Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

*Факультет программной инженерии и компьютерной техники*

Системы искусственного интеллекта

Лабораторная работа №3

Линейная регрессия

Группа: P3324

Выполнил:

Круглов Егор Ильич

Преподаватель:

Королёва Юлия Александровна

Санкт-Петербург

2024г.

**Оглавление**

[Введение 3](#_Toc178536176)

[Описание метода 4](#_Toc178536177)

[Ход работы 5](#_Toc178536178)

[Результаты моделей 14](#_Toc178536179)

[Анализ и сравнение результатов: 15](#_Toc178536180)

# Введение

В данной лабораторной работе целью является реализация метода линейной регрессии для анализа данных о жилье в Калифорнии. Лабораторная работа включает подготовку данных, их предобработку, разделение на обучающие и тестовые выборки, а также обучение нескольких моделей линейной регрессии с различными наборами признаков.

Задание:

1. Получите и визуализируйте (графически) статистику по датасету (включая количество, среднее значение, стандартное отклонение, минимум, максимум и различные квантили).
2. Проведите предварительную обработку данных, включая обработку отсутствующих значений, кодирование категориальных признаков и нормировка.
3. Разделите данные на обучающий и тестовый наборы данных.
4. Реализуйте линейную регрессию с использованием метода наименьших квадратов без использования сторонних библиотек, кроме NumPy и Pandas (для использования коэффициентов использовать библиотеки тоже нельзя). Использовать минимизацию суммы квадратов разностей между фактическими и предсказанными значениями для нахождения оптимальных коэффициентов.
5. Постройте три модели с различными наборами признаков. Для каждой модели проведите оценку производительности, используя метрику коэффициент детерминации, чтобы измерить, насколько хорошо модель соответствует данным.
6. Сравните результаты трех моделей и сделайте выводы о том, какие признаки работают лучше всего для каждой модели.
7. Бонусное задание: ввести синтетический признак при построении модели

# Описание метода

Метод линейной регрессии используется для моделирования зависимости между одной зависимой переменной (в данной работе — индекс успеваемости студентов) и одной или несколькими независимыми переменными (время подготовки, предыдущие оценки и т.д.). Принцип работы метода основан на нахождении такой прямой линии (гиперплоскости в случае многомерных данных), которая минимизирует сумму квадратов отклонений предсказанных значений от фактических. Для этого используется метод наименьших квадратов.

Линейная регрессия представляет собой метод нахождения линейной зависимости между зависимой переменной y и набором независимых переменных X. Модель можно выразить как:

где y — предсказанное значение, θ0 — свободный член, а θ1, …, θn— коэффициенты, которые необходимо оптимизировать.

Алгоритм градиентного спуска применяется для минимизации функции стоимости J(θ) которая является среднеквадратичной ошибкой:

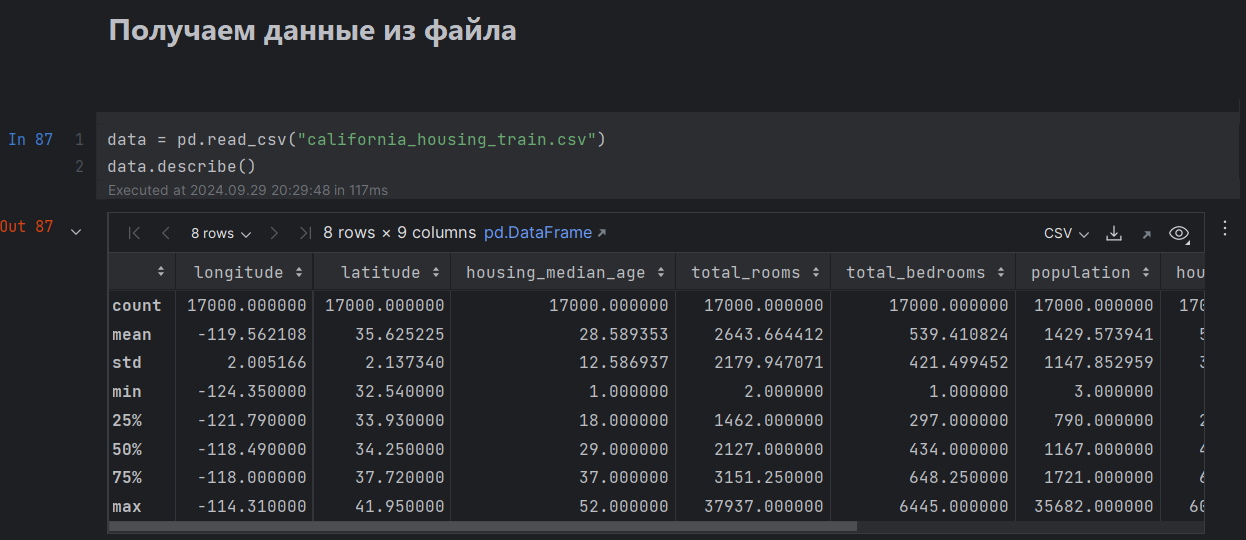
где m — количество обучающих примеров, y^i — предсказанное значение, а yi — фактическое значение.

