Norwegian School of Economics
Bergen, Spring 2021

Interpretable decision support for international stock index investments

Using heterogenous machine learning stacking for more accurate and resilient mid-horizon predictions

Aleksej Hoffärber & Karo Rönty

Supervisor: Lars Jonas Andersson

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

It remains challenging to beat the wisdom of the crowd by predicting future stock index returns. With the rise in financial machine learning, many new approaches are introduced but predictive accuracy, financial performance, and interpretability are rarely connected. We show that it remains possible to predict future MSCI index returns five years ahead and across seven country indices. We train six base-models that differ in methods and data pools to reduce generalization errors. The stacked combinations of these models reduce MAPE by up to 60.5% if compared out-of-sample to the historical mean forecast benchmark, while passing the Diebold-Mariano significance tests. In financial terms, our stacked models outperform the equal-weight portfolio by 1.4% to 2.1% in yearly CAGR and by 34% to 72% in Sharpe Ratios. Our research fills a gap for portfolio optimization, by providing more reliable inputs for future returns, and potential for fundamental research, by introducing a method that combines predictive accuracy with interpretability.

Acknowledgements

We want to express our genuine gratitude to our supervisor Lars Jonas Andersson for his continuous support during our research period. Not only did we share a common vision to combine four interdisciplinary areas with this research, but also a mutual interest to question the applicability of machine learning concepts to finance. Hence, we could unite our results on predictive and financial performance with significance testing and model-agnostic interpretability, a feat that may encourage more interdisciplinary research in the area of financial machine learning.

Our sincere thanks go to our friends and study colleagues at NHH, Egor Zmaznev and Janik Weigel, for questioning and discussing our ideas and concepts. We also want to thank Justin Harlan for his valuable insights on how to make our research more understandable for a wider audience.

Table of Contents

List of Figures

List of Tables

Abbreviations

AICc	–	Small-Sample Corrected Akaike-Information Criteria
ARIMA	–	Auto-Regressive Integrated Moving Average
BRICS	–	Brazil, Russia, India, China, and South Africa
CAGR	–	Compound Annual Growth Rate
CAPE	–	Cyclically Adjusted Price-to-Earnings Ratio
EMH	–	Efficient Market Hypothesis
EN	–	Elastic Net
ETF	–	Exchange-Traded Fund
LASSO	–	Least Absolute Shrinkage and Selection Operator
LSTM	–	Long-Short Term Memory Network
MAE	–	Mean-Absolute Error
MAPE	–	Mean-Absolute-Percentage Error
MSCI	–	Morgan Stanley Capital International
OLS	–	Ordinary Least Squares
RMSE	–	Root-Mean-Squared Error
VAR	–	Vector Auto-Regressive Model
XGBoost	–	eXtreme Gradient Boosted Decision Trees

1. Introduction

Efforts to predict stock market returns are the core of financial research and as old as the stock market itself. Accurate predictions directly affect how assets are selected based on expected future returns, and influence the financial returns gained throughout the period. Financial instruments have become more accessible for retail investors due to passively managed funds, robo-advisors, apps, and commission-fee reductions. This change in accessibility makes portfolio optimization a crucial part of investment decisions, especially if future returns are predicted using inaccurate models, limiting potentially achievable returns for the investor. A popular benchmark choice for return predictions is the historical mean of past returns. Due to its simplicity, it is less sensitive to noise than more complicated approaches, such as ARIMA, Random Forest or XGBoost. The Efficient Market Hypothesis (EMH) assumes that future expectations of asset values are fully reflected in present asset prices. In practice, this assumption is often proven by the historical mean outperforming sophisticated models out-of-sample, despite using additional features.

On the contrary, fundamental analysis, which focuses on predicting the future intrinsic value of an asset, showed that factors, such as dividend-yields, price-to-earnings ratio, and interest rates can lead to accurate return predictions (Guoying & Ping, 2017). Fama (2014), the main contributor behind the EMH, agreed that factors can explain future returns in the mid-term. The back and forth on market efficiency and the feasibility of return predictions led to a consensus, that return predictions are possible; yet, the uncertainty remains to which degree (Welch & Goyal, 2008) and whether they contribute to profit gains.

The primary objective of our research is to predict the five-year compound annual growth rate (5-year CAGR) for seven stock indices more accurately than naïve and mean forecasts. Our secondary objective is to validate if predictive performance is related to profit gains. Hence, we implement stacked models, which combine predictions of heterogenous base-models, such as time-series, regularized regression, and machine learning methods. To increase diversity, we use pooled and single-country datasets to check if using data of multiple MSCI indices can enhance the learning of feature-target relationships. Due to the mid-term return interest, we focus on fundamental and macroeconomic features instead of

technical ones. We opt for CAGR prediction because of its versatility in comparing different assets over multi-year periods, making it important for mid-term investors (Kyriakou et al., 2020). We validate our results out-of-sample across ten years, a span twice the length of the forward prediction to guarantee leakage-free predictions. To validate financial performance, we use a long-only strategy with different rebalancing cycles, which are based on returns predicted by our models. We add feature effect measurements to demonstrate that multi-layered models can remain interpretable for investors and researchers.

Our results show that our stacking methods lead to more accurate predictions and higher predictive resilience. These improvements lead to a MAPE reduction of up to 60.5% across the seven MSCI indices, which are mostly statistically significant according to the Diebold-Mariano test (1995). Specifically, the pooled models are the strongest drivers for resilience and directional accuracy. Yet, our approach does not result in above-benchmark performance for one index. Also, our findings display that stacked models outperform the benchmarks by 1.4% to 2.1% in yearly CAGR and by 33% to 72% in Sharpe Ratios, which remain statistically significant at the 10% level based on equality tests from Wright et al. (2014). To balance the limitations of ex-post backtesting, we illustrate the effect of CAPE, dividend-yields, and 10-year government bond yields on CAGR predictions. These feature effects support theories, such as the capital asset pricing model (CAPM) and theories based on the predictive power of fundamental factors (Campbell & Shiller, 1988; Radha, 2020).

Our research considers challenges from economics, econometrics, finance, and machine learning. We have written this paper so that it encourages more collaboration between finance and machine learning by remaining understandable to both research areas. Thus, Chapter 2 displays the importance of accurate predictions for asset management, investors, and the stock market. Chapter 3 analyses stock market price setting mechanics and reviews under which circumstances mid-horizon predictions (MHP) are feasible. Chapter 4 examines research in financial machine learning that applies fundamental theory to predict mid-term returns. Chapter 5 introduces the data set, while chapter 6 discusses the model selection as well as data pooling and heterogenous stacking. Chapter 7 provides the predictive and financial results and compares our findings with relevant domain research. Chapter 8 discusses limitations and defines implications for future research.

2. Importance of Accurate Mid-Term Predictions

More accurate predictions are crucial for institutional and private investors, as well as for the efficiency of stock markets. Financial machine learning research has contributed to the growth of new approaches, but those focus mainly on improving predictive results. The required degree of know-how to understand new approaches makes it increasingly difficult for private investors, regulators, and research neighboring fields to contextualize these findings because of a lack in consideration for risk and returns or missing model interpretability. Therefore, our secondary objectives are focused on deriving a connection between predictive, risk, and return performance. Additionally, we derive feature effects to illustrate how CAGR predictions are affected by different fundamental and macroeconomic features.

For decades, investors have been interested in methods that increase returns for a given portfolio while minimizing the risk. While portfolio optimization finds those assets that balance risk and return in an optimal way, the accurate and quantitative pre-selection of assets is often overlooked (W. Wang et al., 2020). Portfolio optimization techniques require the modeler to predict future return and risk (Ho et al., 2015), which is usually done using the historical mean of past returns. The lower the accuracy of future returns, the less reliable are the results of asset allocations based on portfolio optimization (W. Chen et al., 2021). Any increase in accuracy leads to an increase in the probability that the given optimization and the recommended portfolio is correct (W. Wang et al., 2020). Furthermore, newer portfolio optimization techniques focus on predicting risks due to their higher autocorrelation (Scalable Capital, 2016), while future returns are still approximated using inflexible standard benchmarks.

Technological advancements have led to a broad data access and an accelerated decision-making process (Cervelló-Royo & Guijarro, 2020). This development requires portfolio managers and investors to combine domain expertise and modeling knowhow to remain successful generating wealth (Sorensen, 2019). Because domain expertise and data access in the field of research are not as extensive as the ones in the industry, this paper focuses on innovation in the modeling area. The prediction superiority of machine learning models

was already proven but only in short-term, arbitrage-oriented high-frequency trading (HFT), and day-ahead price predictions (Hou et al., 2020).

To the institutional and private investors interested in mid and long-term investments, short-term gains do not matter as much. Furthermore, it is recommended to hedge over a longer horizon instead of against sudden crashes (McQuinn et al., 2020) to mitigate risk. Hence, predicting future events needs to consider economic and fundamental data due to their superior mid- and long-term return prediction accuracy. Increasing liquidity in stock indices investing, facilitated by exchange-traded funds and robo advisors (Baumann et al., 2018), shows that current asset management providers have enough motivation to optimize their asset pre-selection methods to create more value for the investor. If MHP lead to more accurate results, equities and securities with above-market return potential can be picked more confidently (Barak & Modarres, 2015). If risk is minimized simultaneously this can lead to better investment strategies (Jiang et al., 2020; Zopounidis et al., 2015), which benefit the investor and the asset manager.

From the macroeconomic perspective, a slight increase in accuracy does not directly translate into profits (Leung et al., 2000), but can facilitate the comparison of indices between different markets. In our setting, this means an increased capital flow to those indices that are projected to perform well in the selected horizon. This capital allocation benefits both institutional and private investors as investments are channeled to the best prospects due to accurate return predictions. Therefore, the business sector has a higher incentive to create sustainable and effective profit streams in order to receive funding from investors at a fair price. In sum, wealth aggregation for investors and ease of financing access for corporates can be enhanced and made more sustainable by increasing market efficiency. Even though the stock market is often classified as an “almost perfect market”, efficiency is often classified as an ongoing issue for regulators (Rounaghi & Nassir Zadeh, 2016).

The application of machine learning methods for MHP remains fairly new despite the list of benefits. Most approaches are validated only statistically without using feature effects (H. Wang et al., 2019). Moreover, a validation on a limited number of stocks or indices fails to consider the complexity of the global financial market and to validate the predictive robustness across markets, which in turn affects asset selection based on feature effects.

3. Predictability of Stock Indices

There is a consensus in research that return predictions are possible and lead to meaningful explanations using fundamental features (Campbell & Thompson, 2008). Still, it remains widely uncertain in which circumstances and to what degree these predictions are viable out-of-sample (Welch & Goyal, 2008) and whether they are robust across multiple countries. Based on previous research and our results, we classify mid-term horizon predictions (MHP) as possible, despite the nature of financial time-series data, the economic theory of efficient markets, and behavioral economic theory. Hence, we focus on extracting mid-term value drivers, which are of key interest for mid- to long-term investors.

3.1 The Nature of Financial Time-Series Data

Dealing with financial time-series data is regarded as a complex matter. Cross-sectional time-series using multiple features per country may lead to dimensionality issues during modeling, which requires either multivariate models to identify the relationships or feature selection to reduce the feature space. More problems await forecasters due to the inherent noise, non-stationarity, and chaotic behavior (Manish & Thenmozhi, 2011). Noisy relationships occur if the past information contained in the selected features is not enough to predict future relationships, leading to correlation in the error term. Non-stationarity occurs if the distribution of features and targets varies over time, which is regularly observed for different features, e.g. the cyclically-adjusted price-to-earnings ratio (Umlauft, 2020). Chaotic behavior occurs if relationships among features and target are time-varying or easily influenced by unforeseen global shocks that lead to extreme reactions, such as bull and bear markets or value-growth cycles. Those lead to a price-setting behavior in the short-term that is more similar to a random walk, even though price-setting becomes more deterministic in the long-run (Manish & Thenmozhi, 2011). Additionally, the information efficiency of financial markets leads to fast strategy adaptations, meaning that relationships or specific patterns are unlikely to persist, as they would be exploited by traders. Also, the technological evolution leads to faster information spread (Sharma & Thaker, 2015) in news, further reducing the persistence.

3.2 Efficient Market Hypothesis and Random Walk

The Efficient Market Hypothesis (EMH) provides an explanation for the random nature of prices. In short, information spread and rational expectations always lead to the stock price reverting back to its inherent fundamental value in the long-run (Andersson & Lauvsnes, 2007). EMH went through an evolution of evidence, highlighting that markets themselves vary in efficiency over time and across markets (Alvarez-Ramirez et al., 2012); therefore, giving forecasting practitioners hope for explainable returns.

Many studies analyzed the random behavior of stock price changes (Kendall & Hill, 1953; Working, 1934). The EMH explains these random fluctuations as slight prices changes around their true, intrinsic values (Fama, 1965a). EMH assumes that all stock market participants share the same access to public information (Fama, 1965a). Therefore, stock market prices cannot be predicted, because price setting is the result of rational behavior (Fama, 1965b) or the competition between market participants (Samuelson, 1965). Consequently, the security price on the market will be a close estimate of its intrinsic value (Sharma & Thaker, 2015). The EMH predicts that fundamental features cannot predict future returns, because all information is reflected in the price leaving only noise for the models, which would lead to non-significant out-of-sample performance. Later, EMH began to separate three forms of market efficiency. The weak form suggests historical data cannot be used for future price prediction, as it follows a random walk. Only fundamental data or news can be used for potentially outperforming predictions, but the latter is also assumed to be random (Andersson & Lauvsnes, 2007). The semi-strong form indicates that all public information is considered by the market, leaving only private insider information potentially impactful. The strong form defines that all information is incorporated in the present prices.

On the one side, the weak form of market efficiency was identified for different markets using monthly returns, considering randomness, autocorrelation, and cointegration tests (Sharma & Thaker, 2015), meaning that prices or past returns do not yield any explanatory power. Reviews of long-horizon prediction (LHP) practices classified their increase in return explanation as unrealistic because of missing consideration of transaction costs, and insufficient out-of-sample testing (Moreno & Olmeda, 2007). On the other side, BRICS indices have shown long-memory behavior, meaning that returns and momentum depend

on lagged prices (Aye et al., 2014), disproving the weak form of the EMH. Further research points out that the random nature of stock prices was only observed in the short-run (Manish & Thenmozhi, 2011). In 2008, Campbell and Thompson validated an explanatory effect of fundamental and macroeconomic features in linear regression models that can be used for return prediction. Fisher (2005) classified that the predictive power of different features depends on intervals, frequency, and context (country, market timing), introducing the term of scale specific predictability. This consensus led to research, showing that MHP and LHP based on monthly data and fundamental features can beat previous benchmarks (Radha, 2020; H. Wang et al., 2019), with new approaches being introduced ever since.

3.3 The Influence of Behavioral Finance

Behavioral finance studies investor biases and delivers arguments against stock market efficiency and rational expectations. Direct implications for our methodology are the focus on stock indices and MHP intervals. Behavioral finance presented challenges to the EMH by questioning rational behavior (Oztekin et al., 2016). This criticism led to an emergence of the alternative market hypothesis (AMH), which considers markets to be evolutionary systems in which participants, their behavior, and feature-target relationships evolve over time. Markets often need to adjust to shocks, which lead to investor overreaction (De Bondt & Thaler, 1987). In addition, information bias during information uncertainty (Tversky & Kahneman, 1973) leads to adverse decision-making by placing too much weight on initial or current information (Tversky & Kahneman, 1974). Both characteristics can be observed in different markets (Kaestner, 2011). While these biases support the focus on longer prediction intervals, the two main behavioral finance biases are disposition effect and herding.

3.3.1 Disposition Effect & Loss Aversion

Disposition shows that investors tend to sell good-performing stocks too early and hold on to below-market performers for too long. This is explained by loss aversion (Kahneman & Tversky, 1979) and regret aversion (Thaler & Johnson, 1990), indicating that emotional attachment to a stock impairs rational decision-making (Peterson, 2012, p. 203). Because disposition effects are connected to individual stocks, we focus on stock indices to reduce

fluctuations due to emotional attachments. These aggregations helps to make returns solely dependent on macroeconomic and fundamental disclosures (Sezer et al., 2019).

3.3.2 Herd Behavior

Herding is apparent if individual investors mimic the actions of other investors (Shiller, 2000, p. 147) or prioritize opinions of other investors over their own prior beliefs (Devenow & Welch, 1996). This trait can lead to erroneous behavior, such as increase in liquidity, information bubbles (Devenow & Welch, 1996), and crashes (Shiller, 2000, p. 155). If investors are afraid about their relative performance to the market (Scharfstein & Stein, 1990), they will more likely opt for standard practices, such as relying on historical averages as method for return predictors. Our approach considers herding and focuses on a prediction period of five years ahead, to smooth out short-term fluctuations that follow a more random pattern, thus increasing mid-term performance.

3.4 Mid-Term Prediction Window

Kyriakou et al. (2020) showed that prediction length is negatively correlated with encountered noise in the context of financial time-series, but just up to a length of five years. Their applied model led to stronger and more significant explanatory power, while correcting for overlapping returns that result by averaging over five years of returns. Additionally, the researchers showed that five-year forecasts show a stronger explanation due to being less volatile. Campbell and Shiller (1988) showed early on, but only for the S&P 500, that averaging returns over several years increases accuracy of predictions for stock indices. Overall, the explanatory effects of fundamental features increases with longer horizons (Jin et al., 2014), which is shown in more detail in the next section. But, possible changes in economic, political systems or investor motivation need to be accounted for accordingly (Nti et al., 2020) by not extending the prediction interval indefinitely. We follow the guidelines from Radha (2020) and focus on five-year returns in order to balance explanatory effects, regime changes, and interests of mid-term investors. Additionally, this interval selection includes at least one business cycle to extract mid-term value-drivers (Umlauft, 2020), which remains important for the eventual application of feature effect measures.

