IRSTI 50.01 https://doi.org/10.26577/jpcsit2024-v2-i4-a5

A. Dildabek , Z. Abdiakhmetova*

Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: zukhra.abdiakhmetova@gmail.com

USING SYNTHETIC DATA TO IMPROVE DATA PROCESSING ALGORITHMS IN BUSINESS INTELLIGENCE

Abstract. The growing volumes of data require the development of effective methods for its processing to solve practical problems. This study is devoted to the use of synthetic data to improve data processing algorithms in business analysis tasks. Synthetic data has a number of benefits, including increasing the amount of data available to train models and ensuring privacy when working with sensitive financial and medical data. The paper examines the potential of synthetic data generated by CTGAN and TVAE methods for regression problems. The study uses two datasets–Health Insurance and Boston Housing–to evaluate the performance of machine learning models, such as linear regression, random forest, and gradient boosting. The results suggest that synthetic data can significantly improve algorithm performance, especially for small or unbalanced datasets, although challenges remain in achieving quality comparable to real-world data. The study highlights the practical importance of synthetic data for optimizing business processes and opens up new opportunities for further study of data generation methods and their application.

Key words: Synthetic data, data processing, CTGAN, TVAE, Linear Regression, RandomForestRegressor, GradientBoostingRegressor

1. Introduction

In today's world, where the volume of data is growing exponentially, synthetic data is becoming an essential tool for overcoming the limitations of data privacy and accessibility. Synthetic data is created artificially, mimicking the properties of real data, which allows it to be used in the tasks of analysis, modeling and training algorithms without the risk of disclosing personal information.

There are many methods for generating synthetic data, such as Conditional Tabular GAN (CTGAN) and Tabular Variational AutoEncoder (TVAE), which have proven to be effective approaches for processing tabular data [1], [2]. These methods allow you to create high-quality, synthetic datasets that preserve the basic properties of the original data, which is critical for prediction and classification tasks. The literature emphasizes the importance of synthetic data in business intelligence and healthcare. Studies demonstrate their versatility and value for optimizing business processes [3], [4]. In addition, the potential of synthetic data in improving the accuracy of machine learning models by increasing the amount of data available is highlighted [5]. However, despite the benefits, the challenges associated with synthetic data generation remain significant. These include ensuring privacy, preventing attacks on data privacy, and keeping synthetic data consistent with real-world datasets. It is important that synthetic data meets the requirements of the tasks for which it is intended, making its use a reliable and effective tool. The purpose of this work is to evaluate the use of CTGAN and TVAE methods for generating synthetic data and their impact on the performance of machine learning models. Two datasets were used for this purpose: Boston Housing and Health Insurance Dataset. The main attention is paid to their application for regression and optimization of business processes, as well as to the comparison of their effectiveness with real data.

2. Materials and Methods

2.1. Datasets

Two datasets were chosen for the study: Health Insurance Dataset and Boston Housing Dataset. The choice of these datasets was determined by several factors that are important for achieving the goals of the experiment. First, the two datasets differ in scope. The Health Insurance dataset contains 2700 lines, while the Boston Housing dataset includes only 507 lines. The difference in size allows you to study the impact of data growth in different conditions, from medium-sized datasets to relatively small ones. This is critical because the real-world data that data scientists encounter often has different scales. The ability to evaluate the effectiveness of methods on sets of different sizes makes the study more comprehensive. Secondly, both datasets are classic regression problems that are widely used in machine learning research. The Health Insurance Dataset presents premium data and includes a variety of characteristics that affect the cost of insurance, making it an excellent example for analyzing risk factors. The purpose of the analysis of this dataset was to study the factors that affect premium rates and build models for their accurate prediction [6], [7].

The Boston Housing Dataset is a classic dataset for estimating housing values, including characteristics such as crime rates in a neighborhood, distance from work centers, school quality, and many other socioeconomic and physical parameters. The goal of analyzing this dataset was to predict housing values based on various characteristics, which is a common task in economics and urban planning. The presence of various types of features in datasets, such as numeric, categorical, and even Boolean data, allows us to explore how synthetic data generation methods cope with different types of information. This helps to identify potential problems and benefits of the approach, as well as draw conclusions about the applicability of synthetic data for specific types of business problems.

2.2. Generation of synthetic data

Two methods were used to increase the volume of original data: CTGAN (Conditional Tabular Generative Adversarial Networks) and TVAE (Tabular Variational Autoencoder) [8]. These methods were chosen because they have proven themselves in tabular data generation tasks and can effectively reproduce the structure of the original data while ensuring that it is realistic.

