ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 191

Proceedings of the 3rd Annual International Conference on Public and Business Administration (AICoBPA 2020)

Predicting VN - Index Value by KNN Algorithm of Machine Learning

Tran Kim Toai*
VSB-Technical University of Ostrava 17, Listopadu
15/2172, 708 33 Ostrava-Poruba Czech Repulic,
tran.kim.toai.st@vsb.cz;
Faculty of Economics, No 1 Vo Van Ngan Street, Linh
Chieu Ward, Ho Chi Minh University of Technology and
Education
Vietnam
*toaitk@hcmute.edu.vn

Bui Tien Thinh
Faculty of Economics, No 1 Vo Van Ngan Street, Linh
Chieu Ward, Ho Chi Minh University of Technology and
Education
Vietnam
*thinhbt@hcmute.edu.vn

Phan-Anh-Huy Nguyen
Faculty of Economics
HCMC University of Technology and Education
Ho Chi Minh city, Vietnam
huynpa@hcmute.edu.vn

Roman Senkerik
Faculty of Applied Informatics, Tomas Bata University in
Zlin, T. G. Masaryka 5555, 760 01, Zlin
Czech Republic
senkerik@utb.cz

***Abstract*—The world has entered the stage of rapid development of technology, especially the fourth industrial revolution with outstanding changes and developments in information technology. Artificial Intelligence (AI) is one of the most mentioned names in this period. AI is part of computer science, developing technology in the direction of automation, self-learning. As a result, it takes a solid knowledge to be able to operate any AI system. There have been many applications of artificial intelligence in the fields of science, technology and economics - finance. From previous decades, the application of algorithms to predict values and variables in economics has been implemented and improved over many different stages. This paper aims to predict VNINDEX value by the application of KNN algorithm machine learning to assess the changes of price indexes and stock variables in general and the VNIndex stock exchange in particular. Research result shows that the outcome of a buy - or - sell decision at the point of view. With a predict signal value of 1, investors should execute buy and sell orders that are advised when predict signal yields -1. Overall, the result indicates that 51% of the stock market price are correctly predicted by the KNN algorithm machine learning.**

Keywords—machine learning, forecasting, KNN

I. Introduction

The world has entered the stage of rapid development of technology, especially the fourth industrial revolution with development of Artificial Intelligence (AI). One of the current trends of technology development is AI – integrated applications that are the foundation of the 4.0 technology revolution.

Machine Learning is a part of AI, born from the ability to recognize existing patterns and from theories on computers that can be learned by themselves without programming. Currently, almost every industry that is operating with large amounts of data recognizes the importance of machine learning.

There have been many applications of artificial intelligence in the fields of science, technology and economics - finance. From previous decades, the application of algorithms to predict values and variables in economics has been implemented and improved over many different stages. However, there is no paper utilized machine learning to improve the VNindex forecasting. To solve this problem, this study aims to predict vn - index value by the application of machine learning to assess the changes of price indexes and stock variables of VN - Index stock exchange.

The purpose of the research is to analyze, evaluate and apply KNN algorithm in predicting VN - Index and stock prices. The authors will attempt to find out the advantages and disadvantages of the algorithm. Through the research, the authors hope to contribute to increasing the accuracy of forecasting the upward-and-downward trend of stock prices, short-term risk prevention, increasing profits.

