# Stock Market Forecasting Using LASSO Linear Regression Model

Sanjiban Sekhar Roy[1], Dishant Mittal[1], Avik Basu[1], and Ajith Abraham[2,3]

[1] School of Computing Science and Engineering[1], VIT University
Vellore, Tamilnadu, India
[2] IT4Innovations, VSB - Technical University of Ostrava, Czech Republic
[3] Machine Intelligence Research Labs (MIR Labs), Washington 98071, USA
{s.roy,dishant.mittal2011,avik.basu2011}@vit.ac.in,
ajith.abraham@ieee.org

**Abstract.** Predicting stock exchange rates is receiving increasing attention and is a vital financial problem as it contributes to the development of effective strategies for stock exchange transactions. The forecasting of stock price movement in general is considered to be a thought-provoking and essential task for financial time series' exploration. In this paper, a Least Absolute Shrinkage and Selection Operator (LASSO) method based on a linear regression model is proposed as a novel method to predict financial market behavior. LASSO method is able to produce sparse solutions and performs very well when the numbers of features are less as compared to the number of observations. Experiments were performed with Goldman Sachs Group Inc. stock to determine the efficiency of the model. The results indicate that the proposed model outperforms the ridge linear regression model.

**Keywords:** Stock price prediction, LASSO regression.

## 1 Introduction

Prediction of stock price is a crucial factor considering its contribution to the development of effective strategies for stock exchange transactions. The Stock market plays a crucial role in the country's economy. This is due to the fact that stock market helps in flourishing the commerce and industry that ultimately has an effect on the country's economy. Whenever the company requires funds for expanding its business or if it is setting a new venture it has two options. Either a loan can be taken from a financial organization or shares can be issued through the stock market. A company can issue its shares that are in part ownership. For issuing shares for investment in the stocks, a company must get listed in the stock exchange and after this they can accumulate the funds needed for its business. Another important function that the stock market plays is that it provides a generic platform for the sellers and buyers of stocks listed on the stock market. The buyers and sellers are basically retail and institutional investors. These people are the traders who provide funds for the



A. Abraham, P. Krömer, & V. Snášel (eds.), *Afro-European Conf. for Ind. Advancement*,
Advances in Intelligent Systems and Computing 334, DOI: 10.1007/978-3-319-13572-4_31

businesses by investing in stocks. If the stock's future price can be predicted, it can preclude significant losses and can certainly increase the profits.

Recently, the prediction of the stock price has gathered significant interests among investors and in incorporating variable historical series into computer algorithms in order to produce estimations of estimated price fluctuations. Due to the blaring environment the prediction of the stock price becomes very complex. Traders often rely on technical indicators based on stock data which can be collected daily. Though the usage of these indicators provides them some information about the prices, but still it is difficult to have an accurate prediction of daily to weekly trends.

For a person who is not well trained trading of stocks is risky. However a neat pile in quick intraday deals can be made if one has a fixation on spotting the trends in the market. There was a generalized mindset in recent times when depending on the beliefs of people trading was considered as game of buying and selling of stocks. Now, some new tools have been devised by the investors by utilizing a method known as technical analysis for predicting future prices from historical price data. On general technical analysis is based on technical indicators. A technical indicator for the stock price is a function that returns a worth for given stock price in some given span of time in history. Information on whether a trend will continue or whether a stock is oversold and overbought can be got from such technical indicators [1].

Apart from the technical analysis, there is a method known as fundamental analysis, which is concerned with the company that underlies the stock itself. It assesses a company's past performance as well as the trustworthiness of its accounts. Many performance ratios are produced that aid in evaluating the rationality of a stock. With the arrival of the digital computer, stock market prediction has since progressed into the technological world. Artificial neural networks (ANNs) and genetic algorithms are involved in some of the most noticeable techniques. ANNs can be considered as approximations of mathematical functions. The use of ANN mimics how the human brain functions, by serving computers with the immense data to mimic human thinking. The feed forward network using the backward propagation of errors algorithm to update the network weights is the most accepted form of ANN for stock market prediction that is currently in use [10]. These networks are commonly referred to as Back propagation networks. Time delay neural network (TDNN) or the time recurrent neural network (TRN) is another type of ANN that is more convenient in forecasting the stock price.

We propose a system which is based on generalized linear regression model and use it for stock market forecasting. In this paper, we present a model that we implemented for the prediction of stock price based on the LASSO method which outperforms the ridge method and the artificial neural network model in terms of accuracy.

