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

Anti-Money Laundering (AML) Problem

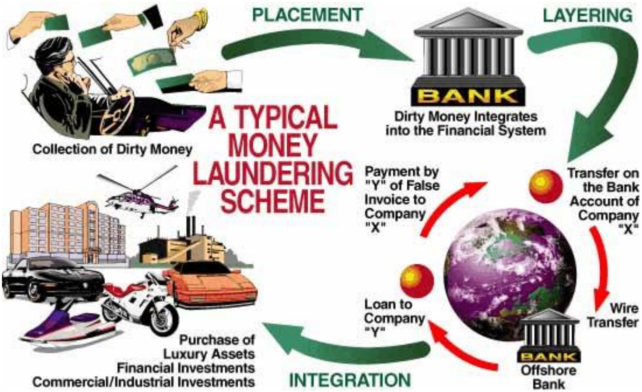


Figure: From United Nations Office on Drugs and Crime (UNODC)

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.

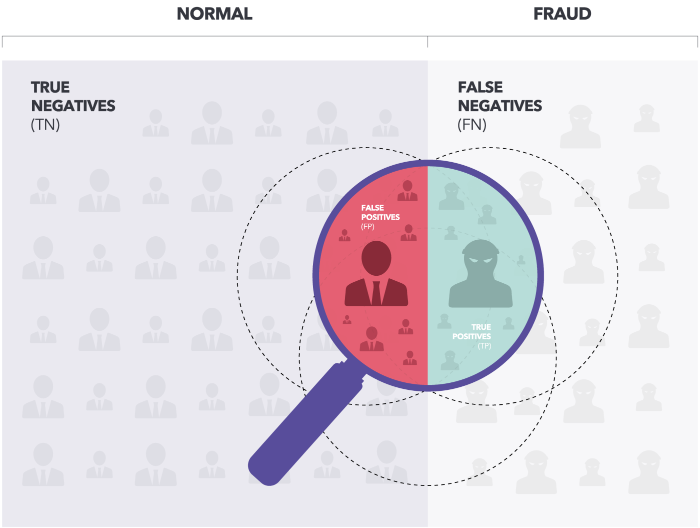
The problem of applying effective controls

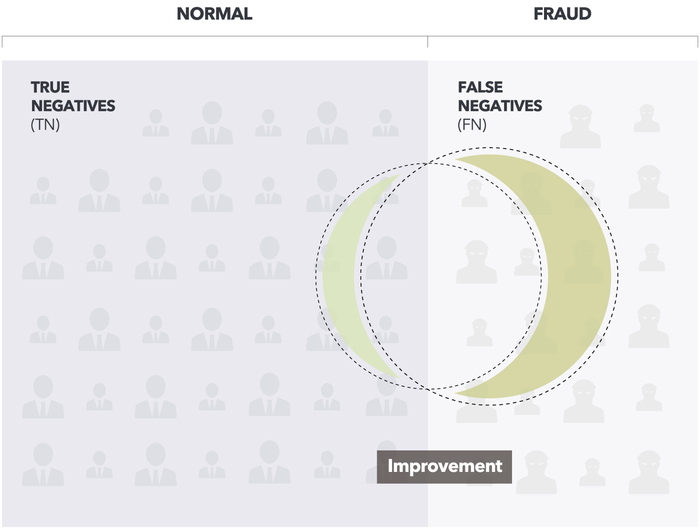
- PRIVACY: Financial institutions protect the financial information of their customers [2].
- ACCESS: Third party providers and researchers find it difficult to obtain financial datasets for developing and testing better controls.
- COSTLY: Even inside a financial organisation, it is difficult to develop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

