Using Generative Adversarial Networks for synthetic data generation

Abstract

During the Model Development process and later in the Independent Validation step, it is essential to have sufficient data to enable proper validation of models. In certain use cases and for specific portfolios, some data points are naturally scarce, e.g. confirmed fraud cases in fraud prevention or defaults in low-default portfolios, resulting in pronounced class imbalances. Consequently, most, if not all, minority class observations are employed for model parameter estimation, making it nearly impossible to retain an independent test dataset. Here, we present an approach to generate synthetic data using Generative Adversarial Networks as an alternative to more traditional methods like over- and under-sampling or SMOTE. By allowing conditional sampling, this technique helps overcome data scarcity in minority classes, while simultaneously enabling broader validation insights. Moreover, it offers the possibility to remove personally identifiable information (PII), making it particularly suitable for sensitive applications where data privacy is paramount.

Authors & Affiliations

Mr Martin Lazo[1], Mr Jakob Kisiala[2]
[1]True North Partners, London, United Kingdom. [2]True North Partners, Edinburgh, United Kingdom