## 2 Related Works

Deng et al. [1] introduced ***a stock price prediction model, which extracts features from time series data and social networks for prediction of stock prices and***

***calculates its performance. The stock price movements were modeled as a function of these input features and was solved as a regression problem in a Multiple Kernel Learning regression framework by them.*** Yoo et al. [2] in their work explored various global events and their concerns on forecasting stock markets. They found that integrating event information with the prediction model plays very significant roles for more exact prediction. Schumann et al. [3] presented two models, namely ARIMA and ANN for the prediction of the stock price. Pakdaman Naeini et al. [4] presented two kinds of neural networks, an Elman recurrent network and a feed forward multilayer Perceptron (MLP) that were in turn utilized to measure a company's stock value based on the record of its stock share value. They demonstrated that the application of MLP neural network is more capable of calculating stock value changes rather than Elman recurrent network and linear regression method. Aseervatham et al. [5] claimed that the ridge logistic regression reaches the same performance as the Support Vector Machine. Ticknor [6] presented a Bayesian regularized artificial neural network as a unique method to estimate the financial market behavior. Data Sets from the corporations like Goldman Sachs Group Inc. and Microsoft Corp. were used to perform experiments. Nair et al. [7] proposed an automated decision tree-adaptive neuro-fuzzy hybrid automated stock market prediction system. Ajith Abraham et al [8] proposed a genetic programming method for forecasting of stock market prices. They experimented on Nasdaq stock market and S and P cnx nifty data. They combined two multiobjective optimization algorithms with techniques like svm and neural networks to get the best results. Yuehui Chen et al. [9] introduced flexible neural tree method for organized representation of stock market data, later they used genetic programming for optimization of this method.

## 3 Generalized Linear Models

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 model in which linear regression fits a linear model with coefficients $w = (w_1, \ldots., w_p)$ 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 $n \geq p$.

### 3.1 Mathematical Formulation

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

$$i = 1,\ldots,N \tag{1}$$

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

$$X = \begin{bmatrix} -x_1^T - \\ -x_2^T - \\ \vdots \\ -x_N^T - \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{4}$$

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

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

which implies

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

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 cost 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 $\boldsymbol{\lambda}$ is the regularization parameter. After minimizing the cost function, we get the coefficients as

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

## 4 Suggested Linear Model

Another modification of the least square method is the LASSO model which stands for Least Absolute Shrinkage and Selection Operator. The suggested model is used for the estimation of sparse coefficients. It is valuable in some backgrounds 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 appropriate group of weights which are not zero can be recovered under certain conditions. Mathematically, a linear model trained with $l_1$ prior as regularize is comprised in it.

To fit the coefficients, the algorithm that is used for the implementation in the class LASSO is coordinate descent [11].

The new objective for LASSO can be defined as

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

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 $\lambda$ is regularization parameter and $\|w\|_1$ is the $l_1$-norm of the parameter vector.

## 5 Experimentation Results

The research data utilized for predicting stock market prices in this study was gathered for Goldman Sachs Group, Inc. (GS). The total number of instances considered for this study were 3686 trading days, from 4 May 1999 to 3 January 2014. Each of the samples composed of daily information including low price, high price, opening price, close price, and trading volume. The training data set was selected as the first 70% of the samples, while the testing data consisted the remaining 30% of the samples. The LASSO model was used to predict the price of the chosen stock for the future days.

A comparison study was performed to test the efficiency of the model suggested in this study to another linear regression model that is named as Ridge and a Bayesian regularized artificial neural network by Jonathan L. Ticknor. For this method, the dataset was gathered for Goldman Sachs Group, Inc. In total number of instances considered for this study were 734 trading days, from 4 January 2010 to 31 December 2012. The training data set was chosen as the first 80% of the samples and the remaining 20% was used as testing dataset.

### 5.1 Algorithm

```
LASSO REGRESSION()

   1. data ←read ('data.csv')

   2. (train_features,train_stock_price)←
      training_function()

   3. (test_features,test_stock_price)←testing_function()

   4. Model←LASSO_train(train_features
      ,train_stock_price, lambda)

   5. stock_price_predict← LASSO_predict(train_features)

   6. MAPE ← mean [abs{(test_stock_price -
      stock_price_predict)/test_stock_price}] * 100

   7. RMSE ← sqrt [mean{(test_stock_price -
      stock_price_predict)^2}]
```

### 5.2 Results and Discussion

The performance of the LASSO Linear regression method was measured by computing root mean square error (RMSE) and the mean absolute percentage error (MAPE). These performance metrics have been used in a number of studies and ensures an effective means of deciding the robustness of the model for predicting daily. It can be represented as :

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - p_i)^2}{n}} \tag{12}$$

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.

