Fraudulent Transactions

Task Develop a model for predicting fraudulent transactions for a financial company and use insights from the model to develop an actionable plan. Data for the case is available in CSV format having 6362620 rows and 10 columns.

Different Steps

- **1. Importing libraries and importing data**
- **2. Data Cleaning**
- **3. Data Visualization and EDA**
- **4.Feature Engineering**
- **5.Data Processing**
- **6.Model**
- **7.Results**

1. Importing libraries and importing data

- Import all required libraries:

❖Pandas

It **provides ready to use high-performance data structures and data analysis tools**.

❖NumPy

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

❖Matplotlib.pyplot

Matplotlib is a python library used to create 2D graphs and plots by using python scripts.

❖Seaborn

It is used for data visualization and exploratory data analysis.

2. Data Cleaning

- **Data cleaning** is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

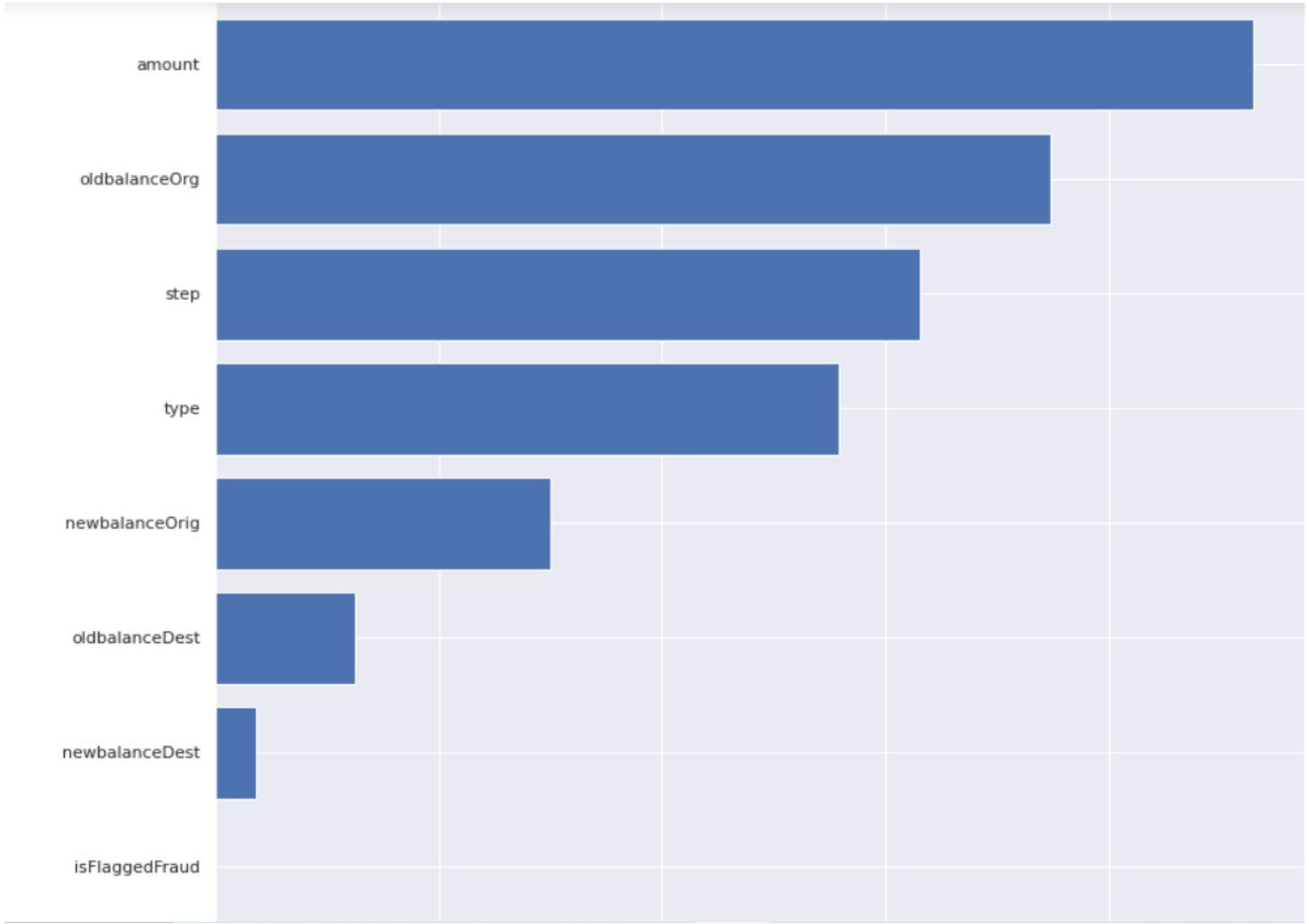
- After figure out the data I have dropped two columns ['nameOrig', 'nameDest’] from dataset.

