Synthetic Tabular Data Generation

MSc. Data Science Thesis - Interim Report

Author: Roshan Pandey
Supervisor: Nic Palmarini

Synthetic Tabular Data Generation

1. Introduction

The rate at which data is growing, it also increases the concerns about the ethical use of the data and the privacy of the individuals involved in the process of gathering this information. Also, there are several sectors where data is still scarce and that makes it hard to train machine learning and deep learning models. To tackle these problems, one solution is to generate synthetic data which is statistically close to the original data. This way it is possible to maintain the privacy of the people and also if the data is not sufficient enough to train deep neural networks we can always generate more synthetic data without losing the consistency of the data.

2. Objective

We have behavioural data from time-use survey [1][7] of 4238 families from all over the UK. Data is consist of what each individual in a family does at a certain point in time of a day, they have recorded their activities every 10 minutes throughout the day. It also consists of demographic information about the individuals. So, the goal is to generate synthetic data from this existing data which will protect the privacy of the individuals and at the same time, the data will keep the consistency same as the real data.

3. Overview of Progress

3.1 Data Collection

The dataset is open-source and was provided by the UK National Innovation Center of Aging (NICA). A time-use survey [8] is a statistical survey which aims to report data on how, on average, people spend their time. However, the data was originally collected by conducting the surveys for 9388 individuals from 4238 households generating a total of 16533 records. Survey/interview was conducted in various settings such as face-to-face interviews, telephone interviews, and diary entries. Files are in “spss” format, so using pandas and pyreadstat to read the files in the Jupyter Notebook environment.

3.2 Literature Review

The literature review has been done by going through various work done in the past. Most of the previous work is done using generative adversarial networks (GAN) [2] but it has been used mostly for image data whereas this project aims to generate tabular data using GAN and its variations [3]. There are some open-source packages that can generate tabular data such as Synthetic Data Vault (SDV) [4] which is a “Synthetic Data Generation ecosystem of libraries that allows users to easily learn single-table, multi-table and time-series datasets to, later on, generate new Synthetic Data that has the same format and statistical properties as the original dataset” [5] but they don't handle categorical data very well so this work will try to overcome that limitation and will provide a more robust framework to generate synthetic tabular data without having to worry about categorical data with a large number of categories present in any particular column.

3.3 Exploratory Data Analysis and Data Manipulation

Have also done basic exploratory data analysis to find any specific pattern in the data, however, haven’t found much information at this point. Now trying to convert categorical data into numeric values. For this task, I will be using “learned embedding” [6] as there are many categories and when one-hot encoding is performed it is increasing the number of columns to a greater extent and also data becomes very sparse.

Most of the data processing and manipulation is being done in python and its packages like pandas, NumPy, matplotlib, seaborn, sklearn, Keras, TensorFlow etc. For data analysis, Jupyter Notebook as an environment is being used.

Git is being used for version control and managing the progress of the project. It makes it easy to track all the changes that have been done throughout the project lifecycle and also keeps the backup of the codebase.

As for data and project management, a subset of cookiecutter is being used. Cookiecutter is a tool that lets us create a Python project from existing templates. However, it’s not limited to Python - there are templates for Java and PHP projects, Sublime Text plugins, and more. We start by selecting an existing template, for example, a Flask website or a Python module.

