Avatarization report for unknown project

Octopize MD

2024-11-13

Contents

Objective of this document

This automatically generated report presents the privacy and utility indicators following the unknown project avatarization.

Summary

Privacy metric	Value	Target
Hidden rate	93.31 %	> 90 %
Median local cloaking	12.0	> 5
Closest distance ratio	1.0	> 0.3
Closest rate	97.01 %	> 90 %
Column direct match protection	95.23 %	> 50 %
Row direct match protection	44.4	> 90

Signal metric	Value	Target
Correlation differences	0.3 %	< 10 %
Distribution differences	0.08	< 0.1

The definitions for each of those metrics can be found below.

Privacy metrics

Attack scenario

The metrics used to evaluate privacy and described below measure the **three criteria** defined by the GDPR:

- **Singling out**, which is the risk to identify an individual in a dataset.
- **Linkability**, which is the ability to link individuals with another dataset.
- **Inference**, which is the possibility to deduce, with significant probability, the value of an individual using the anonymized dataset.

An attack scenario goes as follows:

- The attacker has both datasets: the original dataset and the anonymized one. Note: this is the worst-case scenario and is **highly improbable in practice**.
- The attacker uses a *distance-based attack* and tries to link each individual with its most resembling synthetic data.
- The individuals usually most vulnerable to this type of attack are those with remarkable characteristics, also called *outliers*.

By nature, the Avatar method makes it possible to quantify the re-identification risk associated with the singling out criterion.

Hidden rate

Hidden rate: 93.31 % (Target: > 90%)

The **hidden rate** refers to the singling out criteria. It is the probability that an attacker makes a mistake when linking an individual with its most similar synthetic individual.

Note: a hidden rate below 90% **does not mean** that 10% of the individuals are not protected.

Local cloaking

Median local cloaking: 12.0 (Target: > 5)

The **local cloaking** refers to the singling out criteria. It is the median number of avatars that look more like an individual than the avatar generated.

Note: the higher the local cloaking, the better.

Closest distances ratio

Closest distances ratio: 1.0 (Target: > 0.3)

A closest distances ratio is the distance between a synthetic record and its closest original record divided by the distance to its second closest original record.

The closest distances ratio metric is the median of all those closest distances ratio.

Because this metric does not use the links between original and synthetic records, it can be used to evaluate all kinds of synthetic datasets.

Note: the ratio is bounded between 0 and 1. The higher the closest distances ratio, the better.

Closest rate

Closest rate: 97.01 % (Target: > 90%)

The closest rate evaluates the percentage of individuals combining a low distance to closest and a low closest distances ratio.

Because this metric does not use the links between original and synthetic records, it can be used to evaluate all kinds of synthetic datasets.

Column direct match protection

Column direct match protection: 95.23 % (Target: > 50%)

The **column direct match protection** measures how unlikely each column in the data is to be used as a direct identifier. To compute this metric, it considers both the percentage of correct univariate matches and the number of unique values each variable has. This is to account for the fact that variables with low cardinality will be more likely to be matched than variables with high cardinality, but that this match does not represent a risk of re-identification. The level of protection against a direct match protection is computed for each variable and the **column direct match protection** outputs the minimum value (i.e. the column direct match protection for the variable the most likely to be used as a direct identifier).

Note: A **column direct match protection** of 0 means that a direct identifier has been kept in the data and that the data is **not** anonymized. A **column direct match protection** below the target does not necessarily mean that the data is at risk but it is recommended to check the quality of the avatarization with respect to the variable(s) showing the smallest direct match protection value(s).

Row direct match protection

Row direct match protection: 44.4 % (Target: > 90%)

The row direct match protection measures the percentage of original individuals that are not present in the anonymous dateset. The higher, the better.

Because this metric does not use the links between original and synthetic records, it can be used to evaluate all kinds of synthetic datasets.

Signal retention metrics

Projections

Comparison of datasets structure similarity using factor analysis methods (PCA, FAMD, MCA).

