Synthetic Tabular Data Generation

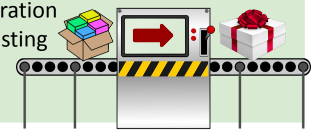
Author: Roshan Pandey, Supervisor: Nic Palmarini
Newcastle University

roshan-pandey/synthetic-data-gen

Introduction

As the data is growing at an exponential pace, people's **privacy issues** are also increasing. To tackle this issue, synthetic data is generated with similar **statistical properties** to real data. In this work, **Synthetic Data Vault (SDV)** will be used. SDV used **Generative Adversarial Network (GAN)** under the hood.

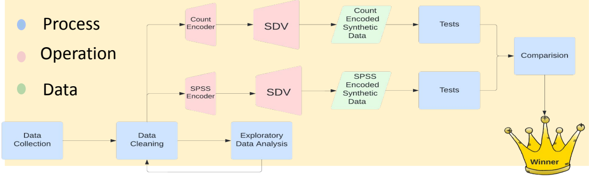
This work is done in collaboration with the subject matter Expert at NICA. This collaboration allowed interaction with existing innovators in the market such as AINDO.



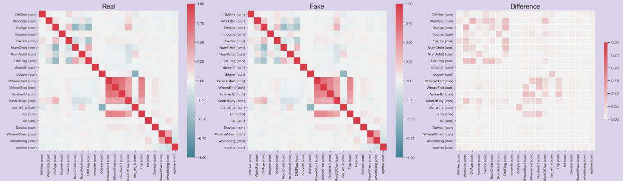
Methodology

Below figure shows the whole process followed to create synthetic data in different data configurations:

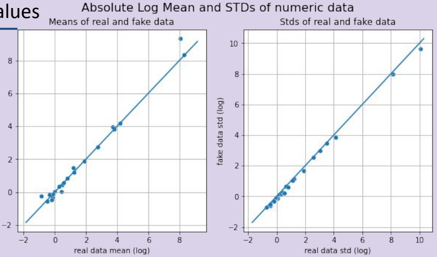
- Config 1: SPSS encoded, 1st merged Activity log and individual info then generated.
- Config 2: SPSS encoded, Generated both tables separately.
- Config 3: Count encoded, 1st merged Activity log and individual info then generated.
- Config 4: Count encoded, Generated both tables separately.



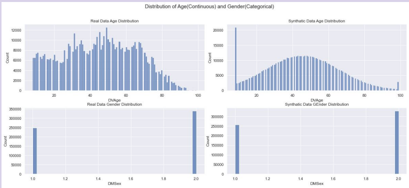
Results



The above figure shows the **correlation plot** of Real and Synthetic data and last is the heat map of the difference between the correlation values of real and synthetic data. We can observe that both data have similar correlation values

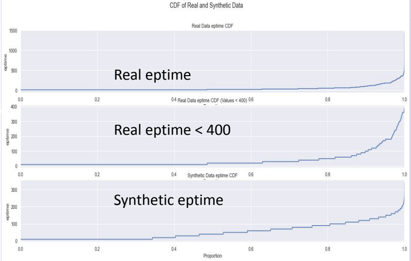


This plot shows the log of sum of **mean** and **standard deviation** of all the **numerical features** of real vs synthetic data. It can be seen that both are in a straight lines and hence we can say that they have their mean and standard deviation close to each other.



1st row represents the **distribution of *Age*** on the LHS, it can be seen that the distribution is slightly close to **skewed normal distribution** hence, the model has produced data which is much more similar to **normal distribution**. 2nd row is a **distribution of categorical variable *Gender*** and they both are very similar.

Results



This plot shows the **cumulative distribution** of **eptime** - Time spent on a particular activity. It can be observed that the CDF of both real and synthetic data is very similar to each other.

Conclusion & Future Scope

Overall, the synthetic data generated using **SDV** was very close to real data in terms of **statistical properties**. However, there are few features whose distributions are deviating by a huge margin. Few things that can be tried in the future are:

- **Wasserstein GAN** can be used if building a data generator from scratch.
- SDV offers **non-GAN** techniques for data generation. Those can be tried as well.
- Apart from *SPSS* and *Count* encoding, there are many more encoding techniques, that are worth trying.

References

[1] GitHub. 2020. GitHub - sdv-dev/SDV: Synthetic Data Generation for tabular, relational and time-series data.. [online] Available at: <https://github.com/sdv-dev/SDV> [Accessed 20 May 2022].
[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems, 27.
[3] Arjovsky, M., Chintala, S. and Bottou, L., 2017, July. Wasserstein generative adversarial networks. In International conference on machine learning (pp. 214-223). PMLR.
[4] Probability distribution - [Accessed 30 July 2022] En.wikipedia.org. Available at: <https://en.wikipedia.org/w/index.php?title=Probability_distribution&oldid=1102516344> .
[5] Cumulative distribution function - [Accessed 30 July 2022]. En.wikipedia.org. Available at: <https://en.wikipedia.org/w/index.php?title=Cumulative_distribution_function&oldid=1104350887>.

r.k.pandey2@newcastle.ac.uk
nic.palmarini@newcastle.ac.uk