Московский Авиационный Институт
(Национальный Исследовательский Университет)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу «Машинное обучение»

Data Mining и исследование данных

Студент: Артамонов О. А.
Группа: М8О-308Б-19
Оценка: _____
Подпись: _____

Москва, 2022

Задание

Определить задачу, которую мы собираемся решать и найти для нее соответствующие данные. Провести анализ найденных данных.

Описание и структура датасета

Рассматривается датасет «Go To College». Задача - для каждого школьника (американского) определить вероятность того, что он продолжит обучение в колледже. Если мы будем знать, что человек не будет учиться дальше, то преподавателям и психологам следует поговорить с ним, помочь найти себя и определиться с будущими действиями. С помощью машинного обучения мы хотим выявлять учеников, которым нужна такая помощь.

Сведения, которые мы знаем о каждом ученике:

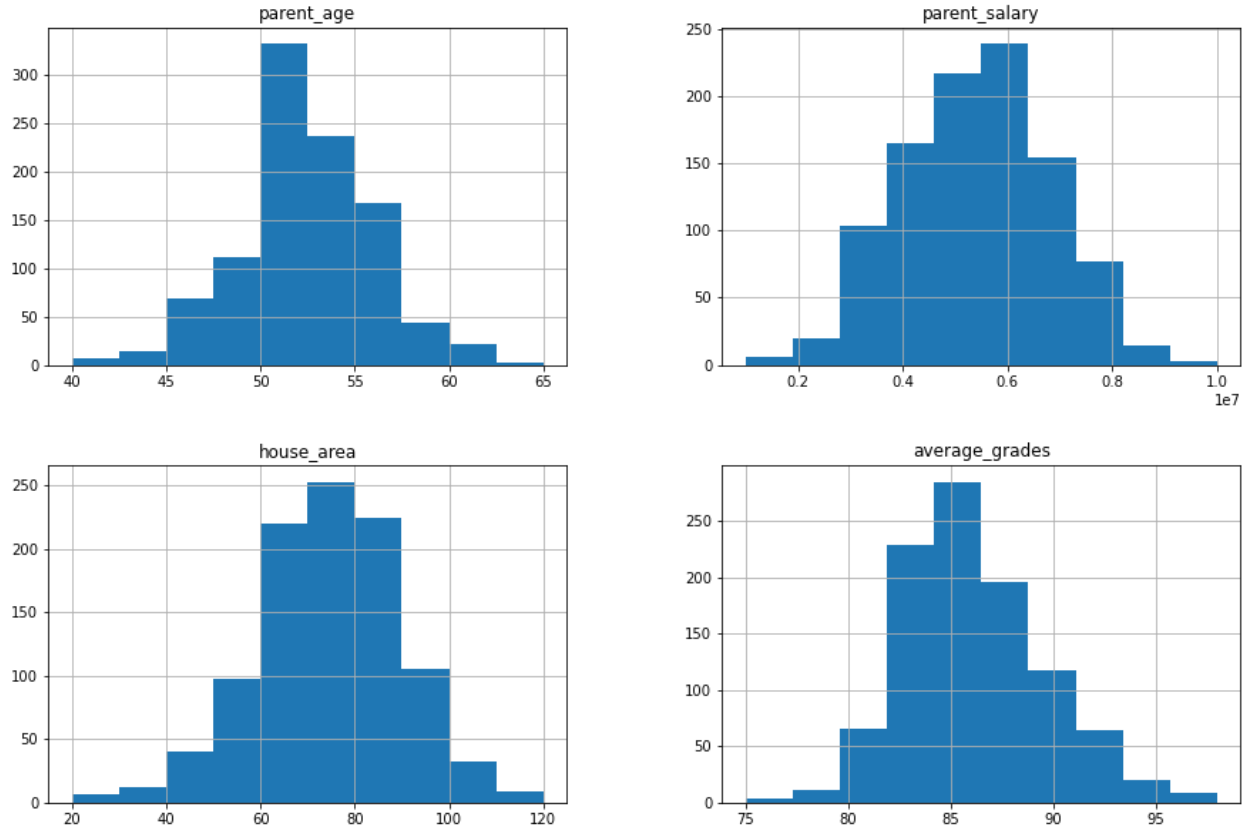
1. **type_school** - тип школы, в которую ходит ученик
2. **school_accreditation** - аккредитация школы (A / B)
3. **gender** - пол ученика
4. **interest** - заинтересованность в учебе
5. **residence** - место проживания (город / пригород)
6. **parent_age** - возраст родителей
7. **parent_salary** - зарплата родителей
8. **house_area** - площадь родительского дома
9. **average_grades** - средний балл (от 0 до 100)
10. **parent_was_in_college** - учились ли родители в колледже
11. **in_college** - пошел ли ученик в колледж - *таргет*

