taiwan-credit-data-preproc

August 21, 2022

1 Taiwan Credit: Data Preprocessing for ML and Neural Network

[1]:

```
from google.colab import drive
drive.mount('/content/gdrive/')
```

```
Drive already mounted at /content/gdrive/; to attempt to forcibly remount, call
drive.mount("/content/gdrive/", force_remount=True).
```

[2]:

```
%cd /content/gdrive/MyDrive/Github/ml-blog
```

```
/content/gdrive/MyDrive/Github/ml-blog
```

[3]:

```
!pip install xlrd==1.2.0
# !pip install matplotlib==3.5.3
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: xlrd==1.2.0 in /usr/local/lib/python3.7/dist-
packages (1.2.0)
```

[4]:

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


filename = 'default of credit card clients.xls'

DATA = os.path.relpath('/content/gdrive/MyDrive/Github/ml-blog/credit/data/' +␣
 ↪filename)

df = pd.read_excel(DATA, 'Data', index_col=[0], header=[1], na_values='NA')
df.head()
```

[4]:

```
    LIMIT_BAL  SEX  EDUCATION  MARRIAGE  AGE  PAY_0  PAY_2  PAY_3  PAY_4  \
ID
1       20000    2          2         1   24      2      2     -1     -1
```

```
2      120000    2          2          2   26      -1       2       0       0
3       90000    2          2          2   34       0       0       0       0
4       50000    2          2          1   37       0       0       0       0
5       50000    1          2          1   57      -1       0      -1       0

    PAY_5  …  BILL_AMT4  BILL_AMT5  BILL_AMT6  PAY_AMT1  PAY_AMT2  PAY_AMT3  \
ID         …
1      -2  …          0          0          0         0       689         0
2       0  …       3272       3455       3261         0      1000      1000
3       0  …      14331      14948      15549      1518      1500      1000
4       0  …      28314      28959      29547      2000      2019      1200
5       0  …      20940      19146      19131      2000     36681     10000

    PAY_AMT4  PAY_AMT5  PAY_AMT6  default payment next month
ID
1          0         0         0                           1
2       1000         0      2000                           1
3       1000      1000      5000                           0
4       1100      1069      1000                           0
5       9000       689       679                           0

[5 rows x 24 columns]
```

[5]:

```
X = df.iloc[:, :23]
Y = df.iloc[:, 23]
X.head(), Y.head()
```

[5]:

```
(    LIMIT_BAL  SEX  EDUCATION  MARRIAGE  AGE  PAY_0  PAY_2  PAY_3  PAY_4  \
 ID
 1       20000    2          2          1   24       2       2      -1      -1
 2      120000    2          2          2   26      -1       2       0       0
 3       90000    2          2          2   34       0       0       0       0
 4       50000    2          2          1   37       0       0       0       0
 5       50000    1          2          1   57      -1       0      -1       0

     PAY_5  …  BILL_AMT3  BILL_AMT4  BILL_AMT5  BILL_AMT6  PAY_AMT1  \
 ID         …
 1      -2  …        689          0          0          0         0
 2       0  …       2682       3272       3455       3261         0
 3       0  …      13559      14331      14948      15549      1518
 4       0  …      49291      28314      28959      29547      2000
 5       0  …      35835      20940      19146      19131      2000

     PAY_AMT2  PAY_AMT3  PAY_AMT4  PAY_AMT5  PAY_AMT6
 ID
 1        689         0         0         0         0
 2       1000      1000      1000         0      2000
 3       1500      1000      1000      1000      5000
 4       2019      1200      1100      1069      1000
 5      36681     10000      9000       689       679

 [5 rows x 23 columns], ID
 1    1
 2    1
 3    0
 4    0
 5    0
 Name: default payment next month, dtype: int64)
```

[6]:
```python
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
```

[7]:
```python
ohe_cols = ['SEX', 'EDUCATION', 'MARRIAGE', 'PAY_0', 'PAY_2', 'PAY_3',
            'PAY_4', 'PAY_5', 'PAY_6']

