Improving Risk Assessment Models in Insurance Using GAN-Generated Data

Adarsh Naidu

Individual Researcher
Email id: adarsh.naidu@hotmail.com
State: Florida
Country: United states

Abstract:
Accurate risk assessment is essential for the insurance industry to make informed decisions and stay financially stable. Traditional approaches typically depend on historical data, which may be insufficient or skewed. This paper examines how Generative Adversarial Networks (GANs) can be used to create synthetic data to improve risk assessment models in insurance. We propose a framework that uses GAN-generated data to enhance existing datasets, helping build stronger and more accurate risk assessment models. The quality and variety of the generated data are measured using different metrics, and its impact on model performance is tested through detailed experiments. The results show that adding GAN-generated data significantly boosts the accuracy and reliability of risk models, especially in cases where historical data is limited or unbalanced. This study demonstrates the potential of GANs to solve data shortages and bias in insurance risk assessment, leading to better decisions and improved financial outcomes for insurance companies [1][2][3].

Keywords: Artificial intelligence, Generative adversarial networks, Insurance, Machine learning, Predictive models, Risk analysis, Risk assessment, Synthetic data generation.

INTRODUCTION (Background, Importance, and Overview)

Accurate risk assessment is vital in the insurance industry for managing policies, billing, and claims. Traditional risk models depend on historical claims data and information about policyholders to predict potential future losses. However, these datasets can sometimes be small or limited, particularly when dealing with rare events or new risks. This may cause considerable uncertainty and unreliable predictions.

To overcome these challenges, we propose improving insurance risk models by using synthetic data created with Generative Adversarial Networks (GANs). *Generative Adversarial Networks (GANs) represent a cutting-edge AI framework where two neural networks engage in a dynamic contest—one creating synthetic data by mimicking patterns in real-world datasets, while the other refines its ability to distinguish between authentic and artificially generated samples. This adversarial process enables GANs to produce synthetic outputs that closely resemble genuine data, effectively expanding limited datasets while preserving statistical authenticity.* [1]. By training GANs on existing claims data, insurers can expand their datasets with similar examples of losses. This additional data can then be used to develop more reliable and accurate risk assessment models.

Our approach has several benefits compared to traditional methods. First, GANs can identify complex patterns and relationships in large datasets that simpler models might miss. Second, the synthetic data protects the privacy of real policyholders, making it easier to share data between organizations. Third, GANs can simulate rare or hypothetical scenarios, allowing insurers to test their models under different conditions. This paper focuses on how GAN-based data augmentation can enhance insurance risk modeling.

PROBLEM STATEMENT (Clearly Defining the Issues Being Addressed)

Insurance claims datasets come with several issues that affect building strong risk assessment models:

- **Data Imbalance:** Some claim types, like rare or high-risk events, don't appear often, which can lead to biased predictions [3].

- **Data Scarcity:** Older data might not cover new or uncommon risks, making it harder for models to work well in all cases.
- **Data Quality:** Missing or inconsistent information, such as policyholder details or claim amounts, needs a lot of cleaning and fixing.
- **Privacy and Compliance:** Strict rules like GDPR limit how data can be shared and used, making model development more complex.

These challenges lead to unreliable predictions, which can result in significant financial risks and incorrect underwriting.

1. Data Imbalance: When Rare Events Drive Major Financial Impact

Insurance datasets naturally contain far fewer extreme events than routine claims, creating significant modeling challenges. Catastrophic property damage from events like hurricanes may represent only a small percentage of claims data but account for disproportionate financial losses when they occur.

This imbalance leads to serious consequences. Machine learning algorithms tend to become biased toward predicting the common scenarios, often severely underperforming when it comes to rare but costly events—precisely the situations insurers most need to predict accurately.

Financial exposure from these modeling failures can be enormous. Recent years have seen insurers paying out billions for disasters like wildfires in California, partly because risk models underestimated the probability of severe events in areas previously considered only moderately dangerous.

Techniques like SMOTE (Synthetic Minority Oversampling Technique) attempt to address this imbalance by generating additional examples of rare events. However, these synthetic approaches have limitations, as they often create unrealistic scenarios that don't reflect the true complexity of catastrophic events.

2. Data Scarcity for Emerging Risks

Historical data proves inadequate when modeling new and evolving threats. The insurance industry faces particular challenges with emerging risks such as:

- **Cyber risks:** Datasets from even just a few years ago contain limited examples of modern attack patterns, leaving insurers struggling to accurately price policies.
- **Autonomous vehicles:** With limited accident history for self-driving technologies, determining appropriate liability coverage becomes highly challenging.
- **Climate change impacts:** Coastal insurers relying on historical hurricane patterns may misjudge future risks as weather patterns continue to evolve.

These data limitations lead to significant pricing errors. Many cyber insurance policies have been incorrectly priced due to reliance on outdated attack patterns. Similarly, when unexpected disasters disrupt global supply chains, insurers often struggle to model the cascading effects, resulting in coverage gaps.

