Improving Synthetic Property & Casualty Data Generation through Expert Input in Generative Adversarial Networks

Improving Synthetic Property & Casualty Data Generation through Expert Input in Generative Adversarial Networks

submitted in partial fulfillment for the degree of master of science

Jan Janiszewski
10004378

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on June 26, 2023

	UvA Supervisor	**External Supervisor**
Title, Name	Erman Acar	External Supervisor
Affiliation	UvA Supervisor	External Supervisor
Email	e.acar@uva.nl	supervisor@company.nl

ABSTRACT

The availability of high-quality customer risk data is crucial for insurances which aim to expand their product portfolio or enter new markets. It allows actuaries to accurately price their new customers. The insurance industry therefore presents a compelling context for the application of synthetic data generation methods, such as generative adversarial networks (GANs), to address the challenge of data sharing while complying with GDPR regulations. This research aims to address this issue by proposing an generative model which combines expert knowledge and claim data together to generate synthetic customer risk data for model training. The approach is based on a GAN which is trained on two inputs, the insurer's customer exposure risk data and expert knowledge, in order to generate a dataset which generates synthetic user data. The proposed research has the potential to contribute to the development of GANs which can generate synthetic data which is more realistic than is currently the case, making insurance data more open for sharing.

KEYWORDS

GAN, GLM, MTPL, Data Science,synthetic data, generative adversarial networks, actuarial science

GITHUB REPOSITORY

https://github.com/AfairiJJ/thesis

1 INTRODUCTION

As artificial intelligence (AI) and data science continue to evolve, their implications have been transformative across multiple industries, including insurance. These advancements have particularly revolutionized underwriting and pricing, among other areas, thus introducing a myriad of novel applications. Nonetheless, despite the proliferation of such innovations, significant challenges relating to data privacy and reliability remain [6].

Insurance data, especially those encompassing customer risk profiles, are highly confidential and competitive, posing considerable challenges to actuarial departments and hindering industry expansion into new markets or products. Furthermore, stringent privacy regulations like the General Data Protection Regulation (GDPR) complicate data sharing among insurers and limit academic research due to the lack of available realistic data. While acquiring customer risk data from other insurers or market providers is feasible, the substantial effort required for data obfuscation and GDPR compliance render such exchanges less advantageous [4, 13].

These circumstances underscore the urgent need for reliable and privacy-compliant synthetic data generation methods. Among various AI models, Generative Adversarial Networks (GANs) present a compelling solution, particularly in their ability to generate high-quality synthetic data. Notably, the Multi Categorical Wasserstein Generative Adversarial Network with Gradient Penalty (MC-WGAN-GP) demonstrates a significant potential to create realistic, multi-categorical data, a characteristic inherent in insurance databases [4, 12].

This paper contributes to the burgeoning field of synthetic data generation in insurance, primarily by exploring the impact of integrating actuarial expert input during the training of MC-WGAN-GP models. As far as we know, the application of such a neuro-symbolic approach to GAN training in an insurance context is innovative and has the potential to enhance synthetic data generation's quality and reliability [20, 21].

Our research provides a detailed examination of the MC-WGAN-GP model for synthetic insurance data generation and scrutinizes the effectiveness of including actuarial expert knowledge across different dataset sizes. The insights derived from this study could serve as valuable input for future AI applications within the insurance industry, particularly concerning data privacy issues.

The subsequent sections of this paper comprise a review of relevant literature, an overview of the various models and their applications in insurance, a description of our research methodology, and a comprehensive analysis of our experimental results. The paper concludes with a discussion on the potential implications of our findings and recommendations for future research directions.

2 RESEARCH QUESTIONS

This research aims to investigate the potential of enhancing GAN model training through the inclusion of actuarial expert knowledge. In pursuit of this aim, we proposed two main research questions:

- *Research Question 1*: Can a GAN model trained on insurance claim data accurately replicate the underlying distribution and relationship between the dependent (claim count) and independent variables as measured by Poisson Deviance, GINI, RMSE, and MAE?
- *Research Question 2*: Does the inclusion of actuarial expert knowledge into the input data during the training of the GAN model lead to improved preservation of the distribution and relationships in the synthetic data generated by the model?

Based on these research questions, we have the following working hypotheses:

- *Hypothesis 1.1*: The XGBoost model built on the data generated by the MC-WGAN-GP provides predictions on the outcome variable (claim count) that are significantly different from those made by a dummy model predicting the average claim count for all respective policies in the dataset, as evaluated by Poisson Deviance, GINI, RMSE, and MAE.
- *Hypothesis 1.2*: The XGBoost model built on the data generated by the MC-WGAN-GP provides predictions on the outcome variable (claim count) that are statistically indistinguishable from those made by a model trained on the original dataset, as evaluated by Poisson Deviance, GINI, RMSE, and MAE.
- *Hypothesis 2*: The XGBoost model built on the data generated by the MC-WGAN-GP with expert input included in the training provides more accurate predictions (as evaluated by Poisson Deviance, GINI, RMSE, and MAE) than the XGBoost model built on the data generated by the MC-WGAN-GP without expert input included.

3 RELATED WORK

The proliferation of machine learning (ML) and artificial intelligence (AI) has stimulated a burgeoning interest in generating robust synthetic data across various insurance sub-disciplines, including pricing, underwriting, and claim management [1]. Our study builds upon this body of work by exploring the intersection of three crucial areas: AI in actuarial pricing, data synthesis in insurance, and knowledge representation in AI. This intersection forms the bedrock of our methodology, which is based on the understanding that a robust application of these areas could improve the precision, efficacy, and interpretability of insurance models.

3.1 AI in Actuarial Pricing

Traditionally, Generalized Linear Models (GLMs) have been widely used for actuarial pricing in non-life insurance, particularly for large datasets [10, 18]. Despite their flexibility in dealing with various error distributions and their capacity to use link functions, GLMs are intrinsically linear and, as such, struggle to capture intricate non-linear relationships. As a consequence, researchers have been exploring the potential of deep learning to augment GLM-based actuarial pricing [17, 22–24].

Recent research has delved into the application of Artificial Neural Networks (ANNs) in actuarial pricing, striking a balance between model interpretability and prediction accuracy. This strategy typically involves training a feed-forward ANN alongside a GLM to bolster the latter's performance [22–24].

Guided by these recent developments, we choose to employ a boosting-based model. This decision is informed by empirical evidence that demonstrates the superiority of boosting-based models over GLM and ANN-based models in predicting insurance pricing [17]. Furthermore, this choice aligns with our aim to compare the quality of synthetic data to real data.

3.2 Data Synthesis in Insurance

In the face of stringent data privacy regulations and confidentiality constraints, there is a growing need for alternative strategies to leverage real-world data in the insurance industry. One such strategy is the generation of synthetic data, with early studies using traditional statistical methods like resampling to create synthetic datasets [5, 9]. However, with the advent of ML and AI, more sophisticated techniques have emerged.

Introduced by Goodfellow et al., Generative Adversarial Networks (GANs) have revolutionized synthetic data generation [**?**]. The GAN architecture, composed of a generator and a discriminator network in a competitive setting, has enabled the creation of high-quality synthetic data. Despite this, their use in insurance remains limited.

Kuo [12] was one of the pioneers in the application of GANs in insurance, demonstrating the potential of the CTGAN algorithm to generate synthetic insurance data. The study found that CTGAN-generated data successfully mimic the distributions of real data. Building on this work, Côté et al. [4] evaluated different GAN architectures, with the MC-WGAN-GP model emerging as the most effective in capturing both individual variable distributions and their correlations.

In line with these findings, our approach to data synthesis leverages the MC-WGAN-GP model to generate synthetic data, as it has proven effective in mimicking both the distributions of single variables and their relationships [4].

3.3 Knowledge Representation in AI

The fusion of expert knowledge within machine learning models presents a compelling research avenue, known for its propensity to amplify guidance during learning, heighten model interpretability, and elevate model performance [14, 16, 19–21]. This is particularly indispensable in sectors such as insurance, where the clarity of model rationale and the justification for decisions play pivotal roles [8].

