Московский авиационный институт
(национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»

Кафедра 806 «Вычислительная математика и программирование»

Лабораторная работа №0 по курсу «Искусственный интеллект»

Студент: М. А. Волков
Преподаватели: Д. В. Сошников
С. Х. Ахмед
Группа: М8О-307Б-19
Дата:
Оценка:
Подпись:

Москва, 2022

Лабораторная работа №0

Задача: В данной лабораторной работе вы выступаете в роли предприимчивого начинающего стартапера в области машинного обучения. Вы заинтересовались этим направлением и хотите предложить миру что-то новое и при этом неплохо заработать. От вас требуется определить задачу которую вы хотите решить и найти под нее соответствующие данные. Так как вы не очень богаты, вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще. Вы готовитесь представить отчет ваши партнерам и спонсорам, от которых зависит дальнейшая ваша судьба. Поэтому тщательно работайте. И главное, день промедления и вас опередит ваш конкурент, да и сплагиаченная работа отразится на репутации По сути в данной лабораторной работе вы выполняете часть работы BI системы.

1 Ход работы

Я выбрал набор данных Beginner classification dataset [1] для выполнения лабораторной работы. В описании датасета предлагают самим "предстваить"целевую переменную. В качестве примера представим, что целевая переменная - успех изучения нового хобби.

Признаки в наборе данных:

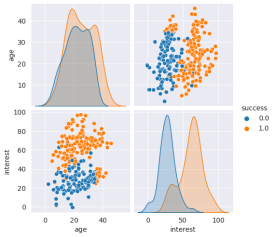
1. age — возраст человека, который решил изучить новое хобби
2. interest — показатель его заинтересованности.
3. success — успех или неуспех изучения нового хобби.

Перед выявлением зависимостей между признаками следует проверяю целостность набора данных:

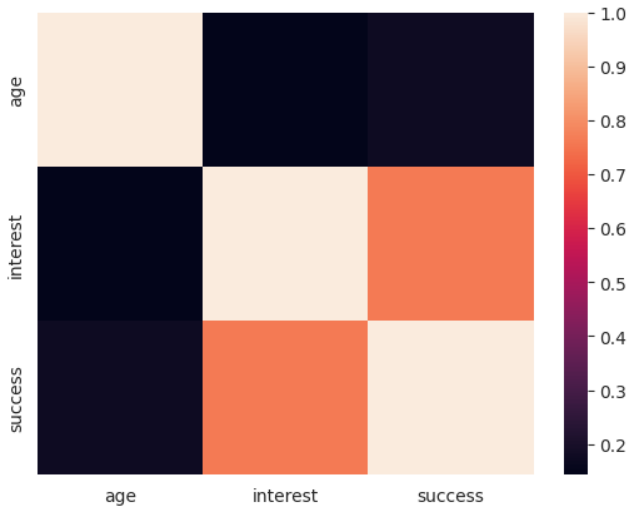
```
RangeIndex: 297 entries,0 to 296
Data columns (total 3 columns):
#   Column    Non-Null Count  Dtype
--- ------    -------------- -----
0   age       297 non-null    float64
1   interest  297 non-null    float64
2   success   297 non-null    float64
dtypes: float64(3)
memory usage: 7.1 KB
```

В наборе нет неполных данных, а все признаки - числовые.

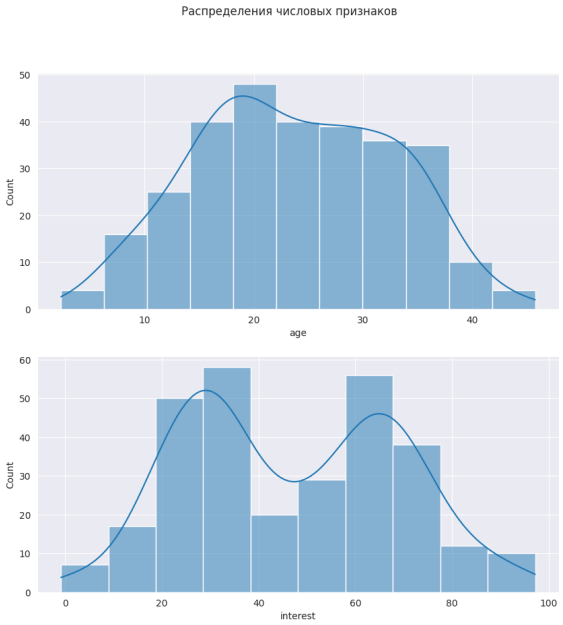
Построю графики для каждой пары признаков. Синим отмечен успех, оранжевым - неуспех:



Построю корреляционную матрицу для признаков:

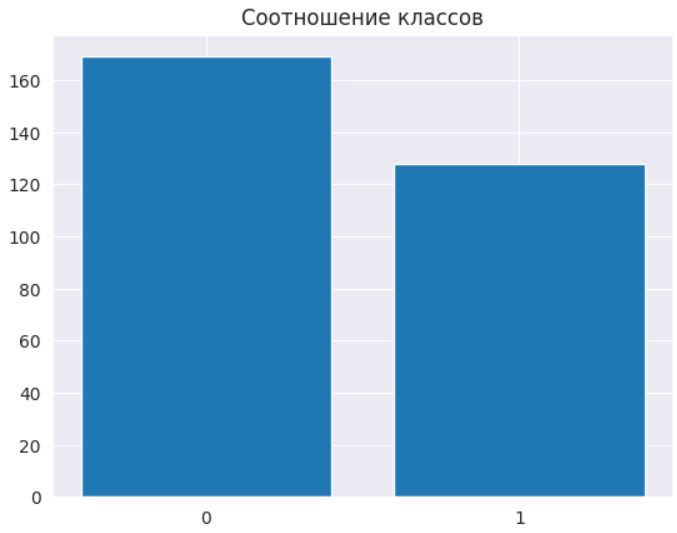


Так же построю гистограммы для числовых признаков:



Выбросов не было обнаружено, так как датасет довольно маленький.

Соотношение классов объектов:



Объектов разных классов примерно одинаковое количество, oversampling не требуется. Данные готовы к обучению.

2 Выводы

В ходе выполнения лабораторной работы я освежил в памяти курс математической статистики: гистограмму, корелляцию и корреляционную матрицу для наборов данных. Так же я изучил библиотеку Pandas, она оказалась очень удобной для анализа данных.

Трудно было найти подходящий набор данных, который подходил бы под параметры для обучения линейных моделей. В ходе своих поисков я также пробовал провести анализ и обучение на датасете для определения качества, но он почти весь сосоял из образцов одного класса - "непригодная"вода, из-за чего просто предсказание, что все образцы воды "плохие"можно добиться высокой точности. Oversampling тоже представлялся сложным занятием, так как количество признаков в датасете равнялось 20 и даже небольшое изменение одного из них приравнивалось к "плохой"воде, судя по данным из датасета.

Был проанализирован набор данных Beginner's classification dataset [1], результаты получились закономерные: успех изучения нового хобби напрямую зависимостей от возраста и заинтересованности отдельно взятого человека.

Список литературы

[1] *Beginner's Classification Dataset*
URL: `https://www.kaggle.com/datasets/sveneschlbeck/beginners-classification-datase`
(дата обращения: 30.08.2022).

[2] *Exploratory data analysis with Pandas — mlcourse.ai*
URL: `https://mlcourse.ai/book/topic01/topic01_pandas_data_analysis.html`
(дата обращения: 30.08.2022).