num_cols = ['LIMIT_BAL', 'AGE', 'BILL_AMT1', 'BILL_AMT2', 'BILL_AMT3',␣
 ↪'BILL_AMT4',
            'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3',
            'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6']
```

[8]:
```python
column_transform = ColumnTransformer(
    [('category', OneHotEncoder(handle_unknown='ignore'), ohe_cols),
     ('nums', StandardScaler(), num_cols)],
     remainder='drop')
```

[9]:
```python
column_transform.fit(X)
```

[9]:
```
ColumnTransformer(transformers=[('category',
                                 OneHotEncoder(handle_unknown='ignore'),
                                 ['SEX', 'EDUCATION', 'MARRIAGE', 'PAY_0',
                                  'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5',
                                  'PAY_6']),
                                ('nums', StandardScaler(),
                                 ['LIMIT_BAL', 'AGE', 'BILL_AMT1', 'BILL_AMT2',
                                  'BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5',
                                  'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2',
                                  'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5',
                                  'PAY_AMT6'])])
```

[10]:
```python
column_transform.get_feature_names_out()
```

[10]:
```
array(['category__SEX_1', 'category__SEX_2', 'category__EDUCATION_0',
       'category__EDUCATION_1', 'category__EDUCATION_2',
       'category__EDUCATION_3', 'category__EDUCATION_4',
       'category__EDUCATION_5', 'category__EDUCATION_6',
       'category__MARRIAGE_0', 'category__MARRIAGE_1',
       'category__MARRIAGE_2', 'category__MARRIAGE_3',
       'category__PAY_0_-2', 'category__PAY_0_-1', 'category__PAY_0_0',
       'category__PAY_0_1', 'category__PAY_0_2', 'category__PAY_0_3',
       'category__PAY_0_4', 'category__PAY_0_5', 'category__PAY_0_6',
       'category__PAY_0_7', 'category__PAY_0_8', 'category__PAY_2_-2',
       'category__PAY_2_-1', 'category__PAY_2_0', 'category__PAY_2_1',
       'category__PAY_2_2', 'category__PAY_2_3', 'category__PAY_2_4',
       'category__PAY_2_5', 'category__PAY_2_6', 'category__PAY_2_7',
       'category__PAY_2_8', 'category__PAY_3_-2', 'category__PAY_3_-1',
       'category__PAY_3_0', 'category__PAY_3_1', 'category__PAY_3_2',
       'category__PAY_3_3', 'category__PAY_3_4', 'category__PAY_3_5',
       'category__PAY_3_6', 'category__PAY_3_7', 'category__PAY_3_8',
       'category__PAY_4_-2', 'category__PAY_4_-1', 'category__PAY_4_0',
       'category__PAY_4_1', 'category__PAY_4_2', 'category__PAY_4_3',
       'category__PAY_4_4', 'category__PAY_4_5', 'category__PAY_4_6',
       'category__PAY_4_7', 'category__PAY_4_8', 'category__PAY_5_-2',
       'category__PAY_5_-1', 'category__PAY_5_0', 'category__PAY_5_2',
       'category__PAY_5_3', 'category__PAY_5_4', 'category__PAY_5_5',
       'category__PAY_5_6', 'category__PAY_5_7', 'category__PAY_5_8',
       'category__PAY_6_-2', 'category__PAY_6_-1', 'category__PAY_6_0',
       'category__PAY_6_2', 'category__PAY_6_3', 'category__PAY_6_4',
       'category__PAY_6_5', 'category__PAY_6_6', 'category__PAY_6_7',
       'category__PAY_6_8', 'nums__LIMIT_BAL', 'nums__AGE',
       'nums__BILL_AMT1', 'nums__BILL_AMT2', 'nums__BILL_AMT3',
       'nums__BILL_AMT4', 'nums__BILL_AMT5', 'nums__BILL_AMT6',
       'nums__PAY_AMT1', 'nums__PAY_AMT2', 'nums__PAY_AMT3',
       'nums__PAY_AMT4', 'nums__PAY_AMT5', 'nums__PAY_AMT6'], dtype=object)
```

[11]:
```
X_prep = pd.DataFrame(column_transform.transform(X).toarray(),␣
 ↪columns=column_transform.get_feature_names_out(), index=X.index)
```

[12]:
```
X_prep.head()
```

[12]:
```
    category__SEX_1  category__SEX_2  category__EDUCATION_0  \
ID
1               0.0              1.0                    0.0
2               0.0              1.0                    0.0
3               0.0              1.0                    0.0
4               0.0              1.0                    0.0
5               1.0              0.0                    0.0

    category__EDUCATION_1  category__EDUCATION_2  category__EDUCATION_3  \
ID
1                     0.0                    1.0                    0.0
2                     0.0                    1.0                    0.0
```

