Received: 21 October 2019 | Revised: 26 March 2020 | Accepted: 30 March 2020
DOI: 10.1002/for.2690

RESEARCH ARTICLE

WILEY

Market timing using combined forecasts and machine learning

David A. Mascio[1] | **Frank J. Fabozzi**[2] | **J. Kenton Zumwalt**[3]

[1]Finance Department, Stetson University, Deland, Florida, USA

[2]EDHEC Business School, Nice, France

[3]Finance and Real Estate Department, Colorado State University, Fort Collins, Colorado, USA

Correspondence
Frank J. Fabozzi, EDHEC Business School, 393, Promenade des Anglais - BP311606202 Nice Cedex 3 - France.
Email: frank.fabozzi@edhec.edu

Abstract

Successful market timing strategies depend on superior forecasting ability. We use a sentiment index model, a kitchen sink logistic regression model, and a machine learning model (least absolute shrinkage and selection operator, LASSO) to forecast 1-month-ahead S&P 500 Index returns. In order to determine how successful each strategy is at forecasting the market direction, a "beta optimization" strategy is implemented. We find that the LASSO model outperforms the other models with consistently higher annual returns and lower monthly drawdowns.

KEYWORDS

beta optimization, combined forecast, machine learning, market timing strategy, sentiment index

1 | INTRODUCTION

Active money managers are continually seeking ways to "beat the market" and, generally, market timing and/or stock selection strategies are used. To use market timing to beat the market, the fund manager must successfully forecast future market directions and switch monies between common stocks and bonds or cash equivalents as the markets fluctuate over time. Originally, the market timing objective was to be long common stocks (100%) during bull markets and long cash equivalents (100%) during bear markets; risk-adjusted returns were yet to be developed. Interestingly, Phillips and Lee (1989) pointed out that market timers may view risk differently from many fund managers and suggest that "A market timer is not concerned with risk in a portfolio sense; risk to a market timer is being in or out of the market at the wrong time" (p. 15). Treynor and Mazuy (1966) pointed out that the payoffs from successful market timing resemble that of a call option as managers shift funds into stocks (bonds or cash) when the market is forecast to increase (decrease).

The most crucial facet of market timing is the accuracy of market forecasts. Fama et al. (1969) posited that because equity market prices "fully reflect" available information markets are efficient and stock prices follow a "random" walk. This "efficient market hypothesis" implies that investors cannot consistently outperform the market over time after adjusting for risk, transaction costs, and management fees, and that trying to do so is a futile effort. Early empirical studies of market timing report mixed results, especially when the frequency of switching is considered. However, as discussed later in this paper, more recent studies provide evidence that some fund managers outperform the market on a risk-adjusted basis and generate positive alphas over time.

To use market timing to generate positive alphas, fund managers must be successful in forecasting future market directions. Because the issue is such a hot-button topic, numerous research efforts with a focus on the efficacy of different forecasting procedures have been published. Over time, there have been substantial improvements in the forecasting environment. More timely and precise data are available at both the macroeconomic and microeconomic levels, and more powerful statistical and econometric techniques are available. Also, the speed of analysis has increased significantly.[1] Because of these technological

[1]Sprothen (2016) in an article in the the *Wall Street Journal* reported that the current computing speeds allow artificial intelligence models to run much more efficiently compared to when the models were developed.

 © 2020 John Wiley & Sons, Ltd.

innovations, it is not surprising that more current studies may generate results different from earlier studies.

The focus of this paper is how or if forecasting models can be useful in timing the market. There are several investor sentiment indexes used for forecasting returns. Five of the best-known ones have been tested by Mascio and Fabozzi (2019), and the one found to perform the best is the sentiment index model by Huang et al. (2015).[2] The forecasting literature suggests that combining forecasts offer the potential to provide a superior forecasting model.[3] In this paper, we test this hypothesis. Moreover, predictive analytics such as machine learning has been proposed as a model for forecasting returns. We also test whether machine learning provides superior return forecasting compared to the best single-investor sentiment index and the combined model.

The models are compared in two ways. First, how many market movements, both up and down, does each model correctly forecast? Second, how effective are the models at predicting the large down market movements? For example, although the models may exhibit similar forecast accuracy in the number of months correctly forecast, one may be superior in forecasting the down markets. Avoiding the worst months can enhance portfolio returns substantially.[4]

The paper proceeds as follows. Section 2 reviews the issues associated with market timing. Section 3 outlines the relationship between stock returns and (1) investor sentiment, (2) macroeconomic models, and (3) their combinations. Section 4 presents the data and methodology, while Section 5 summarizes the empirical results. Concluding remarks are in Section 6.

2 | MARKET TIMING LITERATURE REVIEW

An early study by Sharpe (1975) used an annual investment period to examine market timing and concludes that the timer must be correct approximately 70% of the time to outperform the market. Merton (1981) used the term *macroforecaster* to denote a market timer that predicts whether stocks (bonds) will outperform bonds (stocks). He also observed that successful market timers will generate portfolio returns that are "virtually indistinguishable" from successful option strategies. Henriksson and Merton (1981) developed a model to identify and specify gains from both market timing and stock selection. They observe that, generally, a researcher will not have access to fund managers' market timing forecasts but may use portfolio returns to infer the forecasts.[5] In an interesting contrast to Sharpe, Droms (1989) showed that a market timer only has to be correct 51% of the time with perfect monthly timing.[6]

Bollen and Busse (2001) suggested that fund managers may utilize intramonth portfolio adjustments in their timing strategies and use daily returns in their analysis. Interestingly, 34.2% (33.3%) of the funds outperform (underperform) their benchmark, suggesting a fairly uniform distribution. Fung and Hsieh (2001) examined *market timers* and *trend followers*; a market timer forecasts the future price direction while a trend follower looks for specific price patterns. They observed, like Merton (1981), that the risk for these types of investors cannot be captured by standard linear-factor models because their payoffs are nonlinear. Lam and Li (2004) found that daily switching could generate excess annual returns of approximately 80% for timers with low transaction costs. In addition, fund managers need to outperform the market only 60% of the time.

In a later paper Bollen and Busse (2005) extended the Henriksson and Merton (1981) model using the Carhart (1997) factors as explanatory variables. They observe that relative portfolio performance seems to persist over time, even though many earlier studies report negative excess returns. Their top decile of funds generates statistically significant excess returns of 25–39 basis points per quarter, and they attribute this outperformance to momentum and manager skill.

Examining the market-timing ability of mutual funds managers, Jiang et al. (2007) concluded that the average performance attributable to timing skills is positive.[7] Managers' response to public information such as aggregate earnings-to-price ratios and aggregate dividend yields appear to be the source of the excess returns.[8] Jiang et al. (2007) examined industry shifts in mutual fund port-

[2]This model was selected based on its superior risk-adjusted returns and prediction accuracy relative to the other four models tested.
[3]This literature is reviewed in Section 4.
[4]See Amin et al. (2004) for examples associated with avoiding the worst months, and Savor and Wilson (2014) on average announcement excess daily return.

[5]They point out that the inferences will "provide noisy estimates of the forecasts." They did not test the model empirically.
[6]More frequent switching generates greater transactions costs; the market timer must be successful enough in predicting the market to overcome these costs. Droms (1989) also points out that the periods under consideration can greatly affect the outcomes. He indicates that during the 1970s and 1980s market timers benefited substantially from high interest rates.
[7]The overall return could amount up to 0.6% annually.
[8]Jiang et al. (2007) examined five macroeconomic variables and suggested that further research should consider a wider range of variables. The variables they consider are (1) short-term interest rate (1-month T-bill yield), (2) term spread premium (10-year T-bond yield minus the 1-month T-bill yield), (3) credit premium (Moody's Baa-rated yield minus Aaa-rated corporate bonds), (4) aggregate dividend yield, and (5) aggregate earnings-to-price ratio of the S&P 500 Index.

folios over the business cycle. They reported that fund managers shift their industry composition substantially as they respond to macroeconomic information. Interestingly, Jiang et al. (2007) concluded that "on average, actively managed U.S. domestic equity funds possess positive timing ability."

Cremers and Petajisto (2009) developed an *Active Share* measure to evaluate mutual fund performance. Active Share shows how individual securities are over/underweighted relative to the benchmark. To outperform a benchmark, a fund must differ from the benchmark through factor timing and/or security selection. They find that three features are related to superior performance: (1) high active shares, (2) smaller asset funds, and (3) the best prior-year performance. Cremers and Petajisto found that the return above the benchmark was 6.5% per year net of fees and expenses.

Both market timing and stock picking skills were considered by Kacperczyk et al. (2014) in their evaluation of fund managers. Skill is defined as a "general cognitive ability to pick stocks or time the market." They report that a subset of smaller, more active funds generate abnormal returns by successfully performing both tasks. During recessions these successful managers hold more cash and rotate into defensive industries; portfolio betas tend to be lower. During economic expansions, successful managers invest in cyclical industries. Managers with the necessary cognitive skills focus more on microeconomic factors during economic expansions and on macroeconomic information during recessionary periods.

Jagannathan and Korajczyk (2017) pointed out that the capital asset pricing model (CAPM) divides total variation into systematic and idiosyncratic components. While idiosyncratic risk is assumed to be unpredictable, this is the component that generates the abnormal return or the *alpha* for a portfolio. They define skill as the "the ability to forecast the idiosyncratic returns of assets (which are unconditionally unforecastable)."[9]

Gasbarro et al. (2015) tested their contention that fund managers employ market timing to enhance portfolio performance.[10] They also reported a strong association between success in market timing and successful stock selection. They referred to their estimated coefficients as *timing trade betas*, and those coefficients that are statistically significant provide evidence of market timing.[11]

[9]They pointed out that the skill component is also known as "the risk-adjusted return, the abnormal return, Jensen measure, or alpha of the portfolio."