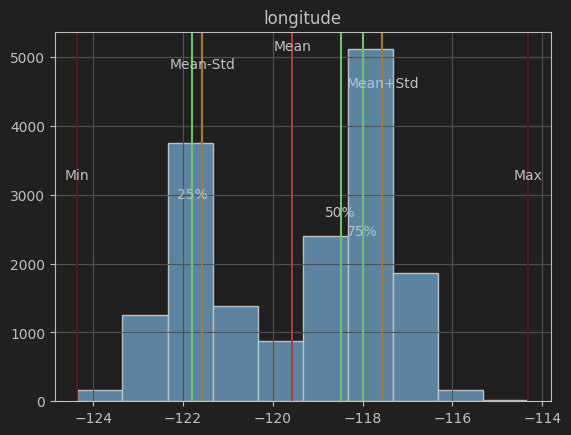
Процесс обновления коэффициентов на каждом шаге выглядит следующим образом:

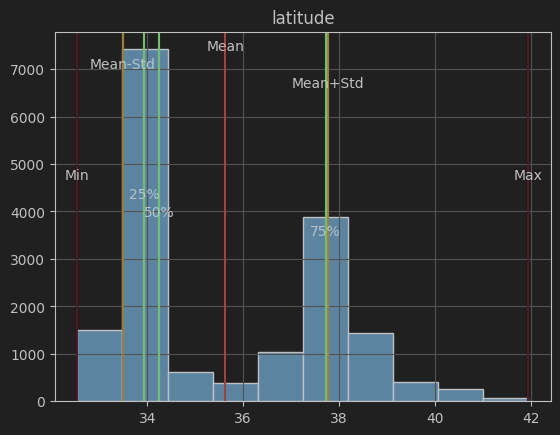
где α — шаг обучения (learning rate).

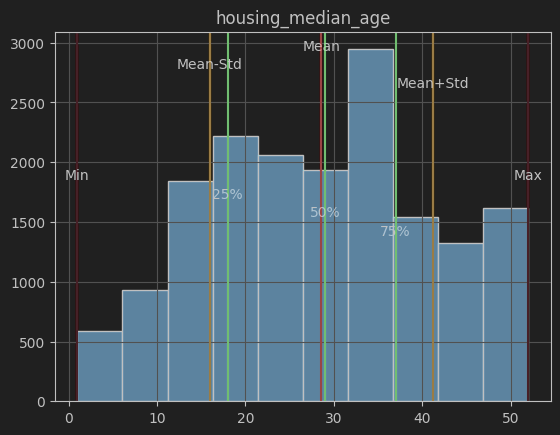
# Ход работы

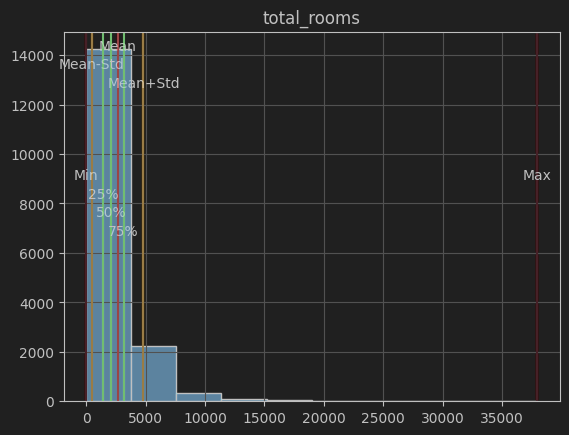
1. Считывание и визуализация данных, расчет основных статистических показателей (среднее, стандартное отклонение и т.д.).

