BUYER BEWARE: UNDERSTANDING THE TRADE-OFF BETWEEN UTILITY AND RISK IN CART BASED MODELS USING SIMULATION DATA

Berlin,
7-8. Oktober, 2024

Jonathan Latner, PhD
Dr. Marcel Neuenhoeffer
Prof. Dr. Jörg Drechsler

SECTION 1: INTRODUCTION

OVERVIEW

- It is well established that there is a trade-off between utility and privacy when generating synthetic data

- Utility in CART based synthesizers is high (Little et al., 2022; Danker and Ibrahim, 2021)

- Privacy in CART based synthesizers is also high (Little et al., 2022)

- It seemed that CART models are less sensitive to this trade-off than other SDGs (i.e. higher utility, lower risk)

- Using simulation data (Reiter et al., 2014), we show that synthetic data from CART models are disclosive

  – Disclosive in ways that are not observable with common privacy metrics

  – It is possible to increase protection (by reducing utility), but you have to choose to do so

  – More generally: If you did not know there was a problem, why would correct it?

WHATS THE GOAL OF SYNTHETIC DATA?

- Synthetic data can accelerate development by replacing sensitive values with synthetic ones with minimal distortion of the statistical information contained in the original data set. (Jordan et al., 2022; Nowak et al., 2016)

- Low disclosure risk (R)

- High data utility (U)

- Visualize the trade-off using the R-U confidentiality map (Duncan et al., 2004)

WHATS THE PROBLEM?

- High data utility – It must be similar to and different from the original data.
  - At the extreme, if the goal is high utility, why not just release the original data?
- Low disclosure risk – Synthetic data is not automatically private.
  - At the extreme, if the goal is low privacy risk, why should we release any data?
- Many measures of utility and privacy exist
  - Therefore, its not clear if data have high utility or low risk
  - 2 problems
    - More specifically, how can we map R-U trade-off if there are multiple measures of both?
    - More generally, how do we know if the data have high levels of utility and low levels of privacy?

WHAT DO WE KNOW?

- SDG with CART is a good option
- Utility
  - Drechsler and Reiter (2011) found that CART models offered the best results in terms of preserving the information from the original data.
  - Other comparisons also found CART is superior (Little et al., 2022; Danker and Ibrahim, 2021)
- Privacy
  - Some evidence also suggests CART is superior (Little et al., 2022)
  - However, other evidence indicates that CART-based synthesis simply replicates most of the original records (Manrique-Vallier and Hu, 2018)

WHAT DO WE NOT KNOW?

- How are CART models so good at minimizing trade-off between risk and utility?

SECTION 2: DATA AND METHODS

DATA AND METHODS

- Data - simulated data with unique record (Reiter et al., 2024)

- Utility measures (synthpop - Raab et al., 2021)

  - Voas Williamson

  - Freeman-Tukey

  - Jensen-Shannaon divergence

  - Kolmogorov-Smirnov statistic

  - Propensity score mean-squared error

  - Bhattacharyya distances

- Privacy measures (synthpop - Raab et al., 2024)

  - Identity disclosure measure

  - Attribute disclosure measure

  - Replicated uniques

SIMULATE DATA WITH A UNIQUE RECORD

Borrowing from Reiter et al. (2014), we set the first 999 observations to be a random sample from a multinomial distribution for all combinations of $var1(0, 1)$, $var2(0, 1)$, $var3(0, 1)$, $var4(0, 1)$ except (var1=1,var2=1,var3=1,var4=1), which we set to be the 1000th observation.
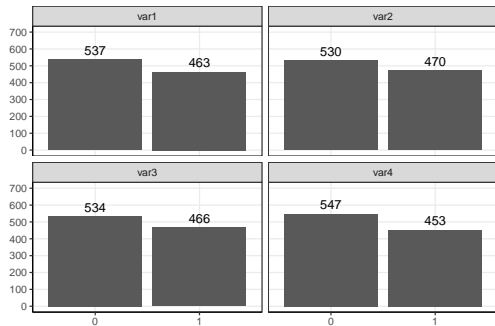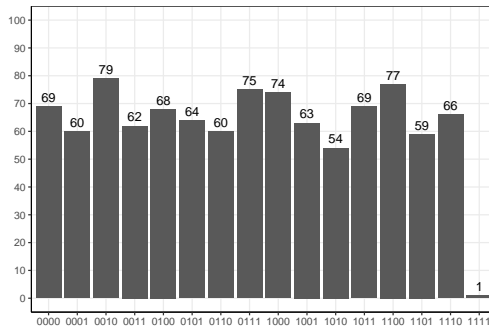

