Time series forecasting using a hybrid ARIMA and Random Forest Model

Franciska Englert, Milan Badics

Abstract. The purpose of this paper is to predict stock prices using various models and identify the best amongst them. A hybrid model was developed combining linear ARIMA and the non-linear Random Forest (RF) regression. The performance of ARIMA-RF is compared with performance of ARIMA, RF, and ensemble models. The different methods are evaluated in terms of statistical metrics.

Keywords: ARIMA · Random Forest · Time series forecasting · Combined forecast.

1 Introduction

Time series forecasting has become an important tool over the past decades, being used in various domains of applied science and engineering. Past observations are collected and analyzed to develop the model which is then used to give predictions for future values. This modeling approach is especially useful when little is known about the underlying data generating process. The evolution of financial time series is one example of such datasets, having a lot of noise and being dependent on various factors including general behavioural and economic conditions, political events, or the movement of other financial markets [5].

A wide range of linear and non-linear models has been used in the past to forecast time series. Early studies included exponential smoothing, discriminant analysis, vector autoregression, and autoregressive integrated moving average (ARIMA) [2]. These linear methods were not able to give satisfactory results in many cases due to the non-linearity of the examined datasets. Hence time series prediction shifted towards machine learning methods. Artificial neural networks, Support Vector Machines, and Random Forest regression are all suitable for modeling non-linear relationships and have been used for such problems [8,3,4].

One of the major developments in time series forecasting was the application of model combination or ensemble modeling. The idea behind this approach is that different models can capture different aspects of the data. Empirical findings have suggested that this is an effective way to improve forecast accuracy especially when the models combined are relatively different [10].

Many attempts have been made to such model combinations. Zhang [9] described a hybrid methodology, combining ARIMA and artificial neural networks. Applying the combined model to three well-known data sets (Wolf's sunspot data, the

Canadian lynx data, and the British pound to US dollar exchange rate data), they concluded that the hybrid method outperformed both ARIMA and ANN when used in isolation. Khashei [6] described a similar ARIMA-ANN methodology, improving the forecasting performance of Zhang's model.

Various papers studied the effect of using different non-linear methods in the hybrid model. Pai and Lin [7] proposed a methodology to exploit the strength of ARIMA and Support Vector Machines for stock prices forecasting and concluded that using the combination improves model performance. Kumar [5] compared the forecasting accuracy of different hybrid models on the SP CNX Nifty index. In their study, an ARIMA-ANN, ARIMA-SVR and ARIMA-RF model were built and tested against simple methodology. They found that the hybrid forecasting resulted in higher accuracy compared to single models in all cases and that ARIMA-SVR performed the best amongst all of them.

In this study, our aim is to build a Random Forest and ARIMA hybrid model and test it on twelve highly liquid stocks. The rest of the paper is organized as follows. The ARIMA, RF, and hybrid methodology is introduced in Section 2. Empirical results from twelve stock index returns are reported in Section 3. Section 4 proposes further research ideas while Section 5 contains the concluding remarks.

2 Methodology

2.1 ARIMA

In the autoregressive moving average model, we assume that the future value y_t is a linear function of several previous variables and random errors. That is, the underlying data generating process has the form

$$\begin{aligned} y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \\ \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ... - \theta_q \epsilon_{t-q} \end{aligned} \tag{1}$$

where y_t is the true value of the time series at time t, and the error terms ϵ_t are assumed to be independent, identically distributed variables with mean zero and variance σ^2. The integers p and q are called the orders of the model, ϕ_1 ... ϕ_p, θ_1, ... θ_q are parameters to be estimated.
The time series model is based on the assumption that the process y_t is stationary. For many real-life time series this assumption does not hold, hence data transformation is needed. In our case, we take the log-return of the stock prices and build a model on this transformed time series. The best model for forecasting is identified using Akaike Information Criteria (AIC).

