Рубежный контроль №1 ИУ5-62Б Заузолков

Денис Вариант №10

Оглавление

1) Формулировка задачи:

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Подключение библиотек

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

plt.style.use('fivethirtyeight')
%matplotlib inline
```

Загрузка данных в Python

In [2]:

```
data = pd.read_csv('rk №1/dc-wikia-data.csv', low_memory=False)
```

2) Основные характеристики датасета

In [3]:

```
data.head()
```

Out[3]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	GSM	ALIVE	APPEARANCES	APPE
0	1422	Batman (Bruce Wayne)	\/wiki\/Batman_(Bruce_Wayne)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	3093.0	1
1	23387	Superman (Clark Kent)	\/wiki\/Superman_(Clark_Kent)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	2496.0	1986
2	1458	Green Lantern (Hal Jordan)	\/wiki\/Green_Lantern_(Hal_Jordan)	Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	NaN	Living Characters	1565.0	1959
3	1659	James Gordon (New Earth)	\/wiki\/James_Gordon_(New_Earth)	Public Identity	Good Characters	Brown Eyes	White Hair	Male Characters	NaN	Living Characters	1316.0	1987,
4	1576	Richard Grayson (New Earth)	\/wiki\/Richard_Grayson_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NaN	Living Characters	1237.0	1

In [4]:

```
data.shape
```

Out[4]:

```
(6896, 13)
```

In [5]:

```
data.dtypes
```

Out[5]:

```
page_id                int64
name                  object
urlslug               object
ID                    object
ALIGN                 object
EYE                   object
HAIR                  object
SEX                   object
GSM                   object
ALIVE                 object
APPEARANCES          float64
FIRST APPEARANCE      object
YEAR                 float64
dtype: object
```

In [6]:

```
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
page_id - 0
name - 0
urlslug - 0
ID - 2013
ALIGN - 601
EYE - 3628
HAIR - 2274
SEX - 125
GSM - 6832
ALIVE - 3
APPEARANCES - 355
FIRST APPEARANCE - 69
YEAR - 69
```

In [7]:

```
data.describe()
```

Out[7]:

	page_id	APPEARANCES	YEAR
count	6896.000000	6541.000000	6827.000000
mean	147441.209252	23.625134	1989.766662
std	108388.631149	87.378509	16.824194
min	1380.000000	1.000000	1935.000000
25%	44105.500000	2.000000	1983.000000
50%	141267.000000	6.000000	1992.000000
75%	213203.000000	15.000000	2003.000000
max	404010.000000	3093.000000	2013.000000

3) Обработка пропусков в данных

In [8]:

```
data.drop(data[data['APPEARANCES'].isnull()].index, inplace=True)
```

In [9]:

```
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
page_id - 0
name - 0
```

```
urlslug - 0
ID - 1883
ALIGN - 566
EYE - 3426
HAIR - 2093
SEX - 114
GSM - 6477
ALIVE - 2
APPEARANCES  -  0
FIRST APPEARANCE - 60
YEAR - 60
```

In [10]:

```
from sklearn.impute import SimpleImputer
# Импьютация константой
imp = SimpleImputer(missing_values=np.nan, strategy='constant', fill_value='NA')
data_imp = imp.fit_transform(data)
data_imp
```

Out[10]:

```
array([[1422, 'Batman (Bruce Wayne)', '\\/wiki\\/Batman_(Bruce_Wayne)',
        ..., 3093.0, '1939, May', 1939.0],
       [23387, 'Superman (Clark Kent)',
        '\\/wiki\\/Superman_(Clark_Kent)', ..., 2496.0, '1986, October',
        1986.0],
       [1458, 'Green Lantern (Hal Jordan)',
        '\\/wiki\\/Green_Lantern_(Hal_Jordan)', ..., 1565.0,
        '1959, October', 1959.0],
       ...,
       [345590, 'Apollo (Roman God) (New Earth)',
        '\\/wiki\\/Apollo_(Roman_God)_(New_Earth)', ..., 1.0, 'NA', 'NA'],
       [15050, 'Ben Lo (New Earth)', '\\/wiki\\/Ben_Lo_(New_Earth)', ...,
        1.0, 'NA', 'NA'],
       [205584, 'Auctioneer II (New Earth)',
        '\\/wiki\\/Auctioneer_II_(New_Earth)', ..., 1.0, 'NA', 'NA']],
      dtype=object)
```

In [11]:

```
data_imp.shape
```

Out[11]:

```
(6541, 13)
```

In [12]:

```
result = pd.DataFrame(data_imp,columns = ['page_id','name','urlslug','ID','ALIGN','EYE','HAIR','SEX','GSM','ALIV
result
```

Out[12]:

	page_id	name	urlslug	ID	ALIGN	EYE	HAIR	SEX	GSM	ALIVE	APPEARANCES
0	1422	Batman (Bruce Wayne)	\/wiki\/Batman_(Bruce_Wayne)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NA	Living Characters	3093.0
1	23387	Superman (Clark Kent)	\/wiki\/Superman_(Clark_Kent)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NA	Living Characters	2496.0
2	1458	Green Lantern (Hal Jordan)	\/wiki\/Green_Lantern_(Hal_Jordan)	Secret Identity	Good Characters	Brown Eyes	Brown Hair	Male Characters	NA	Living Characters	1565.0
3	1659	James Gordon (New Earth)	\/wiki\/James_Gordon_(New_Earth)	Public Identity	Good Characters	Brown Eyes	White Hair	Male Characters	NA	Living Characters	1316.0
4	1576	Richard Grayson (New Earth)	\/wiki\/Richard_Grayson_(New_Earth)	Secret Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NA	Living Characters	1237.0
...	...	...	...	...	...	...	...	...	...	...	...
6536	16094	Mark Antaeus (New Earth)	\/wiki\/Mark_Antaeus_(New_Earth)	Public Identity	Good Characters	Blue Eyes	Black Hair	Male Characters	NA	Deceased Characters	1.0
6537	128000	Jerome Cox (New Earth)	\/wiki\/Jerome_Cox_(New_Earth)	Public Identity	Bad Characters	NA	NA	Male Characters	NA	Living Characters	1.0
6538	345590	Apollo (Roman God) (New Earth)	\/wiki\/Apollo_(Roman_God)_(New_Earth)	NA	Good Characters	NA	NA	Male Characters	NA	Living Characters	1.0

6539	15050	Ben Lo (New Earth)	\/wiki\/Ben_Lo_(New_Earth)	Public Identity	Good Characters	Brown Eyes	Black Hair	Male Characters	NA	Living Characters	1.0
6540	205584	Auctioneer II (New Earth)	\/wiki\/Auctioneer_II_(New_Earth)	Secret Identity	Bad Characters	NA	White Hair	Male Characters	NA	Living Characters	1.0

6541 rows × 13 columns

Для дальнейшего построения моделей машинного обучения следует использовать количественный признак "APPEARANCES" вместе с категориальными признаками, у которых несколько уникальных значений ('ID','ALIGN','EYE','HAIR','SEX','ALIVE')

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js