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

Лабораторная работа №0 по курсу
«Машинное обучение»

Сбор и анализ данных

Студент: Гаптулхаков Руслан Рамилевич
Группа: М80 - 308Б -19
Дата: 06.05.2022
Оценка: _____
Подпись: _____

Москва, 2022

1. Постановка задачи

Собрать или найти готовые данные и провести исследовательский анализ. Написать отчёт по результатам исследования.

2. Описание датасета

Представим, что спустя несколько сотен лет мы сможем премещаться в паралельные изме рения. Нам нужно предсказать смог ли успешно пассажир переместиться в другое измерен ие во время столкновения космического корабля с пространственно-временной аномалией. Чтобы помочь вам сделать эти прогнозы, вам дается набор личных записей, извлеченных из поврежденной компьютерной системы корабля. Этот датасет по хож на классический обучающий датасет "Титаник", но в новой тематике.

train.csv — личные записи примерно двух третей (~8700) пассажиров, которые будут использоваться в качестве обучающих данных.

- PassengerId — уникальный идентификатор для каждого пассажира. Каждый идентификатор принимает форму gggg_pp, где gggg указывает группу, с которой путешествует пассажир, а pp — его номер в группе. Люди в группе часто являются членами семьи, но не всегда.
- HomePlanet — планета, с которой вылетел пассажир, обычно планета его постоянного проживания.
- CryoSleep - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- Cabin — номер каюты, в которой находится пассажир. Принимает форму палуба/число/сторона, где сторона может быть либо P для левого борта, либо S для правого борта.
- Destination — планета, на которую будет высаживаться пассажир.
- Age - возраст пассажира.
- VIP - оплатил ли пассажир специальное VIP-обслуживание во время рейса.
- RoomService, FoodCourt, ShoppingMall, Spa, VRDeck — сумма, которую пассажир выставил в счет за каждое из многочисленных роскошных удобств космического корабля «Титаник».
- Name - имя и фамилия пассажира.
- Transported — был ли пассажир перенесен в другое измерение. Это цель, столбец, который вы пытаетесь предсказать.

test.csv — личные записи оставшейся трети (~ 4300) пассажиров, которые будут использоваться в качестве тестовых данных. Ваша задача состоит в том, чтобы предсказать значение Перевезено для пассажиров в этом наборе.

Ссылка на датасет: https://www.kaggle.com/competitions/spaceship-titanic/overview

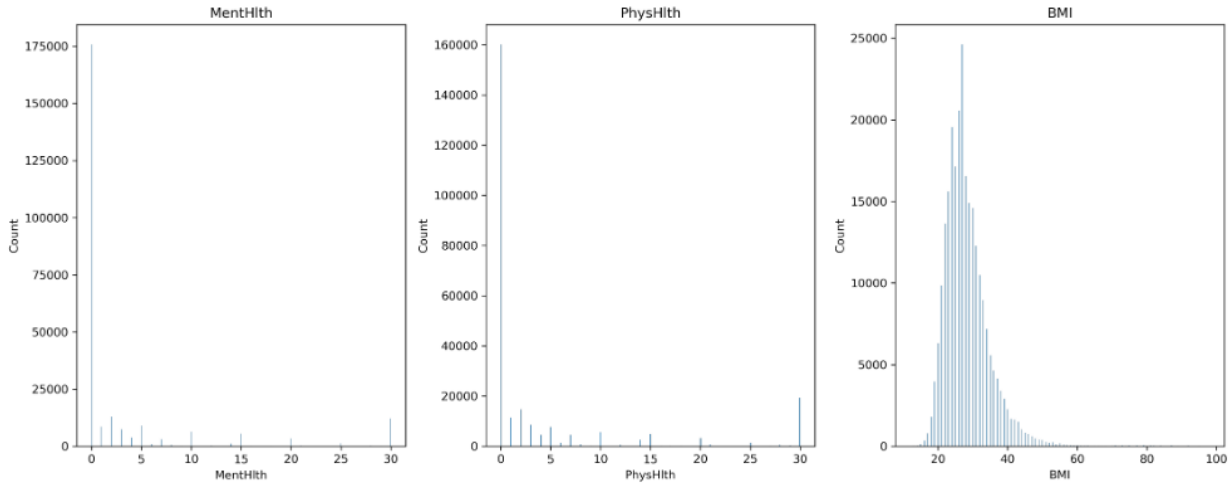
3. Количественные признаки

Имеем три количественных признака - количество ментальных и физических проблем и ИМТ.

