Reply to the Reviewer's (Reviewer 1) Comments on *Predicting the Direction of Stock Market Price Using Tree-Based Classifiers*

June 16, 2018

Before we begin addressing the specific points and nuances raised by the reviewer, we thank the reviewer for very helpful comments and explain the main purpose of the paper.

The aim of the current work is to predict the direction that the price of a stock of a certain company will move towards. This is effectively a gain/loss scenario where we do not try to understand 'how much we have gained' or 'how much we have lost', rather only 'if we have gained anything' or 'if we have lost anything'. This is a deviation from the popular idea of predicting the monetary returns of a stock after a certain time period. This is the problem statement and this is what we have tried to address by using methods in machine learning. Admittedly, we had previously used the word 'returns' in an ambiguous sense in the manuscript and have clarified it in the revised submission. However, with this point clarified, let us address the general comment.

Commment: "The proportion of positive and negative data are in range of 45:55". Here you mean trading days with positive returns? Averaged across all your stocks (in all markets)? This is potentially misleading. You should include the median returns over various n-day periods in your (testing) dataset. If stock prices are stochastic but have positive drift, wouldn't sign predictability get easier with increasing forecast horizon?"

Response: We have rephrased the sentence to avoid any confusion. However, the purpose here is to generalize and develop a method to predict with reasonable efficiency whether the price of a stock will increase or decrease on day $n+t$ as compared to day n. It seems that the mean accuracy is a more representative measure of the efficacy of the system as compared to the median returns. Certainly, in case of a positive drift in the prices of a stock it may be possible to have a better sign predictability, but that does not seem to have a conflict with our method of prediction.

In addition, it appears from the data that most stock prices do not exhibit a strictly negative or a strictly positive drift. Therefore, the purpose of selecting stocks of companies like Nike, Toyota, Facebook, Amazon, Apple etc. is to demonstrate that regardless of the background or domain of a company, regardless of specific fluctuations in the prices over time, the efficacy of our methods hold and these do not succumb to diminishing accuracies.

We now proceed to respond to other specific comments by the reviewers and how we have addressed them in the revised submission.

Summary:

This paper investigates the ability of two machine learning (ML) techniques (Random Forests and Gradient Boosted Trees) to forecast future stock prices. Specifically, forecasting is reformulated as a classification problem; essentially is the stock price expected to go up or down over a given time period. Model inputs are based on signals drawn from technical analysis, to which smoothing is applied by the authors. The ML techniques are presented in some detail with examples. The data set comprises 14 (mostly US) stocks with forecasts made over 3-90 days. Results are presented in terms of statistics such as accuracy, recall, etc. The ML models are held to outperform algorithms used in existing literature. Exploring how the wide range of machine learning techniques can be applied in a financial/economic domain, and how they can be used to complement traditional models, is an interesting area.

Main Comments:

Comment: The abstract should be rewritten to highlight novelty and the contributions of the paper.

Response: We have now re-written the abstract highlighting the novel points in a manner relevant to NAJEF.

Comment: While the discussion of the EM hypothesis is well placed, anomalies could also be discussed. See suggested references below.

Response: We have included the anomalies as obtained from the suggested references as part of the introduction (Pages 1 and 2). Thank you again.

Comment: For me, too much of the paper is devoted to outlining the ML techniques. These are well established and fully documented elsewhere. Readers unfamiliar with the techniques should be directed to primary sources, or authoritative texts. Provide a brief description, giving an intuitive sense of how the ML techniques work. Focus on the benefits such models promise in relation to traditional approaches and to the problem at hand.

