Leveraging Machine Learning in Portfolio Construction

FT Wilshire NxtGen Index Series

January 2025 | Wilshire Indexes Research

Wilshire Indexes

1. Executive Summary

AI and Machine Learning techniques are increasingly being used in finance.

Exponential improvements in computer power and wide-spread availability of very large datasets has meant that AI and Machine Learning techniques are increasingly being applied to various problems in finance.

One important application is in portfolio construction. The FT Wilshire NxtGen Index series aims to solve the classic portfolio optimization problem first set out by Markowitz in the 1950s. The key advance has been achieved by leveraging Machine Learning to yield better estimates of predicted return.

In this paper we introduce various performance and implementation properties of the FT Wilshire NxtGen Index series. We also provide some insights into how performance improvements arise.

Table of Contents

2. Introduction

Modern portfolio theory began with the Noble Prize-winning work by Harry Markowitz in the 1950s. He framed the now famous mean-variance problem:

> "Find the set of portfolio weights that maximizes risk adjusted return."

Since then, this has become the bedrock of many investment approaches that deviate from vanilla market cap weighted investing.

He proposed that the solution to this problem could be achieved via portfolio optimization, using three ingredients:

1. Predicted stock return
2. Predicted portfolio risk
3. Constraints to control diversification, concentration, turnover etc.

Over time, many attempts have been made to realize this program with varying degrees of success. Indeed, what has become apparent is that, whilst techniques for predicting risk have advanced over the years, the development of robust estimates for predicting return has lagged far behind.

Predicting risk has become a tractable problem - fundamental or statistical factor model-based approaches can deliver robust out-of-sample results for volatility and tracking error because stock correlation tends to be stable through time.

Predicting return has remained a challenging problem - traditional linear approaches like CAPM or fundamental factor models have struggled to produce good out-of-sample results.

However, in recent years, increasing computing power and the ability to manage and maintain large data sets means more sophisticated statistical techniques may be applied to obtain more robust predictions for return. These "AI" or "Machine Learning (ML)" approaches are notable because they:

- adjust for over-fitting in high dimensional data sets
- capture non-linear relationships

In this note we introduce Wilshire's NxtGen Index Series that leverages the recent advances in machine learning techniques to solve this classic portfolio optimization problem. In Section 2 we outline the methodology behind the indexes. In section 3 we present the results for various US and Global indexes. Finally in section 4 we provide some insights into how Machine Learning improves return prediction.

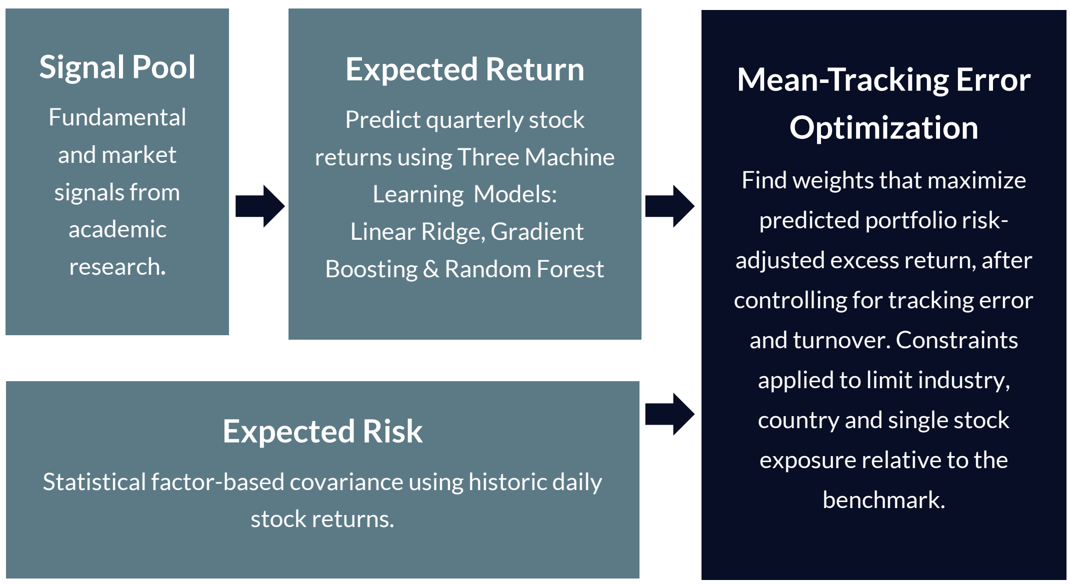
3. Methodology

The NxtGen Index Series uses an optimization approach to create portfolio weights that maximize predicted risk adjusted return whilst simultaneously limiting tracking error relative to a market capitalization benchmark.