II. Literature Review

Currently there are many studies on the volatility of the stock market on the global market and in Vietnam. The authors referred to researches on volatility of Vietnam's stock market through the VN - Index. Emandoust and Bolandraftar [1] provided the theoretical framework for the application of KNN in general economics event and particularly forecasting stock market. Following the work of Emandoust and Bolandraftar [1], a number of authors attempted to apply KNN to predict stock market. Khan et al. [2] attempted to predict the future trends of stock market with data taken from London, New York, and Karachi stock exchange. Khan et al. applied KNN, Naïve Bayes and support vector machine (SVM). The empirical result suggested that KNN was the best selection as it provided the highest accuracy and lowest error in terms of stock market trend. Similarly, Khedr et al. [3] attempted to build a two – stage model based on market sentiment analysis of financial market and historical stock market prices to predict stock market future trends with better performance in terms of errors and accuracy. The first stage is to analyze news sentiment using naïve Bayes algorithm. The second stage involves the combination of recent financial information and past stock market prices to predict stock market future prices using both KNN and SVM. The model's overall performance is improved in terms of accurate results compared to previous studies by taking into account various types of market information and historical stock prices. Puspitasari and Rustam [4] applied both SVM and KNN to predict stock prices of Indonesia Stock Exchange. First, the authors applied feature selection method to select crucial financial indicators. Next, SVM was applied to classify stock data in terms of profit and loss, the output served to identify suitable **nearest neighbor from the training set. Next, KNN is utilized to forecast stock price, KNN's performance is based on the value of Root Mean Square Error (RMSE) and relative error**. Overall, the model proved satisfactory in terms of predicting stock market price based on pre-determined financial indicators. Other researches involving the application of KNN in stock market price prediction such as Alkhatib et al. [5] and Tanuwijaya and Hansun [6] also indicated KNN's performance in terms of improving accuracy and errors of stock price prediction.

III. Theoretical Framework

A. *Definition of KNN (K Nearest Neighbors)*

K Nearest Neighbors (KNN) is a simple supervised machine learning algorithm [7]. It calculates the distance of a new data point to all other training data points. Then the nearest data points of K which can be any integer is chosen. Finally, it assigns data points to the layer to which most of the K data points belong.

KNN holds several advantages: First off, its implementation is easy. KNN is a lazy learning algorithm (instance – based learning) and therefore it does not require training in advance to stock price forecasting. This feature allows new data to be added seamlessly. Additionally, there are only two parameters needed to perform the KNN, i.e. the value of K and the distance function (for example: Euclid or Manhattan, etc.).

A simple implementation of the KNN regression is to calculate the mean of the numerical target of the closest K nearby. Another approach is to use the average of the inverse distance of the nearest K. KNN regression uses distance functions the same as KNN classification (see figure 1).

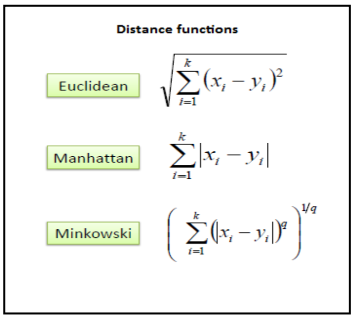
Distance functions
Euclidean
$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
Manhattan
$\sum_{i=1}^{k}|x_i - y_i|$
Minkowski
$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Fig. 1. Distance measurements of KNN.

In KNN implementation, the three distance measurements are to be applied only with continuous variables. However, with categorical variables, the Hamming measurement which is a measure of the number of versions that the corresponding symbols differ in the two strings of equal length must be applied.

Determining the data is the first step of selection of optimal K value. In general, large K values are more accurate because they reduce overall noise. However, a drawback is the blurriness of individual boundaries in the feature space. Cross validation is another method of redefining good K value using independent data sets to validate K value. Overall, the optimal K for most data sets is 10 or more.

B. *Mechanism of KNN*

The mechanism of KNN can be expressed by a simple example. Suppose there is a data set with two variables as shown in the figure 2 below.

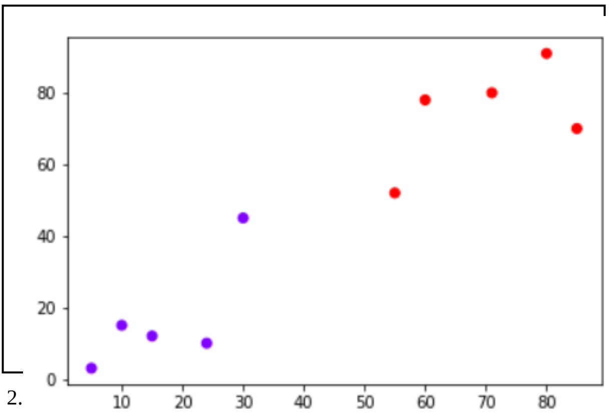

Fig. 2.

