IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 11, June 2020)

Stock Market Prediction using Bayesian Optimized K-Nearest Neighbor

Zahra Malwi Paril Ghori

Abstract – **Stock markets are complex systems due to their non-stationary nature, as the parameters are constantly changing, such as economic conditions and changes in company policy. This paper proposes a model based on Bayesian optimized K-Nearest Neighbor (KNN) for the price prediction of New York stock Exchange. Different configurations of KNN are tested using a six years series (January 2010 to December 2016). Three attributes of dataset; open, high and low values are used for the input of the KNN. The results show a good behaviour of Bayesian optimised KNN with low-performance errors in both learning and prediction.**

Keywords – **Bayesian Optimization, KNN, ORCL.**

I. INTRODUCTION

The original idea for this work arose from the desire to adequately solve the problems that arise when trying to predict future movements of the stock market from its historical data. A preliminary bibliographic review (see Section 2) suggests that the information contained in the historical data of the stock market, if it exists, does not seem to be very relevant, therefore it could This market would be considered, basically, a stochastic system. The studies and theories of various authors seem to confirm the idea of the limited amount of usable information in such data, although some studies manage to extract some utility from them.

Along the same lines, different publications seem to confirm that the use of traditional machine learning techniques on historical stock market data have reported promising results, although few have consistently obtained positive results for long periods of time.

At that point, the following question arises: Is information not found in the historical data because it does not exist, or because the techniques used are not appropriate for it?

From these previous works the following hypotheses have been advanced:

- Although most of the time and for most of the quoted values, the movement of prices can be considered to a great extent random, it is possible that, at specific moments and for specific values, there may be a certain amount of usable information.
- The problem could be posed in dynamic terms where the useful information or market inefficiencies, as it is named in the literature, once detected are of short use before being 'self-destruct' by general knowledge.
- The automatic learning techniques to use, see their performance degraded in a certain way in this type of problems due to the presence of some factors that contribute to learning difficulties.
- If a machine learning technique were available capable of solving problems of pattern classification, in the presence of these factors, the problem of prediction of movement could be faced with certain guarantees of prices on the stock market.

Professionals dedicated to buying / selling shares in order to obtain profits generated by the difference in the purchase price and the sale price, use different techniques in order to try to predict the future behaviour of the market. These techniques can be grouped into:

1. Fundamental Analysis [1], analysis of the (tangible) economic foundations of each company and,
2. Technical Analysis [2] [3] [4], study of past historical prices.

From the academic / scientific sphere, the real efficacy of these methods has been questioned and discussed by Malkiel, B.G., 1999 [5].

The basic criticism of the usefulness of these techniques is based on the Efficient Market Hypothesis (EHR).

The efficient market idea is associated with the concept of equality of conditions, first postulated by Cardano in 1565, as a fundamental principle of games of chance. Throughout the 20th century were shaping the modern definition of efficient market [6] [7], [8], [9], [10], [11]. Thus, it is said that a market

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 11, June 2020)

is efficient when there is sufficient liquidity and economic rationality on the part of the agents, so that any type of relevant information is absorbed by prices instantly, generating a random behavior in them, which makes their systematic forecast impossible.

In [12] and mainly in [8] the existence of three levels of market efficiency is proposed, taking into account the degree of information that intervenes in the formation of prices:

Weak form: the price of the securities reflects all the historical information. It is not possible to use strategies based on the analysis of historical series to achieve returns that exceed the market itself. The appearance of new information, new news, randomly shifts the price up or down. To appear

These are unpredictably over time, the change in prices also reflects this random variation. This hypothesis implies that the Technical Analysis does not provide any utility.

Semi-strong form: the market players have all the past information and, also, all the information made public about a listed security. Prices are instantly adjusted to any information that is made public. The semi-strong hypothesis implies that the Fundamental Analysis will not be able to achieve higher-than-market returns.