[10]The methodology was outlined in Gasbarro, Cullen, and Monroe (2015).

[11]A statistically positive (negative) timing trade beta identifies trades designed to increase (reduce) the portfolio beta by tilting the fund's portfolio towards higher (lower) beta stocks.

3 | STOCK RETURN FORECASTING

In this section, we describe how stock returns are impacted by investor sentiment. In addition, we review macroeconomic models and how they relate to stock returns.

3.1 | Investor sentiment and stock return forecasting

Fisher and Statman (2003) contended that "consumer confidence predicts economic activity" and asked: "does consumer confidence also predict stock returns?" and "what is the relationship between confidence and investor sentiment?"(p. 115). Their analysis reveals a significantly positive relation between changes in consumer confidence and contemporaneous stock returns. However, consumer confidence appears be a contrarian indicator, as an increase in confidence causes the bidding up of stock prices which, in turn, results in negative future stock returns. Moreover, a positive relation is observed between changes in consumer confidence and changes in investor sentiment for individual investors, but not with changes in institutional investor sentiment.

Brown and Cliff (2004) examined the relations between several technical sentiment indicators and survey-type investor sentiment measures.[12] Surveys by the American Association of Individual Investors and Investors Intelligence are used to distinguish between individual and professional investors. They find that market returns predict future individual and institutional investor sentiment, but find little evidence that the sentiment measures can predict future stock returns.

Amenc et al. (2004) used an "almost exhaustive set" of models to examine the risk-adjusted performance of hedge fund managers.[13] Assuming a normal distribution, positive alphas are generated for all the models. However, the average hedge fund alpha is not significantly positive. Interestingly, large funds outperform small funds, and newer funds outperform older funds, high incentive funds outperform low incentive funds, and market neutral funds outperform the average of other funds.[14]

[12]Kalman filters, principal component analysis, and vector autoregression procedures are used in the analysis. Trading volume, type of trade, derivatives, and *other* are the technical indicators.

[13]The CAPM and four CAPM-adjusted models are used. The four CAPM-adjusted models are (1) a payoff distribution pricing model, (2) a CAPM with multiple rewarded-risk factors, (3) an implicit factor model using principal component analysis, and (4) a multi-index model. Control variables include yield on 3-month Treasury bills, dividend yield, term spread of 10-year yield minus 3-month yield on Treasuries, and a credit spread of AAA bond yield minus Baa bond yield.

[14]Additionally, negative betas are associated with short-selling funds and administrative fees appear to be irrelevant.

Investor sentiment is the focus of M. Baker and Wurgler (2006, 2007). They developed two investor sentiment indexes based on their belief that investor sentiment is different for different types of companies. They conjectured that stocks that are more difficult to value and arbitrage are more sensitive to investor sentiment. Hard-to-value and arbitrage stocks are “low capitalization, younger, unprofitable, high-volatility, non-dividend paying, growth companies or firms in financial distress.” They reported that for the hard-to-value and arbitrage firms future returns are especially low (high) when sentiment is high (low). The more stable, value stocks are less affected by investor sentiment.

Baker et al. (2012) and Huang et al. (2015) each developed similar models designed to predict future market returns. Using a partial least squares methodology to create a sentiment index that they labeled the Sentiment Partial Least Squares (SPLS), they reported that the explanatory power of their model was five to six times larger than the sentiment index of M. Baker and Wurgler (2006, 2007). Their results are both statistically and economically significant.

3.2 | Macroeconomic models and stock return forecasting

As was pointed out earlier, the objectives of different statistical and econometric models can be quite diverse. In addition, the model may be designed to explain or to predict the information under consideration. If a model can successfully explain the information of interest, can it be successful in predicting future values? A review of several relevant research efforts is presented below in a roughly chronological sequence. The models represent different objectives, data sets, and methodologies.

Developing an aggregate consumption-wealth ratio, Lettau and Ludvigson (2001) reported that deviations from the trend could explain a significant proportion of next quarter's returns; positive (negative) trend deviations precede large positive (negative) excess returns. They conclude that expected returns vary over time and that fluctuations in the aggregate consumption-wealth ratio are useful in forecasting both real and excess stock returns.

Ludvigson and Ng (2007) used 209 quarterly macroeconomic activity variables and 172 financial variables in a dynamic factor analysis framework to examine excess stock market returns. They identified the first factor as a *volatility factor* and the second factor as a *risk premium factor*. When they combine these two factors with a consumption–wealth variable developed by Lettau and Ludvigson (2001), they found that the model predicts 16% of the one-quarter-ahead excess market returns. They conclude that the stability and statistical significance of the factors represents an improvement over the earlier “low-dimensional forecasting regressions.”

The Philadelphia Federal Reserve Bank provides data on the Business Conditions Index developed by Aruoba et al. (2009)—hereafter ADS. The Business Conditions Index is now known as the ADS Index. This index uses four variables observed at different frequencies in a dynamic factor model to create an index of macroeconomic conditions. The variables and their frequencies are GDP (quarterly), employment (monthly), initial jobless claims (weekly), and the slope of the yield curve term premium (daily).[15] To accommodate the different frequencies of the observations among the variables, ADS used a Kalman filter process. The model generates an index with the peaks and troughs consistent with those identified by the NBER, although the ADS index tends to reach its peaks and troughs earlier than the NBER.

3.3 | Forecasting market conditions

Gilchrist et al. (2016) used the prices of outstanding corporate bonds to create the GZ Credit Spread Index. This index is used to examine the relation between credit spreads and economic activity.[16] Moreover, they split the index into two components: (1) a component that captures expected default risk of individual firms and (2) a residual component, the excess bond premium (EBP), that captures credit market sentiment. They observe that during the 2007–2009 global financial crisis the creditworthiness of cash-market financial intermediaries weakened substantially, resulting in an increase in the credit spread. The decline in the risk-bearing capability of this sector substantially reduced the supply of credit and caused serious adverse consequences for the macroeconomy.[17]

Daily information was used by Da et al. (2015) to develop a sentiment index called FEARS, an acronym for “Financial and Economic Attitudes Revealed by Search.” The index is based on information found by screening the Internet for terms such as *recession*, *unemployment*, and *bankruptcy* that may cause anxiety for investors. They relate changes in asset prices, volatility, and fund flows to the FEARS index.[18] Overall, Da et al. (2015) concluded that

[15] The slope of the yield curve is measured as the 10-year yield minus the 3-month yield on US Treasuries.

[16] The GZ credit spread index is used to forecast (1) the growth of private (nonfarm) payroll employment, (2) the change in the civilian unemployment rate, and (3) the growth of manufacturing industrial production. Their empirical results showed that “both the excess bond premium and the predicted GZ credit spread contain significant independent explanatory power for all three economic indicators, at both the 3- and 12-month forecast horizons.”

[17] Gilchrist and Zakrajsek (2012) did not attempt to forecast movements in the stock market.

[18] Google Trends in the Search Volume Index (SVI) is used: http://www.google.com/trends/

aggregate market returns are somewhat predicted by the FEARS index.

Baker et al. (2016) developed the Economic Policy Uncertainty (EPU) index using newspaper coverage, tax law expirations, and economic forecaster disagreement to examine their relation with the micro- and macroeconomic factors.[19] They reported that stock price volatility was positively related to the EPU index, while investment and employment in government-dependent sectors were negatively related to the EPU index. Three measures of equity uncertainty—the CBOE VIX, an equity market uncertainty index, and a stock-jump measure—were positively related to the EPU.[20]

Jurado et al. (2015) examined macroeconomic uncertainty using both macroeconomic and financial time series data.[21] They pointed out that the volatility of the stock market is a popular proxy for uncertainty, but conclude that periods of macroeconomic uncertainty occur less frequently than suggested by popular uncertainty proxies. In addition, they evaluated the relation between macroeconomic uncertainty and business cycles using principal component analysis.[22] Twelve principal components explain approximately 54% of the total variation in the data. The first principal component is strongly related to stock market information and explains 37% of the total variation in the data, the next two principal components are related to real economic activity and risk, and combined explain 8%. Bond market term spreads explain 3%. Three episodes of macroeconomic uncertainty are identified, all occurring during recessionary periods.[23]

Two recent papers employ data science tools (big data, artificial intelligence, machine learning) to forecast stock returns. Wang and Leung (2018) and Abe and Nakayama (2018) used big data and sophisticated deep learning procedures. Wang and Leung used a factor selection algorithm in a neural network (a type of artificial intelligence tool) to generate an efficient financial forecasting model. Ninety-five financial variables were used to predict 1-month-ahead stock prices. They created long/short portfolios based on the top and bottom return deciles of each company in their sample. The deep reinforcement learning model (a type of neural network) substantially outperformed competing models on all four relevant measures (portfolio return, standard deviation of return, the Sharpe ratio, and maximum drawdown). They concluded that users of their model should "expect the deep reinforcement algorithm to improve their model performance."

Abe and Nakayama (2018) used deep neural network (DNN) to forecast 1-month-ahead returns for Japanese stocks. Twenty-five traditional financial variables are used as input values and future stock returns are the output values. Three measures of performance are used: "the rank correlation between the actual out-of-sample returns and their predicted scores, directional accuracy, and the performance of a simple long-short portfolio strategy." They compared their results with forecasts from two machine learning techniques: support vector regression (SVR) and random forest (RF). They then combined the forecasts of the DNN, the SVR, and the RF in an ensemble methodology. They reported that the ensemble model outperformed each of the single models. Long–short portfolios are constructed using tertile and quintile portfolios. They buy (sell) the top (bottom) stocks to get a net-zero investment. They found that the ensemble methodology outperforms the individual models for the correlations and for the quintile portfolio, but not for the tertile portfolio. They concluded that further improvements in deep-learning models will enhance prediction accuracy in the future.

