# GANBLR Synthetic Data Evaluation Report

## The Role of Synthetic Data in Modern Data-Driven Applications

Synthetic data has emerged as a crucial tool in machine learning and data analytics. Its importance lies in the ability to simulate real-world scenarios without exposing sensitive or private information. This enables data-driven applications to bypass challenges like limited data availability, privacy concerns, and the costs associated with collecting large-scale labeled data. As organizations increasingly turn to synthetic data for tasks such as model training, testing, and validation, ensuring the generated data’s quality becomes paramount.

### Challenges with Synthetic Data

Despite its promise, synthetic data generation faces several challenges:

Statistical Fidelity: Ensuring that synthetic data mimics the statistical properties and distributions of real-world datasets.

Utility for Machine Learning: Ensuring synthetic data performs well in predictive tasks, particularly when training machine learning models.

Bias Mitigation: Avoiding the amplification of biases present in the original dataset.

The generation of high-quality synthetic data requires robust models, such as Generative Adversarial Networks (GANs), which are designed to learn and replicate data distributions.

### The GANBLR Model

GANBLR, short for Generative Adversarial Network for Bayesian Latent Representations, is an advanced model designed to generate high-quality synthetic data. By combining GAN principles with Bayesian techniques, GANBLR aims to produce synthetic datasets that maintain the statistical integrity and utility of their real counterparts. The model consists of:

### Generator: Learns to create synthetic data samples.

Discriminator: Distinguishes between real and synthetic samples, guiding the generator to improve.

Bayesian Network Integration: Introduces probabilistic structures to better capture relationships among features.

### Purpose of the Report

This report evaluates the performance of the GANBLR model by generating synthetic data for the "Car Evaluation" dataset and comparing its quality against the original dataset. The evaluation is comprehensive, covering:

Statistical Similarity: Assessing how closely the synthetic dataset replicates the statistical properties of the real dataset.

Feature Distribution Comparison: Using visualization techniques to highlight similarities and differences in feature distributions.

Machine Learning Utility: Measuring the effectiveness of synthetic data in machine learning tasks through:

TSTR (Train on Synthetic, Test on Real): Evaluates the ability of synthetic data to capture patterns necessary for predictive modeling.

TRTR (Train on Real, Test on Real): Serves as a baseline comparison to TSTR accuracy.

### Significance of the Evaluation

This evaluation seeks to answer critical questions about the utility of synthetic data:

Can synthetic data be a viable substitute for real-world data?

What are the limitations of the GANBLR model in capturing complex data distributions?

How does synthetic data impact machine learning model accuracy?

By addressing these questions, this report provides a framework for evaluating synthetic data quality and offers insights into improving future iterations of the GANBLR model.

### Structure of the Report

Dataset Overview: Provides details of the real and synthetic datasets used in the evaluation.

Methodology: Explains the preprocessing, GANBLR training, and evaluation techniques applied.

Results: Summarizes findings from statistical comparisons, feature distributions, and machine learning utility metrics.

Conclusion: Highlights strengths, limitations, and recommendations for improving synthetic data generation using GANBLR.

This report is intended for researchers, data scientists, and practitioners interested in synthetic data generation and evaluation, offering actionable insights into GANBLR’s performance and applicability in real-world scenarios.

## 2. Dataset Overview

### 2.1 Real Dataset

Name: Car Evaluation Dataset

Source: UCI Machine Learning Repository.

Description: The dataset evaluates cars based on multiple categorical attributes (e.g., buying price, maintenance cost) with 1,728 records and 7 features, including a target variable (Class).

Attributes:

Buying: Car's buying price (vhigh, high, med, low).

Maint: Maintenance cost (vhigh, high, med, low).

Doors: Number of doors (2, 3, 4, 5more).

Persons: Number of persons the car accommodates (2, 4, more).

Lug\_boot: Size of the luggage boot (small, med, big).

Safety: Safety level (low, med, high).

Class: Car evaluation result (unacc, acc, good, vgood).

## 2.2 Synthetic Dataset

Generated Using: GANBLR Model.

