Mobile Money Transactions for Fraud Detection Research

CS 131 Project Team 6

Eric Zhao, Xiaoke Ran
Dao Cao, Nicholson Alforque

Introduction

Financial fraud poses a significant threat to the integrity of financial systems worldwide, necessitating robust mechanisms for detection and prevention. In pursuit of this goal, researchers and analysts rely heavily on datasets that accurately reflect the complexities of real-world financial transactions. However, such datasets are often scarce, limiting the efficacy of fraud detection methodologies.

To address this challenge, we present a novel dataset derived from the PaySim simulator, meticulously crafted to mirror the intricacies of genuine financial activities while incorporating fraudulent behaviors for research purposes. This synthetic representation of mobile money transactions offers a unique resource for studying and evaluating fraud detection techniques.

Dataset Description

The dataset encapsulates a month's worth of mobile money transactions, synthesized from aggregated data obtained from the financial logs of a mobile money service operating in an African country. Comprising various transaction types such as CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER, it spans a simulated period of 30 days, consisting of 744 time steps, with each step equivalent to one hour.

```
sed -n '1,1060436 p' Synthetic_Financial_datasets_log.csv > Financial_datasets_partial.csv
```

Meta Data:

```
Rows: 1,060,435
Columns: 11
$ step           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ type           <chr> "PAYMENT", "PAYMENT", "TRANSFER", "CASH_OUT", "PAYMENT"…
$ amount         <dbl> 9839.64, 1864.28, 181.00, 181.00, 11668.14, 7817.71, 71…
$ nameOrig       <chr> "C1231006815", "C1666544295", "C1305486145", "C84008367…
$ oldbalanceOrg  <dbl> 170136.0, 21249.0, 181.0, 181.0, 41554.0, 53860.0, 1831…
$ newbalanceOrig <dbl> 160296.36, 19384.72, 0.00, 0.00, 29885.86, 46042.29, 17…
$ nameDest       <chr> "M1979787155", "M2044282225", "C553264065", "C38997010"…
$ oldbalanceDest <dbl> 0, 0, 0, 21182, 0, 0, 0, 0, 0, 41898, 10845, 0, 0, 0, 0…
$ newbalanceDest <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 4…
$ isFraud        <dbl> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ isFlaggedFraud <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```

Dataset Structure

step: Represents a unit of time in the real world, with 1 step equating to 1 hour. The total simulation spans 744 steps, equivalent to 30 days.

type: Transaction types include CASH-IN, CASH-OUT, DEBIT, PAYMENT, and TRANSFER.

amount: The transaction amount in the local currency.

nameOrig: The customer initiating the transaction.

oldbalanceOrg: The initial balance before the transaction.

newbalanceOrig: The new balance after the transaction.

nameDest: The transaction's recipient customer.

oldbalanceDest: The initial recipient's balance before the transaction. Not applicable for customers identified by 'M' (Merchants).

newbalanceDest: The new recipient's balance after the transaction. Not applicable for 'M' (Merchants).

isFraud: Identifies transactions conducted by fraudulent agents aiming to deplete customer accounts through transfers and cash-outs.

isFlaggedFraud: Flags large-scale, unauthorized transfers between accounts, with any single transaction exceeding 200,000 being considered illegal.

Cmd page