```
3                     0.0                    1.0                    0.0
4                     0.0                    1.0                    0.0
5                     0.0                    1.0                    0.0

    category__EDUCATION_4  category__EDUCATION_5  category__EDUCATION_6  \
ID
1                     0.0                    0.0                    0.0
2                     0.0                    0.0                    0.0
3                     0.0                    0.0                    0.0
4                     0.0                    0.0                    0.0
5                     0.0                    0.0                    0.0

    category__MARRIAGE_0  …  nums__BILL_AMT3  nums__BILL_AMT4  \
ID                        …
1                    0.0  …        -0.667993        -0.672497
2                    0.0  …        -0.639254        -0.621636
3                    0.0  …        -0.482408        -0.449730
4                    0.0  …         0.032846        -0.232373
5                    0.0  …        -0.161189        -0.346997

    nums__BILL_AMT5  nums__BILL_AMT6  nums__PAY_AMT1  nums__PAY_AMT2  \
ID
1         -0.663059        -0.652724       -0.341942       -0.227086
2         -0.606229        -0.597966       -0.341942       -0.213588
3         -0.417188        -0.391630       -0.250292       -0.191887
4         -0.186729        -0.156579       -0.221191       -0.169361
5         -0.348137        -0.331482       -0.221191        1.335034

    nums__PAY_AMT3  nums__PAY_AMT4  nums__PAY_AMT5  nums__PAY_AMT6
ID
1        -0.296801       -0.308063       -0.314136       -0.293382
2        -0.240005       -0.244230       -0.314136       -0.180878
3        -0.240005       -0.244230       -0.248683       -0.012122
4        -0.228645       -0.237846       -0.244166       -0.237130
5         0.271165        0.266434       -0.269039       -0.255187

[5 rows x 91 columns]
```

[13]:

```
X_prep.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30000 entries, 1 to 30000
Data columns (total 91 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   category__SEX_1        30000 non-null  float64
 1   category__SEX_2        30000 non-null  float64
 2   category__EDUCATION_0  30000 non-null  float64
```

```
3   category__EDUCATION_1  30000 non-null  float64
4   category__EDUCATION_2  30000 non-null  float64
5   category__EDUCATION_3  30000 non-null  float64
6   category__EDUCATION_4  30000 non-null  float64
7   category__EDUCATION_5  30000 non-null  float64
8   category__EDUCATION_6  30000 non-null  float64
9   category__MARRIAGE_0   30000 non-null  float64
10  category__MARRIAGE_1   30000 non-null  float64
11  category__MARRIAGE_2   30000 non-null  float64
12  category__MARRIAGE_3   30000 non-null  float64
13  category__PAY_0_-2     30000 non-null  float64
14  category__PAY_0_-1     30000 non-null  float64
15  category__PAY_0_0      30000 non-null  float64
16  category__PAY_0_1      30000 non-null  float64
17  category__PAY_0_2      30000 non-null  float64
18  category__PAY_0_3      30000 non-null  float64
19  category__PAY_0_4      30000 non-null  float64
20  category__PAY_0_5      30000 non-null  float64
21  category__PAY_0_6      30000 non-null  float64
22  category__PAY_0_7      30000 non-null  float64
23  category__PAY_0_8      30000 non-null  float64
24  category__PAY_2_-2     30000 non-null  float64
25  category__PAY_2_-1     30000 non-null  float64
26  category__PAY_2_0      30000 non-null  float64
27  category__PAY_2_1      30000 non-null  float64
28  category__PAY_2_2      30000 non-null  float64
29  category__PAY_2_3      30000 non-null  float64
30  category__PAY_2_4      30000 non-null  float64
31  category__PAY_2_5      30000 non-null  float64
32  category__PAY_2_6      30000 non-null  float64
33  category__PAY_2_7      30000 non-null  float64
34  category__PAY_2_8      30000 non-null  float64
35  category__PAY_3_-2     30000 non-null  float64
36  category__PAY_3_-1     30000 non-null  float64
37  category__PAY_3_0      30000 non-null  float64
38  category__PAY_3_1      30000 non-null  float64
39  category__PAY_3_2      30000 non-null  float64
40  category__PAY_3_3      30000 non-null  float64
41  category__PAY_3_4      30000 non-null  float64
42  category__PAY_3_5      30000 non-null  float64
43  category__PAY_3_6      30000 non-null  float64
44  category__PAY_3_7      30000 non-null  float64
45  category__PAY_3_8      30000 non-null  float64
46  category__PAY_4_-2     30000 non-null  float64
47  category__PAY_4_-1     30000 non-null  float64
48  category__PAY_4_0      30000 non-null  float64
49  category__PAY_4_1      30000 non-null  float64
50  category__PAY_4_2      30000 non-null  float64
```

