Proceedings of The National Conference
On Undergraduate Research (NCUR) 2016
University of North Carolina Asheville
Asheville, North Carolina
April 7-9, 2016

A Cross-Correlation-Based Stock Forecasting Model

Sungil Kim
Electrical and Computer Engineering
Auburn University
341 War Eagle Way
Auburn, Alabama 36849 USA

Faculty Advisor: Dr. Michael Baginski

Abstract

Researchers are continuously seeking to develop and improve the stock forecasting models by analyzing the past value of a company and predicting future performance based on past data trends. Many previous studies of technical analyses of stocks have focused heavily on forecasting a single stock price based on its own past data. This type of analysis is susceptible to stock market volatility and not very effective for intraday trading. In addition, it is difficult to apply a method used for a particular stock market sector to other market sectors. In this study, a cross-correlation-based forecasting model is proposed, using sets of closely related stocks to forecast future stock performance. For highly correlated pairs with a time delay of K days, the two stocks are assumed to exhibit a similar pattern in the short term—that is, predicting stock B’s price based on stock A’s price will reflect stock A’s future performance K days earlier. The forecasting model generates a buy or sell signal depending on the performance of one stock that influences the other stock with the lag. The accuracy of the developed model is measured using US stocks from the energy sector, which is more volatile than other indexes (i.e., S&P 500), and the technology sector. The proposed model accurately forecasts outcomes 87.2% of the time and generates 3.2% profit per dollar over the span of the 47-day forecast interval. This result shows that the developed forecasting model is ideal for high-risk, high-return investments.

Keywords: Stock Market, Cross-Correlation, Forecasting

1. Introduction

Researchers and investors have studied how to accurately predict the stock market to generate higher returns. Their approach can be categorized into three types of stock market analysis: fundamental, sentimental, and technical analyses. A fundamental analysis determines whether the current price of a company is underestimated or overestimated by looking at the finances of a company (i.e., earning per share, debt ratio, cash flow). If the current price is underestimated, fundamental analysts assign this particular stock as buy. The main drawbacks of this method are that (1) it is difficult to analyze companies from emerging markets or relatively young companies due to the lack of the sufficient financial reports [1], (2) it is time-consuming and difficult to automate the process, and (3) it is susceptible to unknown parameters that affect the stock market prices and market volatility.

A sentimental analysis is based on the assumption that investors’ sentiment affects the general trend of a stock. Previous study [2] has shown that news and social media coverage impacts stock prices. However, this type of analysis is highly susceptible to volatility due to the lack of a strong statistical relationship with volatility [3].

Unlike fundamental and sentimental analyses, technical analyses concern only the past data of a stock. Typically, a technical analysis is applied on a single stock. This approach is less time-consuming. However, during a period of unstable economy, a technical analysis based on a single stock is not effective. For example, the unexpected financial crisis, such as United Kingdom withdrawal from the European Union, directly translates to increase in volatility. When technical analysis is used in this case, it is very challenging to know the extent in which stocks are going down. On the other hand, when a pairwise approach is used, the amount of decrease in highly correlated stocks would vary. However, their trends (going down or up) are the same, and it is hypothesized that a technical analysis using a pair of stocks is more effective in highly volatile, unstable market condition. Unlike the previous study that uses a pairwise technical analysis [4], the proposed method in this paper is automatic. The approach from [4] requires users to input each different time delays and chooses which lag to use; rather, the proposed algorithm here automatically calculates such delays. Furthermore, the result from highly traded US stocks is shown in this paper while previous study [4] shows less frequently traded Australian stocks For low volume stocks, it might be impossible to buy and sell the stocks with a high quantity.

Cross-correlating two stocks works as follows: When the price of stock A is related to the price of stock B but there is a time delay of K days, predicting stock B's price based on stock A's price will reflect its future performance K days earlier. For highly correlated pairs, the two stocks are assumed to exhibit a similar pattern in the short term. For a long-term investment, an algorithm must be continuously run that buys stock B whenever stock A shows a marked increase in price if the correlation is strong and delayed by K days.

