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 29, 2023

	UvA Supervisor	External Supervisor
Title, Name	Erman Acar	Georg Maerz, Isabella Marinelli
Affiliation	UvA Supervisor	External Supervisor, External Supervisor
Email	e.acar@uva.nl	marz.g.com@gmail.com, isabella.marinelli.95@gmail.com

ABSTRACT

Insurance data privacy is a pivotal concern that often limits the potential for robust predictive modeling. This thesis presents a novel approach to generate synthetic insurance datasets using Generative Adversarial Networks (GANs), specifically the Modified Wasserstein GAN with Gradient Penalty (MC-WGAN-GP), and explores the effect of integrating expert knowledge in the data generation process. The generated synthetic datasets' performance was evaluated by training XGBoost models and comparing the prediction results to those obtained from a model trained on the original dataset. Poisson Deviance was used as the main performance metric, with the Area Under the Receiver Operating Characteristic Curve (AUC) as a secondary metric.

The results affirm the potential of the MC-WGAN-GP in generating synthetic datasets that yield comparable predictive accuracy to models trained on original data. Moreover, the inclusion of expert input during the GAN training phase significantly enhanced the predictive performance of the models. However, the benefits of expert input did not extend to smaller subsamples of the original dataset.

These findings illuminate the potential for using GANs to mitigate privacy concerns in the insurance industry, and the value of integrating expert knowledge into the synthetic data generation process. The study also underscores the need for additional research to further refine these methods and assess their applicability across different datasets and predictive models.

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, 15].

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, 14].

This thesis contributes to the burgeoning field of synthetic data generation in insurance, primarily by exploring the impact of integrating actuarial expert knowledge during the training of MC-WGAN-GP models. As far as we know, the application of such a near-symbolic approach to GAN training in an insurance context is a not well investigated area and this thesis aims to fulfil this gap [22, 23].

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.

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[1]?
- *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.
- *Hypothesis 1.2*: Compared to Hypothesis 1.1, the XGBoost model built on the data generated by the MC-WGAN-GP yields predictions that are statistically comparable to those made by a model trained on the original dataset, as evaluated by Poisson Deviance.
- *Hypothesis 2.1*: The XGBoost model built on the data generated by the MC-WGAN-GP with expert knowledge included in the training provides more accurate predictions (as evaluated by Poisson Deviance) than the XGBoost model built on the data generated by the MC-WGAN-GP without expert knowledge included.

[1] Poisson Deviance furnishes an analysis of the fit offered by the model under examination, assuming Poisson Distribution of the data.

- *Hypothesis 2.2*: The effect described in Hypothesis 2.1 is even more prominent on a dataset subsample of the original dataset ($N = 5,000$[2]).

The subsequent sections of this thesis 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 thesis concludes with a discussion on the potential implications of our findings and recommendations for future research directions.

2 RELATED WORK

The proliferation of machine learning (ML) and AI has stimulated an increasing 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.

2.1 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, 20]. 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 [19, 24, 25, 27].

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 [24, 25, 27].

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 [19]. Furthermore, this choice aligns with our aim to compare the quality of synthetic data to real data.

2.2 Data Synthesis in Insurance

With the increasingly rigid data privacy regulations and confidentiality requirements, it is becoming more crucial to explore and implement alternative methods that can effectively utilize real-world data within 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. [11], 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 [14] 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].

2.3 Knowledge Representation for Insurance

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 [16, 18, 21–23]. 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 [22, 23]. As delineated by Prentzas and Hatzilygeroudis [21], 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 [16, 18].

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 [21, 22].

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 [12] 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).

[2] where N is the dataset size

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 [16, 18, 21–23]. Moreover, we developed a system to assess customer risk based on expert rules, aligning with prior studies in insurance.

3 METHODOLOGY

3.1 Research Design

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

We design five modeling pipelines and compare them across three different scenarios of training set sizes (see Figure 1). 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 integrates 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 incorporates 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 remains constant, allowing for unbiased evaluation.

Since our dependent variable (claim count) is Poisson distributed, performance evaluation of the models is done using Poisson Deviance [8]. We do not use other measures often used in GAN research, such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), since they do not apply to the domain of Poisson distributed variables [4, 8, 14].

3.2 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 [26]. However, for accessibility and replicability, the methodology was also designed to run on a more modest machine, such as the MacBook Air M2 with Apple M2 silicon processor. Although the MacBook Air M2's computational capabilities are lower compared to the supercomputer, the training process is still feasible with extended durations.

3.3 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*).

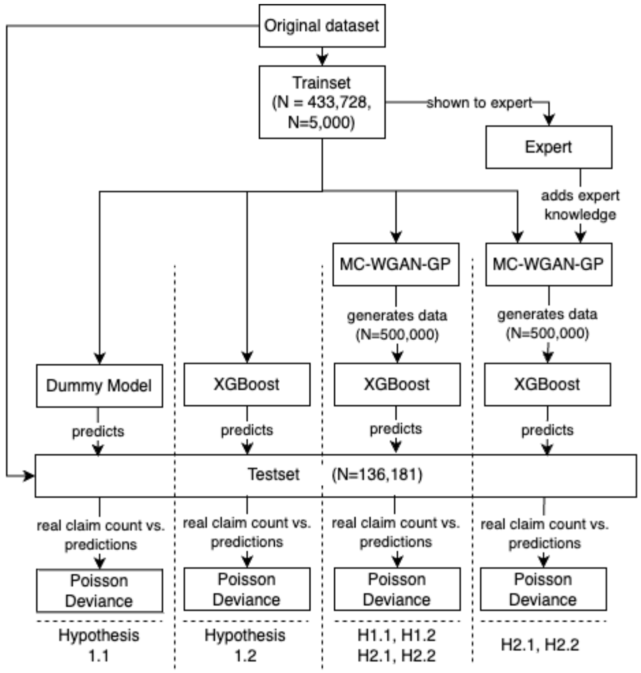


Figure 1: Structure of the different pipelines for each hypothesis

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, the variables *Exposure* and *ClaimNb* 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 with each other except for *Exposure*, policy ID, *IDpol*, and claim count (*ClaimNb*). 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).