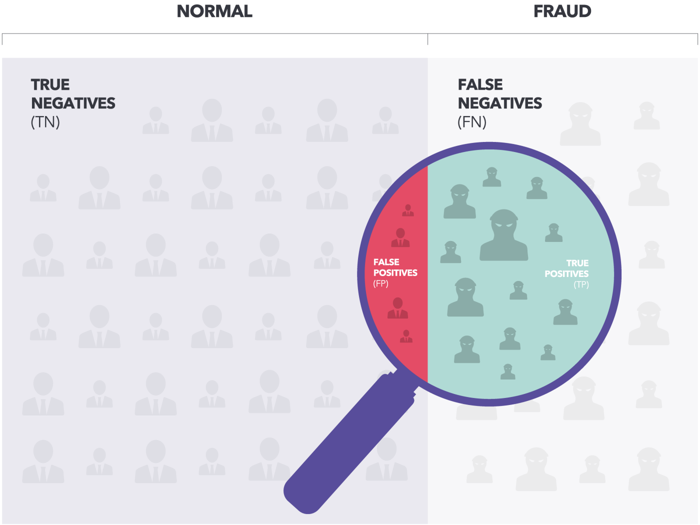
The problem of applying effective controls

- 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.

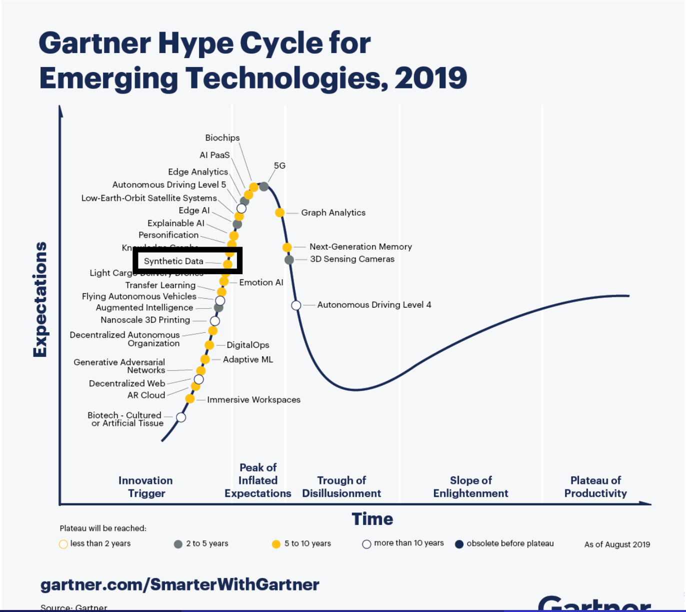

NORMAL
FRAUD
TRUE NEGATIVES (TN)
FALSE NEGATIVES (FN)
FALSE POSITIVES (FP)
TRUE POSITIVES (TP)


NORMAL
FRAUD
TRUE NEGATIVES (TN)
FALSE NEGATIVES (FN)
FALSE POSITIVES (FP)
TRUE POSITIVES (TP)


NORMAL
FRAUD
TRUE
NEGATIVES
(TN)
FALSE
NEGATIVES
(FN)
Improvement


NORMAL
FRAUD
TRUE
NEGATIVES
(TN)
FALSE
NEGATIVES
(FN)
FALSE
POSITIVES
(FP)
TRUE
POSITIVES
(TP)

Agenda


Gartner Hype Cycle for Emerging Technologies, 2019
Expectations
Biochips
AI PaaS
Edge Analytics
Autonomous Driving Level 5
Low-Earth-Orbit Satellite Systems
Edge AI
Explainable AI
Personification
Synthetic Data
Light Carg
Transfer Learning
Flying Autonomous Vehicles
Augmented Intelligence
Nanoscale 3D Printing
Decentralized Autonomous Organization
Generative Adversarial Networks
Decentralized Web
AR Cloud
Biotech - Cultured or Artificial Tissue
5G
Graph Analytics
Next-Generation Memory
3D Sensing Cameras
Emotion AI
Autonomous Driving Level 4
DigitalOps
Adaptive ML
Immersive Workspaces
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
Time
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 considerable 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