Over the years, the AI and ML landscapes have seen remarkable progress in terms of assimilating expert knowledge into learning models, with a particular emphasis on rule-based systems [20, 21]. As delineated by Prentzas [19], these methods primarily fall into two clusters: rule-based reasoning (RBR) and case-based reasoning (CBR). RBR offers a generalized comprehension of the domain, whereas CBR encapsulates detailed knowledge. While rule-based systems generate solutions from the ground up, case-based systems take advantage of established scenarios to tackle analogous new cases. Given the diverse strengths and weaknesses of both RBR and CBR, composite or integrated methods that merge the two have led to innovative and potent results [14, 16].

Hybrid methodologies can be segmented into three primary categories: sequential, co-, and embedded processing. Sequential processing entails the successive integration of different knowledge representation techniques, culminating in an information flow from the preliminary to the final representation. Co-processing refers to a cooperative approach where the assimilated components concurrently work towards the final output. Conversely, embedded processing involves embedding a component based on a particular representation into one or more components predicated on another [19, 20].

Within the insurance practice, numerous strategies for incorporating expert knowledge have been examined. As an illustration, Byczkowska-Lipińska et al. [2] proposed an expert knowledge-driven system to assess insurability potential in medical insurance, founded on expert rules. Hsieh and Wang [11] further extended this research by introducing the Linguistic Descriptions Evaluating Algorithm, a life insurance risk assessment tool hinged on a multitude of linguistic approaches (e.g., linguistic logic, uncertainty numbers modeling, fuzzifications, and defuzzification schemes).

With the rise of Deep Learning algorithms, fresh methods for incorporating expert knowledge into ANNs have come to the forefront. A notable example is Neurules, introduced by Prentzas and Hatzilygeroudis [21]. Neurules are neurosymbolic rules that merge expert knowledge with neural networks, fusing production rules with symbolic representation units within the neural framework [20]. However, the exploration of expert knowledge integration within GANs, particularly within the realm of synthetic data generation for insurance, remains largely uncharted.

Our knowledge representation method in our AI model is informed by hybrid methodologies that sequentially combine different knowledge representation techniques. We aimed to bolster

our model's interpretability by weaving expert knowledge into the learning process [14, 16, 19–21]. Moreover, we developed a system to assess customer risk based on expert rules, aligning with prior studies in insurance.

4 METHODOLOGY

4.1 Research Design

This research involved designing and training various synthetic data generation pipelines by combining Generative Adversarial Networks (GANs) with actuarial expert input, and comparing the performance of the different GAN models for various training set sizes by means of an XGBoost trained on the generated data (e.g., $N = 5000$, $N = 433,728$).

Four modeling pipelines were designed and compared across three different scenarios of training set sizes. In the first pipeline, a dummy model was established which predicted average claim count for each policy in the risk dataset. In the second pipeline, a baseline model was established, predicting claim count based on an XGBoost model trained on real training data. The third pipeline integrated a GAN trained on the real training dataset to generate synthetic data, which was subsequently used to train the XGBoost model. The third pipeline resembled the second but incorporated additional actuarial expert knowledge in the GAN training process to improve synthetic data generation. Throughout these pipelines, data preprocessing steps, XGBoost model hyperparameters, and the test set remained constant, allowing for unbiased evaluation.

Performance evaluation of the models was done using Poisson Deviance, GINI coefficient, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), which are standard metrics in GAN research and underwriting studies [4, 8, 12, 17].

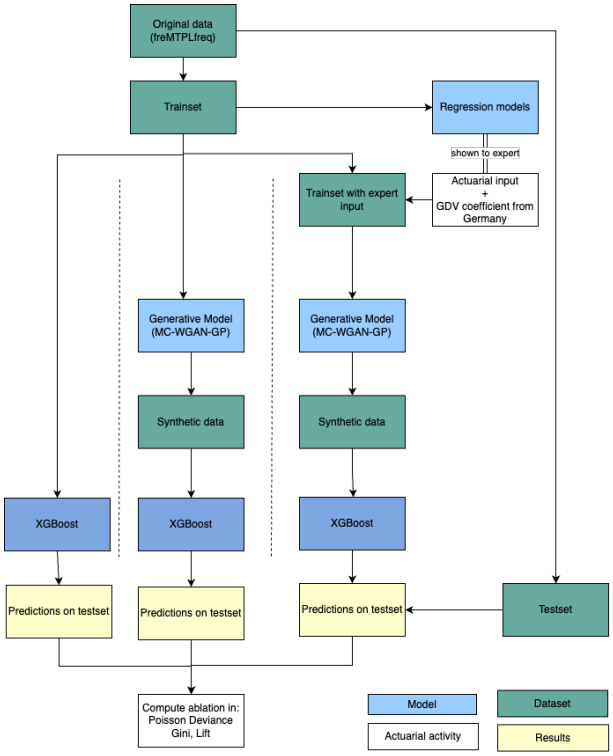


Figure 1: Structure of the pipeline (green: data, blue: models, white: human interaction)

4.2 Hardware and Computational Resources

The research methodology was designed for parallel execution across various computing platforms, ranging from high-performance supercomputers to local systems. The primary implementation language was Python, in combination with the PyTorch framework.

For high-performance computing, the methodology was tested on the Snellius supercomputer, which significantly accelerated the model training process. However, for accessibility and replicability, the methodology was also designed to run on a more modest machine, such as a MacBook M2. Although the MacBook M2's computational capabilities are lower compared to the supercomputer, the training process is still feasible but with extended durations.

4.3 Data

The dataset for this study was derived from the French motor third-party liability (MTPL) insurance portfolio, available on the OpenML platform [7]. This dataset comprises 678,013 car insurance policies and twelve distinct variables. These include the policy number (IDpol), which is typically associated with a customer, car, or a combination of both, as well as the claim count (ClaimNb), measuring the number of claims made by a customer within a specified exposure timeframe. Other variables include the total exposure in years (Exposure), area code (Area), power of the car (VehPower), age of the car in years (VehAge), age of the driver in years (DrivAge), bonus-malus level (BonusMalus), car brand (VehBrand), fuel type (VehGas), population density (Density), and regions in France (Region).

In data preprocessing, particular care was given to ensuring consistency across different variables. Any identified inconsistencies were dealt with appropriate strategies, ensuring the quality of the data.

Notably, two variables deviated from the expected distribution assumptions. Firstly, only 24.80% of the policies exhibited the standard one-year exposure, while the majority had exposure durations of less than a year, which is atypical for insurance policy datasets since policies usually have a standard duration of one year (see Figure 2. As the reasons for this disparity remained unclear, the exposure columns were excluded from the training of the XGBoost models.

Secondly, in 36.67% of the cases, it was unclear whether the policies were unique, as they shared all necessary policy characteristics except for exposure, policy ID, and claim count. It was hypothesized that these non-unique policies represented car fleets, such as leasing cars, rental cars, or company-owned vehicles. Consequently, these policies were retained in the dataset but were grouped together when splitting the dataset into training and test sets to ensure the independence of the test set. No other exceptions or inconsistencies were identified during the analysis (see Appendix A for distribution figures).

