# task2

April 14, 2024

## 0.1 CREDIT CARD FRAUD DETECTION

[120]:

```
import pandas as pd
import numpy as np
data = pd.read_csv("/content/fraudTest.csv")
data.head()
```

[120]:

```
   Unnamed: 0 trans_date_trans_time            cc_num  \
0           0   2020-06-21 12:14:25  2291163933867244
1           1   2020-06-21 12:14:33  3573030041201292
2           2   2020-06-21 12:14:53  3598215285024754
3           3   2020-06-21 12:15:15  3591919803438423
4           4   2020-06-21 12:15:17  3526826139003047

                               merchant        category    amt   first  \
0                 fraud_Kirlin and Sons   personal_care   2.86    Jeff
1                  fraud_Sporer-Keebler   personal_care  29.84  Joanne
2  fraud_Swaniawski, Nitzsche and Welch  health_fitness  41.28  Ashley
3                     fraud_Haley Group        misc_pos  60.05   Brian
4                 fraud_Johnston-Casper          travel   3.19  Nathan

       last gender                       street  …      lat      long  \
0   Elliott      M            351 Darlene Green  …  33.9659  -80.9355
1  Williams      F             3638 Marsh Union  …  40.3207 -110.4360
2     Lopez      F         9333 Valentine Point  …  40.6729  -73.5365
3  Williams      M  32941 Krystal Mill Apt. 552  …  28.5697  -80.8191
4    Massey      M     5783 Evan Roads Apt. 465  …  44.2529  -85.0170

   city_pop                     job         dob  \
0    333497     Mechanical engineer  1968-03-19
1       302  Sales professional, IT  1990-01-17
2     34496       Librarian, public  1970-10-21
3     54767            Set designer  1987-07-25
4      1126      Furniture designer  1955-07-06

                          trans_num   unix_time  merch_lat  merch_long  \
0  2da90c7d74bd46a0caf3777415b3ebd3  1371816865  33.986391  -81.200714
1  324cc204407e99f51b0d6ca0055005e7  1371816873  39.450498 -109.960431
```

```
2  c81755dbbbea9d5c77f094348a7579be  1371816893  40.495810  -74.196111
3  2159175b9efe66dc301f149d3d5abf8c  1371816915  28.812398  -80.883061
4  57ff021bd3f328f8738bb535c302a31b  1371816917  44.959148  -85.884734

   is_fraud
0         0
1         0
2         0
3         0
4         0

[5 rows x 23 columns]
```

[121]:

```
data.isnull().sum()
```

[121]:

```
Unnamed: 0               0
trans_date_trans_time    0
cc_num                   0
merchant                 0
category                 0
amt                      0
first                    0
last                     0
gender                   0
street                   0
city                     0
state                    0
zip                      0
lat                      0
long                     0
city_pop                 0
job                      0
dob                      0
trans_num                0
unix_time                0
merch_lat                0
merch_long               0
is_fraud                 0
dtype: int64
```

[122]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 555719 entries, 0 to 555718
Data columns (total 23 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
```

```
 0   Unnamed: 0             555719 non-null  int64
 1   trans_date_trans_time  555719 non-null  object
 2   cc_num                 555719 non-null  int64
 3   merchant               555719 non-null  object
 4   category               555719 non-null  object
 5   amt                    555719 non-null  float64
 6   first                  555719 non-null  object
 7   last                   555719 non-null  object
 8   gender                 555719 non-null  object
 9   street                 555719 non-null  object
 10  city                   555719 non-null  object
 11  state                  555719 non-null  object
 12  zip                    555719 non-null  int64
 13  lat                    555719 non-null  float64
 14  long                   555719 non-null  float64
 15  city_pop               555719 non-null  int64
 16  job                    555719 non-null  object
 17  dob                    555719 non-null  object
 18  trans_num              555719 non-null  object
 19  unix_time              555719 non-null  int64
 20  merch_lat              555719 non-null  float64
 21  merch_long             555719 non-null  float64
 22  is_fraud               555719 non-null  int64
dtypes: float64(5), int64(6), object(12)
memory usage: 97.5+ MB
```

[123]:
```
data.describe()
```

[123]:
```
          Unnamed: 0        cc_num            amt            zip  \
count  555719.000000  5.557190e+05  555719.000000  555719.000000
mean   277859.000000  4.178387e+17      69.392810   48842.628015
std    160422.401459  1.309837e+18     156.745941   26855.283328
min         0.000000  6.041621e+10       1.000000    1257.000000
25%    138929.500000  1.800429e+14       9.630000   26292.000000
50%    277859.000000  3.521417e+15      47.290000   48174.000000
75%    416788.500000  4.635331e+15      83.010000   72011.000000
max    555718.000000  4.992346e+18   22768.110000   99921.000000

