Problem Statement

You are provided with a mock-up dataset that simulates a simplified fraud detection project you will encounter on the job on a daily basis.

The object is to develop a machine learning model to detect frauds using the information in the dataset. Each true positive benefits the companies by its transaction amount. Each false positive costs the company $200 because it takes time to investigate. There is no restriction on methodologies you can use. The success of your project will be evaluated based on the following criteria:

1. You need to perform data analysis on the raw data and present with at least one data visualization.
   * What insights can you derive from the raw data?
   * Why is the project worthwhile?
   * What are the potential challenges you may be facing during the model development process?
2. Define proper performance metrics and provide business and statistical rationales of your selections.
   * Why are these metrics appropriate?
3. Derive at least one new features from the data.
   * Why can these features potentially help your model?
4. Set up an experimental framework and perform hyperparameter optimization.
   * Does your design lead to a good approximation to the performance in production in the future?
5. Perform a post-modeling analysis to convince the interviewers about the usability of your model.

* What is the model performance?
* What are the key features that drives the predictions?
* What threshold do you propose to use for the model and why?