As the dependent variable, the claim count was further examined. The dataset revealed that 95% of the policies (N=643,953) had no

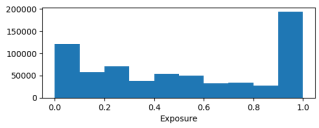


Figure 2: Distribution of exposure values up to one year

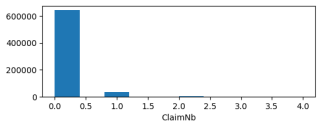


Figure 3: Distribution of claim values below four

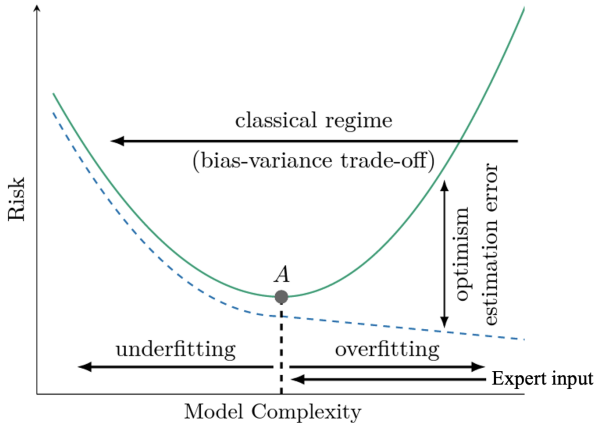


Figure 4: Bias-variance tradeoff with the assumed effect of the introduction of expert knowledge into the model training.

claims filed (i.e., ClaimCount = 0), while 4.75% (N=32,178) of the policies had one claim filed. A small proportion of policies, amounting to 0.25% (N=1,882), had more than one claim filed, with the maximum number of claims reaching 16 (see Figure **??**).

4.4 Data Preprocessing

A standardized data preprocessing pipeline was implemented, based on the recommendations from previous research [4, 17]. This included transformations and standardizations on several variables to ensure data quality and comparability across different modeling pipelines.

The common data preparation steps encompassed the following transformations:

- **IDs:** The policy number (IDpol) was dropped from the dataset as it did not contribute to the modeling process.
- **Claim Cound:** The claim count (ClaimNb; also named Frequency in the communication with the actuary) was capped at values exceeding four claims; to facilitate GAN training, it was also converted into a categorical variable for the training.
- **Exposure:** For XGBoost training, the exposure variable (Exposure) was not utilized. However, to support GAN training, exposure was used with values exceeding one year capped.
- **Area:** The categorical alphabetic representation of the area variable (Area) was transformed into a continuous variable.
- **Vehicle Age:** To avoid excessive skewness in the data, the age of the vehicle (VehAge) was capped at 20 years.
- **Vehicle Power:** To mitigate potential outliers, the power of the vehicle (VehPower) was capped at values exceeding nine.
- **Driver Age:** To limit the influence of extreme values, the age of the driver (DrivAge) was capped at 90 years.
- **BonusMalus:** As per prior recommendations, bonus-malus levels (BonusMalus) exceeding 150 were capped.
- **Density:** To alleviate the impact of skewed distributions, a logarithmic transformation was applied to the density variable (Density).

Following the data preprocessing steps, the numerical features were standardized, ensuring a consistent scale across all models. Categorical variables were encoded using one-hot vectors, with the dimensionality of the vector corresponding to the number of unique categories, as suggested by previous research [4].

After preprocessing, the data was partitioned into training, validation, and test sets in the ratio of 60:20:20. To ensure reproducibility and consistency, a fixed seed value was employed for any random operations conducted during data splitting.

4.5 Incorporating Domain Expertise

The proposed rule integration technique in this work encapsulates elements from both the embedded processing method and the co-processing strategy. Initially, actuarial expertise was converted into a GLM which outlines the relationship between different independent variables and the claim count (akin to the case-based approach). Subsequently, the expert had the opportunity to incorporate rule-based changes directly into the developed model. The resultant synthesis was incorporated into the dataset for the GAN training process. Lastly, the domain expert was given the opportunity to append additional group-based rules directly onto the data.

In risk models, overfitting, especially in the presence of imbalanced ratios between positive and negative outcomes, leads to biases and estimation errors, such as optimism estimation error. We postulate that integrating expert input into the GAN model would help to reduce overfitting, thereby improving the quality of the synthetic data, specifically in accurately representing the relationship between the dependent and independent variables (see Figure 4; [25]). Further, the integration of expert knowledge was anticipated to provide additional insight into the relationship between claim count and independent variables, which might not be directly discernible from the original dataset. An ancillary insight from this study pertains to the capability of GANs in extracting information from auxiliary data added to a dataset.

The process of incorporating actuarial expert knowledge into our research framework followed a series of steps to ensure effective integration. The expert input was integrated into the data provided for GAN training, combining both embedded processing and co-processing approaches.

4.5.1 Scope Definition. The first phase involved identifying variables for which the actuarial expert could provide insights regarding their relationship with the dependent variable (claim count). Variables including Density, Driver Age, Bonus Malus, and Vehicle Age were defined as being within the scope of expert knowledge.

4.5.2 Idea Generation. During this phase, several Generalized Linear Models (GLMs) were trained on the training data to explore potential relationships between each respective variable and claim count. The expert actuary guided the selection of models to be trained. For instance, for vehicle age, multiple polynomial GLMs with a log link and different degrees were trained to encapsulate the polynomial relationship between claim count and vehicle age. An illustration of such a model, showing the relationship $ClaimCount = \beta_0 + \beta_1 * VehicleAge + \beta_2 * VehicleAge^2$, can be found in Figure 5. A range of GLMs were trained to explore potential relationships between the variables and claim count during the idea generation phase. Through analysis and visualization of these models, the expert was able to identify the most suitable representation that resonated with her domain knowledge and expectations.

4.5.3 Representation Selection. The expert then selected the model that best represented her understanding of the relationship between the independent variable (e.g., vehicle age) and claim count. This selection process resulted in the identification of the most appropriate GLM representation (refer to Figure 6 for vehicle age).

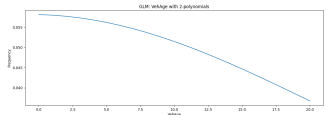


Figure 5: One of multiple potential GLMs to capture the relationship between vehicle age and claim count

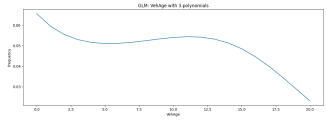


Figure 6: The final GLM chosen to capture the relationship between vehicle age and claim count

4.5.4 Representation Adjustment. During this phase, the expert made modifications to the selected relationship representation for the chosen variables. For example, for the variable vehicle age, the expert expressed that the propensity to file a claim would not increase for drivers of young vehicles (below five years of age) and should remain stable at an average claim count of 0.05 for vehicles less than five years old. This information was implemented as a rule that overwrote the model's predictions for this specific group. This step allowed the expert to fine-tune the selected models to better mirror her expertise. By defining rules and modifications for certain variable ranges or categories, she was able to introduce tailored changes that accounted for specific patterns or behaviors observed in the data.

4.5.5 Additional Rules. The expert was also given an opportunity to provide extra rules, which were subsequently added to the data. Through this, we were able to capture additional insights that might not have been accounted for through purely data-driven methodologies.

All the information gathered was standardized and mapped to the corresponding customers based on their respective values, leading to the addition of multiple new columns (at least one per variable) to the GAN training dataset, thereby incorporating the actuarial knowledge.

Besides the actuarial expert knowledge, attempts were made to incorporate car accident statistics provided by the German Insurance Federation (GDV) into the GAN training data. These statistics were aggregated based on driver age, region, vehicle power, and bonus malus. However, preliminary studies indicated that this data was not apt as expert knowledge due to its limited relevance to the dataset at hand and incorrect alignment between the German GDV values and the dataset's values based on French insurance standards. Therefore, this data was not further used in the analysis.

By involving the expert actuary in the selection and adjustment of models, we ensured that the representations were consistent with her domain knowledge and assumptions. The expert's involvement in the scope definition stage allowed us to identify the variables for which she could provide valuable insights. This focused approach ensured that the expert input was targeted and relevant to the variables under investigation.

4.6 Utilization of the Generative Adversarial Network

Informed by the research conducted by Cote et al. [4], our investigation concentrated on enhancing the performance of the Multicategorical Wasserstein Generative Adversarial Network with Gradient Penalty (MC-WGAN-GP). The architectural parameters and training guidelines were partially derived from the methodologies proposed by Camino et al. [3], Cote et al. [4].

