

Data and machine learning

This is one in a series of white papers intended to help investors educate themselves on funds and investment techniques in order to make the right decisions regarding their assets.

In this paper we provide a guide to machine learning techniques which are behind the statistically based algorithms that drive much of the innovation in fund management. The success of any machine learning technique is completely reliant on its input data and so we also look at some of the different types of data applied to AI decision making in finance.

1 Machine learning in investment

Machine learning techniques used in finance and investment combine the classical trio of supervised, unsupervised and reinforcement learning in addition to the disruptors of deep learning, adversarial learning and transfer/meta learning [Koshiyama et al, 2020; 2019; 2018; Treleaven et al, 2019].

1.1 Classic Machine Learning (ML) algorithms

ML algorithms can learn without explicit programming and adapt when exposed to new data. Classic techniques include

- **Supervised learning** – creates a prediction function based on a large set of training examples with labelled inputs and outputs. This class includes many different techniques, for example regression, random forest and neural networks.
- **Unsupervised learning** – infers information about a data set without labelled outputs. Very often this entails inferring similarities between data points and involves sorting data into clusters.
- **Semi-supervised learning** – unsurprisingly this is a combination of the previous two situations whereby only some of the examples are labelled
- **Reinforcement learning** – rather than being based on a system of optimal classification, reinforcement learning maximises a reward function for a number of future actions based on its knowledge about the system.

Classic techniques can be classified as to whether they are online (i.e., every example is used immediately) or offline where pre-existing data is processed in a batch.

1.2 Disruptor ML

Although newer ML techniques in FinTech can be termed as disruptors, some are fairly well-established ML techniques, for example deep learning. Some disruptor ML is less “explainable” compared to classical techniques which can pose a challenge from both a regulatory and transparency point of view.

- **Deep learning** - deep learning algorithms attempt to model complex data by abstracting it using multi-layer neural networks. In general, neural networks with more than three layers are referred to as “deep” but many more layers than this can be used in practice. The layers between the input and output are referred to as hidden. These hidden layers mean that the operation of deep learning ML algorithms can suffer from a lack of explainability.

- **Adversarial learning** - adversarial machine learning uses “competing” ML systems to disrupt the operation of another ML algorithm, i.e., “the adversary”.
- **Transfer/meta learning** – these paradigms encapsulate knowledge learned across many tasks and transfer it to new domains. Specifically this can be applied to the problem of few-shot learning where there are too few labelled samples to train a classic machine learning model. Transfer learning applies an existing model to a new problem with similar parameters. Meta-Learning generalises models to unseen tasks; this can sometimes be referred to as training a model to learn.

1.3 Combination

The interaction of classic ML and disrupter ML constantly yields new models. Examples include:

- **Long Short-Term Memory (LSTMs)** - a Recurrent Neural Network (RNN) architecture with feedback which can be used in prediction of time series data. However, the literature shows limited success in stock prediction thus far.
- **Generative adversarial networks (GANs)** – an implementation of Adversarial Learning, described above, in which two neural networks compete for prediction accuracy.
- **Transformers** – these are particularly applicable for text understanding and translation as ML models which learn relationships between words in sentences, for example the first word “red” in the English expression “red flag” translates to the second word in the Italian translation “[bandiera rossa](#)”. An example is Generative pre-trained transformer 3 (GPT-3) an autoregressive language model that uses deep learning to produce human-like text.
- **Bidirectional encoder representations from transformers (BERT)** - a neural network-based technique for natural language processing pre-training. A popular variant is FinBERT, a pre-trained language model for analysing financial texts.

1.4 Decision making using ML

Many machine learning methods, once trained, produce decisions in quicker time than the competing techniques using traditional numerical methods (see for example the work by [\[Bayer et al. \(2019\)\]](#)). For methods where speed is critical, machine learning decision making can form part of a completely automated trading system. Indeed some FinTech participants do operate in a completely automated manner. A competing pressure is that of explainability of decisions, both for the purposes of regulation and investor transparency. Explainable AI is in its infancy as a field and so many funds use the AI to create investment strategies, but with the decision to trade ultimately controlled by human intervention. This can sometimes go under the name of augmented intelligence.

