ETH zürich 1st ASTIN Bulletin Conference

Harnessing Generative Models for Synthetic Non-Life Insurance Data

Claudio Giorgio Giancaterino

16/01/2026

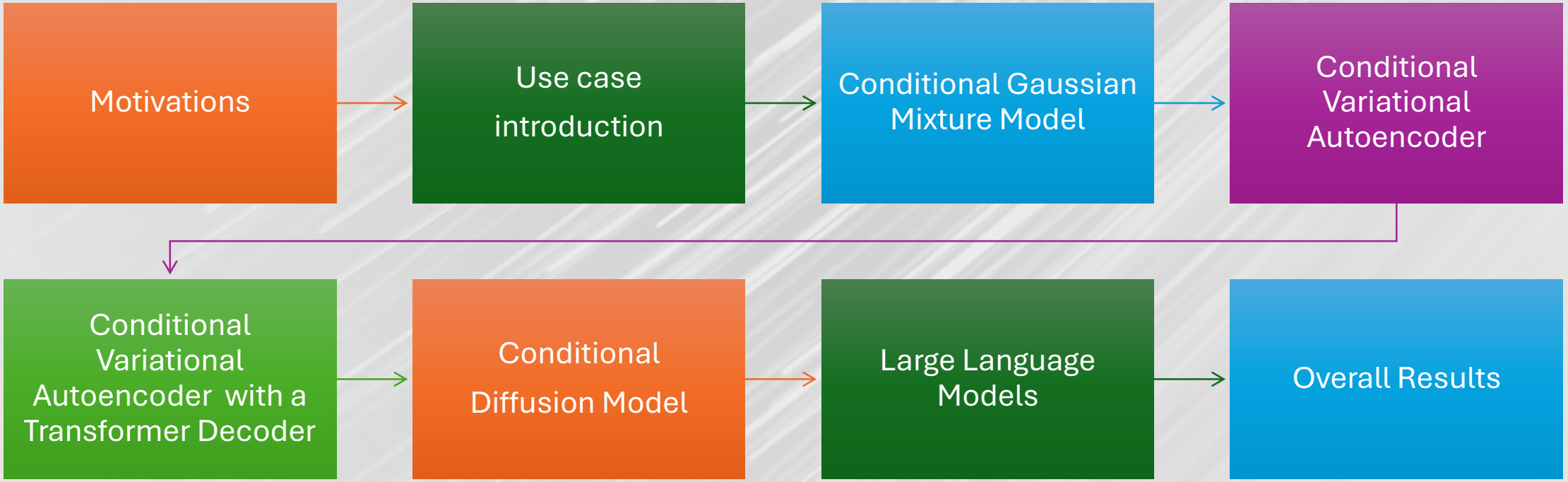
MySelf:

Actuary during the day &
AI Scientist in the free time

Reach Me:

MyLinks

Agenda


Motivations
Use case introduction
Conditional Gaussian Mixture Model
Conditional Variational Autoencoder
Conditional Variational Autoencoder with a Transformer Decoder
Conditional Diffusion Model
Large Language Models
Overall Results

Motivations

- Insurance Use Case
- Employ a wide range of Generative Models

Google Nano Banana

Data Scarcity in Insurance Research

The Problem: Access to realistic datasets is a major barrier to advancing actuarial research and developing insurance analytics tools.

- Most insurance data is confidential and proprietary
- Data masking and bureaucratic processes prevent disclosure
- Limited open resources lack diversity for robust modelling

The Solution: Develop simulated datasets using generative models

The Anatomy of Insurance Non-Life Risk Data: Statistical Components in the Pricing dataset

Risk Premium Components:

Frequency = Number of Claims / Number of Exposures

Severity = Losses / Number of Claims

Risk Premium = Frequency × Severity

The risk premium equals the product of expected claim frequency and expected cost per claim

Datasets employed

Datasets are retrieved from the "CASdatasets" R Package

Dataset 1: ausprivauto0405 - Automobile claim datasets in Australia
It is a data frame of 9 columns and 67,856 rows:
Exposure, VehValue, VehAge, VehBody, Gender, DrivAge, ClaimOcc, ClaimNb, ClaimAmount

Dataset 2: swmotorcycle - Swedish Motorcycle Insurance dataset
It is a data frame of 9 columns and 64,548 rows:
OwnerAge, Gender, Area, RiskClass, VehAge, BonusClass, Exposure, ClaimNb, ClaimAmount

Unlocking Data Quality: Leveraging Claim Occurrence for Stable Generative Modelling

What is ClaimOcc ?