Как выглядит датасет:

	type_school	school_accreditation	gender	interest	residence	parent_age	parent_salary	house_area	average_grades	parent_was_in_college	in_college
0	Academic	A	Male	Less Interested	Urban	56	6950000	83.0	84.09	False	True
1	Academic	A	Male	Less Interested	Urban	57	4410000	76.8	86.91	False	True
2	Academic	B	Female	Very Interested	Urban	50	6500000	80.6	87.43	False	True
3	Vocational	B	Male	Very Interested	Rural	49	6600000	78.2	82.12	True	True
4	Academic	A	Female	Very Interested	Urban	57	5250000	75.1	86.79	False	False

Распределение количественных признаков

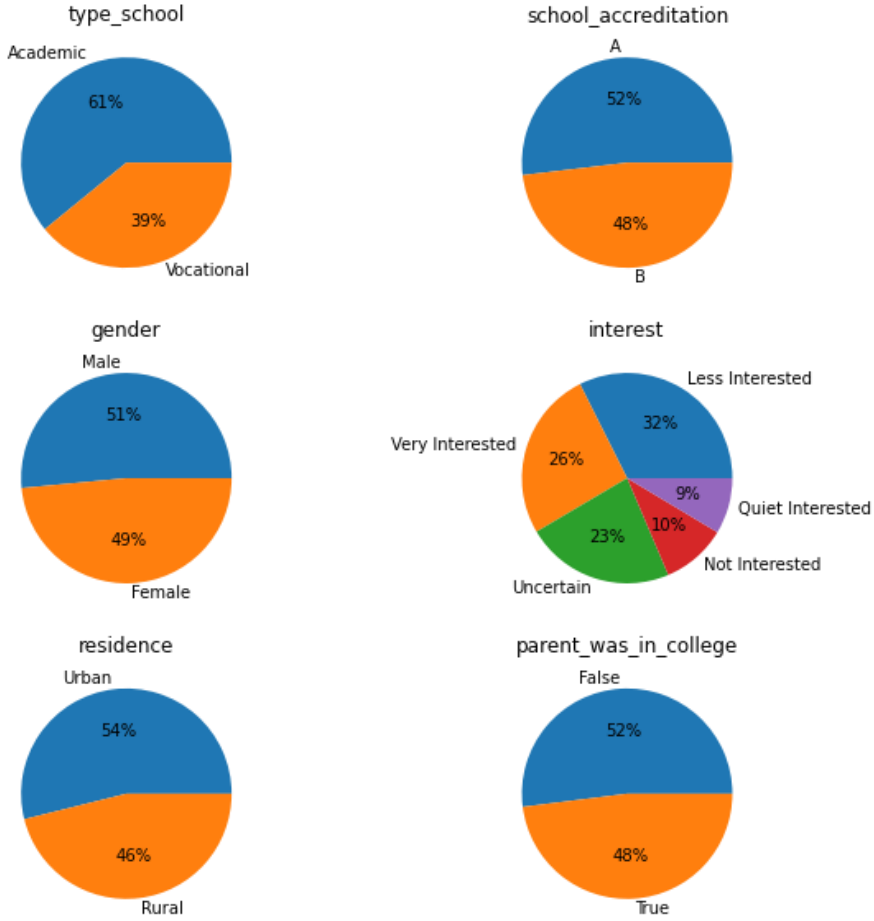
Поделим все признаки на количественные и категориальные. Изучим распределение количественных признаков.



Распределение всех количественных фичей нормальное. Когда будем обучать модели, имеет смысл привести все распределения к стандартному нормальному со средним 0 и отклонением 1.

Распределение категориальных признаков

Посмотрим на распределение категориальных фичей.

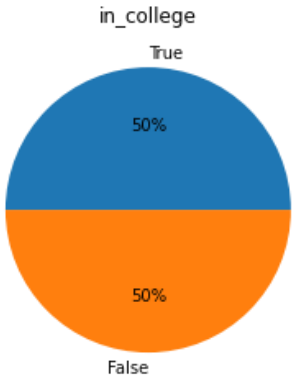


В целом, категориальные фичи распределены более-менее равномерно.

Интересно посмотреть на заинтересованность людей в учебе. Большинство учеников не сильно заинтересовано в учебе (32+10=42%).
35% учеников заинтересованы в продолжении обучения (26% - сильно заинтересованы, 9% - более-менее заинтересованы).
23% пока не определились.

