otv7dvntk

February 10, 2024

1 1. Demonstrate using code and explain how did would you identify potential fraudulent activities in financial transactions.

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sb
```

[2]:
```
df = pd.read_csv("financial_anomaly_data.csv")
df
```

[2]:
```
                Timestamp TransactionID AccountID    Amount   Merchant  \
0       01-01-2023 08:00       TXN1127      ACC4  95071.92  MerchantH
1       01-01-2023 08:01       TXN1639     ACC10  15607.89  MerchantH
2       01-01-2023 08:02        TXN872      ACC8  65092.34  MerchantE
3       01-01-2023 08:03       TXN1438      ACC6     87.87  MerchantE
4       01-01-2023 08:04       TXN1338      ACC6    716.56  MerchantI
…                    …             …       …        …         …
217436               NaN           NaN       NaN       NaN        NaN
217437               NaN           NaN       NaN       NaN        NaN
217438               NaN           NaN       NaN       NaN        NaN
217439               NaN           NaN       NaN       NaN        NaN
217440               NaN           NaN       NaN       NaN        NaN

       TransactionType     Location
0             Purchase        Tokyo
1             Purchase       London
2           Withdrawal       London
3             Purchase       London
4             Purchase  Los Angeles
…                  …            …
217436             NaN          NaN
217437             NaN          NaN
217438             NaN          NaN
217439             NaN          NaN
217440             NaN          NaN

[217441 rows x 7 columns]
```

[3]:
```
df=df.dropna(axis=0, how="any")
df
```

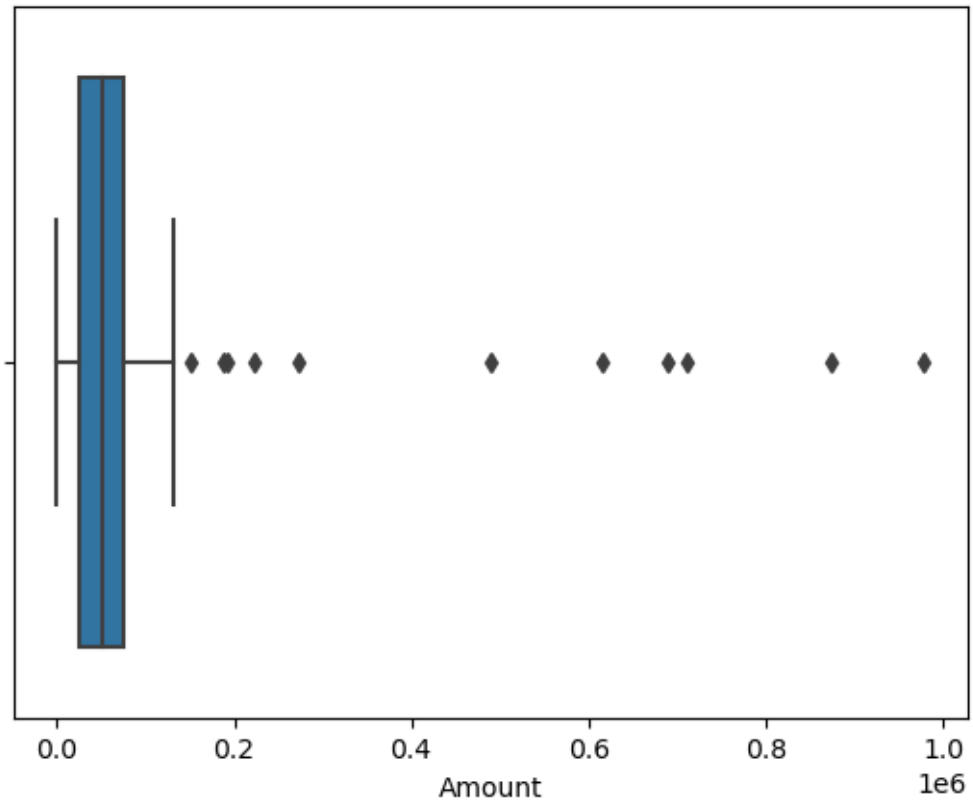
[3]:
```
                Timestamp TransactionID AccountID    Amount   Merchant  \
0       01-01-2023 08:00       TXN1127      ACC4  95071.92  MerchantH
1       01-01-2023 08:01       TXN1639     ACC10  15607.89  MerchantH
2       01-01-2023 08:02        TXN872      ACC8  65092.34  MerchantE
3       01-01-2023 08:03       TXN1438      ACC6     87.87  MerchantE
4       01-01-2023 08:04       TXN1338      ACC6    716.56  MerchantI
…                    …             …       …        …         …
216955  31-05-2023 23:55       TXN1286      ACC6  62536.88  MerchantA
216956  31-05-2023 23:56       TXN1015      ACC5  68629.69  MerchantG
216957  31-05-2023 23:57       TXN1979     ACC15   8203.57  MerchantF
216958  31-05-2023 23:58       TXN1845     ACC14  77800.36  MerchantF
216959  31-05-2023 23:59       TXN1807      ACC3  65004.99  MerchantG

