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 July 4, 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, MTPL, Synthetic Data, Generative Adversarial Networks, Actuarial Science, MC-WGAN-GP

GITHUB REPOSITORY

https://github.com/Afairi/Afairi.io-Research-GAN

1 INTRODUCTION

As artificial intelligence (AI) and data science continue to evolve, their implications have been transformative across multiple industries, including insurance. These advances have particularly revolutionized underwriting and pricing, among other areas, thus introducing a myriad of novel applications. However, 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 such as the General Data Protection Regulation (GDPR) complicate data sharing between insurers and limit academic research due to the lack of available realistic data. Although 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. In particular, the Multi-Categorical Wasserstein Generative Adversarial Network with Gradient Penalty (MC-WGAN-GP) demonstrates significant potential to create realistic, multicategorical 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. To our knowledge, the application of such an approach to GAN training in an insurance context is a not well-explored area and this thesis aims to fill this gap [23, 24].

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

This research aims to investigate the potential to improve the training of the GAN model through the inclusion of actuarial expert knowledge. In pursuit of this objective, we propose 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 data set, 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 data set, 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, based on the assumption of a Poisson distribution of data, provides a robust measure of the goodness-of-fit of the model under investigation. Essentially, this metric affords a quantitative evaluation of how well the chosen model aligns with the observed data. A lower Poisson Deviance value indicates a more favorable fit of the model to the data, thereby denoting superior model performance. For an in-depth explanation, see Section 3.9

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

The subsequent sections of this thesis comprise a review of the 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 of 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 on 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 basis 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 nonlife insurance, particularly for large data sets [10, 21]. 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 nonlinear relationships. As a consequence, researchers have explored the potential of deep learning to augment GLM-based actuarial pricing [20, 25, 26, 28].

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 improve the performance of the latter [25, 26, 28].

Taking advantage of these recent developments, we choose to employ an XGBoost model. This decision is informed by empirical evidence that demonstrates the superiority of XGBoost models over GLM and ANN-based models in predicting insurance pricing [20]. Furthermore, this choice aligns with our aim of comparing 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 efficiently utilize real-world data within the insurance industry. One such strategy is the generation of synthetic data, with early studies using traditional statistical methods such as resampling to create synthetic data sets [5, 9]. However, with the advent of ML and AI, more sophisticated techniques have emerged.

Introduced by Goodfellow et al. [11], 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 the CTGAN-generated data successfully mimic the distributions of the real data. Building on this work, Cote 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 utilizes the MC-WGAN-GP model to generate synthetic data, as it has been shown to be 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, increase model interpretability, and increase model performance [17, 19, 22–24]. This is particularly essential in sectors such as insurance, where clarity of the rationale of the model and the justification of decisions play a pivotal role [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 [23, 24]. As delineated by Prentzas and Hatzilygeroudis [22], these methods fall primarily into two groups: 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 [17, 19].

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. On the contrary, embedded processing involves embedding a component based on a particular representation into one or more components predicated on another [22, 23].

In 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 the potential of insurability in medical insurance, based on expert rules. Hsieh and Wang [12] further extended this research by introducing the Linguistic Descriptions Evaluating Algorithm, a life insurance risk assessment tool based on a multitude of linguistic approaches (e.g., linguistic logic, uncertainty number modeling, fuzzifications, and defuzzification schemes).

[2]where N is the data set size

Our knowledge representation method in our AI model is informed by hybrid methodologies that sequentially combine different knowledge representation techniques. We aim to improve the interpretability of our model by weaving expert knowledge into the learning process [17, 19, 22–24]. Additionally, we developed a system to assess customer risk based on expert rules.

3 METHODOLOGY

3.1 Research Design

This research involved designing and training various synthetic data generation pipelines by combining GANs with the knowledge of actuarial experts and comparing the performance of the different GAN models for various sizes of the training set ($N = 5,000$; $N = 433,728$) using an XGBoost trained on the generated data.

We design five modeling pipelines and compare them in two different scenarios of different training set sizes (see Figure 1). In the first pipeline, a dummy model was established which predicted the average claim count for each policy in the risk data set. 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 data set to generate synthetic data, which was subsequently used to train the XGBoost model. The third pipeline resembled the second, but incorporates additional knowledge of actuarial experts in the GAN training process to improve synthetic data generation. Throughout these pipelines, the data pre-processing steps, the hyperparameters of the XGBoost model, and the test set remain constant, allowing for unbiased evaluation.

Since our dependent variable (claim count) is Poisson distributed, model performance evaluation is done using Poisson Deviance [8]. We do not use other measures often used in GAN research, such as the Root Mean Squared Error (RMSE) and the 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 [27]. However, for accessibility and replicability, the methodology was also designed to run on a more modest machine, such as the MacBook Air M2 with an 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 data set for this study was derived from the French motor third-party liability (MTPL) insurance portfolio, available on the OpenML platform [7]. This data set comprises 678,013 vehicle insurance policies and twelve distinct variables. These include the policy number (*IDpol*), which is typically associated with a customer, vehicle, or a combination of both, as well as the claim count (*ClaimNb*),

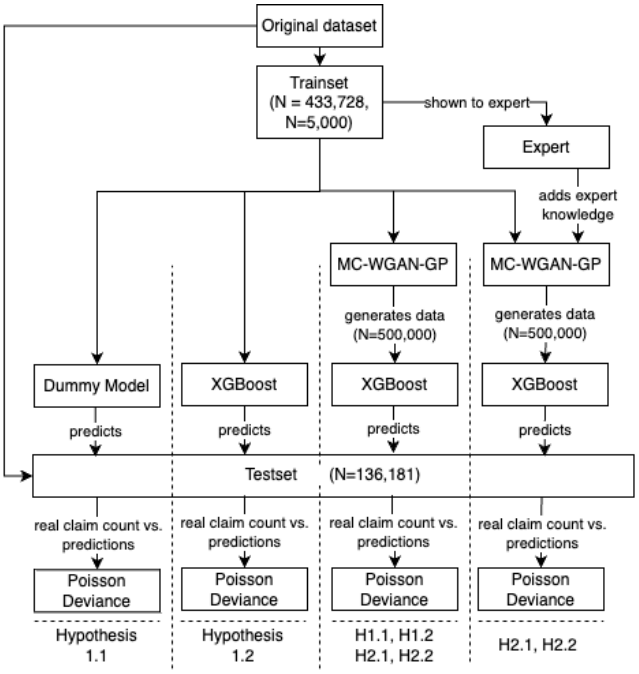

Figure 1: Structure of the different pipelines for each hypothesis

measuring the number of claims made by a customer within a specified exposure time frame (also named *Frequency* in the actuarial context). Other variables include total exposure in years (*Exposure*), area code (*Area*), power of the vehicle (*VehPower*), age of the vehicle in years (*VehAge*), age of the driver in years (*DrivAge*), bonus-malus level (*BonusMalus*), vehicle brand (*VehBrand*), fuel type (*VehGas*), population density (*Density*), and regions in France (*Region*).

In data pre-processing, particular care was taken to ensure consistency between different variables. Any identified inconsistencies were dealt with using appropriate strategies, ensuring the quality of the data.

In particular, the variables *Exposure* and *ClaimNb* deviated from the expected distribution assumptions. First, 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 data sets, 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 the necessary characteristics of the policy with each other except for *Exposure*, policy ID (*IDpol*), and claim count (*ClaimNb*). It was hypothesized that these non-unique policies represented vehicle fleets, such as leasing vehicles, rental vehicles, or company-owned vehicles. Consequently, these policies were retained in the data set but grouped when the data set was split 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).

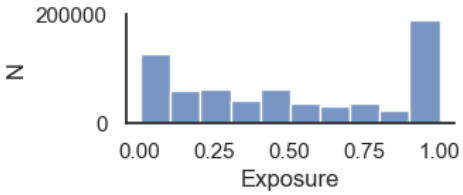

Figure 2: Distribution of exposure values up to one year

The dependent variable *ClaimNb* was further examined. The data set 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 Appendix A for a distribution of *ClaimNb* values below 5).

3.3.1 Data Preprocessing. A standardized data preprocessing pipeline was implemented, based on the recommendations of previous research [4, 20]. This included transformations and standardizations of several variables to ensure data quality and comparability between different modeling pipelines.

The common data preparation steps included the following transformations:

- **Policy ID (*IDpol*):** *IDpol* was dropped from the data set because it did not contribute to the modeling process.
- **Claim Count (*ClaimNb*):** The claim count was capped at values exceeding 4 claims; to facilitate GAN training, it was also converted into a categorical variable for training.
- ***Exposure*:** For XGBoost training, the *Exposure* was not used. However, to support GAN training, exposure was used with values exceeding one year capped.
- ***Area*:** The categorical alphabetic representation of the *Area* variable was transformed into a continuous variable.
- **Vehicle Age (*VehAge*):** To avoid excessive skewness in the data, the vehicle's age was capped at 20 years.
- **Vehicle Power (*VehPower*):** To mitigate potential outliers, *VehPower* was capped at values exceeding 9.
- **Driver Age (*DrivAge*):** To limit the influence of extreme values, the age of the driver (*DrivAge*) was capped at 90 years.
- ***BonusMalus*:** As per prior recommendations, *BonusMalus* levels exceeding 150 were capped.
- ***Density*:** To alleviate the impact of skewed distributions, a logarithmic transformation was applied to the *Density*.

Following the data pre-processing steps, the numerical features were standardized, ensuring a consistent scale across all models. Categorical variables were encoded using one-hot vectors, 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 a ratio of 60:20:20. To ensure reproducibility and consistency, a fixed seed was used for any random operations conducted during data splitting (seed=1).

3.4 Incorporating Expert Knowledge

The proposed rule integration technique in this work encapsulates elements of 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 resulting synthesis was incorporated into the data set for the GAN training process. Lastly, the domain expert was given the opportunity to add additional group-based rules directly to the data.

In risk models, overfitting, especially in the presence of imbalanced ratios between positive and negative outcomes, leads to biases and estimation errors. We postulate that integrating expert knowledge into the GAN model would help reduce overfitting, thus improving the quality of synthetic data, specifically in accurately representing the relationship between dependent and independent variables [29]. Additionally, expert knowledge integration was expected to provide additional information on the relationship between claim count and independent variables, which may not be directly discernible from the original data set. An ancillary insight from this study relates to the capability of GANs to extract information from auxiliary data added to a data set.

Besides the actuarial expert knowledge, attempts were made to incorporate vehicle 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 data set at hand and incorrect alignment between the German GDV values and the data set's values based on French insurance standards. Therefore, this data was not further used in the analysis.

