Macroeconomic Indicators Alone can Predict the Monthly Closing Price of Major U.S. Indices: Insights from Artificial Intelligence, Time-Series Analysis and Hybrid Models

Article *in* Applied Soft Computing · July 2018

DOI: 10.1016/j.asoc.2018.07.024

Macroeconomic Indicators Alone can Predict the Monthly Closing Price of Major U.S. Indices: Insights from Artificial Intelligence, Time-Series Analysis and Hybrid Models

Bin Weng[a], Waldyn Martinez[b], Yao-Te Tsai[c], Chen Li[d], Lin Lu[e], James R. Barth[f], Fadel M. Megahed[g]

[a]*Department of Industrial & Systems Engineering, Auburn University, AL, 36849, USA — Email: bzw0018@auburn.edu*
[b]*Farmer School of Business, Miami University, Oxford, OH 45056, USA — Email: martinwg@miamioh.edu*
[c]*Department of International Business, Feng Chia University, Taiwan 40724, R.O.C — Email: yaottsai@fcu.edu.tw*
[d]*Department of Agricultural Economics, Auburn University, AL, 36849, USA — Email: czl0053@auburn.edu*
[e]*Department of Industrial & Systems Engineering, Auburn University, AL, 36849, USA — Email: lzl0032@auburn.edu*
[f]*Raymond J. Harbert College of Business, Auburn University, AL 36849, USA — Email: barthjr@auburn.edu*
[g]*Corresponding Author. Farmer School of Business, Miami University, OH, 45056, USA — Email: fmegahed@miamioh.edu | Phone: (+1)513-529-4185*

Abstract

This paper proposes a two-stage approach that can be used to investigate whether the information hidden in macroeconomic variables (alone) can be used to accurately predict the one-month ahead price for major U.S stock and sector indices. Stage 1 is constructed to evaluate the hypothesis that the price for different indices is driven by different economic indicators. It consists of three phases. In phase I, the data is automatically acquired using freely available APIs (application programming interfaces) and prepared for analysis. Phase II reduces the set of potential predictors without the loss of information through several variable selection methods. The third phase employs four ensemble models and three time-series models for prediction. The prediction performance of the seven models are compared using the Mean Absolute Percent Error (and two additional metrics). If the hypothesis were to be true, one expects that the performance of the ensemble models to outperform the time-series models since the information in the economy is more important than the information in previous prices. In Stage 2, a hybrid approach of the recurring neural network used for time-series prediction (i.e. the LSTM) and the ensemble models is constructed to examine the secondary hypothesis that the residuals from the time-series models are not random and can be explained by the macroeconomic indicators. To test the two hypotheses, the monthly closing prices for 13 U.S. stock and sector indices and the corresponding values for 23 macroeconomic indicators were collected from 01/1992 10/2016. Based on the case study, the four ensembles prediction performance were superior to that of the three time-series models. The MAPE of the best model for a given index was $< 1.87\%$. The Stage 2 results also show that the three evaluation metrics (RMSE, MAPE and MAE) can be typically improved by 25-50% by incorporating the information hidden in the macroeconomic indicators (through the ensemble approach). Thus, this paper shows that, for the analysis period and the indices studied, the macro-economic indicators are leading predictors of the price of 13 U.S. sector indices.

Keywords: ARIMA, Deep Learning, Ensembles, GARCH, Long Short-Term Memory (LSTM) Networks

1. INTRODUCTION

The prediction of stock prices has continued to fascinate both academia and business. The question driving early stock market research was: "To what extent can the past history of a common stock's price be used to make meaningful predictions concerning the future price of the stock?" [1]. The *Efficient Market Hypothesis* (EMH) [1] and the *Random Walk Theory* [2] provided a theoretical foundation for tackling this question. These models posited that stock prices cannot be forecasted since they are driven by new information and not present/past prices. Thus, prices will follow a random walk and cannot be predicted accurately.

There has been an increasing number of studies [3, 4, 5, 6, 7] that provide evidence contrary to what is suggested by the EMH and random walk theory. These studies show that the stock market can be predicted to some degree and therefore, questioning the EMH's underlying assumptions. Moreover, many practitioners refer to two main examples, which demonstrate that the stock market can be accurately predicted: (a) Warren Buffet's ability to consistently beat the S&P index [8, 9]; and (b) the successful prediction of the 2008 Stock Market crash based on the "housing bubble", which was popularized by the *New York Times Bestseller* book (turned movie): "The Big Short: Inside the Doomsday Machine" [10].

The literature on stock market prediction can be divided, based on the type of prediction models used, into: statistical time-series models and machine learning techniques [11]. Based on the review of [12], autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroscedasticity (GARCH) are the most commonly used time-series approaches for stock prediction. Despite their widespread use, limitations of these models include: (a) needing the model to be prespecified [13]; (b) an increased effect of estimation error as the models become more complex [13]; and (c) sub-par predictive ability when compared to machine learning methods [14, 15, 16]. On the other hand, the machine learning (ML) techniques can be categorized into: (a) non-voting approaches, which include artificial neural networks (ANNs) [17], support vector machines (SVM) [18, 19, 20], and classification & regression trees (CART) [20, 21]; and (b) voting/ensembles [22, 23, 24] and hybrid methods [25, 26, 27]. The reader is referred to [14, 15] for detailed reviews of ML stock market prediction methods.

Based on the above discussion and the reviews of [12, 14, 15, 17], there are four important observations to be made. First, most (if not all) of the stock market prediction papers used some form of the previous price as a predictor/feature. In our estimation, this can be explained by the following logic: if the market can be predicted, then the previous price (or a feature based on it, e.g., through a technical indicator) should explain some of the variation in prices/returns. Second, only a small subset of ML papers considered using macroeconomic predictors [28, 29, 30, 13, 19] (see Table 1 for a summary of their contributions). In our estimation, this can be explained by the following: (a) majority of stock market prediction papers focus on short-term predictions, and (b) macroeconomic indicators are released, at best, monthly. Thus, any paper focusing on the short-term prediction cannot utilize an invariant predictor. The third observation relates to the papers discussed in Table 1. These papers typically showed that macroeconomic indicators can be strong predictors of price (when machine learning models are used). However, these papers generally: (a) had a single index, and (b) utilized both macroeconomic indicators and past prices as predictors so it is not clear how generalizable the results are (to other indices and whether macroeconomic indicators alone can be predictors of future prices). Fourth, the use of ensemble and hybrid-based approaches improves the prediction results through voting/averaging, which is an expected result based on the data mining literature. Based on these observations, this paper will investigate the utility of macroeconomic variables (including those highlighted in [10]) in predicting the one-month ahead price for major U.S stock and sector indices. *The overarching hypothesis is that the price for different indices can be predicted fairly accurately by different economic indicators.* Such effects will be quantified/validated using a novel soft computing approach.

Table 1: A summary of ML papers using macroeconomic predictors

Ref.	**Period**	**Macro Economic Factors**	**Index**	**Target**	**Models Used**	Prediction Outcome
[13]	1976 - 1999	M1, Production Price Index (PPI), CPI, T-Bill, Deposit Rate	S&P Index	Sign of excess stock returns	ANN	CORR: 0.0714; RMSE: 1.21
[28]	1983 - 1990	M1, Gold Price (GP)	S&P Index	Stock movement	ANN	Accuracy: 75%
[29]	1981 - 1991	Industrial Production Index (IPI), GDP, Bond, Consumer Price Index (CPI), Bill rate, Montreal Exchange Index	Company	Return movement	Boltzmann Machine	Accuracy: 66.7%
[30]	1984 - 1994	GP, Oil Price (OP), Bond, Foreign Currency	S&P Index	Volatility	ANN	RMSE 35-days pred. 0.003432 (log transf.)
[19]	1990 - 2002	Exchange Rate of: USD to JPY	NIKKEI 225 Index	Stock Movement	RWH, Linear & Quadratic Discriminant Analyses, SVM	Hit Ratio: 73%
[31]	2000 - 2008	Industrial Production Rate, Inflation Rate Exchange Rate, Unemployment Rate, Oil Price, GDP, M1, M2	Company	Stock Price	Linear Regression ANN	MAPE: ANN = 1.42 & LR= 3.93
[21]	2000 - 2007	A total of 14 and 17 features in their two best models (these included features from fundamental & technical analyses)	Taiwan Stock Index	Stock Movement	Multilayer Perceptron ANN, Principal Component Analysis Genetic Algorithms, & CART	Accuracy: $\approx$ 79%

In this paper, the main research questions are: (a) to examine whether macroeconomic factors can predict the 1-month ahead price of four major U.S stock indices (*the Dow Jones Industrial Average Index,* \$DJI, *the NYSE Composite Index,* \$NYA, *the NASDAQ Composite Index,* \$IXIC, and the *S&P 500 Index,* \$GSPC) and the nine U.S. sector indices; and (b) if the answer to question (a) is "yes", then which factors are the most predictive to each index. To examine these research questions, a two-stage experimental-based framework is proposed.

The first stage is comprised of two main phases. In phase I, an automated data acquisition procedure is developed to capture the monthly values for the different macroeconomic factors (i.e., the independent variables) and closing price for different stock indices (i.e., the response variables). In phase II, four ensemble models and three time-series models are used for predicting the closing price of the different indices. The ensemble models chosen for the analysis are: (i) quantile regression random forest (QRF), (ii) quantile regression neural network ensemble (QRNN), (iii) bagging regression ensemble (BAG_{Reg}), and (iv) boosting regression ensemble ($BOOST_{Reg}$). These have been chosen since they are the most commonly used ensembles for continuous predictions. The performance of these ensembles are then compared with ARIMA and GARCH models, as well as a deep long-term memory recurrent neural network (LSTM) for time-series forecasting [32] (see recent applications to stock predictions in [33, 34, 35]). If the overarching hypothesis in this paper is true (i.e., medium term index prices are driven by macro-economic factors), then one would expect that the performance of the ensemble methods would outperform the time-series methods since the information affecting the medium-term price is in the economy (and not contained completely in past prices).

