Evaluation of fidelity, utility, and privacy aspects of synthetic data sets

Gabriel Sperrer*
Vienna University of Technology
gabriel.sperrer@student.tuwien.ac.at

Maximilian Kleinegger*
Vienna University of Technology
maximilian.kleinegger@student.tuwien.ac.at

*All authors contributed equally to this research.

1 Introduction

This report aims to provide insights into this project. It provides all the necessary information about the assignment, the datasets and methods used, and the results that could be obtained. Simply put, it presents all the challenges and results we had to conquer to obtain synthesized data. Due to the open-ended exercise description and the absence of specific limits, we decided to compare various data generation techniques. This approach enables us to thoroughly evaluate each technique's potential for a given dataset.

2 Overview

Our goal is to utilize a synthetic data generation tool to create synthetic tabular data based on a dataset containing sensitive data. Afterward, the synthetic data's fidelity, utility, and privacy aspects should be evaluated. However, already existing tools should be leveraged and not implemented in any way. Therefore, the assignment primarily builds upon existing libraries to generate and analyze the synthetic data.

The dataset that we're inspecting is based on COMPAS. COMPAS is a commercial risk-assessment tool built by Northpointe and is meant to assist US judges in court in their decision-making by rating the likelihood of a criminal offender becoming a recidivist, a term describing a criminal repeating an undesirable behavior. It received international attention after a detailed analysis by ProPublica in 2016 claimed that the algorithm contained significant bias against black offenders [4].

For this exercise's purpose, we use the same data that ProPublica used in their analysis, which is openly available on GitHub[1]. They received the data through a public request to Broward County in Florida[3].

3 Methodology and Approach

This section explains the dataset used and the approach used to gain the results of this experiment.

3.1 The COMPAS dataset

We use the raw data as given in (*compas-scores-raw.csv*). It consists of the following features: *Person_ID, AssessmentID, Case_ID, Agency_Text, LastName, FirstName, MiddleName, Sex_Code_Text, Ethnic_Code_Text, DateOfBirth, ScaleSet_ID, ScaleSet, AssessmentReason, Language, LegalStatus, CustodyStatus, MaritalStatus, Screening_Date, Scale_ID, DisplayText, RecSupervisionLevel, RecSupervisionLevelText* and the target variables: *RawScore, DecileScore, ScoreText, AssessmentType, IsCompleted, IsDeleted.*

The dataset contains three scores that are based on different scales for each person, described by the identifier of the scale *Scale_ID* and the corresponding *DisplayText* as a textual description of the scale. For our inspection, we decided to use the "Risk of Recidivism" (ID 18).

Since the exercise aims to generate synthetic data that could be used in place of the

[1] https://github.com/propublica/compas-analysis/ (Accessed at 04.05.2024)

original dataset without privacy risk, we've started by pre-processing the dataset. First, we removed all directly identifying features that are not relevant to our experiment:

- The identifiers: *Person_ID, AssessmentID, Case_ID*
- The name: *FirstName, MiddleName, LastName*

Second, we reduced the target values to the *ScoreText* since it's the one that was also used by ProPublica and, as they claimed, the one that influences the decision-making by judges most.

The remaining columns that we use for our evaluation are:

- *Agency_Text*, describing the context where people were assessed by COMPAS
- *Sex_Code_Text*, gender
- *Ethnic_Code_Text*, describing the ethnicity of the offender
- *DateOfBirth*, age
- *Language*, either "English" or "Spanish"
- *LegalStatus*, the procedural status of an individual
- *CustodyStatus*, the custodial arrangement
- *MaritalStatus*, marital status
- *Screening_Date*, the date and time of the assessment
- *ScoreText*, a category describing the risk score of an offender (either "Low", "Medium", or "High")

For further preparation of the dataset, we cleaned the data to make it suitable for the generation of synthetic data and the evaluation of the usability of that data. For the date columns *DateOfBirth* and *Screening_Date*, there are multiple possible ways of encoding depending on the information you want to retain. One example could be potential cyclical correlations (there could be, for example, higher crime rates on Sundays). In this case, we found no such correlations. There is, however, a strong correlation in the data between the date of birth and the *DecileScore* ($r = 0.4473$) as seen in figure 1. This is the reason we encoded the dates as UNIX timestamps, which preserve this correlation.

