Financial Synthetic Data is the New Oil for FinCrime Analytics

Edgar Lopez-Rojas, PhD
FinCrime Analytics Consultant and Researcher

19 May, 2020

Ealax
IT Consulting Services
Agenda

1. Introduction
2. Our Approach
3. Case Study: PaySim
4. Conclusions
5. References
Anti-Money Laundering (AML) Problem

Figure: From United Nations Office on Drugs and Crime (UNODC)
The problem of applying effective controls

- **PRIVACY**: Financial institutions protect the financial information of their customers [2].
The problem of applying effective controls

- **PRIVACY**: Financial institutions protect the financial information of their customers [2].

- **ACCESS**: Third party providers and researchers find it difficult to obtain financial datasets for developing and testing better controls.
The problem of applying effective controls

- **PRIVACY**: Financial institutions protect the financial information of their customers [2].
- **ACCESS**: Third party providers and researchers find it difficult to obtain financial datasets for developing and testing better controls.
- **COSTLY**: Even inside a financial organisation, it is difficult to develop effective controls without going through many cycles of trial and error.
The problem of applying effective controls

- **PRIVACY**: Financial institutions protect the financial information of their customers [2].

- **ACCESS**: Third party providers and researchers find it difficult to obtain financial datasets for developing and testing better controls.

- **COSTLY**: Even inside a financial organisation, it is difficult to develop effective controls without going through many cycles of trial and error.

- **EVIDENCE**: Nearly 90% of the top financial institutions have been fined due to lack of effective controls.
Introduction

Measuring Effective Controls
**Introduction**

**Measuring Effective Controls**

---

**NORMAL**

**TRUE NEGATIVES (TN)**

**FALSE NEGATIVES (FN)**

---

**FRAUD**

**FALSE POSITIVES (FP)**

**TRUE POSITIVES (TP)**
Agenda

1. Introduction
2. Our Approach
3. Case Study: PaySim
4. Conclusions
5. References
Gartner Hype Cycle for Emerging Technologies, 2019

- Biochips
- AI PaaS
- Autonomous Driving Level 5
- Edge AI
- Explainable AI
- Personalization
- Knowledge Graphs
- Light-Cargo Delivery Drones
- Transfer Learning
- Flying Autonomous Vehicles
- Augmented Intelligence
- Nanoscale 3D Printing
- Decentralized Autonomous Organization
- Generative Adversarial Networks
- Decentralized Web
- AR Cloud
- Immersive Workspaces
- DigitalOps
- Adaptive ML
- Next-Generation Memory
- 3D Sensing Cameras
- Autonomous Driving Level 4
- Graph Analytics

Plateau will be reached:
- less than 2 years
- 2 to 5 years
- 5 to 10 years
- more than 10 years
- obsolete before plateau
As of August 2019

gartner.com/SmarterWithGartner

Source: Gartner
Why Synthetic Data?

There are many benefits of using synthetic datasets:

- **Data is ready** and available.
Why Synthetic Data?

There are many benefits of using synthetic datasets:

- **Data is ready** and available.
- **Privacy** of customers is not affected.
Why Synthetic Data?

There are many benefits of using synthetic datasets:

- **Data is ready** and available.
- **Privacy** of customers is not affected.
- Results can be disclosed to, and **compared by**, other researchers.
Why Synthetic Data?

There are many benefits of using synthetic datasets:

- **Data is ready** and available.
- **Privacy** of customers is not affected.
- Results can be disclosed to, and **compared by**, other researchers.
- Different scenarios can be modeled for **experimentation** using well controlled parameters.
Why Synthetic Data?

There are many benefits of using synthetic datasets:

- **Data is ready** and available.
- **Privacy** of customers is not affected.
- Results can be disclosed to, and compared by, other researchers.
- Different scenarios can be modeled for experimentation using well controlled parameters.
- We can also use it for **Training non experts** in a field to become familiar with diverse scenarios before they ever seen it.
Using synthetic data to develop effective controls

- **Machine Learning (ML)** brings powerful capabilities for classification of malicious behaviour [3].
Using synthetic data to develop effective controls

- **Machine Learning (ML)** brings powerful capabilities for classification of malicious behaviour [3].
- Unfortunately it is very dependent on **quality data** to train the models.
Using synthetic data to develop effective controls

- **Machine Learning** (ML) brings powerful capabilities for classification of malicious behaviour [3].
- Unfortunately it is very dependent on **quality data** to train the models.
- Can we generate a **synthetic version** of the required data? [4].
Using synthetic data to develop effective controls

- **Machine Learning** (ML) brings powerful capabilities for classification of malicious behaviour [3].
- Unfortunately it is very dependent on quality data to train the models.
- Can we generate a **synthetic version** of the required data? [4].
- Is it **good enough**?
Using synthetic data to develop effective controls

- **Machine Learning** (ML) brings powerful capabilities for classification of malicious behaviour [3].
- Unfortunately it is very dependent on quality data to train the models.
- Can we generate a synthetic version of the required data? [4].
- Is it good enough?
- Can we measure the hidden crime? [1, 6]
Why Synthetic Data for ML?

The three biggest drawbacks of using ML for AML are:

- The lack of **labelled data** due to the **hidden crime**.
Why Synthetic Data for ML?

The three biggest drawbacks of using ML for AML are:

- The lack of **labelled data** due to the **hidden crime**.
- the **class imbalance** problem. Criminal data is considerably less than other data.
Why Synthetic Data for ML?

The three biggest drawbacks of using ML for AML are:

- The lack of **labelled data** due to the **hidden crime**.
- the **class imbalance** problem. Criminal data is considerable less than other data.
- The evolving threat of Financial Crime that makes training **datasets obsolete** quite fast.
Agenda

1. Introduction
2. Our Approach
3. Case Study: PaySim
4. Conclusions
5. References
Simulation to generate proper synthetic data

Figure: PaySim Simulator [4]
Privacy preserving method

**Diagram:**

1. Obtain data sample
2. Extract parameters

**OWNER of the data**

3. Add fraud parameters
4. Run the Simulator
5. Apply fraud control methods

**RESEARCHER**

6. Summarise Results
7. Repeat for another fraud scenario
Agenda

1. Introduction
2. Our Approach
3. Case Study: PaySim
4. Conclusions
5. References
Figure: Triple-Helix AML [5]
Financial Synthetic Data is the new Oil for Machine Learning Engines in FinCrime Analytics

- Any questions?
- edgar@ealax.com

Would you like to use Synthetic Data for your FinCrime Analytics?
Agenda

1. Introduction
2. Our Approach
3. Case Study: PaySim
4. Conclusions
5. References


