Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

Работа №2

по курсу «Технологии машинного обучения»

Выполнил

студент группы ИУ5-64Б

Шпак И.Д.

Москва, 2021

1 Исходное задание

1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
   - обработку пропусков в данных; кодирование категориальных признаков;
   - масштабирование данных.

2 Код программы

[6]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

[7]:
```
data = pd.read_csv('~/Downloads/wiki_movie_plots_deduped.csv', sep=",")
data.head()
```

[7]:

	Release Year	Title	Origin/Ethnicity	Director	Cast	Genre	Wiki Page
0	1901	Kansas Saloon Smashers	American	Unknown	NaN	unknown	https://en.wikipedia.org/wiki/Kansas_Saloon_Sm…
1	1901	Love by the Light of the Moon	American	Unknown	NaN	unknown	https://en.wikipedia.org/wiki/Love_by_the_Ligh…
2	1901	The Martyred Presidents	American	Unknown	NaN	unknown	https://en.wikipedia.org/wiki/The_Martyred_Pre…
3	1901	Terrible Teddy, the Grizzly King	American	Unknown	NaN	unknown	
4	1902	Jack and the Beanstalk	American	George S. Fleming, Edwin S. Porter	NaN	unknown	

```
3  https://en.wikipedia.org/wiki/Terrible_Teddy,_…
4  https://en.wikipedia.org/wiki/Jack_and_the_Bea…
```

```
                                                 Plot
0  A bartender is working at a saloon, serving dr…
1  The moon, painted with a smiling face hangs ov…
2  The film, just over a minute long, is composed…
3  Lasting just 61 seconds and consisting of two …
4  The earliest known adaptation of the classic f…
```

[8]:
```
data.shape
```

[8]:
```
(34886, 8)
```

[9]:
```
data.dtypes
```

[9]:
```
Release Year         int64
Title               object
Origin/Ethnicity    object
Director            object
Cast                object
Genre               object
Wiki Page           object
Plot                object
dtype: object
```

[10]:
```
data.isnull().sum()
```

[10]:
```
Release Year           0
Title                  0
Origin/Ethnicity       0
Director               0
Cast                1422
Genre                  0
Wiki Page              0
Plot                   0
dtype: int64
```

[13]:
```
data_del0_1 = data.dropna(axis = 1 , how = "any")
data.shape,data_del0_1.shape
```

[13]:
```
((34886, 8), (34886, 7))
```

[14]:
```
data_del0_2 = data.dropna(axis = 0 , how = "any")
data.shape,data_del0_2.shape
```

[14]:
```
((34886, 8), (33464, 8))
```

[15]:
```
data_new_3 = data.fillna("Unknown")
```

[16]:
```
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
num_cols = []
total_count = data.shape[0]
for col in data.columns:
    #
    temp_null_count = data[data[col].isnull()].shape[0]
    dt = str(data[col].dtype)
    if temp_null_count>0 and (dt=='object'):
        num_cols.append(col)
        temp_perc = round((temp_null_count / total_count) * 100.0, 2)
        print('     {}.        {}.                {}, {}%.'.format(col, dt,␣
↪temp_null_count, temp_perc))
```

```
    Cast.        object.                1422, 4.08%.
```

[17]:
```
cat_temp_data = data[['Cast']]
cat_temp_data['Cast'].unique()
```

[17]:
```
array([nan, 'May Clark', 'William Craven, Florence Lawrence', …,
       'Ata Demirer, Tuvana Türkay, Ülkü Duru',
       'YouTubers Shanna Malcolm, Shira Lazar, Sara Fletcher and Ashley
Clements',
       'Halit Ergenç, Tuba Büyüküstün, Mehmet Günsür, Nejat İşler'],
      dtype=object)
```

[18]:
```
cat_temp_data.head()
```

[18]:
```
  Cast
0  NaN
1  NaN
2  NaN
3  NaN
```

```
4  NaN
```

[19]:
```
cat_temp_data[cat_temp_data['Cast'].isnull()].shape
```

[19]:
```
(1422, 1)
```

[21]:
```
imp = SimpleImputer(missing_values=np.nan, strategy='constant',fill_value='NA')
data_imp = imp.fit_transform(cat_temp_data)
data_imp
```

[21]:
```
array([['NA'],
       ['NA'],
       ['NA'],
       …,
       ['Ata Demirer, Tuvana Türkay, Ülkü Duru'],
       ['YouTubers Shanna Malcolm, Shira Lazar, Sara Fletcher and Ashley
Clements'],
       ['Halit Ergenç, Tuba Büyüküstün, Mehmet Günsür, Nejat İşler']],
      dtype=object)
```

[22]:
```
data_imp[data_imp=='NA'].size
```

[22]:
```
1422
```

[23]:
```
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
cat_enc = pd.DataFrame({'c1':data_imp.T[0]})
cat_enc
```