There are two indexes based on the large and small cap segments of the FT Wilshire 5000 index: FT Wilshire NxtGen US Large and FT Wilshire NxtGen US Small. For global indexes, various carve-outs of FT Wilshire Global Equity Markets Series (GEMS) provide suitable starting universes for a NxtGen index. For example: FT Wilshire NxtGen Developed Large and FT Wilshire NxtGen Emerging Large.

Figure 1 gives a high-level overview of the index construction methodology.

Figure 1: NxtGen Index Series Construction



The signal pool splits into 12 distinct categories as is highlighted in Figure 2. The first five categories consist of components of the familiar five factors of Value, Size, Momentum, Quality and Low Risk that academics have studied over the years. The remaining seven categories represent the fruits of other relevant peer-reviewed and published academic research.

Figure 2: Fundamental and market characteristics

Category (#)	List of Fundamental and Market Characteristics
Value (11)	Book to Price, Cash Flow to Price, CFO to Price, Dividend Yield, Earnings to Price, Enterprise Component of Book to Price, Invertible Enterprise Multiple Leverage Component of Book to Price, Long-Term Reversal, Momentum Reversal, Sales to Price
Size (1)	Log Market Capitalization
Momentum (1)	Price Momentum
Profitability (8)	Gross Profit to Assets Ratio, Operating Profit to Equity Ratio, Profit Margin, Return on Assets (ROA), Return on Capital Employed, Return on Equity (ROE), Return on Net Operating Assets, Tax Expense Surprise
Low Risk (7)	12-Month Lottery, Idiosyncratic Volatility, Log Price, Market Beta, Market Beta 60-Day, Return Volatility, Semi-Variance
Accounting Conservatism (12)	Accounts Receivable Accrual, Accrual, Accrual CFO, Change in Deferred Revenue, Change in Net Non-Cash Working Capital, Growth in Long-Term NOA, Inventory Accrual, NOA Scaled, Percentage Operating Accrual, Percentage Total Operating Accrual, Tax Income Ratio, Total Accrual
Default Risk (11)	Cash Flow Volatility, Cash Ratio, Cash to Assets, Change in Financial Liabilities, Convertible Debt to Total Debt, Debt Coverage Ratio, Earnings Volatility, Leverage, Current Ratio, Long-Term Debt to Book, Market Leverage
Growth (12)	Cash Flow Growth, Change in Gross Margin, Change in Profit Margin, Change in ROA, Consecutive Quarters with Earnings Increases, Earnings Consistency, Earnings Growth, Earnings Surprise, EBIT Growth, Growth in Revenue minus SG&A, Revenue Growth, Revenue Surprise
Investment Conservatism (16)	Abnormal Capital Investment, Abnormal Capital Investment Scaled, Asset Growth, Change in Current Operating Assets, Change in Current Operating Liabilities, Change in Long-Term Investments, Change in Net Financial Assets, Change in Net Non-Current Operating assets, Change in Non-Current Operating Assets, Change in Non-Current Operating Liabilities, Goodwill Growth, Goodwill to Assets, Investment Growth, Investment to Assets Ratio, Investment to Capital Ratio, Net Non-Current Operating Assets
Issuance (9)	Composite Debt Issuance, Composite Equity Issuance, Earnings Distributed to Equity Holder, Growth in Debt, Net Cash Distributed to Equity Holder, Net External Financing, Payout, Payout Yield, Repurchase Binary
Liquidity (6)	Amihud's Measure for Illiquidity, Average Share Turnover, Average Traded Value, Change of Traded Value to Market Value of Equity, Traded Value to Market Value of Equity, Traded Value Volatility
Productivity (11)	Asset Turnover, Average R&D, Change in Asset Turnover, Firm Age, Firm Productivity, F-score, Operating Leverage, Overproduction, R&D Capital to Assets Ratio, R&D Reporting Biases, R&D to Sales Ratio

NxtGen uses three machine learning models trained on this set of fundamental financial characteristics to give improved estimates for return.

