Synthetic Data Generation

1 Problem Statement

Synthetic data generation for minority class Using Generative Adversarial Network.

2 Explanation

In this problem we have transaction data where target is having two category, 1. whether it a fraud or 2 not fraud. Where we have 492 frauds out of 284,807 transactions. Which makes dataset highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. When we go for making a model using this data set then model will be skewed towards majority class as it have more data from majority class it will learn majority class pattern mostly. Due to this it become very important to do some method so that our data will become balanced and one of technique is to generate synthetic data.

3 Literature Review

As we discussed importance of data balancing. Synthetic data generation mostly preferred as removing data point from majority class will lead data loss.

For synthetic data generation any techniques have been proposed and one of main technique used these days are SMOTE. In SMOTE we generate data based on nearest neighbor and this way we can miss any hidden pattern of data. Keeping this in view GAN models are used.

4 Data

Data contains 30 features and a target.

These are: ['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Amount'] and ['Class']

where 492 are frauds out of 284,807

5 Deliverable

Model which will generate synthetic minority class.

6 Evaluation

Confusion metric is used to for validation in model.

7 Data Ingestion

Data is taken from Data folder. We are reading this data using pandas library as pandas data frame.

8 Data Analysis

Data contains some duplicate point .
Need to modify data representation because of skewness of data.

9 Data Munging

Not Required

10 Data Exploration

Amount columns in data is transformed using log to make make it normally distributed.
To center and scale all data, the middle 99.8% was used, so outliers don't pull too much.

11 Feature Engineering

Not required to preserve actual pattern of data.

12 Modeling

Vanilla GAN model is used to generate synthetic data.

13 Optimization

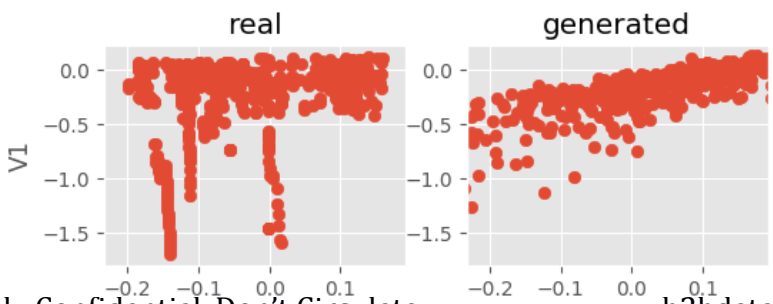
Not Required

14 Prediction

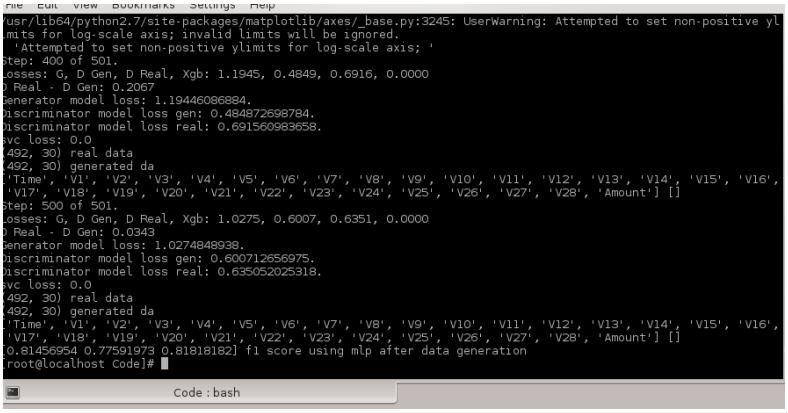
Synthetic data was generated.

F1 score used to predict using generated data : 0.81

15 Visual Analysis



Plot of actual data and synthetic generated data.



16 Results

Here we used GAN model to generate minority class data. And we can see f1 score is 0.81 which indicates good result.