Using K-Nearest Neighbor (kNN) Algorithm for Stock Price Prediction

Boluwatife Aiki-Raji

Brittany Miller

Samman Bikram Thapa

Introduction

“Security prices fully reflect all available information” (Fama, 1991). In her 1991 paper on Effective Capital Markets, Fama reiterates that a security prices at any time will “fully reflect” all available information. Based on this assumption she makes, the world today is dedicated to researching on ways to predict future stock prices based on their current information. And for this, many researches and data mining is done based on the data from stock market. In this paper, we attempt to do such analysis but with an emphasis of using a machine learning algorithm. We applied k-nearest neighbor algorithm in order to predict stock prices for a sample of seven major companies listed on the NASDAQ stock market to assist investors, management, decision makers, and users in making correct and informed investments decisions. According to the results, the kNN algorithm is mildly robust with a good accuracy; consequently the results were rational and also reasonable. In addition, depending on the actual stock prices data; the prediction results were close and fairly parallel to actual stock prices.

Even though, a lot of businesses are vested in researching on predicting the Stock market, financial data is considered as complex data to forecast and or predict. Predicting market prices are seen as problematical, and as explained in the efficient market hypothesis (EMH henceforth) (Fama 1991). The EMH is considered as bridging the gap between financial information and the financial market; it also affirms that the fluctuations in prices are only a result of newly available information; and that all available information reflected in market prices. The EMH assert that stocks are at all times in equilibrium and are difficult for inventors to speculate. Furthermore, it has been affirmed that stock prices do not pursue a random walk and stock prediction needs more evidence (Gallagher and Taylor, 2002).

In addition to purchasing and selling stocks and shares in stock markets, each stock is not only characterized by its price, but also by other variables such as closing price which represents the most important variable for predicting next day price for a specific stock. There is a relationship and specific behavior exists between all variables that affect stock movements overtime. Different economic factors, such as political stability, and other unforeseeable circumstances are variables that have been considered for stock price predictions (Fama 1991). The following table summarizes main variable that affect stock movements:

Variable	**Description**
Price	Current Price
Open	Opening price of a stock for the day
Close	Closing price of a stock for the day
High	Highest price of a stock for the day
Low	Lowest price of a stock for the day

Figure 1. Main variables that affect stock movement

In stock predictions, a set of pure technical data, fundamental data, and derived data are used in prediction of future values of stocks. The pure technical data is based on previous stock data while the fundamental data represents the company's' activity and the situation of market. When we combine these information about a company and its stock, we assume that we should be able to yield a prediction of the future price of the stock. In classification approaches using machine learning algorithms, a data set is divided into training data set and testing set. kNN uses similarity metrics to compare a given test entity with the training data set. Each data entity represents a record with n features. In order to predict a class label for unknown record, kNN selects k records of training data set that are closest to the unknown records.

The rest of the article will be structured as follows; section 2 will exemplify the review of some literatures that have already been published on using kNN for stock prediction, section 3 will describe the research methodology used and analysis that were conducted, section 4 will show the description of the data used and the results we obtained, and finally the conclusion is seen in section 6.

2. Literature Review

There are a lot of researches going on in the field of data mining and future prediction. Since financial securities can yield a lot of profit from small investment in time and capital, a lot of researches are being done especially in this field. Since there are a large amount of financial information sources in the world that can be valuable research areas, getting hands on to these data and using these data is not a difficult ask anymore. Stock prediction becomes increasingly important especially if number of rules could be created to help making better investment decisions in different stock markets. Hellstrom and Holmstrom used a statistical analysis based on a modified kNN to determine where correlated areas fall in the input space to improve the performance of prediction for the period 1987-1996 (Hellstorm, Holmstrom 1998). The study developed potential guidelines to mine pairs of stocks, stock-trading rules, and markets; it also showed that such approach is useful for real trading. Moreover, Qian and Rasheed adopted kNN as prediction techniques much like in this paper (Qian and Rasheed, 2007).

3. Research Methodology And Analysis

We implemented the kNN algorithm from scratch on python 2.7 to conduct the experiments for the paper. For implementing the algorithm, we took help from *"http://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/".*

3.1 About kNN

kNN is a instance-based, competitive learning, and lazy learning algorithm.

Instance based algorithms, sometimes called memory-based learning, are those algorithms that, instead of performing explicit generalization, use the instances seen in the training as a comparison standard. For kNN, the entire training dataset is the model. When a prediction is required for a unseen data instance, the kNN algorithm will search through the training dataset for the k-most similar instances.

kNN is a competitive learning model because a majority vote is performed among the selected k records to determine the class label and then assigned it to the query record.

kNN is considered a lazy learning that does not build a model or function previously, but yields the closest k records of the training data set that have the highest similarity to the test (i.e. query record).