4 | DATA AND METHODOLOGY

Section 4.1 begins by briefly reviewing the literature, focusing on the combining of forecasts, and Section 4.2 presents the five forecasting models that are combined into the kitchen sink model and the machine learning LASSO model. Section 4.3 outlines the data, and Section 4.4 explains the logistic regression model. Sections 4.5 and 4.6 present the kitchen sink and LASSO models, respectively. The final section presents the certainty equivalents from logistic regressions.

4.1 | Combining forecasts

Combining forecasts can be complex, and there is a substantial body of literature dealing with this issue. While a comprehensive review of forecasting is beyond the scope of this paper, selected relevant articles are briefly reviewed here. Elliott and Timmermann (2008) observed that forecasting was only useful when it could improve financial and/or economic decisions. Wallis (2011) reviewed research on combining forecasts and agreed with Bates and

[19]They screened newspaper coverage for words that suggested economic uncertainty. The second uncertainty variable is the dollar amount of scheduled tax expirations obtained from the Congressional Budget Office. The Philadelphia Federal Reserve Bank's measure of economic forecaster disagreement is the model's third component. Weights are assigned to each factor and the results are summed to complete the EPU Index.

[20]The correlations between the EPU and the VIX and the stock-jump measure are 0.578 and 0.575, respectively. The correlation between the EPU and the equity market uncertainty index is 0.743. The equity market uncertainty index is created in a similar fashion as the EPU index, using stock market related words in the screens. Stock-jumps are defined as ±2.5% stock market jumps.

[21]The macroeconomic time series data include 25 financial indicators.

[22]They focused on industrial production, employment, and hours as reflecting the business cycle.

[23]The period examined runs from July 1960 through December 2012. The three recessionary periods are 1973–1974, 1981–1982, and 2007–2009.

Granger (1969) that a linear combination of two competing forecasts results in better accuracy than the individual forecasts. Forty years ago, Makridakis and Hibon (1979) showed that generally simple forecasting models outperformed more sophisticated approaches and that a simple average of forecasts outperformed individual techniques. Lobo (1991) compared analysts' forecasts of annual corporate earnings from I/B/E/S and Value Line with statistical forecasts. He concluded that forecast accuracy could be considerably improved by combining analysts' forecasts with statistical model forecasts. Kim et al. (2001) examined combined forecast accuracy assuming analysts have two types of information: common and private. When analysts' forecasts are combined, the common information is overweighted relative to the private information.

Wallis (2011) pointed out that to produce a better forecast it is useful to have "individual forecasts based on different information sets," but that "aggregating forecasts is not the same as aggregating information sets." Investigating whether a simple average of expert forecasts is the optimal way to combine forecasts, Genre et al. (2013) compared the average of expert forecasts with a number of time series, recursive, and principal component models. Although a number of combination strategies improve the forecast relative to the benchmark, the gains are modest and they could not identify one particular strategy that dominated across different horizons or variables. Elliott et al. (2013) used a complete subset regression procedure to examine estimation errors in forecasting stock returns. They combined forecasts from all possible linear regression models. In addition, model complexity is related to the tradeoff between the bias and the variance of forecast errors. Their results indicate "that combinations of subset regressions can produce more accurate forecasts than conventional approaches based on equal-weighted forecasts."

In summary, it appears that the combining of forecasts can improve forecast accuracy. This study compares the forecast accuracy of three models, but also outlines the implications for fund managers that are concerned with risk/return tradeoffs and portfolio drawdowns

4.2 | Forecasting models

In this paper we combine the forecasts of the following five models: (1) Investor Sentiment Index (SI), introduced by Huang et al. (2015); (2) the ADS Business Conditions Index (ADS), created by Aruoba et al. (2009); (3) the GZ spread index (GZ), by Gilchrist and Zakrajsek (2012); (4) the Financial Uncertainty Index (FUI), developed by Jurado et al. (2015); and (5) the Economic Policy Uncertainty Index (EPU), developed by Baker, Bloom, and Davis (2015).[24,25] Two procedures are used to combine these five models. The first combines the five indices into a kitchen sink index, and the second uses the least absolute shrinkage and selection operator (LASSO), which is a machine learning technique used to choose the best forecasting index each period.

4.3 | Data

The five forecasting indexes are used to predict the direction of the S&P 500 Large Cap Stock Index (SPX) 1-month ahead. The forecast for each index is F_{kt}, where k refers to the particular forecasting model, $k \in (1 : 5)$ and t is the month. The forecasts are rolled forward 1 month at a time, and the return in the following month, R_{t+1}, is regressed on the forecast as follows:

$$R_{t+1}^{m} = \alpha + \beta_k F_{k,t} + \epsilon_{t+1}. \tag{1}$$

The results for each forecasting model are compared on both a cross-sectional and time series basis. In the present paper, the forecasts are then combined as explained in the next section.

In order to match the same date range for all the forecast indexes, the first month of the study is January 1985, and the last month is April 2014, resulting in 312 monthly observations in the sample period. This period differs from (Mascio & Fabozzi, 2019) in order to eliminate the high interest rate period of the late 1970s and the early 1980s, and to determine how well the models forecast in a low inflationary environment.[26] The actual data came from several sources, and include periods of strong growth, recession, and low inflation along with two sharp stock market selloffs. The S&P 500 Index monthly data came from Bloomberg. Data for the improved Sentiment Index (SI) came directly from Huang et al. (2015),[27] and data for the ADS Business Conditions Index came directly from the Philadelphia Federal Reserve Bank.[28] The GZ Credit Spread Index data come directly from Gilchrist et al. (2016).[29] The Economic Policy Uncertainty (EPU) Index is available online,[30] and the Financial Uncertainty Index

[24] See Mascio and Fabozzi (2019) for a detailed description of each forecast index.

[25] This model replaces the Investor Sentiment Index by M. Baker and Wurgler (2007) from Mascio and Fabozzi's (2019) paper. The primary reason for the substitution is to have a more diverse group of forecasting models to create a lower correlation amongst the models.

[26] Mascio and Fabozzi (2019) examined the time period from 1/1973 to 3/2014, in which the yield on the US 10-year Treasury note during the 1970s averaged nearly 12.5%.

[27] We appreciate Dashan Huang's responsiveness in providing the entire data set from their 2015 paper.

[28] The web address link is https://www.philadelphiafed.org/research-and-data/real-time-center/business-conditions-index/

[29] The Fed Notes Web address link is https://www.federalreserve.gov/econresdata/notes/feds-notes/2016/files/ebp.csv

[30] The data are available from https://www.policyuncertainty.com/

(FUI) data are available from the Board of Governors of the Federal Reserve System.[31]

4.4 | Logistic regression

Similar to Mascio and Fabozzi (2019), we use the five predictive indexes to create a beta optimization strategy. Each model's information is used to forecast an "up" or "down" market for the next month for the S&P 500 Index (SPX1). The dependent variable (SPX1) has two mutually exclusive outcomes estimated using a logistic regression. Specific to this experiment, when the market is forecasted to be up (down) in the subsequent month, the parameter is assigned a value of 1 (0). The five models are the forecasting factor, $F_{k,t}$, where k refers to the models, $k = 1, \dots 5$, and t designates the month.

The prediction procedure begins by first categorizing the market direction as "Up" if the SPX performance in the next month is positive, and "down" if the SPX performance is negative, resulting in an up or down binary variable x_{t+1} being created for each month t. Next, we line up the current month indexes (factor) $F_{k,t}$ with the next month SPX direction x_{t+1}. Lastly, we use the previous 24 months $(t-25, t-1)$ of observations to generate a rolling estimation window.[32] The probability of an upmarket is denoted by $x = 1$ as $p(x)$. The following is the logistic regression model:

$$\log \frac{p(x_{t+1} = 1)}{1 - p(x_{t+1} = 1)} = \alpha_k + \beta_k F_{k,t}, \tag{2}$$

where $F_{k,t}$ is an index. Equivalently,

$$p(x_{t+1} = 1) = \frac{1}{1 + e^{-(\alpha_k + \beta_k F_{k,t})}}. \tag{3}$$

The maximum likelihood approach is used to estimate the parameters, and the index values at the end of the month use estimated logit model parameters. The probability of the SPX having a positive return for the next month is determined by the previous 24 months of observations. An up market (down market) is forecasted when $p(x = 1) \geq 0.5$ ($p(x = 1) < 0.5$.). That is, if $p(x) \geq 0.5$, we predict $x = 1$. Once the logit forecasts are calculated for each predictive model, either a long or short position is taken in the SPX,[33] depending on the model forecast being up or down for the next month. Then the performance of each portfolio is calculated based on each model's results.

[31] https://www.federalreserve.gov/econresdata/workingpapers.htm

[32] Similar to Mascio and Fabozzi (2019), we tested the following estimation windows: 24, 48, 72, 96, 120, during the sample period of 351 months. The criterion was based on accuracy of the predication, mean monthly forecasted returns and statistical significance (p-value). All results are available upon request.

[33] Trading costs are ignored for all timing models. Discount broker Charles Schwab offers trading accounts with no transaction fees (see https://www.schwab.com/onesource).