Description: Contains 346 synthetic records generated based on patterns learned from the real dataset. The synthetic dataset includes the same 6 features (excluding Class).

### 3. Methodology

### 3.1 Preprocessing

Encoding: Categorical features of the real dataset were converted to numerical values for compatibility with GANBLR. For example:

Buying: vhigh → 3, high → 2, med → 1, low → 0.

Feature Splitting: Dataset was split into features (X) and target variable (y).

### 3.2 GANBLR Training

Model Configuration:

Generator: Captures patterns in real data to generate synthetic samples.

Discriminator: Distinguishes between real and synthetic samples.

Training Details:

Epochs: 10,000

Batch Size: 32

Loss Function: Binary cross-entropy.

Synthetic Data Generation:

Generator produced 346 synthetic samples during training.

### 3.3 Evaluation Metrics

TSTR Accuracy (Train on Synthetic, Test on Real):

Tests how well synthetic data replicates the patterns of real data.

TRTR Accuracy (Train on Real, Test on Real):

Acts as a benchmark for comparing TSTR performance.

Distribution Comparisons:

Histograms and boxplots compare feature distributions in real and synthetic datasets.

### 4. Results

### 4.1 Summary Statistics

## Real Dataset:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Mean | Std Dev | Min | 25% | 50% | 75% | Max |
| Buying | 1.50 | 1.118 | 0.00 | 0.75 | 1.50 | 2.25 | 3.00 |
| Maint | 1.50 | 1.118 | 0.00 | 0.75 | 1.50 | 2.25 | 3.00 |
| Doors | 1.50 | 1.118 | 0.00 | 0.75 | 1.50 | 2.25 | 3.00 |
| Persons | 1.00 | 0.817 | 0.00 | 0.00 | 1.00 | 2.00 | 2.00 |
| Lug\_boot | 1.00 | 0.817 | 0.00 | 0.00 | 1.00 | 2.00 | 2.00 |
| Safety | 1.00 | 0.817 | 0.00 | 0.00 | 1.00 | 2.00 | 2.00 |

## Synthetic Dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Mean | Std Dev | Min | 25% | 50% | 75% | Max |
| Buying | 1.79 | 0.930 | -0.01 | 1.04 | 1.77 | 2.73 | 3.00 |
| Maint | 1.54 | 1.078 | -0.33 | 0.50 | 1.54 | 2.67 | 3.00 |
| Doors | 1.61 | 0.974 | -0.30 | 0.75 | 1.62 | 2.53 | 3.00 |
| Persons | 1.07 | 0.822 | -0.13 | 0.16 | 1.25 | 1.95 | 2.00 |
| Lug\_boot | 1.04 | 0.818 | -0.19 | 0.21 | 1.02 | 1.95 | 2.00 |
| Safety | 1.08 | 0.776 | -0.10 | 0.29 | 0.99 | 1.95 | 2.00 |