Linear Ridge – is a linear regression model, except that a penalty term is added to penalize the coefficients becoming too large. In comparison to the linear regression model, the linear ridge model is less likely to overfit in-sample.

Gradient Boosting - is a decision tree model, which builds trees in a sequential manner, where each new tree helps to correct errors made by the previous tree. As a decision tree model, gradient boosting captures non-linear relationships and interactions between variables. Learning rate, max depth and number of trees are main parameters to mitigate overfitting bias.

Random Forest - as another kind of decision tree model, makes predictions using a collection of decision trees, where each tree is built independently from "randomly" selected data points and features. The idea of averaging the prediction results of trees reduces bias in overfitting, while retaining the predictive power of the trees in the forest.

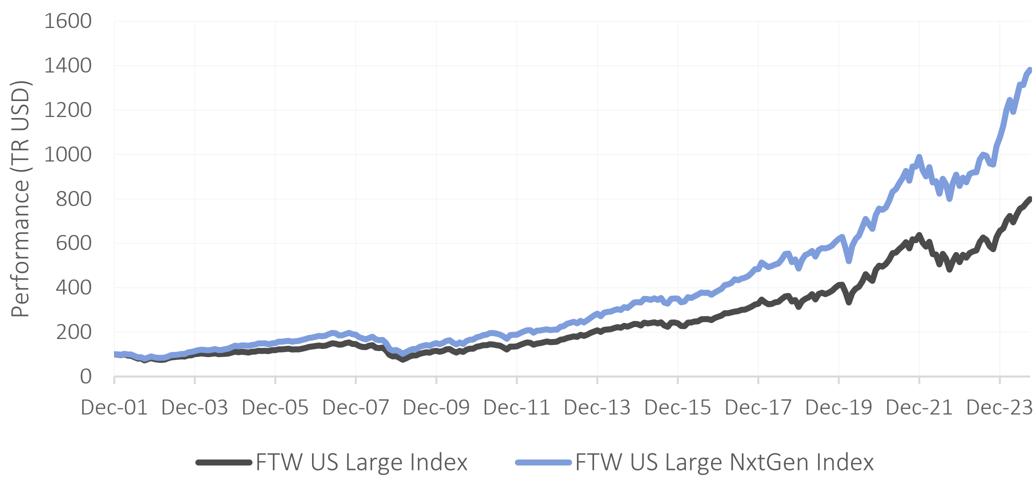
The models are trained on and make out-of-sample predictions of excess return for **all** stocks in Wilshire's Global Equity Market Series (GEMS) universe. Hence one set of predicted returns is used for the entire series.

4. NxtGen Index Series

Performance

Figure 3 shows the performance of the FT Wilshire US Large NxtGen index versus the FT Wilshire US Large Cap index between Dec 2001 to Sept 2024. This index was launched in July 2022. As remarked the returns used in this model represent a subset of predictions made on the global universe. Hence the set predicted returns used for the entire series have also been live for over two years.

Figure 3: Performance: FT Wilshire US Large Cap Index and FT Wilshire US Large NxtGen Index

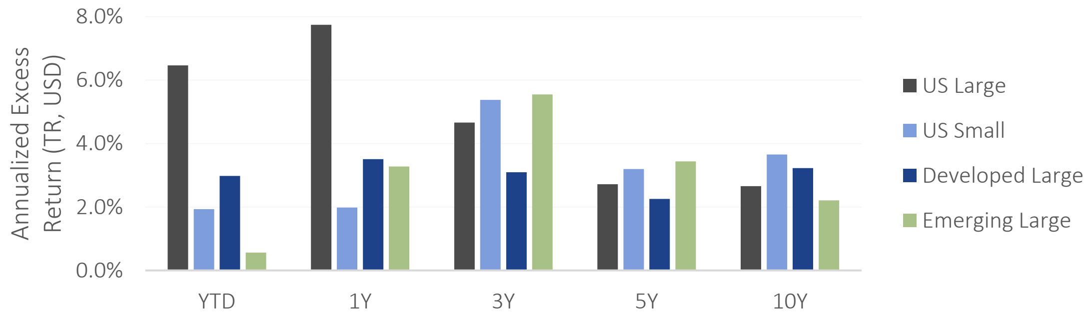


Source: Wilshire Indexes. Data as of September 2024.

This performance behavior is typical of all the other indexes in the series. Figure 4 shows the excess return for each index over various time horizons compared to their cap weighted parents.