Strong form: investors, in addition to past information and public information, have insider information on assets. This hypothesis implies that the price reflects as much information as possible and no one can outperform the market.

II. Machine Learning for Stock Market Prediction

Following are papers, in chronological order, of machine learning techniques that have been used in the context of investing in the stock market.

As input data to the classifiers, the following have been used.

In one of the first publications of applications of Multilayer Perceptor (MLP) in this field, Kimoto [13] applied this technique to the prediction of the Tokyo Stock Market Index.

In [14], the authors review the applications of MLP to investment in markets.

On the other hand, in [15] it is affirmed to be able to reach a 63% accuracy of the stock prices in the Tokyo Stock Market.

In [16], it is concluded that the Support Vector Machines (SVM) technique can be useful in predicting the direction of change (rise or fall) of the Nikkei Index (Japanese Stock Market Index). However, they point out that making accurate predictions is very costly, given that the market appears complex, evolutionary, non-linear dynamic and is affected by interactions with political news, economic conditions and subjective expectations of the different actors of the system.

In [17], [18], [19], [20], different techniques are presented (LCS, XCS) that, making use of the concept of continuous adaptation of the predictive model to the dynamism of the market, allow us to obtain useful information from this.

In [21], the authors try to elucidate whether the failures published in the prediction of timing (choice of the best moment to carry out an operation in the market), making use of gen programming ethics, are due to errors or deficiencies in the prediction technique or to the intrinsic randomness of the market.

In [22], an attempt is made to predict the Dow Jones Industrial Average Index using MLP, Decision Trees and K-NN. The study concludes that not all periods evolve equally randomly. The authors make use of the percentage of correct answers to measure the goodness of the prediction, reaching values of up to 65%.

In [23], an attribute selection algorithm is proposed for prediction problems in Asian Stock Markets. As input data, the authors use 23 technical indicators, calculated from data from 1991, for the markets of Korea and Taiwan. The objective is to predict the direction of the index movement (up or down). The proposed algorithm is based on a voting system, where making use, as an attribute selection technique, of a Wrapper and different classifiers (SVM, K-NN, MLP, Random Forest and Logistic Regression), the optimal subset of attributes is obtained. Here, the optimal subset is understood as the one that allows obtaining the best results in the classification phase. As a measure of the success of the classification, the authors make use of accuracy. The article concludes that the proposed solution (use of an attribute selection phase prior to classification) obtains better results than using all indicators as input to the classifier.

In [24], the Markov Blanket Random Forest method is presented to predict the stock market. The authors state that it outperforms the passive Buy and Hold strategy and other regression methods used in the experiment (Linear Regression, SVM and MLP).

In [25], a system for prediction of the stock market based on neural networks MLP and Elman recurrent network (ERN) is presented.

In [26], K-NN is used to predict the evolution of five companies on the Jordanian Stock Market during 2009. As input data to the algorithm, the ticks data are used daily of each value.

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 11, June 2020)

In [27], predictions are made on the Singapore Stock Market. Gradient Boosted Random Forest (GBR) has been used as a regression technique. The presented method outperforms the Buy and Hold strategy in economic performance.

III. Bayesian Optimized KNN and Prediction of the Shares Prices

For the development of the model, Oracle Corporation raised, which is simultaneously listed on the New York Stock Exchange. This information was obtained from the website - https://www.nyse.com/index. The input variable is the daily closing price in United States dollars (USD), and as the only output you have the price to be predicted for the next day. K-Nearest Neighbor is used to predict stock prices. The KNN algorithm has been used to solve many problems. These include economic and financial factors, many of which highlight their use in time series forecasts and their ability to detect and exploit data nonlinearity, even in situations where noise data is incomplete or present. They also stand out for their effectiveness in solving complex problems where pattern or behavior recognition is important.