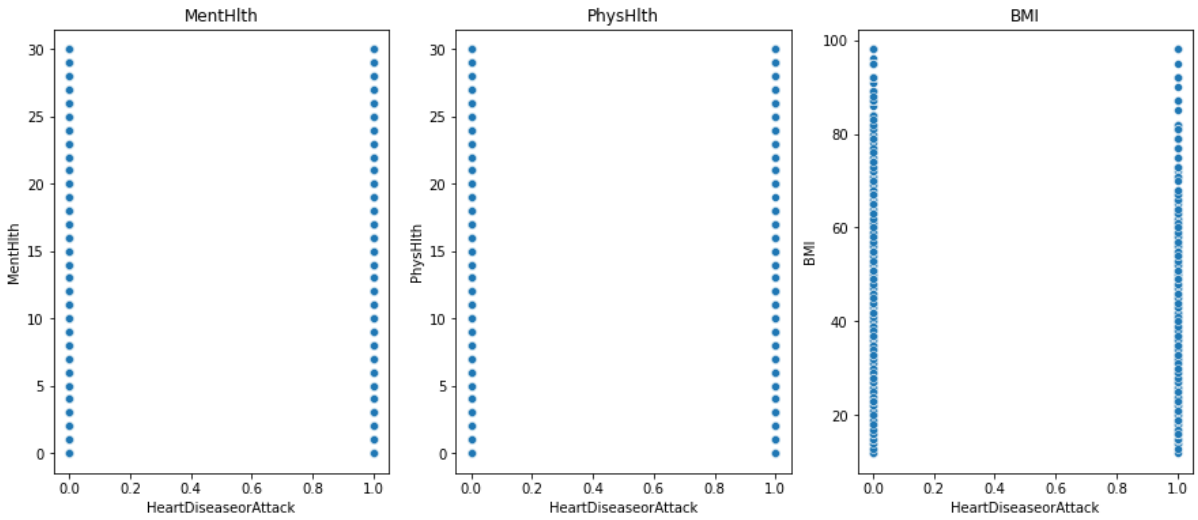
Посмотрим на статистическое описание признаков

	MentHlth	PhysHlth	BMI
count	253680.000000	253680.000000	253680.000000
mean	3.184772	4.242081	28.382364
std	7.412847	8.717951	6.608694
min	0.000000	0.000000	12.000000
25%	0.000000	0.000000	24.000000
50%	0.000000	0.000000	27.000000
75%	2.000000	3.000000	31.000000
max	30.000000	30.000000	98.000000

