Group Number: 16
Names: @Sawyer Mckenney, @Samuel Meaux, @David Savić, @Abdirahman Abdi

1) Project Description

The goal of this project is to evaluate and compare supervised and unsupervised fraud detection methods for identifying fraudulent financial transactions under extreme class imbalance. Specifically, we aim to answer the question "How do supervised fraud detection models compare to unsupervised anomaly detection methods in identifying rare fraudulent financial transactions and under what conditions does each approach perform best?"

Fraud detection appears to be a big problem in the finance world, where undetected fraud can lead to significant monetary losses, while false positives can negatively impact customer experience. Financial institutions, payment processors, and online merchants can all benefit from effective fraud detection systems. By comparing supervised and unsupervised approaches, this project provides insight into trade-offs between model accuracy, interpretability, and reliance on labeled data, which is often scarce or costly to obtain.

At a high level, we will perform predictive modeling and anomaly detection analysis using transaction-level data. Supervised learning methods such as logistic regression, random forests, and gradient boosting will be trained using labeled fraud data. In contrast, unsupervised methods (working on these – outliers?) will be applied to identify anomalous transactions without using labels. This will also help us identify patterns that we may not have previously noticed in the labeled data. Model performance will be systematically compared to understand the strengths and limitations of each approach in real-world fraud detection scenarios.

2) Data Sources

Credit Card Fraud Detection
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud (284k transactions with 30 features and binary fraud label)

The Fraud Dataset Benchmark (FDB):
https://github.com/amazon-science/fraud-dataset-benchmark (This is a compellation of data sets that we found on amazon science)
Based on our limited research, common variables used to detect fraud include transaction amount, transaction type, and currency.

3) Potential Bias / Limitations (Required)

After opening the data, we found that out of nearly 285,000 transactions, only 492 are labeled as fraud. That's less than 1%. If the model simply marks everything as "not fraud," it would have a 99.8% accuracy rate while failing its primary mission. To mitigate this, we will move away from "Overall Accuracy" and instead measure the True Positive Rate. This allows us to ignore the 99.8% of safe transactions and focus strictly on how many of the 492 known fraudulent cases we detected.

This data set has hidden the real names of the columns (using PCA) so to the naked eye it is just a jumbled mess. We lose the ability to have a real person in the loop to use common sense to catch things that a model wouldn't or to identify edge cases to build into the program. A $500 charge at a bar might not be flagged by a model, but a normal person knows $500 at a bar is incredibly high (usually). To mitigate, we can rank the columns to see which ones are the most "active" when fraud happens. Even if we don't know exactly what a particular column label is, we can at least identify that it's a primary "tripwire" for catching a fraudster.

This data is from 2013, before the everyday use of AI. Bad actors adapt quickly with the introduction of new technologies. Tactics, techniques, and other procedures that were common in 2013 are likely to be no longer relevant. A model trained on this data may only be relevant to 2013-style fraud and not relevant to today's fraud. For mitigation, we are treating this as a "proof of concept" to show our logic and system work. We'll acknowledge in our report that for a real-world mission, we would need to test this against a more recent dataset to see if these old patterns still hold up.