Capstone 1 Project : Data Wrangling
Fraud Detection in Mobile Payment Data
Springboard Data Science Career Track

Eric Cruz

The dataset is synthetic and was created for research purposes, and comes as a single file in csv format. See link to the Jupyter Notebook used to explore the data using pandas and matplotlib:

https://github.com/cruzer42/Capstone-Project-1-for-Springboard/blob/master/Capstone%20Project1%20Data%20Wrangling.ipynb

There was no wrangling required with respect to missing values, as determined by the following procedure:

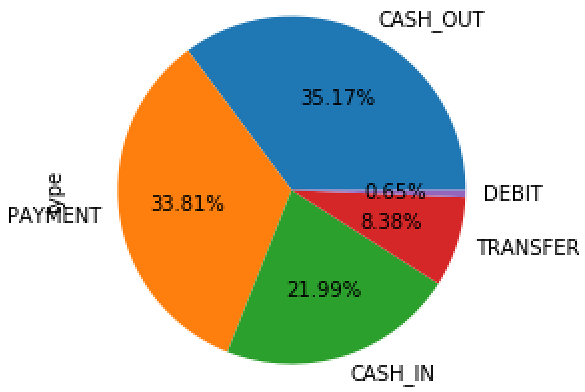
A) The csv file was converted to a pandas dataframe with pd.read_csv(), and examined with functions info(), head(), and describe().
B) A check for missing values was performed with pd.isnull().sum() and none were detected
C) However, zero values were examined and analyzed at a high level, as they may provide insights into the fraud detection patterns. In other words, zero balances have special meaning rather than indicating missing information.
D) A histogram for each column in the dataframe was generated to illustrate the imbalanced nature of the fraud case frequency in the data, and the time series frequency distribution.

Note the following description of each of the 11 columns, as well as observations about zero values and frequency distributions for each column as applicable:

1) step - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).
    a) Note that this is the dataset author's description posted on Kaggle, but the actual number of unique values is 743, with values 1 through 743.
    b) The highest frequency is around 50,000 transactions for that hour, and there is a cluster of other intervals with a similar number of transactions. The frequency range gets as low as 2 and the distribution of frequencies appears fairly smooth.
2) type - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
    a) See frequency of the 5 type values:

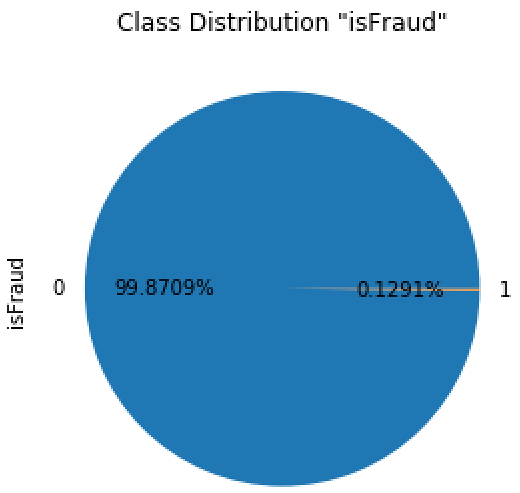
CASH_OUT	2,237,500
PAYMENT	2,151,495
CASH_IN	1,399,284
TRANSFER	532,909
DEBIT	41,432

b) A pie chart was generated using matplotlib to visualize the above distribution