The proposed forecasting model discussed generates buy and sell signals along with corresponding trade dates and takes the following inputs: a pair of stocks, range of dates, correlation coefficient $(-1 < \rho < 1)$ threshold, and maximum number of tries. The model first retrieves data from two stocks in a specified range of dates. It then calculates the cross-correlation and finds (1) whether the two stocks are strongly correlated ($\rho > \rho$-threshold) and (2) whether the time delay (lag) is not zero, meaning a time delay occurred between the two stocks. After these two conditions are met, the forecasting model generates a buy or sell signal depending on the performance of one stock that influences the other stock with the lag. If either of the two conditions fails, the algorithm either adjusts the range of dates or changes the pair of stocks when it reaches the maximum number of tries.

2. Methods

The proposed forecasting model retrieves previous data of a pair of stocks and compute cross-correlation. It then checks two conditions: (1) the correlation coefficient is greater than the threshold (> 0.7) and (2) the time delay (lag) is non-zero. The function of "Cross-Correlate" has two outputs: how strong two stocks are correlated and when they have such optimal correlation (lag). Therefore, the algorithm automatically rejects when it detects lag of zero. Because this method focuses only on a long position of stocks, excluding a short position, the algorithm rejects when the time delay is negative. However, when a short position is in interest, simple modification in condition checking will result in pair of stocks that have high negative correlation. When either of these conditions is not met, the algorithm either modifies the time interval or change stock pairs if an attempt to modify the time interval has reached the maximum number of tries. When both conditions are satisfied, the model starts generating buy or sell signals and calculates the buy or sell dates depending on the lag. The algorithm flowchart and pseudocode are shown in Figure 1 and Figure 2, respectively.

The model can be applied to any set of stocks from various sectors, although it is more effective to target a pair such that the lag is less than two weeks with a high expected short-term gain. Furthermore, stocks with a previous history of drastic change are the main targets, as they often accompany more buy opportunities. Although the short term is relative, we consider it as between one to two months. Stock selection can be done from a set of provided stocks. They are initially picked from intuitions. For example, company A that produces a component that company B uses likely has significant correlation with company B. The model does not include the predicted price as the trend (either predicted up or down) is more important and the actual profit can be calculated after the lag.

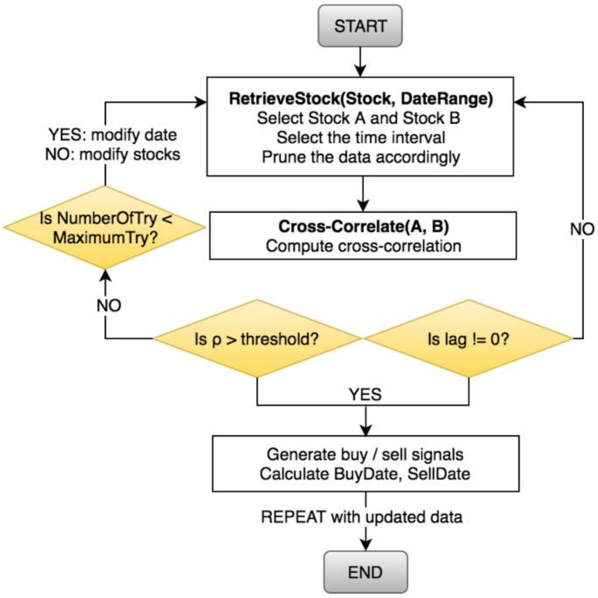
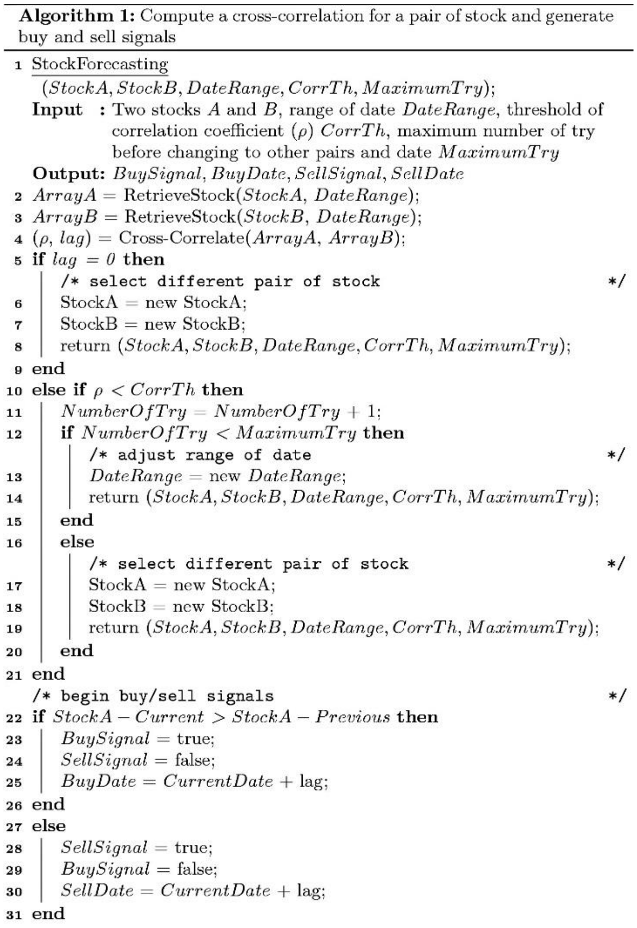