2 Augmented Intelligence and Explainable AI

Partly driven by the difficulty of understanding, and by extension regulating, AI systems which operate as a “black-box”, interest has been growing in the areas of Augmented Intelligence and

Explainable AI (XAI). This is especially true in areas which are heavy with regulation such as healthcare and finance.

Explainable AI means that the system is designed in such a way in addition to the required output (e.g. whether to invest in a stock), it also generates signals which can be used to explain why this decision has been made. One drawback of this is that the performance of such systems has up to now been poorer than purely black-box systems. However they can still be useful in some areas. [\[Alkhalidi, 2021\]](#).

In contrast Augmented Intelligence describes a system which has been designed to work alongside human experts. Quite often these experts will have some control over the operation of the system, e.g. the ability to limit the system to a certain solution space. Alternatively the final decision (e.g. whether to invest in a stock) is made by a human but this is informed by signals generated by an AI system.

3 Data

AI algorithms and machine learning have allowed more information to be extracted from existing data sources. However, the increasing availability of Big data is equally important, possibly more so, in recent developments in FinTech [\[Recce, 2020\]](#). This is down to both the digitisation and standardisation of existing data sources and the opening up of new data sources via APIs or other access methods. We present an overview of different data types in the following sections, considering both traditional and alternative data types [\[Taysom et al, 2021\]](#).

3.1 Traditional Data

Traditional financial data includes market data, published economic and corporate reporting. These can be subdivided into the following sections

- **Market data** – Asset price data (and derivatives thereof), order volume and flow data. This can include the structure of the limit order book on traditional exchanges.
- **Corporate “Fundamental” data** – Corporate reporting data (including reported ESG data and analyst recommendations)
- **Economic “Macro” and political “Facts” data** – Reported economic data (for example GDP and inflation data).
- **Calendar and Events data** - information about holidays, non-trading days, et. all. as well as corporate splits, mergers, acquisitions, dividends, maturities and similar.
- **Analyst and Agency Ratings** - buy, hold, sell analyst scores as well as agency ratings e.g. AAA

3.2 Alternative Data

Emerging new datasets are referred to as alternative data or ‘alt data’. These can be considered ‘digital residue’ from a world increasingly online and connected [\[Recce, 2020\]](#). One key area is data extracted from text – such as internal business data, political data and data from media, reports and letters. Once extracted, Natural Language Processing (NLP) and sentiment analysis is applied.

Several different taxonomies of the emerging data space exist – for instance by source [Kolanovic & Krishnamachari, 2017] and attributes (e.g. the popular 3Vs, or [Monk, Prins & Rook's, 2019] 6-dimensional schema). As per [Taysom et al, 2021], we divide the data types into the following sections:

- **Satellite and Geographic Information Systems (GIS) location** data – arial images, shipping data, mobile phone location data.
- **Consumer data** – Anonymized credit card receipts, in-app smartphone purchases, website traffic, IoT IP addresses, online search data, survey data.
- **Personal data** – Social media connections, work experience and education data.
- **Themes and sentiment** – Themes and sentiment extracted from news, corporate reporting, social media, academic literature, online search data.
- **Interpersonal and interactions** – Analysis of sentiment and psychometric profiling based on different forms of communications.
- **Internal data** – Organisations extracting themes from their own internal data such as emails and notes, and also metadata from internal work and previous decisions.

4 Financial Science

The Financial Science research initiative was set up to investigate and optimise an algorithmic investment strategy focused on resilience and long term capital growth. Its operational procedures centre on an evolving research pipeline which feeds latest study findings and empirical results into the development process. In many funds the algorithmic approach is to generate signals which provide a model of properties of the future price trajectory such as growth or volatility. Literature has shown that even very sophisticated machine learning models can struggle to predict these values with any great accuracy. Despite these inaccuracies, the portfolio allocation function of such funds will then use these values to infer the desirability of stocks as investment vehicles.

The Financial Science approach takes the alternate methodology of allocating scores to stocks to directly reflect the desirability of the stock as part of a portfolio. One advantage is that score generation can take into account any number of different quantities that can reflect a stock's appeal. A second advantage is that the allocation function can then directly optimise for the attractiveness of the stocks as investment vehicles, rather than having to infer this from other quantities.

