A Gradient Boosting model to predict the directional change in stock market returns

Hongjai Rhee

School of Business

Ajou University, Suwon, Korea

hrhee@ajou.ac.kr

Version: Feb 18, 2021

Abstract

This study examines the directional predictability of stock returns. To predict this directionality, we apply a gradient boosting machine learning model comprising a minimal set of covariates. We show that the parsimonious measures of past returns, one each for the individual stock of interest and for the overall market index, greatly enhance the predictive model fit compared to conventional logit models and a benchmark machine learning model. This finding supports the proposition of directional predictability advanced in previous literature. We demonstrate that the findings are universal for many KOSPI and NASDAQ test stocks.

Keywords: Asset returns, Directional predictability, Gradient boosting model, Markov model

Classification Codes: C14, C58, G17

1. Introduction

Are stock or asset returns predictable? The theoretical answer, based on the efficient market hypothesis (EMH), is a definitive no. Although many empirical studies have challenged this hypothesis, the countervailing evidence seems weak or problematic to some extent (Choi, Jacewitz, and Park 2016). To sidestep this problem, we turn our focus to the signs of daily returns, rather than their exact levels. Are the signs predictable? Christoffersen and Diebold (2006) explored this problem for U.S. equity returns and concluded that *"sign dependence is not likely to be found via analysis of sign autocorrelations, runs tests, or traditional market timing tests, because of the special nonlinear nature of sign dependence."*

The academic exploration of sign predictability is well summarized by Becker and Leschinski (2018). Moreover, Leung, Daouk, and Chen (2000) tested the efficacy of trading strategies driven by the probabilities estimated from various classification models, with respect to the signs of returns. Linton and Whang (2007) also found statistical evidence for directional predictability using a graphical device called the quantilogram.

Notably, most research on sign change has employed several macroeconomic covariates or risk measures to predict the signs of returns (Zhong and Enke 2019; Pönkä 2017; Wang 2014). By contrast, we consider only the information contained in the returns. If the EMH holds, the returns should reflect the effects of all publicly known variables. Specifically, we use the current sign and the discounted indices of both individual stock returns and market returns. The main research questions are: how and how much do these parsimonious measures of history help predict the future direction of individual stock returns? With similar conditioning information, Rhee (2021) showed that the return levels are not predictable at all, although explainable to a great extent. However, the effect of these measures on the predictability of return signs remains unclear. We explore this problem here.

Recent studies have applied machine learning algorithms to solve economic and financial problems. Gogas and Papadimitriou (2021), Varian (2014), and Athey and Imbens (2019) provide a great history, perspective, and possibilities in relation to the applicability of machine learning in such research. Nonetheless, machine learning remains underutilized in economic studies. This is probably because economic researchers are still unfamiliar with the machine learning framework, and do not want to sacrifice the advantage of intuitive model interpretation through parametric modes only to achieve a mediocre increase in model fit by employing complex machine learning models. In this study, using gradient boosting (GB) ensemble trees, we demonstrate that the fit actually increases substantially without incurring an insurmountable level of difficulty in model interpretation.

The remainder of this paper is organized as follows. Section 2 describes the data and preprocessing of the discounted index of returns. Section 3 briefly introduces the GB tree model. Section 4 summarizes the findings regarding the model's predictive performance and the marginal impacts of covariates. Finally, Section 5 presents the conclusions of this paper.

2. Data

Let r(t) be the rate of return on day t for a sample stock. We classify r(t) as a positive or negative sign indicator such that:

$$S(t) = 1 \text{ if } r(t) \geq 0$$

$$= 0 \text{ otherwise.}$$

For example, Figure 1 shows the daily rates of return of Samsung Electronics, which is the largest stock by market value on the Korea Stock Price Index (KOSPI), for the period January 2011 to December 2020.

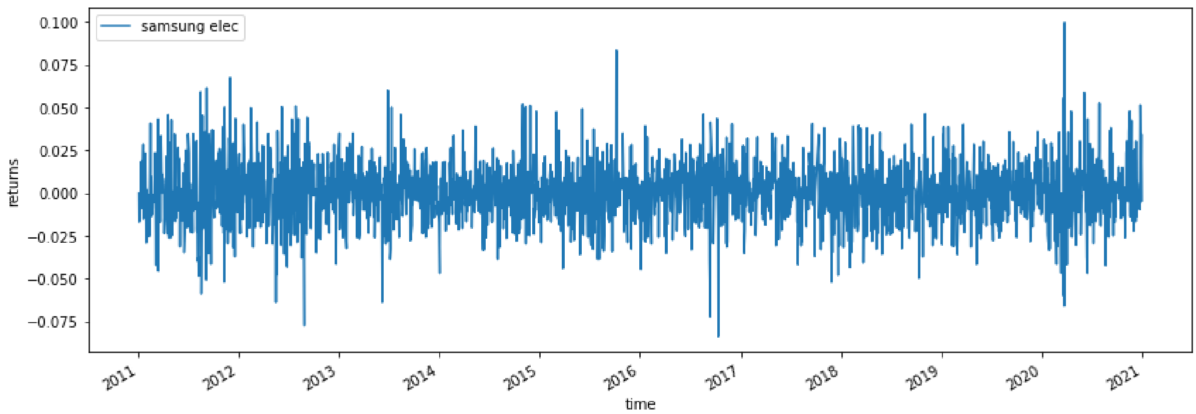

Fig. 1 Rates of return for Samsung on the KOSPI

Figure 2 plots the state (or class) variables for the last 100 days. S = 1 (i.e., nonnegative returns) comprises approximately 56% of the sample.