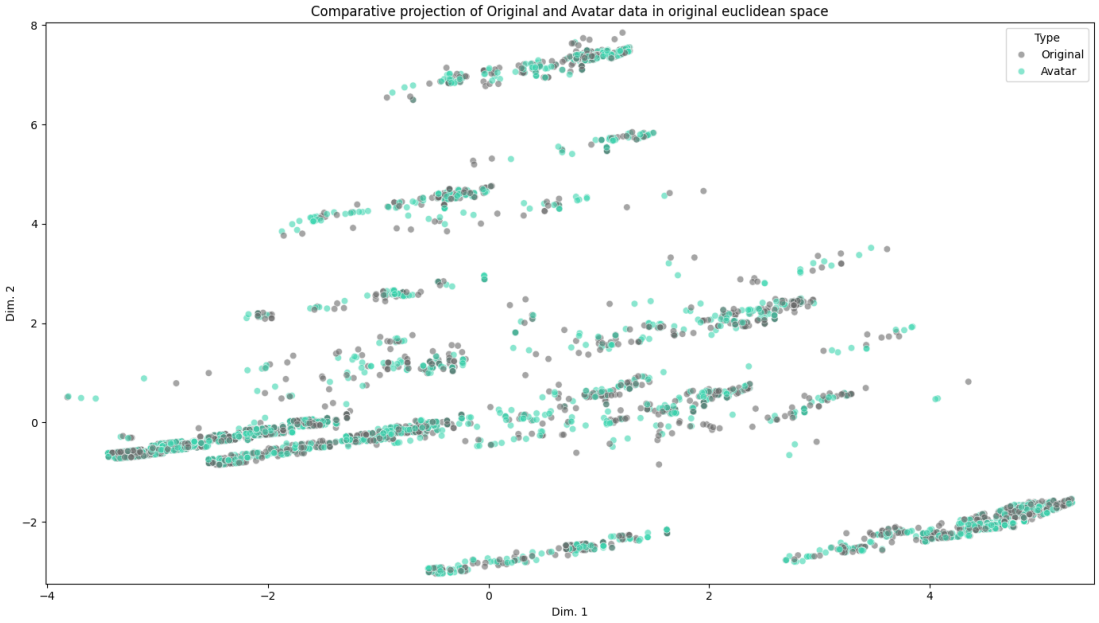

Figure 1: projections of original (in grey) and avatar data (in green) in the space defined by original data (first two dimensions)

The superposition of scatter shows that **the original structure is kept** in the avatarized dataset.

Note: the method purposely recenters outliers because they are the most likely to be re-identified. For ease of readability, the graph is produced with a maximum of 10,000 records taken at random from the dataset.

Contributions

Analysis of variable contribution to the construction of the first two projection dimensions.

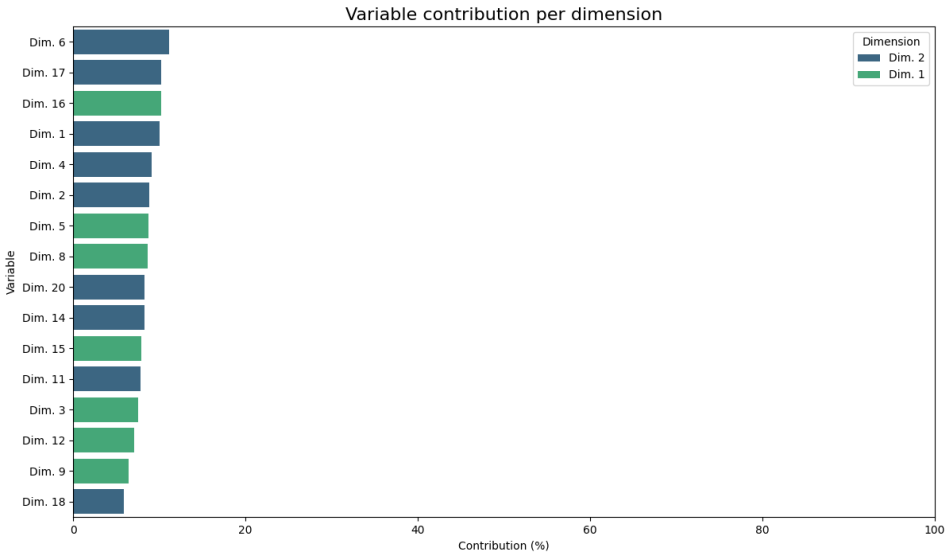

Figure 2: percentage of variable contribution to the construction of Dim. 1 (green) and Dim. 2 (blue).

Note: to facilitate visualization, only variables with a contribution greater than 5% are shown.

Hellinger distance mean

Hellinger distance mean: 0.08 (Target: < 0.1) with a standard deviation of **0.1**

Distance between the original and the avatar distribution is mathematically computed using Hellinger distance. The metric returns the mean of Hellinger distance computed for each variable.

The Hellinger distance fluctuate between 0 and 1 where 0 represents no difference between the two distributions and 1 represents a difference, at any point, one from the other.

We consider that a Hellinger distance lower than represents a very good signal retention while distances between and represent tolerable signal retention.