4. Literature Review

After discussing the rationale for mid-horizon predictions (MHP) and stock indices as asset classes, we review key literature to identify which models, features, and targets are most suitable for our objective. First, we examine why stacking is a suitable ensemble model to increase predictive resilience and accuracy. Next, we consider the latest fundamental research to discuss a suitable feature set and target specification. Lastly, we review a financial machine learning methodology that is the closest to our eventual methodology, which helps us to define our research question and gap we fill with our contribution.

Beside bagging and boosting, stacking is the third ensemble model type and builds aggregated predictions by fitting multiple heterogenous base-models (Allende & Valle, 2017). Stacking allows for flexibility in model choices, but also for combining models which use different data pools (Ribeiro & dos Santos Coelho, 2020). Despite the advantages of stacking, its usage in return forecasting and finance is rare and limited in its application to international assets. Jiang et al. (2020) use mainly daily price and liquidity and economic features from 2003 to 2019 to predict index returns up to 30 days ahead using stacking. The applied portfolio of first-layer base-models is vast in models but centered around two major model families, such as decision trees and neural networks, namely random forest (RF), extremely randomized trees, eXtreme Gradient Boosting (XGBoost), and various specifications of neural networks. In the second layer, either linear regression or least-absolute shrinkage and selection operator (LASSO) are applied to combine the results into a weighted return prediction. Across three U.S. stock indices, the stacked models outperform every base-model slightly, while the research highlights a strong performance dependency on meta-model selection, which combines base-model forecasts. Stacking provides minor improvements if compared to base-models. Conclusively, we build upon their learnings and use a more mid-term-oriented feature set. Furthermore, we consider a set of base-models that includes more than two model families and more than one data set specification to increase heterogeneity across base-model results.

As discussed previously, MHP are possible if features are in accordance with fundamental value theory. This theory considers that assuming weak market efficiency, mid- and long-

run excess returns can be forecast using macroeconomic measures or fundamental metrics derived from financial statements. If returns are calculated over several years instead of months, stocks and stock index returns are more likely to be predictable (Campbell & Shiller, 1988). Fama and French (1988) add that negative autocorrelation between stock prices makes 3- to 5-year returns more predictable, with the effect being weaker for large firms, as listed in stock indices. Kyriakou et al. (2020) support this claim by using fundamental features to compare the predictive power for one-year and five-year returns in a non-parametric setting. Because the latter target is less volatile and the method relies on features that explain the stock market in the mid-term, the five-year forecasts entail a stronger and more significant explanatory power. Radha (2020) adds a more fundamental view on return forecasts and filled a monumental research gap: due to his focus on MHP instead of LHP for more than 10 year ahead and countries aside the U.S., Radha identifies that mid-term country yield (CY-M) can be forecasted accurately. The model only considers the cyclically adjusted price-to-earnings ratio (CAPE), real-adjusted exchange rate, and five-year price return momentum, while explaining most of the countries return variance despite its linear regression setting. Another contribution of this research is the transformation of CY-M on a CAGR basis. CAGR makes the comparison of assets throughout time easier and facilitates the asset selection in a given strategy. Additionally, monthly CAGR values give the equity-return expectation for the selected period in advance, which is an important indicator for long-term investors. Hence, we consider Radha (2020) and Kyriakou et al. (2020) choices in feature selection and target transformation.

Wang et al. (2019) use linear time-series and machine learning models in a combined approach with fundamental features to predict 10-year ahead returns. In terms of model choice, features, and target this paper is the closest to our methodology. The authors confirmed that using fundamental features to predict 10-year returns directly is likely to perform worse than the historical mean, regardless of the model. To combat this issue the paper proposes a two-step approach in which the inverse CAPE is predicted first using a ML-VAR hybrid. The features and their lags are predicted using methods, such as VAR, GBM, RF, GRU, Support Vector Regression, and simple ensembles of these models. In essence, the authors replace the linear autoregressive core of the VAR with aforementioned machine learning models to enhance their forecasts for the inverse CAPE by allowing for

non-linear relationships. The eventual return is calculated by taking the respective return parts into account, namely the valuation expansion, earnings growth, and dividend yields. In sum, this approach reduces forecast errors by 50% from 1960 onwards if the best ensemble model is compared to benchmarks and machine learning models. In detail, the RMSE of every machine learning based model decreased by approximately 50%, to a level between 0.038 and 0.026. Nevertheless, the authors only predict the returns of the S&P 500 instead of analyzing this increase in accuracy using multiple international indices, mainly due to shorter time-series availability for other countries. In sum, these results show us that MHP based on fundamental and macroeconomic features may lead to a decrease in forecast errors. Yet, it remains unanswered, if stacking can lead to improvements in accuracy and resilience for MHP and whether the results yield any profits.

With Wang et al. (2019) and Radha (2020) we include two research papers that are associated with research in private asset allocation companies. On the one hand, this connection allows us to incorporate methods that are aligned with the status quo of financial machine learning research in practice. On the other hand, we assume a bias between the performance and disclosure of publicly available return prediction models and unpublished models used in the investment industry. If exploration leads to more efficient markets, it is desirable to keep better methods private (Leung et al., 2000), because it directly translates into capital gains. Considering previous research conducted in the field, our motivation and contribution to the return forecasting and asset management literature is three-fold:

1. Primarily, we use stacking for mid-horizon predictions of future returns. For this we apply heterogenous base-models and data set variations to determine if feature-target relationships are predicted more accurately using more training data.
2. Secondly, we seek to understand to what degree statistical accuracy affects financial portfolio performance by designing a rebalanced long-only portfolio. Investments are based on the predicted CAGR of the individual stock indices.
3. Thirdly, we design our predictive models to stay interpretable. This allows us to visualize the feature effects to shed light into the black box of our stacked models.

5. Data

Our scope in countries, features, and data history is an important pillar for stacking. Firstly, we validate our results across a wider range of indices than reviewed publications, such as Wang et. al (2019) and Jiang et al. (2020) that refer to U.S. indices, such as the S&P 500 or the Dow 30, Nasdaq, and S&P 500, respectively. Secondly, the set-up of a cross-sectional time-series dataset allows us to use data pooling. The eventual data set includes multiple features across 41 countries and for a maximum of 70 years. Thirdly, our data segmentation and pre-processing process makes our approach feasible, protects against data leakage, and corrects for the data overlapping that is introduced by our target.

5.1 Target and Features

We gather monthly data for 57 different stock indices from 1 January 1950 until 1 August 2020, with 14,239 monthly observations in total. We use MSCI as a source to obtain price data of the main stock index for different countries, including markets classified as developing and emerging markets. The gathered data is in local currency with dividends included. We use local currency and do not convert to USD to not introduce further variance that would weaken the signal. The target is calculated using the price data as follows:

$$CAGR_t = \left(\frac{p_{t+60}}{p_t}\right)^{\frac{1}{5}}$$

where p is the price of the stock index and t the current time period. The monthly CAGR values can be interpreted as the predicted 5-year trend of compounded equity-returns (Radha, 2020). This means whether returns are rising, declining, stagnating, and to what extent, while defining a measure that facilitates the comparison among assets.

5.1.1 Fundamental Features

In addition to the return data, we use features supported by research with a superior performance in explaining future returns. Firstly, CAPE is the main fundamental feature we use to measure valuation expansion and is obtained from Barclays. It is calculated by dividing

the current price with an inflation-adjusted 10-year average of past earnings to smooth out business cycles. CAPE is a reliable and strong feature, due to smoothing the variance of the earnings over multiple business cycles (Umlauft, 2020). Despite being a feature that summarizes past information, its predictive power is based on the assumption that long-term market valuations revert back to their mean. Therefore, after a cycle of overvaluation it is necessary for an index to decrease in valuation and to eventually revert back to its historical mean (Poterba & Summers, 1987). This explains why CAPE is reliable and averages a correlation of up to 70% for future returns of five years (Radha, 2020). We confirm these findings partly, with a correlation analysis among CAGR leads and CAPE, as seen in Figure A1. Secondly, we use the dividend yield of a stock index, because this feature has been shown to explain up to 30% of future multiyear stock returns, leading to a stronger explanation than the lags of returns (Campbell & Shiller, 1988), because the present price represents the present value of future dividends.

5.1.2 Macroeconomic Features

Mid-term predictions of returns are influenced by macroeconomic conditions, as shown repeatedly in factor models from Fama and French, Campbell and "other notable researchers" (Leung et al., 2000). We obtain the macroeconomic features from OECD. Firstly, the unemployment rate has been shown to be negatively correlated with the stock market, since companies depend on the macroeconomic environment with the unemployment rate being one of the most important measures. The reason is that current prices consider future conditions of the economy, to which the labor market is a key contributor. Therefore, we expect future return expectations to be high if the unemployment rate is expected to peak in the close future, unless this is captured by other features such as those related to valuation. Secondly, we include the 10-year government bond yield as it is often used as the risk-free rate for longer investment horizons (Damodaran, 2008). If added with the equity risk premia, which is unknown in advance, the risk-free rate results in the expected return of stocks. Thirdly, we include the 3-month government bond yield as indicator for the current macroeconomic situation. Lastly, we include the consumer price index (CPI) as a feature, since inflation directly affects the discount rate of stocks with the expectation of increasing inflation affecting stocks negatively.

5.2 Data Preparation

Initially, we calculate our target from actual stock index prices and express it as CAGR to reduce variance and facilitate the comparability asset returns for mid-horizon predictions, as already used by Radha (2020) and Wang et al. (2019). Adversely, this transformation introduces a higher degree of autocorrelation, especially if the target were to be used as a lagged feature. Model specific transformations are explained in Table B1.

Thereafter, we split the data into training, leakage, and test sets, with each of the countries having the same dates in the test set and leakage sets in order to avoid any possible leakage caused by correlation between the stock markets. Leakage may be apparent if past observations in the training set are correlated with observations in the test set, meaning that the test set is not a reliable representation of out-of-sample data and would lead to artificially lower forecast errors (López de Prado, 2018, p. 103). Only our training set differs in length due to the different dates when the data became available for the different countries. Furthermore, we add a leakage set of five years between the training and test sets to avoid leakage. Otherwise, we would partially defeat the purpose of the training-test separation and our model would be tested on previously seen data because of the overlapping nature of the 5-year CAGR. To test our results on a longer financial market period, we extend the test set period to ten years, as illustrated in Figure 2. In case of any leakage issues, this expansion may show any defects in model performance after five years. Our test set ranges until 2015, with the last prediction being the CAGR from August 2015 until August 2020.

Lastly, we exclude countries from our single-models if they do not have enough observations for the features. This greatly reduces our effective set of countries, because the computation of the CAGR and the introduction of a leakage set reduces the available data by 15 years. If the CAPE, dividend-yields, 10-year government bond yields are incomplete, we exclude the MSCI index, as seen in Table A1. The seven remaining countries, for which CAGR is predicted, are Australia, Canada, Germany, Netherlands, Switzerland, United States, and United Kingdom. We exclude countries only in the single-country set, because they may lead to valuable explanations in our pooled model set-up. In other words, the countries with not enough data are used only in the training of the pooled models.

6. Methodology

Our framework applies four base-learners and improves the stacking model by Jiang et al. (2020) by using diverse learners and datasets. Our combination of linear time-series, regularized regression, and ensemble regression-tree models was not applied to MHP before. To analyze financial performance, we employ a rebalanced long-only strategy to build a portfolio based on the predicted CAGR. To balance the complexity and to understand the drivers behind predictive performance, we implement accumulated local effects for base- and stacked models to analyze the feature effects among models. Remarks on replicability of our research can be found in Appendix E.

6.1 Stacking

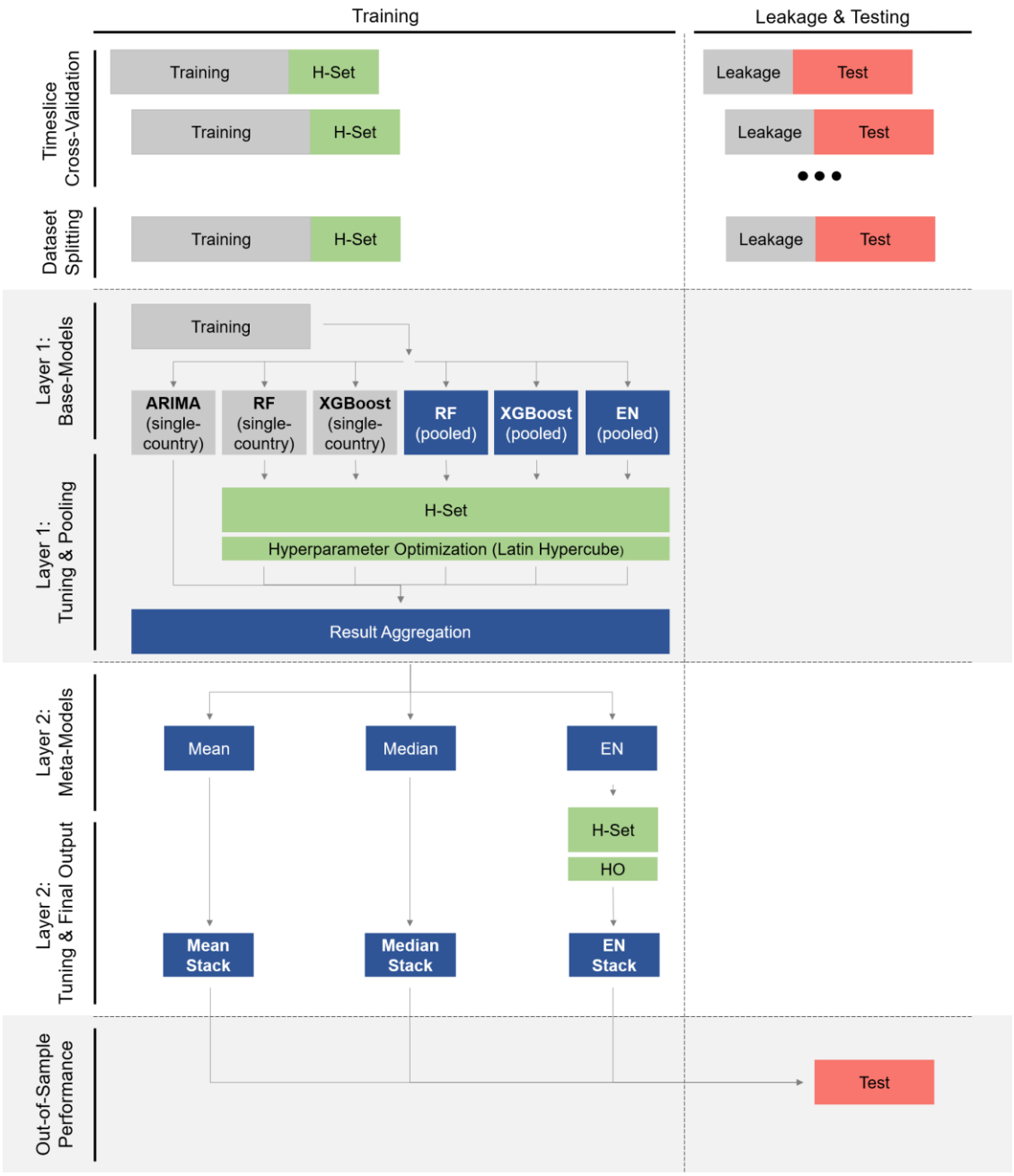
Allende & Valle (2017) describe the two benefits of stacking as combining strengths and weaknesses of multiple models in order to create more robust predictions. Firstly, it increases the chance of opting for a good base-model across an array of different base-models. Those are compared in-sample by the corresponding meta-model. Secondly, by weighting the different base-models we can increase forecast accuracy due to prediction aggregation. This follows the same principle as known in homogenous ensemble models, only that heterogenous models deal better with noise (Ribeiro & dos Santos Coelho, 2020), because of their different forecasts and loss minimization functions. Those two factors lead to reduced errors out-of-sample. In order to reduce these error the base-models must be uncorrelated in their forecast errors, so that the base learners can extract the most information (Jiang et al., 2020), while still coping with the nature of financial time-series data. Our model training and evaluation approach is shown below in Figure 2.

6.2 Data Pool Variety

Pooled models allow us to use all available data to create accurate feature-target relationships, as opposed to single-country models that only use country data to estimate own feature-target relationships. Pooling combines all available data into one basket, fitting one

common model for all countries inside the data basket. Using one-country indicators as dummy variables allows for estimating future predictions for single-countries based only

Figure 1: Stacked model training and evaluation process



on its country-specific test set features, despite fitting one model only. Using stacking together with pooling allows us to test if the additional observations lead to more accurate predictions by decreasing the adverse effect of fundamental data availability. The detailed process of pooling and model training is illustrated in Figure C1.

The advantages are that the relationships between features and the target can be learnt using much more data than in a single-country setting, which is beneficial in a financial context where spurious correlations might last for decades. Secondly, access to more data is key to models using fundamental and macroeconomic features, because their monthly availability and our data segmentation and test set extension limit the chance to understand those relationships. Thirdly, from a finance point-of view, pooling is beneficial because it introduces countries as subgroups of the complete group, the latter representing the global financial market for stock indices. This flexibility is needed because it was proven many times that financial markets may influence each other, but the relationships are expected to change and were not explored yet for mid-horizon CAGR predictions. Fourthly, pooling can help in dealing with overfitting, because one model is fit for all countries. Because of a greater access to data, the resulting relationships are representative across most of stock indices, due to the nature of error minimization. A downside is that hyperparameters need to be optimized on a larger sample because pooled models are more prone to overfitting since we introduce 41 dummy variables. This increases the already high computation time due to the bigger training data size as compared to the single-country models.

Only machine learning models can employ hyperparameter optimization to deal with additional complexity. We use pooling with RF, XGBoost, and Elastic Net, while using ARIMA, RF, and XGBoost in the single-country setting, totaling the number of available models for stacking to six base-learners.

