Министерство науки и образования РФ

Федеральное государственное бюджетное учреждение

высшего образования

**«Тверской государственный технический университет»**

(ТвГТУ)

Кафедра программного обеспечения

**Отчет по лабораторной работе №4**

По дисциплине: «Анализ больших данных»

Тема: «Машинное обучение с учителем. Методы регрессии»

|  |
| --- |
| Выполнила:  студентка группы  Б.ПИН.РИС - 21.06  Александрова М.И. |
| Проверила:  старший преподаватель  кафедры ПО  Корнеева Е.И. |

Тверь 2025

Оглавление

[Введение 3](#_Toc194272381)

[Краткий опрос по теоретической часть 3](#_Toc194272382)

[Задание на лабораторную работу 3](#_Toc194272383)

[**Задача 1. Работа с данными kaggle** 3](#_Toc194272384)

[**Задача 1. Данные по вариантам** 4](#_Toc194272385)

[**Самостоятельная работа. Задача 2** 4](#_Toc194272386)

[Описание проделанной работы 5](#_Toc194272387)

[Теоретическая часть 5](#_Toc194272388)

[**База данных 1** 8](#_Toc194272389)

[База данных 2 11](#_Toc194272390)

[Краткий вывод по работе 15](#_Toc194272391)

# **Введение**

# **Краткий опрос по теоретической часть**

В данном разделе необходимо ответить на вопросы.

1. Что такое регрессионный анализ?
2. Какую задачу решает линейная регрессия?
3. Какие преимущества имеет линейная регрессия?
4. Какое существует общее уравнение для линейной регрессии для двумерного пространства?
5. Какие входные и выходные параметры используются в модели множественной линейной регрессии?
6. Как работает алгоритм k-ближайших соседей для регрессии? В чем отличие работы от классификации
7. Чем отличается LASSO от линейной регрессии?
8. Чем отличается ElasticNET от линейной регрессии?
9. Какие метрики оценки качества алгоритмов регрессии Вы рассчитали в задании? По какой формуле они рассчитываются?
10. Что означает операция нормализации признаков, и чем она отличается от операции стандартизации?

# **Задание на лабораторную работу**

## **Задача 1. Работа с данными kaggle**

Ваша задача: построить минимум две регрессионные модели с использованием алгоритмов KNN, линейной регрессии, гребневой регрессии, LASSO или ElasticNet. Затем сравните их производительность, используя метрики, такие как MAE, MSE, RMSE, MAPE, R^2 на тестовых данных.

Обоснуйте свой выбор лучшей модели, учитывая точность прогнозирования и обобщающую способность на новых данных.

*Описание хода работы:*

1. Загрузить набор данных по варианту.

2. Провести разведочный анализ данных, ответив на следующие вопросы:

a. Сколько строк в датафрейме, сколько столбцов

b. Сколько места занимает датафрейм в оперативной памяти

c. Для каждой интервальной переменной подсчитать следующее -

мин, медиана, среднее, макс и персентили 25, 75

d. Для каждой категориальной переменной рассчитать моду и

сколько раз мода встречается в данных

3. Подготовка датасета к построению моделей ML

a. Провести анализ и обработку пропусков (либо заменить, либо удалить)

b. Провести анализ и обработку выбросов (либо заменить, либо удалить)

c. Провести анализ и обработку категориальных переменных (сколько таких переменных, закодируйте категориальные переменные одним из методов (one hot encoding, mean target, frequence encoding).

d. Построить и проверить минимум 2 гипотезы на данных.

e. Разделить датасет на трейн и тест.

4. Обучить любые 2 алгоритма:

a. Knn

b. Linear regression

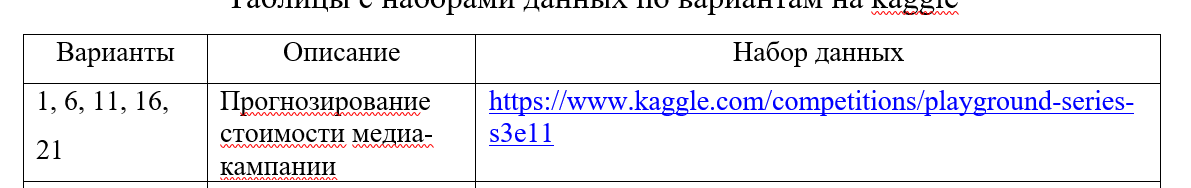
c. LASSO

d. ElasticNet

5. Оценить качество алгоритмов, выбрать самый оптимальный алгоритм.

