HANDBOOK OF ARTIFICIAL INTELLIGENCE AND BIG DATA APPLICATIONS IN INVESTMENTS

I. MACHINE LEARNING AND DATA SCIENCE APPLICATIONS IN INVESTMENTS

This book can be found at cfainstitute.org/ai-and-big-data

1. ON MACHINE LEARNING APPLICATIONS IN INVESTMENTS

Mike Chen, PhD
Head, Alternative Alpha Research, Robeco

Weili Zhou, CFA
Head, Quant Equity Research, Robeco

Introduction

In recent years, machine learning (ML) has been a popular technique in various domains, ranging from streaming video and online shopping recommendations to image detection and generation to autonomous driving. The attraction and desire to apply machine learning in finance are no different.

- "The global AI fintech market is predicted to grow at a CAGR of 25.3% between 2022 and 2027" (Columbus 2020).
- "A survey of IT executives in banking finds that 85% have a 'clear strategy' for adopting AI in developing new products and services" (Nadeem 2018).

Putting aside the common and widespread confusion between artificial intelligence (AI) and ML (see, e.g., Cao 2018; Nadeem 2018), the growth of ML in finance is projected to be much faster than that of the overall industry itself, as the previous quotes suggest. Faced with this outlook, practitioners may want answers to the following questions:

- What does ML bring to the table compared with traditional techniques?
- How do I make ML for finance work? Are there special considerations? What are some common pitfalls?
- What are some examples of ML applied to finance?

In this chapter, we explore how ML can be applied from a practitioner's perspective and attempt to answer many common questions, including the ones above.[1]

The first section of the chapter discusses practitioners' motivations for using ML, common challenges in implementing ML for finance, and solutions. The second section discusses several concrete examples of ML applications in finance and, in particular, equity investments.

Motivations, Challenges, and Solutions in Applying ML in Investments

In this section, we discuss reasons for applying ML, the unique challenges involved, and how to avoid common pitfalls in the process.

Motivations

The primary attraction of applying ML to equity investing, as with almost all investment-related endeavors, is the promise of higher risk-adjusted return. The hypothesis is that these techniques, explicitly designed for prediction tasks based on high-dimensional data and without any functional form specification, should excel at predicting future equity returns.

Emerging academic literature and collective practitioner experience support this hypothesis. In recent years, practitioners have successfully applied ML algorithms to predict equity returns, and ML-based return prediction algorithms have been making their way into quantitative investment models. These algorithms have been used worldwide in both developed and emerging markets, for large-cap and small-cap investment universes, and with single-country or multi-country strategies.[2] In general, practitioners have found that ML-derived alpha models outperform those generated from more traditional linear models[3] in predicting cross-sectional equity returns.

[1]Readers interested in the theoretical underpinnings of ML algorithms, such as random forest or neural networks, should read Hastie, Tibshirani, and Friedman (2009) and Goodfellow, Bengio, and Courville (2016).

[2]There are also numerous academic studies on using ML to predict returns. For example, ML techniques have been applied in a single-country setting by Gu, Kelly, and Xiu (2020) to the United States, by Abe and Nakayama (2018) to Japan, and by Leippold, Wang, and Zhou (2022) to China's A-share markets. Similarly, in a multi-country/regional setting, ML has been applied by Tobek and Hronec (2021) and Leung, Lohre, Mischlich, Shea, and Stroh (2021) to developed markets and by Hanauer and Kalsbach (2022) to emerging markets.

[3]For linear equity models, see, for example, Grinold and Kahn (1999).

In addition to predicting equity returns, ML has been used to predict intermediate metrics known to predict future returns. For example, practitioners have used ML to forecast corporate earnings and have found ML-derived forecasts to be significantly more accurate and informative than other commonly used earnings prediction models.[4] Another use of ML in equity investing developed by Robeco's quantitative researchers has been to predict not the entire investment universe's return but the returns of those equities that are likely to suffer a severe price drop in the near future. Investment teams at Robeco have found that ML techniques generate superior crash predictions compared with those from linear models using traditional metrics, such as leverage ratio or distance to default.[5]

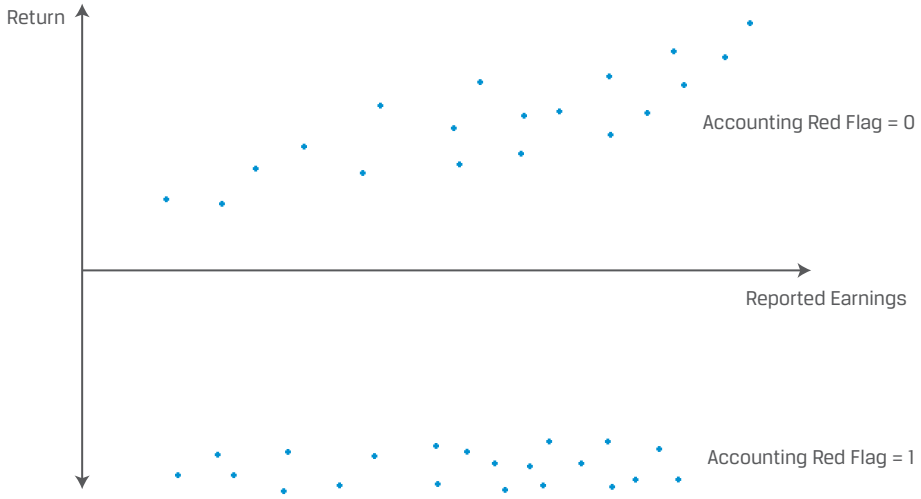
What drives the outperformance of ML over other known quantitative techniques? The main conclusion from practitioners and academics is that because ML algorithms do not prespecify the functional relationship between the prediction variables (equity return, future earnings, etc.) and the predictors (metrics from financial statements, past returns, etc.), ML algorithms are not constrained to a linear format as is typical of other techniques but, rather, can uncover interaction and nonlinear relationships between the input features and the output variable(s).

Interaction and nonlinear effects

The interaction effect occurs when the prediction outputs cannot be expressed as a linear combination of the individual inputs because the effect of one input depends on the value of the other ones. Consider a stylistic example of predicting equity price based on two input features: reported earnings and an accounting red flag, where the red-flag input is binary: 0 (no cause of concern) and 1 (grave concern). The resulting ML output with these two inputs may be that when the red-flag input is 0, the output is linearly and positively related to reported earnings; in contrast, when the red-flag input is 1, the output is a 50% decrease in price regardless of the reported earnings. **Exhibit 1** illustrates this stylistic example.

ML prediction can also outperform the traditional linear model prediction due to nonlinear effects. There are many empirically observed nonlinear effects that linear models cannot model. For example, there is a nonlinear relationship between a firm's credit default swap (CDS) spread and its equity returns because equity can be framed as an embedded call option on a firm's assets, thereby introducing nonlinearity.[6] **Exhibit 2** illustrates this example. Many ML algorithms, particularly neural networks,[7] explicitly

Exhibit 1. Illustration of the Interaction Effect between Accounting Red Flags and Equity Returns



Source: Robeco.

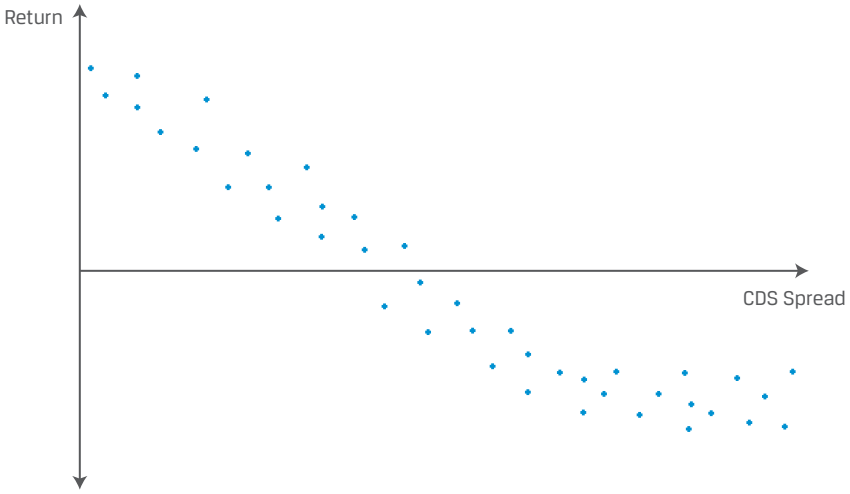
[4]This conclusion was replicated and supported also by academics. For example, see Cao and You (2021).

[5]For more information on using ML for crash prediction, see Swinkels and Hoogteijling (2022).

[6]For more details, see Birge and Zhang (2018).

[7]Neural networks incorporate nonlinearity through activation functions in each neuron. Without activation functions, neural networks, regardless of their depth, reduce down to traditional linear models commonly used in finance. For more on activation functions, see Goodfellow et al. (2016).

Exhibit 2. Illustration of the Nonlinear Relationship between a CDS and Equity Returns



Source: Robeco.

introduce nonlinearity into the model setup, facilitating nonlinearity modeling.

Empirically, academics and practitioners have found that the interaction effect accounts for a large portion of ML techniques' outperformance, while the jury is still out on whether nonlinear effects contribute positively to the outperformance. A well-known study in the field by Gu, Kelly, and Xiu (2020, p. 2258) found that "the favorable performance of [ML algorithms] indicates a benefit to allowing for potentially complex interactions among predictors." In the study, comparing the performance of purely linear models with that of generalized linear models, the authors note that "the generalized linear model … fails to improve on the performance of purely linear methods (R^2_{OOS} of 0.19%). The fact that this method uses spline functions of individual features, but includes no interaction among features, suggests that univariate expansions provide little incremental information beyond the [purely] linear model" (Gu et al. 2020, p. 2251). However, other studies have found that both interaction and nonlinear effects contribute positively to ML models' outperformance (see, e.g., Abe and Nakayama 2018; Swinkels and Hoogteijling 2022; Choi, Jiang, and Zhang 2022).

Find relationships from the data deluge

Another attraction of applying ML to financial markets is the promise of having the algorithm discover relationships not specified or perhaps not known by academics and practitioners—that is, *data mining*, which historically has been a pejorative in quantitative finance circles. Another term being used, perhaps with a more positive connotation, is "knowledge discovery."

With the ongoing information and computing revolution and the increased popularity of quantitative finance, the amount of financial data is growing at a rapid pace. This increased amount of data may or may not embody relevant information for investment purposes. Since many of the data types and sources are new, many investors do not have a strong prior opinion on whether and how they can be useful. Thus, ML algorithms that are designed to look for relationships have become attractive for practitioners and academics in the hope that, even without specifying a hypothesis on the economic relationship, the ML algorithm will figure out the link between inputs and outputs. Although ML algorithms have built-in mechanisms to combat overfitting or discovering spurious correlations between input features and output predictions,[8] caution must still be taken to avoid discovering nonrepeatable and nonreplicable relationships and patterns. Later in this chapter, we will address this issue further and consider other challenges practitioners face when implementing ML in live portfolios. But first, we will discuss what makes the financial market different from other domains in which ML has shown tremendous success.

Unique Challenges of Applying ML in Finance

When applying ML for investments, great care must be taken because financial markets differ from domains where

[8]Examples include *k*-fold cross-validation, dropout, and regularization. For deeper discussions, see Hastie et al. (2009) and Goodfellow et al. (2016).

ML has made tremendous strides. These differences can mitigate many of the specific advantages ML algorithms enjoy, making them less effective in practice when applied in real-life situations. A few of these differences follow.

Signal-to-noise ratio and system complexity

Financial data have a low signal-to-noise ratio. For a given security, any one metric is generally not a huge determinant of how that security will perform. For example, suppose Company XYZ announced great earnings this quarter. Its price can still go down after the announcement because earnings were below expectations, because central banks are hiking interest rates, or because investors are generally long the security and are looking to reduce their positions. Compare this situation with a high signal-to-noise domain—for example, streaming video recommendation systems. If a person watches many movies in a specific genre, chances are high that the person will also like other movies in that same genre. Because financial returns compress high-dimensional information and drivers (company-specific, macro, behavioral, market positioning, etc.) into one dimension, positive or negative, the signal-to-noise ratio of any particular information item is generally low.

It is fair to say that the financial market is one of the most complex man-made systems in the world. And this complexity and low signal-to-noise ratio can cause issues when ML algorithms are not applied skillfully. Although ML algorithms are adept at detecting complex relationships, the complexity of the financial market and the low signal-to-noise ratio that characterizes it can still pose a problem because they make the true relationship between drivers of security return and the outcome difficult to detect.

Small vs. big data

Another major challenge in applying ML in financial markets is the amount of available data. The amount of financial data is still relatively small compared with many domains in which ML has thrived, such as the consumer internet domain or the physical sciences. The data that quantitative investors traditionally have used are typically quarterly or monthly. And even for the United States, the market with the longest available reliable data, going back 100 years, the number of monthly data points for any security we might wish to consider is at most 1,200. Compared with other domains where the amount of data is in the billions and trillions, the quantity of financial data is minuscule. To be fair, some of the newer data sources, or "alternative data," such as social media posts or news articles, are much more abundant than traditional financial data. However, overall, the amount of financial data is still small compared with other domains.

The small amount of financial data is a challenge to ML applications because a significant driver of an ML algorithm's power is the amount of available data (see Halevy, Norvig, and Pereira 2009). Between a simple ML algorithm trained on a large set of data versus a sophisticated ML algorithm trained on a relatively smaller set of data, the simpler algorithm often outperforms in real-life testing. With a large set of data, investors applying ML can perform true cross-validation and out-of-sample testing to minimize overfitting by dividing input data into different segments. The investor can conduct proper hyperparameter tuning and robustness checks only if the amount of data is large enough. The small amount of financial data adds to the challenges of applying ML to financial markets mentioned earlier—the financial markets' high system complexity and low signal-to-noise ratio.