```
froyal@EliteBook:~/Documents/Code/CS131/Project$ wc -l Financial_dataset_partial.csv
1060436 Financial_dataset_partial.csv
froyal@EliteBook:~/Documents/Code/CS131/Project$ head -n 10 Financial_dataset_partial.csv
step,type,amount,nameOrig,oldbalanceOrg,newbalanceOrig,nameDest,oldbalanceDest,newbalanceDest,isFraud,isFlaggedFraud
1,PAYMENT,9839.64,C1231006815,170136.0,160296.36,M1979787155,0.0,0.0,0,0
1,PAYMENT,1864.28,C1666544295,21249.0,19384.72,M2044282225,0.0,0.0,0,0
1,TRANSFER,181.0,C1305486145,181.0,0.0,C553264065,0.0,0.0,1,0
1,CASH_OUT,181.0,C840083671,181.0,0.0,C38997010,21182.0,0.0,1,0
1,PAYMENT,11668.14,C2048537720,41554.0,29885.86,M1230701703,0.0,0.0,0,0
1,PAYMENT,7817.71,C90045638,53860.0,46042.29,M573487274,0.0,0.0,0,0
1,PAYMENT,7107.77,C154988899,183195.0,176087.23,M408069119,0.0,0.0,0,0
1,PAYMENT,7861.64,C1912850431,176087.23,168225.59,M633326333,0.0,0.0,0,0
1,PAYMENT,4024.36,C1265012928,2671.0,0.0,M1176932104,0.0,0.0,0,0
froyal@EliteBook:~/Documents/Code/CS131/Project$ head -n 10 Financial_dataset_partial.csv | column -t -s
column: option requires an argument -- 's'
Try 'column --help' for more information.
froyal@EliteBook:~/Documents/Code/CS131/Project$ head -n 10 Financial_dataset_partial.csv | column -t -s,
step  type      amount    nameOrig     oldbalanceOrg  newbalanceOrig  nameDest     oldbalanceDest  newbalanceDest  isFraud  isFlaggedFraud
1     PAYMENT   9839.64   C1231006815  170136.0       160296.36       M1979787155  0.0             0.0             0        0
1     PAYMENT   1864.28   C1666544295  21249.0        19384.72        M2044282225  0.0             0.0             0        0
1     TRANSFER  181.0     C1305486145  181.0          0.0             C553264065   0.0             0.0             1        0
1     CASH_OUT  181.0     C840083671   181.0          0.0             C38997010    21182.0         0.0             1        0
1     PAYMENT   11668.14  C2048537720  41554.0        29885.86        M1230701703  0.0             0.0             0        0
1     PAYMENT   7817.71   C90045638    53860.0        46042.29        M573487274   0.0             0.0             0        0
1     PAYMENT   7107.77   C154988899   183195.0       176087.23       M408069119   0.0             0.0             0        0
1     PAYMENT   7861.64   C1912850431  176087.23      168225.59       M633326333   0.0             0.0             0        0
1     PAYMENT   4024.36   C1265012928  2671.0         0.0             M1176932104  0.0             0.0             0        0
```

wc -l : count how many lines
head -n 10: output first 10 lines
column -t -s: format as table
delimiter as white space

Easier to view what data we are working with

Cmd page

Fraud 1386 and non-Fraud 1059049

```
froyal@EliteBook:~/Documents/Code/CS131/Project$
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR > 1 && $10 == 1 {print}' Financial_dataset_partial.csv | wc -l
1386
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR > 1 && $10 == 0 {print}' Financial_dataset_partial.csv | wc -l
1059049
froyal@EliteBook:~/Documents/Code/CS131/Project$
```

Fraud mean and non-fraud mean

```
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR > 1 && $10 == 1 {print}' Financial_dataset_partial.csv | awk -F, '{sum +
= $3}; {mean = sum / 1386}; END{print mean}'
1.19823e+06
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR > 1 && $10 == 0 {print}' Financial_dataset_partial.csv | awk -F, '{sum +
= $3}; {mean = sum / 1059049}; END{print mean}'
157087
froyal@EliteBook:~/Documents/Code/CS131/Project$
```

Fraud amount total

```
froyal@EliteBook:~/Documents/Code/CS131/Project$
froyal@EliteBook:~/Documents/Code/CS131/Project$
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR > 1 && $10 == 1 {print}' Financial_dataset_partial.csv | awk -F, '{sum +
= $3}; END{print sum}'
1.66075e+09
froyal@EliteBook:~/Documents/Code/CS131/Project$
```

Cmd page

Fraud 1386 and non-Fraud 1059049

