### Project-1 Proposal: Credit Card Fraud Detection (Zaki Alawami)

#### Question/need:

* **How to determine if a credit card transaction is fraudulent or genuine?**

I will analyze a credit card transactions dataset, highlight any possible insights, and build a predictive classification model to predict if a transaction is possibly fraud or valid.

* Credit card companies, Banks, and Merchants can save billions of dollars annually that is lost to credit card fraud using such a model. Also, and more importantly credit card holders and customers will have higher level of trust in their credit card providers if they can determine and stop fraudulent transactions before processing the charges.

#### Data Description:

* The following Kaggle dataset will be used:

[Kaggle credit card fraud dataset](https://www.kaggle.com/pardhasaradhireddy/credit-card-fraud-detection-models/data/)

* + - fraudTest.csv
    - fraudTrain.csv
* The dataset has 21 columns/features, one target (valid/fraud), and is about 1.3 million records (hopefully this will work on my laptop without crashing!)
* This is sample record/transaction:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| trans\_date\_trans\_time | cc\_num | merchant | category | amt | first | last | gender | street | city | state | zip | lat | long | city\_pop | job | dob | trans\_num | unix\_time | merch\_lat | merch\_long | is\_fraud |

* I suspect that features like transaction time, transaction amount, and maybe the merchant details as opposed to the customer buying history may provide some indications that the transaction is fraudulent.
* The problem with this dataset is that it is heavily biased/imbalanced, as the number of training fraud records is very small (0.5% of total records), so I am not sure if this will be sufficient to create a successful model!
* The model will try to predict if a given transaction is fraudulent or not (Classification modeling problem)

#### Tools:

* I will utilize Python, Jupyter, Pandas, Matplotlib, Plotly, and Scikit-learn.
* I will add additional tools as necessary and as discovered in the process.

#### MVP Goal:

* A Jupyter Notebook with some EDA Statistics/charts describing the dataset
* An initial (rudimentary) classification predictive model (or at least data preprocessing necessary to train a first version of the model)