A Machine Learning Model for Healthcare Stocks Forecasting in the US Stock Market during COVID-19 Period

Prapatchon Jariyapan¹, Jittima Singvejsakul² and Chukiat Chaiboonsri³

1 King Mongkut's Institute of Technology Ladkrabang, Business School
2 Department of Agricultural Economy and Development, Faculty of Agriculture, Chiang Mai University, Chiang Mai 50200, Thailand
3 Faculty of Economic, Chiang Mai University, Chiang Mai 50200, Thailand
E-mail: Jittimavsk@gmail.com

Abstract This paper study the nowcasting and forecasting for the healthcare stock price in the united states during the Covid-19 period including the google trend data information. The data is collected in monthly data from 2015 to 2020 which are five interested stock price indexes in the healthcare sector. Empirically, the finding reveals that the Bayesian structural time series analysis can be used to investigate the stock price indexes with the google trend data is becoming useful for the prediction in term of current movement. In term of the machine learning algorithms, the unsupervised learning k-Mean algorithm is employed to cluster the cycle regimes of the stock market which provided three regimes such as Bull market, Sideways and Bear market. There are twenty-nine months stand for bull market, thirty-seven months are predictively provided sideways market and five months are referred as the bear market. Additionally, the supervised learning algorithms by using the Linear Discriminant Analysis (LDA), k-Nearest Neighbors (kNN) and Support vector machine (SVM) are used to investigate the cycle regimes of healthcare stock in next five year. The results indicated that LDA is chosen by the highest coefficient validation which represented the regimes of stock in the healcare sector of the unites states of America will stay on the sideways periods in the next five years. Thus, the finding in this paper can be the useful information for investor to manage their portfolio especially, in healthcare sector during the Covid-19 period.

1. Introduction
The stock market forecasting is regarded as an important choice for the investors to make their decision. The forecasting in the stock market has been emphasized for many years because it can earn the profit. Interesting, there are many factors that affect the consequence of the return in the stock market including the financial crisis, random walk behavior, trader expectation and the unexpected situation. That the problem from COVID-19 pandemic, happened on 31 December 2019 from the identity of the World Health Organization (WHO) for the first case in China. Many countries around the world face this problem, suffering both the economic and health sector. Thus, the stock index in the many countries has been decreased significantly such as the US stock market fell by 20% as well as the Nikkei (Tokyo stock market) and Shangha (China stock market) has undergo a dropped price in their stock index since mid-March 2020. The composite stock index in US also decline with the concern in the effect of the COVID-19 pandemic on the economy. However, the COVID-19 pandemic is helping the improvement of innovation in biotech toward the biopharma industry and the approach of the healthcare sector reform. The healthcare sector has been the leader return sector during the COVID-19 pandemic, regarding many investors. Interestingly, many people worldwide have experienced with the
COVID-19, the need for vaccine or treatment has become increasing. Then, the increasing effort to handle COVID-19 could also improve the perception of people for the biopharma industry, which in recent years has drawn scrutiny because of rising drug prices. As well as the report from the U.S.A that a more positive view of the healthcare sector before the pandemic began. Therefore, the healthcare stock has outperformed other sector in the stock market. This momentum could continue, especially as the outbreak helps accelerate growth trends in healthcare and, long term, potentially improves sentiment toward the sector. However, the events affect the change of stock price have different characteristics. Unexpected situation that can happen any time, like the COVID-19 pandemic, economic embargoes and the trend of investor expectation may drastically influence the stock price index.

Historically, there several studies are considered the machine learning approach to predict the stock market. The study of A Hybrid Machine Learning System for Stock Market Forecasting which introduced Genetic Algorithm (GA) and Support Vector Machines (SVM) by [1] to predict the stock market price. The accuracy of estimate is found by the different between the actual price and the forecast price. The study found that the hybrid machine learning provided better result than the stand-alone machine learning. [2] study a machine learning model for stock market prediction to maximize investor’s gain. The machine learning is proposed to predict the stock market price by using the particle swarm optimization (PSO) and Least square support vector machine learning (LS-SVM). The study showed that the proposed model outperforms in term of the accuracy of the prediction and the PSO has the potential optimization of LS-SVM. In order to study the 21 internal and external shocks of the European market using the Least Absolute Shrinkage and Selection Operator (LASSO) which outperform the traditional regression methods to the problem of limited data and multi-collinearity. The results indicated that the European stock market was affect by gold market before the pandemic announcement by WHO while Germany and France are the most predictors in the stock market after the announcement. [3] study the nowcasting and forecasting for macroeconomic cycle in Thailand by using machine learning algorithms to understand the Thailand economic cycles. The study employed The Bayesian structural time-series (BSTS), k-Mean clustering, Linear Discriminant Analysis (LDA) and k-Nearest Neighbors (kNN) which is the unsupervised and supervised machine learning. The results showed that the finding from these two approaches is different, the supervised machine learning found the expansion period are the predictive sign of Thailand but the unsupervised machine learning or kNN found that the recession period are the signs for Thailand cycles for next three years.