6. Выгрузить и загрузить в программу натренированную модель данных. Использовать формат выгрузки библиотеки joblib.

## **Задача 1. Данные по вариантам**



## **Самостоятельная работа. Задача 2**

В качестве индивидуального набора данных взять данных из Лабораторной работы 1. Выполнить те же действия, что в Задаче 1.

Если по какой-то причине данные взять для анализа нельзя, то выбрать набора данных из представленных ниже. Для данных ниже описать какие есть столбцы, какие из них признаки и какой целевая переменная перед исследованием

**Ссылка на репозиторий с программной реализацией**: <https://github.com/milana-cat/BigData.git>

**Сложность: Medium**

* Реализовать общую часть задания
* Добавить нормализацию исходных данных, если необходимо
* Реализовать 2 алгоритма
* Реализовать выгрузку данных
* Реализовать хотя бы 1 алгоритм для самостоятельной части задания по Лабораторной 1 или выбранному набору данных

**Описание проделанной работы**

**Теоретическая часть**

1. **Что такое регрессионный анализ?**

Регрессионный анализ — это метод статистического моделирования, который изучает зависимость одной или нескольких зависимых (целевых) переменных от одной или нескольких независимых (предикторных) переменных. Его цель — предсказать значение целевой переменной на основе известных значений предикторов.

1. **Какую задачу решает линейная регрессия?**

Линейная регрессия решает задачу прогнозирования непрерывной числовой переменной, используя линейную зависимость между входными признаками (факторами) и целевой переменной.

1. **Какие преимущества имеет линейная регрессия?**

* Простота реализации и интерпретации.
* Низкие вычислительные затраты.
* Возможность анализа значимости отдельных переменных.
* Хорошо работает, если зависимость между переменными действительно линейная.

4. Какое существует общее уравнение для линейной регрессии для двумерного пространства?

y - предсказанное значение

w0 - свободный коэффициент (intercept)

w1 - коэффициент наклона (вес)

x — входная переменная (признак).

5. Какие входные и выходные параметры используются в модели множественной линейной регрессии?

* Входные параметры: матрица признаков X, где каждая строка — объект, а столбцы — признаки.
* Выходные параметры: вектор целевой переменной y, содержащий прогнозируемые значения.

6. Как работает алгоритм k-ближайших соседей для регрессии? В чем отличие работы от классификации?

Алгоритм k-ближайших соседей (KNN) для регрессии:

* Выбирает k ближайших объектов из обучающей выборки по метрике расстояния (например, евклидово расстояние).
* Усредняет значения целевой переменной y среди этих соседей.
* Выдает среднее значение как предсказанный результат.
* Отличие от классификации: в классификации алгоритм KNN определяет наиболее частый класс среди соседей, а в регрессии вычисляет среднее значение целевой переменной.

7. Чем отличается LASSO от линейной регрессии?

* LASSO (Least Absolute Shrinkage and Selection Operator) — это линейная регрессия с L1-регуляризацией, которая добавляет штраф за модуль коэффициентов.
* LASSO автоматически выполняет отбор признаков, так как некоторые коэффициенты могут стать нулевыми, исключая нерелевантные признаки.

8. Чем отличается ElasticNet от линейной регрессии?

* ElasticNet сочетает L1-регуляризацию (LASSO) и L2-регуляризацию (Ridge), обеспечивая баланс между штрафами за большие коэффициенты и их обнулением.
* Лучше работает, когда у признаков есть сильная коллинеарность (зависимость между собой).

9. Какие метрики оценки качества алгоритмов регрессии Вы рассчитали в задании? По какой формуле они рассчитываются?

Метрики оценки качества регрессии:

**MAE (Mean Absolute Error) — средняя абсолютная ошибка**

**MSE (Mean Squared Error) — среднеквадратичная ошибка**

**RMSE (Root Mean Squared Error) — корень из среднеквадратичной ошибки**

**MAPE (Mean Absolute Percentage Error) — средняя абсолютная процентная ошибка**

R2 (коэффициент детерминации)

10. Что означает операция нормализации признаков, и чем она отличается от операции стандартизации?