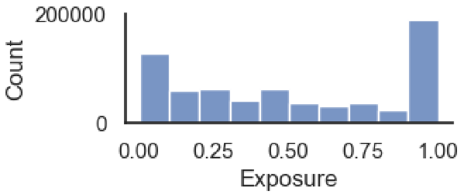


Figure 2: Distribution of exposure values up to one year

As the dependent variable, the *ClaimNb* was further examined. The dataset revealed that 95% of the policies (N=643,953) had no claims filed (i.e., *ClaimNb* = 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 3 for *ClaimNb* $<$ 5).

3.4 Data Preprocessing

A standardized data preprocessing pipeline was implemented, based on the recommendations from previous research [4, 19]. 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 Count:** 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.

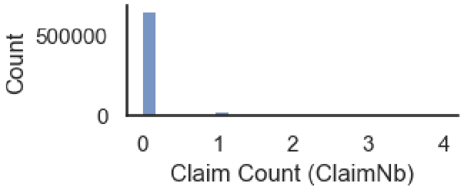


Figure 3: Distribution of claim values <= 4

- **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.

3.5 Incorporating Expert Knowledge

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 knowledge 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 [28]. 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.

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. However, preliminary studies indicated that this data was not suitable 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.

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

3.5.1 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 (*DrivAge*), *Bonus Malus*, and Vehicle Age (*VehAge*) were defined as being within the scope of expert knowledge.

3.5.2 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 $ClaimCount = \beta_0 + \beta_1 \times VehAge + \beta_2 * VehAge^2$, can be found in Figure 4. 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.

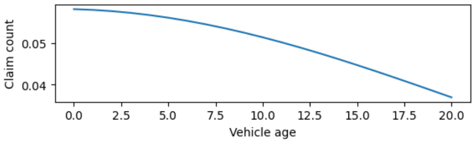


Figure 4: One of multiple potential GLMs to capture the relationship between vehicle age and claim count

3.5.3 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 5 for vehicle age).

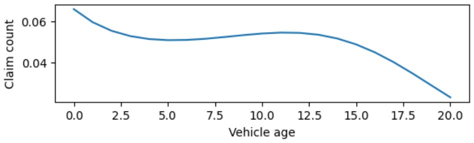


Figure 5: The final GLM chosen to capture the relationship between vehicle age and claim count

3.5.4 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 (refer to Figure 6 for vehicle age).

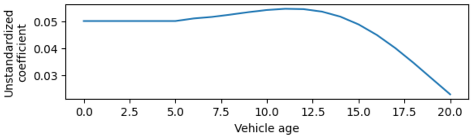


Figure 6: The final relationship representation with Vehicle age $<= 5.00$ **overwritten by actuarial decision to keep the coefficient at** 0.05

3.5.5 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.

Expert Knowledge Integration into GAN To incorporate the expert knowledge into the training of the Generative Adversarial Network (GAN), all the gathered expert knowledge was standardized and systematically mapped to the corresponding customers in our dataset. This mapping was based on the values for the independent variables that each customer possessed and resulted in the creation of multiple new columns within the GAN training dataset - at least one for each variable.

The enriched training dataset was then utilized to train the critic within our GAN model. This approach ensured that the critic was being trained with both the original data and the added expert knowledge, thereby bolstering its performance and accuracy. Importantly, this method only indirectly trained the generator with the expert knowledge; the generator did not have direct access to this knowledge.

However, the generator did benefit from this knowledge via the improved feedback it received from the enhanced critic. This indirect mode of knowledge inclusion harnessed the adversarial dynamic within the GANs, resulting in the generator being guided towards generating synthetic data that not only retained the inherent patterns present in the original dataset but also incorporated the additional insights provided by the expert knowledge. Thus, the synthetic data produced by the generator was a reflection of both the original data structure and the expert's specialized insights.

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 knowledge was targeted and relevant to the variables under investigation.

3.6 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].

3.7 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].

3.8 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].

3.9 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.

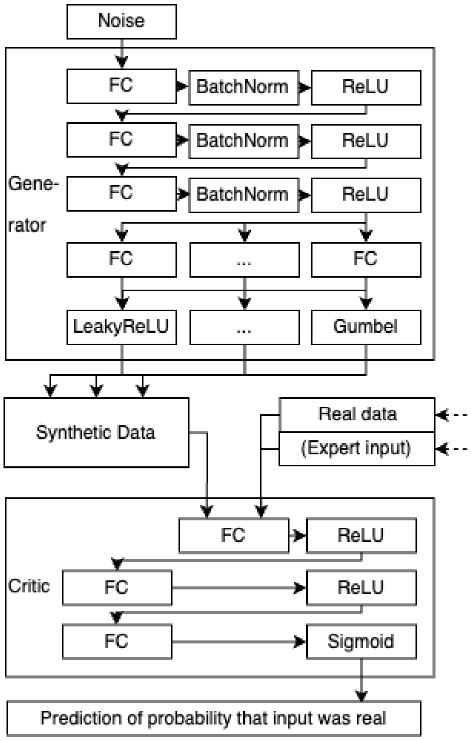


Figure 7: The architecture of the final version of the MC-WGAN-GP; in brackets, elements which are only used in certain scenarios

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.

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 D contains the optimal parameters identified for each dataset size.

3.10 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 knowledge, 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 [19]. 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 [17] and an alternative source [13]. 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 knowledge on the quality of the synthetic data produced by the GAN models, thereby highlighting the benefits and potential limitations of this approach.