Response: We have followed this suggestion carefully. The main sources where the techniques are explored and emphasized are cited in the main text along with a short introduction to the analytical and intuitive aspects of the models developed. While aspects of the algorithms and methods can be found through other sources, our intention amidst all the work was to break the usual treatment of ML as *black-boxes*. The example that we elaborately explained is based on real data samples from the data that we used. We broke down the methods by means of examples to maintain complete transparency of the work that we have done. Upon the reviewer's recommendation, we have moved some of the writing on the theoretical aspects of the algorithms, and the example, to a supplementary document (Sections 2, 3, 4 and 5 in the supplementary material). We have retained the material as a supplementary document as we think that it is highly relevant in the context of the paper.

Comment: While perhaps common in 'Information System' journals, my recommendation is to remove the outline algorithms and examples (e.g. Tables 2, 3, 4, etc). While one figure may be useful to aid understanding of a tree, the extended example and corresponding figures should be removed (or at least relegated to a separate appendix).

To appear in the finance/economics journal, the paper should be targeted to such an audience. Figure 2 is too basic to merit inclusion. Similarly figure 1 adds little value and could be easily summarised in a few lines.

Derivation of established formula/results (e.g. Chebyshev's Inequality) and basic definitions (e.g. an indicator variable) need not be included and certainly not proved. We have moved the material in these sections to a supplementary document (Sections 3-5 in the supplementary document).

Similarly, it is unnecessary to provide formula or examples for well-established technical indicators [incidentally, why include day 1 in the RSI formula; 15 prices are required for 14 daily returns]. Rather, focus on why these indicators are considered to have predictive ability. See also comments below.

Response: We have shortened the overall presentation of the theorems to aid better readability for economics/finance journals. The appropriate references are cited and we have deleted the discussion and proof of theorems from the main paper, as suggested. However, we have moved them to the supplementary reading material (Sections 2 and 3 in the supplementary file).

Comment: The paper does not adequately discuss feature selection or reference supporting literature to justify the choice of input attributes believed to have predictive ability. What do the models under investigation tell us about variable importance? Do certain technical indicators prove more useful for prediction? If these models really are outperforming existing benchmarks, from a finance perspective we want to understand why.

Response: The following is added to the main text to address this comment.

In recent years, the stock market analysis and prediction have been studied with the aid of methods such as machine learning and text mining. Data mining studies use daily stock data. For example, prediction studies based on support vector machines (SVMs) (Cao and Tay, 2001; Ince and Trafalis, 2007, Atsalakis and Valavanis, 2009, etc.) have been conducted to determine pattern categories. In addition, artificial neural networks (ANNs) (Kimoto et al., 1990; Kohara et al., 1997) have been employed to achieve good predictions even in the case of complex relationships of variables. Typically, autoregressive integrated moving average (ARIMA) model (Pai and Lin, 2005; Wang and Leu, 1996, Moscowitz et al. 2012) are used for identifying and predicting time series variation. Notwithstanding, since behavior and individualized responses play a significant role in dictating the stock turnovers and prices,, a few studies have engaged with word analysis of news articles (Mittermayer, 2004; Nikfarjam et al., 2010; Nyberg, 2011 for the US; Kim et al., 2014, etc.) and its predictive ability. However, most of these studies have some limitations for short-term prediction. First, without filtering for outliers, the predictions based on all historical data leads to potential errors. Second, although the total completion price is determined by a variety of factors such as the foreign purchase closing price and domestic selling completion amount, this set needs to be expanded in order to reduce omitted variables bias. Variables of importance may include, categories of financial ratios, macro, labour market and housing variables and measures of sentiment and leverage (Black et al. 2014; Cochrane, 2008, etc). With respect to the current paper, it is important to note that the main purpose is to implement two distinct methods on stock data and highlight their advantage over other non-ensemble techniques within the machine learning approaches for analyzing and predicting stock prices. This does not warrant conducting a regression analysis. Therefore, our engagement with feature extraction and assigning of importance to respective variables will follow available wisdom, except that the outcomes will be more efficient due to the choice of models.

References:

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. Expert Systems with Applications, 36(3), 5932-5941.