Stationarity vs. adaptive market, irrationality

Finally, what makes financial markets challenging for ML application in general is that markets are nonstationary. What we mean is that financial markets adapt and change over time. Many other domains where ML algorithms have shined are static systems. For example, the rules governing protein unfolding are likely to stay constant regardless of whether researchers understand them. In contrast, because of the promised rewards, financial markets "learn" as investors learn what works over time and change their behavior, thereby changing the behavior of the overall market. Furthermore, because academics have been successful over the last few decades in discovering and publishing drivers of market returns—for example, Fama and French (1993)—their research also increased knowledge of *all* market participants and such research changes market behavior, as noted by McLean and Pontiff (2016).

The adaptive nature of financial markets means not only that ML algorithms trained for investment do not have a long enough history with which to train the model and a low signal-to-noise ratio to contend with but also that the rules and dynamics that govern the outcome the models try to predict also change over time. Luckily, many ML algorithms are adaptive or can be designed to adapt to evolving systems. Still, the changing system calls into question the validity and applicability of historical data that can be used to train the algorithms—data series that were not long enough to begin with. To further complicate the issue, financial markets are man-made systems. Their results are the collective of individual human actions, and human beings often behave irrationally—for example, making decisions based on greed or fear.[9] This irrationality characteristic does not exist in many of the other domains in which ML has succeeded.

[9]There have been various studies on how greed and fear affect market participants' decision-making process. See, for example, Lo, Repin, and Steenbarger (2005).

How to Avoid Common Pitfalls When Applying ML in Finance

This section addresses some potential pitfalls when applying ML to financial investment.[10]

Overfitting

Because of the short sample data history, overfitting is a significant concern when applying ML techniques in finance. This concern is even stronger than when applying traditional quantitative techniques, such as linear regression, because of the high degrees of freedom inherent in ML algorithms. The result of overfitting is that one may get a fantastic backtest, but out-of-sample, the results will not live up to expectations.

There are some common techniques used in ML across all domains to combat overfitting. They include cross-validation, feature selection and removal, regularization, early stopping, ensembling, and having holdout data.[11] Because of the lower sample data availability in finance, some of these standard techniques might not be applicable or work as well in the financial domain as in others.

However, there are also advantages to working in the financial domain. The most significant advantage is human intuition and economic domain knowledge. What we mean by this is that investors and researchers applying machine learning can conduct "smell tests" to see whether the relationships found by ML algorithms between input features and output predictions make intuitive or economic sense. For example, to examine the relationship between inputs and outputs, one can look at Shapley additive explanation (SHAP) value, introduced by Lundberg and Lee (2017). SHAP value is computed from the average of the marginal contribution of the feature when predicting the targets, where the marginal contribution is computed by comparing the performance after withholding that variable from the feature set versus the feature set that includes the variable.

Exhibit 3 plots SHAP values from Robeco's work on using ML to predict equity crashes,[12] where various input features are used to predict the probability of financial distress of various stocks in the investment universe. The color of each dot indicates the sign and magnitude of a feature, where red signals a high feature value and blue denotes a low feature value. Take Feature 25, for example. As the feature value increases (as indicated by the color red), the ML algorithm predicts a higher probability of financial distress. And as the feature value decreases (as indicated by the color blue), the ML algorithm predicts a lower probability of financial distress. With this information, experienced investors can apply their domain knowledge to see whether the relationship discovered by the ML algorithm makes sense. For example, if Feature 25, in this case, is the leverage ratio and Feature 1 is distance to default,[13] then the relationship may make sense. However, if the features are flipped (Feature 25 is distance to default and Feature 1 is the leverage ratio), then it is likely that the ML algorithm made a mistake, possibly through overfitting.

Another approach to mitigate overfitting and having ML algorithms find spurious correlations is to try to eliminate it from the start. *Ex post* explanation via SHAP values and other techniques is useful, but investors can also apply their investment domain knowledge to curate the input set to select those inputs likely to have a relationship with the prediction objective. This is called "feature engineering" in ML lingo and requires financial domain knowledge. As an example, when we are trying to predict stock crash probability, fundamental financial metrics such as profit margin and debt coverage ratio are sensible input features for the ML algorithm, but the first letter of the last name of the CEO, for example, is likely not a sensible input feature.

Replicability

There is a debate about whether there is a replicability crisis in financial research.[14] The concerns about replicability are especially relevant to results derived from applying ML because, in addition to the usual reasons for replicability difficulties (differences in universe tested, *p*-hacking, etc.), replicating ML-derived results also faces the following challenges:

- The number of tunable variables in ML algorithms is even larger than in the more traditional statistical techniques.
- ML algorithms are readily available online. Investors can often download open-sourced algorithms from the internet without knowing the algorithm's specific version used in the original result, random seed, and so on. In addition, if one is not careful, different implementations of the same algorithm can have subtle differences that result in different outcomes.

To avoid replicability challenges, we suggest ML investors first spend time building up a robust data and code

[10]For additional readings, see, for example, López de Prado (2018); Arnott, Harvey, and Markowitz (2019); Leung et al. (2021).

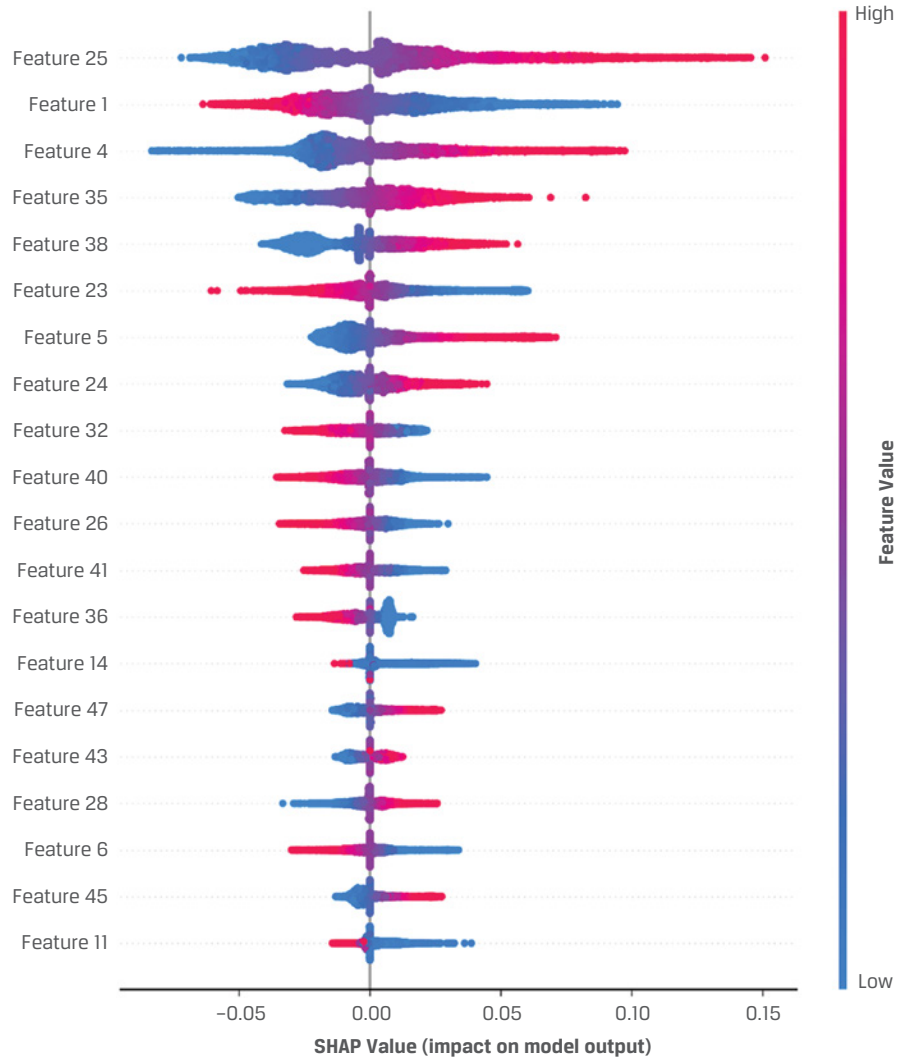
[11]See Hastie et al. (2009) and Goodfellow et al. (2016) for more discussion on these techniques.

[12]For more details, see Swinkels and Hoogteijling (2022).

[13]See Jessen and Lando (2013) for more discussion on distance to default.

[14]See Harvey, Liu, and Zhu (2016) and Hou, Xue, and Zhang (2020) for more discussion on the replicability crisis in finance.

Exhibit 3. SHAP Value between Input Features and ML Output Predicting Financial Distress



Source: Swinkels and Hoogteijling (2022).

infrastructure to conduct ML research and experiments. This includes, but is not limited to, the following:

- A robust system for code version control, a way to specify and fix all parameters used in the algorithm, including ML algorithm version number, investment universe tested, hyperparameters, feature sets used, and so on
- Documentation of all the tested iterations and hypotheses, including those that failed to show good results (in other words, showing the research graveyards)

Such documentation listed above may not be possible regarding publicly disclosed results, but it should at least be attempted when discussing and verifying results within the same organization.

Lookahead bias/data leakage

Lookahead bias is another commonly known issue for experienced quant investors that applies to ML. An example would be that if quarterly results are available only 40 days after the quarter end, the quarterly data should be used only when available historically[15] and not at the quarter end date.

[15]This is called "point-in-time" in the quant investment industry.

Other cases can be more subtle. For example, if one conducts research over a period that includes the tech bubble and its subsequent crash (2000–2002) and the investment universe tested does not include information technology (IT) firms, then it is incumbent upon the investor to provide a sensible reason why IT firms are not included.

Related to lookahead bias and a more general problem directly related to ML is the problem of data leakage. Data leakage occurs when data used in the training set contain information that can be used to infer the prediction, information that would otherwise not be available to the ML model in live production.

Because financial data occur in time series, they are often divided chronologically into training, validation, and testing sets. ML predictive modeling aims to predict outcomes the model has not seen before. One form of data leakage can occur if information that should be in one set ends up in another among the three sets of data (training, validation, and testing). When this occurs, the ML algorithm is evaluated on the data it has seen. In such cases, the results will be overly optimistic and true out-of-sample performance (that is, the live performance of the algorithm) will likely disappoint. This phenomenon is called "leakage in data."

Another type of data leakage is called "leakage in features." Leakage in features occurs when informative features about the prediction outcome are included but would otherwise be unavailable to ML models at production time, even if the information is not in the future. For example, suppose a company's key executive is experiencing serious health issues but the executive has not disclosed this fact to the general public. In that case, including that information in the ML feature set may generate strong backtest performance, but it would be an example of feature leakage.

Various techniques can be applied to minimize the possibility of data leakage. One of the most basic is to introduce a sufficient gap period between training and validation sets and between validation and testing. For example, instead of having validation sets begin immediately after the end of the training set, introduce a time gap of between a few months and a few quarters to ensure complete separation of data. To prevent data leakage, investors applying ML should think carefully about what is available and what is not during the model development and testing phases. In short, common sense still needs to prevail when applying ML algorithms.

Implementation gap

Another possible pitfall to watch out for when deploying ML algorithms in finance is the so-called implementation gap, defined as trading instruments in the backtest that are either impossible or infeasible to trade in live production. An example of an implementation gap is that the ML algorithm generates its outperformance in the backtest mainly from small- or micro-cap stocks. However, in live trading, either these small- or micro-cap stocks may be too costly to trade because of transaction costs or there might not be enough outstanding shares available to own in the scale that would make a difference to the strategy deploying the ML algorithm. As mentioned, implementation affects all quant strategies, not just those using ML algorithms. But ML algorithms tend to have high turnover, increasing trading cost associated with smaller market-cap securities. Similar to small- or micro-cap stocks, another example of an implementation gap is shorting securities in a long–short strategy. In practice, shorting stocks might be impossible or infeasible because of an insufficient quantity of a given stock to short or excessive short borrowing costs.[16]

Explainability and performance attribution

A main criticism of applying machine learning in finance is that the models are difficult to understand. According to Gu et al. (2020, p. 2258), "Machine learning models are often referred to as 'black boxes,' a term that is in some sense a misnomer, as the models are readily inspectable. However, they are complex, and this is the source of their power and opacity. Any exploration of the interaction effect is vexed by vast possibilities for identity and functional forms for interacting predictors."

Machine learning for other applications might not need to be fully explainable at all times; for example, if the ML algorithm suggests wrong video recommendations based on the viewers' past viewing history, the consequences are not catastrophic. However, with billions of investment dollars on the line, asset owners demand managers that use ML-based algorithms explain how the investment decisions are made and how performance can be attributed. In recent years, explainable machine learning has emerged in the financial domain as a focus topic for both practitioners and academics.[17]

The fundamental approach to recent explainable machine learning work is as follows:

1. For each input feature, f_i, in the ML algorithm, fix its value as x. Combine this fixed value for f_i with all other sample data while replacing feature f_i with the value x. Obtain the prediction output.
2. Average the resulting predictions. This is the partial prediction at point x.

[16]For additional readings, see, for example, Avramov, Chordia, and Goyal (2006); López de Prado (2018); Hou et al. (2020); Avramov, Cheng, and Metzker (2022).

[17]In addition to the SHAP values discussed in Lundberg and Lee (2017), recent works in the area of explainable machine learning include Li, Turkington, and Yazdani (2020); Li, Simon, and Turkington (2022); and Daul, Jaisson, and Nagy (2022).

3. Now range the fixed value x over feature f_i's typical range to plot out the resulting function. This is called the "partial dependence response" to feature f_i.

This response plot, an example of which is illustrated in **Exhibit 4**, can then be decomposed into linear and nonlinear components.

