Forecasting Stock Prices

Eric Ofori

11/1/2020

Title: *Forecasting Stock Prices Using Machine Learning Methods*

Introduction

Accurately forecasting stock prices is probably almost impossible, otherwise everyone on Wall Street would be rich. Trends in stock prices are nonlinear and non-stationary time-series, which makes forecasting stock prices a challenging and difficult task in the financial market. Conventional time series models have been used to forecast stock prices, and many researchers are still devoted to the development and improvement of time-series forecasting models. The most well-known conventional time series forecasting approach is autoregressive integrated moving average (ARIMA), which is employed when the time-series data is linear and there are no missing values. Statistical methods, such as traditional time series models, usually address linear forecasting models and variables must obey statistical normal distribution. Therefore, conventional time series methods are not suitable for forecasting stock prices, because stock price fluctuation is usually nonlinear and non-stationary.

Further, most conventional time-series models utilize one variable (the previous day's stock price) only, when, there are actually many influential factors, such as market indexes, technical indicators, economics, political environments, investor psychology, and the fundamental financial analysis of companies that can influence forecasting performance. In practice, researchers use many technical indicators as independent variables for forecasting stock prices. How to select the key variables from numerous technical indicators is a critical step in the forecasting process. Investors usually prefer to select technical indicators depending on their experience or feelings for forecasting stock prices despite this behavior be highly risky. However, choosing unrepresentative indicators may result in losing profits for investors. Therefore, selecting the relevant indicators to forecast stock prices is one of the important issues for investors. Financial researchers must identify the key technical indicators that have higher relevance to the stock price by indicator selection. Therefore, proposed models must incorporate indicator selection in the stock forecasting process to enhance forecasting accuracy.

Forecasting stock prices has become a hot topic with the advent of more efficient and precise machine learning methods. To overcome the shortcomings of traditional time series models for estimating stock prices, I investigate and compare two of such as machine learning approaches: ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression.

Empirical Framework

There is a set of methods proposed for regression in which the target value is likely to be a linear combination of the input variables. In mathematical notion, if Y is the predicted value, then its value can be obtained from equation 2 as h(x). Ordinary least squares method is one of the generalized linear regression models in which linear regression fits a linear model with coefficients w = (w1,. . . , wp) to minimize the residual sum of squares between the witnessed responses in the dataset, and the responses forecasted by the linear approximation. Mathematically, it solves a problem of minimizing the expression of type as shown in equation 3. The linear regression can take in its fit method arrays X, y and will accumulate the coefficients w of the linear model in its coef_member. However, the freedom of the model terms decides coefficient assessments for Ordinary Least Squares. This method calculates the least squares solution using a singular value decomposition of X. If X is a matrix of size (n, p) then this method has a cost of O(np^2), if $n \geq p$.

Mathematical Formulation

Consider the set of training vectors $(x_i, y_i), x_n$ belongs to R_n, y_n belongs to R,

$$i = 1, 2, 3, ..., N \tag{1}$$

The hypothesis or the linear regression output is given by

$$h(x) = \sum_{j=0}^{d} w_j x_j = w^T x \tag{2}$$

where w is the weight vector and d is the dimensionality of the problem or the number of features.
Also, $x_0 = 1$ has been added to make equation (2) valid.
The cost function or the squared error function is defined as

$$J(w) = \frac{1}{N} \sum_{i=0}^{N} (h(x_i) - y_i)^2 = \frac{1}{N} ||Xw - y||^2 \tag{3}$$

where

$$X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_N^T \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{4}$$