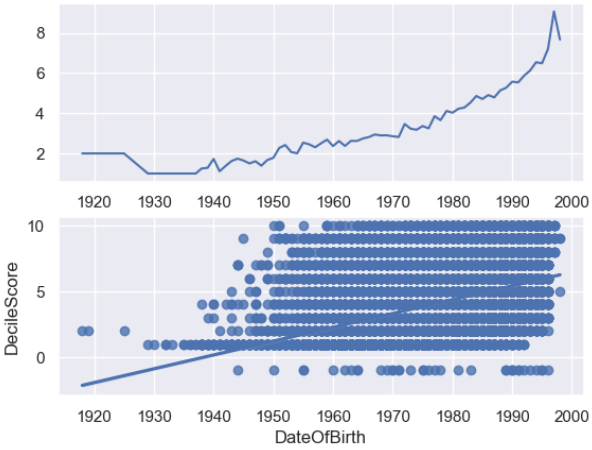

Figure 1: *DateOfBirth* versus *DecileScore*

The column *Ethnic_Code_Text* contained both *African-Am* and *African-American*, so we unified those into *African-American*. As a last step, we removed *NaN* values in the *ScoreText* column.

The resulting data that we use for our inspection consists of 14 columns, of which all except the two date columns were of categorical type. For simplicity, we converted the *Sex_Code_Text* to a binary representation. All further columns are kept as they were for the generation of the synthetic data.

3.2 Experiment

This section describes the experiment. First, we use a RandomForest with 250 trees and a max-depth of 15 as a classifier for evaluation. Secondly, we used six Data synthesizers: four from Synthetic Data Vault (SVD) [5] and two from DataSynthesizer[2]. With each synthesizer, we generated data and afterward evaluated the data with SDMetrics [1], a tool to evaluate synthetic data by providing metrics that measure the statistical difference between generated and original data. Furthermore, we trained our classifier with the data and compared the *accuracy, recall, precision*

[2] `https://github.com/DataResponsibly/DataSynthesizer` (Accessed at 04.05.2024)

and *f1-score* with a baseline evaluation, which is trained on the original data. It's important to note that all classifiers are evaluated on the original test data; therefore, no synthetic data is used here. Lastly, we evaluate privacy and inference risks using Anonymeter [2].

3.2.1 Synthesizers

In this section, we explain the data synthesizers used. SVD provides four synthesizers, mainly *GaussianCopulaSynthesizer, CTGANSynthesizer, TVAESynthesizer* and *CopulaGANSynthesizer* and DataSynthesizer provides a *Bayesian Network (BN)* which operates in an *independent* or *correlated* attribute mode.
SVD data synthesizers use different methods to train a model and afterward generate synthetic data. The different implementations use the following methods:

- *GaussianCopulaSynthesizer (GCS)* uses classic, statistical methods,
- *CTGANSynthesizer (CTGANS)* uses GAN-based, deep learning methods,
- *TVAESynthesizer (TVAES)* uses a variational autoencoder (VAE)-based, neural network techniques,
- *CopulaGANSynthesizer(CGANS)* uses a mix of classic, statistical methods, and GAN-based deep learning methods

All synthesizers use default parameters, except we enforce that min-max values of the original numeric distributions are not violated and mostly decrease the number of epochs to train a model.

DataSynthesizer alternatively learns a *BN* with two modes. The *independent (IBN)* attribute mode assumes complete independence between all attributes, when creating and training the network, whereas the *correlated (CBN)* attribute mode assumes non-independence during the creation and training of the network, which leads to different networks and therefore to different outcomes. Here we only used the necessary parameters like which mode we want to deploy and which attributes are categorical.

3.2.2 Metrics

To evaluate the generated synthetic data, we use various metrics to obtain comparable results for our experiment.

Synthetic data *fidelity* is an evaluation of how well the synthetic data captures the mathematical properties of the real data. To evaluate the fidelity of the synthetic data that we generate through the various methods, we rely on the methods provided by *SDMetrics*. In more detail, we calculate the following metrics, which should cover all statistical aspects of our data differences:

- *Qualitity report*, which contains the average percentage of how similar are the column shapes and how similar are the column pair trends. From 0.0 to 1.0, where 1.0 means that they represent the original data perfectly and 0.0 the opposite.
- *CSTest* computes if the synthetic data is significantly different than the original data concerning column shapes. From 0.0 to 1.0, where 1.0 means it is not significantly different and 0.0 it is.
- *DiscreteKLDivergence, ContinuousKLDivergence* computes the respective KL-Divergence for each compatible pair of columns and then calculates the average across all these pairs. From 0.0 to 1.0, where 1.0 means that both datasets diverge between two probability distributions, whereas 0.0 means they do not.
- *SVCDetection, LogisticDetection* calculate how difficult it is to tell apart the real data from the synthetic data. From 0.0 to 1.0, where 1.0 means that the data cannot be differentiated and 0.0 the opposite. We use both metrics to have a more reliable source because both should normally show the same trend.

To compare *utility* we used well-known metrics to compare the effectiveness of machine learning models. In our case, we rely on *accuracy, recall, precision* and *f1-score* for multi-class classification to compare the outcomes of our RandomForests.