Black, AJ, Klinkowska, O, McMillan, DG and McMillan, FJ. (2014), 'Predicting stock returns: Do commodities prices help?', Journal of Forecasting, 33, 627-639.

Cao, L. and Tay, F. E. (2001). Financial forecasting using support vector machines. Neural Computing & Applications, 10(2):184–192.

Cochrane, J. (2008), 'The dog that did not bark: A defense of return predictability. Review of Financial Studies, 21, 1533-1575.

Ince, H. and Trafalis, T. B. (2007). Kernel principal component analysis and support vector machines for stock price prediction. IIE Transactions, 39(6):629–637.

Kim, Y., Jeong, S. R., and Ghani, I. (2014). Text opinion mining to analyze news for stock market prediction. Int. J. Advance. Soft Comput. Appl, 6(1).

Kimoto, T., Asakawa, K., Yoda, M., and Takeoka, M. (1990). Stock market prediction system with modular neural networks. In Neural Networks, 1990., 1990 IJCNN International Joint Conference on, pages 1–6. IEEE.

Kohara, K., Ishikawa, T., Fukuhara, Y., and Nakamura, Y. (1997). Stock price prediction using prior knowledge and neural networks. Intelligent systems in accounting, finance and management, 6(1):11–22.

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. Journal of financial economics, 104(2), 228-250.

Pai, P. F. and Lin, C. S. (2005). A hybrid arima and support vector machines model in stock price forecasting. Omega, 33(6):497–505.

Wang, J.-H. and Leu, J.-Y. (1996). Stock market trend prediction using arima-based neural networks. In Neural Networks, 1996., IEEE International Conference on, volume 4, pages 2160–2165. IEEE.

Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. In System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on,

pages 10–pp. IEEE.

Nikfarjam, A., Emadzadeh, E., and Muthaiyah, S. (2010). Text mining approaches for stock market prediction. In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, volume 4, pages 256–260. IEEE.

Nyberg, H. (2011). Forecasting the direction of the US stock market with dynamic binary probit models. International Journal of Forecasting, 27(2), 561-578.

Regarding Feature Importance: Typically, when the feature space has too many variables for confounding, dimensionality reduction becomes a necessity. The feature space needs to be broken into non-overlapping subspaces so that the learning algorithms are trained efficiently to discriminate between classes (feature dependent). However, in this particular class of problems, where the number of features aren't too many, such an exercise doesn't facilitate the computational efficiency of the machine learning algorithms we applied. Notwithstanding, in order to understand the contribution of features toward effective discrimination (may be an academic exercise), we have included this in Section 3.5. As expected, all the features/technical indicators are significant enough. However percentage contribution of OBV steadily increases with the increase in the size of the trading window (please see Table 1 of the revised manuscript), in comparison to the other technical indicators. This is observed to be consistent across all stocks considered in our experiments.

Comment: A complete description of the data should be presented before the results. You indicate that the start date is when the company went public. Does this mean different companies are examined for different periods? The reasons for selecting the 14 companies, or why they span different counties are not clear. Was this in part to aid comparison with the results of existing papers?

Response: We have provided a description of the data in Section 3.5 in the main paper. Further description of the data is not of additional help as the paper does not specifically analyze the data as part of the forecasting problem. Our methods facilitate a *greedy* approach in order to make the best case prediction of gains or losses. We applied our methods on various datasets, such that, we selected stocks of companies from various backgrounds including dot-coms, social media platforms, electronics, sports goods, and automobile. Despite the diversity of the stocks, the results have been good and encouraging. It suggests efficiency of our methods.

Similarly, you need to clarify:

(a) The time period(s) over which training and test sets were chosen (and why).

Response: From the day that the stocks went public till 3rd March 2017 (the date we did the final test runs). The reason for selecting the entire range of data was to be able to work with a lot of data and to see how accurate and resilient a direction-prediction method based on ML can be.

(b) If a range of different training/testing configurations were trialled.