       TransactionType       Location
0             Purchase          Tokyo
1             Purchase         London
2           Withdrawal         London
3             Purchase         London
4             Purchase    Los Angeles
…                  …              …
216955      Withdrawal  San Francisco
216956        Transfer         London
216957        Purchase         London
216958        Purchase       New York
216959      Withdrawal    Los Angeles

[216960 rows x 7 columns]
```

[4]:
```
df.dtypes
```

[4]:
```
Timestamp           object
TransactionID       object
AccountID           object
Amount             float64
Merchant            object
TransactionType     object
Location            object
dtype: object
```

[5]:
```
sb.boxplot(x=df['Amount'])
```

[5]:
```
<Axes: xlabel='Amount'>
```



[6]:
```
df['Amount'].describe()
```

[6]:
```
count    216960.000000
mean      50090.025108
std       29097.905016
min          10.510000
25%       25061.242500
50%       50183.980000
75%       75080.460000
max      978942.260000
Name: Amount, dtype: float64
```

[7]:
```
df['Location'].nunique()
```

[7]:
```
5
```

[8]:
```
df['TransactionType'].nunique()
```

[8]:
```
3
```

[9]:
```
df['Merchant'].nunique()
```

[9]: 10

[10]:
```
df['Timestamp'].nunique()
```

[10]: 216960

[11]:
```
df['AccountID'].nunique()
```

[11]: 15

[12]:
```
df['TransactionID'].nunique()
```

[12]: 1999

[13]:
```
df.columns
```

[13]:
```
Index(['Timestamp', 'TransactionID', 'AccountID', 'Amount', 'Merchant',
       'TransactionType', 'Location'],
      dtype='object')
```

[4]:
```
df=pd.get_dummies(df, columns=['AccountID', 'Merchant', 'TransactionType',␣
↪'Location'], drop_first=True)
```

[5]:
```
df
```

[5]:

	Timestamp	TransactionID	Amount	AccountID_ACC10	\
0	01-01-2023 08:00	TXN1127	95071.92	False	
1	01-01-2023 08:01	TXN1639	15607.89	True	
2	01-01-2023 08:02	TXN872	65092.34	False	
3	01-01-2023 08:03	TXN1438	87.87	False	
4	01-01-2023 08:04	TXN1338	716.56	False	
…	…	…	…	…	
216955	31-05-2023 23:55	TXN1286	62536.88	False	
216956	31-05-2023 23:56	TXN1015	68629.69	False	
216957	31-05-2023 23:57	TXN1979	8203.57	False	
216958	31-05-2023 23:58	TXN1845	77800.36	False	
216959	31-05-2023 23:59	TXN1807	65004.99	False	

	AccountID_ACC11	AccountID_ACC12	AccountID_ACC13	AccountID_ACC14	\
0	False	False	False	False	
1	False	False	False	False	
2	False	False	False	False	
3	False	False	False	False	
4	False	False	False	False	
…	…	…	…	…	
216955	False	False	False	False	
216956	False	False	False	False	