4. Project Plan

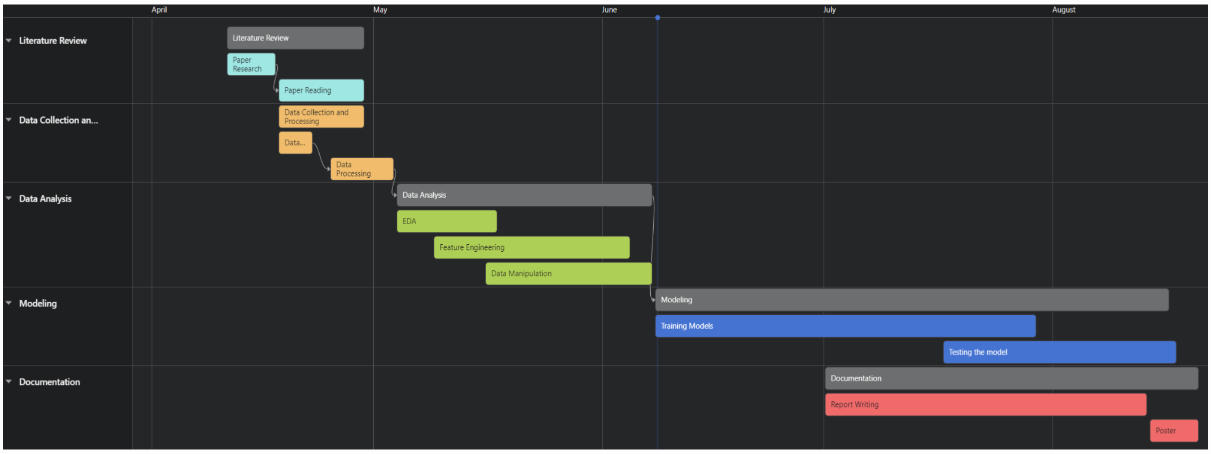

Fig 1: Gantt Chart of Synthetic Data Generation Project

In figure 1, the complete timeline of the project is shown. The literature review started around the 11th of April and it lasted till the end of April. Data was collected from NICA in the mid of April 2022 and soon after that, some minor data manipulation was done so that exploratory data analysis can be done. EDA took around a month of time from the 2nd of May till the 3rd of June. Now that there is a proper understanding of the data training and testing of various models can be started which will take approximately 2 months of time from the start of June till the end of July and then final testing of consistency of synthetic data will be done by training models on synthetic data and testing their accuracy on real data and vice versa. Also, the heatmap of the correlation of the variables will be plotted for the ease of visualization of the closeness of both real and synthetic data. Once there is a good amount of details and clarity about the development of the project documentation will start from mid of July and will last until the mid to late August 2022.

5. References

[1] Morris, S., Humphrey, A., Alvarez, P. and D'Lima, O., 2016. *The UK Time Diary Study 2014 - 2015*. London.

[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] Ashrapov, I., 2020. *Tabular GANs for uneven distribution*. arXiv preprint arXiv:2010.00638.

[4] N. Patki, R. Wedge and K. Veeramachaneni, "The Synthetic Data Vault," *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016, pp. 399-410, doi: 10.1109/DSAA.2016.49.

[5] 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].

[6] Developers, G., 2020. *Embeddings: Translating to a Lower-Dimensional Space | Machine Learning Crash Course | Google Developers.* [online] Google Developers. Available at: <https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space> [Accessed 27 April 2022].

[7] Gershuny, J., Sullivan, O. (2017). *United Kingdom Time Use Survey, 2014-2015*. Centre for Time Use Research, University of Oxford. [data collection]. UK Data Service. SN: 8128, http://doi.org/10.5255/UKDA-SN-8128-1

[8] Beta.ukdataservice.ac.uk. 2017. *UK Data Service › Study*. [online] Available at: <https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8128> [Accessed 22 April 2022].

Data Management Plan

0. Project title, author, Version and date

Project: Synthetic Tabular Data Generation		
Author: Roshan Pandey	Version: 1.0	Date: 7th June 2022

1. Description of the data

1.1 Type of Study

This data was collected to understand the behaviour of people based on their demographic data such as age, sex, location etc.

1.2 Assessment of existing data

Data is publically available at (UK Data Service › Study) however, one needs to be registered for UK data services hence, this data was provided by the UK National Innovation Center of Aging for various purposes such as visualization, predicting what people might do at a certain point of time in a day, and to generate synthetic data which will have a very close correlation with the real one.

1.3 Types of data

Data explained in the previous section (1.2) was collected from the survey conducted for 4238 families from all over the UK with a total of 9388 individuals aged 8 years and above. Data is consist of what each individual in a family does at a certain point in time in a day, they have recorded their activities every 10 minutes throughout the day. Data comprises both quantitative as well as qualitative data.