Response: As we have already explained in Section 3, we have worked with different trading windows. These are intervals of 3, 5, 10, 30, 60, and 90 days. The results for these are presented through Tables 2, 3, A.4-A.7 in the manuscript, and include various 'configurations' as suggested. The longer trading windows offer more accurate prediction across models and configurations.

(c) The number of attributes used (preferably with summary statistics). For example, was PROC calculated over different periods? PCA is referenced earlier, was it used? Were all features smoothed in the same manner? Using what alpha value?

Response: PCA was not used. The features used are described in Section 3.2. If t is the trading window, then all the attributes which are calculated over different time intervals (like PROC) were calculated over a time period of t days. For each experiment run, thus, six features were used.

(d) The ML parameters (e.g. number of trees, number of variables used for each split).

Response: We thank the reviewer for this comment. We used 100 trees with each classifier. The more the number of trees, the better is the prediction, as proven by Chebyshev's inequality. In trees in random forests, one variable is used in each node for a split whereas as function is regressed on the variables in a node in a tree in gradient boosted trees. The Gini impurity criteria then decides which is the best split at a node. We have explained points (c) and (d) in Section 3.4 in the revised manuscript.

(e) The sources on randomness in a random forest (along with their importance) should be clarified. In random forests, a random subset of features is selected for each branching decision.

Response: This is definitely an interesting an relevant question. It is for the issue pointed at by this question that we created an elaborate illustrative example. As explained in Section 3.3 in the manuscript, the number of features used to build a tree are lesser than the total number of features in the data. If there are m features in the data, then typically, $\sqrt{m}$ features are used to grow each tree. Additionally, a subset of the number of samples can also be used to create a decision tree. These are the primary *sources* of randomness in random forests as opposed to plain decision tree classifiers where all features are used in a single tree.

Comment: Results should be presented against a suitable baseline drawn from existing literature. A model that always predicts an upward trend may prove instructive. Can you provide a statistical justification that these models are superior? Are the results consistent if repeated over different time periods?

Response:

- The statistical justification is provided using the Chebyshev's inequality. It is a proof of convergence of the algorithm.

- The paper discusses the merit of using random forests and XGBoost systems as compared to other techniques used under non-ensemble procedures. Indeed, it has been shown in several studies that trained SVM applied on Korean stock market (Kim, 2003), or DT plus ANN applied on Taiwan market (Tsai and Wong, 2009) reach 56% and 67% accuracy overall.

Compared to this the present techniques attain up to about 90% accuracy. In addition, the reason for choosing random forests over decision trees is because Random Forests uses a significant amount of voting-based conclusions as compared to that of decision trees. It runs a *bagging* based routine by using a large number of de-correlated Decision Trees (consider growing forests in random fashion) to classify a predicted class. This course of operation is highly suitable for the stock data and associated classification, as it meticulously examines the feature space to make better judgments over which class to finalize as the expected outcome. We have duly included this in the literature review section.

- Moreover, as we have clarified in Section 4.3 in the revised paper (Page 11), Figure 4 (in the revised manuscript) was a visualization of a comparative analysis done by us for understanding which classifier is superior. Hence, we have included the results of other classifiers: logistic regression, SVM and ANN, based on the same preprocessing steps as those used before implementing random forests and xgboost models.

Comment: Could other performance statistics be included? For example AUC or Brier scores (you mention ROC curves in the appendix but do not use them).

Response: Thank you for this suggestion, we have done the needful. The updated results are presented through Tables 2, 3, A.4-A.7. An exhaustive set of ROC plots are included in the supplementary file in Sections 5 and 6.

Comment: Section 6.3 felt out of place and left me confused. Why consider pharmaceutical stocks? How is your study of sign predictability related to "why certain stocks did not succumb to the aggravated economic crisis"?