Similarly, one can estimate the pairwise-interaction part of the ML algorithm, computed using joint partial prediction of features f_i and f_j, by subtracting the partial prediction of each feature independently. An example of the pairwise-interaction result is shown in **Exhibit 5**.

Exhibits 3–5 allow ML investors to understand how the input features affect the output prediction. To conduct performance attribution of an ML portfolio and decompose it into the various parts (linear, interaction, and nonlinear), one can extract the partial dependence responses and form portfolios from them. With this approach, one can get return attribution, as shown in **Exhibit 6**.

Sample ML Applications in Finance

We have seen the common pitfalls when applying ML to finance and the strategies to mitigate them. Let us now look at examples of ML applied to financial investing.[18]

Predicting Cross-Sectional Stock Returns

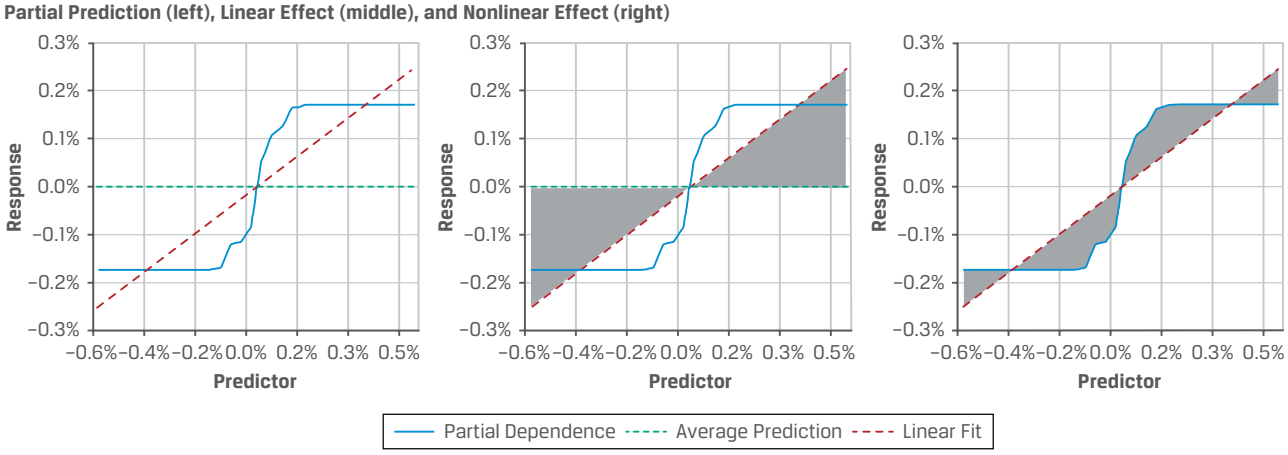
In this section, we discuss specifically using ML to predict cross-sectional stock returns.

Investment problem

Perhaps the most obvious application of ML to financial investments is to directly use ML to predict whether each security's price is expected to rise or fall and whether to buy or sell those securities. Numerous practitioners and academics have applied ML algorithms to this prediction task.[19]

The ML algorithms used in this problem are set up to compare cross-sectional stock returns. That is, we are interested in finding the relative returns of securities in our investment universe rather than their absolute returns.[20] The ML algorithms make stock selection decisions rather than country/industry timing and allocation decisions. Stock selection is an easier problem than the timing and allocation problem, because the algorithms have more data to work with.[21]

Exhibit 4. Linear and Nonlinear Decomposition of an ML Algorithm's Output to Input Feature f_i



Source: Li, Turkington, and Yazdani (2020).

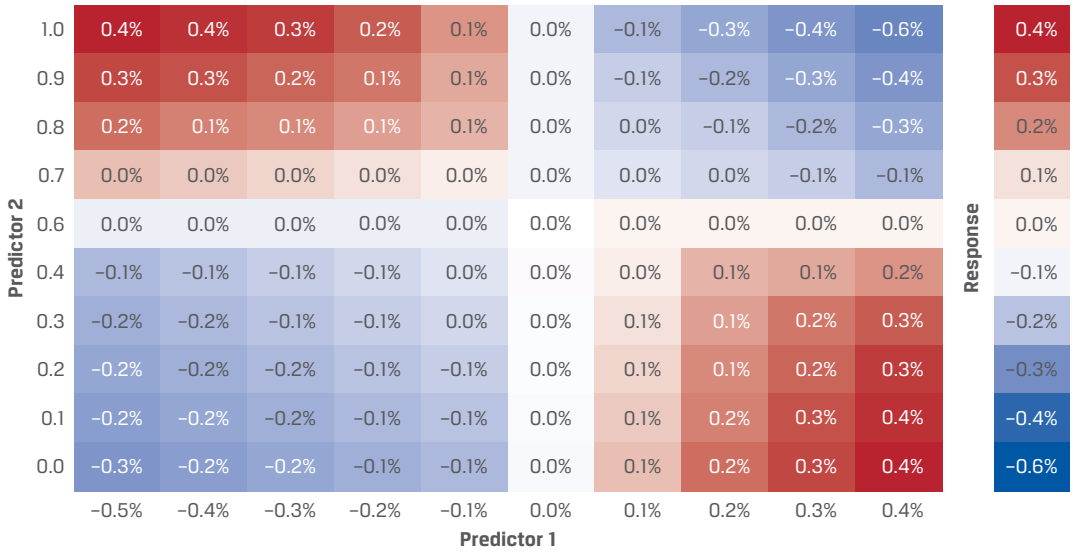
[18]For more general applications of ML to finance, see López de Prado (2019).

[19]This problem has been studied in numerous recent papers—for example, Abe and Nakayama (2018); Rasekhschaffe and Jones (2019); Gu et al. (2020); Choi, Jiang, and Zhang (2022); Hanauer and Kalsbach (2022).

[20]In addition to cross-sectional returns, Gu et al. (2020) also study the time-series problem.

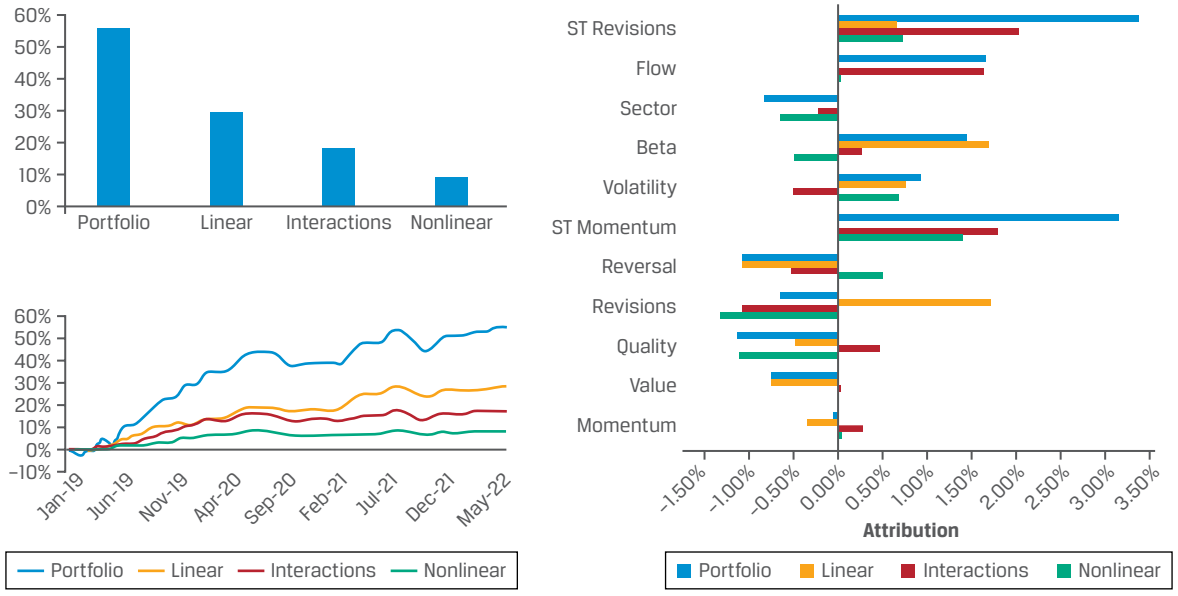
[21]However, when compared with other fields, the amount of data here is still miniscule, as noted before.

Exhibit 5. Example of the Pairwise-Interaction Effect



Source: Li, Simon, and Turkington (2022).

Exhibit 6. Example ML Portfolio Return Attribution



Source: Robeco.

Methodology

This problem is set up with the following three major components:

1. Investment universe: US, international, emerging market, and so on
2. ML algorithms: (boosted) trees/random forests, neural networks with *l* layers, and so on, and ensembles thereof
3. Feature set: typical financial metrics that are used in linear models, such as various price ratios (value), profitability (quality), and past returns (momentum)

Note that for Item 3, by using a very limited feature set, the ML investor is essentially applying her domain knowledge and imposing a structure on the ML algorithm to counter the limited data challenge discussed in the previous section.

Results

There are five consistent results that various practitioner and academic studies have found. First and foremost, there is a statistically significant and economically meaningful outperformance of ML algorithm prediction versus the traditional linear approach. For example, in forming a long–short decile spread portfolio from (four-layer) neural network–generated stock return prediction, a well-known study in equity return prediction (Gu et al. 2020) found that the strategy has an annualized out-of-sample Sharpe ratio of 1.35 under value weighting and 2.45 under equal weighting. For comparison, the same portfolio constructed from ordinary least squares (OLS) prediction with the same input features produced a Sharpe ratio of 0.61 and 0.83 for value weighting and equal weighting, respectively. **Exhibit 7** shows the value-weighting results.

Exhibit 7. Out-of-Sample Performance of Benchmark OLS Portfolio vs. Various ML Portfolios, Value Weighting

	OLS-3+H				PLS				PCR			
	Pred.	Avg.	Std. Dev.	Sharpe Ratio	Pred.	Avg.	Std. Dev.	Sharpe Ratio	Pred.	Avg.	Std. Dev.	Sharpe Ratio
Low (L)	−0.17	0.40	5.90	0.24	−0.83	0.29	5.31	0.19	−0.68	0.03	5.98	0.02
2	0.17	0.58	4.65	0.43	−0.21	0.55	4.96	0.38	−0.11	0.42	5.25	0.28
3	0.35	0.60	4.43	0.47	0.12	0.64	4.63	0.48	0.19	0.53	4.94	0.37
4	0.49	0.71	4.32	0.57	0.38	0.78	4.30	0.63	0.42	0.68	4.64	0.51
5	0.62	0.79	4.57	0.60	0.61	0.77	4.53	0.59	0.62	0.81	4.66	0.60
6	0.75	0.92	5.03	0.63	0.84	0.88	4.78	0.64	0.81	0.81	4.58	0.61
7	0.88	0.85	5.18	0.57	1.06	0.92	4.89	0.65	1.01	0.87	4.72	0.64
8	1.02	0.86	5.29	0.56	1.32	0.92	5.14	0.62	1.23	1.01	4.77	0.73
9	1.21	1.18	5.47	0.75	1.66	1.15	5.24	0.76	1.52	1.20	4.88	0.86
High (H)	1.51	1.34	5.88	0.79	2.25	1.30	5.85	0.77	2.02	1.25	5.60	0.77
H – L	1.67	0.94	5.33	0.61	3.09	1.02	4.88	0.72	2.70	1.22	4.82	0.88
	ENet+H				GLM+H				RF			
	Pred	Avg	Std. Dev.	Sharpe Ratio	Pred	Avg	Std. Dev.	Sharpe Ratio	Pred	Avg	Std. Dev.	Sharpe Ratio
Low (L)	−0.04	0.24	5.44	0.15	−0.47	0.08	5.65	0.05	0.29	−0.09	6.00	−0.05
2	0.27	0.56	4.84	0.40	0.01	0.49	4.80	0.35	0.44	0.38	5.02	0.27
3	0.44	0.53	4.50	0.40	0.29	0.65	4.52	0.50	0.53	0.64	4.70	0.48
4	0.59	0.72	4.11	0.61	0.50	0.72	4.59	0.55	0.60	0.60	4.56	0.46
5	0.73	0.72	4.42	0.57	0.68	0.70	4.55	0.53	0.67	0.57	4.51	0.44
6	0.87	0.85	4.60	0.64	0.84	0.84	4.53	0.65	0.73	0.64	4.54	0.49
7	1.01	0.87	4.75	0.64	1.00	0.86	4.82	0.62	0.80	0.67	4.65	0.50
8	1.16	0.88	5.20	0.59	1.18	0.87	5.18	0.58	0.87	1.00	4.91	0.71
9	1.36	0.80	5.61	0.50	1.40	1.04	5.44	0.66	0.96	1.23	5.59	0.76
High (H)	1.66	0.84	6.76	0.43	1.81	1.14	6.33	0.62	1.12	1.53	7.27	0.73
H – L	1.70	0.60	5.37	0.39	2.27	1.06	4.79	0.76	0.83	1.62	5.75	0.98

(*continued*)

Exhibit 7. Out-of-Sample Performance of Benchmark OLS Portfolio vs. Various ML Portfolios, Value Weighting (*continued*)