It is a binary indicator which indicates occurrence of a claim

If ClaimOcc = 0 → ClaimNb = 0 and ClaimAmount = 0

If ClaimOcc = 1 → ClaimNb ≥ 1 and ClaimAmount > 0

Why use ClaimOcc?

Structural relationship: drives logical dependency between ClaimNb and ClaimAmount

Class imbalance: separates rare claim events from common no-claim cases

Model stability: improves synthetic data quality by conditioning

Role in Generative Models

CGMM: Fits separate Gaussian mixture models for each ClaimOcc value

CVAE & CTVAE: Concatenates ClaimOcc to the latent encoding for conditional generation

CDM: Uses ClaimOcc as a condition signal to guide the denoising process

Synthetic Data Generation Trials

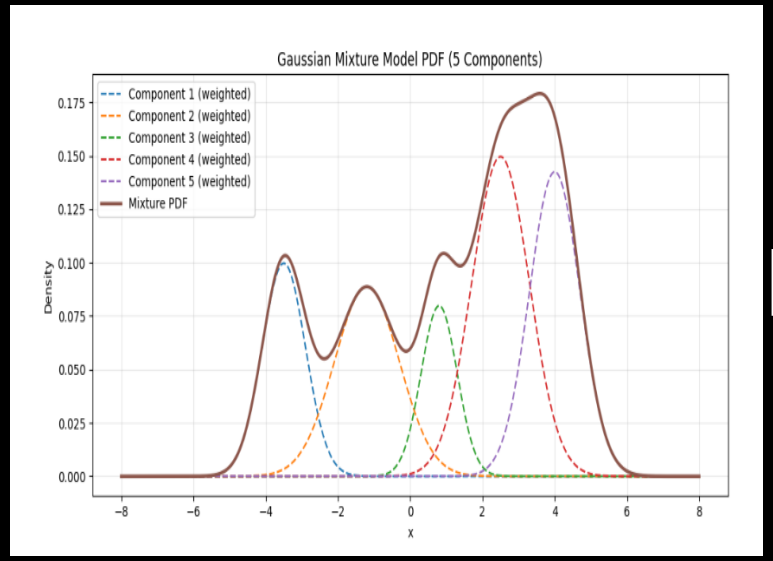
Trial 1: 80% Training → 80% Generation
Baseline Quality Assessment
What is the baseline quality of synthetic insurance data when a generative model is trained on 80% of the original data? How well does it preserve statistical distributions and enable accurate GLM predictions?

Trial 2: 60% Training → 80% Generation
Data Augmentation with Limited Training
Can a generative model trained on only 60% of the data successfully generate synthetic samples to reach 80% dataset size while maintaining statistical fidelity and predictive performance?

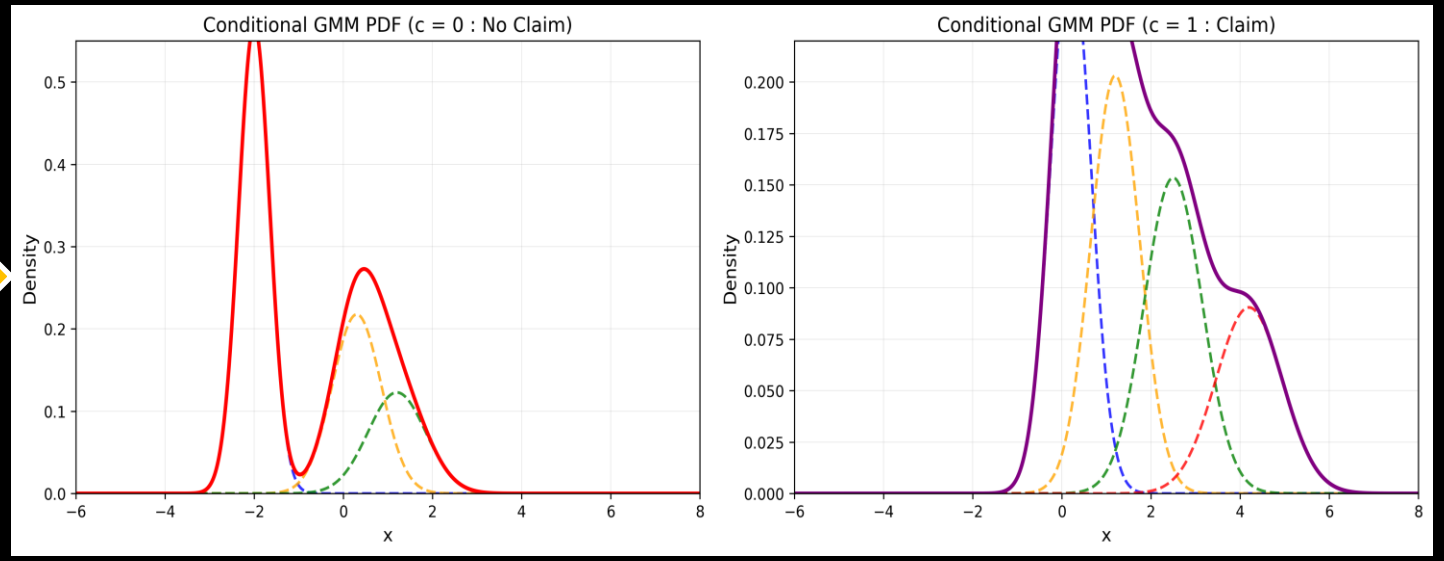
Trial 3: 80% Gender-Masked → 80% Generation
Fairness-Aware Generation
Can a gender-masked generative model generate unbiased synthetic insurance data while maintaining predictive utility, and what is the fairness-accuracy trade-off?