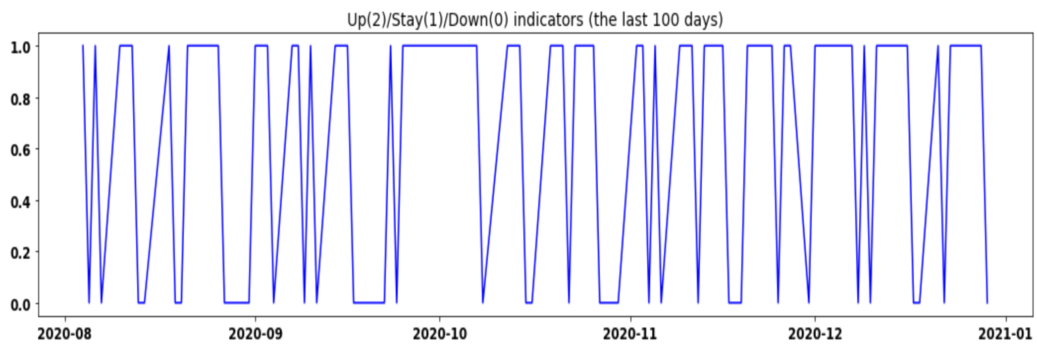

Fig. 2 Sign indicators of daily returns for Samsung on the KOSPI

Discrete indicators can be gleaned by a two-state Markov transition model. To show that the indicators are random, we first applied a run test for the series. The p-value for the null hypothesis of randomness is 0.142. Therefore, the sign series appears almost random at the 5% significance level. Table 1 summarizes the cross-tabulation of the daily transition of signs (or states).

S(t-1)	**down (0)**	**S(t) up (1)**	**All**
down (0)	528	585	1113
up (1)	585	727	1312
All	1113	1312	**2425**

Table 1 Day-to-day transition of return signs (Samsung)

Let us preview the prediction of states using GB trees that is explored in the next section. The decision tree classifier, which is the building block of the GB model, predicts the most frequent class in each circumstance. Therefore, the predicted class becomes 1 for all sample cases. Figure 3 shows a decision tree with no covariates other than the current signs. Note that the sample sizes are the same as in Table 1.

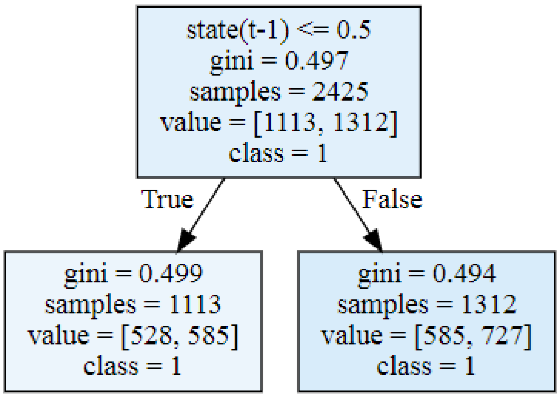

Fig. 3 Decision tree with no other covariates

Therefore, a simple Markov model cannot predict the signs of returns. We need other covariates that diversify the conditioning circumstances for the prediction.

We consider two types of information: what is the sign of the return now and how did it end up there? To obtain a parsimonious measure of the latter information, we construct the discounted rate of return for an individual stock as follows:

$$\mathrm{z}_t = \sum_{s=t}^{t-T} \delta^{t-s} r_s, \tag{1}$$

where $\delta \in [0,1]$ refers to the time discount rate to be incorporated in the model. We find that $\delta = 0.90$ maximizes the GB model's predictive fit. A higher z_t implies that the stock is now moving upward. Likewise, we construct a discounted stream of market returns (i.e., KOSPI) to represent the impact of environments.

$$\mathrm{Z}_t = \sum_{s=t}^{t-T} \delta^{t-s} R_s. \tag{2}$$

Then, the problem is reduced to make the following:

$$\mathrm{E}[s_{t+1} | s_t; z_t, \mathrm{Z}_t]. \tag{3}$$

Figure 4 displays shows the discounted stream for a recollection horizon T, set as 30 days.