3. Data Visualization and EDA

- Exploratory Data Analysis (EDA) is a process of describing the data by means of statistical and visualization techniques in order to bring important aspects of that data into focus for further analysis

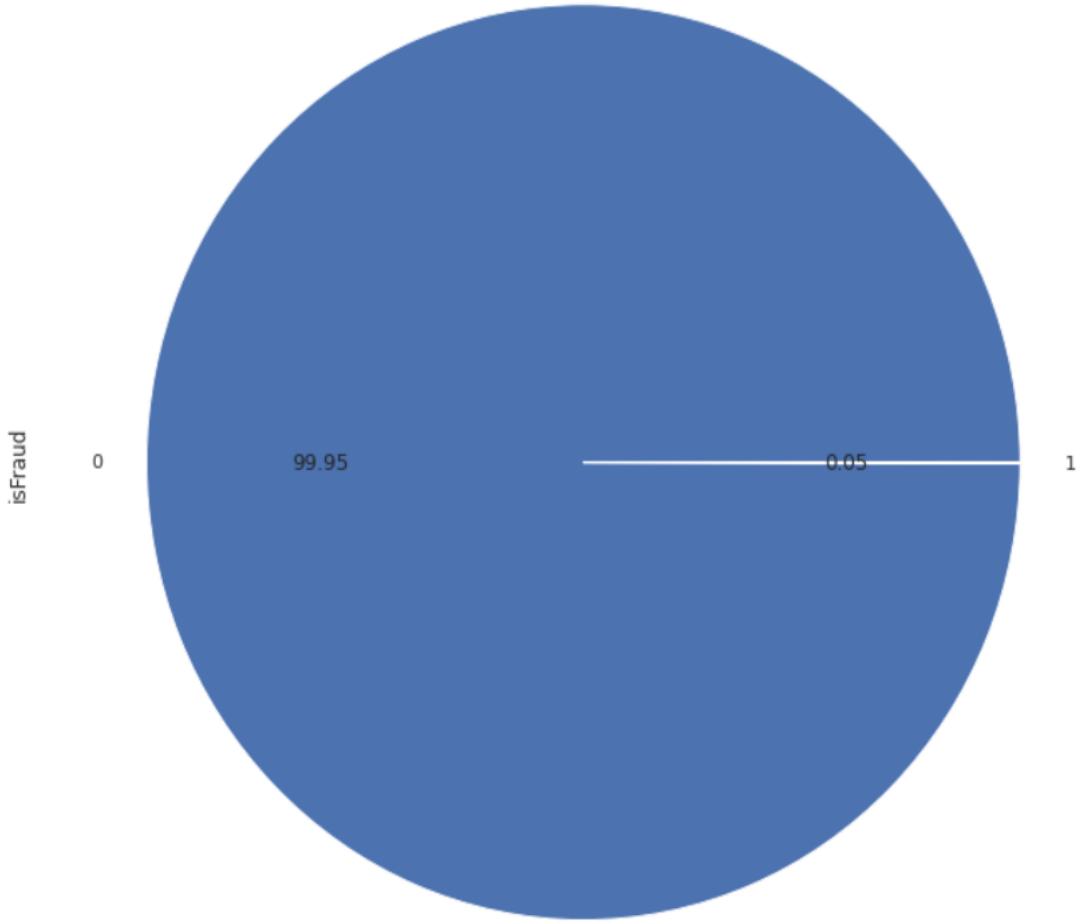
visualize the mutual information scores

- As we can see the amount feature tells the most about the transaction being freudulent or not but the isFlaggedFraud does not so it might be better to just drop that feature

