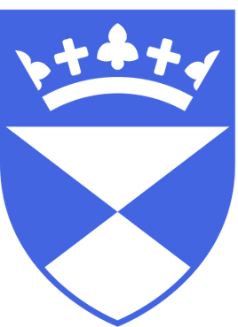
Synthetic Data Generation and evaluation



Anna Marek

Industry supervisor: Thanos Tsiaras

Academic supervisors: Dr Jianguo Zhang and Andrew Cobley

MSc Data Engineering, School of Science and Engineering

University of Dundee

10th September 2019

# Declaration

"I declare that the special study described in this dissertation has been carried out and the dissertation composed by me, and that the dissertation has not been accepted in fulfilment of the requirements of any other degree or professional qualification."

# Certificate

"I certify that Anna Marek has satisfied the conditions of the Ordinance and Regulations and is qualified to submit this dissertation in application for the degree of Master of Science."

# Acknowledgements

# Abstract

# Table of Contents

[Declaration 9](#_Toc11240425)

[Certificate 10](#_Toc11240426)

[Acknowledgements 11](#_Toc11240427)

[Abstract 12](#_Toc11240428)

[Table of Contents 13](#_Toc11240429)

[List of Figures 14](#_Toc11240430)

[References 15](#_Toc11240431)

[Appendix 15](#_Toc11240432)

# List of Figures

# Introduction

Dataset overview

Literature overview

Synthea for synthesising health records

Justify choosing random forest

It is well understood that inferences from synthetic data will only be valid if the models

used to synthesize the data correspond to those that can be considered as having generated

the original data. It is important for staff synthesizing the data to assess how well this

condition is fulfilled by their synthetic dataset, and this can be done by so-called general

and specific measures of utility. The former are summaries of differences between the

distributions of the original and the altered data while the latter compare the differences

between results from particular analyses.1

Published evaluations of synthetic data using

specific utility measures, usually for just a few selected analyses, have highlighted differ-ences in the quality of syntheses (Reiter (2005a), Dreschler and Reiter (2009), Kinney

et al. (2011), Miranda and Vilhuber (2016), Nowok (2015)).

KL divergence – Woo et al. 2009

Methods of dealing with unbalanced datasets:

<https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18>

# Methods

Synthetic data generation (problem, methodology, results)

## Data Quality Assessment

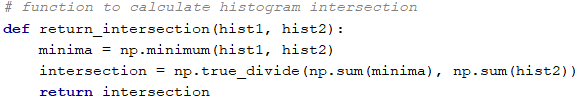
### Histograms and scatterplots2

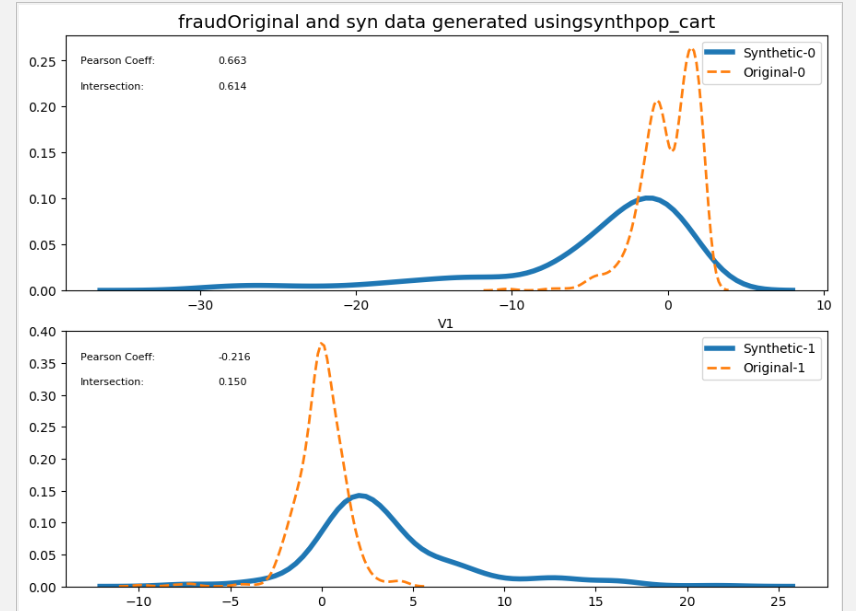
The distribution of data is assessed using histograms and scatterplots.