                 lat           long      city_pop     unix_time  \
count  555719.000000  555719.000000  5.557190e+05  5.557190e+05
mean       38.543253     -90.231325  8.822189e+04  1.380679e+09
std         5.061336      13.721780  3.003909e+05  5.201104e+06
min        20.027100    -165.672300  2.300000e+01  1.371817e+09
25%        34.668900     -96.798000  7.410000e+02  1.376029e+09
50%        39.371600     -87.476900  2.408000e+03  1.380762e+09
75%        41.894800     -80.175200  1.968500e+04  1.385867e+09
max        65.689900     -67.950300  2.906700e+06  1.388534e+09
```

```
           merch_lat      merch_long       is_fraud
count  555719.000000  555719.000000  555719.000000
mean       38.542798     -90.231380       0.003860
std         5.095829      13.733071       0.062008
min        19.027422    -166.671575       0.000000
25%        34.755302     -96.905129       0.000000
50%        39.376593     -87.445204       0.000000
75%        41.954163     -80.264637       0.000000
max        66.679297     -66.952026       1.000000
```

[124]:

```
data['city_pop'].fillna(data['city_pop'].median(), inplace=True)
data['unix_time'].fillna(data['unix_time'].median(), inplace=True)
data['merch_lat'].fillna(data['merch_lat'].median(), inplace=True)
data['merch_long'].fillna(data['merch_long'].median(), inplace=True)
data['is_fraud'].fillna(0, inplace=True)
```

[125]:

```
data.dropna(subset=['unix_time', 'merch_lat', 'merch_long', 'is_fraud'],␣
 ↪inplace=True)
```

[126]:

```
# Check for missing values in the entire dataset
missing_values = data.isnull().sum()
print(missing_values)
```

```
Unnamed: 0               0
trans_date_trans_time    0
cc_num                   0
merchant                 0
category                 0
amt                      0
first                    0
last                     0
gender                   0
street                   0
city                     0
state                    0
zip                      0
lat                      0
long                     0
city_pop                 0
job                      0
dob                      0
trans_num                0
unix_time                0
merch_lat                0
merch_long               0
is_fraud                 0
```

```
dtype: int64
```

[127]:
```
X = data.drop('is_fraud', axis=1)
y = data['is_fraud']
```

[128]:
```
data['Unnamed: 0'],unnamed_name=pd.factorize(data['Unnamed: 0'])
print(unnamed_name)
```

```
Index([     0,      1,      2,      3,      4,      5,      6,      7,      8,
            9,
       …
       555709, 555710, 555711, 555712, 555713, 555714, 555715, 555716, 555717,
       555718],
      dtype='int64', length=555719)
```

[129]:
```
data['cc_num'],cc_name=pd.factorize(data['cc_num'])
print(cc_name)
```

```
Index([2291163933867244, 3573030041201292, 3598215285024754, 3591919803438423,
       3526826139003047,   30407675418785,  213180742685905, 3589289942931264,
       3596357274378601, 3546897637165774,
       …
       3550412175018089,     586100864972,  372965408103277,  180020605265701,
        347399333635231,    4883407061576,    4295296907373, 4087542780207162,
       3588001568691267, 2242176657877538],
      dtype='int64', length=924)
```

[130]:
```
data['category'],category_name=pd.factorize(data['category'])
print(category_name)
```

```
Index(['personal_care', 'health_fitness', 'misc_pos', 'travel', 'kids_pets',
       'shopping_pos', 'food_dining', 'home', 'entertainment', 'shopping_net',
       'misc_net', 'grocery_pos', 'gas_transport', 'grocery_net'],
      dtype='object')
```

[131]:
```
data['trans_date_trans_time'],time_name=pd.
  ↪factorize(data['trans_date_trans_time'])
print(time_name)
```

