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»

Студентка гр.0305 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Петракова М. А.

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Гатауллин Р. И.

Санкт – Петербург

2023

**Цель работы**

Получения и закрепление навыков предобработки данных и применения методов машинного обучения для решения задач регрессии.

**Задача**

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

**Теоретическая часть**

Задача регрессии – это обучение с учителем. Имеется обучающая выбора: множество пар (объект, ответ) – S = {Xi, yi}, i ∈ 1:n.

Переменная yi ∈ Y = {y1, …, yn} – ответ, зависимая переменная. Каждый объект представлен набором m признаком – Xi = {xi1, …, xim}, i ∈ 1:n, m≥1. Признаки xij – факторы, независимые переменные.

Требуется найти линейную функцию (модель, функция регрессии)

f: X -> Y, аппроксимирующую значения зависимой переменной y на обучающей выборке S для всех объектов X. Функцию регрессии ищем в виде:

f(X, W) = w0 + w1X1 + … + wnXn.

Оценки для yi запишем в виде:

yi = f(Xi, W) + ϵi = w0 + w1x1i + … + wnxni + ϵi, где ϵi – независимые нормальные случайные величины с нулевым математическим ожиданием. Задача: найти W.

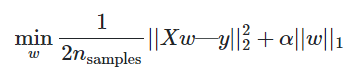
Обычный метод наименьших квадратов (Линейная регрессия).

Линейная регрессия подгоняет линейную модель к минимизации остаточной суммы квадрата между наблюдаемого целевого признака в наборе данных и предсказанного целевого признака по линейной аппроксимации. Математически это решение проблемы в следующем виде:



Лассо (LASSO).

Лассо — это линейная модель, которая оценивает разреженные коэффициенты. Это полезно в некоторых контекстах из-за своей тенденции отдавать предпочтение решениям с меньшим количеством ненулевых коэффициентов, эффективно уменьшая количество функций, от которых зависит данное решение. Математически он состоит из линейной модели с добавленным членом регуляризации. Целевая функция, которую необходимо минимизировать:



Ридж (Ridge).

Ридж-регрессия (гребневая регрессия) — это метод оценки коэффициентов моделей множественной регрессии, когда независимые переменные сильно коррелированы. Метод обеспечивает повышенную эффективность в задачах оценки параметров за счет допустимой степени смещения. Подобно регрессии Лассо (оператор наименьшего абсолютного сжатия и выбора) вводит штрафной коэффициент, накладывая ограничение, но в отличие от нее принимает квадрат, а не величину коэффициентов.

**Выполнение работы**

Предварительно была произведена подготовка данных. Построена матрица корреляции и определены признаки, корреляция признака “RainTomorrow” с которыми наибольшая (по модулю) (рис.1). Ими оказались Sunshine, Humidity3pm, Cloud3pm.

