Synthetic Tabular Data with GAN and Transformers:

The experiment aims to determine which model is more effective in generating synthetic tabular data that is both high in quality and useful for practical applications, while also maintaining data privacy and addressing data availability issues.

The problem of generating high-quality synthetic tabular data is important for several reasons:

1. Data Privacy and Security : Synthetic data allows organizations to use data that mimics real-world datasets without exposing sensitive or confidential information. This is crucial in sectors like healthcare and finance, where data privacy regulations are strict[1][5].

2. Data Availability: Many industries face challenges with limited data availability due to privacy concerns or the rarity of certain data types. Synthetic data can fill these gaps by providing large amounts of data that would otherwise be difficult or impossible to collect[3][4].

3. Cost and Efficiency: Traditional data collection is often expensive and time-consuming. Synthetic data generation reduces these costs and accelerates development workflows by providing readily available datasets for analysis and machine learning model training[1][3].

4. Quality and Diversity: Synthetic data can be tailored to meet specific quality and format requirements, ensuring that datasets are diverse and representative. This is particularly beneficial for creating balanced datasets in machine learning applications, which can improve model performance[1][4].

5. Innovation and Testing: By using synthetic data, companies can test new ideas and develop products without risking personal data. This allows for more experimentation and innovation in data-driven projects[5].

Dataset :
This dataset captures comprehensive metrics and demographics related to player behavior in online gaming environments. It includes variables such as player demographics, game-specific details, engagement metrics, and a target variable reflecting player retention.
Analyzing outliers, null values, and correlations in synthetic tabular data is crucial for ensuring its quality and utility. Outliers can significantly skew analyses and model predictions, so identifying and understanding them helps maintain data integrity.
Checking for null values is essential because missing data can lead to biased results or reduced model performance. Proper handling, such as imputation or removal, ensures that the dataset remains robust and reliable.
Correlation analysis, both for numerical and categorical variables, helps preserve the relationships present in the original data. This is vital for synthetic data, as it must accurately reflect these relationships to be useful in tasks like predictive modeling or statistical analysis.
By comparing these aspects between the original and synthetic datasets, you can assess how well the synthetic data mimics the real data's structure and behavior. This ensures that the synthetic data is not only a privacy-preserving alternative but also a practical substitute for real-world applications.

Data Preparation :

1. Data Division:
- Divided the original dataset into training and testing sets. This is crucial for evaluating the performance of your synthetic data generation models.

2. Synthetic Data Generation:
- CTGAN: A GAN-based model specifically designed for generating tabular data. It captures the complex relationships between features.
- REaLTabFormer: A transformer-based model that can also generate synthetic tabular data, potentially offering advantages in terms of capturing long-range dependencies and complex feature interactions.

3. Blending Datasets:
- Created a new dataset by blending the synthetic datasets generated by CTGAN and REaLTabFormer. This approach can help leverage the strengths of both models, potentially resulting in a more robust synthetic dataset.

Exploratory Data Analysis (EDA) :

1. Descriptive Statistics:
- Mean, Median, Mode: Calculating these measures helps understand the central tendency of the data. Comparing these statistics between the original and synthetic datasets can provide insights into how well the synthetic data mimics the original data.

2. Histograms:
- Visualizing the distribution of each feature using histograms can help identify differences in distribution between the original and synthetic datasets. Looking for similarities and differences in shape, spread, and skewness.

3. Categorical variable distribution :
- Used Bar plots to verify the distribution of the categorical data.

Examining the frequency distributions and relative proportions of each category to see if they align with the original data.

4. Quantile distribution :
- This helps in assessing whether the synthetic data preserves the distributional properties of the original data.

Evaluating for ML Downstream Task Classification:

1. Model Evaluation:

By comparing the model's results on synthetic versus real data, We can gauge the effectiveness and realism of the synthetic data. Additionally, this evaluation will provide insights into the generalization ability of the models when applied to actual data, highlighting potential gaps or overfitting issues.

2. Feature Importance:

To further understand the quality of the synthetic data, it's crucial to analyze feature importance and how well the synthetic features replicate the relationships found in the original data. Additionally, consider employing advanced metrics tailored to synthetic data evaluation, such as distributional similarity measures, to ensure the synthetic data preserves not only the individual feature characteristics but also the intricate interdependencies among features.

Conclusion :

The REaLTabFormer model performed better in generating synthetic tabular data compared to CTGAN. This conclusion is supported by the superior performance of REaLTabFormer across various utility measures in experiments, as highlighted in the research findings.
REaLTabFormer's ability to capture relational structures more effectively than baseline models contributes to its enhanced performance. This is likely due to the transformer architecture's strength in modeling complex relationships and dependencies within data, which is crucial for generating realistic synthetic datasets. In contrast, while CTGAN is a robust GAN-based model for synthetic data generation, it may not capture these intricate relationships as effectively as REaLTabFormer

Future work :

The world is growing rapidly and leaving graces in the form of data. Currently, 80% data is in the form of text and images. For future work, using these models to generate synthetic data of text or image format can be an interesting experiment.

Citations:

[1] https://syntheticus.ai/blog/the-benefits-and-limitations-of-generating-synthetic-data
[2] https://developer.nvidia.com/blog/generating-synthetic-data-with-transformers-a-solution-for-enterprise-data-challenges/
[3] https://syntheticus.ai/guide-everything-you-need-to-know-about-synthetic-data
[4] https://neptune.ai/blog/the-advantages-of-synthetic-data-over-real-data
[5] https://mitsloan.mit.edu/ideas-made-to-matter/what-synthetic-data-and-how-can-it-help-you-competitively
[6] https://github.com/worldbank/REaLTabFormer?tab=readme-ov-file
[7] https://github.com/sdv-dev/CTGAN
[8] How to evaluate the quality of the synthetic data – measuring from the perspective of fidelity, utility, and privacy | AWS Machine Learning Blog (amazon.com)
[9] https://mostly.ai/what-is-synthetic-data
[10] https://research.ibm.com/blog/what-is-synthetic-data