

**МИНОБРНАУКИ РОССИИ**  
**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**  
**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**  
**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**  
**Кафедра Вычислительной техники**

**ОТЧЕТ**  
**по лабораторной работе № 4**  
**по дисциплине «Машинное обучение»**  
**Тема: «Регрессия»**

Студент гр. 0308

Сабурова Е.А.

Преподаватель

Гатауллин Р.И.

Санкт-Петербург

2023

## **Оглавление**

<b>1. Цель работы .....</b>	<b>3</b>
<b>2. Задание .....</b>	<b>3</b>
<b>3. Краткая теоретическая информация по теме работы .....</b>	<b>3</b>
<b>4. Ход работы .....</b>	<b>6</b>
<b>5. Выводы .....</b>	<b>10</b>

## **1. Цель работы**

Получение и закрепления навыков предобработки данных и применения методов машинного обучения для решения задач регрессии.

## **2. Задание**

### **1. Предварительная обработка данных**

a. Визуализация значимых признаков (диаграммы рассеяния, ящики с усами, гистограммы)

b. Очистка данных (удаление пропусков, нормализация, удаление дубликатов)

### **2. Обучение моделей и подбор параметров (где применимо):**

a. Линейная регрессия

b. LASSO

c. Ридж-регрессия

### **3. Оценка моделей**

a. Вывод метрик

b. Построение графиков

## **3. Краткая теоретическая информация по теме работы**

Регрессия – это задача машинного обучения с учителем, которая заключается в предсказании некоторой непрерывной величины.

Для использования регрессионных моделей нужно, чтобы в датасете были характеристики объектов и “правильные” значения целевой переменной. Задача регрессии основывается на предположении, что значение целевой переменной зависит от значения признаков. Регрессионная модель принимает набор значений и выдает предсказание значения целевой переменной.

В качестве регрессионных моделей часто берут аналитические функции, например, линейную. Примеры регрессионных задач - предсказание цены акции, оценка объекта недвижимости.

## **Обучение моделей машинного обучения, использованные в работе**

### Линейная регрессия

Линейная регрессия - это контролируемый метод машинного обучения, который используется инструментом Обучение с использованием AutoML (автоматизации большей части процесса обучения) и находит линейное уравнение, лучше всего описывающее корреляцию зависимых переменных с независимыми.

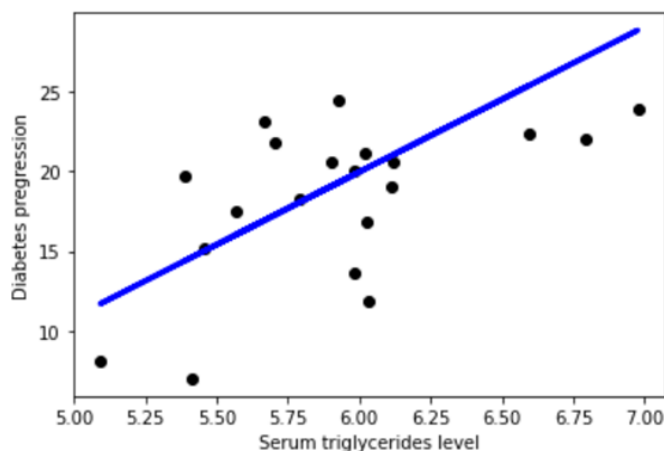


Рисунок 1 – изображение работы линейной регрессии

### LASSO (Least absolute shrinkage and selection operator)

LASSO регрессия – это метод линейной регрессии с добавлением L1-регуляризации, направленной на уменьшение весов признаков и отбор наиболее важных из них.

Особенность LASSO заключается в том, что она может приводить к «разреженным» моделям, где некоторые веса становятся точно нулевыми.

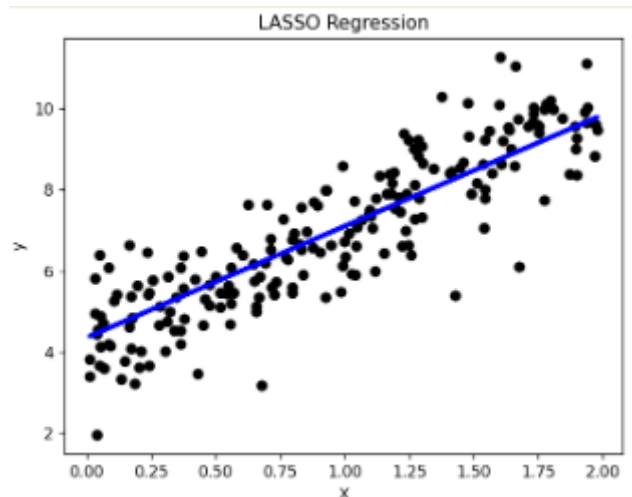


Рисунок 2 – изображение LASSO

### Ридж-регрессия

Ридж-регрессия – это метод, схожий с LASSO, только вместо L1-регуляризации используется L2.