Generative Adversarial Networks (GANs) comprise a pair of adversarial neural networks: the generator, which produces synthetic data intended to emulate the real data, and the discriminator, which determines the authenticity of the data points. The mutual competition between these networks facilitates the improvement of the synthetic data's quality, achieving higher resemblance to the actual data [4].

4.7 Wasserstein Generative Adversarial Network

To address the common issue of training instability in conventional GANs, the Wasserstein Generative Adversarial Network (WGAN) was introduced. This network variant employs a critic operating with real values instead of a binary classification discriminator. The training procedure was further stabilized by implementing a gradient penalty that ensures the critic's outputs remain within a predefined range [4].

4.8 Multicategorical Wasserstein Generative Adversarial Network with Gradient Penalty

In the context of this research, we utilized the MC-WGAN-GP to process tabular insurance policyholder data. Our MC-WGAN-GP model expands on the WGAN-GP structure and incorporates the method of handling multi-categorical variables suggested by Camino et al. [3]. In our design, every categorical variable is processed through a dense layer, followed by softmax activation in the generator. The outputs from these layers are subsequently concatenated to yield the final generator output [4].

Despite experimenting with various modifications to the model and the training process (such as modifying the final layer and introducing a gradient penalty layer), these modifications failed to yield significant improvements. This highlights the robustness of the original MC-WGAN-GP model [4]. The model successfully learned the structure and variability within the policyholder data, which set the stage for the generation of realistic synthetic data. The generated synthetic data can then be utilized to train predictive models, such as XGBoost, to predict claim counts.

It's worth mentioning that while this study presents a methodology for generating synthetic insurance data using GANs, the effectiveness of the results is contingent on numerous factors, such as the specifics of the employed dataset and the optimization of the model's hyperparameters. The selection of these hyperparameters is largely empirical, and the process is often informed by previous research [4].

4.9 Hyperparameter Optimization

In the endeavor of optimizing hyperparameters, the model architecture suggested by Camino et al. [3] was utilized as a reference point.

To determine the optimal hyperparameters for the MC-WGAN-GP, a grid search technique was applied. The selection of the best hyperparameters was mainly informed by previous research [4], given the resource constraints associated with training, and can be found in the study's Github repository. The validation of these hyperparameters was accomplished using the XGBoost model, which was previously defined on the validation set. The MC-WGAN-GP that generated the data leading to the most accurate GLM predictions was identified as the optimal model.

The scope of our grid search included parameters such as batch size, loss penalty, batch normalization decay for the discriminator and generator, discriminator's leaky parameter, noise size, the ratio of critic updates per generator update, L2 regularization for the discriminator, and sizes of the discriminator's hidden layers. Notably, the standard leaky ReLU activation function, used as the final layer of the critic, was replaced with a sigmoid function for certain grid search iterations.

Experiments included changes to the model and training protocol, such as replacing the last layer of the generator from a leaky ReLU layer to a Sigmoid layer and omitting the gradient penalty in some iterations. Attempts were also made to balance the positive and negative cases in the dataset by undersampling policies without claims for the initial 500 epochs of the GAN training process. Unfortunately, these changes failed to enhance the model's data generation quality.

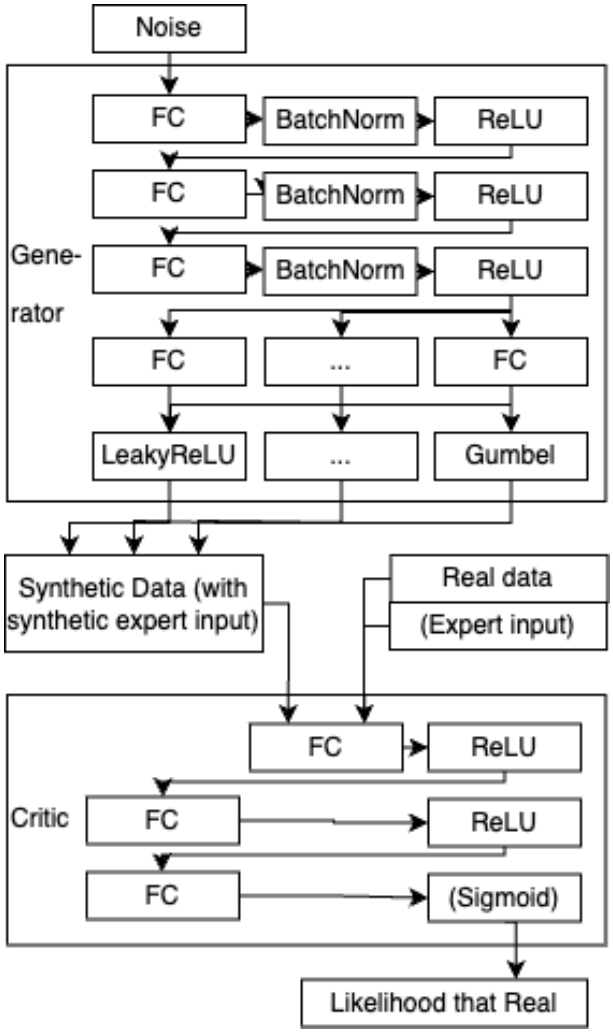


Figure 7: The architecture of the final version of the MC-WGAN-GP; in brackets, elements which are only used in certain scenarios

Training of the MC-WGAN-GP was carried out for 15,000 epochs in mini-batches, utilizing the binary cross-entropy loss function. Both the generator and discriminator shared the same zero-sum objective function, operating under a 3:1 training-validation split. For every two updates made to the discriminator, a single update was applied to the generator [4].

The final structure of the MC-WGAN-GP generator used in this study consisted of three fully connected layers, supplemented with batch normalization and ReLU, followed by a fully connected layer. The terminal layer utilized a leaky ReLU for continuous variables and a Gumbel function for categorical variables. The critic's final architecture comprised two fully connected layers, enhanced with a ReLU function and terminated by a sigmoid activation function (refer to Figure 7 for details).

Hyperparameter tuning was carried out across three different dataset sizes—large, medium, and small—to assess the scalability

and efficiency of our model. Further details regarding these datasets are available in the Data subsection of the Methodology.

The performance results of the model with different hyperparameters on large, medium, and small datasets will be presented in the subsequent results section.

Appendix ?? contains the optimal parameters identified for each dataset size.

4.10 XGBoost Model

To assess the quality of the different datasets, we trained an XGBoost model, a popular choice for structured and tabular data, on these datasets. Various versions of the same XGBoost model, all retaining consistent hyperparameters, were trained on different datasets. These datasets included the synthetic data generated by the GAN model with and without expert input, as well as the original dataset.

In line with common practice in the insurance sector, we assumed the claim count to follow a Poisson distribution [17]. As such, an XGBoost model presuming a Poisson distribution was trained on the input data, maintaining the same model architecture for both synthetic and real datasets.

Instead of undertaking a comprehensive grid search for the XGBoost hyperparameters, we made use of the findings from previous research that had already identified optimal parameters for training the XGBoost model on our dataset. Therefore, we conducted only a limited grid search, comparing the performance of the hyperparameters suggested by Martínez de Lizarduy Kostornichenko [15] and an alternative source (XXX). The hyperparameters from the latter study exhibited superior predictive performance and were hence selected for further analysis.

This systematic approach to model development and evaluation provides a robust method to compare the quality of the datasets generated by our GAN models. By training and testing the XGBoost model on the synthetic and real datasets, we can objectively assess the resemblance of the generated data to the original dataset. Moreover, this comparison allows us to evaluate the impact of expert input on the quality of the synthetic data produced by the GAN models, thereby highlighting the benefits and potential limitations of this approach.

4.11 Evaluation

Our proposed methodology was evaluated through a comparative analysis of the performance of identical XGBoost models trained on three distinct datasets - the original dataset, a synthetic dataset generated by the MC-WGAN-GP with expert input, and a synthetic dataset generated by the MC-WGAN-GP without expert input. In the case of the original dataset, 100 different bootstrap versions were used for both training and testing the models. For the synthetic data, the GAN was used to generate 100 distinct versions of 100,000 synthetic policies, each of which was used to train the model. The trained models were subsequently evaluated against the 100 bootstrap versions of the test set.