We need to minimize the objective function to get the optimal value of the weight vector.
Minimizing the objective function,

$$\nabla J(w) = \frac{2}{N} X^T (Xw - y) = 0 \tag{5}$$

which implies

$$X^T X w = X^T y \tag{6}$$

hence

$$w = X^+ y \tag{7}$$

where

$$X^+ = (X^T X)^- 1 X^T \tag{8}$$

Putting the value of w from (7) the optimal hypothesis is obtained.
The traditional least square method can be modified to Ridge regression model.
The objective function for Ridge regression can be specified as

$$J(w) = \frac{1}{N}\sum_{i=0}^{N}(h(x_i) - y_i)^2 + \lambda\sum_{j=1}^{d} w_j^2 = \frac{1}{N}||Xw - y||^2 + \lambda\sum_{j=1}^{d} w_j^2 \tag{9}$$

where λ is the regularization parameter. After minimizing the cost function, we get the coefficients as

$$w = (X^T X + \lambda I)^- 1X^T y \tag{10}$$

Another modification of the least square method is the LASSO model. It is valuable due to its affinity to prefer solutions with fewer parameter values, efficiently decreasing the number of variables upon which the given solution is dependent.
The objective function for LASSO can be defined as

$$J(w) = \frac{1}{N}\sum_{i=0}^{N}(h(x_i) - y_i)^2 + \lambda\sum_{j=1}^{d} |w_j| = \frac{1}{N}||Xw - y||^2 + \lambda\sum_{j=1}^{d} |w_j| \tag{11}$$

The added term corresponds to l_1-norm. The lasso estimate thus explains the minimization of the least square penalty with $\lambda||w||_1$ added where λ is regularization parameter and $||w||_1$ is the l_1-norm of the parameter vector.

Data and Methods

The research data utilized for forecasting stock market prices in this study was obtained from Yahoo Finance using the quantmod package in R. The total number of instances considered for this study were 2130 trading days, from December 13th 2010 to June 7th 2019. The data is composed of daily information including adjusted price and volume traded for Apple, adjusted price for stock market indexes; Dow Jones Industrial Average (DJI), Nasdaq Composite (NASDAQ), S&P 500 Index (S&P). The training data set was selected as the first 80% of the sample, while the testing data consisted the remaining 20% of the sample.

I choose an P_t autoregressive AR(N) model of daily prices as the benchmark because of its widespread use in the literature and its relatively good performance in predicting time-series information, particularly prices (e.g. Conejo et al. 2005; Weron 2006; Misiorek et al. 2006).
Specifically, I specify an N^{th} order autoregressive AR(N) model as

$$AP_t = \alpha + \sum_{i=1}^{N}\beta_i AP_{t-i} + \sum_{i=0}^{kN}\gamma_{ki} D_{kt-ki} + \epsilon_t \tag{12}$$

where AP_t is daily adjusted price of Apple, D_k is a vector of exogenous variables including; volume traded for Apple, adjusted price for DJI, NASDAQ, and S&P.
From (12), we note that number of observations $N <$ number of parameters p, where $N = 1704$ and $p = 10649$; thus, cannot be estimated by OLS. I therefore estimate a ridge and LASSO regression and comapre results from the two estimations.

The performance of the ridge and LASSO regression methods were measured by computing root mean square error (RMSE) and the mean absolute percentage error (MAPE).
The RSME is represented as:

$$RSME = \sqrt{\frac{\sum_{i=1}^{N}(y_i - p_i)^2}{N}} \tag{13}$$

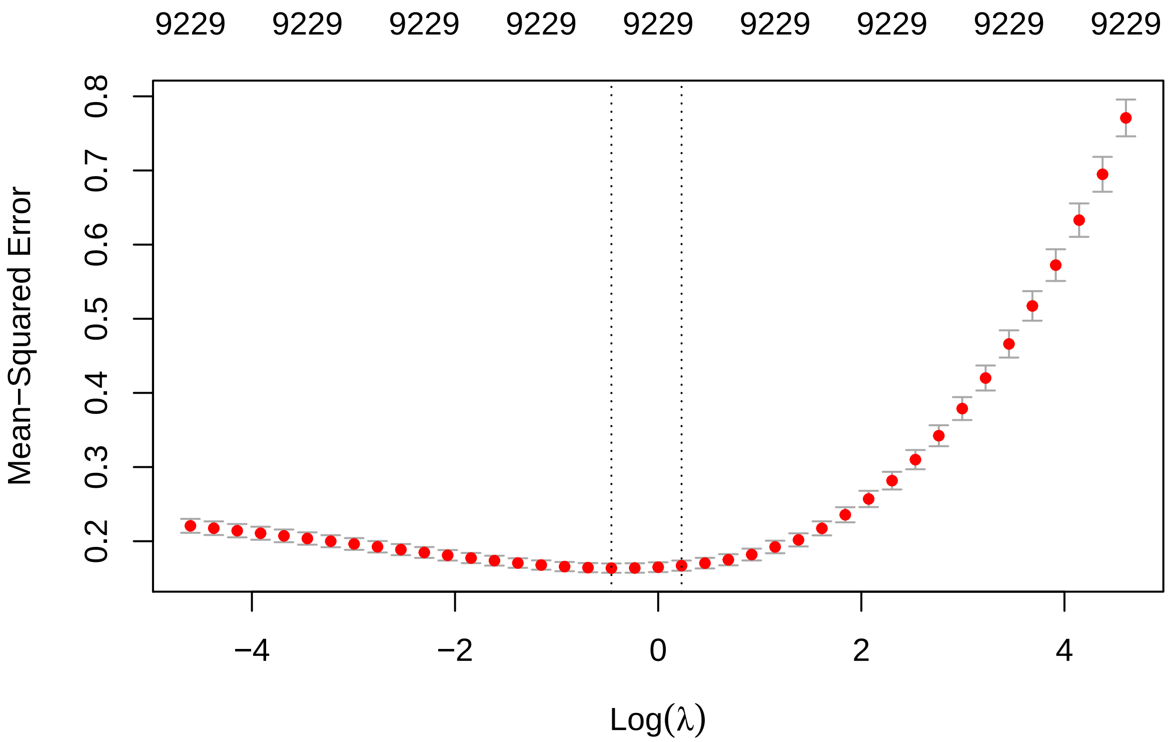
MAPE is found by calculating the absolute value of the variation between the actual stock price and the expected stock price. The MAPE is represented as:

$$MAPE = \frac{\sum_{i=1}^{N} \frac{|y_i - p_i|}{y_i}}{N} 100\% \tag{14}$$

where N is the total number of trading days p_i is the predicted stock price on day i and y_i is the actual stock price on the same day.