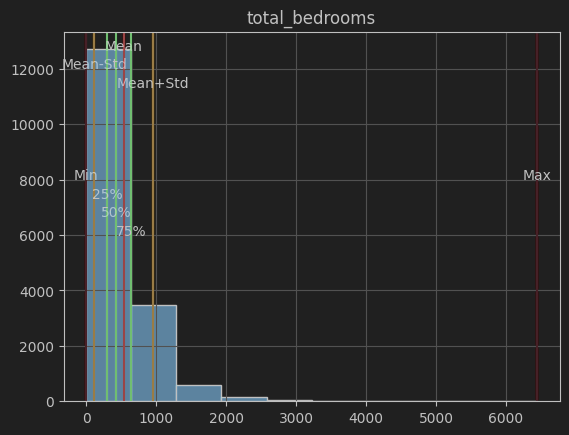


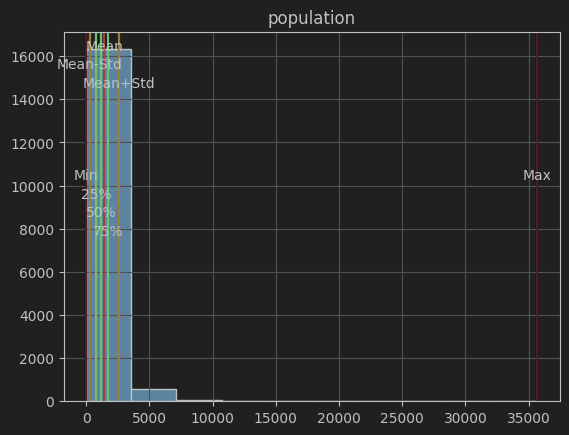


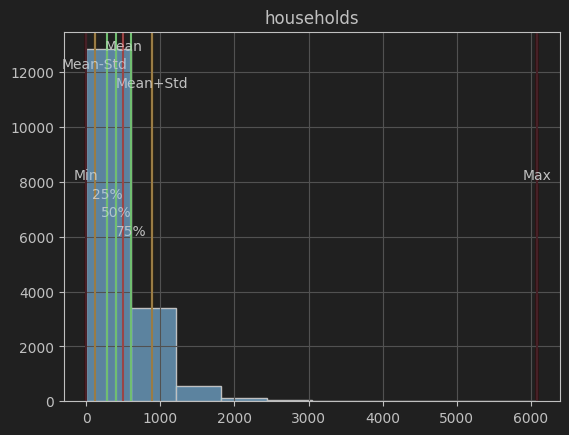


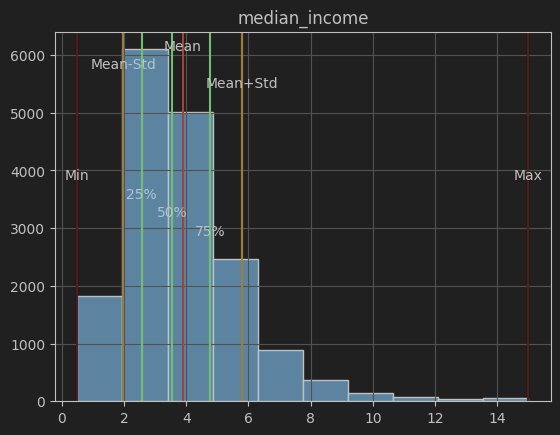


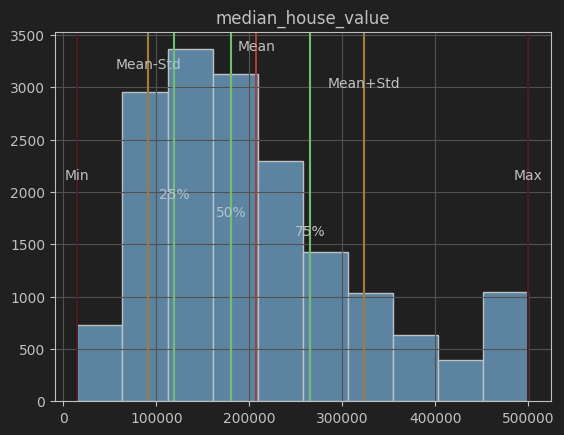