A. Learning Algorithm

Learning rules or algorithms are mechanisms by which all network parameters are adapted and changed. In the case of KNN, this is a supervised learning algorithm; That is, the parameter change is performed so that the output of the system is as close as possible to the output provided by the supervisor, or to the desired output. KNN has no specific training phase and uses all training data during classification or regression tasks. This is a nonparametric learning algorithm because it assumes nothing about the underlying data.

The selection of the best target for the proposed work was determined using traditional internal and external predictive valuation methods described in the next section.

B. Prediction Model with KNN

The K nearest neighbor algorithm is simple, but it can give interesting results if the data range is large enough. It is a widely used classification method in many areas and is also one of the top ten data mining algorithms. In general, the neighboring houses have similar characteristics. We can classify and classify them. Algorithms use the same logic to try to group closely spaced elements.

KNN is an example of instance based learning. It works in situations where each instance can be defined by an n-dimensional vector, where n is the number of attributes used to describe each instance and rank is a discrete value. The training data is stored and when a new sample is found, it is compared with the training data to find the nearest neighbors.

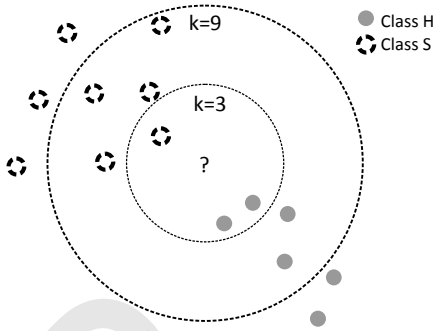

Figure 1: Operation of a simple KNN with k = 3 and k = 9

The nearest neighbor is the closest at the Euclidean distance. The distance between two elements $A = \langle a_1, \dots, a_n \rangle$ and $B = \langle b_1, \dots, b_n \rangle$ is calculated as follows:

$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad (1)$$

Ordered by the k nearest neighbors of the new instance, the classification will be the class with the largest occurrence among them.

Here is pseudo code representing the algorithm:

Pseudo-Code for KNN

Requires 3 parameters: a set of examples X, a given x *and* $k \in \{1, \dots k\}$
For each example $x_i \in X$
Calculate the distance between x_i *and* x: $\delta(x_i, x)$
End for
For $j \in \{1, \dots k\}$ *do*

$$KNN(j) \leftarrow \arg\min \delta(x_i, x) i \in 1, \dots n$$
$$\delta(x_i, x) \leftarrow +\infty$$

End for

Determine the class of x from the class of examples whose number is stored in the KNN.

C. Prediction Model with Bayesian Optimization of KNN

First, there are three parameters to consider: the sample data, the number of nearest neighbors from which to select (k), and the point we want to estimate (x). Then, for each sample, we estimate the distance between the X pivot point and the x point; of all studies, and we checked if the distance between them was less than indicated in the list of nearest neighbors. In this case, the point is added to the list. If the number of elements in the list is greater than k, the last value is removed from the list. The

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 11, June 2020)

algorithm itself is not very complicated and can give good search results if the sample is not very large. However, when we talk about data mining, the number of people to be evaluated is usually very large, so an optimization algorithm is needed.

The main idea of Bayesian optimization (BO) is to consistently build a probabilistic substitution model to try to derive an objective function. New observations are made iteratively and the model is updated, reducing its uncertainty makes it possible to work with known and less expensive models that are used to construct utility functions that determine the next evaluation point. The various steps in the BO methodology are described below.

First, you need to choose the old model, taking into account the possible function space. For this, various parametric approaches can be used, such as the Beta-Bernoulli Bandit or a linear (generalized) model, or nonparametric models, such as the Student's or Gaussian t-process [28].

Then several times until certain stop criteria [29]:

The assumptions and probabilities of observations so far have been combined to produce a posterior distribution. This is done using Bayes' theorem, hence the name.