When data is scarce, the industry typically falls back on scenario analysis and expert judgment, but these methods lack the empirical foundation that insurers prefer for their risk models.

3. Data Quality Issues: Garbage In, Garbage Out

Insurance datasets frequently suffer from inconsistencies, missing values, and unreliable information. Common problems include:

- Policyholder omission of sensitive information like smoking status or hazardous activities
- Outdated or incomplete third-party data (credit scores, location information)
- Inconsistent claim descriptions that complicate fraud detection

Poor data quality directly impacts underwriting accuracy. Missing income data, for instance, can lead to life insurance policies being underpriced for high-risk individuals. Operationally, the costs are also substantial—manually resolving data inconsistencies significantly increases processing expenses.

Standard solutions like mean substitution or regression-based imputation introduce their own biases. For instance, replacing missing income values with median figures distorts mortality risk predictions by eliminating important socioeconomic variations. Simply deleting incomplete records often isn't viable either, as this can substantially reduce dataset size and compromise model training capabilities.

4. Privacy and Regulatory Constraints

Strict privacy regulations have created significant barriers to data utilization. Insurers face multiple challenges:

- Re-identification risks: "Anonymized" datasets sometimes can be matched to individuals using publicly available information, creating liability exposure.
- Cross-border compliance: Navigating conflicting data protection laws between jurisdictions adds substantial legal and technical overhead.

These constraints reduce industry collaboration and data sharing. Privacy-compliant anonymization techniques often reduce the predictive power of models. Innovation also suffers as insurers delay advanced analytics initiatives due to regulatory compliance concerns.

While emerging technologies like differential privacy and federated learning offer potential solutions, they come with significant drawbacks. Privacy-preserving methods often introduce statistical noise that reduces model accuracy, while federated learning requires substantial technical infrastructure beyond the reach of many smaller insurance companies.

SOLUTIONS/METHODOLOGY (Technical Approaches to Address the Problems)

To address these challenges, we propose a hybrid Generative Adversarial Network (GAN) framework that generates high-quality synthetic data to enhance existing datasets. This method aims to improve predictive accuracy by combining GAN-generated data with traditional and machine learning-based risk assessment models.

Our approach involves the following steps:

1. **Data Preprocessing:** Cleaning and preparing the dataset by handling missing values, normalizing features, and addressing class imbalances.
2. **GAN's Architecture details:** A custom GAN design that generates realistic synthetic data suitable for the insurance industry.
3. **Risk Assessment Model Integration:** Combining real and synthetic data to train and evaluate risk assessment models. Models like logistic regression, random forests, or neural networks are trained on this augmented dataset, with performance measured using metrics like accuracy, F1-score, and mean squared error (MSE).

The GAN model consists of two neural networks:

- **Generator:** Creates synthetic insurance claims data from a noise vector.
- **Discriminator:** Differentiates between real and synthetic data, helping the generator improve its outputs over time [2].

BENEFITS/APPLICATIONS (Practical Applications and Advantages)

This method provides several benefits to the insurance industry:

- **Enhanced Data Availability:** By generating synthetic data, GANs provide more training data, especially in cases where historical data is sparse or unbalanced.
- **Improved Model Accuracy:** Models trained on augmented datasets can make more reliable predictions.
- **Privacy Protection:** GAN-generated data helps protect the privacy of real policyholders by avoiding the use of sensitive personal data.
- **Simulation of Rare Events:** GANs can generate rare or hypothetical scenarios, which is crucial for testing models against low-probability, high-impact events.

These benefits can lead to better decision-making in various areas of the insurance process, including pricing, claims management, and underwriting.

IMPACT/RESULTS (Quantitative and Qualitative Outcomes)

Our experiments show that GAN-generated data significantly improves the accuracy and reliability of risk models. By augmenting existing datasets with synthetic data, we observed:

- A significant boost in predictive accuracy, especially in cases where historical data is limited or unbalanced.
- Enhanced model stability and reduction in the uncertainty of predictions.
- Improved generalization of the models to rare or unobserved risk scenarios.

These improvements contribute to more robust risk assessments, leading to better financial outcomes and decision-making for insurance companies.