Note that the NxtGen indexes outperform over each period. This contrasts with traditional smart beta indexes that have disappointed for much of the last 10 years. Indeed, over this period each NxtGen index has managed to outperform its cap weighted counterpart by well over 2.0% per annum.

Figure 4: Annualized[1] excess return over various time horizons: NxtGen Indexes



Source: Wilshire Indexes. Data as of September 2024.

Figure 5 shows the annualized performance and risk statistics for all launched indexes for the entire period of December 2001 to September 2024. It is noteworthy that for the Large Cap indexes, the excess return improves as we move from more to less efficient markets: US Large at 2.67%p.a., Developed at 3.60%p.a. and finally Emerging at 4.34%p.a.

Figure 5: Return, Risk and Implementation Statistics: Market Cap and NxtGen Indexes

	US LC	US LC NxtGen	US SC	US SC NxtGen	Dev LC	Dev LC NxtGen	EM LC	EM LC NxtGen
Geo. Ret.	9.57%	12.23%	9.87%	12.30%	8.70%	12.30%	9.54%	13.88%
Volatility	15.01%	14.06%	18.82%	17.94%	15.39%	14.18%	20.38%	19.67%
Return/Risk	0.64	0.87	0.52	0.69	0.57	0.87	0.47	0.71
Max DD	54.79%	52.37%	59.06%	61.47%	57.37%	53.30%	65.11%	-63.19%
Excess	-	2.67%	-	2.43%	-	3.60%	-	4.34%
Tracking Err	-	3.47%	-	3.49%	-	3.35%	-	3.62%
Info Ratio	-	0.77	-	0.70	-	1.07	-	1.20
Avg. N	576	145	1412	230	1941	408	918	203
Eff. N	118	79	882	133	314	148	143	101
Active Share	-	71.75%	-	84.69%	-	82.03%	-	75.94%
Top 10 wgt	21.97%	24.89%	2.99%	15.89%	12.05%	17.77%	19.95%	21.10%
2-Way T/O	5.87%	71.94%	26.48%	58.53%	10.26%	83.62%	20.65%	106.29%

Source: Wilshire Indexes. Data as of September 2024.

All indexes have a lower volatility, and most have improved drawdowns compared to their benchmarks. Tracking errors are maintained at the

[1] All time periods show annualized data apart from YTD which is excess return since 29th Dec 2023

3.5%p.a. level, making for healthy information ratios (excess/tracking error) approaching or above one.

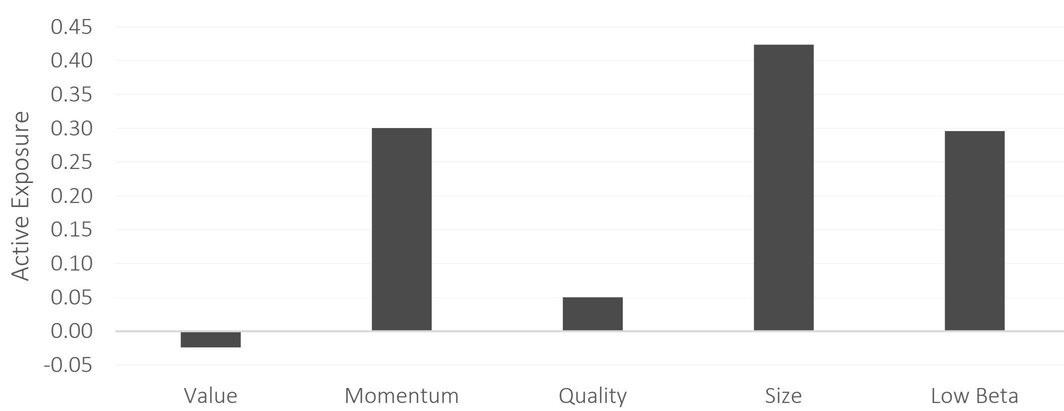
Note that each NxtGen index contains approximately one fifth of their parent's constituents. Transaction costs are controlled by ensuring that annual two-way turnover is in the range 60-110%p.a.

Factor Exposures

Some insight into the performance of the NxtGen indexes can be gained by examining their factor exposures (for factor definitions see [Factor Index Series Methodology](#)). Figure 6 shows the average active factor exposures over the last 10 years for FTW US Large NxtGen Index. Note positive exposures to all the traditional factors except to value. This is unsurprising as it has been well documented that value has underperformed over the last 10 years. Note also the positive exposure to momentum, the strongest performing factor of the last 10 years.