The prediction attribute of the most similar instances is summarized and returned as the prediction for the unseen instance. The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. Other other types of data such as categorical or binary data, Hamming distance can be used. In the case of regression problems, the average of the predicted attribute may be returned. In the case of classification, the most prevalent class may be returned.

In this paper, we implement kNN in the following steps:

1. Handle data: Get data using yahoo finance and save and the load the dataset from CSV and split into test/train datasets
2. Similarity: Calculate the distance between two data instances

3. Neighbors: Locate k most similar data instances
4. Response: Use a majority vote for the class labels of k nearest neighbors and generate a response from a set of data instances
5. Accuracy: Summarize the accuracy of predictions
6. Main: Tie it all together

3.2 Mathematical Calculations and Visualizations Models

Accuracy:

Our data is structured such that we see “up” or “down” values for each day if its stock price has gone up or down respectively. So, since there are only two possible outcomes of the process, so we directly compare the predicted outcome with the actual outcome to calculate the accuracy of the model. We can see this implemented in the getAccuracy() function in our code.

Visualization:

For visualization, we have constructed a graph, for each stock, that shows the actual stock prices on a given day with the prediction for that day in figure 3.1. We also have a graph for each stock that shows the predicted against actual change in stock price movement. And finally, we also have created, for each stock, we show their stock performance using their return.

4. Data Description, Results, And Analysis

In this article, data from the NASDAQ stock exchange was analyzed and a brief data analysis was presented to provide the reader with the fundamental concepts of data attributes. Also, the obtained results of prediction of the NASDAQ stock exchange are provided.

4.1 Data Description

The sample data was extracted from the NASDAQ stock exchange from yahoo finance. The study sample included stock data of seven selected companies listed on the NASDAQ stock exchange. The selected stocks are shown in table 4.1. As a sample training dataset from the period Jan 1, 2011 to April 16, 2017, each of these companies has six attributes including Date, Opening price, High, Low, Adjusted Closing price, and state change of the stock as shown in table 4.1.2 (this only has some randomly selected data to show how the data looks like since showing all data will take lots of space). Closing price is the main factor that affects the prediction process for a specific stock based on kNN algorithm. In our case, we have used the Adjusted Closing price, which depicts the corporate actions on a stock's closing price for that day.

All of these data are received from Yahoo finance, however, we calculate the "Price Change" value for each stock for each day. We subtract the adjusted closing price for the stock for the previous day from the adjusted closing price of the stock for that day. And if the remaining number is positive, we give it a "up" label signifying that the stock price has gone up, else we label the state change as "down" to signify decrease in stock value.

The kNN algorithm is applied on 20294 records.

Table 4.1.1: The Stocks used that are listed in NASDAQ

Stock name	**Company**
AMTD	TD Ameritrade

AMZN	Amazon.com Inc
DIS	The Walt Disney Company
SBUX	Starbux
TWLO	Twilio
TWTR	Twitter Inc
YHOO	Yahoo Inc

Table 4.1.2: Sample data

Date	Open	High	Low	Yesterday Close	Price Change
2014-05-20	57.040001	57.450001	56.900002	56.770033	up
2016-08-25	59.139999	59.400002	58.689999	57.891185	up
2017-03-30	28.67	28.889999	28.120001	28.620001	down
2016-12-08	31.049999	31.780001	30.34	30.6	down
2012-11-19	15.46	15.55	15.39	13.65676	down
2015-06-08	38.450001	38.75	38.009998	36.873237	up
2012-05-10	17.91	18.139999	17.709999	15.527802	down

4.2 Analysis And Results

The results of the predicted stock price trend for each individual company used in the sample are presented as graphs along with the actual trend from fig 4.2.10 to 4.2.16. The actual trend can be seen in blue and the predicted trend can be seen in red. From fig 4.2.0 to 4.2.6, the stock chart for each stock can be seen. However, only this chart does not help much in determining the trend for analysts. And from fig 4.2.20 to 4.2.26, the predicted vs actual trend can be seen but from a zoomed in perspective, and this is the most helpful chart among the three as this shows more details on the trend. The same data is zoomed out in fig 4.2.10 to 4.2.16.

The table 4.2 shows the final result of the accuracy testing of the algorithm. Overall, the average accuracy of kNN from these dataset is 70.236%, which is a fairly good number for financial data prediction.

