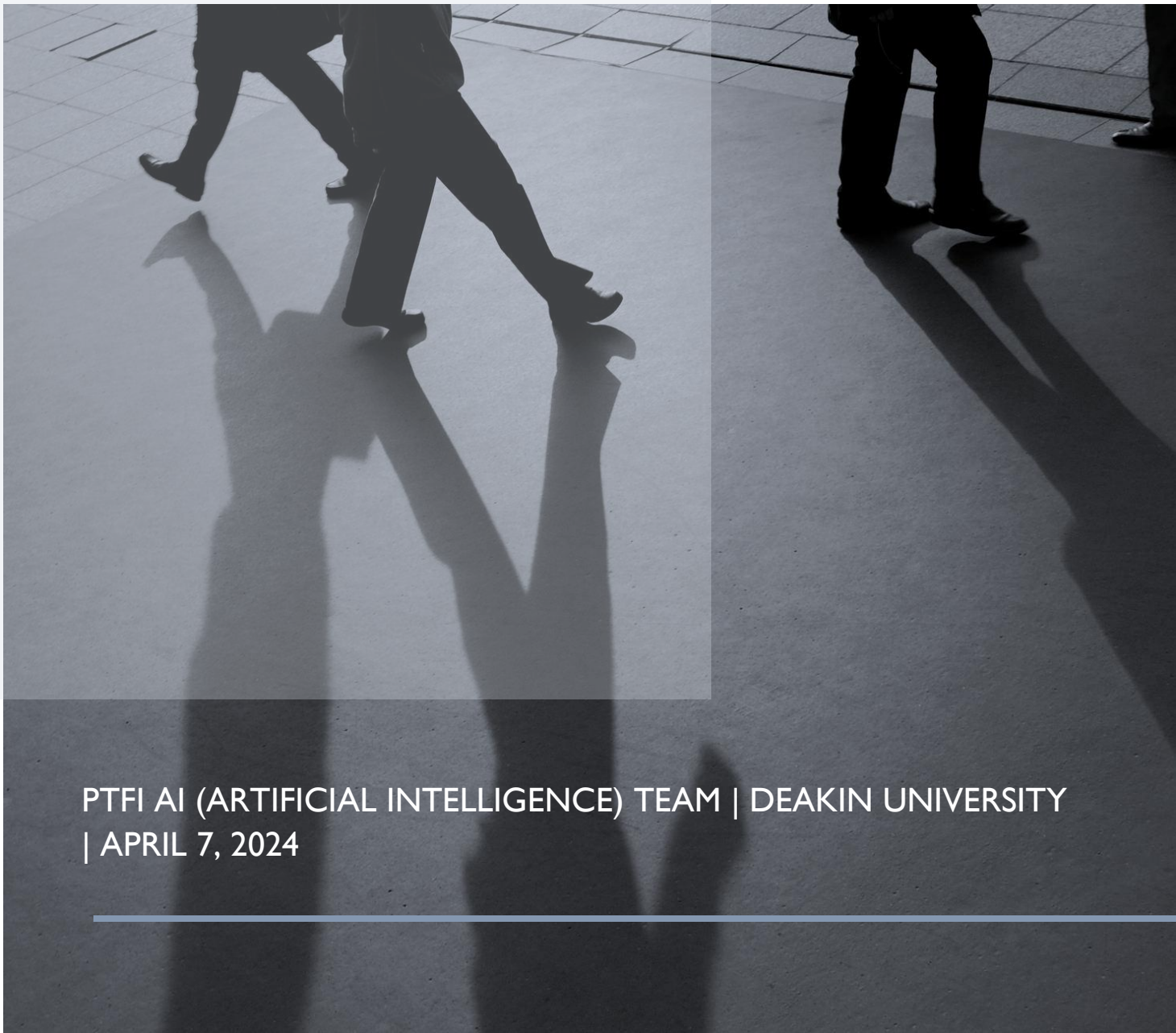
SYNTHETIC DATA ANALYSIS

PRIVACY TECHNOLOGIES FOR FINANCIAL INTELLIGENCE: A DATA BYTES COMPANY PROJECT



PTFI AI (ARTIFICIAL INTELLIGENCE) TEAM | DEAKIN UNIVERSITY
| APRIL 7, 2024

BACKGROUND

Creating synthetic datasets for financial crime when real datasets are unavailable due to privacy concerns and regulations can be a viable solution. However, it's crucial to note that the effectiveness of synthetic datasets depends on the quality of the generation process. The synthetic data should accurately capture statistical properties and patterns present in real financial data to ensure that models trained on such data are relevant and reliable in real-world applications.