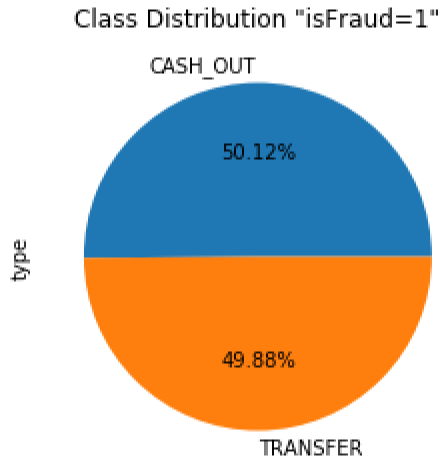


3) amount - amount of the transaction in local currency.
    a) Summary stats are included with the describe() output for all columns. The dollar amount ranges from about $92.5MM to zero, although the zero values appear to have special meaning for the 16 cases out of over 6MM observations. All are flagged as isFraud=1, so may need to be excluded or otherwise treated.
4) nameOrig - customer who started the transaction
    a) This column contains an account number, and the frequency of values was obtained using describe() on that column of the dataframe. The highest frequency for an account was 3
    b) The percentage of unique accounts is 99.8536 or 9,313 non-unique accounts out of 6,362,620
5) oldbalanceOrg - initial balance before the transaction
    a) The ratio of zero values is 0.33 or 2,102,449 of 6,362,620
6) newbalanceOrig - new balance after the transaction
    a) The ratio of zero values is 0.57 or 3,609,566 of 6,362,620
7) nameDest - customer who is the recipient of the transaction
    a) The top frequency for recipient accounts is 113, and there is a cluster of frequencies tailing off smoothly.
    b) The percentage of unique accounts is 42.7868 or 3,640,258 non-unique accounts out of 6,362,620
8) oldbalanceDest - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).
    a) The ratio of zero values is 0.43 or 2,704,388 of 6,362,620
    b) Need to classify orig and dest accounts by “M” and review zero values for each
9) newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
    a) The ratio of zero values is 0.38 or 2,439,433 of 6,362,620
    b) Need to classify orig and dest accounts by “M” and review zero values for each

10) isFraud - This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system.
    a) The ratio of zero values is 0.998709 or 6,354,407 of 6,362,620
    b) The number of isFraud cases is 8,213
    c) A pie plot was generated using matplotlib to visualize the distribution
    d) Note the distribution of isFraud=1 types is only among 2 categories, Cash Out and Transfer. These are split fairly evenly, but recall that only 8.4% of transactions are Transfer type, while 35.2% are Cash Out.
    e) Also note the 16 Amount=0, and the complete set of 16 isFlaggedFraud=1 may need special treatment or exclusion