3.11 Evaluation

The proposed methodology's efficacy was appraised via comparative performance analysis of identical XGBoost models, each trained on one of three distinct datasets - the original dataset, a synthetic dataset generated by the MC-WGAN-GP incorporating expert knowledge, and another synthetic dataset produced by the MC-WGAN-GP without such expertise. For the original dataset, we created 100 unique bootstrap versions for both model training and evaluation. In the case of synthetic data, the GAN was employed to generate 100 separate iterations of 100,000 synthetic policies, with each used for model training. Subsequent evaluations of these trained models were performed against the 100 bootstrap versions of the test set.

To ensure the reliability of our results, a two-sample, one-sided t-test was adopted to verify both hypotheses. The Poisson deviance was computed for each of the 100 bootstrap versions of the test set for every variable pair under consideration. A p-value below the conventional significance threshold of 0.05 was interpreted as grounds to reject the null hypothesis in favor of the alternative.

Our choice of a one-sided test hinged upon the study's principal goal, namely, to determine if the GAN can produce synthetic data that does not underperform the original dataset in terms of predictive capacity (Hypothesis 1) and whether the inclusion of expert knowledge can augment the synthetic data's quality (Hypothesis 2).

Worth noting is that the t-test's independence and normality assumptions were considered met in our scenario. The bootstrap versions of the test set were independent, thus satisfying the independence requirement. Invoking the Central Limit Theorem justified the normality assumption given the large number of bootstrap versions.

For hypothesis testing, our attention centered on the models that demonstrated superior performance during the hyperparameter search. In this context, we define $P_{ne,l}$ and P_0 as the Poisson Deviance for predictions made by the XGBoost model trained on synthetic data sans expert knowledge, and the original data, respectively. Hypothesis 1, which solely concerns the MC-WGAN-GP trained on the large dataset devoid of expert knowledge, frames the null and alternative hypotheses as follows:

$$H_0 : P_{ne,l} = P_0$$

$$H_a : P_{ne,l} < P_0$$

For Hypothesis 2, we introduce $P_{e,s}$, $P_{e,m}$, and $P_{e,l}$, representing Poisson Deviances from predictions by the model trained on synthetic data, crafted by the GAN trained with expert knowledge for small, medium, and large datasets, respectively. In a similar vein, $P_{ne,s}$, $P_{ne,m}$, and $P_{ne,l}$ symbolize the Poisson Deviances from predictions by the model trained on synthetic data devoid of expert knowledge for the small, medium, and large datasets, respectively. Hence, for Hypothesis 2, the null and alternative hypotheses are as follows:

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

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

Hypothesis 2 is deemed validated only if all conditions delineated in the alternative hypothesis are satisfied for all datasets.

Given the nature of the evaluation, potential limitations encompass the inherent assumptions of the selected hypothesis testing methodology, reliance on the XGBoost model's predictive accuracy, and the use of a singular, albeit intricate, method for synthetic dataset generation. However, these limitations are mitigated by the rigor of the evaluation procedure and the robustness of the chosen metrics, affording a comprehensive appraisal of the model's performance.

This thorough evaluation procedure is designed to provide substantial evidence supporting or disputing the utility of GANs in synthesizing insurance datasets, and the impact of integrating expert knowledge in this process. These findings can serve as a basis for future research to further refine the data synthesis process and extend the applications of GANs in the insurance sector.

3.12 Poisson Deviance

The assessment of our model's predictive efficacy was undertaken through the application of the Poisson Deviance. This statistical measure is frequently used in evaluating the quality of fit in Poisson regression models, especially in contexts involving count data [8]. Poisson Deviance furnishes a comparison between the fit offered by the model under examination and that offered by a 'perfect' model. The latter serves as an idealized benchmark, representing a model that would perfectly predict the observed data.

Within the framework of our investigation, the computation of the Poisson Deviance took the form of the following equation:

$$D = 2 \sum \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right]$$

, where y_i represents the observed data (ground truth) and $\hat{\mu}_i$ stands for the model prediction.

A lower Poisson Deviance value signifies a superior model fit, with a value of zero indicating an impeccable fit. Notably, while the Poisson Deviance is sensitive to the mean of the distribution, it does not heavily penalize errors that exhibit a bias above the mean.

4 RESULTS

The findings of our study provide evidence in support of Hypotheses 1.1 and 2.1. A comparative analysis of the XGBoost models, particularly their performance metrics, revealed nuanced insights, particularly related to the effect of synthetic data generation and the inclusion of expert knowledge.

Hypothesis 1.1 posited that the XGBoost model built on data generated by the MC-WGAN-GP would produce significantly different predictions on the outcome variable, namely the claim count, compared to a dummy model predicting the average claim count for all respective policies in the dataset. As evidenced by our results, this hypothesis is supported for Poisson Deviance [$t(99) = -1.61$, $p < 0.001$]. We found a significant difference between the models' performance, with the XGBoost model trained on GAN data producing a Poisson Deviance of 31.536 ($SD = 0.251$), which was statistically different from the dummy model [$M = 32.909$, $SD = 0.279$].

Hypothesis 1.2, on the other hand, predicted that the XGBoost model built on the synthetic data would have performance metrics statistically indistinguishable from those of a model trained on the original dataset. Contrary to this hypothesis, we found that the Poisson Deviance were indeed statistically different [$M = 31.536$, $SD = 0.251$ for the model trained on GAN data versus $M = 30.377$, $SD = 0.272$ for the model trained on the original dataset, $t(99) = -40$, $p < 0.001$].

When it comes to the use of expert knowledge in the GAN training phase, **Hypothesis 2.1** expected the XGBoost model built on synthetic data generated with expert knowledge to provide more accurate predictions than the model built on data generated without expert knowledge. Our results indeed indicate a significant improvement in the model's performance when it was trained on synthetic data generated with expert knowledge. The Poisson Deviance decreased to 31.457 ($SD = TODO$), better than the value observed for the model trained without expert knowledge, $M = 31.536$, $SD = TODO$, $t(99) = -6.876$, $p < 0.001$.