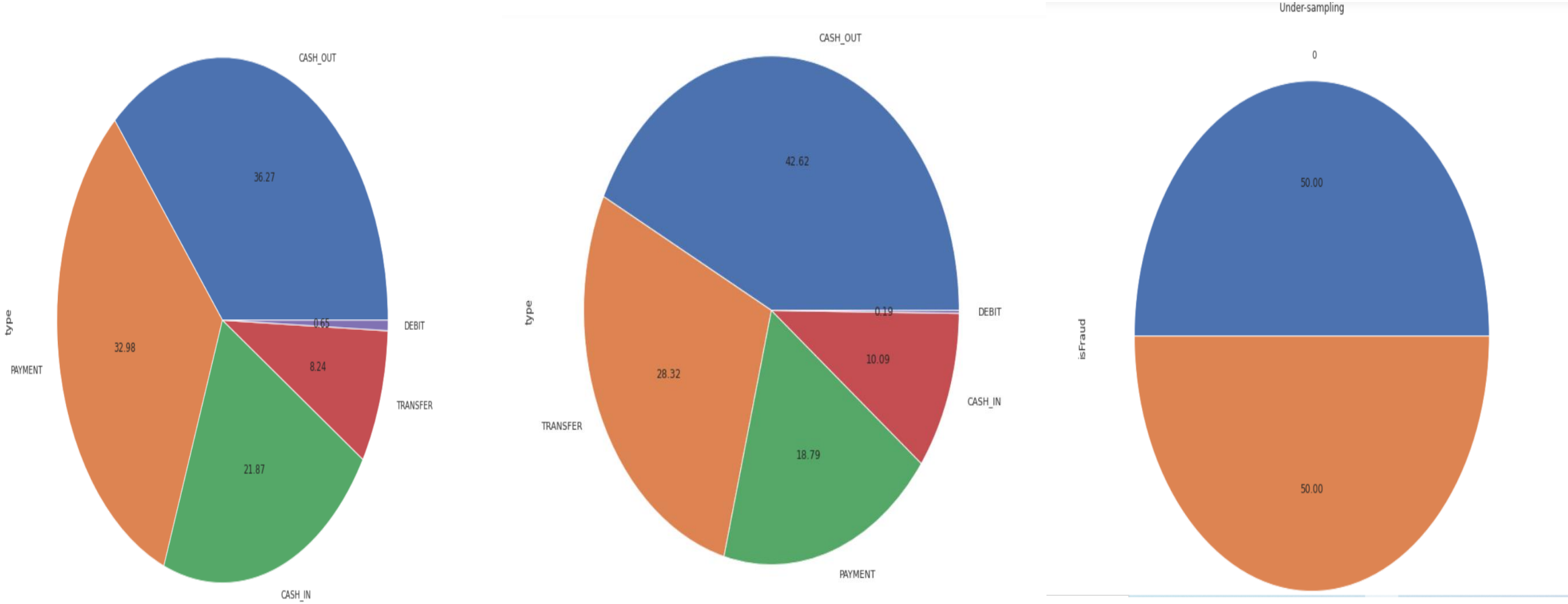


distribution of the isFraud feature

- This is just a proportion of the data but we can assume that the other 500,000 chunks behave more or less the same. So unless the features correctly describe a fraudulent transaction we can expect some bias in our model



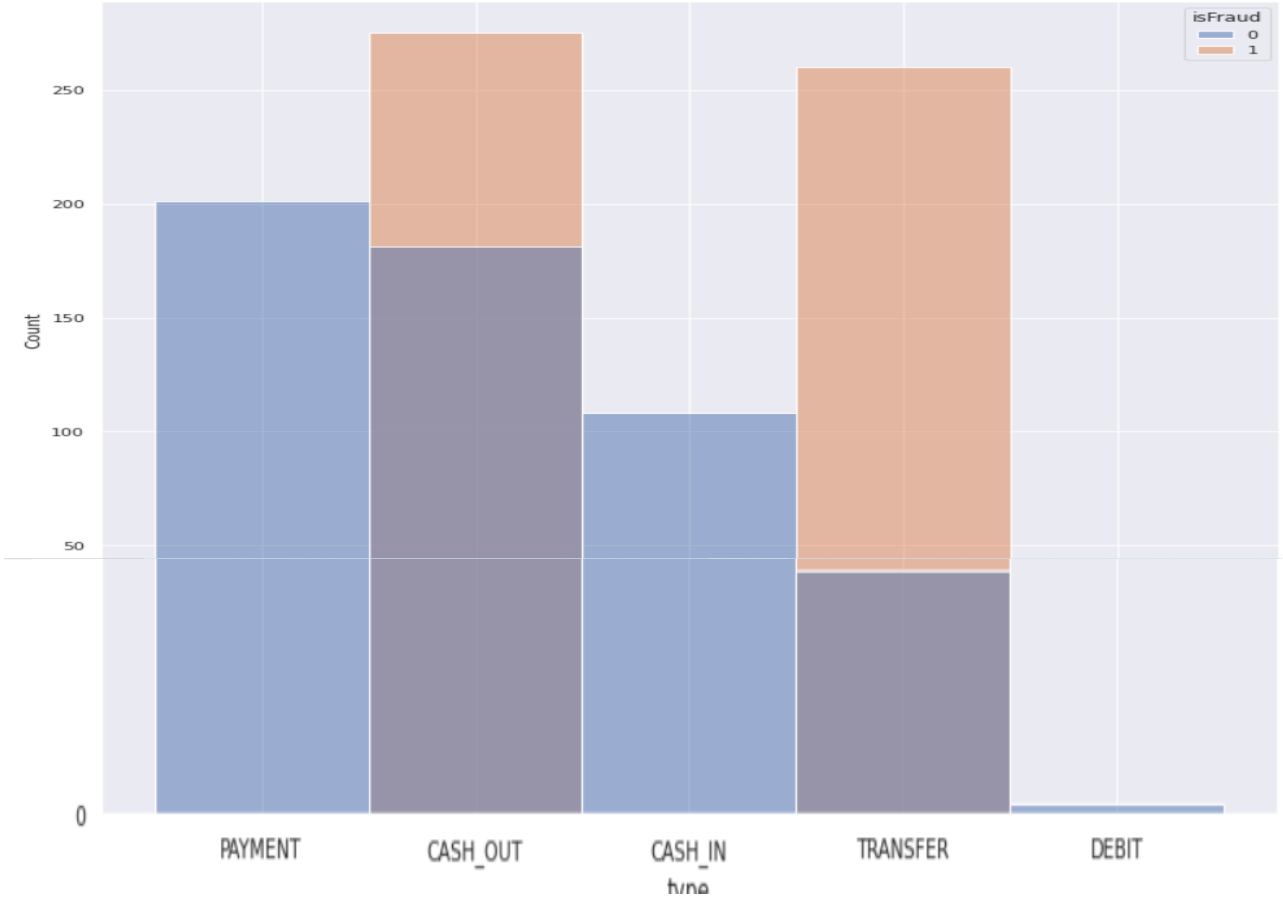
Balance the data out for EDA purposes



- As we can see the balanced data does a good job of representing the imbalanced data because they both represent a wide set of entries