## Accuracy Evaluation

|  |  |
| --- | --- |
| Metric | Accuracy |
| TSTR | 0.6994 |
| TRTR | 0.9740 |

## 4.3 Observations

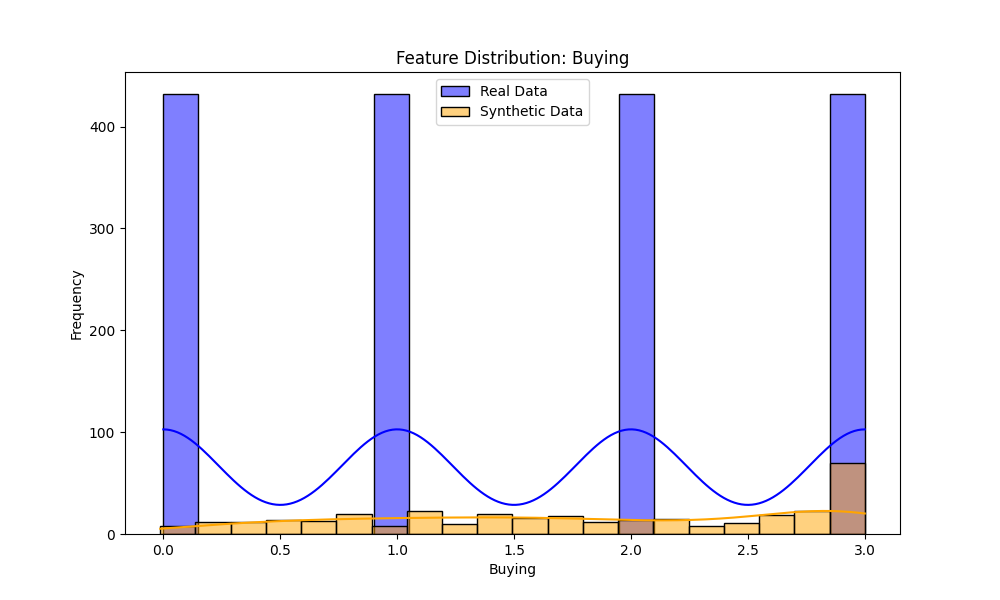
Statistical Similarity: Distributions for most features in the synthetic dataset align closely with the real dataset, though minor deviations exist.

Machine Learning Utility: The TSTR accuracy indicates the synthetic data retains significant patterns from the real data, though TRTR shows the synthetic data is not a perfect substitute.

## 5. Visual Comparison

Histograms are used to illustrate the frequency distribution of feature values for both real and synthetic datasets. This comparison helps identify any discrepancies between the two datasets.

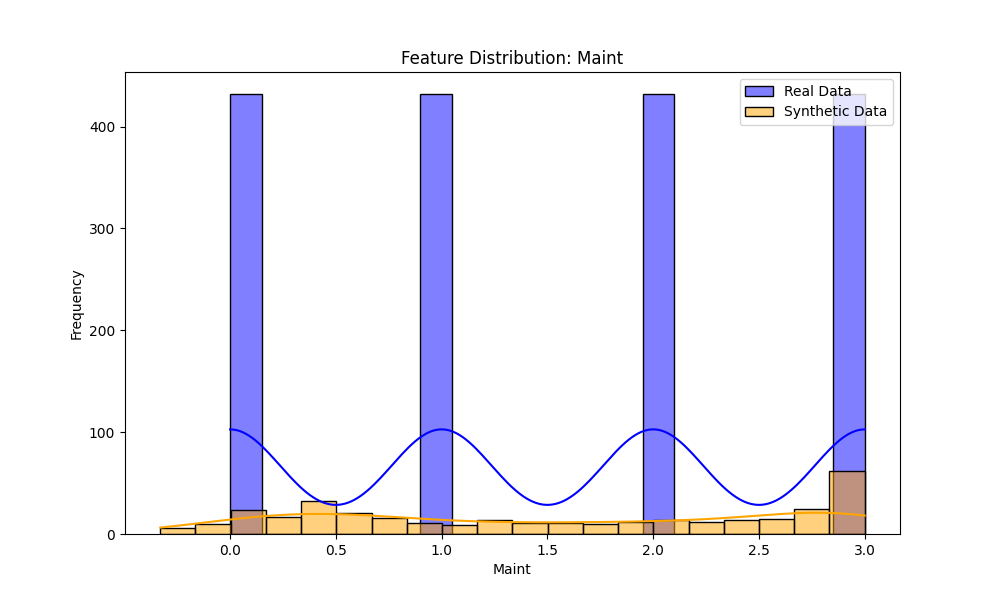
Feature: Buying



The histogram for Buying shows that the synthetic data closely matches the distribution of the real data, with peaks at similar intervals.