```
216957           False            False            False            False
216958           False            False            False             True
216959           False            False            False            False

        AccountID_ACC15  AccountID_ACC2  …  Merchant_MerchantG  \
0                 False           False  …               False
1                 False           False  …               False
2                 False           False  …               False
3                 False           False  …               False
4                 False           False  …               False
…                 …             …  …                 …
216955            False           False  …               False
216956            False           False  …                True
216957             True           False  …               False
216958            False           False  …               False
216959            False           False  …                True

        Merchant_MerchantH  Merchant_MerchantI  Merchant_MerchantJ  \
0                     True               False               False
1                     True               False               False
2                    False               False               False
3                    False               False               False
4                    False                True               False
…                    …                 …                 …
216955               False               False               False
216956               False               False               False
216957               False               False               False
216958               False               False               False
216959               False               False               False

        TransactionType_Transfer  TransactionType_Withdrawal  \
0                          False                        False
1                          False                        False
2                          False                         True
3                          False                        False
4                          False                        False
…                          …                          …
216955                     False                         True
216956                      True                        False
216957                     False                        False
216958                     False                        False
216959                     False                         True

        Location_Los Angeles  Location_New York  Location_San Francisco  \
0                      False              False                   False
1                      False              False                   False
2                      False              False                   False
```

```
3                    False              False                   False
4                     True              False                   False
…                     …                 …                       …
216955               False              False                    True
216956               False              False                   False
216957               False              False                   False
216958               False               True                   False
216959                True              False                   False

        Location_Tokyo
0                 True
1                False
2                False
3                False
4                False
…                …
216955           False
216956           False
216957           False
216958           False
216959           False

[216960 rows x 32 columns]
```

[6]:
```
from sklearn.preprocessing import RobustScaler

# create a StandardScaler object to perform Z-score normalization on the␣
 ↪'Amount' column
scaler = RobustScaler()

# reshape the 'Amount' column because the RobustScaler expects a 2D array
df['Amount'] = scaler.fit_transform(df['Amount'].values.reshape(-1, 1))

df
```

[6]:
```
               Timestamp TransactionID    Amount  AccountID_ACC10  \
0       01-01-2023 08:00       TXN1127  0.897414            False
1       01-01-2023 08:01       TXN1639 -0.691256             True
2       01-01-2023 08:02        TXN872  0.298053            False
3       01-01-2023 08:03       TXN1438 -1.001537            False
4       01-01-2023 08:04       TXN1338 -0.988968            False
…                    …             …       …                …
216955  31-05-2023 23:55       TXN1286  0.246963            False
216956  31-05-2023 23:56       TXN1015  0.368772            False
216957  31-05-2023 23:57       TXN1979 -0.839286            False
216958  31-05-2023 23:58       TXN1845  0.552115            False
216959  31-05-2023 23:59       TXN1807  0.296306            False
```

```
        AccountID_ACC11  AccountID_ACC12  AccountID_ACC13  AccountID_ACC14  \
0                 False            False            False            False
1                 False            False            False            False
2                 False            False            False            False
3                 False            False            False            False
4                 False            False            False            False
…                 …              …                …              …
216955            False            False            False            False
216956            False            False            False            False
216957            False            False            False            False
216958            False            False            False             True
216959            False            False            False            False

        AccountID_ACC15  AccountID_ACC2  …  Merchant_MerchantG  \
0                 False           False  …               False
1                 False           False  …               False
2                 False           False  …               False
3                 False           False  …               False
4                 False           False  …               False
…                 …             …  …                 …
216955            False           False  …               False
216956            False           False  …                True
216957             True           False  …               False
216958            False           False  …               False
216959            False           False  …                True

        Merchant_MerchantH  Merchant_MerchantI  Merchant_MerchantJ  \
0                     True               False               False
1                     True               False               False
2                    False               False               False
3                    False               False               False
4                    False                True               False
…                    …                 …                 …
216955               False               False               False
216956               False               False               False
216957               False               False               False
216958               False               False               False
216959               False               False               False

