NayaOne

NayaOne Data Science Project

Improved GAN Model

Company overview

NayaOne provides banks with a single point of access to hundreds of fintechs and datasets, through our Digital Sandbox. NayaOne work with a broad range of financial institutions, from banks, regulators, and scaling fintechs; through NayaOne's data, technology integrations, and platform the institutions are able to test and develop new innovative solutions.

Project overview

Personally identifiable information (PII) is protected by GDPR; while a fantastic regulation for protecting individuals, can hamper innovation - new technology providers are unable to purchase or even view real data. This lack of access to data in turn prevents the technology company from being able to either develop their products or demonstrate their value to possible customers, such as banks, without months of legal work.

It is in this space that synthetic data has its place, giving statistically accurate data but with no PII. Whilst there are many approaches to generating synthetic data (see H. Surendra, and S. Mohan, "A review of synthetic data generation methods for privacy preserving data publishing", International Journal of Scientific & Technology Research, vol. 6, no. 3, pp. 95-101, 2017.), GAN models are particularly interesting.

For this project you will use the data described below and try to and improve on the open source CTGAN described here and here. The goal is to produce data that better matches the original data, whilst successfully protecting the original data. A useful tool used in industry to quickly see attributes and correlations of a dataset is pandas profiling, found here. This tool will allow you to tell the difference between datasets quickly without deep statistical analysis; however, statistical analysis is still encouraged where possible.

In particular the CTGAN struggles with strings, if this could be successfully generated the data that could be given to fintechs and technology providers working with NayaOne would increase.

A final note on the project is that you are aiming to improve the model, not simply optimise the epochs and batch size, the variables should be kept constant wherever possible.

NayaOne

Project deliverables

- Mathematical model and code for the technical aspects of the project
- Report on method used to perform the data generation and the results, including the level of accuracy obtained
- Goal 1 - improve on the CTGAN for numeric and categorical data
- Goal 2 - enable the same GAN to work with string data

Input data

To allow you to try and develop the best GAN possible two datasets should be considered for this project.

The first is a financial loans dataset. This dataset is particularly category and numerically heavy and thus is the easier of the two datasets to work with. We recommend using the lc_loan.csv dataset or combining both the test and train dataset for more data.

https://www.kaggle.com/datasets/husainsb/lendingclub-issued-loans?select=lc_loan.csv

The second dataset is a list of complaints. This dataset is particularly string focused.

https://www.kaggle.com/datasets/kaggle/us-consumer-finance-complaints