Hist plot for fraudulent transactions

- We can see that almost no fraudulent transactions happen by CASH_IN or PAYMENT or DEBIT. They primarily happen by CASH_OUT and TRANSFER. This does make sense as one would try to tranfer the money to their offshore account for example or just take the cash out directly for a less chance of someone tracing the fraud back to them



Pandas_profiling library to generate some useful insights

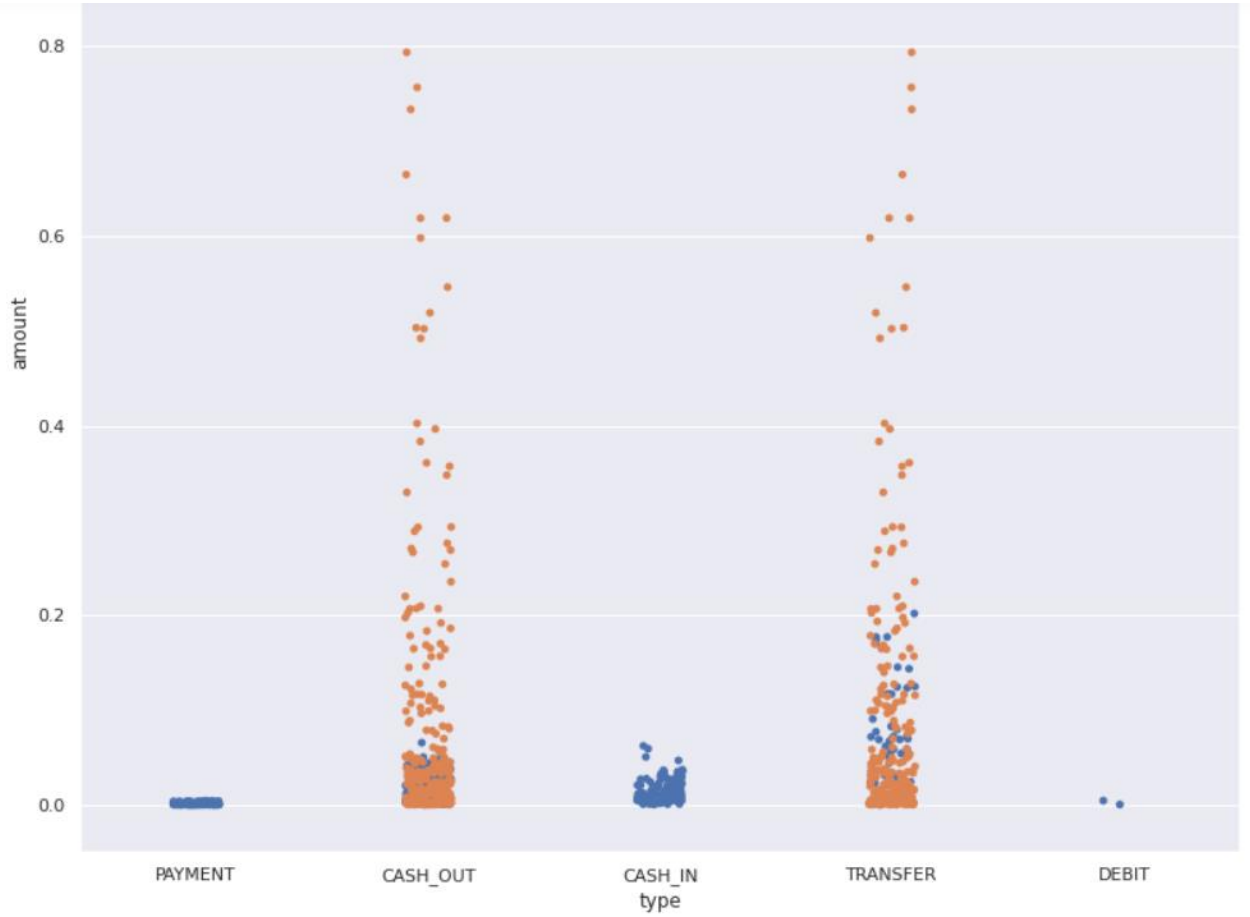
```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

amount Vs isfraud

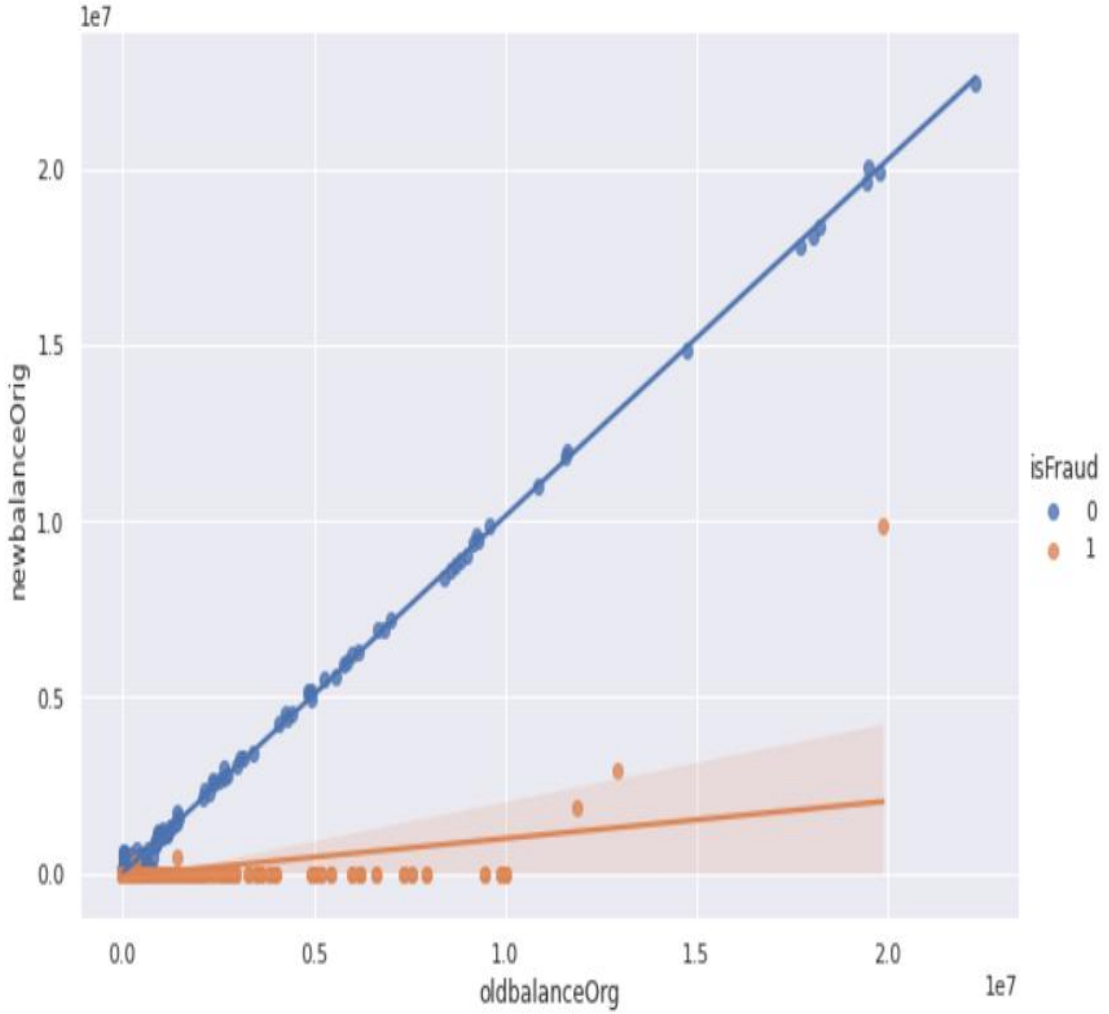
- Here we can clearly see that the fraudulent transactions tend to come at lower amounts but also very high amounts. This shows that people who commit fraud usually will also try to steal a higher amount at one time. Maybe they think that once they do it, they