[23]:

	c1
0	NA
1	NA
2	NA
3	NA
4	NA
…	…
34881	Director: Russell Crowe\r\nCast: Russell Crowe…
34882	Ahmet Kural, Murat Cemcir
34883	Ata Demirer, Tuvana Türkay, Ülkü Duru
34884	YouTubers Shanna Malcolm, Shira Lazar, Sara Fl…
34885	Halit Ergenç, Tuba Büyüküstün, Mehmet Günsür, …

```
[34886 rows x 1 columns]
```

[24]:
```
le = LabelEncoder()
cat_enc_le = le.fit_transform(cat_enc['c1'])
cat_enc['c1'].unique()
```

[24]:
```
array(['NA', 'May Clark', 'William Craven, Florence Lawrence', …,
       'Ata Demirer, Tuvana Türkay, Ülkü Duru',
       'YouTubers Shanna Malcolm, Shira Lazar, Sara Fletcher and Ashley
Clements',
       'Halit Ergenç, Tuba Büyüküstün, Mehmet Günsür, Nejat İşler'],
      dtype=object)
```

[25]:
```
np.unique(cat_enc_le)
```

[25]:
```
array([    0,     1,     2, …, 32180, 32181, 32182])
```

[26]:
```
le.inverse_transform([0, 1, 2, 3])
```

[26]:
```
array(['"Manamantha"', "'Fatty' Arbuckle / Buster Keaton",
       "'Fatty' Arbuckle, Al St. John", "'Fatty' Arbuckle, Buster Keaton"],
      dtype=object)
```

[27]:
```
ohe = OneHotEncoder()
cat_enc_ohe = ohe.fit_transform(cat_enc[['c1']])
cat_enc.shape
```

[27]:
```
(34886, 1)
```

[28]:
```
cat_enc_ohe
```

[28]:
```
<34886x32183 sparse matrix of type '<class 'numpy.float64'>'
        with 34886 stored elements in Compressed Sparse Row format>
```

[29]:
```
cat_enc_ohe.todense()[0:10]
```

[29]:
```
matrix([[0., 0., 0., …, 0., 0., 0.],
        [0., 0., 0., …, 0., 0., 0.],
        [0., 0., 0., …, 0., 0., 0.],
        …,
        [0., 0., 0., …, 0., 0., 0.],
        [0., 0., 0., …, 0., 0., 0.],
        [0., 0., 0., …, 0., 0., 0.]])
```

[30]:
```
pd.get_dummies(cat_temp_data, dummy_na=True).head()
```

[30]:

	Cast_"Manamantha"	Cast_'Fatty' Arbuckle / Buster Keaton	\
0	0	0	
1	0	0	
2	0	0	
3	0	0	
4	0	0	

	Cast_'Fatty' Arbuckle, Al St. John	Cast_'Fatty' Arbuckle, Buster Keaton	\
0	0	0	
1	0	0	
2	0	0	
3	0	0	
4	0	0	

	Cast_'Fatty' Arbuckle/Buster Keaton	\
0	0	
1	0	
2	0	
3	0	
4	0	

	Cast_(Korean dubbed) Kim Il, Choi Jeong-ho	\
0	0	
1	0	
2	0	
3	0	
4	0	

	Cast_(Korean dubbed) Lee Taemin, Sunny	\
0	0	
1	0	
2	0	
3	0	
4	0	

	Cast_(voices of) Judy Garland, Robert Goulet, Red Buttons	\
0	0	

```
1                                                   0
2                                                   0
3                                                   0
4                                                   0

   Cast_(voices of) Kelsey Grammer, Ian Holm, Paul Scofield, Patrick Stewart,
Julia Ormond, Peter Ustinov  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0

   Cast_(voices) Jonathan Taylor Thomas, Matthew Broderick, Jeremy Irons, James
Earl Jones, Whoopi Goldberg, Moira Kelly, Nathan Lane  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0

   …  Cast_Öner Erkan, Kadir Çermik & Damla Sönmez  \
0  …                                             0
1  …                                             0
2  …                                             0
3  …                                             0
4  …                                             0

   Cast_İsmail Hacıoğlu, Erkan Can & Uğur Polat  \
0                                               0
1                                               0
2                                               0
3                                               0
4                                               0

   Cast_Şafak Sezer, Alp Kırşan & Ahmet Mümtaz Taylan  \
0                                                   0
1                                                   0
```

```
2                                                    0
3                                                    0
4                                                    0

   Cast_Şahan Gökbakar, Gülsen Özbakan & Efe Babacan  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0

   Cast_Şahan Gökbakar, Zeynep Çamcı & Emirhan Çelik  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0

   Cast_Şahin K, Nuri Alço & Coşkun Göğen  \
0                                       0
1                                       0
2                                       0
3                                       0
4                                       0

   Cast_Şerif Sezer, Mark Dacascos & Zeynep Beşerler  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0

   Cast_Şevket Emrulla, Nilüfer Açıkalın & İlker İnanoğlu  \
0                                                        0
1                                                        0
2                                                        0
3                                                        0
4                                                        0
```

	Cast_Şeyma Uzunlar, Vahide Gördüm & Sevinç Baş	Cast_nan
0	0	1
1	0	1
2	0	1
3	0	1
4	0	1