6.3. Heterogenous Base-Learner

Base-models define the first layer of our approach. We choose ARIMA, Elastic Net, RF, and XGBoost as base-models. This mix combines different flexibility and regularization levels in order to balance the bias-variance tradeoff. The heterogeneity of models is

described by using different data sets (pooling vs. single-country), the application of lags in (only in time-series regression), and loss-minimization principles. All the employed machine learning models use Latin Hypercube for hyperparameter optimization, so that values for hyperparameters appear once in the whole grid and are evenly spread across the parameter range, increasing computational speed while decreasing overfitting tendencies (Urban & Fricker, 2010). The chosen hyperparameters are listed in Table C2.

6.3.1 ARIMA with exogenous features

Autoregressive Integrated Moving Average (ARIMA) models combine three parts: The autoregressive (AR) part defines the relationship of the target as being dependent on its past, so-called lagged, observations and an error term. Integrated (I) transforms a time-series into a stationary process using differencing. The Moving Average (MA) defines the target as being a linear combination of the white noise process to capture short-memory. ARIMAX adds features and assumes that information not contained in the regression is fully described by η_t, with residual Gaussian errors, leading to a multivariate model:

$$Y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \eta_t$$

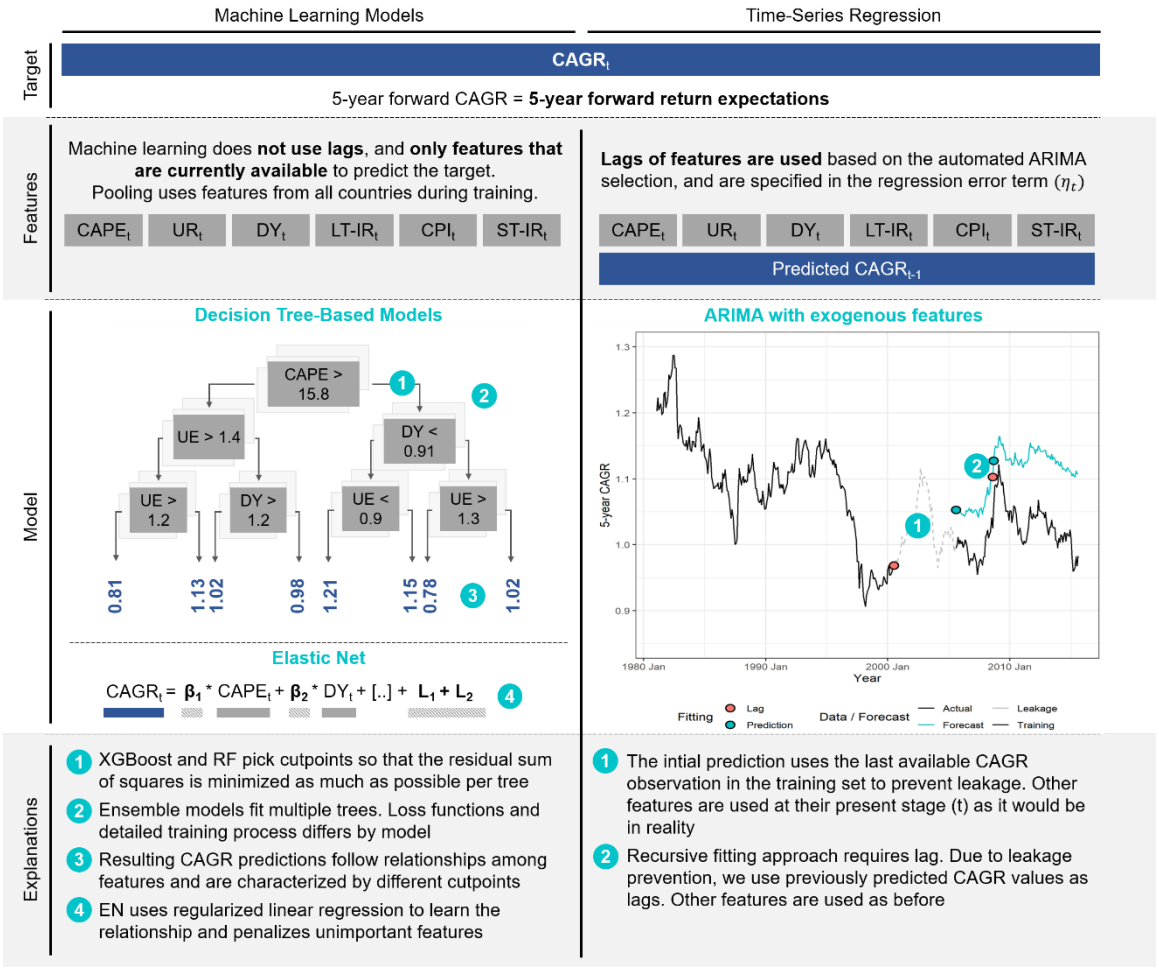
$$\eta_t = \phi_1 \eta_{t-1} + \cdots + \phi_p n_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

ε_t is a Gaussian white noise series with a mean of zero. The variance of ε_t is assumed to be constant, since otherwise the model suffers from heteroscedasticity. If a time-series were to be non-stationary or to suffer from heteroscedasticity, this would lead to unstable coefficients due to rising variance and low out-of-sample performance. We apply the Hyndman-Khandakar algorithm to ensure stationarity, significant coefficients, and to find an optimal in-sample AICc minimizing model (Hyndman & Khandakar, 2008). Because of stationarity among the features and the target, we force the algorithm to use $d \geq 1$, as implied by the Augmented-Dickey-Fuller and Kwiatkowski–Phillips–Schmidt–Shin tests. The fitted ARIMA models and its coefficients are listed in Table C1.

In order to account for data leakage, we use a recursive approach without re-estimation, meaning that the model is not re-estimated throughout the prediction and validation process. Therefore, we use predicted CAGR as lags for the present CAGR prediction. We

refrain from using any actual CAGR lags, because this would lead to data leakage, as visualized in Figure 2.

Figure 2: Training process for used base-models



6.3.2 Elastic Net Model

Elastic Net (EN) is a regularization method and pioneered by Zou & Hastie (2005). EN adds two parameters to the OLS regression to account for variable selection and parameter shrinkage in order to improve its predictive capabilities for high-dimensional feature spaces. The L_2-norm ($|\beta|^2$), also used in Ridge regression, penalizes the size of parameter estimates to shrink them towards zero while decreasing the model complexity without fully removing parameters. The L_1-norm ($|\beta|_1$), also used in LASSO, introduces continuous shrinkage and feature selection but suffers from lower performance compared to Ridge regression if features are correlated (Zou & Hastie, 2005). The α hyperparameter is used

to balance between the L_1 and L_2 norm, respectively, with $\alpha = 1$ representing a Ridge regression. The elastic net optimization for a feature matrix X, is written as:

$$\hat{\beta} = arg\,\min_{\beta}|y - X\beta|^2, \quad s.t. \;\; (1-\alpha)|\beta|_1 + \alpha|\beta|^2$$

$$where \quad |\beta|_1 = \sum_{j=1}^{p}|\beta_j|, \quad |\beta|^2 = \sum_{j=1}^{p}\beta_j^2$$

6.3.3 Random Forest

Random Forest is a machine learning approach pioneered by Breiman (2001). RF builds upon regression trees, while decreasing their tendencies of overfit. Normal decision trees use a greedy portioning algorithm known as recursive binary splitting to classify the feature space in partitions (Yuan et al., 2020), which are local optimization problems and making simple trees prone to overfitting, tendencies that can be exaggerated on financial data. The objective of decision trees is to find X_j and s, so that the residual sum of squares (RSS) is minimized by the greatest amount possible, considering all choices in features and cut-points (James et al., 2013, p. 306). The observations within a space R_j are represented by a leaf and its prediction is the mean of the i observations within this space.

$$\text{Obj} = \sum_{j=1}^{J}\sum_{i\in R_j}\left(y_i - \hat{y}_{R_j}\right)^2$$

$$R_1(j,s) = \{X|X_1 < s\} \quad \text{and} \quad R_2(j,s) = \{X|X_1 \geq s\}$$

RF creates multiple random training samples from the training using bootstrapping with replacement, to achieve a similar distribution across the bagged samples (Weng, 2017, p. 43). A possible downside of such resampling is the strong similarity between the trees, which leads to low diversity and therefore less robust performance (Ribeiro & dos Santos Coelho, 2020). RF circumvents this problem by introducing randomness in features by restricting the feature space *J*, to decorrelate the individual trees (López de Prado, 2018, p. 98). The ensemble model result for $\hat{y}_{R_j}$ is based on the average of all sub models.

RF are advantageous because they create forecasts while using a non-additive and two-step randomization, leading to reduced generalization errors (Breiman, 2001). Due to the relatively low number of hyperparameters, RF are easy to compute. The *number of trees* specify the sum of single trees in the forest. *Maximum depth* specifies the depth of a single tree, with smaller depths controlling for overfitting. *Number of random features* randomizes the feature space, with lower samples controlling for overfitting by decorrelating the features and trees, leading to a lower variance.

6.3.4 Extreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting uses sequential boosting to reduce the generalization error of decision trees. It applies a gradient descent approach to find an additive model that minimizes the error. The novelty of XGBoost is the change of the loss function to a second-order Taylor polynomial. This leads to a faster training process, because new trees are only considered if they lead to an explanatory gain (Ribeiro & dos Santos Coelho, 2020). Future models are fitted on the residuals of previous trees, a process which reduces the problem of underfitting and bias (Cervelló-Royo & Guijarro, 2020). Observations that lead to large errors during training receive a high weight so that consecutive models reduce errors associated with the difficult to fit observations. This continues until the specified maximum number of trees is reached.

This may lead to overfitting since weights are selected for individual observations. Therefore, XGBoost adds a new regularization term shrinks the added weights to address overfitting. Also, feature subsampling is used to decorrelate trees by reducing the feature space (T. Chen & Guestrin, 2016), to address overfitting. XGBoost uses K additive trees to predict the target. The weights ω are assigned in the training process for every tree in f_k, leading to a continuous score for every prediction. This score is summed up across all trees in the ensemble and is added to the final prediction $\hat{Y}_i$ (T. Chen & Guestrin, 2016).

$$\mathcal{L}(\phi) = \sum_{i}^{N} l(\hat{Y}_i, Y_i) + \sum_{k}^{K} \Omega(f_k)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2$$

$\mathcal{L}$ is the objective function, consisting of a differentiable loss function l, which measures the distance between the prediction and actual target. The penalization term Ω decreases weak-learner complexity of single trees. Hyperparameter γ adds a threshold for splitting a tree, based on how the resulting tree minimizes the loss (Jiang et al., 2020). λ is shrinkage term for the newly added weights. Further hyperparameters include the *learning rate*, which scales newly added weights ω by factor η. *Maximum number* of trees specifies the number of trees that are fitted sequentially and controls for overfitting. *Maximum depth* specifies the depth of a single tree, with smaller depths controlling for overfitting. *Number of random features* randomizes the feature space, with lower samples controlling for overfitting by decorrelating the features.

6.4 Explanation for Base-Learner Choice

ARIMA models are often used in financial forecasts (Feuerriegel & Gordon, 2018) despite their assumption of a linear relationship between targets and features, which is not optimal for relationships observed in financial time-series data. These models are suited for our stacking approach since they can capture target dependencies over time-spans (Bukhari et al., 2020) and apply a different loss minimization principle from our tree-based models, which are based on maximum-likelihood.

We opt for Elastic Net, because it combines the LASSO and Ridge penalties and is proven to usually dominate individual LASSO, Ridge or OLS models. The reason is the convex combination of the penalty terms resulting in a unique optimization minimum. This means that the model profits form computational speed similar to that of a LASSO model, while not suffering from low performance, if the features are correlated (Zou & Hastie, 2005). These factors are essential, if considering the long computational time for pooled models.

RFs application in hybrid long-term return predictions (Cervelló-Royo & Guijarro, 2020; H. Wang et al., 2019) makes them an important part of our stacking architecture. Financial data often leads to overfitting due to the relatively low signal-to-noise ratio (López de Prado, 2018, p. 99). Additionally, the high momentum on time-series makes bagging-based models important because these reduce the variance of the predictions.

XGBoost addresses underfitting and follows a different loss minimization. Consequentially, RF and XGBoost models explain feature-target relationships differently (Jiang et al., 2020). Considering the set of machine learning models, we stabilize predictions by using methods that extrapolate strongly (EN) and minimally (XGBoost, RF), leading to robust predictions if the range of features extends outside of the range learned during training.

6.5 Meta-Learner

Our approach uses a two-layer training approach to combine the outputs from the base-models to a final robust prediction. For this we use Mean, Median and Elastic Net as meta-models. A meta-model combines the actual forecasts of layer one and treats them as an input to predict the eventual target (Ribeiro & dos Santos Coelho, 2020). The meta-model is trained on the same training data as the base-models and determines error minimizing weights for models in layer one, as seen before in Figure 1.

The choice of an appropriate meta-model is crucial for the out-of-sample performance of the aggregated stacked model. Complicated meta-models do not necessarily translate to improved accuracy if compared with more simple and regularization-based models, such as Ridge Regression, LASSO, or Elastic Net (Jiang et al., 2020). Often easier methods, such as averages or trimmed means are preferred, due to the noisy nature of financial time-series data (Allende & Valle, 2017, p. 224). Additionally, more flexible meta-models might overestimate the predictive contribution of flexible base-models, because of their strong in-sample fit. If these base-models show better in-sample performance than simpler, possibly less accurate models, these flexible base-models are favored during the training process. Consequentially, this may impact out-of-sample performance because feature-target relationships are subject to fluctuate in financial time-series data.

Due to the absence of clear guidelines or evaluation principles apart from accuracy, we validate multiple meta-models. This list is constructed in accordance with previous applications, versatility, and weighting characteristics, as outlined by Allende and Valle (2017).

6.6 Evaluation Procedure

Prediction performance measures are necessary to understand the improvement against benchmarks and base-models. In contrast, these improvements do not necessarily translate into profits and must be considered together with risk (Yuan et al., 2020), because investors show different affinities for risk. Hypothetically, if we consider a strategy with comparatively low returns and low risk, leverage can be used to generate higher returns while adjusting risk exposure to be similar to other portfolios (Guasoni & Mayerhofer, 2019).

Statistical measures facilitate the comparison of models and their suitability across markets and their phases. Our selection includes Mean Absolute Error (MAE), Root-Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), because those are often used in forecasting evaluation and return prediction research (Lyhagen et al., 2015; H. Wang et al., 2019). Generally, the lower the scores, the better the out-of-sample accuracy. We include MAE as a measure of absolute bias to specify whether our forecasts have any problems with approximating the data in- and out-of-sample. RMSE specifically penalizes high errors. MAPE is a popular measure because it is scale invariant (Lyhagen et al., 2015), unit free, and presented as percentage. Primarily, we use the MAPE and list RMSE and MAE in our Appendix (see Table D1 and D2). Additionally, we apply the Diebold-Mariano test, because of its flexibility in treating forecast errors (Diebold & Mariano, 1995), to show whether our forecasts are significantly better than the benchmark.

Furthermore, we implement a backtesting strategy using a rebalanced long-only portfolio, due to its more common occurrence than a buy-and-hold strategy (Hou et al., 2020). Based on the CAGR predictions, we classify the MSCI indices into longing (highest predicted CAGRs) and zero (lowest predicted CAGRs). We perform the rebalancing by making the investment in each of the indices equal at the beginning of a new rebalancing period. At the beginning of a rebalancing period, we invest into three stock indices that are predicted to outperform the rest, with no investments for the remaining indices. We benchmark the performance against an equal-weight strategy across all seven stock indices to test the financial effect of setting asset weights based on forecasts. The comparison to the mean and naïve benchmarks is secondary, because those strategies pre-select assets and our primary

interest is in testing the asset pre-selection based on our method. We allow the model to rebalance either once, twice or not at all, meaning that the index selection can vary.

Because we pick the three best-performing indices, our strategy can be assumed to be riskier than the equal-weight strategy. Therefore, we add the Sharpe Ratio, which is often used for portfolio choices (Bao, 2009). It is based on the risk-free rate of return, as calculated by subtracting the risk-free rate from the actual index return divided by the standard deviation. We use the U.S. 10-Year Treasury Bond Rate as the risk-free rate due to our mid-term focus and its higher yield rates compared to treasury bills of shorter maturities. Additionally, we employ the equality test methodology described by Wright et al. (2014) to statistically validate our Sharpe Ratio results. This method uses a more general assumption to cope with non-IID returns. It estimates the covariance matrix of returns by correcting for heteroscedasticity and autocorrelation, which is important because in the non-IID scenario the variance of present returns may also depend on lagged returns (Lo, 2002).

6.7 Interpretation of Base- and Stacked Models

We use Accumulated Local Effects (ALE) to interpret feature importance across all models. ALE considers the conditional distribution among the features to circumvent the inclusion of unlikely feature combinations (Molnar, 2021). Also, ALE applies differencing to extract the pure effect of a feature without mixing its effect with correlated features. These single, local affects are then aggregated to facilitate interpretation (Apley & Zhu, 2020).

The benefits of using ALE are three-fold. Firstly, we select a method that is accepted as research tool for interpreting feature results of machine learning models. Secondly, we opt for ALE because of its capability to deal with correlated data, which Feature Importance, Partial Dependence, and Individual Conditional Expectation cannot deal with as effectively (Molnar, 2021) and may lead to biased feature effect metrics. Thirdly, we opt for a model-agnostic measure to compare ALE scores among our stacked models. This requires high flexibility, so that we retain the feature effects in the second layer, which uses the base-model predictions as features.