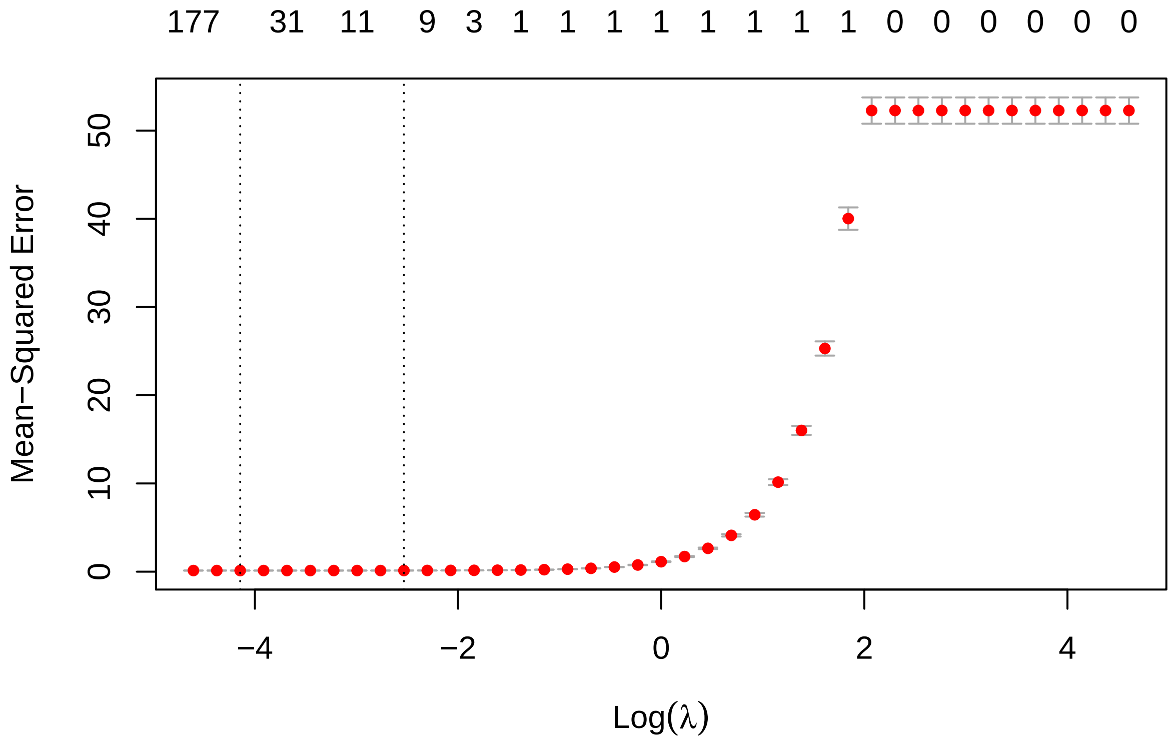
Results and Discussion

Figure 1: CV Plot for Ridge Regression method



The best lambda for ridge regression method is 0.6309573

Figure 2: CV Plot for LASSO Regression method



The best lambda for ridge regression method is 0.0158489

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

The number of predictors included the LASSO model is 99.

Table 1: LASSO Regression Predictors and Coefficients

Predictor	Coefficient
GSPC.Adjusted.lag1362	-0.00006
AAPL.Volume.lag1143	0.00000
AAPL.Volume.lag912	0.00000
AAPL.Volume.lag41	0.00000
AAPL.Volume.lag170	0.00000
AAPL.Volume.lag1137	0.00000
AAPL.Volume.lag1842	0.00000
AAPL.Volume.lag169	0.00000
AAPL.Volume.lag1778	0.00000
AAPL.Volume.lag645	0.00000
AAPL.Volume.lag1283	0.00000

Predictor	Coefficient
AAPL.Volume	0.00000
AAPL.Volume.lag990	0.00000
AAPL.Volume.lag906	0.00000
AAPL.Volume.lag209	0.00000
AAPL.Volume.lag628	0.00000
AAPL.Volume.lag343	0.00000
AAPL.Volume.lag1833	0.00000
AAPL.Volume.lag67	0.00000
AAPL.Volume.lag650	0.00000
AAPL.Volume.lag194	0.00000
AAPL.Volume.lag1151	0.00000
AAPL.Volume.lag219	0.00000
AAPL.Volume.lag629	0.00000
AAPL.Volume.lag225	0.00000
AAPL.Volume.lag1176	0.00000
AAPL.Volume.lag1823	0.00000
AAPL.Volume.lag1148	0.00000
AAPL.Volume.lag1258	0.00000
AAPL.Volume.lag899	0.00000
AAPL.Volume.lag1257	0.00000
AAPL.Volume.lag1813	0.00000
AAPL.Volume.lag1050	0.00000
AAPL.Volume.lag296	0.00000
AAPL.Volume.lag1284	0.00000
AAPL.Volume.lag1611	0.00000
AAPL.Volume.lag1161	0.00000
AAPL.Volume.lag908	0.00000
AAPL.Volume.lag1785	0.00000
AAPL.Volume.lag1152	0.00000
AAPL.Volume.lag1771	0.00000
AAPL.Volume.lag12	0.00000
AAPL.Volume.lag63	0.00000
AAPL.Volume.lag1164	0.00000
AAPL.Volume.lag1506	0.00000
AAPL.Volume.lag931	0.00000
AAPL.Volume.lag1144	0.00000
AAPL.Volume.lag644	0.00000
AAPL.Volume.lag859	0.00000
AAPL.Volume.lag1296	0.00000
AAPL.Volume.lag651	0.00000
AAPL.Volume.lag1770	0.00000
AAPL.Volume.lag176	0.00000
AAPL.Volume.lag647	0.00000
AAPL.Volume.lag177	0.00000
AAPL.Volume.lag1251	0.00000
AAPL.Volume.lag655	0.00000
AAPL.Volume.lag463	0.00000
AAPL.Volume.lag1633	0.00000
AAPL.Volume.lag1394	0.00000
AAPL.Volume.lag1647	0.00000
AAPL.Volume.lag1661	0.00000
AAPL.Volume.lag1634	0.00000