The scope of this analysis was to investigate options for generating synthetic data for applying privacy enhancing technologies. A prior attempt to generate data using the Faker library was successful but the data lacked the statistical properties of the source data. Three further methods were researched: mimesis, mockaroo and SDV. Two of the methods are python libraries (mimesis and SDV) while mockaroo is a web application.

The report covers some background of the three methods, the application of the methods and the results achieved and an evaluation of these results. It is noted that there are many other methods to generate synthetic data, however, due to time constraints only the three methods selected were investigated.

METHOD 1: SYNTHETIC DATA VAULT (SDV)

The Synthetic Data Vault Project was first created at MIT's Data to AI Lab in 2016. According to the SDV GitHub, the **Synthetic Data Vault** (SDV) is a Python library designed to be your one-stop shop for creating tabular synthetic data. The SDV uses a variety of machine learning algorithms to learn patterns from your real data and emulate them in synthetic data[1]. In particular, this research will focus on CTGAN which is a collection of Deep Learning based synthetic data generators for single table data, which are able to learn from real data and generate synthetic data with high fidelity[2].

What is GAN?

GANs (Generative Adversarial Networks) consist of two neural networks—the generator and the discriminator. The generator creates synthetic data, while the discriminator evaluates whether it's real or fake. These networks iteratively improve each other by minimizing their respective loss values during training.

CTGAN learns from real data catching the underlying statistical patterns. CTGAN generates synthetic data with similar statistical properties using the learned knowledge. GANs excel in image generation, data augmentation, and anomaly detection. Challenges include mode collapse and training instability. Researchers aim to improve stability and explore conditional GANs.

Which type of data was generated?

In this analysis, an insurance fraudulent claims detection dataset was generated. Insurance fraud involves deceptive practices related to insurance. It encompasses illegitimate claims, intentionally misrepresenting information, and organised manipulation of the claims process. Opportunistic fraud occurs when individuals overstate losses, while premeditated fraud involves planned schemes by professional fraudsters. Insurance fraud is a serious offense, punishable by imprisonment and fines. Detecting and preventing benefits all consumers by reducing overall costs and maintaining fair premiums.

Banks in Australia offer various financial services where health insurance is a major part of their business. In Australia, private health insurance plays a significant role in the financial services sector. For more information on the dataset, please see Fraudulent Claims Detection Dataset (kaggle.com).

Results

The results indicated that evaluation visualisation demonstrates a remarkable similarity between the first insurance dataset and the synthetic data generated by CTGAN. CTGAN, a machine learning-based approach, effectively learns intricate patterns from the real data and realistically replicates them in the synthetic dataset. This method is particularly well-suited for our project's goal of generating high-quality synthetic data closely mirrors the characteristics of the original data records.

While CTGAN has proven effective in generating synthetic insurance records, its versatilitiy extends beyond this specific use case. PTFI can explore how to leverage CTGAN for other financial datasets, opening up possibilities for creating realistic synthetic data.

METHOD 2: MOCKAROO

Mockaroo is a useful web-based tool that generates realistic datasets for data science projects, software development and testing. It allows users to create custom data sets with various data types, including names, addresses, phone numbers, email addresses, and dates. By defining the structure and format of the data, users can generate synthetic data in formats like CSV, JSON, SQL, and Excel.

Data scientists and data analysts commonly use Mockaroo. It's ideal for enriching the data sets for a better data modelling, testing, and presentations or demonstrations. We can quickly create realistic data without manual input or complex generation processes1.

Which type of data was generated?

The generated data was transaction data for Anti Money Laundering (AML) purposes. This data, known as the Synthetic Anti-Money Laundering Transaction Data (SAML-D) dataset, was created using a novel AML transaction generator. The aim was to enhance the features present in the dataset, thereby helping researchers in evaluating their models and developing more advanced monitoring methods.