Стоит отметить, что в отличие от L1, L2-регуляризация склонна уменьшать веса, но не устанавливает их в точно нулевые значения. Это позволяет этой регрессии сохранить все признаки, даже если некоторые из них могут быть менее важными.

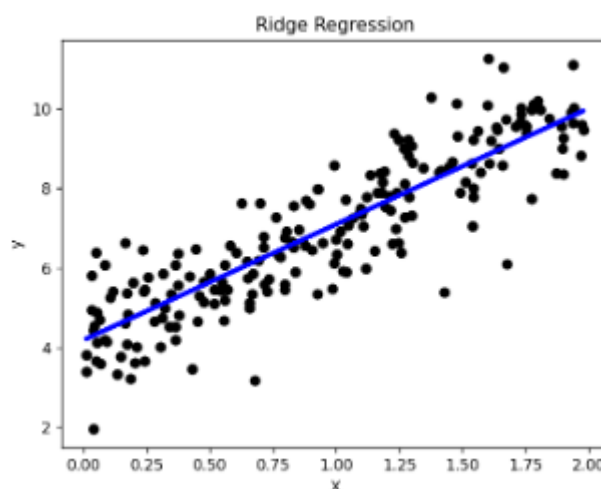


Рисунок 3 – изображение ридж-регрессии

## **4. Ход работы**

В лабораторной работе рассматривается кластеризация для набора данных о погоде в Индии.

Для определения значимых признаков была построена матрица из точечных

графиков. Матрица представлена на Рисунке 4. По матрице видно, лишь несколько признаков обладают линейной зависимостью.

Визуализация исходных данных представлена на Рисунке 4.