4.5 | The kitchen sink model

The forecasting literature strongly suggests that a combination of the forecasts of the five models will outperform individual forecasts. Hence a kitchen sink variable, denoted ALL, is created by combining the five forecasts using the following multivariate logistic regression:

$$\log \frac{p(x_{t+1} = 1)}{1 - p(x_{t+1} = 1)} = a + \sum_{k=1}^{5} b_k F_{k,t}. \tag{4}$$

Once these results are calculated, the long (short) positions are taken in the SPX and performance is recorded.

4.6 | The LASSO model

Tibshirani (1996) introduced LASSO regression procedures to increase prediction accuracy and to ease interpretation issues. This machine learning procedure is designed to identify the most important predictor variables from a larger set of candidate variables. LASSO performs covariate selection related to stepwise regression and shrinks large coefficients associated with a ridge regression. The LASSO selection criteria model combines ordinary least squares (OLS) estimation procedures with a penalty function. LASSO shrinks the sum of the coefficients, and reduces other less significant coefficients to 0. Elliot and Timmermann (2016) emphasized that economic forecasting is a decision problem and the loss function should be linked to the economic costs of prediction errors.

The OLS regression assumes the observations are independent or conditionally independent given a set of predictors. Tibshirani (1996) built the model based on a given a set of data $(x_i, y_i), i = 1, 2, \dots, N$, where $x_i = (x_n, \dots, x_{ip})^T$ represent the forecast variables and y_i are the responses. Assume x_{ij} are standardized where $\frac{\Sigma_i x_{ij}}{N} = 0$, and $\frac{\Sigma_i x_{ij}^2}{N} = 1$. The coefficients are represented by $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$. The solution to the following model is a *quadratic programming problem* that demonstrates linear inequality constraints:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left\{ \sum_{i=1}^{N} \left(y - i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} + \lambda_i \sum_{j=1}^{N} \hat{w}_{i,j} |b_{i,j}|, \tag{5}$$

where $\lambda_i \sum_{j=1}^{N} \hat{w}_{i,j} |b_{i,j}|$ is the penalty function associated with the LASSO parameter λ_i,[34] $t \geq 0$ is the tuning parameter subject to $\sum_j |\beta_j| \leq t$. Generality is not lost if $\bar{y} = 0$ when α is deleted, and t is solved when α is $\hat{\alpha} = \bar{y}$. The overall solution is determined when $t \geq 0$ controls the

[34] LASSO parameter is selected based on a cross-validation methodology.

degree of shrinkage applied to the estimates. If $\hat{\beta}_j^o$ is the least square estimate and $t_0 = \Sigma|\hat{\beta}_j^o|$, when $t < t_0$ the solutions will decrease towards 0, and in some instances the coefficients will be equal to 0.

4.7 | Certainty equivalents from logistic regression

We use the certainty equivalent (CEQ) to indicate investors' risk/return preferences. To calculate the CEQ, we use the methodology of DeMiguel et al. (2009). This method determines the tradeoff between an investor risk tolerance towards investing in stocks or the risk-free asset. We use the S&P 500 Index (SPX) as a comparison portfolio to the sentiment index, the kitchen sink model, and the LASSO model to determine the economic significance of the expected returns. A mean–variance investor that is assigned a risk aversion of $\gamma = 1, 2, 3, 4, 5$ can choose among a set of portfolios that are entirely invested in stocks based on their return and variance expectations.

The risk aversion of investors will determine their peak utility of each forecast variable in the out-of-sample period. We compute the CEQ return for each k strategy:

$$\widehat{\mathrm{CEQ}} = \hat{\mu}_k - \frac{\gamma}{2}\hat{\sigma}_k^2, \tag{6}$$

where $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are, respectively, the out-of-sample mean excess performance and volatility for each strategy k, and γ is an investor's risk aversion. To determine if the CEQ returns are statistically different from the SPX returns, we use the Sharpe ratio test for equality developed by Lo (2002).[35] This test calculates the difference between each model's Sharpe ratio and the S&P 500 Index. When the p-value exceeds 0.050, we accept the null hypothesis that the Sharpe ratios are not statistically different.

5 | EMPIRICAL RESULTS

5.1 | Descriptive statistics and performance of predictors

As explained in Section 4, each of the five indices is used to predict the direction of the market 1-month ahead. The forecasts from the five models are combined in an equally weighted model, ALL and a LASSO model. Table 1 presents the correlations between the SPX, the Up movement, and the three models. As can be seen, the correlation between the Up movement and the 1-month-ahead returns SPX1 is 0.763, suggesting the SPX has positive returns 76% of the time. The SI index and the LASSO model has a correlation of 0.224 and 0.094, respectively, with the SPX1. Although not reported here, it should be noted that the correlations between the individual forecasting models and the equally weighted ALL index model are greater than or equal to 0.38.

TABLE 1 Correlation matrix of monthly return predictions

	SPX1	UP	SI	ALL	LASSO
SPX1	1.000				
UP	0.763	1.000			
SI	0.224	0.069	1.000		
ALL	0.011	0.004	0.235	1.000	
LASSO	0.094	0.051	0.256	0.316	1.000

This table illustrates the correlations from 01/1985 to 04/2014 between the 1-month-ahead S&P 500 Index (SPX1) return and the UP movement of the S&P 500 Index. The combination of all five forecasting indices is represented as the ALL and LASSO models. SI (Huang et al. 2015) is the Sentiment Index, which is the best performer of the five forecasting models.

Annual holding period returns are presented in Table 2. In a number of years the returns of each index are identical to the SPX returns. In 11 of the years the holding period returns for all three models were the same as the SPX. For example, in 1993 the SPX return was 9.8%. All of the models generated identical returns because when the model predicts an Up month all indices returns will be the same. However, there are instances when the models substantially outperformed or underperformed the SPX. For example, in 1989, the SPX generated returns of 10.6% while the ALL model would have realized a loss of 16.2%. In years where the SPX exhibits substantial losses, sometimes the models perform better and some worse than the SPX. In 2001 and 2002, the SPX had losses of 17.3% and 24.3%, respectively. The LASSO index showed a small loss of 1.7% in 2001 but a substantial gain of 26.3% in 2002. The following year, 2003, the SPX gained 32.2% while the LASSO model generated a loss of 5.7%. Interestingly, in that year the ALL model generated returns equal to the SPX. In 2008 and 2009, the SPX generated a loss of 40.1% and then a gain of 30%. In these years LASSO generated 10.2% and an impressive 74.6% return.

Overall, the SI matched the SPX in 23 of the 27 years and outperformed (underperformed) in three (one) of the years. The ALL and LASSO models were the same as the SPX in 15 and 18 years, respectively. The ALL model was better in five and worse in seven of the years. The LASSO outperformed the SPX in 4 years and underperformed in 5 years. Of the three models, the LASSO was the superior model in forecasting the negative return periods.

[35]The statistical test calculates a hypothesis test between Sharpe ratio pairs of given assets.

TABLE 2 Annual holding period returns for each forecasting model

	SI	ALL	LASSO	SPX
1988	0.149	0.061	0.149	0.149
1989	0.106	−0.162	0.106	0.106
1990	0.045	0.197	0.045	0.045
1991	0.189	0.039	0.086	0.189
1992	0.073	−0.008	0.073	0.073
1993	0.098	0.098	0.098	0.098
1994	−0.023	−0.023	−0.023	−0.023
1995	0.352	0.352	0.352	0.352
1996	0.236	0.236	0.236	0.236
1997	0.247	0.140	0.247	0.247
1998	0.305	0.545	0.305	0.305
1999	0.090	0.090	0.090	0.090
2000	−0.061	−0.104	−0.020	−0.020
2001	−0.039	−0.073	−0.017	−0.173
2002	−0.243	−0.243	0.263	−0.243
2003	0.322	0.322	−0.057	0.322
2004	0.044	0.044	0.044	0.044
2005	0.084	0.084	0.084	0.084
2006	0.124	0.124	0.124	0.124
2007	−0.042	−0.042	−0.042	−0.042
2008	0.146	0.368	0.102	−0.401
2009	0.300	0.017	0.746	0.300
2010	0.370	0.423	−0.084	0.198
2011	0.020	0.020	0.020	0.020
2012	0.141	0.141	0.135	0.141
2013	0.190	0.190	0.190	0.190
2014	0.050	0.050	−0.036	0.050

This table illustrates annual holding period returns from 01/1988 to 04/2014) for the S&P 500 Index (SPX), the LASSO selection criteria model. The combination of all five forecasting indices is represented as the ALL model. SI (Huang et al. 2015) is the Sentiment Index.

TABLE 3 Summary statistics of monthly portfolio returns

	SI	ALL	LASSO	SPX
Minimum	−0.1458	−0.1100	−0.1458	−0.1694
Quartile 1	−0.0168	−0.0189	−0.0175	−0.0179
Median	0.0110	0.0101	0.0117	0.0112
Arithmetic mean	0.0091	0.0086	0.0098	0.0073
Geometric mean	0.0082	0.0077	0.0089	0.0064
Quartile 3	0.0360	0.0348	0.0351	0.0348
Maximum	0.1694	0.1694	0.1694	0.1116
SD	0.0415	0.0419	0.0416	0.0421
Skewness	−0.1710	0.1255	−0.0875	−0.6024
Kurtosis	0.9803	0.6813	0.8901	1.1801

This table illustrates the summary statistics and return distributions using monthly holding period returns on an annual basis from 01/1988 to 04/2014. The buy and hold portfolio is represented by the S&P 500 Index (SPX), and the model selection portfolio is the LASSO model (LASSO). SI (Huang et al. 2015) is the Sentiment Index. The combination of all five forecasting indices is represented as the ALL model.

