Neural generative techniques for synthetic data in insurance

London – 15/06/2023

Aurelien Couloumy
acouloumy@novaa-tech.com

With the contribution of:
M. BEN CHEIKH LEHOCINE (CCR),
A. KAIS (CCR Re),
E. LAVERGE (CCR)

Insurance
Data
Science

Generate synthetic data to handle insurance data issues

- Insurance data can be a **source of problems**, regarding for example:

Quality

- Missing data, incoherent values, ineffective data quality techniques for law behaviour setup, technical pricing or reserving calculations.

Exhaustivity

- Limited labelling budget, lack of data regarding emerging risks, new stress test for capital modelling, rare events scenario for natcat, fraud.

Privacy

- Restricted use of medical or geotracked data for actuarial calculations, HDS storage, third parties (broker, MGA) share.

- The generation of synthetic data **could help** to overcome such problems:

Data **imputation**

Data **augmentation**

Data **anonymisation**

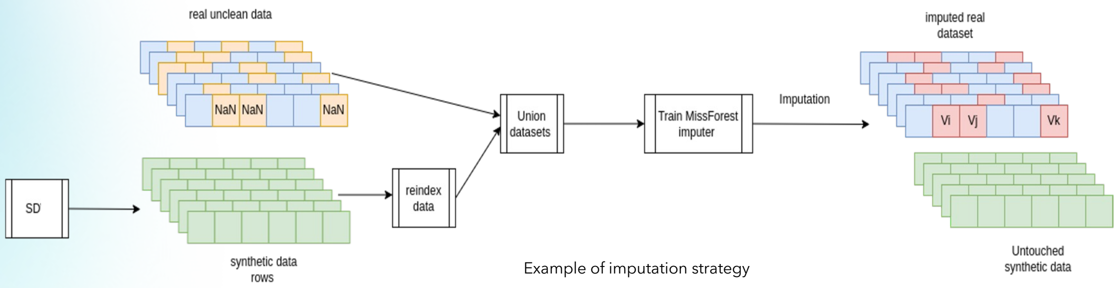
- The study aims at presenting **neural generative approaches** for exploiting such ideas, as well as case studies highlighting benefits and drawbacks.

Application

Baseline, GANs and hybrid strategies

Data imputation

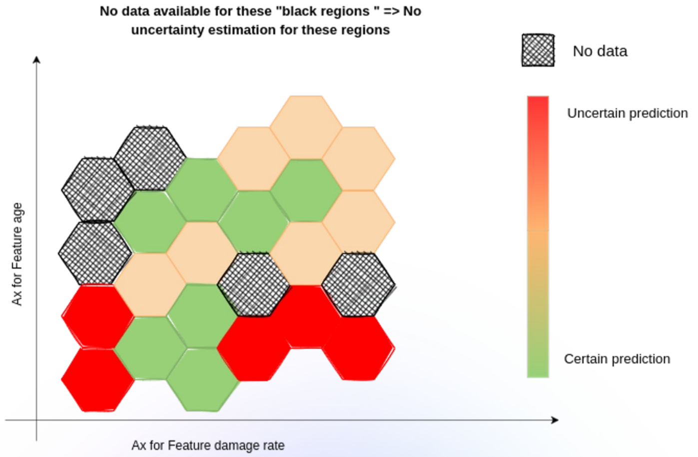
- Use of **sampling based neural generative** techniques (CTGAN [1] , TVAE, CopulaGAN, etc.) to generate synthetic data.
- **Nan injection** to ensure MCAR/MAR hypothesis.
- **Imputation strategies** and different NaNs to run sensitivity tests:
    - Univariate simple imputers
    - Multivariate KNN/iterative imputers
    - Multivariate similarity imputation with synthetic data
    - Multivariate synthetic data with iterative imputers



Example of imputation strategy

Data augmentation

- Many existing data augmenters to manage Imbalanced classifiers (SMOTE) or any other non tabular tasks (LLMs).
- Rarely take into account regressors cases and do not allow to force data constraints (regarding uncertainty).
- Cumulative use of **Bayesian Neural Networks (BNN) [2]** to identify model uncertainty [3] areas, then to define constraints to generate **synthetic adversarial data** using CTGAN.