Bayes' theorem: If A and B are two events for which the conditional probability $P(B|A)$ is known, then the probability $P(B|A)$ is defined as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

Where $P(A)$ is the prior probability, $P(B|A)$ is the probability of event B depending on the occurrence of event A, and $P(A|B)$ is the posterior probability.

Then some of the utility functions are maximized in the next model to determine the next scoring point, and new observations are collected to repeat until the completion criterion.

Since the KNN approach uses sampling techniques for continuous parameters, it gives less accurate results with loss of data. The proposed work analyzes an algorithm that can set the KNN parameter.

The algorithm proposed in this article refers to the optimal value of K. Its parameters are: i) weight, C; and ii) core functions. The weights are a trade-off between some classification errors and the correct classification of others, while the kernel is used to directly tune KNN parameters and select subsets of features.

BO-KNN Algorithm

Input: n, m, C, γ, and termination criterion
Output: Optimal value for KNN parameter
Begin
Initialize n solutions
call KNN algorithm to evaluate n solutions
T = Sort ($k_1, \dots, k_n$)
while number of iteration ≠10 do
for i = 1 to m do
select k according to its weight
sample selected k
store newly generated solutions
call KNN algorithm to evaluate newly generated solutions
end
T = Best (Sort $k_1, \dots, k_{n+m}$), n)
end
End

In algorithms, n is the size of the solution file, m is the number of models used to generate the solution, C is the smoothing or soft edge parameter, γ is the kernel function parameter, called the margin or width parameter, and, finally, the termination condition for the best value of the KNN parameter (K).

This article explores the needs of swing traders in the US stock market whose main goal is to get an overview of the future value of stocks in the coming days; Thus, a one-year forecast horizon was chosen, which showed that the optimized Bayesian KNN was correctly formed with data from the previous two years.

IV. Simulation And Results

The performance of proposed algorithms has been studied by means of MATLAB simulation.

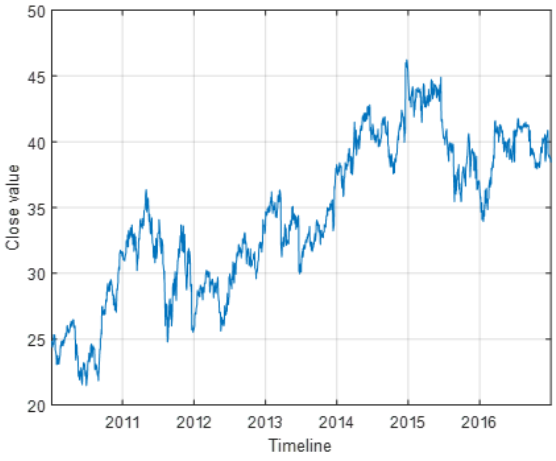

Figure 2: ORCL stock price from January 2010 to December 2016

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 11, June 2020)