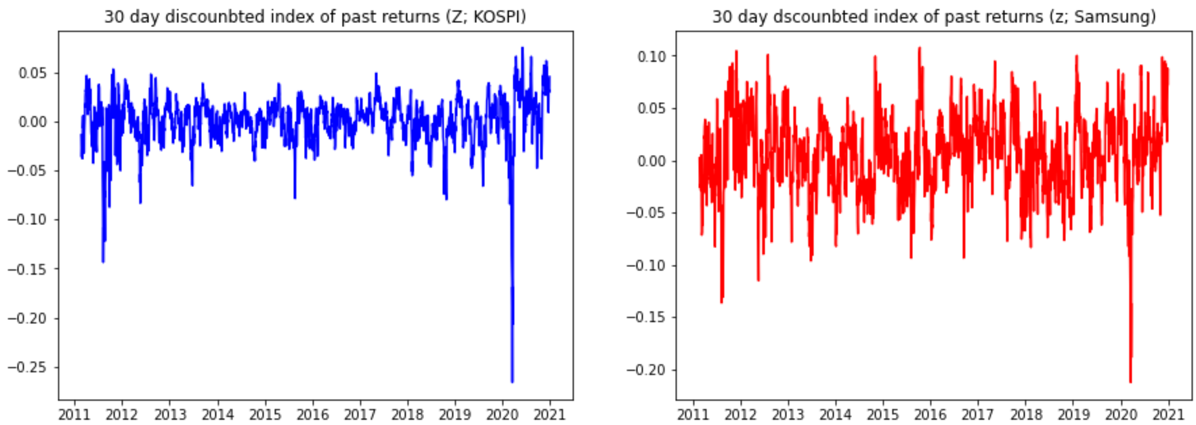

Fig. 4 Discounted index of returns.

We expect that these variables will affect the transition probability. As such, our model is similar to the covariance-dependent Markov models (e.g., Durland and McCurdy 1994) with a transition probability contingent on the measure of past returns.

3. Model

We now apply the GB regression model (Friedman 2001), which consists of bootstrapped data and a set of decision trees. The GB model is an upgraded version of the random forest (RF) model (Breiman 1996), and is another class of ensemble machine learning algorithms (Soybilgen and Yazgan 2020). To illustrate the model briefly, we consider the following mini dataset:

s(t+1)	s(t)	z(t)	Z(t)
0	0	0.05	0.02
1	0	-0.01	0.01
2	0	0.02	0.05
0	1	-0.05	-0.03
1	1	0.01	0.01
1	2	-0.01	0.02
2	1	0.015	0.03
2	2	0.01	-0.05
0	1	-0.02	-0.02
0	2	-0.03	-0.05

Table 2 Mini dataset for illustration

A simple decision tree model finds the best possible big tree that minimizes the objective loss function (usually, the cross-entropy function for a discrete choice model). Figure 5 shows the best tree. As the tree bifurcates, the covariate space is partitioned into subspaces. The prediction is calculated based on the most frequent class in the corresponding subspace.