To validate the insights gained from the first stage, a hybrid approach of the LSTM and the ensemble models will be constructed and utilized in the second stage. In the hybrid approach, the LSTM model (chosen given its nonparametric nature) is used to predict the closing price of the different indices (i.e. the same approach from stage 1). The residuals from this model are then used as target for prediction (i.e. the dependent variable) for the four ensemble models, then the predictions from the LSTM model are adjusted by adding the corrections predicted by the ensembles. Just as the first stage, the macroeconomic indicators are used to predict the 1-month ahead residuals from the LSTM model. Thus, the hybrid approach is used to test the following secondary hypothesis: *the errors/residuals from the time-series models are not completely random and can be explained by the macro economic indicators.*

The remainder of this paper is organized as follows. Section 2 presents the macroeconomic factors that are used in this study and provides a justification for their selection. In Section 3, the proposed two-stage approach is detailed. The results of the experimental study for Stages 1 and 2 are showed in Sections 4 and 5, respectively. Finally, the main contributions of the paper, its practical implications, and suggestions for future work are highlighted in Section 6.

2. JUSTIFICATION FOR THE MACROECONOMIC INDICATORS USED

Researchers list several different macroeconomic factors that could potentially have an impact on stock market movements including oil prices [36, 37, 38], housing prices [39], interest rates [40], foreign markets [41], and inflation [21]. Ref. [42] explored the effects of important macroeconomic variables on stock market returns. From the results, they concluded that industrial production, risk premium change, yield curve twist, and inflation all have significant effects on the variability of stock returns, but macroeconomic factors do not have significant influence on stock price changes. Some researchers have interest not only in stock returns and prices, but also in the relationship between macroeconomic factors and trading volume. For example, [43] paid specific attention to trading volume from 1980 to 1996 and utilized 17 macroeconomic factors to analyze their relationship with high trading volume during the same time frame. Ref. [43] observed that the Consumer Price Index (CPI), the Producer Price Index (PPI), the Monetary Aggregates, the Employment Report, the Balance of Trade, and the Housing Starts are strong risk indicators for the stock market. Ref. [13], for instance, selected 31 financial and economic factors to forecast stock market returns with neural network models. In addition, Ref. [31] found that the Inflation Rate, Money Supply (M1), and Growth Rates of Industrial Production to be predictive of stock price of individual stocks. The discussed references have only focused on major indices. It is not clear what macroeconomic factors can help predict the different sector indices.

Based on the discussion above, a list of 23 macroeconomic factors is generated for this study. Table 2 highlights these predictors, the source from which they were extracted from and some of the papers that utilized them in their analyses. Note that some potential factors, which were not used in the prior literature, were added since this paper also investigates predicting the nine sector prices. These factors were hypothesized to be potentially relevant. The data pertaining to these factors is collected from January, 1992 to October, 2016 on a monthly basis.

Table 2: A list of potentially predictive macroeconomic factors

Macroeconomic indices (monthly) used for prediction			
Oil Price(1) [31, 30]	US unemployment rate(2) [21, 31]	US trade balance(3) [21]	US consumer price index CPI(2) [13, 29, 21]
US auto Sales(3) [44]	Gold Price(5) [28, 30]	US monetary amount (M1)(5) [13, 21, 31]	US monetary supply (M2)(5) [21, 31]
US industrial production index (IPI)(5) [13, 28, 29, 21, 31]	Effective federal fund rate(5) [44]	US inflation_rate(5) [31]	
Macroeconomic indices (monthly) used for association			
Oil production(1) [38]	Oil supply(1) [38]	US house price index(5) [39]	US housing starts(8) [43]
US manufacturing PMI(4) [45]	US house sold(6) [43]	US employment change rate(2) [31]	
Other potential macroeconomic indices (monthly)			
US housing market index (HMI)(6)	US mortgage rate 15 years(7)	US mortgage rate 30 years(7)	US auto production(3)
US consumer sentiment(5)			
Public databases used			
(1): US Energy Information Administration (EIA)		(2): US Bureau of Labor Statistics (BLS)	
(3): US Bureau of Economic Analysis (BEA)		(4): Institute of Supply Management (ISM)	
(5): Federal Reserve Economic Data (FRED)		(6): National Association of Home Builders (NAHB)	
(7): Federal Home Loan Mortgage Corporation (Freddi Mac)		(8): US Census Bureau (CB)	

3. TWO-STAGE APPROACH TO DEMONSTRATE THE UTILITY OF MACROECONOMIC INDICATORS IN PREDICTING MONTHLY STOCK PRICES

Figure 1 presents the process to build up the model. In Stage 1, the data from several different online resources are first collected. The data acquisition phase is divided into two main steps, where the dependent indices' monthly closing prices and the independent macroeconomic predictors (used in the ensemble models) are obtained. Then, in phase II the variables are selected using three machine learning models and consolidated into one final set of features. Phase III compares the seven prediction models and evaluates them using: MAPE, MAE and RMSE. In stage 2, the residuals for the LSTM model are merged with the lagged macroeconomic indicators. For this new dataset, the same variable selection process will be repeated on the macroeconomic factors. Then, the four ensemble models will be re-applied to predict the residuals. By combining the price prediction from the LSTM with the error prediction from the ensembles, a final price prediction is obtained. This prediction is then used to evaluate whether the macroeconomic predictors explain the errors in the time-series model.

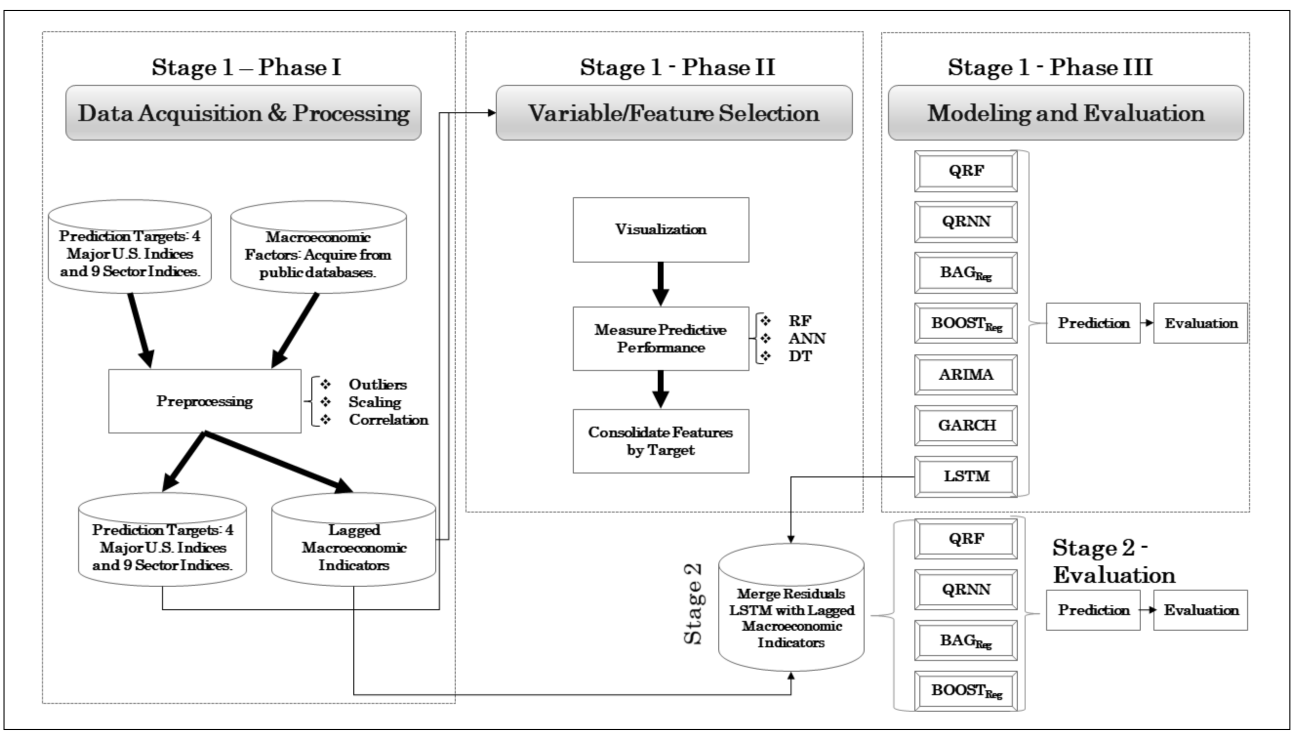


Figure 1: An Overview of the Proposed Method

3.1. Stage 1: Methods Used to Evaluate the Overarching Hypothesis

As discussed in Section 1, the purpose of stage 1 is to evaluate the main hypothesis: "*the price for different indices is driven by different economic indicators*". To evaluate the hypothesis, this study proposes utilizing a traditional data analytic framework comprised of three phases. These phases are highlighted in the subsections below.

3.1.1. Phase I: Data Preparation And Acquisition

In this phase, the monthly closing price data for each of the indices is obtained using the *quantmod* package in R [46]. These closing prices constitute the response variables for our seven models (4 ensembles and three time-series models). In addition, the *quandl* package in R [47] and simple web scraping are utilized to obtain the monthly values for the 23 macroeconomic indicators. These indicators represent the set of potential predictors used by the ensemble methods. The reader is referred to our code, see the *Supplementary Materials* section, for more details on this phase.

3.1.2. Phase II: Variable/Feature Selection

As discussed earlier, phase II is only applicable to the ensemble methods. Here, the goal of is twofold: identify and eliminate the irrelevant/redundant features for the different sector indices and provide intuition on to which factors are the most predictive. To achieve this goal, the first step involved using the following data mining techniques for feature selection: decision trees, random forests and artificial neural networks. These techniques have been shown to be suitable for feature selection in the machine learning literature [13, 21, 48, 49]. A modified Leave-one-out cross validation (LOOCV) is applied to the three methods to minimize the bias associated with the random sampling of the training and testing data samples (see Ref. [50] for details). In the second step, the importance of the predictors for different indices is measured using the sensitivity analysis approach detailed in [51]. This process is performed for each of three data mining feature selecting approaches and a final set of candidate features is generated for phase III. The approach is performed using the *Caret Package* in R. Readers are referred to [52] for more details on the selection process and on how the level of importance of each variable is calculated.

3.1.3. Phase III: Predictive Modeling

In this phase, the prediction performances of the four ensemble models and the three time-series based models are compared. A separate comparison is made for each of the sector indices. Recall that the purpose of the comparison is to evaluate whether the ensemble models (using the set of macroeconomic indicators from phase II as the only candidate predictors) can outperform the time-series models (i.e. where the price history is the only explanatory variable).