won't be able to do it again using the same accounts so they better get a big amount of money. People who do not commit fraud don't need to do high amount transactions because they can always go and do another transaction, that is why their transaction amounts are lower in comparison to people who commit fraud.

Type of Transaction Vs Amount

- **Here we can again see the types of transactions used for fraud as we've seen in the graph above but we can also see the amount of a transaction in regards to the type of the transaction. We already knew that Cash and Transfer transactions were the types that were used in fraud but we can also see that because those are the ones used for fraud that they are also the types that include the highest money amount**



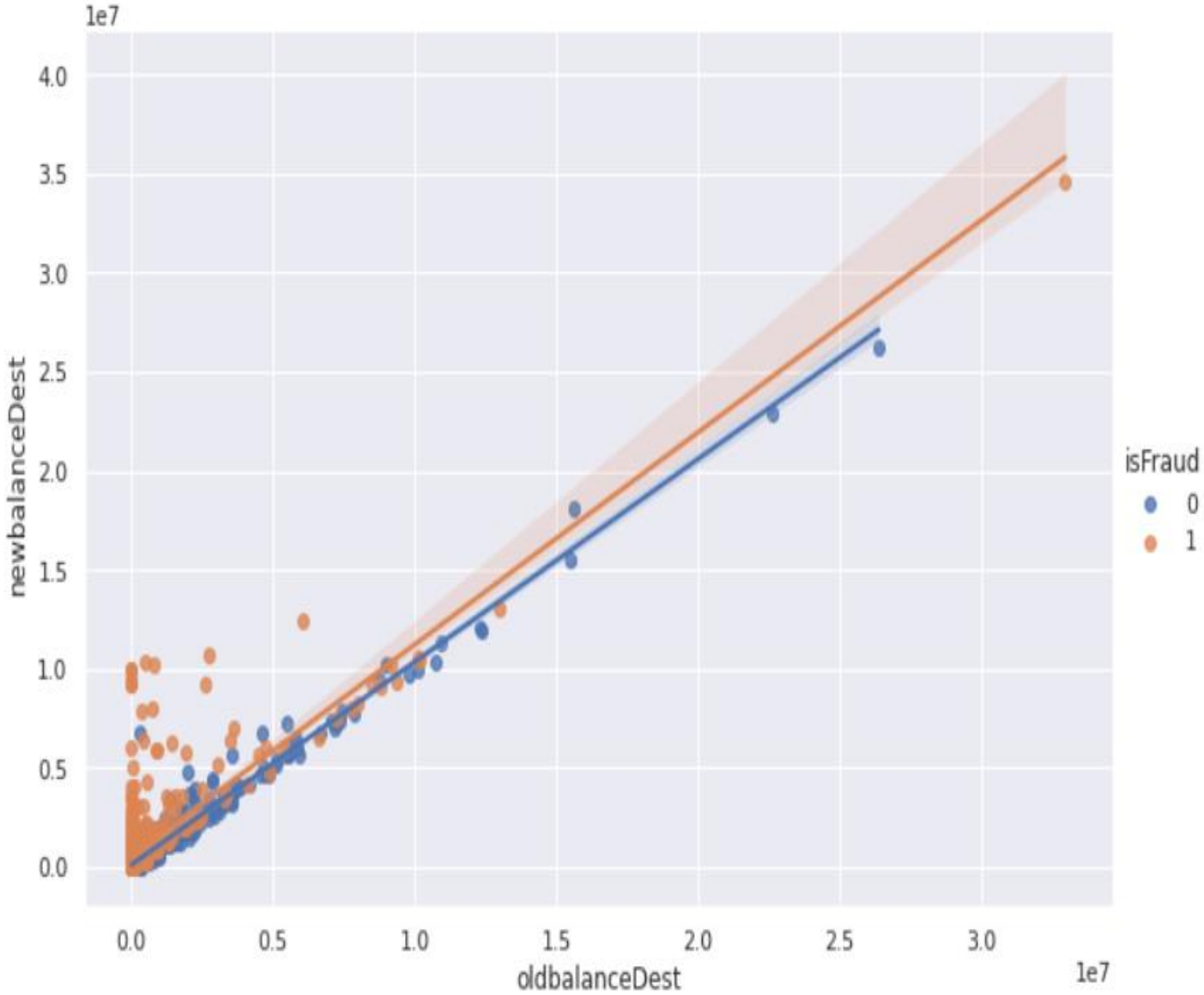
