<https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8>

The dataset from Kaggle https://www.kaggle.com/mlg-ulb/creditcardfraud can actually give us an insight into the specific characteristics that make a transaction more or less likely to be fraudulent. The data set has 31 features, 28 of which have been anonymized and are labeled V1 through V28. The remaining three features are the time and the amount of the transaction. This dataset is basically a classification problem with two classes, it takes value 1 in case of fraud and 0 otherwise.

In order to create an accurate model to solve this problem, we need a huge amount of data about different fraudulent and non-fraudulent transactions. So, I believe this problem lends itself to a data-driven solution.

To analyze the data, we can do exploratory data analysis. I will also run a simple decision tree or Naive Bayes classifier to get an understanding of the distribution of the data. This will also give me an idea about if the data is skewed or not and to check missing values. After this, I would try PCA for dimensionality reduction and for handling any noisy data if it is present. By projecting the data set into 2-D space, we will be able to produce a scatter plot showing the clusters of fraudulent and non-fraudulent transactions. I will use logistic regression to make predictions. I would also try multi-layer perceptron to improve accuracy.