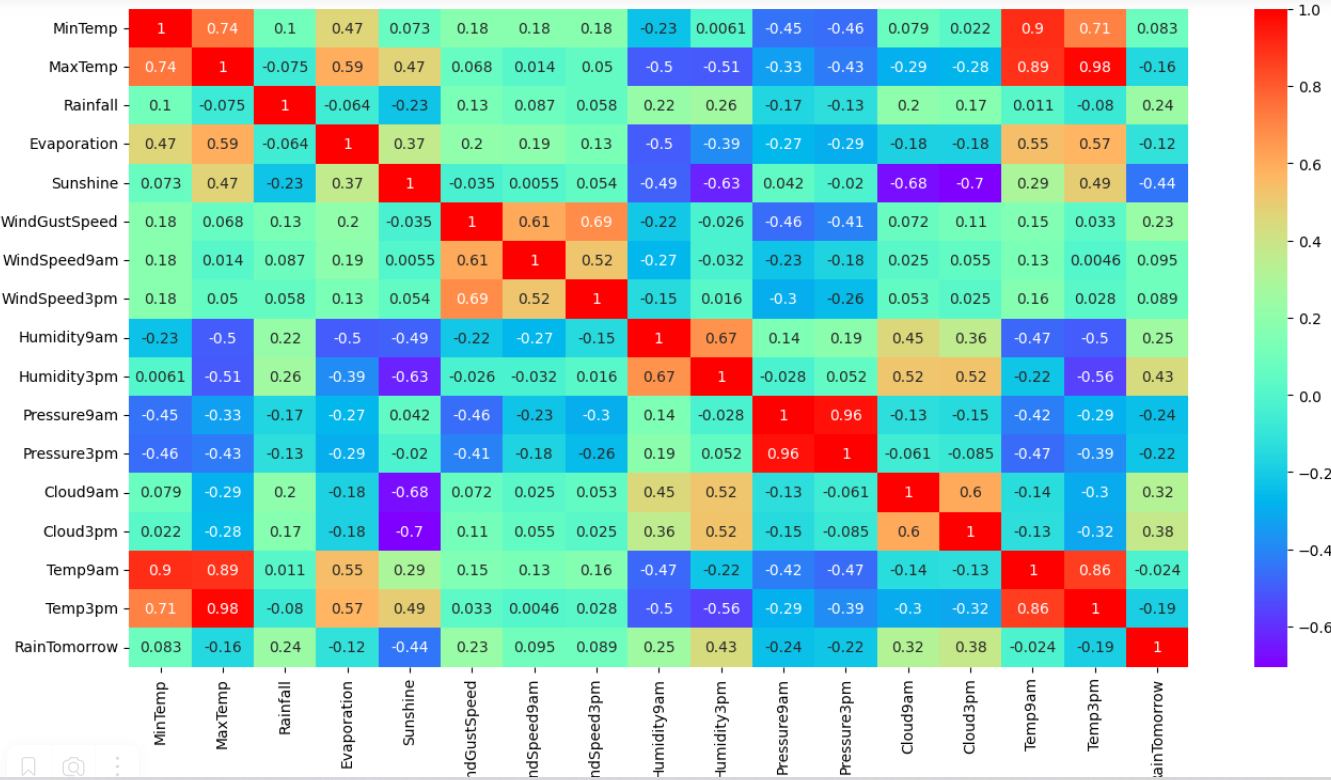


Рисунок 1

Построены гистограммы для признаков: RainTomorrow (рис.2),

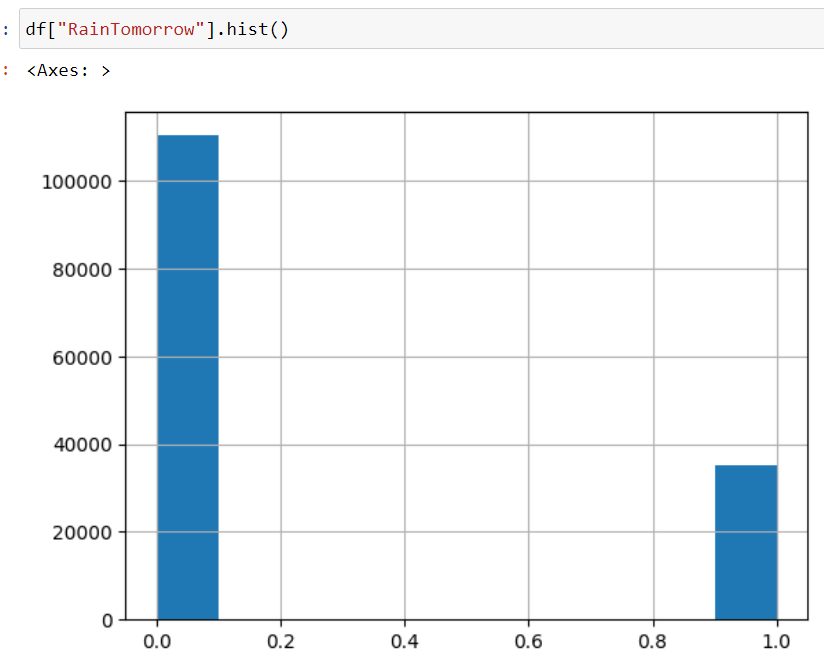


Рисунок 2

Humidity3pm (рис.3),

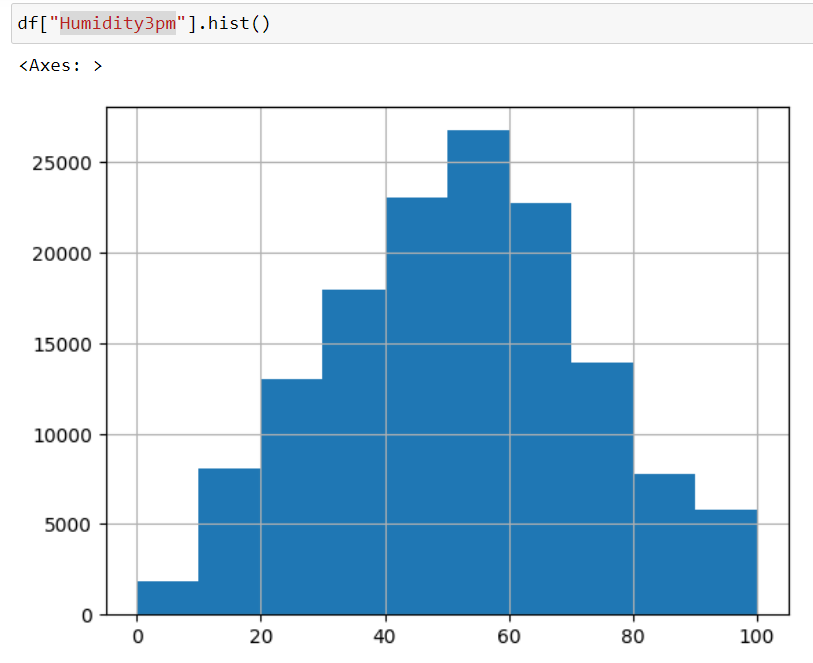


Рисунок 3

Sunshine (рис.4),

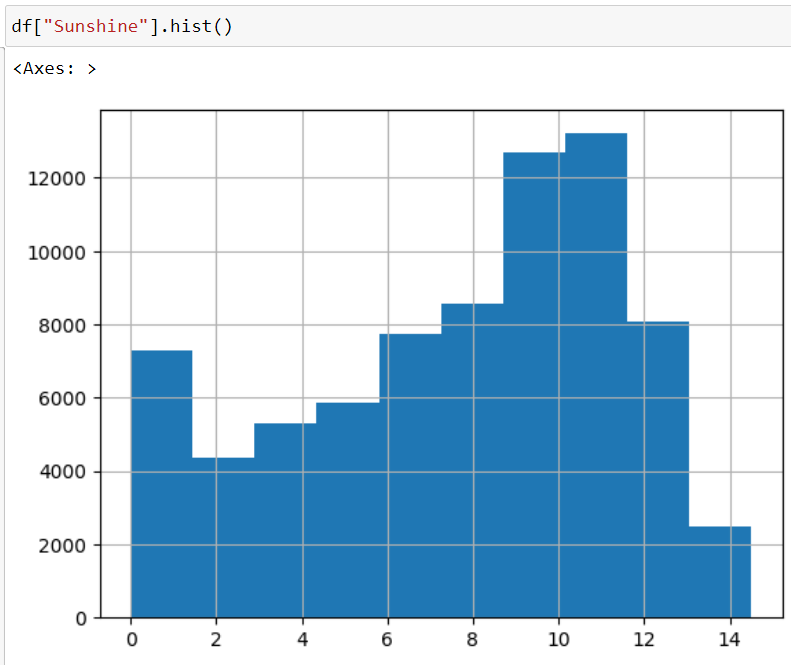


Рисунок 4

Cloud3pm (рис.5).