Lastly, to evaluate *privacy* we used following attack based metrics provided by Anonymeter [2]:

- *Privacy risks* based on how well somebody could be singled out through the synthetic dataset based on the original one. We calculate it univariate and multivariate, with the maximum number of columns available. In this case the lower the score the better.

- *Inference risks* calculates the chance of guessing private details from a dataset, showing how vulnerable it is to privacy breaches and if it's secure from unauthorized access. In this case the lower the score the better.

Important to note is, that we use the test set as a control group, which should suffice other than that we simply use all needed or default parameters. However, we decided against using *linkability* risks as a metric for one good reason: When performing linkability tests, we want to evaluate if you can link records from one dataset to another. Because we assume that the original data never gets released and we do not have any other datasets we can use for this, we do not need it.

4 Results

To achieve comparable results for all the synthetic datasets we generated, we used the same quality metrics across all methods and compared firstly concerning fidelity and utility and secondly concerning privacy. The following section highlights the most interesting findings.

For all data we generated, we ran the diagnostic report as recommended by *SDMetrics* to ensure that the data is generally valid. The scores were 100% for each dataset, so they are not shown individually for each synthesizer.

4.1 Fidelity and utility evaluation

The evaluation of fidelity and utility for the various synthetic datasets, for which the corresponding results can be seen in table 1 and 2 respectively, provides insightful distinctions across different synthesis methods. The *GaussianCopulaSynthesizer (GCS)* demonstrated superior fidelity with column shapes very close to the original data, achieving a shape quality of 99% and a trend quality of 70%, alongside high *CSTest* and *KL Divergence* scores, suggesting minimal statistical divergence from the original data. However, its utility lagged significantly behind with only 51.50% *accuracy* and a 37.50% *f1-score*.

Conversely, the *CopulaGANSynthesizer (CGANS)* and the *CTGANSynthesizer (CTGANS)* provided a balanced profile of both high fidelity and utility. *CGANS* and *CTGANS* produced competitive fidelity metrics, closely mirroring original data trends and distributions, while also achieving high utility scores, particularly in precision and f1-scores, nearly reaching the baseline's performance set by models trained on actual data.

DataSynthesizer, when using the correlated mode (*CBN*), yielded promising results with a quality report indicating strong adherence to the original data trends and an excellent utility performance with an *accuracy* of 88.89% and an *f1-score* of 86.74%, closely rivaling the baseline model. This suggests that *CBN*'s method of considering attribute correlations during synthesis contributes effectively to maintaining both the statistical properties and the predictive qualities of the data as seen in figure 2.

The variation in fidelity and utility outcomes highlights the influence of different synthetic data generation techniques. Moreover, also the difference in generating different data distributions is seen in figure 3. Techniques that effectively capture and reproduce the complex interdependencies between data attributes, such as *CGANS* and *CBN*, tend to offer synthetic data that is not only statistically similar to the original but also retains its utility in practical machine-learning scenarios.

Additionally, it should be mentioned that the Bayesian Network outperformed GAN-based deep learning methods in terms of training time. This aspect becomes particularly important as the size of the original data increases.

Method	Qualityreport		CSTest	KL Divergence		Detection	
	Shape	Trends		Continuous	Discrete	SVC	Logistic
GCS	0.99	0.70	0.99	0.99	0.94	0.25	1.0
CTGANS	0.93	0.66	0.96	0.95	0.91	0.40	0.63
TVAES	0.93	0.68	0.99	0.90	0.94	0.28	0.58
CGANS	0.94	0.66	0.96	0.96	0.90	0.38	0.58
IBN	0.97	0.67	0.98	0.97	0.86	0.12	0.74
CBN	0.97	0.70	0.90	0.91	0.84	0.59	0.83

Table 1: Fidelity metrics

Method	Accuracy	Precision	Recall	f1
Baseline	89.28%	90.38%	86.49%	87.33%
GCS	51.50%	35.55%	43.08%	37.50%
CTGANS	86.86%	87.54%	82.47%	84.14%
TVAES	59.50%	37.50%	60.29%	46.06 %
CGANS	86.05%	87.12%	81.22%	83.21%
IBN	55.00%	29.36%	32.74%	25.65%
CBN	88.89%	89.77%	86.10%	86.74%