```
froyal@EliteBook:~/Documents/Code/CS131/Project$
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR > 1
1386
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR > 1
1059049
froyal@EliteBook:~/Documents/Code/CS131/Project$
```

Fraud mean and non-fraud mean

```
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR
= $3}; {mean = sum / 1386}; END{print mean}'
1.19823e+06
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR
= $3}; {mean = sum / 1059049}; END{print mean}'
157087
froyal@EliteBook:~/Documents/Code/CS131/Project$
```

Fraud amount total

```
froyal@EliteBook:~/Documents/Code/CS131/Project$
froyal@EliteBook:~/Documents/Code/CS131/Project$
froyal@EliteBook:~/Documents/Code/CS131/Project$ awk -F, 'NR
= $3}; END{print sum}'
1.66075e+09
froyal@EliteBook:~/Documents/Code/CS131/Project$
```

awk -F, 'NR > 1 && $10 == 1 {print} ' file.csv | wc -l:

Count the lines with value 1 in column 10

awk -F, 'NR > 1 && $10 == 0 {print} ' file.csv | wc -l:

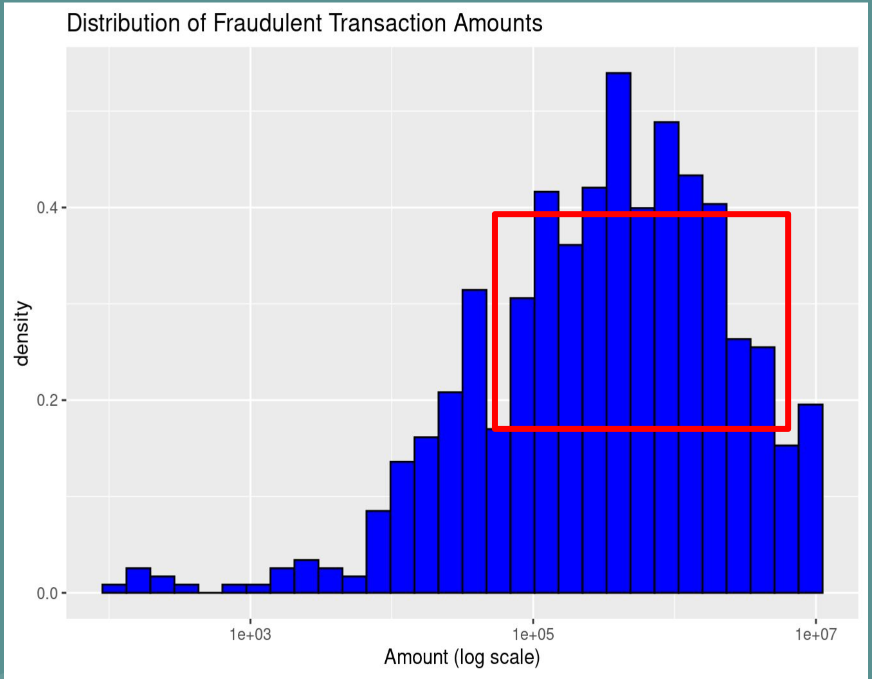
Count the lines with value 0 in column 10

{Sum += $3}; {mean = sum / numOfLines} ; END{print mean}

{Sum += $3}; END {print sum}

Comparison

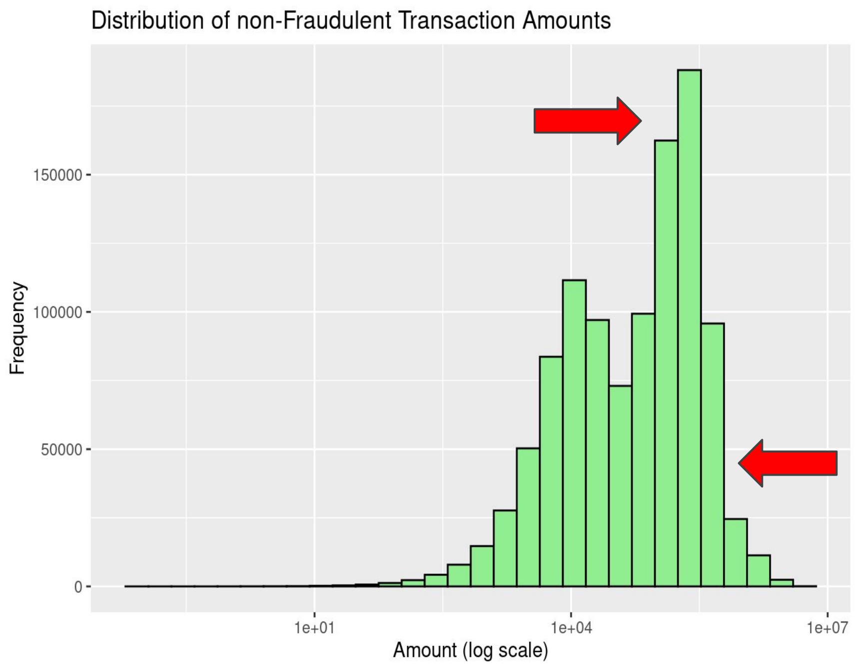
Fraud Transactions



- The histogram shows several peaks, which suggests that there are multiple common amounts for fraudulent transactions. Most of the density is concentrated in the middle range of the scale. This could indicate specific amounts that fraudsters frequently target are between 100,000-10,000,000