Table 4.2: Final result and accuracy

Stock	**Data used**	**Accuracy (in %)**
AMTD	Train: 2538 Test: 1308	72.314
AMZN	Train: 2575 Test: 1271	74.186
DIS	Train: 2536 Test: 1310	67.293
SBUX	Train: 2576 Test: 1270	68.530
TWLO	Train: 138 Test: 64	65.278
TWTR	Train: 562 Test: 300	69.611
YHOO	Train: 2535 Test: 1311	74.440
		Average = 70.236

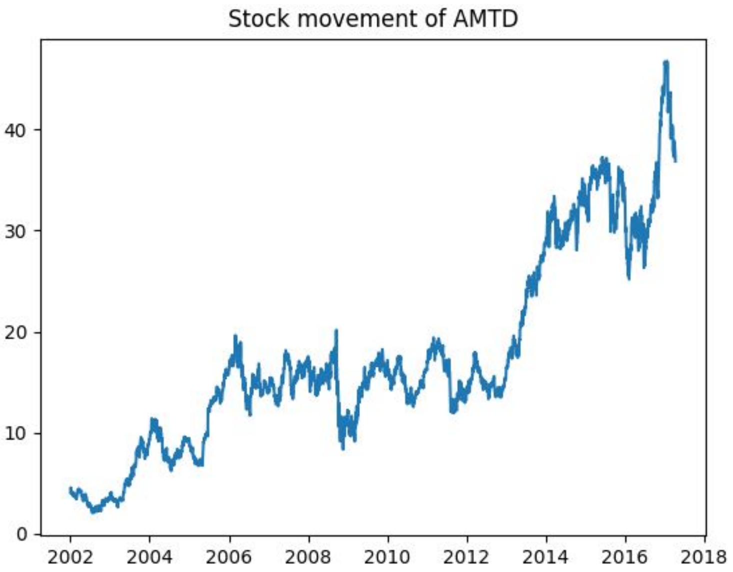

Fig 4.2.0

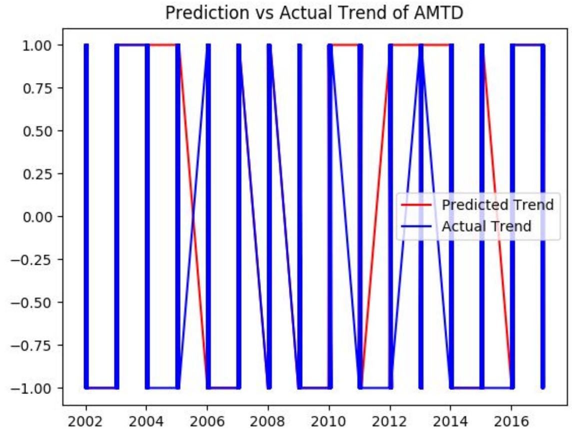

Fig 4.2.10

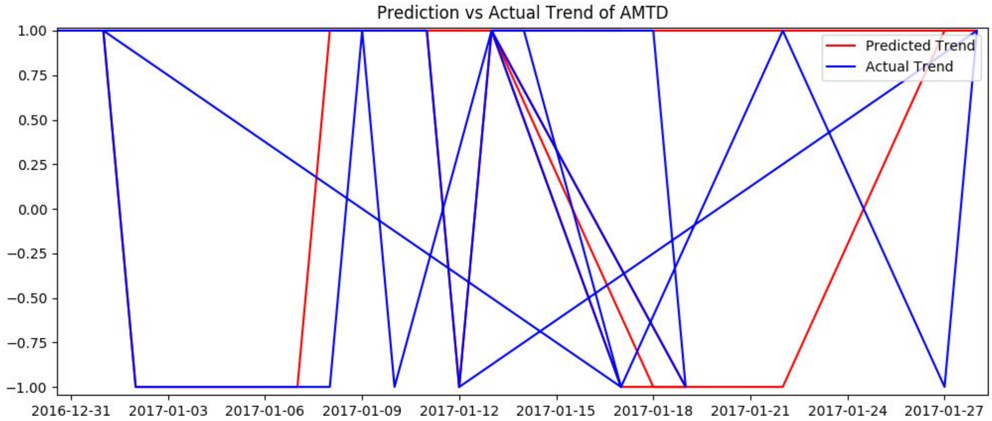

Fig 4.2.20

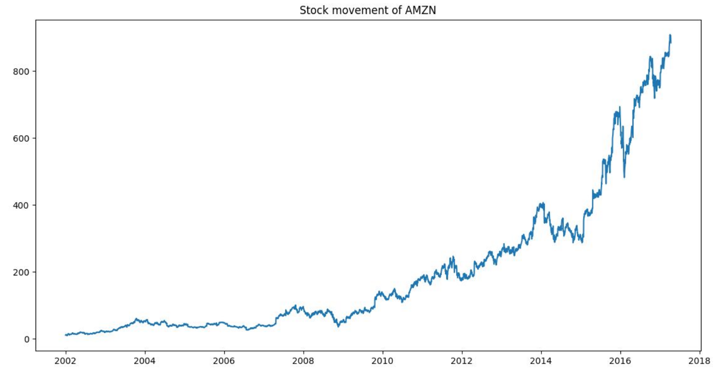

Fig 4.2.1

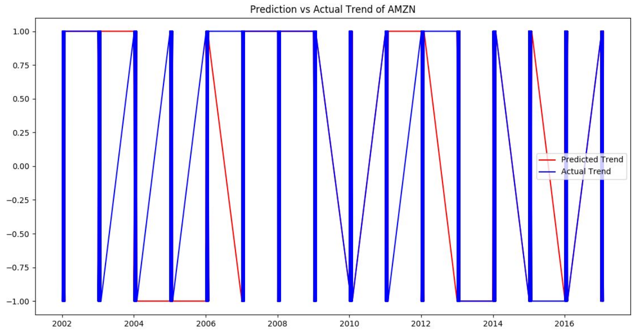

Fig 4.2.11

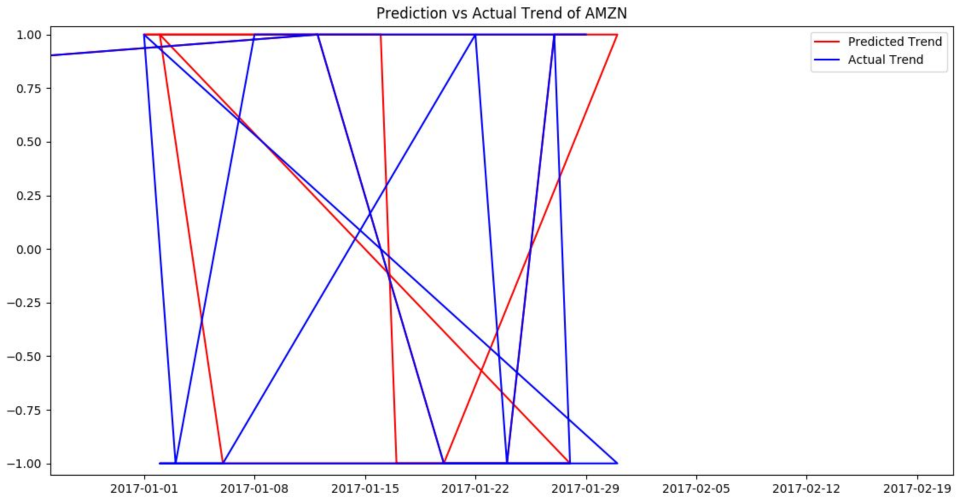

Fig 4.2.21