5.2 | Logistic regression results and statistical tests

The summary statistics for monthly portfolio returns for the three models are presented in Table 3. It can be seen that the minimum return for the SI and LASSO was −14.58%; the SPX had a loss of 16.94% and the best performer was the equally weighted ALL model, with a loss of 11.0%. The LASSO model generates greater returns than both of the other models and the SPX with higher median, arithmetic, and geometric mean returns of 1.17%, 0.98%, and 0.89%, respectively. The standard deviations for all of the models are similar at approximately 4.2%. The skewness value for the LASSO model, at −0.0875, is much better than the SPX at −0.6024. This suggests the LASSO model results are reasonably symmetric while the SPX exhibits moderate skewness. Interestingly, the ALL model exhibits a positive skew of 0.1255 that is also reasonably symmetric. The kurtosis values are approximately 1.0 for the SI, SPX, and the LASSO, but for the ALL model it is 0.6813. Hence all distributions are less peaked than a normal distribution.

A number of tests are used to provide insights into each forecast model's probability distribution. Results from the Shapiro–Wilks test[36] for normality as shown in Figure 1[37] show none of the models' distributions are Gaussian in nature and the Bartlett test[38] indicates that the models' variances are not equal. Interestingly, however, the nonparametric Kruskal–Wallis[39] test reveals that we cannot reject the null hypothesis that our probability estimates are from the same distribution. Finally, the nonparametric Dunn test[40] shows that the null hypothesis of equal probabilities cannot be rejected.

5.3 | Forecasting errors and model performance

Monthly forecasting errors are shown in Table 4. The objective of the forecasting models is to correctly predict the direction of the 1-month-ahead SPX1: an UP (DOWN) forecast indicates that a positive (negative) return is expected. First, consider the two rows at the bottom of the table. These two rows show the number of incorrect and correct forecasts for the 312 months of this study. It is seen that the LASSO and SI models correctly predict the SPX's market direction in 160 and 161 of the months, while

[36]The Shapiro and Wilk (1965) test statistic is given as $W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$, where x is the sample mean and a_i are the constants.

[37]Figure 1 illustrates a Q-Q plot of each forecast index's probability estimates.

[38]Snedecor and Cochran (1989) updated the original Bartlett test in their book *Statistical Methods*.

[39]Kruskal–Wallis (1952) is a rank sum test of all estimates from 1 to N. The test will result in an X^2 value, and does not determine a relationship between each estimate in a pairwise order.

[40]See Dunn (1961).

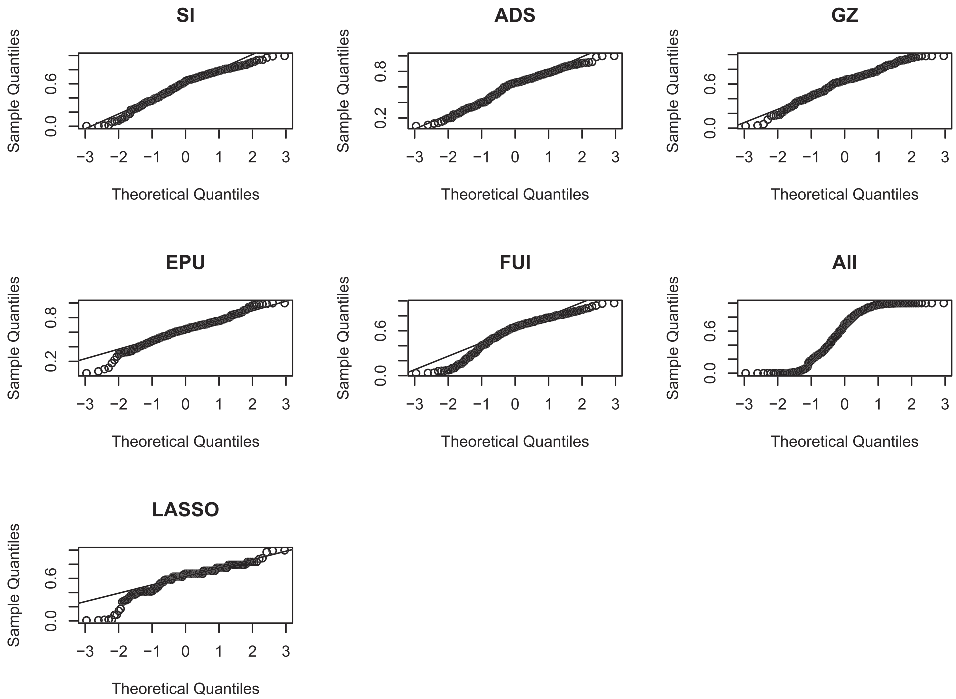

FIGURE 1 Normality test of the probability estimates. This figure represents the results of the Shapiro normality tests using ggplots on the probability estimates of the five forecast model portfolios, and the combined indices from 01/1985 to 04/2014. SI, Huang, Jiang, Tu, and Zhou Sentiment Index; ADS, Arubold–Diebold–Scotti Business Conditions Index; GZ, Gilchrist–Zakrajsek Credit Spread Index; EPU, Baker, Bloom, and Davis Economic Policy Uncertainty Index; FUI, Jurado, Ludvigson, and Ng Financial Uncertainty Index. The combination of all five forecasting indices kitchen sink is represented as the ALL model, and the model selection index (LASSO). All estimates are calculated using a 24-month rolling window

the ALL model lags with 153 correct predictions. The values in the columns represent the errors associated with the type of error. ER1 represents an error of predicting the market will go down but the market goes up. ER2 is the reverse; that is, predicting the market will go up but the market actually goes down. For example, consider the year 2000. There are four months when the market went up and eight months when the market went down. The ALL model mispredicts three of the four up months and seven of the eight down months. The SI mispredicts three of the up months and six of the down months. The LASSO model misses all of the down months. The total row at the bottom of the table shows the number of incorrect predictions overall, revealing that the ALL model index and the LASSO index mispredicted fewer of the down markets than the SI model.

Table 5 presents the annualized performance analysis. It is seen that the SPX generated an annual return of 7.9%. All three models generated higher returns, with LASSO as the best performer, with an 11.3% annual return.

The standard deviations for all three models and the SPX are similar at approximately 14.5%. LASSO is the best performer when considering three annualized return/risk measures: the Sharpe ratio, the Sortino ratio, and the omega ratio. All three models demonstrate slightly lower loss deviations compared to the SPX. The maximum drawdown measure shows that all three models outperform the SPX. In this case, the LASSO model reduces drawdown to less than one-half of the drawdown of the SPX, SI, and ALL. LASSO at 0.230 has the lowest drawndown and down capture compared to the other models.

5.4 | Statistical significance of model probability estimates

Table 6 illustrates the results of monthly logistic regressions using the maximum likelihood approach. The observation period is between 01/1985 and 04/2014. All reported statistics are on a rolling 24-month estimation window. The first columns display each coefficient (β %), which results in a binary forecast of the S&P 500 Index (SPX) as Up or Down in the following month. It should be remembered that an Up month is predicted when $p(x = 1) \geq 0.5$, and a Down month is predicted when ($p(x = 1) < 0.5$). If $p(x) \geq 0.5$, we forecast $x = 1$; if $p(x) < 0.5$, $x = 0$ is predicted. The middle columns show the Newey–West t-statistics, and the far right columns are the McFadden pseudo-R^2.

According to Hastie et al. (2015), typical significance tests assume that the null is a random variable with little relationship on the final outputs. Therefore, applying a traditional significance test to the LASSO model would result in biased outcomes, because LASSO already selects the best variables from the penalty function. However, there are some recent papers that have developed methodologies to extract significance of the variables within the model selection framework (see, e.g., Hastie et al., 2015).

Each β coefficient is a forecast of the next month's market (SPX) direction. Of the 27 years in the sample period there are 11 years in which the forecast indices differ in their prediction. For example, in 1990 the LASSO (0.471) model correctly forecasts a Down market for the year,

TABLE 4 Model monthly forecasting errors

Year	SPX		ALL		LASSO		SI	
	UP	DOWN	ER1	ER2	ER1	ER2	ER1	ER2
1988	6	3	1	3	0	3	0	3
1989	8	4	2	4	0	4	0	4
1990	5	7	2	4	0	7	0	7
1991	9	3	1	3	1	3	0	3
1992	8	4	1	4	0	4	0	4
1993	8	4	0	4	0	4	0	4
1994	7	5	0	5	0	5	0	5
1995	10	2	0	2	0	2	0	2
1996	10	2	0	2	0	2	0	2
1997	9	3	1	3	0	3	0	3
1998	9	3	1	2	0	3	0	3
1999	7	5	0	5	0	5	0	5
2000	4	8	3	7	0	8	3	6
2001	6	6	3	4	4	1	3	4
2002	4	8	0	8	4	0	0	8
2003	9	3	0	3	6	1	0	3
2004	9	3	0	3	0	3	0	3
2005	5	7	0	7	0	7	0	7
2006	11	1	0	1	0	1	0	1
2007	7	5	0	5	0	5	0	5
2008	4	8	2	4	4	2	1	6
2009	9	3	4	1	0	2	0	2
2010	7	5	2	3	3	4	1	4
2011	5	7	0	7	0	7	0	7
2012	9	3	0	3	1	3	0	3
2013	10	2	0	2	0	2	0	2
2014	2	1	0	1	1	1	0	1
Total	197	115	23	100	24	92	8	107
Incorrect			123		116		115	
Correct			153		160		161	

This table summarizes each model's monthly forecasting errors, on an annual basis. The buy and hold portfolio is represented by the S&P 500 Index (SPX), and the model selection portfolio is the LASSO model (LASSO). SI (Huang et al. 2015) is the Sentiment Index. The combination of all five forecasting indices is represented as the ALL model. Each forecasting portfolio, with the exception of the SPX, has two forecasting error columns: ER1 and ER2. ER1 refers to the error of predicting that the market would go down, but the market went up: Market Up–Predicted Down. ER2 represents when the prediction was for an up market, but the market declined: Market Down–Predicted Up. The bottom two rows illustrate the total "Incorrect" and "Correct" forecasting errors.

while the SI and ALL models forecast an Up market. Overall, ignoring statistical significance here, the models are predicting an upward trend for the market with probabilities greater than 0.50 for 144 of the possible 189 forecast-months. The SI predicts 9 down years, while the ALL and LASSO models predict 5 and 6 down years, respectively. All models correctly anticipate the bursting of the tech bubble in 2000 with probabilities below 0.50. Both the SI and LASSO correctly predicted the 2008 down year, but, unexpectedly, the ALL model predicted an up year for 2008.