2.2 Random Forest regression

The Random Forest (RF) algorithm was originally proposed by Breiman [1]. The idea is that instead of growing a single decision tree, let's have numerous trees and average out the votes given by these.

Let **x** be an observed covariate vector of dimension p drawn from the random vector **X**, and let the numerical response **y** be an instance of the random variable **Y**. Then we can describe our dataset as (x, y) points. Let's bootstrap this with replacement. That is, if θ_k is a random vector taken from the set $\{1, 2, ..., N\}^p$, our i^{th} observation in the bootstrapped dataset will be defined as the $\theta_k(i)^{th}$ observation in the original dataset. Then a decision tree can be built using these data points. Each tree grown by the algorithm can be associated with a classifier $h(x, \theta_k)$. Then the prediction given by the Random Forest regression for x is

$$h(x) = \frac{1}{K} \sum_{i=1}^{K} h(x, \theta_i) \tag{2}$$

We can define the margin function as

$$mg(X, Y) = \frac{1}{K} \sum_{k=1}^{K} I(h(X, \theta_k) = Y) - \max_{j \neq Y} \frac{1}{K} \sum_{k=1}^{K} I(h(X, \theta_k) = j) \tag{3}$$

with $I(.)$ being the indicator function. This measures the extent to which the average number of votes at (X, Y) for the right class exceeds the average vote for other classes.The larger the margin, the more confidence we have in the classification.

Let the prediction error be

$$PE^* \coloneqq \mathbb{P}_{X,Y}(mg(X, Y) < 0) \tag{4}$$

It can be proved [1] that as the number of trees increases, this prediction error will converge almost surely. Therefore, adding more trees to the Random Forest regression will not result in overfitting but instead will give a limiting generalization error.

In addition to this, to keep residual correlation low, at each node we will randomly pick m variables out of the p covariates and use only those to grow the tree. In the corresponding literature, m is usually chosen to be approximately $\sqrt{p}$.

2.3 The hybrid model

As seen in Zhang (2003, [9]) a time series can be formed using a linear autocorrelation structure and a non-linear component. That is, the time series Y_t can be written as

$$Y_t = L_t + N_t \tag{5}$$

where L_t and N_t correspond to the linear and non-linear component respectively. The linearity of the data set can be captured using the ARIMA model. Hence, if $\hat{L}_t$ is the forecast value for Y_t then the residual $r_t = Y_t - \hat{L}_t$ will capture the non-linearity of the time series. These residuals can be forecasted using an appropriate Random Forest model. If the forecast from RF is denoted by $\hat{N}_t$ the combined forecast for Y_t will be

$$\hat{Y}_t = \hat{L}_t + \hat{N}_t \tag{6}$$

3 Empirical results

To compare the models built, it is necessary to evaluate them on out-of-sample data. This is the best way to get a picture of their forecasting performance when applied in real life. In order to do this, a rolling window was used to train the model accompanied with 1-week-ahead forecasting.