The process of incorporating the knowledge of actuaries in our research framework followed a series of steps to ensure effective integration. Expert knowledge was integrated into the data provided for GAN training, combining both embedded processing and co-processing approaches (see Appendix C for a description of the introduced expert knowledge, including all rules and adjustments).

3.4.1 Scope Definition. The first phase involved identifying variables for which the actuarial expert could provide information about 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.

3.4.2 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 the claim count. The expert actuary guided the selection of models to be trained. For example, 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 3. A range of GLMs were trained to explore potential relationships between variables and claim count during the idea generation phase. Through the analysis and visualization of these

models, the expert was able to identify the most suitable representation that resonated with her domain knowledge and expectations.

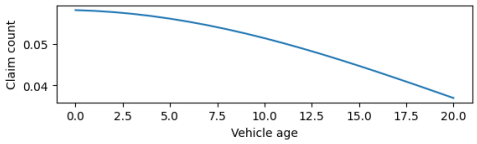

Figure 3: One of multiple potential GLMs to capture the relationship between vehicle age and claim count

3.4.3 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 the claim count. This selection process resulted in the identification of the most appropriate GLM representation (see Figure 4 for the example of vehicle age).

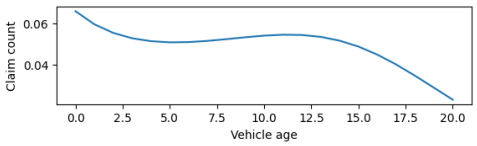

Figure 4: The final GLM chosen to capture the relationship between vehicle age and claim count

3.4.4 Representation Adjustment. During this phase, the actuarial expert implemented modifications to the representations of the established relationships for the selected variables. Consider the case of the *VehAge* variable. Drawing from her experience, the actuary posited that the likelihood of filing a claim does not increase for relatively new vehicles (those under 5 years of age). Consequently, she suggested maintaining a constant average claim count of 0.05 for vehicles within this age bracket. This expert insight was codified as a rule that superseded the model's original predictions for this particular subgroup. Such modifications provided the opportunity for the expert to fine-tune the selected models to better reflect her professional insights. By formulating rules and adjustments for certain variable ranges or categories, she was able to incorporate nuanced changes that catered to specific patterns or tendencies discerned in the data.

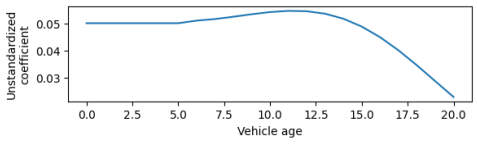

Figure 5: The final relationship representation with Vehicle age <= 5.00 overwritten by actuarial decision to keep the coefficient at 0.05

3.4.5 Additional Rules. The actuary could propose additional rules, which were later incorporated into the data set. This provision enabled us to capture further insights that may have been overlooked or inadequately represented by strictly data-driven approaches. For example, the expert recommended the introduction of a variable designed to differentiate policies of customers who possessed a *BonusMalus* below 100 from those with a *BonusMalus* exceeding 100. This distinction provided a refined perspective on the potential impact of *BonusMalus* on claim count (see Appendix C for all additional rules).

3.5 Integration of Expert Knowledge in the GAN

To incorporate the expert knowledge into the GAN training, all the gathered expert knowledge was standardized and systematically assigned to the corresponding customers in our data set. 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 data set - at least one for each variable.

The enriched training data set 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 through 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 data set, but also incorporated the additional insights provided by the expert knowledge. Thus, the synthetic data produced by the generator were 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 focused on enhancing the performance of the Multicategorical Wasserstein Generative Adversarial Network with Gradient Penalty (MC-WGAN-GP). Architectural parameters and training guidelines were partially derived from the methodologies proposed by Camino et al. [3], Cote et al. [4].

GANs comprise a pair of adversarial neural networks: the generator, which produces synthetic data intended to emulate the real data, and the critic, which determines the authenticity of the data points. Mutual competition between these two networks facilitates

the improvement of the data quality of the synthetic data, achieving greater similarity to the actual data [4].

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

3.6.2 Multicategorical Wasserstein Generative Adversarial Network with Gradient Penalty. In the context of this research, we used 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 multicategorical variables suggested by Camino et al. [3]. In our design, every categorical variable is processed through a dense layer, followed by a softmax activation layer in the generator. The outputs of these layers are subsequently concatenated to yield the final output of the generator [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 produce 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 synthetic data generated can then be utilized to train predictive models, such as XGBoost, to predict claim counts.

It is 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 data set 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].

3.6.3 Hyperparameter Optimization. In the effort to optimize the 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 GitHub repository. The validation of these hyperparameters was performed 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 critic and generator, critic's leaky ReLU's parameter, noise size, the ratio of critic updates per generator update, L2 regularization for the critic, and sizes of the critic's hidden layers. In particular, 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.

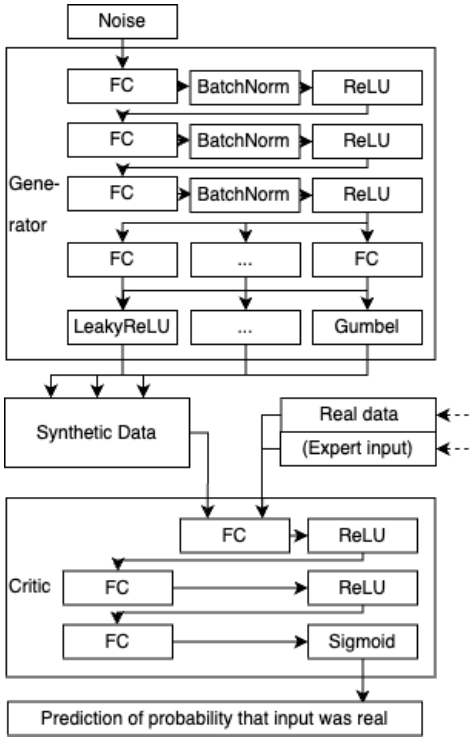

Figure 6: The architecture of the final version of the MC-WGAN-GP; in brackets, elements which are only used in certain scenarios

Experiments included changes to the model and training protocol, such as replacing the last generator layer 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 data set by undersampling policies without claims for the initial 500 epochs of the GAN training process. Unfortunately, these changes did not enhance the quality of the model's data generation.

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 the critic shared the same zero-sum objective function, operating under a 3:1 training-validation split [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 final architecture of the critic comprised two fully connected layers, enhanced with a ReLU function and terminated by a sigmoid activation function (refer to Figure 6 for details).

Hyperparameter tuning was performed across two different data set sizes, large and small, to assess the scalability and efficiency of our model. More details on these data sets are available in the 3.3 section.

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

Appendix D contains the optimal parameters identified for each of the hypotheses.

3.7 XGBoost Model

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

In line with common practice in the insurance sector, we assumed that the claim count follows a Poisson distribution [20]. 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 data sets.

Instead of undertaking a comprehensive grid search for the XGBoost hyperparameters, we used the findings of previous research that had already identified optimal parameters to train the XGBoost model on our data set. Therefore, we conducted only a limited grid search, comparing the performance of the hyperparameters suggested by Martínez de Lizarduy Kostornichenko [18] with those by König and Loser [13] on the large training data set. Appendix E contains the optimal parameters identified.

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

3.8 Evaluation

3.9 Poisson Deviance

The assessment of the predictive efficacy of our model was carried out by applying the Poisson Deviance. This statistical measure is frequently used to evaluate 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 calculation 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 $\hat{\mu}_i$ represents the prediction of the model.

A lower Poisson Deviance value indicates a more favorable fit of the model to the data, thereby denoting superior model performance; a Poisson Deviance value of 0 indicates 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.

3.9.1 Evaluation Design. The proposed methodology's efficacy was appraised via comparative performance analysis of identical XGBoost models, each trained on one of three distinct data sets - the original data set, a synthetic data set generated by the MC-WGAN-GP that incorporates expert knowledge, and another synthetic data set produced by the MC-WGAN-GP without such expertise. For the original data set, we created 100 unique bootstrap versions for both model training and evaluation. In the case of synthetic data, the GAN was used to generate 100 separate iterations of 100,000 synthetic policies, each of which was 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 one-sided, two-sample t-test was adopted to verify both hypotheses. Poisson deviance was computed for each of the 100 bootstrap versions of the test set for each pair of variables 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 on the study's principal goal, namely, to determine if the GAN can produce synthetic data that do not underperform the original data set in terms of predictive capacity (Hypothesis 1) and whether the inclusion of expert knowledge can increase the quality of the synthetic data (hypothesis 2).

Notably, the independence and normality assumptions of the t-test were considered met in our scenario. The bootstrap versions of the test set were independent, satisfying the independence requirement. Invoking the Central Limit Theorem justified the normality assumption given the large number of bootstrap versions [16].

For hypothesis testing, our attention was focused 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 without expert knowledge and the original data, respectively. Hypothesis 1, which solely concerns the MC-WGAN-GP trained on the large data set without 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, designed by the GAN trained with expert knowledge for small, medium, and large data sets, respectively. In a similar vein, $P_{ne,s}$, $P_{ne,m}$, and $P_{ne,l}$ symbolize the Poisson Deviances from the predictions of the model trained on synthetic data without expert knowledge for the small, medium, and large data sets, 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 considered validated only if all the conditions delineated in the alternative hypothesis are satisfied for all data sets.

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 data set 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 performance.

This thorough evaluation procedure is designed to provide substantial evidence supporting or disputing the utility of GANs in synthesizing insurance data sets 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.

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 the 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 data set. 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 performance of the models, 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 synthetic data would have performance metrics statistically indistinguishable from those of a model trained on the original data set. Contrary to this hypothesis, we found that the Poisson Deviances were statistically different with a Poisson Deviance of 31.536 ($SD = 0.251$) for the model trained on GAN data versus a Poisson Deviance of 30.377 ($SD = 0.272$) for the model trained on the original data set, $t(99) = -40$, $p < 0.001$. See Figure 7 for a comparison of the model performances.

When it comes to the use of expert knowledge in the GAN training phase, **Hypothesis 2.1** expected that the XGBoost model built on synthetic data generated with expert knowledge would 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, $M = 31.536$, $SD = TODO$, $t(99) = -6.876$, $p < 0.001$.

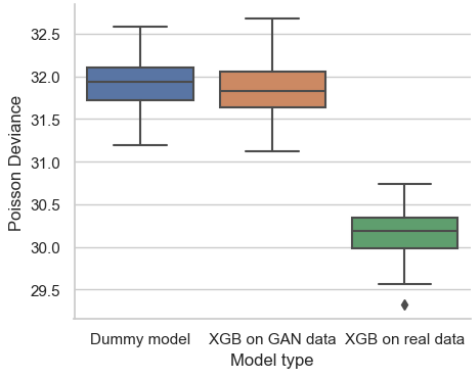

Figure 7: Research Question 2: Comparison of the Poisson Deviance across the XGBoost models trained on different data sets.