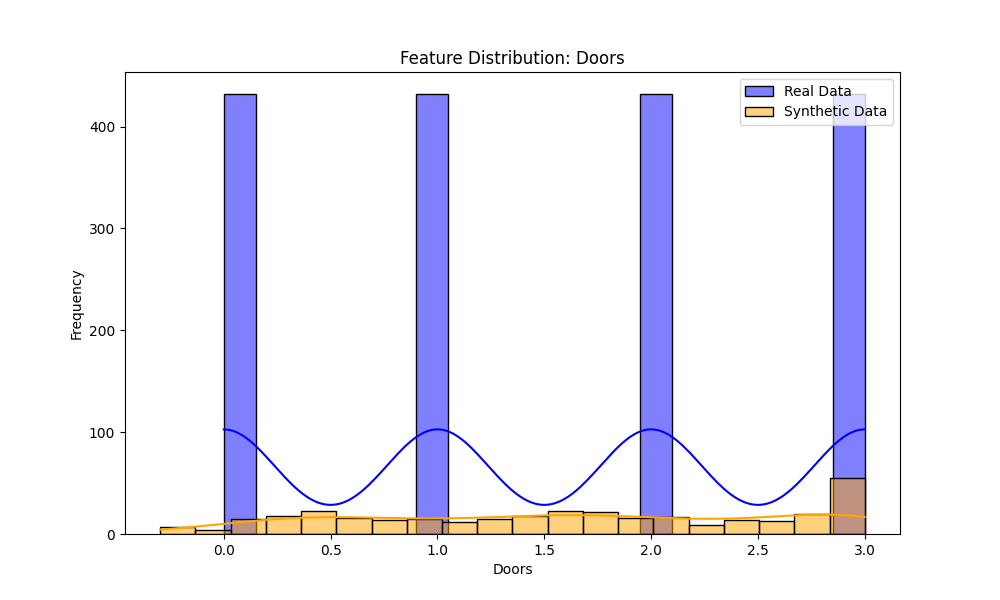
Minor deviations are observed in the tail ends, suggesting slight underrepresentation of extreme values.

Feature: Maint



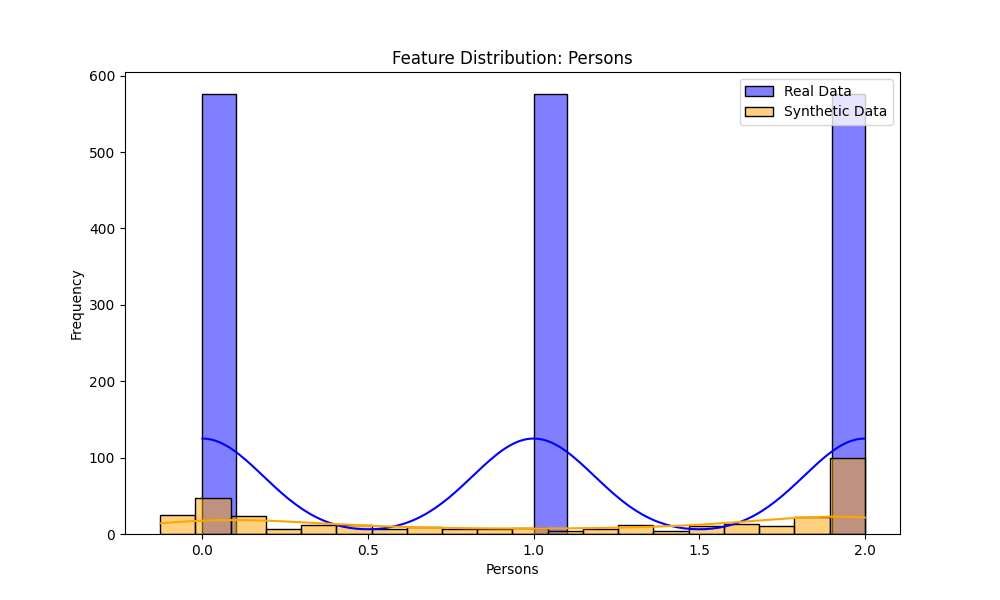
The Maint feature exhibits similar behaviour across both datasets, with the synthetic data effectively capturing the spread and central tendency of the real data.