The Ensemble Models

Ensemble methods are effective fusion methods to improve the prediction accuracy of classifiers. Instead of selecting a single model, the idea of an ensemble model is to use a vote or an average of various models (often referred to as weak learners) for a specific prediction. There is sufficient empirical evidence pointing to ensemble performance being generally superior to that of individual classifiers [53, 54, 55, 56, 57, 58, 59, 60, 61]. Ensemble methods are commonly used in machine learning to decrease the bias (boosting method) and variance (bagging method) of predictions. The ensemble methods analyzed here include the Quantile Regression Forest (QRF) [62], which is a generalization of the random forests (RFs) [60]. The QRF ensemble gives an approximation of the conditional mean of a response variable, but can also estimate different quantiles of the response variable for given values of the predictors. RF and QRF ensembles have been widely and successfully used in stock market prediction studies [14, 63, 64]. The QRF method is implemented in the quantregForest R package. The reader is referred to [62] for more details.

Similar to QRF, the Artificial Neural Networks (ANNs) are widely employed in a variety of computational data analytics including classification, regression, and pattern recognition. ANNs can provide a more reliable prediction result for high-dimensional nonlinear data and has been a popular approach for stock market prediction (see e.g., [14, 31, 16, 65, 66, 67]. An ensemble of Quantile Regression Neural Networks (QRNN)[68] is also used as a prediction model. Within the QRNN ensembles, the sigmoid function has been used as the activation function. The QRNN ensembles were trained and optimized by adjusting the number of neurons for each target to optimize the prediction performance. The reader is referred to

[68] for more details on the QRNN R package implemented in this study.

Bagging, short for bootstrap aggregation, is another strong ensemble method considered in this study. Given a training set S of size n, bagging uses bootstrapping to generate a new training set S_b of size $n_b = n$ and fits a weak learner to the data. This process is repeated B times, and the final prediction aggregation is an average of the predicted values for regression problems (referred to as BAG_{Reg} here). For this analysis, a regression tree of depth = 1 (regression stump) is used as weak learner. The BAG_{Reg} algorithm is implemented using the "fitensemble" package in Matlab.

The fourth ensemble considered here is a boosting regression ensemble (BOOST_{Reg}). Boosting refers to the idea of converting a weak learning algorithm into a strong learner, that is, taking a classifier that performs slightly better than random chance and boosting it into a classifier with arbitrarily high accuracy. At every iteration, the BOOST_{Reg} ensemble: (a) fits a new learner, i.e., a regression tree of depth 1, and then (b) computes the difference between the observed response and the aggregated prediction of all learners grown previously while minimizing the mean-squared error criterion. The "fitensemble" package in Matlab is used for implementing BOOST_{Reg}.

The Time-Series Models Used for Benchmarking

In this paper, we hypothesize that the price of U.S. stock indices can be better predicted by the macroeconomic predictors alone than by solely using the past prices. To evaluate this hypothesis, three time-series models are used for comparison.

First, a parameter-optimized ARIMA is used to fit the data. Based on the ARIMA model, there are several parameters that need to be estimated. Those parameters are: (a) the number of lags, which is the AR component of the model. The number of lags refer to the number of previous prices of the index that will be used for forecasting the 1-month ahead index); (b) the degree of differencing, which is used to stabilize the series when nonstationarity is suspected (i.e., the I component of the model), and (c) the moving average (MA component) used to correct the prediction through incorporating the previous errors in the series. The "forecast" package in R is used for implementing the ARIMA model [69]. The parameters are automatically optimized to optimally fit the current series. The reader is referred to [70] for more information on this particular ARIMA implementation and to [71] for a complete coverage of ARIMA models.

The second time-series model used in this study is the GARCH model, which often outperforms ARIMA models the analyzed data exhibits a high level of volatility and uncertainty. The specification of a GARCH model is dependent on the estimation of the autoregressive and moving average terms. A GARCH(1,1) model is implemented here for comparison purposes. These specific parameters were chosen as a result of an exploratory inspection of the autocorrelation and partial autocorrelation functions of the predicted indices. Note that we suspect that a GARCH model would be better suited for predicting a series with more volatility (i.e, a 1-day ahead prediction as opposed to 1-month ahead prediction). However, the model is included here for completion since it is commonly used in sock market prediction problem. The reader is referred to [72, 70] for more information on the R GARCH implementation used in this study.

The third time-series model analyzed here is a deep long short-term memory network (Deep LSTM). A 2-layer stacked LSTM with 6 units per layer was chosen for the analysis since it yielded the most consistent results in the preliminary trials on the data. As a result, this LSTM architecture with a single dense output neuron was fit to all predicted indices. The "keras" package in Python was used to fit the models using the ADAM stochastic optimization criterion [73] and the MSE loss function.

The Metrics Used for Assessing the Prediction Performance of the Seven Models

To evaluate the performance of the implemented methodologies, three different evaluation metrics are used in this study: root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE). These three metrics are suitable for regression analysis based on previous literature. The reader is referred to [14] for more details. The MAPE is used as the primary evaluation metric in this paper since it can be easily interpreted. For this analysis, we use ensembles 500 weak learners.

3.2. Stage 2: Methods Used to Evaluate the Secondary Hypothesis

As discussed in Section 1, the purpose of stage 2 is to explain why better predictions can be obtained using the macroeconomic predictors alone when compared to just utilizing the previous prices through a time-series approach. To achieve this goal, a hybrid approach is used to test whether a time-series methodology to predict stock indices can be further optimized by using macroeconomic factors as predictors. The main hypothesis here is that "*the errors/residuals from the time-series models are not entirely random and can be explained by the macro economic indicators.*". Hybrid models have shown to be useful in complementing time-series formulations [74, 75, 76] of ARIMA models given their inherent inability to capture highly nonlinear patterns. In the hybrid formulation proposed here, the residuals from the Deep LSTM model are analyzed for further optimization. The Deep LSTM model was chosen given its nonparametric nature and nonlinear predictive ability. More specifically, given the Deep LSTM prediction $\hat{y}_t$ of index y at a specific time period t is used to obtain the residual e_t at period t. The residual e_t is then used as the response variable for the 4 ensemble methods analyzed here at time t, that is, using the macroeconomic indicators of time period $t-1$ as predictors. The estimated $\hat{e}_t$ serves as a correction to the Deep LSTM prediction to form the hybrid prediction h, that is, $\hat{h}_t = \hat{y} + \hat{e}_t$. The RMSE, MAPE and MAE metrics are used on predictions $\hat{h}_t$ for evaluation. An improvement on the metrics would suggest that the residuals are not completely random and validate the conjecture that macroeconomic predictors can be used to further improve the time-series formulation.

4. STAGE 1: RESULTS AND DISCUSSION

In this section, the Stage 1 results for the proposed method are presented. First, the phase I results are highlighted, where irrelevant and redundant features that do not contribute to, or have a minimal contribution to, the predictive models are identified. Then, the results of the seven prediction models (four ensembles and three time-series models are presented). The performance of these models are compared using three metrics as mentioned in Section 3.1.3. To facilitate the replication of our results, our code and a detailed tabular view of our results are presented at https://github.com/martinwg/stockpredict.git.

4.1. Variable/Feature Selection

To predict stock price fluctuations and trends of U.S. major indices and sector indices, we started by selecting the most important macroeconomic factors as input predictors. Firstly, we assumed that the factors that could affect stock market should stand in distinctive levels among different categories of indices or sectors. As mentioned in Section 3.1.2, four data mining methods with a modified LOOCV approach were applied and evaluated for the selection. We then applied our feature selection approach for each category of target. To select the variables, we use the importance score metric for regression ensembles. The importance score metric is based on the total decrease in node impurities (using decision trees as weak learners), as measured by the sum of squares error (SSE) from splitting on the particular variable, averaged over all trees. We selected the factors with importance scores greater than 0.6 as the predictors. We discuss the results of major indices and sector indices separately.

4.1.1. Important Factors for U.S. Major Indices

We discuss here the macroeconomic factor influences on four U.S. major indices, and differentiate the factor importance levels based on our results. Table 3 shows the influential factors with importance scores greater than 0.6 for each of the indices. We list the factors in a descending order based on their influence to the model. There are several additional discussions to be made from Table 3:

(A) Four major indices are affected by different sets of socioeconomic factors, and this verifies our assumption.

(B) The three main indicators with the highest importance scores are IPI, Money Stock M2, and CPI. The results show that IPI, Money Stock M2 and CPI could impact the stock prices of all four indices to a great extent; however, some differences still exist among these three indicators. The change of IPI could have more influence on the stock price to the Dow Jones Industrial Average Index ($DJI), NYSE Composite Index ($NYA), and S&P 500 Index ($GSPC) independently compared to the Money Stock M2 and CPI. The fluctuation of CPI could affect the stock price of the NASDAQ Composite Index ($IXIC) most, compared to the influence of the other two factors.

(C) Based on our variable selection rules, the factors that affect the stock price of the NASDAQ Composite Index ($IXIC) are very different from the other three indices. For example, except for the NASDAQ Composite Index ($IXIC), the stock price of the other three major indices are strongly influenced by IPI, Money Stock M2 and CPI as their importance scores are greater than 0.8 and the score differentiation is

Table 3: Important Factors For U.S. Major Indices & Sectors

Index/Sector	**Important Factors**			
DJI	IPI	M2	CPI	House Price Index
	M1	Gold Price	15 Year Mortgage Rate	30 Year Mortgage Rate
GSPC	IPI	M2	CPI	House Price Index
	M1	Gold Price		
IXIC	CPI	M2	IPI	M1
	15 Year Mortgage Rate	30 Year Mortgage Rate	House Price Index	
NYA	IPI	M2	CPI	House Price Index
	M1	Gold Price	Oil Price	
Materials	CPI	M2	M1	Housing Starts
	Gold Price	Oil Production	15 Year Mortgage Rate	
Energy	CPI	M2	Housing Starts	M1
	Oil Price	Gold Price	15 Year Mortgage Rate	30 Year Mortgage Rate
	House Sold			
Financial	M2	Housing Starts	House Price	Unemployment Rate
	M1	CPI	Consumer Sentiment	
Industrials	CPI	M2	M1	Oil Production
	IPI	Housing Starts	15 Year Mortgage Rate	Gold Price
	30 Year Mortgage Rate			
Technology	M2	15 Year Mortgage Rate	CPI	30 Year Mortgage Rate
	M1			
Utilities	CPI	M2	M1	Housing Starts
	IPI	Oil Production	15 Year Mortgage Rate	Gold Price
	30 Year Mortgage Rate	House Sold	Federal Fund Rate	
Consumer Staples	M2	CPI	M1	Oil Production
	15 Year Mortgage Rate	30 Year Mortgage Rate	Housing Starts	Gold Price
	IPI	Federal Fund Rate		
Healthcare	M1	M2	Oil Production	CPI
	IPI	15 Year Mortgage Rate	30 Year Mortgage Rate	Gold Price
	House Sold			
Consumer Discretionary	M1	M2	CPI	Oil Production
	15 Year Mortgage Rate	IPI	30 Year Mortgage Rate	Gold Price
	Housing Starts			

small, but less or equal to 0.8 for the NASDAQ Composite Index ($IXIC). Gold Price is also considered as an input predictor for the Dow Jones Industrial Average Index ($DJI), NYSE Composite Index ($NYA), and S&P 500 Index ($GSPC), but not for the NASDAQ Composite Index ($IXIC).