Lastly, **Hypothesis 2.2**, asserting that the positive effect of expert knowledge would be more prominent on a subsample of the original data set, 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 size of the data set is reduced.

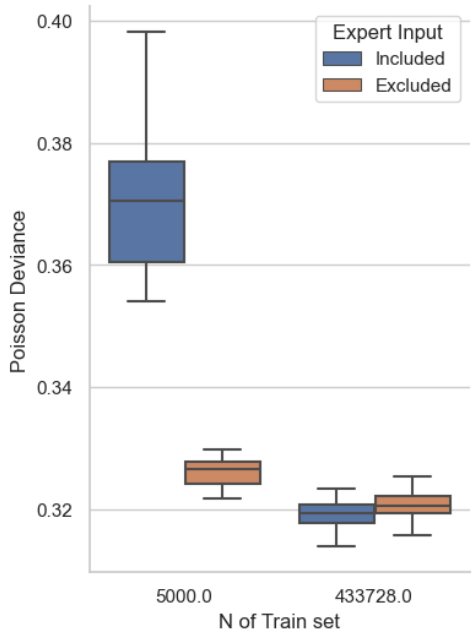

Figure 8: Research Question 2: Comparison of the Poisson Deviance across the XGBoost models trained on different data sets.

Figure 8 provides a visual representation of Poisson Deviances in the different models. As depicted, there are observable differences

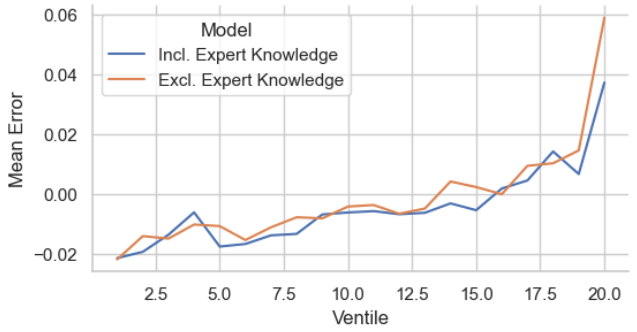

Figure 9: MAE per ventile for the XGBoost for GANs trained on large data set

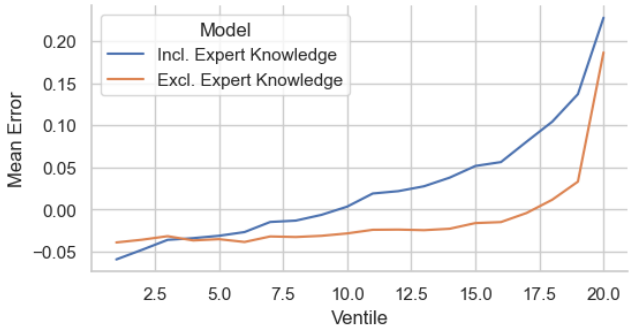

Figure 10: MAE per ventile for the XGBoost for GANs trained on small data set

in the models' performance depending on the type of data on which they were trained, which substantiates our numerical findings.

To delve deeper into the dynamics of our models, we conducted an in-depth analysis centered around the variable distributions within a sample generated by the GAN models. Notably, the GAN model that was trained on data devoid of expert input appeared to falter when it came to generating policies associated with 3 or 4 claims. In stark contrast, the GAN model that incorporated expert knowledge demonstrated the capability to successfully generate such policies (Refer to Appendix A for the detailed distribution of *ClaimNb*).

Further, we performed an analysis that focused on the mean error per ventile. We partitioned the data into ten distinct ventiles based on the predictions rendered by each model, with the highest and lowest claim count predictions falling into the 10th and 1st ventiles, respectively. Subsequently, we computed the residuals between the model predictions and the actual claim counts. These residuals were then collated by prediction ventile group, yielding average residuals per ventile for each model.

The analysis revealed noteworthy findings. With the larger dataset, the model which integrated expert knowledge outperformed its counterpart, particularly in its predictions for the ventiles with exceptionally high projected claim counts. Notably, this model did not overestimate the claim count for these groups to the same extent as the model devoid of expert input (see ventiles 19 and 20 in Figure 9). This observation corroborates our underlying hypothesis suggesting that the incorporation of expert knowledge can mitigate the risk of model overfitting.

In contrast, an examination of the smaller dataset yielded a starkly different pattern. The GAN model bereft of expert input resulted in an XGBoost model that exhibits a consistent tendency toward underprediction of claim counts. This behavior suggests potential underfitting within the GAN, which in turn might have impeded the generation of rare claim events. On the contrary, the XGBoost model constructed from GAN data incorporating expert input manifested a pronounced tendency to overestimate claim counts. This overestimation was particularly evident in ventiles associated with higher predicted claim counts (refer to ventile 10 to 20 in Figure 10), which might suggest a scenario of overfitting in which the model generated an inordinately large number of customers with filed claims.

The confidence intervals for the aforementioned comparisons were calculated, all supported findings remaining consistent within the established intervals (see Table 2). A 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 data sets 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 utilizing MC-WGAN-GP, our study underscores the capabilities of GANs in producing synthetic insurance data sets 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. Although not perfectly capturing all nuances of the original data set's intricate relationships, the quality of prediction was sufficiently high. This underscores the potential of GANs in addressing data privacy concerns while maintaining 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 evidenced 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 the incorporation of expert knowledge on model performance was attenuated when the size of the data set was reduced. This suggests the existence of a certain data set size threshold necessary for the optimal manifestation of

Training set size	Data source	Poisson Deviance
433,728	Dummy model	0.319 (0.319, 0.319)
433,728	Synthetic	31.536 (31.480, 31.590)
433,728	**Real data**	**0.321 (0.321, 0.322)**

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	0.317 (0.316, 0.318)
433,728	**Synthetic**	**Yes**	**0.319 (0.319, 0.320)**
5,000	**Synthetic**	**No**	**0.321 (0.320, 0.321)**
5,000	Synthetic	Yes	0.326 (0.326, 0.327)

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

expert knowledge benefits. We observed an unexpected decline in model performance when the MC-WGAN-GP was trained on smaller data sets supplemented with expert knowledge, emphasizing the importance of adequate data set size.

Understanding the relationship between the effectiveness of expert knowledge integration and data set size emerges as a key area for future exploration. Identifying the critical thresholds in data set 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 instance, 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 inclusion of expert knowledge and data set sizes. Additionally, the systematic and comprehensive integration of expert knowledge greatly enhanced the modeling process by refining the relationships between dependent and independent variables. We also established a streamlined process to capture the knowledge of actuaries, a strategy that could be 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 data sets for research purposes. Our study's exploration of incorporating expert knowledge 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 successfully navigates the intersection of synthetic data generation, data privacy, and the insurance industry, offering notable insights. Using the MC-WGAN-GP, we have demonstrated the potential of synthetic insurance datasets to closely emulate the inherent structure and relationships of the original data, thus underlining the importance of GANs in data privacy endeavors.

Crucially, our findings elevate the importance of expert knowledge in bolstering the quality of synthetic data generation, setting the stage for further exploration of integration of domain-specific expertise in machine learning methodologies. We revealed an intricate interplay between the effectiveness of this integration of expert knowledge and the size of the training dataset, thereby identifying a new direction for future research.

Our research can have effects that traverse the confines of the insurance industry, contributing to the broader discourse on synthetic data generation and data privacy. These findings offer potential avenues for generating non-confidential datasets and promote improved data sharing practices among various stakeholders.

The integration of expert knowledge into GAN training, as demonstrated in this research, heralds a future of enhanced data-driven decision making across various industries. Future research, based on our findings, should contribute to expanding the boundaries of understanding GAN applications, not just within the insurance industry but also across broader spheres.

In summary, this research lays the foundation for a future where synthetic data generation, powered by expert knowledge, is not only feasible but delivers optimal results. This study paves the way for the systematic and beneficial integration of domain expertise into GAN training, promising significant improvements in 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. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council.* European Union. 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] Alexandre Liapounoff. 1900. Sur une proposition de la théorie des probabilités. *Izvestija Rossijskoj akademii nauk. Serija matematičeskaja* 13, 4 (1900), 359–386.

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

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

[19] 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.

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

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

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

[23] 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.

[24] 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.

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

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

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

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

[29] 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 FULL DATASET SCENARIO

In order to understand the impact of expert knowledge on synthetic data generation, we carried out a comparative analysis of the variable distributions across three specific datasets in the full dataset scenario, which includes training, validation, and test set ($N = 678013$). These datasets were the original data, the synthetic data created by the GAN with expert input, and the synthetic data produced by the GAN without expert input. The figures and tables in this Appendix depict the distribution profiles of various variables across these datasets. To ensure visual comparability, we calibrated the size of the synthetic data samples generated by the GANs to correspond with the size of the real dataset for this specific examination. Additionally, any synthetic data points that deviated beyond the boundaries of the real dataset (such as instances where *DrivAge* < *18*) were modified to align within the range defined by the real dataset.

A.1 *ClaimNb*

Data	$N(ClaimNb = 0)$	$N(ClaimNb = 1)$	$N(ClaimNb = 2)$	$N(ClaimNb = 3)$	$N(ClaimNb = 4)$
Real	643953	32178	1784	82	16
Generated: GAN with expert input	639395	36286	2182	149	1
Generated: GAN without expert input	639708	37090	1215	0	0