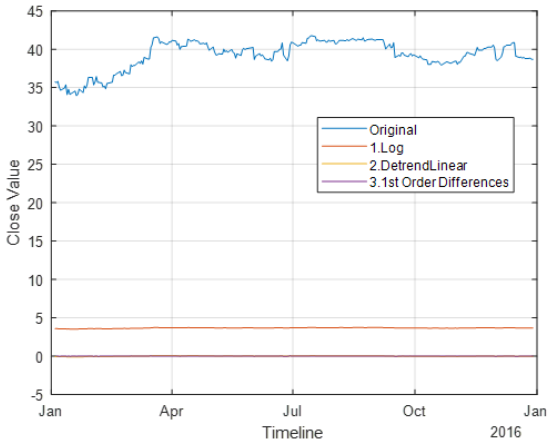

Figure 3: ORCL stock price in year 2016

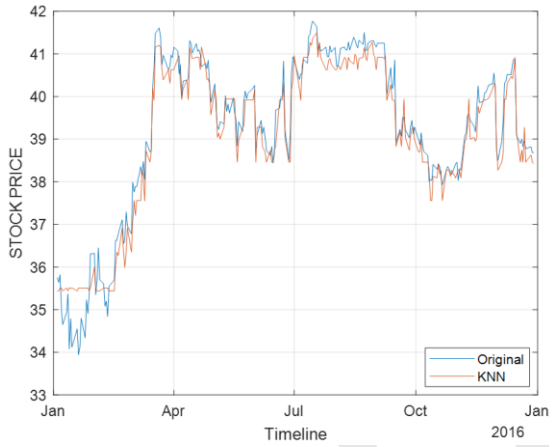

Figure 4: ORCL stock price in year 2016 using KNN

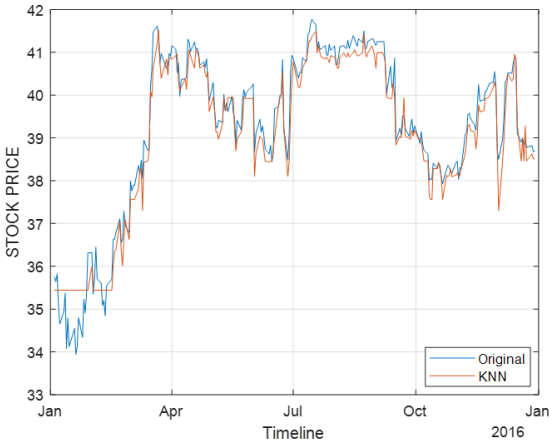

Figure 5: ORCL stock price in year 2016 using Bayesian optimized KNN

Table 1: Comparative results

Evaluation Measures	KNN	Bayesian optimized KNN
Mean Absolute Error	0.3159	0.2910
Mean Squared Error	0.1639	0.1507
Root Mean Squared Error	0.4049	0.3882
Mean Absolute Relative Error	0.0082	0.0076
Mean Squared Relative Error	$1.1612e^{-04}$	$1.0690e^{-04}$
Root Mean Squared Relative Error	0.0108	0.0103
Mean Absolute Percentage Error	0.8219	0.7586
Root Mean Squared Percentage Error	1.0776	1.0339

V. CONCLUSION

The successful use of Bayesian optimized K-Nearest Neighbors to prediction the price of the New York Stock Exchange demonstrates their applicability in emerging markets. The advantages of KNN are concluded as they are simpler models to implement and allow to obtain low prediction errors both inside and outside the sample. The effect of including the Bayesian optimization improved the performance of the KNN. Although the applications shown were in actions of different economic sectors (energy and financial), it was found that practically the same BO optimized KNN structure using only the price series, can reliably represent the actions used.

REFERENCE

[1] Graham, B., Cottle, S., Tatham, C. and Dodd, D.L., 1962. Security analysis: Principles and technique.

[2] Kaufman, P.J., 2013. *Trading Systems and Methods,+ Website* (Vol. 591). John Wiley & Sons.

[3] Schwager, J.D., 1996. *Technical Analysis: Schwager on Futures*. J. Wiley.

[4] Sewell, M., 2008. Technical analysis. *London: University Collegge London*.

[5] Malkiel, B.G., 1999. *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company.

[6] Malkiel, B.G., 1989. Efficient market hypothesis. In *Finance* (pp. 127-134). Palgrave Macmillan, London.

[7] Fama, E.F., 1991. Efficient capital markets: II. *The journal of finance*, *46*(5), pp.1575-1617.

[8] Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), pp.383-417.

[9] Mandelbrot, B., 1966. Forecasts of future prices, unbiased markets, and" martingale" models. *The Journal of Business*, *39*(1), pp.242-255.

[10] Samuelson, P.A., 2016. Proof that properly anticipated prices fluctuate randomly. In *The world scientific handbook of futures markets* (pp. 25-38).

[11] Fama, E.F., 1965. The behavior of stock-market prices. *The journal of Business*, *38*(1), pp.34-105.