To calculate the isolated effect of a feature (x_i) on our target (y), we divide the feature space into intervals. For predictions within an interval, ALE computes the prediction differences if we replace the actual feature observation with the upper and lower interval limit, while keeping all other feature effects equal. The feature effects are computed within an interval, so that similarity on conditional distribution of x_i is satisfied and accumulated along all the intervals. Specifically, we apply ALE on our stacked models, using the training set data as input to get the isolated effect of each feature on the target as predicted by the combination of base-models. We exclude the country-specific dummy variables from examination as they are less relevant and only used in the pooled models and extra parameters that are estimated during training of the ARIMA model. We consider only indices that are used for the predictions. Technically, we use a total of 100 intervals for each feature which directly affects the complexity of the illustrated effects.

7. Results

Our stacked models show that heterogeneity in base-models leads to more resilience. Our stacked models beat the benchmark by up to 60.5% across the indices, while remaining statistically significant. Based on ALE, we confirm a strong effect for fundamental features and the 10-year bond-yield. Our predictive gains translate into above-benchmark yearly CAGRs and Sharpe Ratios, which remain statistically significant only at the 10% level for the three best performing portfolios, including the median-stacked model.

7.1 Predictive Performance of Base-Models

As shown in Figure 3, the prediction performance of the applied models varies by country and by model flexibility. In general, we observe a close fit across all models with strong differences for model performance for Germany, Netherlands, and USA. These three countries also show the lowest significance ratings across all models. For the Netherlands and Germany, ARIMA and some single-country base-models show strong deviations from the actual values. The pooled models show a robust and accurate fit across all countries, while ARIMA tends to over- or underestimate CAGR often, as seen in Table 1.

Table 1 shows the MAPE reduction per model-country combination, compared to the historical mean benchmark. We use the median since the accuracy scores are very right-tailed, and the median is more outlier-robust. On average, the pooled models led to a MAPE reduction of 36.4% to 58.1%. The single-country XGBoost outperforms its pooled counterpart and leads to a median MAPE reduction of 52.3%, compared to 36.4% for the pooled model. Whereas RF displays lower median MAPE than its pooled counterpart, beating the benchmark by 42.5%, compared to 54.0%. The relative underperformance of ARIMA shows that using previously predicted CAGR lags with a Dynamic Regression set-up leads to no explanatory gains for future returns compared to other base-models. ARIMA shows 5.3% higher median MAPE than the benchmark. ARIMAs underperformance and pooled RF resilient performance across all indices highlight the difficulty of beating the standard benchmark across multiple stock indices. The outperformance of pooled RF remains if we consider RMSE or MAE to test accuracy. According to those metrics, the base-models beat

the historic mean more often. Also, the ARIMA results based on those metrics are slightly less inferior, as shown in Table D1 and D2. Referring to the Diebold-Mariano test, we classify most of the base-model forecasts as significantly better than the benchmark.

A correlation analysis, as seen in Table D4, supports the previous findings by highlighting that more complex and pooled models also predict the direction of the forward return expectations more accurately. Pooled XGBoost and RF achieve the highest correlation scores, but also lead to the highest fluctuation, with EN being less correlated despite its similar accuracy. Across all countries, all single-country models show low correlation, showing their missing ability to anticipate future return directions.

Table 1: MAPE reduction, compared to historical mean model

	Base						Meta		
	Pooled			Single-Country			Stacking		
Country	XGB_pool	RF_pool	EN_pool	ARIMA	XGB_s	RF_s	EN	Mean	Median
AUS	62.5%***	80.5%***	62%***	-7.4%	62.6%***	71.4%***	76.3%***	71.7%***	72.8%***
CAN	21.1%**	61.9%***	50.2%**	-201.1%	30.9%***	55.5%***	36.5%***	17.1%**	46.6%***
DEU	27.8%*	45%*	39.6%**	-5.3%	52.3%**	19.7%	44.1%**	41.9%**	51.2%**
NLD	10.1%*	7.4%	-66.0%	-490.7%	-249.0%	-127.0%	-94.0%	-145.4%	-106.4%
CHE	49.2%**	54.9%***	68.1%**	68.5%**	56.6%**	63.5%***	70.4%***	72.6%***	69.7%***
GBR	74.6%***	38%**	74.9%***	69.1%***	84.3%***	58.7%***	84.9%***	86.1%***	85.4%***
USA	-88.6%	23.4%**	-78.8%	38.4%**	0.7%*	-34.2%	6.5%*	10.3%**	24.2%**
MEDIAN[1)]	**36.4%**	**54.0%**	**58.1%**	**-5.3%**	**52.3%**	**42.5%**	**57.4%**	**59.1%**	**60.5%**

Error reduction is calculated by: 1 - ($MAPE_{model}$ / $MAPE_{mean}$).
[1)] The median calculation computes first the median $MAPE_{model}$ and $MAPE_{mean}$

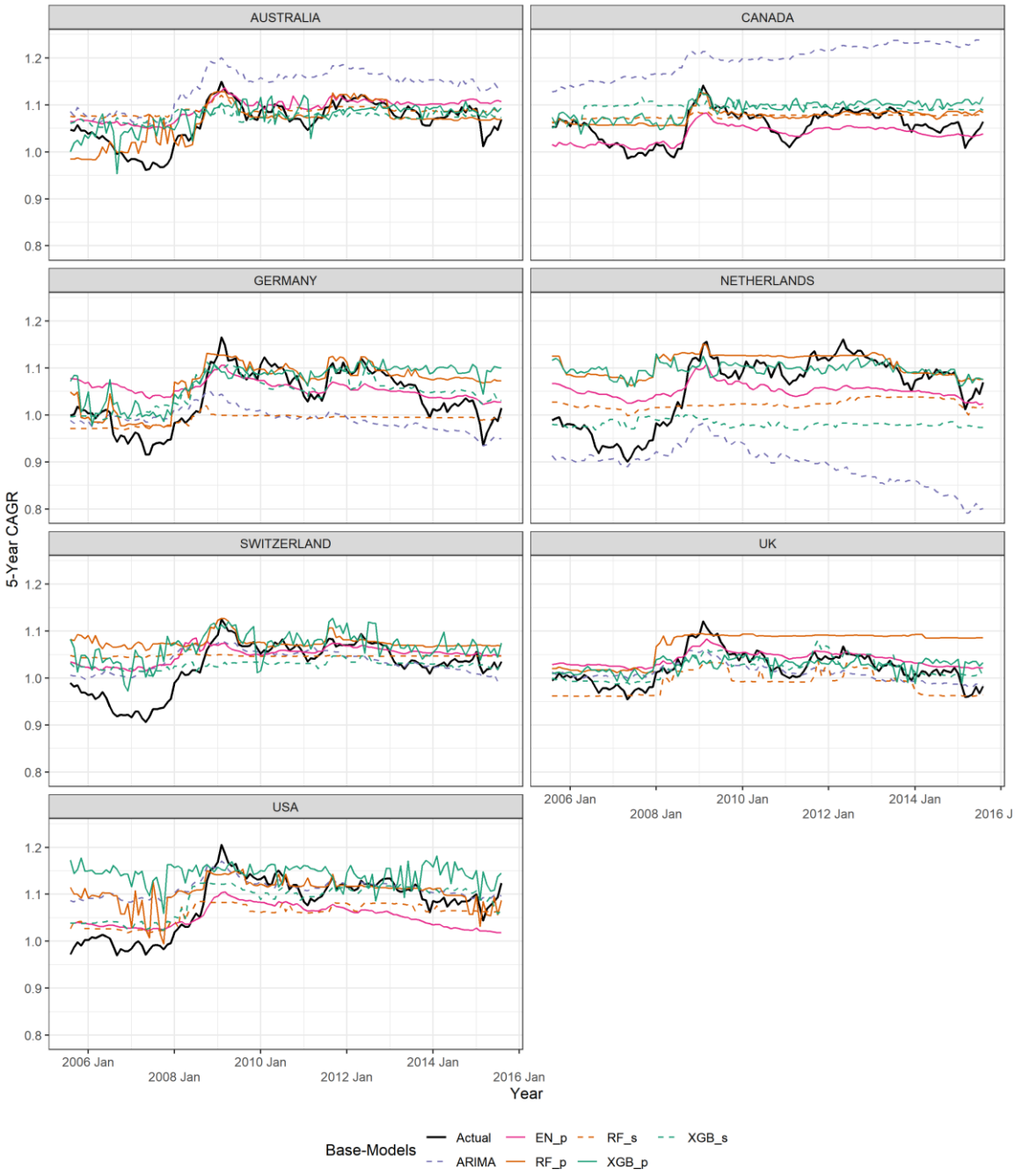
Significance values refer to Diebold-Mariano test (1995), whether the forecast errors are significantly different from the historical mean benchmark: Significance level at 90% (*), 95% (**), 99% (***)

A country level analysis shows a balanced picture: Pooled XGBoost performs the best in the Netherlands, whereas pooled RF leads to best predictive performance for Australia and Canada. Single-country XGBoost shows the strongest performance for Germany and UK. Despite the low forecast accuracy across all countries, ARIMA leads to superior results in Switzerland and USA, but underperforms for four out of seven indices. These differences in performance emphasize the need for heterogenous models, due to the dependence of out-of-sample performance on the target, country, prediction interval, and chosen data set.

Upon closer investigation of CAGR trends in Figure 3 below, the biases of the predictions become visible. Firstly, if actual five-year forward return growth expectations are below 1, meaning that the future returns are not expected to increase, the employed models are not

able to approximate this trend fully, leading to overestimates in the period from 2006 to 2009 and lower accuracy. Secondly, if the actual CAGR values range across a wide interval, as seen in Germany, Netherlands and USA, the models show lower prediction performance, because they cannot fit the strong fluctuations. In contrast, bear markets are well fitted with most of the machine learning base-models, as seen in 2009 and 2012.

Figure 3: CAGR prediction across base-models and stock indices



7.2 Predictive Performance of Stacking

Differences in predictive performance diminish once the base-models are used in a stacking framework. While results become more robust across the indices by reducing the generalization error, the results do not indicate overly reduced MAPE. This is also confirmed by higher statistical significance for mean- and median-stacked models, as seen in Table 1. Elastic Net stacking (0.024) and mean-stacking (0.023) show average MAPE on par with pooled RF (0.026) and Elastic Net (0.023). However, the median-stacked model shows a slight improvement (0.022), if compared to the most accurate pooled models (0.023). Despite the weighting, stacked models are capable in predicting country-specific bear- and bull-market phases as well as predicting individual long-run CAGR trends. It is important to highlight the ability of stacked models to accurately determine the initial CAGR prediction, despite a prior leakage set and strong variations for the initial prediction among the base-models, especially for Australia, Canada, Germany, and the UK (Figure 4). Higher average correlation compared to base-models with the actual CAGR confirms (Table D4), that directional performance also improved, as can be seen during the COVID pandemic.

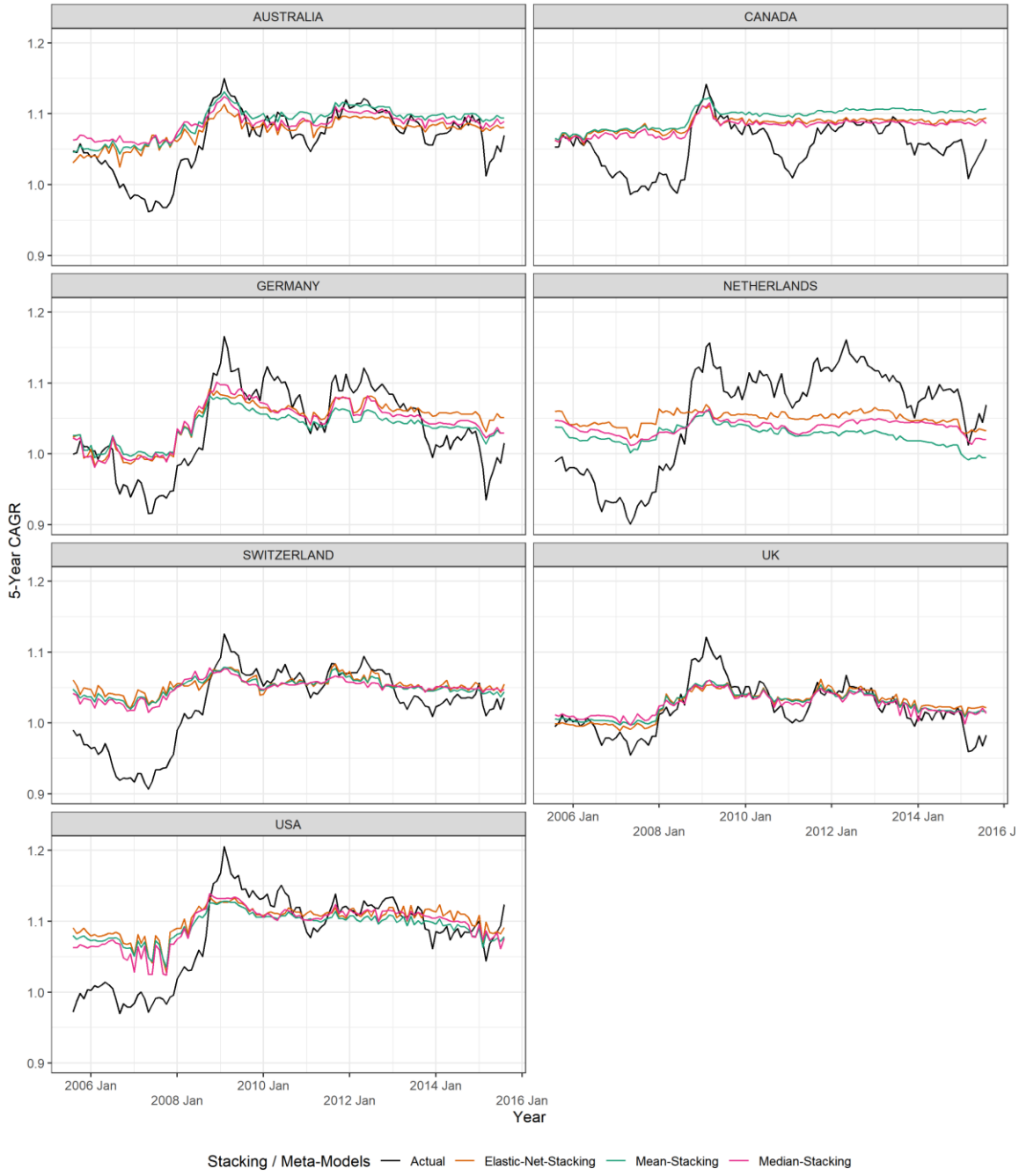
Relative to the benchmark, the stacked models show strong results, especially for Australia, Canada, Germany, Switzerland, and UK. The downside is the lack in predictive outperformance for the Netherlands, and slightly less accurate results for the Germany, and the U.S.

Table 2: MAPE across base-, meta- and benchmark Models

	Base						Meta			Benchmark	
	Pooled			Single-Country			Stacking				
Country	XGB_pool	RF_pool	EN_pool	ARIMA	XGB_s	RF_s	EN	Mean	Median	Mean	Naive
AUS	**0.023**	**0.012**	**0.023**	0.066	**0.023**	**0.018**	**0.015**	**0.017**	**0.017**	0.061	**0.029**
CAN	**0.035**	**0.017**	**0.022**	0.135	**0.031**	**0.020**	**0.028**	**0.037**	**0.024**	0.045	0.051
DEU	**0.040**	**0.031**	**0.034**	0.058	**0.026**	**0.045**	**0.031**	**0.032**	**0.027**	0.056	0.112
NLD	**0.025**	**0.026**	0.046	0.163	0.096	0.063	**0.053**	0.068	0.057	0.028	0.159
CHE	**0.037**	**0.033**	**0.023**	**0.023**	**0.031**	**0.026**	**0.021**	**0.020**	**0.022**	0.072	**0.061**
GBR	**0.020**	**0.048**	**0.019**	**0.024**	**0.012**	**0.032**	**0.012**	**0.011**	**0.011**	0.077	**0.046**
USA	0.048	**0.019**	0.045	**0.016**	**0.025**	0.034	**0.024**	**0.023**	**0.019**	0.025	0.115
MEDIAN	**0.035**	**0.026**	**0.023**	0.058	**0.026**	**0.032**	**0.024**	**0.023**	**0.022**	0.056	0.061

0.048 = MAPE of country-model combination, less accurate as historical benchmark
0.025 = more accurate as historical benchmark, best base-model
0.019 = more accurate as historical benchmark, best stacking model

Figure 4: CAGR prediction across meta-models and stock indices



7.3 Drivers of Model Performance

Data-segmentation and model characteristics lead to the superiority of machine learning models in predicting mid-horizon CAGRs. Data segmentation contributed to a lower than possible out-of-sample performance in two ways. Firstly, a longer test set contributed to

less data being available for model training. Secondly, the effect of the financial crisis of 2007/2008 on the target could not be learned due to the need for a leakage set. Because the feature-target relationships can change over time, significant crises are an important contributor to accuracy resilience and lead to a lower predictive performance especially in the beginning of the test set.

We observe specific characteristics among our models and used data that improve stacking architectures: Firstly, there are strong differences in assumed functional flexibility between the applied base-models. As pointed out before, chaotic and noisy relationships among the features and targets violate requirements that must be fulfilled for good ARIMA performance, even though ARIMA captures linear dependence among features and target. Additionally, there might be an insufficient amount of autocorrelation between predicted CAGR lags, which is introduced by constructing the 5-year forward CAGR. This leads to ARIMA performing worse than other base-models and the standard benchmark as measured by the median MAPE. Additionally, it is argued that because of the need for differencing in linear regression models, valuable short-memory information is discarded (Brooks, 2019), therefore leading to inferior out-of-sample predictions (López de Prado, 2018). Secondly, flexibility is not the sole contribution to accurate 5-year CAGR predictions. XGBoost and RF impose the most flexibility because of their ensemble structure as well as their flexible loss function, while EN also shows robust and high accuracy, which is on par with pooled RF. This means that EN regularization and feature selection set-up is generally better at anticipating the first prediction on the test set and long-run CAGR trend. Its high correlation between predicted and actual CAGR values indicates that it is superior to pooled XGBoost and RF in predicting month-to-month CAGR changes. Thirdly, machine learning models apply feature selection and optimize hyperparameters using Latin Hypercube as a method to make the cross-validation grid in order to reduce the overfitting tendencies, which ARIMA does not. Lastly, pooled models can consider a wider array of feature-target relationships among the 41 countries, which are subject to change based on the nature of financial time-series data. Consequentially, those models identify the relationships between the features and the target more effectively. This is visible by comparing XGBoost and Random Forest in their single-country and pooled setting, the latter contributing to a strong decrease in MAPE. Pooling helps to balance the low data availability for fundamental and

macroeconomic data and contributes to performance gains, as seen by the median MAPE. Conversely, predictive performance for the Netherlands shows that feature-target relationships are not equal among the indices. However, all three pooled models outperformed single-country models for the Netherlands, showing that other financial markets can lead to explanatory power and enhanced predictive performance.