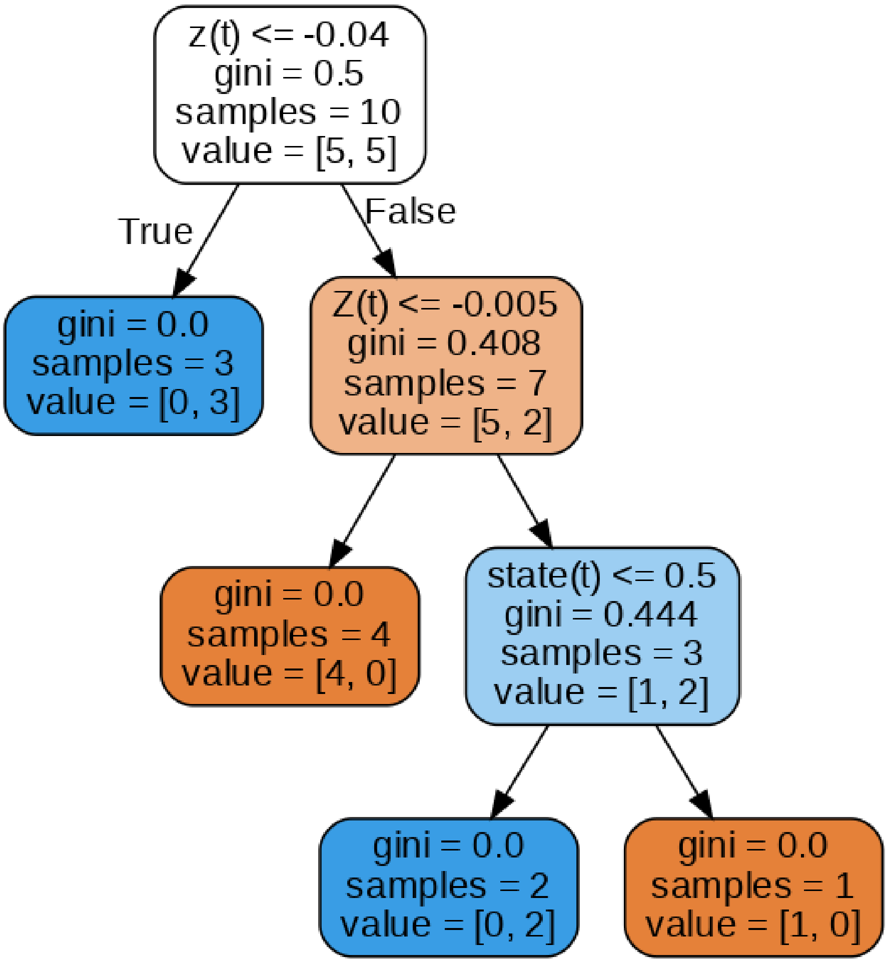

Fig. 5 The optimal decision tree for the mini data

By contrast, the GB and RF models use many weak trees, to avoid overfitting (Marquering and Verbeek 2005). Figure 6 depicts a weak tree. The difference between the GB and RF models is the method of constructing trees in the forest. While the RF model produces many randomized trees at the same time, the GB model sequentially optimizes the next tree (in feature and depth) to compensate for the weakness of the existing trees. The optimization is guided by an objective loss function's gradient information. It is well documented that a well-tuned GB model often outperforms the RF model and conventional parametric models (Rossi 2018). For details on implementing a GB model, see Natekin and Knoll (2013).

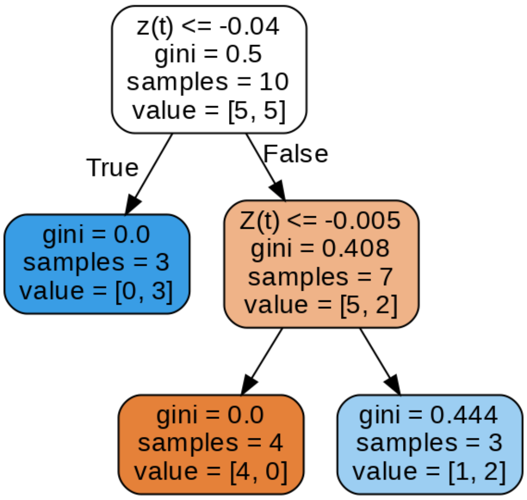

Fig. 6 A weak decision tree for the mini data (depth = 2)

4. Results

4.1. Hyper-parameters

Using a grid search over the key hyper-parameters, we find that the model with 220 trees and a learning rate of 0.015 roughly maximizes the test hit rates while maintaining the training hit rates within an acceptable range. Figures 7 and 8 show the contour plots of the hit rates for the test (the last 425 observations for Samsung), and training samples (the first 2000 observations for Samsung), respectively.