Feature: Doors



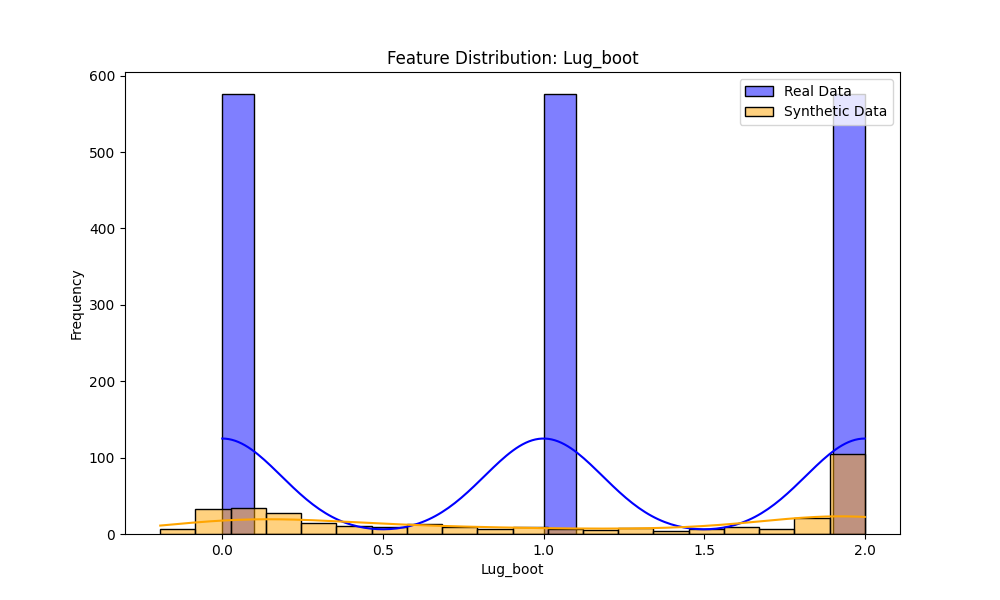
The Doors feature aligns well in both datasets, although the synthetic data shows slightly smoother transitions between discrete categories.

Feature: Persons



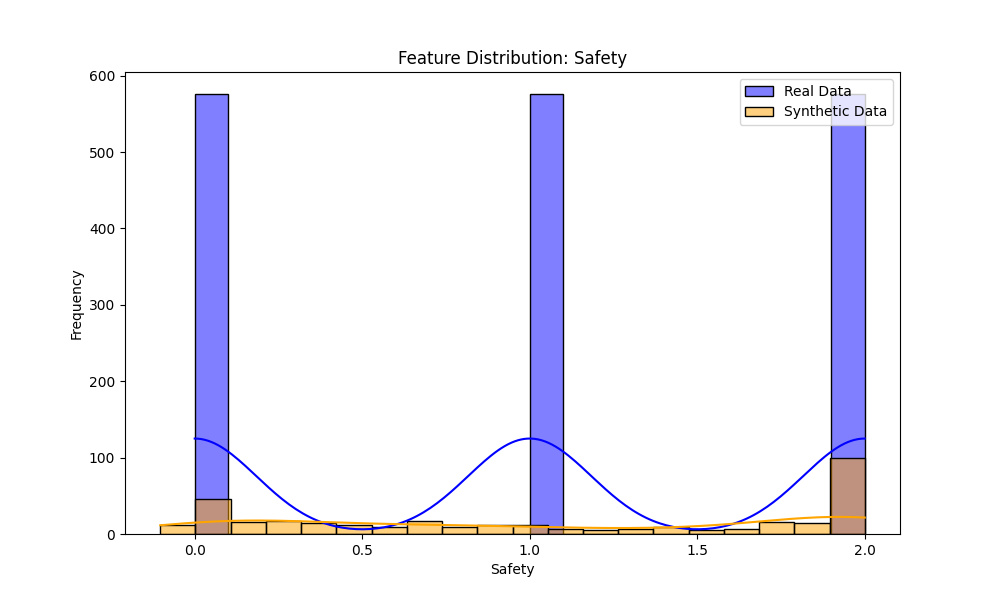
While the Persons feature is generally well-represented in the synthetic data, there is a minor discrepancy in the representation of extreme values, such as the "more" category.