(D) The change of some macroeconomic factors could not have an obvious effect on the stock price of four indices, because of their low importance scores. For example, Manufacturing PMI, Employment Change, and Consumer Sentiment are the least powerful factors for all of the four major indices, since their importance scores are close to 0.

(E) The other factors, except for those discussed above, could still impact the stock price of the four indices in different levels. For example, house market and oil market are the two major markets that are associated with the the stock prices of the four indices significantly. Another interesting finding is that Inflation Rate, which is an important macroeconomic indicator related with CPI and currency, has low level impact on all the four indices. In addition, the change of employment condition does not result in much fluctuation based on the feature importance score.

4.1.2. Important Factors for U.S. Stock Major Sectors

Similar to Subsection 4.1.1, we capture the results in Table 3, and discuss the relationship between macroeconomic factors and nine U.S. major sector indices. We followed the S&P Dow Jones Indices stock sector classification rule to categorize the stock sectors including Materials, Energy, Financial, Industrials, Technology, Utilities, Consumer Staples, Consumer Discreet, and Healthcare. Unlike the influence of macroeconomic factors on the four major U.S. indices, the relationship between the factors and sector indices are primarily different for each sector.

For the sector index of Materials, Energy, and Utilities, the stock price is sensitive to the changes of CPI, Money Stock M2, Money Stock M1 and Housing Starts. This means that the stock price of the three sector indices maybe responding to the changes of these three macroeconomic factors quickly. This finding is very different when compared to the four major indices whose most important factors are IPI, Money Stock M2 and CPI. One explanation can be that the companies in these three sectors are related to infrastructure. The change of consumption and financial conditions may result in the fluctuation of these sector indices directly.

The stock prices of the Industrials, Consumer Staples, Healthcare, and Consumer Discreet sector indices are mainly impacted by four macroeconomic factors, which are CPI, Money Stock M2, Money Stock M1, and Oil Production. By comparing the importance scores among these four factors, CPI, Money Stock M2, and Money Stock M1 seem to be driving the stock price of three sector indices (Industrials, Consumer Staples and Consumer Discreet) to a high level. However, the macroeconomic factors affecting the stock price of the Health Care sector index is somewhat different when compared to the other three sectors. Specifically, the Money Stock M1, Money Stock M2, and Oil Production are the three most influential factors. Based on these results, it seems reasonable to posit that these four sectors are closely related to daily-life consumer behavior.

The results of the financial sector and technology sector indices seem to indicate that they are affected by a smaller subset of macroeconomic factors. The trends in the prices of these sections imply that they are correlated. In our estimation, this "correlation" makes sense since technology companies are disrupting the Technology sector (e.g., Apple Pay, Google Wallet and PayPal). The results show that no macroeconomic factors had an importance score greater than 0.8. We hypothesize that this may be explained by the fact that our analysis was limited to the past 14 years. This somewhat small sample size may be insufficient to capture an emerging pattern, especially since the impact of technology on the financial sector can be seen as a recent phenomenon. Some other observations that pertain to the selected features include: (a) the financial sector is affected by consumer sentiment, and (b) changes in the housing market result in some fluctuations for both sector indices.

4.2. Outcomes from the Prediction Models

As explained in Section 4.1.1 and 4.1.2, we use the RMSE, MAPE, MAE metrics as the criteria to evaluate the performance of the quantile regression forest (QRF), quantile regression neural network (QRNN), Bagging Regression (BAG_{Reg}), Boosting Regression (BOOST_{Reg}), ARIMA, GARCH(1,1) and Deep LSTM models. In this section, the results of the prediction models are explained in two parts. The first part discusses the parameter settings for the ensemble machine learning models. Then, the prediction results are discussed.

4.2.1. Parameters Settings

To maximize the performance of outputs, the first step is to set up the optimal parameters for the two models. The quantile regression forest is a generalization of random forests that can measure conditional quantiles to improve the information learned as discussed in Section 3. The size of terminal nodes needs to be set when applying a QRF model. The larger the size, the smaller the trees to be grown (and thus less time is need for training and execution). Based on a comparison of various tree sizes, our prediction model uses a node size = 10 for the 13 indices. The quantile regression neural network uses bootstrap aggregation to create an ensemble of models. We adjust the number of neurons for each target to optimize the output. We use MAPE as the primary evaluation criterion for the comparison. Figures 2 and 3 show how the number of neurons affect the MAPE values for the four major indices and nine sector indices, respectively.

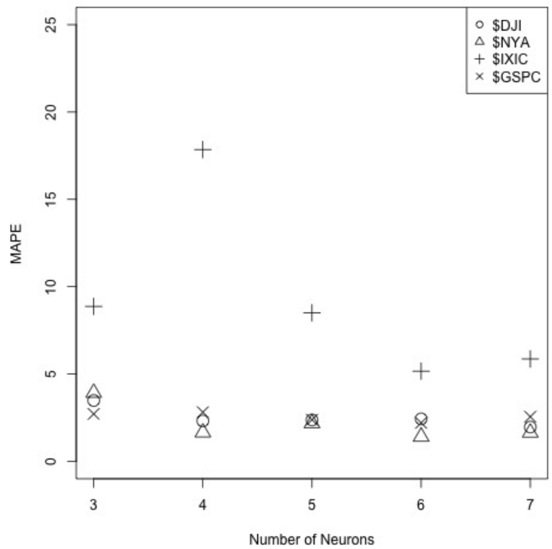


Figure 2: Performance of QRNN for Different Number of Neurons: Indices

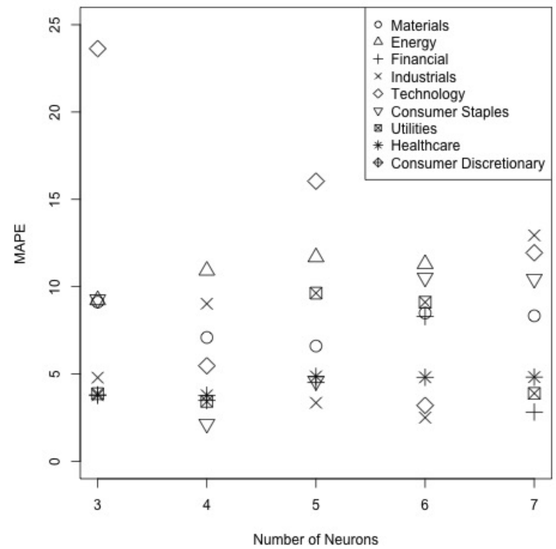


Figure 3: Performance of QRNN for Different Number of Neurons: Sectors

From these two figures, there three noteworthy observations. First, there exist significant differences in the MAPE values among the major indices and sector indices with different number of neurons. Second, when we compare the four major indices, the S&P 500 Index ($GSPC) consistently has a higher MAPE, and the number of neurons has a more profound impact on its performance. Third, the optimal number of neurons used for the QRNN is four or six for our 13 indices since the improvement when more than six neurons are used is negligible (and sometimes negative). For the four major indices, six neurons results in the best combined computational efficiency and MAPE. This is also true for the Materials, Industrials and Technology sector indices. For the remaining indices, we use four neurons.

4.2.2. Experiment Results

To demonstrate the feasibility and effectiveness of the proposed methods, we perform experiments on predicting four major indices and nine major sectors. Our experimental results are presented in Table 4. We depict our predictions for the four major indices in

Figure 4. Note that the straight line is the actual stock market price and the dotted lines are the prediction value using the QRF, QRNN, BAG$_{Reg}$, BOOST$_{Reg}$ models. Similar approaches and figures are also generated for the nine different sectors using the QRF, QRNN, BAG$_{Reg}$, BOOST$_{Reg}$ models. For the sake of conciseness, we refer the reader to our Github website (https://github.com/martinwg/stockpredict.git).

