Санкт-Петербургский национальный исследовательский университет ИТМО

Факультет программной инженерии и компьютерной техники

Направление подготовки 09.03.04 Программная инженерия

Дисциплина «Система искусственного интеллекта»

Лабораторная работа №4

Вариант №3

Студент

Бобрусь Александр Владимирович

Группа P33091

Преподаватель

Авдюшина Анна Евгеньевна

Санкт-Петербург, 2023 г.

Задание

- Выбор датасетов:
  - Студенты с **четным** порядковым номером в группе должны использовать набор данных о жилье в Калифорнии Скачать тут
  - Студенты с **нечетным** порядковым номером в группе должны использовать про обучение студентов
- Получите и визуализируйте статистику по датасету (включая количество, среднее значение, стандартное отклонение, минимум, максимум и различные квантили).
- Проведите предварительную обработку данных, включая обработку отсутствующих значений, кодирование категориальных признаков и нормировка.
- Разделите данные на обучающий и тестовый наборы данных.
- Реализуйте линейную регрессию с использованием метода наименьших квадратов без использования сторонних библиотек, кроме NumPy и Pandas (для использования коэффициентов использовать библиотеки тоже нельзя). Использовать минимизацию суммы квадратов разностей между фактическими и предсказанными значениями для нахождения оптимальных коэффициентов.
- Постройте **три модели** с различными наборами признаков.
- Для каждой модели проведите оценку производительности, используя метрику коэффициент детерминации, чтобы измерить, насколько хорошо модель соответствует данным.
- Сравните результаты трех моделей и сделайте выводы о том, какие признаки работают лучше всего для каждой модели.
- Бонусное задание
  - Ввести синтетический признак при построении модели

Код программы

https://github.com/BobrAll/ITMO/tree/main/term%205/Artificial%20intelligence%20systems

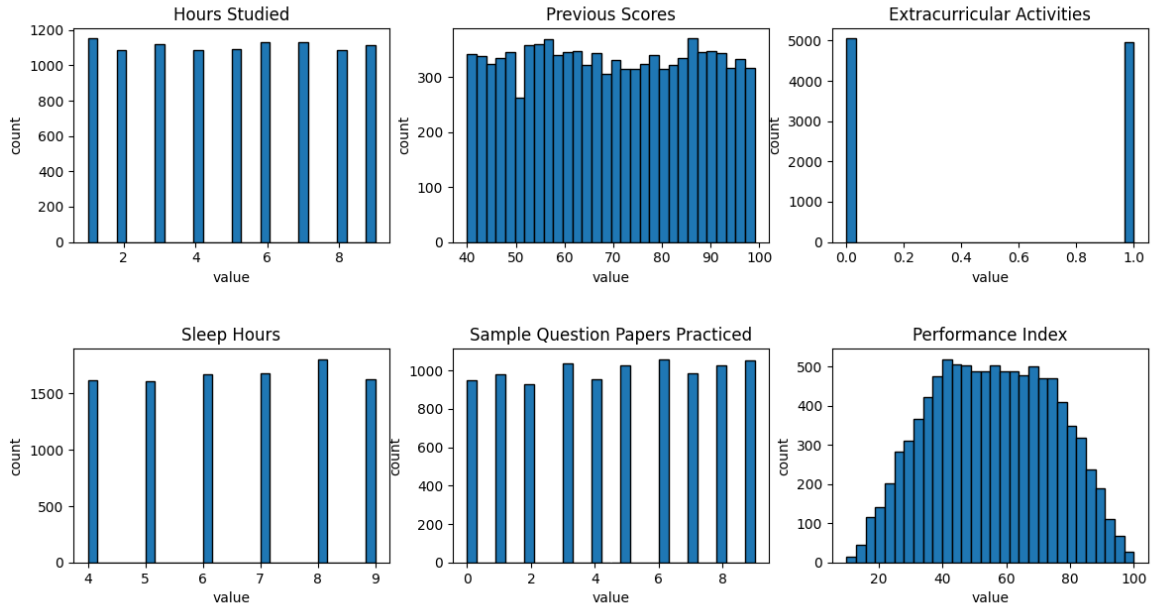
Заменим в выборке Yes/No на 1/0.