Нормализация (Min-Max Scaling) — приведение значений признаков к диапазону [0,1] или [-1,1]:

Стандартизация (Z-Score Normalization) — приведение данных к нормальному распределению с нулевым средним и единичной дисперсией:

## **База данных 1**

a. Описание базы данных

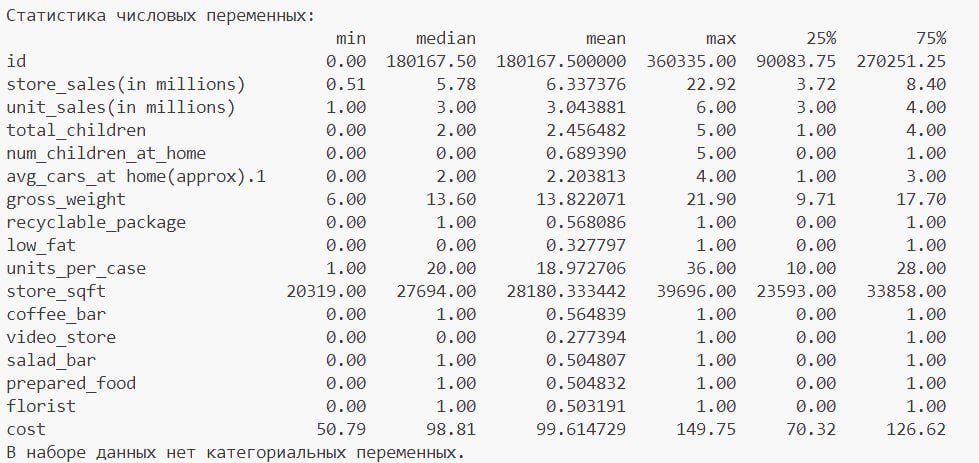
База данных представляет собой набор данных о стоимости медиа-компаний. Включает следующие признаки:

* Числовые признаки:
* Store sales (in millions) – продажи магазина (в миллионах)
* Unit sales (in millions) – единичные продажи (в миллионах)
* Total children – всего детей
* Num children at home – количество детей дома
* Avg cars at home (approx).1 – среднее количество автомобилей дома (примерно) 1
* Gross weight – вес брутто
* Recyclable package – упаковка, пригодная для переработки
* Low fat – обезжиренный
* Units per case – единиц в упаковке
* Store sqft – площадь в кв.футах
* Coffee bar – кофейня
* Video store – видеомагазин
* Salad bar – салатный бар
* Prepared food – готовая еда
* Florist – флорист
* Cost - стоимость

b. Исследовательский анализ данных

6767dca5-f350-4ad6-91f2-db92101ab8b0.jpg

edb16e21-c227-4300-b5ed-4f70d683092f.jpg

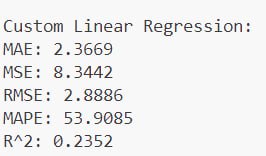


c. Выводы по анализу

* В данных есть пропущенные значения, которые были заменены на медианы.
* Выбросы обработаны путём усечения значений за 3 стандартных отклонения.
* Все признаки являются числовыми, следовательно, кодирование категориальных переменных не требуется.