3.1 Data sets

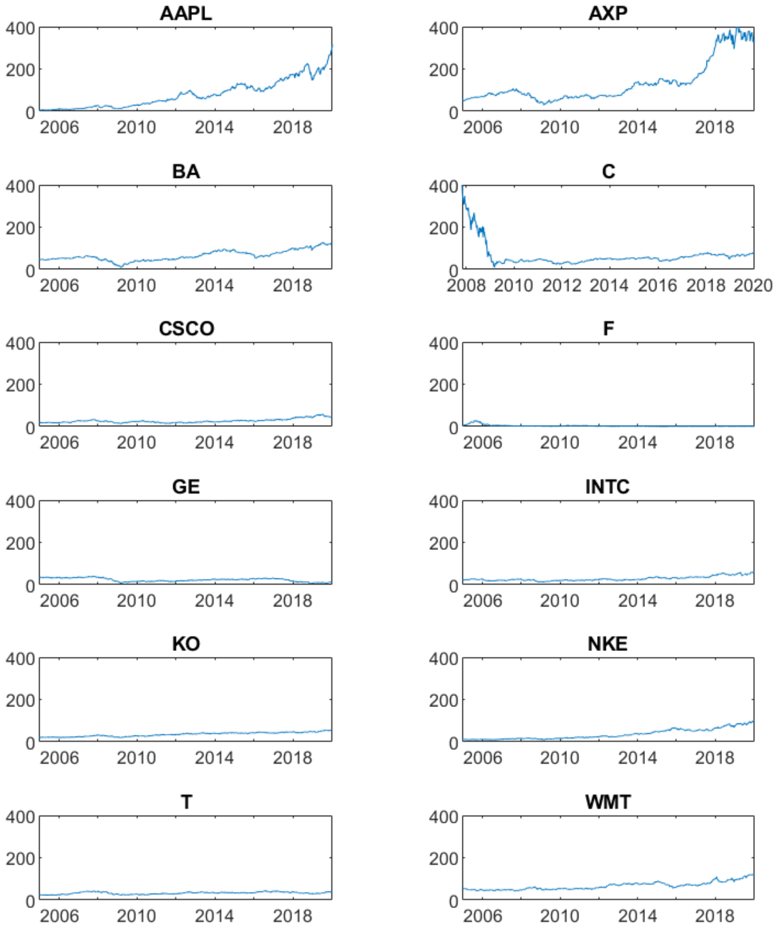

Fig. 1: Evolution of time series for the twelve stocks examined

Our underlying data is weekly stock market returns between December 2004 and January 2020. We examine twelve liquid stocks in this period: Apple (AAPL), American Express (AXP), Boeing (BA), Citigroup (C), Cisco (CSCO), Ford (F), General Electric (GE), Intel (INTC), Coca-Cola (KO), Nike (NKE), ATT (T) and Walmart (WMT). Having a total number of 717 observations, a window size of 50 was used for prediction. Therefore the out-of-sample data has size 667. Evolution of the time series can be seen on Figure 1.

It can be seen that these time series follow a similar pattern except for the Citigroup data, which had a major fall in the period 2006 to 2009.

3.2 ARIMA model

Before developing the ARIMA models, it is necessary to transform the time series. We take the log-return of the stock prices and then build the ARIMA methodology on the transformed dataset. The study experimented with past 4 lags, the best model among these was chosen using Akaike's Information Criterion. We only perform the optimization procedure once, using the first 50 data points.

3.3 RF regression

In this study, the RF model was not optimized. The lag was chosen to be 4 ($p = 4$) and the number of explanatory variables was $m = 2$. A total number of $K = 100$ trees were grown to minimize computational time. To keep individual error low, trees were grown to maximal depth.

3.4 ARIMA-RF model

The input to the RF model was the residuals of the fitted ARIMA model. We used past 4 lags of the residuals of ARIMA model to develop ARIMA-RF model. Here too, the number of trees grown was 100. The predicted residuals obtained by the RF model were added to the forecasts of the ARIMA model to get the forecasts of the hybrid methodology.

3.5 Forecast evaluation

In addition to the three models built, results will be evaluated using the ensemble model as well. Forecasts given by the ensemble model can be obtained by averaging the predictions of ARIMA and RF models. The aim of this is to see whether the hybrid methodology captures the nonlinear - linear pattern of the dataset better than using only simple combination.
We used three statistical metrics such as MAE, RMSE, NMSE to evaluate the forecasting accuracy of the models developed. MAE, RMSE and NMSE measure the deviation between the predicted and actual values. The smaller the values of MAE, NMSE and RMSE, the better the forecasts are. Summary of these statistical metrics can be studied on Table 1.