- CTGAN uses a generative adversarial network approach, which allows it to generate data that is similar in distribution to the original. This is especially important for simulating complex dependencies between features.
- TVAE is a variant autoencoder designed to work with tabular data. It encodes the original data into a latent representation and then decodes it back to generate new synthetic instances, which also allows for high accuracy of feature imitation [9], [10].

Experiments were conducted with different values for the number of learning epochs: 300, 1000, and 10,000 epochs. These values were chosen to evaluate the effect of training depth on the quality of the generated data and to determine the optimal training time for each method on the size of the original dataset, which made it possible to assess how different amounts of added data affect the performance of models.

2.3. Regression models and the evaluation process

Three regression algorithms were selected to evaluate the effectiveness of the generated synthetic data:

1. Linear Regression: Classical linear regression is a basic method for analyzing data. It makes it easy to interpret the results and is a good starting point for assessing data quality.

2. RandomForestRegressor: An ensemble method based on building multiple decision trees. This technique is well suited for handling unbalanced and noisy data, making it useful in conditions of data growth.

3. GradientBoostingRegressor: A powerful ensemble algorithm that trains a series of decision trees to minimize error. This method was chosen because it often shows high accuracy on small samples, and it was important to see if synthetic data augmentation could improve its performance[11], [12].

The model evaluation process involved training and testing on the original data and then on the combined datasets (original + synthetic data). Thus, it was possible to visually see how the synthetic increase in data affects the quality of the predictions. For each combination of the dataset, the synthetic data generation method and the regression algorithm, estimates were carried out using standard metrics such as MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) and R^2. The results of the experiment made it possible to assess not only the change in the accuracy of predictions, but also to identify potential limitations in the use of synthetic data to optimize big data processing in business processes.

The study created and evaluated synthetic datasets using CTGAN and TVAE models, with the original datasets being health insurance rates (2,700 rows) and Boston real estate data (507 rows). Regression models, including linear regression, random forest, and gradient boosting, were applied to both the source data and synthetic data generated at different numbers of epochs (300, 1000, and 10000). Below are the detailed results.

Performance measures obtained using synthetic and raw data for the health insurance betting dataset show significant differences in model accuracy. When linear regression was applied to the original dataset, a coefficient of determination of R^2 of 0.74 was achieved. However, for synthetic data generated using CTGAN, estimates ranged from 0.54 at 300 epochs to 0.71 at 10000 epochs. The data generated by TVAE showed slightly better consistency with estimates of 0.74, 0.73, and 0.68 for the 300, 1,000, and 10,000 epochs, respectively, indicating comparable performance to the original data in some cases[13]. For the random forest model, the original data gave a high score of 0.93, whereas synthetic data from CTGAN and TVAE showed a decrease in performance, especially when using CTGAN at 300 epochs, resulting in a score of 0.61. Increasing the number of epochs improved the score to 0.80 and 0.88 for CTGAN and TVAE at 10000 epochs, respectively. TVAE outperformed CTGAN overall for random forest, indicating its suitability to capture the necessary features to better predict the model. Gradient boosting showed similar trends, with a score of 0.88 for the original dataset, while CTGAN-generated datasets showed lower results, ranging from 0.58 to 0.78. However, TVAE was able to generate data that resulted in a comparable performance of 0.84 at 300 epochs, which shows promising results in terms of trait representation (see Table 1).

Table 1 – Results of regression analysis on the Health Insurance dataset

Count epochs	Original Data	Ctgan 300	Ctgan 1000	Ctgan 10000	Tvae 300	Tvae 1000	Tvae 10000
Linear Regression	0.74	0.54	0.66	0.71	0.74	0.73	0.68
RandomForest Regressor	0.93	0.61	0.75	0.80	0.88	0.83	0.83
GradientBoosting Regressor	0.88	0.58	0.72	0.78	0.84	0.81	0.80

The Boston real estate dataset, consisting of 507 rows, was also tested using the same approach. A linear regression model applied to the original dataset yielded an R^2 coefficient of determination of 0.71. In contrast, the synthetic data generated by CTGAN showed a decrease in performance, with estimates ranging from 0.41 at 300 epochs to 0.65 at 10000 epochs. TVAE showed similar results, with the highest score of 0.63 at 10,000 epochs, indicating that the quality of synthetic data improves with increasing training time, but still lags behind the original data in terms of model performance [14]. For the random forest, the original dataset yielded a score of 0.87, whereas synthetic data from CTGAN and TVAE showed performance degradation across all epochs, with the highest score of 0.77 for both models at 10,000 epochs. This suggests that while increasing the number of epochs improves the quality of synthetic data, there remains a gap compared to the use of real data. Results for gradient boosting showed an R^2 coefficient of determination of 0.89 for the original dataset. Synthetic datasets generated by CTGAN and TVAE showed lower performance, with the highest scores of 0.77 and 0.70 at 10000 epochs, respectively (see Table 2).

