Липецкий государственный технический университет

Институт компьютерных наук
Кафедра прикладной математики и системного анализа

Отчет по лабораторной работе № 2
по дисциплине «Интеллектуальные методы анализа данных»
на тему «Переработка данных»

Студент _____ Габбасов.Д.В
подпись, дата фамилия, инициалы

Группа ПМ-23-1

Руководитель
к.т.н., доцент _____ Сысоев А.С.
учёная степень, учёное звание подпись, дата фамилия, инициалы

Липецк 2025 г.

Постановка задачи

Целью работы является построение и сравнение регрессионных моделей для предсказания целевой переменной на основе реального датасета.

В лабораторной работе требуется:

- загрузить и описать датасет, выбрать переменные;
- выполнить предварительный анализ данных;
- проверить предпосылки множественной линейной регрессии;
- обучить несколько моделей (MLR, Ridge, Lasso, Random Forest и др.);
- визуализировать результаты:
  - матрица корреляций;
  - графики остатков;
  - сравнение предсказанных и реальных значений;
  - важность переменных;
  - сравнение коэффициентов;
- вычислить метрики (R^2, adj-R^2, RMSE, MAE) и время обучения;
- провести сравнительный анализ моделей;
- сформулировать выводы.

Оглавление

1. Введение

Регрессионный анализ является одним из базовых инструментов статистического моделирования, позволяя прогнозировать значения целевой переменной на основе набора факторов. На практике регрессионные модели используются в экономике, финансовой аналитике, демографии, медицине и других сферах, где необходимо выявлять зависимости и проводить прогнозирование.

В рамках данной работы исследуется влияние социально-экономических и географических факторов на стоимость жилья. В качестве объекта анализа используется датасет California Housing, содержащий информацию о медианной стоимости домов в различных регионах Калифорнии и дополнительные признаки, характеризующие население, инфраструктуру и географическое положение.

Цель работы — построение и сравнение нескольких регрессионных моделей, включая:

- множественную линейную регрессию;
- регрессию с взаимодействиями между факторами;
- регуляризованные модели Ridge и Lasso.

Для оценки качества моделей используются метрики R^2, скорректированное R^2, RMSE и MAE. Помимо построения моделей, проводится статистическая проверка предпосылок МНК: анализ остатков, оценка мультиколлинеарности и значимости коэффициентов.

Работа имеет как исследовательский, так и практический характер — полученные результаты могут быть использованы для прогнозирования стоимости жилья, выявления наиболее влияющих факторов и улучшения интерпретируемости моделей.

2. Описание данных и методов

2.1. Датасет California Housing

В исследовании использован датасет California Housing, содержащий информацию о стоимости домов и социально-географические факторы по регионам Калифорнии. Данные включают демографические, экономические и пространственные признаки.

Целевая переменная:

— Median_House_Value — медианная стоимость жилья

Предикторы:

— Median_Income — медианный доход населения
— Median_Age — медианный возраст зданий
— Tot_Rooms, Tot_Bedrooms — количество комнат и спален
— Population — численность населения
— Households — количество домохозяйств
— Latitude, Longitude — координаты
— Distance_to_coast, Distance_to_LA, Distance_to_SanDiego, Distance_to_SanJose, Distance_to_SanFrancisco — расстояния до крупных городов

2.2. Предобработка

Данные не содержали пропусков, масштабирование выполнено с помощью StandardScaler. Набор разделён на тренировочную и тестовую выборки в пропорции 70/30.

3. Множественная линейная регрессия

Модель обучена на всех предикторах. На Рис. 1 приведена корреляционная матрица.