	GBRT+H				NN1				NN2			
	Pred	Avg	Std. Dev.	Sharpe Ratio	Pred	Avg	Std. Dev.	Sharpe Ratio	Pred	Avg	Std. Dev.	Sharpe Ratio
Low (L)	−0.45	0.18	5.60	0.11	−0.38	−0.29	7.02	−0.14	−0.23	−0.54	7.83	−0.24
2	−0.16	0.49	4.93	0.35	0.16	0.41	5.89	0.24	0.21	0.36	6.08	0.20
3	0.02	0.59	4.75	0.43	0.44	0.51	5.07	0.35	0.44	0.65	5.07	0.44
4	0.17	0.63	4.68	0.46	0.64	0.70	4.56	0.53	0.59	0.73	4.53	0.56
5	0.34	0.57	4.70	0.42	0.80	0.77	4.37	0.61	0.72	0.81	4.38	0.64
6	0.46	0.77	4.48	0.59	0.95	0.78	4.39	0.62	0.84	0.84	4.51	0.65
7	0.59	0.52	4.73	0.38	1.11	0.81	4.40	0.64	0.97	0.95	4.61	0.71
8	0.72	0.72	4.92	0.51	1.31	0.75	4.86	0.54	1.13	0.93	5.09	0.63
9	0.88	0.99	5.19	0.66	1.58	0.96	5.22	0.64	1.37	1.04	5.69	0.63
High (H)	1.11	1.17	5.88	0.69	2.19	1.52	6.79	0.77	1.99	1.38	6.98	0.69
H − L	1.56	0.99	4.22	0.81	2.57	1.81	5.34	1.17	2.22	1.92	5.75	1.16
	NN3				NN4				NN5			
	Pred	Avg	Std. Dev.	Sharpe Ratio	Pred	Avg	Std. Dev.	Sharpe Ratio	Pred	Avg	Std. Dev.	Sharpe Ratio
Low (L)	−0.03	−0.43	7.73	−0.19	−0.12	−0.52	7.69	−0.23	−0.23	−0.51	7.69	−0.23
2	0.34	0.30	6.38	0.16	0.30	0.33	6.16	0.19	0.23	0.31	6.10	0.17
3	0.51	0.57	5.27	0.37	0.50	0.42	5.18	0.28	0.45	0.54	5.02	0.37
4	0.63	0.66	4.69	0.49	0.62	0.60	4.51	0.46	0.60	0.67	4.47	0.52
5	0.71	0.69	4.41	0.55	0.72	0.69	4.26	0.56	0.73	0.77	4.32	0.62
6	0.79	0.76	4.46	0.59	0.81	0.84	4.46	0.65	0.85	0.86	4.35	0.68
7	0.88	0.99	4.77	0.72	0.90	0.93	4.56	0.70	0.96	0.88	4.76	0.64
8	1.00	1.09	5.47	0.69	1.03	1.08	5.13	0.73	1.11	0.94	5.17	0.63
9	1.21	1.25	5.94	0.73	1.23	1.26	5.93	0.74	1.34	1.02	6.02	0.58
High (H)	1.83	1.69	7.29	0.80	1.89	1.75	7.51	0.81	1.99	1.46	7.40	0.68
H − L	1.86	2.12	6.13	1.20	2.01	2.26	5.80	1.35	2.22	1.97	5.93	1.15

Notes: OLS-3+H is ordinary least squares that preselect size, book-to-market, and momentum using Huber loss rather than the standard l2 loss. PLS is partial least squares. PCR ENet+H, GLM+H, RF, GBRT+H, and NN1 to NN5 are neural networks with one to five hidden layers.

Source: Gu et al. (2020).

The second consistent result is that what drives the outperformance of ML-based prediction versus that of linear models is that ML algorithms not only are limited to linear combinations of feature sets but can formulate higher-order functional dependencies, such as nonlinearity and interaction. To test whether higher-order effects can contribute to security prediction, one can compare linear machine learning models (for example, LASSO and RIDGE) with models that consider nonlinear and complex interaction effects (such as trees and neural networks). Investors at Robeco have found that higher-order machine learning models outperform their simpler linear competitors. The outperformance of higher-order machine learning models was also confirmed by academics,[22] as evident from Exhibit 7.

Between nonlinearity and interactions, interactions have the greater impact on model performance. This also can be seen in Exhibit 7, where the performance of the generalized linear model with Huber loss (GLM+H) is inferior to those models that consider interaction, boosted trees, and neural networks.

Third, ML investors have found that the features that most determine prediction outcomes are remarkably consistent regardless of the specific ML algorithms used. This is somewhat of a surprise because the various ML algorithms, such as boosted trees and neural networks, use dissimilar approaches to arrive at their outcomes. However, the similar importance assigned to certain input features confirms that these are salient characteristics that drive cross-sectional stock returns. From Robeco's own experience and various published studies, the characteristics that dominate cross-sectional returns are found to be short-term reversals, stock and sector return momentum, return volatility, and firm size.

Fourth, simple ML algorithms outperform more complicated ML algorithms. This result is very likely because, since there are not much data to train on, the simpler models, due to their parsimonious nature, are less likely to overfit and thereby perform better out of sample. We confirm this observation from Exhibit 7, where the best out-of-sample Sharpe ratio is achieved by a neural network with four hidden layers.

Fifth, the more data there are, the better the ML prediction algorithm performs. This is fundamental to the nature of ML algorithms, as observed by Halevy et al. (2009) for general machine learning applications and confirmed by Choi, Jiang, and Zhang (2022) in a financial context.

Predicting Stock Crashes

In this section, we discuss the example of using ML to predict stock crashes.

Investment problem

Rather than predicting the rank order of future returns, another application for ML algorithms may simply be to predict the worst-performing stocks—that is, those most likely to crash.[23] The results of this prediction can be applied in such strategies as conservative equities,[24] where the securities most likely to crash are excluded from the investment universe. That is, win by not losing.

Methodology

A crash event for equities is defined as a significant drop in a stock's price relative to peers; thus, it is idiosyncratic rather than systematic when a group of stocks or the market crashes. The ML prediction is set up as follows:

1. Investment universe: US, international, emerging market, and so on
2. ML algorithms: logistic regression, random forest, gradient boosted tree, and ensemble of the three
3. Feature set: various financial distress indicators—such as distance to default, volatility, and market beta—in addition to traditional fundamental financial metrics

As one can see, the problem setup is very similar to that of cross-sectional stock return prediction. The main differences are the objective function of the ML algorithm (ML prediction goal) and the input feature set (feature engineering). Care should be taken in determining both components to ensure the prediction algorithm solves the stated problem and performs well out of sample.

Results

The performance of the portfolio of stocks with the highest distress probability versus that of the market is shown in **Exhibit 8**. We see that the performance of the ML algorithm is greater than that obtained using traditional approaches for both developed and emerging markets.

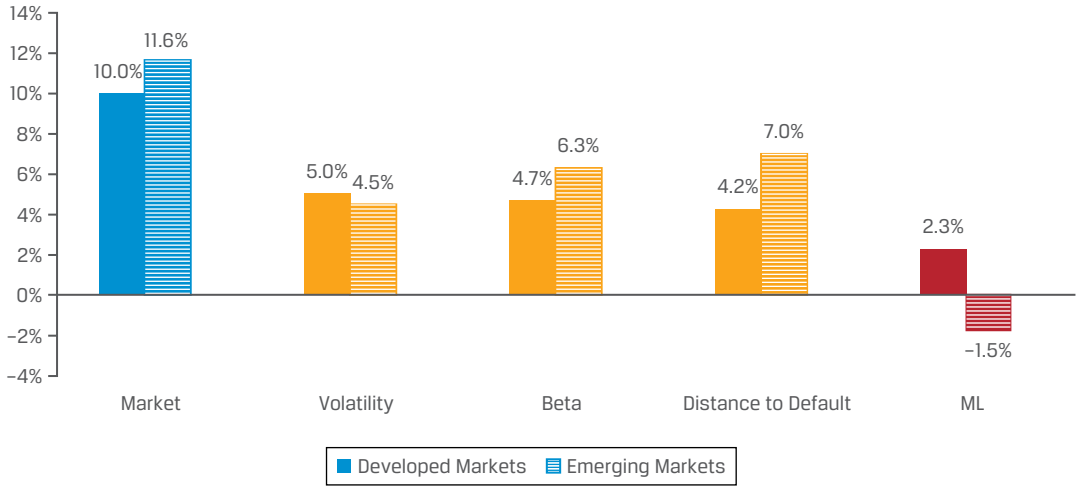
Looking at the sector composition of the likely distressed stocks, shown in **Exhibit 9**, we see that the ML algorithm choices are reasonable, as technology stocks dominated during the bursting of the dot-com bubble in the early 2000s and financial stocks dominated during the Global Financial Crisis of 2008. Overall, we see a wide sector dispersion for the likely distressed stocks, indicating that the prediction return is mostly from stock selection rather than sector allocation.

[22]For more discussion, see Choi, Jiang, and Zhang (2022) and Gu et al. (2020).

[23]This ML prediction task is studied in Swinkels and Hoogteijling (2022).

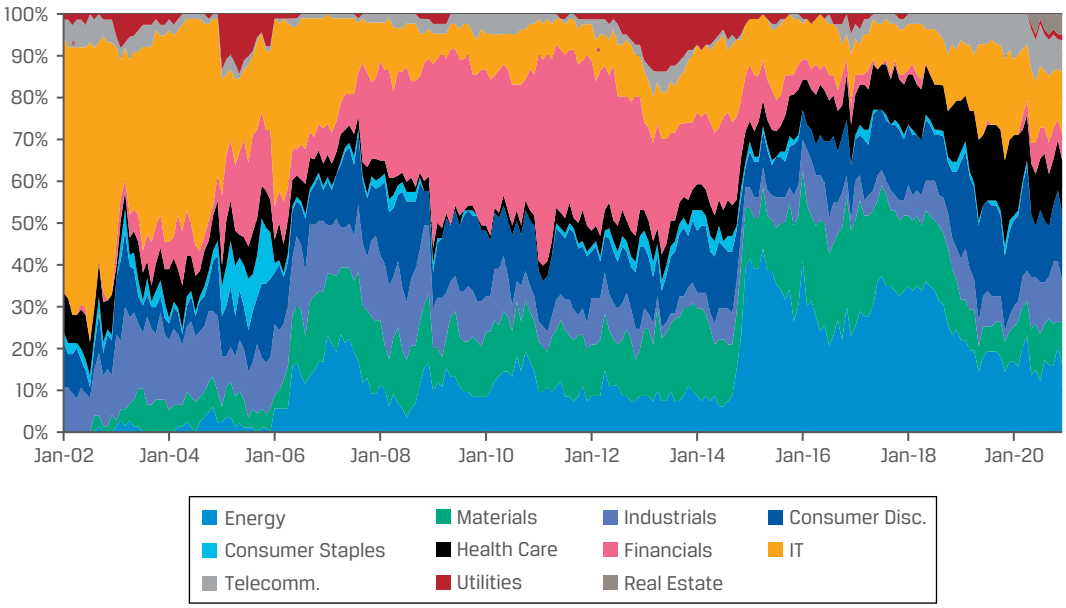
[24]Also called low-volatility equity, among other names.

Exhibit 8. Market Return vs. Return of a Portfolio Consisting of Likely Distressed Stocks, Estimated under Various Prediction Approaches



Source: Swinkels and Hoogteijling (2022).

Exhibit 9. Sector Composition of the ML-Predicted Likely Distressed Stocks



Source: Swinkels and Hoogteijling (2022).

Predicting Fundamental Variables

In this section, we discuss the example of using ML to predict company fundamentals.

Investment problem

Predicting stock returns is notoriously hard. As mentioned previously, there are potentially thousands of variables (dimensions) that can affect a stock's performance—investor sentiment, path dependency, and so on. The various factors, endogenous and exogenous, ultimately get translated into only a one-dimensional response—higher return (up) or lower return (down). An easier task may be to use ML algorithms to predict company fundamentals, such as return on assets and corporate earnings. Company fundamentals also have the additional beneficial characteristic of being more stable than stock returns, making them

better suited for ML predictions. Because of these reasons, company fundamentals prediction is another popular application of ML algorithms in the financial domain. In this section, we look at the findings from a popular study[25] where ML algorithm-predicted earnings forecasts are compared with those from more traditional approaches.

Methodology

The investment universe is the US stock market, excluding the financial and utility sectors. The study was conducted over the period 1975–2019.

Six different ML models were tested: three linear ML models (OLS, LASSO, and ridge regression) and three nonlinear ML models (random forest, gradient boosting regression, and neural networks). Six traditional models were used as a benchmark to compare against ML models: random walk, autoregressive model, models from Hou, van Dijk, and Zhang (2012; HVZ) and So (2013; SO), the earnings persistence model, and the residual income model. Various ensembles of these models were also tested.

The feature set is composed of 28 major financial statement line items and their first-order differences. So, there are 56 features in total.

Results

The results are shown in **Exhibit 10**. Consistent with the results for cross-sectional stock returns, Cao and You (2021) found that machine learning models give more accurate earnings forecasts. The linear ML models are more accurate than the benchmark traditional models (by about 6%), and the nonlinear ML models are more accurate than the linear ML models (by about 1%–3% on top of the linear model). Not only are the ML models more accurate; the traditional models autoregression, HVZ, SO, earnings persistence, and residual income were not more accurate than the naive random walk model. The ensemble models, traditional or machine learned, were more accurate than the individual models alone, with the order of accuracy preserved: ML ensemble beating traditional methods, ensemble and nonlinear ML ensemble beating linear ML ensemble.

The ML models' better performance can be attributed to the following:

- They are learning economically meaningful predictors of future earnings. One can make this conclusion by examining feature importance through such tools as Shapley value.
- The nonlinearity and interaction effects are useful in further refining the accuracy of forward earnings predictions, as evidenced by the higher performance of nonlinear ML models compared with linear ML models.

The takeaway from this study is the same as in the previous two examples. That is, ML models can provide value on top of traditional models (especially due to nonlinearity and interaction components), and ensembling is one of the closest things to a free lunch in ML, much like diversification for investing.

NLP in Multiple Languages

So far, we have discussed problems where ML algorithms are used for prediction. Another major category for ML applications is textual language reading, understanding, and analysis, which is called "natural language processing" (NLP). Modern NLP techniques use neural networks to achieve the great capability improvements they have made in recent years. NLP is discussed in other chapters of this book, so we will not discuss the techniques extensively here, but we will discuss one NLP application that can be interesting for practitioners.