Table 3: Comparison of distributions of *ClaimNb*

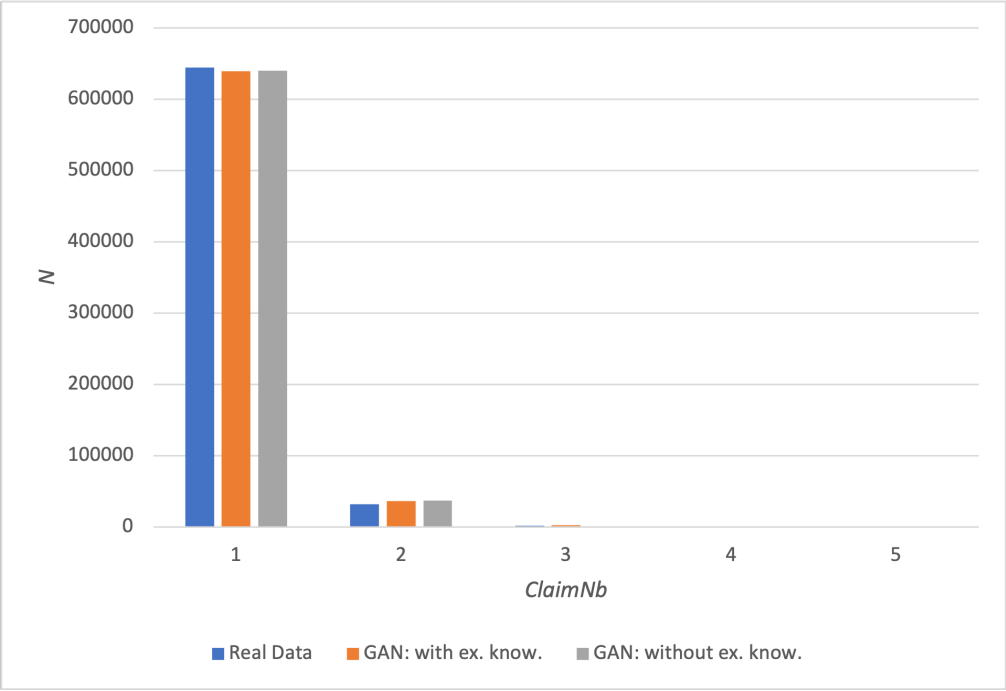

Figure 11: Comparison of distributions of *ClaimNb*

A.2 *Other variables*

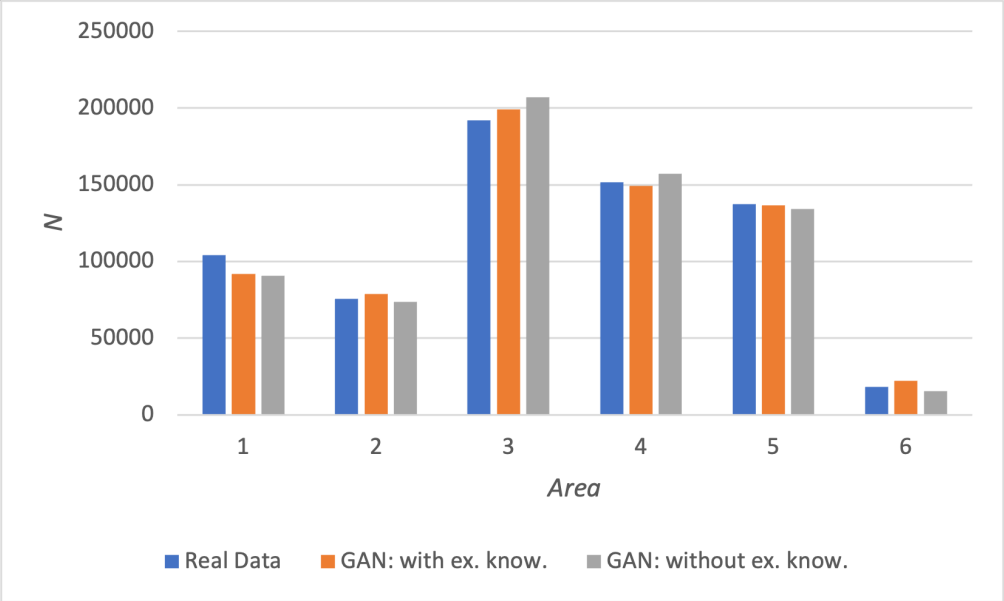

Figure 12: Comparison of distributions of *Area*

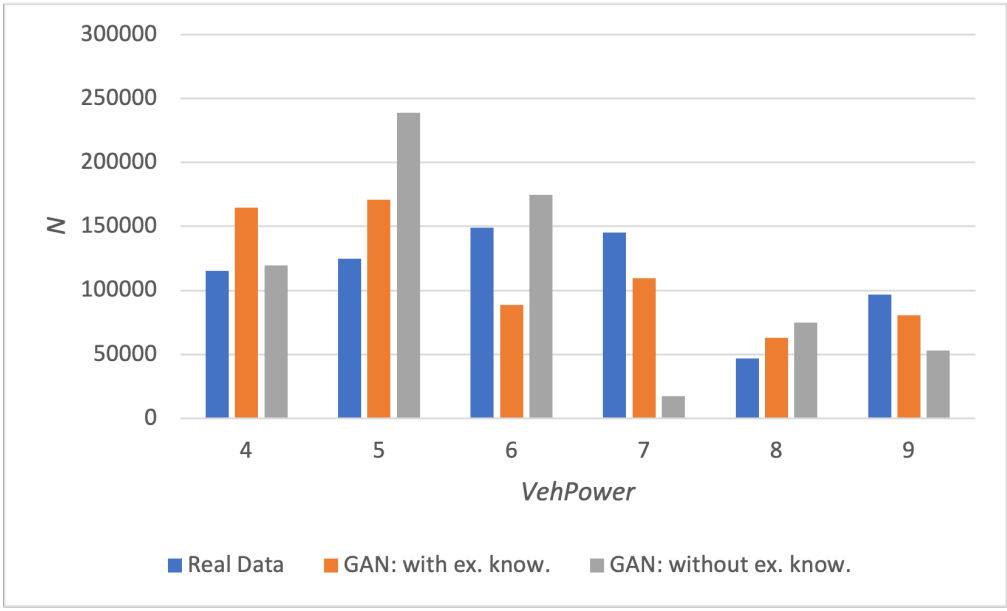

Figure 13: Comparison of distributions of *VehPower*

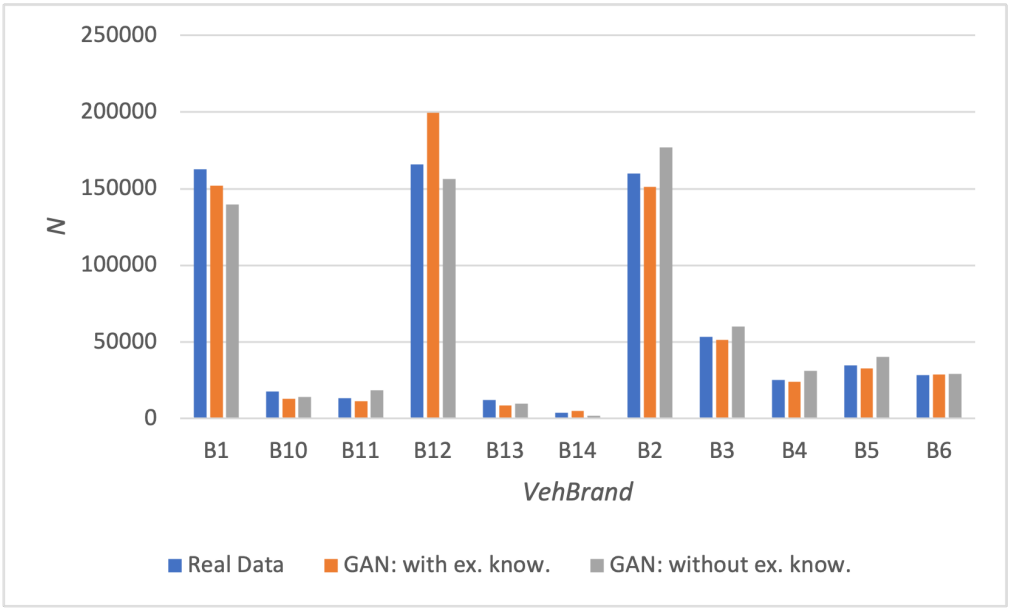

Figure 14: Comparison of distributions of *VehBrand*

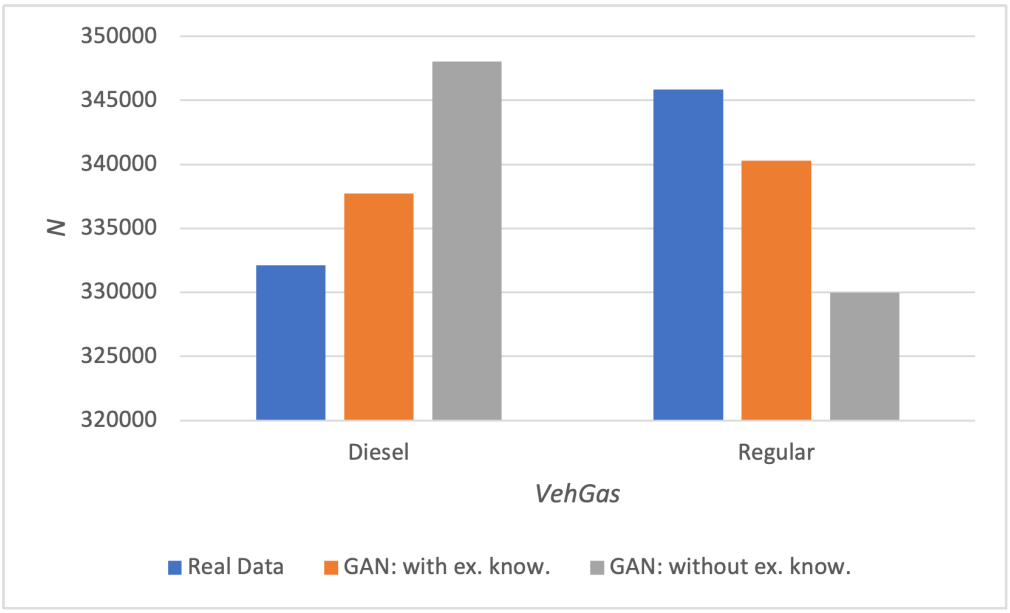

Figure 15: Comparison of distributions of *VehGas*

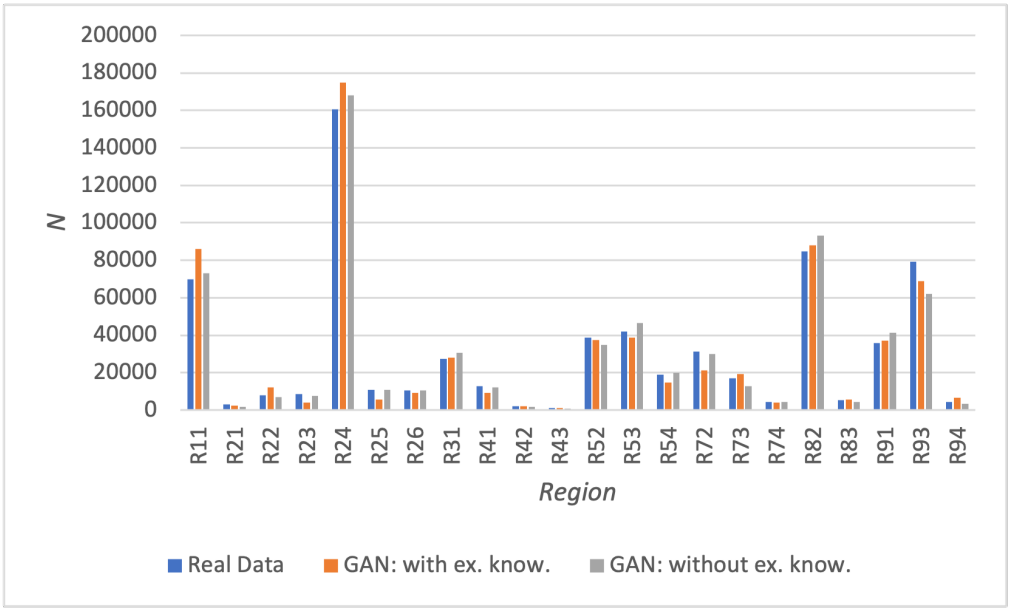

Figure 16: Comparison of distributions for *Region*

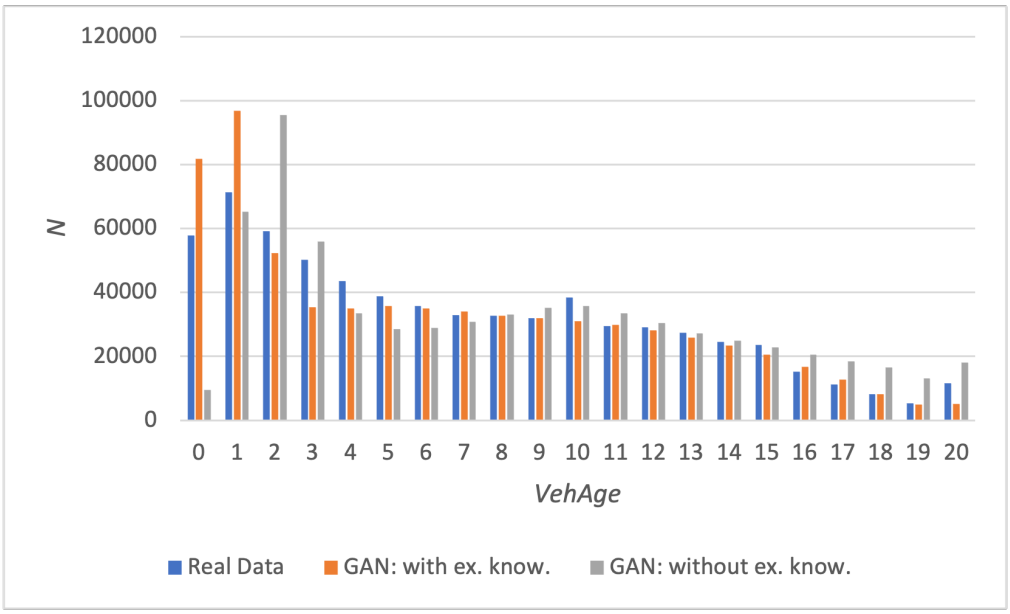

Figure 17: Comparison of distributions for *VehAge*

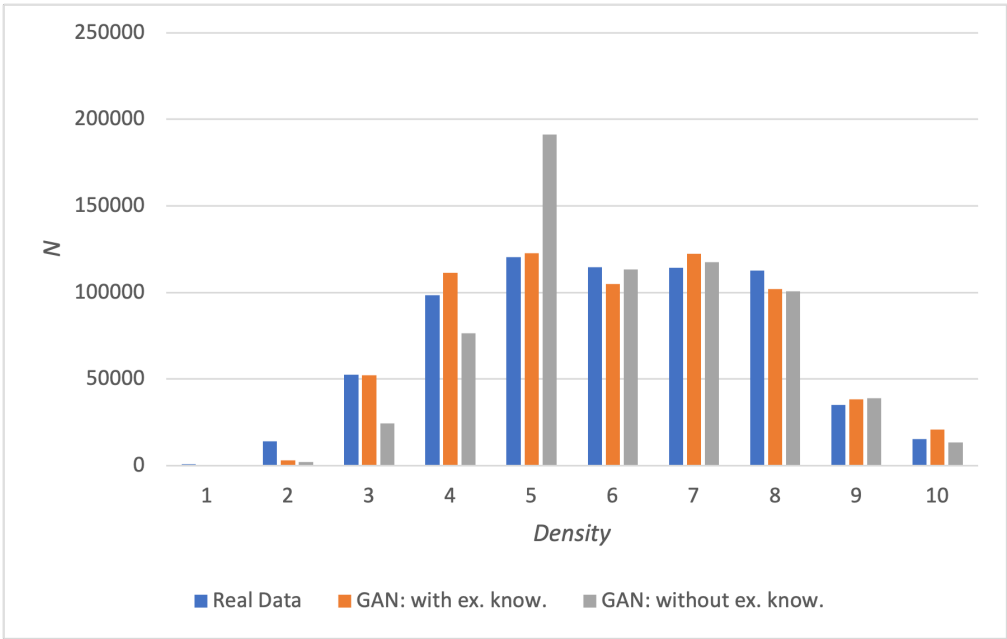

Figure 18: Comparison of distributions for *Density*

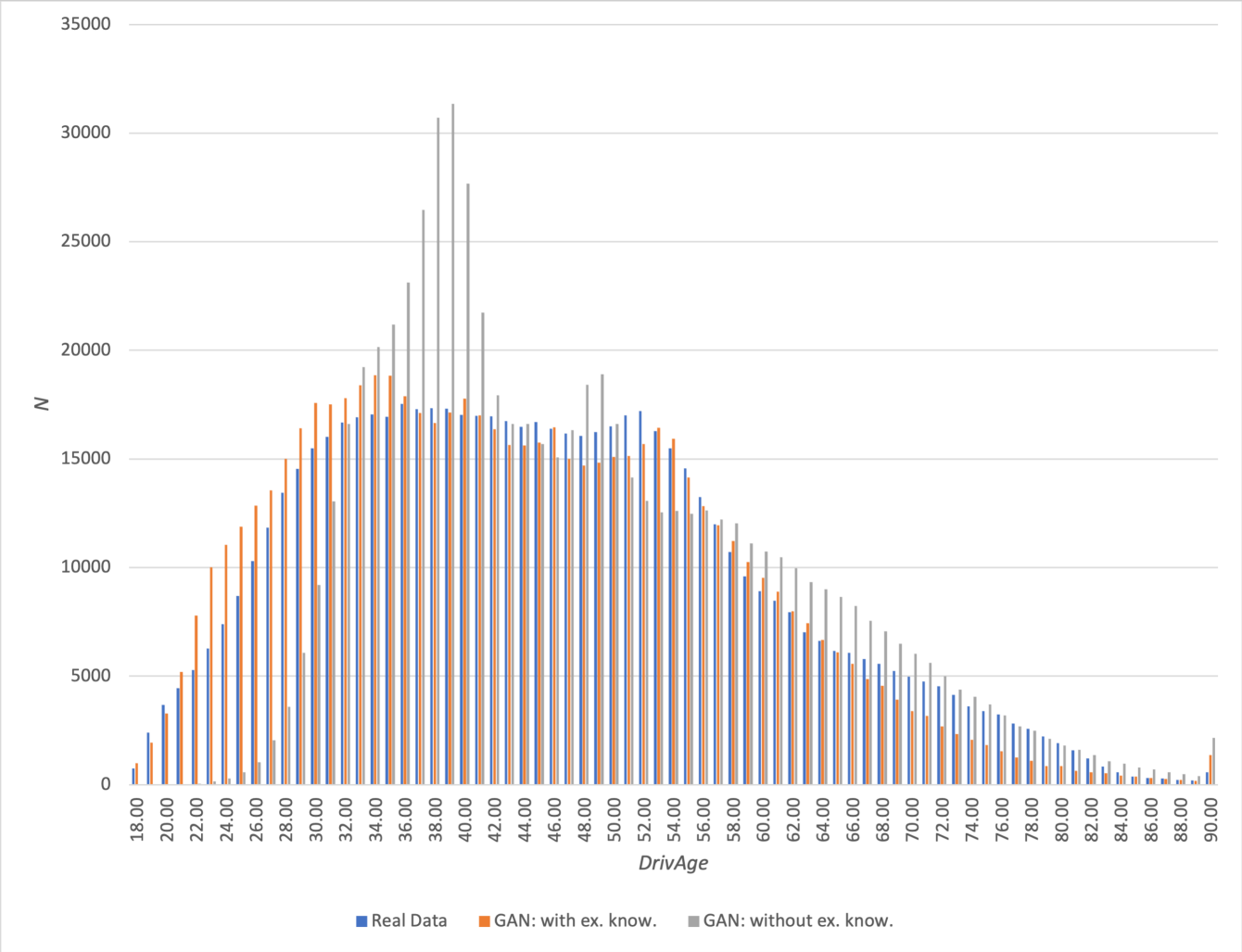

Figure 19: Comparison of distributions of *DrivAge*

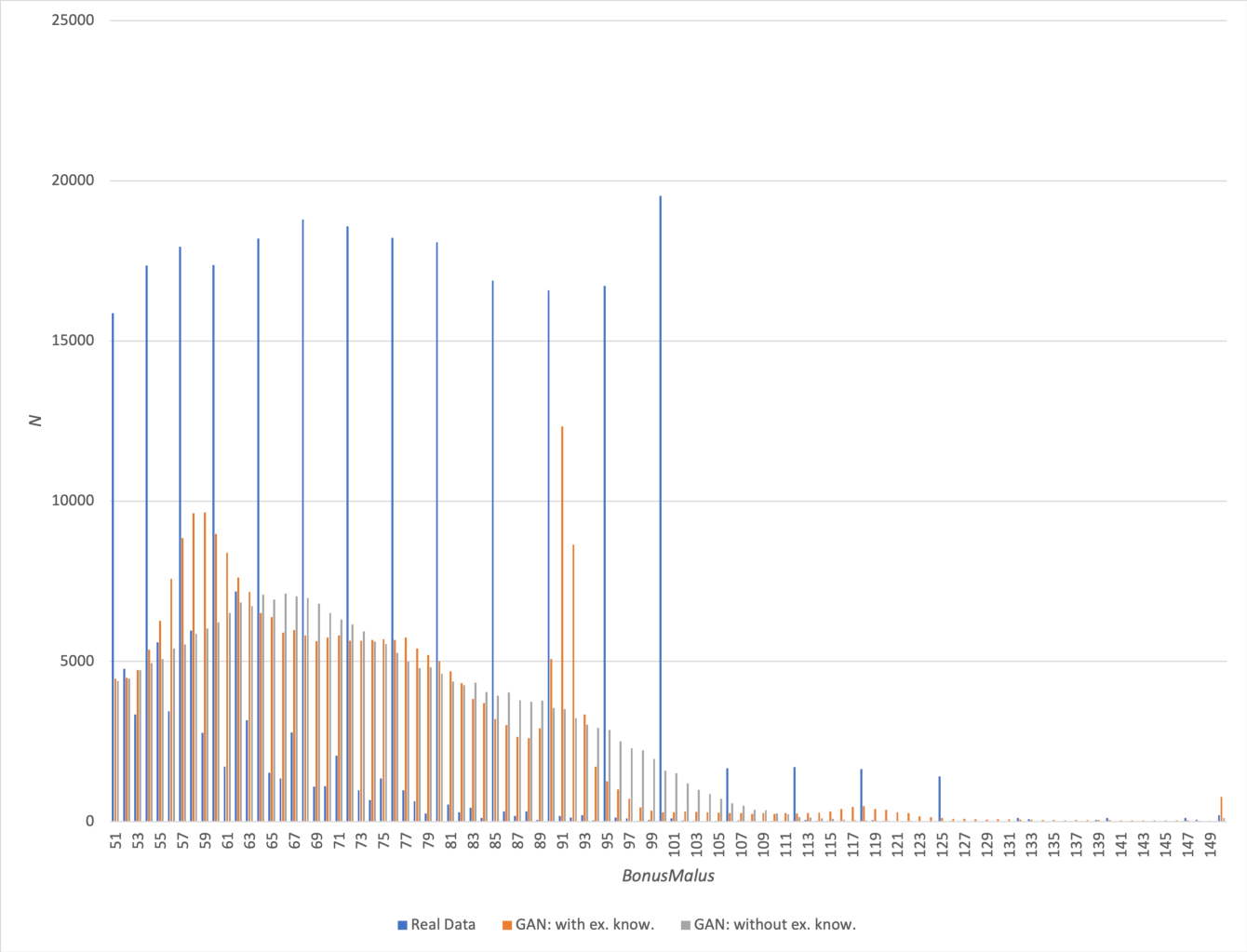

Figure 20: Comparison of distributions of *BonusMalus*, excluding $BonusMalus = 50$ **for visualisation purposes**

Appendix B COMPARISON OF RELATIONSHIPS BETWEEN INDEPENDENT VARIABLES AND *CLAIMNB* FOR FULL DATASET SCENARIO

In order to understand the impact of expert knowledge on synthetic data generation, we carried out a comparative analysis of the relationships between the independent variables and *ClaimNb* across three specific datasets in the full dataset scenario, which includes training, validation, and test set (N = 678013). These datasets were the original data, the synthetic data created by the GAN with expert input, and the synthetic data produced by the GAN without expert input. The figures and tables in this Appendix depict the relationships between various variables across these datasets. To ensure visual comparability, we calibrated the size of the synthetic data samples generated by the GANs to correspond with the size of the real dataset for this specific examination. Additionally, any synthetic data points that deviated beyond the boundaries of the real dataset (such as instances where *DrivAge* < *18*) were modified to align within the range defined by the real dataset.