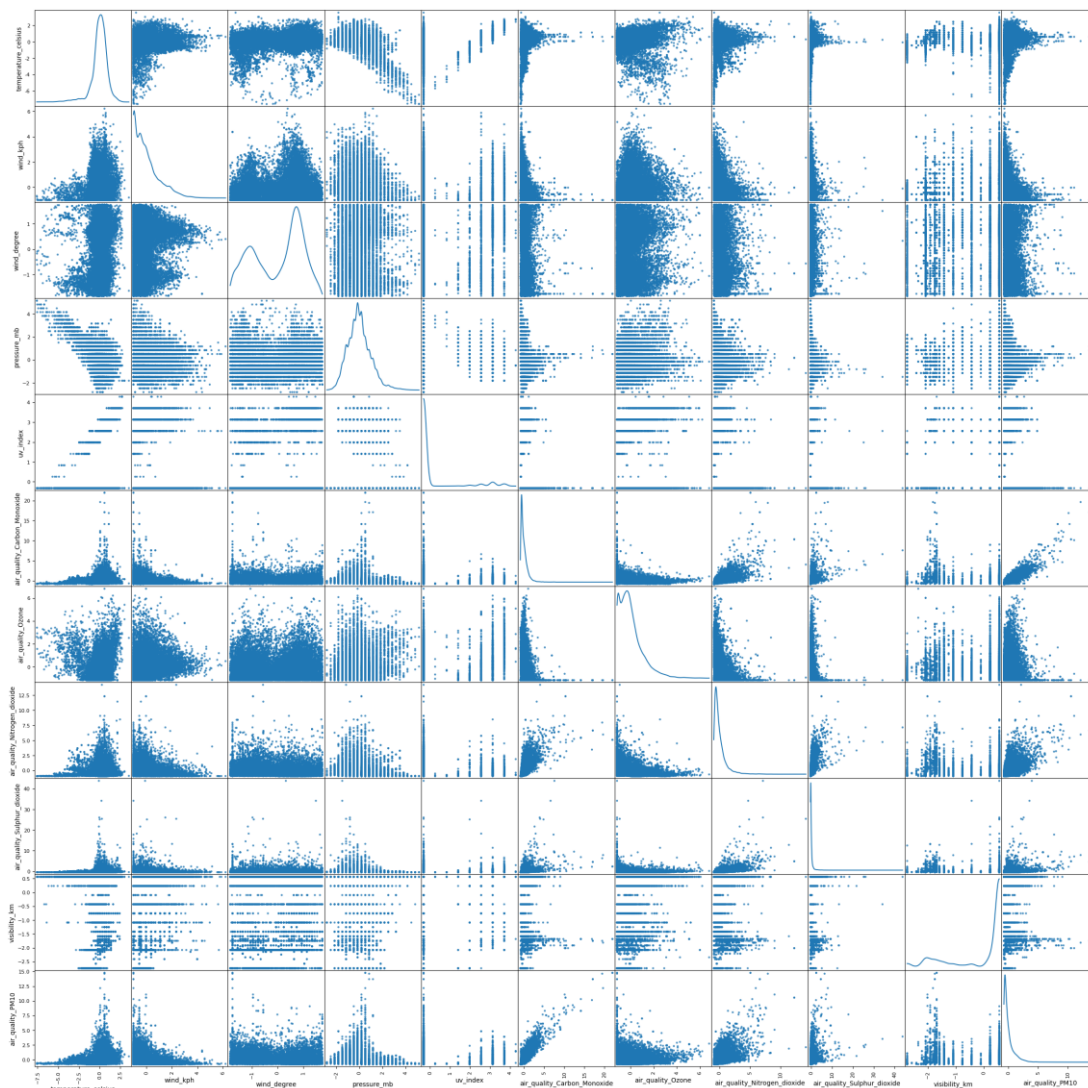


Рисунок 4 – Матрица точечных графиков

Для поиска целевой переменной, попробуем найти непрерывный числовой график. На Рисунке 5 представлен набор графиков числовых признаков. Из них самым подходящим признаком, с точки зрения логики и вида графика – один из коэффициентов качества воздуха.