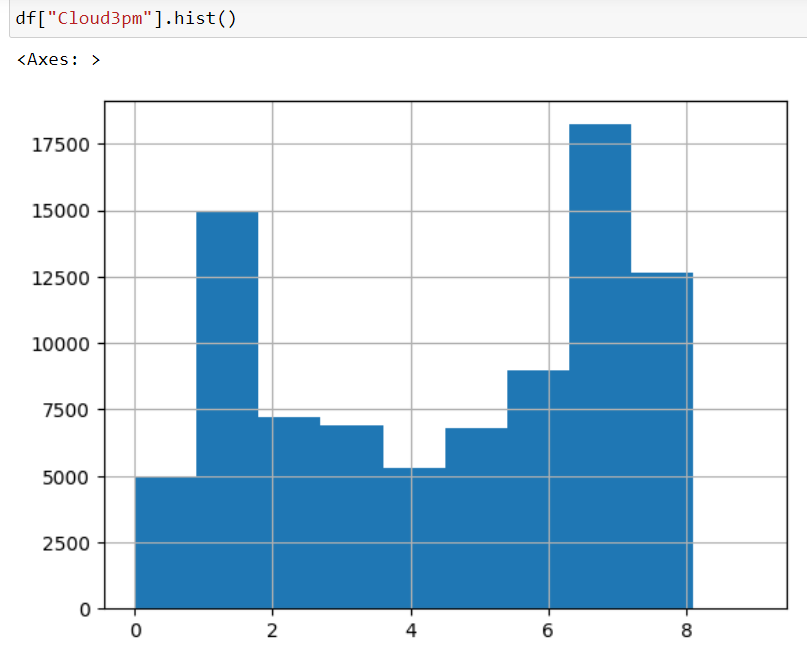


Рисунок 5

Построены диаграммы рассеяния: Sunshine+RainTomorrow (рис.6),

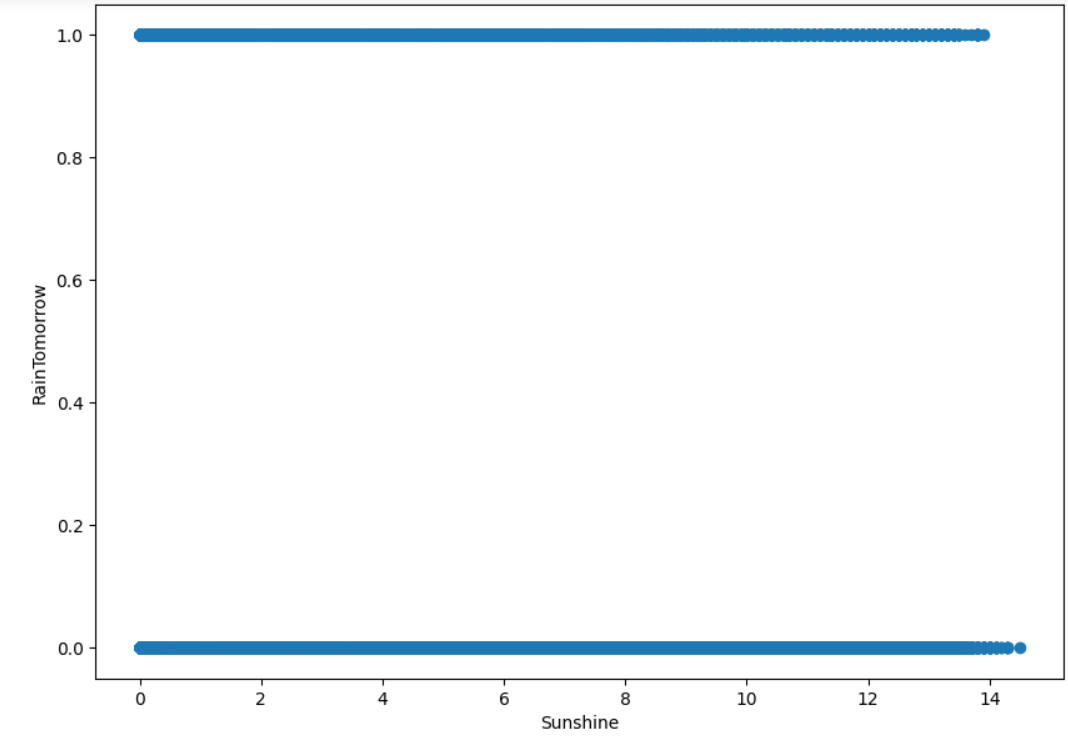


Рисунок 6

В среднем, при очень высоких значениях признака RainTomorrow=0.