        TransactionType_Transfer  TransactionType_Withdrawal  \
0                          False                        False
1                          False                        False
2                          False                         True
3                          False                        False
4                          False                        False
…                          …                          …
```

```
216955                   False                        True
216956                    True                       False
216957                   False                       False
216958                   False                       False
216959                   False                        True

        Location_Los Angeles  Location_New York  Location_San Francisco  \
0                      False              False                   False
1                      False              False                   False
2                      False              False                   False
3                      False              False                   False
4                       True              False                   False
…                       …                  …                       …
216955                 False              False                    True
216956                 False              False                   False
216957                 False              False                   False
216958                 False               True                   False
216959                  True              False                   False

        Location_Tokyo
0                 True
1                False
2                False
3                False
4                False
…                    …
216955           False
216956           False
216957           False
216958           False
216959           False

[216960 rows x 32 columns]
```

[16]:
```
df.columns
```

[16]:
```
Index(['Timestamp', 'TransactionID', 'Amount', 'AccountID_ACC10',
       'AccountID_ACC11', 'AccountID_ACC12', 'AccountID_ACC13',
       'AccountID_ACC14', 'AccountID_ACC15', 'AccountID_ACC2',
       'AccountID_ACC3', 'AccountID_ACC4', 'AccountID_ACC5', 'AccountID_ACC6',
       'AccountID_ACC7', 'AccountID_ACC8', 'AccountID_ACC9',
       'Merchant_MerchantB', 'Merchant_MerchantC', 'Merchant_MerchantD',
       'Merchant_MerchantE', 'Merchant_MerchantF', 'Merchant_MerchantG',
       'Merchant_MerchantH', 'Merchant_MerchantI', 'Merchant_MerchantJ',
       'TransactionType_Transfer', 'TransactionType_Withdrawal',
       'Location_Los Angeles', 'Location_New York', 'Location_San Francisco',
       'Location_Tokyo'],
```

```
      dtype='object')
```

[7]:
```
features=['Amount', 'AccountID_ACC10',
       'AccountID_ACC11', 'AccountID_ACC12', 'AccountID_ACC13',
       'AccountID_ACC14', 'AccountID_ACC15', 'AccountID_ACC2',
       'AccountID_ACC3', 'AccountID_ACC4', 'AccountID_ACC5', 'AccountID_ACC6',
       'AccountID_ACC7', 'AccountID_ACC8', 'AccountID_ACC9',
       'Merchant_MerchantB', 'Merchant_MerchantC', 'Merchant_MerchantD',
       'Merchant_MerchantE', 'Merchant_MerchantF', 'Merchant_MerchantG',
       'Merchant_MerchantH', 'Merchant_MerchantI', 'Merchant_MerchantJ',
       'TransactionType_Transfer', 'TransactionType_Withdrawal',
       'Location_Los Angeles', 'Location_New York', 'Location_San Francisco',
       'Location_Tokyo']
```

2 Model: Isolation Forest

[8]:
```
from sklearn.ensemble import IsolationForest
```

[9]:
```
# create an isolation forest model with 100 trees and 0.01 contamination rate
model = IsolationForest(n_estimators=100, contamination=0.01, random_state=42)
```

[10]:
```
# fit the model to the data
model.fit(df[features])
```

[10]:
```
IsolationForest(contamination=0.01, random_state=42)
```

[11]:
```
# predict the anomaly scores for each data point
scores = model.decision_function(df[features])
```

[12]:
```
# label the data points as normal (1) or anomalous (-1)
labels = model.predict(df[features])
```

[13]:
```
# add the scores and labels to the original dataframe
df["score"] = scores
df["label"] = labels
```

[14]:
```
# filter the dataframe to show only the anomalous transactions
if_anomalies = df[df["label"] == -1]
```

[15]:
```
if_anomalies
```

[15]:

	Timestamp	TransactionID	Amount	AccountID_ACC10	\
70	01-01-2023 09:10	TXN656	0.935041	False	
143	01-01-2023 10:23	TXN1891	0.968465	False	
210	01-01-2023 11:30	TXN1648	0.748945	False	
348	01-01-2023 13:48	TXN291	-0.973774	False	

