Project Proposal: INFO 7390 Advances in Data Sciences

Overview:

We propose to create a solution that deciphers a fraudulent transaction from a genuine one.

Dataset:

Credit Card Fraud Detection Dataset

<https://data.world/vlad/credit-card-fraud-detection/discuss/credit-card-fraud-detection/mm4wiyjv>

Data includes:

-V1 – V28: These are features determined from the original dataset that is confidential after performing PCA.

-Timestamp: We have data from every second from the first transaction.

-Amount: Amount of transaction

-Class: Determining if the transaction is false or true (y-variable)

Approach:

-Scrape link to download Dataset

-Perform under sampling and over sampling (only 492 out 284,807 are true)

-Apply various classification models (Logistic Regression, Random Forest, Support Vector Machine, Naïve base Classifier)

-Select best parameters

-Compute Confusion Matrix and ROC curve for all models (with under sampling and over sampling)

-Perform Auto – ML to check for best model

-Present final pipeline