Cloud3pm (рис.7),

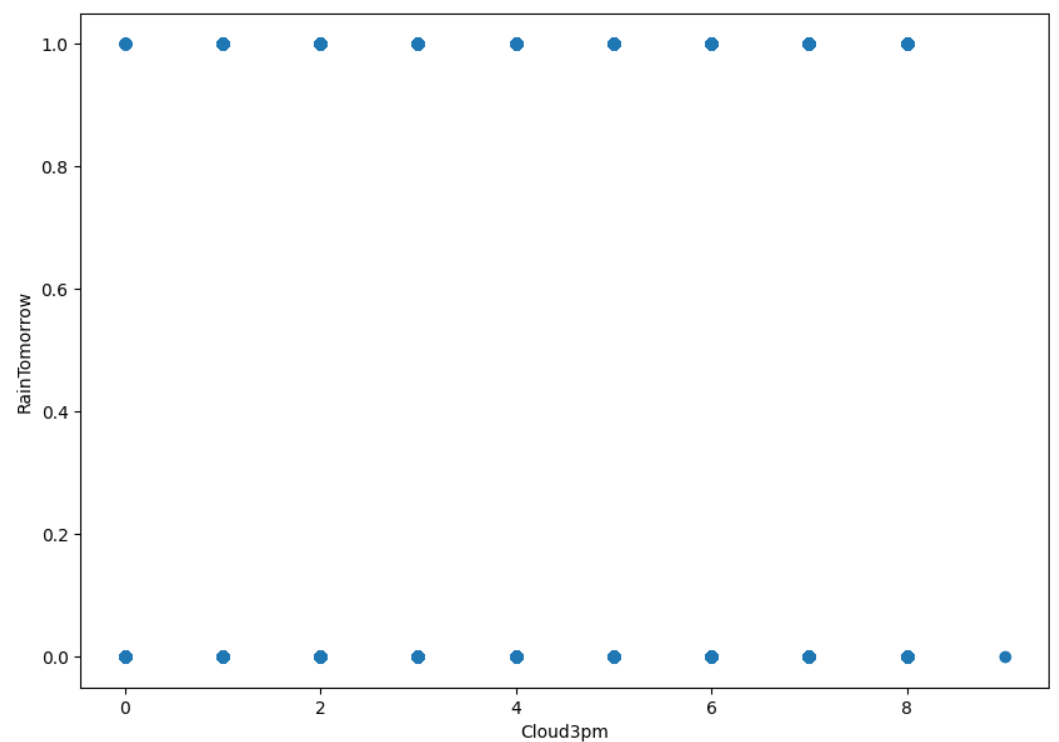


Рисунок 7

Humidity3pm (рис.8).