```
436     01-01-2023 15:16        TXN534 -0.916788            False
…                  …           …       …                 …
216703  31-05-2023 19:43        TXN386 -0.909379            False
216729  31-05-2023 20:09       TXN1332 -0.989948            False
216735  31-05-2023 20:15        TXN175 -0.854780            False
216783  31-05-2023 21:03       TXN1656 -0.263071            False
216816  31-05-2023 21:36       TXN1764 -0.853206            False

        AccountID_ACC11  AccountID_ACC12  AccountID_ACC13  AccountID_ACC14  \
70                False            False             True            False
143               False            False             True            False
210               False            False             True            False
348               False             True            False            False
436               False             True            False            False
…                 …                …                …                …
216703            False            False            False            False
216729            False            False             True            False
216735            False            False            False            False
216783            False            False             True            False
216816             True            False            False            False

        AccountID_ACC15  AccountID_ACC2  …  Merchant_MerchantI  \
70                False           False  …               False
143               False           False  …               False
210               False           False  …               False
348               False           False  …               False
436               False           False  …               False
…                 …              …  …                 …
216703            False           False  …               False
216729            False           False  …                True
216735            False           False  …               False
216783            False           False  …               False
216816            False           False  …               False

        Merchant_MerchantJ  TransactionType_Transfer  \
70                    True                     False
143                  False                     False
210                  False                     False
348                  False                     False
436                  False                     False
…                    …                        …
216703                True                      True
216729               False                      True
216735               False                     False
216783               False                     False
216816               False                      True
```

```
        TransactionType_Withdrawal  Location_Los Angeles  Location_New York  \
70                            True                  True              False
143                           True                 False              False
210                           True                 False               True
348                           True                 False              False
436                           True                 False              False
…                              …                     …                  …
216703                       False                 False               True
216729                       False                  True              False
216735                        True                 False              False
216783                        True                 False               True
216816                       False                  True              False

        Location_San Francisco  Location_Tokyo     score  label
70                       False           False -0.001907     -1
143                       True           False -0.005280     -1
210                      False           False -0.005911     -1
348                       True           False -0.008430     -1
436                      False            True -0.002368     -1
…                         …              …       …    …
216703                   False           False -0.009844     -1
216729                   False           False -0.002820     -1
216735                    True           False -0.002920     -1
216783                   False           False -0.000818     -1
216816                   False           False -0.000440     -1

[2169 rows x 34 columns]
```

3 Model: Local Outlier Factor(LOF)

[16]:
```
from sklearn.neighbors import LocalOutlierFactor
```

[17]:
```
# create a local outlier factor model with 20 neighbors and 0.01 contamination␣
↪rate
model = LocalOutlierFactor(n_neighbors=20, contamination=0.01)

# fit the model to the data and predict the labels for each data point
labels = model.fit_predict(df[features])

# add the labels to the original dataframe
df["label"] = labels

# filter the dataframe to show only the anomalous transactions
lof_anomalies = df[df["label"] == -1]
```

[18]:
```
lof_anomalies
```

[18]:

```
              Timestamp TransactionID    Amount  AccountID_ACC10  \
241     01-01-2023 12:01       TXN1514 -0.929115            False
272     01-01-2023 12:32       TXN1055 -0.992809            False
401     01-01-2023 14:41        TXN642  0.985675            False
445     01-01-2023 15:25        TXN602 -0.932506             True
452     01-01-2023 15:32       TXN1939 -0.957251            False
…                  …           …       …                 …
216445  31-05-2023 15:25        TXN862  0.859373            False
216595  31-05-2023 17:55        TXN132  0.992980            False
216729  31-05-2023 20:09       TXN1332 -0.989948            False
216758  31-05-2023 20:38          TXN4  0.890300             True
216840  31-05-2023 22:00       TXN1318 -1.000534            False