In order to derive the most reliable results, a two-sample, one-sided t-test was utilized to verify both hypotheses. For each pair of variables under comparison, the Poisson deviance was calculated for each of the 100 bootstrap versions of the test set. A p-value less than the conventional significance level of 0.05 was taken as evidence to reject the null hypothesis in favor of the alternative.

The choice of a one-sided test was predicated on the study's main objective of ascertaining whether the GAN can generate synthetic data that is not worse than the original dataset in terms of predictive performance (Hypothesis 1), and if expert input can enhance the quality of the generated synthetic data (Hypothesis 2).

It is noteworthy to mention that the t-test assumptions of independence and normality were deemed to be met in this case. The bootstrap versions of the test set were independent of each other, satisfying the independence assumption. The Central Limit Theorem justifies the normality assumption as the number of bootstrap versions is sufficiently large.

For hypothesis testing, we focused on the best-performing models as determined through the hyperparameter search. In this context, let's denote P_a as the Poisson Deviation from the predictions made by the XGBoost model trained on the synthetic data without expert input, and P_0 as the Poisson Deviation from the predictions made by the XGBoost model trained on the original data. For Hypothesis 1, which pertains solely to the MC-WGAN-GP trained on the large dataset without expert input, the null and alternative hypotheses can be represented as follows:

$$H_0 : P_a = P_0$$
$$H_a : P_a < P_0$$

For Hypothesis 2, we define $P_{e,s}$, $P_{e,m}$, and $P_{e,l}$ as the Poisson Deviations from predictions made by the model trained on synthetic data generated by the GAN trained with expert input for small, medium, and large datasets, respectively. Similarly, $P_{ne,s}$, $P_{ne,m}$, and $P_{ne,l}$ represent the Poisson Deviations from predictions made by the model trained on synthetic data without expert input for the small, medium, and large datasets, respectively. Thus, for Hypothesis 2, the null and alternative hypotheses are as follows:

$$H_0 : P_{e,s} = P_{ne,s}; P_{e,m} = P_{ne,m}; P_{e,l} = P_{ne,l}$$
$$H_a : P_{e,s} < P_{ne,s}; P_{e,m} < P_{ne,m}; P_{e,l} < P_{ne,l}$$

Hypothesis 2 would only be supported if the conditions specified in the alternative hypothesis are met for all datasets.

Given the nature of the evaluation, potential limitations include the assumptions inherent in the chosen hypothesis testing method, the reliance on the predictive accuracy of the XGBoost model, and the use of a single, albeit complex, synthetic dataset generation method. Nonetheless, these limitations are counterbalanced by the rigorousness of the evaluation procedure and the robustness of the chosen metrics, which allow for a comprehensive assessment of the model's performance.

By employing this detailed evaluation procedure, we aim to establish clear evidence for or against the potential of GANs in synthesizing insurance datasets, and the effect of including expert input in this process. Future research can build on these findings to further refine the data synthesis process and expand the applications of GANs in the insurance industry.

5 RESULTS

Our investigation comprised of two focal research questions. Initially, we analyzed whether the MC-WGAN-GP, trained on real insurance claim data, generates data resembling its training dataset.

The comparison of the Poisson deviance for our baseline scenario (predicting an average claim count for all customers; $M = 31.909$, $SD = 0.279$) and the model trained on data produced by the baseline MC-WGAN-GP ($M = 31.536$, $SD = 0.275$) resulted in statistically significant support ($t(99) = 9.497$, $p < 0.001$) for our hypothesis. This suggests that the boosting model trained on the data generated by the MC-WGAN-GP predicts the outcome variable (claim count) more effectively than the baseline model (see Table 1. The comparison of RMSE for the boosting model trained on synthetic data ($M = 0.238$, $SD = 0.004$) with the results obtained by Kuo [12] ($M = 0.242$) indicates that our findings are consistent with the previous research.

Subsequently, we aimed to investigate whether the inclusion of expert knowledge in the MC-WGAN-GP training data would enhance the quality of the data it generates. For this analysis, we conducted a comparative study across different models trained on various dataset sizes.

In the context of the large dataset, the Poisson Deviation for the XGBoost model trained on synthetic data produced by the GAN (trained on data including expert knowledge, $M = 31.457$, $SD = 0.259$) was compared with the XGBoost model trained on synthetic data generated by the GAN (trained without expert knowledge, $M = 31.536$, $SD = 0.275$). The results demonstrate that the inclusion of expert input led to statistically significant improvements in predicting claim counts ($t(99) = -2.063$, $p = 0.020$).

Conversely, with the small dataset, the inclusion of expert input did not result in significant improvements in the predictions of the XGBoost model ($M = 37.214$, $SD = 0.1.454$) compared to its counterpart trained on traditional GAN synthetic data, $M = 32.714$, $SD = 0.312$, $t(99) = 30.115$, $p = 1.000$. However, these results should be interpreted carefully since none of the two MC-WGAN-GP models performed better than predicting the average claim count for each case ($M = 31.909$, $SD = 0.279$; see Table 2.

Looking at the distributions of key variables of the models, we can see that they are not very similar yet, a result likely since we don't optimize the GAN for generation of the correct data distribution but for generation of correct relationship between the independent variables and the dependent variable (see Figures 8 until 10 for first insights since not all results are available yet).

THESE FIGURES ARE NOT FINISHED YET!

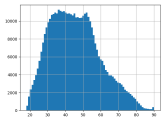


Figure 8: Distribution for age for the real data

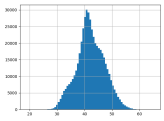


Figure 9: Distribution for age for the generated data with expert input

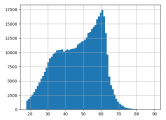


Figure 10: Distribution for age for the generated data without expert input

Looking at the relationship between key variables of the models and the dependent variables, we can see the relationships have been correctly picked up by the models with expert input but not by the models without expert input (see Figure **??**-13 for first insights since not all results are available yet).

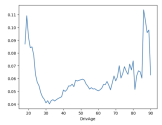


Figure 11: Relationship between age and frequency for the predictions based on real data

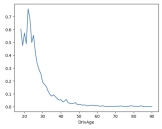


Figure 12: Relationship between age and frequency for the predictions based on generated data with expert input

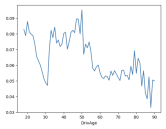


Figure 13: Relationship between age and frequency for the predictions based on generated data without expert input

6 DISCUSSION

Our study was motivated by the objective of bridging gaps in the domains of artificial intelligence in underwriting, data generation in insurance, and knowledge representation in AI. Two central research questions framed our study.

Firstly, we investigated the possibility of training a Modified Wasserstein Generative Adversarial Network with Gradient Penalty (MC-WGAN-GP) to comprehend the relationship between dependent and independent variables in a motor third-party liability (MTPL) policy risk dataset. The results derived offered an affirmative answer to this question, substantiating our hypothesis.