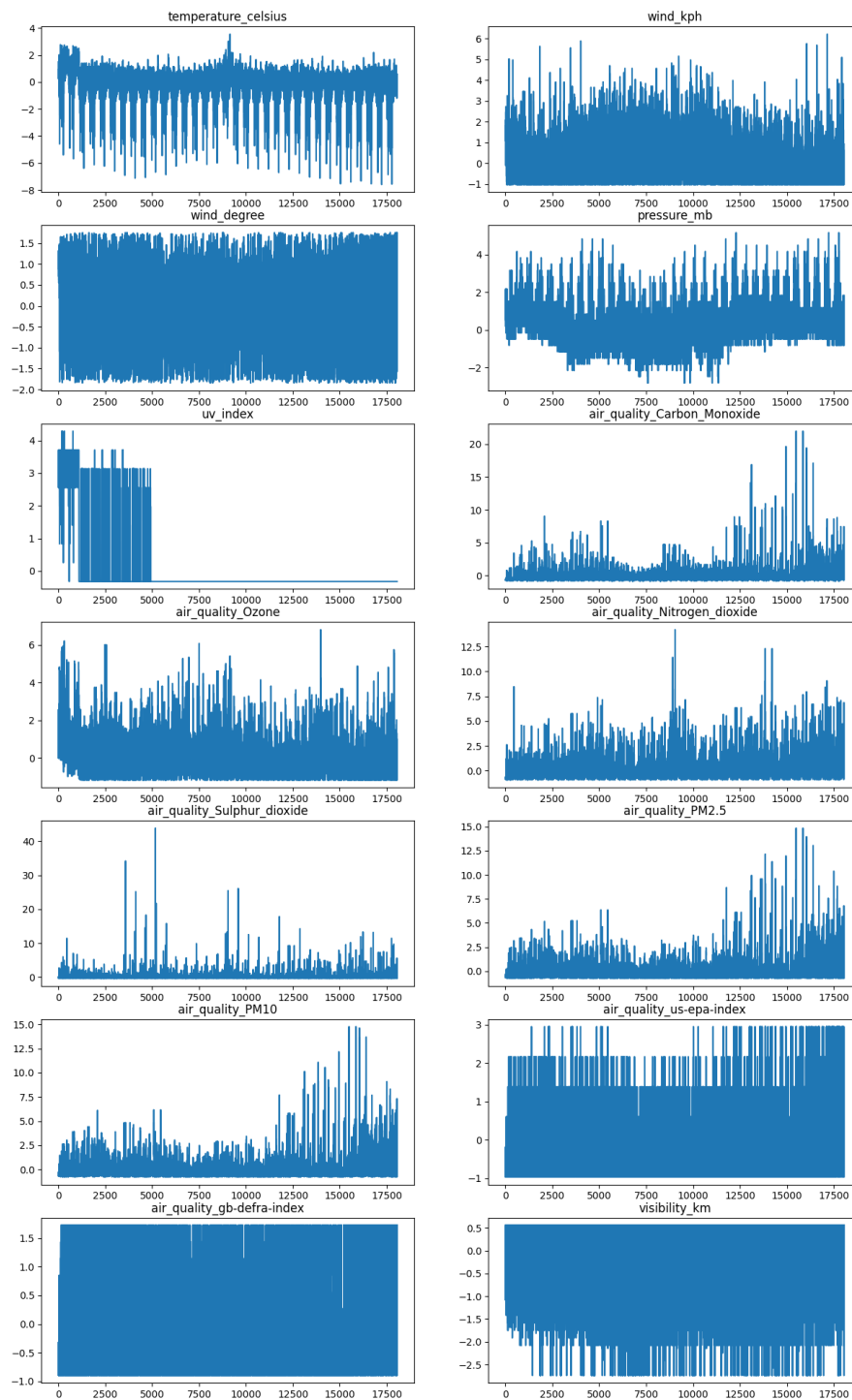


Рисунок 5 — набор графиков числовых признаков

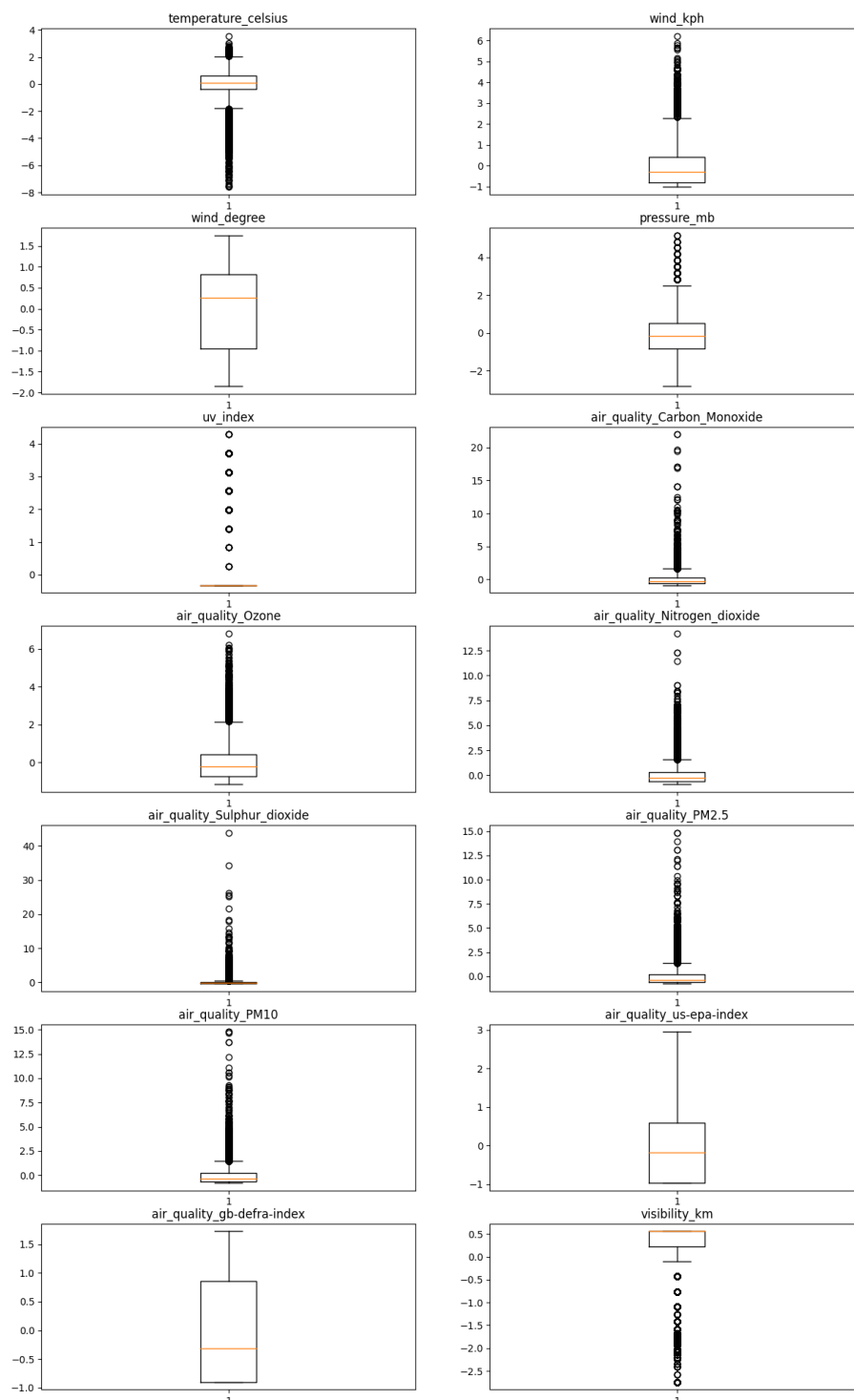


Рисунок 6 – набор графиков ящиков с усами

### ***Обучение моделей и подбор параметров***

Линейная регрессия обучалась на стандартизованном наборе данных.