Conditional Gaussian Mixture Model (CGMM)

What is it?
A Gaussian Mixture Model assumes that the data are generated by a mixture of K Gaussian distributions, each with unknown parameters and corresponding to a cluster. Every Gaussian has its own mean μ_k, covariance Σ_k, and mixing weight π_k.



ChatGPT

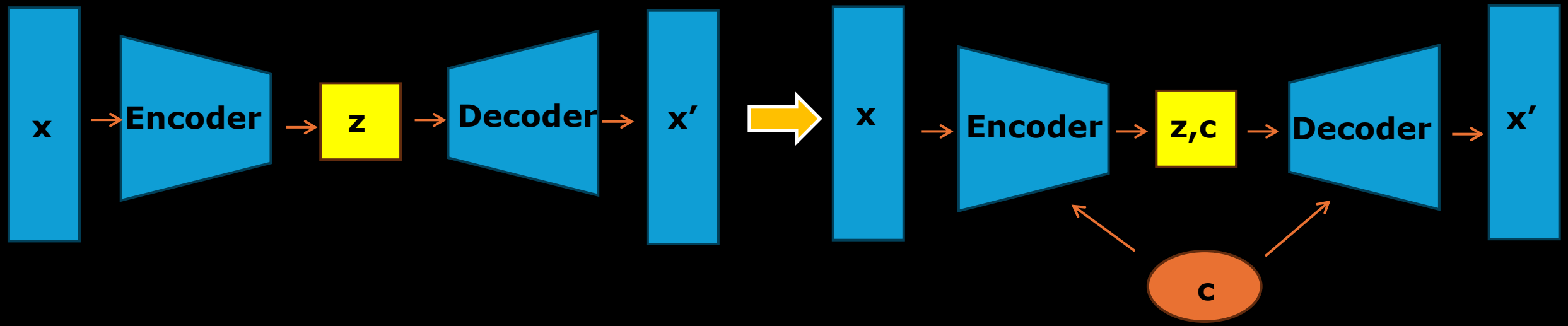


$$Loss = -\sum_{n=1}^{N} log\left(\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)\right)$$

$$Loss = -\sum_{n=1}^{N} log\left(\sum_{k=1}^{K} \pi_k(c_n) N(x_n|\mu_k(c_n), \Sigma_k(c_n))\right)$$

Conditional Variational Autoencoder (CVAE)