1.4 Format and scale of data

File format is spss (.sav) and python packages such as pyreadstat and pandas are used to read the file into dataframe format. There are a total of 7 files each containing data of the survey in different structures such as long and wide formats and information related to individuals, their household data etc. The long format contains 587632 rows and 50 columns and the wide format contains 16533 rows and 2335 columns and the overall size of the data is around 162MB. Spss files are a commonly used format in the industry so it's easy to share and use in any environment.

The data was collected in 2014-2015 so it is quite old now, however, the shift in people's behaviour is a very slow process and it was collected from a wide range of families with varying family sizes, different occupations, age ranges etc. so it can still be used for various analysis and if synthetic data can be generated using currently available data, it will be possible to generate more synthetic data using the newer version of data when available.

2. Data collection/generation

2.1 Methodologies for data collection/generation

From my end, I received the data from the UK NICA. However, the data was originally collected by conducting the surveys for 9388 individuals from 4238 households generating a total of 16533 records. Survey/interview was conducted in various settings such as face-to-face interviews, telephone interviews, and diary entries.

2.2 Data quality and standards

As stated in the previous section 2.1 the data was collected from various interviews and surveys so the all the entries are done manually by human beings and they have followed a set of rules to make all the entries so the consistency of the data can be assured but some outliers can be expected because of human error. Also, there are not many null values.

3. Data management, documentation and curation

3.1 Managing, storing and curating data

Data has been downloaded on a local machine and is also available on one drive in case of data loss from the local machine. To store the code files and other useful resources related to the project such as plots, model weights, and dependencies' details GitHub is being used.

3.2 Metadata standards and data documentation

The data is well documented and details of the data can be found in the book "What We Really Do All Day Insights from the Center for Time use Research" by Jonathan Gershuny and Oriel Sullivan. Also, an online pdf can be found here NatCen Reports, Questionnaire and Methodology (ukdataservice.ac.uk) to learn more about the data.

4. Data security and confidentiality of potentially disclosive information

4.1 Main risks to data security

The is available to anybody who is registered with UK Data Services so there is no concern about the data security and confidentiality. The is already sanitised before it was shared with NICA which means all the vulnerable personal details like name or email address have been encoded with unique serial numbers.

5. Data sharing and access

5.1 Suitability for sharing

Yes, the data is suitable to be shared within Newcastle University as this data was provided by NICA and they are registered with UK Data Services.

5.2 Discovery by potential users of the research data

The data and its documentation (metadata) can easily be found on the internet by searching "The UK Time Diary Study 2014 - 2015". DOI for the data and its details is here http://doi.org/10.5255/UKDA-SN-8128-1

5.3 Data preservation strategy and standards

Before sharing this data UK Data Services sanitise the data and remove all the sensitive fields so, from a research perspective, it can be shared within the organisation which is registered with UK Data Services for as long the data is available on the platform here (United Kingdom Time Use Survey, 2014-2015 | Centre for Time Use Research).

5.4 Restrictions or delays to sharing, with planned actions to limit such restrictions

There is no restriction or delays in sharing of the data used for this project within the Newcastle University or members registered with the UK Data Service. For code base that will be created during the project can only be shared with NICA and Newcastle University. It will be available in the Git repository. Upon completion of this project, if it can be made open source, the repository will be public. So, there is a delay in sharing the code base.

6. Responsibilities and Resources

Most of the resources like a computer screen, wifi connectivity and learning materials have been provided by NICA. For compute power, the local machine (Laptop) is been used.	
7. Relevant institutional, departmental or study policies on data sharing and data security	
Policy	**URL or Reference**
Data Management Policy and Procedures	https://www.ncl.ac.uk/media/wwwnclacuk/research/files/ResearchDataManagementPolicy.pdf
Information Security	https://services.ncl.ac.uk/itservice/policies/InformationSecurityPolicy-v2_1.pdf
Other	