Investment problem

Investing is a global business, and much of the relevant information for security returns is written in the local language. In general, a global portfolio may invest in 20–30 different countries,[26] while a typical investor may understand only two or three languages, if that. This fact presents a problem, but fortunately, it is a problem that we can attempt to solve through ML algorithms.

From the perspective of Western investors, one language of interest is Chinese. The Chinese A-share market is both large and liquid. But understanding Chinese texts on A-share investing can be challenging because Chinese is not an alphabetical language and it follows very different grammatical constructs than English. In addition, since retail investors dominate the Chinese A-share market,[27] a subculture of investment slang has developed, where terms used are often not standard Chinese, thereby compounding the problem for investors without a strong local language understanding.

In a paper by Chen, Lee, and Mussalli (2020), the authors applied ML-based NLP techniques to try to understand investment slang written in Mandarin Chinese by retail investors in online stock discussion forums. We discuss the results of that paper here.

[25]For more details, see Cao and You (2021). The problem of ML company fundamentals prediction was also examined in Alberg and Lipton (2017).

[26]The MSCI All Country World Index (ACWI) covers stocks from 26 countries.

[27]Some studies have concluded that in the A-share market, retail trading volume can be up to 80% of the total. In recent years, institutional market trading has proportionally increased as the Chinese market matures.

Exhibit 10. Prediction Accuracy Results from Cao and You (2021)

	Mean Absolute Forecast Errors				Median Absolute Forecast Errors			
		Comparison with RW				Comparison with RW		
	Average	DIFF	*t*-Stat	%DIFF	Average	DIFF	*t*-Stat	%DIFF
Benchmark model								
RW	0.0764				0.0309			
Extant models								
AR	0.0755	−0.0009	−2.51	−1.15%	0.0308	−0.0001	−0.22	−0.24%
HYZ	0.0743	−0.0022	−3.63	−2.82%	0.0311	0.0002	0.64	0.76%
EP	0.0742	−0.0022	−2.79	−2.85%	0.0313	0.0004	1.02	1.42%
RI	0.0741	−0.0023	−3.15	−3.07%	0.0311	0.0002	0.66	0.74%
SO	0.0870	0.0105	5.19	13.78%	0.0347	0.0039	5.50	12.56%
Linear machine learning models								
OLS	0.0720	−0.0045	−5.04	−5.83%	0.0306	−0.0002	−0.60	−0.73%
LASSO	0.0716	−0.0048	−5.32	−6.31%	0.0304	−0.0004	−1.11	−1.43%
Ridge	0.0718	−0.0047	−5.19	−6.11%	0.0305	−0.0003	−0.87	−1.08%
Nonlinear machine learning models								
RF	0.0698	−0.0066	−6.44	−8.64%	0.0296	−0.0012	−3.10	−3.97%
GBR	0.0697	−0.0068	−6.08	−8.86%	0.0292	−0.0016	−4.23	−5.34%
ANN	0.0713	−0.0051	−5.38	−6.67%	0.0310	0.0001	0.24	0.38%
Composite models								
COMP_EXT	0.0737	−0.0027	−3.89	−3.58%	0.0311	0.0002	0.56	0.66%
COMP_LR	0.0717	−0.0047	−5.25	−6.16%	0.0305	−0.0004	−1.02	−1.33%
COMP_NL	0.0689	−0.0075	−6.99	−9.87%	0.0292	−0.0017	−3.92	−5.55%
COMP_ML	0.0693	−0.0071	−7.12	−9.35%	0.0294	−0.0015	−3.75	−4.81%

Notes: RW stands for random walk. AR is the autoregressive model. HVZ is the model from Hou et al. (2012). SO is the model from So (2013). EP is the earnings persistence model. RI is the residual income model. RF stands for random forests. GBR stands for gradient boost regression. ANN stands for artificial neural networks. COMP_EXT is an ensemble of traditional models. COMP_LR is an ensemble of linear ML models. COMP_NL is an ensemble of nonlinear ML models. COMP_ML is an ensemble of all ML models.

Source: Cao and You (2021).

Methodology

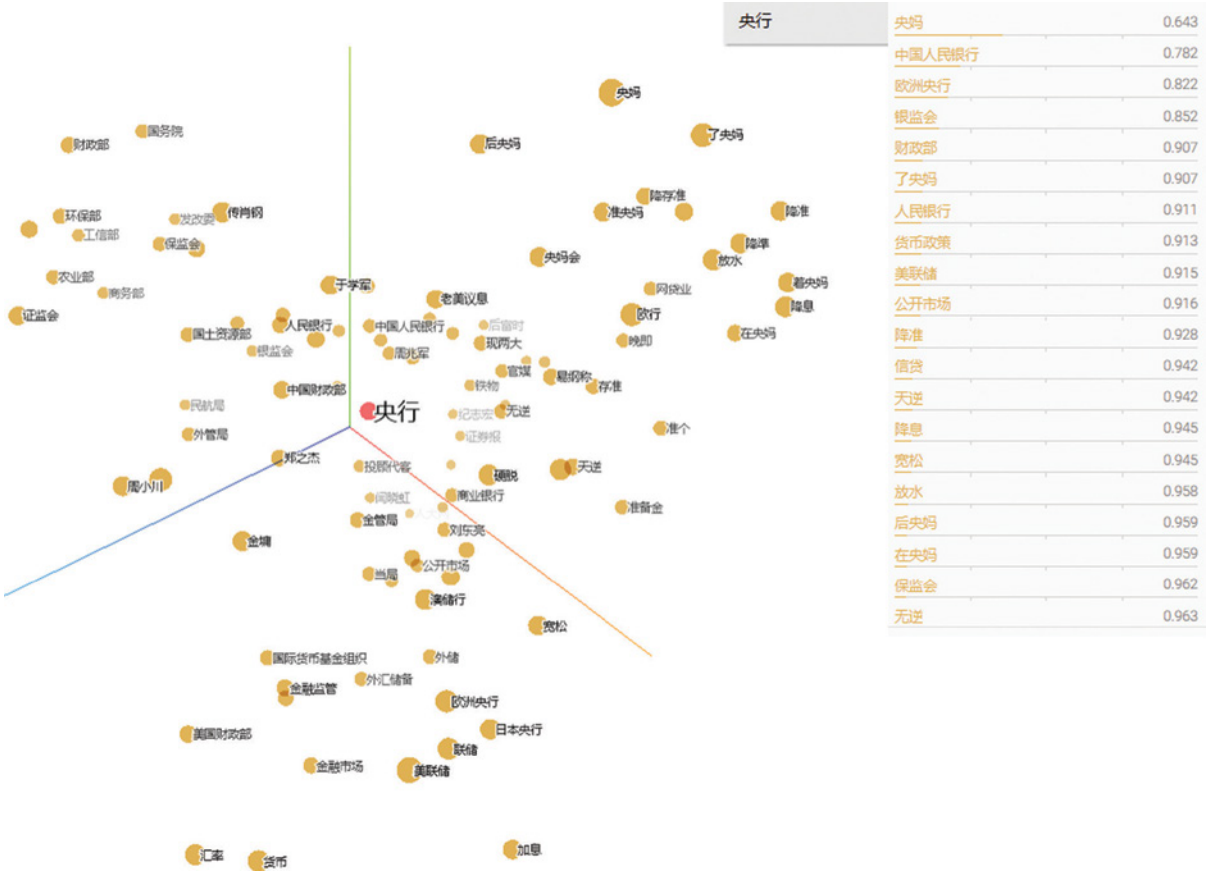
1. Download and process investment blogs actively participated in by Chinese A-share retail investors.
2. Apply embedding-based NLP techniques and train on the downloaded investment blogs.
3. Starting with standard Chinese sentiment dictionaries, look for words surrounding these standard Chinese sentiment words. By the construct of the embedding models, these surrounding words often have the same contextual meaning as the words in the standard sentiment dictionaries, whether they are standard Chinese or slang. An example of this is shown in **Exhibit 11**.

Results

Using this technique, we can detect investment words used in standard Chinese and the slang used by retail investors. In Exhibit 11, the red dot illustrates the Chinese word for "central bank." The column on the right side of Exhibit 11 shows the words closest to this word, and the closest is the word that translates to "central mother" in Chinese. This is a slang word often used by Chinese retail investors as a substitute for the word "central bank" because central banks often take actions to calm down market tantrums, much like a mother does to her children when they have tantrums.

The embedded words also exhibit the same embedded word vector arithmetic made famous by the following example (see Mikolov, Yih, and Zweig 2013): King – Man + Woman ≈ Queen. For example, **Exhibit 12** shows the following embedded Chinese word relationship: Floating red – Rise + Fall ≈ Floating green.[28]

Exhibit 11. Embedded Chinese Words from A-Share Investor Blogs Projected onto a 3-D Space



Source: Chen et al. (2020).

Exhibit 12. Chinese Word Embedding Still Preserves Vector Arithmetic

飘红 – 涨 + 跌 ≈ 飘绿

Source: Chen et al. (2020).

[28]In contrast to most of the world's markets, in the Chinese stock market, gains are colored red whereas losses are colored green.

This study illustrates that ML techniques not only are useful for numerical tasks of results prediction but can also be useful in other tasks, such as foreign language understanding.

Conclusion

In this chapter, we gave an overview of how ML can be applied in financial investing. Because there is a lot of excitement around the promise of ML for finance, we began the chapter with a discussion on how the financial market is different from other domains in which ML has made tremendous strides in recent years and how it would serve the financial ML practitioner to not get carried away by the hype. Applying ML to the financial market is different from applying ML in other domains in that the financial market does not have as much data, the market is nonstationary, the market can often behave irrationally because human investor emotions are often a big driver of market returns, and so on. Given these differences, we discussed several common pitfalls and potential mitigation strategies when applying machine learning to financial investing.

In the second half of the chapter, we discussed several recent studies that have applied ML techniques to investment problems. Common findings of these studies are as follows:

- ML techniques can deliver performance above and beyond traditional approaches if applied to the right problem.
- The source of ML algorithms' outperformance includes the ability to consider nonlinear and interaction effects among the input features.
- Ensembling of ML algorithms often delivers better performance than what individual ML algorithms can achieve.

We showed that in addition to predicting numerical results, ML could also help investors in other tasks, such as sentiment analysis or foreign language understanding. Of course, the applications discussed here are only a small subset of what ML can do in the financial domain. Other possible tasks include data cleaning, fraud detection, credit scoring, and trading optimization.

Machine learning is a powerful set of tools for investors, and we are just at the beginning of the journey of applying ML to the investment domain. Like all techniques, machine learning is powerful only if applied to the right problems and if practitioners know the technique's limits. Having said that, we believe one can expect to see a lot more innovation and improved results coming out of this space going forward.

References

Abe, M., and H. Nakayama. 2018. "Deep Learning for Forecasting Stock Returns in the Cross-Section." In *Advances in Knowledge Discovery and Data Mining,* edited by D. Phung, V. Tseng, G. Webb, B. Ho, M. Ganji, and L. Rashidi, 273–84. Berlin: Springer Cham.

Alberg, J., and Z. Lipton. 2017. "Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals." Cornell University, arXiv:1711.04837 (13 November). https://arxiv.org/abs/1711.04837.

Arnott, R., C. Harvey, and H. Markowitz. 2019. "A Backtesting Protocol in the Era of Machine Learning." *Journal of Financial Data Science* 1 (1): 64–74.

Avramov, D., S. Cheng, and L. Metzker. 2022. "Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability." *Management Science* (29 June).

Avramov, D., T. Chordia, and A. Goyal. 2006. "Liquidity and Autocorrelations in Individual Stock Returns." *Journal of Finance* 61 (5): 2365–94.

Birge, J., and Y. Zhang. 2018. "Risk Factors That Explain Stock Returns: A Non-Linear Factor Pricing Model." Working paper (9 August).

Cao, K., and H. You. 2021. "Fundamental Analysis via Machine Learning." Emerging Finance and Financial Practices eJournal.

Cao, Larry. 2018. "Artificial Intelligence, Machine Learning, and Deep Learning: A Primer." *Enterprising Investor* (blog; 13 February). https://blogs.cfainstitute.org/investor/2018/02/13/artificial-intelligence-machine-learning-and-deep-learning-in-investment-management-a-primer/.

Chen, M., J. Lee, and G. Mussalli. 2020. "Teaching Machines to Understand Chinese Investment Slang." *Journal of Financial Data Science* 2 (1): 116–25.

Choi, D., W. Jiang, and C. Zhang. 2022. "Alpha Go Everywhere: Machine Learning and International Stock Returns." Working paper (29 November). Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489679.

Columbus, Louis. 2020. "The State of AI Adoption In Financial Services." *Forbes* (31 October). www.forbes.com/sites/louiscolumbus/2020/10/31/the-state-of-ai-adoption-in-financial-services/?sh=1a4f7be62aac.

Daul, S., T. Jaisson, and A. Nagy. 2022. "Performance Attribution of Machine Learning Methods for Stock Returns Prediction." *Journal of Finance and Data Science* 8 (November): 86–104.

Fama, E., and K. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Grinold, R., and R. Kahn. 1999. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*. New York: McGraw Hill.

Gu, S., B. Kelly, and D. Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *Review of Financial Studies* 33 (5): 2223–73.

Halevy, A., P. Norvig, and F. Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 (2): 8–12.

Hanauer, M., and T. Kalsbach. 2022. "Machine Learning and the Cross-Section of Emerging Market Stock Returns." Working paper (5 December).

Harvey, C., Y. Liu, and H. Zhu. 2016. "… and the Cross-Section of Expected Returns." *Review of Financial Studies* 29 (1): 5-68.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer.

Hou, K., M. van Dijk, and Y. Zhang. 2012. "The Implied Cost of Capital: A New Approach." *Journal of Accounting and Economics* 53 (3): 504–26.

Hou, K., C. Xue, and L. Zhang. 2020. "Replicating Anomalies." *Review of Financial Studies* 33 (5): 2019–133.

