Data Driven Automated Algorithmic Trading

Gabriel Gauci Maistre

Abstract — Various existing stock market price forecasting methods were analysed in this report. Three methods were applied towards the problem making use of Technical Analysis, these were Time Series Analysis, Machine Learning, and Bayesian Statistics. Through the results of this report, it was found that the Efficient Market Hypothesis remains true, that past data does not contain enough useful information to forecast future prices and gain an advantage over the market. However, the results proved that Technical Analysis and Machine Learning could still be used to guide an investors decision. It was also found that the Random Walk Hypothesis was not necessarily true, as some stocks showed signs of auto and partial correlation. A common application of technical analysis was demonstrated and shown to produce limited useful information in beating the market. Based on the findings, a number of automated trading algorithms were developed using machine learning and backtested to determine their effectiveness.

Keywords — machine learning, time series analysis, probabilistic, bayesian, statistics, inference

1. Introduction

The stock market retains its status as a prime location for investors to invest in the market and earn a profit, however this is not always easy due to the constantly thriving and changing nature which follows the stock market. Investors are constantly presented with numerous profit potential opportunities, however without intensive planning and analysis, these opportunities could easily turn into losses. This means that it is crucial for every investor to carry out stock market ananlysis prior to any investment by monitoring past price movements in order to forecast future trends. Even though past data is not a clear indication of future movement, it is still proven to provide some useful insite.

2. Related work

There is a growing demand for forecasting interest rates, as financial researchers, economists, and players in the fixed income markets seek to find the best method to get ahead of the market. A study was carried out to develop an appropriate model for forecasting the short-term interest rates, implicit yield on 91 day treasury bill, overnight MIBOR rate, and call money rate. The short-term interest rates are forecasted using univariate models such as the Random Walk, ARIMA, ARMA-GARCH, and ARMA-EGARCH. The appropriate model for forecasting is determined considering a six-year period from 1999. Radha et al. showed evidence that GARCH models are best suited for fore- castint when applied towards time series having volatility clustering effects. It was their firm belief that ARIMA-EGARCH is the most appropriate forecasting model for these circumstances. Darrat et al. set out to investigate with the use of new daily data, whether prices in the two Chinese stock exchanges (Shanghai and Shenzhen) follow a random walk process as required by market efficiency. Two different approaches were applied, the standard variance-ratio test, and a model-comparison test that compares the ex post forecasts from a naive model with those obtained from several alternative models such as ARIMA, GARCH, and ANNs. To evaluate ex post forecasts, Darrat et al. made use of several procedures including root-mean-square error (RMSE), mean absolute error (MAE), uncertainty coefficient, and encompassing tests. It was concluded that the model-comparison approach yielded results which were quite strongly rejected the RWH in both Chinese stock markets when compared with the variance-ratio test. Darret et al. recommended the use of ANNs, as their results showed strong support for the model as a potentially useful factor for forecasting stock prices in emerging markets.

Support vector machines (SVM), are a class of machine learning algorithms that have become incredibly popular in the past few years. SVMs are very similar to classifiers in the sense that they also classify data by drawing a line, called a decision boundary, to separate them. However, SVMs go a step further by calculating a vector from the data point with the smallest margin to the decision boundary. This is called a support vector. A vast majority of academics tend to predict the price of stocks in financial markets, however most models used are flawed and only focus on the accurate forecasting of the levels of the underlying stock index. There is a lack of studies examining the predictabil- ity of the direction of stock index movement. Given the notion that a prediction with little forecast error does not necessarily translate into capital gain, the authors of this research attempt to predict the direction of the S&P CNX NIFTY Market Index of the National Stock Exchange, one of the fastest growing financial exchanges in developing Asian countries. Machine learning models such as random forest and SVMs, differ widely from other models, and are making strides in predicting the financial markets. Kumar et al. tested classification models to predict the direction of the markets, by applying models such as linear discriminant analysis, logistic regression, ANNs, random forest, and SVM. Their evidence shows that SVMs outperform the other classification methods in terms of predicting the direction of the stock market movement, and that the random forest model outperforms other models such as ANNs, discriminant analysis, and logistic regression.

Decision tree learning uses a decision tree as a predictive model which maps observations about an item, represented in the branches, to conclusions about the items target value represented in the leaves. Tree models where the target variable can take a finite set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values, typically real numbers, are called regression trees. In their research, Creamer et al. developed a automated trading algorithm making use of multiple stocks relying on a layered structure consisting of a machine learning algorithm, an online learning utility, and a risk management overlay. The machine learning algorithm which they made use of was an Alternating Decision Tree (ADT) implemented with Logitboost. Their algorithm was able to select the best combination of rules derived from well known technical analysis indicators, and the best peramaters of the indicators in question. Additionally, their online learning layer was also able to combine the output of several ADTs, suggesting a short or long position. Finally, the risk management layer in which they implemented, was able to validate the trading signal once it exceeds a specified non-zero threshold and limit the application of their trading strategy when it is not profitable. They tested the expert weighting algorithm with data of 100 randomly selected companies of the S&P 500 index dur- ing the period 20032005. They found that their algorithm generated abnormal returns during the test period. Their experiments show that the boosting approach was able to improve the predictive capacity when indicators were combined and aggregated as a single predictor. Furthermode, their results indicated that the combination of indicators of different stocks were adequate in order to reduce the use of computational resources, while still maintaining an adequate predictive capacity.