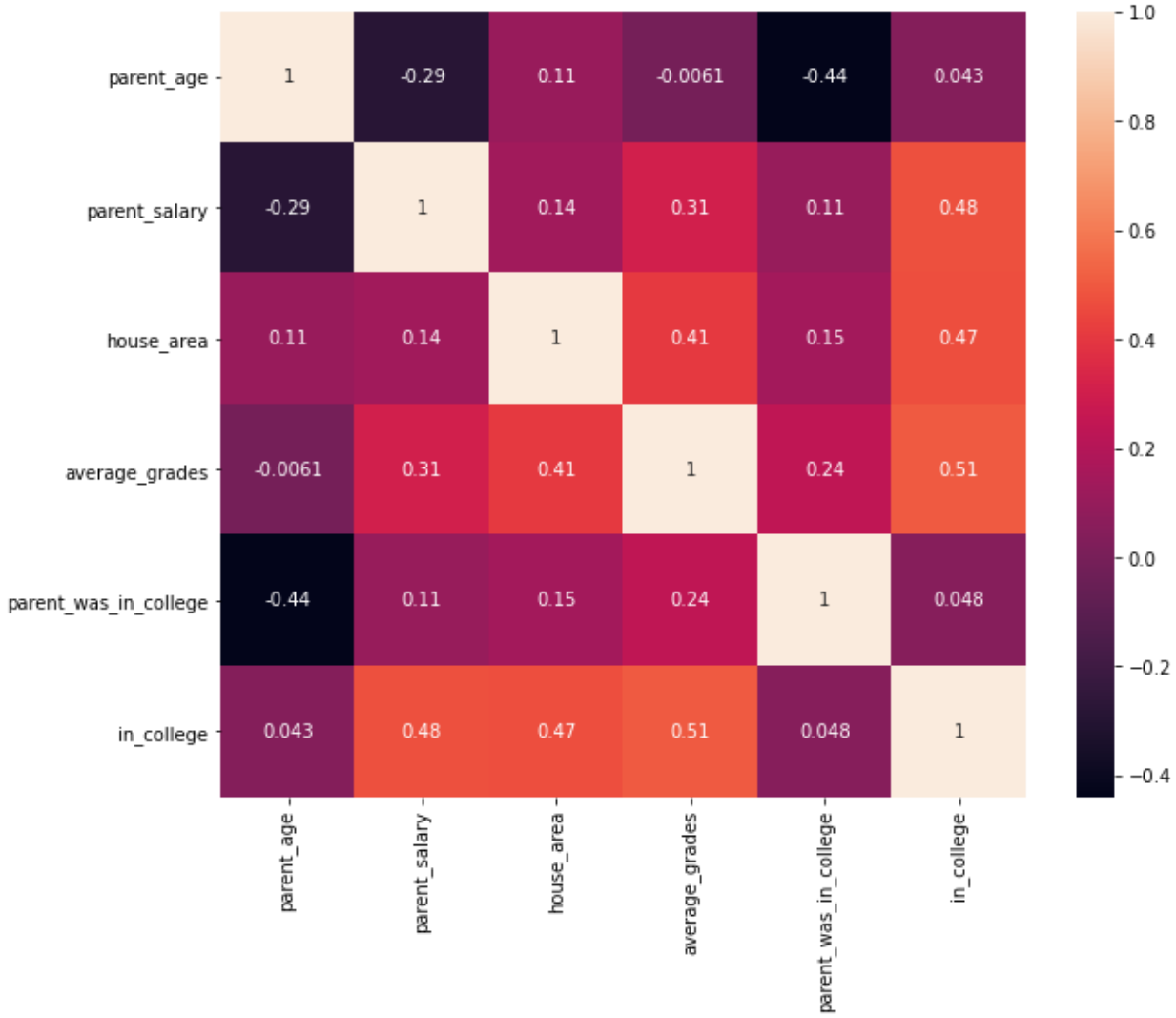
Распределение таргета и корреляция

Посмотрим на распределение таргета.



Классы идеально сбалансированы. Значит веса при обучении использовать не придется.

Построим корреляционную матрицу для численных признаков и таргета.



Зарплата родителей, площадь дома и средний балл ученика хорошо коррелируют с таргетом. В датасете нет линейно зависимых фичей, поэтому выкидывать ничего не будем.

Интересное наблюдение: сильнее всего коррелируют между собой фичи 'parent_age' и 'parent_was_in_college'. Но только на основе корреляции мы не можем строить причинно-следственные связи.

Выводы

Я рассмотрел датасет 'Go to college' и провел его анализ.

В ходе анализа получили следующие результаты:
1. В датасете нет пропущенных данных
2. Все количественные фичи имеют нормальное распределение
3. Категориальные фичи распределены более-менее равномерно

4. Классы сбалансированы
5. В датасете нет линейно зависимых фичей
6. Таргет хорошо коррелирует с несколькими количественными фичами.

Датасет готов к дальнейшей работе, приступаю к обучению моделей.