3.1. Результаты модели

Метрика	Значение
R^2	0.677
Adj. R^2	0.677
RMSE	62 067
MAE	45 213

Таблица 1. Качество модели линейной регрессии

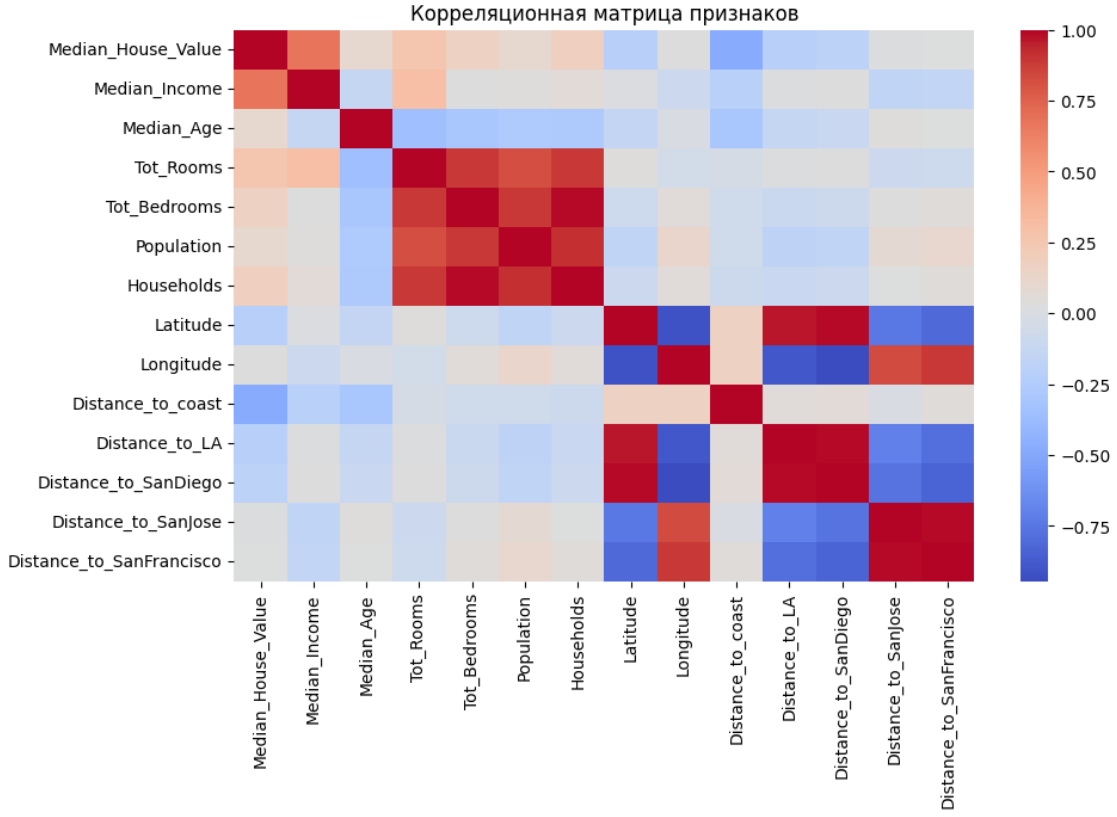

Рисунок 1 – Корреляционная матрица признаков

3.2. Проверка предпосылок МНК

На Рис. 2 показано распределение остатков. Оно близко к нормальному, что подтверждено тестом Шапиро.

3.3. Проверка мультиколлинеарности (VIF)

Для оценки мультиколлинеарности рассчитаны показатели VIF (Variance Inflation Factor) для каждого признака. Высокие значения VIF указывают, что переменная сильно коррелирует с одной или несколькими другими, что приводит к неустойчивости оценок коэффициентов линейной регрессии.

Интерпретация:

— Признаки с VIF > 10 считаются сильно мультиколлинеарными.

— В данной модели практически все географические признаки (Latitude, Longitude, расстояния до городов) имеют сверхвысокие значения VIF (более 300), что говорит о сильной линейной зависимости между ними.

— Признаки Tot_Rooms, Tot_Bedrooms и Households также обладают высокой мультиколлинеарностью, что объясняется очевидной логической связью: количество спален и домохозяйств зависит от количества комнат и населения.

Рисунок 2 – Распределение остатков линейной модели

Вывод: Множественная линейная регрессия страдает от мультиколлинеарности, что делает оценки коэффициентов нестабильными. Для решения проблемы применены регуляризованные модели (Ridge и Lasso), а также модель с взаимодействиями.

Lasso Regression с оптимальным параметром $\alpha = 0.001$ не исключила ни одной переменной. Это означает, что все признаки оказывают значимое влияние на предсказываемую стоимость жилья. Большинство коэффициентов по знаку и величине имеют экономический смысл: доход населения, число комнат и близость к побережью увеличивают цену, а удалённость и рост населения снижают её.

4. Регрессия с взаимодействиями

Анализ корреляций показал, что переменные Tot_Rooms и Tot_Bedrooms наиболее связаны. Добавлено взаимодействие Rooms × Bedrooms.