The objective is to identify a new data point with 'X' as "Blue" or "Red". The data point coordination is as follow: 45 for x and 50 for y. Assuming that K is 3. The KNN algorithm begins by measuring the distance of X from other data points. Then KNN attempts to identify the 3 closest points to point X as shown in the figure 3 below.

The final step of the KNN is the assignment of a new grade to the class that the majority of the selected nearest points belong to. Based on figure 2, two out of the three closest points lie in the "Red" layer while the remaining point is in the "Blue" layer. As a result, new data points will be categorized as "Red”.

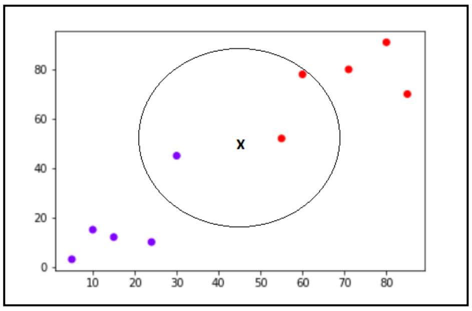

Fig. 3. New data points.

C. *Measurements of the Selection of Algorihm Model*

Once the working mechanism of the KNN algorithm is understood, a new question is: How to choose the optimal number of Neighbors? The number of neighbors (K) in the KNN is a super parameter which must be selected at the beginning of model building with K being a control variable for the model.

Researches indicate that each data set is unique, therefore, there is no fixed value of optimal Neighbors number that can be used for all data sets. In case of a small number of neighbors, big data will have a greater influence on the result and a large amount of data makes it computationally expensive. Researches also show that a small number of most suitable Neighbors will have low deviations but high variance and a large number of Neighbors will have a smoother decision boundary which means lower variance but higher deviation.

When the value of K is reduced to 1, the prediction of becomes less stable. Conversely, when the value of K is increased, the prediction is more stable due to the majority vote of its K neighbors, and therefore, more likely to make a more accurate prediction.

To select K in accordance with existing data, the authors attempt to initiate the KNN algorithm multiple times, each time is with a specific K value. The authors will identify the K value that results in minimum value of errors while maintaining the forecasting ability of the algorithm when data is provided.

D. *KNN and Superiority of the Forcast*

Data science usually starts with linear models, but in its way, K Nearest Neighbors is the most widely used conceptual model. KNN models show that things with similar features tend to be good. This is not an insight, but these practical implementations can be extremely compelling and especially for someone who accesses an unknown data set, which can be handled nonlinearly without any technical skills or data model complexity. However, the drawback here is that: while the model can be established in a quick manner, the forecasting process will be slow, as when KNN predicts the result for a new value, it will search in all data points within the training set to find the nearest point. Therefore, for large datasets, the KNN may be slow when compared to other regressions with longer required time to match but simpler prediction and calculation process.

K Nearest Neighbors was proposed sixty years ago, but because of the need for large and computational memory, this method was not popular for a long time. With advances in parallel processing and with increasingly efficient memory and computing capability, such methods have recently been more widely used. Unfortunately, it can still be quite computationally expensive when it comes to large training data sets because we need to calculate the distance for each sample.

Also, when looking at low-dimensional spaces and having enough data, the KNN works very well for accuracy, because we have enough nearby data points to have a good answer. As the number of dimensions increases, the algorithm performs poorly, this is due to the fact that distance measurement becomes meaningless as the size of the data increases significantly.

