**Capstone 2 Proposal**

**Predicting likelihood of fraud transactions in an imbalanced dataset.**

Financial frauds are a wide-ranging term for theft including involving a payment card, such as a credit card or debit card or loans. In the United States, the ratio of the financial fraud happening is limited to %0.1 which means that 99.9% of the transactions are safe. This still causes huge financial losses in the USA and all around the world. This losses can be scaled up to tens of thousands of million dollars annually.

In this project a fraud prediction model will be developed. This project may potentially benefit financial institutions saving thousands of US dollars and each individuals who are victims of incidents.

The problem with working this data set that it is imbalanced. Before getting started, the ratio of transactions to not to be fraud is more than 99%. In order to correctly address this dataset following steps will be followed:

Part I

1. Cleaning
2. Exploratory Data Analysis

Part II

1. Cleaning Dataset
2. Encoding (One hot encoding or pandas.get\_dummies())
3. Scaling
4. Modelling
5. Hyperparameter Tuning
6. Model Evaluations

**Dataset**

Data set is acquired form Kaggle. Below you can find the details about the dataset.

“Paysim synthetic dataset of mobile money transactions. Each step represents an hour of simulation. This dataset is scaled down 1/4 of the original datasets which is presented in the paper "PaySim: A financial mobile money simulator for fraud detection".” Kaggle(2019)

<https://www.kaggle.com/ntnu-testimon/paysim1>