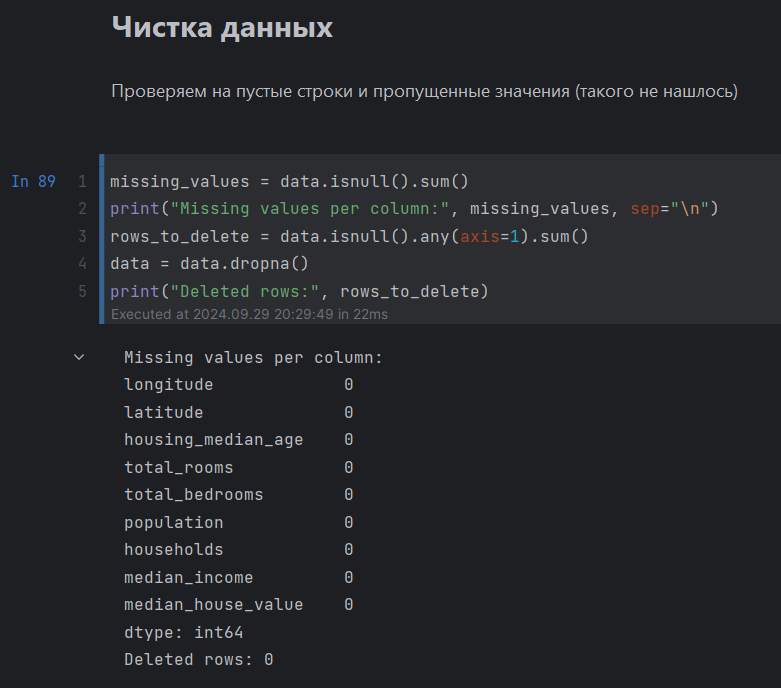


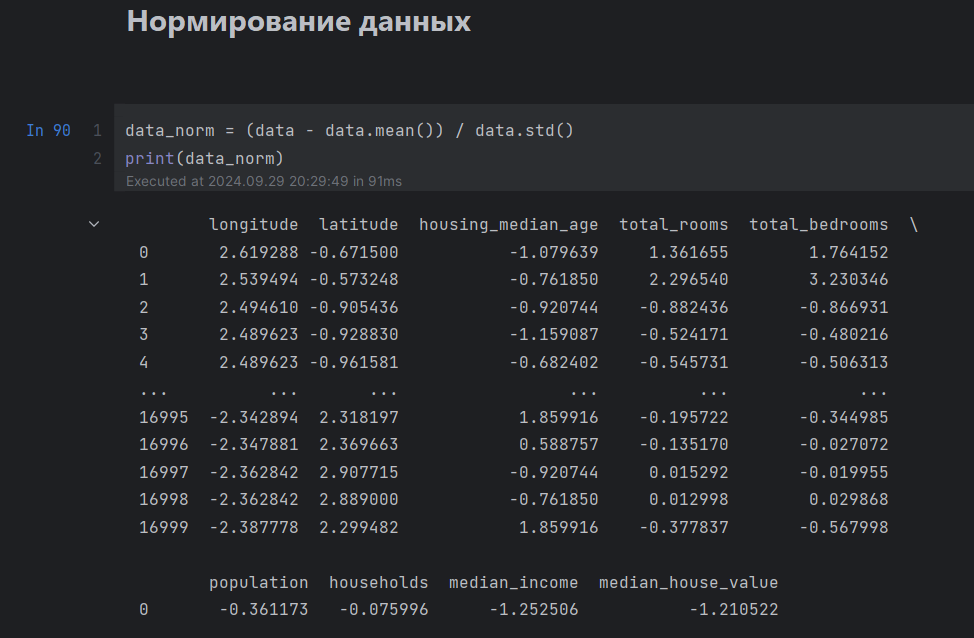




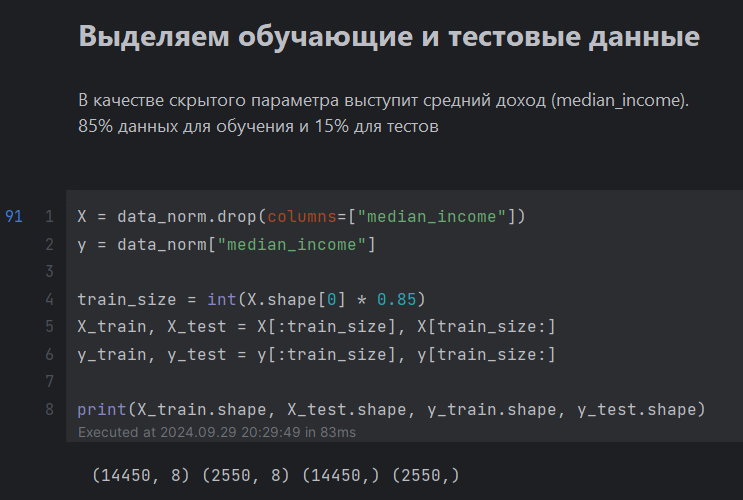


1. Предобработка данных: удаление строк с пропущенными значениями, кодирование категориальных признаков