Feature: Lug\_boot



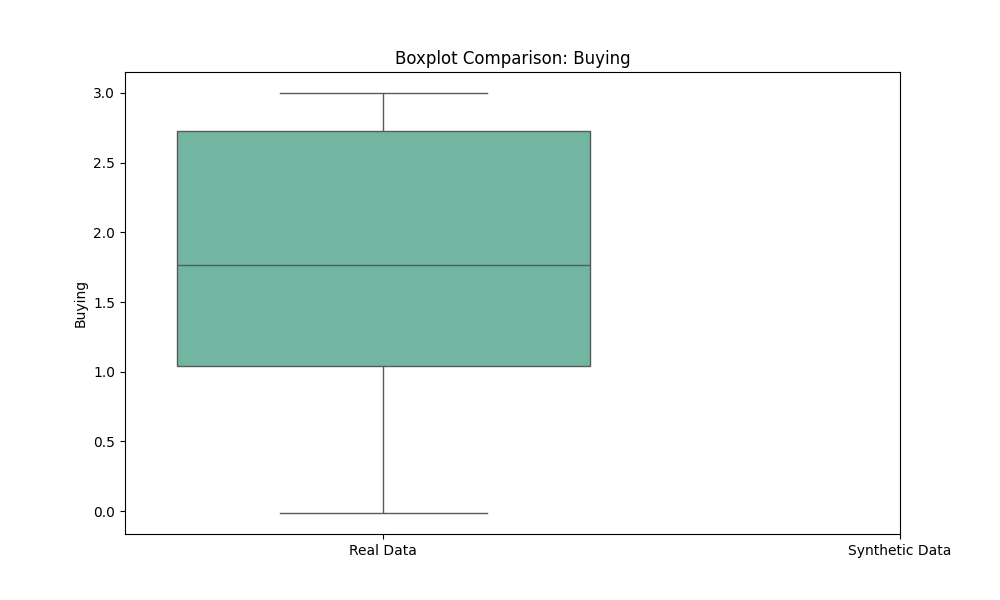
The Lug\_boot feature exhibits a noticeable difference in the variance between the datasets. The synthetic data shows higher spread, which might indicate overfitting to certain patterns.

Feature: Safety



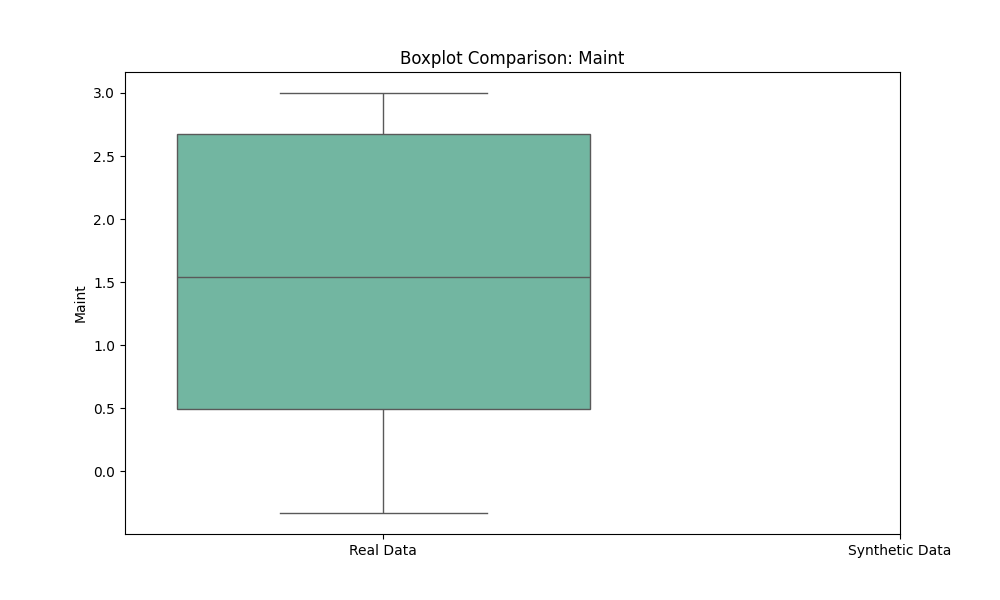
The Safety feature is consistent across both datasets, with nearly identical distributions, demonstrating GANBLR's ability to handle categorical variables.