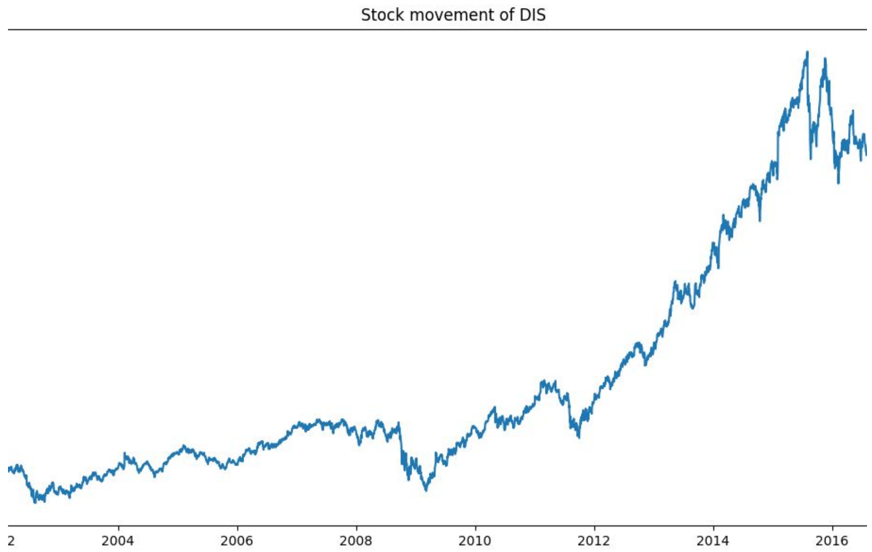

Fig 4.2.2

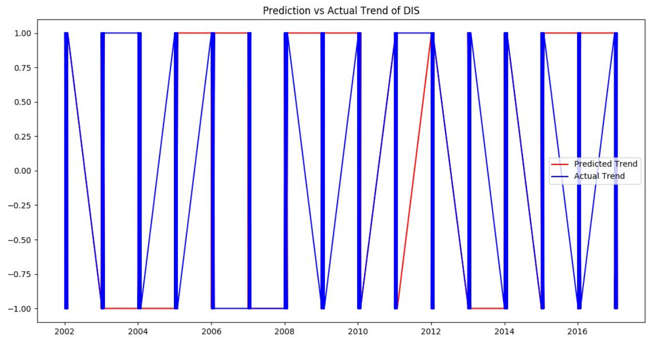

Fig 4.2.12

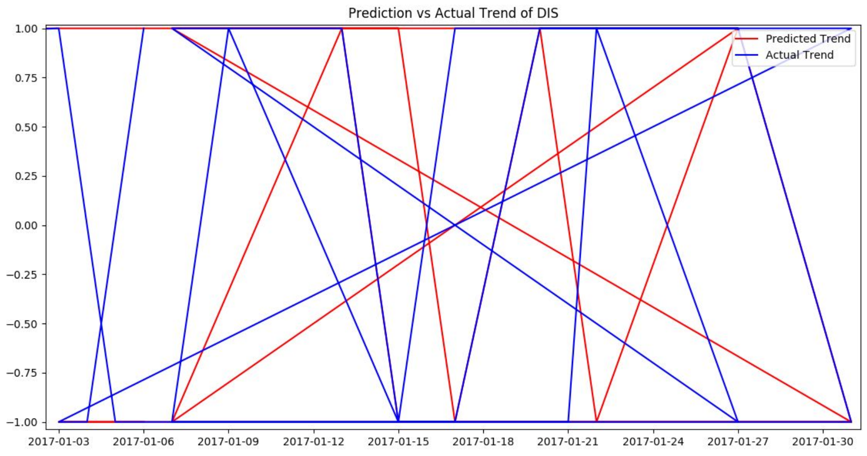

Fig 4.2.22

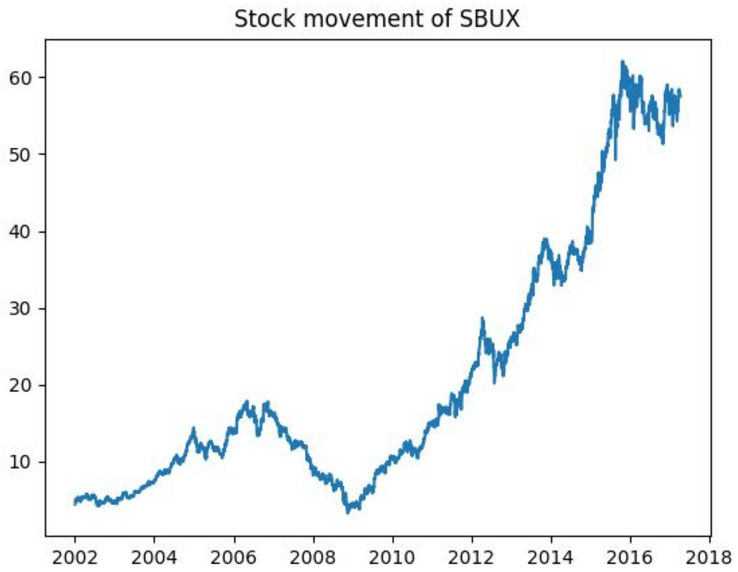

Fig 4.2.3

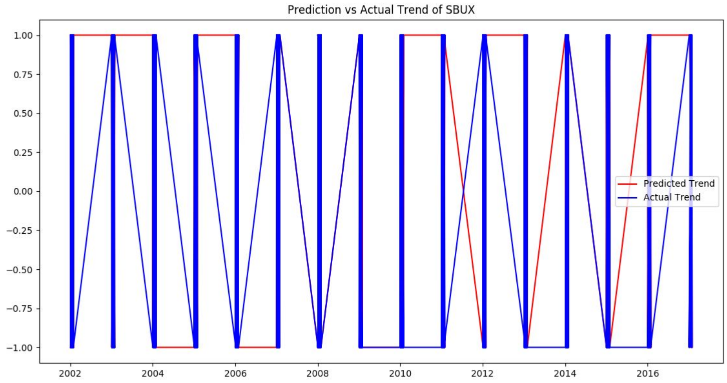

Fig 4.2.13

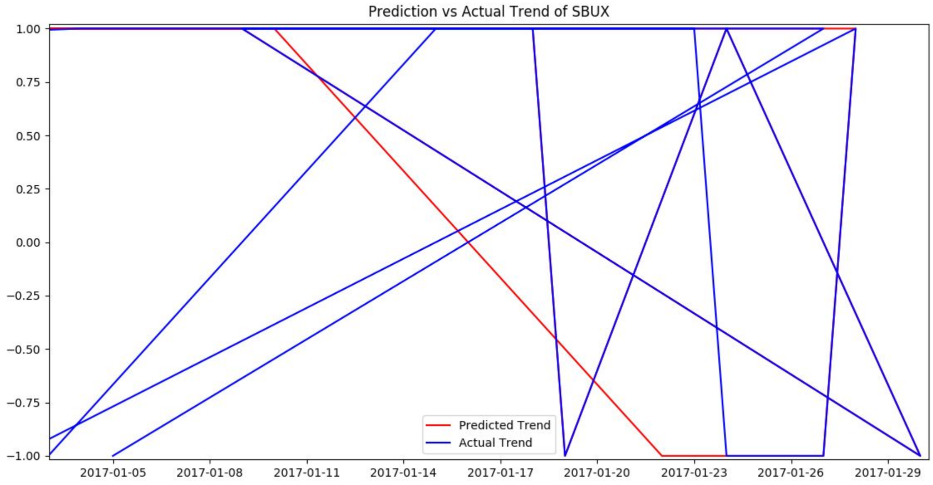

Fig 4.2.23

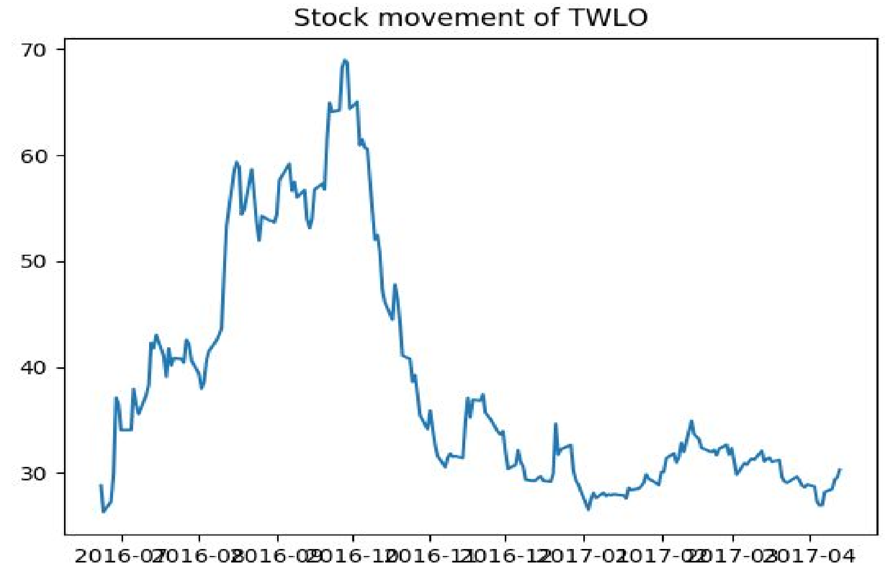

Fig 4.2.4

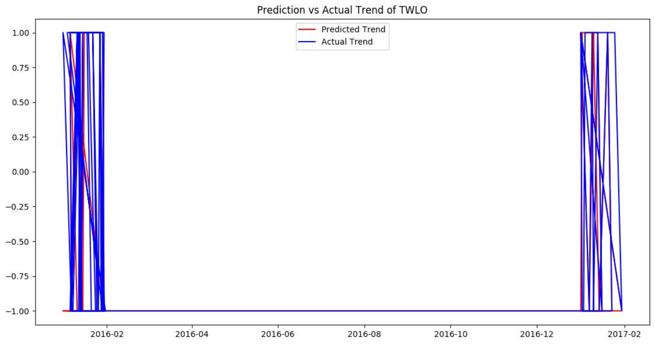

Fig 4.2.14

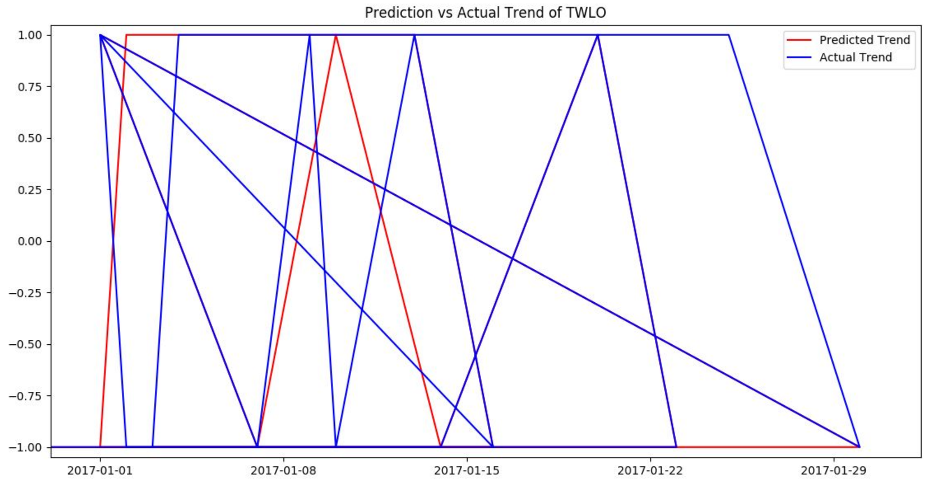

Fig 4.2.24

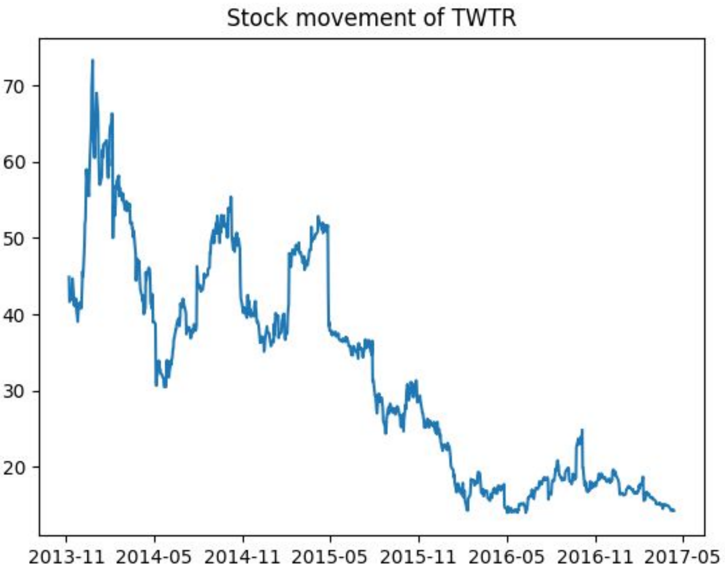

Fig 4.2.5

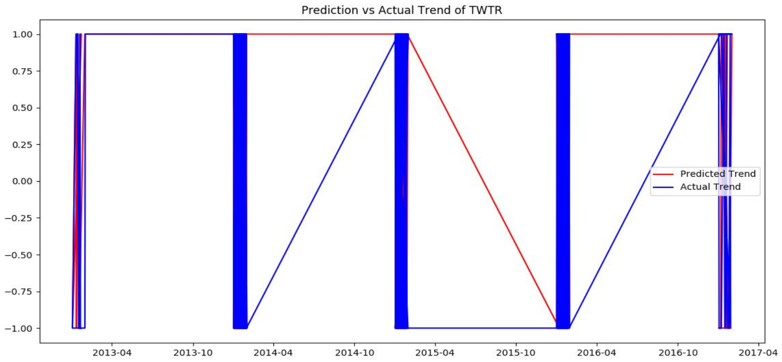

Fig 4.2.15

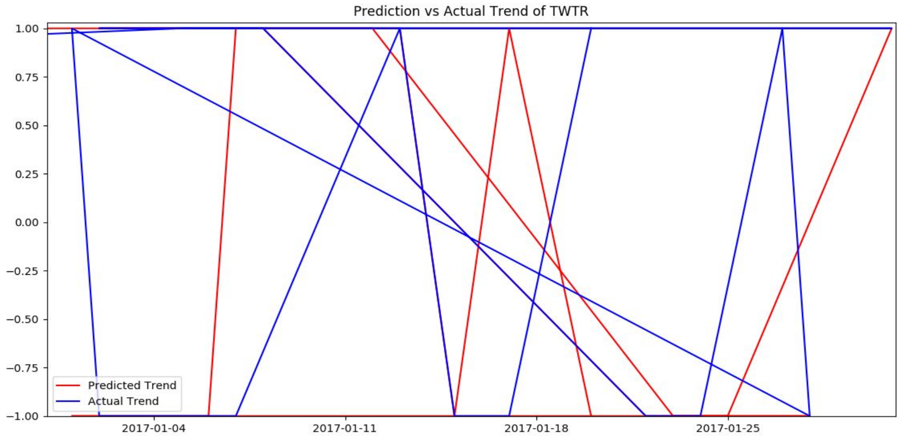

Fig 4.2.25

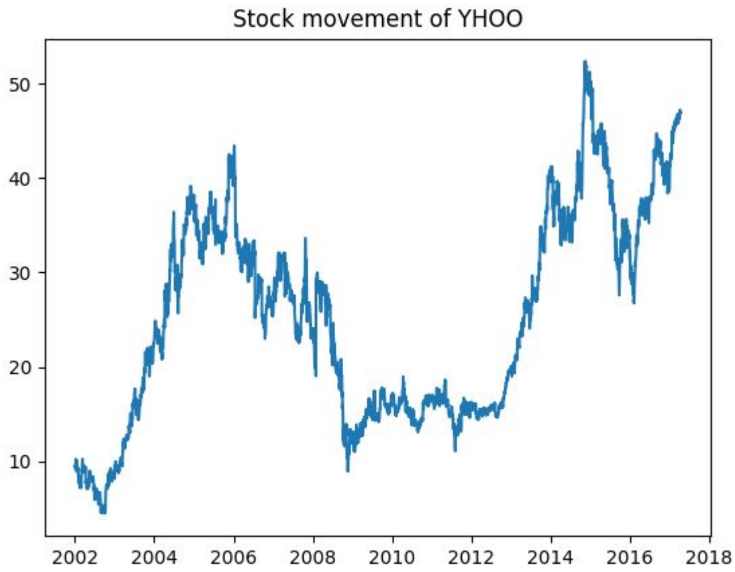

Fig 4.2.6