и нормировка. Нормирование данных.

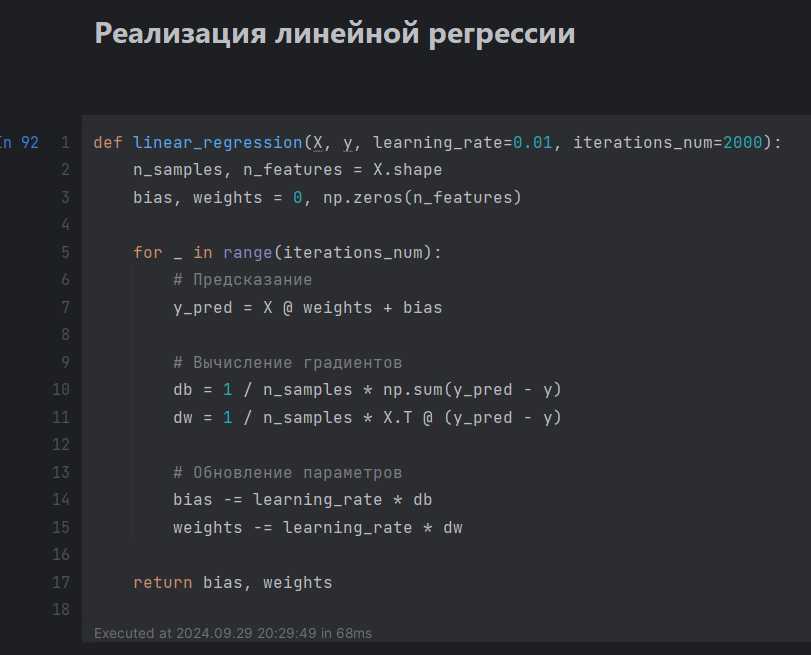




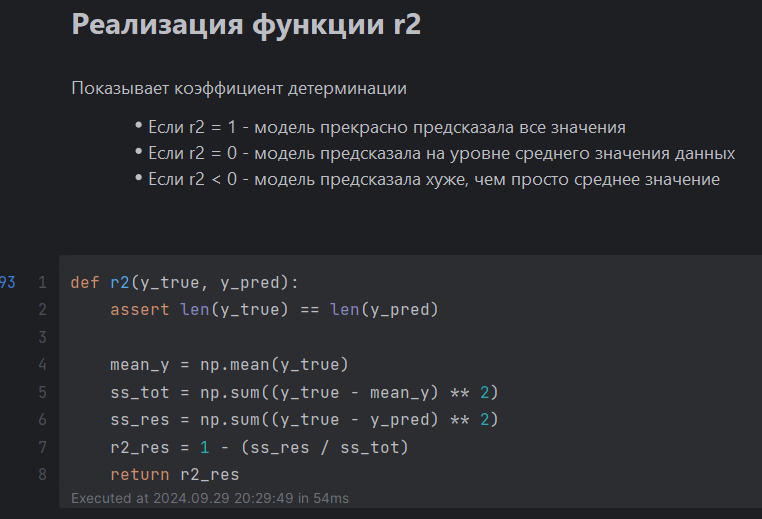
1. Выбор скрытого параметра. Разделение данных на обучающую и тестовую выборки в соотношении 85:15.