If combined, the heterogeneous strengths and weaknesses among base-models are balanced and lead to increased performance and robustness, whilst also fitting country-specific CAGR trends on the test period accurately. EN-stacking performs on par or slightly worse than the median-stacking, due to two factors. Firstly, the EN is trained on the training set and therefore selects base-models according to their in-sample fit. This leads to a stronger weight towards single-country and pooled models, such as XGBoost and RF, which perform well in-sample but not across all countries which are better covered by ARIMA or EN. Thus, simpler models with a lower performance are shrunken in their contribution or not selected at all, decreasing the overall heterogeneity of EN-stacking, as shown in Table D3. Secondly, the training-leakage-test split shows that the training and test sets differ strongly, an effect which is enhanced by the leakage period, conclusively leading to difficulties in selecting appropriate weights only based on the training and the corresponding validation sets that are formed inside the training set.

These two factors do not influence the performance of the mean- and median-stacked models. Both models use a simple statistic to aggregate the result, which leads to higher out-of-sample performance due to the changing relationships among targets and features between the test and training periods (Allende & Valle, 2017). On top, using the median instead of the mean leads to less sensibility to outliers, as was shown for predictions for Australia, Canada, and Netherlands. However, if the number of underfitting models reaches a threshold of half of the base-models, the aggregated performance across all stacked models deteriorates, as seen for the Netherlands. Conversely, if the threshold is not surpassed, predictions become more accurate and robust, as shown for UK, with a decrease in MAPE by up to 86.1%. These tendencies show why heterogeneity and balancing different model types, such as pooling vs. single-country or linear vs. ensemble machine learning models remain important to reduce the generalization error.

7.4 Financial Performance

As seen in Table 3 and Figure 5, pooled models generate lower performance metrics than the mean and naïve prediction-based portfolios, with pooled RF performing worse than the equal-weight benchmark. These three pooled models also lead to statistically insignificant different Sharpe Ratio results. Meanwhile single-country RF, mean-stacking, and EN-stacking surpass the benchmark once the portfolio is rebalanced more than once but do not lead to statistically significant results. For these cases, outperformance increases with rebalancing intervals, because these models change their asset selection during the rebalancing cycles. Hence, the effect of rebalancing is positive on both metrics because the higher predictive performance of these models led to the selection of indices that performed better than the equal-weight benchmark. However, the statistical significance for the Sharpe Ratio equality increases only for single-country RF until it reaches a significance level at 5%, while also leading to the strongest financial performance metrics across all portfolios.

Table 3: Financial performance across rebalancing windows

		Predictive	Financial					
		Median	Yearly CAGR			Sharpe Ratio[1,2)]		
	Model	MAPE	none	once	twice	none	once	twice
Pooled	XGB_pool	0.035	**5.331**	**5.667**	**4.934**	**0.244**	**0.267**	**0.214**
	RF_pool	0.026	**5.331**	**4.292**	**4.895**	**0.244**	0.171	0.211
	EN_pool	0.023	**5.278**	**5.878**	**5.805**	**0.231**	**0.277**	**0.273**
Single	ARIMA	0.058	**6.915**	**6.915**	**6.915**	**0.362***	**0.362***	**0.362***
	XGB_s	0.026	**6.915**	**6.915**	**6.915**	**0.362***	**0.362***	**0.362***
	RF_s	0.032	**5.503**	**6.692**	**7.148**	**0.274**	**0.357***	**0.387****
Stacked	EN	0.024	**5.686**	**6.106**	**6.183**	**0.283**	**0.309**	**0.313**
	Mean	0.023	**5.686**	**6.106**	**6.183**	**0.283**	**0.309**	**0.313**
	Median	0.022	**6.915**	**6.915**	**6.915**	**0.362***	**0.362***	**0.362***
Trad.	Mean	0.056	**5.912**	**5.912**	**5.912**	**0.297**	**0.297**	**0.297**
	Naive	0.061	**5.503**	**5.503**	**5.503**	**0.274**	**0.274**	**0.274**
	Benchmark (Equal-Weight)		4.818	4.818	4.818	0.211	0.211	0.211

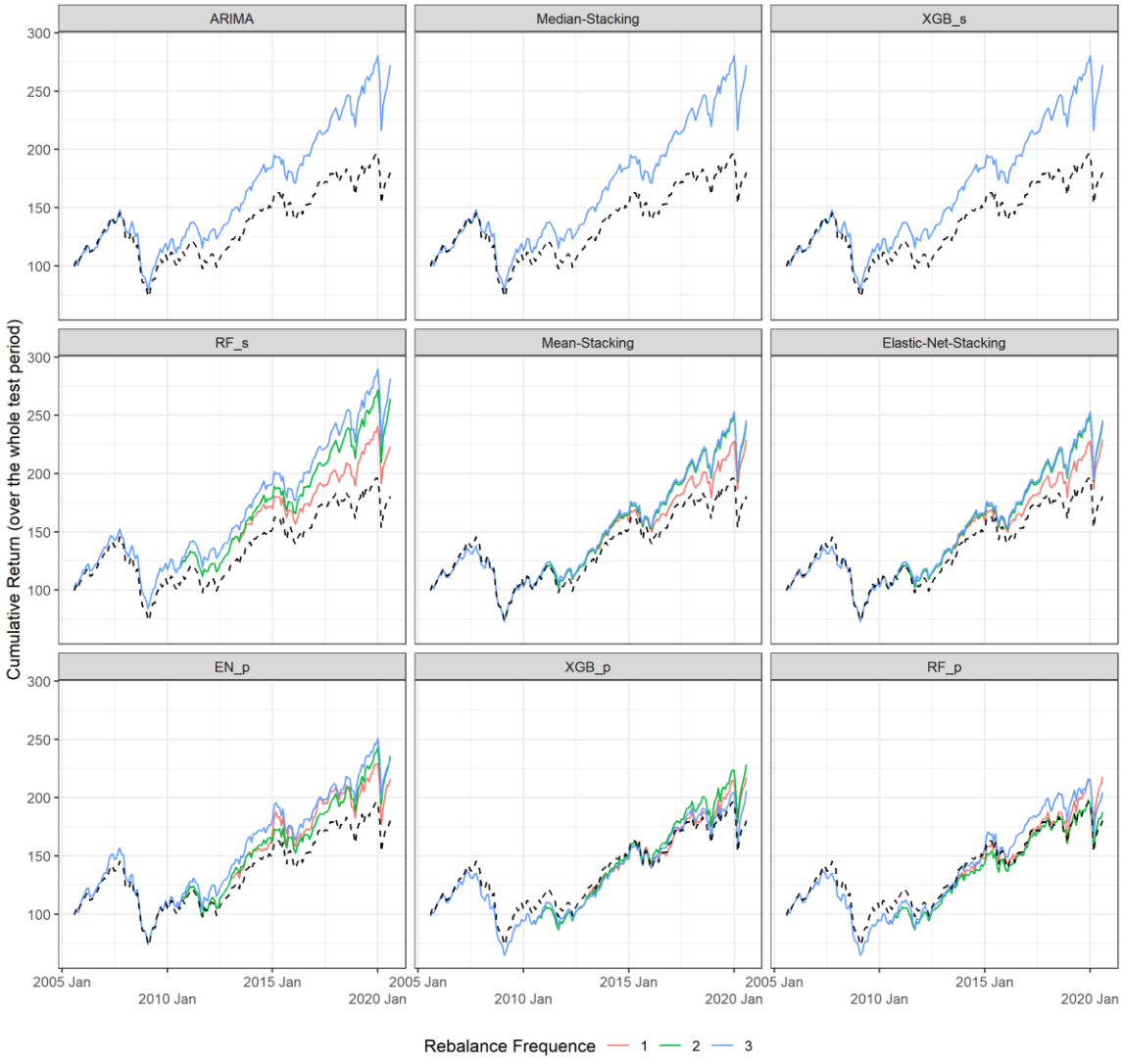
5.278 = higher yearly CAGR or annualized Sharpe Ratio than equal-share strategy
6.915 = higher yearly CAGR than benchmark and historical mean based strategy
0.562 = higher annualized Sharpe Ratio than benchmark and historical mean based strategy

[1)] Significance values refer to Wright et al. (2014), with correction for non-normal returns. Based on that we use a generalized method and estimate heteroscedasticity and autocorrelation consistent covariance matrices. We test for equality of our portfolios and equal-weight benchmark: Levels at 90% (*), 95% (**), 99% (***)

[2)] U.S. 10-Year Treasury Bond Rate acts as risk-free rate for Sharpe Ratio calculation

ARIMA, single-country XGBoost, and median-stacking show the same asset selection between their portfolios and across the backtesting period, eventually resulting in the same financial performance metrics. Nevertheless, these models beat the mean and naïve-based portfolios considerably, outperforming by 1% in yearly CAGR and Sharpe Ratio across all rebalancing variations. The three models lead only to statistically significant different Sharpe Ratios from the equal-weight benchmark at a 10% significance level, despite showing approximately 72% higher Sharpe Ratios.

Figure 5: Cumulative portfolio returns across rebalancing frequencies



7.5 Drivers of Financial Performance

We observe that predictive performance influences financial performance to a certain degree but is not its sole contributor. Pooled models, which showed a superior predictive performance compared to most single-country and benchmark models, underperformed in terms of Sharpe Ratio and do not necessarily lead to higher performance with more frequent rebalancing. Additionally, ARIMA-based predictions that struggled to beat previous benchmarks, lead to Sharpe Ratios similar to median-stacking model and single-country XGBoost as well as to the same selection in stock indices across three rebalancing variations. Therefore, the rank between the predicted indices plays a role for validating the financial performance in our set-up. If the resulting order among models is similar, despite a lower overall accuracy, the models perform similarly on our limited data set.

Because the data availability on fundamentals and macroeconomic features for further countries is limited as seen during preprocessing, checking the results for a larger set of country stock indices remains difficult. This becomes evident in the missing statistical significance Sharpe Ratio results. Additionally, our backtests show only a first indication on financial performance because the backtesting period is only 15 years long and shorter than of comparable research with 30 to 60 years. (Hou et al., 2020; H. Wang et al., 2019). Because of our five-year ahead forecasts, we only have a maximum of three instances in which we can rebalance the portfolio, leading to insignificant results on a past sample. General limitations of backtesting are highlighted in the next chapter. However, we observe that rebalancing contributes to better portfolio performance, if initial CAGR predictions are worse than later predictions. Nevertheless, rebalancing leads to transaction costs, which are discussed in the following chapter.

These findings help to understand why international mid-term horizon predictions are of limited interest in research. Nevertheless, we observe that our robust median-stacking model is on par with the best models across both metrics, without showing a strong variation in predictive accuracy across the stock indices, as both base-models tend to. Therefore, in our sample financial performance is not solely driven by predictive performance.

7.6 Feature Effects

Feature effects based on ALE are analyzed ex-ante and before the backtesting simulation, which is important to prevent reengineering of models (López de Prado, 2018, p. 153). Our stacked models find feature-target relationships, that are in line with fundamental theory.

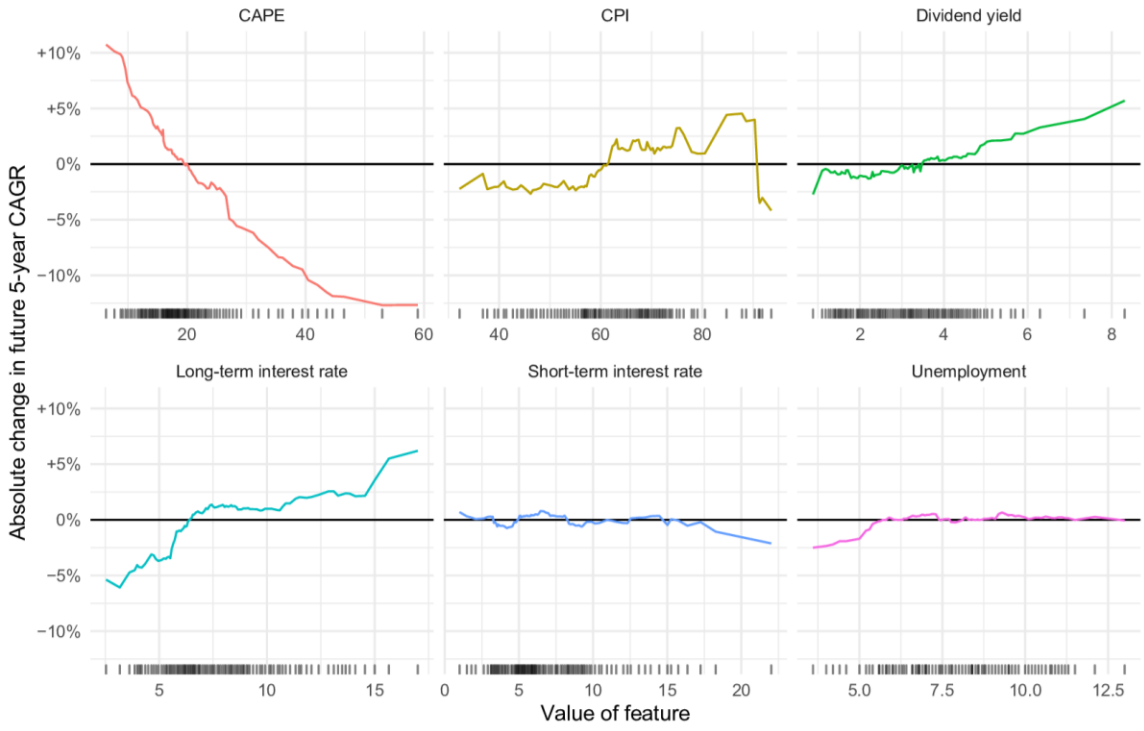
As seen in Figure 6, The ALE for the EN-stacked model shows that a lower valuation, as indicated by a low CAPE or a high dividend yield, and a higher long-term interest rate leads to higher returns. In addition to CPI, these three features are the most important since the range of the ALE indicates the relative feature importance throughout the feature values. CAPE has a non-linear relationship with the future returns, where a value of CAPE that was at the upper or lower range has a smaller effect on the target. The smallest CAPE of 6.30 in the training set increases the returns by 10.7% as compared to the mean CAPE of around 20 among the seven countries, while the maximum CAPE of 58.95 decreases the 5-year future returns by 12.7%. The long-term interest rate, which is the second most influential feature, only affects the future 5-year CAGR between 6.2% and -5.4%. The effect of the dividend yield is small but increasing and continues with a steady pace until the end of the range. The short-term interest and unemployment rates on the other hand do not show a meaningful effect on the future 5-year CAGR if outliers are disregarded, indicating that they do not have additional information for forecasting mid-term returns.

Our results are in accordance with past research that links valuations and returns (Campbell & Shiller, 1988; Radha, 2020). The high importance of CAPE can be explained by its ability to capture the discrepancy between the price and the underlying fundamental value of the stocks which make up the country-specific indices. The effect of the dividend yield was validated by the Gordon Growth Model (Gordon, 1959), which states that the present value of a stock is the sum of the discounted future dividends. Since the dividend yield is computed by using only the current dividend of each time point, the lower importance of the feature can be expected due to the higher variance of non-smoothed dividends.

In addition, the link between risk-free rates and the expected return of capital assets was shown in the theory regarding CAPM (Sharpe, 1964), which defines that higher risk-free rates lead to higher equity returns due to the expected return being a sum of the risk-free

rate and the equity-specific risk premium, assuming that they do not interact with each other. The interest rates of the 10-year government bonds reflect the expectations regarding the state of the economies and therefore also the underlying stocks of the indices. The results for the ALE based on the two other meta-models showed very similar results and are therefore shown in Figures D1 and D2 in the Appendix.

Figure 6: ALE for the stacked Elastic Net model



In conclusion, our models discovered patterns between the features and the future mid-term returns, thus, opening the “black box” of the machine learning models. In addition, our results shed light on the magnitudes and possible non-linearities of these relationships.

7.7 Comparison to Previous Research Results

The outperformance of benchmarks is not novel and often reviewed across research (Welch & Goyal, 2008). Hence, we focus on accuracy and resilience of our results, due to a lack of financial validation of machine learning based approaches. Wang et al. (2019) predicted the 10-year CAGR of S&P 500 and show a RMSE of 0.026 for their ensemble models.

Even though we use the MSCI index for the U.S., we can compare accuracies based on the similarities in methodology and features. The RMSE of our median-stacking with 0.040 is higher, despite having a similar training set length (33 vs. 29 years), but strongly different training intervals. For instance, their period from 1926 to 1959 includes the major financial crisis of 1929, while our training set includes comparably weaker recessions based on drawdown, limiting the CAGR range our base-models can use to train. Additionally, our shorter CAGR interval leads to lower out-of-sample RMSE because of higher variance of the target. Both the differences in training samples and target explain why the RMSE errors of our historical mean benchmarks are higher than of Wang et al. (2019). We deliberately choose a five-year interval in order to design a realistic model that can consider changes in regimes, policies or behavioral changes. Conclusively, the researchers show an error reduction of 50% on a test set of 58 years, leading to significant results at the 1% level. This is in accordance with the RMSE reduction of 44.7% to 52.1% (Table D3), we observe across seven MSCI indices, while our U.S. results are only significant to the 5% significance level and show an error reduction of just up to 39.1%.