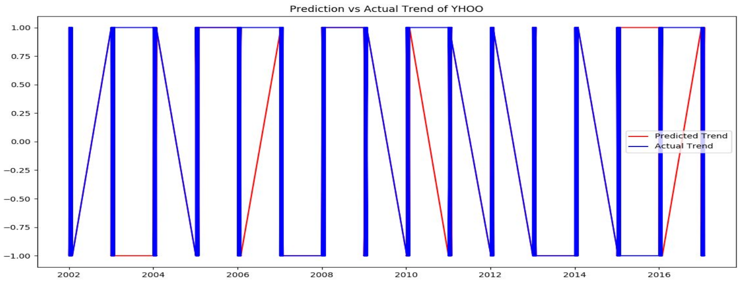

Fig 4.2.16

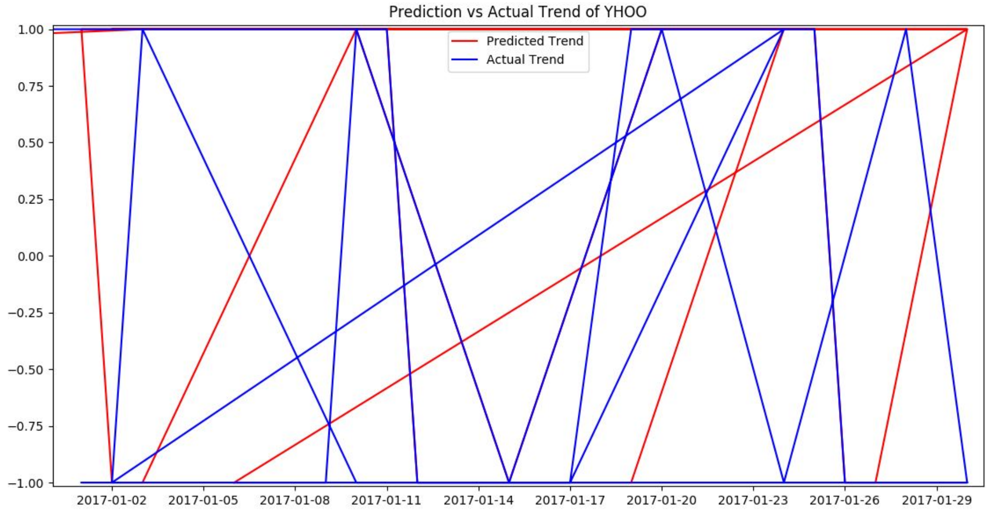

Fig 4.2.26

As depicted in the figures (4.2.10-4.2.16) above, the prediction and the actual trend overlap in a lot of areas. The average accuracy of 70.236% is visible in the graphs. This overlapping of actual and predicted trends shows that the values of errors are very small which indicate that the actual value and predicted value are close. This yields a high accuracy of using the kNN algorithm in predicting stock values.

5. Conclusion

In this paper, a prediction process for seven listed companies on the NASDAQ Stock Market was carried out. Consequently, a robust model