START
RetrieveStock(Stock, DateRange)
Select Stock A and Stock B
Select the time interval
Prune the data accordingly
YES: modify date
NO: modify stocks
Is NumberOfTry < MaximumTry?
Cross-Correlate(A, B)
Compute cross-correlation
NO
NO
Is ρ > threshold?
Is lag != 0?
YES
Generate buy / sell signals
Calculate BuyDate, SellDate
REPEAT with updated data
END

Figure 1. algorithm flowchart



Algorithm 1: Compute a cross-correlation for a pair of stock and generate buy and sell signals

```
StockForecasting
  ($StockA, StockB, DateRange, CorrTh, MaximumTry$);
  Input  : Two stocks $A$ and $B$, range of date $DateRange$, threshold of
           correlation coefficient ($\rho$) $CorrTh$, maximum number of try
           before changing to other pairs and date $MaximumTry$
  Output: $BuySignal, BuyDate, SellSignal, SellDate$
$ArrayA$ = RetrieveStock($StockA$, $DateRange$);
$ArrayB$ = RetrieveStock($StockB$, $DateRange$);
($\rho$, $lag$) = Cross-Correlate($ArrayA$, $ArrayB$);
if $lag = 0$ then
    /* select different pair of stock                                */
    StockA = new StockA;
    StockB = new StockB;
    return ($StockA, StockB, DateRange, CorrTh, MaximumTry$);
end
else if $\rho < CorrTh$ then
    $NumberOfTry = NumberOfTry + 1$;
    if $NumberOfTry < MaximumTry$ then
        /* adjust range of date                                      */
        $DateRange$ = new $DateRange$;
        return ($StockA, StockB, DateRange, CorrTh, MaximumTry$);
    end
    else
        /* select different pair of stock                            */
        StockA = new StockA;
        StockB = new StockB;
        return ($StockA, StockB, DateRange, CorrTh, MaximumTry$);
    end
end
/* begin buy/sell signals                                            */
if $StockA - Current > StockA - Previous$ then
    $BuySignal$ = true;
    $SellSignal$ = false;
    $BuyDate = CurrentDate +$ lag;
end
else
    $SellSignal$ = true;
    $BuySignal$ = false;
    $SellDate = CurrentDate +$ lag;
end
```