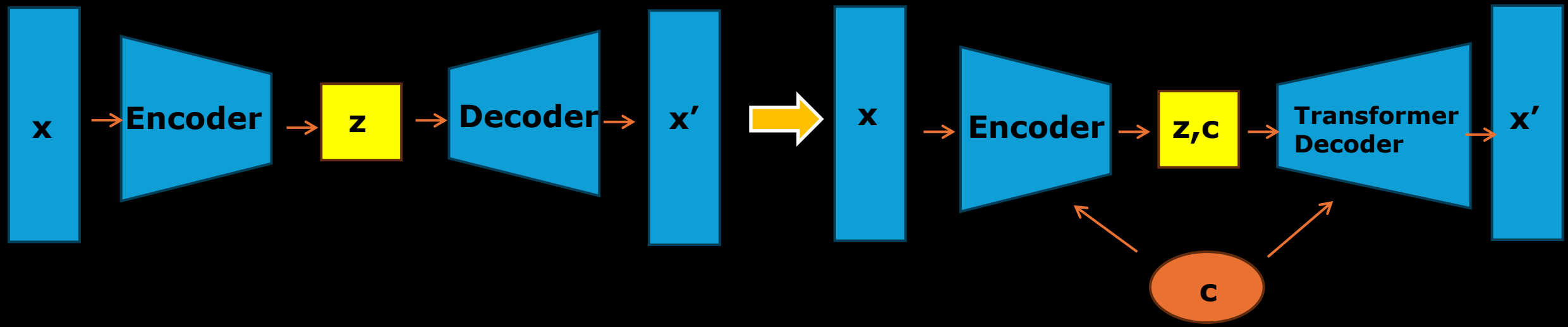
What is it?
A Variational Autoencoder (VAE) is a type of neural network that learns to represent input data as a probability distribution in a lower-dimensional space and then decode it to generate similar input data.



$$Loss = -E_{q_\phi(z|x)}[log_{p_\theta}(x|z)] + KL(q_\phi(z|x)||p(z)) \Rightarrow Loss = -E_{q_\phi(z|x,c)}[log_{p_\theta}(x|z,c)] + KL(q_\phi(z|x,c)||p(z,c))$$

Conditional Variational Autoencoder with a Transformer-based Decoder (CTVAE)

What is it?
It's a hybrid architecture that leverages:
-VAEs for learning a smooth, continuous latent space.
-Transformers for decoding sequences with long-range dependencies and attention mechanisms.



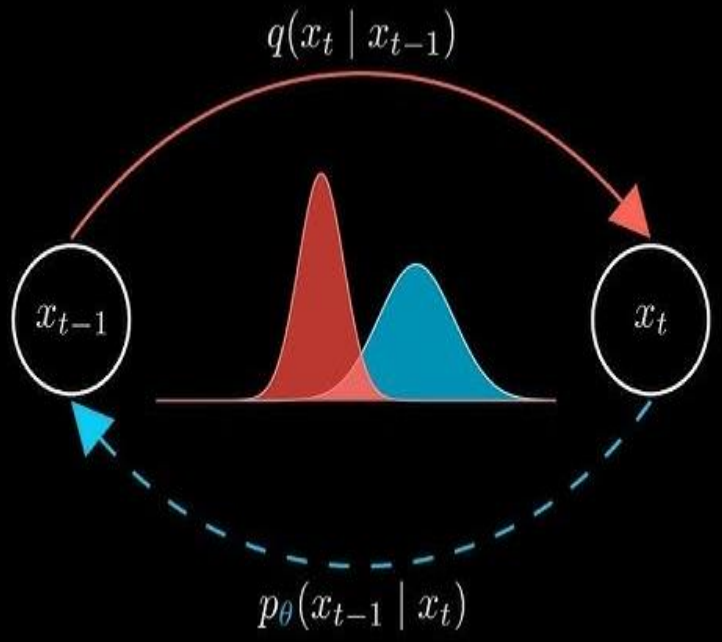
$$Loss = -E_{q_\phi(z|x)}[log_{p_\theta}(x|z)] + KL(q_\phi(z|x)||p(z))$$