Table 2 – Results of regression analysis on the Boston Housing dataset

Count epochs	Original Data	Ctgan 300	Ctgan 1000	Ctgan 10000	Tvae 300	Tvae 1000	Tvae 10000
Linear Regression	0.71	0.41	0.57	0.65	0.58	0.50	0.63
RandomForest Regressor	0.87	0.47	0.60	0.77	0.75	0.63	0.72
GradientBoosting Regressor	0.89	0.44	0.63	0.77	0.75	0.62	0.70

3. Results

The results for both datasets suggest that while the synthetic data generated by CTGAN and TVAE may approach the performance of the original dataset to a certain extent, a performance gap remains, especially for complex models such as random forest and gradient boosting [15]. The figure shows a comparison of R² coefficients of determination for linear regression, random forest, and gradient boosting models trained on source and synthetic data for the Health Insurance dataset (Figure 1). As you can see from the graph, TVAE shows better performance than CTGAN in most cases, especially when increasing the number of epochs to 10,000. However, the performance of the synthetic data is still lower than that of the original data, which is confirmed by the results on all models.

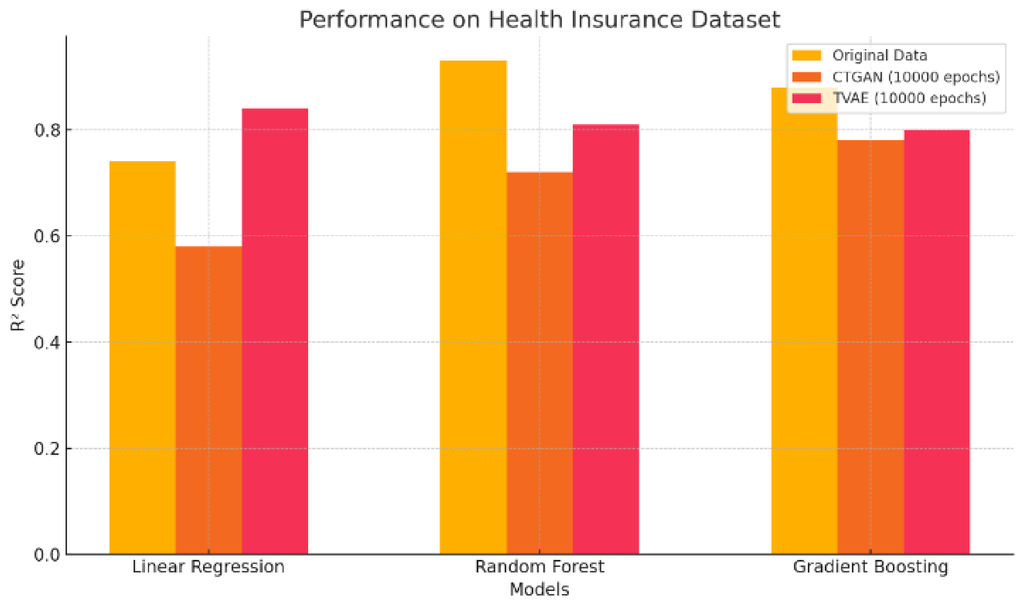

Figure 1 – Compare the performance of machine learning models on source and synthetic data for the Health Insurance dataset

The following image shows the results for the Boston Housing dataset (Figure 2). Similar to the first dataset, TVAE performed more consistently, but the performance of the synthetic data was inferior to the original for all three models. Complex models such as gradient boosting are particularly sensitive to data quality, as evidenced by the marked decrease in the R² coefficient of determination when using synthetic data generated by CTGAN.

These results highlight that the effectiveness of synthetic data is highly dependent on the generation methods used and the number of training epochs. Increasing the number of epochs results in improved performance, as can be seen in both graphs. However, to close the performance gap, it may be necessary to further optimize data generation methods, including CTGAN and TVAE, as well as study how to adapt them to specific regression tasks.