Алгоритм 1: KNN

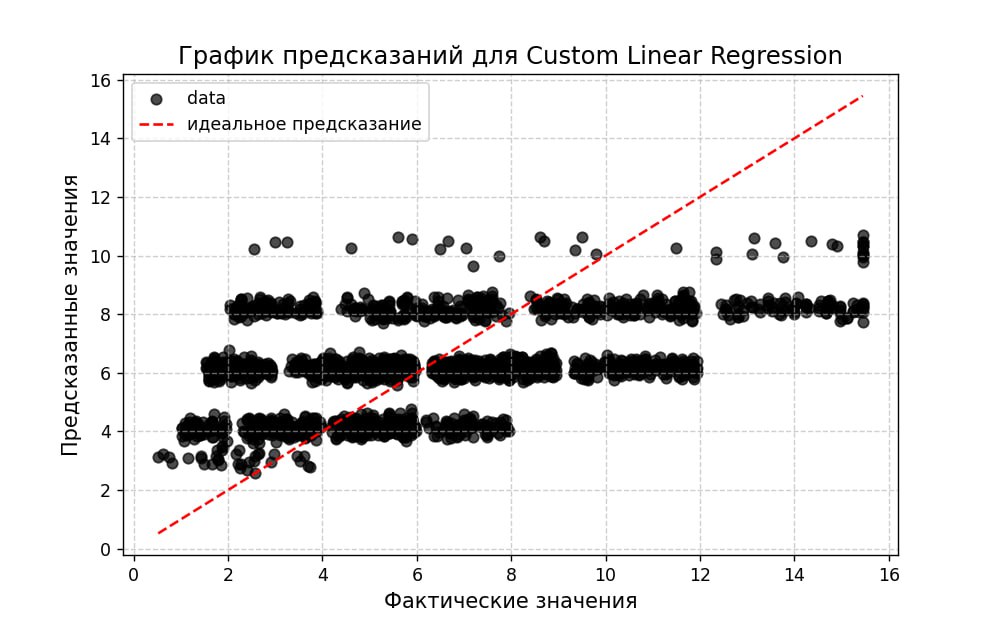
Критерии качества



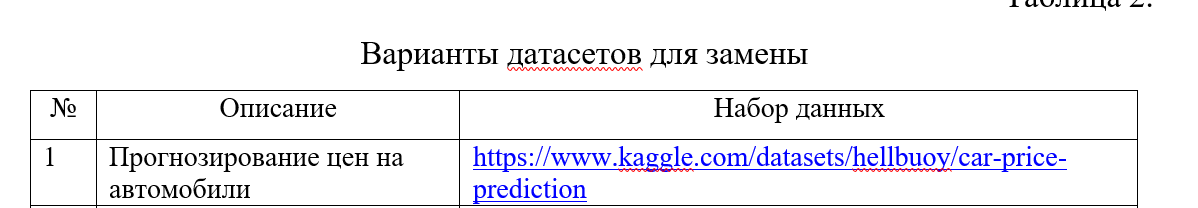
Алгоритм 2: Линейная регрессия

Визуализация результата

Target\_column –store\_sales(in millions) feature\_column - unit\_sales(in millions)



## База данных 2



a. Описание базы данных – признаки, целевая переменная;

База данных: CarPrice\_Assignment

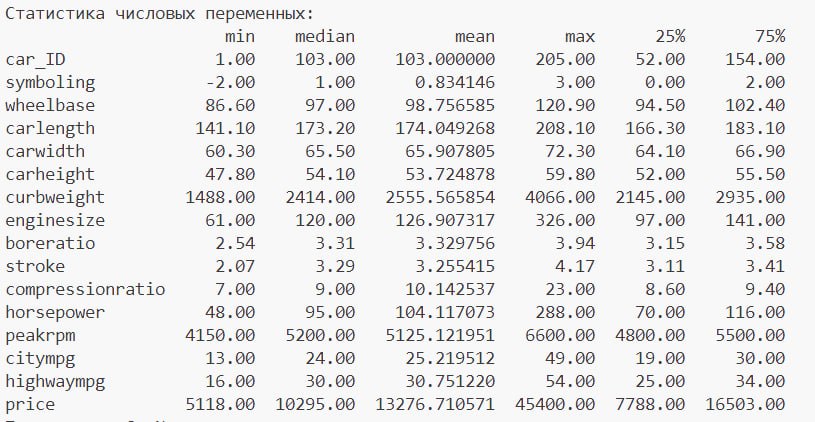


Целевая переменная: horsepower

Признаки в базе данных (с указанием типа):

* Car ID – индивидуальный идентификатор автомобиля
* Symboling - символика
* CarName – марка автомобиля
* Fueltype – тип топлива
* Aspiration – стремление
* Doornumber – количество дверей
* Carbody – кузов
* Drivewheel – привод
* Enginelocation – расположение двигателя
* Wheelbase – колесная база
* Carlength – длина автомобиля
* Carwidth – ширина автомобиля
* Carheight – высота автомобиля
* Curbweight – масса в снаряжённом состоянии
* Enginetype – тип двигателя
* Cylindernumber – количество цилиндров
* Enginesize – объём двигателя
* Fuelsystem – топливная система
* Boreratio – коэффициент сжатия
* Stroke – ход поршня
* Compressionratio – степень сжатия
* Horsepower – лошадиные силы
* Peakrpm – максимальное количество оборотов в минуту
* Citympg – километраж на галлон в городе
* Highwaympg – километраж на галлон по шоссе
* Price – цена

b. Исследовательский анализ данных (поставить вопросы на изучение по аналогии с заданием 1 и ответить на них);