IV. DESIGNING FORECASTING MODEL OF VN - INDEX

A. *Designing Regression Model to Forecast Stock Prices with Groups of Macroeconomic Variables*

The theoretical background of financial investment, Investment Analysis and Portfolio Management by Reilly et al. [8] provides typical evidence of economic variables affecting stock prices. Based on this foundation, the authors consider the group of macroeconomic indicators affecting this stock price to be input variables for the model in the process of designing a forecast model.

The success of designing predictive models depends on a researcher's ability to understand the issue being studied. Recognizing which variables play an important role in the market is a prerequisite for the process of designing a forecasting model

B. *Economic Data Input*

For the training and testing of KNN algorithm in stock market prediction, the data set will be divided into two separate parts: 70% of the data set will be for training purposes while the remaining 30% of the data set is for testing and accuracy evaluation purposes.

NASDAQ: The NASDAQ Composite is a stock market index of popular stocks listed on the Nasdaq stock exchange.

DOWJONES: The Dow Jones Industrial Average tracks the stock exchange price of 30 large companies whose names are listed on the US stock exchanges.

S&P 500: The S&P 500 is a common stock index of the 500 largest market capitalization companies listed on the NYSE or NASDAQ.

GOLD: Gold price

USD: Exchange rate of US Dollar and Vietnamese currency

VN - Index: The VN - Index is an index that reflects stock price movements, and is calculated according to the average market capitalization method of all stocks listed on the Ho Chi Minh Stock Exchange.

C. Step – by – step Execution

The summary of KNN application in this paper can be summarized as in the figure 4 below:

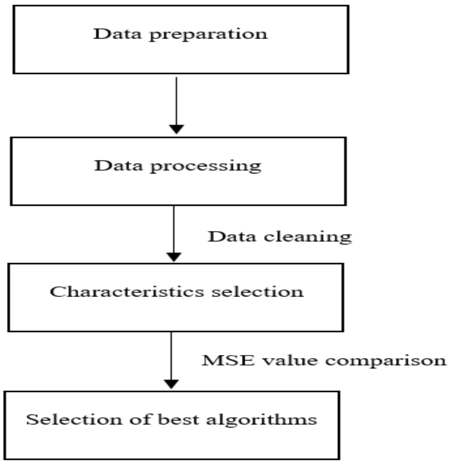

Fig. 4. Summary of KNN application.

The first thing we need to do is include the support libraries in the model. Here the authors use libraries including Sklearn, Numpy, Mathplotlib, Pandas, and Pyplot. Then the authors put the data in the same format and range with the tool MinMaxScaler

Step 1: Prepare the input data set and filter the data

As shown in Figure 2, we have 4550 values for each input variable. The value used is calculated from the column “Close” (closing price). of each variable and the blank data has been removed by precursor method. Data used in the project is downloaded from finance.yahoo.com and some other sources in the range from 1/2000 to 2/2020 (about 20 years) with data content including: Date (date, month), Open, High, Low, Close, Volume and are standardized in the same format to use in their entirety. project.

Step 2: Pretreatment:

Data cleaning: The authors first filtered the missing data and then used the previous data to replace it.

Data integration: The authors integrated data files. After the data set was converted to a clean data set, the data set was divided into training and testing set for evaluation.

Step 3: Select the property:

In this step, the selected data properties will be supplied to the predictive model. Price at closing moment is chosen as the feature to be selected

Step 4: Algorithms and selection of algorithm with the best value:

The authors used a number of algorithms and made a general comparison to choose the most suitable algorithm. The authors also used 10-fold Cross Validation and evaluated based on MSE index (Mean Square Error) to evaluate the degree of error of each algorithm model (with 0 being the best model).

The algorithm models used for testing include:

- Linear Regression (LR):
- Lasso:
- Elastic Net (EN)
- K Nearest Neighbors (KNN):

Decision Tree Regressor (CART):

- Support Vector Regressor (SVM)

After running the experimental model, the authors decided to choose the best model in this case which was K Nearest Neighbors algorithm with the mean quadratic error MSE = -0.009984 (0.002852).