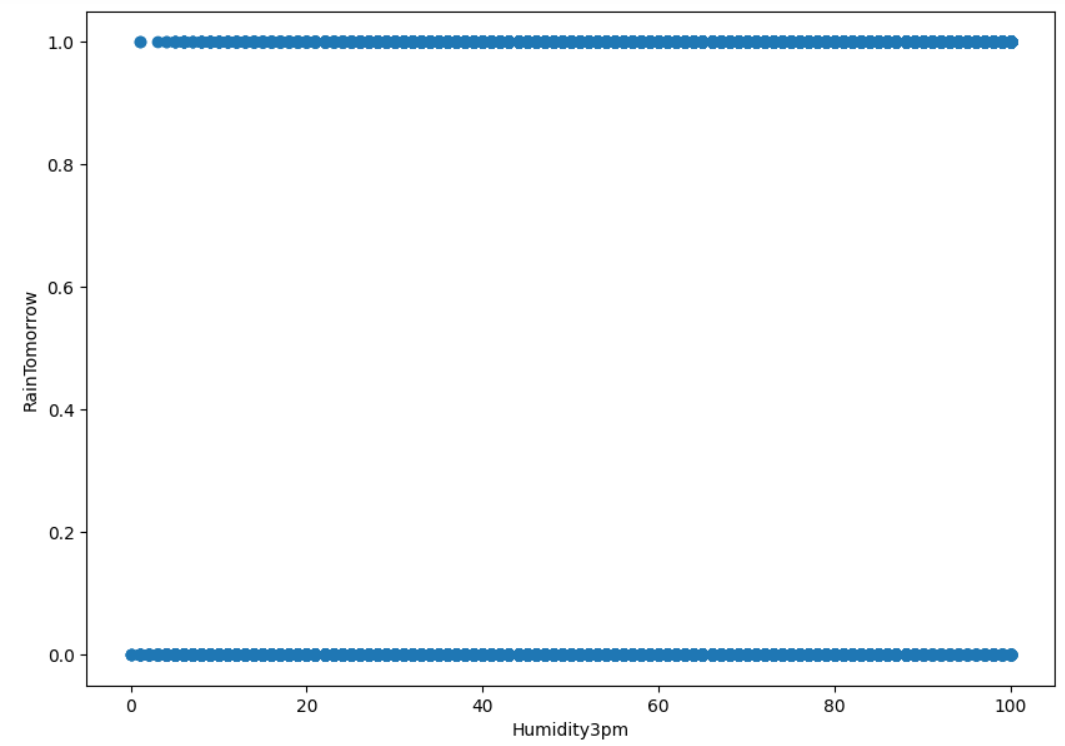


Рисунок 8

Построены ящики с усами: Humidity3pm (рис.9),

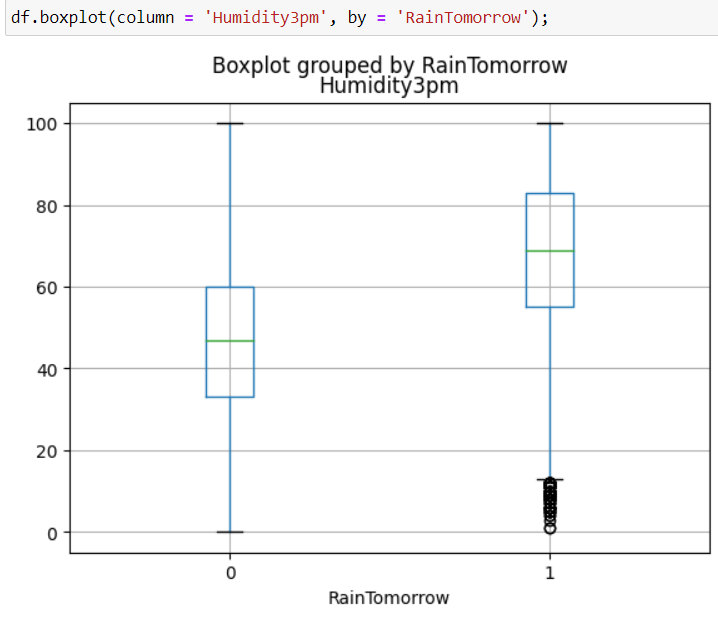


Рисунок 9

Sunshine (рис. 10),

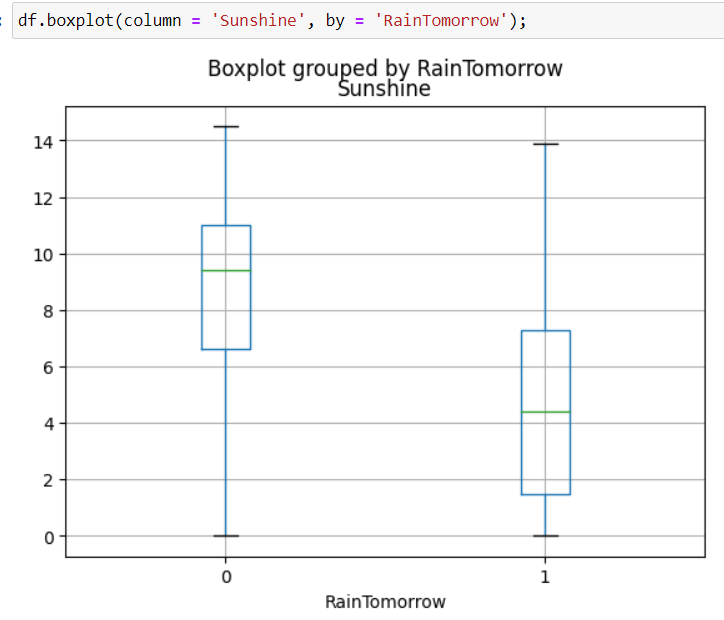


Рисунок 10

Cloud3pm (рис. 11).

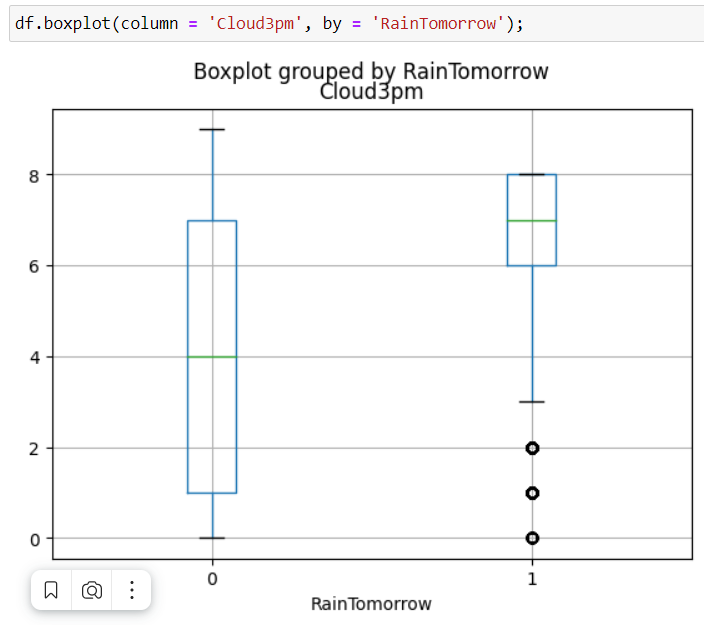


Рисунок 11

Видно, что значения признаков с RainTomorrow хорошо коррелируют. Проведем дальнейшие этапы подготовки: удаление пропусков, замена строковых данных на числовые, нормализация.

Разделение выборки на тренировочную и тестовую (рис.12):

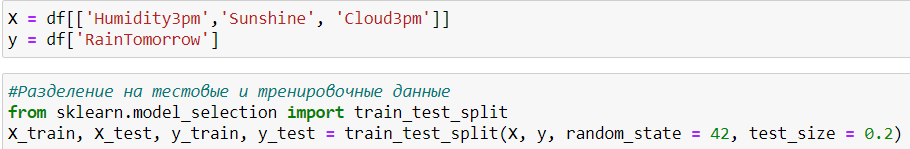


Рисунок 12

1. Линейная регрессия

Обучение модели (рис.13).

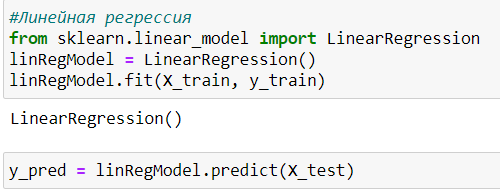


Рисунок 13

Оценки среднеквадратической и абсолютной ошибки (рис.14).