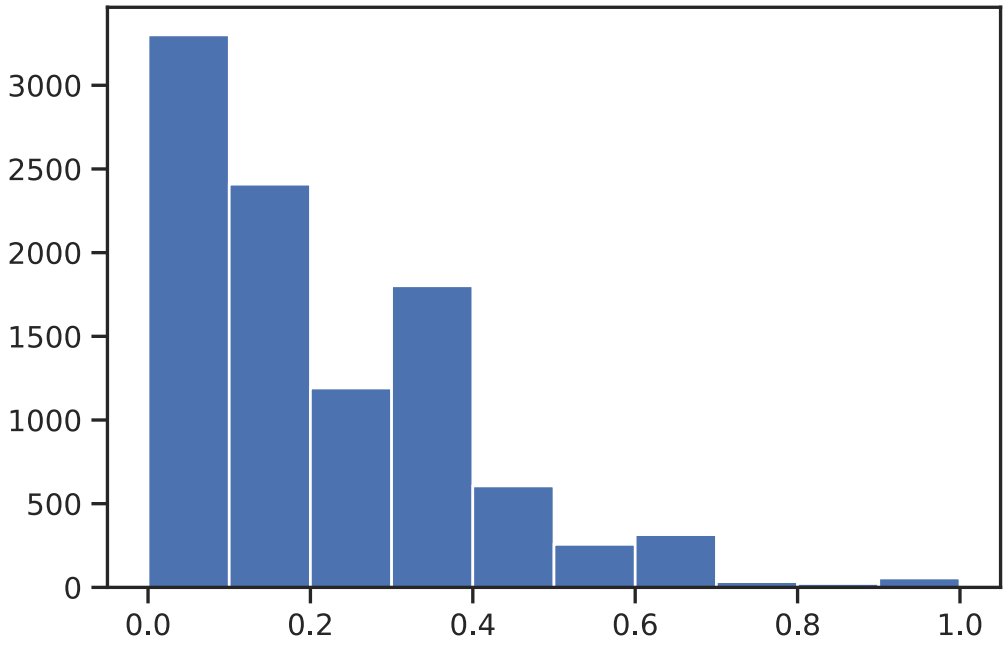
[5 rows x 32183 columns]

[31]:
```
from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer


data = pd.read_csv("~/Downloads/SampleSuperstore.csv",sep = ",")
data.dtypes
```

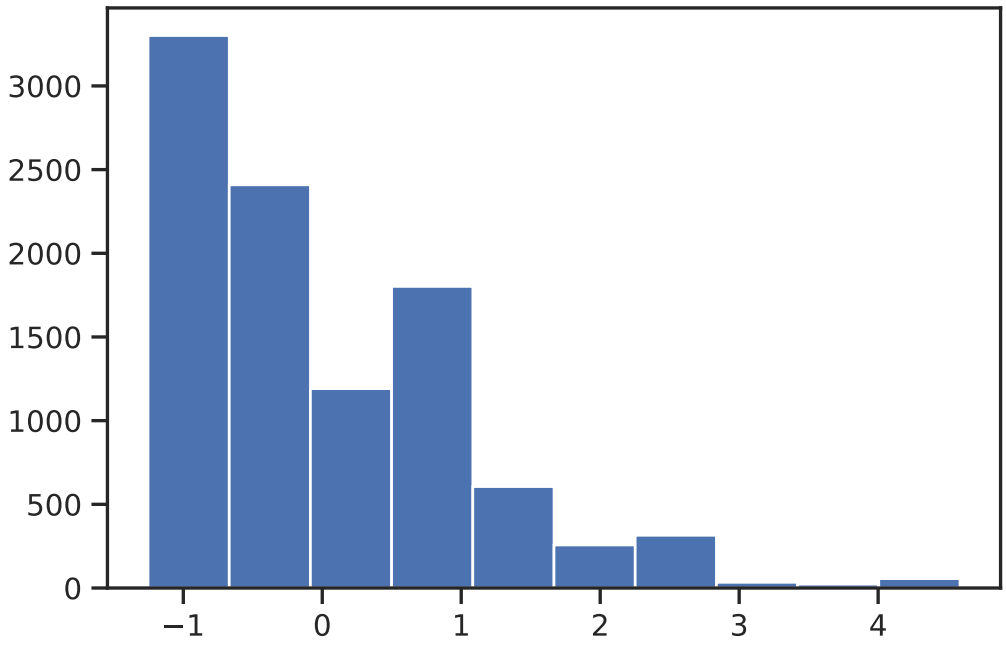
[31]:
```
Ship Mode        object
Segment          object
Country          object
City             object
State            object
Postal Code       int64
Region           object
Category         object
Sub-Category     object
Sales           float64
Quantity          int64
Discount        float64
Profit          float64
dtype: object
```

[74]:
```
sc1 = MinMaxScaler()
sc1_data = sc1.fit_transform(data[['Quantity']])
plt.hist(sc1_data)
plt.show()
```



[77]:
```
sc2 = StandardScaler()
sc2_data = sc2.fit_transform(data[['Quantity']])
```

[78]:
```
plt.hist(sc2_data)
plt.show()
```



[ ]:
```

```