей на основе существующих. Также была осуществлена фильтрация данных для удаления нерелевантных записей, что повысило качество данных для последующего анализа.

Для предсказания стоимости дома был выбран и реализован метод линейной регрессии. Построение регрессионной модели позволило количественно оценить влияние таких факторов, как состояние дома, имеется ли вид на море и другие на стоимость дома.

Результаты регрессионного анализа и выявленные зависимости были представлены в виде таблиц, что обеспечило наглядную интерпретацию полученных выводов.

Таким образом, все поставленные в рамках работы задачи были выполнены. Полученные выводы могут быть использованы для дальнейших исследований в данной предметной области.

Список литературы

1. GitHub: исходный код лабораторной работы. – URL: Лабораторная работа №3.2 (дата обращения: [09.10.2025]). – Текст: электронный.