Histograms are plotted for each individual feature, the original and synthetic data histograms are presented on the same graph for a better comparison. Pearson coefficient and intersection values are also displayed.

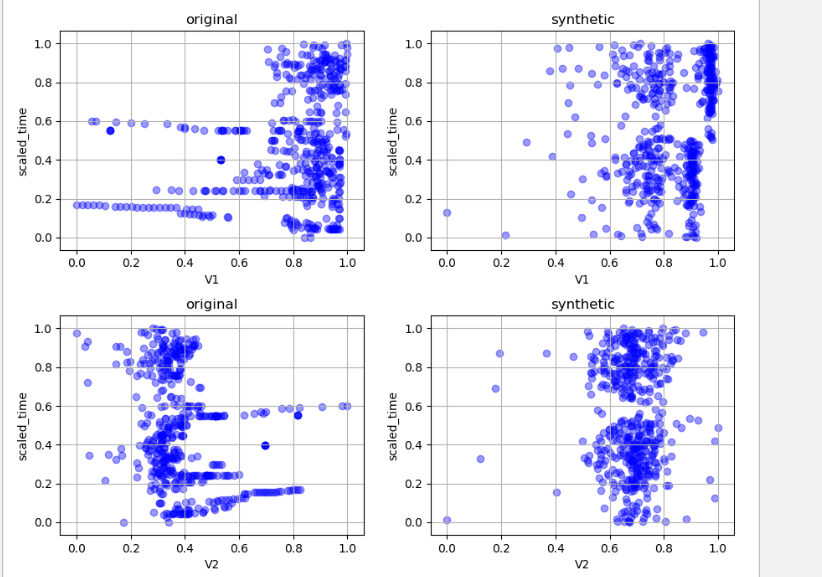
Pearson correlation coefficient measures the linear relationship between two datasets and is calculated using scipy.stats.pearsonr() function.

Intersection is used to calculate the similarity of two normalized histograms and can be defined as follows.





Scatterplots are calculated for the combination of each two features. The scatter plots for original and synthetic data are displayed next to each other. The data is rescaled to values between 0 and 1 for easier comparison.



### Pairwise Pearson Correlation2

<https://github.com/datasciencecampus/Synthetic_data>

This method looks at the relationships between the variables and provides a way of verifying whether the relationships seen in the original dataset have been preserved in the synthetic dataset.