Lastly, **Hypothesis 2.2**, asserting that the positive effect of expert knowledge would be more prominent on a subsample of the original dataset, was not corroborated by the results, as shown by Poisson Deviance, $t(99) = 30.115$, $p = 1.0$. This suggests that while expert knowledge improves the quality of synthetic data, its impact may not be as pronounced when the dataset size is reduced.

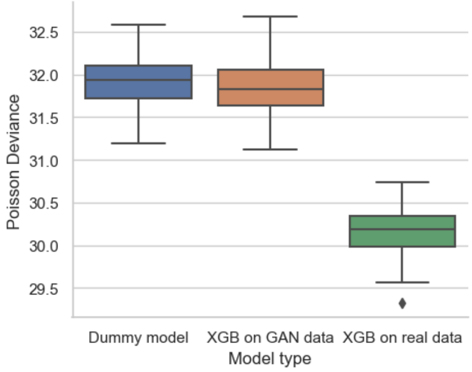


Figure 8: Research Question 2: Comparison of the Poisson Deviance across the XGBoost models trained on different datasets.

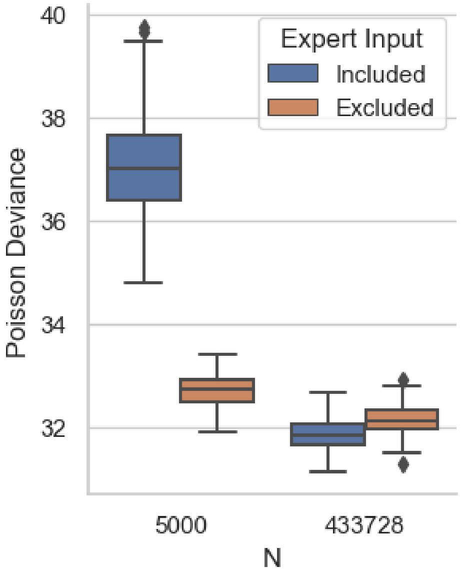


Figure 9: Research Question 2: Comparison of the Poisson Deviance across the XGBoost models trained on different datasets.

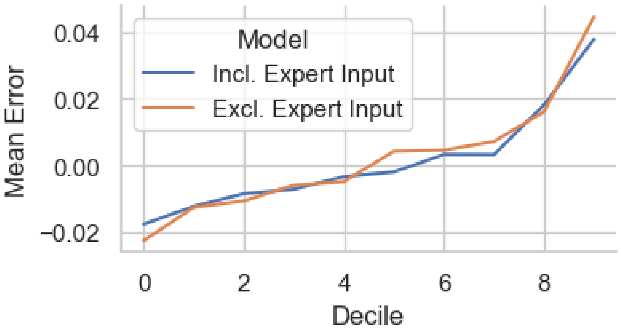


Figure 10: MAE per decile for the XGBoost for GANs trained on large dataset

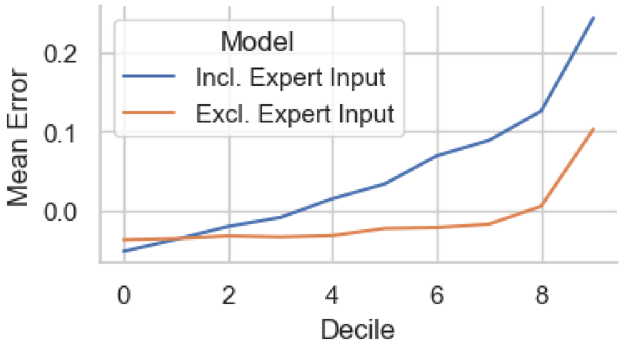


Figure 11: MAE per decile for the XGBoost for GANs trained on small dataset

Figure 9 provides a visual representation of the Poisson Deviance and AUC metrics across the different models. As depicted, there are observable differences in the models' performance depending on the type of data they were trained on, which substantiates our numerical findings.

In order to gain a deeper insight into the inner workings of our models, we performed an exploratory analysis centered around the mean error per decile. The results of this analysis indicated that, for the larger dataset, the model which included expert knowledge was superior at predicting deciles exhibiting extremely low or high claim counts. This was especially noticeable in the three deciles with the lowest claim count and the single decile with the highest claim count. The fact that the XGBoost, constructed based on the GAN model that incorporated expert knowledge, demonstrated enhanced performance for extreme claim predictions, provides support for our hypothesis that the inclusion of expert knowledge can safeguard the model against overfitting (refer to Figure 10 for more details).

Conversely, when considering the smaller dataset, we observed a reversal of the aforementioned effect. The GAN model that did not incorporate expert knowledge, led to the generation of an XGBoost model that tended to slightly underpredict claim count. This could potentially indicate that the GAN may have underfitted, thereby leading to difficulties in producing rare claim occurrences. On the other hand, the XGBoost model derived from the GAN data trained with expert knowledge showed a stark overestimation of claim counts. This could signify a case of overfitting, wherein an excessive number of customers filing claims were generated (refer to Figure 11 for more details).

Confidence intervals for the aforementioned comparisons were calculated, with all the supported findings remaining consistent within the established intervals (see Table 2. The detailed interpretation of these results, along with relevant visual representations, is provided in the following sections. We acknowledge that our results are constrained by the assumptions of the t-test and the reliance on the predictive accuracy of the XGBoost model, among other limitations. However, the results shed light on the potential of GANs in synthesizing insurance datasets and the possible benefits of incorporating expert knowledge.

5 DISCUSSION

The present research makes a substantial contribution to the growing corpus of research focusing on synthetic data generation and data privacy, specifically within the insurance industry's context. By leveraging a novel implementation of Generative Adversarial Networks (GANs) – specifically, the MC-WGAN-GP variant – our study underscores the capabilities of GANs in producing synthetic insurance datasets that uphold the inherent structure and relationships evident in the original data. When the synthetically generated data were utilized to train XGBoost models, they facilitated claim count predictions reflective of the dependent and independent variables' interplay to a significant extent. Though not perfectly capturing all nuances of the original dataset's intricate relationships, the quality of prediction was sufficiently high. This underscores GANs' potential in addressing data privacy concerns while preserving data utility.