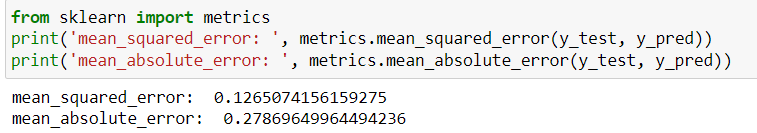


Рисунок 14

Сравнение данных с предсказанными значениями (рис.15), по оси х – Humidity3pm. Серый цвет – заданные значение RainTomorrow (y\_test), черный – предсказанные. Видно, что линия тренда поддерживается. При меньших значениях признака результат близок к 0.

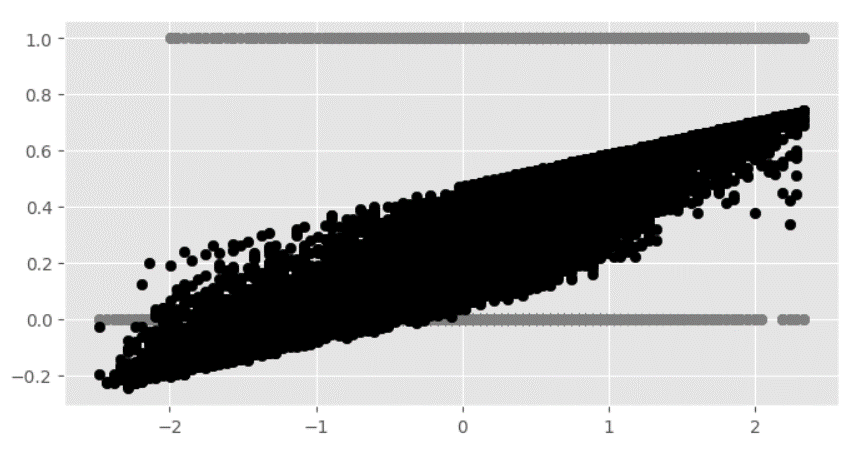


Рисунок 15

Сравнение данных с предсказанными значениями (рис.16), по оси х – Sunshine. При высоких значениях признака результат стремится к 0.

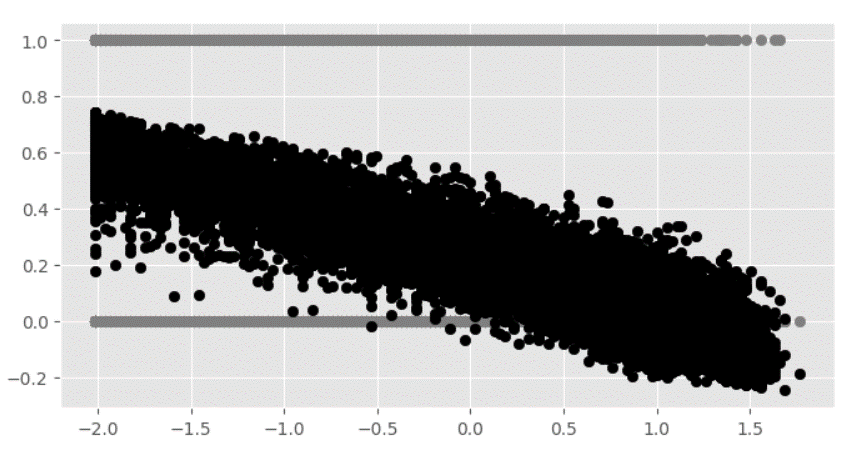


Рисунок 16

Сравнение данных с предсказанными значениями (рис.17), по оси х – Cloud3pm. При высоких значениях признака результат ближе к 1.

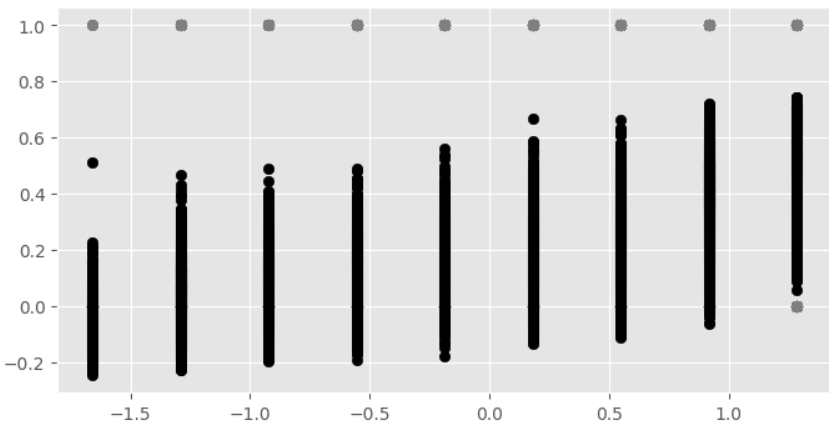


Рисунок 17

1. LASSO

C помощью GRID-search находим оптимальные параметры alpha и max\_iter (рис.18).

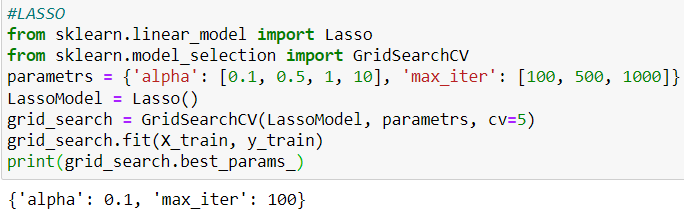


Рисунок 18

Обучение модели (рис.19).

