# DOPPELGÄNGER

# SYNTHETIC DATA GENERATOR

## INDIVIDUAL REPORT FOR **KAR CHAUDHURI ANIRBAN (A0108517H)**

GRADUATE CERTIFICATE IN PATTERN RECOGNITION SYSTEMS – PRS GROUP 19

### Your personal contribution to project?

All 4 members contributed significantly to these projects and proposed good business use cases of synthetic data before we agreed to proceed with Nirav's idea on life expectancy and handwritten digits. I researched into and initially suggested synthetic data generation and simulation for air pollution, fraud detection and covid-19 probability predictions. I took the lead role in modelling and coding: data pre-processing, exploratory data analysis and developing well-trained AI models. I also researched & prototyped the use of both unsupervised algorithms (PCA, Variational Autoencoders) for synthetic numeric data and images respectively and supervised discriminator (XGBoost Regressor, Random Forest Regressor, Convolutional Neural Network classifier). Given high amount of noise in the data, hyperparameter tuning was a tedious, time-consuming process. I did extensive research by reading thesis papers to experiment with various model architectures and parameters. In the end, I managed to ensure optimal training and test scores for both case studies. I advised and guided teammates on common visualisations, data pre-processing techniques best practices and statistical tests to understand underlying data trends. Apart from being lead coder and modeller, I donned the role of a mentor too. Nirav is our project coordinator and business domain expert who guided us with directions whenever we were confused. Taksh is our DevOps guru who dealt with CI/CD pipeline and Prashant is solution/system architect who designed Rest API containing our data processing and model prediction generation and User-Interface too. I aided them too where possible. We all contributed significantly to final report creation.

### What you have learnt from the project?

The most challenging aspect of the project was generating good quality synthetic data, but it was a tremendously beneficial learning journey. The different data types like numeric, image & time-series require quite distinct approaches & modelling techniques, and even within the same data type different data sets have their own unique pre-processing challenges. One must decide carefully an appropriate missing value imputation technique and winsorize outliers as they may distort synthetic data trend. Data visualisation is important to understand quality of synthetic data as well as determine which predictor variables significantly impact target variable. One can also create a 'Generator-Validator-Evaluator' type solution framework that can be applied towards most types of data – albeit with customized selection of the pre-processing techniques and ML models depending on the nature of the real data set.

I was inquisitive and fascinated about learning and being able to use all three ML techniques: Unsupervised (PCA as a Generator), Semi-supervised (Auto-Encoders as Validator) and Supervised (CNN as Evaluator) within the same project. I was also personally quite fascinated to learn about Variational Auto-Encoders (VAE). Lectures taught us about Auto-Encoders and how they capture and abstract the key patterns within a dataset but VAEs take this one step further by normalizing the abstracted latent space into a distribution. By perturbing this latent distribution, it is possible to generate completely new data which is still consistent with the patterns of the original data. This technique has been well documented, and we adopted this approach to generate our synthetic image data. The cost function of the VAE, that involved creating a customisable cost function in Keras involving summing up reconstruction and KL divergence loss is also worth taking note of as it tells us significance on how reconstructed images are different from original ones. Principal Component Analysis, on the other hand is a linear dimensionality reduction technique that was used for numeric data by creating linear, orthogonal features but don’t work well for non-linear images due to complex representations between pixels.

As for the supervised learning models, I was most fascinated by convolutional neural network. I tried Conv LSTM too but that was computational slower and was not as accurate as pure CNN. I experimented with various techniques to avoid overfitting: dropout rate, regularisation, decaying learning rate, initialisation of weights and different kernel filter sizes. A 5*5 filter gave optimal performance. Between Random Forest and XGBoost, two ensemble techniques, I discovered that Random Forest reduces variance better while XGBoost is better for reducing bias but also tend to overfit more. PCA inverse transform reconstructed data patterns with almost similar summary statistics but with greater multicollinearity. XGBoost has regularisation parameters that handled multicollinearity well. Large number of trees slow computational time so larger depth of tree is good consideration as well.

**How you can apply this in future work-related projects?**

The Synthetic Data Generator has several possible applications in my workplace. I am working with marketing analytics handling customer data containing personal identifiable information, transaction amount, payment methods, merchants and outlets visited. We collect data on daily basis.

A generative model that extracts these data patterns in a synthetic data set without exposing the details of the actual client related to payment instrument and amount, we can then make this synthetic data available to our market analysts. Cluster analysis based on recency, frequency and monetary values of customers by marketing analysis to segregate customers, create visualisations to understand KPIs of changing customer behaviours will be key to understand customer churn. We to be able to research and experiment with latest recommendation system algorithms that are representative of latest customer trends and able to recommend relevant products. This will ensure optimal customer engagement, thus reducing probability of customer churn. Marketing analysts can also then create predictive models to predict customer churn too apart from profiling based on K-means clustering. Customers who have high probability of churning within next month should be identified. Suitable recommended merchants and outlets based on distance (10km radius around a coordinate point for instance) should be identified and marketing campaign resources and strategies can be finalised too targeting these customers.

There are  dozens of similar use-cases that can be worked upon if we can synthesise our historical data such that it can be made available to our data analysts without compromising the privacy, confidence and trust of our clients.

--- END ----