Выведем статистику:

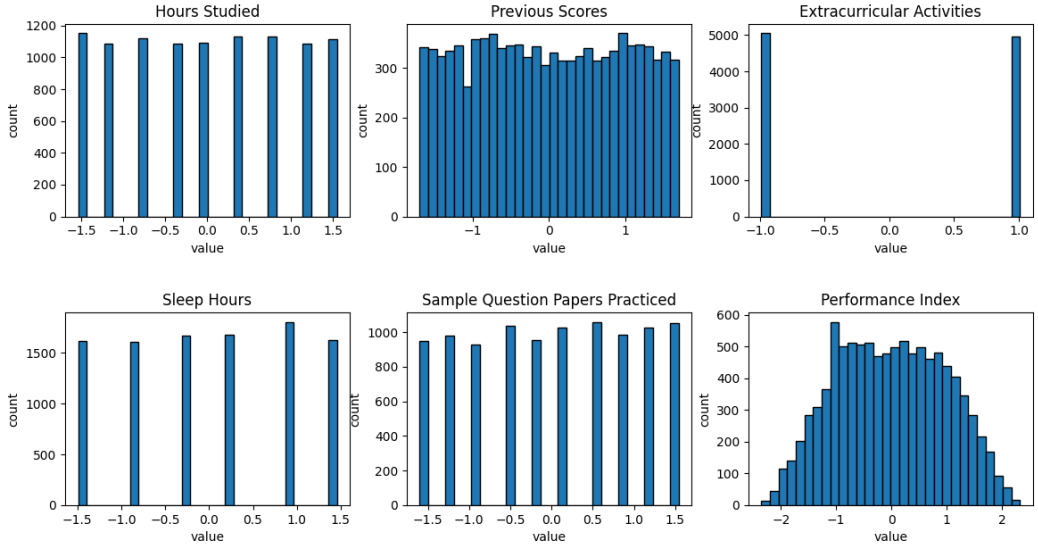
```
"/Users/bobr/PycharmProjects/Sai lab1/venv/bin/python" /Users/bobr/PycharmProjects/Sai lab1/main.py
       Hours Studied  Previous Scores  Extracurricular Activities  \
count   10000.000000     10000.000000                10000.000000
mean        4.992900        69.445700                    0.494800
std         2.589309        17.343152                    0.499998
min         1.000000        40.000000                    0.000000
25%         3.000000        54.000000                    0.000000
50%         5.000000        69.000000                    0.000000
75%         7.000000        85.000000                    1.000000
max         9.000000        99.000000                    1.000000

        Sleep Hours  Sample Question Papers Practiced  Performance Index
count  10000.000000                      10000.000000       10000.000000
mean       6.530600                          4.583300          55.224800
std        1.695863                          2.867348          19.212558
min        4.000000                          0.000000          10.000000
25%        5.000000                          2.000000          40.000000
50%        7.000000                          5.000000          55.000000
75%        8.000000                          7.000000          71.000000
max        9.000000                          9.000000         100.000000
```

Построим графики:



После нормализации:



Производим обучение моделей после разбиения на обучающую и тестовую выборки и оцениваем результат по коэффициенту детерминации:
1 модель: используем все признаки.
2 модель: для наглядности возьмем признаки, которые, на наш взгляд, не будут (или не значительно будут) коррелировать с искомой величиной (Sleep Hours, Extracurricular Activities)
3 модель: возьмем признаки, которые должны хорошо отражать искомую величину (Hours Studied, Sample Question Papers Practiced, Previous Scores)
4 модель: синтезируем новый признак, используя существующие (Hours Studied + Previous Scores * 2.5)

Результат:

```
Determination coef. (model #1) =  0.9889895699393596
Determination coef. (model #2) =  0.004826482833158452
Determination coef. (model #3) =  0.9868902473144947
Determination coef. (model #4) =  0.9857087336564307
```

Вывод

В ходе выполнения данной работы я научился строить регрессионную модель зависимости данных, а также научился синтезировать признаки для повышения производительности при работе с моделями.