```
Index(['2020-06-21 12:14:25', '2020-06-21 12:14:33', '2020-06-21 12:14:53',
       '2020-06-21 12:15:15', '2020-06-21 12:15:17', '2020-06-21 12:15:37',
       '2020-06-21 12:15:44', '2020-06-21 12:15:50', '2020-06-21 12:16:10',
       '2020-06-21 12:16:11',
       …
       '2020-12-31 23:57:18', '2020-12-31 23:57:50', '2020-12-31 23:57:56',
       '2020-12-31 23:58:04', '2020-12-31 23:58:34', '2020-12-31 23:59:07',
       '2020-12-31 23:59:09', '2020-12-31 23:59:15', '2020-12-31 23:59:24',
       '2020-12-31 23:59:34'],
      dtype='object', length=544760)
```

[132]:
```
data['amt'],amt_name=pd.factorize(data['amt'])
print(amt_name)
```

```
Index([   2.86,   29.84,   41.28,   60.05,    3.19,   19.55,  133.93,   10.37,
          4.37,   66.54,
       …
       2149.66,  537.02, 1309.21,  256.67,  500.31,  850.87,  516.74,  255.42,
        302.79, 1164.37],
      dtype='float64', length=37256)
```

[133]:
```
data['merchant'],merchant_name=pd.factorize(data['merchant'])
print(merchant_name)
```

```
Index(['fraud_Kirlin and Sons', 'fraud_Sporer-Keebler',
       'fraud_Swaniawski, Nitzsche and Welch', 'fraud_Haley Group',
       'fraud_Johnston-Casper', 'fraud_Daugherty LLC', 'fraud_Romaguera Ltd',
       'fraud_Reichel LLC', 'fraud_Goyette, Howell and Collier',
       'fraud_Kilback Group',
       …
       'fraud_Rippin, Kub and Mann', 'fraud_Rempel PLC',
       'fraud_Leannon-Nikolaus', 'fraud_Monahan, Hermann and Johns',
       'fraud_Block-Hauck', 'fraud_Hagenes, Hermann and Stroman',
       'fraud_Hermann-Gaylord', 'fraud_Mante Group', 'fraud_Corwin-Gorczany',
       'fraud_McCullough Group'],
      dtype='object', length=693)
```

[134]:
```
data['zip'],zip_name=pd.factorize(data['zip'])
print(zip_name)
```

```
Index([29209, 84002, 11710, 32780, 49632, 14816, 95528, 57374, 16858, 76678,
       …
       40502, 13795, 87417, 66958, 65745, 98118, 52658, 73044, 99921, 38668],
      dtype='int64', length=912)
```

[135]:
```
data['lat'],lat_name=pd.factorize(data['lat'])
print(lat_name)
```

```
Index([33.9659, 40.3207, 40.6729, 28.5697, 44.2529, 42.1939,  40.507, 43.7557,
       41.0001, 31.6591,
       …
       38.0174, 42.0695,  36.741, 39.8616, 36.5276, 47.5412, 40.7067,  35.833,
       55.4732, 34.6323],
      dtype='float64', length=910)
```

[136]:
```
data['long'],long_name=pd.factorize(data['long'])
print(long_name)
```

```
Index([          -80.9355,           -110.436,           -73.5365,
                 -80.8191, -85.01700000000001,           -76.7361,
                -123.9743,           -97.5936,           -78.2357,
                 -96.8094,
       …
                 -84.4854,           -75.7967,           -108.351,
                 -97.1825,           -93.9359,           -122.275,
                 -91.2268,            -97.436,          -133.1171,
                 -89.8855],
      dtype='float64', length=910)
```

[137]:

```
data['city_pop'],city_name=pd.factorize(data['city_pop'])
print(city_name)
```

```
Index([333497,    302,  34496,  54767,   1126,    520,   1139,    343,   3688,
          263,
       …
       296965,   3800,   6910,    314,   2693, 837792,   1071,  20226,   1920,
        14462],
      dtype='int64', length=835)
```

[138]:

```
data['is_fraud'],fraud_name=pd.factorize(data['is_fraud'])
print(fraud_name)
```

```
Index([0, 1], dtype='int64')
```

[139]:

```
data['first'],first_name=pd.factorize(data['first'])
print(first_name)
```

```
Index(['Jeff', 'Joanne', 'Ashley', 'Brian', 'Nathan', 'Danielle', 'Kayla',
       'Paula', 'David', 'Samuel',
       …
       'Katelyn', 'Wesley', 'Sonya', 'Collin', 'Tommy', 'Guy', 'Dennis',
       'Bruce', 'Evan', 'Nicole'],
      dtype='object', length=341)
```

[140]:

```
data['last'],last_name=pd.factorize(data['last'])
print(last_name)
```

```
Index(['Elliott', 'Williams', 'Lopez', 'Massey', 'Evans', 'Sutton', 'Estrada',
       'Everett', 'Obrien', 'Jenkins',
       …
       'Bridges', 'Raymond', 'Davidson', 'Osborne', 'Webster', 'Freeman',
       'Bartlett', 'Santiago', 'Bates', 'Robbins'],
      dtype='object', length=471)
```

[141]:

```
data['street'],street_name=pd.factorize(data['street'])
print(street_name)
```

```
Index(['351 Darlene Green', '3638 Marsh Union', '9333 Valentine Point',
       '32941 Krystal Mill Apt. 552', '5783 Evan Roads Apt. 465',
       '76752 David Lodge Apt. 064', '010 Weaver Land', '350 Stacy Glens',
       '4138 David Fall', '7921 Robert Port Suite 343',
       …
       '742 Sellers Ferry', '4481 Maldonado Hollow',
       '53199 Laurie Mills Apt. 864', '7908 Derrick Mount',
       '13128 Hall Station Suite 588', '6386 Bailey Hill Apt. 421',
       '007 Tonya Isle Suite 299', '537 Brian Island', '5942 Thomas Park',
       '1327 Rose Causeway Apt. 610'],
      dtype='object', length=924)
```

[142]:

```
data['job'],job_name=pd.factorize(data['job'])
print(job_name)
```

```
Index(['Mechanical engineer', 'Sales professional, IT', 'Librarian, public',
       'Set designer', 'Furniture designer', 'Psychotherapist',
       'Therapist, occupational', 'Development worker, international aid',
       'Advice worker', 'Barrister',
       …
       'Medical technical officer', 'Charity officer', 'Administrator, arts',
       'Occupational therapist', 'Solicitor, Scotland', 'Sports administrator',
       'Artist', 'Engineer, water', 'Operational investment banker',
       'Software engineer'],
      dtype='object', length=478)
```

[143]:

```
data['dob'],dob_name=pd.factorize(data['dob'])
print(dob_name)
```

```
Index(['1968-03-19', '1990-01-17', '1970-10-21', '1987-07-25', '1955-07-06',
       '1991-10-13', '1951-01-15', '1972-03-05', '1973-05-27', '1956-05-30',
       …
       '1962-12-30', '1968-07-06', '1956-02-02', '2002-03-17', '1968-02-05',
       '1936-12-23', '1998-08-02', '1969-11-08', '1997-06-17', '1959-03-03'],
      dtype='object', length=910)
```

[144]:

```
data['trans_num'],trans_name=pd.factorize(data['trans_num'])
print(trans_name)
```

```
Index(['2da90c7d74bd46a0caf3777415b3ebd3', '324cc204407e99f51b0d6ca0055005e7',
       'c81755dbbbea9d5c77f094348a7579be', '2159175b9efe66dc301f149d3d5abf8c',
       '57ff021bd3f328f8738bb535c302a31b', '798db04aaceb4febd084f1a7c404da93',
       '17003d7ce534440eadb10c4750e020e5', '8be473af4f05fc6146ea55ace73e7ca2',
       '71a1da150d1ce510193d7622e08e784e', 'a7915132c7c4240996ba03a47f81e3bd',
       …
       'a7105564935ea3977dc61ff9ced3bf5e', '9fc9f6f9be3182d519a61a119cf97199',
       'a8310343c189e4a5b6316050d2d6b014', 'bd7071fd5c9510a5594ee196368ac80e',
       '6d04313bfe4b661b8ca2b6a499a320fe', '9b1f753c79894c9f4b71f04581835ada',
       '2090647dac2c89a1d86c514c427f5b91', '6c5b7c8add471975aa0fec023b2e8408',
       '14392d723bb7737606b2700ac791b7aa', '1765bb45b3aa3224b4cdcb6e7a96cee3'],
      dtype='object', length=555719)
```

[145]:

```
data['gender'],gender_name=pd.factorize(data['gender'])
print(gender_name)
```

```
Index(['M', 'F'], dtype='object')
```

[146]:

```
data['city'],city_name=pd.factorize(data['city'])
print(city_name)
```

```
Index(['Columbia', 'Altonah', 'Bellmore', 'Titusville', 'Falmouth',
       'Breesport', 'Carlotta', 'Spencer', 'Morrisdale', 'Prairie Hill',
       …
       'Lexington', 'Kirkwood', 'Kirtland', 'Morrowville', 'Seligman',
       'Seattle', 'Wever', 'Guthrie', 'Craig', 'Senatobia'],
      dtype='object', length=849)
```

[147]:

```
data['state'],state_name=pd.factorize(data['state'])
print(state_name)
```

```
Index(['SC', 'UT', 'NY', 'FL', 'MI', 'CA', 'SD', 'PA', 'TX', 'KY', 'WY', 'AL',
       'LA', 'GA', 'CO', 'OH', 'WI', 'VT', 'AR', 'NJ', 'IA', 'MD', 'MS', 'KS',
       'IL', 'MO', 'ME', 'TN', 'DC', 'AZ', 'MT', 'MN', 'OK', 'WA', 'WV', 'NM',
       'MA', 'NE', 'VA', 'ID', 'OR', 'IN', 'NC', 'NH', 'ND', 'CT', 'NV', 'HI',
       'RI', 'AK'],
      dtype='object')
```

[148]:

```
x=data.iloc[:,0:-1]
y=data.iloc[:,-1]
print(x)
print(y)
```

```
        Unnamed: 0  trans_date_trans_time  cc_num  merchant  category    amt  \
0                0                      0       0         0         0      0
1                1                      1       1         1         0      1
2                2                      2       2         2         1      2
3                3                      3       3         3         2      3
4                4                      4       4         4         3      4
…            …                    …     …       …       …    …
555714      555714                 544755     757       296         1   5224
555715      555715                 544756     136        35         4   5504
555716      555716                 544757     607        39         4   5273
555717      555717                 544758     350       163         3   2243
555718      555718                 544759     250       173         8  13727

        first  last  gender  street  …  zip  lat  long  city_pop  job  dob  \
0           0     0       0       0  …    0    0     0         0    0    0
1           1     1       1       1  …    1    1     1         1    1    1
2           2     2       1       2  …    2    2     2         2    2    2
3           3     1       0       3  …    3    3     3         3    3    3
4           4     3       0       4  …    4    4     4         4    4    4
…       …   …     …     …  …  …  …   …       …  …  …
555714     89   325       0     757  …  747  747   746       694   16  746
555715    102   109       0     136  …  136  136   136       135  127  136
555716    285   353       1     607  …  600  600   599       565   38  597
555717    204   232       0     350  …  348  348   347       338  262  347
555718      9   184       0     250  …  249  249   249       242  205  249

        trans_num   unix_time  merch_lat  merch_long
0               0  1371816865  33.986391  -81.200714
1               1  1371816873  39.450498 -109.960431
2               2  1371816893  40.495810  -74.196111
3               3  1371816915  28.812398  -80.883061
4               4  1371816917  44.959148  -85.884734
…           …         …        …         …
555714     555714  1388534347  39.946837  -91.333331
555715     555715  1388534349  29.661049  -96.186633
555716     555716  1388534355  46.658340 -119.715054
555717     555717  1388534364  44.470525 -117.080888
555718     555718  1388534374  36.210097  -97.036372

[555719 rows x 22 columns]
0         0
1         0
2         0
3         0
4         0
         ..
555714    0
555715    0
555716    0
555717    0
555718    0
Name: is_fraud, Length: 555719, dtype: int64
```

[149]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

[150]:

```
feature_names = x.columns
```

[155]:

```
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression()
clf.fit(x_train, y_train)
```

```
[155]: LogisticRegression()
```

```
[156]: y_pred = dtree.predict(x_test)
```

```
[157]: from sklearn.metrics import accuracy_score
        accuracy = accuracy_score(y_test, y_pred)
        print("LogisticRegression:")
        print("Accuracy:", accuracy)
```

```
LogisticRegression:
Accuracy: 0.9961131505074498
```

```
[ ]:
```