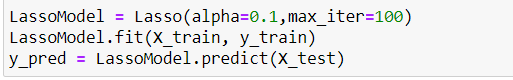


Рисунок 19

Оценки среднеквадратической и абсолютной ошибки (рис.20).

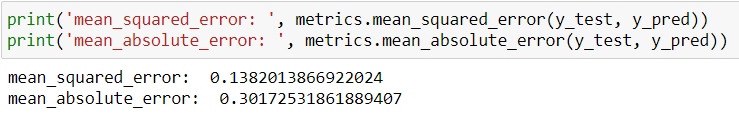


Рисунок 20

Сравнение данных с предсказанными значениями (рис.21), по оси х – Humidity3pm. Видно, что линия тренда поддерживается. При меньших значениях признака результат близок к 0. По сравнению с линейной регрессией, не имеет значений < 0.

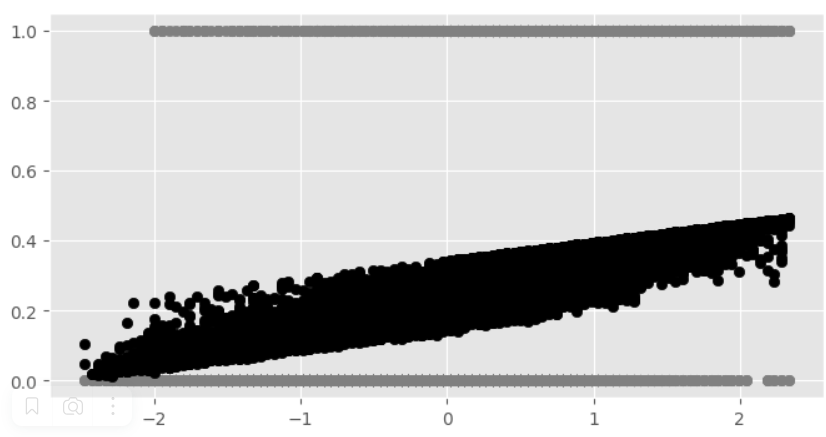


Рисунок 21

Сравнение данных с предсказанными значениями (рис.22), по оси х – Sunshine. При высоких значениях признака результат стремится к 0.

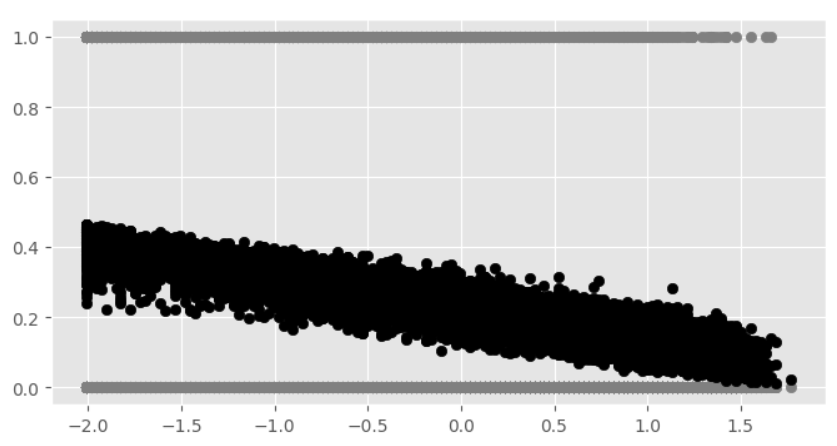


Рисунок 22

Сравнение данных с предсказанными значениями (рис.23), по оси х – Cloud3pm. Показатели колеблются около середины.

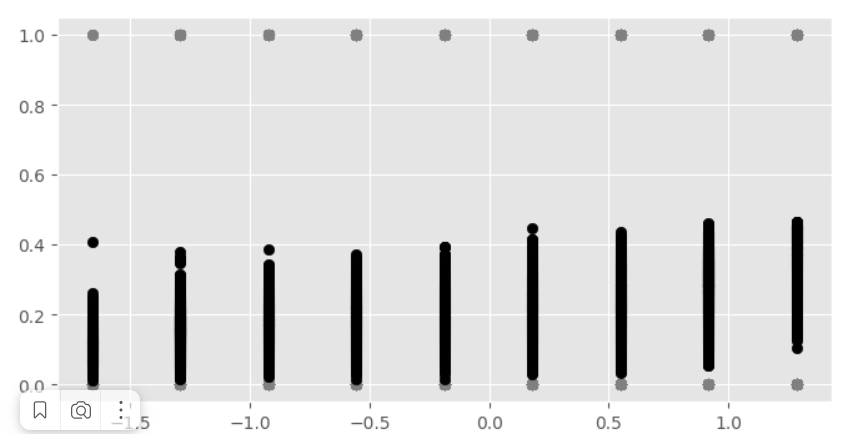


Рисунок 23

1. Ridge (гребневая регрессия)

C помощью GRID-search находим оптимальные параметры alpha и max\_iter (рис.24).

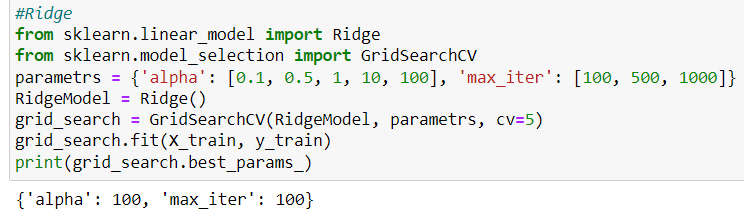


Рисунок 24

Обучение модели (рис.25).