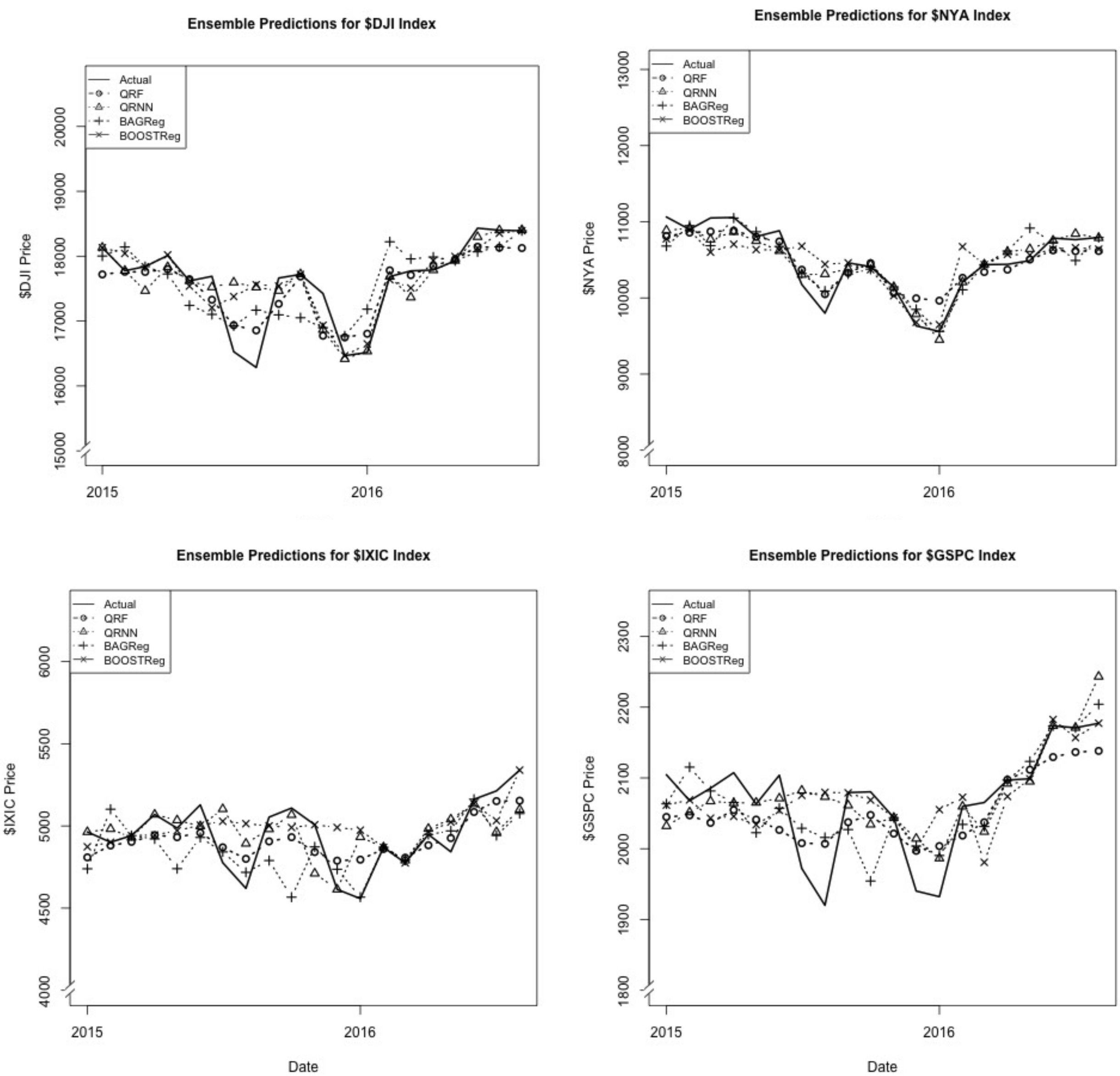


Figure 4: Experiment Results of 4 Major Indices Using the QRF, QRNN, BAG$_{Reg}$ and BOOST$_{Reg}$ models

Based on Figure 4 and Table 4, there are several interesting results that should be noted. First, the ensemble models have excellent predictive performances. The average MAPE

Table 4: Performance of Ensemble Methods for Major/Sector Indices

Target	Prediction Model	Measurements			Target	Prediction Model	Measurements		
		RMSE	MAPE (%)	MAE			RMSE	MAPE (%)	MAE
Dow Jones Industrial Average	QRF	287.67	1.14	201.60	**NASDAQ Composite**	QRF	112.80	1.87	94.12
	QRNN	318.67	1.40	247.44		QRNN	186.98	2.58	128.59
	BAG_{Reg}	168.03	0.61	110.19		BAG_{Reg}	145.75	2.28	113.27
	$BOOST_{Reg}$	210.00	0.64	113.10		$BOOST_{Reg}$	146.05	2.17	109.54
	ARIMA	696.40	3.15	538.73		ARIMA	227.81	3.71	185.40
	GARCH (1,1)	691.43	3.11	531.34		GARCH (1,1)	291.30	5.85	290.95
	Deep LSTM	583.98	3.06	636.06		Deep LSTM	212.92	4.16	205.53
NYSE Composite	QRF	184.07	1.43	148.61	**S&P 500**	QRF	45.41	1.84	38.75
	QRNN	273.36	1.96	205.78		QRNN	24.79	0.81	17.15
	BAG_{Reg}	133.72	0.91	95.46		BAG_{Reg}	43.73	1.37	29.30
	$BOOST_{Reg}$	178.28	1.32	138.67		$BOOST_{Reg}$	27.94	1.08	22.75
	ARIMA	324.42	2.56	266.25		ARIMA	70.63	2.83	58.05
	GARCH (1,1)	326.51	2.55	265.36		GARCH (1,1)	85.70	4.17	85.54
	Deep LSTM	327.22	4.72	487.94		Deep LSTM	67.62	3.51	71.92
Industrials	QRF	1.21	1.67	0.90	**Consumer Staples**	QRF	1.00	1.59	0.79
	QRNN	3.10	4.55	2.47		QRNN	3.15	5.76	2.86
	BAG_{Reg}	1.87	1.69	0.90		BAG_{Reg}	1.02	1.54	0.77
	$BOOST_{Reg}$	0.50	0.51	0.28		$BOOST_{Reg}$	0.80	0.80	0.40
	ARIMA	1.91	2.73	1.45		ARIMA	1.46	2.38	1.18
	GARCH (1,1)	3.07	5.68	3.06		GARCH (1,1)	3.23	6.47	3.22
	Deep LSTM	1.90	5.15	2.73		Deep LSTM	1.47	6.18	3.08
Technology	QRF	1.33	2.56	1.12	**Utilities**	QRF	1.25	2.13	0.97
	QRNN	2.25	4.16	1.76		QRNN	4.13	7.56	3.51
	BAG_{Reg}	1.37	2.62	1.13		BAG_{Reg}	1.23	2.01	0.92
	$BOOST_{Reg}$	0.99	1.67	0.70		$BOOST_{Reg}$	0.76	0.95	0.42
	ARIMA	2.11	4.21	1.80		ARIMA	1.91	3.43	1.55
	GARCH (1,1)	1.96	3.88	1.65		GARCH (1,1)	2.35	5.14	2.32
	Deep LSTM	0.78	3.38	0.81		Deep LSTM	1.76	8.77	3.97
Materials	QRF	1.78	3.06	1.34	**Healthcare**	QRF	2.53	2.70	1.91
	QRNN	3.85	7.27	3.15		QRNN	6.59	8.00	5.51
	BAG_{Reg}	1.89	3.24	1.41		BAG_{Reg}	2.75	2.58	1.83
	$BOOST_{Reg}$	1.24	1.35	0.56		$BOOST_{Reg}$	1.13	0.93	0.65
	ARIMA	2.55	4.83	2.12		ARIMA	2.93	3.18	2.21
	GARCH (1,1)	2.56	4.83	2.12		GARCH (1,1)	4.54	6.43	4.53
	Deep LSTM	2.49	6.90	3.48		Deep LSTM	2.83	4.49	3.13
Energy	QRF	2.33	2.80	1.80	**Discretionary**	QRF	1.54	1.67	1.29
	QRNN	6.21	7.92	5.03		QRNN	5.66	6.70	5.20
	BAG_{Reg}	2.68	3.13	1.97		BAG_{Reg}	1.51	1.58	1.22
	$BOOST_{Reg}$	0.90	1.03	0.66		$BOOST_{Reg}$	0.72	0.67	0.51
	ARIMA	3.79	4.72	3.11		ARIMA	3.05	3.03	2.32
	GARCH (1,1)	3.80	4.76	3.13		GARCH (1,1)	4.92	6.40	4.91
	Deep LSTM	3.97	10.81	7.05		Deep LSTM	2.77	2.87	2.98
Financial	QRF	0.65	2.62	0.48					
	QRNN	1.15	4.73	0.86					
	BAG_{Reg}	0.72	2.84	0.52					
	$BOOST_{Reg}$	0.47	1.60	0.29					
	ARIMA	0.78	3.45	0.64					
	GARCH (1,1)	0.79	3.45	0.64					
	Deep LSTM	0.78	4.18	0.82					

across all models and indices is 2.53%. If we divide this average MAPE across the four major stock indices and the sector indices, the corresponding average MAPEs are 1.46% and 3.01%.

Perhaps, what is even more impressive is that the best model for a given index performs no worse than 1.87% (which is the QRF for the NASDAQ Composite Index). This means that our best model predicts, on average, within 2% of the actual price for the next month. This result is significantly better than the reported values in the literature (see e.g., [28, 13, 66, 77]). Second, the BOOST$_{Reg}$ model has the best overall performance. Third, a closer examination of Figure 4 shows that the prediction performance varies among different time periods. We hypothesize that this might be an indication that some macroeconomic factors might actually lag the stock market movement. While this is a reasonable justification, this is an area that need to be further studied in future studies. Fourth, and a not obvious result, the QRNN's performance is dependent on the number of input features/predictors; this result can be seen by combining the results from Tables 3 and 4. Finally, a comparison to the three time-series methodologies; ARIMA, GARCH(1,1) and Deep LSTM shows that the time-series formulation generally perform worse than the ensemble methods. This supports our overarching hypothesis: "*the price for different indices is driven by different economic indicators*". Note that our conclusions from this experiment are limited to the time horizon and the indices examined.

From a more holistic perspective, we chose 23 macroeconomic factors to evaluate their impact on 13 stock indices. Using a structured variable selection approach for each index, we obtained the subset of the most important predictors. The inclusion criterion was having factors with importance scores greater than 0.6. Based on our variable selection approach, we have determined that different sectors and indices are affected by somewhat different subsets of macroeconomic factors. While this is not a surprising result, it is not obvious from the analysis of the literature since most approaches typically predicted one target (i.e. a stock or an index, see Table 1 in Section 1). From a prediction perspective, our average MAPE results and our best case performances clearly demonstrate the accuracy of our method for predicting the one-month ahead index prices.

5. STAGE 2: EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the evidence supporting the secondary hypothesis: "*the errors/residuals from the time-series models are not entirely random and can be explained by the macro* *economic indicators*" is evaluated. As explained in Section 3, the results for the proposed hybrid Deep LSTM-Ensemble formulation are presented here. Recall that the residuals from the Deep LSTM model e_t are used as Target for the four ensembles analyzed. The predicted residuals $\hat{e}_t$ are then used to correct the errors in the Deep LSTM prediction.