Simulation to generate proper synthetic data

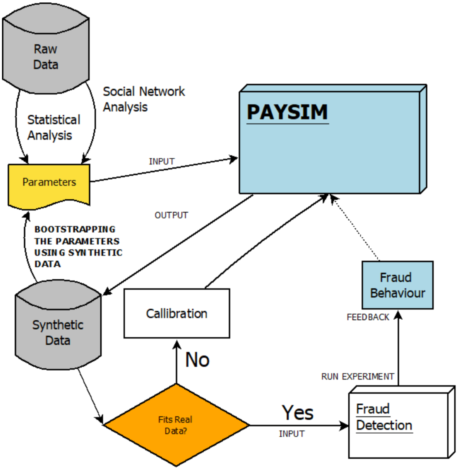
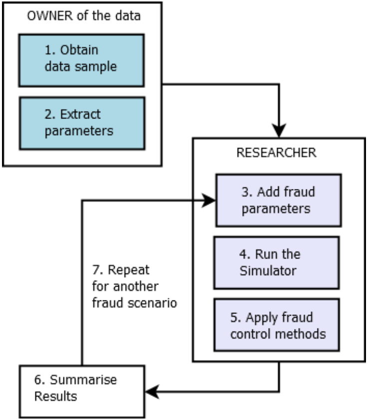


Figure: PaySim Simulator [4]

Privacy preserving method



Agenda

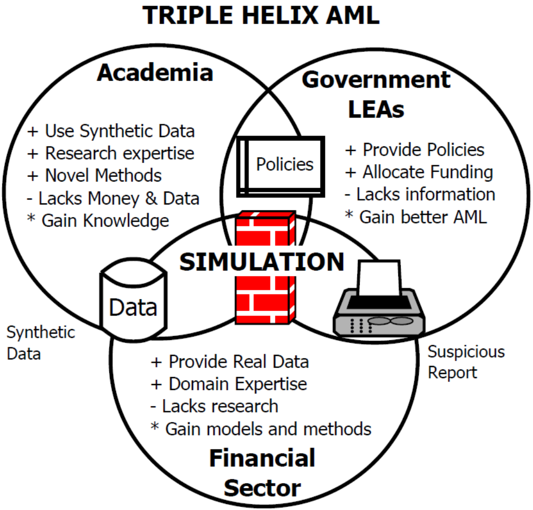
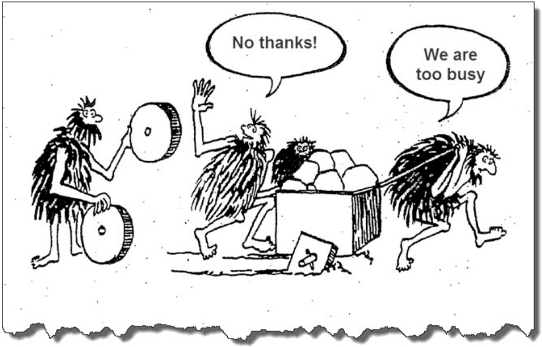


Figure: Triple-Helix AML [5]

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] Michael Levi, Peter Reuter, and Terence Halliday. Can the AML system be evaluated without better data? *Crime, Law and Social Change*, 2018.

[2] Edgar A. Lopez-Rojas, Dincer Gultemen, and Erjon Zoto. On the gdpr introduction in eu and its impact on financial fraud research. In *The 30th European Modeling and Simulation Symposium-EMSS, Budapest, Hungary*, 2018.

[3] Edgar Alonso Lopez-Rojas and Stefan Axelsson. Money Laundering Detection using Synthetic Data. In Julien Karlsson, Lars ; Bidot, editor, *The 27th workshop of (SAIS)*, pages 33–40, Örebro, 2012. Linköping University Electronic Press.

[4] Edgar Alonso Lopez-Rojas, Ahmad Elmir, and Stefan Axelsson. PaySim: A financial mobile money simulator for fraud detection. In *The 28th European Modeling and Simulation Symposium-EMSS*, Larnaca, Cyprus, 2016.

[5] Edgar Alonso Lopez-Rojas and Erjon Zoto. Triple Helix Approach for Anti-Money Laundering (AML) Research Using Synthetic Data Generation Methods. In *The 10th International Conference on Society and Information Technologies: ICSIT 2019*, 2019.

[6] Michael J. O'Loughlin, Mary K. Driskell, and Gregory Diehl. Financial simulation: Combining cost information in systems analysis. In *Winter Simulation Conference Proceedings*, pages 578–581, 1990.