The Mean Absolute Percentage error (MAPE) metric is first found by calculating the absolute value of the variation between the actual stock price and the expected stock price. The MAPE value is calculated using the following equation.

MAPE formula:

$$\text{MAPE} = \frac{\sum_{i=1}^{n}\frac{|y_i - p_i|}{y_i}}{n} \times 100\% \tag{13}$$

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.

**Table 1.** RMSE and MAPE of training and test set (Ridge vs LASSO)

| **Method** | **Training RMSE** | **Test RMSE** | **Training MAPE** | **Test MAPE** |
|---|---|---|---|---|
| Ridge | 1.7648 | 3.2272 | 1.3028 | 1.8065 |
| LASSO | 1.1403 | 2.5401 | 0.9304 | 1.4726 |

**Table 2.** Forecast accuracy comparison with Jonathan L. Ticknor

| Method | Training MAPE (%) | Testing MAPE (%) |
|---|---|---|
| Bayesian Regularized ANN | 1.5235 | 1.3291 |
| LASSO | 0.1806 | 0.6869 |

Table 1 presents the experimental results for the two methods (Ridge and LASSO) chosen for prediction over the Goldman Sachs (GS) dataset .The RMSE and MAPE values were calculated for the training and testing dataset to monitor the effectiveness of the models. It was deduced that the testing set MAPE of LASSO regression is less than the testing set MAPE of the Ridge regression method indicating that LASSO method is better than Ridge. The same results are reflected in the graphs from Fig. 1-6. This can be said according to the proximity of the regression line for each graph to the data points.