	Performance metrics			Performance metrics		
	MAE	RMSE	NMSE	MAE	RMSE	NMSE
		AAPL			**AXP**	
ARIMA-RF	0.0378	0.0522	0	0.0331	0.0458	0
ARIMA	0.0362	0.0495	0	0.0319	0.0448	0
RF	0.0373	0.0515	0	0.0324	0.0442	0
Ensemble	0.0364	0.0501	0	0.0317	0.0440	0
		BA			**C**	
ARIMA-RF	0.0339	0.0539	0	0.0488	0.0539	0
ARIMA	0.0323	0.0518	0	0.0461	0.0850	0
RF	0.0327	0.0442	0	0.0486	0.0859	0
Ensemble	0.0321	0.0513	0	0.0467	0.0844	0
		CSCO			**F**	
ARIMA-RF	0.0299	0.0417	0	0.0570	0.0881	0
ARIMA	0.0294	0.0412	0	0.0543	0.0856	0
RF	0.0299	0.0415	0	0.0568	0.088	0
Ensemble	0.0292	0.0409	0	0.0548	0.0861	0
		GE			**INTC**	
ARIMA-RF	0.0325	0.0498	0	0.0319	0.0427	0
ARIMA	0.0309	0.0479	0	0.0310	0.0417	0
RF	0.0322	0.0498	0	0.0314	0.0417	0
Ensemble	0.0310	0.0484	0	0.0307	0.0412	0
		KO			**NKE**	
ARIMA-RF	0.0179	0.0241	0	0.0292	0.0410	0
ARIMA	0.0169	0.0232	0	0.0279	0.0394	0
RF	0.0176	0.0239	0	0.0288	0.0409	0
Ensemble	0.0171	0.0233	0	0.0280	0.0397	0
		T			**WMT**	
ARIMA-RF	0.0230	0.0410	0	0.0210	0.0288	0
ARIMA	0.0223	0.0303	0	0.0200	0.0278	0
RF	0.0228	0.0309	0	0.0203	0.0279	0
Ensemble	0.0223	0.0303	0	0.0199	0.0276	0

Table 1: Forecasting performance (in terms of MAE, RMSE, NMSE) for the twelve stocks

Apart from statistical metrics, another way of evaluating the results is to plot the cumulative sum of errors for the four models developed. Two examples are given in this paper, on Figure 2 and Figure 3 the resuls for stocks AAPL and C are reported respectively. In our case, the accuracy statistics and cumulative sum plots provide us with mixed results.

The non-linear Random Forest model was outperformed by the linear ARIMA model in all twelve cases. This is a phenomenon that was discussed by Kumar [5] as well, when evaluated on the S&P CNX Nifty Index. Our hybrid ARIMA-RF methodology gave the worst predictions on these stocks. This might be due to poor model parameter selection and the lack of model optimization for Random Forest regression.

For five stocks (AXP, BA, CSCO, INTC, WMT) the Ensemble model performed better than ARIMA, leaving the other seven stocks with the linear ARIMA performing the best amongst them.

From the cumulative sum of error plots we can conclude similar results. It is interesting to note though, that in many cases the hybrid methodology had the smallest cumulative sum of errors up to a certain time after which the ARIMA model outperformed it greatly.