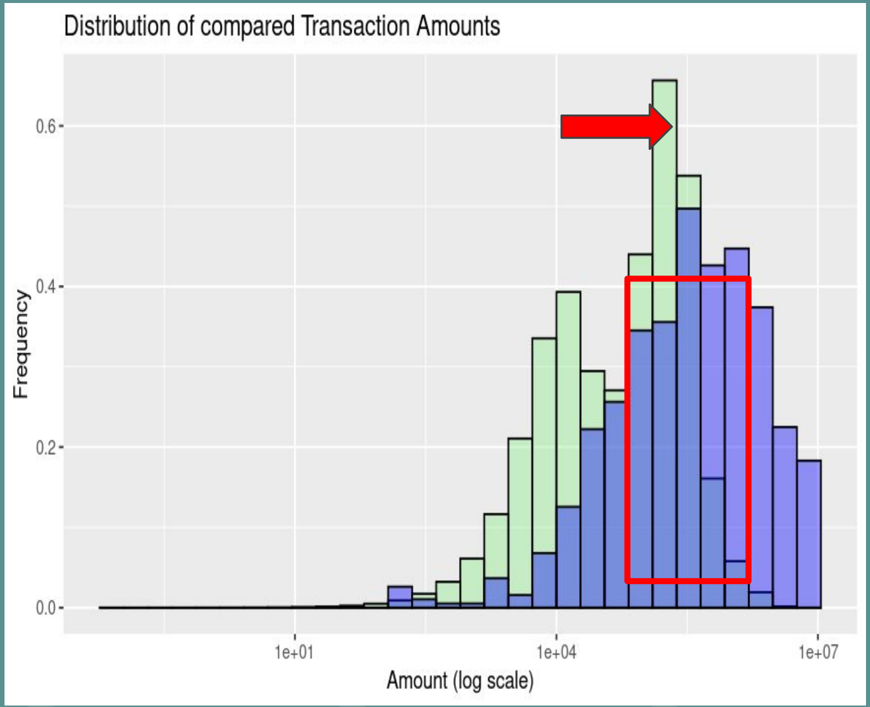
⟶ Fraudulent usually target to account with large balance.

non- Fraud Transactions



- The distribution amount arrange mostly from 1,000-1,000,000.
- The tallest peak occurs between approximately 100,000-1,000,000. This indicates that the most frequent non-fraud transaction amounts fall within this range.
- After the last peak, the number of high-value transactions gradually decreases rather than stopping suddenly. This indicates that while transactions involving large amounts are rare, they still happen and are generally not fraudulent.

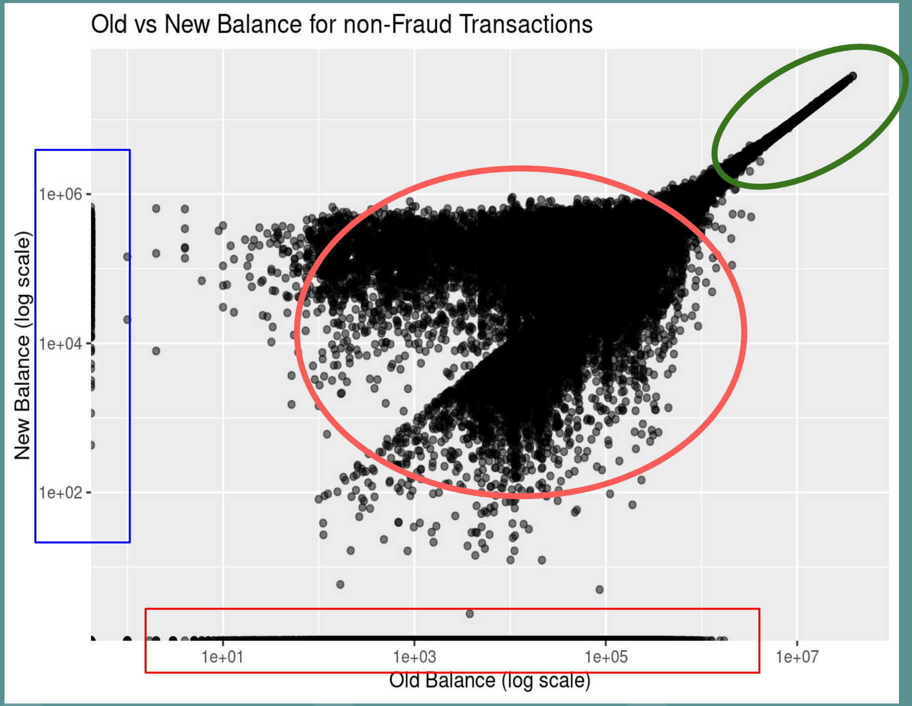
Compared Graphs



- In the graph, green is non-fraud, blue is fraud.
- The fraudulent transactions overlap non-fraudulent in the middle range of the graph suggesting that it is difficult to say that large transactions are fraudulent.
- Fraudulent transactions are less common in very low amounts. Small transactions are rarely targeted for fraud because the potential gain is not worth the effort.

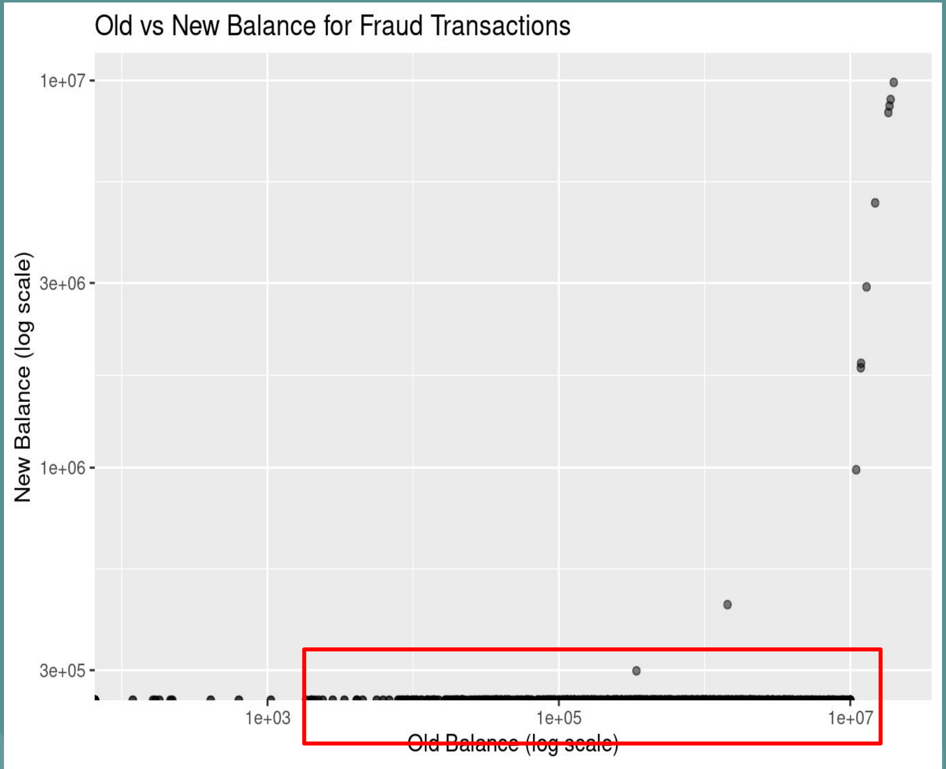
The graph shows a pattern close to normal distribution for both non-fraud and fraudulent transactions, but they often look similar to legitimate ones. This means more advanced methods are needed to correctly identify fraud. Identifying common amounts in fraudulent transactions could help improve fraud detection algorithms.

