Alternative prediction methods in the stock exchanges of Thailand

Kanchana Chokethaworn, Chukiat Chaiboonsri and Satawat Wannapan¹
Faculty of Economics, Chiang Mai University, Thailand

¹Email: lionz1988@gmail.com

Abstract. This paper is conducted to substantially do the two alternatives between the traditional statistical methods such as ARMA and HW models, and artificial intelligence (AI) contains KNN and ELM, respectively. To scope the main object of the paper, SET indexes are collected as the main financial variable, which are 5,472 daily observed days during 9 September 1997 to 11 June 2018. Technically, the cross-entropy (CE) analysis, MSE and RMSE calculations are computationally employed to clarify the resolution of the two computations. The empirical results state that the AI prediction can be a substitution replacing the traditional estimations, and this can strongly confirm that machine learning (ML) algorithms are continuously interested, and they are recently becoming a powerful tool for modern econometric forecasting.

1. Introduction
Undeniably, Thailand economic system has been driven by financial sectors along thirty years after the great financial crisis in 1998. The financial market and banking sectors in Thailand are recently becoming stronger as well as skilled investors are continuously increased. They inevitably need a sharp tool to guarantee their prediction. However, an enormous amount of information causes the forecasting to be elusive, and many disadvantages of P-value are likely to make predictive results become suspicious. Moreover, economists and financial predictors have been struggling to employ traditionally statistical computations to provide their results for stock selections and risk classifications. Consequently, the development of methods to more precisely predict and make the financial market’s trends become sensible is very crucial in financial analyses.

Interestingly, a novel concept called “Machine Learning (ML)” is to use a type of computation-al artificial intelligence (AI) that learns when exposed to new data and let the machine deal with data by itself. The implemented machine learning classifiers named K-Nearest Neighbor (KNN) and Extreme Learning Machine (ELM) have been employed to overcome restrictions in traditional statistical forecasting. The performance of these classifiers has been compared using different matrices that include mean absolute error, root mean square error and accuracy validations. Considering into the computation, although ML has been being continuously applied to predict stock markets, but the challenging question is how we can substantially present the differences between AI computations and traditional manual estimations, especially in the financial econometric field. Consequently, this paper is determinedly conducted to solve the question. Moreover, since the advancement and availability of forecasting technologies and software, AI and traditional computations can be quickly provided for clarifying the prediction in one machine.
Historically, machine learning algorithms have been frequently used to be the predictor for financial sectors since several years ago. For instance, [1], [2], [3], [4], and [5]. On the other hand, the prediction by traditional statistics, employing the ARMA model, has been famous for several decades, especially in financial econometrics. For example, [6], [7], [8]. As seen in literature, the two total different methods for stock forecasting is rare to see that they are either substituted, or which one is sensible to apply for real predictions, or they are both to be used. Hence, the first propose of this paper is to evaluate the forecasting estimations between an AI algorithm and traditionally statistical tool in the stock exchange of Thailand by using both MSE and RMSE based on out-of-sample testing, and the second is to compute the cross-checking application called “frequenting cross-entropy calculation” for giving a suggestion to possible select the collection of econometric tools for financial forecasting.

2. Research methodologies

The method of the paper is divided into two major parts, including data classification and data prediction. The former section is to employ the data stationary test, ADF unit testing, based on Bayesian inference. From the Bayes’ theorem, we obtain the posterior density $\pi(\theta|y)$ that refers to a probability distribution of parameters $\theta$ as [9],

$$\pi(\theta|y) \propto L(y|\theta) \pi(\theta).$$  \hspace{1cm} (1)

As shown in Equation (1), $\pi(\theta)$ is the prior density for $\theta$. Bayesian modeling, as mentioned, requires a joint distribution, which is conveniently factored into a prior distribution for the parameters, and the complete-data likelihood function expressed in Equation (2).

$$L(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y_i^2}{2\sigma^2}\right).$$ \hspace{1cm} (2)

where $y$ refers to the time-series data of $n$ observations, $y = (y_1, y_2, ..., y_n)$ and $\theta$ stands for the estimated parameters. According to [10] and [11], Bayesian statistics considers hypotheses regarding multiple parameters by adapting Bayes factor comparisons. Expressly, the ADF test analyzes the null hypothesis that a time-series data is I(1) against the alternative I(0), assuming that the dynamics in the data have an ARMA structure [12]. The ADF test is relied on estimating the regression test, which is presented in equation (3).