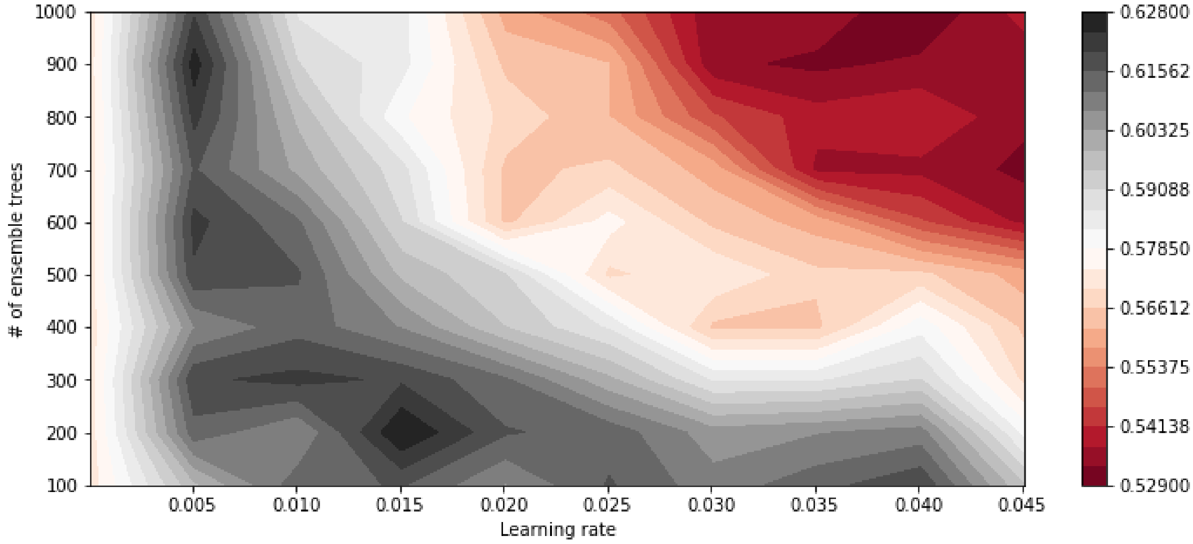

Fig. 7 Contour plot of the hit rates for the test sample

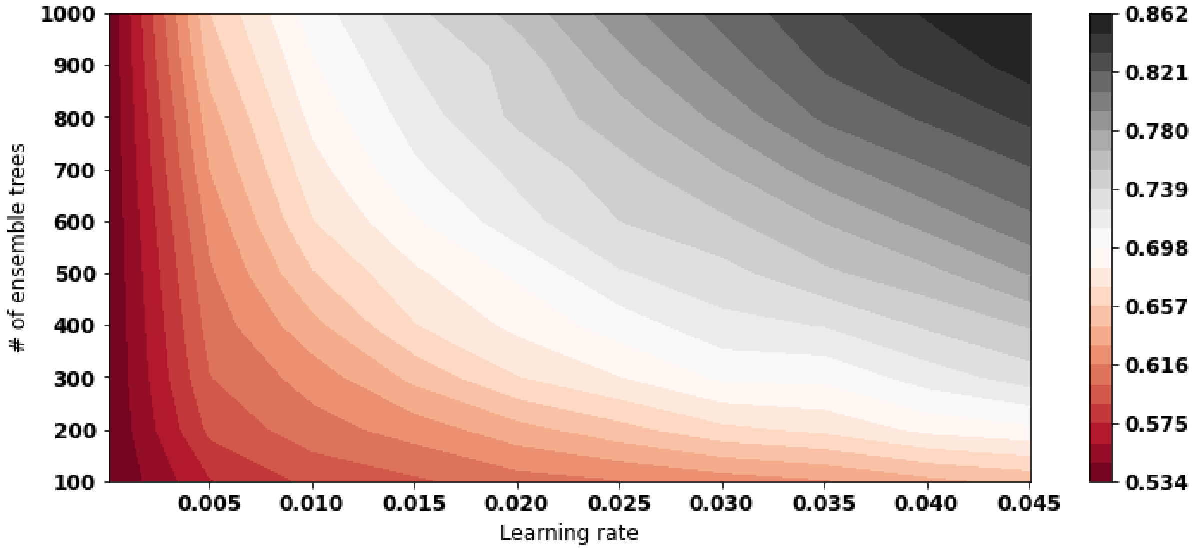

Fig. 8 Contour plot of the hit rates for the training sample

4.2. Feature importance

Figure 9 illustrates the feature importance. These statistics represent an increase in the cross-entropy loss function when each covariate (or feature) is discarded from the data or completely randomized. We find that the effects of the two discounted indices of past returns are mostly comparable, while those of the current signs are minor.

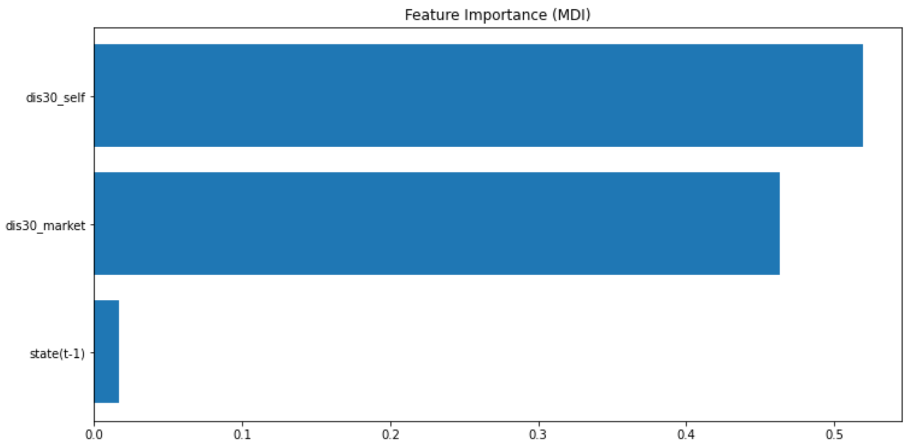

Fig. 9 Importance of covariates in the loss function

4.3. Model performance

To demonstrate the model fits, we calculate the in-sample and out-of-sample hit rates. The hit rate is 0.624 for the training sample (Table 3) and 0.626 for the test sample (Table 4).

	Predicted Label			Hit rate
True Label	0	1	total	
0 (Down)	379	552	931	40.7%
1 (Up)	200	869	1069	81.3%
total	579	1421	**2000**	**62.4%**

Table 3 Training sample hit rates by GB model

	Predicted Label			Hit rate
True Label	0	1	total	
0 (Down)	79	103	182	43.4%
1 (Up)	56	187	243	77.0%
total	135	290	**425**	**62.6%**

Table 4 Test sample hit rates by GB model

Note that the in-sample fit has been purposefully compromised to balance the in-sample and out-of-sample fits. Nevertheless, the GB model outperforms conventional logistic regression (in-sample = 0.572, out-of-sample = 0.574). Judging from the confusion matrix of the logit model in Table 5, both the qualitative and quantitative fit of the GB model are striking. This advantage is probably due to the nonlinearity of the tree-based prediction utilized in the GB model.

	Predicted Label			Hit rate
True Label	0	1	total	
0 (Down)	1	181	182	0.6%
1 (Up)	0	243	243	100.0%
total	1	424	**425**	**57.4%**

Table 5 Out-of-sample hit rates by logit model

As another benchmark, we compare our results with those of Becker and Leschinski (2018), who reported the predictive hit rates from various classification models for a long time series of U.S. stock market returns (summarized in Table 6).

