Abstract

In this competition, we aimed to understand the dataset and identify the most important features for characterizing and flagging fraudulent transactions. We decided to use as many features as possible for a more holistic approach to the problem. After analyzing various methodologies, we decided to use a K-nearest neighbours (KNN) model for our final product. One of the challenges we faced was preprocessing the data, as KNN works best with continuous and discrete features. We encoded categorical and Bernoulli features and used a robust scaler to account for outliers. We also used a PCA transform to reduce dimensionality while retaining 95% of the original variability of the data.

We also considered the bigger picture of fraud, as it is not just data to explore but also a struggle that people and businesses face. We investigated the benefits of creating an accurate model and the real-life differences it can make. For example, banks spend millions of dollars on extracting and storing data, so a model can be more inexpensive. Managing fraud detection also comes down to managing trust between customers and businesses, and ensuring a smooth experience by only declining suspicious transactions and contacting customers for confirmation.

In conclusion, further research is needed to highlight which features are most telling of a fraudulent transaction. From the data given and external resources, we believe that location, time, transaction ID, and flag type are the best indicators. By having a greater understanding of which features matter, we can create better models with reduced noise that can help prevent fraud altogether.