Посмотрим на распределение этих величин



Также построим точечный график зависимости этих величин от таргета



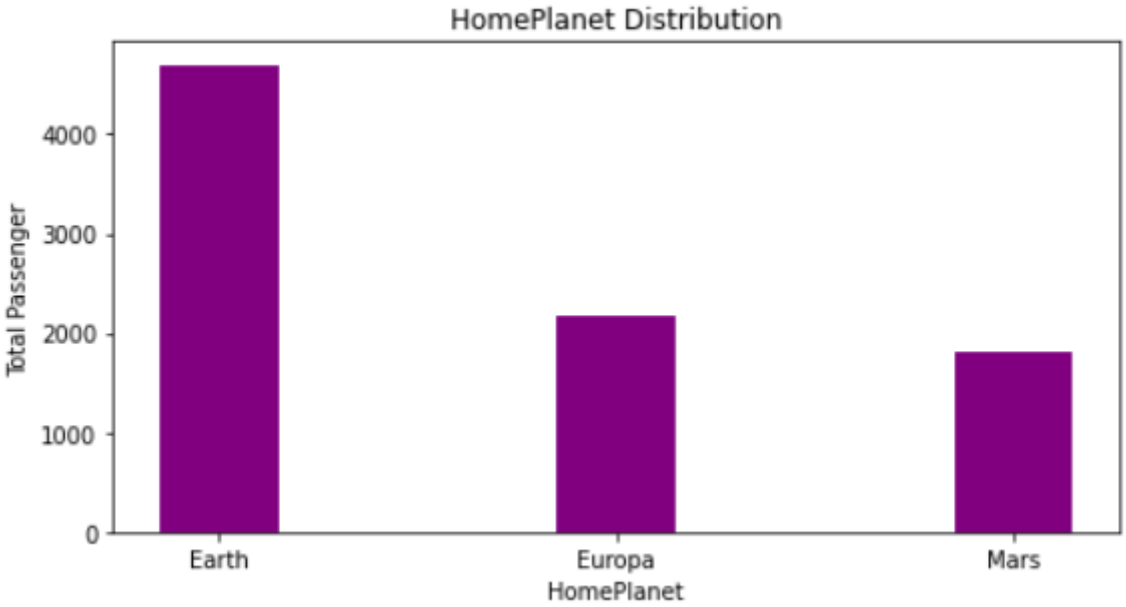
Видим, что 75-ый перцентиль у фичей с количеством ментальных и физических проблем равен 2 и 3 соответственно. Это значит, что только у 25% опрошенных наблюдалось большее количество проблем. Большая же часть людей сталкивалась с ними не часто.

По ИМТ видим, что здесь, наоборот, больше часть людей (примерно 70%) имеет ИМТ ниже среднего. Высокие значения индекса (больше 30) встречаются редко - примерно в 25% случаев. Это хорошо видно по точечному графику.

Распределение ИМТ похоже на нормальное.

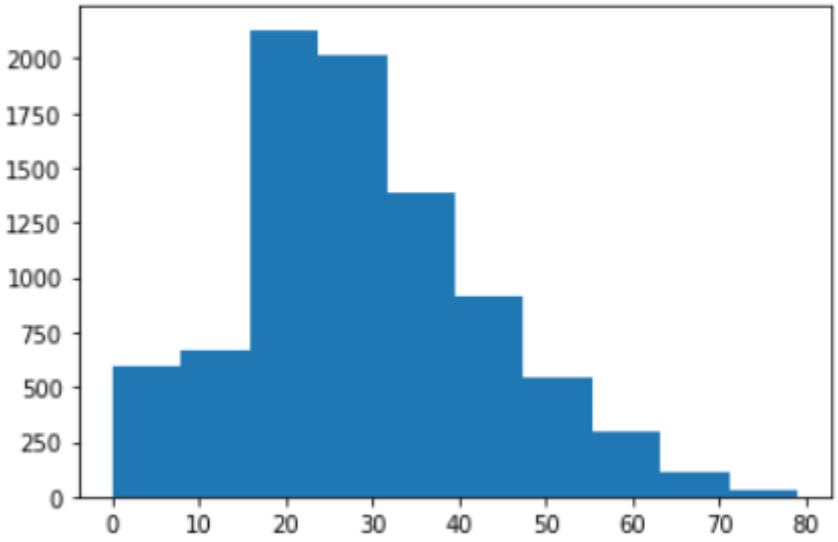
4. Категориальные признаки

Первое, что стоит сделать нужно посмотреть на колличество данных форму их представления. Затем я провери есть ли незаполненные поля в таблице и заполнил их ближайшими значениями.
Тут мы видим, на каких планетах живкт наши пассажиры:



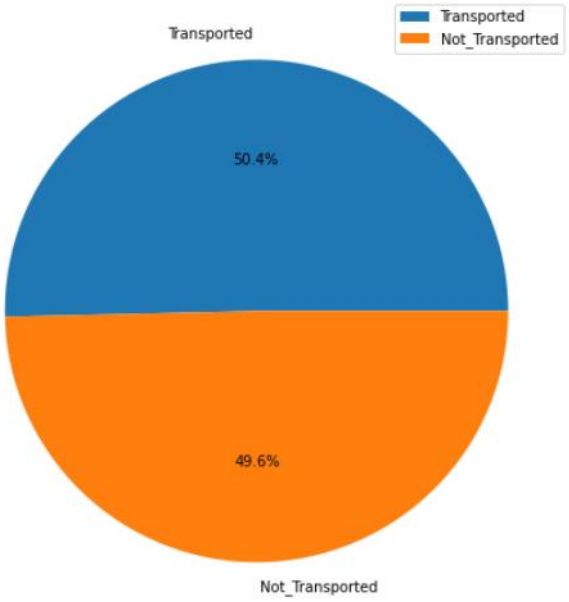
Больше всего людей зивёт на Земле.

Далее посмотри на возрастное распределение:



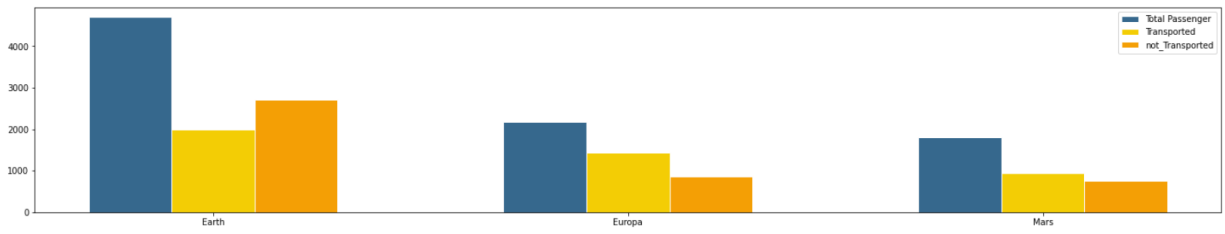
В среднем пассажиры корабля имели 20-30 лет.

Теперь посмотрим на распределение целевой пременной:

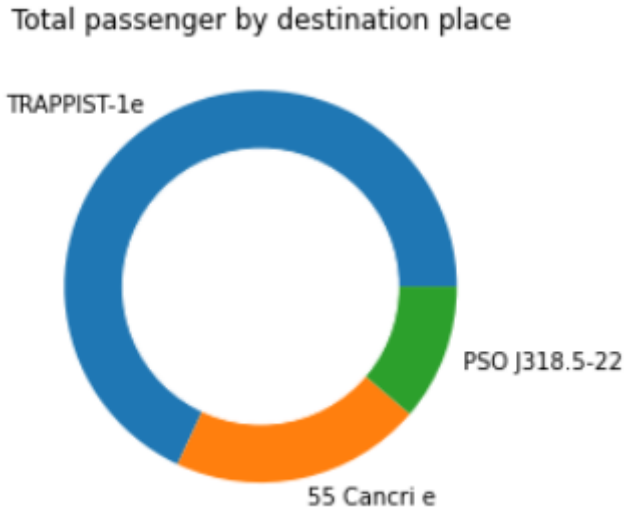


Из pie-диаграмммы видно, целевые признаки сбалансированные. Этот факт нам наруку.

Тут мы видим пропорциональное соотношение по планетам отправки между группами целевого признака. То есть сколько всего отправились с данной планеты, сколько успешно перебралось, сколько не перебралось.