- **Because of fraudulent activity we can see that most types of fraud will not reflect on the new balance of the account owner**

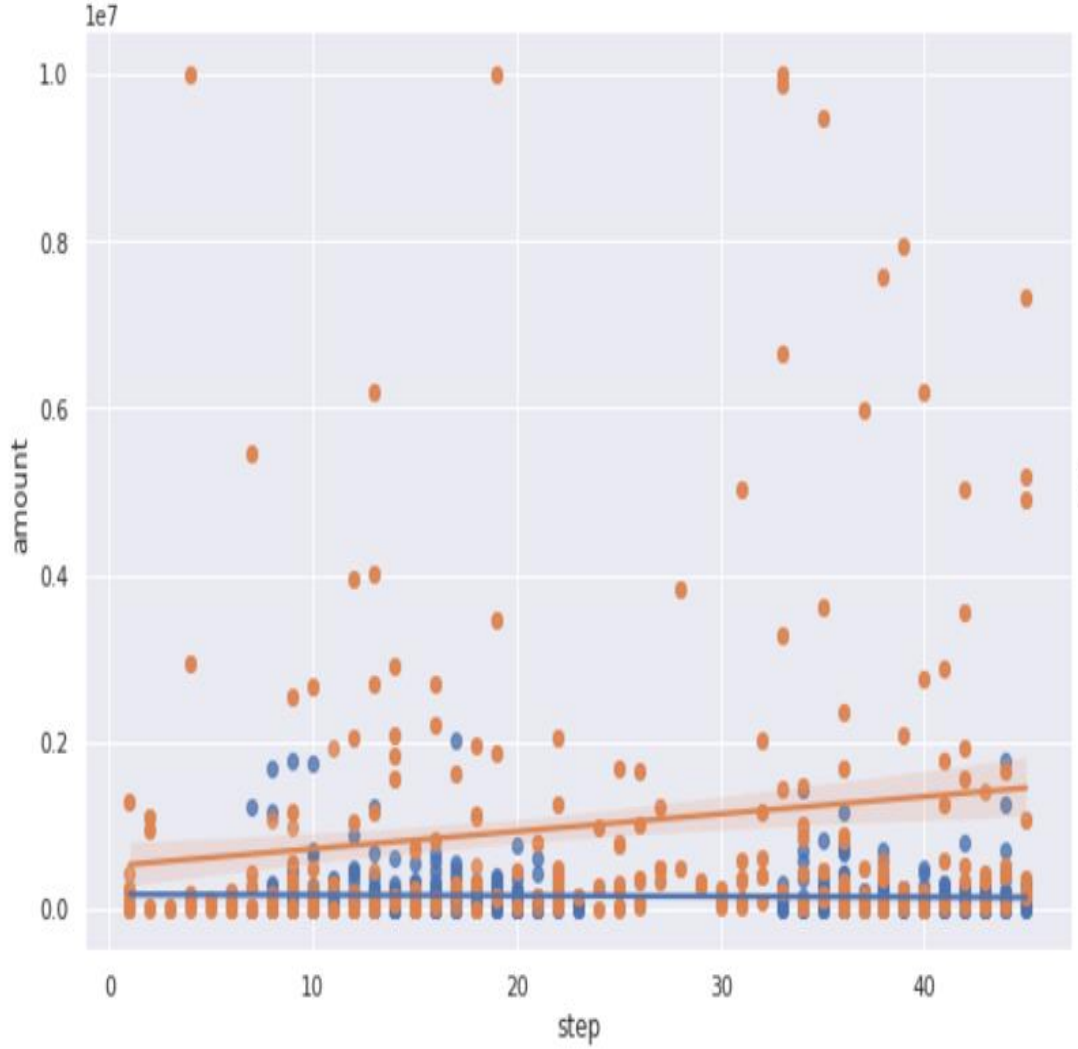


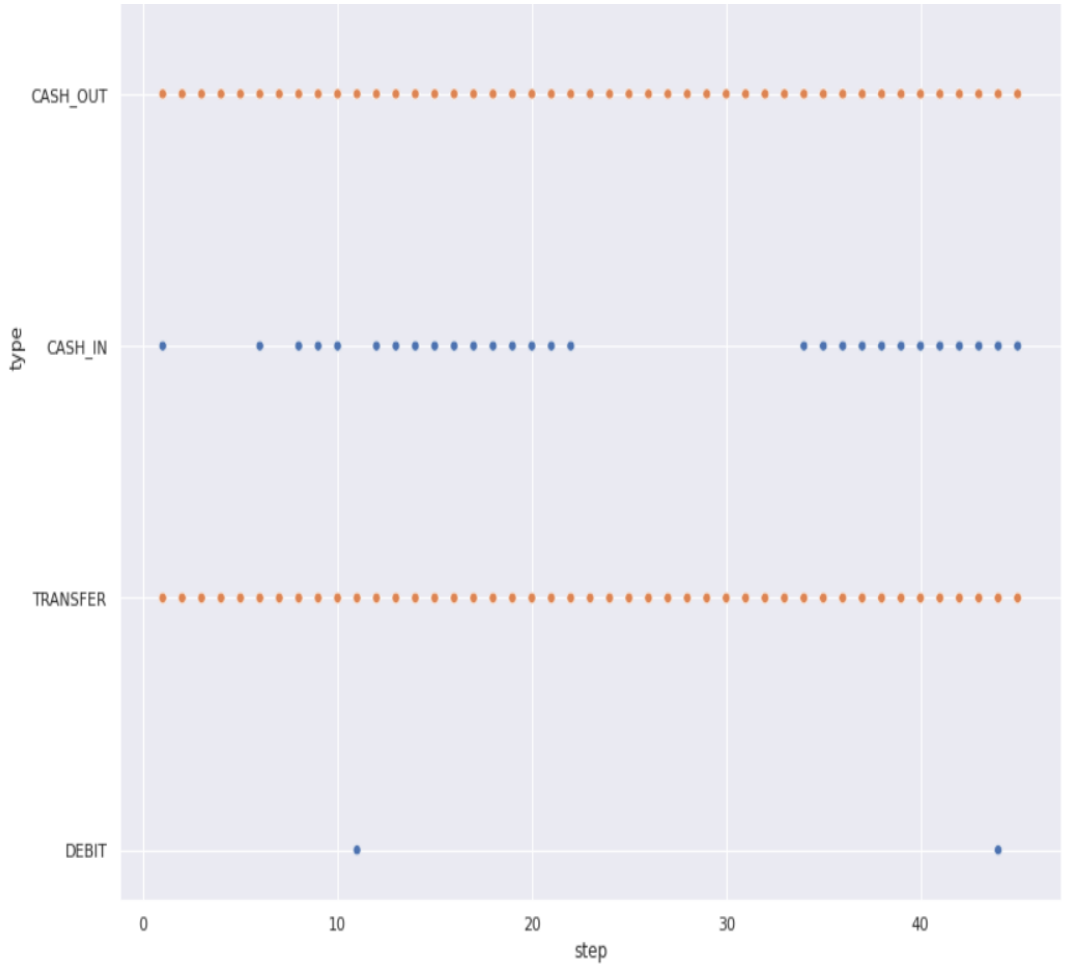
- But here you can see that fraudulent activity will not be reflected on the thiefs old account balance because they stole the money

- Now....something I find a bit odd is the Step variable. The step might have a hight Mutual Information (MI) score but that doesn't mean that it's useful. So lets look at it in a better light

- By definition the 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).







- **There doesn't seem to be a correlation. Because why would there be ? Fraud isn't predictable by time, is it ? So lets drop this column**

Look onto Outliers

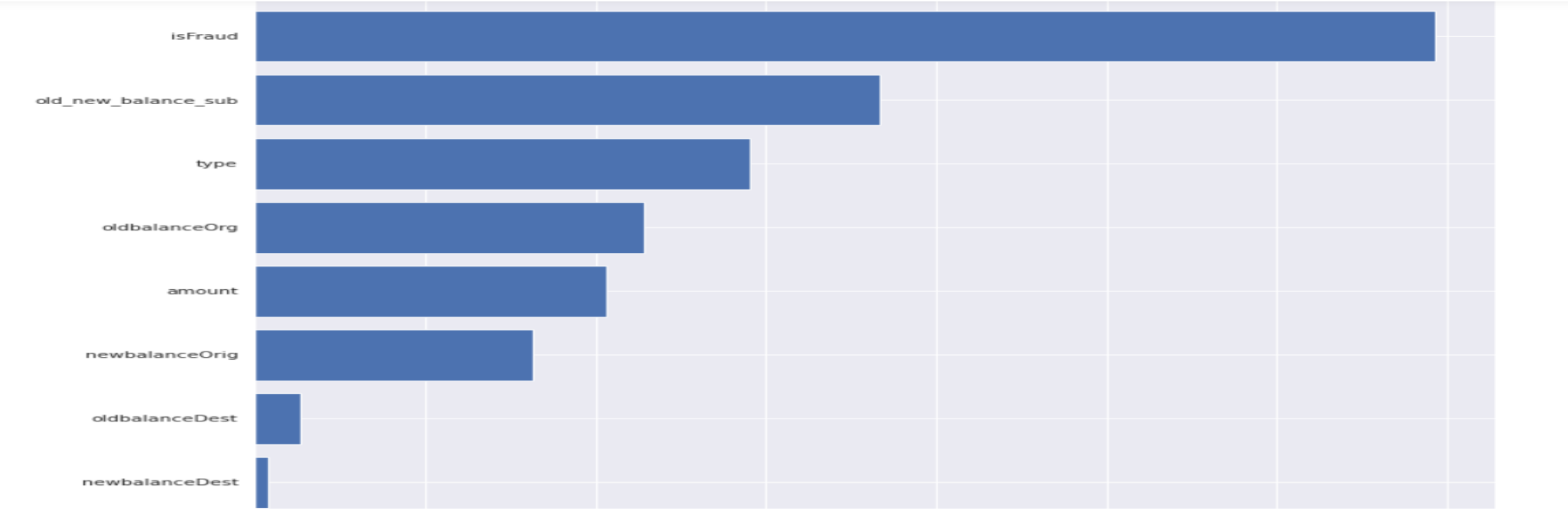
While there are some outliers here most of them look like they are fraud so we''ll keep them

4.Feature Engineering

- Feature engineering is **a machine learning technique that leverages data to create new variables that aren't in the training set**.

- **I am specifically going to look at oldbalanceOrg and newbalanceOrig to see if we can maybe get the ratio or subtraction as a usefule new feature**

calculate the MI scores



As we can clearly see we have successfully created a new feature that has a large effect on the dependant variable

5.Data Processing

- It is a series of calculations or actions that a computer performs on a given set of data to produce a desired result.

- **when data is collected and translated into usable information**.

6.Model

- **The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from**.

- A machine learning model is **a file that has been trained to recognize certain types of patterns**

7.Results

- **Using the SGDClassifier we were able to reach a good model performance but this was only tested with a chunk of the data**
- **Now let's have a final walkthrough of what we did:**

1. We loaded the data and tried to understan it
2. We cleaned the data and removed faetures that were unimportant
3. We visualized the data which helped us further understand the data at hand
4. We generated a new feature which turned out to be a great predicter
5. We built a model that classified the fraudulent transactions
6. Imbalanced Data leads to a very low recall score which means that the model fails to classify the fraudulent Transactions so we will have to find a way to balance the data. Right now we do Random Undersampling to achieve that which gets us a good result in terms of accuracy and recall but we loose a lot of data