Dataset Analysis

Dataset

Before performing data analysis we did basic cleaning of the data such as dropping NaN values and duplicate rows however, we found none even in the initial uncleaned dataset. There are 9 features and 1 label in total with 6 features being numerical, 3 being nominal, and the label being a binary classifier. The dataset is large with 6.3512×10^{06} rows of data.

Classifier

The primary observation when observing the dataset is that the label has disproportionately more of the not fraud class than the fraud class (99.8% vs 0.2%). Perform balancing of the two classes will be needed so that the fraud class is not buried. It is also worth noting that all of the fraud classes occur at lower dollar amounts and higher step values.

Numerical Features

All of the numerical features are technically right skewed. This is primarily because the features are money and thus have no cap on how high it can go but cannot go below 0. This leads to many values that are far higher than the rest (outliers) causing a skew when in reality, the majority of the values are closer to each other than would be indicated by the level of skew. This is a consistent aspect of all the numerical features except “step”. “step” is right skewed as well but most of its values are more widely distributed in general, never really reflecting a normal distribution.

Nominal Features

Of the 3 categorical features, the type features has 5 different values with two making up the majority of the feature (cash_out and payment). The other 2 features - nameOrig and nameDest - are highly cardinal with a lot of distinct values. This makes sense as these two features are more of identifiers of individuals in the dataset than actual data values.

Correlation

There are some correlated features in the dataset, some of which are very highly correlated: newbalanceOrig X oldbalanceOrig - 1.00; newbalanceDest X oldbalanceDest - 0.98; amount X newBalanceDest - 0.46; amount X oldbalanceDest - 0.29. These correlated features all make sense as they all have a cause-effect relationship on the features they are correlated with (the old balance will obviously affect the new balance). Other features have little to no correlation.

Analysis done with the Dataprep python Library:

Jinglin Peng, Weiyuan Wu, Brandon Lockhart, Song Bian, Jing Nathan Yan, Linghao Xu, Zhixuan Chi, Jeffrey M. Rzeszotarski, and Jiannan Wang. DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python. *SIGMOD 2021*.