The second question explored the impact of incorporating expert knowledge into MC-WGAN-GP training on enhancing the model's ability to comprehend the relationship between dependent

Training set size	Data source	Poisson Deviance	MAE	RMSE	GINI
433,728	Dummy model	31.909 (31.854, 31.965)	0.101 (0.101, 0.101)	0.236 (0.236, 0.237)	1.000 (0.000, 0.000)
433,728	Synthetic	31.536 (31.480, 31.590)	0.108 (0.107, 0.109)	0.238 (0.238, 0.239)	0.159 (0.158, 0.161)
433,728	Real data	30.377 (30.325, 30.428)	0.102 (0.102, 0.102)	0.234 (0.234, 0.234)	0.214 (0.213, 0.214)

Table 1: Comparison of the different pipelines for research question 1 on their main metrics with confidence intervals (in brackets)

Training set size	Data source	Expert input	Poisson Deviance	MAE	RMSE	GINI
433,728	Synthetic	No	31.536 (31.480, 31.590)	0.108 (0.107, 0.109)	0.238 (0.238, 0.239)	0.159 (0.158, 0.161)
433,728	Synthetic	Yes	**31.457 (31.406, 31.508)**	0.107 (0.106, 0.108)	0.237 (0.236, 0.238)	0.151 (0.149, 0.153)
5,000	Synthetic	No	**32.714 (32.651, 32.776)**	0.099 (0.099, 0.100)	0.240 (0.240, 0.240)	0.279 (0.278, 0.281)
5,000	Synthetic	Yes	37.214 (36.924, 37.504)	0.136 (0.134, 0.139)	0.248 (0.246, 0.249)	0.365 (0.359, 0.372)

Table 2: Comparison of the different pipelines for research question 2 on their main metrics with confidence intervals (in brackets)

and independent variables. The response to this question was partially positive, as we found that introducing expert knowledge to a GAN trained on smaller datasets (e.g., N=5,000) did not statistically significantly improve its capability. However, for larger datasets, expert knowledge noticeably improved the predictions made by the XGBoost model trained on the data.

The addition of expert knowledge as proposed in our hypothesis protected the model from overfitting. Interestingly, our study also suggests that expert knowledge becomes beneficial only when a minimum dataset size is accessible. Training MC-WGAN-GP on smaller datasets with expert knowledge appeared to deteriorate the model's performance. A plausible explanation for this is the high correlation between the expert knowledge and the dependent variable (claim count). Given the tendency of GANs towards instability, the strong correlation could have destabilized the GAN, as per Durall et al. (2020).

Our study, despite supporting both hypotheses, also reveals that integrating expert knowledge did not improve the model performance to the anticipated extent. This limitation could stem from various sources. Notably, while we introduced expert knowledge into the GAN training, it was primarily integrated into the critic and only evaluated against the real dataset. The generator, which did not receive any direct expert knowledge, gained from this knowledge indirectly through improved feedback from the critic. If either the critic or the generator failed to accurately learn the relationship between the expert knowledge and claim count, the use of expert knowledge could even impair model performance by potentially amplifying the amount of random error experienced by the generator during backpropagation.

We propose that future research focus on three key areas. Firstly, having gained initial insights into how expert knowledge can improve GAN model performance, research should investigate when expert knowledge enhances model performance and when it leads to instability and mode collapse. We have already established that sufficient data availability is one of the prerequisites.

Secondly, researchers should investigate the most effective method for integrating additional information. For instance, future studies could explore methods to have a more direct influence on the GAN. One promising approach is the addition of an extra layer to the generator to incorporate the expert knowledge. However, the challenge lies in managing the potential problems with backpropagation of the loss function during the training of the generator. Lastly, research should focus on approaches to gather expert knowledge once the effective inclusion of this knowledge into the GAN is established.

Our research had some constraints, including the bias of expert knowledge based on the German market, which may not align with the French MTPL market, and the lack of diverse variables in the dataset for the actuarial expert to provide information. Future research should ensure a correlation between expert input and the dataset used for validation. We hope for more data sharing by insurers to enrich future research.

Despite these limitations, our study offers numerous strengths. For instance, through the use of various metrics and multiple dataset size scenarios, we gained an in-depth understanding of the workings of MC-WGAN-GP with and without expert knowledge. We also identified Poisson Deviation as the optimal performance measurement metric and uncovered MAE and RMSE to be potentially not suitable to measure model quality in Poisson distributed models. Moreover, we confirmed that GANs can pick up information on relationships between claim frequency and policy-related variables, at least within insurance datasets.

Furthermore, the expert input was integrated systematically and thoroughly, enhancing the modeling process by fusing domain knowledge to refine the relationships between independent and dependent variables. Involving an actuarial expert in the model selection and adjustment stages aligned our representations with her expertise and assumptions. By identifying the variables that she could provide valuable insights into during the scope definition stage, we ensured a targeted and relevant expert input.

This expert knowledge, once standardized and mapped, was integrated seamlessly into the dataset, enhancing compatibility with our overall modeling framework. By appending expert-derived variables as new columns, we expanded the feature space and incorporated the refined relationships into the modeling process. The integration of expert knowledge played a crucial role in enhancing the modeling process. The incorporation of the actuarial expert's insights allowed us to capture nuances and specific patterns that

could have been overlooked by purely data-driven methods, thereby enhancing the accuracy and comprehensiveness of our modeling results.

Although our research focused on a broad research question, it provided valuable insights into the incorporation of expert knowledge into the GAN. We established that including expert knowledge into the critic, rather than the generator, improved the performance of the generator. Furthermore, we developed a streamlined process for acquiring actuarial expert knowledge, a methodology that could also be applicable to other AI research areas.

In conclusion, our research offers a significant step forward in understanding how to generate synthetic data using GANs. By enhancing the quality of tabular GAN models, particularly the MC-WGAN-GP, through the integration of expert knowledge and employing various information integration strategies, we believe our findings can augment risk modeling and decision-making processes in the insurance industry. Future studies building upon our research can delve deeper into the questions and ideas we have posed, thus pushing the boundaries of AI application in the insurance industry even further.

This research contributes to the development of more accurate and reliable data generation processes in the underwriting industry, facilitating data sharing among insurers and the research community. It also has the potential to create non-confidential publicly available datasets for research purposes. Moreover, our study may lead to improved methods for insurance pricing, benefiting both insurers and policyholders. Additionally, the exploration of neuro-symbolic input in GAN training contributes to the fields of data science and actuarial studies. The idea of incorporating neuro-symbolic input has the potential to advance GAN models in other industries, opening up new possibilities for data-driven decision making.

7 CONCLUSION

In this paper, we presented, implemented, and compared the use of expert input for different dataset sizes to synthesize insurance data. Training the MC-WGAN-GP on X data with/without expert input synthesized the most realistic data. It generated data which was very similar (accounting for both univariate and multivariate relationships) to the real data. Future work can start from the point that expert input as provided right now did not improve the model but can have the potential to improve MC-WGAN-GP training if injected at some other point or in another way. In any case, the MC-WGAN-GP is a promising tool for synthesizing and protecting private and important data as has been seen in the comparison of the models to the baseline case.

REFERENCES

[1] Ramnath Balasubramanian, Ari Libarikian, and Doug McElhaney. 2021. https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance

[2] Liliana Byczkowska-Lipińska, Mariusz Szydło, and Piotr Lipiński. 2009. *Expert Systems in the Medical Insurance Industry.* Springer Berlin Heidelberg, Berlin, Heidelberg, 189–199. https://doi.org/10.1007/978-3-642-04462-5_19

[3] Ramiro Camino, Christian Hammerschmidt, and Radu State. 2018. Generating multi-categorical samples with generative adversarial networks. *arXiv preprint arXiv:1807.01202* (2018).

[4] Marie-Pier Cote, Brian Hartman, Olivier Mercier, Joshua Meyers, Jared Cummings, and Elijah Harmon. 2020. Synthesizing property & casualty ratemaking datasets using generative adversarial networks. *arXiv preprint arXiv:2008.06110* (2020).

[5] Hubert Dichtl, Wolfgang Drobetz, and Martin Wambach. 2017. A bootstrap-based comparison of portfolio insurance strategies. *The European Journal of Finance* 23, 1 (2017), 31–59.

[6] European Parliament and Council of the European Union. [n. d.]. *Regulation (EU) 2016/679 of the European Parliament and of the Council.* https://data.europa.eu/eli/reg/2016/679/oj

[7] Matthias Feurer, Jan N. van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Mueller, Joaquin Vanschoren, and Frank Hutter. [n. d.]. OpenML-Python: an extensible Python API for OpenML. *arXiv* 1911.02490 ([n. d.]). https://arxiv.org/pdf/1911.02490.pdf

[8] Tobias Fissler, Christian Lorentzen, and Michael Mayer. 2022. Model comparison and calibration assessment: user guide for consistent scoring functions in machine learning and actuarial practice. *arXiv preprint arXiv:2202.12780* (2022).