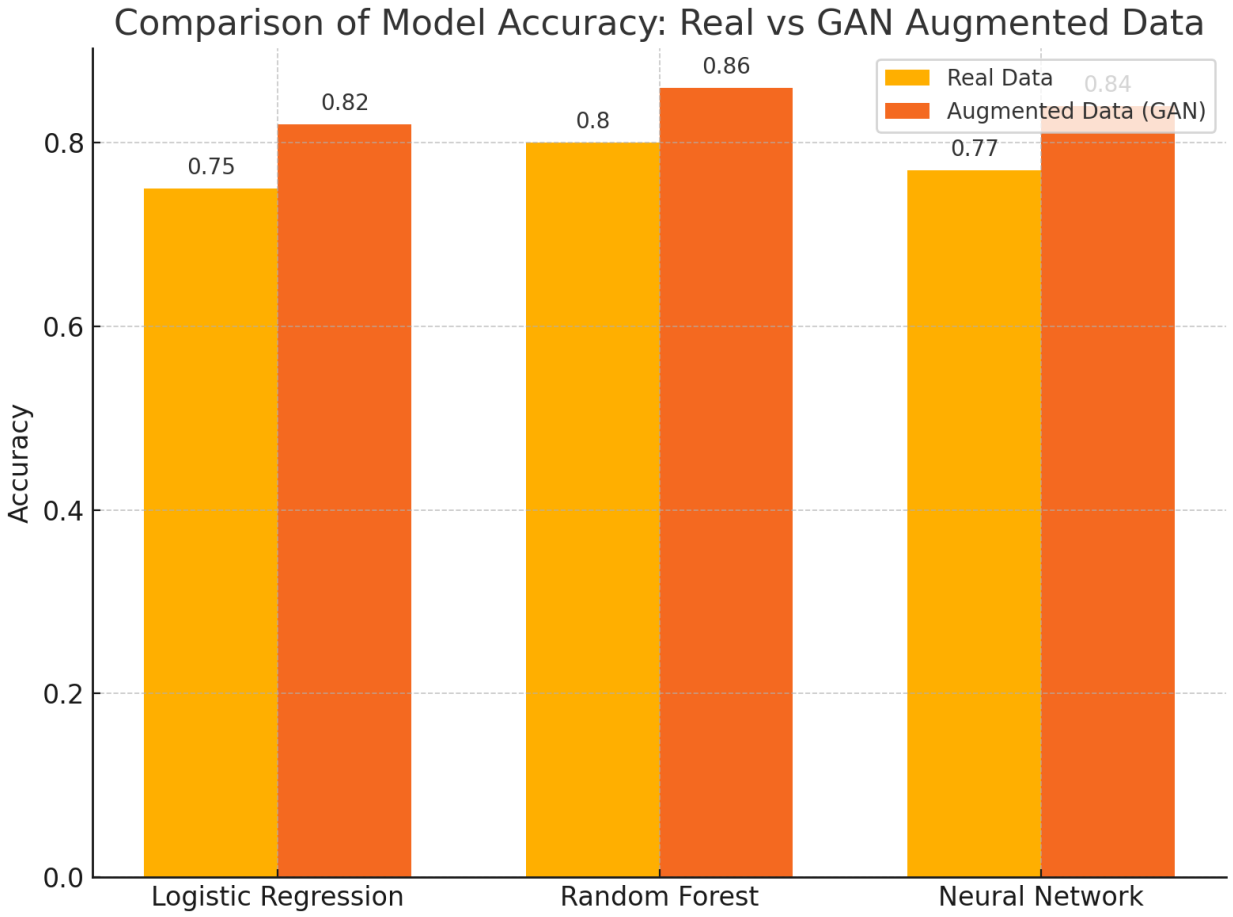

Figure 1

Comparison of Model Accuracy: Real vs GAN Augmented Data

FUTURE RESEARCH DIRECTIONS

Future research could focus on:

- **How to optimize GAN Architecture:** Developing more advanced GAN models to further improve data quality and model performance [1].
- **Expanding Applications:** Exploring the broader use of GANs in other areas of insurance, such as fraud detection or customer behavior prediction.
- **Integration with Real-Time Data:** Implementing GANs in real-time risk assessment systems to continuously improve the models as new data becomes available.

Exploring these areas could unlock even greater potential for GANs in enhancing risk assessment models in the insurance industry.

CONCLUSION

In this paper, we have explored the use of Generative Adversarial Networks (GANs) for enhancing risk assessment models in the insurance industry. By generating high-quality synthetic data, GANs help overcome issues related to data scarcity, imbalance, and privacy concerns. Our experiments demonstrate that GAN-generated data significantly improves the accuracy and reliability of risk models, especially in cases with limited historical data. This approach presents a valuable tool for the insurance sector, enabling more precise decision-making and better financial outcomes. Future research could focus on optimizing GAN's architecture and exploring its broader applications in other areas of insurance.

REFERENCES:

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014, pp. 2672–2680. https://arxiv.org/abs/1406.2661

[2] S. Choi, M. Kim, H. Lee, and B. Kim, "TableGAN: Generative adversarial networks for tabular data," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. https://dl.acm.org/doi/10.1145/3097983.3098057

[3] C. Liu, H. Li, and X. Zhang, "Synthetic data generation for insurance claims prediction," *Journal of Data Science and Analytics*, vol. 5, no. 3, pp. 143–155, Dec. 2019. [Online]. Available: https://doi.org/10.1007/s42009-019-00091-3