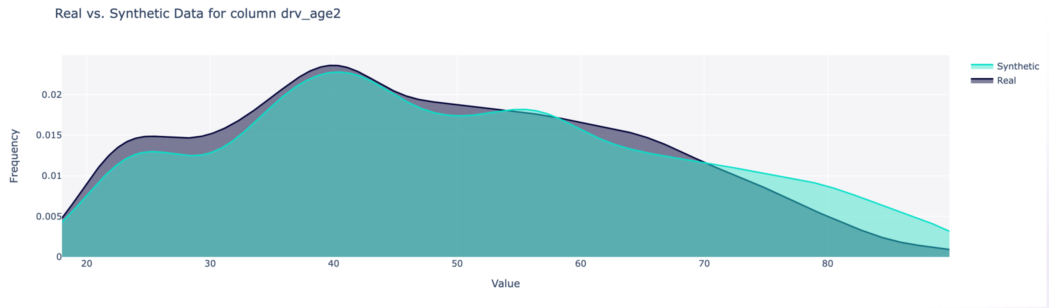


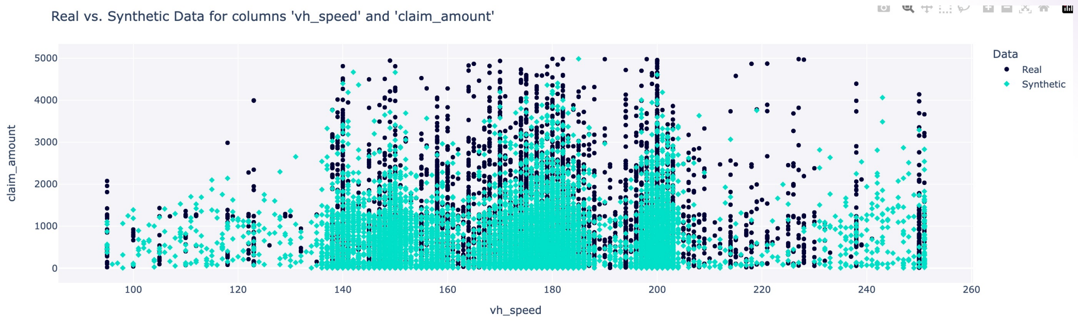
[1] Lei et al, Oct 2019. Modeling tabular data using conditional GAN. https://arxiv.org/abs/1907.00503
[2] N. G. Polson, V. Sokolov et al., (2017) Deep learning: a Bayesian perspective, Bayesian Analysis, vol. 12, no. 4, pp. 1275-1304. https://arxiv.org/pdf/1706.00473.pdf
[3] Y Gal, (2016) Uncertainty in Deep Learning, http://www.cs.ox.ac.uk/people/yarin.gal/website//thesis/thesis.pdf

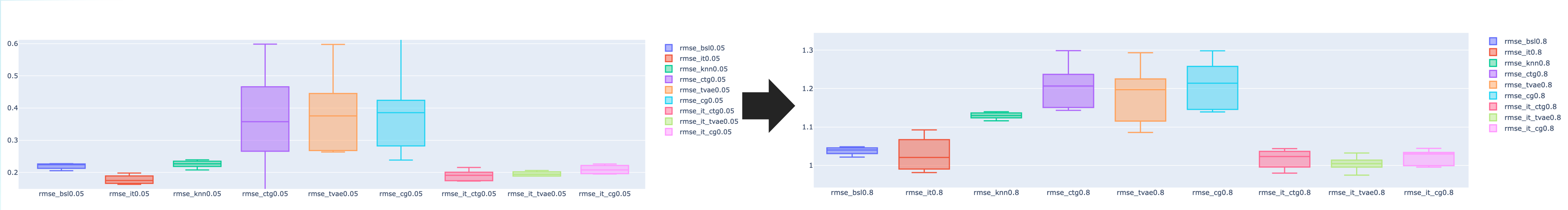
Improve the before and after modelling

Data imputation

- Experimentation on a motor pricing dataset.
- Charts and measures help to **ensure conservation of feature distributions** (overall quality, shapes, pair trends) [4].
- At first glance, similarity imputations with synthetic data **do not beat popular approaches**.
- When Nan % increases, **interest of mixing techniques** also increases.







