**Project 1**

**IEEE-CIS Fraud Detection**

**Abstract:**

This project is inspired from IEEE-CIS Fraud Detection Kaggle competition, where we have to solve a binary classification to predict fraudulent transaction. The dataset is cleaned, feature engineered, modeling and ensembling is done.

**Methodology:**

1. Data Cleaning:

* The training and testing dataset is separated in transaction and identity. So the dataset is merged on TransactionID for training and testing.
* There are some mismatch in training and testing features. So, the columns of test dataset is renamed to match the columns of the training data.
* There are a lots of null value. The feature containing more the 100000 null value is removed.
* The null values of the feature containing less null values will be filled with mode values.
* Memory size is reduced by changing datatype to manage this large volume of data.

1. Exploratory Data Analysis:

* Some EDA is done to make sense of the data.

1. Feature Selection:

* Correlation is measured related to Fraud to find out the best feature.
* Features with less correlation is dropped from the dataset.

1. Handling Class Imbalance:

* The dataset contains huge class imbalance issue. To tackle this issue, I’ve tried using Up-Sampling and Down-Sampling.

1. Modeling:

* XGBoost and LightGBM seems gave the state of the art accuracy. So, finetuned hyper-parameter is used to train the dataset.

1. Ensemble:

* Weighted Average ensemble is apply on two models prediction

Result Analysis:

* Up-Sampling and Down-sampling cased the worst result. They were both 0.50 on public leaderboard.
* XGBoost model gave the best score which is 0.915064 but took too much time, around 25 mins 31 sec.
* LightGBM too way less time but score was also a bit less. Score on public leaderboard is 0.896678.
* Weighted average ensemble was applied on the models (XGBoost, LightGBM) but the score was still 0. 915064