Jessen, C., and D. Lando. 2013. "Robustness of Distance-to-Default." 26th Australasian Finance and Banking Conference 2013 (16 August).

Leippold, M., Q. Wang, and W. Zhou. 2022. "Machine Learning in the Chinese Stock Market." *Journal of Financial Economics* 145 (2): 64–82.

Leung, E., H. Lohre, D. Mischlich, Y. Shea, and M. Stroh. 2021. "The Promises and Pitfalls of Machine Learning for Predicting Stock Returns." *Journal of Financial Data Science* 3 (2): 21–50.

Li, Y., Z. Simon, and D. Turkington. 2022. "Investable and Interpretable Machine Learning for Equities." *Journal of Financial Data Science* 4 (1): 54–74.

Li, Y., D. Turkington, and A. Yazdani. 2020. "Beyond the Black Box: An Intuitive Approach to Investment Prediction with Machine Learning." *Journal of Financial Data Science* 2 (1): 61–75.

Lo, A., D. Repin, and B. Steenbarger. 2005. "Greed and Fear in Financial Markets: A Clinical Study of Day-Traders." NBER Working Paper No. w11243.

López de Prado, M. 2018. "The 10 Reasons Most Machine Learning Funds Fail." *Journal of Portfolio Management* 44 (6): 120–33.

López de Prado, M. 2019. "Ten Applications of Financial Machine Learning." Working paper (22 September).

Lundberg, S., and S. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." NIPS 17: Proceedings of the 31st Conference on Neural Information Processing Systems: 4768–77.

McLean, R., and J. Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance* 71 (1): 5–32.

Mikolov, T., W. Yih, and G. Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." Proceedings of NAACL-HLT 2013: 746–51. www.aclweb.org/anthology/N13-1090.pdf.

Nadeem, Monisa. 2018. "Clearing the Confusion: AI vs. Machine Learning vs. Deep Learning." Global Tech Council (25 November). www.globaltechcouncil.org/artificial-intelligence/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning/.

Rasekhschaffe, K., and R. Jones. 2019. "Machine Learning for Stock Selection." *Financial Analysts Journal* 75 (3): 70–88.

So, E. 2013. "A New Approach to Predicting Analyst Forecast Errors: Do Investors Overweight Analyst Forecasts?" *Journal of Financial Economics* 108 (3): 615–40.

Swinkels, L., and T. Hoogteijling. 2022. "Forecasting Stock Crash Risk with Machine Learning." White paper, Robeco.

Tobek, O., and M. Hronec. 2021. "Does It Pay to Follow Anomalies Research? Machine Learning Approach with International Evidence." *Journal of Financial Markets* 56 (November).

2. ALTERNATIVE DATA AND AI IN INVESTMENT RESEARCH

Ingrid Tierens, PhD, CFA
Managing Director, Goldman Sachs Global Investment Research

Dan Duggan, PhD
Vice President, Goldman Sachs Global Investment Research

Artificial Intelligence (AI) and alternative data are not new concepts in investment research or finance. AI can be broadly defined as using computers to mimic human problem solving. Alternative data can be broadly defined as any data in a nonstructured format. The arrival of computers in financial firms and the arrival of nonstructured data in digital format meant that entrepreneurial finance professionals started to see opportunities to address financial use cases by leveraging these tools instead of primarily relying on the human brain, helped by pen, paper, an abacus, or a calculator. If you define investment decision making as trying to make objective sense of a wide spectrum of diverse data to allocate capital, the appeal of AI and alternative data to improve investment decisions should not come as a surprise. Although early attempts to codify human decision making may not meet today's definition of AI and alternative data, because they may have been either too ambitious or have had a marginal impact by today's standards, they paved the way for an innovation cycle that is by now well under way in the financial services industry, well past its early adoption cycle, and affecting all facets of how financial firms conduct their business. From enhancing investment insights to constructing better portfolios, optimizing trading decisions, streamlining client service, reducing operational issues, better aligning products to client needs, and extracting better business intelligence insights, AI and alternative data are leaving their fingerprints everywhere.

Embedding AI and alternative data into the day-to-day activities of an investment analyst is not mission impossible. It can be compared with executing a New Year's resolution, such as learning a new language, getting in better shape, or running your first marathon. Resolutions become reality through commitment, persistence, and resilience. The journey of Goldman Sachs Global Investment Research (GIR) over the past five-plus years illustrates that you do not need expensive "equipment" (i.e., an unlimited budget) and an army of "coaches with name recognition" (i.e., hard-to-find talent) to make a difference. Hard work by open-minded people who collectively buy into the mission, bring different components to the table, learn from each other, and collaborate to execute on the joint mission to produce leading-edge investment insights will lead to a noticeable impact. This chapter demonstrates that you can start small and that investment analysts are not bystanders but play a crucial role in making AI and alternative data part of their lexicon and research process.

Where Can AI and Alternative Data Be Additive? Investment Research Use Cases

Alternative data and AI, including machine learning (ML) and natural language processing (NLP), are not an end in and of themselves but are additional tools in the research toolkit to create differentiated investment insights. The key to their use in investment research is appreciating that alternative data and AI do not define the investment thesis but, instead, help prove or disprove it. AI and nontraditional data can be additive across the spectrum of investment strategies, as long as the AI and data efforts are aligned with a particular strategy. Comparing and contrasting systematic and fundamental investment strategies can be enlightening in this regard.

At first glance, AI and alternative data sound as if they may only provide an advantage to systematic strategies because of their quantitative nature. Within the AI and alternative data spectrum, however, there are many niche datasets and techniques that may provide added value for fundamental strategies that can successfully incorporate data and data science expertise in their investment process. Let us take a closer look at use cases for each strategy.

For a systematic strategy where breadth matters more than depth, datasets and data analysis techniques need to be applicable across many securities. Therefore, it should not come as a surprise that systematic managers focus their time and effort more on alternative datasets or techniques that can be applied to a large investment universe. Factors extracted from text have been added to many quant investment processes, because NLP analysis can be easily repeated across company filings, news articles, and other documents that are available for large sets of securities. In addition, sophisticated econometric techniques can be

helpful to convert individual alphas into more solid multifactor alphas and, in turn, into more robust portfolios. But even within systematic strategies, AI and alternative data use may differ significantly, especially when the investment horizon is taken into account. For example, intraday news sentiment indicators may be helpful for high-frequency trading strategies, but relatively less value adding for an investment professional focused on an investment horizon spanning weeks or months, even if the signals can be applied to thousands of securities.

For a fundamental strategy where depth matters more than breadth, a portfolio manager or analyst will go deep into specific use cases and will put a higher premium on precision than a systematic manager dealing with a diversified portfolio and for whom each holding has less of an impact on performance. Moreover, niche datasets can be more useful for fundamental analysts or portfolio managers to complement the information set they draw from, thus providing a fuller mosaic view. For example, consumer sentiment via social media, brick-and-mortar foot traffic, and even consumer search trends around product cycles offer different angles to create a more complete picture for consumer-focused sectors. And while sectors with more digital presence—for example, retail and technology, media, and telecommunications—were early targets for AI and alternative data applications, GIR has seen many use cases in both single-stock and macro research that may not seem like obvious candidates for successful AI or alternative data use. Examples include leveraging app downloads in the medical device space and localized market share analysis for membership in managed care organizations or even quarry production.

Which Alternative Data Should I Pay Attention To? Data Sourcing in a World of Overwhelming Data Supply

Keeping up with the supply of alternative and big data is a Herculean task. There is an overabundance of providers because the barriers to entry in this space have become very low. Providers of "traditional" data tend to be established organizations that typically offer curated datasets and have many years of experience with data due diligence and support. With alternative data, the onus of due diligence has shifted from the data producer more to the data consumer. There are clear parallels with the production and consumption of news, where the due diligence and fact checking have similarly shifted from the news provider more to the news consumer.

The investment needed to bring onboard vendors, ingest data, and test the data can outweigh the benefits of a new data source, and the licensing costs can further tip the scale. Given how time consuming the due diligence process can be, the data acquisition efforts within GIR are demand driven. Research analysts as subject matter experts in their field typically have a good idea of the type of data that may be useful for the investment thesis they are pursuing. While they may not have locked in a specific data source, may not be aware of the most scalable way to obtain the data, or may not be comfortable manipulating unstructured data, working backwards from their use case and initial data ideas provides a good starting point. Ideas for nontraditional data can emerge when analysts question whether they have the full mosaic of data to assess their thesis and whether the datasets they have used so far continue to provide a comprehensive picture.

In addition, new data sources do not necessarily need to offer orthogonal information but can also be value adding if they offer legacy data in a more scalable and more timely fashion. The questions that today's analysts are trying to answer are, in essence, no different from what their predecessors tried to address, such as the following: How is a new product launch perceived by the marketplace? What do price trends and inventories look like? How do changing demographics affect the economy? What is different or alternative is that the analyst no longer needs to visit stores, run in-person surveys, or pursue other manual paths to get answers. A continually increasing amount of relevant information is now available in digital format on a much larger scale and in a timelier fashion or more frequently than was the case in the past. The qualifier "alternative" in alternative data may be a bit of a misnomer from that perspective.

Some investment managers who are very data driven have a separate, dedicated data-scouting effort, possibly an extension of the team that interacts with market data vendors and brokers to surface new and novel datasets. GIR has not gone down that path, because it has found that having the subject matter experts—the analysts—take at least co-ownership of identifying relevant data sources for their specific use cases outweighs the scalability a data scout may bring to the table. Where research analysts often need help is in how to efficiently analyze the relevant data, especially as datasets have become more complex and harder to wrangle. GIR's Data Strategy team, a dedicated team with a more focused analytical and quantitative background, collaborates with single-stock and macro research teams to help them achieve that objective through its GS Data Works initiative.

Use Cases for Alternative Data at Goldman Sachs Global Investment Research

With time and people in short supply, GIR's Data Strategy team cannot chase every data idea, so it prioritizes data that can be relevant for multiple teams or have a high probability of being used by at least one team on an ongoing basis. Data budget constraints obviously also play a role, especially in a research setting where success in execution of research ideas mostly accrues to the benefit of third parties. Fortunately, the explosion of data includes a diverse and abundant number of publicly available datasets that have proven useful for research purposes when properly combined with other data sources already being considered. **Exhibit 1** provides an overview of alternative data types used across the Goldman Sachs research division.

Exhibit 1. Types of Alternative Data Used in Goldman Sachs Research

Source: Goldman Sachs Global Investment Research.

- If the research use case has a geographic component, geospatial data may enter the equation. Datasets comprise not only dedicated geospatial measurements, such as mobile foot traffic data, satellite imagery, and government population and demographic data, but also a wide variety of information with inherent geographical importance, such as store locations or electric vehicle charging stations. GIR's understanding of brick-and-mortar retail sales is greatly enhanced by analyzing locations of retailers and their competitors. Similarly, leveraging public environmental service data (e.g., location-specific landfill capacities and lifetimes) provides a deeper understanding of the competitive pricing landscape in the environmental services industry.
- If the research use case has an online component, digital information can be additive. Datasets include app downloads, online point-of-sale information, website visits, product counts and pricing, and search trends. Example use cases include quantifying consumer interest (e.g., search intensity and active user counts) to better understand user engagement and assessing product launches or distress situations through social media sentiment, app downloads, and product-specific data.
- If the research use case can be addressed by searching through text, NLP techniques may uncover additional insights. This category is quite broad, covering a wide range of unstructured data, from earnings call transcripts and company filings to tweets and blog posts. David Kostin, the chief US equity strategist at Goldman Sachs, publishes a quarterly S&P 500 Beige Book, which leverages NLP to identify relevant themes each earnings cycle, one of the many research areas where NLP has proven to be additive.

The three categories in Exhibit 1 are prone to overlap, as many datasets span multiple dimensions. Ultimately, the combination of dimensions provides a more complete mosaic to better answer most research questions. GIR thus often uses more than one category to enhance its mosaic. For example, a company's pricing power can be derived not only from many product codes and prices (digital) but potentially also from its local market share in certain geographic regions (geospatial). In addition, commentary about the company's pricing power during earnings calls may be informative for a covering analyst (NLP).

Which Component of an AI Effort Is the Most Time Consuming? The Underappreciated Power of Doing the Data Grunt Work

Do not expect AI to provide insights if you do not understand the data you apply AI to. Data scientists will be successful only if they appreciate the nuances, strengths, and weaknesses of the data that go into their modeling efforts. Cases of algorithmic bias caused by algorithms trained on biased data have made the news in areas outside investments, but that does not mean investment applications are immune to being misguided because of data issues. This is an area where investment professionals who may not have data science skills but are subject matter experts can make a real difference. It is unrealistic to expect a data scientist to understand each investment use case, just as it is unrealistic to expect each research analyst to be a data science expert. However, if the two sides have an open mind to learn from each other and iterate, tremendous synergies and unique opportunities to expand each other's skill sets will surface.

Before making final decisions on sourcing a particular dataset, trying to identify a small test case is strongly recommended to give a good flavor of the data without time-consuming technological or due diligence hurdles. Again, this is where the subject matter expert can play a pivotal role. GIR's data strategists have witnessed more than once that a few targeted questions from an experienced analyst during a meeting with a data provider highlighted a data shortcoming, which led to shelving the data source. Even after onboarding a new data source, you need to stay alert because the data may evolve over time, especially if the provider has limited experience providing data or is unfamiliar with how the financial services industry may use its data. Goldman Sachs has dealt with situations where underlying inputs to the data were removed, added, or edited. If, in addition, the data you are sourcing are derived from underlying inputs, getting enough transparency in the algorithms used to create the derived data will add more complexity to the task of familiarizing yourself with the data. The following are a few important considerations depending on the type of data you are looking at.

