**1. Introduction**

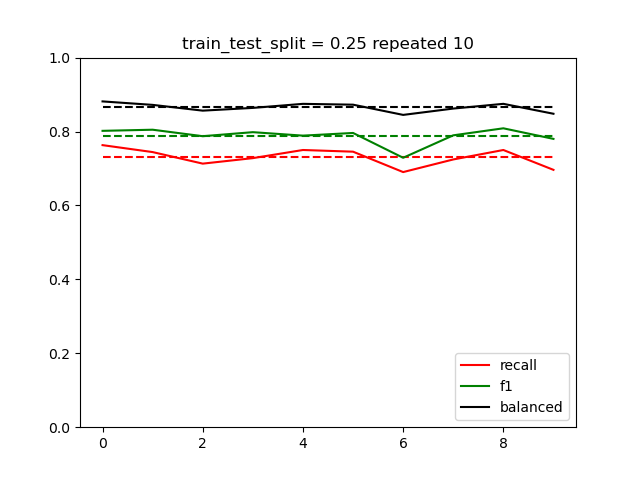
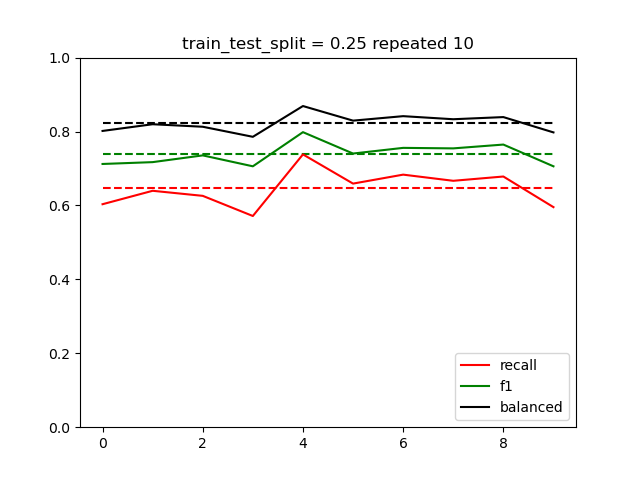
***[A].*** ***Purpose***

On semester of lectures, we have been exploring multiples of different Machine Learning & Deep Learning techniques and be able to use these highly studied models easily with convenient python modules like Scikit-learn, Tensorflow and Pytorch. Being able to use it fluently gives us the possibility to apply this in real word application. By extension of the course, this paper will explore broad aspect of Machine Learning implementation and data science in general by applying what we have leaned to two different famous datasets.

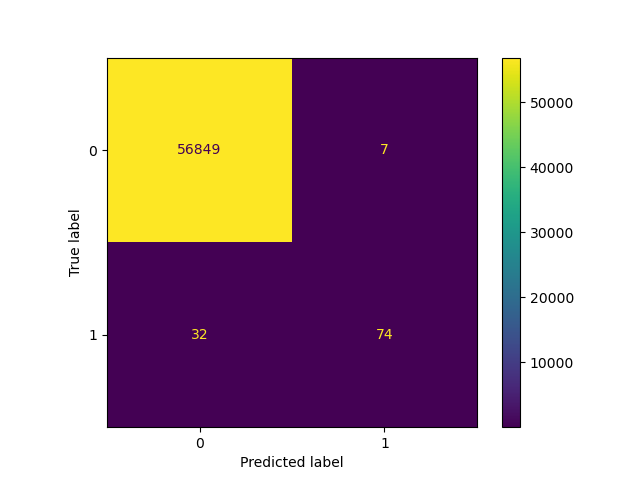
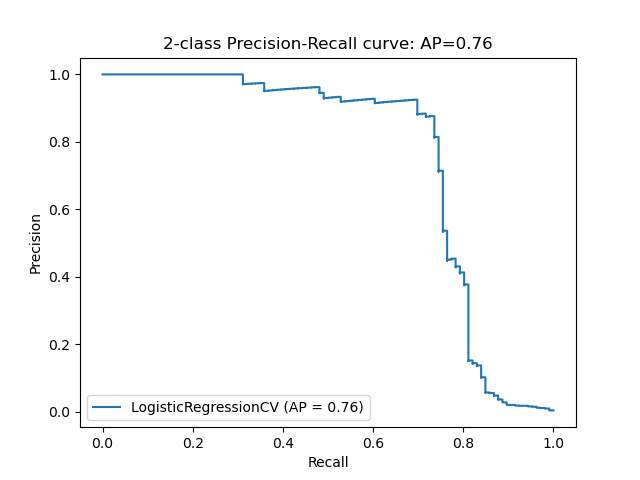
**2. Credit Card Fraud Detection**