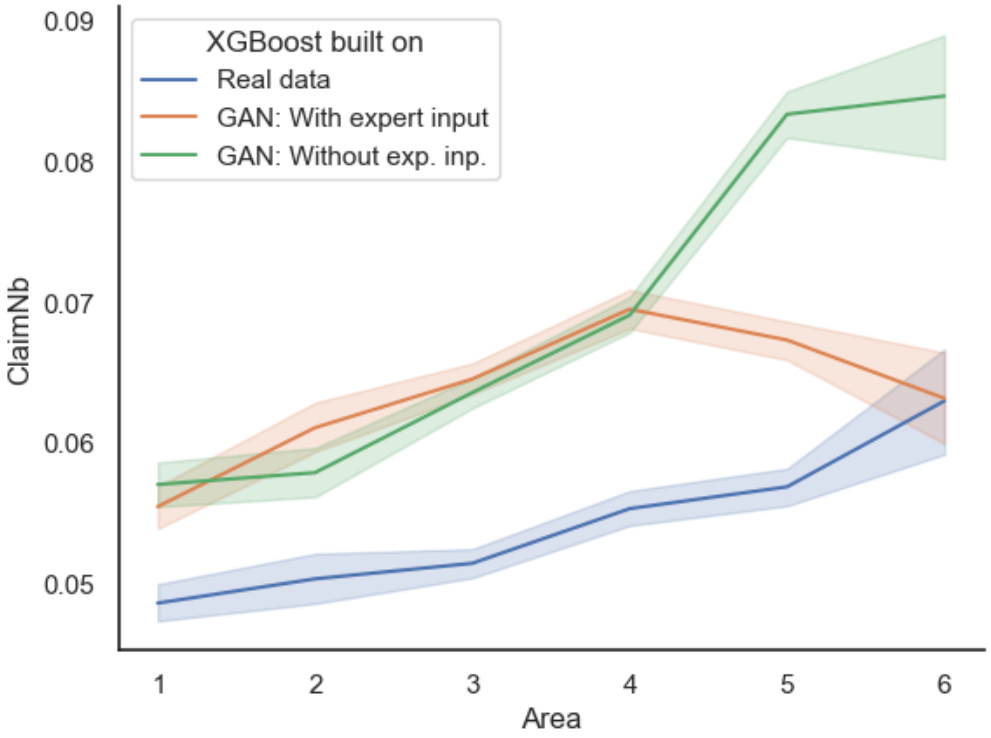

Figure 21: Enter Caption

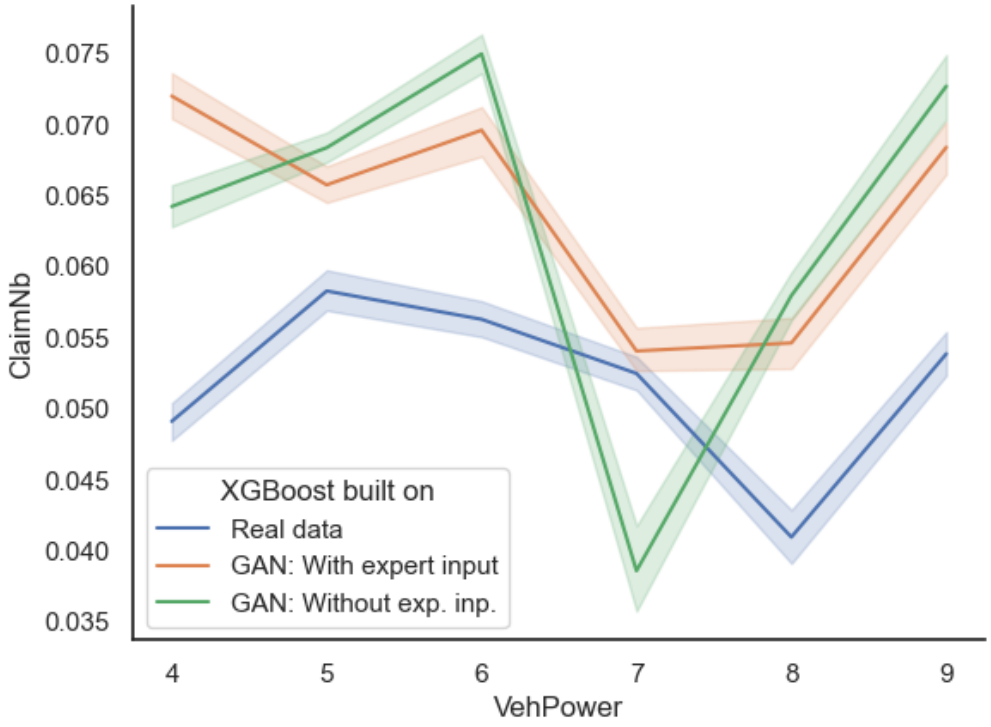

Figure 22: Enter Caption

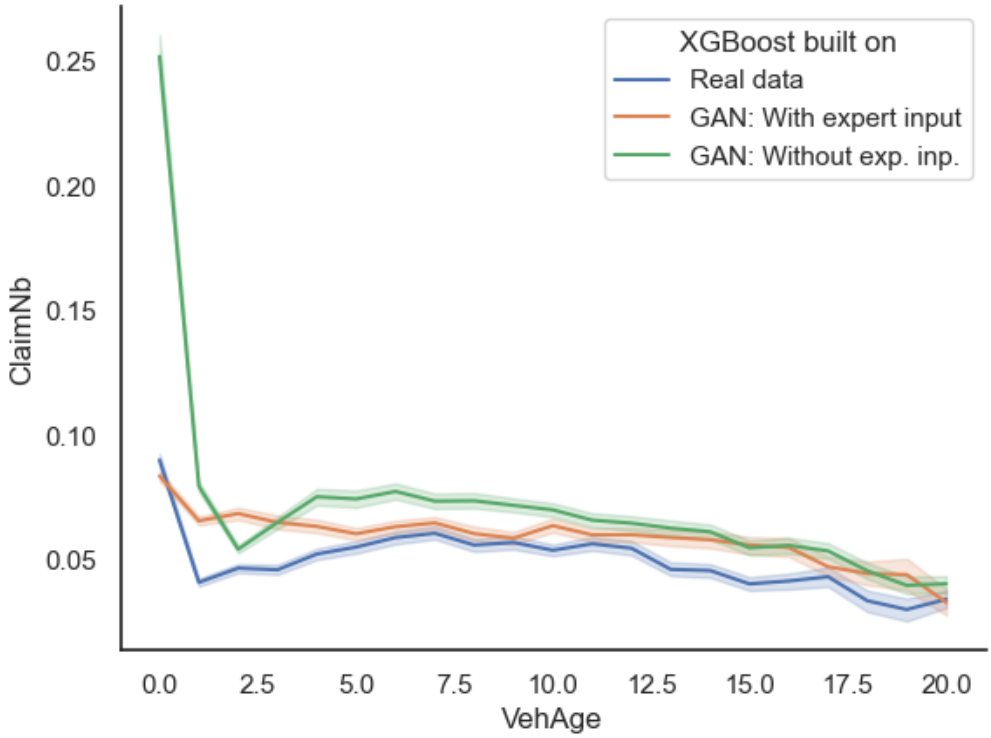

Figure 23: Enter Caption

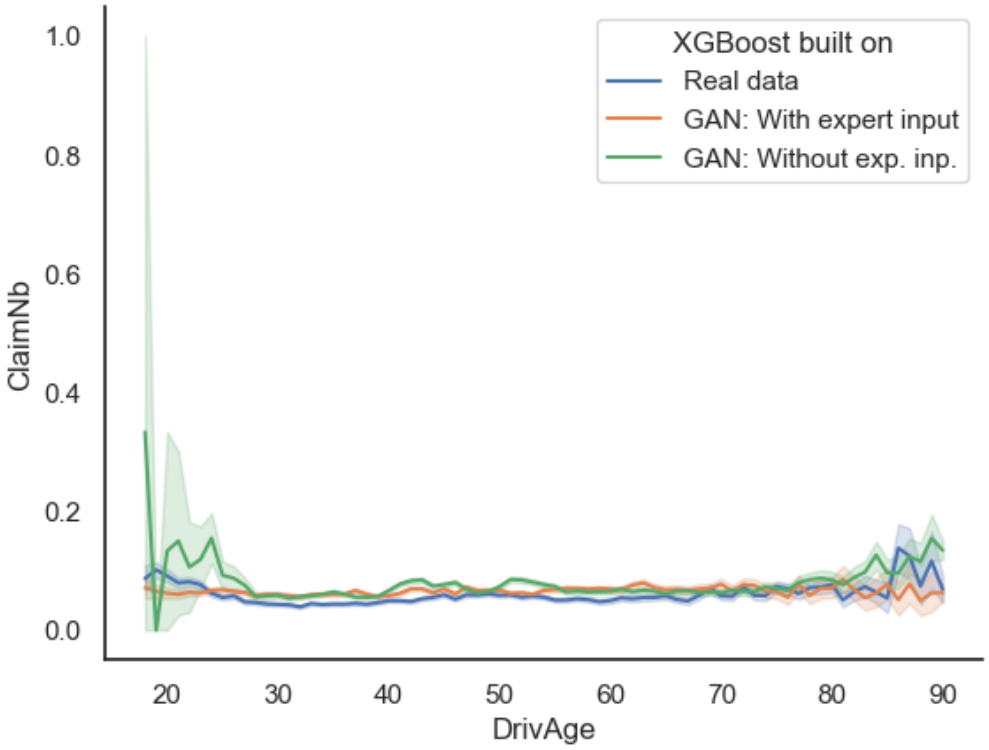

Figure 24: Enter Caption

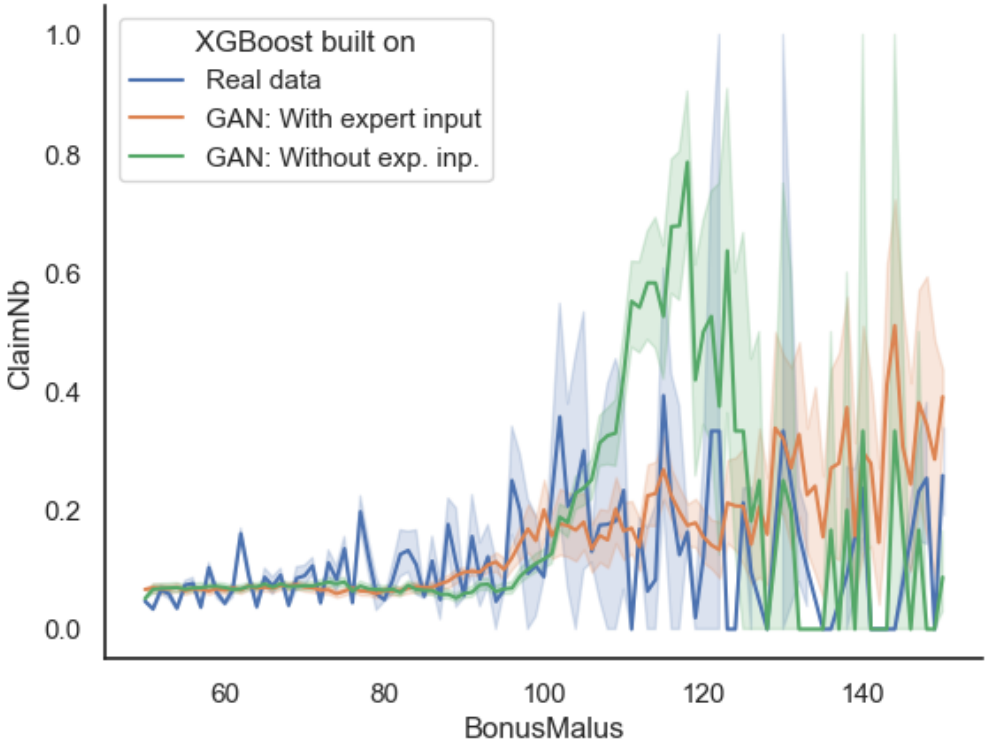

Figure 25: Enter Caption

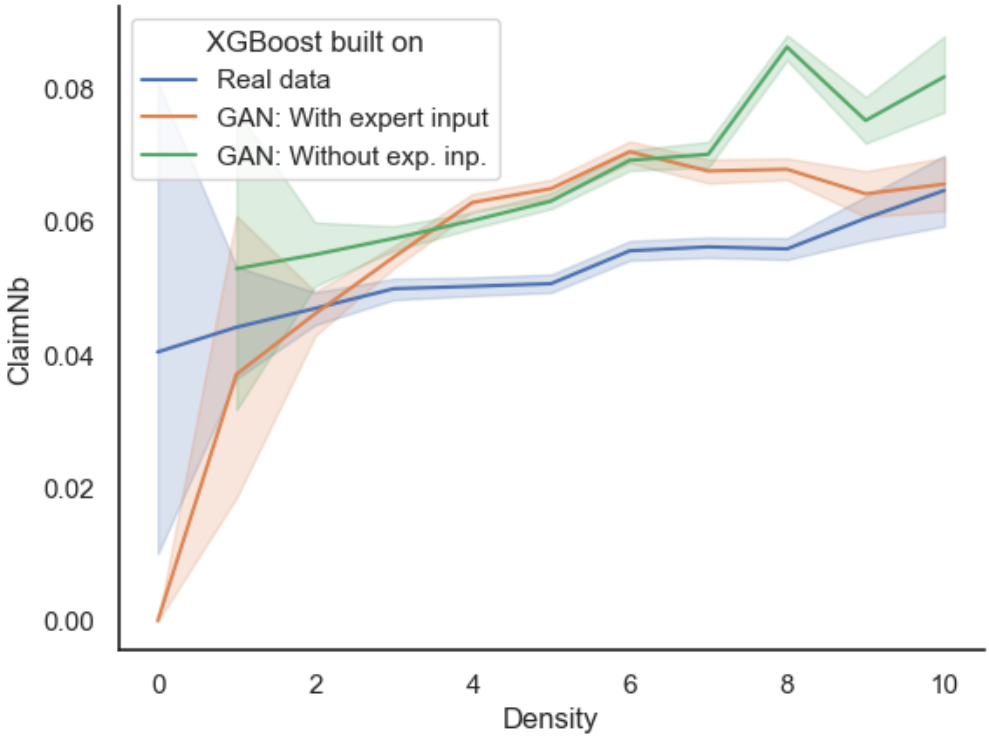

Figure 26: Enter Caption

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.

C.1 Scope Definition

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

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 $\hat{ClaimNb} = 0.05$.
- Additional Rules: No additional rules have been suggested.

See Figure 27 for the final representation.

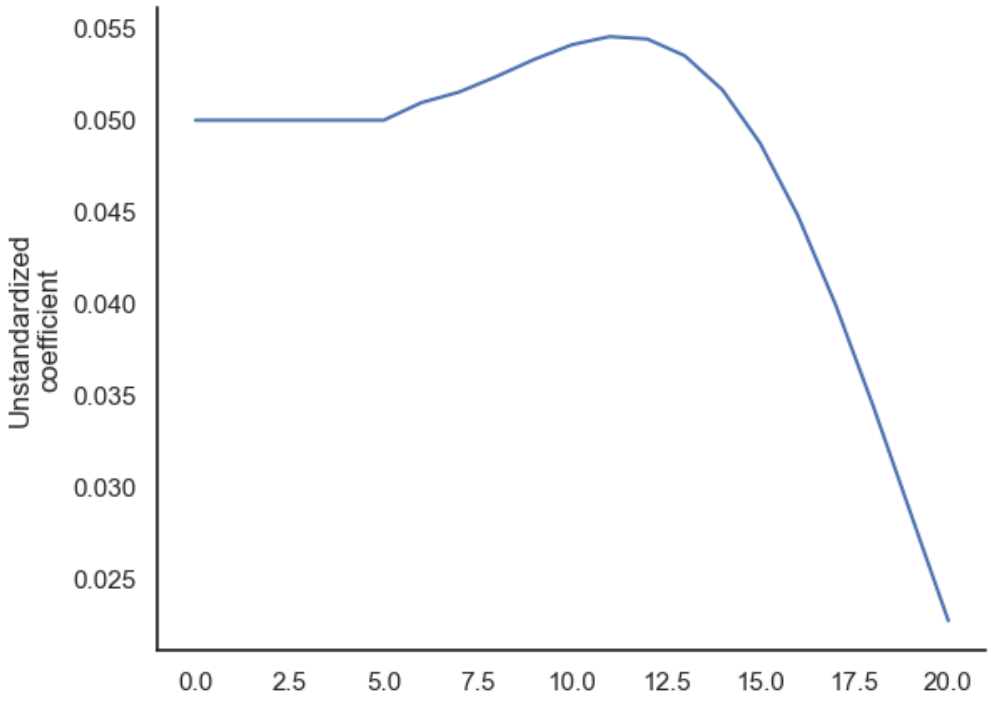

