Reply to the Reviewer's (Reviewer 2) Comments on *Predicting the Direction of Stock Market Price Using Tree-Based Classifiers*

June 16, 2018

Summary: In this paper the authors studied the prediction of stock market direction using two ensemble learning algorithms: random forest and XGboost. After preprocessing the data six technical indicators are calculated and the two algorithms are implemented using scikit learn and other software packages taking the technical indicators as input features. Discussion of results are made in the case of data taken from 10 specific companies and it is concluded that these two methods outperform logistic regression, neural network ,and SVM.

The following are some of my observations and suggestions:

Comment: The paper is too long. It can be shortened by omitting sections 4.1-4.4,,5.1-5.3 (which contain detail description of decision trees, random forest, XGboost, along with a worked out example, some derivations,etc.) as these are standard material available in data mining /machine learning text books. Section 4.5 and 4.6 about chebychev inequality and its proof can also be omitted. However , a very brief description of random forest and XGboost with the algorithms may be given in the appendix. Again the appendix A with key definitions also are to be omitted. Also the presentation in sections 1,2,and 3 may be shortened.

Response: First of all, the authors thank the reviewer for reading our work intensively and for the very useful comments. We have addressed each issue raised by the reviewer and the detailed responses follow.

We have shortened the paper and have moved the proofs. However, it is worth mentioning that the documentation on the ML techniques applied to the problem at hand is original and therefore, should be made accessible to the readers in some form. Consequently, the parts which contain mathematical notations and proofs have relegated to a supplementary file. In the supplementary part, we request retaining the proof of the Chebyshev's inequality applied to random forests since it provides the theoretical basis of why such classifiers work.

We have also merged the sections containing detail description of decision trees, random forest, XGBoost, along with a worked out example, some derivations, etc. with the supplementary file. Once again, for supporting future research in this area, print or online publication of these materials should be quite useful for the readers.

Comment: The authors claim that random forest and XGboost outperform SVM, ANN, logistic regression based on the values of performance metrics for RF and XGboost. But the corresponding values for SVM and other methods have not been computed for the data set used by the authors. To justify this claim a table containing the values of performance metrics(such as accuracy ,etc) for the three methods SVM, RF, XGboost for the same data set be prepared and the values be compared. This may further be illustrated using ROC curves for the three methods for the same data set.

Response: A sincere thanks to the reviewer for this suggestion.

As per suggestion, we have presented the results for a variety of companies from different backgrounds. Table 2 offers an example and data for other companies are available in Appendix A. The results are encouraging and indicate that these methods work well in a general sense. We have cited sources in the literature for the other methods and have presented more extensive results, in support of the comparative analysis. Fig.4 in the main text presents a comparative study to this effect and for our claim.

A clarification that we have made in Section 4.3 of the revised manuscript is that the comparisons that we have made with the other methods were done by us. Perhaps what was unclear from the previous version of the paper is that the comparisons were not done only with results available in literature. Such an ambiguity should not be prevalent here on.

Further, following the Reviewer's suggestions, we have presented a representative sample of ROC plots for SVM in Section 7 of the supplementary file, after repeating the experiments carefully with SVM. These plots are consistent with what we have presented in Figure 4 in the main test. Note that it is a small section and is not as exhaustive as our analysis using RF and XGBoost as the application of SVM is not the cornerstone of the paper. Notwithstanding, the representative sample of ROC curves should suffice to facilitate the readers' understanding of the efficacy of SVM applied to this problem.

Comment: Authors mention that return on the stock can be predicted using the model. But the model predicts only upward and downward movement. How can one determine the return on stock ? Similarly with out a quantitative predicted value of stock how can one make efficient portfolio management?

Response: We sincerely thank the reviewer for this comment. We understand that this sentence is ambiguous and taken literally, does not make sense in the purview of our work. We have changed this statement appropriately to remove the confusion.

Comment: Stock price prediction is a more general problem then stock direction prediction as the former gives more information than later. So in the title from forecasting to classification does not seem reasonable. This part may be omitted.

Response: We have changed the title.

Comment: The size of the data set should be mentioned in sec.7. Deep learning method may be explored only if available data size is big.

Response: We have mentioned in the revised manuscript that we have used the data (closing prices) from the date that the stocks went public, along with the range of sizes of the datasets, in Section 3.5.

Closing Remarks: We thank the reviewer again for specifically pointing out what needs to be changed. We have made corrections in every statement pointed out and have also revised other sections in line with these comments.