[9] Andrea Gabrielli and Mario V. Wüthrich. 2018. An individual claims history simulation machine. *Risks* 6, 2 (2018), 29.

[10] Mark Goldburd, Dan Khare, Anand amd Tevet, and Dmitriy Guller. 2020. Generalized Linear Models for Insurance Rating.

[11] Chih Hsun Hsieh and Paul P Wang. 2011. Linguistic evaluation system and insurance. *New Mathematics and Natural Computation* 7, 03 (2011), 383–411.

[12] Kevin Kuo. 2019. Generative synthesis of insurance datasets. *arXiv preprint arXiv:1912.02423* (2019).

[13] Xenofon Liapakis. 2018. A GDPR Implementation Guide for the Insurance Industry. *International Journal of Reliable and Quality E-Healthcare (IJRQEH)* 7, 4 (2018), 34–44.

[14] Cynthia Marling, Mohammed Sqalli, Edwina Rissland, Hector Muñoz-Avila, and David Aha. 2002. Case-based reasoning integrations. *AI magazine* 23, 1 (2002), 69–69.

[15] Viktor Martínez de Lizarduy Kostornichenko. 2021. *Comparative performance analysis between Grandient Boosting models and GLMs for non-life pricing.* Master's thesis.

[16] Héctor Munoz-Avila, David W Aha, Len Breslow, and Dana Nau. 1999. HICAP: An interactive case-based planning architecture and its application to noncombatant evacuation operations. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence.* 870–875.

[17] Alexander Noll, Robert Salzmann, and Mario V Wuthrich. 2020. Case study: French motor third-party liability claims. *Available at SSRN 3164764* (2020).

[18] Pietro Parodi. 2014. *Pricing in general insurance.* CRC press.

[19] Jim Prentzas and Ioannis Hatzilygeroudis. 2007. Categorizing approaches combining rule-based and case-based reasoning. *Expert Systems* 24, 2 (2007), 97–122.

[20] Jim Prentzas and Ioannis Hatzilygeroudis. 2011. Neurules-a type of neuro-symbolic rules: An overview. In *Combinations of Intelligent Methods and Applications: Proceedings of the 2nd International Workshop, CIMA 2010, France, October 2010.* Springer, 145–165.

[21] Jim Prentzas and Ioannis Hatzilygeroudis. 2016. Assessment of life insurance applications: an approach integrating neuro-symbolic rule-based with case-based reasoning. *Expert Systems* 33, 2 (2016), 145–160.

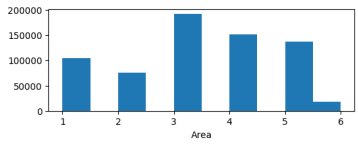
[22] Ronald Richman and Mario V Wüthrich. 2022. LocalGLMnet: interpretable deep learning for tabular data. *Scandinavian Actuarial Journal* (2022), 1–25.

[23] Jürg Schelldorfer and Mario V Wuthrich. 2019. Nesting classical actuarial models into neural networks. *Available at SSRN 3320525* (2019).

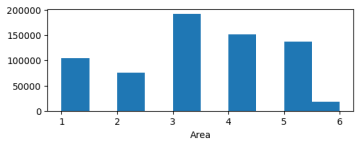
[24] Mario V Wüthrich and Michael Merz. 2019. Yes, we CANN! *ASTIN Bulletin: The Journal of the IAA* 49, 1 (2019), 1–3.

[25] Yasin Yazici, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, and Vijay Chandrasekhar. 2020. Empirical analysis of overfitting and mode drop in gan training. In *2020 IEEE International Conference on Image Processing (ICIP).* IEEE, 1651–1655.

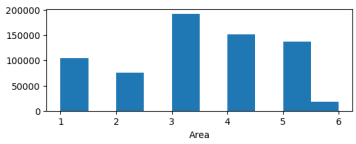
Appendix A COMPARISON OF DISTRIBUTIONS FOR DIFFERENT DATASETS (BIG DATASET SCENARIO)



(a) Real data

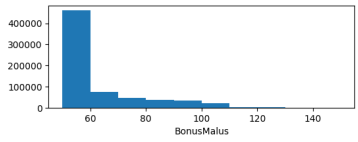


(b) Synthetic with expert input

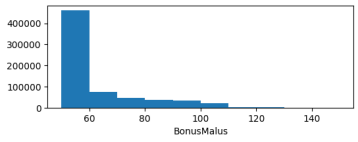


(c) Synthetic without expert input

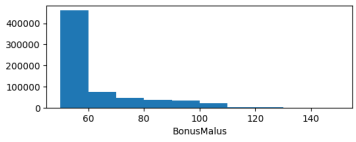
Figure 14: Distribution of area variable for the different datasets (big dataset scenario)



(a) Real data

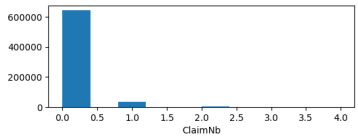


(b) Synthetic with expert input

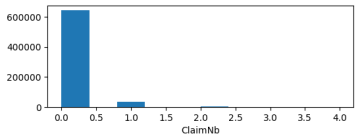


(c) Synthetic without expert input

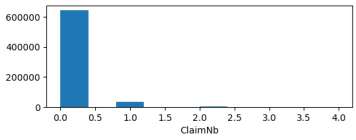
Figure 15: Distribution of bonus malus variable for the different datasets (big dataset scenario)



(a) Real data

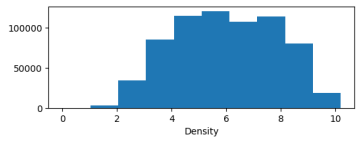


(b) Synthetic with expert input

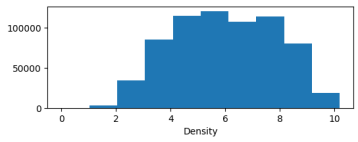


(c) Synthetic without expert input

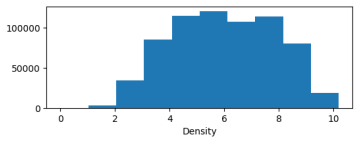
Figure 16: Distribution of claim count variable for the different datasets (big dataset scenario)



(a) Real data



(b) Synthetic with expert input



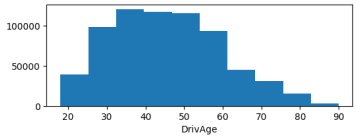
(c) Synthetic without expert input

Figure 17: Distribution of density variable for the different datasets (big dataset scenario)

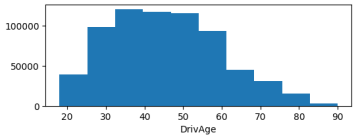
Appendix B EXPERT INPUT PROVIDED

After being shown

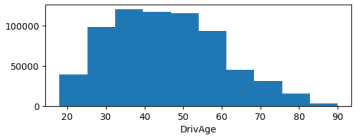
- Density: GLM1, do not adjust
- BonusMalus: GLM2 -> Group above/below 100, take average of both groups
- BonusMalus: GLM2 passt
- VehPower: Take only average, no model -> No need to model
- DrivAge: GLM5



(a) Real data

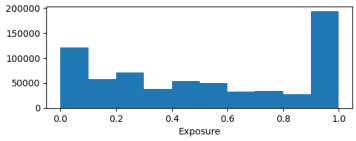


(b) Synthetic with expert input

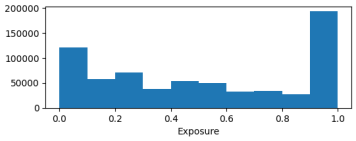


(c) Synthetic without expert input

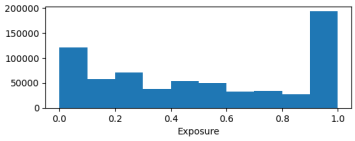
Figure 18: Distribution of driver age variable for the different datasets (big dataset scenario)