$$y_t = c + \alpha D_t + \phi y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + \varepsilon_t,$$ \hspace{1cm} (3)

The latter part is to provide the various computations of traditional econometric models and AI algorithms. Fundamentally, the analysis of time series with ARMA (p,q) models, possibly overlapping uses in financial data. The ARMA model including the measurement of errors can be expressed as the following equation [13].

$$(x_t - \mu) = \sum_{j=1}^{p} \beta_j (x_{t-j} - \mu) + \sum_{j=0}^{q} \alpha_j \varepsilon_{t-j}$$ \hspace{1cm} (4)

Another traditional statistical regression is the Holt-Winters exponential smoothing (HW) method. This application is extended to be able to accommodate multi-seasonal patterns [14]. The relationship of these three components can be expressed by the following equations.

$$\mu_t = \alpha \left(\frac{X_t}{I_{t-s}}\right) + (1-\alpha)(\mu_{t-1} + T_{t-1}).$$ \hspace{1cm} (5)

$$T_t = \gamma (\mu_{t-1} - \mu_{t-1}) + (1-\gamma)T_{t-1}.$$ \hspace{1cm} (6)

$$I_t = \delta \left(\frac{X_t}{\mu_t}\right) + (1-\delta)I_{t-s}.$$ \hspace{1cm} (7)

$$\hat{X}_t(h) = (\mu_t + h T_t)I_{t-s+h}.$$ \hspace{1cm} (8)
It is notable that the three equations require constant smoothing, \( \alpha \), \( \gamma \), and \( \delta \). They take values between 0 and 1. The goal of Holt-Winters smoothing of series is to use the same for forecasting future time series values [15].

Speaking to Machine Learning (ML) methods, the analysis of machine learning algorithms is well known as computational learning theory. In this paper, R software is the major application for the computation. There are essential approaches and crucial processes for ML described as follows:

- The decision tree learning.
- An artificial neural network (ANN) learning algorithm.
- Inductive logic programming (ILP).
- Support vector machines (SVMs).
- Clustering analysis for unsupervised learning, and a common technique for statistical data analysis.
- Bayesian networks for performing algorithm inferences and learning.
- Representation learning (RL) for underlying factors of variation that explain the observed data [16].
- Genetic Algorithm (GA).
- Rule-based machine learning is a contrast model to other machine learners that commonly verifies a singular model for predicting [17].
- Feature selection approaches [18].

For doing the computation of machine learning approaches in this paper, namely, \( k \)-nearest neighbors (kNN) and Extreme Learning Machine (ELM) are selected to be the algorithms in AI computations for financial data. The former is a non-parametric method for both classification and regression problems. Data is performed as its \( k \)-nearest to others in the feature space. For solving the problem, data observations are assigned to the majority class of its nearest observations.

The latter is the extreme learning machine (ELM). ELM was originally developed for the single-hidden-layer feedforward neural networks and then extended to the “generalized” SLFNs which may not be neuron alike [19].

\[ f(x) = h(x)\beta, \]  

(9)

where \( h(x) \) is the hidden-layer output corresponding to the input sample \( x \) and \( \beta \) is the output weight vector between the hidden layer and the output layer [20]. For proposing the learning algorithm for SLFNs, a simply extreme learning machine can be summarized as follows [21]:

**Algorithm ELM:** Given a training set \( \mathcal{N} = \{x_i, t_i\} | x_i \in \mathbb{R}^n, t_i \in \mathbb{R}^m, i = 1, \ldots, N \}, \) an activated function \( g(x) \) and hidden node number \( \tilde{N} \) can follow the upcoming steps. Step 1 is to randomly assign input weight \( w_i \) and bias \( b_i \), \( i = 1, \ldots, \tilde{N} \). Step 2 is to calculate the hidden layer output matrix \( H \). Step 3 is to calculate the output weight \( \beta \), \( \beta = H^T T \), where \( T = [t_1, \ldots, t_N]^T \) [20], [21].

For selecting AI and traditional statistical predicting models is provided by the value of the mean square error (MSE) and root mean square error (RMSE). Additionally, this section regards the model validation by employing the frequenting cross-entropy analysis, which can help to select the predicting model easier. The formula is explained in Equation (12).

\[ H(X, Y) = -\sum p(x) \times \log 2(g(y)), \]  

(10)

where \( H(X, Y) \) is cross entropy, which can measure the value of relatedness between the random variable \( X \) and \( Y \). \( P(x) \) is the probability of occurrence of outcome \( x \) of variable \( X \), and \( P(y) \) is the probability of occurrence of outcome \( y \) of variable \( Y \) [22].
3. Empirical results

3.1. Descriptive data
Fundamentally, collected financial trends were observed as daily data during 1997 to 2018 (5,742 observations). Data was transformed to be the rate of log return values, and it is confirmed to be the normal distribution by Jarque-Bera normality testing. The graphical feature and normal distribution testing are shown in Figure 1.