TABLE 5 Annualized performance analysis

	SI	ALL	LASSO	SPX
Annualized return	0.110	0.097	0.113	0.079
Annualized SD	0.144	0.145	0.144	0.146
Annualized Sharpe (Rf = 0%)	0.779	0.667	0.783	0.542
Sortino ratio (MAR = 0%)	0.389	0.345	0.393	0.255
Omega ($L = 0\%$)	1.829	1.694	1.834	1.557
Tracking error	0.032	0.049	0.044	0.000
Annualized tracking error	0.110	0.170	0.153	0.000
Information ratio	0.341	0.099	0.104	0.220
Semi deviation	0.030	0.029	0.030	0.032
Gain deviation	0.027	0.028	0.027	0.024
Loss deviation	0.027	0.024	0.026	0.031
Maximum drawdown	0.404	0.432	0.230	0.526
Up capture	0.928	0.685	0.732	1.000
Down capture	0.670	0.408	0.387	1.000

This table summarizes annualized performance and risk statistics for the buy and hold portfolio, represented by the S&P 500 Index (SPX); the model selection portfolio is the LASSO model (LASSO), and the combination of all five forecasting indices is represented as the ALL model. SI (Huang et al. 2015) is the Sentiment Index.

When statistical significance is considered, the ALL model is clearly the better performer, with six coefficients that are statistically different from zero with t-statistics greater than 2.00. The SI only records one year that is statistically significant.

The pseudo-R^2 values provide similar results. The average pseudo-R^2 for the ALL model (0.78) is substantially greater than the SI model. In addition, it has the greatest maximum value, at 0.89, and the greatest minimum value, at 0.56.

5.5 | Prediction accuracy and significance of the combined model

The primary focus of this section is to discuss the forecast ability of the SI, ALL, and LASSO models. Four different time periods are used to evaluate the statistical significance of the "kitchen sink" (ALL) and the LASSO models for predicting the 1-month-ahead returns of the S&P 500 Index. The first is the 3-month period before and after the 1987 stock market crash. The second is the recessionary period (2000–2003) following the dot.com bubble and the 9/11 terrorist attacks. Third is the financial crisis and recession between 2008 and 2009, and, lastly, the recovery period from 2011 to 2014.

In our sample period, the worst month of returns for the S&P 500 Index was October 1987, at −21.73%. The ALL model forecasted correctly this large market drawdown with the probability of an Up market at only 0.00505. Conversely, the LASSO model during the same period forecasts the probability of an Up market at 75%—an inaccurate prediction.

TABLE 6 Predictive logistic regression estimation results

	β (%)			t-statistic		Pseudo-R^2	
	SI	ALL	LASSO	SI	ALL	SI	ALL
1987	0.364	0.942	0.667	1.000	1.945	0.618	0.605
1988	0.644	0.758	0.625	0.184	1.693	0.295	0.556
1989	0.695	0.892	0.667	−0.897	1.603	0.416	0.817
1990	0.800	0.897	0.471	−1.265	2.139	0.643	0.825
1991	0.252	0.974	0.688	−1.242	2.295	0.665	0.857
1992	0.727	0.820	0.708	−1.504	1.780	0.488	0.681
1993	0.680	0.884	0.667	−0.305	1.429	0.381	0.802
1994	0.582	0.877	0.625	0.332	1.866	0.131	0.789
1995	0.662	0.983	0.708	−0.962	1.885	0.338	0.873
1996	0.795	1.000	0.875	−0.399	1.989	0.631	0.852
1997	0.814	0.844	0.792	−0.168	2.024	0.670	0.728
1998	0.777	1.000	0.708	−0.195	1.363	0.596	0.888
1999	0.858	1.000	0.667	0.452	1.980	0.754	0.749
2000	0.363	0.320	0.302	−0.688	1.422	0.626	0.825
2001	0.546	0.193	0.417	−0.604	1.927	0.031	0.755
2002	0.411	0.539	0.494	0.058	2.106	0.424	0.789
2003	0.788	0.043	0.581	−2.624	2.228	0.617	0.780
2004	0.970	0.994	0.708	−1.783	1.932	0.685	0.790
2005	0.467	0.335	0.625	1.171	1.684	0.221	0.752
2006	0.570	0.996	0.667	−0.957	1.927	0.097	0.779
2007	0.176	0.783	0.750	−1.900	2.099	0.522	0.609
2008	0.231	0.903	0.199	−1.154	1.879	0.457	0.836
2009	0.448	0.041	0.021	−1.606	1.657	0.632	0.874
2010	0.940	0.961	0.667	−1.796	1.638	0.600	0.836
2011	0.455	0.990	0.839	0.341	1.555	0.262	0.883
2012	0.546	0.997	0.583	−0.501	1.916	0.030	0.795
2013	0.838	0.979	0.792	−0.718	1.743	0.716	0.865

This table illustrates each coefficient (%), Newey–West t-statistic, and pseudo (McFadden) R^2 associated with the logistic regression model using a maximum likelihood approach following the same methodology of Mascio and Fabozzi (2019). A kitchen sink variable, denoted by ALL, is created by combining the five forecasting models using the following multivariate logistic regression:

$$\log \frac{p(x_{t+1}=1)}{1-p(x_{t+1}=1)} = a + \sum_{k=1}^{5} b_k F_{k,t}.$$

Each observation is the last month (December) of each year during the forecasting period (01/1985–04/2014). These results are based on the combined prediction of the five forecasting model portfolios from Mascio and Fabozzi (2019). SI (Huang et al., 2015) is the Sentiment Index. The LASSO model is denoted *LASSO*; its estimates are only shown under column 3 in the β (%) reporting section. All statistics are estimated using a 24-month rolling estimation window (t-statistics above ±2.00, 2.50, and 3.00 are significant at the 10%, 5%, and 1% levels, respectively).

Both the ALL and LASSO models do especially well at forecasting the returns of the S&P 500 Index between the period of March 2000 and February 2003. During this period, which includes the recession of 2001, the ALL and LASSO models have a mean monthly forecast of an Up market at 41.15% and 41.87%, respectively. They accurately forecast the negative monthly returns of the SPX nearly 75% of the time. During this 35-month period, the SPX had 10 months of negative returns greater than 5.00%.[41] The LASSO model predicted all 10 negative months accurately, with the exception of August 2002, while the ALL model correctly forecasted six of the 10 months.

May 2008 to February 2009 represents the worst 10 months of consecutive return for the SPX during our sample period. The highest (lowest) monthly return for the period was realized in July 2008 (September 2008) of 1.21% (−16.9%). The total drawdown of the SPX was −60.6%. Both the ALL and the LASSO model accurately predicted each negative monthly return during the period, with the exception of September 2008, when the ALL model forecasted an Up month. The LASSO (ALL) portfolios had a mean monthly probability of an Up market of 41.19% (19.95%), resulting in a 85% prediction accuracy for both models.

In the recovery following the financial crisis, the SPX total return between 9/2011 and 3/2014 was 52.89%. The LASSO (ALL) model correctly forecasted the monthly market return 76.01% (61.29%) of the time. From November 2011 to February 2012, both the LASSO and the ALL model accurately predicted the 12.41% return of the SPX each month during the period.

Another noteworthy time period is between March 2012 and August 2013, when the SPX was up 15.94%. Once again, the ALL and the LASSO models correctly predicted the monthly returns on the market more than 65% of the time.

5.6 | Persistency of the estimates using autocorrelation

Determining the persistency of the probability estimates follows a similar autoregressive time series (AR) process to Mascio and Fabozzi (2019). The reported statistics have autocorrelation monthly lags ranging from 1 and 20. We use an AR(1) model to compare the probability estimates of the SI, the kitchen sink (ALL) model, and the LASSO model. The first-order autocorrelation model is represented as a standard linear difference equation:

$$X_t = \rho X_{t-1} + \epsilon, \tag{7}$$

where $t = 0, 1, 2, \ldots$ and ϵ_t is the error term computed from the time series. The difference equation relates X_t, which is the original value at time t, and a lagged parameter at X_{t-1}.

Figure 2 presents three probability estimates of the autocorrelation function (acf) for the predictive models. The plots display the lag periods in months on the horizontal axis and the acf on the vertical axis using an AR(1) process.

[41] Top ten negative SPX return months from March 2000 to February 2003; 10/2000 (−8.00%), 2/2001 (−9.22%), 7/2001 (−6.41%), 8/2001 (−8.17%), 3/2002 (−6.14%), 5/2002 (−7.24%), 6/2002 (−7.90%), 8/2002 (−11.00%).