        AccountID_ACC11  AccountID_ACC12  AccountID_ACC13  AccountID_ACC14  \
241               False            False            False            False
272               False            False            False            False
401               False            False            False             True
445               False            False            False            False
452               False            False            False            False
…                 …                …                …                …
216445            False            False            False            False
216595            False            False            False             True
216729            False            False             True            False
216758            False            False            False            False
216840            False            False            False            False

        AccountID_ACC15  AccountID_ACC2  …  Merchant_MerchantI  \
241               False           False  …               False
272               False           False  …               False
401               False           False  …               False
445               False           False  …               False
452                True           False  …               False
…                 …               …  …  …                …
216445            False           False  …               False
216595            False           False  …               False
216729            False           False  …                True
216758            False           False  …                True
216840            False           False  …               False

        Merchant_MerchantJ  TransactionType_Transfer  \
241                  False                      True
272                  False                     False
401                  False                      True
445                  False                     False
452                  False                     False
…                    …                         …
216445               False                     False
```

```
216595                False                      False
216729                False                       True
216758                False                      False
216840                False                      False

        TransactionType_Withdrawal  Location_Los Angeles  Location_New York  \
241                          False                 False              False
272                           True                 False              False
401                          False                 False               True
445                           True                 False               True
452                          False                  True              False
…                                …                     …                  …
216445                        True                  True              False
216595                       False                 False               True
216729                       False                  True              False
216758                       False                 False              False
216840                       False                 False              False

        Location_San Francisco  Location_Tokyo     score  label
241                      False           False  0.039339     -1
272                       True           False  0.023460     -1
401                      False           False  0.020040     -1
445                      False           False  0.057650     -1
452                      False           False  0.014635     -1
…                            …               …        …    …
216445                   False           False  0.007211     -1
216595                   False           False  0.035176     -1
216729                   False           False -0.002820     -1
216758                   False           False  0.057230     -1
216840                   False           False  0.033391     -1

[2170 rows x 34 columns]
```

[19]:
```
common_anomalies = pd.merge(if_anomalies, lof_anomalies, how='inner')
combined_anomalies = pd.merge(if_anomalies, lof_anomalies, how='outer')
```

[20]:
```
common_anomalies
```

[20]:
```
            Timestamp TransactionID    Amount  AccountID_ACC10  \
0   03-01-2023 05:44        TXN657  0.979939            False
1   03-01-2023 20:33       TXN1002 -0.914648            False
2   04-01-2023 02:36        TXN741 -0.923990            False
3   04-01-2023 09:10       TXN1261 -0.936354            False
4   06-01-2023 16:11         TXN97 -0.937480            False
..               …            …       …                …
78  24-05-2023 05:27        TXN610 -0.929457            False
79  26-05-2023 02:42        TXN805 -0.962079            False
```

```
80  29-05-2023 11:56       TXN1401 -0.966966            False
81  30-05-2023 06:36        TXN903  0.975310            False
82  31-05-2023 20:09       TXN1332 -0.989948            False

    AccountID_ACC11  AccountID_ACC12  AccountID_ACC13  AccountID_ACC14  \
0             False            False            False            False
1             False            False             True            False
2             False            False            False            False
3             False            False            False            False
4             False            False             True            False
..              …                …                …                …
78            False            False             True            False
79            False             True            False            False
80            False            False             True            False
81            False            False            False            False
82            False            False             True            False

    AccountID_ACC15  AccountID_ACC2  …  Merchant_MerchantI  \
0             False           False  …               False
1             False           False  …               False
2             False           False  …               False
3             False           False  …               False
4             False           False  …               False
..              …               …  …                 …
78            False           False  …               False
79            False           False  …               False
80            False           False  …               False
81            False           False  …               False
82            False           False  …                True

    Merchant_MerchantJ  TransactionType_Transfer  TransactionType_Withdrawal  \
0                False                      True                       False
1                False                     False                        True
2                False                     False                        True
3                False                     False                        True
4                False                     False                        True
..                 …                         …                           …
78               False                     False                        True
79               False                      True                       False
80                True                     False                        True
81                True                      True                       False
82               False                      True                       False

    Location_Los Angeles  Location_New York  Location_San Francisco  \
0                   True              False                   False
1                  False              False                   False
2                  False               True                   False
```