Response: From a machine learning and methodology point of view, it is an interesting problem to address. The reasons we want to address stocks of pharmaceutical companies are two. First, it is for the purpose of diversification – we wanted to ensure that we include stocks of different types of companies and we have set no priors to that. However, the use of pharma stocks may be considered as a case study. And second, the low fluctuations in the stocks provides an interesting premise for automation methods for stock trading suggestions and strategies. The results confirm our choice of methods and the higher-than-usual accuracy in turn verifies the nature of the stocks of pharmaceutical companies.

The *sign predictability* is related to the stocks' variance over time. Naturally, the gains or losses of stocks which are stable would be easier to predict than stocks that are relatively noisy. A lower variance in the data implies a better predictive capability of ML classifiers, and from an economic point of view, it accounts for the stability. This is included in section 4.

As the writeup may feel a bit out of place, we have moved the entire section on the discussion of the ML methods applied to pharmaceutical stocks to a section in the supplementary file in Section 5. We have not removed it completely as we believe that it does provide some interesting insights between the lines of finance and machine learning.

Comment: Comparison of the accuracy achieved in this work to the accuracies achieved in available literature." On what basis has figure 15 been compiled? The caption is confusing. Were the

3 additional models run using identical data and features? If these were to be used as comparator models this should be made clear in the earlier methodology.

Response: We have clarified the confusion in the Results section (Section 4.3). The additional classifiers that we have mentioned are those of logistic regression, SVM and ANN. We have implemented these classifiers and have reported the best-case average accuracy across different stocks for a performance baseline. In comparison to this, we observe that the best-case performance by random forests and decision trees beat that of the remaining classifiers. Reiterating what we have already clarified in Section 4.3, the feature extraction and preprocessing were the same for logistic regression, SVM and ANN as they were for RF and XGBoost. Since SVM has been a popular classifier, we have presented a representative sample of ROC curves for this classifier in Section 7 of the supplementary file.

Comment: Over what dates is figure 16 plotted? Does the large stock price drop represent a stock split (and if so why haven't you accounted for this)? What constitutes a 'buy' signal or a 'sell' signal? Are there particular RF voting thresholds that must be breached to trigger a new signal? If so, why are there consecutive buys (and consecutive sells)?

Perhaps if this was clarified you could attempt to assess the economic impact of a trading strategy based on the ML models.

Response: This figure (now figure 7 in the revised submission) was plotted from the first day that the AAPL stock went public. As a representative sample, we have plotted the subsequent gains/losses with time intervals of 90 days. The idea of 'buy' and 'sell' signals is a simplistic notion for gains and losses, respectively. We have changed the terms regarding this as this is not an end-to-end trading strategy tool: we're calling them as buy or sell indicators, and the whole perspective as 'trading indications', as the information of predicted gains and losses might be useful for making trade decisions. Reiterating what we have mentioned earlier, we are not specifically trying to handle fluctuations in data over time as we are not posing this as a time-series analysis. The random forest classifier is able to handle such dips and rises in data and that speaks for the efficacy of our approach.

Minor Comments:

Comment: Rephrase "Market risk, strongly correlated with forecasting errors, needs to be minimized to ensure minimal risk in investment". I assume the point you want to make here is that minimising forecast error would minimise risk.

Response: Thank you, we have made the necessary changes in Section 1, Introduction.

Comment: "measuring the change in price of a stock compared to its price t days back" seems like a clumsy way of saying the "t-day return".

Response: Thank you, we have made the necessary changes in the last paragraph of Section 1, Introduction.

Comment: Wouldn't investors be more interested in establishing which assets might provide positive excess returns?

Response: Indeed that most work on the subject is on forecasting prices such that investors can have excess gains. But reiterating what we have stated in the Introduction (Section 1) of the revised manuscript, that is not the problem that we're addressing here. This is a deviation from the traditional perspective of predicting prices and is only about predicting gains or losses.

Comment: "Wisdom of Crown" should be "Wisdom of Crowds"?

Response: Thank you for this comment. We have rephrased this in paragraph 1 of Introduction (Section 1).