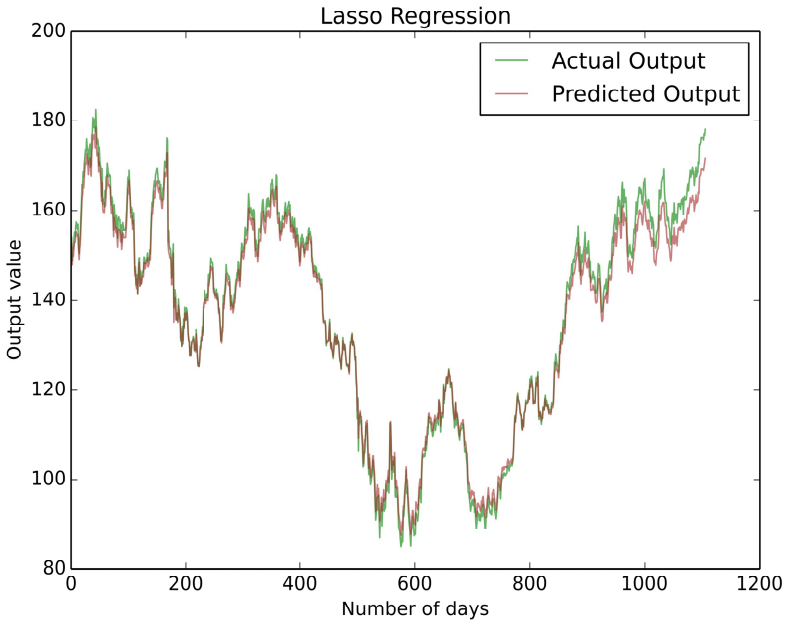


**Fig. 1.** Future day prediction (LASSO)

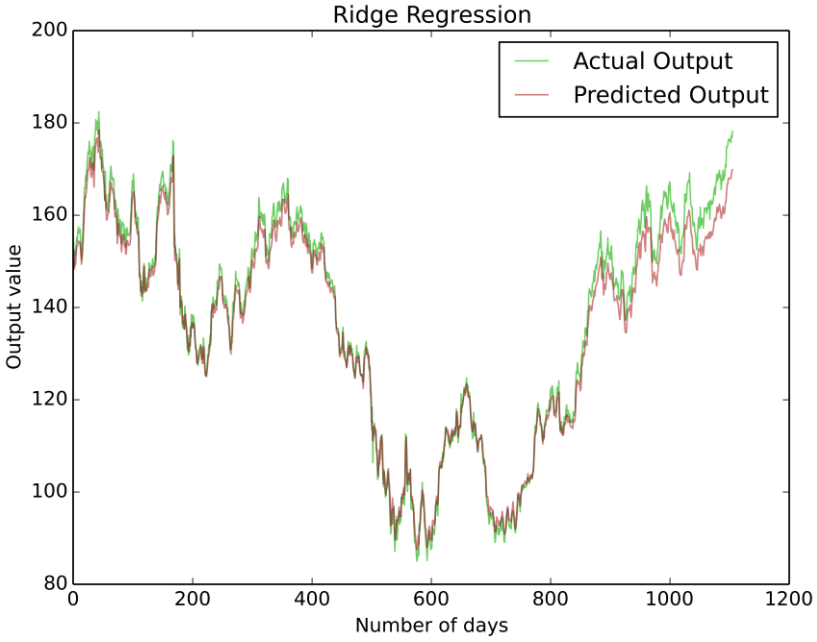


**Fig. 2.** Future day prediction (Ridge)


**Fig. 3.** Target vs predicted stock price using training set (LASSO)

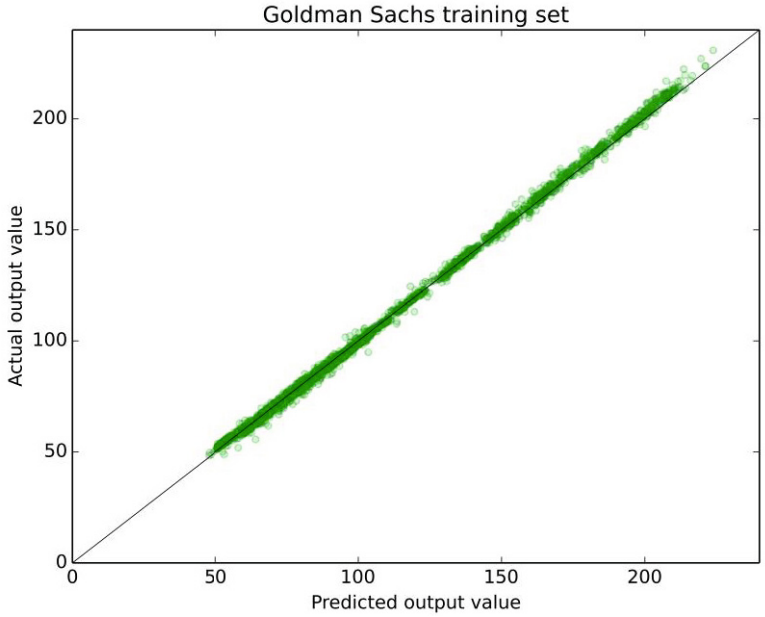


**Fig. 4.** Target vs predicted stock price using training set (Ridge)

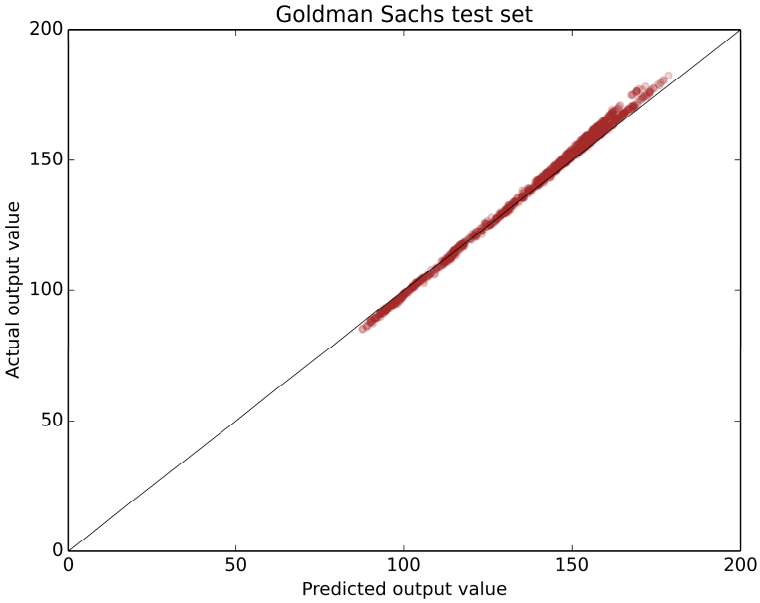


**Fig. 5.** Target vs predicted stock price using test set (LASSO)

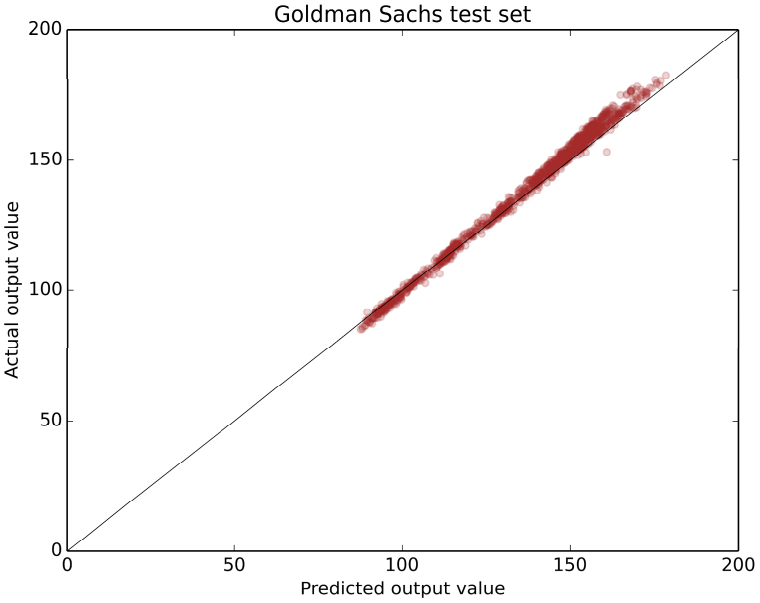


**Fig. 6.** Target vs predicted stock price using test set (Ridge)

## 6 Conclusions

Empirical results indicated that the model outperformed the ridge linear regression model and Bayesian regularized artificial model. The model resulted in a MAPE of 0.6869 with respect to 1.3291 that Bayesian artificial neural network method produced for the mentioned dataset. A MAPE value of 1.4726 and RMSE value 2.5401 was produced by using LASSO algorithm, whereas MAPE value 1.8065 and RMSE value 3.2272 was reported by utilizing ridge regression algorithm with respect to the mentioned dataset for 3686 instances. To evaluate the effectiveness of this model, the network was successfully compared with a Bayesian regularized artificial neural network model. Forecast of stock market drifts is very important for the development of effective trading policies.

## References

1. Deng, S., Takashi, M., Kei, S., Tatsuro, S.: Akito Sakurai.: Combining technical analysis with sentiment analysis for stock price prediction. In: IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC), pp. 800–807. IEEE (2011)
2. Yoo, P.D., Kim, M.H., Jan, T.: Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In: International Conference on Computational Intelligence for Modeling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, vol. 2, pp. 835–841 (2005)