Figure 27: Chosen representation for the relationship between vehicle age and claim count

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 that best represents 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 28 for the final representation.

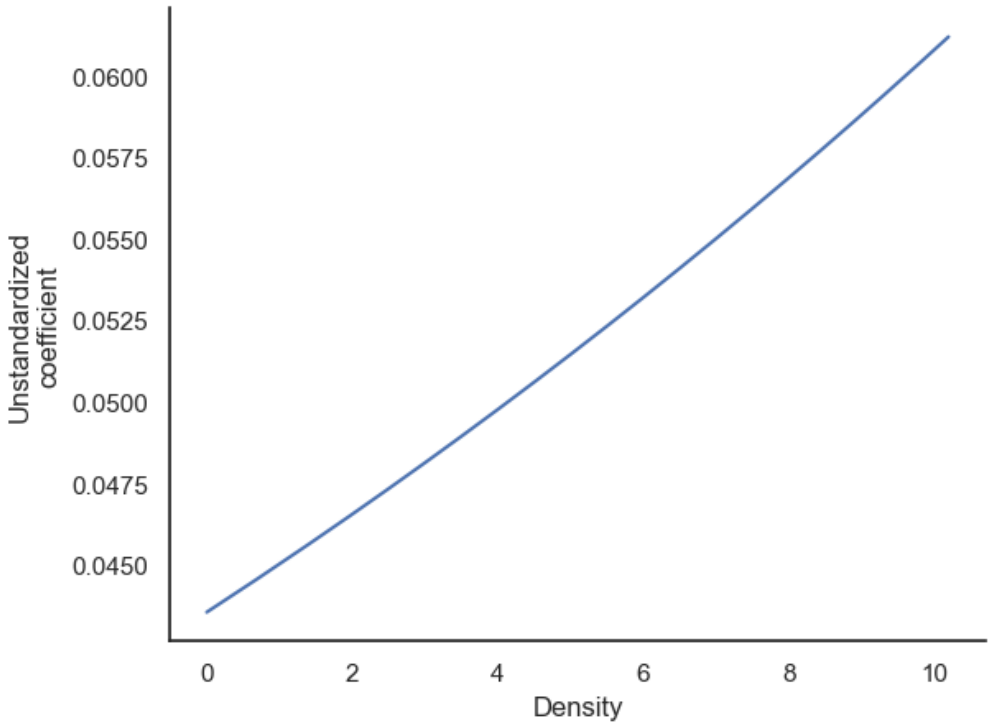

Figure 28: Chosen representation for the relationship between vehicle age and claim count

C.4 *BonusMalus*

- 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 that best represents the relationship is $ClaimNb = \beta_0 + \beta_1 \times BonusMalus + \beta_2 \times BonusMalus^2$.
- Representation Adjustment: The were no adjustments to the model after it was built

See Figure 29 for the final representation of *BonusMalus*.