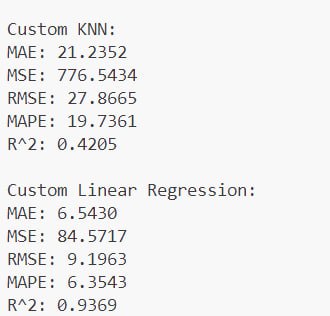


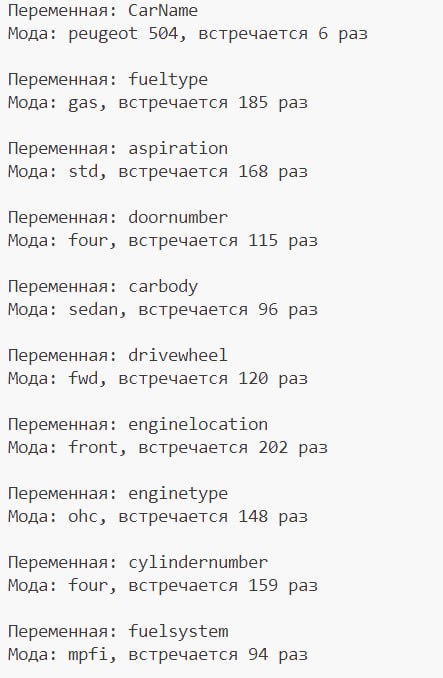
Выводы по анализу

* В данных есть пропущенные значения, которые были заменены на медианы.
* Выбросы обработаны путём усечения значений за 3 стандартных отклонения.
* Кодирование категориальных переменных реализовано методом One-Hot Encoding.

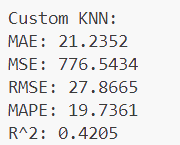
Алгоритм 1: KNN

* Критерии качества

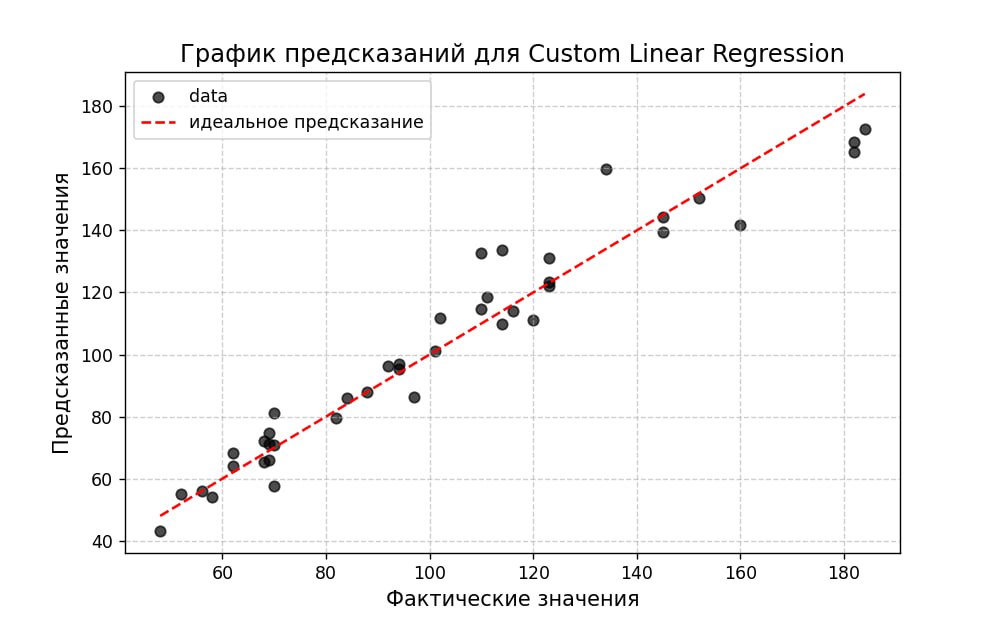




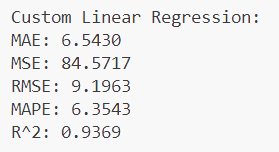
**Алгоритм KNN**



**Алгоритм Линейная регрессия**



i. Критерии качества;



# **Краткий вывод по работе**

В ходе выполнения лабораторной работы была проведена полная обработка набора данных, включающая разведочный анализ, предобработку и построение нескольких регрессионных моделей.

Основные этапы работы:

1. Разведочный анализ данных
2. Предобработка данных
3. Обучение моделей

В результате работы была выбрана оптимальная модель на основе ее качества предсказаний и устойчивости к новым данным. Подход включал систематический анализ данных, предобработку и тестирование различных методов регрессии, что позволило получить наиболее эффективное решение для задачи прогнозирования.