5.1. *Phase I: Variable/Feature Selection*

Similar to Stage 1, the three data mining methods utilizing a modified LOOCV approach were used for variable selection. The results from this phase are depicted in Table 5, which shows the most influential macroeconomic factors for predicting the Deep LSTM model's residuals. The reader should note that the most important predictors are quite different than the predictors showed in Table 3. For example, *Consumer Sentiment*, which was not commonly selected in Stage 1 now consistently ranks as the most important prediction.

Table 5: Important Factors For U.S. Major Indices & Sectors using Deep LSTM Residuals as target

Index/Sector	**Important Factors**			
DJI	Consumer Sentiment IPI	Manufacturing PMI Auto Sales	Oil Production	Inflation Rate
GSPC	Consumer Sentiment Auto Sales	Manufacturing PMI Employment Change	Oil Production	Inflation Rate
IXIC	Inflation Rate Oil Production	Effective Federal Fund Rate	HMI	Unemployment Rate
NYA	Oil Production Effective Federal Fund Rate	15 Year Mortgage Rate Money Stock M1	30 Year Mortgage Rate	Consumer Sentiment
Materials	Oil Production Employment Change	Inflation Rate Consumer Sentiment	Manufacturing PMI HMI	Gold Price
Energy	Oil Production Employment Change	Inflation Rate Consumer Sentiment	Manufacturing PMI HMI	Gold Price
Financial	Consumer Sentiment Employment Change	Manufacturing PMI Auto Sales	Inflation Rate	Oil Production
Industrials	Consumer Sentiment Employment Change	Manufacturing PMI Employment Rate	Inflation Rate	Oil Production
Technology	Consumer Sentiment Auto Sales	Employment Change CPI	HMI Auto Production	Manufacturing PMI
Utilities	Inflation Rate Unemployment Rate	Manufacturing PMI Housing Starts	Oil Production	Employment Change
Consumer Staples	Auto Sales Oil Supply	Consumer Sentiment	Unemployment Rate	Manufacturing PMI
Healthcare	Oil Production Manufacturing PMI	Auto Sales	Consumer Sentiment	Employment Change
Consumer Discretionary	Consumer Sentiment Manufacturing PMI	Oil Production Auto Sales	Housing Starts	IPI

5.2. *Experiment Results*

Table 6 compares the results of the deep LSTM model with those of the four hybrid models (one combining the LSTM with each ensemble). For a given sector, the top row presents the baseline performance, where the single LSTM model was applied (i.e., the same results from Table 4 of Section 1). The following four rows capture the performance of each hybrid model. The results show that there is a substantial improvement in performance using a hybrid approach compared to the time-series formulation. In fact, the improvement is typically between 25-50%. This result supports our secondary hypothesis, which in turn validates our earlier observation that the information hidden in the macroeconomic factors is more predictive than that contained in the previous prices of the indices.

Table 6: Performance of Ensemble Methods for Major/Sector Indices using Deep LSTM Residuals as target

Target	**Prediction Model**	**Measurements**			**Target**	**Prediction Model**	**Measurements**		
		RMSE	**MAPE (%)**	**MAE**			**RMSE**	**MAPE (%)**	**MAE**
Dow Jones Industrial Average	Deep LSTM	583.98	3.06	636.06	**NASDAQ Composite**	Deep LSTM	212.92	4.16	205.53
	LSTM-QRF Hybrid	481.66	2.01	347.07		LSTM-QRF Hybrid	142.64	2.45	120.44
	LSTM-QRNN Hybrid	600.10	2.68	461.15		LSTM-QRNN Hybrid	199.14	3.46	167.66
	LSTM-BAG$_{Reg}$ Hybrid	512.83	2.18	375.36		LSTM-BAG$_{Reg}$ Hybrid	164.59	2.78	136.55
	LSTM-BOOST$_{Reg}$ Hybrid	377.76	1.36	236.66		LSTM-BOOST$_{Reg}$ Hybrid	137.93	2.24	110.07
NYSE Composite	Deep LSTM	327.22	4.72	487.94	**S&P 500**	Deep LSTM	67.62	3.51	71.92
	LSTM-QRF Hybrid	234.53	1.79	184.20		LSTM-QRF Hybrid	47.26	1.75	35.90
	LSTM-QRNN Hybrid	346.40	2.64	270.76		LSTM-QRNN Hybrid	55.51	1.96	40.02
	LSTM-BAG$_{Reg}$ Hybrid	246.95	1.83	188.91		LSTM-BAG$_{Reg}$ Hybrid	48.68	1.81	36.92
	LSTM-BOOST$_{Reg}$ Hybrid	168.21	1.29	133.74		LSTM-BOOST$_{Reg}$ Hybrid	23.38	0.66	13.78
Industrials	Deep LSTM	1.90	5.15	2.73	**Consumer Staples**	Deep LSTM	1.47	6.18	3.08
	LSTM-QRF Hybrid	1.44	2.11	1.11		LSTM-QRF Hybrid	1.03	1.71	0.84
	LSTM-QRNN Hybrid	2.07	3.27	1.71		LSTM-QRNN Hybrid	1.46	2.50	1.22
	LSTM-BAG$_{Reg}$ Hybrid	1.43	2.07	1.09		LSTM-BAG$_{Reg}$ Hybrid	1.06	1.75	0.86
	LSTM-BOOST$_{Reg}$ Hybrid	0.95	0.57	1.09		LSTM-BOOST$_{Reg}$ Hybrid	0.41	0.73	0.36
Technology	Deep LSTM	0.78	3.38	0.81	**Utilities**	Deep LSTM	1.76	8.77	3.97
	LSTM-QRF Hybrid	0.52	2.32	0.43		LSTM-QRF Hybrid	1.25	2.13	0.97
	LSTM-QRNN Hybrid	0.86	3.71	0.68		LSTM-QRNN Hybrid	4.13	7.56	3.51
	LSTM-BAG$_{Reg}$ Hybrid	0.54	2.36	0.43		LSTM-BAG$_{Reg}$ Hybrid	1.23	2.01	0.92
	LSTM-BOOST$_{Reg}$ Hybrid	0.15	0.66	0.12		LSTM-BOOST$_{Reg}$ Hybrid	0.76	0.95	0.42
Materials	Deep LSTM	2.49	6.90	3.48	**Healthcare**	Deep LSTM	2.83	4.49	3.13
	LSTM-QRF Hybrid	1.68	3.21	1.41		LSTM-QRF Hybrid	1.80	2.86	1.41
	LSTM-QRNN Hybrid	2.53	4.85	2.11		LSTM-QRNN Hybrid	2.68	2.86	1.99
	LSTM-BAG$_{Reg}$ Hybrid	1.71	3.29	1.44		LSTM-BAG$_{Reg}$ Hybrid	1.83	2.04	1.42
	LSTM-BOOST$_{Reg}$ Hybrid	0.47	0.86	0.38		LSTM-BOOST$_{Reg}$ Hybrid	0.42	0.47	0.33
Energy	Deep LSTM	3.97	10.81	7.05	**Discretionary**	Deep LSTM	2.77	2.87	2.98
	LSTM-QRF Hybrid	2.53	3.20	2.13		LSTM-QRF Hybrid	2.12	2.18	1.66
	LSTM-QRNN Hybrid	4.28	5.67	3.69		LSTM-QRNN Hybrid	3.03	3.46	2.62
	LSTM-BAG$_{Reg}$ Hybrid	2.69	3.34	2.22		LSTM-BAG$_{Reg}$ Hybrid	2.09	2.12	1.62
	LSTM-BOOST$_{Reg}$ Hybrid	1.20	1.30	0.87		LSTM-BOOST$_{Reg}$ Hybrid	0.79	0.71	0.55
Financial	Deep LSTM Hybrid	0.78	4.18	0.82					
	LSTM-QRF Hybrid	0.56	2.52	0.46					
	LSTM-QRNN Hybrid	0.89	3.85	0.70					
	LSTM-BAG$_{Reg}$ Hybrid	0.55	2.41	0.45					
	LSTM-BOOST$_{Reg}$ Hybrid	0.14	0.49	0.09					

6. CONCLUSIONS AND FUTURE WORK

6.1. An Overview of the Impacts and Contributions of this Paper

The overarching goal behind this paper was to investigate if macroeconomic indicators are drivers for the monthly prices of the main stock and sector indexes in the U.S. To investigate this hypothesis, a two-stage approach was proposed. The first stage was comprised of three phases. In phase I, the data from 01/1992 to 10/2016 was acquired, covering the monthly values of 13 major indexes and 23 potentially relevant macroeconomic indicators. Phase II involved the use of variable selection methodology to reduce the subset of potential predictors without the loss of information. The results from the variable selection suggested that the important subset of important macroeconomic predictors can change according to the target index. In phase III, four ensemble approaches (QRF, QRNN, Bag_{Reg}, and Boost_{Reg}) and three time-series methods (ARIMA, GARCH and LSTM) were evaluated in terms of their ability to predict the price for each of the 13 indexes. The evaluation was primarily performed using MAPE. The phase III results showed that the use of macroeconomic indicators (alone, via an ensemble) are more predictive than the information contained in historical prices (alone). To the best of our knowledge, this is a novel and important result, which has not been reported prior in the literature.

Based on the result from phase III, the second stage was used to test whether the result can be further explained. Accordingly, a novel hybrid method was proposed to investigate whether the residuals from the LSTM model (i.e. one of the three time-series models) can be explained by the macroeconomic indicators. The four ensembles, using only the macroeconomic indicators as explanatory variables, were then applied to predict the one-month ahead error from the LSTM model (i.e. the bias). After the error was predicted, a hybrid additive prediction as made (i.e. the price from the LSTM + the bias from each ensemble). The results from stage II show that the three evaluation metrics (RMSE, MAPE and MAE) can be typically improved by 25-50% by incorporating the information hidden in the macroeconomic indicators (through the ensemble approach).

6.2. Practical Implications from our Work

The ability to accurately predict the stock price, and consequently compute the estimated return, is the "dream" of every investor. In this paper, we presented an ensemble-based approach for predicting the one-month ahead price of 13 U.S. indexes. Based on our reported results, where the MAPE of the best model for a given index was $< 1.87\%$, we believe that our approach has the potential to be informative for investors. As such, we have "packaged" our approach in an interactive decision support system (DSS) that can be used by investors.

