A Review of *Identification of Nonlinear Determinants of Stock Indices Derived by Random Forest Algorithm*

Background and Motivation

Traditional financial market research on stock indices has been dominated by regression-based models typically composed of a fairly limited set of macroeconomic factors such as GDP, interest rates, GDP changes, and inflation rate. These models have a firm foundation in economics and tend to assume a linear relationship between variables (Al-Shubiri, 2010; Al-Tamimi et al., 2011). These models have proven to be very robust and do provide good interpretability but as with many linear modeling approach they can miss nonlinear interactions common in the global financial markets of which stock indices are composed. The intent of the article is to address this need.

According to Tratkowski's (2020) article there are two important limitations in the current literature. He asserts that the explanatory variables tend to be narrowly selected and based on economic theory, and that the predominant modeling approaches in the literature to not account for nonlinear interactions between variables. The author proposes Random Forest algorithms as an effective modeling approach to addressing the latter issue of nonlinear interactions and identifying nonlinear determinants in stock indices.

The author's motive in this research is to provide additional tools for making stock index theoretical work more robust. He also notes that the practical application of these approaches could provide competitive advantage and risk management tools to fund managers and investors by more accurately identifying nonlinear determinants. Central to the research is discovering what variables serve as the most influential on changes in stock market indices.

Methods Used

Tratkowski (2020) applies the supervised Random Forest algorithm to the problem. Random forests use a set of easily trained decision trees using a random sample of the data (Breiman, 2001; Ho, 1998). Each individual tree may be fairly inaccurate but the net effect of each tree "voting" on the predicted outcome provides an outcome that is more accurate that the individual models. This also serves to improve generalization of the model and to provide insight into the importance of each variables' contribution to the model. This is the key strength of Random Forest in that it provide this explainability and insight into the relative power of each variable without making any assumptions about model linearity.

The dataset used in the study used 20 years of daily observations January 2000-2020 composed of data from the 20 largest companies on the Warsaw Stock Exchange (WIG20), the 30 largest public companies in Germany (DAX) , and 600 large companies from 17 European countries (Stoxx Europe 600). 209 variables were included in the study with the same number of observations for each variable. The variables covered macroeconomic indicators, bond information, exchange rates, and commodity prices.

Data preparation was performed with z-score standardization. The prediction/response variable selected was the future one-month price change discretized to predict 1 if price increased more than 2%, 0 if price changed -2 to 2%, and -1 if the price changed by more than -2%. A rolling 5-year window cross validation was used in place of train/test data split to accommodate the lack of

stationarity in stock market time series data. Overfitting was reduced using Recursive Feature Elimination (RFE) algorithm (Geneure et. al., 2010) which selected the top 5 most important features for each month and eliminated the other features. The selected 5 features are used to fit the model for that month.

Significance of the Work

The determinants noted in the Polish Stock Exchange (WIG20) data demonstrate the interdependency of global financial markets. The three highest impact factors are measures from China, Germany, and the US. The top 6 most important factors represent data from 5 different countries. Manufacturing and production measures in the United States are also noted to be significantly impactful demonstrating the importance of the US productivity on global finance. The China Prime Lending Rate demonstrated a dominance in the factors for each market.

In practice for fund managers and investors this demonstrates clearly the importance of monitoring foreign policy changes and productivity numbers. In academia the addition of Algorithmic Modeling in contract to Data Modeling can provide a focus on results and outcomes versus carefully selected parameters and linear models (Breiman, 2001).

Connection to Other Work

The article seeks to take a most algorithmic approach to stock market determinant modeling in place of the common linear economic models of determinants. Studies specifically highlighted in the references as being in the traditional vein include those by Al-Shubiri (2010), Al-Tamimi et al. (2011), and Garefalakis et al. (2013). The author particularly mentions these studies using earnings and dividends as factors used in comparison to external drivers of index value. Tratkowski bases his work on prior work of Ho (1998), who developed the Random Forest algorithm and Breiman (2001), who extended it.

This paper is significant in that it takes a purely algorithmic approach to estimating determinants allowing the modeling of non-linear effects in dynamic global economic markets.

Relevance to Capstone Project

This paper is directly relevant to my capstone project on using the Wisdom of the Crowd as implemented in Random Forest models to model financial systems. I did not know of the existence of Recursive Feature Elimination (RFE) before this and it's a great approach I intend to add to my Random Forest models over time series data. I also am now aware of Rolling Window Cross Validation, and this knowledge could have made my final project in my UWF Time Series class more powerful.

My capstone project will not study the same dataset or type of data as used in this study, however this does provide validation and tools I can use in estimating determinants using Random Forest models in my final project.

References

Al-Shubiri, F. N. (2010). Analysis of the determinants of market stock price movements: An empirical study of Jordanian commercial banks. *International Journal of Business and Management, 5*(10), 137.

Al-Tamimi, H. A. H., Alwan, A. A., & Abdel Rahman, A. A. (2011). Factors affecting stock prices in the UAE financial markets. *Journal of Transnational Management, 16*(1), 3–19.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Chen, S. S. (2009). Predicting the bear stock market: Macroeconomic variables as leading indicators. *Journal of Banking & Finance, 33*(2), 211–223.

Demir, C. (2019). Macroeconomic determinants of stock market fluctuations: The case of BIST-100. *Economies, 7*(1), 8.

Garefalakis, A., Dimitras, A. I., & Lemonakis, C. (2013). The determinants of stock market development: Evidence from European Union. *International Journal of Business and Management, 8*(19), 101–114.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844.

Jeon, J. H. (2020). Macro and non-macro determinants of Korean tourism stock performance: A quantile regression approach. *The Journal of Asian Finance, Economics, and Business, 7*(3), 149–156.

Tratkowski, G. (2020). Identification of nonlinear determinants of stock indices derived by random forest algorithm. *International Journal of Management and Economics, 56*(3), 209–217.