Our investigation also brought to light the significant role of expert knowledge incorporation in enhancing the quality of synthetic data during GAN training, as evinced by the augmented predictive accuracy of models trained on such data. This serves to accentuate the integral role of domain-specific knowledge in bolstering the performance of machine learning techniques, including GANs.

The research further illuminated potential drawbacks in integrating expert knowledge into GAN training. Specifically, the study revealed that the beneficial impact of expert knowledge incorporation on model performance was attenuated when the dataset size was reduced. This suggests the existence of a certain dataset size threshold necessary for the optimal manifestation of expert knowledge benefits. We observed an unexpected decline in model performance when the MC-WGAN-GP was trained on smaller datasets supplemented with expert knowledge, emphasizing the importance of adequate dataset size.

Understanding the relationship between the effectiveness of expert knowledge integration and dataset size emerges as a key area for future exploration. Identifying the critical thresholds in dataset size that influence expert knowledge's beneficial effects could be advantageous. Moreover, alternative methods of integrating expert knowledge into the training process warrant exploration. For

Training set size	Data source	Poisson Deviance
433,728	Dummy model	31.909 (31.854, 31.965)
433,728	Synthetic	31.536 (31.480, 31.590)
433,728	**Real data**	**30.377 (30.325, 30.42)**

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

Training set size	Data source	Expert knowledge	Poisson Deviance
433,728	Synthetic	No	31.536 (31.480, 31.590)
433,728	**Synthetic**	**Yes**	**31.457 (31.406, 31.508)**
5,000	**Synthetic**	**No**	**32.714 (32.651, 32.776)**
5,000	Synthetic	Yes	37.214 (36.924, 37.504)

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

instance, the direct inclusion of expert knowledge through the addition of a dedicated layer in the generator architecture represents a promising alternative. However, potential complications associated with the backpropagation phase of the generator's training require careful consideration and navigation.

Despite these challenges, the study offers several strengths, including gaining a comprehensive understanding of the functioning of MC-WGAN-GP under different scenarios of expert knowledge inclusion and dataset sizes. Additionally, the systematic and comprehensive integration of expert knowledge greatly enhanced the modeling process by refining relationships between dependent and independent variables. We also established a streamlined process for capturing actuarial expert knowledge, a strategy potentially applicable to other AI research areas.

The findings have the potential to shape more accurate and reliable data generation processes in the insurance industry, thereby facilitating data sharing among insurers and researchers and creating non-confidential datasets for research purposes. Our study's exploration of neuro-symbolic input in GAN training may also lead to advancements in GAN models across various industries, thereby paving the way for more sophisticated data-driven decision-making processes.

6 CONCLUSION

Our research provides pivotal insights into synthetic data generation using Generative Adversarial Networks (GANs), with an emphasis on data privacy in the context of the insurance industry. We leveraged a novel application of the MC-WGAN-GP variant of GANs to generate synthetic insurance datasets, with an aim to retain the underlying structure and relationships inherent in the original data.

A key finding from our study highlights the role of expert knowledge in enhancing the quality of synthetic data generation, providing a fresh perspective on integrating domain-specific expertise into machine learning techniques. We also uncovered a complex interplay between the efficacy of expert knowledge integration and the size of the dataset used, shedding light on an area that warrants further research.

The insights derived from this research have implications beyond the realm of insurance, contributing to broader discussions on synthetic data generation, machine learning, and data privacy. Moreover, our findings have the potential to guide the creation of non-confidential datasets and to facilitate more robust data sharing practices among stakeholders. The possibility of using neuro-symbolic input in GAN training opens up new avenues for data-driven decision making across various industries. Future research building on our findings will play a vital role in extending the understanding of GAN applications in the insurance industry, and beyond.

In essence, our research has sown the seeds for a future where - through use of expert knowledge - synthetic data generation is not only feasible but also optimally beneficial, offering invaluable lessons on integrating domain expertise into GAN training for improved results.

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] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

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

[13] Daniel König and Friedrich Loser. 2020. https://www.kaggle.com/code/floser/glm-neural-nets-and-xgboost-for-insurance-pricing

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

[15] 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.

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

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

[18] 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.

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

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

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

[22] 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.

[23] 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.

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

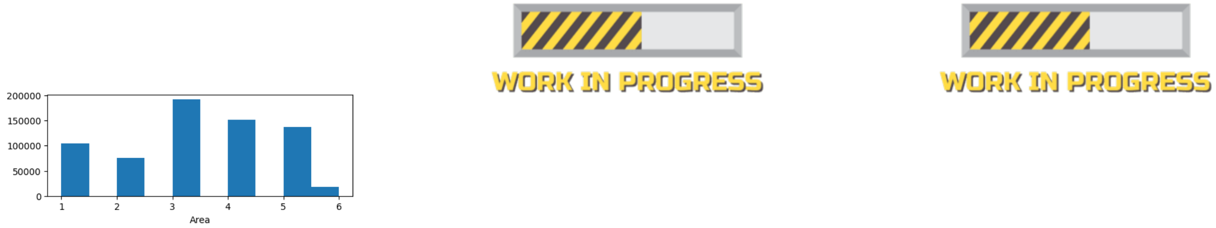
[25] Jürg Schelldorfer and Mario V Wuthrich. 2019. Nesting classical actuarial models into neural networks. *Available at SSRN 3320525* (2019).