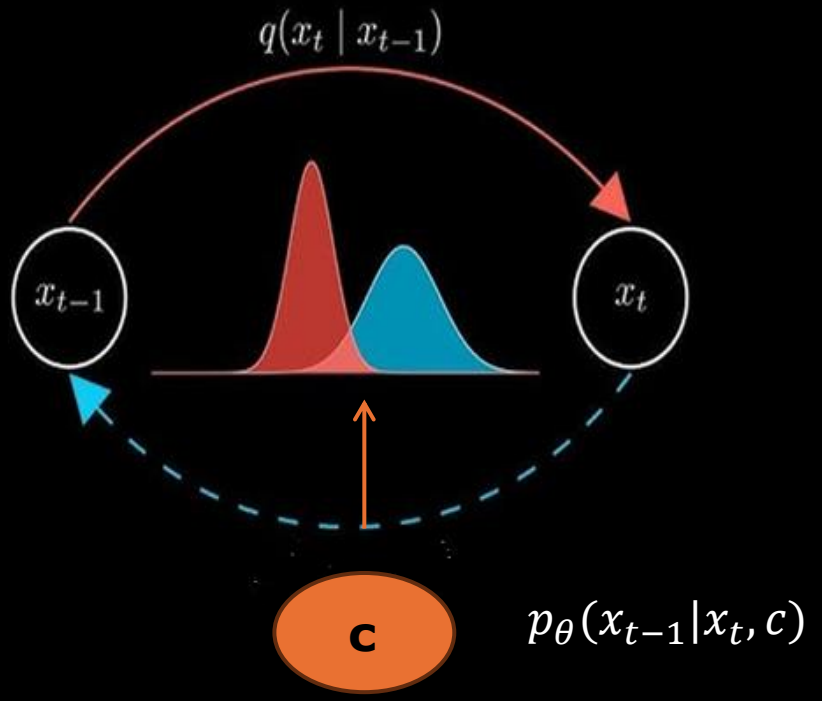
$$Loss = -E_{q_\phi(z|x,c)}[log_{p_\theta}(x|x_{<t}, z, c)] + KL(q_\phi(z|x,c)||p(z,c))$$

Conditional Diffusion Model (CDM)

What is it?
A Diffusion Model learns to generate data by simulating a gradual noising-and-denosing process.

source





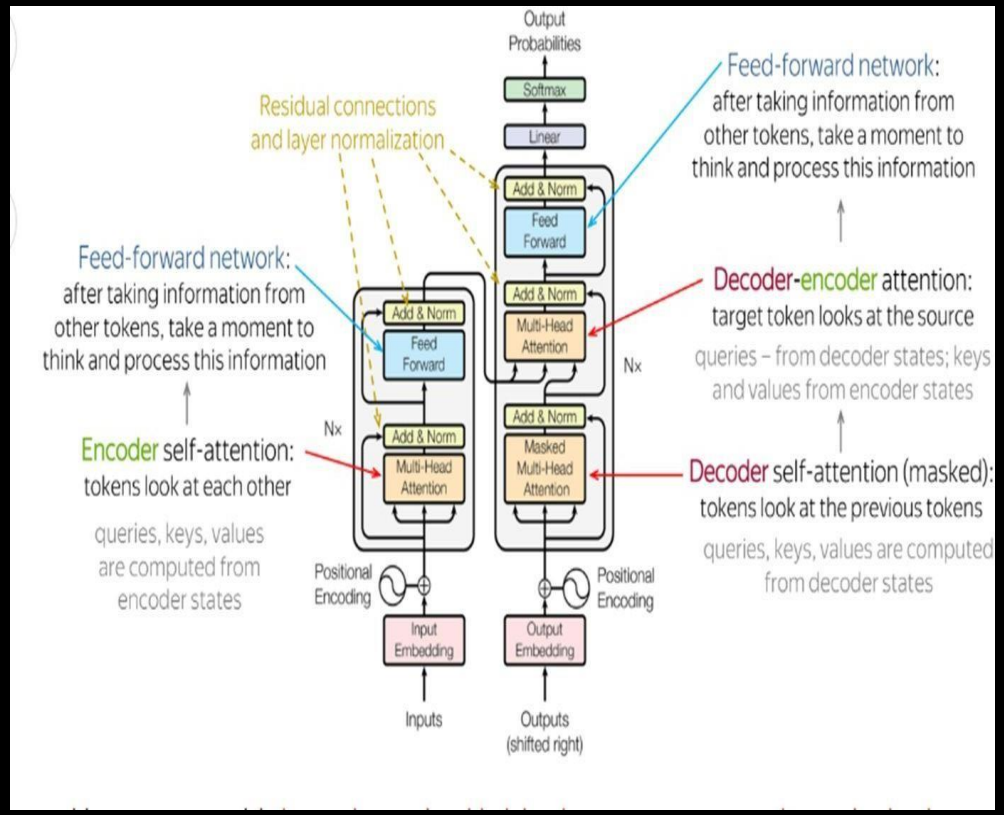
$$Loss = E_{x_0,t,\varepsilon}[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2]$$

$$Loss = E_{x_0,t,\varepsilon}[\|\varepsilon - \varepsilon_\theta(x_t, t, c)\|^2]$$

Large Language Model (LLM)