Old vs New Balance for Non-Fraud Transactions



- Explain plot in 4 different parts:
- Blue Box: most likely the transaction is account with no balance and deposit some amount.
- Red circle area: randomly deposit and withdraw transaction activity.
- Green circle area: large amount account does not have much change in balance after transaction.
- Red Box: account balance cleared out.

Old vs New Balance for Fraud Transactions



- This graph shows a bottom line of points which means the new balance changed significantly after the transaction. The balance was depleted to nearly zero indicate an easy-to-see sign of scams that is your account will be drained of money as soon as the scammer takes over the account.

- The density is concentrated in range between 10,000-10,000,000

Conclusion

- **Consistency in Fraudulent Behavior**: Fraudulent activities within the dataset are more predictable, with a higher chance of detection by the depletion withdrawn in high-balance accounts.

- **Fraudulent Transactions Amounts:** The preferred transaction amounts for fraud are less varied and tend to cluster around specific values, potentially indicative of fraudsters targeting certain amount thresholds to avoid detection.

- **Significance of Large Transactions in Fraud:** There is a notable presence of large transaction amounts in fraudulent activities, suggesting that when fraud occurs, it tends to be significant in size indicating that fraudsters commonly target accounts with high balances.

- **Challenges in Fraud Detection**: Despite these patterns, there is an overlap in the transaction amount distributions for fraud and non-fraudulent transactions, especially in the middle ranges, which complicates the identification of fraud based solely on the transaction amount.

Fraud Detection: Effective fraud detection strategies must consider both the amount and the impact on account balances. The marked depletion of balances following certain transactions is a strong fraud indicator.

Consistency in Fraudulent Behavior

Fraudulent Transactions Amounts

Significance of Large Transactions in Fraud

Challenges in Fraud Detection

Fraudulent Detection

Thank you.

References

Financial Fraud Detection Dataset:
https://www.kaggle.com/datasets/sriharshaeedala/financial-fraud-detection-dataset/data

An In-Depth Look at the Fraud Investigation Process:
https://financialcrimeacademy.org/the-fraud-investigation-process/

Financial Fraud Enforcement Task Force (FFETF):
https://www.fincen.gov/financial-fraud-enforcement-task-force-ffetf

Detecting Financial Statement Fraud:
https://www.investopedia.com/articles/financial-theory/11/detecting-financial-fraud.asp