3. Schumann, M., Lohrbach, T.: Comparing artificial neural networks with statistical methods within the field of stock market prediction. In: Proceeding of the Twenty-Sixth Hawaii International Conference on in System Sciences, vol. 4, pp. 597–606. IEEE (1993)
4. Naeini, M.P., Taremian, H., Hashemi, H.B.: Stock market value prediction using neural networks. In: 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 132–136. IEEE (2010)
5. Aseervatham, S., Antoniadis, A., Gaussier, E., Burlet, M., Denneulin, Y.: A sparse version of the ridge logistic regression for large-scale text categorization. Pattern Recognition Letters 32(2), 101–106 (2011)
6. Ticknor, J.L.: A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications 40(14), 5501–5506 (2013)
7. Nair, B.B., Minuvarthini, M., Sujithra, B., Mohandas, V.: Stock market prediction using a hybrid neuro-fuzzy system. In: 2010 International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom), pp. 243–247. IEEE (2010)
8. Abraham, A., Grosan, C., Han, S.Y., Gelbukh, A.: Evolutionary multiobjective optimization approach for evolving ensemble of intelligent paradigms for stock market modeling. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, pp. 673–681. Springer, Heidelberg (2005)
9. Chen, Y., Yang, B., Abraham, A.: Flexible neural trees ensemble for stock index modeling. Neurocomputing 70(4), 697–703 (2007)
10. Pathak, A.: Predictive time series analysis of stock prices using neural network classifier. International Journal of Computer Science and Engineering Technology, 2229–3345 (2014)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-Learn: Machine Learning in Python. JMLR Journal of Machine Learning Research, 2825–2830 (2011)