What is it?
Language Models can be defined as a probability distribution over sequences of words.
Given a vocabulary of words, a Language Model assigns a probability to each sequence of words, and the purpose is to predict the next word in the sequence.
Large Language Models like BERT and GPT are trained on a vast amount of text data, to learn the structure, meaning, and usage of language.

source



$$Loss = -\frac{1}{T}\sum_{t=1}^{T} log\, p_\theta(x_t|x_{<t})$$

Validation by Consistency records

The consistency test validates the **logical relationships** between three insurance claim variables in the synthetic data:

-ClaimNb (Number of claims)
-ClaimOcc (Claim occurrences)
-ClaimAmount (Claim amount)

No Claims: All three = 0 (no claims occurred)
Claims Exist: ClaimOcc =1; ClaimNb >0; ClaimAmount >0

Validation by Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) test compares the distributions of original and synthetic data for each feature. A **high p-value (>0.05)** indicates that the two distributions are statistically similar, while a **low p-value (<0.05)** suggests significant differences.

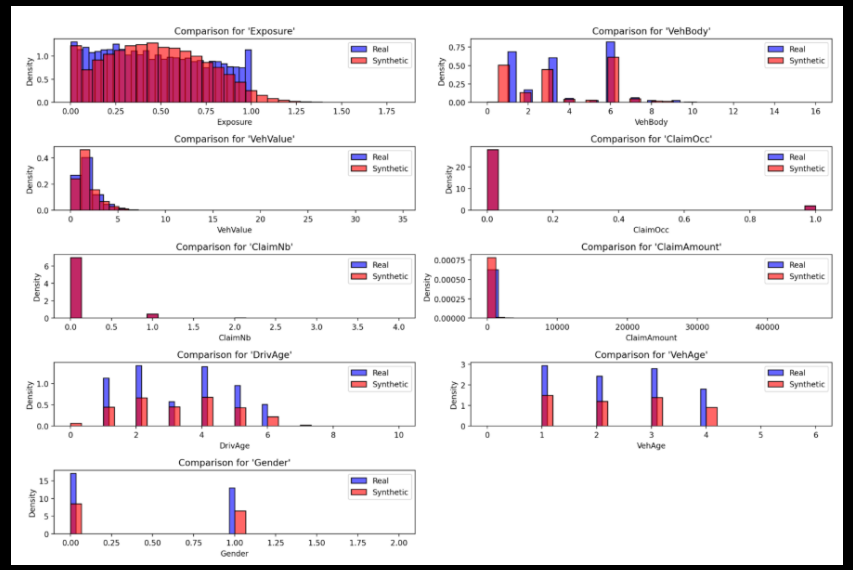
```python
# Kolmogorov-Smirnov test
for column in X_train.columns:
    original = X_train[column].values
    generated = new_samples_df[column].values
    statistic, p_value = ks_2samp(original, generated)
    print(f"KS Test for {column}: Statistic={statistic}, P-value={p_value}")
```

```
KS Test for Exposure: Statistic=0.0480120030007502, P-value=7.598641486479458e-54
KS Test for VehValue: Statistic=0.013934733683420854, P-value=6.314917316823904e-05
KS Test for VehAge: Statistic=0.003638409602400583, P-value=0.8707136778711597
KS Test for VehBody: Statistic=0.0032445611402850713, P-value=0.9407888567993886
KS Test for Gender: Statistic=0.0014628657164290626, P-value=0.9999999949094549
KS Test for DrivAge: Statistic=0.032501875468867236, P-value=6.669794205601761e-25
KS Test for ClaimOcc: Statistic=0.0, P-value=1.0
KS Test for ClaimNb: Statistic=0.00018754688672173447, P-value=1.0
KS Test for ClaimAmount: Statistic=0.01121530382595648, P-value=0.00242672366876259
```