The DSS requires no coding by an investor, and is hosted on: http://shiny.eng.auburn.edu/eco-stock/. In our estimation, our DSS has several features that do not exist in current systems (see e.g., [78]). First, it allows the investor to "pull" all the data needed, with a few clicks. This is only possible since our macroeconomic data is scraped from several public repositories. Second, we present some visualizations that are typically used in stock market analysis. For example, we provide the investor with an "interactive technical analysis chart". While we do not use *technical analysis* in our model, we believe that this analysis is useful from an exploratory data analysis viewpoint. Third, our predictions of actual price instead of movement (i.e. up or down) is insightful, especially since our prediction error is small. In our estimation, this DSS increases the appeal behind our method.

6.3. Limitations and Future Work

Despite the excellent predictive performance of our proposed methodology, there are a number of limitations that need to be highlighted. First, the utility of our two-stage approach is only examined for one-month ahead predictions. It is not clear whether the results from testing our primary and secondary hypotheses will be the same if longer time periods are investigated. Second, our analytical framework cannot be applied for intervals that are smaller than one month (e.g., daily forecasting) since macroeconomic factors are released monthly. Third, the scope of this work was limited to major U.S. stock and sector indexes. Based on our scope, it is not clear if the insights from this study will remain valid in the case of price prediction of major indexes in non-US stock exchanges. The reader should note that the utilization of our approach in an emerging market may (and potentially should) result in a different initial set of macroeconomic factors. For example, in the case of an emerging market, it may be necessary to include other macroeconomic factors such as exchange rates. Fourth, the research team did not consider any additional data sources (e.g., predictors derived from technical analysis and fundamental analysis, *Twitter sentiment*, *Wikipedia traffic volume*, etc.). In our analyses, these omissions were justified since the addition of these sources would have, at best, led to a minor (practically insignificant) improvement. However, if any of the three assumptions above were changed, it is unclear whether ignoring these potential predictors can be justified. Note that the discussion in this section focuses on limitations that pertain to price prediction (i.e., not stock movement) and to the utilization of the studied ensembles. Thus, the observations in this section do not reflect on the literature that had fundamentally different objectives, utilized models, and/or assumptions.

In our estimation, there are three major opportunities for future research. First, researchers can examine the impact of predicting the price at different time-points. This is an important direction since it can provide insights pertaining to answering some of the gaps/limitations in our work. For example, the choice of a different time interval may lead to determining that multiple data sources are needed. Second, it seems logical to extend our work into a prescriptive trading engine, which uses our predictions to minimize investment risk and maximize the returns. For this second opportunity, researchers should examine multiple measures of risk [79]. In addition, the variation in price predictions from our four ensembles can potentially be used to quantify the uncertainty in a single index's price forecast. The third stream is related to applied soft computing theory. This paper demonstrated the potential advantages from combining traditional ensemble methodology with time-series based methods. The main advantage is to provide a possible explanation/justification for when the main assumption of time-series based methods (i.e. unstructured errors) is violated. Thus, there is an opportunity to extend our work to other problems of similar structure. One example, that the authors are familiar with, can be the use of wearables to detect fatigue in humans [80].

In summary, this paper proposed a novel two-stage framework that was used to show: (a) utilizing only macroeconomic indicators, one can predict the one-month ahead price of major stock and sector indexes; and (b) macroeconomic indicators can be used to explain the error from time-series models used in forecasting the one-month ahead index price. Three major research streams are highlighted to highlight some future research opportunities. Our code and data are made available at https://github.com/martinwg/stockpredict.git to encourage the reproducibility of our work and future research. Finally, a decision support system is presented to encourage the utilization of our work by practitioners and researchers from other disciplines. The system is hosted online at: http://shiny.eng.auburn.edu/eco-stock/.

7. REFERENCES

[1] E. F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34–105.
URL http://www.jstor.org/stable/2350752

[2] P. Cootner, The random character of stock market prices, M.I.T. Press, 1964.
URL https://books.google.com/books?id=jW9gT8U6dqQC

[3] B. G. Malkiel, The efficient market hypothesis and its critics, The Journal of Economic Perspectives 17 (1) (2003) 59–82.

[4] V. L. Smith, Constructivist and ecological rationality in economics, The American Economic Review 93 (3) (2003) 465–508.

[5] J. R. Nofsinger, Social mood and financial economics, The Journal of Behavioral Finance 6 (3) (2005) 144–160.

[6] R. R. Prechter Jr, W. D. Parker, The financial/economic dichotomy in social behavioral dynamics: the socionomic perspective, The Journal of Behavioral Finance 8 (2) (2007) 84–108.

[7] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science 2 (1) (2011) 1–8.

[8] A. Kiersz, Here's how badly warren buffett beat the market, http://www.businessinsider.com/warren-buffett-berkshire-hathaway-vs-sp-500-2015-3 (03 2015).

[9] C. J. Loomis, Buffett beats the sp for the 39th year, http://fortune.com/2012/02/25/buffett-beats-the-sp-for-the-39th-year/ (02 2012).

[10] M. Lewis, The Big Short: Inside the Doomsday Machine (movie tie-in), WW Norton & Company, 2015.

[11] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, S.-P. Guo, Stock index forecasting based on a hybrid model, Omega 40 (6) (2012) 758–766.

[12] S.-H. Poon, C. W. Granger, Forecasting volatility in financial markets: A review, Journal of economic literature 41 (2) (2003) 478–539.

[13] D. Enke, S. Thawornwong, The use of data mining and neural networks for forecasting stock market returns, Expert Systems with Applications 29 (4) (2005) 927–940.

[14] G. S. Atsalakis, K. P. Valavanis, Surveying stock market forecasting techniques–part ii: Soft computing methods, Expert Systems with Applications 36 (3) (2009) 5932–5941.

[15] E. Hajizadeh, H. D. Ardakani, J. Shahrabi, Application of data mining techniques in stock markets: A survey, Journal of Economics and International Finance 2 (7) (2010) 109.

[16] K. S. Vaisla, A. K. Bhatt, An analysis of the performance of artificial neural network technique for stock market forecasting, International Journal on Computer Science and Engineering 2 (6) (2010) 2104–2109.

[17] Ö. Ican, T. B. Çelik, Stock market prediction performance of neural networks: A literature review, International Journal of Economics and Finance 9 (11) (2017) 100.

[18] K.-j. Kim, Financial time series forecasting using support vector machines, Neurocomputing 55 (1) (2003) 307–319.

[19] W. Huang, Y. Nakamori, S.-Y. Wang, Forecasting stock market movement direction with support vector machine, Computers & Operations Research 32 (10) (2005) 2513–2522.

[20] P. Ou, H. Wang, Prediction of stock market index movement by ten data mining techniques, Modern Applied Science 3 (12) (2009) 28.

[21] C.-F. Tsai, Y.-C. Hsiao, Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches, Decision Support Systems 50 (1) (2010) 258–269.

[22] C.-F. Tsai, Y.-C. Lin, D. C. Yen, Y.-M. Chen, Predicting stock returns by classifier ensembles, Applied Soft Computing 11 (2) (2011) 2452–2459.

[23] N. F. Da Silva, E. R. Hruschka, E. R. Hruschka, Tweet sentiment analysis with classifier ensembles, Decision Support Systems 66 (2014) 170–179.

[24] S. Pyo, J. Lee, M. Cha, H. Jang, Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the korean stock markets, PloS one 12 (11) (2017) e0188107.

[25] L.-J. Kao, C.-C. Chiu, C.-J. Lu, C.-H. Chang, A hybrid approach by integrating wavelet-based feature extraction with mars and svr for stock index forecasting, Decision Support Systems 54 (3) (2013) 1228–1244.

[26] J. Patel, S. Shah, P. Thakkar, K. Kotecha, Predicting stock market index using fusion of machine learning techniques, Expert Systems with Applications 42 (4) (2015) 2162–2172.

[27] H. J. Sadaei, R. Enayatifar, M. H. Lee, M. Mahmud, A hybrid model based on differential fuzzy logic relationships and imperialist competitive algorithm for stock market forecasting, Applied Soft Computing 40 (2016) 132–149.

[28] G. Grudnitski, L. Osburn, Forecasting s&p and gold futures prices: An application of neural networks, Journal of Futures Markets 13 (6) (1993) 631–643.

[29] L. Kryzanowski, M. Galler, D. W. Wright, Using artificial neural networks to pick stocks, Financial Analysts Journal 49 (4) (1993) 21–27.

[30] S. A. Hamid, Z. Iqbal, Using neural networks for forecasting volatility of s&p 500 index futures prices, Journal of Business Research 57 (10) (2004) 1116–1125.

[31] R. G. Ahangar, M. Yahyazadehfar, H. Pournaghshband, The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in tehran stock exchange, International Journal of Computer Science and Information Security 7 (2) (2010) 38–46.

[32] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[33] R. Akita, A. Yoshihara, T. Matsubara, K. Uehara, Deep learning for stock prediction using numerical and textual information, in: Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on, IEEE, 2016, pp. 1–6.

[34] H. Jia, Investigation into the effectiveness of long short term memory networks for stock price prediction, arXiv preprint arXiv:1603.07893.

[35] K. Chen, Y. Zhou, F. Dai, A lstm-based method for stock returns prediction: A case study of china stock market, in: Big Data (Big Data), 2015 IEEE International Conference on, IEEE, 2015, pp. 2823–2824.

[36] P. Sadorsky, Oil price shocks and stock market activity, Energy Economics 21 (5) (1999) 449–469.

[37] J. Park, R. A. Ratti, Oil price shocks and stock markets in the us and 13 european countries, Energy Economics 30 (5) (2008) 2587–2608.

[38] L. Kilian, C. Park, The impact of oil price shocks on the us stock market, International Economic Review 50 (4) (2009) 1267–1287.

[39] K. E. Case, J. M. Quigley, R. J. Shiller, Comparing wealth effects: the stock market versus the housing market, Advances in Macroeconomics 5 (1) (2005) 1–23.

[40] A. A. Rahman, N. Z. M. Sidek, F. H. Tafri, Macroeconomic determinants of malaysian stock market, African Journal of Business Management 3 (3) (2009) 95.

[41] Y. Hamao, R. W. Masulis, V. Ng, Correlations in price changes and volatility across international stock markets, Review of Financial Studies 3 (2) (1990) 281–307.

