LATEST VERSION:

**Synthesising - Noor**

We used 2 data synthesising methods to generate a total of 8000 rows which is about 38% of the total dataset. These rows were split evenly between 2 data generation methods, each with their own unique approach.

Method 1 generated 4000 rows using remapping and mathematical sampling. This works by remapping the existing data and then applying mathematical sampling techniques to generate new rows which fit to the structure of the original data. It preserves the authenticity of the fraud column by ensuring it remains untouched by the remapping process. Method 1 also allows for quick and efficient data generation and is ideal when needing to generate data quickly with minimal resources. It has a low computational cost.

Now onto Method 2 which also generated 4000 rows but using CTGAN, a more advanced approach. CTGAN is a sophisticated machine learning method based on neural networks. It learns the underlying patterns and complex relationships in the data. A key strength of this method is its ability to capture intricate relationships between variables – particularly the fraud column – making it an ideal method for accurate and high-quality data generation. CTGAN is best suited for larger datasets handling complex data structures where it is critical to ensure realism i.e. generated data which can replicate the real-world patterns.

**Data Merging - Noor**

On the right, is a diagram of our merged dataset. During the merging process, we combined three (cleaned and enriched) original datasets and two synthetic datasets, resulting in a final dataset of over 20,000 rows, or more precisely, 20,877 rows.

This includes 1,000 rows from Dataset 1, about 1,500 rows from Dataset 2, about 10,000 rows from Dataset 3, and 4,000 rows each from the two synthetic datasets.

To maintain consistency and have a form of identifier, we assigned a unique index for each row. Each row also has a fraud label which is a numerical boolean, with 1 representing fraud and 0 representing not fraud. This fraud label is crucial for ML & AI training & testing. The merged dataset is now ready for the AI team to begin their model training.

**Live Demo - Noor**

Now let’s perform a live demo of the data merger. This is our python script to load and merge our datasets.

Clicking run and we see it finished in less than a second. Now let’s open this resulting output file.

We can see it’s a large tabular database.

Some key features of this resulting dataset are:

It has about 20,800 rows and each of those rows has an index, which is like an identifier

Each row also contains a fraud column, 1 means it's fraudulent and 0 means not fraudulent. This is crucial for the AI to train and learn about fraudulent claims.

Original Versions:

Data Synthesising - Noor

The synthesising process created 2 datasets each with 4000 rows - so 8000 in total which would make up 38% of the final dataset. This was done using 2 methods. Method 1, Remapping & Mathematical Sampling, works by remapping the existing data and then using mathematical sampling to create new rows. It preserves the authenticity of the fraud column by ensuring it remains untouched by the remapping process. It allows quick and efficient data generation and is an ideal approach when needing to generate data quickly using minimal resources.

Method 2 CTGAN, uses Machine Learning Neural Networks to learn patterns. CTGAN is a more advanced method where the model learns the complex relationships between variables and generates realistic synthetic data. It is particularly useful for larger, complex datasets that need to maintain accurate relationships between variables, the fraud column in this case.

Data Merging - Noor

On the right is a diagram depicting the shape of the merged dataset.

In the merging process, we brought together three cleaned and 2 synthetic datasets and combined them into a dataset with over 20,000 rows. To have a form of identifier, we assigned a unique index to each row. The merged dataset is ready for the AI team to train their model on.

Live Demo - Noor

Now let’s perform a live demo of the data merger. This is our python script to load and merge our datasets.

Clicking run and we see it finished in less than a second. Now let’s open this resulting output file.

We can see it’s a large tabular database.

Some key features of this resulting dataset are:

It has about 20,800 rows and each of those rows has an index, which is like an identifier

Each row also contains a fraud column, 1 means it's fraudulent and 0 means not fraudulent. This is crucial for the AI to train and learn about fraudulent claims.