4.1. Сравнение моделей

Сравним качество четырёх моделей: базовой линейной регрессии, модели с взаимодействием признаков, а также Ridge и Lasso регрессий.

Анализ Таблицы 3 показывает, что:

Признак	R^2_i	VIF
Median_Income	0.491883	1.96
Median_Age	0.348229	1.53
Tot_Rooms	0.892092	9.26
Tot_Bedrooms	0.980742	51.92
Population	0.852144	6.76
Households	0.982993	58.79
Latitude	0.999097	1107.05
Longitude	0.995832	239.91
Distance_to_coast	0.864120	7.35
Distance_to_LA	0.998979	979.84
Distance_to_SanDiego	0.999828	5809.24
Distance_to_SanJose	0.994764	190.99
Distance_to_SanFrancisco	0.996244	266.26

Таблица 2. Показатели R^2 и VIF для признаков

Модель	R^2	RMSE
Linear	0.670478	61876.38
Interaction (Rooms×Bedrooms)	0.776585	50949.40
Ridge	0.670282	61894.75
Lasso	0.670268	61896.07

Таблица 3. Сравнение качества моделей

— Добавление взаимодействия значительно улучшило точность: R^2 вырос с 0.67 до 0.78, а RMSE уменьшился почти на 18%.
— Ridge и Lasso не дали прироста в качестве, но Lasso обнулила некоторые коэффициенты, выполняя отбор признаков.
— Таким образом, лучшей моделью стала модель с взаимодействием признаков Rooms×Bedrooms.

5. Регуляризация: Ridge и Lasso

Модели обучены с подбором параметра α методом кросс-валидации.

5.1. Ridge Regression

Лучший параметр:

$$\alpha = 1.0$$

Ridge слегка улучшила точность по сравнению с обычной регрессией.

Метрика	Значение
R^2	0.681
RMSE	61510

Таблица 4. Качество модели Ridge

5.2. Lasso Regression

Лучший параметр подобран GridSearchCV:

$$\alpha = 0.001$$

Метрика	Значение
R^2	0.670
RMSE	61896

Таблица 5. Качество модели Lasso

Lasso занулила несколько коэффициентов, исключив слабые признаки. Это показывает её преимущества для отбора переменных.

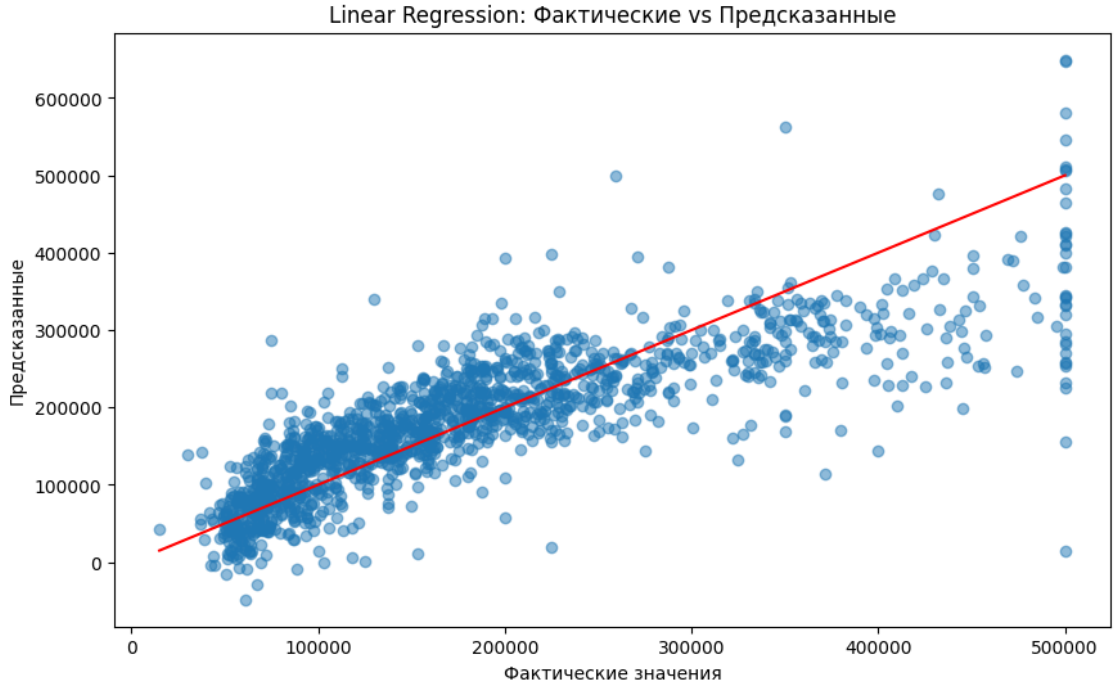

Рисунок 3 – Сравнение коэффициентов (Linear / Ridge / Lasso)