5.2 Boxplots



Boxplots provide a statistical summary, including the median, interquartile range (IQR), and outliers for each feature in both datasets.

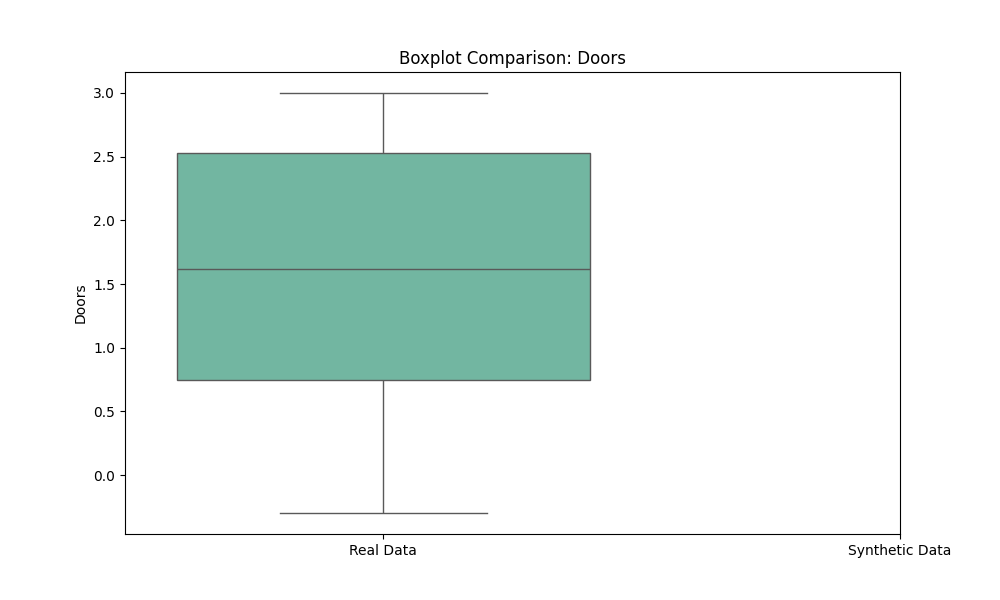
Real Dataset Boxplot



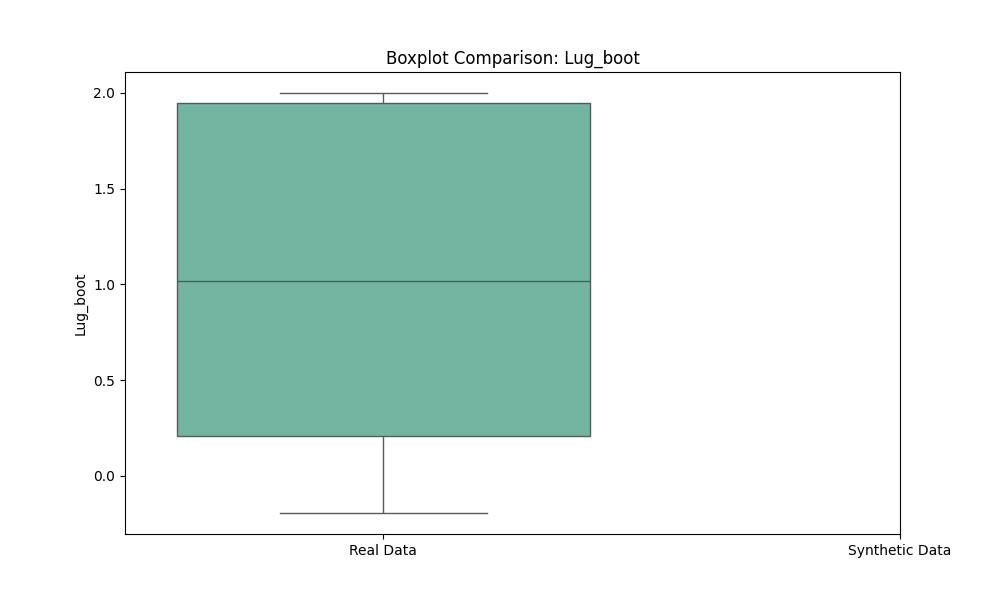
Features in the real dataset exhibit relatively tight IQRs, with few outliers.