Siami-Namini et al. (2019) predict stock index prices for indices in Hong Kong, Japan, and the U.S based on monthly macroeconomic and price data. The RMSE results are not fully comparable due to a different objective that relies on one-period-ahead-forecasts. Additionally, their rolling forecast includes test set observations into the training set to make predictions for the subsequent period. Despite these observations, our research agrees with their implications that univariate ARIMA models can be outperformed. They suggest a decrease in RMSE of 84% to 87% if long-memory adjusted recurrent neural networks (LSTM) are applied instead of ARIMA. While our results are not as strong, they also imply that non-linear methods can be superior for return forecasts.

Our research confirms findings from Jiang et al. (2020), that stacking performance depends on the meta-model and mainly adds resilience instead of predictive accuracy. We extend these findings by showing that higher resilience leads also to statistically significant forecast, by improving the signal-to-noise ratio compared to the historical mean.

8. Discussion

Our primary objective is the accurate and resilient forecast of stock indices returns five years ahead. Thus, we implement a stacked forecasting approach that combines forecasts of six base-models into an aggregated CAGR prediction. A portfolio strategy was implemented based on the results of the CAGR predictions and benchmarked against an equal-weight portfolio. In this final chapter, we discuss the limitations of our predictive and financial results. Also, we highlight consequences of our research for market efficiency, fundamental, and portfolio optimization research.

8.1 Limitations in Predictive Performance

The predictive performance of our stacked and base-models is greatly driven by the selected features, data availability, and the selected test set period, which may lead to look-ahead bias, survivorship bias, and lower reliability of our results, respectively. Additionally, the increased accuracy and robustness of stacking leads to high computational effort.

8.1.1 Features

On the one side, we select the features based on fundamental research to predict mid-term returns. Therefore, our research is subject to some degree of look-ahead bias due to overlapping periods between our research and the past research that indicates these features as useful (Campbell & Shiller, 1988). However, this bias is reduced since research on said features was pioneered before our test set begins. On the other side, we observe via ALE that some features could have been omitted from the feature set, which might have led to a decrease in out-of-sample performance.

8.1.2 Data Availability

Financial data can be assumed to be more available for economies that have large stock markets and are overall further evolved, possibly implying that the stock markets of these economies have also seen higher-than-normal returns. However, survivorship bias should not affect our results since we cannot assume that our base- or stacked models perform

better when returns are high. In addition, our benchmark for the financial performance consists of the equal weighted portfolio of the same countries we test the accuracy of, making the possible effect of the survivorship bias even smaller.

8.1.3 Test Set Period

Due to the time series nature and the limited number of years, the model accuracy and financial performance were only tested on a limited sample. While the different countries act as a mitigating factor for overfitting, financial data is correlated both across countries and time, which reduces the trustworthiness of our results, despite the significance test results. Therefore, we opted for a test set of 10 years to have a larger sample for testing without retraining the models for the latter five years. This allows us to include more periods with unforeseen events, such as the financial crisis and the recent COVID pandemic, that are be more difficult for the models to predict. Since the models performed rather well despite excluding these periods, we expect them to perform better with the data included.

8.1.4 Complexity of Pooling and Stacking

Developing a stacking framework requires domain knowledge and technical methodology know-how on strengths and weaknesses to build a truly heterogenous approach. The training time increases because different base-models, especially pooling, need to be trained and optimized. The second-layer needs, if a meta-model is used, training capacity resulting in high computational effort (Ribeiro & dos Santos Coelho, 2020). Because predictions are used as inputs in the second layer, generating feature effects from features requires the introduction of model-agnostic methods, to bridge both layers effectively.

8.2 Limitations of Financial Performance Results

The limitations associated with predictive performance also influence our financial results. Especially, the application of backtesting can lead to erroneous results and conclusions if the findings are not seen in context of the validation period. We deem transaction costs as rather unimportant for our set-up, mainly due to a low number of rebalancing periods and low fees for stock indices investments via exchange-traded funds (ETFs).

8.2.1 Challenges of Backtesting

Despite the popularity of backtesting in research, we must highlight that running a simulation on past observations is not regarded as a proper research tool, because it is carried out ex-post (López de Prado, 2018, p. 153). The past does not repeat itself (Ibid., p. 151) and backtesting can often lead to reverse engineering strategies with knowledge of the present that would not been available at the beginning of our testing period in 2005. The only key contribution a backtest validates are the scale of achieved returns and resilience to transaction costs. Therefore, even significant findings of financial outperformance do not guarantee similar results if the predictive models are used in reality. To limit the misclassification of our results, we circumvent the seven most popular backtesting fallacies, as defined by Luo et al. (2014), by focusing on stock indices, accounting for data leakage, and refraining from portfolio strategies that rely on shorting.

8.2.2 Significance of Financial Results

Predicting five-year ahead CAGR has a range of disadvantages, especially if evaluating the statistical significance of our results. Combining 60 months of returns into a single observation leads to strongly autocorrelated targets, meaning that we only have three uncorrelated CAGR predictions in our test set for each country (August 2005, 2010, and 2015). Therefore, we focus on predicting feature-target relationships using machine learning without lags, with only the ARIMA models using predicted CAGR lags. Nevertheless, our financial results can only be relied on with caution. Despite all the six base-model strategies and three meta-model portfolio strategies beating the equal-weight benchmark, none of the models show significant outperformance throughout all rebalancing cycles. As discussed, the low number of stock indices makes financial results too much dependent on the predicted order, rather than accuracy. Therefore, our financial results must be seen as bound to the selected simulation period from August 2005 to August 2020.

8.2.3 Transaction Costs

We do not account for transaction cost because these are not a cost driver in our scenarios as seen in active, daily trading strategies. Furthermore, transaction costs vary by market,

individual, and especially whether the investor is an institution or a retail investor. Our low rebalancing frequency, the low transaction costs for stock indices, and that we invest in only three out of seven indices are factors for lower transaction costs. A significant driver of transaction costs are expense ratios and bid-ask spreads with the latter only being of significant factor if traded ETFs shows high trading frequency (Angel et al., 2016), as the major stock indices we selected. Therefore, we can safely assume transaction costs of 0.6% per trade for private investors and significantly lower for institutional investors. Considering the significance level of our results, our median-stacking based strategy outpaced those costs in the span of one year, meaning that outperformance does not diminish due to the transaction costs, but only for the given backtesting scenario we selected.

8.3 Efficient Market Hypothesis – Quo vadis?

Considering the limitations of backtesting and the insignificant results for stacked models for significance levels at 5% or lower, we cannot safely claim that our CAGR predictions lead to a general exploitation of excess returns outside the selected time period. However, our predictive models show significantly lower forecast errors if compared to the historical mean benchmark. Thus, our stacked models are able to identify feature-target relationships that lead to explanation out-of-sample. This leads to an increase in the signal-to-noise ratio, as these relationships are also viable across multiple stock indices, despite a leakage-set, a prolonged test set, and changing financial market characteristics.

Nevertheless, the predictive and directional performances are not equal across the indices. Based on MAPE, stacked forecasts for the Netherlands performed worse than the benchmark, with only some base-models beating the benchmark lightly. Countries such as Canada and the U.S led to relatively small reduction in RMSE and MAPE for stacked models. On the other side, the pooled models outperformed the benchmark considerably for those three indices. Based on these differences we cannot infer market efficiency differences between the countries. Having used only a small number of features, we expect that further features may be more explanatory for the one or another stock index and that performance accuracies are mainly driven by feature selection.

We conclude that all MSCI stock indices except for the Netherlands show a rather weak-form of market efficiency, due to the statistically significant predictive performance out-of-sample which is driven by the fundamental features and our stacking approach.

8.4 Contribution of Our Predictive Model

Our contributions are three-fold and span the fields of portfolio optimization and fundamental research. Firstly, we showed that return predictions for stock indices can be more accurately predicted than by the current practices in portfolio optimization. Since the 1980s, ETFs gained strongly in popularity among investors, leading to 30% of total assets being locked in passive investments (Ben-David et al., 2014), and showing that passive stock indices are an important part of modern portfolios. Together with research milestones in risk prediction, more accurate risk and return predictions can support investment managers to provide more resilient portfolios, with retail investors, companies, and markets profiting from the gain in efficacy.

Secondly, we showed that the application of data pooling can increase accuracy while not showing overfitting tendencies. This is crucial for further fundamental and finance research because pooling helps to decrease the problem of data availability. By including data from other economies in the training process, we help the selected base-models to learn relationships more effectively, but also focusing on relationships that lead to significant error minimization across multiple countries.

Thirdly, our model confirms the long-standing fundamental research over the last decades and its uncovered relationships between excess returns and fundamental features. We show that complex machine learning and stacked models can be interpretable and accurate. In fact, this interpretability helps to validate our results in light of the long-standing research.

8.5 Implications for Further Research

Similar to our contributions, the possibilities for further research are two-fold and cover the areas of fundamental and regulatory research and portfolio optimization. Firstly, our flexible approach gives room to include more features or more countries. These features could include derivatives of the current ones, such as a cyclically adjusted version of the dividend-yield. Despite the data challenges faced with fundamental time-series data, reducing the test set length, or opting for different aggregation levels can increase the range of training data to even include emerging countries. Also, solving this data problem could help to increase the significance of the financial results, which depend on the sample length (Lo, 2002). On the contrary, using more features is not always the solution and can lead to overfitting, especially if features with no explanatory effect are used. Therefore, removing redundant features as shown by ALE could improve results. ALE could also be used to improve the out-of-sample accuracy with some of the models such as XGBoost, by constraining a monotone relationship between a feature and the target based on the observed relationships that are backed up by previous research. For example, this could include forcing the effect of higher valuations and lower long-term interest rates to have a strictly decreasing effect on the target. In addition, the application of our methodology to different assets or sectors can yield valuable insights in the predictive drivers between those stylistics. This can supplement current understanding but also show how those relationships change throughout time, giving more indication about market efficiency in niche markets.

Secondly, in order to evaluate the effect of more accurate predictions on portfolio optimization, these predictions need to be combined with portfolio selection techniques. The research efforts between both areas increases, but often relies on combining one type of machine learning (W. Chen et al., 2021) or deep learning model (W. Wang et al., 2020). Moreover, the research is only limited to short-term predictions using price data and technical indicators, leaving the space open for research aimed at mid- to long-term investors.

9. Conclusion

Considering relevant research on financial machine learning from Wang et al. (2019) and stacking from Jiang et al. (2020), we design a stacking framework to research if we can predict five-year ahead future returns of seven MSCI country indices more accurately than mean or naïve benchmarks. In addition, we investigate if the predictive accuracy of our stacked models influences financial performance, which is validated from 2005 to 2020.

The core of our heterogenous stacking framework is defined by four different model methodologies, namely ARIMA with exogenous features (ARIMA), Elastic Net (EN), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). Additionally, we apply pooling to extend the training set observations and limiting the problem of monthly data availability. This results in three single-country and three pooled models: ARIMA, XGBoost, RF, and Pooled XGBoost, Pooled RF, and Pooled EN. Based on six base-learner predictions, we construct a two-layered approach that weights the predictions according to the specific meta-model method, namely Median, Mean, and Elastic Net. Overall, we achieve a median MAPE reduction between 57.4% and 60.5% compared to the historical mean benchmark and among the seven stock indices, which is in accordance with results of previous research carried out for fewer indices. Particularly the pooled models show a strong performance across the stock indices, indicating that the increase in training data supported the models in finding feature-target relationships that were effective for forecasts out-of-sample. ARIMA performs the worst, mainly due to the lack in autocorrelation between features and target. Despite its linearity assumption for underlying feature-target relationships, Elastic Net performed well and mainly because of its pooled set-up. The three stacked models lead to statistically significant forecast improvements for six out of seven MSCI indices.

During the financial validation we use a long-only portfolio, which invests into three indices that are expected to perform the best based on the 5-year forward CAGR prediction. An equal-weight portfolio is used as benchmark in this historical backtest simulation. While most of the models beat the benchmark, only portfolios based on ARIMA, single-country XGB, and the median-stacked model show statistically significant improved Sharpe Ratios at the 10% level, based on the equality test defined by Wright et al. (2014).

Yet, these models outperformed the equal-weight benchmark by 2.1% in yearly CAGR by 72% in annualized Sharpe Ratio. The same performance metrics among these three models is a result of predicting the CAGR orders of the stock indices identically, also showing a limitation of our research. As the base-model with the worst out-of-sample MAPE performance, the ARIMA-based portfolio strategy shows identical profits as the single-country XGB and median-stacked model. Using only seven stock indices results in a setting in which financial performance is more driven by predicting a suitable order instead of accuracy. Therefore, we observe that predictive performance is only partly a driver for financial performance in our sample. The results could be validated with more stock indices and a longer data history to obtain more reliable information on outperformance and significance.

Generally, backtested results are often criticized due to their ex-post simulation. Therefore, our results must be seen as bound to the selected validation period. To offer another way of validation, we implement accumulated local effects (ALE) to gain an ex-ante view into the feature effects and their relationship with the target. This model-agnostic interpretation method helps our contribution to be contextualized within the fundamental research area. Our results confirm long-standing research findings on the relationship of CAPE, dividend-yields, and 10-year government bond yields in regard to future returns.

In essence, our interdisciplinary financial machine learning approach combines statistical significance testing and interpretability. Hence, its major contribution is the increase in future return predictability which can be used as an asset pre-selection step to research more robust portfolio optimization. Also, the inclusion of ALE should encourage further research to validate previously identified feature-target relationships for different assets, industries, and financial markets. On this note, using pooling to limit the effect of data unavailability may support future research in identifying reliable relationships out-of-sample that lead to statistically significant predictions compared to a benchmark.

Appendix

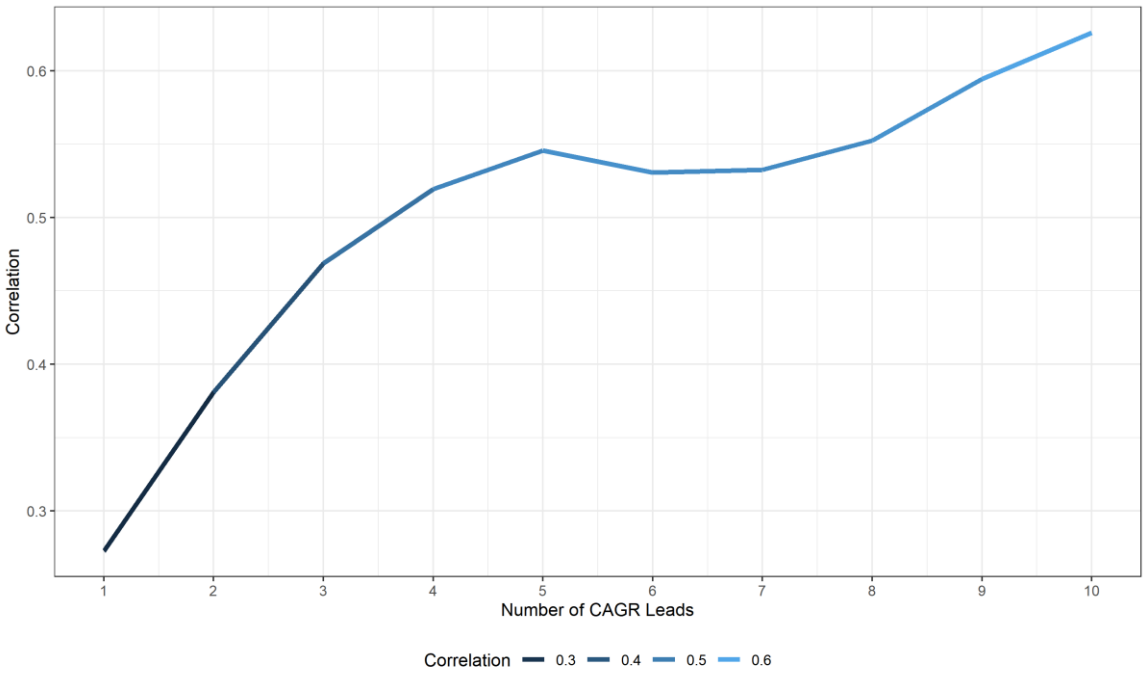
A. Data

The following tables and figures expand on index selection and feature correlation.

Table A1: Data availability for all MSCI country indices (before data segmentation)

Country	5-year CAGR	CAPE	Long-term interest rate	Unemp-loyment	Dividend yield	Short-term interest rate	CPI	MEAN
CANADA	40.7	33.7	60.6	60.6	51.6	59.6	65.6	53.2
USA	40.7	33.7	62.3	60.6	42.5	51.2	60.6	50.2
SWITZERLAND	40.7	33.7	60.6	**0.0**	42.6	41.6	60.6	46.6
GERMANY	40.7	33.7	59.3	24.5	42.6	55.6	60.6	45.3
UK	40.7	33.7	55.5	32.3	50.6	29.6	60.6	43.3
NETHERLANDS	40.7	33.7	56.6	32.5	42.6	33.6	55.3	42.1
AUSTRALIA	40.7	33.7	46.1	37.4	28.6	47.6	**0.0**	39.0
BELGIUM	40.7	**0.0**	60.6	32.5	42.6	57.7	60.6	49.1
FRANCE	40.7	16.5	55.5	32.5	27.6	45.6	60.6	39.8
NORWAY	40.7	**0.0**	30.6	26.4	**0.0**	36.6	60.6	39.0
HONG KONG	40.7	33.7	**0.0**	**0.0**	42.2	**0.0**	**0.0**	38.8
NEW ZEALAND	28.7	**0.0**	45.6	**0.0**	**0.0**	41.7	**0.0**	38.6
IRELAND	22.7	**0.0**	44.6	32.5	**0.0**	31.6	60.5	38.4
SWEDEN	40.7	33.7	28.7	32.5	**0.0**	33.6	60.6	38.3
...				...				...