It is undeniable that there is lack of observation in the time series data which is the major concern for analyzing in the economics. The big data is becoming more important which is currently being used in different industries, including the stock market trading. Big data has impacted on the financial sector in term of the changing investment, the huge volume of data is engaged in different way of trading. It helps the trader to make better decision to invest the stocks which are consistent in the process. Consequently, the major stocks in healthcare sector such as ACHC (Acadia Healthcare Company), BIO (Bio-Rad Laboratories), INO (Inovio Pharmaceuticals), QDEL (Quidel Corporation) and SAGE (SAGE Therapeutics) from the Nasdaq stock market in US are employed to statistically explore up-to-date sets of stock market variable with google trend by Nowcasting calculation and forecasting the stock market cyclical movement of monthly data from 2015 to 2020 by using machine learning model. In order to modernize econometric investigation, the machine learning model is used to estimate with the flexibility for handling with some losing information, the huge volumes of data and the decreasing problem of multicollinearity problem. Therefore, the machine learning model including supervised and unsupervised approach are used to investigate the healthcare stocks cyclical stage and validating the data for better forecasting.

2. Data review
For the data review, Table 1 represent the stock price in the healthcare sector from the Nasdaq index to be the observed data for the learning model. The range of data is in the period of 2015 to 2020 which the data were transformed to log return.
Table 1. The log return of stock price in the healthcare sector information

| Healthcare stocks                     | Code  | Range       | Source          |
|---------------------------------------|-------|-------------|-----------------|
| Acadia Healthcare Company             | ACHC  | 2015 – 2020 | Nasdaq Index    |
| Bio-Rad Laboratories                  | BIO   | 2015 – 2020 | Nasdaq Index    |
| Inovio Pharmaceuticals Inc            | INO   | 2015 – 2020 | Nasdaq Index    |
| Quidel Corporation                    | QDEL  | 2015 – 2020 | Nasdaq Index    |
| Sage Therapeutics                     | SAGE  | 2015 – 2020 | Nasdaq Index    |

In order to predict the economic cycle in the machine learning model, the behavior factor is needed to be estimated. For the technology of databases in the google trends is important one to be used to be the one factor in data prediction which are represent in Table 2.

Table 2. The collected information of behavior factor searching databases from google trend

| Behavior factors                  | Code       | Range       | Source          |
|-----------------------------------|------------|-------------|-----------------|
| Investment                        | Invest     | 2015 – 2020 | Google Trends   |
| Return on investment              | ROI        | 2015 – 2020 | Google Trends   |
| Nasdaq                            | Nasdaq     | 2015 – 2020 | Google Trends   |
| Health stock                      | Health stock| 2015 – 2020| Google Trends   |
| Stock market and COVID-19         | Stock &Covid| 2015 – 2020| Google Trends   |

3. Research methodology

3.1. The Bayesian structural time series (BSTS) analysis

The Bayesian structural time series (BSTS) is proposed [4] for the adjusted non-seasonally in order to predict the united states unemployment benefits with the behavioral indexes from google trends as the effect factor in the BSTS analysis. Furthermore, [5] introduced a Bayesian structural time series Forecasting (BSTSF) model applying the time series data for a sustainable technical analysis in the regression model. The BSTS model can be explain based on the Bayes theorem as follows [6]

\[
p(\theta | x) \propto P(x | \theta)P(\theta),
\]

where \( x \) is observed information and \( \theta \) is the model parameter. \( P(x | \theta) \) refer the prior and likelihood functions. \( P(\theta | x) \) is the posterior function. The observed data \( x \) given \( \theta \) is the updated learning for a likelihood. The prior setting is the normal distribution which can be expressed as the Bayesian regression as follows

