Synthetic Data Generation to Train Machine Learning Algorithms

Keywords: Synthetic data generation, machine learning

Project description

- The motivation comes from the fact that data availability is sometimes limited, distorting or preventing the calibration of machine learning algorithms
- The goal of this IDP is to create a synthetic and flexible data generation methodology for empirically validated data and a specific economic problem set
- The data generation process is developed with Python
- The exact task will be the following:
  - Pick a specific economic use case for machine learning – an example is the sales prediction across stores (https://www.kaggle.com/c/rossmann-store-sales/data) – and specify the data items required
  - Compare Monte Carlo simulations (parametric) to Bootstrapping procedures (non-parametric) as adequate data generation processes
  - For the Monte Carlo simulations, derive assumptions (mainly the (joint) distributions) and cross-validate the assumptions empirically (e.g. by data from kaggle.com)
  - Pick an adequate bootstrapping procedure (e.g. Bayesian bootstrap, Smooth bootstrap etc.) which is motivated by the literature
  - Generate additional observations to improve the calibration of the ML algorithm
  - Evaluate if the data generation improved the calibration and thus the predictability of the ML algorithm (e.g. by applying the RMSE, F1 score etc.)

What we are looking for

- Strong analytical and project management skills
- Determination and passion for your areas of expertise
- IT skills required for the IDP
- Interest to learn something about finance

What we offer

- Kick-off session including introduction to relevant finance topics
- Experience with IDPs
- Open dialogue and support
- Potential for publication and/or evaluation of future use cases
- Both single and group projects are possible

Interested?

Please send an e-mail with CV, academic transcript and your preference for this project to lisa.knauer@tum.de.

Questions?

In case of any (e.g. topic related) questions, please contact Lisa Knauer (lisa.knauer@tum or call +49 89 289 25483).