D. KNN algorithm and the VN - Index Price Prediction

1) Enter libraries to be used for data reading and algorithm training.

2) Economic data input: The economic data input includes VN - Index which was taken from Bao Viet Security and ranging from 2000 to 2019. The database includes 5 collums and 4550 values.

3) Defining variables: Predictor variable (X): The prediction variable is set to be by the signal between "Open - Close" and 'High-Low' (High- Low). X is a part of target variable forecasting.

Target variable (Y): The target variable is the security's price to close or fall on the next trading day. When today's close is greater than tomorrow's close, a buy signal (1) is activated, otherwise for a sell signal (-1).

4) Dividing database: The database is then divided into training data set and test data set. Data from 2000 to 2013 is to train KNN while 2013-2019 data is to test KNN prediction. A variable called "split" is used to segment data. The variables X_train and Y_train are defined for training. The variables X_test and Y_test are for testing purposes.

5) Application of the KNN model: Using the K Neighbor Classifier, the authors set a k value of 3. To be more specific, an average buy or sell signal of the 3 closest values will be used to predict the target variable. Additionally, increasing k value will lead to reduction of the variance and increase of the bias. The "fit" function allows the authors to adapt the training data to this KNN model. The function "precision_score" shows the authors the accuracy score of the model. The result shows that the model's prediction accuracy on the test data is correct 51% of the time.

V. Results and Discussion

A. The Interaction between VN - Index and Other Indicators

From the calculation steps, the authors got the correlation coefficient results between the indicators in the following data set. See table 1 below.

TABLE I. WEIGHTED MATRIX

	VNI	NASDAQ	S&P500	DOWJ ONES	GOLD	USD
VNI	1	0.62	0.67	0.72	-0.53	-0.6
NASDA Q	0.72	1.0	0.97	0.91	-0.78	-0.66
S&P500	0.67	0.97	1.0	0.93	-0.74	-0.66
DOWJO NES	0.72	0.91	0.93	1.0	-0.82	-0.76
GOLD	- 0.53	-0.78	-0.74	-0.82	1.0	0.85
USD	-0.6	-0.66	-0.66	-0.76	0.85	1.0

Through the table above, it is easy to see that the indicators VNI, NASDAQ, S&P500, DOWJONES, GOLD and USD have a mutual impact on moderate to high levels (value 0.5-0.97) and here there are both negative and positive correlations in the indices.

B. The Quadratic Error of Algorihmic Models

TABLE II. VALUE OF MSE

Name of the algorithm	MSE
Linear Regression	0.056810
Lasso	0.057255
Elastic Net	0.057030
K Neighbor Nearest Regression	0.002852
Decission Tree Regression	0.016006

From the table 2 above, the authors selected the model of KNN algorithm with the least square error which is the model used for the calculation process.

C. Modifying the KNN Model

TABLE III. K VALUE AND ASSOCIATED ERROR VALUE

K Value	Error
1	0.004629
3	0.003780
5	0.003894
7	0.004324
9	0.004884
11	0.005498
13	0.006811
15	0.008189
17	0.009566
19	0.010972
21	0.012031

From the table 3, in order to better evaluate the models, the authors performed tuning for the algorithms and selected the values k=3, with error result -0.00378 then the KNN model will produce the best results with the data set

D. Ensemble Method and Adjustment

TABLE IV. ENSEMBLE ALGORITHM VALUE

Method	Error
Ada boost	0.154569
Gradient Boosting	0.038451
Random forest	0.017936
Extra Tree	0.005306

From the table 4 above, the authors selected ensemble algorithm as Gradient boosting to adjust and have the following corrective results.

TABLE V. NUMBER OF VARIABLES AND ERRORS VALUE

Number of variables	Error
50	0.058391
100	0.038434
150	0.031481
200	0.027976
250	0.025901
300	0.024477
350	0.023342
400	0.022740

From the table 5 above, the authors noticed with variable 400, GBM gives the best value.