[26] SURF. 2021. Dutch National Supercomputer Snellius. https://www.surf.nl/en/dutch-national-supercomputer-snellius

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

[28] 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 WORK IN PROGRESS: COMPARISON OF DISTRIBUTIONS FOR FULL DATASET SCENARIO

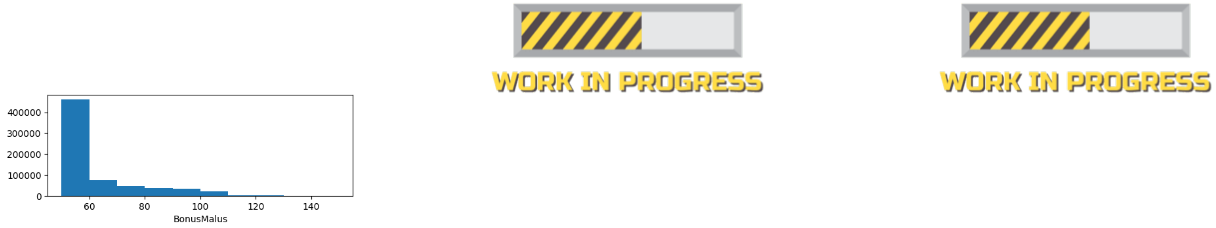


(a) Real data

(b) Synthetic with expert input

(c) Synthetic without expert input

Figure 12: Distribution of area variable for the different datasets (big dataset scenario)

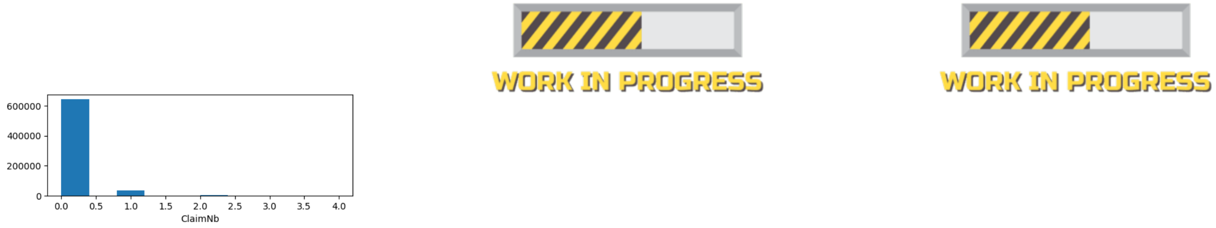


(a) Real data

(b) Synthetic with expert input

(c) Synthetic without expert input

Figure 13: Distribution of bonus malus variable for the different datasets (big dataset scenario)

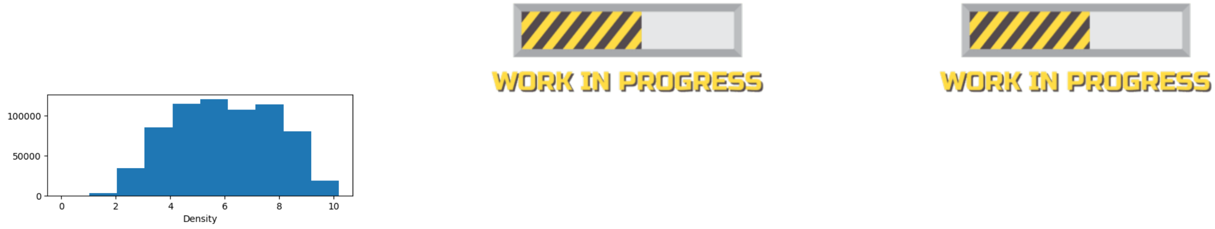


(a) Real data

(b) Synthetic with expert input

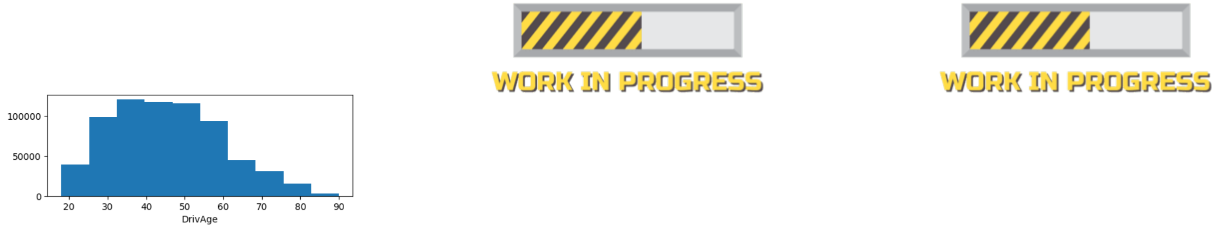
(c) Synthetic without expert input

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



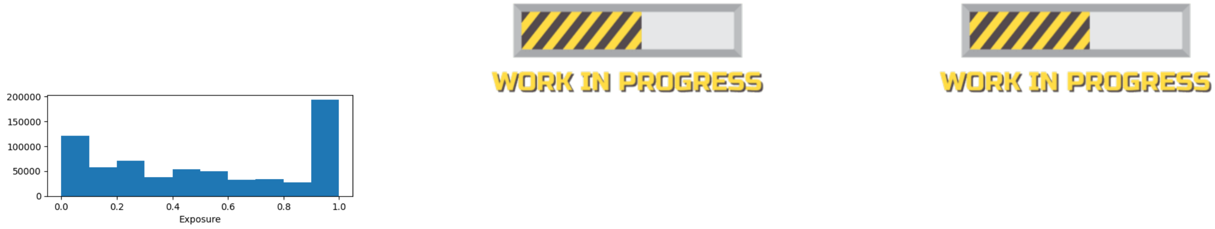
(a) Real data **(b) Synthetic with expert input** **(c) Synthetic without expert input**

Figure 15: Distribution of density 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 driver age variable for the different datasets (big dataset scenario)