Predictor	Coefficient
AAPL.Volume.lag458	0.00000
AAPL.Volume.lag489	0.00000
DJI.Adjusted.lag92	0.00000
DJI.Adjusted.lag1617	0.00000
DJI.Adjusted.lag89	0.00000
DJI.Adjusted.lag987	0.00000
DJI.Adjusted.lag94	0.00000
GSPC.Adjusted.lag939	0.00000
DJI.Adjusted.lag948	0.00000
DJI.Adjusted.lag983	0.00000
DJI.Adjusted.lag1682	0.00000
DJI.Adjusted.lag988	0.00000
DJI.Adjusted.lag86	0.00000
DJI.Adjusted.lag69	0.00000
DJI.Adjusted.lag938	0.00000
DJI.Adjusted.lag908	0.00001
DJI.Adjusted.lag939	0.00001
DJI.Adjusted.lag907	0.00001
DJI.Adjusted	0.00001
DJI.Adjusted.lag1609	0.00001
DJI.Adjusted.lag1610	0.00001
AAPL.Adjusted.lag43	0.00001
DJI.Adjusted.lag1608	0.00001
GSPC.Adjusted.lag948	0.00002
GSPC.Adjusted.lag940	0.00002
NDAQ.Adjusted.lag920	0.00024
GSPC.Adjusted	0.00030
NDAQ.Adjusted.lag982	0.00045
NDAQ.Adjusted.lag959	0.00045
NDAQ.Adjusted.lag932	0.00202
NDAQ.Adjusted.lag960	0.00213
AAPL.Adjusted.lag3	0.00290
NDAQ.Adjusted.lag921	0.00304
AAPL.Adjusted.lag4	0.00693
AAPL.Adjusted.lag1637	0.02343
AAPL.Adjusted.lag1	0.93867