```
3                  False               True                    False
4                  False              False                    False
..                   …                 …                       …
78                 False              False                    False
79                 False               True                    False
80                 False              False                     True
81                 False              False                     True
82                  True              False                    False

    Location_Tokyo     score  label
0            False -0.002557     -1
1             True -0.000929     -1
2            False -0.024142     -1
3            False -0.025608     -1
4             True -0.003382     -1
..             …       …    …
78            True -0.003382     -1
79           False -0.003294     -1
80           False -0.011881     -1
81           False -0.002576     -1
82           False -0.002820     -1

[83 rows x 34 columns]
```

[21]:
```
combined_anomalies
```

[21]:
```
             Timestamp TransactionID    Amount  AccountID_ACC10  \
0     01-01-2023 09:10        TXN656  0.935041            False
1     01-01-2023 10:23       TXN1891  0.968465            False
2     01-01-2023 11:30       TXN1648  0.748945            False
3     01-01-2023 13:48        TXN291 -0.973774            False
4     01-01-2023 15:16        TXN534 -0.916788            False
…                  …           …       …                …
4251  31-05-2023 13:51       TXN1578 -0.952660            False
4252  31-05-2023 15:25        TXN862  0.859373            False
4253  31-05-2023 17:55        TXN132  0.992980            False
4254  31-05-2023 20:38          TXN4  0.890300             True
4255  31-05-2023 22:00       TXN1318 -1.000534            False

      AccountID_ACC11  AccountID_ACC12  AccountID_ACC13  AccountID_ACC14  \
0               False            False             True            False
1               False            False             True            False
2               False            False             True            False
3               False             True            False            False
4               False             True            False            False
…                   …                …                …                …
4251             True            False            False            False
```

```
4252            False            False            False            False
4253            False            False            False             True
4254            False            False            False            False
4255            False            False            False            False

      AccountID_ACC15  AccountID_ACC2  …  Merchant_MerchantI  \
0               False           False  …               False
1               False           False  …               False
2               False           False  …               False
3               False           False  …               False
4               False           False  …               False
…                 …               …    …                 …
4251            False           False  …               False
4252            False           False  …               False
4253            False           False  …               False
4254            False           False  …                True
4255            False           False  …               False

      Merchant_MerchantJ  TransactionType_Transfer  \
0                   True                     False
1                  False                     False
2                  False                     False
3                  False                     False
4                  False                     False
…                    …                       …
4251               False                      True
4252               False                     False
4253               False                     False
4254               False                     False
4255               False                     False

      TransactionType_Withdrawal  Location_Los Angeles  Location_New York  \
0                           True                  True              False
1                           True                 False              False
2                           True                 False               True
3                           True                 False              False
4                           True                 False              False
…                            …                     …                  …
4251                       False                 False              False
4252                        True                  True              False
4253                       False                 False               True
4254                       False                 False              False
4255                       False                 False              False

      Location_San Francisco  Location_Tokyo     score  label
0                      False           False -0.001907     -1
1                       True           False -0.005280     -1
```

```
2                    False           False -0.005911     -1
3                     True           False -0.008430     -1
4                    False            True -0.002368     -1
…                    …              …       …    …
4251                 False            True  0.016850     -1
4252                 False           False  0.007211     -1
4253                 False           False  0.035176     -1
4254                 False           False  0.057230     -1
4255                 False           False  0.033391     -1