Table 2: Utility metrics

Figure 2: Comparison of GCS (left) and CBN (right) Column Pair Trends

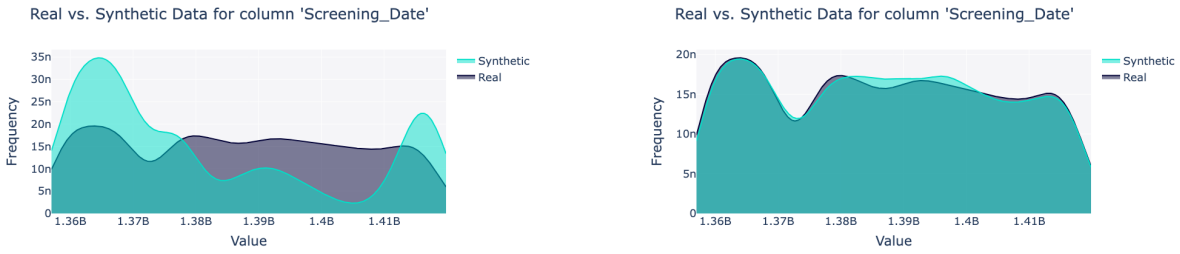

Figure 3: Comparison of TVAES (left) and CBN (right) data generated for *Screening_Date*. CBN does a great job mimicking the original distribution, while TVAES fails to generate accurate data.

4.2 Privacy evaluation

As you see in table 3 all data synthesizers perform a good job in generating data that secures privacy. If the attacker is only in possession of one attribute it is not possible to identify anything or anybody with our data. However, if the attacker has multiple columns available, in our case all of them, the pri-

vacy risks increases, but it is still very unlikely that a single out attack would be successful. All data synthesizers return a privacy risk below 0.02. Interesting to note is the excellent performance of the *CopulaGANSynthesizer*, which synthetic data does not allow singling out attacks.

Metric	GCS	CTGANS	TVAES	CGANS	IBN	CBN
Privacy Risks (univariate)	0.0	0.0	0.0	0.0	0.0	0.0
Privacy Risks (multivariate)	0.028	0.040	0.144	0.0	0.020	0.0164

Table 3: Privacy risks

Secondly, we examine the inference risks associated with the generated synthetic data. We compare the extent of inference risks present in different attributes from the dataset. As shown in Table 4, certain attributes pose higher inference risks compared to others, while some attributes pose no inference risks at all. Particularly, attributes such as *Sex_Code_Text, ScaleSet_ID, Language, LegalStatus, CustodyStatus*, and *RecSupervisionLevel* exhibit higher risks across multiple data synthesizer techniques. However, some techniques generate data distributions that are more susceptible to inference attacks. It is noteworthy that *Bayesian Networks*, regardless of the mode, performs the best, whereas the others perform significantly worse, with *TVAESynthesizer* being the least effective.

Attribute	GCS	CTGANS	TVAES	CGANS	IBN	CBN
Agency_Text	0.0	0.0	0.122	0.0	0.0	0.0
Sex_Code_Text	0.038	0.059	0.003	0.003	0.0	0.0
Ethnic_Code_Text	0.0	0.046	0.0	0.037	0.0	0.0
DateOfBirth	0.0	0.0	0.0	0.012	0.0	0.0
ScaleSet_ID	0.203	0.054	0.0	0.132	0.095	0.0
AssessmentReason	0.0	0.0	0.0	0.0	0.0	0.0
Language	0.0	0.0	0.338	0.193	0.0	0.0
LegalStatus	0.0	0.065	0.020	0.0	0.0	0.020
CustodyStatus	0.0	0.0	0.028	0.0	0.005	0.0
MaritalStatus	0.062	0.035	0.0	0.0	0.0	0.0
Screening_Date	0.0	0.0	0.0	0.0	0.0	0.0
RecSupervisionLevel	0.012	0.0	0.082	0.024	0.083	0.143
Scale_ID	0.0	0.0	0.0	0.0	0.0	0.0
ScoreText	0.073	0.0	0.047	0.0	0.0	0.0

Table 4: Inference risks

Furthermore, because including too many graphs and plots would exceed the scope of this report, we have only included the necessary results. However, we encourage everyone interested to take a closer look at all metric results through our plots in the Jupyter Notebook.

5 Conclusions

In conclusion, we demonstrated that careful selection and application of synthetic data generation techniques can produce synthetic datasets with high fidelity and utility in machine learning applications while maintaining strong privacy protection. This was only possible by trying different synthetization techniques and comparing the outcomes to the original data. The variations in the observed performance across synthesizers underscore the importance of matching the synthesis method to the specific characteristics of the datasets and the requirements of the task. Techniques that incorporate the un-

derlying data structure and interdependencies, such as CopulaGANSynthesizer and the correlated Bayesian Network mode of DataSynthesizer, were particularly effective for mimicking the COMPAS dataset.

References

[1] DataCebo, Inc. *Synthetic Data Metrics*, 05 2024.

[2] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data, 2023.

[3] Jeff Larson, Julia Angwin, Surya Mattu, and Lauren Kirchner. How we analyzed the compas recidivism algorithm. ProPublica, 2016.

[4] Jeff Larson, Julia Angwin, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, 2016.

[5] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016.