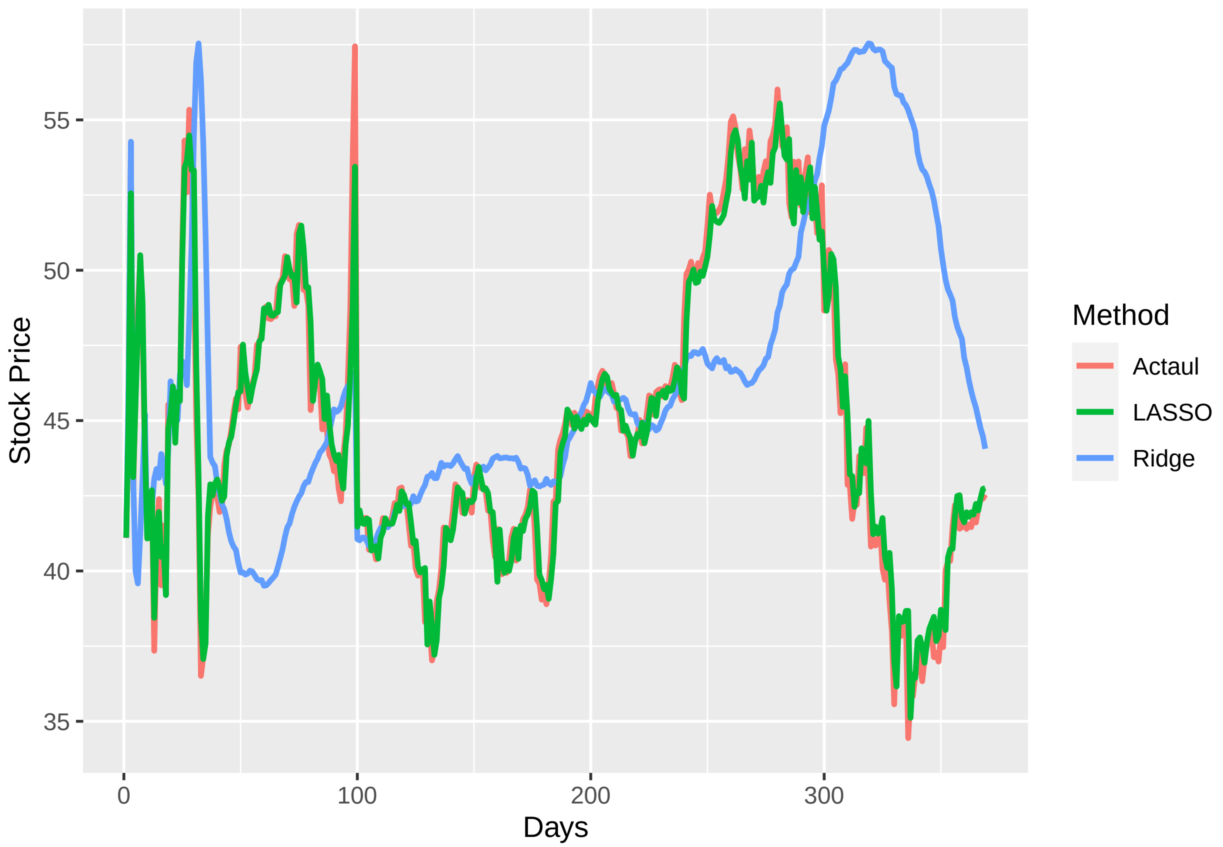
Table 1 presents results of performance measures for both methods. The RMSE for LASSO regression method is 0.9005794 which is less than the RMSE for ridge regression method 7.0922849.

Also, the MAPE for LASSO regression method is 1.3963121 which is less than the MAPE for ridge regression method 11.3515078.

Table 2: Performance Measures

Method	RMSE	MAPE
Ridge	7.0922849	11.351508
LASSO	0.9005794	1.396312

Figure 3: Actual versus Predicted Stock Prices



Conclusion and Future Work

LASSO regression method outperforms ridge regression in predicting Apple stock prices shown by the values of RMSE and MAPE for both methods.
Future work would need to explore more sources of data such financial text data such news the Wall Street Journal. Also, it will be important to explore other machine learning methods such as nueral networks. However, this study provides ample evidence of the importance of such methods especially in situations where OLS is not feasible; in predicting stock prices.

References

1. Bessembinder H. Quote-based competition and trade execution costs in NYSE-listed stocks. Journal of Financial Economics. 2003; 70: 385–422

2. Box GEP, Jenkins GM. Time series analysis: forecasting and control. San Francisco: Holden-Day; 1970

3. Kao LJ, Chiu CC, Lu CJ, Yang JL. Integration of nonlinear independent component analysis and support vector regression for stock price forecasting. Neurocomputing. 2013; 99: 534–542

4. Leigh W, Pussell R, Ragusa JM. Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. Decision Support Systems. 2002; 32: 361–377

5. Lütkepohl H (2005) New introduction to multiple time series analysis. Springer, Berlin

6. Pathak, A.: Predictive time series analysis of stock prices using neural network classifier. International Journal of Computer Science and Engineering Technology, 2229–3345 (2014)

7. Tsai CF, Lin YC, Yen DC, Chen YM. Predicting stock returns by classifier ensembles. Applied Soft Computing. 2011; 11: 2452–2459.