Hoffman et al. introduce the No-U-Turn-Sampler, an extension to the Hamiltonian Monte Carlo (HMC), which is an MCMC algorithm that avoids the random walk behavioir and sensitivity to correlated paramaters that plague many MCMC methods by taking a series of steps informed by first-order gradient information. In their paper, they claim HMCs performance to be highly sensitive to two user-specified parameters: a step size, and a desired number of steps. NUTS is an improvement from HMC as it eliminates the need to set a number of steps, and works by build a set of likely candidate points that spans a wide swath of the target distribution, stopping automatically when it starts to double back and retrace its steps. They achieved all of this by making use of a recursive algorithm.

3. Conclusion

Three financial forecasting methods were presented in this report, two of which showed little to no potential of ever producing any statistically significant result when the correct methodology was applied. The third method, machine learning, showed some potential in the tests carried out, which is why this method was built into an automated algo- rithmic strategy to trade with. The algorithm proved to be successful in forecasting future prices, using both classification and regression methods. However, the backtest- ing proved this method to fail in forecasting price falls. Once this factor was removed from the equation, the algorithms were very successful and reported a profit by the end of the test. This is however not always ideal as stocks which could fall in price could be catastrophic to the strategy. A stop loss would be ideal in insuring that no positions are held in downward falling stocks. It was also evident that regression methods were more successful in forecasting future price movements when compared to classification methods.

If there is anything that this report shows, is that profitable stock market prediction is an extremely tough problem. Even though the strategies reported a profit by the end of the backtest, they still did not beat the market. Whether it is at all possible to use such methods to outperform the markets returns, ultimately remains an open question. These findings support the Efficient Market Hypothesis, proving that casual investors are better off investing in passive buy and hold strategies consisting of index funds and ETFs. However, there was some evidence found showing that the Random Walk Hypothesis does not hold true for all cases, as some stocks did show signs of repeating trends.

Acknowledgements

The author would like to express his special thanks of gratitude to my supervisor, Alan Gatt, for the patient guidance, encouragement, and advice he has provided throughout my time as his student. He would also like to thank Luke Vella Critien, for guiding him towards the right parth in the early stages of my research and for recommending Alan Gatt as my tutor. The author's gratitude is also extended to Emma Galea and Miguel Attard for their valuable input while carrying out his research. Completing this work would have been all the more difficult were it not for the support and friendship provided by the other members of the Malta College of Arts, Sciences, and Technology, and the institute of Information and Technology. The author is indebted to them for their help. Finally, the author wishes to thank his family who have supported me all throughout the final year of his bachelors.

References

[1] K. R. Wilson and V. V. Yakovlev, „Ultrafast rainbow: tunable ultrashort pulses from a solid-state kilohertz system", *J. Opt. Soc. Am. B*, vol. 14, pp. 444–448, 1997.

[2] J. Comly and E. Garmire, „Second harmonic generation from short pulses", *Appl. Phys. Lett.*, vol. 12, no. 7-9, 1968.

[3] O. E. Martinez, „Achromatic phase matching for second harmonic generation of femtosecond pulses", *IEEE J. Quant. Electron.*, vol. QE-25, pp. 2464–2468, 1989.

[4] G. Szabo and Z. Bor, „Broadband frequency doubler for femtosecond pulses", *Appl. Phys. B*, vol. 50, pp. 51–54, 1990.

[5] J.-Y. Zhang, J. Y. Huang, H. Wang, K. S. Wong, and G.K. Wong, „Second-harmonic generation from regeneratively amplified femtosecond laser pulses in BBO and LBO crystals", *J. Opt. Soc. Am. B*, vol. 15, pp. 200–209, 1998.

[6] K. Hayata and M. Koshiba, „Group-velocity-matched second-harmonic generation: an efficient scheme for femtosecond ultraviolet pulse generation in periodically domain-inverted β-BaB2O4", *Appl. Phys. Lett.*, vol. 62, pp. 2188–2190, 1993.

[7] G. Y. Wang and E. M. Garmire, „High-efficiency generation of ultrashort second-harmonic pulses based on the Cherenkov geometry", *Opt. Lett.*, vol. 19, pp. 254–256, 1994.