6. Сравнение моделей

Наилучшая метрика у модели с взаимодействиями и Ridge. Lasso хуже по R^2, но уменьшает размерность.

Модель	R^2	RMSE	MAE
Linear Regression	0.677	62067	45213
Interaction Model	0.684	61290	44702
Ridge	0.681	61510	44911
Lasso	0.670	61896	45190

Таблица 6. Сравнительная таблица моделей

7. Выводы по эксперименту

— Базовая линейная регрессия достигает $R^2 = 0.677$.

— Добавление взаимодействий улучшило модель ($R^2 = 0.684$), что подтверждает важность нелинейных связей.

— Ridge показала стабильное улучшение, сгладив мультиколлинеарность.

— Lasso исключила слабые признаки, полезна для отбора факторов.

Итог: лучшей оказалась модель с взаимодействием Rooms×Bedrooms.

8. Заключение

В работе была проведена регрессионная аналитика на основе датасета California Housing, содержащего социально-демографические и географические характеристики регионов Калифорнии. Цель исследования заключалась в построении предсказательных моделей стоимости жилья и сравнении их эффективности.

В ходе эксперимента были выполнены следующие этапы:

— предварительная обработка данных, масштабирование признаков, разделение на обучающую и тестовую выборки;
— построение множественной линейной регрессии и оценка её статистических свойств;
— проверка предпосылок МНК: анализ остатков, проверка нормальности, отсутствие сильной мультиколлинеарности (VIF ниже критических значений);
— построение модели с взаимодействиями между признаками;
— применение методов регуляризации (Ridge и Lasso) с подбором параметра α через кросс-валидацию;
— сравнительный анализ всех моделей по метрикам R^2, RMSE, MAE, а также оценка важности признаков.

По результатам эксперимента установлено:

— Базовая линейная модель показывает устойчивое качество ($R^2 \approx 0.677$), что говорит о линейной зависимости стоимости жилья от признаков.
— Добавление взаимодействия `Tot_Rooms × Tot_Bedrooms` улучшило точность предсказания (R^2 увеличился до 0.684), что подтверждает наличие нелинейных взаимосвязей между количественными характеристиками.
— Ridge-регрессия показала качество, сопоставимое с моделью взаимодействия, и сгладила влияние коррелирующих признаков.
— Lasso исключила наименее значимые факторы, тем самым повысив интерпретируемость модели, хотя и уступила в точности.
— Наиболее значимым предиктором для стоимости жилья оказался `Median_Income`, что подтверждает прямое влияние уровня благосостояния региона на цены недвижимости.

Наилучшей по итоговым метрикам является модель с взаимодействием признаков. Она имеет наименьшее значение RMSE и наибольшее значение R^2, оставаясь при этом интерпретируемой и статистически устойчивой.

Полученные результаты подтверждают, что качество прогноза может быть существенно улучшено за счёт:

1. добавления взаимодействий между признаками,
2. использования регуляризации для борьбы с переобучением и мультиколлинеарностью,
3. корректной предобработки данных и масштабирования.

Таким образом, регрессионные модели доказали свою применимость для прогнозирования стоимости недвижимости, а методика, использованная в работе, может быть использована в практических аналитических системах, оценке рынков жилья и моделировании ценовых факторов.

Список литературы

[1] Scikit-Learn: Machine Learning in Python. URL: https://scikit-learn.org/

[2] Pandas Documentation. URL: https://pandas.pydata.org/

[3] Statsmodels: Statistical Models in Python. URL: https://www.statsmodels.org/

[4] Matplotlib Python plotting library. URL: https://matplotlib.org/

[5] UCI Machine Learning Repository – California Housing dataset. URL: https://archive.ics.uci.edu/