![Figure 1. Descriptive information and normality testing.](image_url)

3.2. The result of data stationary checking based on Bayesian inference
The ADF unit-root testing based on Bayes statistics is employed to guarantee that suspicious information would not be occurred during the long-run period. However, the result stated that the empirical value calculated from the Bayes factorial comparison between the non-stationary model (Model1) and stationary model (Model2) is less than 0.01, which equals 0.000866. The result is also represented in Table 1.

| Index | Hypothesis | Posterior odds ratio | Interpretation | Result |
|-------|------------|----------------------|----------------|--------|
| SET   | $H_0 (M_1)$: Non-stationary data (Model1) | 0.000866 | Strong evidence of $M_j$ | Stationary data |

![Table 1. The Bayesian ADF testing for data stationary.](image_url)

3.3. The forecasting resolutions estimated by traditional econometric models and AI algorithms
Interestingly, the empirical result of the evaluation shows the AI prediction can be substituted the traditional estimation. In other words, both MSE and RMSE values estimated from the KNN regression are near to the points of the counterpart, which are 0.00000719 and 0.00268160, respectively (as seen details in Table 2). These values are close to the predictive results estimated by the ARMA model and Holt-Winters, which are 0.000007205, 0.002684199, 0.0000073, and 0.0027049, respectively. However, the NNT model is not suitable for predicting in this case.
Table 2. The forecast evaluation by MSE and RMSE based on out-of-sample testing.

| Times D/M/Y | SET index | Traditional Forecasting Approach | AI Forecasting Approach |
|-------------|-----------|---------------------------------|------------------------|
|             |           | ARMA model                       | Holt-Winters           | KNN regression | NNT (ELM) |
| 8/1/18      | 0.00513   | -0.00000107764607840043          | 0.0004013              | 0.00026474    | -0.0085881 |
| to 9/3/18   | -0.000093 | -0.00000000000000000023          | 0.0004013              | 0.00000784    | -0.0016536 |
| Out of Samples | MSE | 0.000007205 | 0.0000073 | 0.00000719 | 0.0000291 |
|             | RMSE     | 0.002684199 | 0.0027049 | 0.00268160 | 0.0054026 |

3.4. The forecasting resolutions estimated by traditional econometric models and AI algorithms
Another essential part of the paper is the cross-entropy (CE) analysis, which is employed to clarify the collection of predicting models. In the case of traditional modeling prediction, the result presented in Table 3 stated the values (P,Q) of CE is 0.33487, and the cross values from the ARMA model is closer the CE point than the cross value of Holt-Winters. The ARMA model should be selected. In the case of AI forecasting, the result showed the values (P,Q) of CE is 0.35059, and the cross values from the KNN algorithm is closer the CE point than the cross value of the NNT algorithm. Thus, the KNN model should be selected and it can potentially substitute the traditional models.

Table 3. Cross entropy checking for traditionally statistical models in the regression.

| ARMA model | Holt-Winters |
|------------|--------------|
| pi         | -pi*LOG(pi,2) |
| 0.125      | -0.375       |
| 0.875      | -0.168564443 |
| Entropy    | 0.5436       |

Cross entropy (P,Q) (M1,M2): CE = 0.33487***

| KNN algorithm | NNT (ELM) algorithm |
|---------------|---------------------|
| pi            | -pi*LOG(pi,2)      |
| 0.208333      | -0.471465501       |
| 0.791667      | -0.266819365       |
| Entropy       | 0.7383              |

Cross entropy (P,Q) (M1,M2): CE = 0.35059***

4. Conclusions
The complicated issue regarding the possibility of the replacement AI algorithms to do the econometric prediction is successfully clarified in this paper. To answer the argument which states these artificial computations can be substantially substituted traditional statistical applications, the comparison by using MSE and RMSE based on out-of-sample testing and the cross-entropy (CE) analysis are employed. Experimentally, the financial data trends of Thailand Stock Exchange (SET) were observed as daily details during 9 September 1997 to 11 June 2018. This data can be typed as the big data analysis. Empirically, the results were obvious that the trend using AI algorithms to do an efficient prediction is recently being interested. The AI calculations can be effectively replaced the counterpart to compute linear forecasting. This can be supported by the error terms estimated by both AI algorithm (KNN) and ARMA model are very similar. Moreover, the CE analysis simultaneously confirmed the KNN algorithm and ARMA model are selected to be the most efficient choice for predicting the collected data. However, the restriction of the paper is to estimate only univariate
variables. It is interested and it should be the future study on employing AI algorithms and traditional models to estimate the multi-dimensional prediction in the big-data analysis.