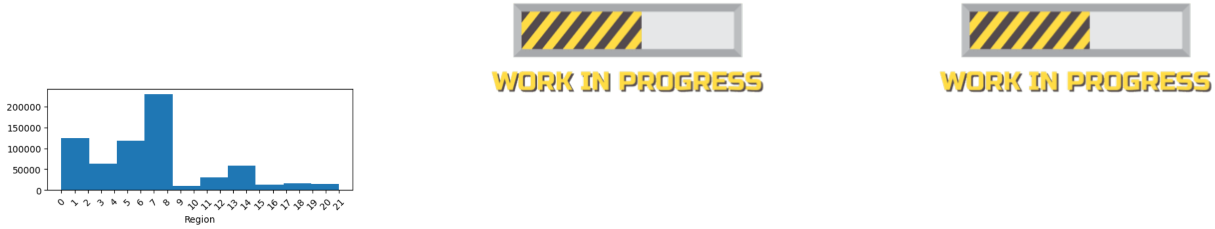
(a) Real data **(b) Synthetic with expert input** **(c) Synthetic without expert input**

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

Appendix B APPENDIX NOT FINISHED YET!! STILL WORK IN PROGRESS

Appendix C SUMMARY OF EXPERT KNOWLEDGE PROVIDED

In multiple sessions, the actuarial expert provided her knowledge on MTPL vehicle insurance to us. The summary of the sessions is shared in this Appendix. The summary is ordered into the different steps of expert knowledge inclusion defined in the *Incorporating Expert Knowledge* section in the *Method* section: Idea Generation, Representation Selection, Representation Adjustment, Additional Rules.

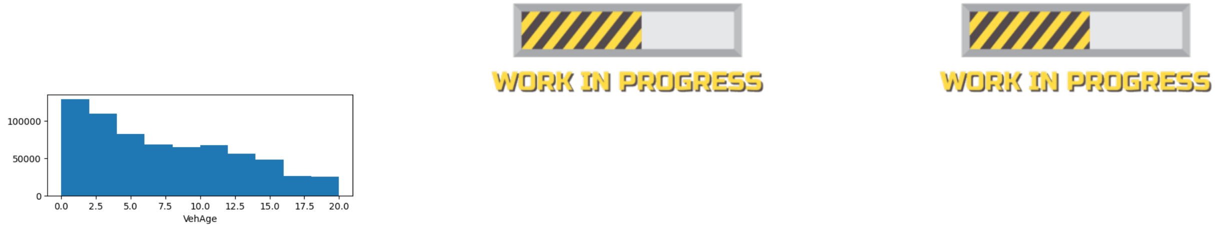


(a) Real data

(b) Synthetic with expert input

(c) Synthetic without expert input

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

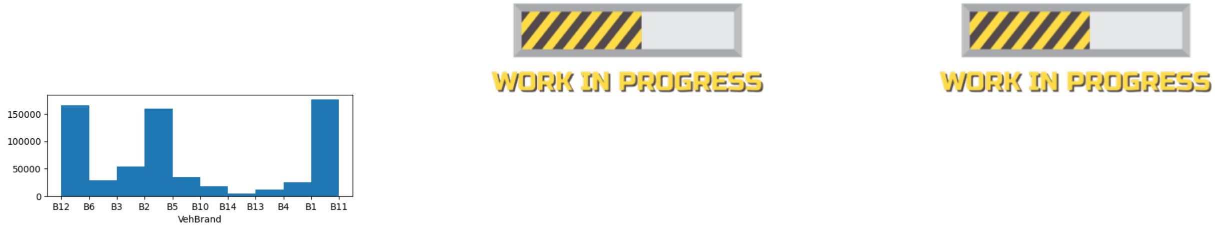


(a) Real data

(b) Synthetic with expert input

(c) Synthetic without expert input

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



(a) Real data

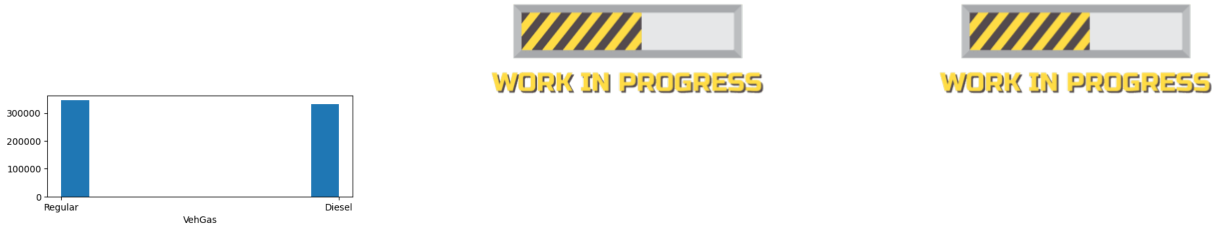
(b) Synthetic with expert input

(c) Synthetic without expert input

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

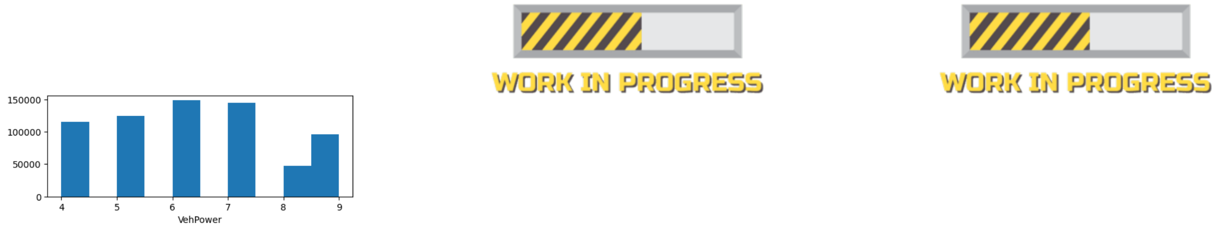
C.1 Scope Definition

The actuary suggested that she can provide expert knowledge on the following variables: *VehAge, Density, BonusMalus, VehPower, DrivAge.*



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

Figure 21: 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 22: Distribution of vehicle power variable for the different datasets (big dataset scenario)

C.2 *VehAge*

- Idea Generation: The actuary proposed that the relationship between *VehAge* and *ClaimNb* can be modeled using a polynomial Generalized Linear Model (GLM). The degree of the polynomial (that is, the highest power of *VehAge* in the model) should not be greater than 4. The four different models were trained on the training set and shown to the expert.
- Representation Selection: The actuary chose that the model best representing the relationship between *VehAge* and *ClaimNb* is $ClaimNb = \beta_0 + \beta_1 \times VehAge + \beta_2 \times VehAge^2 + \beta_3 \times VehAge^3$.
- Representation Adjustment: The actuary decided that for vehicles below 5 years of age, the expected *ClaimNb* should be set at $\widehat{ClaimNb} = 0.05$.
- Additional Rules: No additional rules have been suggested.