Our aim for using this dataset is that it is the closed to a realistic AML dataset which was created by credible researchers. This dataset underlying characteristics will aid us in our understanding of AML data and help us with creating our own AML synthetic dataset.

For more information on the dataset, please see https://www.kaggle.com/datasets/berkanoztas/synthetic-transaction-monitoring-dataset-aml?resource=download

Results

The results indicated that the dataset used for the comparison, "SAML-D" dataset had certain characteristics in the distribution of there data, which was lacking in the mockaroo dataset. Although we would be able to add minor tweaks in order to re-engineer the dataset with the necessary characteristics, an exact replication of the characteristics would be nearly impossible. But at the end of the day the mockaroo dataset can still be used for certain applications such as flagging fraud transactions by manipulating data at random intervals.

METHOD 3: MIMESIS

Mimesis is a robust data generator for Python that can produce a wide range of synthetic data in various languages. This tool is useful for populating testing databases, creating fake API endpoints, filling pandas DataFrames, generating JSON and XML files with custom structures, and anonymizing production data, among other purposes[3].

Mimesis is similar to Faker which was used to generate the prior synthetic dataset. However, Mimesis boasts a quicker run time, 35 different locales (languages), supports custom data providers and custom field handlers and has a variety of data from different providers. Similar to Faker, the data generated is purely random and does not consider the statistical properties of a source dataset.

Which type of data was generated?

The dataset used to compared the generated data with was *The Anti Money Laundering Transaction Data (SAML-D)*. The data was generated via a novel AML transaction generator, creating the SAML-D dataset with enhanced features and typologies, aiming to aid researchers in evaluating their models and developing more advanced monitoring methods[4]. For more information on the dataset, please see https://www.kaggle.com/datasets/berkanoztas/synthetic-transaction-monitoring-dataset-aml?resource=download.

AML data was selected due to the importance and complexity of privacy and sharing sensitive data between organizations and government agencies. Money laundering is at the forefront of financial crime and generating datasets for applying privacy technology and experimentation using machine learning and AI approaches provides real world significance and real value to society.

Results

The results when comparing the generated data to the source data using Mimesis were underwhelming. The source dataset contained 12 columns, of which, most of the generated data columns did not represent similar distributions or summary statistics as the source data. For example, the average transaction amount in the source data was $8,762 with a standard deviation of $25,615 indicating a wide spread of data whereas the average amount in the generated data was $1,000 with a standard deviation of $292.

Another example was the frequency of payment types where 85% of transactions in the source data were from 4 out of the 7 payment types. However, around 55% of transactions were from these payement types in the generated dataset.

Additionally, not all columns could be generated using Mimesis. Numpy operations were required to generate columns for account numbers and the laundering flag. Using purely random data to apply in a real world scenario will unlikely generalize well and will most likely produce poor results.

RECOMMENDATIONS

Given the information above, SDV is the recommended option due to the ability to replicate similar statistical properties from the source dataset. There are uses cases where Mockaroo and Mimesis are uaseful for generating synthetic data but for PTFI, SDV stands out the best.

The full notebooks for the analysis of SDV, Mockaroo and Mimesis can be located on the PTFI GitHub:

https://github.com/DataBytes-Organisation/Privacy-Technologies-for-Financial-Intelligence/blob/develop/notebooks/

REFERENCES

1. SDV Development Team. (2024). Synthetic Data Vault [GitHub repository]. GitHub. https://github.com/sdv-dev/SDV
2. SDV Development Team. (Year). CTGAN: Conditional Tabular GAN [GitHub repository]. GitHub. https://github.com/sdv-dev/CTGAN
3. Mimesis Developers. (2024). Mimesis Documentation. Retrieved from https://mimesis.name/en/master/
4. Berkanoztas, B. (2024). Synthetic Transaction Monitoring Dataset (AML) [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/berkanoztas/synthetic-transaction-monitoring-dataset-aml?resource=download
5. Unknown, (2024). Fraudulent Claims Detection Dataset [Data set]. Kaggle. Retrieved from Fraudulent Claims Detection Dataset (kaggle.com)