\[
y = \beta_0 + \sum_{j=1}^{n} \beta_jx_j + e
\]

where \( e \) is the normal distribution with mean = 0 and variance = \( \sigma^2 \). The prior of \( \sigma^2 \) is relied on an inverse-chi square (\( Inv - \chi^2 \)) distribution that can be showed as

\[
\sigma^2 \sim Inv - \chi^2(n - p, s^2).
\]
\( n \) and \( p \) are data and parameter sizes, respectively where \( s^2 \) is given by

\[
s^2 = \frac{1}{n-p} (y - \bar{y} \hat{\beta})^T (y - \bar{y} \hat{\beta})
\]

To apply the BSTS model for nowcasting, the Local linear trend without the google trend information is written as

\[
\mu_{t,1} = \mu_t + \delta_t + \eta_{\mu_t},
\]
\[
\delta_{t,1} = \delta_t + \eta_{\delta_t}.
\]

\( \eta_{\mu_t} \) and \( \eta_{\delta_t} \) are assumed to be normal distributions, \( N(0, \sigma_{\mu_t}^2) \) and \( N(0, \sigma_{\delta_t}^2) \), respectively. For the component \( \mu_t \) is the value of the trend at time \( t \) and \( \delta_t \) is the expected increase in \( \mu \) between times \( t \) and \( t+1 \).

The model includes the google trend information which is the \( \beta^T x \) stands for the behavioral factors including in the structural time-series model as represented by following

\[
y_t = \mu_t + \beta^T x_t + \epsilon_t,
\]
\[
\mu_{t,1} = \mu_t + \delta_t + \eta_{\mu_t},
\]
\[
\delta_{t,1} = \delta_t + \eta_{\delta_t}.
\]

3.2. The k-Means clustering algorithm of the unsupervised machine learning

The k-mean algorithm is specified to be a function of the available data type and the purpose of selective analysis [7]. The Euclidean distance is employed to be the criterion that calculate the hyperspace and two-dimensional space [8]. For the development of k-Means algorithms can be represented as five steps by following

- Step 1: clustering an input data and minimizing variance of the data in order to accept the number of group data and cluster the data into \( k \) groups.
- Step 2: taking the first random sampling or the first \( k \) instance of \( k \) elements by initializing the first \( k \) groups.
- Step 3: the arithmetical mean of each cluster formed from the set of data is estimated to be designed as a gradient descent procedure.
- Step 4: the initial cluster is specified class by allocating the k-Means algorithms in each point of the dataset. The nearest cluster assigned each point through a measurement of Euclidean distance.
- Step 5: each point of recorded data is repeatedly assigned by the algorithm

3.3. The supervised machine learning approaches

3.3.1 Linear Discriminant Analysis (LDA). The linear discriminant analysis (LDA) is employed to indicate the linear feature for maximizing the between-class separation of data and minimizing the data within the class, is the basic supervised learning algorithms [9,10]. For the formula of the algorithm can be express by the number of \( n \) which is the clusters of the regimes’ stock market such as Bull market, Sideways and Bear market. Let \( \mu \) be the mean for the samples, \( \mu_i \) and \( N_i \) is the mean and number of the observation data in \( i^{th} \) class, respectively. Then, \( \Sigma_i \) is the scatter matrix of the observation data, the \( S_w \) can be written as follow

\[
S_w = \sum_{i=1}^{k} \Sigma_i,
\]

where the solution of the between-class scatter matrix, \( S_B \), can be expressed as


\[ S_B = \sum_{i=1}^{n} N_i (\mu_i - \mu)(\mu_i - \mu)^T \]  

(11)

with the condition that \( S_W^{-1} S_B \), eigenvalues and eigenvectors are solved.