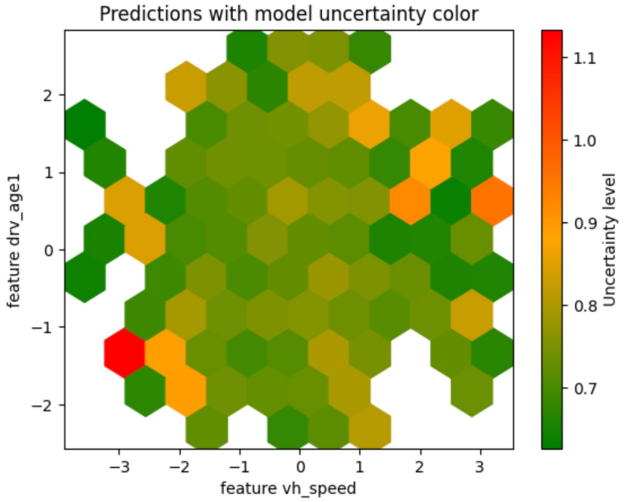
[4] Patki, Neha and Wedge, Roy and Veeramachaneni, Kalyan, 2016, The Synthetic data vault, IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 399-410

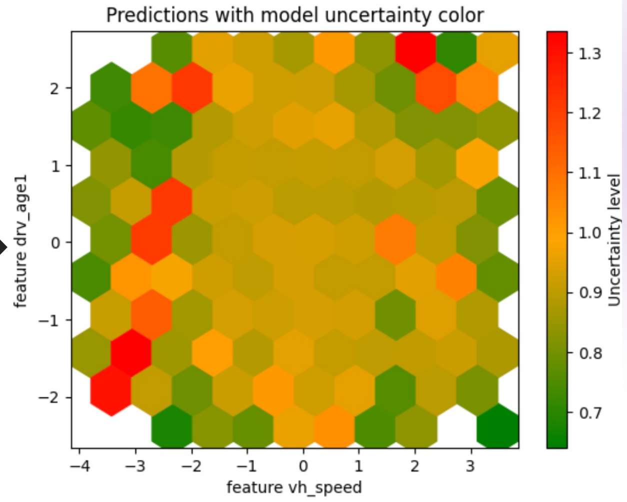
Results

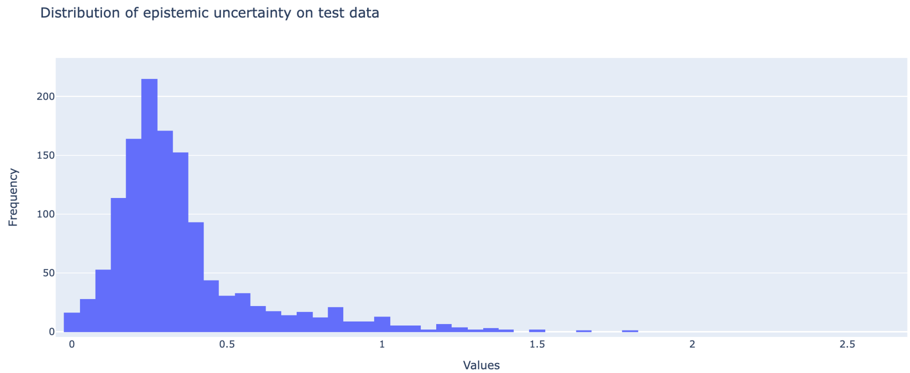
Improve the before and after modelling

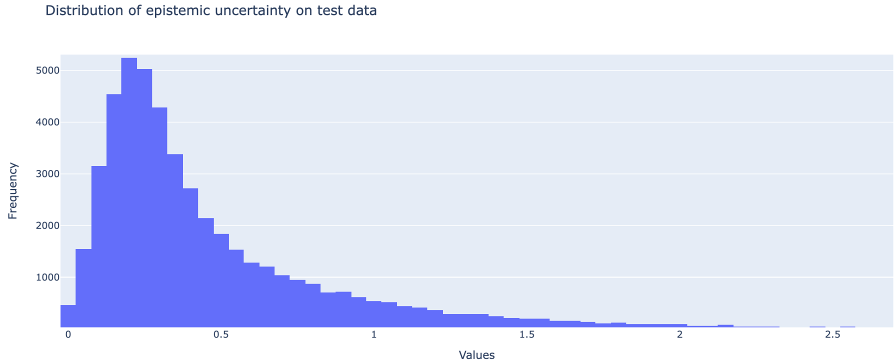
Data
augmentation

- Use of CTGAN to create synthetic data in the blank areas and use of the MC Dropout BNN [5] to estimate the associated uncertainties.
- It allows to **reveal new uncertainty** [6] areas.
- It allows to understand how certain the model would be with **any scenarios (lower bound).**
- Finally, It provides extreme scenarios that may be **source of larger uncertainties**.









[5] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning., Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 1050–1059, New York, New York, USA, 20–22 Jun. PMLR.
[6] A Der Kiureghian and O Ditlevsen. (2009) Aleatory or epistemic? does it matter? Structural Safety, 31 (2):105–112,