[12] Roberts, H., 1967. Statistical versus clinical prediction of the stock market. *Unpublished manuscript*.

[13] Kimoto, T., Asakawa, K., Yoda, M. and Takeoka, M., 1990, June. Stock market prediction system with modular neural networks. In *1990 IJCNN international joint conference on neural networks* (pp. 1-6). IEEE.

[14] Zhang, G., Patuwo, B.E. and Hu, M.Y., 1998. Forecasting with artificial neural networks:: The state of

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 11, June 2020)

the art. *International journal of forecasting*, *14*(1), pp.35-62.

[15] Mizuno, H., Kosaka, M., Yajima, H. and Komoda, N., 1998. Application of neural network to technical analysis of stock market prediction. *Studies in Informatic and control*, *7*(3), pp.111-120.

[16] Huang, W., Nakamori, Y. and Wang, S.Y., 2005. Forecasting stock market movement direction with support vector machine. *Computers & operations research*, *32*(10), pp.2513-2522.

[17] Wong, S.Y. and Schulenburg, S., 2007, July. Portfolio allocation using XCS experts in technical analysis, market conditions and options market. In *Proceedings of the 9th annual conference companion on Genetic and evolutionary computation* (pp. 2965-2972).

[18] Schulenburg, S. and Ross, P., 1999. An evolutionary approach to modelling the behaviours of financial traders. In *Genetic and Evolutionary Computation Conference Late Braking Papers* (pp. 245-253).

[19] Schulenburg, S. and Ross, P., 2000, September. Strength and money: An LCS approach to increasing returns. In *International Workshop on Learning Classifier Systems* (pp. 114-137). Springer, Berlin, Heidelberg.

[20] Schulenburg, S. and Ross, P., 2001, July. Explorations in LCS models of stock trading. In *International Workshop on Learning Classifier Systems* (pp. 151-180). Springer, Berlin, Heidelberg.

[21] Chen, S.H. and Navet, N., 2007. Failure of genetic-programming induced trading strategies: Distinguishing between efficient markets and inefficient algorithms. In *Computational Intelligence in Economics and Finance* (pp. 169-182). Springer, Berlin, Heidelberg.

[22] Qian, B. and Rasheed, K., 2007. Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), pp.25-33.

[23] Huang, C.J., Yang, D.X. and Chuang, Y.T., 2008. Application of wrapper approach and composite classifier to the stock trend prediction. *Expert Systems with Applications*, *34*(4), pp.2870-2878.

[24] Maragoudakis, M. and Serpanos, D., 2010, October. Towards stock market data mining using enriched random forests from textual resources and technical indicators. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 278-286). Springer, Berlin, Heidelberg.

[25] Naeini, M.P., Taremian, H. and Hashemi, H.B., 2010, October. Stock market value prediction using neural networks. In *2010 international conference on computer information systems and industrial management applications (CISIM)* (pp. 132-136). IEEE.

[26] Alkhatib, K., Najadat, H., Hmeidi, I. and Shatnawi, M.K.A., 2013. Stock price prediction using k-nearest neighbor (kNN) algorithm. *International Journal of Business, Humanities and Technology*, *3*(3), pp.32-44.

[27] Qin, Q., Wang, Q.G., Li, J. and Ge, S.S., 2013. Linear and nonlinear trading models with gradient boosted random forests and application to Singapore stock market.

[28] Acuna, D. and Schrater, P., 2008, July. Bayesian modeling of human sequential decision-making on the multi-armed bandit problem. In *Proceedings of the 30th annual conference of the cognitive science society* (Vol. 100, pp. 200-300). Washington, DC: Cognitive Science Society.

[29] Pelikan, M., Goldberg, D.E. and Cantú-Paz, E., 1999, July. BOA: The Bayesian optimization algorithm. In *Proceedings of the genetic and evolutionary computation conference GECCO-99* (Vol. 1, pp. 525-532).