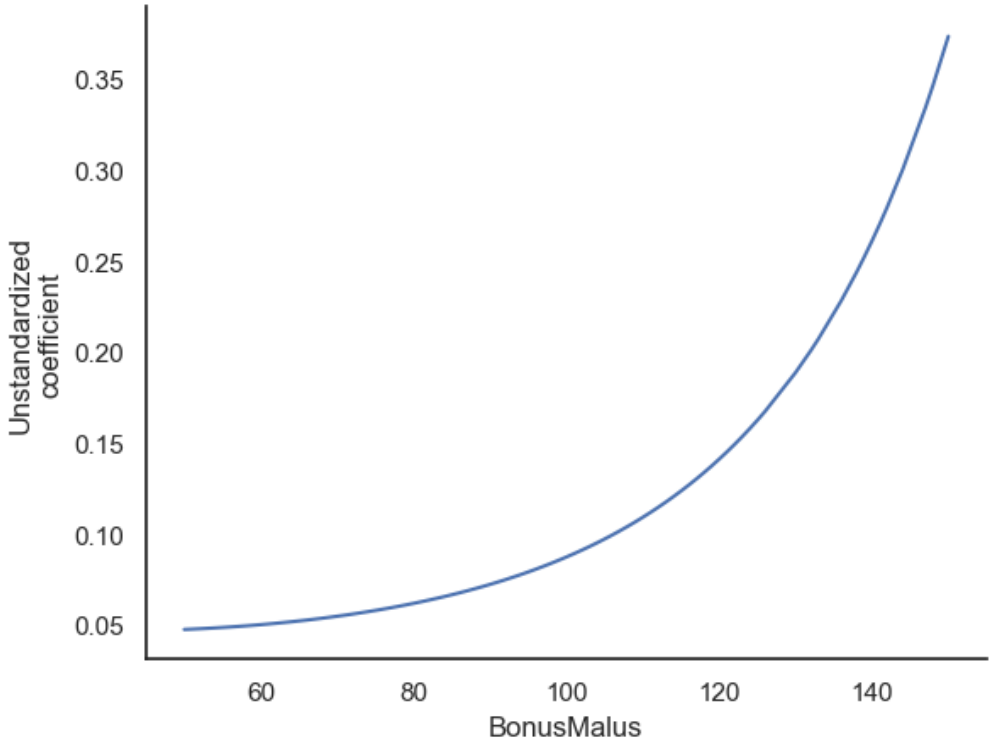

Figure 29: Enter Caption

- Additional Rules: 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 **??**). See Figure 29 for the final representation of the additional *BonusMalus* rule.

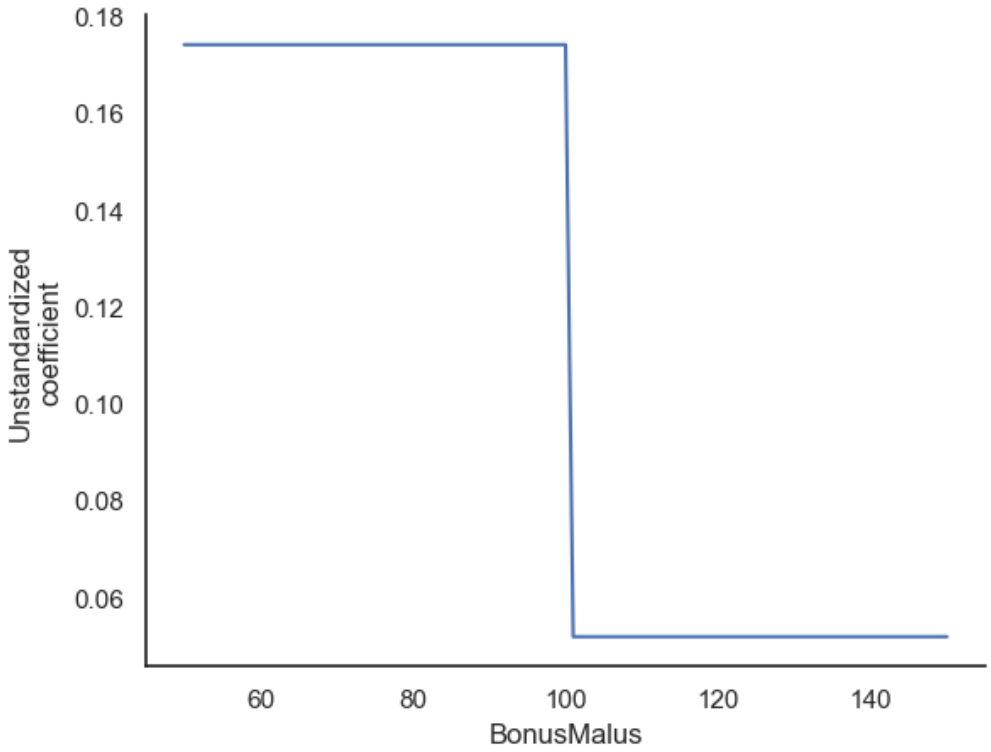

Figure 30: Enter Caption

C.5 *DrivAge*

- Idea Generation: The actuary proposed that the relationship between *DrivAge* and *ClaimNb* can be modeled using a polynomial Generalized Linear Model (GLM). The degree of the polynomial (that is, the highest power of *DrivAge* in the model) should be between 3 and 6. 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 $ClaimNb = \beta_0 + \beta_1 \times DrivAge + \beta_2 \times VehAge^2 + \beta_3 \times DrivAge^3 + \beta_4 \times DrivAge^4 + \beta_4 \times DrivAge^4 + \beta_5 \times DrivAge^5$.
- Representation Adjustment: No adjustment has been made by the adjustment
- Additional Rules: No additional rules have been suggested.

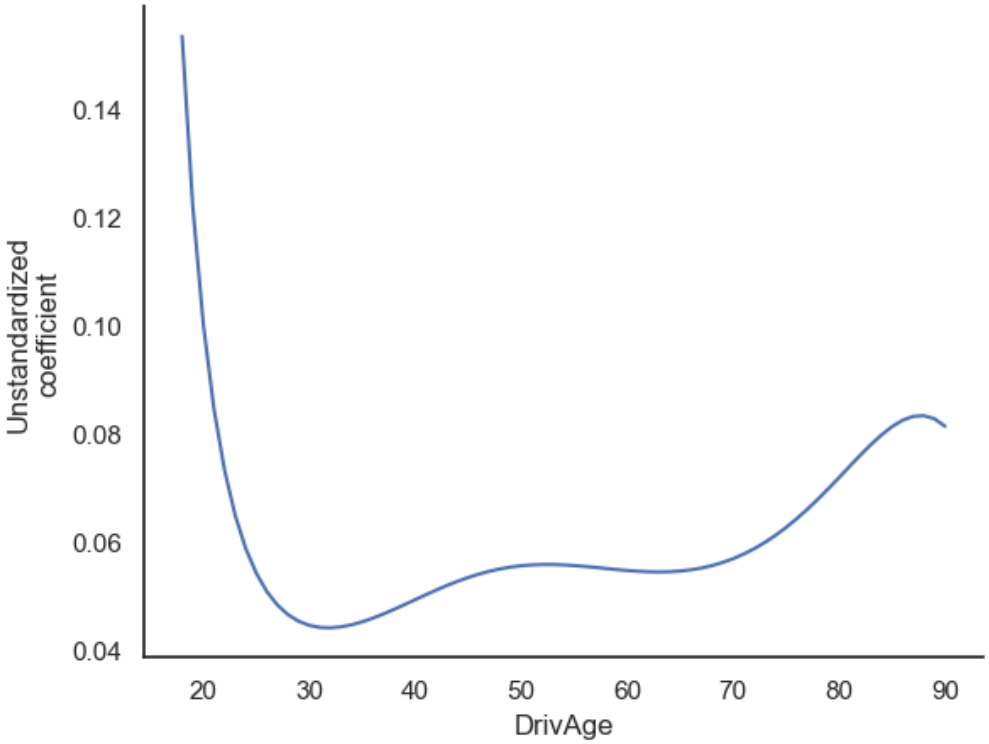

Figure 31: Enter Caption

Appendix D GRIDSEARCH OF GAN HYPERPARAMETERS

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 the respective ganruns.csv file in our GitHub repository (URL: https://github.com/AfairiJJ/thesis/blob/main/config/ganruns.csv).

The optimized hyperparameters were: mini-batch size for training (batch size), generator train gorunds per epoch (Gen. epochs), penalty factor for the loss function (Loss penalty), generator batch norm decay (Gen. bn. decay), critic batch norm decay (Cr. bn. decay), if the sigmoid function was used in the laster layer of the critic (Sigmoid). See Table 4 for the best found GAN hyperparameters.

Further parameters which were adopted from Cote et al. [4] without gridsearch were: Number of epochs (15,000), critic train rounds per epoch (2 per epoch), size of hidden layers in generator (3×100), generator learning rate (0.001), critic learning rate (0.001), generator L2 regularization factor (0), generator learning rate (1), critic learning rate (1),

Hyperparameter	GAN for Hypotheses 1.1, 1.2	GAN for Hypothesis 2.1, 2.2
Batch size	100	500
Gen. epochs	1	2
Loss penalty	10	5
Critic bn. decay	0.010	0.000
Gen. bn. decay	0.900	0.500
Critic leak par.	0.200	0.100
Noise size	75	100
Sigmoid layer	True	True
Critic L2 reg.	0.000	0.000
Critic hidden layers	100	100, 100

Table 4: Best hyperparameters found for the GANs for the different hypotheses

Appendix E GRIDSEARCH OF XGBOOST HYPERPARAMETERS