Validation by Data Visualization

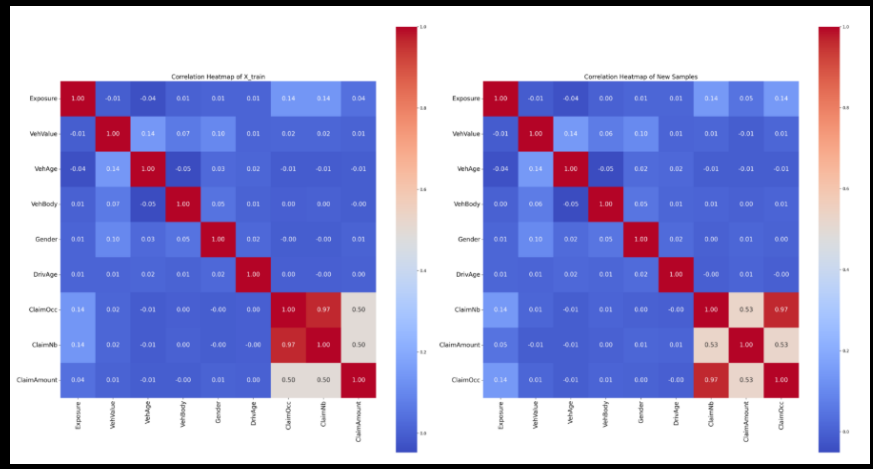
Univariate Analysis

Displays the alignment of the univariate statistical properties for each feature across the real and synthetic data



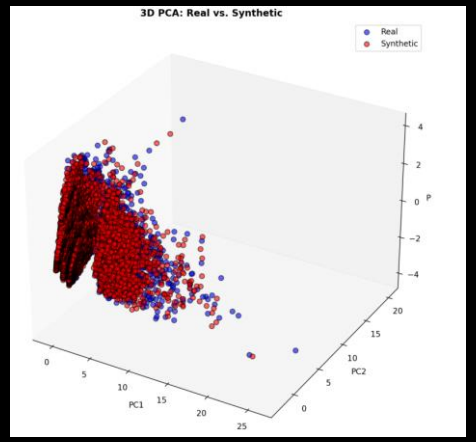
Correlation Matrix

Assesses the structural fidelity of the synthetic data by comparing the correlation matrix to the original data



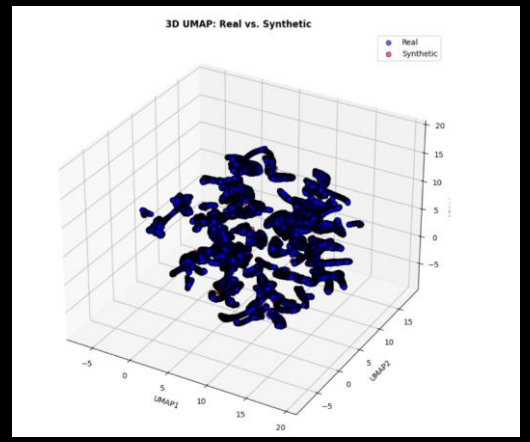
3D PCA

Compares the explained variance captured by the principal components of the synthetic data against the real data



3D UMAP

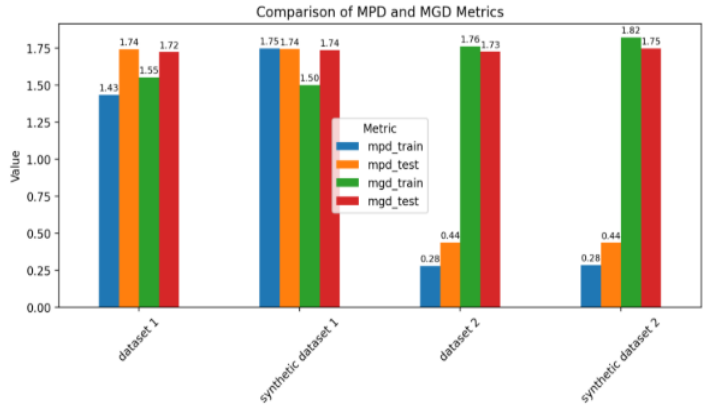
Evaluates the degree to which the synthetic data preserves the local and global topological structure of the original data