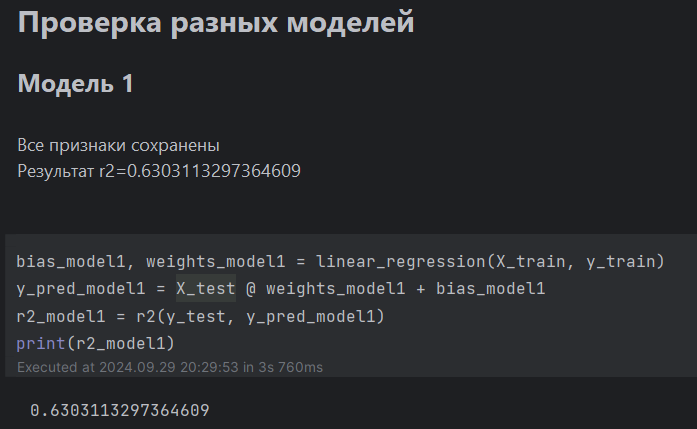
1. Реализация линейной регрессии.

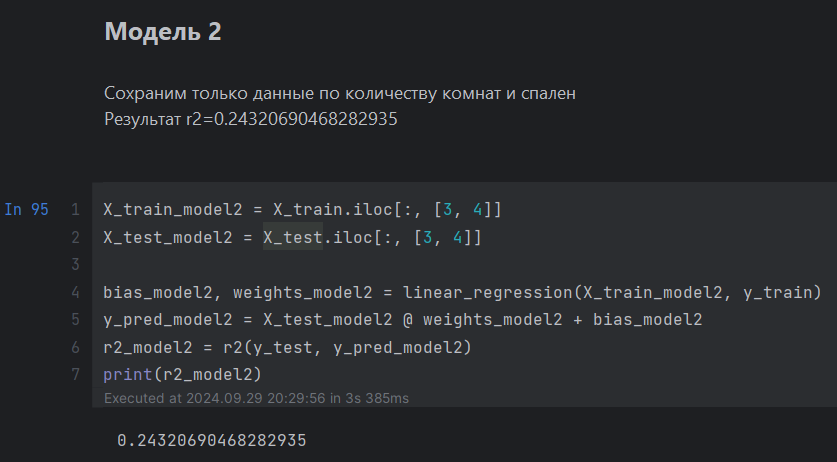


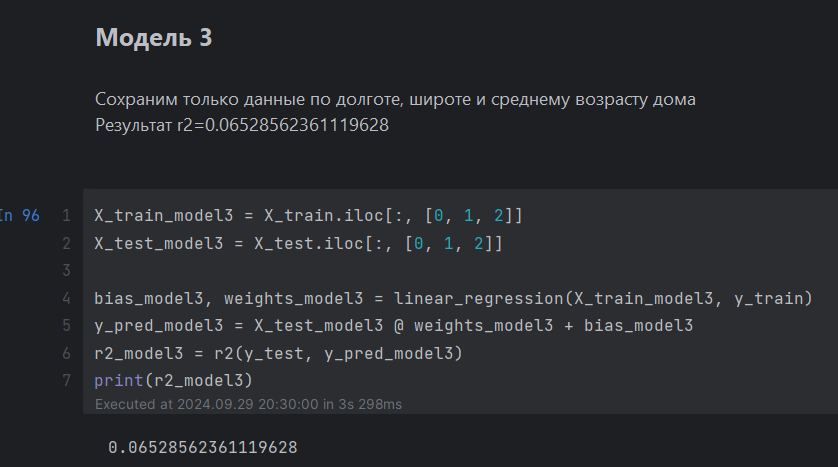
1. Реализация функции для оценки качества обучения моделей (коэффициент детерминации)