Figure A1: Correlation between CAGR leads and CAPE, among selected indices



B. Details on Data Preprocessing

A thorough list on data preprocessing techniques can be found in Table B1. Due to the different properties of the models, we use different cross-validation and imputation methods based on the model type and data pool, aiming for the best model-specific approach.

The pooled and single-country tree-based models, RF and XGBoost, have the built-in capability to handle nonlinearities, feature selection, feature interactions, and outliers. Therefore, it is not required to add additional interaction features. Instead of performing imputation on the missing data or excluding features based on small amounts of data, we can mark the missing data with a value of 1000, so that it is treated appropriately in order to retain the feature-target interactions. ARIMA does not possess such capabilities, causing us to resort on K-Nearest Neighbors (KNN) imputation based on the training sets of each country. Due to the stationarity requirement of ARIMA models, we also force differencing for the models to be between the order of 1 and 5. We do not perform logarithmic transformation on the target because it does not improve performance in-sample.

In addition to performing the KNN imputation for the Elastic Net model, we standardize the features based on the training set to get the appropriate regularization equally for all of the features. EN-stacking does not require imputation due to all of the model-specific predictions being available, therefore requiring only the normalization for the preprocessing.

Table B1: Preprocessing overview across base- and meta-models

		XGB_pool	XGB_s	RF_pool	RF_s	EN_pool	ARIMA_s	EN-Stacking
	Pooling	Yes	No	Yes	No	Yes	No	No
Pre-processing	Imputation	No	No	No	No	KNN	KNN	No
	Feature standardization	No	No	No	No	Yes	No	Yes
	Logarithmic target	No	No	No	No	No	No	No
	Removal of NZV[1)] features	No	Yes	No	Yes	No	No	No
Hyper-parameters	Hyperparameter tuning	LHS[2)]	LHS[2)]	LHS[2)]	LHS[2)]	LHS[2)]	H-K[3)]	LHS[2)]
	Hyperparameter grid size	100	100	25	100	50	No	100
	One standard error rule	No	No	No	No	Yes	No	Yes

[1)] NZV features refers to features with near-zero variance
[2)] Latin Hypercube Sampling
[3)] Hyndman-Khandakar algorithm (2008)

C. Methodology

C.1. Resulting ARIMA Models

Based on the Hyndman-Khandakar algorithm, we select the AICc-minimizing models, as listed in Table C1. Normal CAGR lags are only used for Australia, Netherlands, and the U.S, while seasonal CAGR lags are used across all the other stock indices, showing that predicted CAGR from two years prior showed some explanatory power in-sample.

C.2. Hyperparameter Optimization and Cross-Validation

The available hyperparameters vary greatly by model, with only RF and XGBoost having some similar hyperparameters, in addition to the pooled and the corresponding single-country models sharing all of the same hyperparameters. We tune the hyperparameters for each specific model in-sample, by further splitting the training set into multiple training and validation sets and increasing or decreasing the range of a hyperparameter based on whether the specific optimal hyperparameter value was at its upper or lower limit. This way we can avoid any leakage and still get more optimal hyperparameter combinations. The considered hyperparameters, their ranges and resulting optimal hyperparameter values are available in Table C2.

Our available computational power limits possible hyperparameter grid sizes. With every additional hyperparameter combination, the number of computations increases in polynomial time. To consider extensive ranges for the grid search we use the Latin Hypercube sampling method as described (McKay et al., 1979), which allows us to consider more values for each hyperparameter without extensively increasing the computation time as compared to random or grid search (Urban & Fricker, 2010). We also treat the parameters of the ARIMA models as hyperparameters but apply the Hyndman-Khandakar algorithm to identify the best models in-sample, instead of searching the parameter space of a grid.

The choice of the optimal hyperparameter combinations is based on validation sets that select hyperparameters which lead to the lowest MAPE across the folds. Additionally, for the Elastic Net models we use the one-standard-error rule (Hastie et al., 2008) on the penalty hyperparameter to combat possible overfitting.

Table C1: ARIMA model specification and coefficients

Country	Code	Specification	ARIMA Coefficients					
			p(1)	q(1)	q(2)	P(1)	P(2)	Q(1)
AUSTRALIA	AUS	ARIMA(1,1,1)(2,0,0)[12]	-0.945***	0.914***		-0.166***	0.128***	
CANADA	CAN	ARIMA(0,1,0)(2,0,0)[12]				-0.036**	0.007**	
NETHERLANDS	NLD	ARIMA(1,1,0)(2,0,1)[12]	0.039***			-0.087	0.007***	0.144
GERMANY	DEU	ARIMA(0,1,0)(2,0,0)[12]				0.066***	0.028***	
SWITZERLAND	CHE	ARIMA(0,1,1)(2,0,0)[12]		0.149***		0.106***	0.050***	
USA	USA	ARIMA(1,1,0)(2,0,0)[12]	-0.181***			-0.017***	0.090***	
UK	GBR	ARIMA(0,1,0)(2,0,0)[12]				0.026***	0.128***	

Significance levels at 90% (*), 95% (**), 99% (***)

Table C2: Hyperparameter optimization results for all base- and meta-models

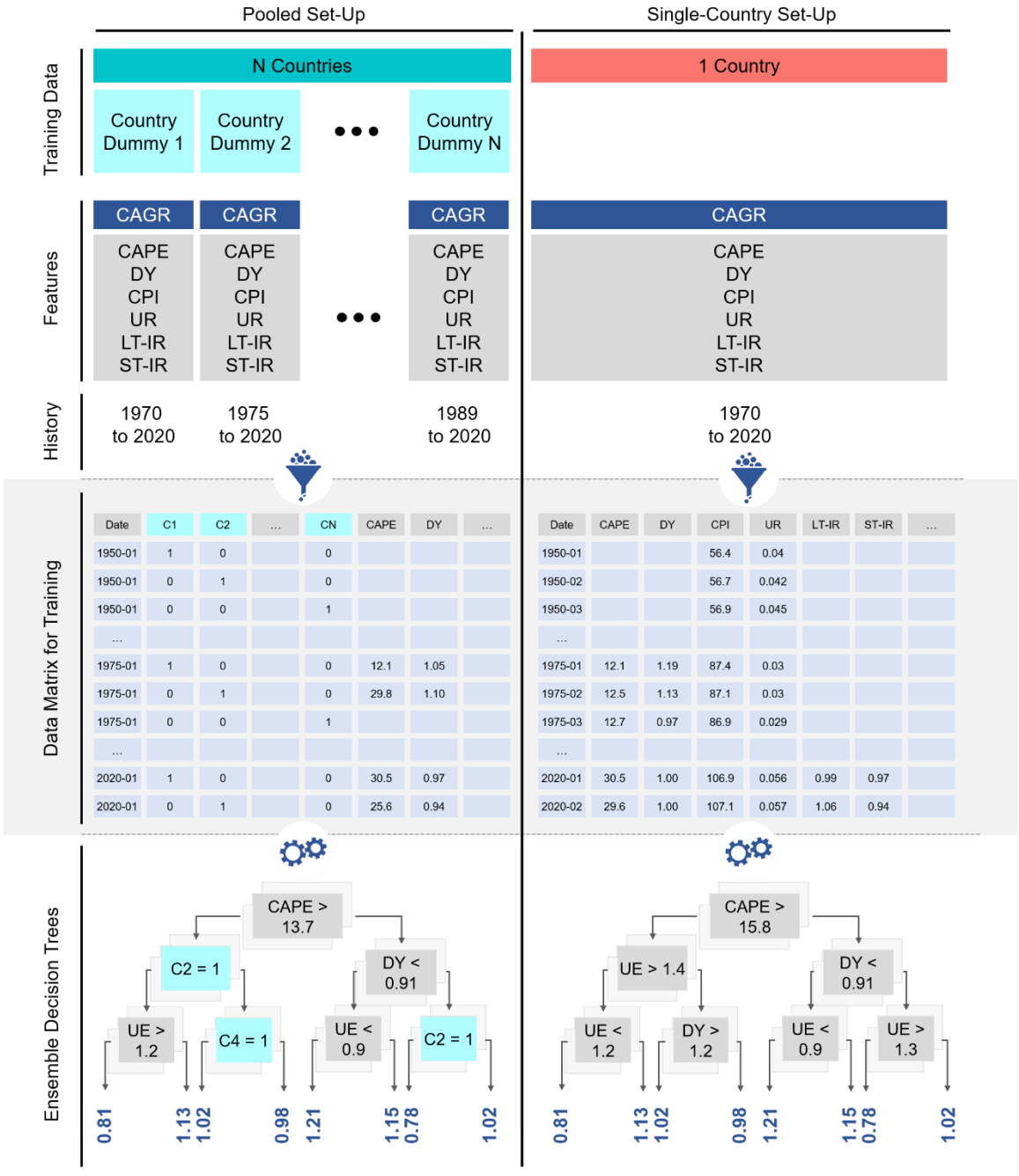
		Stacking	Pooled	Single-Country						
	Hyperparameter			AUS	CAN	DEU	NLD	CHE	GBR	USA
XGB	eta		0.191	0.061	0.262	0.061	0.177	0.151	0.104	0.104
	max_depth		55	28	46	28	19	56	15	15
	gamma		1.2E-05	5.1E-07	1.1E-09	5.1E-07	2.0E-06	4.2E-07	1.0E-08	1.0E-08
	colsample_bytree		0.804	1	1	1	1	1	1	1
	min_child_weight		3	2	4	2	8	7	4	4
	subsample		0.864	0.255	0.423	0.255	0.643	0.96	0.705	0.705
	nrounds		3087	2354	1454	2354	3829	2227	1438	1438
RF	mtry		46	2	2	2	5	3	5	5
	num.trees		700	90	90	90	195	778	268	995
	min.node.size		3	2	2	2	2	2	2	2
EN	alpha	0.408	0.172							
	lambda	4.58E-07	5.00E-03							

Table C3: Base-Model coefficients for the EN-stacked meta-model

Feature	**Intercept**	**XGB_pool**	**XGB_s**	**RF_s**	**RF_pool**	**ARIMA_s**	**EN_pool**
Estimate	1.110	0.026	0.026	0.017	0.017	0.000	0.000

C.3. Comparison of Pooled and Single-Country Set-Up

Figure C1: Training process of pooling and single-country set-up



D. Results

This Appendix chapter lists the RMSE and MAE for the resulting base-, meta-, and benchmark models. Additionally, we list the RMSE improvement to facilitate the comparison of our research results with results from similar research, as outlined in Chapter 7.7.

Correlation between actual and predicted CAGR values are listed in Table D4, and accumulated local effects for mean-stacked and median-stacked models in Figure D1 and D2.

Table D1: RMSE across base-, meta- and benchmark models

	Base						Meta			Benchmark	
	Pooled			Single-Country			Stacking				
Country	XGB_pool	RF_pool	EN_pool	ARIMA	XGB_s	RF_s	EN	Mean	Median	Mean	Naive
AUS	**0.037**	**0.030**	**0.039**	**0.075**	**0.041**	**0.044**	**0.034**	**0.037**	**0.038**	0.085	**0.058**
CAN	**0.045**	**0.032**	**0.027**	0.147	**0.053**	**0.038**	**0.042**	**0.048**	**0.039**	0.061	**0.060**
DEU	**0.054**	**0.044**	**0.053**	**0.071**	**0.038**	**0.069**	**0.038**	**0.040**	**0.036**	0.077	0.134
NLD	**0.079**	**0.078**	**0.063**	0.182	0.103	**0.073**	**0.066**	**0.073**	**0.066**	0.087	0.163
CHE	**0.056**	**0.070**	**0.046**	**0.043**	**0.047**	**0.057**	**0.053**	**0.050**	**0.048**	0.100	**0.075**
GBR	**0.031**	**0.059**	**0.029**	**0.026**	**0.020**	**0.039**	**0.023**	**0.022**	**0.023**	0.081	**0.063**
USA	0.088	**0.052**	**0.051**	**0.054**	**0.034**	**0.046**	**0.049**	**0.045**	**0.040**	0.066	0.128
MEDIAN	**0.054**	**0.052**	**0.046**	**0.071**	**0.041**	**0.046**	**0.042**	**0.045**	**0.039**	0.081	**0.075**

Table D2: MAE across base-, meta- and benchmark models

	Base						Meta			Benchmark	
	Pooled			Single-Country			Stacking				
Country	XGB_pool	RF_pool	EN_pool	ARIMA	XGB_s	RF_s	EN	Mean	Median	Mean	Naive
AUS	**0.029**	**0.020**	**0.031**	**0.073**	**0.032**	**0.032**	**0.024**	**0.027**	**0.027**	0.073	**0.044**
CAN	**0.038**	**0.024**	**0.023**	0.142	**0.041**	**0.029**	**0.033**	**0.041**	**0.029**	0.052	**0.052**
DEU	**0.045**	**0.036**	**0.043**	0.063	**0.031**	**0.057**	**0.032**	**0.034**	**0.030**	0.061	0.121
NLD	**0.056**	**0.055**	**0.056**	0.162	0.092	0.067	**0.059**	0.069	**0.061**	0.062	0.147
CHE	**0.045**	**0.051**	**0.035**	**0.032**	**0.038**	**0.043**	**0.037**	**0.036**	**0.036**	0.085	**0.067**
GBR	**0.025**	**0.052**	**0.024**	**0.022**	**0.015**	**0.033**	**0.017**	**0.016**	**0.017**	0.074	**0.053**
USA	0.068	**0.037**	**0.047**	**0.038**	**0.029**	**0.040**	**0.038**	**0.035**	**0.031**	0.049	0.114
MED	**0.045**	**0.037**	**0.035**	0.063	**0.032**	**0.040**	**0.033**	**0.035**	**0.030**	0.062	0.067

Table D3: RMSE reduction, compared to the historical mean model

	Base						Meta		
	Pooled			Single-Country			Stacking		
Country	XGB_pool	RF_pool	EN_pool	ARIMA	XGB_s	RF_s	EN	Mean	Median
AUS	57.1%***	65.2%***	54.1%***	11.8%	52%***	48.3%***	60.3%***	56.9%***	55.8%***
CAN	26.2%**	48.0%***	55.4%**	-140.3%	12.7%***	37.6%***	30.6%***	20.7%**	36.5%***
DEU	29.3%*	42.4%*	30.7%**	8.0%	50.2%**	9.8%	50.2%**	48.1%**	53.4%**
NLD	8.9%*	9.9%	27.2%	-108.8%	-18.8%	16.2%	24.3%	16.3%	23.8%
CHE	43.5%**	29.7%***	53.9%**	57.4%**	52.5%**	43.2%***	47.4%***	50.3%***	51.7%***
GBR	61.5%***	27.4%**	64%***	67.6%***	74.8%***	52.3%***	71.9%***	73.4%***	72.0%***
USA	-34.1%	20.9%**	22.1%	17.3%**	47.6%*	29.9%	25.9%*	31.8%**	39.1%**
MEDIAN[1)]	**32.5%**	**35.8%**	**42.9%**	**12.2%**	**49.3%**	**43.1%**	**47.6%**	**44.7%**	**52.1%**

Error reduction is calculated by: 1 - ($RMSE_{model}$ / $RMSE_{mean}$).