3.3.2 The k-Nearest Neighbors (kNN) algorithm. The k-Nearest Neighbors (kNN) is a non-form artificial approach for both linear regression problem and classification [10]. In order to cluster the data, which was performed as the other feature space, the observation data are allocated. Hence, the mean value of the nearest neighbors in the linear regression is the majority class to group \( K \) neighbors. The error rate of the misclassified samples from the classification and squared error similar in regression problem can be analyzed by using the k-NN algorithm. First, the process of the algorithms is the calculation for the searching distance \( x_i \) which is the single sample related to all other points in the feature space. In this stage, the \( k \) is controlled from the closet neighbors \( \{ x_i \} \) as well as the measure distance is generally indicated by the Euclidean distance. Then, the next sase is to assume the result of \( y_i \), to be the class membership such as \( y_i \in \{ c_1, c_2, ..., c_c \} \) \( \forall i = 1, 2, ..., n \), where \( c \) is a number of class levels from the majority vote of its kNN. The k-NN regression can be expressed a distance-weighted average as following

\[ y_i = \frac{1}{k} \sum_{j=1}^{k} \sum_{x_i \in \{ x_j \}} x_j \]  

(12)

where \( y_i \) represents the average value of its single nearest neighbor.

3.3.3 The Support vector machine (SVM). For the support vector machine or SVM is the algorithm for applying the regression and classification problem. Generally, in term of the classification the problems are basically modeled as an interval range of data, a binary 0 to 1 which is the logistic regression (Logit model). Hence, the interval of binary 0 to 1 range is assigned to be the position of the hyperplane in the feature space, is the probabilities of class memberships. For the supporting vector is referred to one while the error function is zero, the SVM algorithms are employed to analyze the decision boundary to separate the observation information into two classes by the maximal margin. First, the separate group of the data feature in the space are divided by linearly separation. Then, the point is identified in the input space to support the maximum margin in the vectors. The main properties of the SVM algorithm consists as [11].

\[ e = e_{\text{emp}} + e_{\text{g}} \]  

(13)

where \( e_{\text{emp}} \) is the training error and \( e_{\text{g}} \) is the generalization error.

4. Empirical results

4.1. The Nowcasting prediction by the Bayesian structural time series model

The stock price indexes in the healthcare sector of the united states are represented in Figure 1 in order to explain accurately for the univariate time series. The line prediction shows the movement of the scatter plot which is the predicted rate of the stock market cycles, estimated by its lags, without Google Trend data as the black shade along the line on the left-hand side Figure. However, the estimation of the Nowcasting by BSTS model with google Trend data is presented on the right-hand side of the picture which showed the black shade evidently visualized wider than the prediction without the Google Trend data. According to the power of behavioral factors added into the structural time-series model, the performance of the forecasting is improved in term of the data distribution. Therefore, the efficiency of the forecasting regard to the BSTS model without google trend data is lower than the BSTS estimation with google trend indexes which is the traditional method. The forecasting by the BSTS model with google trend indexes perform better results than the previous case and become useful for the prediction in term of current movement.
4.2. The results of clustering by the k-Means algorithm

In this section the stock market regimes are investigated by the k-means algorithm from 2015 to 2020 which cover the Covid-19 situation period. The stock price indexes from the healthcare sector including the google trend data are added to be the important factor that use to find the stock market regimes. As the study of [12] found that the stock market follows three cycle regimes which are the bull market, sideways and bear market by using the optimal trend following to find these sufficient conditions. Consequently, we explore the stock market regimes by clustering into three part which are bull market, sideways and bear market which is showed in Table 3. The results indicate that there are twenty-nine months stand for bull market, thirty-seven months are predictively provided sideways market and five months are referred as the bear market.

| Description | Frequencies (Months) |
|-------------|----------------------|
| Bull Market | 29                   |
| Sideways    | 37                   |
| Bear Market | 5                    |

Source: Computing

4.3. The stock market regimes forecasting by the machine learning approach

In this part, there are three learning machine algorithms are employed to investigate the stock market regimes in the next five years such as linear discriminant modelling (LDA), Nearest Neighbors (kNN) and Support Vector Machine (SVM). The Kappa’s coefficient is used to validate the machine learning algorithms in order to select the best algorithms for prediction. For the results indicated that LDA that stand for the linear-form learning algorithm is chosen by the highest coefficient validation. Additionally, the accuracies and coefficient are detailed in Table 4.

| Algorithm | Cross-validation method | Accuracy  | Kappa coefficients |
|-----------|-------------------------|-----------|--------------------|
| LDA**     | Cohen’s kappa           | 0.8138095 | 0.6425780          |
| k-NN      | Cohen’s kappa           | 0.5223810 | 0.0758920          |
| SVM       | Cohen’s kappa           | 0.7847619 | 0.5868421          |