[4256 rows x 34 columns]
```

4 2. Why did you choose the given approach over other methods? Which other methods did you evaluate?

Both Isolation Forest(IF) and Local Outlier Factor(LOF) machine learning models were chosen to identify potential anomalies/fraudulent activities. IF is sensitive to global outliers and LOF is better suited to identify local outliers. Both types of outliers could be relevant for detecting anomalies in financial transactions.

A local outlier could be a transaction that has a high amount compared to other transactions from the same user, merchant, or location, but not compared to the overall distribution of transaction amounts. This could indicate a fraudulent transaction or a sudden change in user behaviour. Whereas, a global outlier could be a transaction that has a very low or very high amount compared to all other transactions in the dataset, regardless of the user, merchant, or location. This could also indicate an extreme event.

Morever, in the given dataset, the feature 'Amount' is severely skewed towards the right, and both IF and LOF can handle skewness well.

One-class SVM cannot handle imbalanced datasets well. Also, K-means is sensitive to outliers and noise in the data, as they may affect the cluster centroids and the assignment of the points. Although DBSCAN can find both global and local outliers, it may or may do so depending on the data and the user's definition of outliers. DBSCAN requires two parameters: epsilon and min_samples, which determine the density threshold and the minimum size of a cluster, respectively. Choosing the optimal values for these parameters can be challenging, as they depend on the scale and distribution of the data. That is why one-class SVM, K-means clustering and DBSCAN models were not considered.

5 3. What features did you consider to find potential fraudulent activities? How did you perform feature engineering to improve the model?

Features considered are 'AccountID', 'Amount', 'Merchant', 'TransactionType' and 'Location'.

Except 'Amount', all others are categorical variables and were transformed to dummies.

Feature scaling is generally not necessary for tree based models like IF. Even feature scaling for

LOF may not be necessary in this case as just one numerical feature 'Amount' is present. Moreover, 'Amount' is heavily skewed and feature scaling may alter the relative density or shape of the data, which can affect outlier detection. However, if the model gets trained on just one particular order of magnitude of the feature 'Amount', then it might not perform well on an unseen data containing a drastically different order of magnitude. So the feature 'Amount' was scaled using RobustScaler() function. The RobustScaler() function is also less sensitive to ouliers present in the feature as compared to say, StandardScaler().

6 4. Demonstrate using code and explain how would you predict the spend for all Transaction Types for the month of June.

[30]:
```
from statsmodels.tsa.arima.model import ARIMA

df1 = pd.read_csv("financial_anomaly_data.csv")
df1 = df1.dropna(axis=0, how="any")

# convert the 'Timestamp' column to datetime
df1['Timestamp'] = pd.to_datetime(df1['Timestamp'], format="%d-%m-%Y %H:%M")

# set 'Timestamp' as the index
df1.set_index('Timestamp', inplace=True)

# group by 'TransactionType' and resample to get monthly totals for each␣
 ↪transaction type
df1_monthly = df1.groupby('TransactionType').resample('M').sum()

# fit the ARIMA model for each transaction type
for transaction_type in df1['TransactionType'].unique():
    model = ARIMA(df1_monthly.loc[transaction_type, 'Amount'], order=(1, 1, 1))
    model_fit = model.fit()

# forecast the next month (June 2023)
forecast = model_fit.forecast(steps=1)
print(f'Forecast for {transaction_type}: {forecast}')
```

```
Forecast for Purchase: 2023-06-30    7.255925e+08
Freq: M, dtype: float64
Forecast for Withdrawal: 2023-06-30    7.283862e+08
Freq: M, dtype: float64
Forecast for Transfer: 2023-06-30    7.338594e+08
Freq: M, dtype: float64

C:\Users\pritamaxx\anaconda3\Lib\site-
packages\statsmodels\tsa\statespace\sarimax.py:966: UserWarning: Non-stationary
starting autoregressive parameters found. Using zeros as starting parameters.
  warn('Non-stationary starting autoregressive parameters'
C:\Users\pritamaxx\anaconda3\Lib\site-
```

```
packages\statsmodels\tsa\statespace\sarimax.py:966: UserWarning: Non-stationary
starting autoregressive parameters found. Using zeros as starting parameters.
  warn('Non-stationary starting autoregressive parameters'
```

7 5. How would you test the effectiveness of the model to unseen data?

In case of unsupervised learning, such as this problem, the effectiveness of the model to unseen data can be tested only if the ground truth is known. Then, the performance of the unseen/test set can be measured by metrics such as precision, recall, F1-score, or ROC curve. Cross-validation techniques, such as k-fold or leave-one-out can be used, to fit and evaluate the IF and LOF models on different subsets of the training data itself.