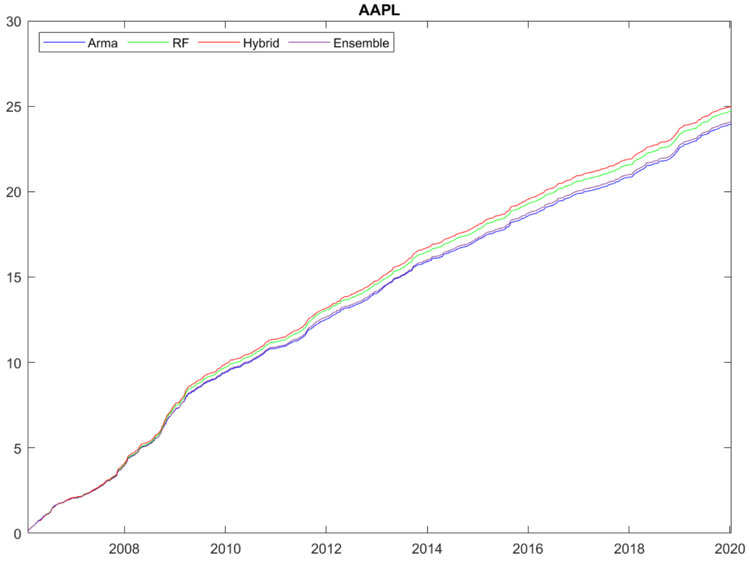

Fig. 2: Cumulative sum of errors for AAPL

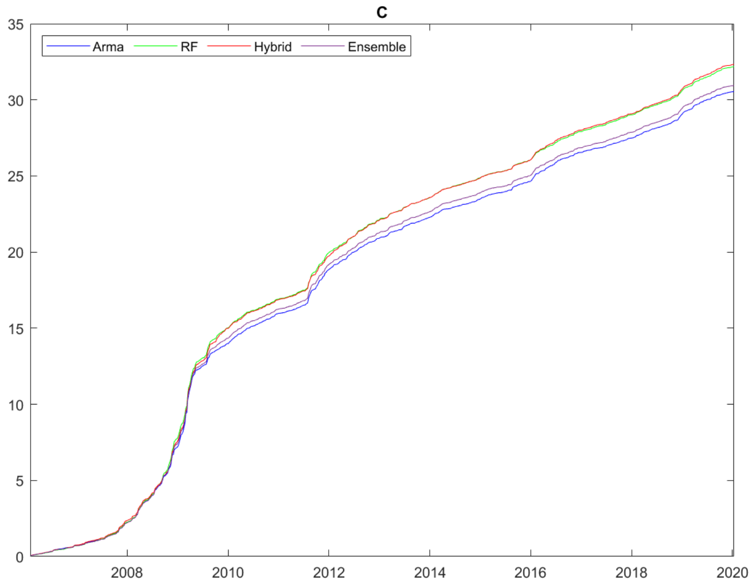

Fig. 3: Cumulative sum of errors for C

4 Further Research

In this study, a hybrid ARIMA and Random Forest model has been developed. The parameters of the Random Forest model were not optimized. Previously, a hybrid ARIMA-ANN and ARIMA-SVR model was built but there is still space for further research. Instead of using RF regression, it is possible to use Gradient boosting. It would be interesting to see which one works better in this setting. Kumar at al (2014) tested the ARIMA-RF model on the SP CNX Nifty index which is an emerging market. Kumar discusses that financial time series forecasting is more accurate in such markets. In this study, we focused on American stocks. Would be interesting to see how much better does the model perform on one data than the other.

5 Conclusion

References

1. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
2. Jan G. De Gooijer, R.J.H.: 25 years of time series forecasting. International Journal of Forecasting **22**, 443–473 (2006)
3. jae Kim, K.: Financial time series forecasting using support vector machines. Neurocomputing **55**, 307–319 (2003)
4. Manish Kumar, M.T.: Forecasting stock index movement: A comparison of support vector machines and random forest. Indian Institute of Capital Markets 9th Capital Markets Conference Paper (2006)

5. Manish Kumar, M.T.: Forecasting stock index returns using arima-svm, arima-ann, and arima-random forest hybrid models. Int. J. Banking, Accounting and Finance **5**, 3 (2014)
6. Mehdi Khashei, M.B.: An artificial neural network (p, d, q) model for timeseries forecasting. Expert Systems with Applications **37**, 479–489 (2010)
7. Ping-Feng Pai, C.S.L.: A hybrid arima and support vector machines model in stock price forecasting. Omega **33**, 497–505 (2005)
8. Zaiyong Tang, Chrys de Almeida, P.A.F.: Time series forecasting using neural networks vs. box- jenkins methodology. Simulation **57**, 303–310 (1991)
9. Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. Neurocomputing **50**, 159–175 (2003)
10. Zhang, G.P.: A neural network ensemble method with jittered training data for time series forecasting. Information Sciences **177**, 5329–5346 (2007)