	Logistic Regression	Generalized Additive Model	Neural Network	Support Vector Machine	Random Forest	Boosted Tree
Hit Rate	51.99%	51.35%	50.51%	51.02%	50.51%	50.92%
No. of covariates	7	10	7	12	13	6

Table 6 Summary of model comparisons in Becker and Leschinski (2018)

The covariates that Becker and Leschinski (2018) used in the boosted tree model in Table 6, which is very similar to our GB model, include self-stock returns, S&P 500 market returns, log realized variance, high-low variance, 12-day moving average of binary stock returns, and the rate-of-change indicator. The authors found that the logistic model fits the best (hit rate = 0.52) for hold-out samples. Although not directly comparable, our GB model produces a better fit using less conditioning information.

4.4. Covariates effects

Finally, we examine the effects of the three covariates in our model. Since the GB model is essentially a black box, it is not straightforward to evaluate the marginal effects as in a parametric model. However, it is still possible to graphically evaluate the marginal effects because of recent advances in data science. Figure 10 shows the partial dependence plot (PDP) for the discounted self-index of returns (z). The plot is obtained from the locus of the model predictions by only changing the covariate of interest for the given data. In Figure 10, the solid line shows the average prediction, and the shaded area shows the variation contingent on the values of the other covariates. Interestingly, the self-index seems to decrease the probability of having positive signs (1) on the next day’s self-returns, particularly when it is exceptionally high (i.e., the right tail).

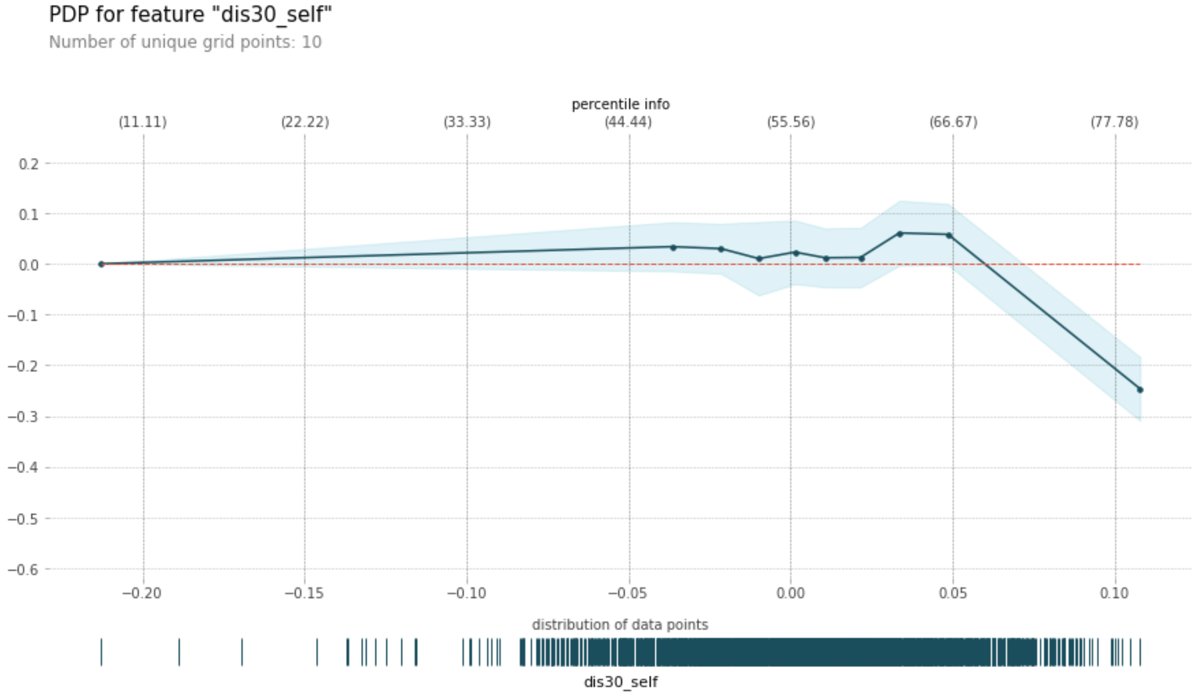

Fig. 10 Marginal effect of discounted self-index of returns (Samsung)

Figure 11 shows the PDP for the discounted market index of returns (Z), which is opposite to that for the self-index. The market index seems to substantially increase the probability of having positive signs (1) on the next day's self-returns, particularly when it is in an exceptional right tail of the distribution.

