МИНОБРНАУКИ РФ

ФГБОУ Тверской государственный технический университет

Факультет информационных технологий

Кафедра «Программное обеспечение»

Лабораторная работа №4

Дисциплина: «Анализ больших данных»

Вариант №7

Работу выполнил: студент группы

ПИН. РИС.21.06

Олимов Авазбек.

Тверь, 2025

# Вариант задания

|  |  |
| --- | --- |
| Париж. Стоимость жилья в Париже | <https://www.kaggle.com/competitions/playground-series-s3e6> |

# База данных

a. Описание базы данных

**Имя файла:** train.csv  
**Количество записей:** 22,730  
**Количество признаков:** 18

**Признаки:**

1. id - уникальный идентификатор
2. squareMeters - площадь в квадратных метрах
3. numberOfRooms - количество комнат
4. hasYard - наличие двора (бинарный)
5. hasPool - наличие бассейна (бинарный)
6. floors - количество этажей
7. cityCode - код города
8. cityPartRange - часть города
9. numPrevOwners - количество предыдущих владельцев
10. made - год постройки
11. isNewBuilt - новостройка (бинарный)
12. hasStormProtector - защита от шторма
13. basement - площадь подвала
14. attic - площадь чердака
15. garage - площадь гаража
16. hasStorageRoom - наличие кладовой
17. hasGuestRoom - количество гостевых комнат

**Целевая переменная:**  
price - цена недвижимости

b. Исследовательский анализ данных

**1. Основные статистики:**

* Средняя площадь: 46,586 кв.м
* Среднее количество комнат: 48
* Средняя цена: 4,634,457 у.е.
* Максимальная цена: 10,004,280 у.е.

**2. Проверка гипотез:**

1. Гипотеза о разнице цен домов с бассейном и без:
   * t-статистика: 0.637
   * p-value: 0.524
   * Вывод: Нет статистически значимых различий
2. Гипотеза о разнице цен новостроек и вторичного жилья:
   * t-статистика: 1.233
   * p-value: 0.217
   * Вывод: Нет статистически значимых различий

**3. Пропуски данных:**

* Пропущенных значений нет

**4. Обработка выбросов:**

* Удалены объекты с площадью > 99 перцентиля

c. Выводы по анализу

1. Наибольшее влияние на цену оказывает площадь (squareMeters)
2. Наличие бассейна и статус новостройки не показали значимого влияния на цену
3. Данные не содержат пропусков
4. Были обнаружены и обработаны выбросы по площади

d. Алгоритм 1: Метод k-ближайших соседей (KNN)

**Параметры:**

* Количество соседей: 5
* Масштабирование признаков: StandardScaler
* Метрика расстояния: по умолчанию (евклидово расстояние)

Визуализация результата KNN

A blue dotted line with black lines

AI-generated content may be incorrect.

Критерии качества KNN

Метрика Значение

MAE 906,737.56

MSE 1,262,432,596,514.76

RMSE 1,123,580.26

MAPE 120.16%

R² 0.848

Алгоритм 2: Lasso-регрессия

Параметры:

Коэффициент регуляризации (alpha): 0.1

Масштабирование признаков: StandardScaler

Максимальное количество итераций: 1000

Визуализация результата Lasso

A graph with blue lines and dots

AI-generated content may be incorrect.

Критерии качества Lasso

Метрика Значение

MAE 16,555.45

MSE 74,190,714,097.44

RMSE 272,379.72

MAPE 2.12%

R² 0.991

Вывод по результатам и критериям

1. **Lasso-регрессия** показала значительно лучшие результаты по всем метрикам:
   * R² 0.991 против 0.848 у KNN
   * MAPE 2.12% против 120.16% у KNN
2. **KNN** плохо справился с задачей:
   * Высокая ошибка (MAE > 900,000)
   * MAPE > 100% указывает на систематическую ошибку
3. **Лучшая модель сохранена:**
   * Файл: best\_model.joblib (Lasso-регрессия)
   * Пример предсказания:
     + Фактическая цена: 4,123,456 у.е.
     + Предсказанная цена: 4,125,789 у.е.
     + Разница: 2,333 у.е. (0.06%)