Title

Harnessing Generative Models for Synthetic Non-Life Insurance Data

Speaker/Company

Claudio Giorgio Giancaterino, Independent

Abstract

Obtaining realistic, publicly accessible datasets is a significant barrier in advancing actuarial research and developing open-source tools for insurance analytics. This study leverages synthetic data and aims to evaluate various Generative Models for producing a standard synthetic non-life insurance premium dataset.

A Conditional Gaussian Mixture Model has been employed as a benchmark. The methodology involved splitting the dataset into two subsets based on the "claim occurence" variable as a binary indicator. For each subgroup, a multivariate Gaussian Mixture Model is fitted, allowing for complex, multi-modal distributions. This benchmark was then compared with advanced Deep Learning architectures, including a Conditional Variational Autoencoder, a Conditional Variational Autoencoder with a Transformer-based Decoder, and a Conditional Diffusion Model. Additionally, the GPT-5.1 Large Language Model was used to generate synthetic datasets via prompt.

The experiments were conducted on two insurance datasets retrieved from the CASdatasets R package following three trials. In the first experiment, used as a baseline quality assessment, a portion of each complete dataset was fed into the generative models, which were tasked with generating an equal number of records. In the second experiment, used to evaluate the data augmentation capacity, a smaller portion of each dataset was used, with the models tasked with producing a larger number of rows, equal to that in the first experiment. In the final trial, with attention to ethical considerations, the gender variable was omitted to protect privacy.

Validation of the generated data included several steps: data visualization with comparison through univariate analysis, PCA, and UMAP representations, evaluation of the consistency of the produced data with the original, and the statistical Kolmogorov-Smirnov test. Predictive modeling of frequency and severity using Generalized Linear Models (GLMs) based on Tweedie distribution were employed to assess the quality of the generated data. Furthermore, the importance of features was analyzed.

This analysis evaluates each model’s ability to accurately capture underlying distributions, preserve complex dependencies, and maintain intrinsic relationships. The findings offer valuable insights for improving synthetic data generation in the insurance field, potentially enhancing risk modeling, pricing strategies in the face of data scarcity, and ensuring regulatory compliance.

Keywords:
Conditional Variational Autoencoder, Conditional Gaussian Mixture Model, Conditional Diffusion Model, Conditional Variational Autoencoder with a Transformer-based Decoder, GPT-5.1 Large

Language Model, PCA, UMAP, GLMs

References:
1. Ian Goodfellow and Yoshua Bengio and Aaron Courville, 2016, Deep Learning, MIT Press.
2. Mario V. Wuthrich, Ronald Richman, Benjamin Avanzi, Mathias Lindholm, Michael Mayer, Jürg Schelldorfer, Salvatore Scognamiglio, 2025, AI Tools for Actuaries, SSRN.
3. David Foster, 2023, Generative Deep Learning, 2nd Edition, O'Reilly.
4. Jake VanderPlas, 2016, Python Data Science Handbook, O'Reilly.
5. Jamotton, Charlotte ; Hainaut, Donatien, 2023, Variational autoencoder for synthetic insurance data, ISBA.
6. Harshvardhan GM, Mahendra Kumar Gourisaria, Manjusha Pandey, Siddharth Swarup Rautaray, 2020, A comprehensive survey and analysis of generative models in machine learning, ScienceDirect.

Biography

Claudio Giancaterino is a qualified Actuary, working during the day at Intesa Sanpaolo Assicurazioni, an Italian Insurance Company based in Milan. In his free time, he is involved in AI & Data Science activities on an independent, freelance basis. Claudio is a member of the Astin Actuarial Group, and the Italian Actuarial Body Association. Previously, he was an assistant professor of Insurance Statistics at the Catholic University of Milan. He collaborated with the IAA and IFoA in several working parties, and attended Kaggle competitions. Claudio hold webinars, workshops and talks at several organisations, conferences and meetups.