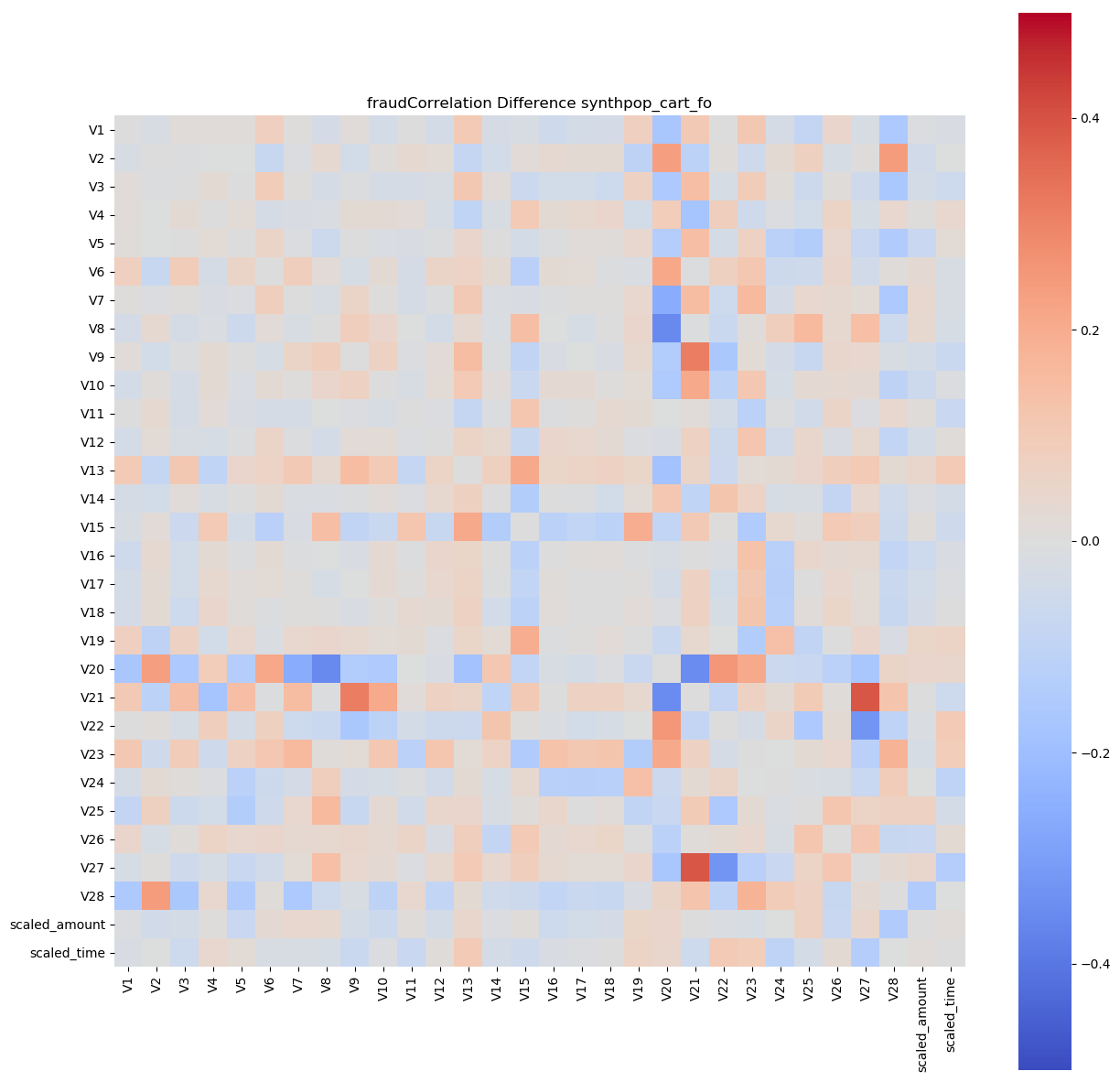
Pearson Correlation between n variables in the original dataset is calculated producing n×n matrix CorrMatrixoriginal, the same is done for synthetic data with the CorrMatrixsynthetic  as a result.

In order to investigate whether the relationships between the variables based on this measure are similar for synthetic and original data correlation matrix different is calculated as follows:

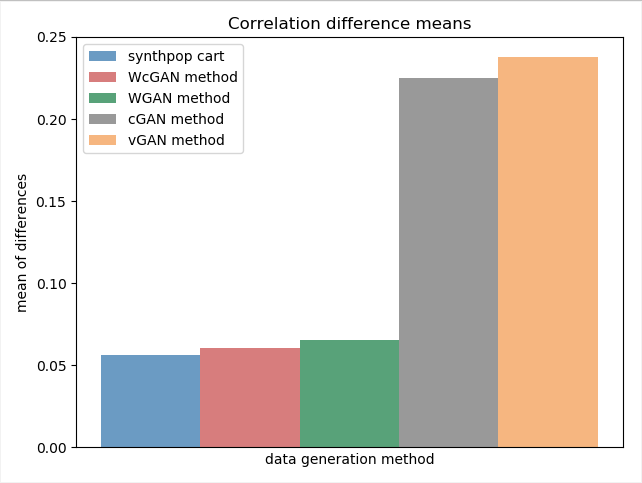
CorrMatrixdifference = CorrMatrixoriginal - CorrMatrixsynthetic

If both datasets are identical the CorrMatrixdifference is equal to zero. Low values suggest a good quality synthetic dataset.

CorrMatrixdifference is visualised in from of a heat map.