See Figure 23 for the final representation.

C.3 *Density*

- Idea Generation and Representation Selection: The actuary proposed that the relationship between *Density* and *ClaimNb* is linear. Therefore, it can be modeled using a monomial Generalized Linear Model (GLM). The model best representing the relationship is $ClaimNb = \beta_0 + \beta_1 \times Density$.
- Representation Adjustment: The were no adjustments to the model after it was built
- Additional Rules: No additional rules have been suggested.

See Figure 24 for the final representation.

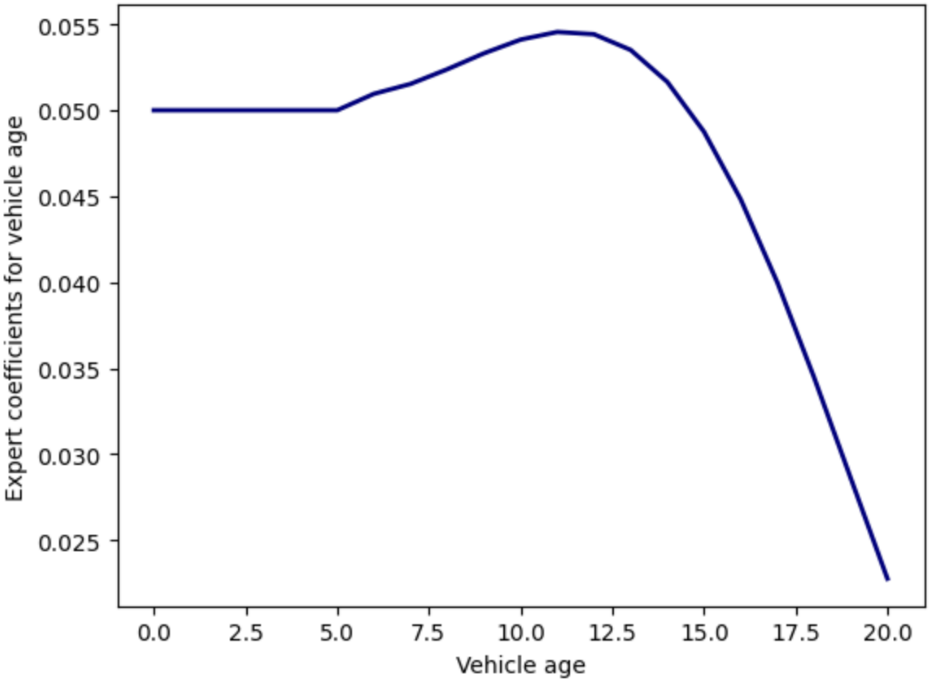


Figure 23: Chosen representation for the relationship between vehicle age and claim count

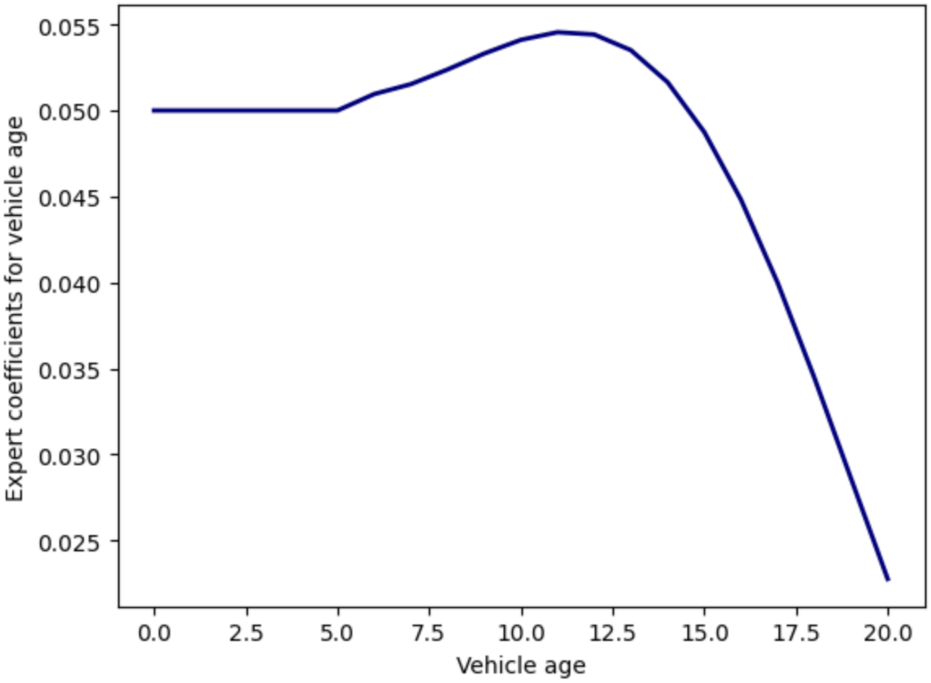


Figure 24: Chosen representation for the relationship between vehicle age and claim count

C.4 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 D GRIDSEARCH

To find the best performing model, a hyperparameter gridsearch was conducted. Due to infrastructural limits (usage of high-capacity Snellius graphical processing units was limited to 50000 system billing units), our gridsearch was limited to only the combinations of parameters that produced the models which best performed in previous research [4]. For the gridsearch hyperparameters searched, see https://github.com/AfairiJJ/thesis/blob/main/config/ganruns.csv

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), it is not included in the hyperparameter search.