Harder to explain is the significant exposure to size. This is surprising given the size factor's poor performance over the last few years, due in no small part to the rise of mega cap stocks like the Magnificent Seven.

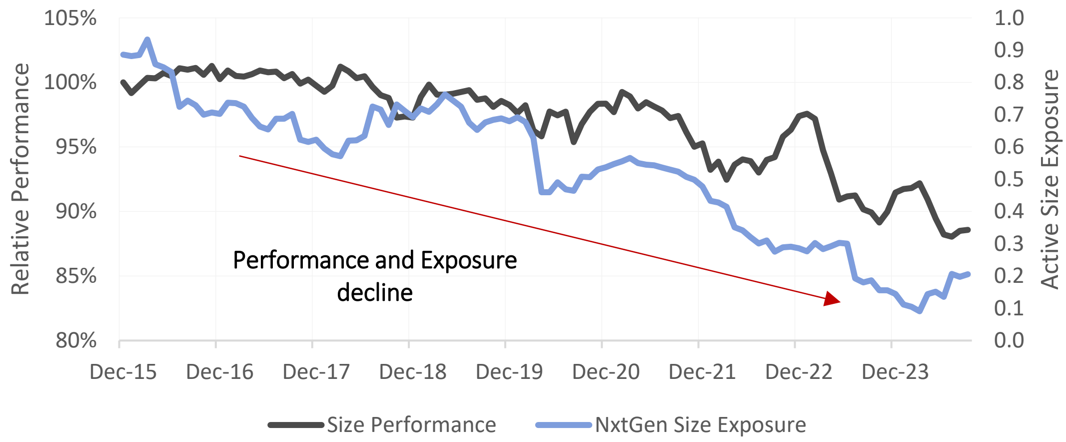
Figure 6: Average Active Factor Exposures for FTW US Large NxtGen Index



Source: Wilshire Indexes. Data as of September 2024.

However, the dynamic nature of NxtGen is made clear by tracking its size exposure through time against the relative performance of the Size factor in Figure 7. The left-hand axis measures the performance of the Size factor (black), whilst the right-hand axis measures active size exposure of the NxtGen index (light blue).

Figure 7: Active Size Exposure for FTW US Large NxtGen vs Size Factor Performance

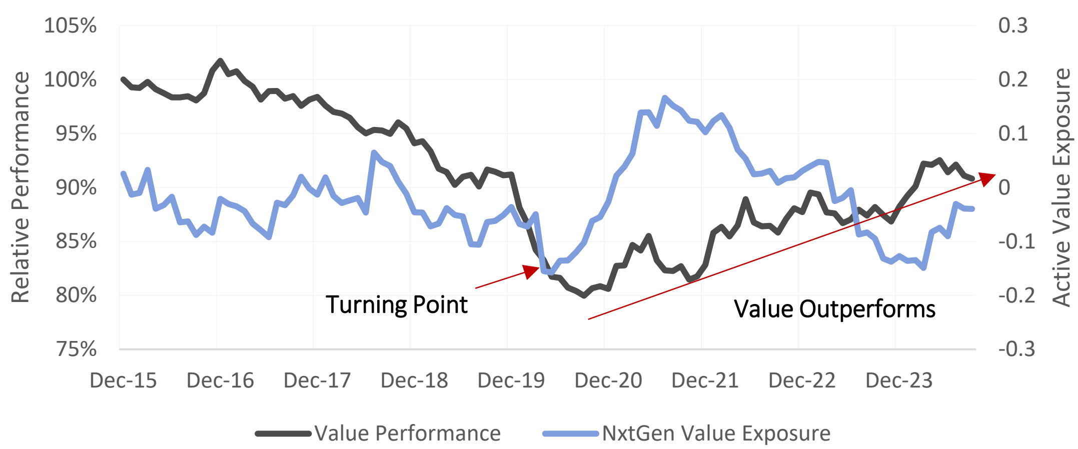


Source: Wilshire Indexes. Data as of September 2024.

The NxtGen Index's exposure to small caps decreases through time tracking the performance of the size factor. Clearly this is an example where Machine Learning detects a poorly performing factor and then gradually rotates away from it.