***[A]. Feature Selection & Explore Imbalanced Dataset***

Credit card fraud detection dataset which has compiled in csv format is easily be accessed by importing into pandas Dataframe. The dataset has already compiled in a way, each column represents features of a datapoint. Column ‘Class’ will be used as binary classification label for the datapoint, and column ‘Time’ will be dropped because of its irrelevance to the purpose of the dataset. The dataset is imbalanced which has 492 frauds where rest 284315 is not a fraud. Because of the Imbalance of the dataset, choosing a right metric to measure performance of the classifier had to be chosen carefully. Despite its good amount of dataset, some of the baseline model shows inconsistency of its performance depends on how dataset has been split into train set and validation set. For the baseline model I used Logistic Regression without any feature engineer, preprocessing nor hyperparameter tuning. The test has been shown spikes(inconsistency) on performance metrics experiment of multiple runs with randomly split dataset ***(Figure 1).***



***(Figure 1) LogisticRegression (no preprocess) (Figure 2) LogisticRegressionCV (Scaled, Normalized)***

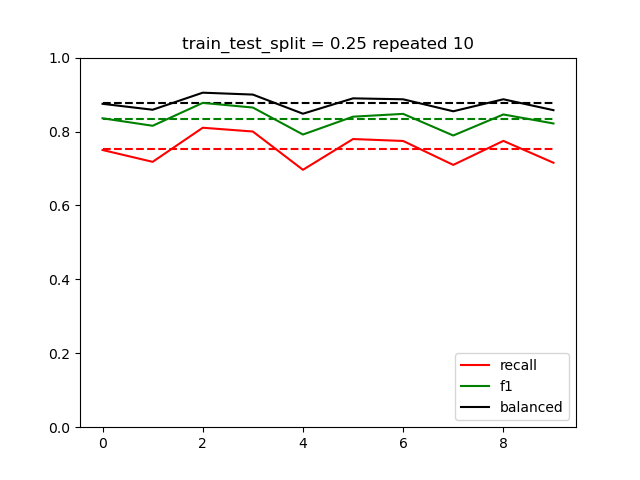


***(Figure 3) CM & Precision-Recall curve LR-CV (Scaled, Normalized) Precision:91.3 F1:79.1 Recall:69.8***

The horizontal dash-line is the overall average of each metrics. Three metrics are used for performance evaluation. This inconsistency happens due to its randomized selection of the train and test set where dataset is imbalanced specifically too few number of ‘Fraud’ case on the entire dataset. By applying some preprocessing method with CrossValidation, it produces more consistent, better performing classifier. By looking at Confusion matrix, model failed to detect 32 fraud and 7 non-fraud has false detected as fraud. Because of the imbalanced characteristic of the dataset, the Recall(69.8%) is pretty low. It means 30.2% of fraud can be skipped ***(Figure 2&3).*** As experiment above, normalizing standardizing and scaling dataset results more consistent, better result, let’s set the baseline here and try different classification techniques.

[B] The Ensemble Classification (Voting Classifier)

There are few different techniques as Ensemble ML model. Among that, this paper will explore Voting method. Voting method is to use various regression or classification model train at the same time and literally vote to resulting a prediction. There is soft method and hard method. Soft method uses predict\_proba(scikit-learn) as it’s value composition which means, the probability of each model’s prediction will equally contribute to the final result. While hard method only uses the winner(one prediction) as its contribution. In recent studies, soft method is generally gives better result since hard method can easily biased where there is few classifier gives absolutely wrong decision where no apparent difference in probabilities ex) One classifier resulting in binary classification 49.9% proba of class 1 and 50.1 proba of class 2 then, 49.9% of probability of class 1 will be forgotten indefinitely.



3. Baseline Classifier

As a baseline classifier, some of the traditional ML algorithm such as Logistic Regression and Decision Tree are going to be used and there will be a Neural Network approach as well. First,