Домашнее задание 0.

Дедлайн: суббота 18.12.2021, 23.59
Фидбек от меня: в течение 19 декабря 2021

Что нужно сделать

Нужно собрать датасет, на котором вы будете практиковаться в машинном обучении.
Для первого блока курса нужны табличные данные.

Требования:

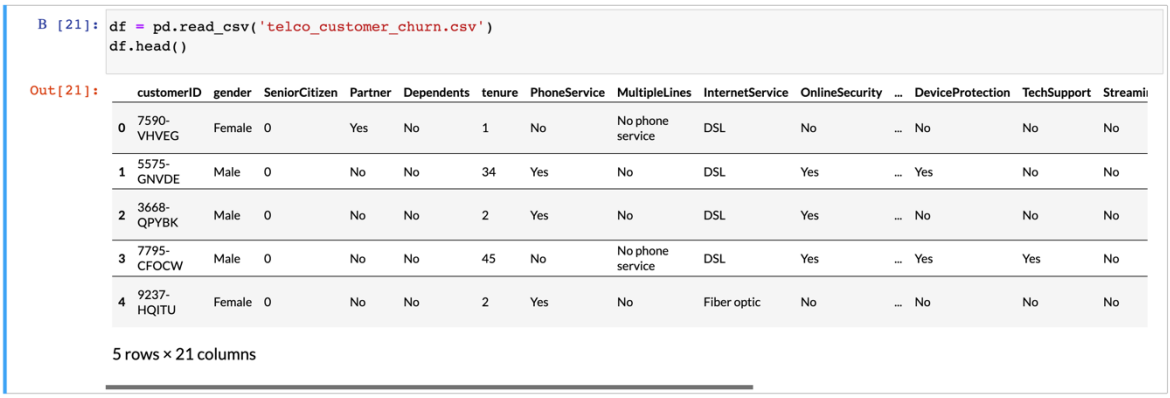
- 500-1000 строк (больше – лучше)
- каждая строка – это объект, для которого делается прогноз
- для каждой строки известно значение целевой переменной (т.е. есть столбец с таргетом)
- в датасете есть 2-3 категориальных переменных минимум
- в датасете есть 2-3

Пример 1:

Датасет по оттоку телеком-компании.

В нем 7043 строки, одна строка – запись об уникальном клиенте (клиент – объект прогноза), для него прогнозируем отток (в датасете есть колонка Churn), разметка по оттоку есть в датасете.

Это задача классификации. Выбирая такой датасет, я буду практиковаться в применении алгоритмов классификации.



```
In [21]: df = pd.read_csv('telco_customer_churn.csv')
         df.head()
```

Out[21]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	Streami
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	No

5 rows × 21 columns

Пример 2:

Датасет по магазинам Walmart.
В нем 421000 строк, одна строка – запись о департаменте в магазине в неделю. Для каждого департамента-магазино-недели известны продажи (целевая переменная), которую я буду алгоритмом прогнозировать.

Это задача регрессии. Выбирая такой датасет, я буду практиковаться в применении алгоритмов регрессии.

```
B [16]: data = pd.read_csv('walmart.csv', sep='\t')
        data.head()
```

Out[16]:

	Store	Date	Dept	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	U
0	1	2010-02-05	1	24924.50	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.
1	1	2010-02-05	2	50605.27	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.
2	1	2010-02-05	3	13740.12	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.
3	1	2010-02-05	4	39954.04	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.
4	1	2010-02-05	5	32229.38	False	A	151315	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.

Комментарии

- Датасет может быть для классификации или регрессии – ограничений нет. Главное, наличие целевой переменной
- Приветствуются датасеты практические, собранные из данных компании.
- Но датасеты могут быть и из открытых источников – выбор за вами