[8] C. Radzewicz, Y. B. Band, G. W. Pearson, and J. S. Krasinski, „Short pulse nonlinear frequency conversion without group-velocity-mismatch broadening", *Opt. Commun.*, vol. 117, pp. 295–303, 1995.

[9] P. Di Trapani, A. Andreoni, G. P. Banfi, C. Solcia, R. Danielius, A. Piskarskas, P. Foggi, M. Monguzzi, and C. Sozzi, „Group-velocity self-matching of femtosecond pulses in noncollinear parametric generation", *Phys. Rev. A*, vol. 51, pp. 3164–3168, 1995.

[10] V. Krylov, A. Kalintsev, A. Rebane, D. Erni, and U. P. Wild, „Noncollinear parametric generation in LiIO3 and β-barium borate by frequency-doubled femtosecond Ti:sapphire laser pulses", *Opt. Lett.*, vol. 20, pp. 151–153, 1995.

[11] P. Di Trapani, A. Andreoni, C. Solcia, P. Foggi, R. Danielius, A. Dubietis, and A. Piskarskas, „Matching of group velocities in three-wave parametric interaction with fs pulses and application to travelling-wave generators", *J. Opt. Soc. Am. B*, vol. 12, pp. 2237–2244, 1995.

[12] P. Di Trapani, A. Andreoni, P. Foggi, C. Solcia, R. Danielius, and A. Piskarskas, „Efficient conversion of femtosecond blue pulses by travelling-wave parametric generation in non-collinear phase matching", *Opt. Commun.*, vol. 119, pp. 327–332, 1995.

[13] T. R. Zhang, H. R. Choo, and M. C. Downer, „Phase and group velocity matching for second harmonic generation of femtosecond pulses", *Appl. Opt.*, vol. 29, pp. 3927–3933, 1990.

[14] A. Andreoni and M. Bondani, „Group-velocity control in the mixing of three non-collinear phase-matched waves", *Appl. Opt.*, vol. 37, pp. 2414–2423, 1998.

[15] P. Di Trapani, A. Andreoni, C. Solcia, G. P. Banfi, R. Danielius, A. Piskarskas, and P. Foggi, „Powerful sub-100-fs pulses broadly tunable in the visible from a blue-pumped parametric ganerator and amplifier", *J. Opt. Soc. Am. B*, vol. 14, pp. 1245–1248, 1997.

[16] R. Danielius, A. Piskarskas, P. Di Trapani, A. Andreoni, C. Solcia, and P. Foggi, „Matching of group velocities by spatial walk-off in collinear three-wave interaction with tilted pulses", *Opt. Lett.*, vol. 21, pp. 973–975, 1996.

[17] R. Danielius, A. Piskarskas, A. Stabinis, G. P. Banfi, P. Di Trapani, and R. Righini, „Travelling-wave parametric generation of widely tunable, highly coherent femtosecond light pulses", *J. Opt. Soc. Am. B*, vol. 10, pp. 2222–2232, 1993.

[18] S. A. Akhmanov, A. S. Chirkin, K. N. Drabovich, A. I. Kovrigin, R. V. Khokhlov, and A. P. Sukhorukov, „Nonstationary nonlinear optical effects and ultrashort light pulse formation", *IEEE J. Quant. Electron.*, vol. QE-4, pp. 598–605, 1968.

[19] R. Danielius, A. Piskarskas, P. Di Trapani, A. Andreoni, C. Solcia, and P. Foggi, „A collinearly phase-matched parametric generator/amplifier of visible femtosecond pulses", *IEEE J. Quant. Electron.*, vol. 34, pp. 459–464, 1998.

[20] S. Sartania, Z. Cheng, M. Lenzner, G. Tempea, Ch. Spielmann, F. Krausz, and K. Ferencz, „Generation of 0.1-TW 5-fs optical pulses at a 1-kHz repetition rate", *Opt. Lett.*, vol. 22, pp. 1562–1564, 1997.

[21] M. Nisoli, S. De Silvestri, O. Svelto, R. Szipoecs, K. Ferencz, Ch. Spielmann, S. Sartania, and F. Krausz, „Compression of high-energy laser pulses below 5 fs", *Opt. Lett.*, vol. 22, pp. 522–524, 1997.

[22] S. A. Akhmanov, V. A. Vysloukh, and A. S. Chirkin, *Optics of femtosecond laser pulses*. New York: American Institute of Physics, 1992.

[23] A. Andreoni, M. Bondani, and M. A. C. Potenza, „Ultra-broadband and chirp-free frequency doubling in β-barium borate", *Opt. Commun.*, vol. 154, pp. 376–382, 1998.

[24] H. Wang, K. S. Wong, D. Deng, Z. Xu, G. K. L. Wong, and J. Zhang, „Kilohertz femtosecond UV-pumped visible β-barium borate and lithium triborate optical parametric generator and amplifier", *Appl. Opt.*, vol. 36, pp. 1889–1893, 1997.