[42] N.-F. Chen, R. Roll, S. A. Ross, Economic forces and the stock market, Journal of Business (1986) 383–403.

[43] M. J. Flannery, A. A. Protopapadakis, Macroeconomic factors do influence aggregate stock returns, Review of Financial Studies 15 (3) (2002) 751–782.

[44] A. Mahajan, L. Dey, S. M. Haque, Mining financial news for major events and their impacts on the market, Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 1 (2008) 423–426.

[45] K. J. Johnson, Mark A.and Watson, Can changes in the purchasing managers index foretell stock returns? an additional forward-looking sentiment indicator, The Journal of Investing 20 (4) (2011) 89–98.

[46] J. Ryan, J. Ulrich, Quantmod: Quantitative financial modelling framework. r package version 0.4-12, https://cran.r-project.org/web/packages/quantmod/quantmod.pdf (2017).

[47] R. McTaggart, G. Daroczi, C. Leung, Quandl: Api wrapper for quandl.com. r package version 2.8.0, https://github.com/joshuaulrich/quantmod (2016).

[48] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann series in machine learning, Elsevier Science, 2014.
URL https://books.google.com/books?id=b3ujBQAAQBAJ

[49] A. Dag, K. Topuz, A. Oztekin, S. Bulur, F. M. Megahed, A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival, Decision Support Systems 86 (2016) 1–12.

[50] S. Arlot, A. Celisse, et al., A survey of cross-validation procedures for model selection, Statistics Surveys 4 (2010) 40–79.

[51] A. Dag, A. Oztekin, A. Yucel, S. Bulur, F. M. Megahed, Predicting heart transplantation outcomes through data analytics, Decision Support Systems 94 (2017) 42–52.

[52] M. Kuhn, Caret package, Journal of Statistical Software 28 (5) (2008) 1–26.

[53] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, V. Vapnik, Boosting and other ensemble methods, Neural Computation 6 (6) (1994) 1289–1301.

[54] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.

[55] L. Breiman, Bias, variance, and arcing classifiers, Technical Report 460.

[56] J. R. Quinlan, Bagging, boosting, and c4. 5, AAAI/IAAI, Vol. 1 (1996) 725–730.

[57] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, Boosting the margin: A new explanation for the effectiveness of voting methods, Annals of Statistics 26 (1998) 1651–1686.

[58] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research (1999) 169–198.

[59] T. G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.

[60] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[61] R. Maclin, D. Opitz, Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research 11 (2011) 169–198.

[62] N. Meinshausen, Quantile regression forests, Journal of Machine Learning Research 7 (Jun) (2006) 983–999.

[63] R. K. Lai, C.-Y. Fan, W.-H. Huang, P.-C. Chang, Evolving and clustering fuzzy decision tree for financial time series data forecasting, Expert Systems with Applications 36 (2) (2009) 3761–3773.

[64] M. Wiesmeier, F. Barthold, B. Blank, I. Kögel-Knabner, Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem, Plant and Soil 340 (1-2) (2011) 7–24.

[65] E. Guresen, G. Kayakutlu, T. U. Daim, Using artificial neural network models in stock market index prediction, Expert Systems with Applications 38 (8) (2011) 10389–10397.

[66] F. A. de Oliveira, C. N. Nobre, L. E. Zarate, Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index–case study of petr4, petrobras, brazil, Expert Systems with Applications 40 (18) (2013) 7596–7606.

[67] R. P. Schumaker, Machine learning the harness track: Crowdsourcing and varying race history, Decision Support Systems 54 (3) (2013) 1370–1379.

[68] J. W. Taylor, A quantile regression neural network approach to estimating the conditional density of multiperiod returns, Journal of Forecasting 19 (4) (2000) 299–311.

[69] R. Hyndman, M. O'Hara-Wild, C. Bergmeir, S. Razbash, E. Wang, forecast: Forecasting functions for time series and linear models. r package version 8.2, https://cran.r-project.org/web/packages/forecast/forecast.pdf (2017).

[70] R. J. Hyndman, Y. Khandakar, et al., Automatic time series for forecasting: the forecast package for R, no. 6/07, Monash University, Department of Econometrics and Business Statistics, 2007.

[71] G. E. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.

[72] A. Ghalanos, Introduction to the rugarch package.(version 1.3-1), Tech. rep., Technical report v. Available at http://cran. r-project. org/web/packages/rugarch (2018).

[73] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[74] G. P. Zhang, Time series forecasting using a hybrid arima and neural network model, Neurocomputing 50 (2003) 159–175.

[75] P.-F. Pai, C.-S. Lin, A hybrid arima and support vector machines model in stock price forecasting, Omega 33 (6) (2005) 497–505.

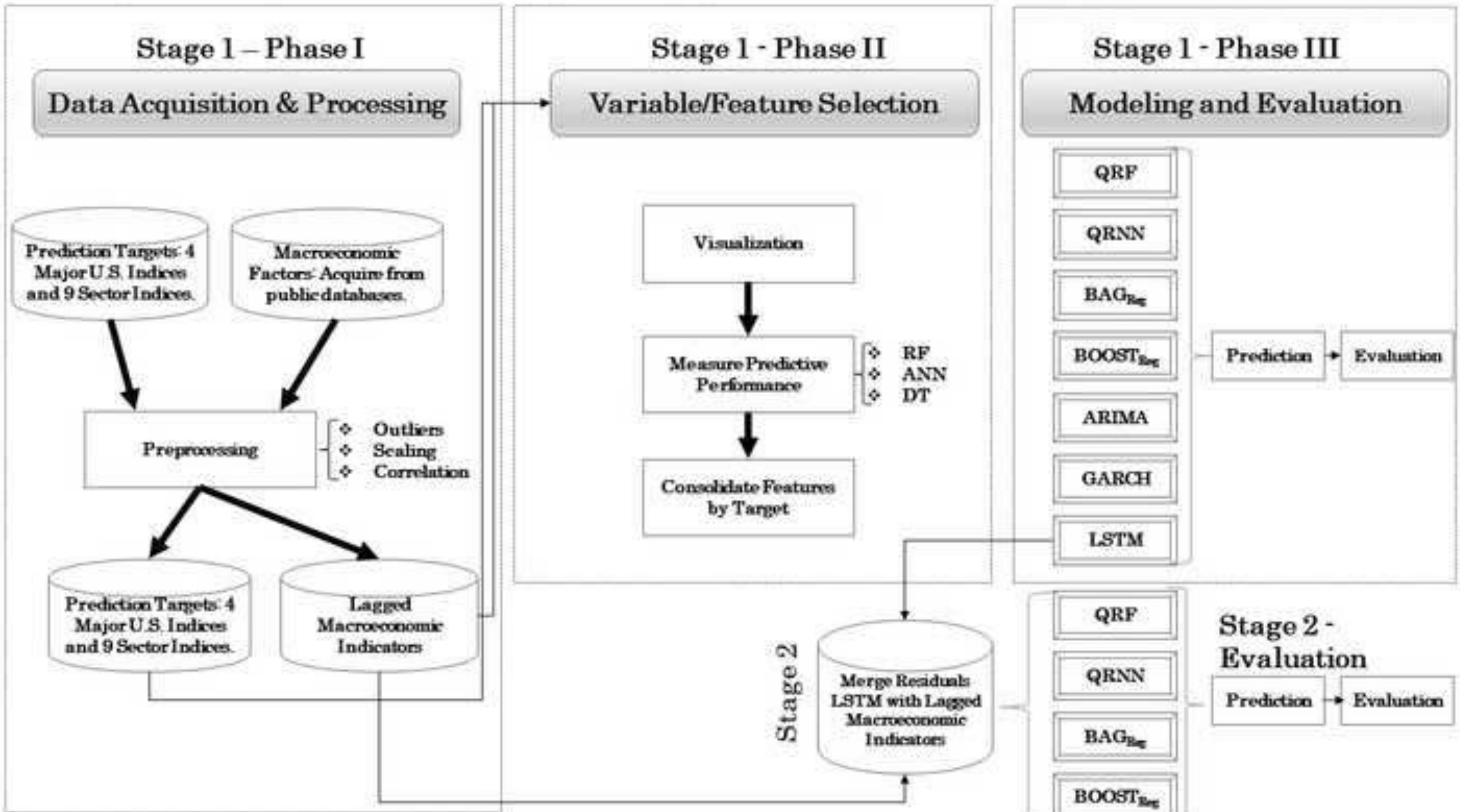
[76] H. Liu, H.-q. Tian, Y.-f. Li, Comparison of two new arima-ann and arima-kalman hybrid methods for wind speed prediction, Applied Energy 98 (2012) 415–424.

[77] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, O. K. Hussain, Support vector regression with chaos-based firefly algorithm for stock market price forecasting, Applied Soft Computing 13 (2) (2013) 947–958.

[78] J. Gottschlich, O. Hinz, A decision support system for stock investment recommendations using collective wisdom, Decision support systems 59 (2014) 52–62.

[79] G. Szegö, Measures of risk, Journal of Banking & Finance 26 (7) (2002) 1253–1272.

[80] Z. S. Maman, M. A. A. Yazdi, L. A. Cavuoto, F. M. Megahed, A data-driven approach to modeling physical fatigue in the workplace using wearable sensors, Applied ergonomics 65 (2017) 515–529.


Stage 1 – Phase I
Data Acquisition & Processing
Prediction Targets: 4 Major U.S. Indices and 9 Sector Indices.
Macroeconomic Factors: Acquire from public databases.
Preprocessing
Outliers
Scaling
Correlation
Prediction Targets: 4 Major U.S. Indices and 9 Sector Indices.
Lagged Macroeconomic Indicators
Stage 1 - Phase II
Variable/Feature Selection
Visualization
Measure Predictive Performance
RF
ANN
DT
Consolidate Features by Target
Stage 1 - Phase III
Modeling and Evaluation
QRF
QRNN
BAG$_{Reg}$
BOOST$_{Reg}$
ARIMA
GARCH
LSTM
Prediction
Evaluation
Stage 2
Merge Residuals LSTM with Lagged Macroeconomic Indicators
QRF
QRNN
BAG$_{Reg}$
BOOST$_{Reg}$
Stage 2 - Evaluation
Prediction
Evaluation