Figure 1: Frequency

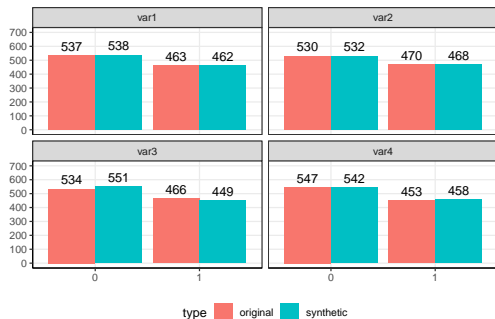

Figure 2: Histogram

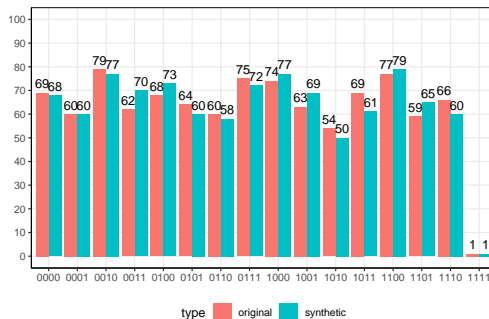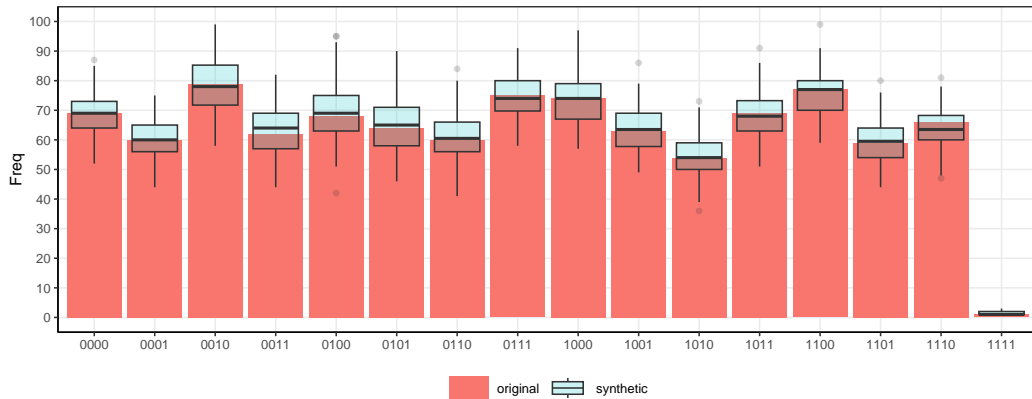# GENERATE SYNTHETIC DATA

Figure 3: Frequency

Figure 4: Histogram

COMPARE HISTOGRAM X 100 SYNTHETIC DATASETS

Figure 5

COMPARING UTILITY MEASURES

Table 1

name	var1	var2	var3	var4	average
Voas Williamson utility measure	0.00	0.02	1.16	0.10	0.32
Freeman-Tukey utility measure	0.00	0.02	1.16	0.10	0.32
Jensen-Shannaon divergence	0.00	0.00	0.00	0.00	0.00
Kolmogorov-Smirnov statistic	0.00	0.00	0.02	0.01	0.01
propensity score mean-squared error	0.00	0.00	0.00	0.00	0.00
Bhattacharyya distances	0.00	0.00	0.01	0.00	0.00

COMPARING PRIVACY MEASURES (SET.SEED = 1237, I.E. UNIQUE = 1)

```
1  > print(t1, plot = FALSE)
2  Disclosure risk for 1000 records in the original data
3
4  Identity disclosure measures
5  from keys: var1 var2 var3
6  For original  ( UiO )  0 %
7  For synthetic ( repU ) 0 %.
8
9  Table of attribute disclosure measures for var1 var2 var3
10 Original measure is  Dorig and synthetic measure is DiSCO
11 Variables Ordered by synthetic disclosure measure
12
13          attrib.orig attrib.syn check1 Npairs check2
14 1 var4          0         0             0
```

```
1  > replicated.uniques (sds, df_ods)
2      var1 var2 var3 var4
3  973    1    1    1    1
4  Uniques and replicated uniques for  1  synthesised data set(s)
5   from keys:  var1 var2 var3 var4
6
7  Uniques in  original data:
8  1 from  1000 records ( 0.1 %)
9  Uniques in synthetic data:
10  1 from  1000 records ( 0.1% )
11
12 Replicated uniques:
13  1
14 as a % of uniques in synthetic  100%
15 as a % of original records (repU) 0.1%
```

COMPARING PRIVACY MEASURES (SET.SEED = 1240, I.E. UNIQUE = 3)

```
> print(t1, plot = FALSE)
Disclosure risk for 1000 records in the original data

Identity disclosure measures
from keys: var1 var2 var3
For original  ( UiO )  0 %
For synthetic ( repU ) 0 %.

Table of attribute disclosure measures for var1 var2 var3
Original measure is  Dorig and synthetic measure is DiSCO
Variables Ordered by synthetic disclosure measure

        attrib.orig attrib.syn check1 Npairs check2
1 var4           0          0             0
```

```
> replicated.uniques (sds, df_ods)
Uniques and replicated uniques for  1  synthesised data set(s)
 from keys:  var1 var2 var3 var4

Uniques in  original data:
 1 from  1000 records ( 0.1 %)
Uniques in synthetic data:
 0 from  1000 records ( 0% )

Replicated uniques:
 0
as a % of uniques in synthetic  NaN%
as a % of original records (repU) 0%
```