The same analysis is applied to the value exposure of the NxtGen index in Figure 8, where this time we see evidence of NxtGen rotating into a factor when that factor's performance improves. Note that the NxtGen value exposure is mostly negative or zero up to a turning point in the middle of 2020, when it begins to trend upwards. Shortly afterwards the value factor begins to outperform.

Figure 8: Active Size Exposure for FTW US Large NxtGen vs Size Factor Performance



Source: Wilshire Indexes. Data as of September 2024.

This is evidence that part of Machine Learning's boost to the performance of the NxtGen Indexes can be interpreted as the "timing" of traditional factors.

Note however, that these results arise from the complex interaction between return/risk prediction and portfolio optimization subject to constraints. In the next section we will better isolate the role that Machine Learning return predictions play in improving performance outcomes.

5. Improving Return Prediction

To better understand how Machine Learning models of predicted return improve performance outcomes, we need to introduce the notion of Information Coefficient (IC). This is the rank correlation of return predictions made for a universe of stocks over a given period with the realized returns over that same period. The higher the correlation the more robust the return predictions.

Figure 9 illustrates the difference between the pattern of predicted and realized returns for low and high information coefficients.

Figure 9: Predicted versus Realized Return: Low and High Information Coefficient

Source: Wilshire Indexes.

For an IC of 1% the relationship between predicted and realized return is almost random. However, for an IC of 70% we see a clear linear relationship emerge.

Here the information coefficients are based on predictions made at the beginning of a 3-month period and realized returns at the end of that period. The ML models are trained on all (in sample) data prior to the 3-month period to give the out of sample predictions. We shift forward by one month and repeat the process giving an IC for each month[2].

[2] 10 years of analysis starting in December 2014

We have seen that the following combination of properties seems to improve the performance of the NxtGen Indexes:

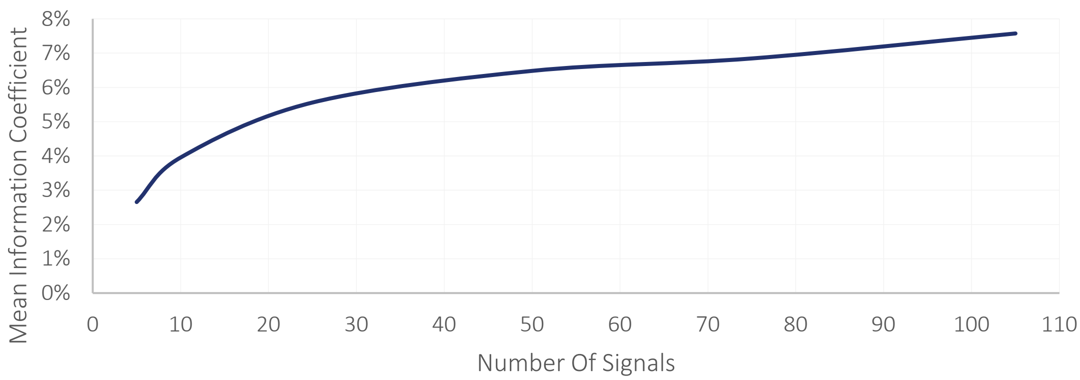
- More input signals – we used **105** financial signals compared to say, the traditional five factors.
- More sophisticated statistical models – we used **linear ridge, gradient boosting and random forest** in sample to avoid overfitting inherent in traditional linear models.
- A wider universe of stocks – we trained our ML model on a **global universe** of stocks rather than a US only universe to obtain excess return predictions for the US universe.

In this section we examine how much each of these factors contribute to the improvement in the predictive power of the ML expected return model.

Number of Input Signals

To quantify the effect of more input signals, we compare the results of training using randomly drawn subsets of 5, 10, 25, 50, 75 and all 105 signals set out in Section 3[3]. Figure 10 displays the mean monthly information coefficient as a function of the number of signals.

Figure 10: Mean Monthly IC verses Number of Signals



Source: Wilshire Indexes.

Note that most improvement occurs moving from a handful of signals up to around 50. After this point there is continual improvement but at a much more incremental level. This makes sense, since if the relationship did not display this law of diminishing returns, one could expect there to be a practically achievable number of signals that would yield perfect predictions with ICs of 100%!

[3] The 5 signals form a subset of the 10 signals, the 10 signals a subset of the 25, etc., all the way up to 50 signals forming a subset of the 75 signals. The average of ten such random nested subsets are used in the calculations.