4.1 Fund operation


The main operational process components for the Financial Science process are therefore:

- **Feature ingestion and factor generation.** These extract multiple features from the data including, but not limited to, stock prices, volatility and key microeconomic ratios and combine them to form factors which represent a relative score of a stock's performance with respect to a stock's evaluation criteria which matches our risk-return profile. This process is under constant improvement from our research pipeline which analyses new features and factors.
- **Pre-trade analysis** examines the stock universe based on criteria including quality of the fundamentals of the relevant companies. The aim is to maximise the size of an available stock universe whilst controlling for risk. There are a number of research pipelines

concerning the effectiveness of these screeners and whether the stock universe can be optimally separated.

- **Signal generation** generated features are combined into a single rating of overall desirability of the stock according to evaluation criteria. Our research pipeline investigates different weights assigned to individual factors and whether they can be optimised, especially under varying market conditions in different sectors across diverse markets.
- **Asset allocation** The score feeds directly into the allocation algorithm which optimises combined scores subject to constraints aspects such as cash ratio and diversification. Our research has evaluated several allocation techniques [Rutkowska et al, 2016; 2020] and these findings have fed directly into the operation of the fund.
- The asset allocation methodology is fully integrated with the risk profile of the fund, and diversification is incorporated into the process, rather than as an add-on. The asset allocation operates with a selected utility function, as this is concave, representing risk aversion, the allocation favours consistent returns.
- **Post-trade analysis** In keeping with the transparent philosophy, the process is subject to a robust reporting regime. Our research-based findings are published as academic papers, industry-orientated white papers and with information directly available for which day-to-day management is supported by reporting tearsheets.
- **Optimisation, versioning and model-drift controls** ensure that the appropriate up to date models are applied and compared.
- **Transparent fees application** provides best value for long term growth.

5 References

- [1] Adriano Koshiyama, Nick Firoozye, and Philip Treleaven. Generative adversarial networks for financial trading strategies fine-tuning and combination. Quantitative Finance, pages 1–17, 2020.
- [2] Adriano Koshiyama and Nick Firoozye. Avoiding backtesting overfitting by covariance-penalties: an empirical investigation of the ordinary and total least squares.cases..The Journal of Financial Data Science, 1(4):63–83, 2019.
- [3] Adriano Soares Koshiyama, Nick Firoozye, and Philip Treleaven. A machine learning-based recommendation system for swaptions strategies. arXiv preprint arXiv:1810.02125, 2018.
- [4] Philip Treleaven, Jeremy Barnett, and Adriano Koshiyama. Algorithms:law and regulation. Computer, 52(2):32–40, 2019.
- [5] Christian Bayer, Blanka Horvath, Aitor Muguruza, Benjamin Stemper, and Mehdi Tomas “On deep calibration of (rough) stochastic volatility models” arXiv.preprint arXiv:1908.08806, 2019.
- [6] Alkhaldi, Nadejda, “What should your company know about explainable AI and its principles?”, Innovation Analyst Published on August 26, 2021
- [7]  Michael Recce - Fundamental Valuation of Companies Using New Data and Quant Methods , 2021
- [8] Taysom et al 2021 TBA
- [9] Kolanovic, M., & Krishnamachari , R. (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. White paper, JP Morgan, Quantitative and Derivatives Strategy.
- [10] Monk, A., Prins, M., & Rook, D. (2019). Rethinking Alternative Data in Institutional Investment. The Journal of Financial Data Science.

[11] Rutkowska, Aleksandra. "Influence of membership function's shape on portfolio optimization results." *Journal of Artificial Intelligence and Soft Computing Research* 6.1 (2016): 45-54.

[12] Bartkowiak, Marcin, and Aleksandra Rutkowska. "Black-Litterman Model with Multiple Experts' Linguistic Views." *International Conference on Soft Methods in Probability and Statistics*. Springer, Cham, 2016.

[13] Bartkowiak, Marcin, and Aleksandra Rutkowska. "Vague expert information/recommendation in portfolio optimization-an empirical study." *Axioms* 9.2 (2020): 38.