E. Comparision between GBM and KNN

The global error of Gradient Boosting on the training data set is close to the global error of all data (0.016 is closed to 0.022).

The global error of KNN on the training data set is close to the global error of all data (0.0027 is closed to 0.0037).

Authors found that the quadratic error (MSE) when using the data set with 2 algorithms KNN and GBM, KNN with

0.0027 <0.0165, choose the most suitable algorithm to predict VN - Index will be KNN (with k=3).

F. Applying the Model

To test the performance of the model, the authors could compare it to the buy and hold strategy. Returns the purchase and hold logs recorded in the "df_returns" column. Furthermore, the "cumsum" function is used to determine cumulative profits. The transaction log returns based on the empirical buy and sell signals of the empirical group recorded in the 'Strategy_Returns' column.

Once again, the authors used the "cumsum" function to determine the cumulative strategic profit. 'Cumulative_df_Returns' and cumulative_Strargety_Returns' are drawn together using Matplotlib.

Initiating this strategy on the authors' test data set (2013-2019), this KNN strategy uses a cumulative log income of 90.099409% while the strategy buys and holds only the profit of 87.609093%. The overall accuracy of the model in terms of stock market prediction is 51%.

In addition to the use of profits, this strategy can be evaluated by other performance indicators such as the Sharpe ratio. In this example, the strategy of the authors reaching the Sharpe ratio of 0.28 is not necessarily absolute. Although the increase exceeds the major between the strategy and the buy and hold, the standard deviation of the profit is high.

The research of KNN application in terms of VN – Index stock market price prediction also illustrates several research recommendations. Firstly, the accuracy of stock market price prediction of KNN algorithm can further be improved by LSTM algorithm. Secondly, due to the rather lack of in – depth data set of VN – Index, the authors suggest that the data set should be further expanded in future research to improve overall accuracy.

VI. Conclusion

Research result shows that the outcome of a buy - or - sell decision at the point of view. With a predict signal value of 1, investors should execute buy and sell orders that are advised when predict signal yields -1. Overall, the result indicates that 51% of the stock market price are correctly predicted by the KNN algorithm machine learning.

References

[1] S.B. Emandoust and M. Bolandraftar, “Application of K – Nearest Neighbor (KNN) Approach for predicting economic events: Theoretical background”, International Journal of Engineering Research and Application, vol. 03, issue 05, pp. 605 – 610, 2013.

[2] W. Khan, M.A. Ghazanfar, M. Assam, “Predicting trend in stock market exchange using machine learning classifiers”, Science International Lahore, No. 28, Issue. 02, 2016, pp. 1363 – 1367, 2016.

[3] A.E. Khedr, S.E. Salama, N. Yaseen, “Predicting stock market behavior using data mining technique and news sentiment analysis” International Journal Intelligent Systems and Applications, vol. 07, pp. 22 – 30, 2017.

[4] D.A. Puspitasari, Z. Rustam, “Application of SVM – KNN using SVR as feature selection on stock analysis for Indonesia stock exchange. AIP Conference Proceedings 2023, 020207, 2018.

[5] K. Alkhatib, H. Najadat, I. Hmeidi, M.K.A. Shatnawi, “Stock price prediction using K – Nearest Neighbor (KNN) Algorithm”, International Journal of Business, Humanities and Technology, vol. 03, issue. 03, pp. 32 – 44, 2013.

[6] J. Tanuwijaya and S. Hansun, “LQ45 Stock Index prediction using K – Nearest Neibors regression”, International Journal of Recent Technology and Enginerring, vol. 08, issue. 03, September 2019, pp. 2388 – 2391, 2019.

[7] D. Aha, D.W Kibler, and M.K. Albert, “Instance-based learning algorithms”. Mach Learn, vol. 6, pp. 37 – 66, 1991.

[8] F.K. Reilly and K.C. Brown, “Investment analysis and portfolio management”. Mason, Ohio: South-Western/Thomson Learning, 2003.