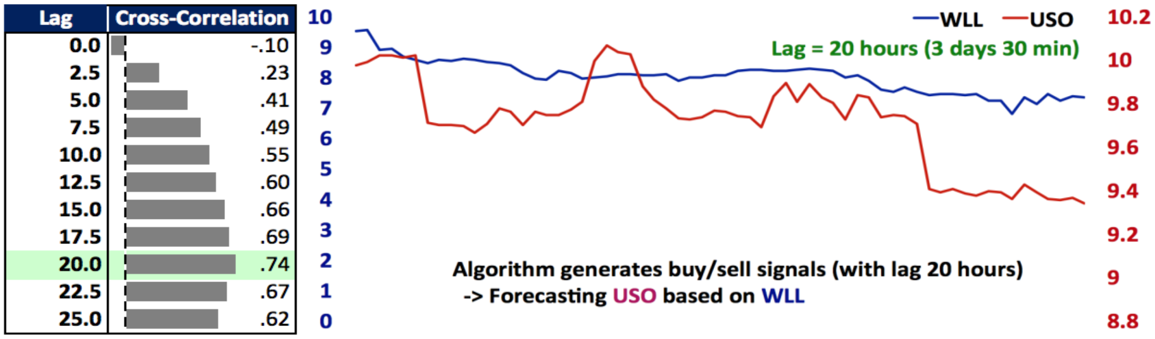
Figure 2. algorithm pseudocode

3. Results

Using US stocks from the energy sector, the lag for buy and sell signals as well as the accuracy and profit per dollar are reported. The reason we use the energy sector is to measure the resistance to volatility as the energy sector has a daily volatility of 3.9% (annualized: 61.9%). The process of selecting a stock that is highly correlated with United States Oil (USO) is shown in Table 1. From the set of stocks provided, the algorithm computes cross-correlation with each provided stock and rejects all except WLL and PGH. Between WLL and PGH, we choose Whiting Petroleum Corporation (WLL) due to the stronger cross-correlation with United States Oil Fund (USO), use data from the previous seven years to compute the cross-correlation, and forecast for 47 days. The lag and two stocks after the lag time K are shown in Figure 3 and Figure 4 for the sell and buy signal, respectively. Note that algorithm fails to generate buy or sell signal for lag of zero, but in Figure 3 and 4, it is shown for demonstration that two stocks are not highly correlated when lag is zero. The results show that the proposed model accurately forecasts the upward trend 15 out of 17 times (88.2%) and the downward trend 26 out of 30 times (86.7%), for a total of 87.2%, as shown in Table 2. Compared to a previous study [4] using a data-mining algorithm with a lagged correlation with 67% accuracy, the proposed model is significantly more accurate. Furthermore, the proposed model generates 3.2% profit per dollar over the span of the 47-day forecast interval.

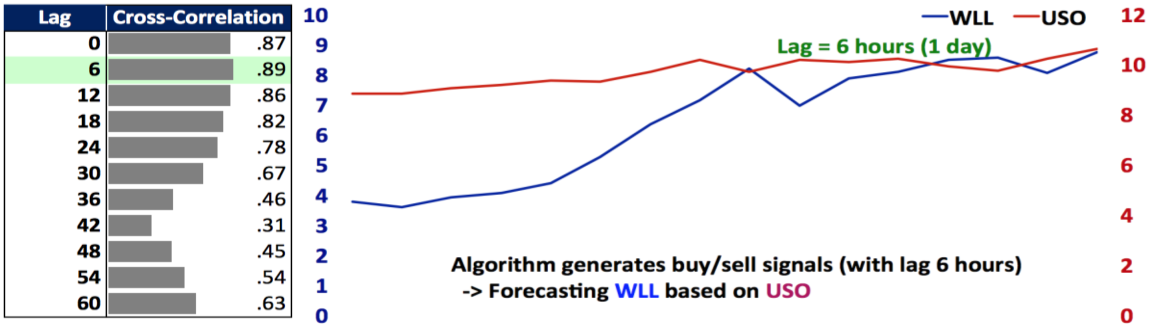
Table 1. Selection Process

Samples for pairing with USO	Cross-Correlation (1: maximum)	Lag (hours)	Accept or Reject
BP	0.57	6	Reject
CHK	0.75	0	Reject
ETE	0.49	0	Reject
PGH	0.78	1	Accept
WLL	0.89	6	Accept