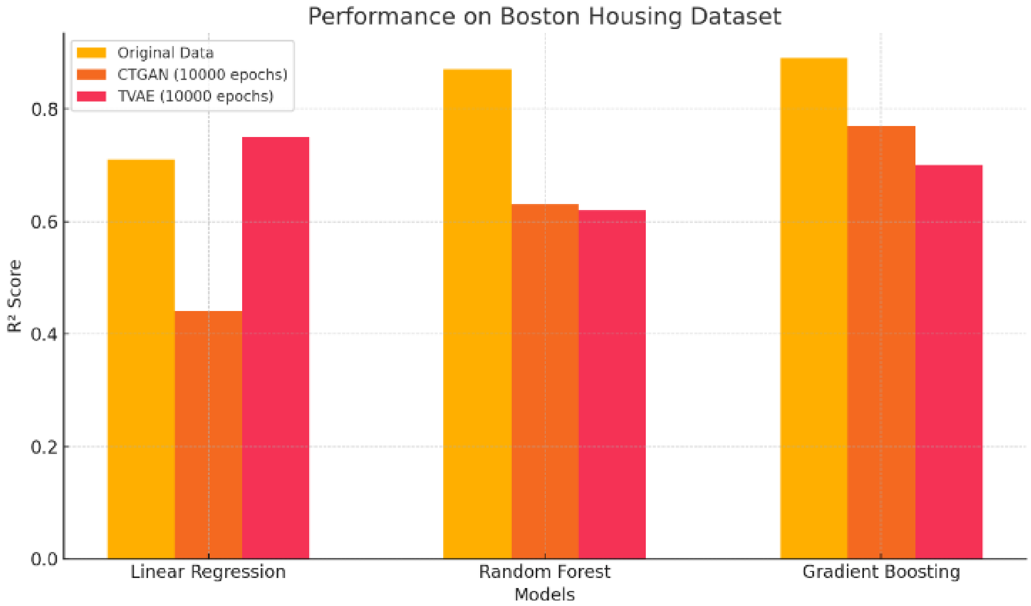

Figure 2 – Compare the performance of machine learning models on source and synthetic data for the Boston Housing dataset

4. Discussion

The results of this study highlight the importance of synthetic data for regression tasks and demonstrate the potential of CTGAN and TVAE methods in reproducing complex relationships between traits. However, despite the successes achieved, the productivity gap between the raw and synthetic data indicates the need for further optimization and adaptation of generation methods [16].

In future research, it is important not only to continue to improve existing methods, such as CTGAN and TVAE, but also to consider applying other approaches, including hybrid models or algorithms adapted to specific data types. Expanding the work to classification problems seems to be a promising direction, since it will allow us to study how synthetic data can be applied to more complex scenarios and tasks other than regression [17], [18]. The addition of new datasets with diverse structures and feature types will also contribute to a deeper understanding of the capabilities and limitations of synthetic data generation methods.

In addition, examining the impact of various characteristics of the source data, such as class imbalances, sample size, and dependency complexity, can help develop methods that better account for these aspects. Thus, expanding the range of tasks and methods will not only increase the applicability of synthetic data in machine learning, but also improve their quality, making them a more reliable tool for real-world scenarios [19], [20].

5. Conclusion

In the course of this study, the use of synthetic data generated by various methods for regression tasks was evaluated. The results demonstrated that synthetic data have the potential to overcome the limitations of real data, especially in cases where access to real data is difficult due to confidentiality concerns or insufficient data [21]. Synthetic data has proven to be a promising tool for expanding datasets and improving machine learning models. However, important quality issues remain, including the potential for loss of key relationships and the introduction of random noise. These aspects require further optimization of generation methods and careful verification of the data obtained before using it [23].

As such, synthetic data can be a reliable solution for applications where privacy and scalability are needed, but further work is needed to improve its quality and versatility.

Author Contributions

Conceptualization, A.D. and Z.A.; Methodology, A.D.; Software, A.D.; Validation, A.D.; Formal Analysis, Z.A.; Investigation, A.D.; Resources, A.D.; Data Curation, A.D.; Writing – Original Draft Preparation, A.D.; Writing – Review & Editing, A.D.; Visualization, A.D.; Supervision, A.D. and Z.A.; Project Administration, A.D., Z.A.

Conflicts of Interest

The authors declare no conflict of interest

References

1. L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling Tabular Data using Conditional GAN,” *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: https://arxiv.org/pdf/1907.00503.

2. N. Patki, R. Wedge, and K. Veeramachaneni, “The Synthetic Data Vault,” in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399–410. doi: 10.1109/DSAA.2016.49.

3. M. Miletic and M. Sariyar, “Challenges of using synthetic data generation methods for tabular microdata,” *Applied Sciences*, vol. 14, p. 5975, 2024. doi: 10.3390/app14145975.