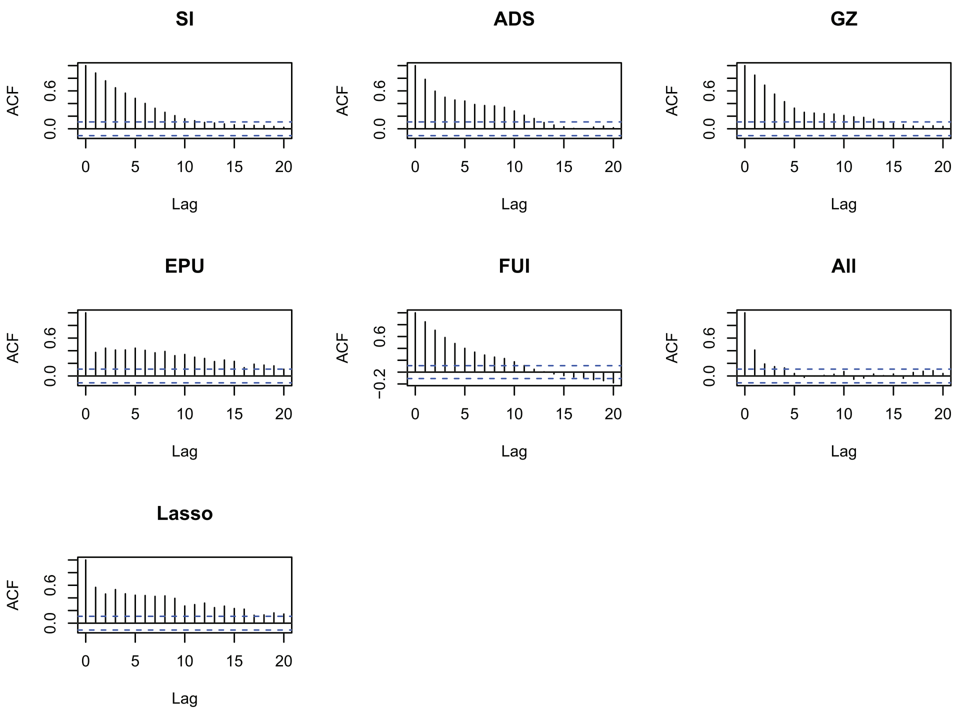

FIGURE 2 Persistence of probability forecast using autocorrelation. The six plots represent the persistence of probability forecasts among the five forecast model portfolios, and the benchmark model from 01/1983 to 04/2014. SI, Huang, Jiang, Tu, and Zhou Sentiment Index; ADS, Arubold–Diebold–Scotti Business Conditions Index; GZ, Gilchrist–Zakrajsek Credit Spread Index; EPU, Baker, Bloom, Davis Economic Policy Index; FUI, Jurado, Ludvigson, and Ng Financial Uncertainty Index; "kitchen sink" (ALL) model; and LASSO model selection index. Horizontal dashes represent the 95% confidence interval. In addition, each plot illustrates the autocorrelation corresponding to each monthly lag based on the following AR(1) model: $X_k = \rho X_{k-1} + \epsilon$ where $k = [0, 1, 2, \ldots, 20]$ [Colour figure can be viewed at wileyonlinelibrary.com]

Both of the forecast models show significant persistency of the probability estimates, but differ in the longevity of the lag period of persistency. The LASSO model exhibits persistency over the entire 20-month period, while the SI and ALL models are persistent for 12 and 4 months, respectively.

5.7 | Forecast portfolio rankings

Table 7 compares the SPX with the ALL and LASSO forecasting models based on the mean annual return over the entire sample period (1988–2014). The models have nearly identical standard deviations, but the LASSO model has a greater average return and a higher Sharpe ratio. The drawdown associated with the LASSO model is approximately one half the drawdown of the other two models and less than one half that of the SPX.

TABLE 7 Annualized return rankings among all strategies

	Return	SD	Sharpe	Drawdown
LASSO	0.113	0.144	0.783	0.230
SI	0.110	0.144	0.779	0.404
ALL	0.097	0.145	0.667	0.432
SPX	0.079	0.146	0.542	0.526

This table ranks each individual forecasting strategy based on its mean annual return between 01/1988 and 04/2014. These results are based on the prediction of the combination of the forecasting models and the model selection strategy. The model selection portfolio is the LASSO and the combination of all five forecasting models is represented as the kitchen sink portfolio (ALL). SI (Huang et al. 2015) is the Sentiment Index. The buy and hold portfolio is the S&P 500 Index (SPX).

5.8 | CEQ performance results

The results of the CEQ returns are reported in Table 8. Similar to Mascio and Fabozzi (2019), the CEQ performance results are consistent with the geometric returns of each forecasting portfolio. Each column represents an investor's risk aversion $\gamma = 1, 2, 3, 4, 5$. The first shows the CEQ returns of the S&P 500 Index (SPX), and below that are displayed the CEQ monthly performance of all three forecast models. Under each model's CEQ returns, in parentheses, is the p-value associated with the Sharpe ratio equality test with the SPX. The base CEQ model is calculated as a result of the logistic regression, with a probability threshold of 0.50.

When an investor's risk aversion is $\gamma = 1$, which is the base case for DeMiguel et al. (2009), we find that the CEQ return for the SPX (0.00350) is lower than all three forecast strategies. The SI and ALL models slightly outperform the SPX, and the LASSO model has substantially higher

CEQ returns and its p-value suggest, that the Sharpe ratio is different from the SPX. For $\gamma = 2, \ldots, 5$, LASSO exhibits higher returns than all three models including the SPX. In addition, the p-value for LASSO is below 0.050 in all cases.

Lastly, to test the robustness of the logistic regression probability cutoff, the cutoff rate is changed from 0.50 to 0.75. The results are also shown in Table 8. At all levels of risk aversion, $\gamma = 1, 2, 3, 4, 5$, we find that LASSO outperforms all three models including the SPX, although only two cases are significant at the 0.05 level. None of the coefficients for the SI and ALL models are significant, and all are negative.

TABLE 8 Certainty equivalent returns for empirical results

	Logistic regression cutoff(0.50)					**Logistic regression cutoff(0.75)**				
	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$
SPX	0.0035	0.00255	0.00159	0.00064	−0.00031	0.00293	0.00196	0.001	0.00003	−0.00093
ALL	0.00366	0.00243	0.00146	0.00049	−0.00049	−0.00083	−0.00181	−0.0028	−0.00379	−0.00477
	(0.45029)	(0.65944)	(0.79871)	(0.89661)	(0.95420)	(0.93508)	(0.97206)	(0.98949)	(0.99655)	(0.99901)
SI	0.00616	0.00522	0.00429	0.00335	0.00242	−0.00347	−0.00446	−0.00545	−0.00643	−0.00742
	(0.00075)	(0.00289)	(0.00958)	(0.02712)	(0.06593)	(0.78215)	(0.88016)	(0.94184)	(0.97518)	(0.99070)
LASSO	0.00772	0.00602	0.00507	0.00411	0.00316	0.00414	0.00318	0.00223	0.00128	0.00032
	(0.00006)	(0.00096)	(0.00365)	(0.01184)	(0.03282)	(0.01050)	(0.02774)	(0.06405)	(0.12968)	(0.23140)

This table illustrates the monthly certainty equivalent (CEQ) performance for the S&P 500 portfolio Index (SPX), the out-of-sample CEQ performance for the five forecast model strategies, and the combined models. We used the same CEQ calculation as DeMiguel et al. (2009):

$$\widehat{\mathrm{CEQ}} = \hat{\mu}_k - \frac{\gamma}{2}\hat{\sigma}_k^2,$$

where $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are mean and variance of the out-of-sample excess returns for each model strategy k and risk aversion γ. The time period is 01/1983–04/2014. These results are based on the out-of-sample prediction of the combined portfolios: the "kitchen sink" (ALL) and the LASSO. SI (Huang et al., 2015) is the Sentiment Index. All statistics are estimated using a 24-month rolling estimation window. p-values are reported in parentheses to represent the difference between the benchmark (SPX) portfolio's Sharpe ratio compared to each forecast index. Any p-value below 0.050 would suggest the Sharpe ratios are different.

6 | CONCLUSION

The objective of this paper is to assess the efficacy of three forecasting models for predicting the 1-month-ahead S&P 500 Index returns. The three models are the sentiment index, a combined *kitchen sink* forecasting model, and a LASSO model. A market timing strategy, beta optimization, is used to change the beta of a portfolio if future market expectations change. Functionally, when the S&P 500 Index is forecasted to be up (down), the portfolio is directed to have a long (short) position.

Since the mid-2000s several indexes have been developed to explain or forecast the overall stock market's direction: Huang et al. (2015) developed an improved sentiment index; Aruoba et al. (2009) developed the ADS Business Conditions Index, designed to measure current business conditions; Baker, Bloom, and Davis (2015) constructed the Economic Policy Uncertainty Index; Gilchrist and Zakrajsek (2012) created the GZ Spread model; and Jurado et al. (2015) constructed a time-varying Financial Uncertainty Index. These five models are combined into an ALL, or kitchen sink, model, as well as in a LASSO model to build either a long or short position in the S&P 500 Index.

A buy-and-hold strategy is used as a benchmark to compare the effectiveness of both the kitchen sink portfolio and LASSO model portfolio at predicting the direction of the market. Over the sample period, the S&P 500 Index was up 197 months and down 115 months. Although the LASSO model was the best at forecasting down months for the S&P 500 Index with 24 correct predictions, more importantly, it was successful at predicting the major market drawdowns during the recession of the early 2000s, and the financial crisis of 2008–2009. Moreover, in volatile markets investors would have benefited more by using the forecasts of the LASSO model to avoid large monthly losses. Also, consistent with prior research, combining the five forecasts improved prediction accuracy, but the timing of the predictions did not produce favorable returns. Conversely, the LASSO model produced the highest average annual returns, the highest certainty equivalent returns, the most statistically significant returns, and the lowest monthly drawdown by nearly 50%.