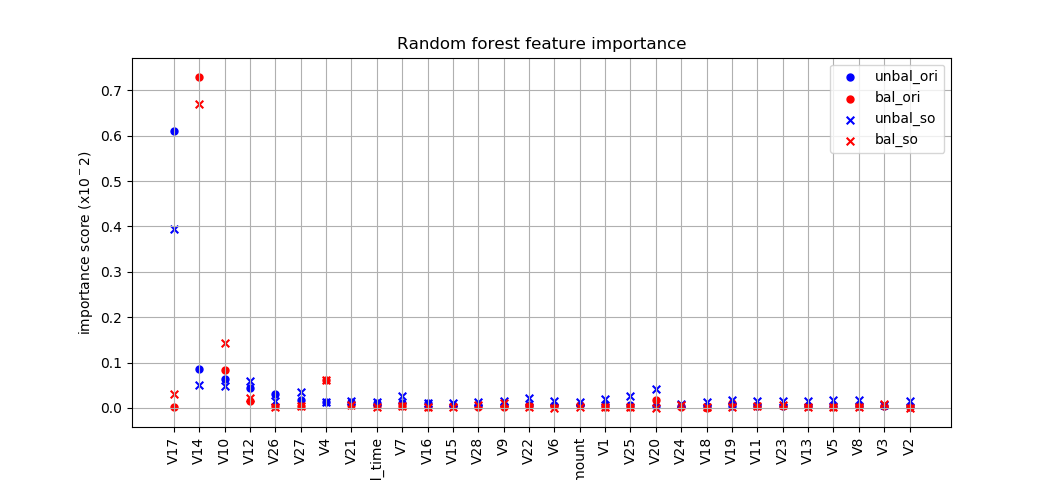
It is quite difficult to compare the matrices based simply on the visual assessment of heat maps and so the mean of the absolute values of CorrMatrixdifference is calculated (Mabs). The lower the Mabs value, the better. Comparing the values for different methods of data synthesis gives an indication of the best solution.



### Random forest feature importance3

Random forest algorithm ranks features based on how much they contribute to changing the purity of the node. This approach investigates how using synthetic dataset to train the model changes the feature importance scores in respect to model trained on the original dataset. The assumption is that a good quality synthetic dataset would lead to the algorithm using similar predictive features.

ONE NUMBER FOR IT ALL??



### Synthetic Ranking Agreement (SRA)4

This method focuses on the importance of any two algorithms trained on the synthetic data having similar relative performance to the same algorithms trained on the original data. This is particularly relevant when researchers want to obtain the best performing algorithm without sharing the original dataset. Choosing algorithms or tuning hyper parameters is a stepwise process and usually done over time, it is therefore important that not only the best, but any of the compared algorithms are seeing similar improvement or deterioration in performance. If the synthetic data influences the performance of algorithms in the same way as the original, the final algorithm chosen relying on synthetic data would be the same as the one chosen using original data.

The property described above can be formalized in the way shown below:

<

* <

MT – performance metric

D1/D1S - train/synthetic train set

D2 / D2S - test/ synthetic test set

A– learning algorithm

A(D1 ) – model

The SRA for synthetic data generated using method G, given k algorithms A1,..,Ak is defined as:

The SRA is described as ‘the empirical probability of a comparison on the synthetic data being ‘correct’’4. The final value is dependent on the type and number of algorithms used. In practice, the closer the SRA value to 1, the better.

### Measure based on the propensity score1

The general idea behind this measure is that the trained classifier would perform worse when trying to distinguish between original and synthetic data if the synthetic data was of a good quality. The synthesised data of bad quality would allow better classifier performance.

To assess this idea mathematically a propensity-score-based general utility measure is used. Propensity score is simply the probability of a unit being assigned to a particular group.

First, n1 original data rows are combined with n2 synthetic data rows resulting in dataset with N rows (N= n1 + n2).

Original and synthetic data rows are assigned with labels, ‘0’ and ‘1’ respectively.

