

# clearbox<sup>AI</sup>

humans && machines

---

SYNTHETIC DATA FOR DATA AUGMENTATION

# Augmenting historical financial data with synthetic time series

C.so Castelfidardo, 30/a  
10129, Torino (Italy)  
info@clearbox.ai

VAT ID: (IT)12161430017

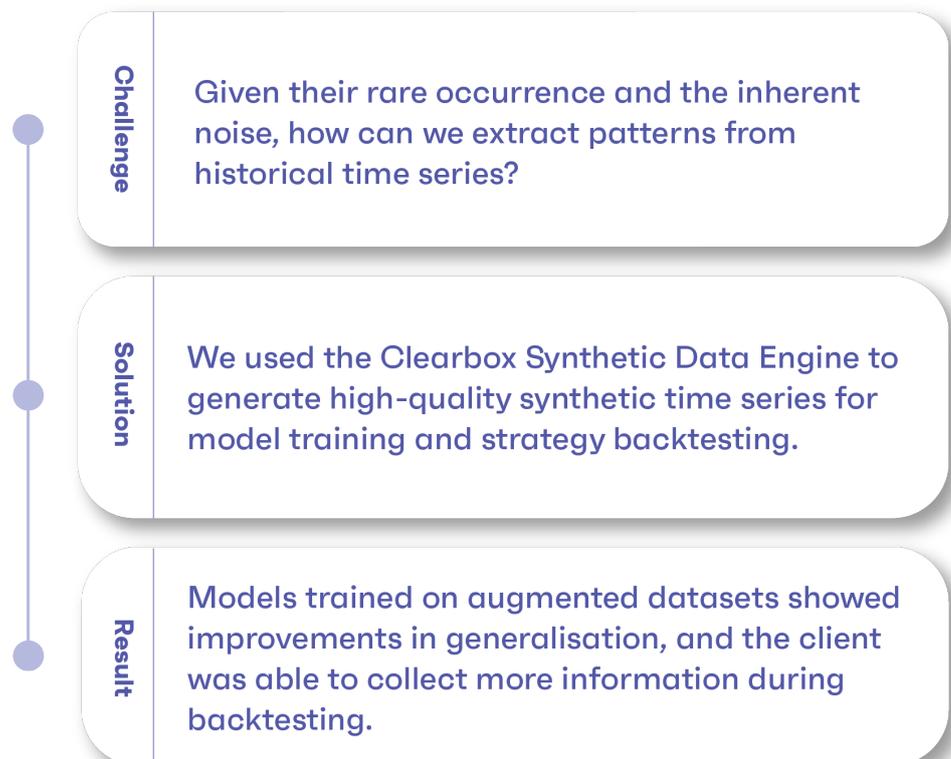
clearbox.ai

## Introduction

**Financial time series** are beneficial to **train machine learning models** deployed as trading agents, market indicators and forecasting tools.

Nevertheless, a few **issues** may appear: historical financial data presents itself with a high degree of **noise**, and forecast targets often represent **rare events**. For these reasons, models, especially deep-learning ones, tend to **overfit** and struggle at **generalising**.

This use case displays our work with a **financial trading company**, showing how **synthetic data** could effectively help **improve models** and that companies can also use it to enhance **strategy backtesting**.



## Challenge

**Machine learning** proved to be a **valuable asset** in the **financial trading** field due to the vast historical financial data available. Yet, a few **critical issues** still affect machine learning models: data often shows much **aleatoric noise**, and the design of many ML agents is to predict events characterised by very **low occurrences**.

Furthermore, markets keep changing due to small and large **scale events**. That's why models defined by many parameters such as deep learning architectures tend to **overfit**, i.e. they perform very well in terms of validation metrics, but they stop accurately working when presented with Out Of Sample test data. In addition, **backtesting models** over short time intervals provides limited information about the model behaviour.

## Solution

We helped the client solve this challenge through our Synthetic Data Engine, **generating multivariate synthetic time series** representing several financial instruments over time. We segmented the original data in an unsupervised way through the **data profiling and assessment tool** while determining **anomalies** and **outliers**. This information **facilitates the parametrisation** of the synthetic output, for example, by focusing on particular market trends.

We then used the Synthetic Data Engine to generate historical time series and create data points to **augment the supervised training process**. To obtain the augmented dataset, we assembled minority samples to **help with class imbalance issues** and **populated data segments** associated with low accuracy.

## Result

The process included the use of synthetic time series in two different areas. First, we used the **synthetic data to improve model training**, helping with target class imbalance and parametric regularisation. Second, we fed the same series to a market simulator to **generate additional backtesting output**.

The use case showed that **synthetic data helped improve model regularisation** - and thus **performance - for Out Of Sample points**, especially for imbalanced prediction targets. At the same time, extending the backtesting analysis to additional synthetic scenarios made **model behaviour analysis more comprehensive**. It made it possible to extract rule-based trading strategies from ML trading agents.

humans && machines

C.so Castelfidardo,30/a  
10129, Torino (Italy)  
info@clearbox.ai

VAT ID: (IT)12161430017

clearbox.ai