Traditional versus Non-Linear Models

Figure 11 shows the monthly statistics for the information coefficients in the US Large Cap universe for each of the Machine Learning models trained on the global universe. For comparison we have also included a "Traditional" factor model that uses simple linear regression on the standard set of five factors: Value, Momentum, Quality, Size and Beta, trained on the US Large Cap universe.

Figure 11: Monthly ICs Statistics: Traditional, Linear Ridge, Gradient Boosting, Random Forest, and Ensemble

	Mean	St. Dev.	Max	Min	Median	% Positive
Traditional	4.02%	17.37%	45.15%	-47.41%	4.91%	61.90%
Linear Ridge	7.36%	18.96%	43.48%	-50.07%	8.81%	68.45%
Gradient Boosting	7.62%	18.39%	42.77%	-48.39%	10.44%	69.64%
Random Forest	6.99%	18.24%	45.80%	-48.96%	7.63%	71.43%
Ensemble	7.84%	19.30%	43.26%	-49.44%	9.63%	69.64%

Source: Wilshire Indexes.

The first thing to note is that the ML models achieve an IC of around 7% compared to 4% for the traditional model. This modest improvement should be considered in the light that an IC of 5% is considered good enough to build a successful portfolio. This also emphasizes that the use of ML techniques represents an *evolution* rather than a *revolution* in portfolio construction.

Interestingly the addition of non-linearity does not seem to yield much (if any) improvement in the forecasting power. Indeed for most years Linear Ridge, Gradient Boosting and Random Forest yield very similar results.

However, for certain years, there are significant differences in the forecasting powers of the models, for example, during the Covid crisis year of 2020 (Linear Ridge 3.06% vs Random Forest -5.09%) and during the following recovery year of 2021 (Linear Ridge -0.2% vs Random Forest 5.10%). This suggests that a more robust estimate may be obtained by combining the three models.

This is indeed the case as can be seen in Figure 11, where single model mean ICs in the range of 7.0 - 7.6% are improved by taking an ensemble of the models to yield 7.8%.

Wider training universes

A key difference between traditional and ML return predictions for the US large cap universe is that the ML models are trained on global stocks rather than just the US large cap universe. Traditional factor approaches tend only to use data that is native to the universe of consideration. In this section we will see how expanding the training set from US Large stocks to US Large plus Small stocks to global stocks affects outcomes.

Figure 12 shows how the widening of training universe affects the IC of stocks in the prediction universe.

Figure 12: Information Coefficient: Training versus Prediction Universe

		Prediction Universe		
		US Large Cap	US Large + Small	Global
Training Universe	US Large Cap	5.39%	4.24%	5.40%
	US Large + Small	7.40%	6.49%	6.39%
	Global	7.84%	6.52%	8.39%

Source: Wilshire Indexes.

In general, as the scope of the training universe increases so does the IC of the universe the predictions are applied to. So, for example, the IC calculated for stocks in the US Large Cap universe increases as the training universe moves from US Large (5.39%), to US Large plus Small (7.40%) and finally to Global (7.84%).

In other words, the returns/signals of stocks outside of the US Large Cap universe contain valuable information that influences the behavior of stocks entirely within the US Large Cap universe.

Also worthy of note is that, when the ML models use US Large Cap as their sole training set, they still deliver better results for mean IC of US Large, than the traditional factor model (5.39% versus 4.02%).

6. Conclusions

Markowitz's classic portfolio optimization problem relies on reliable out-of-sample predictions of stock return. We have shown that machine learning applied to high dimensional data sets can give rise to a robust model of expected return and contrast that with traditional approaches that have had limited success over the years.

We have seen how the fidelity of expected return models can be quantified by the notion of information coefficient (IC) which measures the correlation between predicted and realized return.

There is a monotonic relationship between IC and the number of financial signals used in the training set. However, as the number of signals increases, the improvements become increasingly marginal.

We find that the wider the training universe of stocks the better the IC. For example, training our ML models on a global universe of stocks yields superior results for predictions on a US universe than those obtained by training on the US universe alone.

We see little evidence that non-linear models yield better return predictions on average. However, the combination of different ML models (linear and non-linear) does seem to give rise to modest improvements.

Finally using our ML return predictions as inputs to Mean-Tracking error optimization, we have created a new generation of smart beta, the NxtGen Index Series. Their promising results demonstrate that even modest improvements in ICs lead to profound benefits for portfolio performance.