Noted: ** indicates the chosen algorithm

Source: authors

The data from the stock price indexes and the data set from the google trend data are classifying between the bull market and bear market by the linear learning algorithm. Interestingly, the grouping data is intercept for each other which the algorithm can group the predominant occurrences. Consequently, the forecasting result from the LDA found that the stock price indexes in the healthcare
sector of the United States of America will stay on the sideways periods in the next five years which are showed in Table 5.

| Year/Stages | Bull Market | Bear Market | Sideways |
|-------------|-------------|-------------|----------|
| 2021        | 0           | 0           | 1        |
| 2022        | 0           | 0           | 1        |
| 2023        | 0           | 0           | 1        |
| 2024        | 0           | 0           | 1        |
| 2025        | 0           | 0           | 1        |

Source: Computing

5. Conclusion
The range of time series data observed of stock price indexes from healthcare sector in the United States are covered from the period during 2015 to 2020. That the world economy is currently facing the recession caused by the Covid-19 epidemic occurred in 2020. The machine learning approaches are the tool which can be used for the various type of data for example the big data which is more incentive to be used in the investigation in the real-world problem. Interestingly, the substantiality of the finding in this study reveals the critically predictable trends of the stock market in healthcare sector of the United States fluctuations through the forecasting, clustering and nowcasting of machine learning approaches. The Bayesian structural time series model is employed to calculate the nowcasting with updating the behavioral information from google trend data which can investigate the misreading up-to-date information which is the insignificant assumption of the traditional econometric methods. For the computational clustering of the stock market cycle regimes, the k-Mean algorithm which is the unsupervised learning can be used to tackle the complexity of data and categorized the cycle regimes of the stock market. Hence, the machines learning approaches can predictively calculate the upcoming regimes situation of stock market in healthcare sector in the United States. For the capability of supervised learning to estimate the enormous amount high dimensional data outperform the traditional method.

The Linear discrimination analysis is the flexibility machine learning model to challenge the parametric estimations. The results reveal that the LDA learning algorithm are evident that the stock market regimes are sideways for the upcoming five years.

To conclude, the machine learning performance can cover the major role of method in econometrics such as a clustering computation, classification, parametric or no parametric predictor. The suggestion provides that the United States’ stock market price indexes in healthcare sector do not need the procyclical policies implementation. However, it needs to know the amount stock market in sideways situation which is the useful information for investor to manage their portfolio especially, in healthcare sector during the Covid-19 period.

6. References
[1] Choudhry, R., and Garg, K.: 2008. A Hybrid Machine Learning System for Stock Market Forecasting. Engineering and Technology, (39), pp.315-318.
[2] 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), pp. 17-23.
[3] Chaiboonsri, C., and Wannapan., S.: 2020. Nowcasting and Forecasting for Thailand’s Macroeconomic Cycles Using Machine Learning Algorithms. International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, pp 270-282.
[4] Scott, S.L. & Varian, H.R.: 2014. Predicting the present with Bayesian structural time series. Int. J. Mathematical Modelling and Numerical Optimisation, Vol. 5, Nos. 1/2, 4-23.
[5] Jun, S.H.: 2019. Bayesian structural time series and regression modeling for sustainable technology management. Sustainability, Vol.11, No. 4945; doi:10.3390/su11184945.
[6] Koduvely, H.M. (2015). *Learning Bayesian Models with R*. Packt: Birmingham, UK.

[7] Oyelade, O.J., Oladipupo, O. O., & Obagbuwa, I. C.: 2010. Application of k-Means Clustering algorithm for prediction of students’ academic performance. *International Journal of Computer Science and Information Security*, Vol.7, No.1, 292-295.

[8] Li, Y. & Wu, H.: 2012. A clustering method based on K-Means algorithm. *Physics Procedia*, Vol.25, 1104 – 1109.

[9] Bishop, C.: 1995. *Neural networks for pattern recognition*. Oxford University Press.

[10] Ioffe, S.: 2006. *Probabilistic linear discriminant analysis*. A. Leonardis, H. Bischof, and A. Pinz (Eds.). Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3954. Springer, Berlin, Heidelberg.

[11] Chakraborty, C. & Joseph, A.: 2017. *Machine learning at central banks*. Staff Working Paper No. 647. Bank of England.

[12] Yu, J., and Zhang, Q.: 2019. Optimal trend-following trading rules under a three-state regime switching model. *American Institute of Mathematical Sciences*, 2(1), pp. 81-100.