4. K. A. Pareek, D. May, P. Meszmer, M. A. Ras, and B. Wunderle, “Synthetic data generation using finite element method to pre-train an image segmentation model for defect detection using infrared thermography,” *Journal of Intelligent Manufacturing*, 2024. doi: 10.1007/s10845-024-02326-1.

5. S. Kwatra and V. Torra, “Empirical evaluation of synthetic data created by generative models via attribute inference attack,” in *Privacy and Identity 2023, IFIP AICT 695*, F. Bieker *et al.*, Eds. 2024, pp. 282–291. doi: 10.1007/978-3-031-57978-3_18.

6. D. Harrison and D. L. Rubinfeld, “Hedonic housing prices and the demand for clean air,” *Journal of Environmental Economics and Management*, vol. 5, no. 1, pp. 81–102, 1978. doi: 10.1016/0095-0696(78)90006-2.

7. S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, “How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study,” *International Journal of Production Economics*, vol. 165, pp. 234–246, 2015. doi: 10.1016/j.ijpe.2014.12.031.

8. S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, “Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, pp. 7327–7347, 2022.

9. G. O. Ghosheh, J. Li, and T. Zhu, “A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records,” *ACM Comput. Surv.*, vol. 56, pp. 1–34, 2024.

10. M. Boeckhout, G. A. Zielhuis, and A. L. Bredenoord, “The FAIR guiding principles for data stewardship: Fair enough?” *Eur. J. Hum. Genet.*, vol. 26, pp. 931–936, 2018.

11. Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, and K. El Emam, “Can synthetic data be a proxy for real clinical trial data? A validation study,” *BMJ Open*, vol. 11, no. 4, e043497, 2021.

12. M. Templ and M. Sariyar, “A systematic overview on methods to protect sensitive data provided for various analyses,” *Int. J. Inf. Secur.*, vol. 21, pp. 1233–1246, 2022.

13. V. C. Pezoulas *et al.*, “Synthetic data generation methods in healthcare: A review on open-source tools and methods,” *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2910, 2024. doi: 10.1016/j.csbj.2024.07.005.

14. R. T. Q. Chen, J. Behrmann, D. K. Duvenaud, and J. H. Jacobsen, “Residual Flows for Invertible Generative Modeling,” *Advances in Neural Information Processing Systems*, vol. 32, 2019. [Online]. Available: https://arxiv.org/pdf/1906.02735.

15. S. James, C. Harbron, J. Branson, and M. Sundler, “Synthetic data use: Exploring use cases to optimise data utility,” *Discover Artificial Intelligence*, vol. 1, p. 15, 2021. doi: 10.1007/s44163-021-00016-y.

16. W. Wang, L. Ying, and J. Zhang, “On the Relation Between Identifiability, Differential Privacy, and Mutual-Information Privacy,” *IEEE Trans. Inf. Theory*, vol. 62, pp. 5018–5029, 2016.

17. Sci2s Research Group, “Big Data,” [Online]. Available: https://sci2s.ugr.es/BigData. Accessed: Aug. 26, 2024.

18. ResearchGate, “Advances in MapReduce Big Data Processing Platform: Tools and Algorithms,” [Online]. Available: https://www.researchgate.net/publication/349242793_Advances_in_MapReduce_Big_Data_Processing_Platform_Tools_and_Algorithms. Accessed: Aug. 26, 2024.

19. C. Garrido and A. Morales, “A Review of Big Data Integration Platforms for Scalable Analytics,” in *Big Data and High Performance Computing*, Springer, 2024, pp. 19–32. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-21534-6_2. Accessed: Aug. 26, 2024.

20. E. Navarro and J. Ortiz, “Advances in Big Data Analytics: Challenges and Solutions,” in *Big Data and High Performance Computing*, Springer, 2024, pp. 121–136. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-21534-6_9. Accessed: Aug. 26, 2024.

21. CPRD.com, “Synthetic data,” [Online]. Available: https://www.cprd.com/content/synthetic-data. Accessed: Oct. 4, 2021.

22. B. A. Malin, K. E. Emam, and C. M. O’Keefe, “Biomedical data privacy: Problems, perspectives, and recent advances,” *J. Am. Med. Inform. Assoc.*, vol. 20, pp. 2–6, 2013.

Information About Authors:

Dildabek Aizat is a Master of the Faculty of Artificial Intelligence and Big Date at al-Farabi Kazakh National University (Almaty, Kazakhstan, aizat.dildabek@gmail.com),

Abdiakhmetova Zukhra Muratovna is Deputy Dean for Educational, Methodical and Educational Work, Senior lecturer, PhD (Almaty, Kazakhstan, zukhra.abdiakhmetova@gmail.com).