Comment: "some stocks tend to develop linear trends in the long-run". Do you mean that they exhibit momentum or directional trends?

Response: What we meant were directional trends but we have chosen to omit this sentence in the revised submission as long-term linear trends are not something that we're specifically addressing by our methods.

Comment: "Tree based learning methods are non-metric". Perhaps clarify in terms of discrete/continuous versus nominal/ordinal.

Response: Random forests are *non-metric* classifiers, which means that unlike gradient-based methods, there are no learning-parameters which need to be set, and unlike Bayesian methods, it does not require an assumption of a prior distribution. Additionally, this is one of the reasons that has made random forests a popular classifier for various tasks. Now, this does not relate to discrete/continuous valued attributes or nominal/ordinal, as in it's basic formulation, decision trees and random forests can handle both discrete and continuous data.

We thank the reviewer for this suggestion. We have made the necessary clarification in the last paragraph of **Random Forests** in Section 3.3, Page 6 in the revised submission.

Comment: Boosting a classifier means combining the results of many weak predictors to make a strong prediction". Isn't it more that this? A RF does that.

Response: Thank you for this comment. There's a lot of difference between bootstrap aggregation (*bagging*, which is what happens in random forests) and boosting. What happens in an RF is that a certain number of decision trees are randomly constructed on a subset of the feature-space. However, in boosting, every subsequent learner tries to better classify the misclassified samples of the previous learner. While RF takes into consideration the votes of randomly created tree, boosted tree algorithms move towards a better classification in every step. Additionally, the number of levels in a tree in an RF is usually not greatly constrained (unless we want to incorporate regularization by tree-pruning), whereas the number of levels in a tree in a boosted algorithm is usually very less, a ballpark figure of say between 1 and 5 – this is what we mean by a *weak* learner, that the individual learner's capability itself is only slightly better than an arbitrary guess, but together, whilst constructed carefully, an aggregate of similar weak learners can provide very accurate classifications.

Comment: Why not present all results as percentages?

Response: Generally, accuracy is represented as a percentage whereas the other measures of goodness are represented as fractions or decimals. Trying to keep at par with the general practice, we have not presented all the results as percentages. For instance, the Brier score is usually represented as a number between 0 and 1. And the same applies for sensitivity, specificity, AUC, etc. With all due respect, we do not see how a percentage representation or a fractional representation might affect the interpretability of the results.

Comment: Figure 14 has many series (mostly blue) but only 4 legend items.

Response: Dear Reviewer, we have mentioned in the caption that the blue items correspond to the change in accuracies of the stocks that we have presented in the main analysis (whose results are presented in Tables 2, 3, A.4-A.7). Now, the reason we did not present specific legends corresponding to each of these as the highlight of that figure to show that convergence was accomplished faster for the pharmaceutical stocks.

Comment: Fix Lane (1984) bibliography entry.

Response: Thank you for this comment. We have rectified the entry.

Comment: Review paper to ensure abbreviations are consistently introduced and used.

Response: Thank you. We have ensured this.

Comment: Take care to clarify/standardise use of terms such as ‘accuracy rate’, ‘hit rate’ and ‘success rate’.

Response: Thank you. We have ensured this.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–Part II: Soft computing methods. Expert Systems with Applications, 36(3), 5932-5941.

Christoffersen, P. F., & Diebold, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. Management Science, 52(8), 1273-1287.

Moskowitz, T. J., Ooi, Y. H., & Pedersen, L. H. (2012). Time series momentum. Journal of financial economics, 104(2), 228-250.

Nyberg, H. (2011). Forecasting the direction of the US stock market with dynamic binary probit models. International Journal of Forecasting, 27(2), 561-578.

Authors’ reply:

We thank the reviewer for specifically pointing out what needs to be changed. We have made corrections in every statement pointed out and have also revised other sections where we had used similar ambiguous phrases. We have added the suggested references.