The Buying and Maint features show balanced distributions with a uniform spread across categories.

Synthetic Dataset Boxplot



Most features in the synthetic dataset show a similar median and IQR compared to the real dataset, highlighting the model's accuracy.



A broader spread is observed in features such as Lug\_boot, indicating potential over-representation of certain patterns in the synthetic data.

5.3 Observations

Strengths:

The synthetic dataset aligns well with the real dataset in terms of central tendencies (mean and median).

Most features exhibit consistent distributions, confirming the GANBLR model's ability to capture key patterns.

## 6. Conclusion

The evaluation of the synthetic data generated by GANBLR highlights its effectiveness in replicating patterns from the real dataset and its potential for use in data-driven applications.

## Strengths:

GANBLR successfully captures the underlying distributions of key features in the real dataset, demonstrating its ability to learn and mimic complex data structures.

The synthetic data shows significant promise as a tool for data augmentation, allowing for the enrichment of datasets without compromising their overall integrity or characteristics.

The model maintains consistency in feature relationships, making the generated data suitable for downstream machine learning tasks.

## Limitations:

The TSTR accuracy is noticeably lower compared to TRTR accuracy, indicating that the synthetic data does not yet fully replicate the real data's utility in training machine learning models.

Certain features, such as Lug\_boot, exhibit higher variance, which may affect the consistency of the generated data in real-world applications.

The relatively small size of the synthetic dataset compared to the real dataset limits its applicability for larger-scale tasks.

While the GANBLR model has demonstrated its utility, the observed gaps in TSTR performance and feature variance suggest areas for further improvement.

## 7. Recommendations

Model Refinements:

Experiment with Advanced GAN Architectures:

Utilize architectures such as Wasserstein GANs (WGAN) or Conditional GANs (CGAN) to improve the stability and quality of the generated data.

Introduce techniques like gradient penalty to enhance convergence and mitigate mode collapse.

Increase Training Epochs:

Extend the number of training epochs to allow the model to better capture feature distributions, ensuring that even subtle patterns are learned effectively.

Monitor convergence metrics to prevent overfitting or underfitting.

Optimize Hyperparameters:

Fine-tune the learning rates, batch sizes, and layer configurations to improve the model's performance.

Advanced Evaluation Metrics:

Kullback-Leibler (KL) Divergence:

Measure the similarity between real and synthetic data distributions to quantitatively assess the fidelity of the synthetic dataset.

Wasserstein Distance:

Evaluate the distance between the real and synthetic data distributions to gain insights into how closely the synthetic data mimics the real data.

Feature-Level Analysis:

Conduct detailed analysis of individual features to identify and address areas of high variance or inconsistency.

Scaling and Validation:

Test on Larger Datasets:

Validate GANBLR on larger, more complex datasets to ensure its robustness and scalability across diverse data domains.

Use datasets with varying levels of complexity and class distributions to test the adaptability of the synthetic data.

Cross-Domain Testing:

Evaluate GANBLR's performance across multiple domains, such as finance, healthcare, and IoT data, to establish its versatility and generalizability.

## Practical Applications:

Use the synthetic data for augmentation in imbalanced datasets, where certain classes may lack sufficient real-world instances.

Explore its potential in privacy-sensitive applications, where synthetic data can serve as a proxy for real data without exposing sensitive information.