DATA SELECTION PROPOSAL

Project idea: CREDIT CARD FRAUD DETECTION

Description: The aim of my project is to recognize fraudulent credit card transactions so that the customers will not be paying for those transitions they don't make.

1. **Data set:** https://www.kaggle.com/mlg-ulb/creditcardfraud

   These datasets contain transaction made by credit cards in September 2013 by European cardholders. The datasets represent transactions that occurred in two days, in total we have 492 frauds out of 284,807 transactions.

2. **Methodology:**

   a. **Data Preprocessing:** As my dataset will only contain numerical input variables which are the result of a PCA transformation, I will be using Numerical Preprocessing methods such as rescaling, NaNs,...

   b. **Machine learning model:** Classification: Binary Classification and other methods for unbalanced classification since the fraud class is only 0.172%, indicating that the dataset is highly unbalanced.

   c. **Evaluation Metric**: Features Vi's is the principal components obtained with PCA and the other features not transformed with PCA are 'Time' and 'Amount'. 'Time' shows the seconds elapsed between each transaction and the first transaction in the dataset. 'Amount' represents the transaction value, this feature will be used for dependant cost-sensitive learning. The dataset contains two classes, representing the response variables, 1 indicating fraud transaction and 0 indicating genuine ones. I will measure the accuracy using the Area Under the Precision-Recall Curve (AUPRC). I want to track the false positive, false negative, true negative and true positive from the dataset.

   d. **Final conceptualization:** At the end of my final project, I will build a webapp which identifies and displays the proportion of fraud transactions along with genuine ones.