References
[1] Hegazy O, Soliman O S and Salam M A 2013 A machine learning model for stock market prediction. International Journal of Computer Science and Telecommunications 4(12) pp. 17-23
[2] Cocianu C L and Grigoryan H 2016 Machine learning techniques for stock market prediction: a case of study of OMV petrom. Economic Computation and Economic Cybernetics Studies and Research 50(3) pp. 63-82
[3] Pinheiro L D S and Dras M 2017 Stock market prediction with deep learning: a character-based neural language model for event-based trading In Proceedings of Australasian Language Technology Association Workshop pages 6-15
[4] Ghazanfar M A, Alahmari S A, Aldhafiri Y F, Mustaqueen A, Maqsood M and Azam M A 2017 Using machine learning classifiers to predict stock exchange index. International Journal of Machine Learning and Computing 5(2) pp. 24-29
[5] Macchiariulo A 2018 Predicting and beating the stock market with machine learning and technical analysis Journal of Internet Banking and Commerce 23(1) pp. 1-22
[6] Mathew O O, Sola A F, Oladiran B H and Amos A A 2013 Prediction of stock price using autoregressive integrated moving average filter ((ARIMA (P,D,Q)) Global Journal of Science Frontier Research 13(8) pp. 78-88
[7] Saini N and Mittal A K 2014 Forecasting volatility in India stock market using state space models Journal of Statistical and Econometric Methods 3(1) pp. 115-136
[8] Groda B, Vrbka J, Váchal J, Vochozka M and Horák J 2017 Prediction of stock price developments using the Box-Jenkins method. SHS Web of Conferences 39 pp. 1-7
[9] Takaishi T 2010 Bayesian inference with an adaptive proposal density for GARCH models Journal of Physics: Conference Series 221
[10] Kass R E and Raftery A E 1995 Bayes factors Journal of the American Statistical Association 90 pp. 773-795
[11] Chen C W S, Chen S Y and Lee S 2013 Bayesian unit root test in double threshold heteroskedastic models. Comput Econ 42 pp. 471-490
[12] Said S E and Dickey D 1984 Testing for unit roots in autoregressive moving-average models with unknown order Biometrika 71 pp. 599-607
[13] Ives A R, Abbott K C and Ziebarth N L 2010 Analysis of ecological time series with ARMA(p,q) models Ecology 91(3) pp. 858-871
[14] Shahin A A 2016 Multiple seasonal Holt-Winters exponential smoothing to predict cloud resource provisioning International Journal of Advanced Computer Science and Applications 7(11) pp. 91-96
[15] Lepojević A and Andelković-Pešić M 2011 Forecasting electricity consumption by using Holt-Winters and seasonal regression models Series: Economics and Organization 8(4) pp. 421-431
[16] Bengio Y 2009 Learning deep architectures for AI Now Publishers Inc. pp. 1-3. ISBN 978-1-60198-294-0
[17] Bassel G W, Glaab E, Marquez J, Holdsworth M J and Bacardit J 2011 Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets. The Plant Cell. 23 (9) pp. 3101-3116
[18] Chakraborty C and Joseph A 2017 Machine learning at central banks. Staff Working Paper No. 647. Bank of England
[19] Huang G B, Zhou H, Ding X and Zhang R 2012 Extreme learning machine for regression and multiclass classification. IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS-PART B: CYBERNETICS 42(2) pp. 513-529
[20] Albadr M A A and Tuin S 2017 Extreme learning machine: a review. International Journal of Applied Engineering Research 12(14) pp. 4610-4623

[21] Huang G B, Zhu Q Y and Siew C K 2006 Extreme learning machine: Theory and applications. Neurocomputing 70 pp. 489-501

[22] Shah V H 2007 Machine learning techniques for stock prediction. Foundations of Machine Learning Courant Institute of Mathematical Science New York University