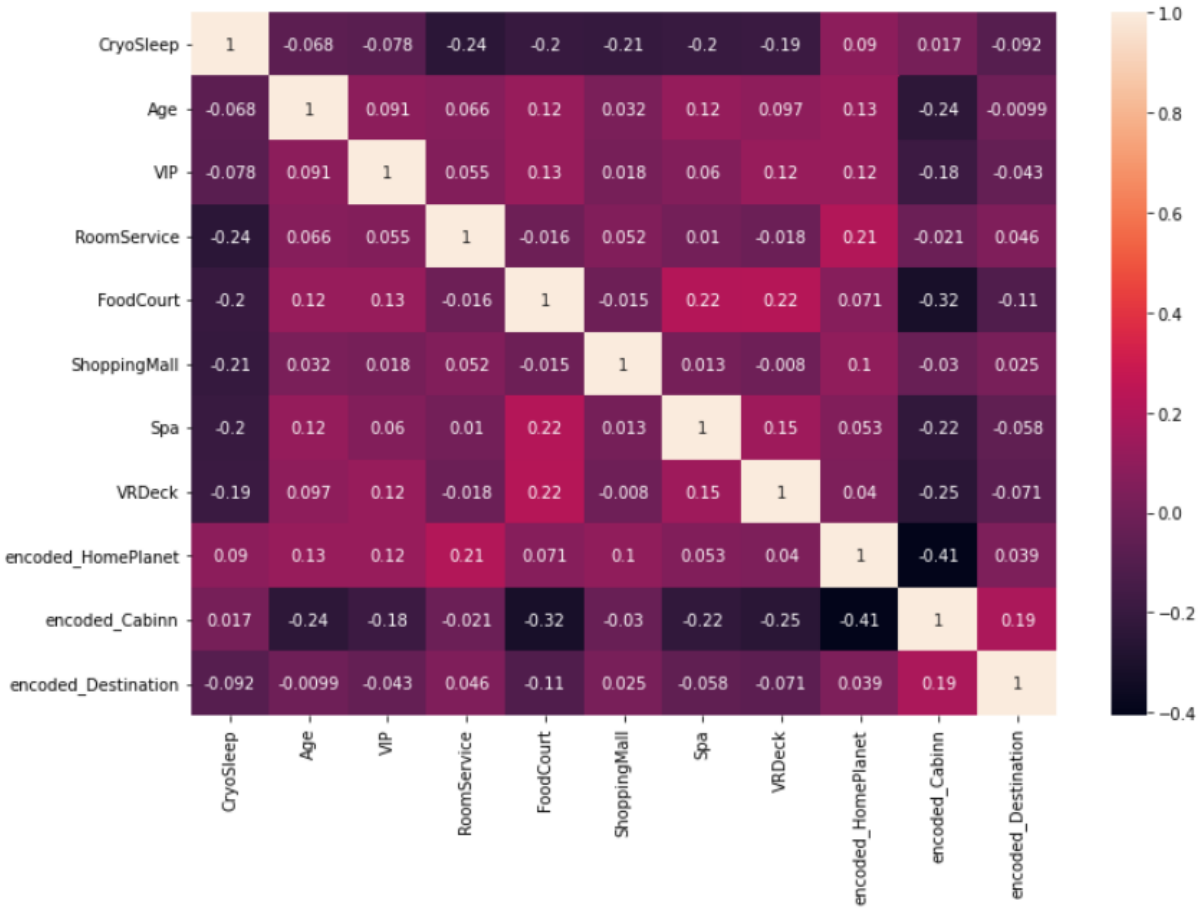


Посмотрим на распределение пассажиров по конечным пунктам назначения:



Далее мы разбили номер каюты на составляющие, чтобы больше извлечь оттуда информации. Затем перевели категориальные признаки в числовые и выкиную бесполезные столбцы.

Посмотрим на корреляционную матрицу.



Выкинем столбец encoded_Cabin так как он сильно коррелирует с encoded_HomePlanet.

5. Вывод

В данной лабораторной работе я исследовал выбранный датасет. В процессе работы были интуитивно заполнены недостающие данные, проанализированны распределения признаков с визуализацией, извлечены дополнительные признаки из уже имеющихся, переведены все признаки в числа.

В конце, я убедился, что целевая переменная зависит от имеющихся признаков. Следовательно, у нас есть все шансы получить хорошую модель.