- For large and complex datasets, one of the most general questions to ask is how comprehensive the data are. Certain data fields may have missing values, which can bias computational estimates such as averages and sums. Such bias has been observed in, for example, pricing data attained via web scraping, the practice of programmatically extracting publicly-available information from websites. The dataset may already have a layer of analysis to account for data problems, such as automated filling around missing data, as well as more sophisticated analysis choices, so it is important to understand the assumptions underlying that analysis layer. Another question to address is whether the data allow you to delve into the dimensions that are relevant for your investment thesis. For example, do you need style-specific breakouts in addition to aggregated counts when you lever Stock Keeping Unit (SKU) counts to assess the success of a new product launch?
- Even small datasets may be built off an analysis layer that relies on sampling, and problems can arise from techniques that either undersample or have inherent selection biases built in, such as point-of-sale, satellite imagery, or mobile foot traffic trends. Point-of-sale information that requires opt-in consumer panels may lever discounts for products in return for data use permission, which may skew demographics beyond what a simple reweighting can remedy. Similarly, parking lot car counts from satellite imagery cannot measure covered parking. Mobile foot traffic data have limited precision in malls and other multitenant structures. Whatever the dataset, the process of its construction is vital to assessing its advantages and limitations to ultimately determine its appropriateness.
- For NLP and text analysis, understanding details around the input data is also vital to interpreting results. For example, third-party aggregators of news articles may not be able to see paywall content, which may introduce subtle biases. If you search across multiple text sources, are source-specific biases and trends easy to separate and identify? Another example is sentiment analysis. Results from Instagram will likely be different from those from Twitter and Reddit. Breaking apart trends and results by source can help identify issues that could otherwise translate into misguided signals. In addition, when NLP is used for sentiment analysis to evaluate important themes, topics, or events, guarding against false positive or false negative results plays an important role and, simultaneously, provides an opportunity for a more nuanced view. For example, understanding the difference between a poor consumer product, on the one hand, and consumer frustrations at out-of-stock items or launch logistics issues, on the other hand, will not only strengthen results but also better inform investable decisions.

The bottom line is that data curation may not sound overly exciting, but it can easily determine success or failure. No matter how sophisticated the data analysis capabilities are, the AI effort will fail—or, possibly worse, create wrong outputs—if the data do not receive at least as much attention as the analysis. Again, this is not an area that GIR outsources to a different part of the organization. While there are parts of the chain that can potentially be

handled by teams with more of an operational background, such as data ingestion and low-level cleaning, data curation requires the attention of the people who are familiar enough with the business use case—that is, the combination of the subject matter research experts and the data scientists. As a research division, GIR not only consumes data but also produces a substantial number of proprietary forecasts and indicators. Wearing both hats has made the Goldman Sachs research division appreciate even more how nontrivial it is to create trustworthy data.

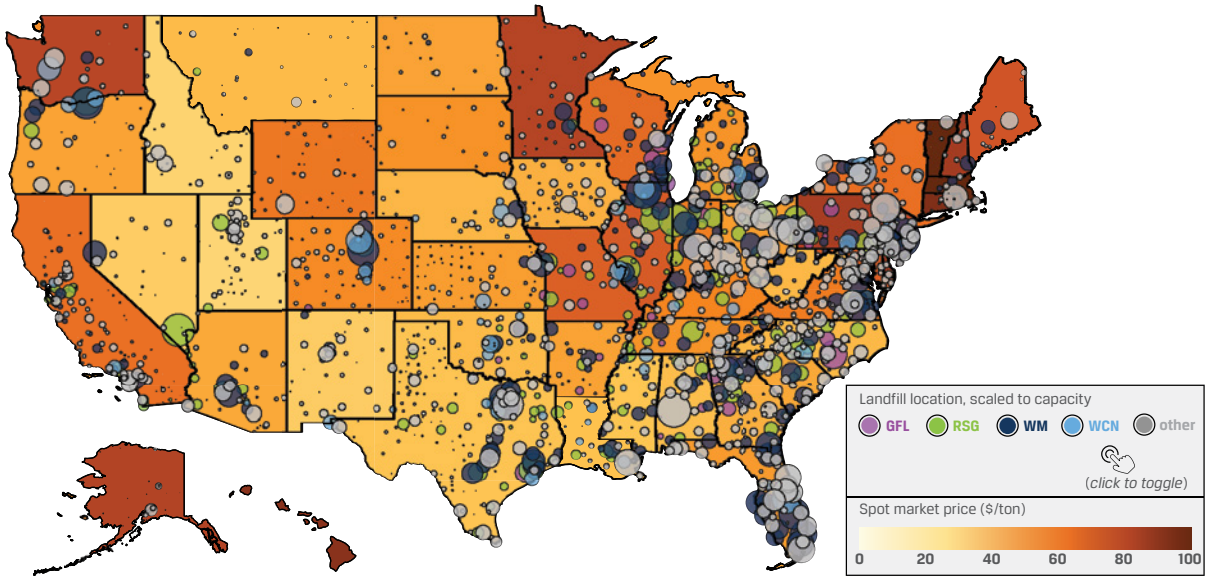
Where Does Data Analysis Come In? Looking at the Spectrum of Analytical Approaches

Like the data landscape, the analysis landscape has expanded dramatically. For data analysis, GIR's Data Strategy team also follows a "pull" approach, as opposed to pushing a particular analytical technique in search of a problem to solve. The team works backwards from the use case at hand and collaborates with research analysts on the following questions: What are the strengths and weaknesses of approaches you have tried in the past? Are the outcomes vastly different when you make minor changes to your assumptions? Do outliers have a significant impact on the results? If any of these answers is yes, it may be time to try more sophisticated approaches that can add robustness to the research findings.

Also like the data landscape, where data fees and required effort are counterbalancing factors to simply load up on more data, explainability and required effort need to be weighed against the promises of AI. If a more straightforward approach to analyzing the data works, GIR tries to avoid adding unnecessary complexity. For example, when tracking inflation as a theme in earnings calls, GIR found that simply counting mentions of inflation was a robust proxy for a much more complex screen consisting of a combination of nearly 200 terms. But when the situation calls for more complex analysis, data scientists need to be able to provide enough intuition so that research analysts are comfortable with the outcomes.

If structured appropriately, analysis frameworks can provide a significant amount of flexibility and scalability. For example, **Exhibit 2** shows the results of an analysis of the pricing power of US environmental service companies based on local market dynamics. The different shades of color indicate different spot price ranges, a proxy of pricing power. The circles indicate landfill locations scaled to capacity. While the datasets leveraged were specific to environmental services, the techniques are not. The environmental data were combined with a more mature geospatial framework that had previously answered similar

Exhibit 2. Leveraging Geospatial Analysis Applied in One Industry to Another Industry



Sources: "Americas Environmental Services: Compounding Unit Profitability amid Building Local Market Share," published by lead equity analyst Jerry Revich, CFA, on 6 April 2021. Analysis by Goldman Sachs Global Investment Research based on data compiled from the *Waste Business Journal,* the New York State Department of Environmental Conservation, the California Water Boards, and other state sources.

questions for suppliers of heavy building materials, which was highlighted in the 2019 "AI Pioneers in Investment Management" report by CFA Institute.[29]

Data science techniques have made remarkable progress, from enabling self-driving cars to providing customized recommendations for a variety of services and products. Use cases for data science in financial services are expanding as well and include, for example, recommendation engines to customize dissemination of research and further advances in trading algorithms that benefit from vast amounts of high-frequency order and execution data. However, many dynamics in financial markets are inherently unstable and adaptive in nature; that is, the amount of data that can be used for modeling and running in-sample and out-of-sample analyses is limited. This reality puts some limitations on how far these techniques can be pushed from a purely investment research perspective.

The term "AI" may also suggest that humans have no role to play in the analysis phase. As mentioned before, humans can add significant value by properly curating data, a necessary input for AI. Similarly, humans will add value by continuously assessing the results of algorithms. Breakpoints in markets, such as the COVID-19 pandemic, keep illustrating that modeling efforts need to be adapted as markets change, and humans play a critical role here. For example, some of the Goldman Sachs proprietary "nowcasting" indicators went through that adaptation cycle soon after COVID-19 became a global phenomenon, and the subject matter experts realized how extreme outliers broke down statistical relationships used in the creation of the indicators.

Finally, the use of NLP techniques for investment research purposes deserves a special mention. GIR's efforts in this field have been geared more toward surfacing specific content in a scalable fashion than interpreting the meaning or sentiment of written words. Applying NLP to a use case where the accuracy of output was essential and consistently finding through human review that the NLP output flagged too many false positives and false negatives provided a good education on some of the boundaries of NLP in a research setting.

How Do I Measure Success for an AI/Alternative Data Effort? Using a Wide Lens Instead of a Narrow Lens

The initial hype around AI and alternative data may have created unrealistic expectations, which, in turn, led to demands for hard evidence that AI and alternative data add to a financial institution's bottom line, especially in cases where new data sources come with a hefty price tag and data science expertise is hard to find. It is helpful to keep a broad perspective and evaluate these efforts, traditional or nontraditional, through multiple lenses:

- Does the approach create alpha? While this is the question that most people would like to see a concrete answer to, it is unlikely that a single dataset or single analysis technique will be an alpha generator. For systematic strategies, it may be possible to backtest a strategy with or without inclusion of the AI component and measure the marginal alpha contribution. The usual caveats related to backtesting apply, but those are no different with or without AI and alternative data.
- Does the AI effort produce unique insights in a more timely or more precise fashion? For fundamental strategies, the direct alpha added by the AI effort may be hard to assess. That said, quantifying the improvement along specific dimensions can be a useful exercise. For example, GIR's nowcasting of port congestion allowed a more real-time confirmation of the shipping disruption Ukraine experienced during the early days of its invasion by Russia and as the war has continued to evolve. GIR's distance measurements of vessels laying high-voltage cable established estimates for revenue-driving business at a very granular level. GIR's market share analyses within multiple industrial businesses, analyzed via geo-clustering, provided the basis for broader statements about company pricing power. These examples also illustrate that the majority of GIR's AI and alternative data efforts are geared toward descriptive analyses that feed into the broader investment thesis, as opposed to being prescriptive analyses that directly lead to an investment recommendation. Knowing what you hope to get out of the process can help focus the effort and suitably ground expectations of success.
- Does AI create scalability that saves valuable analyst time or allows a research hypothesis to be applied across a broader set of securities and/or applied more frequently? A good example is NLP analysis of earnings transcripts, where a GIR analyst can now easily identify themes not only across the stocks in her own coverage universe but also relative to peer companies and assess these themes going back over multiple years. Conversations with GIR analysts that started as an inquiry about the use of alternative data in some cases highlighted much lower hanging fruit, where a more systematized approach could create scalability that allowed the analysts to consider various scenarios that could not have been achieved manually.

[29]"AI Pioneers in Investment Management" (Charlottesville, VA: CFA Institute, 2019). https://www.cfainstitute.org/-/media/documents/survey/AI-Pioneers-in-Investment-Management.pdf.

While asking for a return on the AI investment is absolutely the right thing to do, the hurdles to justify the effort seem somewhat unfair relative to how budgets for traditional data sources and traditional data analysis (for which the costs can also add up) are determined. Trying to identify appropriate key performance indicators for an AI effort has therefore started to make people question their broader data efforts, wondering whether they are extracting sufficient return on investment across the entire data spectrum. This evolution is healthy and may lead to a more level playing field for AI-focused efforts.

Where Do I Start? It Is a Marathon, Not a Sprint

If you are embarking on your AI and alternative data journey in a setting with people who already have years of investment expertise, hiring one or two key people who have a data science background, have hands-on experience digging into data, and are genuinely interested in learning about and solving investment problems is a great starting point. Teaming up those key hires with one or two analysts who are interested in leveraging and analyzing more data, have a solid use case that has staying power, and understand that you need to iterate to get to a useful result should allow you to hit the ground running. Investment analysts who have ideas for new data to incorporate into their process and appreciate that properly analyzing those data may require trial and error are great candidates to start an organization on its path to AI and alternative data adoption and success. The initial use cases ideally do not depend on a complicated dataset that may take too long to bring on board.

Having senior sponsorship from the start is important. Those senior sponsors do not need to have familiarity with AI and alternative data themselves but need to have an appreciation that these approaches can be additive to the investment process. Their role is to provide the trailblazers some room and cover to experiment and iterate, while keeping them accountable by staying on top of (especially the initial) use cases. Once there are specific, tangible outcomes, others can get their heads around what it means in practice to lever alternative data and more sophisticated techniques. At that point, momentum to expand across the organization is created and scaling up the effort across geographies, analyst teams, and asset classes becomes a realistic next step.

Another question that is raised in this context is whether the AI and alternative data effort should be centralized or embedded in each team. As the need for having a dedicated data science expert in each team was *de minimis*, GIR went for a centralized effort that can be thought of as an extension of its research teams across the globe and across asset classes. Its centralized approach has given GIR the benefit of connecting the dots across use cases coming from different research teams.

Other teams in an investment organization can provide leverage, and it is good practice to make people in engineering, legal, and compliance departments aware of the AI and alternative data effort from Day 1, even if they do not have an immediate role to play. As the use cases expand, questions about data storage, firewalls, terms and conditions, and licensing rights for use of specific data sources will increase, which will require proper attention from those experts. In addition, as your use cases become more sophisticated, you may consider building additional functionality in house as opposed to relying on a third party to do some of the processing for you (i.e., build versus buy), which may require a dedicated engineering focus. GIR's own journey has reflected that evolution, where the first years of experience with embedding AI and alternative data provided a wealth of information on what approaches do and, more importantly, do not work and how they can fit into the workflow of a research analyst. GIR's Data Strategy team had to revisit a number of initial approaches on how to best evolve its data ecosystem but is now in a much better position to partner further with colleagues in engineering to make it a reality, as opposed to having the initial enthusiasm translate into engineering effort spent on building systems that may not have been properly thought through.