Lag	Cross-Correlation
0.0	-.10
2.5	.23
5.0	.41
7.5	.49
10.0	.55
12.5	.60
15.0	.66
17.5	.69
20.0	.74
22.5	.67
25.0	.62

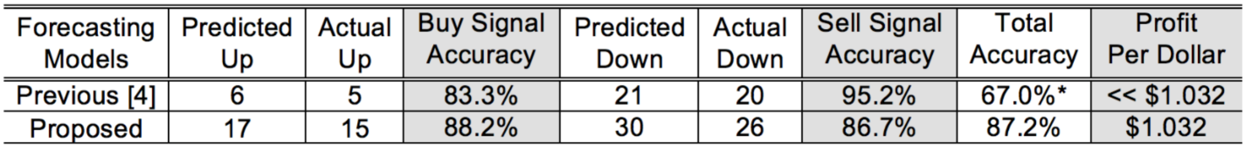
Figure 3. example of a sell signal



Lag	Cross-Correlation
0	.87
6	.89
12	.86
18	.82
24	.78
30	.67
36	.46
42	.31
48	.45
54	.54
60	.63

Figure 4. example of a buy signal

Table 2. results and comparison with previous study (* total accuracy of 67.0% of previous study is for the entire sets tested, and buy/sell signal accuracy is not reported for all data)



Forecasting Models	Predicted Up	Actual Up	Buy Signal Accuracy	Predicted Down	Actual Down	Sell Signal Accuracy	Total Accuracy	Profit Per Dollar
Previous [4]	6	5	83.3%	21	20	95.2%	67.0%*	<< $1.032
Proposed	17	15	88.2%	30	26	86.7%	87.2%	$1.032

4. Conclusion

This research developed a cross-correlation-based forecasting model and demonstrated that a pair of strongly related stocks shows a similar trend in the near future. During the 47-day interval, the proposed model showed an accuracy of 87.2% compared to 67% from the previous study [4]. Furthermore, the sets of stocks targeted from previous study [4] are low volume stocks, and there are not enough buy opportunities to make a meaningful profit. On the other hand, our proposed model has enough buy opportunities for energy sector stocks whose annualized volatility is 61.9% (daily: 3.9%), and the profit is 3.2% over 47 days. This result shows that the cross-correlation-based stock forecasting model is effective for highly volatile, short-term trading. The proposed model provides new insights for researchers, investors, and individuals regarding how cross-correlation can improve the accuracy of forecasting highly volatile stocks. The future work would be broader analysis of stocks from different sectors and comparison of the distribution of investment returns.

5. Acknowledgement

The author appreciates the funding from Auburn University competitive undergraduate research fellowship. Moreover, the author would like to thank Samuel Ginn College of Engineering, Department of Electrical and Computer Engineering, and Honors College.

6. References

1. Ahmed Wafi, Hassan Hassan, and Adel Mabrouk, "Fundamental Analysis Models in Financial Markets - Review Study," *Procedia Economics and Finance*, vol. 30, pp. 939-947, 2015.
2. Kuei-Yuan Wang and Chien-Kuo Chen, "The Influence of Media Coverage on the Stock Returns and Momentum Profits," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS),* July 2012, pp. 943-947.
3. Rapheal Olaniyan, Daniel Stamate, Lahcen Ouarbya, and Doina Logofatu, "Sentiment and stock market volatility predictive modeling - a hybrid approach," in *Data Science and Advanced Analytics (DSAA)*, Oct. 2015, pp. 1-10.
4. Cicil Fonseka and Liwan Liyanage, "A Data Mining Algorithm to Analyse Stock Market Data using Lagged Correlation," in *Information and Automation for Sustainability (ICIAFS),* Dec. 2008, pp. 163-166.