Active investment managers have the ability to modify their strategies by shifting between high and low beta stocks. Prior research has shown that successful market timing strategies can yield option-like payoffs, but overall volatility risk of these strategies has not been accurately quantified. Even though the beta optimization strategy using the LASSO model has worked well over the examined time period, other machine learning models could prove to be superior at forecasting stock returns.

DATA AVAILABILITY STATEMENT

With the exception of the SI data, all other data that support the findings of this study are available from the first author upon reasonable request.

REFERENCES

Abe, M., & Nakayama, H. (2018). Deep Learning for Forecasting Stock Returns in the Cross-Section. In *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 273–284.

Amenc, N., Curtis, S, & Martellini, L. (2004). *The Alpha and Omega of Hedge Fund Performance Measurement*: EDHEC Risk and Asset Management Research Centre.

Amin, K., Coval, J., & Seyhun, H. (2004). Index Options and Stock Market Momentum. *The Journal of Business*, *77*(4), 835–874.

Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-Time Measurement of Business Conditions. *Journal of Business & Economic Statistics*, *27*(4), 417–427.

Baker, S.R., Bloom, N., & Davis, S.J. (2015) Measuring Economic Policy Uncertainty. NBER Working Paper No. 21633.

Baker, S., Bloom, N., & Davis, S. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1–69.

Baker, M., & Wurgler, J. (2006). Investor Sentiment and the Cross-Section of Stock Returns. *Journal of Finance*, *61*(4), 1645–1680.

Baker, M, & Wurgler, J (2007). Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, *21*(2), 129–151.

Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, Local, and Contagious Investor Sentiment. *Journal of Financial Economics*, *104*(2), 272–287.

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468.

Bollen, N. P. B., & Busse, J. A. (2001). On the Timing of Mutual Fund Managers. *Journal of Finance*, *56*(3), 1075–1094.

Bollen, N.P.B & Busse, J.A. (2005). Short-Term Persistence in Mutual Fund Performance. *Review of Financial Studies*, *18*(2), 569–597.

Brown, G. W., & Cliff, M. T. (2004). Investor Sentiment and the Near-Term Stock Market. *Journal of Empirical Finance*, *11*(1), 1–27, available at 9809069v1

Carhart, M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance*, *52*(1), 57–82.

Cremers, K. J. M., & Petajisto, A. (2009). How Active Is Your Fund Manager A New Measure That Predicts Performance. *Review of Financial Studies*, *22*(9), 3329–3365.

Da, Z., Engelberg, J., & Gao, P. (2015). The Sum of All FEARS: Investor Sentiment and Asset Prices. *Review of Financial Studies*, *28*(1), 1–40.

Droms, W. G. (1989). Market Timing as an Investment Policy. *Financial Analysts Journal*, *45*(1), 73–77.

Dunn, O. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, *56*, 52–64.

Elliot, G., & Timmermann, A. (2016). *Economic Forecasting*. Princeton, NJ: Princeton University Press.

Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete Subset Regressions. *Journal of Econometrics*, *177*(2), 357–373.

Elliott, G., & Timmermann, A. G. (2008). Economic Forecasting. Journal of Economic Literature. *Journal of Economic Literature*, *46*(1), 3–56.

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The Adjustment of Stock Prices To New Information. *International Economic Review*, *10*(1), 1–21.

Fisher, K. L., & Statman, M. (2003). Consumer Confidence and Stock Returns. *The Journal of Portfolio Management*, *30*(1), 115–127.

Fung, W., & Hsieh, D. A. (2001). The Risk in Hedge Fund Strategies: Theory and Evidence From Trend Followers. *Review of Financial Studies*, *14*(2), 313–341.

Gasbarro, D, Cullen, G, Le, K, & Monroe, G (2015). Mutual Fund Trades: Timing Sentiment and Managing Tracking Error Variance. *SSRN Electronic Journal*.

Genre, V., Kenny, G., Meyler, A., & Timmermann, A. G. (2013). Combining Expert Forecasts: Can Anything Beat the Simple Average?. *International Journal of Forecasting*, *29*(1), 108–121.

Gilchrist, S., & Zakrajsek, E. (2012). Credit Spreads and Business Cycles. *American Economic Review*, *102*, 1692–1720.

Gilchrist, S., Zakrajsek, E., Favara, G., & Lewis, K. (2016). Recession Risk and the Excess Bond Premium. *Fed Notes*, *April*(8), 1–3.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: CRC Press, Champman and Hall.

Henriksson, R. D., & Merton, R. C. (1981). On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills. *The Journal of Business*, *54*(4), 513–533.

Huang, D., Jiang, F., Tu, J., & Zhou, G. (2015). Investor Sentiment Aligned: A Powerful Predictor of Stock Returns. *Review of Financial Studies*, *28*(3), 791–837.

Jagannathan, R., & Korajczyk, R. A. (2017). Market Timing. In *Portfolio Construction, Measurement, and Efficiency: Essays in Honor of Jack Treynor* (pp. 49–71). Springer International Publishing.

Jiang, G. J., Yao, T., & Yu, T. (2007). Do Mutual Funds Time the Market? Evidence from Portfolio Holdings. *Journal of Financial Economics*, *86*(3), 724–758.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring Uncertainty. *American Economic Review*, *105*(3), 1177–1216.

Kacperczyk, M., Nieuwerburgh, S. V., & Veldkamp, L. (2014). Time-Varying Fund Manager Skill. *Journal of Finance*, *69*(4), 1455–1484.

Kim, O., Lim, S., & Shaw, K. (2001). The Inefficiency of the Mean Analyst Forecast as a Summary Forecast of Earnings. *Journal of Accounting Research*, *39*(2), 329–335.

Lam, K., & Li, W. (2004). Is the Perfect Timing Strategy Truly Perfect?. *Review of Quantitative Finance and Accounting*, *22*(1), 39–51.

Lettau, M., & Ludvigson, S. (2001). Consumption, Aggregate Wealth, and Expected Stock Returns. *Journal of Finance*, *56*(3), 815–849.

Lo, A. (2002). The Statistics of Sharpe Ratios. *Financial Analysts Journal*, *58*(4), 36–52.

Lobo, G. J. (1991). Alternative Methods of Combining Security Analysts' and Statistical Forecasts of Annual Corporate Earnings. *Journal of Forecasting*, *7*(1), 57–63.

Ludvigson, S. C., & Ng, S. (2007). The Empirical Risk-Return Relation: A Factor Analysis Approach. *Journal of Financial Economics*, *83*(1), 171–222.

Makridakis, S., & Hibon, M. (1979). Accuracy of Forecasting: An Empirical Investigation. *Journal of the Royal Statistical Society*, *142*(2), 97–125.

Mascio, D. A., & Fabozzi, F. J. (2019). Sentiment Indices and their Forecasting Ability. *Journal of Forecasting*, *38*(4), 257–276.

Merton, R. C. (1981). On Market Timing and Investment Performance. I. An Equilibrium Theory of Value for Market Forecasts. *Journal of Business*, *54*(3), 363–406.

Phillips, D, & Lee, J (1989). Differentiating Tactical Asset Allocation from Market Timing. *Financial Analysts Journal*, *2*(March-April), 14–16.

Savor, P., & Wilson, M. (2014). Asset Pricing: A Tale of Two Days. *Journal of Financial Economics*, *113*(2), 171–201.

Shapiro, S., & Wilk, M. (1965). An Analysis of Variance Test For Normality (Complete Samples). *Biometrika*, *52*(34), 591–611.

Sharpe, W. F. (1975). Likely Gains from Market Timing. *Financial Analysts Journal*, *31*(2), 60–69.

Snedecor, G., & Cochran, W. (1989). *Statistical Methods* (8th ed.). Iowa City, IA: Iowa State University Press.

Sprothen, V. (2016). Trading Tech Accelerates Toward Speed of Light. *The Wall Street Journal*, October.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288.

Treynor, J, & Mazuy, K (1966). Can Mutual Funds Outguess the Market?. *Harvard Business Review*, *44*(4), 131–136.

Wallis, K. F. (2011). Combining Forecasts: Forty Years Later. *Applied Financial Economics*, *21*(1-2), 33–41.

Wang, Z., & Leung, N. (2018). Factor Selection with Deep Reinforcement Learning for Financial Forecasting. Available at SSRN 3128678.

AUTHOR BIOGRAPHIES

David A. Mascio is Roland and Sarah George Professor of Investments at Stetson University and the founding and managing principal at Della Parola Capital Management in Fort Collins, Colorado. He earned his doctorate in finance at EDHEC Business School in Nice, France in 2018.

Frank J. Fabozzi is a professor of finance at EDHEC Business School in Nice, France, and editor of *The Journal of Portfolio Management* and co-editor and founding editor of *The Journal of Financial Data Science*. He earned a doctorate in economics from the City University of New York in 1972.

J. Kenton Zumwalt is Professor Emeritus at Colorado State University. Dr. Zumwalt has published in a number of leading finance journals. He earned a B.S. in Physics and worked at a nuclear reactor before returning to graduate school, earning an MBA and Ph.D. in Finance at the University of Missouri.

How to cite this article: Mascio DA, Fabozzi FJ, Zumwalt JK. Market timing using combined forecasts and machine learning. *Journal of Forecasting*. 20201–16. https://doi.org/10.1002/for.2690