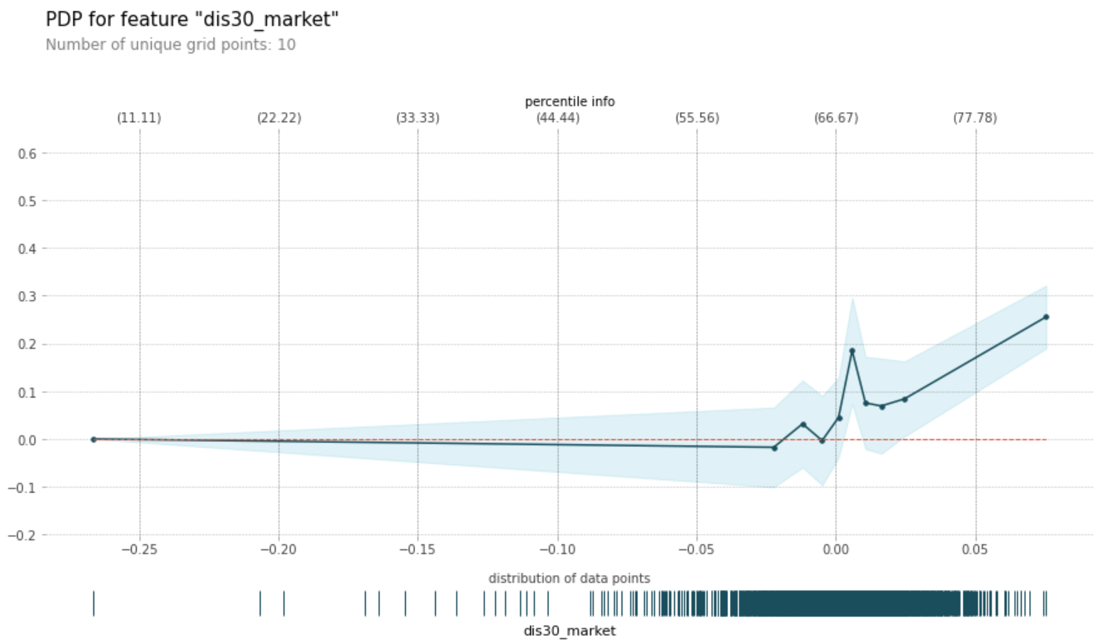

Fig. 11 Marginal effect of the discounted market index of returns (KOSPI)

Figure 12 shows a sort of inertia in return signs: the signs are passively correlated across successive periods when the other covariate effects are controlled for. That is, the probability of having positive signs (1) on the next day's self-returns was slightly higher when the current state was positive.

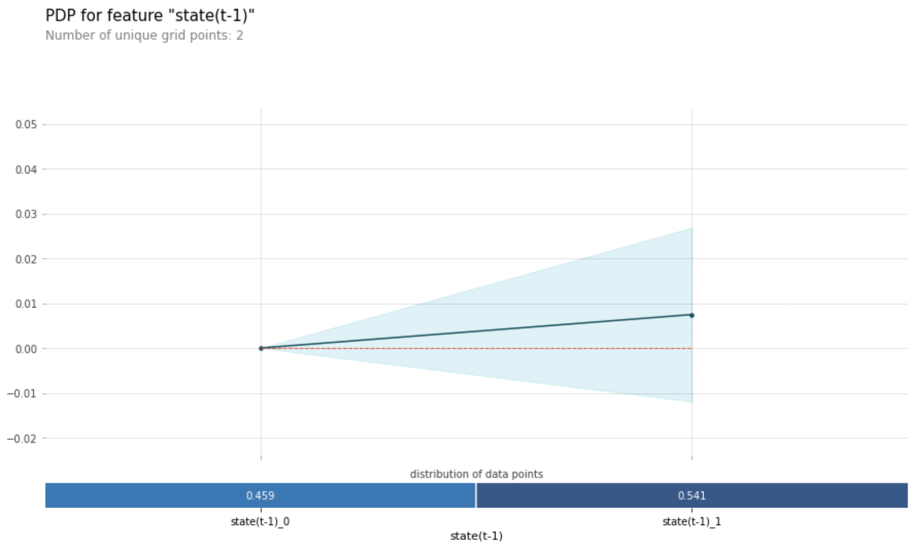

Fig. 12 Marginal effect of the current return signs (Samsung)

Figure 13 shows the interaction PDP for the current state and the discounted self-index. Notably, the right-tail effect of the discounted self-index, as in Figure 10, prevails regardless of the current state.