was constructed for the purpose set out. The data was extracted from yahoo finance. We adopted an efficient prediction algorithm tool of kNN with k=5 to perform such tests on the training data sets we had. According to the results, kNN algorithm was stable and robust with good accuracy, so the results were rational and reasonable. In addition, depending on the actual stock prices data; the prediction results were close to actual prices. Having such rational results for predictions in specific, and for using data mining techniques in real life; this presents a good indication that the use of data mining techniques could help decision makers at various levels when using kNN for data analysis. So, we consider that employing this prediction model, kNN is real and viable for stock predictions. Nevertheless, the accuracy is not as much as we would want, since the decisions made in the stock market can lead to millions of dollars of loss, so there is still a long way to utilize advanced predicting models to help the financial markets and brokerage houses and to move forward and be part of the developed international financial markets. Furthermore, this may weaken the attractiveness of investments in the NASDAQ market which eventually weakens the market return. The study also shows that contemporary data mining techniques offer the world of finance useful stock market movements' prediction analysis.

Reference

Fama, E. F. (1991, December). Efficient Capital Markets: II. The Journal of Finance,25(2).

Gallagher, L. Taylor, M. (2002). Permanent and temporary components of stock prices: Evidence from assessing macroeconomic stocks. Southern Eco J 69, 245–262

Hellstrom, T., Holmstrom, K. (1998). Predicting the stock market. Technical report series. Center of Mathematical Modeling, Malardalen University

Qian, B., Rasheed, K. (2004). Hurst exponent and financial market predictability. Proceedings of the 2nd IASTED international conference on financial engineering and applications. Cambridge, MA, USA, 203–209