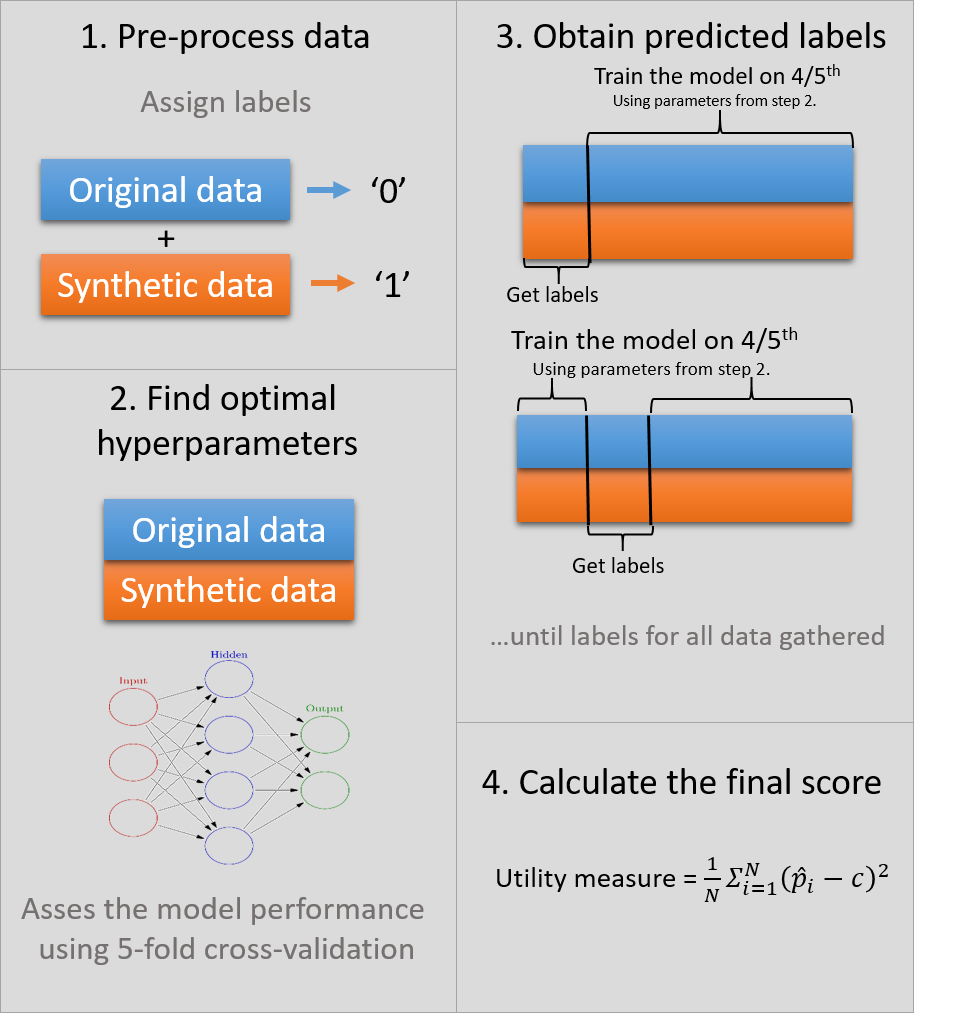
A neural network classifier is trained in a 5-fold nested cross-validation setting to find the best hyperparameters. All the data is used in this process.

Once the optimal parameters are found, the model is retrained using 4 folds of the data and the 5th fold is used to obtain class label (original or synthetic) in a form of probability. Model is then retrained using different 4 folds and 5th fold is used to obtain labels for a different part of the dataset. This is repeated until labels for all folds are obtained.

The propensity score is calculated using the below equation, where c is the proportion of synthetic data (c = n2 /N).

Utility measure =

The utility statistic is equal to 0 when original and synthetic data are identical.



# Results

User interface – python

# Discussion

Issues with SRA – not measuring the magnitude of relationships. Need to synthesise the whole dataset. Many algorithms needed, time consuming - training, data synthesis. When an agency prepares synthetic data for user, they will not know exactly what analysis will be carried out, what methods of data evaluation will be used.

How to connect the results of all the quality measures?

We optimised the hyperparameters of the algorithms we used for synthetic data generation against the aforementioned µ***abs*** metric with grid search. Note that the optimisation process for the deep-learning models is still in progress, as we are currently experimenting with different types of neural networks to further improve performance. – this provides the risk of optimizing for this one particular measure.

# References

1. Snoke J, Raab G, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. April 2016. http://arxiv.org/abs/1604.06651. Accessed May 31, 2019.

2. Kaloskampis I. Synthetic data for public good. 2019:1-24. https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/.

3. Beaulieu-Jones BK, Wu ZS, Williams C, Greene CS. Privacy-preserving generative deep neural networks support clinical data sharing. doi:10.1101/159756

4. Jordon J, Yoon J, van der Schaar M. Measuring the quality of Synthetic data for use in competitions. June 2018. http://arxiv.org/abs/1806.11345. Accessed May 31, 2019.

# Appendix