```
 51  category__PAY_4_3       30000 non-null  float64
 52  category__PAY_4_4       30000 non-null  float64
 53  category__PAY_4_5       30000 non-null  float64
 54  category__PAY_4_6       30000 non-null  float64
 55  category__PAY_4_7       30000 non-null  float64
 56  category__PAY_4_8       30000 non-null  float64
 57  category__PAY_5_-2      30000 non-null  float64
 58  category__PAY_5_-1      30000 non-null  float64
 59  category__PAY_5_0       30000 non-null  float64
 60  category__PAY_5_2       30000 non-null  float64
 61  category__PAY_5_3       30000 non-null  float64
 62  category__PAY_5_4       30000 non-null  float64
 63  category__PAY_5_5       30000 non-null  float64
 64  category__PAY_5_6       30000 non-null  float64
 65  category__PAY_5_7       30000 non-null  float64
 66  category__PAY_5_8       30000 non-null  float64
 67  category__PAY_6_-2      30000 non-null  float64
 68  category__PAY_6_-1      30000 non-null  float64
 69  category__PAY_6_0       30000 non-null  float64
 70  category__PAY_6_2       30000 non-null  float64
 71  category__PAY_6_3       30000 non-null  float64
 72  category__PAY_6_4       30000 non-null  float64
 73  category__PAY_6_5       30000 non-null  float64
 74  category__PAY_6_6       30000 non-null  float64
 75  category__PAY_6_7       30000 non-null  float64
 76  category__PAY_6_8       30000 non-null  float64
 77  nums__LIMIT_BAL         30000 non-null  float64
 78  nums__AGE               30000 non-null  float64
 79  nums__BILL_AMT1         30000 non-null  float64
 80  nums__BILL_AMT2         30000 non-null  float64
 81  nums__BILL_AMT3         30000 non-null  float64
 82  nums__BILL_AMT4         30000 non-null  float64
 83  nums__BILL_AMT5         30000 non-null  float64
 84  nums__BILL_AMT6         30000 non-null  float64
 85  nums__PAY_AMT1          30000 non-null  float64
 86  nums__PAY_AMT2          30000 non-null  float64
 87  nums__PAY_AMT3          30000 non-null  float64
 88  nums__PAY_AMT4          30000 non-null  float64
 89  nums__PAY_AMT5          30000 non-null  float64
 90  nums__PAY_AMT6          30000 non-null  float64
dtypes: float64(91)
memory usage: 21.1 MB
```

```
[16]: X_prep_labels = pd.concat([X_prep, Y], axis=1)
```

```
[17]: X_prep_labels.head()
```

[17]:

```
    category__SEX_1  category__SEX_2  category__EDUCATION_0  \
ID
1               0.0              1.0                    0.0
2               0.0              1.0                    0.0
3               0.0              1.0                    0.0
4               0.0              1.0                    0.0
5               1.0              0.0                    0.0

    category__EDUCATION_1  category__EDUCATION_2  category__EDUCATION_3  \
ID
1                     0.0                    1.0                    0.0
2                     0.0                    1.0                    0.0
3                     0.0                    1.0                    0.0
4                     0.0                    1.0                    0.0
5                     0.0                    1.0                    0.0

    category__EDUCATION_4  category__EDUCATION_5  category__EDUCATION_6  \
ID
1                     0.0                    0.0                    0.0
2                     0.0                    0.0                    0.0
3                     0.0                    0.0                    0.0
4                     0.0                    0.0                    0.0
5                     0.0                    0.0                    0.0

    category__MARRIAGE_0  …  nums__BILL_AMT4  nums__BILL_AMT5  \
ID                        …
1                    0.0  …        -0.672497        -0.663059
2                    0.0  …        -0.621636        -0.606229
3                    0.0  …        -0.449730        -0.417188
4                    0.0  …        -0.232373        -0.186729
5                    0.0  …        -0.346997        -0.348137

    nums__BILL_AMT6  nums__PAY_AMT1  nums__PAY_AMT2  nums__PAY_AMT3  \
ID
1         -0.652724       -0.341942       -0.227086       -0.296801
2         -0.597966       -0.341942       -0.213588       -0.240005
3         -0.391630       -0.250292       -0.191887       -0.240005
4         -0.156579       -0.221191       -0.169361       -0.228645
5         -0.331482       -0.221191        1.335034        0.271165

    nums__PAY_AMT4  nums__PAY_AMT5  nums__PAY_AMT6  default payment next month
ID
1        -0.308063       -0.314136       -0.293382                           1
2        -0.244230       -0.314136       -0.180878                           1
3        -0.244230       -0.248683       -0.012122                           0
4        -0.237846       -0.244166       -0.237130                           0
5         0.266434       -0.269039       -0.255187                           0
```

```
[5 rows x 92 columns]
```

```
[19]: X_prep_labels.to_csv('./credit/data/taiwan-credit-col-transform-FULL.csv',␣
        ↪header=True)
```

```
[ ]:
```