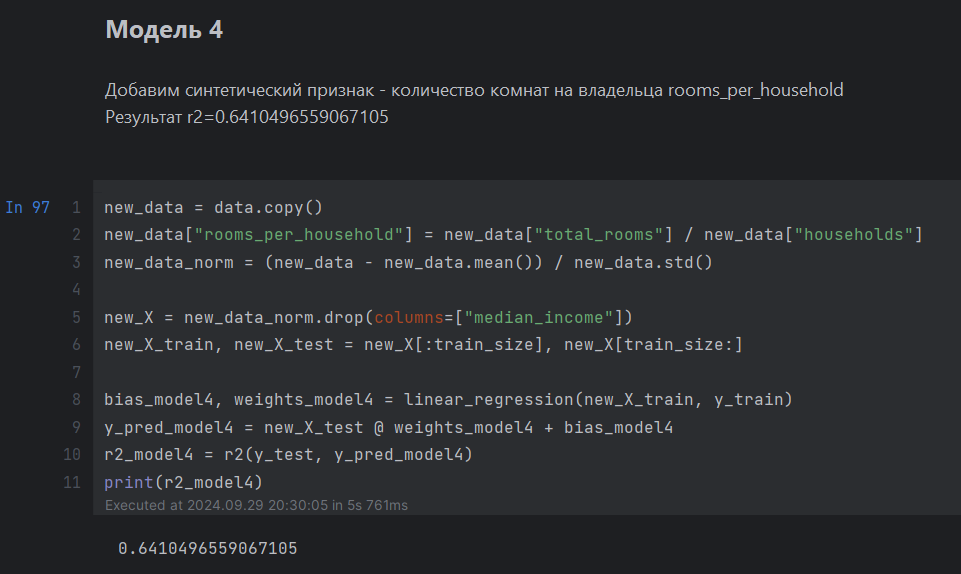
1. Обучение и проверка трех моделей с различными наборами признаков.







1. Добавление синтетического признака и проверка результатов.



# Результаты моделей

Были построены 4 модели со следующими признаками:

1. Модель 1: все признаки
2. Модель 2: только признаки «количество комнат» и «количество спален»
3. Модель 3: только признаки «долгота», «широта» и «средний возраст дома»
4. Модель 4: все признаки с добавлением синтетического признака «кол-во комнат на владельца», который равен отношению кол-ва комнат к числу владельцев.

Все модели тренировались по функции линейной регрессии с параметрами и 2000 итераций.

Для оценки моделей используется функция для вычисления коэффициента детерминации:

где RSS(Residual Sum of Squares) — сумма квадратов остатков, или невязок между фактическими и предсказанными значениями, TSS (Total Sum of Squares) — общая сумма квадратов отклонений фактических значений от их среднего.

*Модель 1:*

Коэффициент детерминации 0.6303113297364609

*Модель 2:*

Коэффициент детерминации 0.24320690468282935

*Модель 3:*

Коэффициент детерминации 0.06528562361119628

*Модель 4:*

Коэффициент детерминации 0.6410496559067105

# Анализ и сравнение результатов:

Модель 1 показывает неплохие результаты (, так как использует все признаки. Такой коэффициент детерминации говорит о том, что данная модель может неплохо предугадывать заработок.

Модель 2 демонстрирует результат похуже (). Все же для предсказания заработка важно не только количество комнат и спален.

Модель 3 с набором других признаков показывает результат еще хуже (), чем Модель 2, несмотря на большее число признаков. Это говорит о том, что параметры координат дома и его среднего возраста хуже коррелируют с средним заработком, чем данные о количестве комнат и спален. Однако, , значит эти данные все же что-то говорят о заработке.

Модель 4, включает все признаки + 1 синтетический. За счет этого она имеет наилучший коэффициент детерминации (). Это не сильно больше, чем Модель 1. Однако данный опыт свидетельствует о том, что признаки о количестве людей и количестве комнат играют большую роль чем в среднем влияют все признаки.