Validation by Predictive Modelling

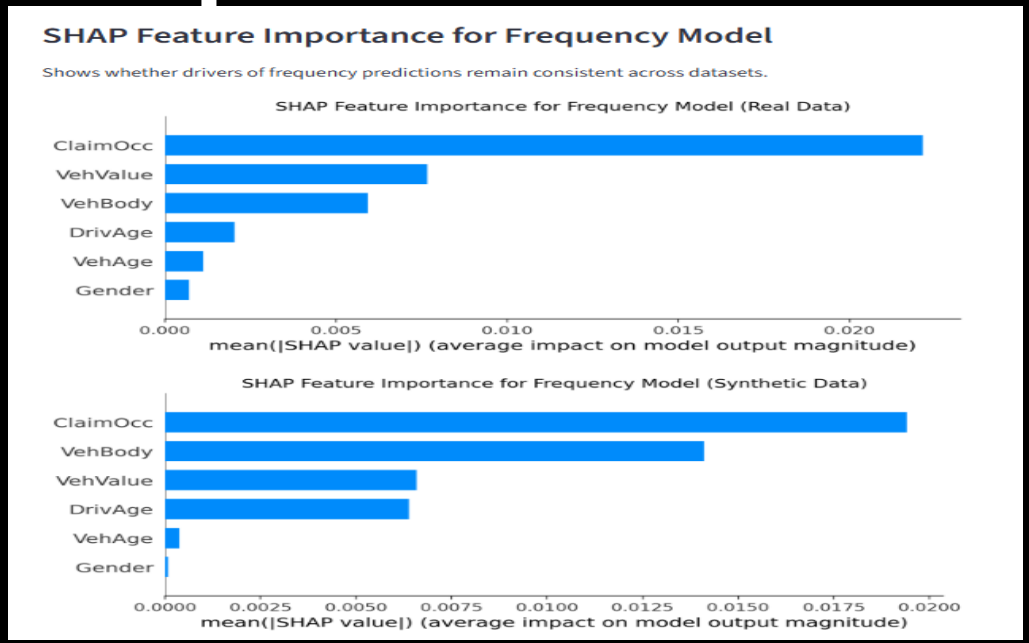
Frequency and Severity prediction
GLM on both synthetic and real data to evaluate performance using deviance metrics.

	mpd_train	mpd_test	mgd_train	mgd_test
dataset 1	1.4345	1.7412	1.5509	1.7221
synthetic dataset 1	1.7457	1.7427	1.4991	1.7361
dataset 2	0.2788	0.4368	1.7615	1.7251
synthetic dataset 2	0.2848	0.4357	1.8222	1.7478



Validation by Feature Importance

SHAP Feature Importance
Shows whether drivers of frequency and severity predictions remain consistent across datasets.



Overall Results

Focusing on predictive modelling performance and the preservation of statistical properties, I recommend the Conditional Gaussian Mixture Model, the Large Language Model, and the Conditional Diffusion Model. The two Conditional Variational Autoencoders show strong training performance, but suffer from severe overfitting. In subsequent trials, model performance degrades, but in some cases it remains acceptable.

Watch the results from the web demo app.

Conclusions

What I've learned

- ✅ Conditional Gaussian Mixture Models are competitive as generative models for tabular data. Good compromise between ease of use and effectiveness.
- ✅ Large Language Models are promising for synthetic data generation due to their performance and ease of use. Complex model implementation isn't required, but detailed prompting and deep data analysis are needed to write a good prompt.
- ✅ Conditional Diffusion Model is competitive in performance, but it requires computational and coding effort.

References

-Jan Goodfellow and Yoshua Bengio and Aaron Courville, 2016, *Deep Learning*, MIT Press.
-Mario V. Wuthrich, Ronald Richman, Benjamin Avanzi, Mathias Lindholm, Michael Mayer, Jürg Schelldorfer, Salvatore Scognamiglio, 2025, *AI Tools for Actuaries*, SSRN.
-David Foster, 2023, *Generative Deep Learning, 2nd Edition*, O'Reilly.
-Jake VanderPlas, 2016, *Python Data Science Handbook*, O'Reilly.
-Jamotton, Charlotte; Hainaut, Donatien, 2023, *Variational autoencoder for synthetic insurance data*, ISBA.
-Harshvardhan GM, Mahendra Kumar Gourisaria, Manjusha Pandey, Siddharth Swarup Rautaray, 2020, *A comprehensive survey and analysis of generative models in machine learning*, ScienceDirect.
Repository: https://github.com/claudio1975/Generative_Modelling
Web_APP: https://huggingface.co/spaces/towardsinnovationlab/Generative_Models_4_Insurance_Data
Article: https://medium.com/@c.giancaterino/stop-waiting-for-data-how-generative-models-are-reshaping-insurance-analytics-ec102a2e5177
Video: https://www.youtube.com/watch?v=DQ58pGWp0NE&list=PLGVZCDnMOq0qmerwB1elTnr5AfYRGm0DF&index=21

?

?

?

?

?

?

?

Thank you

Keep in touch:

- Linkedin
- Newsletter
- Medium
- Website