[1)] The median calculation computes first the median $RMSE_{model}$ and $RMSE_{mean}$

Significance values refer to Diebold-Mariano test (1995), whether the forecast errors are significantly different from the historical mean benchmark: Significance level at 90% (*), 95% (**), 99% (***)

Table D4: Correlation between actual and predicted CAGR

	Base						Meta		
	Pooled			Single-Country			Stacking		
Country	XGB_pool	RF_pool	EN_pool	ARIMA	XGB_s	RF_s	EN	Mean	Median
AUS	0.612	0.752	0.897	**0.903**	0.540	0.760	0.774	0.859	0.871
CAN	0.616	0.770	**0.835**	0.460	-0.171	0.426	0.572	0.637	0.660
DEU	0.724	0.838	0.606	0.377	0.815	0.702	0.833	**0.880**	0.866
NLD	0.326	0.666	0.563	-0.031	-0.062	0.604	0.583	0.399	**0.708**
CHE	0.723	0.433	**0.868**	0.664	0.745	0.210	0.778	0.845	0.855
GBR	0.478	0.610	**0.920**	0.757	0.823	0.833	0.811	0.886	0.892
USA	0.059	0.648	0.773	0.816	0.884	0.875	0.817	0.868	**0.887**

Figure D1: ALE for the median-stacked model

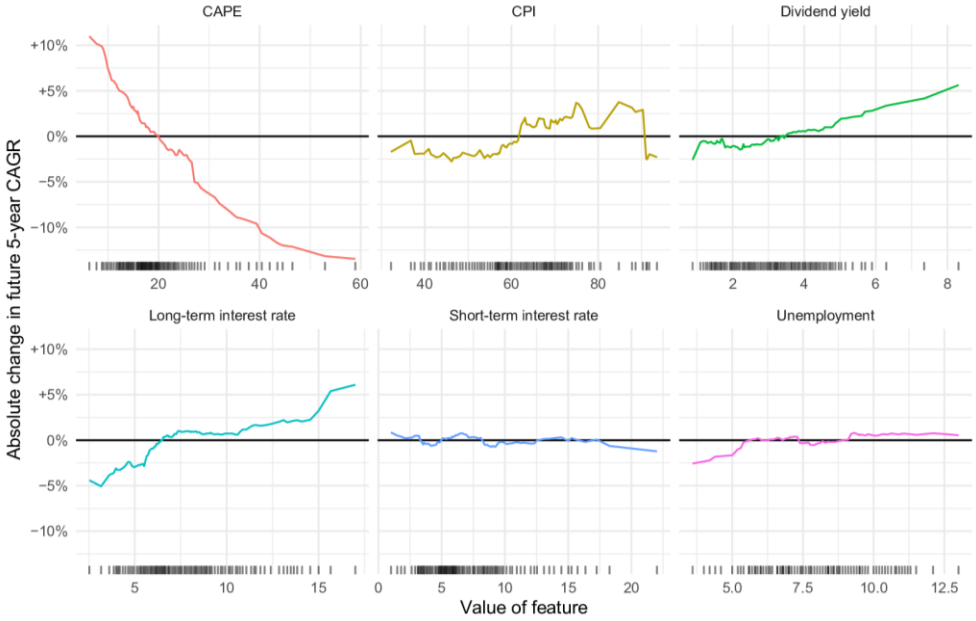
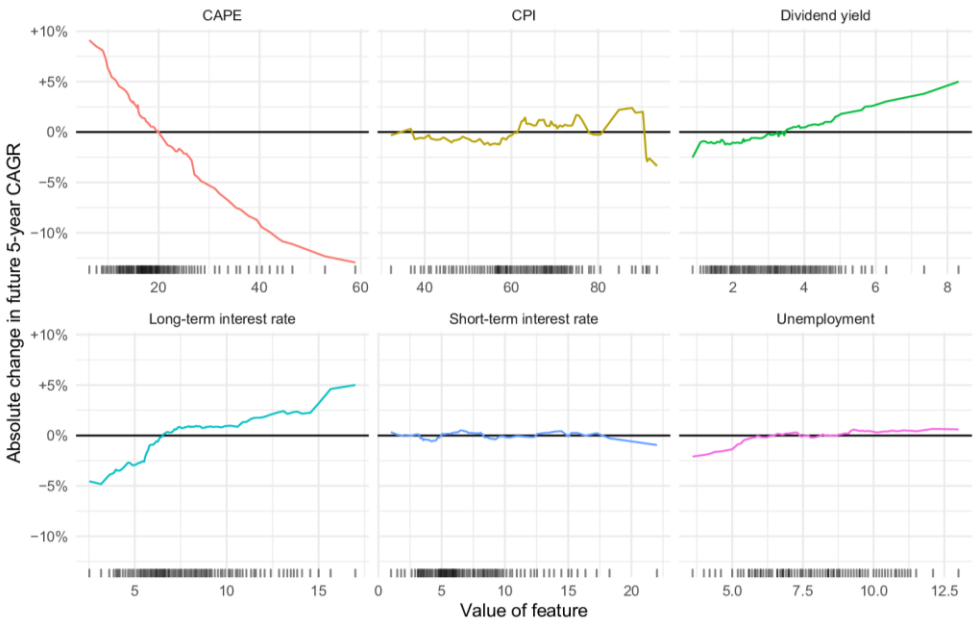


Figure D2: ALE for the mean-stacked model



E. Replicability

We use the R language to conduct the analyses in this paper. The complete code, with all commits and data is accessible via GitHub. The three main frameworks used are the *tidyverse*, *tidyverts*, and *tidymodels* package collections. The data wrangling and visualization are primarily done using functions from the *tidyverse* framework belonging to the *dplyr* and *ggplot2* packages, respectively, while *tsibble* and *fable* from *tidyverts* are used for time series modeling. For machine learning, we use the *tidymodels* framework. Diebold-Mariano significance tests (1995) forecast error differences between the base- and meta-models and the historical mean benchmark are computed using the *multDM* library.

To analyze the financial performance, we use the *SharpeR* library for Sharpe Ratio calculations and significance testing, the latter being based on equality tests defined by Wright at al. (2014). We do not rely on a specific package for the backtesting, but instead wrote the function suitable for our needs ourselves. The Accumulated Local Effects algorithm we use is a slight modification from the *ALEPlot* package.

The results might differ slightly depending on the operating system the code is run on. This is due to the different way the seed values affect the replicability of the functions that involve randomness, such as model fitting and the choice of the hyperparameter grid values inside the hyperparameter value ranges.

References

Allende, H., & Valle, C. (2017). Ensemble methods for time series forecasting. In *Studies in Fuzziness and Soft Computing* (Vol. 349, pp. 217–232). Springer Verlag. https://doi.org/10.1007/978-3-319-48317-7_13

Alvarez-Ramirez, J., Rodriguez, E., & Alvarez, J. (2012). A multiscale entropy approach for market efficiency. *International Review of Financial Analysis*, *21*, 64–69. https://doi.org/10.1016/j.irfa.2011.12.001

Andersson, J., & Lauvsnes, S. O. (2007). *Forecasting stock index prices and domestic credit: Does cointegration help?*

Angel, J. J., Broms, T. J., & Gastineau, G. L. (2016). ETF transaction costs are often higher than investors realize. *Journal of Portfolio Management*, *42*(3), 65–75. https://doi.org/10.3905/jpm.2016.42.3.065

Apley, D. W., & Zhu, J. (2020). Visualizing the Effects of Predictor Variables in Black Box Su-pervised Learning Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1059–1086.

Aye, G. C., Balcilar, M., Gupta, R., Kilimani, N., Nakumuryango, A., & Redford, S. (2014). Predicting BRICS stock returns using ARFIMA models. *Applied Financial Economics*, *24*(17), 1159–1166. https://doi.org/10.1080/09603107.2014.924297

Bao, Y. (2009). Estimation risk-adjusted sharpe ratio and fund performance ranking under a general return distribution. *Journal of Financial Econometrics*, *7*(2), 152–173. https://doi.org/10.1093/jjfinec/nbn022

Barak, S., & Modarres, M. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications*, *42*(3), 1325–1339. https://doi.org/10.1016/j.eswa.2014.09.026

Baumann, M., Baumann, M., & Herz, B. (2018). *Are ETFs Bad for Financial Health? : Some Counterintuitive Examples*.

Ben-David, I., Franzoni, F., Moussawi, R., Aragon, G., Downing, C., Ellul, A., Fardeau, V., Foucault, T., Frehen, R., Glushkov, D., Han, J., Hau, H., Landier, A., Madhavan, A., Mann, D., Martell, R., Menkveld, A., Nestor, R., Pagano, M., … Yan, H. (2014). Do ETFs increase volatility? In *NBER Working Paper Series*. http://www.nber.org/papers/w20071

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brooks, C. (2019). Introductory Econometrics for Finance. In *Introductory Econometrics for Finance*. https://doi.org/10.1017/9781108524872

Bukhari, A. H., Raja, M. A. Z., Sulaiman, M., Islam, S., Shoaib, M., & Kumam, P. (2020). Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access*, *8*, 71326–71338. https://doi.org/10.1109/ACCESS.2020.2985763

Campbell, J. Y., & Shiller, R. J. (1988). Stock Prices, Earnings, and Expected Dividends. *The Journal of Finance*, *43*(3), 661–676. https://doi.org/10.1111/j.1540-6261.1988.tb04598.x

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, *21*(4), 1509–1531. https://doi.org/10.1093/rfs/hhm055

Cervelló-Royo, R., & Guijarro, F. (2020). Forecasting stock market trend: a comparison of machine learning algorithms. *Finance, Markets and Valuation*, *2020*(1), 37–49. https://doi.org/10.46503/nluf8557

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, W., Zhang, H., Mehlawat, M. K., & Jia, L. (2021). Mean–variance portfolio optimization using machine learning-based stock price prediction. *Applied Soft Computing*, *100*. https://doi.org/10.1016/j.asoc.2020.106943

Damodaran, A. (2008). *What is the riskfree rate? A Search for the Basic Building Block*.

De Bondt, W. F. M., & Thaler, R. H. (1987). Further Evidence On Investor Overreaction and Stock Market Seasonality. *The Journal of Finance*, *42*(3), 557–581. https://doi.org/10.1111/j.1540-6261.1987.tb04569.x

Devenow, A., & Welch, I. (1996). Rational herding in financial economics. *European Economic Review*, *40*(3–5), 603–615. https://doi.org/10.1016/0014-2921(95)00073-9

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. In *Source: Journal of Business & Economic Statistics* (Vol. 13, Issue 3).

Fama, E. F. (1965a). Random Walks in Stock Market Prices. *Financial Analysts Journal*, *21*(5), 55–59. https://www.jstor.org/stable/4469865?seq=1#metadata_info_tab_contents

Fama, E. F. (1965b). The Behavior of Stock-Market Prices. In *Source: The Journal of Business* (Vol. 38, Issue 1). https://www.jstor.org/stable/2350752

Fama, E. F. (2014). Two pillars of asset pricing. *American Economic Review*, *104*(6), 1467–1485. https://doi.org/10.1257/aer.104.6.1467

Fama, E. F., & French, K. R. (1988). Permanent and Temporary Components of Stock Prices. *Journal of Political Economy*, *96*(2), 246–273. https://doi.org/10.1086/261535

Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, *112*, 88–97. https://doi.org/10.1016/j.dss.2018.06.008

Fisher, A. J. (2005). Multifrequency News and Stock Returns. *NBER Working Paper 11441*.

Guasoni, P., & Mayerhofer, E. (2019). The limits of leverage. *Mathematical Finance*,

29(1), 249–284. https://doi.org/10.1111/mafi.12172

Guoying, Z., & Ping, C. (2017). Forecast of Yearly Stock Returns Based on Adaboost Integration Algorithm. *Proceedings - 2nd IEEE International Conference on Smart Cloud, SmartCloud 2017*, 263–267. https://doi.org/10.1109/SmartCloud.2017.49

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning - Data Mining, Inference, and Prediction. In *Springer Series in Statistics* (Vol. 2). Springer.

Ho, M., Sun, Z., & Xin, J. (2015). Weighted elastic net penalized mean-variance portfolio design and computation. In *SIAM Journal on Financial Mathematics* (Vol. 6, Issue 1, pp. 1220–1244). Society for Industrial and Applied Mathematics Publications. https://doi.org/10.1137/15M1007872

Hou, X., Wang, K., Zhang, J., & Wei, Z. (2020). An Enriched Time-Series Forecasting Framework for Long-Short Portfolio Strategy. *IEEE Access*, *8*, 31992–32002. https://doi.org/10.1109/ACCESS.2020.2973037

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

James, G., Hastie, T., Tibshirani, R., & Witten, D. (2013). *An introduction to statistical learning : with applications in R*. New York : Springer, [2013] ©2013. https://doi.org/10.1007/978-1-4614-7138-7

Jiang, M., Liu, J., Zhang, L., & Liu, C. (2020). An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica A: Statistical Mechanics and Its Applications*, *541*, 122272. https://doi.org/10.1016/j.physa.2019.122272

Jin, S., Su, L., & Ullah, A. (2014). Robustify Financial Time Series Forecasting with Bagging. *Econometric Reviews*, *33*(5–6), 575–605. https://doi.org/10.1080/07474938.2013.825142

Kaestner, M. (2011). Investors' Misreaction to Unexpected Earnings: Evidence of Simultaneous Overreaction and Underreaction. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.868346

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. In *Econometrica* (Vol. 47, Issue 2).

Kendall, M. G., & Hill, A. B. (1953). The Analysis of Economic Time-Series-Part I: Prices. In *Source: Journal of the Royal Statistical Society. Series A (General)* (Vol. 116, Issue 1). https://www.jstor.org/stable/2980947

Kyriakou, I., Mousavi, P., Nielsen, J. P., & Scholz, M. (2020). Longer-term forecasting of excess stock returns-the five-year case. *Mathematics*, *8*(6), 927. https://doi.org/10.3390/math8060927

Leung, M. T., Daouk, H., & Chen, A. S. (2000). Forecasting stock indices: A comparison

of classification and level estimation models. *International Journal of Forecasting*, *16*(2), 173–190. https://doi.org/10.1016/S0169-2070(99)00048-5

Lo, A. W. (2002). *The Statistics of Sharpe Ratios*.

López de Prado, M. (2018). Advances in Financial Machine Learning. In *Wiley*. John Wiley & Sons, Inc. https://doi.org/10.2139/ssrn.3270269

Lyhagen, J., Ekberg, S., & Eidestedt, R. (2015). Beating the VAR: Improving swedish GDP forecasts using error and intercept corrections. *Journal of Forecasting*, *34*(5), 354–363. https://doi.org/10.1002/for.2329

Manish, K., & Thenmozhi, M. (2011). Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.876544

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *42*(1), 55–61. https://doi.org/10.1080/00401706.2000.10485979

McQuinn, N., Thapar, A., & Villalon, D. (2020). Portfolio Protection? It's a Long (Term) Story…. *The Journal of Portfolio Management*, jpm.2020.1.203. https://doi.org/10.3905/jpm.2020.1.203

Molnar, C. (2021). *Interpretable Machine Learning*. Chapter 5 Model-Agnostic Methods

Moreno, D., & Olmeda, I. (2007). Is the predictability of emerging and developed stock markets really exploitable? *European Journal of Operational Research*, *182*(1), 436–454. https://doi.org/10.1016/j.ejor.2006.07.032

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, *53*(4), 3007–3057. https://doi.org/10.1007/s10462-019-09754-z

Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research*, *253*(3), 697–710. https://doi.org/10.1016/j.ejor.2016.02.056

Peterson, R. (Ed.). (2012). *Inside the Investor's Brain*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119196945

Poterba, J. M., & Summers, L. H. (1987). *Mean Reversion in Stock Prices: Evidence and Implications*.

Quantitative Strategy, & Signal Processing. (2014). Seven Sins of Quantitative Investing. *Deutsche Bank*, 1–52. http://eqindex.db.com/gqs

Radha, S. S. (2020). Using CAPE to forecast country returns for designing an international country rotation portfolio. *Journal of Portfolio Management*, *46*(7), 101–117. https://doi.org/10.3905/jpm.2020.1.160

Ribeiro, M. H. D. M., & dos Santos Coelho, L. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series.

Applied Soft Computing Journal, *86*, 105837. https://doi.org/10.1016/j.asoc.2019.105837

Rounaghi, M. M., & Nassir Zadeh, F. (2016). Investigation of market efficiency and Financial Stability between S&P 500 and London Stock Exchange: Monthly and yearly Forecasting of Time Series Stock Returns using ARMA model. *Physica A: Statistical Mechanics and Its Applications*, *456*, 10–21. https://doi.org/10.1016/j.physa.2016.03.006

Samuelson, P. A. (1965). Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review*, *6*(2), 41–48. http://jrjgb.jj.cqut.edu.cn/__local/6/35/CE/516B5F529EC5AAF4B9C1FFD5C4B_A532FC8A_B39E2.pdf

Scalable Capital. (2016). *The Scalable Capital Investment Process White Paper*. https://scalable.capital

Scharfstein, D. S., & Stein, J. C. (1990). *Herd Behavior and Investment* (Vol. 80, Issue 3).

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2019). *Financial Time Series Forecasting with Deep Learning : A Systematic Literature Review: 2005-2019*. http://arxiv.org/abs/1911.13288

Sharma, A., & Thaker, K. (2015). Market efficiency in developed and emerging markets. *Afro-Asian Journal of Finance and Accounting*, *5*(4), 311–333. https://doi.org/10.1504/AAJFA.2015.073470

Sharpe, W. F. (1964). Capital Asset Prices: a Theory of Market Equilibrium Under Conditions of Risk. *The Journal of Finance*, *19*(3), 425–442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Shiller, R. J. (2000). *Irrational exuberance*. Princeton University Press. https://doi.org/10.5860/choice.37-6377

Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2019). A Comparison of ARIMA and LSTM in Forecasting Time Series. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

Sorensen, E. H. (2019). The golden age of quant. *Journal of Portfolio Management*, *46*(1), 12–24. https://doi.org/10.3905/jpm.2019.1.109

Thaler, R. H., & Johnson, E. J. (1990). Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. *Management Science*, *36*(6), 643–660. https://doi.org/10.1287/mnsc.36.6.643

Tversky, A., & Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability122. In *Cognitive Psychology* (Vol. 5).

Tversky, A., & Kahneman, D. (1974). Judgement under Uncertainty: Heuristics and Biases. *Science 185*, 1124–1130.

Umlauft, T. S. (2020). The Market Value of Equity-to-Gross Domestic Product Ratio as a Predictor of Long-term Equity Returns: Evidence from the U.S. Market, 1951-2019. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3551661

Urban, N. M., & Fricker, T. E. (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Computers and Geosciences*, *36*(6), 746–755. https://doi.org/10.1016/j.cageo.2009.11.004

Wang, H., Ahluwalia, H., Aliaga-Diaz, R. A., & Davis, J. H. (2019). The Best of Both Worlds: Forecasting US Equity Market Returns using a Hybrid Machine Learning – Time Series Approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3497170

Wang, W., Li, W., Zhang, N., & Liu, K. (2020). Portfolio formation with preselection using deep learning from long-term financial data. *Expert Systems with Applications*, *143*, 113042. https://doi.org/10.1016/j.eswa.2019.113042

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*(4), 1455–1508. https://doi.org/10.1093/rfs/hhm014

Weng, B. (2017). *Application of machine learning techniques for stock market prediction*.

Working, H. (1934). A Random-Difference Series for Use in the Analysis of Time Series. *Journal of the American Statistical Association*, *29*(185), 11–24. https://doi.org/10.1080/01621459.1934.10502683

Wright, A., John, C. P. Y., & Yung, S. P. (2014). A test for the equality of multiple Sharpe ratios. *Journal of Risk*, *16*(4), 3–21. www.risk.net/journal

Yuan, X., Yuan, J., Jiang, T., & Ain, Q. U. (2020). Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market. *IEEE Access*, *8*, 22672–22685. https://doi.org/10.1109/ACCESS.2020.2969293

Zopounidis, C., Galariotis, E., Doumpos, M., Sarri, S., & Andriosopoulos, K. (2015). Multiple criteria decision aiding for finance: An updated bibliographic survey. *European Journal of Operational Research*, *247*(2), 339–348. https://doi.org/10.1016/j.ejor.2015.05.032

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. In *J. R. Statist. Soc. B* (Vol. 67, Issue 2).