(a) Real data

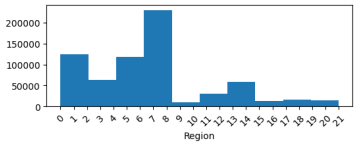


(b) Synthetic with expert input

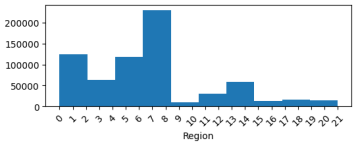


(c) Synthetic without expert input

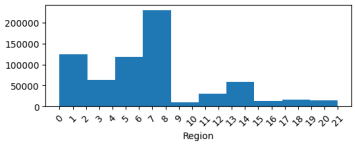
Figure 19: Distribution of exposure variable for the different datasets (big dataset scenario)



(a) Real data

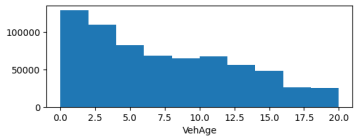


(b) Synthetic with expert input

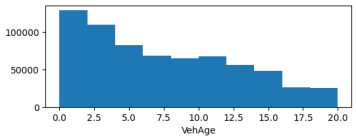


(c) Synthetic without expert input

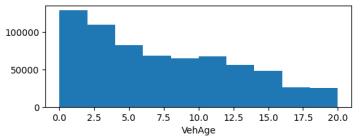
Figure 20: Distribution of region variable for the different datasets (big dataset scenario)



(a) Real data

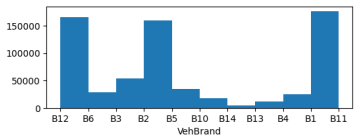


(b) Synthetic with expert input

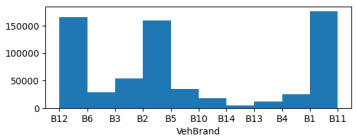


(c) Synthetic without expert input

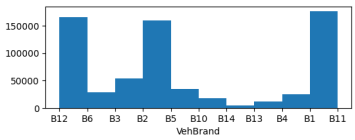
Figure 21: Distribution of vehicle age variable for the different datasets (big dataset scenario)



(a) Real data

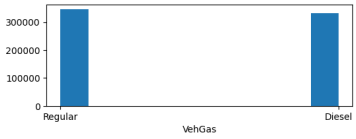


(b) Synthetic with expert input

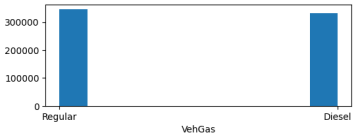


(c) Synthetic without expert input

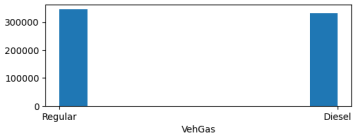
Figure 22: Distribution of vehicle brand variable for the different datasets (big dataset scenario)



(a) Real data

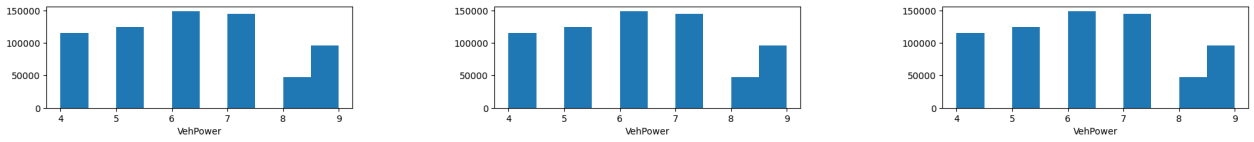


(b) Synthetic with expert input



(c) Synthetic without expert input

Figure 23: Distribution of vehicle gas variable for the different datasets (big dataset scenario)



(a) Real data **(b) Synthetic with expert input** **(c) Synthetic without expert input**

Figure 24: Distribution of vehicle power variable for the different datasets (big dataset scenario)

B.1 Vehicle Age

Upon being provided with regressions representing the potential relationships between vehicle age and claim count, the actuarial expert decided for a representation of the relationship as follows (see Figure 25 for the representation):

- Vehicles with an age below 5 years should be grouped into one group and their average claim count taken
- Vehicles with an age above 5 years should be represented by a GLM with $ClaimCount = \beta_0 + \beta_1 * VehicleAge + \beta_2 * VehicleAge^2 + \beta_3 * VehicleAge^3$

This representation was then translated into a new column in the input data which was mapped to each customers vehicle age and provided a respective expert input coefficient for each customer, based on the coefficients found in Figure 25.

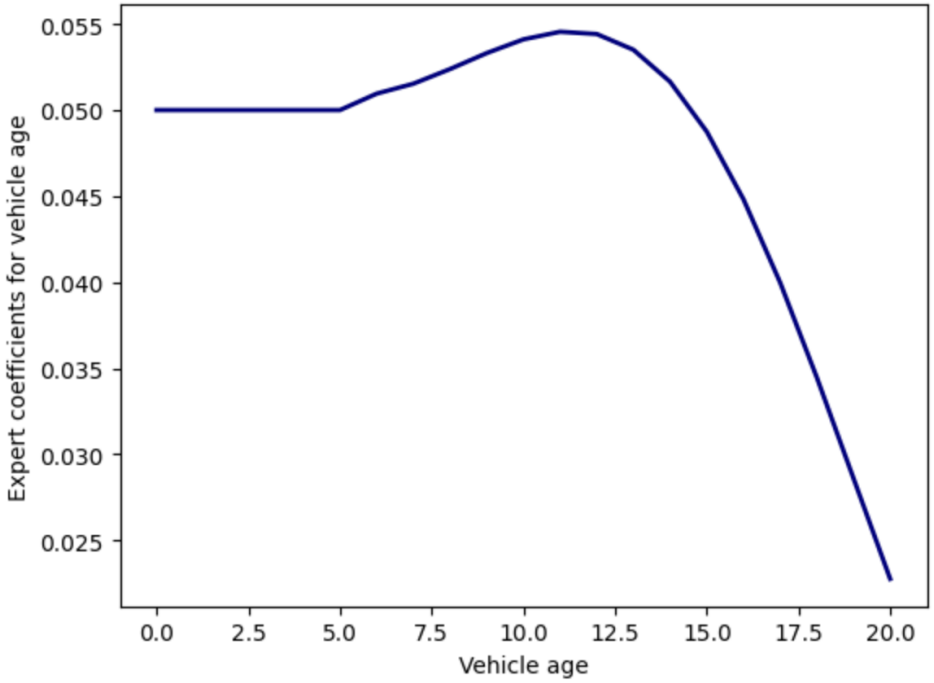


Figure 25: Chosen representation for the relationship between vehicle age and claim count

B.2 Bonus Malus

Additionally, the expert added a variable which indicates if a customer had a bonus malus level below/above 100. This reflects in her eyes the fact that customers above this threshold behave much more risky than customers below this bonus malus level. The split was introduced into the model as grouping the customers into customers with a bonus malus above/below 100 and taking the average claim number. It resulted in a new variable with the coefficients X for customers with a lower bonus malus than 100 and Y for customers with a higher bonus malus than 100 (see Figure **??**).

Appendix C GRIDSEARCH

To find the best performing model, a hyperparameter gridsearch was conducted. Due to infrastructural limits (usage of high-capacity Snellius GPU's was limited to 50000 SBU's), our gridsearch was limited to only the combinations of parameters that produced the models which best performed in terms of RMSE and MAE in previous research (see [4] for the entire hyperparameter search).

Additionally, research on introducing major changes to the model was conducted (see Method section for more information). Since this research was conducted for fewer training epochs (less than 2000) and on local hardware, and did not reveal any significant WC-GAN-GP improvements compared to the already existing architecture (i.e. Poisson Deviance was always higher than 35 for the validation dataset during training of the GAN), its results are not included