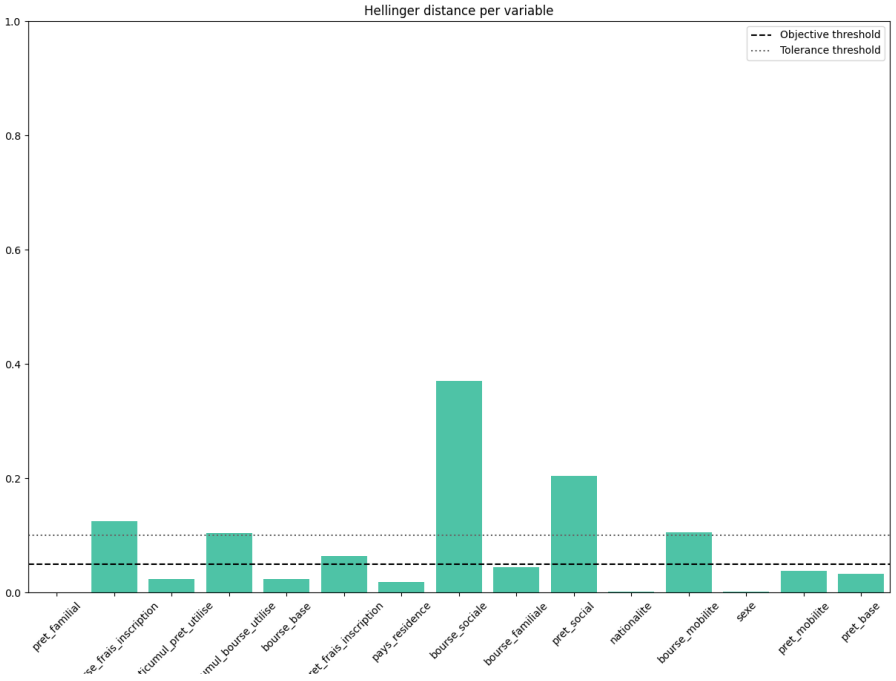

Figure 3: Hellinger distances for all variables with utility thresholds.

Note: variables with a **high proportion of missing data** are likely to have higher Hellinger distance since changing a single value has more impact on the total distribution. This does not mean that the avatarization failed for these variables.

Correlation similarities

Global correlation difference: 0.3% (Target: < 10%)

It computes the average of the absolute variations of Pearson's correlation

To visually compare correlations preserved by avatarization, we represent the correlation matrices of the two datasets.

Note: these matrices only consider continuous variables.

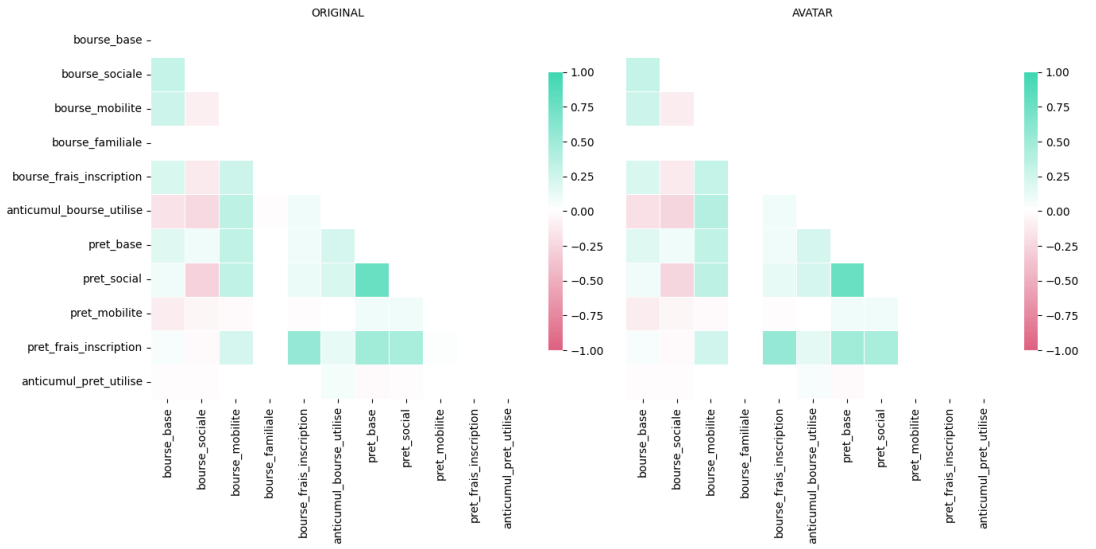

Figure 4: correlation matrices for original and avatar dataset.

We compute the overall mean correlation difference over all continuous variables.

Metadata

- avatarization job id: 2ab4b39e-e250-41b0-8401-39416370adde
- avatarization job created_at: 2024-11-13 12:41:51.184551+00:00
- original dataset id: 492b4ce7-d10f-41e8-9194-82ca66473c6e
- original dataset hash: e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855
- original dataset lines: 36701
- original dataset dimensions: 64
- avatar dataset id: 663d5071-8af8-4cf0-8586-cf0ddeec6fcc
- avatar dataset hash: 150a3c47eecb9f58cf46f90326593e8f0edb2a96e2743aadc6253bff88926be7
- k: 10
- ncp: None
- $use\ categorical\ reduction$: True