Подбор параметров не производился, так как у данной модели нет подбираемых гиперпараметров.

Ridge-регрессия – обучение проходило на стандартизованном наборе



данных. Перед обучением был подобран гиперпараметр  $\alpha = 10$ . Подбор производился при помощи класса RidgeCV.

Lasso-регрессия – обучение проходило на стандартизированном наборе данных. Перед обучением был подобран гиперпараметр  $\alpha = 0.3$ . Подбор производился при помощи класса GridSearchCV.

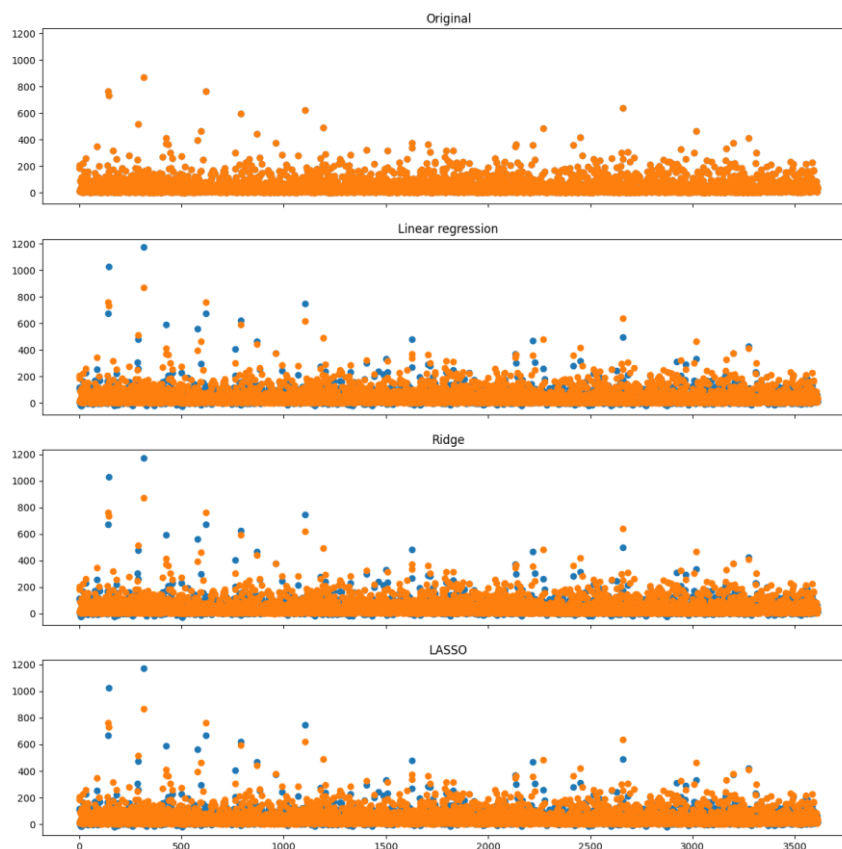
### ***Оценка моделей***

Оценка моделей по метрикам представлена на Рисунке 7.

```
---LINEAR REGRESSION---  
MSE: 1005.9479749889526  
R2: 0.7810129611024038  
  
---RIDGE---  
MSE: 1005.7219151442769  
R2: 0.7810621725698256  
  
---LASSO---  
MSE: 1010.9771392934334  
R2: 0.7799181512051228
```

*Рисунок 7 – Оценка по метрикам*

Оценка моделей по графикам представлена на Рисунке 8.



*Рисунок 8 – оценка моделей по графикам*

## **Выводы**

В результате выполнения работы было обучено 3 модели, решающих задачу регрессии. Лучше всех оказалась модель ридж-регрессии, её результаты оказались немного лучше линейной регрессии и LASSO. Небольшая разница в результатах связана с тем, что в качестве признаков были выбраны столбцы, имеющие линейную зависимость с целевой переменной. Из-за этого влияние регуляризации в ридж и LASSO регрессиях оказывало незначительное влияние на результат.