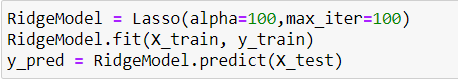


Рисунок 25

Оценки среднеквадратической и абсолютной ошибки (рис.26).

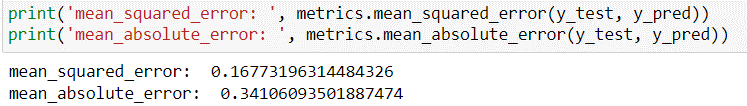


Рисунок 26

Сравнение данных с предсказанными значениями (рис.27), по оси х – Humidity3pm, рис.28 – Sunshine, рис.29 – Cloud3pm.

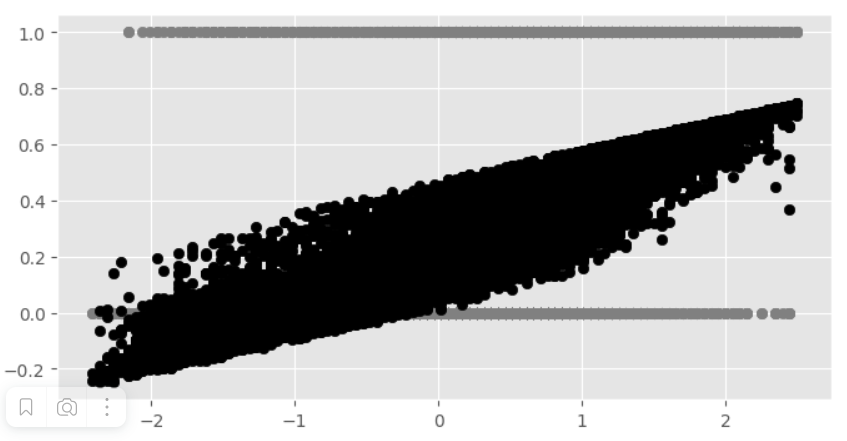


Рисунок 27

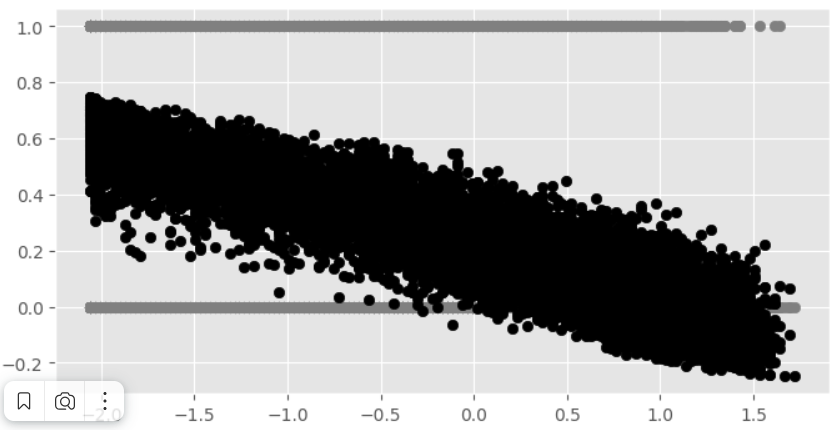


Рисунок 28

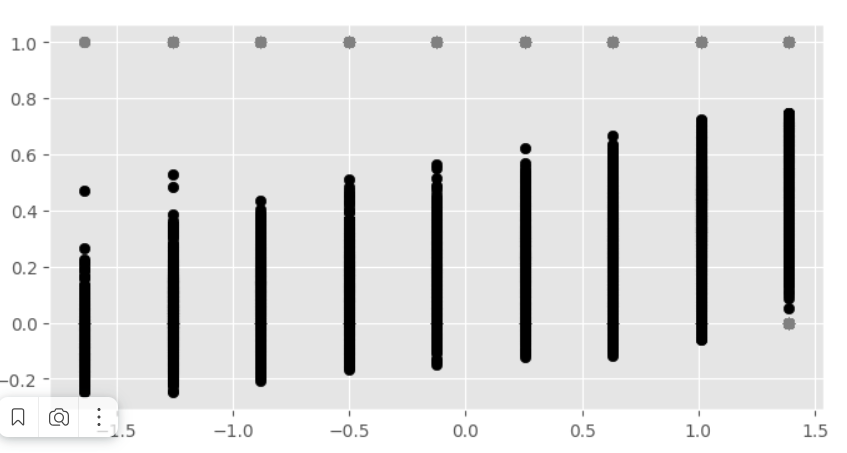


Рисунок 29

Сравнение оценок моделей.

|  |  |  |  |
| --- | --- | --- | --- |
|  | R^2 | mean\_squared\_error | mean\_absolute\_error |
| Линейная регрессия | 0.24214579926466018 | 0.1265074156159275 | 0.27869649964494236 |
| LASSO | 0.17639521113346235 | 0.1382013866922024 | 0.30172531861889407 |
| Ridge | 0.2421688932341155 | 0.13008657513144123 | 0.2825893213334669 |

Наилучшие значения ошибки показала самая простая линейная регрессия и Ridge (приближенно).

**Выводы**

В результате выполнения лабораторной работы были получены и закреплены навыки предобработки данных и применения методов машинного обучения для решения задач регрессии. Были изучены различные методы регрессии и оценки их качества. Определено, что наилучшую оценку получила линейная регрессия, в решении данной поставленной задачи. Сделан вывод о том, что выбор того или иного алгоритма регрессии зависит от применения и задачи.