Legal Notices

1. Intellectual Property

No material published by Wilshire OpCo UK Limited and its affiliates ("Wilshire Indexes") should be construed as granting any license or right to use any copyright, trademark, service mark or other intellectual property right for any purpose whatsoever without the written permission of the lawful owner and an appropriate license.

Wilshire and Wilshire 5000 are registered service marks owned by Wilshire Advisors LLC and are used under license. "FT" is the trade mark of the Financial Times Limited and is used under licence.

"Information" means any data, text, or other material published or made available by Wilshire Indexes whether owned by Wilshire Indexes or a third-party. This includes presentations, factsheets, product files, methodology documents, benchmark statements, policies or other documents relating to the operation, calculation or administration of Wilshire Indexes benchmarks, whether supplied physically or published electronically, for example on Wilshire Indexes websites or the websites or materials of Wilshire Indexes clients that relate to Wilshire Indexes.

All Information is provided for information purposes only and is made available "as is" without warranty or other commitment of any kind.

Use or distribution of Information requires a license from Wilshire Indexes.

No part of the Information contained herein may be copied, stored in a retrieval system, distributed, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission from Wilshire Indexes.

None of the Information may be used to create derivative works or to verify or correct other data or Information. For example (but without limitation), the Information may not be used to create indexes, databases, risk models, analytics, software, or in connection with the issuing, offering, sponsoring, managing or marketing of any securities, portfolios, financial products or other investment vehicles utilizing or based on, linked to, tracking or otherwise derived from the Information or any other Wilshire Indexes data, Information, products or services.

2. Disclaimers

Wilshire Indexes entities and their respective directors, officers, employees, partners and licensors make no claim, prediction, warranty, representation or other commitment whatsoever, expressly or impliedly, as to the accuracy, timeliness, completeness, merchantability of any Information available or of results to be obtained from the use of the Information or the fitness or suitability of the Information for any particular purpose to which it might be put.

Any representation of historical data is provided for informational purposes only and is not a reliable indicator of future performance. The Information may contain back-tested data. Back-tested performance is not actual performance but is hypothetical. There may be differences between back-tested performance results and actual results subsequently achieved by any investment strategy.

To the fullest extent permitted by applicable law, Wilshire Indexes entities and their respective directors, officers, employees, partners, and licensors accept no responsibility or liability for:

a) any loss or damage in whole or in part caused by, resulting from, or relating to any error (negligent or otherwise) or other circumstance involved in procuring, collecting, calculating, analyzing, editing, transcribing, transmitting, communicating, or delivering the Information or from use of the Information; and

b) any direct, indirect, special, consequential, or incidental damages whatsoever, even if Wilshire Indexes (or the relevant person) is advised in advance of the possibility of such damages, resulting from the use of, or inability to use, the Information.

Wilshire Indexes entities and their respective directors, officers, employees, partners, and licensors do not provide investment advice and nothing in the Information or any links thereto, including statistical data and industry reports, should be taken as constituting financial or investment advice or a financial promotion.

Wilshire Indexes entities and their respective directors, officers, employees, partners, and licensors are not responsible for and make no representation regarding the appropriateness or suitability of using, or investing in any financial instrument or entering into any contract linked to, Wilshire Indexes benchmarks or other Information provided by Wilshire Indexes and any decision to engage in such use or to invest in any such instrument or enter into any such contract should not be made in reliance on Wilshire Indexes benchmarks or other Information provided or published by Wilshire Indexes. All Information is impersonal and not tailored to the needs of any person, entity, or group of persons.

To the fullest extent permitted by applicable law, Wilshire Indexes entities and their respective directors, officers, employees, partners and licensors will not be liable in contract or tort (including negligence), for breach of statutory duty, misrepresentation or otherwise, in respect of any inaccuracies, errors, omissions, delays, failures, cessations or changes (material or otherwise) in Wilshire Indexes benchmarks or other Information.

Wilshire Indexes is the trading name of Wilshire OpCo UK Limited, a company registered in England and Wales under registered number 12991656, with its registered office at C/O Tmf Group, 13th Floor, One Angel Court, London EC2R 7HJ, United Kingdom. Wilshire OpCo UK Limited is authorized and regulated by the Financial Conduct Authority as a benchmark administrator, with reference number 985021.