SUMMARY

- CART (synthpop) generates synthetic data with both high utility and low risk

- However and this is the point:

    – We know there is a problem (because we created it)

    – We know that common measures do not capture the problem

- We are also not alone in identifying this problem (Manrique-Vallier and Hu, 2018)

SECTION 4: SOLUTION

- The good news: we know how to solve the problem

- The bad news: why would we solve the problem if we did not know it existed?

THE GOOD NEWS: SOLUTIONS

- minumlevels = 5: Ensures the data are treated as categorical

- cp = 0.05 (default = $1e^{-8}$): prevent large trees (i.e. overfitting)

- minbucket = 75 (default = 5): the minimum number of observations in any terminal node

- Other options also exist

- More generally: It is possible to solve the problem, but you have to know the problem exists

VISUALIZING TREES (DEFAULT VS. MODIFIED)
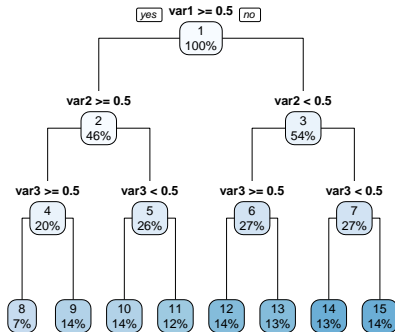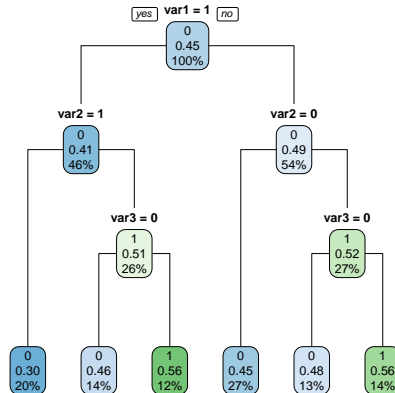
Figure 6: CART (default)

Figure 7: CART (modified)

COMPARE HISTOGRAM X 100 SYNTHETIC DATASETS
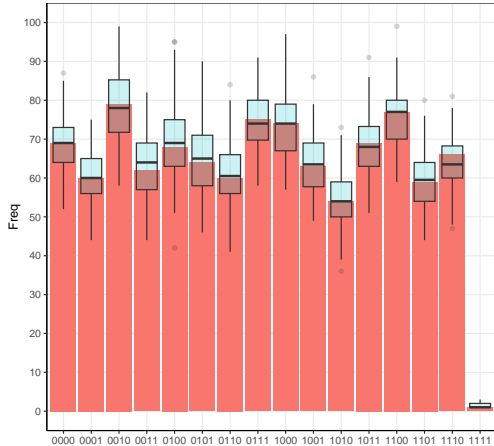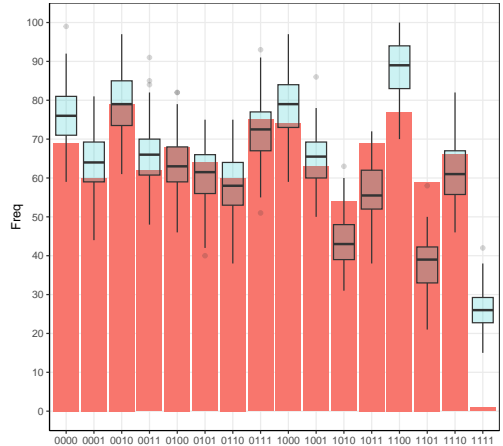
Figure 8: CART (default)

Figure 9: CART (modified)

SUMMARY

- You can solve the risk of disclosure by reducing utility

- The bad news: why would you reduce utility, if you did not know there was a risk of disclosure

- More generally: how do we know privacy measures are effective in identifying the problem?

SECTION 5: CONCLUSION

CONCLUSION

- It has long been understood that there is a trade-off between utility and risk

- It seemed that CART models were less sensitive to this trade-off than other SDGs (i.e. higher utility, lower risk)

- Using a simulated data set, we show that CART models do not protect unique cases

- It is possible to protect unique records,
  - You have to sacrifice utility

- Using common privacy metrics, we show that these do not capture risk in our simulated data
  - How do you know if there is a problem?

- If you did not know there was a problem, why would you sacrifice utility?

CONTACT

Jonathan Latner

jonathan.latner@iab.de

Reproducible code:

- Github: https://github.com/jonlatner/KEM_GAN/tree/main/latner/projects/simulation