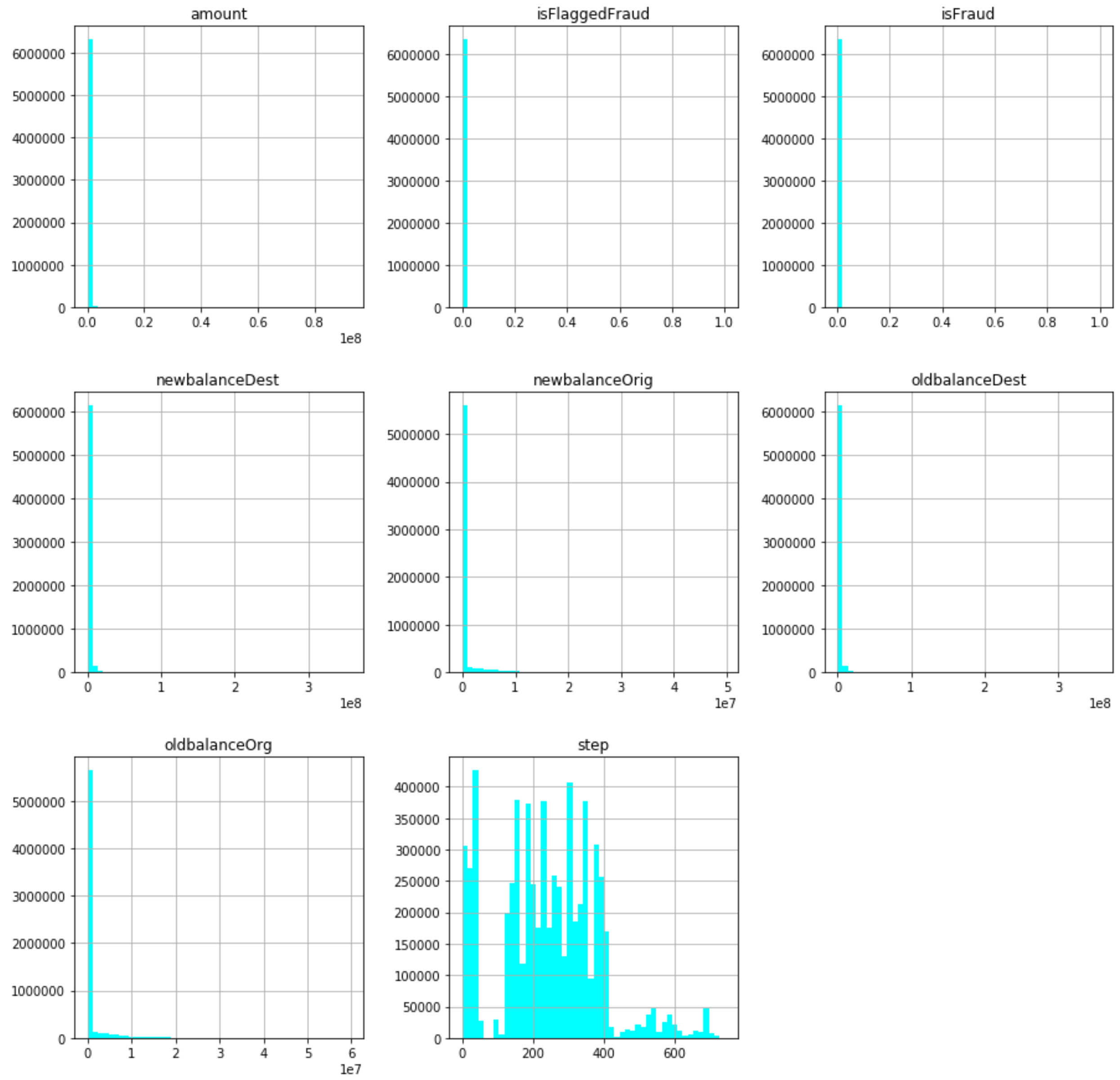


```
CASH_OUT    4116

TRANSFER    4097

Name: type, dtype: int64
```



11) isFlaggedFraud - The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.
    a) The ratio of zero values is 0.999997 or 6,362,604 of 6,362,620
    b) The number of isFlaggedFraud cases is 16, which coincidentally is the same number of transactions with Amount=0 and isFraud. Note that for each set of 16 uncommon values, they also have the feature isFraud=1. This may be a reason to exclude or otherwise treat these with special weight.
    c) All are type TRANSFER and the amounts have a pattern of matching “Orig” balances individually or in combination, and all “Dest” balances are zero.
    d) These relationships can be seen by showing head(16) on the 16 flagged values

Histogram function - Using the hist() function on the pandas dataframe and displaying with matplotlib generates a histogram for each column. Since the thing to be predicted has such a low frequency in the data, it is considered highly imbalanced. Appropriate modeling techniques should be selected taking the imbalance into account.


amount
6000000
5000000
4000000
3000000
2000000
1000000
0
0.0
0.2
0.4
0.6
0.8
1e8
isFlaggedFraud
6000000
5000000
4000000
3000000
2000000
1000000
0
0.0
0.2
0.4
0.6
0.8
1.0
isFraud
6000000
5000000
4000000
3000000
2000000
1000000
0
0.0
0.2
0.4
0.6
0.8
1.0
newbalanceDest
6000000
5000000
4000000
3000000
2000000
1000000
0
0
1
2
3
1e8
newbalanceOrig
5000000
4000000
3000000
2000000
1000000
0
0
1
2
3
4
5
1e7
oldbalanceDest
6000000
5000000
4000000
3000000
2000000
1000000
0
0
1
2
3
1e8
oldbalanceOrg
5000000
4000000
3000000
2000000
1000000
0
0
1
2
3
4
5
6
1e7
step
400000
350000
300000
250000
200000
150000
100000
50000
0
0
200
400
600