Conclusion

The impact of AI and alternative data on investment research is evolutionary, not revolutionary. Goldman Sachs looks at nonstructured, alternative, and big data as other components in the spectrum of data that may be relevant to its research process. It looks at AI, ML, and NLP as other components in the spectrum of analytical tools to make sense of the data. With the lines between traditional and alternative data becoming blurred and an increasingly ambiguous definition of what AI does or does not include, GIR does not draw artificial boundaries across these spectrums and therefore does not assess the value of AI and alternative data differently from its other data and analytical efforts. It draws from the types of data and the types of analyses as needed, often mixing unstructured data with traditional data and having more sophisticated approaches live side by side with more intuitive approaches. Subject matter and data science experts team up as appropriate, while ensuring they draw the best of man plus machine to minimize algorithmic and data biases. As this space matures further and the lines blur even more, Goldman Sachs expects that this integrated, iterative, and collaborative approach will continue to bear fruit for its investment research use cases. And because it also expects that AI and alternative data will simply become part and parcel of the research process, there may be a day when the labels "big" and "alternative" are no longer relevant!

3. DATA SCIENCE FOR ACTIVE AND LONG-TERM FUNDAMENTAL INVESTING

Kai Cui, PhD
Managing Director, Head of Equity Data Science, Neuberger Berman

Jonathan Shahrabani
Managing Director, Chief Operating Officer – Global Research Strategies, Neuberger Berman

Large-scale alternative data (alt-data) and data science are improving the investment techniques in the industry. Active asset managers with a long-term orientation particularly, given their long-term mindset and the resulting lower portfolio turnover, have differentiating opportunities to innovate in data science and big data. Therefore, sustainable, long-shelf-life data insights are as important as (if not more important than) short-term insights related to potential mispricing and time-sensitive advantages.

Long-term fundamental investors in active asset management strategies started embracing alt-data and data science two or three years later than their hedge fund peers. For example, data science has been an increasingly important capability of Neuberger Berman since 2017 across sectors, geographies, and asset classes.

Alt-data allow asset managers to "bear-hug" companies by enhancing our understanding of their operations and organizations independent of and incremental to information

Data Science Integration at Neuberger Berman

At Neuberger Berman, the data science function is deeply integrated and provides significant added value in its practices managing global equity mandates for institutional, retail, and high-net-worth investors. Idea generation is the combined effort and responsibility of both fundamental industry analysts and data scientists.

Given the vast alternative data footprint that is continuously being created by companies, the teams use research templates and data science processes—including data science–integrated scalable financial models, data modeling, and analytical tools—extensively to systematically capture and synthesize this information.

The design and construction of these research templates are a collaboration between both fundamental industry analysts and data scientists. These templates capture the key performance indicators (KPIs) for a given industry and key thesis metrics and data insights for a given company that are material to its longer-term earnings power. Simultaneously, with curated long-term fundamental support, we actively construct and track a spectrum of alt-data metrics to glean a comprehensive view of the company's real-time operational metrics and/or risk toward our long-term growth targets.

Both fundamental industry analysts and data scientists review these data on at least a weekly basis and discuss any notable upward or downward movements.

If a template is not relevant for a given industry or company (or a deep dive is required), fundamental industry analysts and data scientists partner on a custom analysis. As can be expected, custom analyses require extensive collaboration between both parties. For certain industries, such as health care, the background of the fundamental industry analyst takes on increasing importance because of the vast differences between subsectors (biotechnology, pharmaceuticals, life sciences, etc.).

During year-end reviews, fundamental industry analysts and data scientists are evaluated based on their contributions from multiple perspectives, including, but not limited to, investment performance, engagement, and strategy innovations and development. These are important drivers of incentives for both parties.

gleaned from financial statements and management interactions. The depth and breadth of alternative data, if analyzed and integrated correctly, can generate invaluable insights with a longer shelf life for long-term alpha generation.

The Operating Model and Evolution of Data Science Integration

There is no single best operating model for data science integration that fits all active asset managers. Data science must be customized to fit into each firm's own unique culture, organizational structure, core value proposition, strategic prioritization, and even innovative budgeting methods.

Evolution of the Data Science Organization

In this section, we discuss the decision of whether to use a centralized or decentralized data science team and how to evaluate the return on investment of data science initiatives.

To Centralize or Not

While a highly centralized data science team may enjoy more autonomy than decentralized teams and serve well as an innovation hub or research lab, if detached from investment decision making and business opportunities, it tends to fail to drive sustainable innovation that can be integrated or provide incremental value to the core investment platform.[30] In contrast, a completely decentralized data science team can prove economically inefficient and risk duplication of effort, especially when data scientists are less experienced.

The level of data science integration among strategies varies, and we can distinguish three stages of data science integration/engagement with investment teams even within the same firm: (1) engagement to establish credibility, (2) active and deep integration for investment evaluation, and (3) data science as part of the core investment strategy.

The level of centralization depends on the stage of data science integration. Generally, the data science team is centralized in the early to intermediate stages of development. When best practices are established and experience grows, experienced data scientists can make a more direct impact as part of an investment team's decision-making processes. The transition will happen faster as more data scientists and analysts are developed with training in both data science and fundamental investing.

Evaluating Data Science Initiatives' Return on Investment

While it is important that all data science initiatives start with illustrating incremental added value and establishing good engagement relationships with fundamental industry analysts, rapid "time to market" and the early establishment of credibility and process momentum on the road to subsequent stages are essential as well. Notably and importantly, the KPIs for team performance and return on investment (ROI) assessment could be different when planning and prioritizing engagement/integration efforts in different stages of data science and investment process integration.

For example, driving data source coverage, metrics and insights coverage, and the number of high-quality use cases are essential in the early stage to establish credibility. When coming to more active and deeper integration, data science teams need to align performance KPIs with the measurement of investment decision-making impacts and contribution to investment performance. Furthermore, if data science is part of a core investment strategy and/or is driving innovative investment solutions for investors and prospects, additional KPIs and ROI measurements on contribution to strategic development, growth in assets under management, and client engagement are also important, in addition to the contribution to investment performance. Over time, data science efforts evolve with the growing capabilities of team members and the closer partnership with fundamental industry analysts, and we continue to enable the data science team for long-term success focusing on deeper integration into both investment decision making and innovative investment solutions.

ROI assessment of data insights and data sources also needs to be customized to the investment styles of active asset managers while taking into account such factors as investment *capacity*, which is as important as investment performance for innovative data science–integrated investment solutions. For example, alt-data may have proven to generate incremental alpha in sector-focused, cross-sectional systematic strategies, but a data science platform supporting more diversified signals and types of strategies needs further development to support diversified client demands, global exposure, and thus the larger-scale investment capacity requirement needed by a large global asset manager with a long-term orientation.

[30]For more discussion on centralization, see CFA Institute, "T-Shaped Teams: Organizing to Adopt AI and Big Data at Investment Firms" (2021, p. 23). www.cfainstitute.org/-/media/documents/article/industry-research/t-shaped-teams.pdf.

Analysts with Training in Both Data Science and Investments

At the 1997 Worldwide Developer Conference, Steve Jobs was asked "to express in clear terms how, say, Java, in any of its incarnations, expresses the ideas embodied in OpenDoc." His answer was to start with the key business question of how to best serve customers rather than "sitting down with engineers and figuring out what awesome technologies we have and how we are going to market that."[31]

Data scientists often have similar questions for their fundamental investing counterparts, such as, Are the vast amounts of new metrics generated from new data sources and data science techniques fully understood and appreciated by fundamental investment professionals to realize their full potential? The correct answer is also similar: Data science endeavors should start with key investment questions rather than what data and metrics data scientists have already built and how to figure out use cases out of them.

A large portion of the data created daily proves of limited value to data scientists in evaluating investment opportunities. In contrast, we focus on data insights into key controversies and thesis metrics that allow us to have an informed and, often, differentiated view on a company's earnings power relative to market expectations three to five years out.

To this end, data scientists on the team (including some with a prior background as fundamental industry analysts) have been trained internally for both data science and fundamental research skills.[32] This training ensures data scientists have a strong knowledge to initiate meaningful discussions of investment use cases in multiple industries, as well as the key drivers of earnings power. Inevitably, some mistakes have been made along the way, but they created opportunities for data scientists and fundamental analysts to learn to speak the same language and strengthen their bond.

Data scientists with a background and training in fundamental investing have a better chance of cutting through conflicting and noisy data. For example, a high number of open job listings for a given company's salesforce provides a better indication of business momentum in some industries than in others. However, it can also mean employee retention is low and signal underlying issues for an organization. Understanding the nuances of each company and its peers (competitive positioning, geographic footprint, merger and acquisition strategy, etc.) is critical.

Armed with their alternative data findings, data scientists and fundamental industry analysts are able to engage company executives on a deeper level, provide additional insights into the inner workings of the company under study, and enrich our insights into an investee organization. It is our experience that analyzing data in isolation and without context can lead to erroneous conclusions.

Longer-Term Earnings Power Insights vs. Shorter-Term Operational Monitoring

Integrating alt-data into the fundamental investing framework allows investment teams to have an informed and, often, differentiated view on a portfolio company's earnings power three to five years out. At the same time, actively constructing and tracking a spectrum of short-term alt-data metrics to develop a comprehensive view of real-time operational metrics and/or risk toward our longer-term growth targets are equally important.

For example, many investment management firms have access to some form of credit card data. Indeed, there are many third-party providers in the marketplace that scrub this type of data for subscribers and provide an easily digestible format in real time. On the surface, credit card data in raw form and slightly lagged delivery provide less of a timeliness advantage. However, these data may allow data scientists working with long-term investment teams to perform in-depth proprietary research that is better aligned with their core investment disciplines.

Although many large, publicly traded companies are heavily researched and alt-data coverage is not universally and equally distributed, it is still possible, with balanced longer-term earning power insights and shorter-term operational metrics monitoring, to have a differentiated view about earnings power relative to market consensus. A few specific examples will help illustrate this point.

An Athleisure Brand

A working, fundamental thesis postulated that the operating margin of a leading athleisure brand was at an inflection point, driven by growth in the company's digital offering. Presented with this thesis, the data science team

[31]For a transcript and video of the remarks by Steve Jobs, see, e.g., Sebastiaan van der Lans, "Transcript: Steve Jobs at Apple's WWDC 1997," *Sebastiaan van der Lans—On WordPress, Blockchain, Timestamps & Trust* (blog, 3 January 2020). https://sebastiaans.blog/steve-jobs-wwdc-1997/.

[32]For a related discussion on the evolution of investment and data science function integrations, see CFA Institute, "T-Shaped Teams."

recommended a proprietary analysis breaking down the average selling price (ASP) between the company's digital and brick-and-mortar transactions. The fundamental industry analyst agreed with the data science team that such an analysis would be highly valuable, particularly because the management team of the company in question provided minimal disclosure regarding its digital offerings. The results of the analysis showed a significant premium for an extended period of time on digital transactions versus brick-and-mortar transactions.

Taking it one step further, the fundamental industry analyst later proposed tracking and monitoring the ASPs of the company under review versus a major competitor as further validation of the thesis. Indeed, the ASP premium achieved by the company in digital transactions over this major competitor was materially higher over a similar period of time. This analysis contributed to our portfolio management team's initiation of an investment position in the stock, supported by our increasing confidence that the operating margins of the company could expand meaningfully above consensus expectations over the medium term and long term. At the same time, because the company was one of the best covered names by alt-data sources, an array of metrics was actively monitored to glean a comprehensive view of the company's operations, including digital app engagement, consumer interests in various product lines, pricing power, and geographic new initiatives.

A Global Luxury Brand

A fundamental industry analyst began to lose conviction when recommending a well-known global luxury company because of the Chinese government's emphasis on "common prosperity." The growth of the luxury industry in recent years has been dominated by Chinese consumers. As a way to cut across the debate, given the analyst's view that the impact of a government initiative might not be easily spotted in conventional data, the data science team captured multiple operational KPIs.

Although we did see a modest dip in some short-term data metrics, such as social media momentum, we performed an extensive analysis of the company's customer base over a comprehensive period of time, which was more indicative of the company's ability to contribute to our long-term growth targets. Specifically, our analysis revealed that many months after the government's "common prosperity" push, the company's largest cohort of customers (female midlevel wage earners) increased their spending as a percentage of disposable income at this brand. Such an approach ultimately allowed our portfolio management teams to maintain ownership in the stock based on unobvious alt-data metrics that our data science team turned into an investible insight.

A Leading Asset Manager

A fundamental working thesis held that many investors were underestimating the growth of a new retail product offered by a leading asset manager. In addition to tracking regular asset flow data provided by the company, our fundamental industry analyst colleagues and the data science team developed an alternative method for tracking and monitoring traffic in the new product on the web portal of this specific retailer. This analysis contributed to our portfolio management teams' initiation of a position in the stock, bolstered by increased confidence in its earnings power and potential for upside growth.

Later, our ongoing tracking and monitoring of the data allowed our portfolio management team to play defense when necessary. Specifically, a sharp drop-off in web portal traffic observed by our data science team ultimately led to a more cautious outlook by our fundamental industry analyst and portfolio management teams. This, in turn, resulted in our portfolio management team's reduction (and in some cases elimination) of its position in the stock.

Conclusion

Data science and alternative data can help provide insights and added value to a range of investment processes, strategies, and portfolio management teams. However, there exists no single, self-evident methodology or road map to follow. Rather—and critically—distinct investment strategies, investment horizons, and portfolio teams each require their own individual approaches to the use of data science and the range of alternative datasets they can leverage.

The winning formula, in our opinion, is a partnership-driven approach that brings together data scientists, fundamental analysts, and portfolio managers with value-added datasets specifically built to address the firm's specific investment needs.