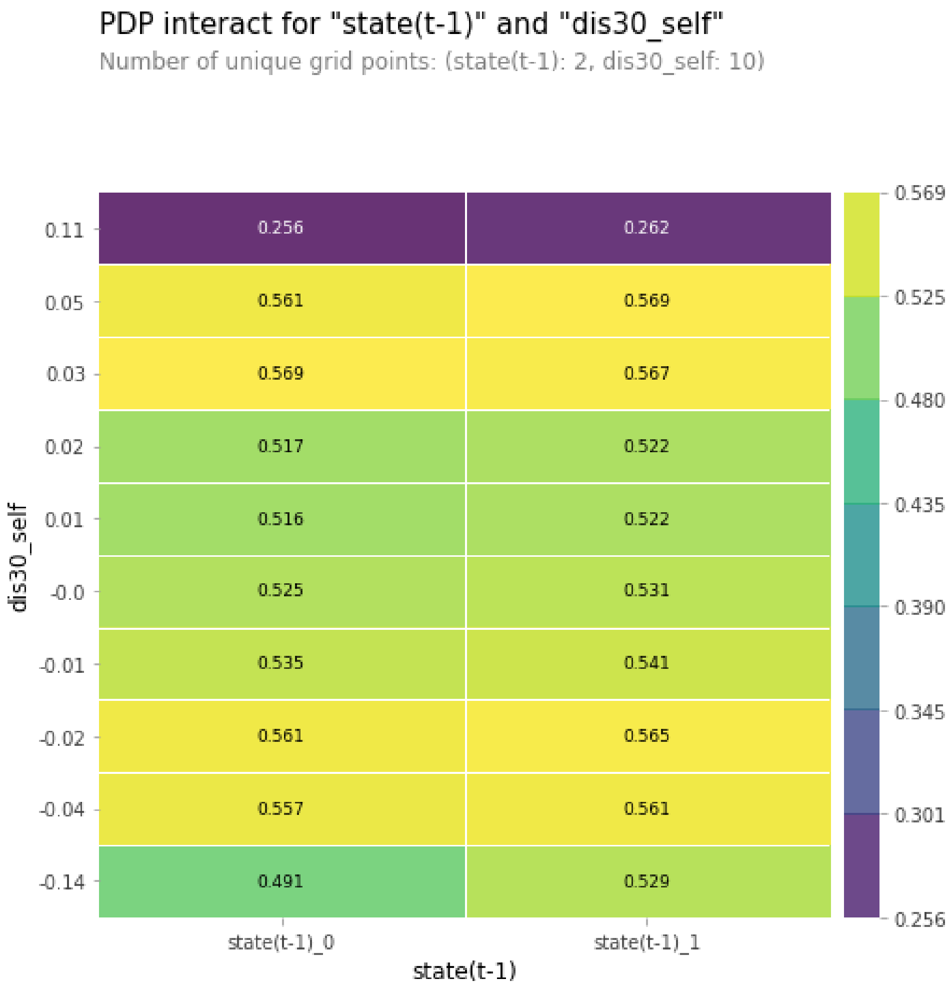

Fig. 13 Interaction of covariate effects

5. Conclusion

In this study, we demonstrated that the GB model can successfully forecast directional changes in stock market returns. The model provides superior prediction compared to a benchmark model, even though it employs only minimal information about the return signs at a certain date and how the stock and market returns have changed up to this point. This superior fit is attributable to the nonlinear, and even nonparametric, nature of the tree-based model.

Furthermore, we showed that it is easy to interpret the marginal effect of covariates using an enhanced visualization toolbox.

To show the model's usefulness beyond our sample stock (Samsung in KOSPI), we repeated the analysis for several other stocks on the KOSPI and NASDAQ. Table 7 summarizes the model fits and feature importance for the test stocks. The sample period is from January 2015 to December 2020. The out-of-sample hit rates range between 0.54 and 0.63. Furthermore, the relative importance of stocks are mostly similar across stocks.

Interestingly, the PDP shows different effects of the self- and market-index of returns, even though the rates of returns react only to exceptional events. Future research can explore this, and if possible, further improve the model's predictability by enclosing another set of covariates.

Market	**Stock**	**Prediction Accuracy**		**Feature Importance**		
		In-sample	Out-of-sample	Self_index	Market_index	State(t-1)
KOSPI	Samsung Elec.	0.624	0.630	0.51	0.47	0.02
	LG Chemical	0.614	0.574	0.49	0.48	0.02
	NAVER	0.654	0.554	0.53	0.43	0.04
	Samsung SDI	0.628	0.589	0.38	0.59	0.03
	Celltrion	0.643	0.584	0.49	0.50	0.01
	Hyundai Motors	0.755	0.567	0.44	0.55	0.01
NASDAQ	Apple	0.809	0.567	0.49	0.49	0.02
	Microsoft	0.608	0.589	0.43	0.55	0.02
	Amazon	0.795	0.548	0.47	0.52	0.01
	Tesla	0.632	0.537	0.49	0.47	0.04
	Google	0.770	0.542	0.47	0.49	0.04

Table 7 Summary of the results for other stock returns

References

Ang, Andrew and Bekaert, Geert (2006), “Stock Return Predictability: Is it There?” *The Review of Financial Studies,* Volume 20, Issue 3, 651–707.

Athey, S. and Imbens, G. W. (2019), “Machine Learning Methods Economists Should Know About,” *Annual Review of Economics*. 11, 685–725.

Becker, Janis & Leschinski, Christian (2018), "Directional Predictability of Daily Stock Returns," *Hannover Economic Papers*, dp-624, Leibniz Universität Hannover.

Breiman, L. (1996), “Bagging predictors,” Machine learning, 24(2), 123–140.

Choi, Y., S. Jacewitz, and J. Park (2016). “A reexamination of stock return predictability”. *Journal of Econometrics* 192(1), 168–189.

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

Durland, J., & McCurdy, T. (1994), “Duration-Dependent Transitions in a Markov Model of U.S. GNP Growth,” *Journal of Business & Economic Statistics*, 12(3), 279-288. doi:10.2307/1392084

Friedman, J. H. (2001), “Greedy function approximation: A gradient boosting machine,” *The Annals of Statistics*, 29(5), 1189–1232.

Gogas, P., Papadimitriou, T. (2021), “Machine Learning in Economics and Finance,” *Computational Economics*, https://doi.org/10.1007/s10614-021-10094-w

Harri Pönkä, (2017), "Predicting the direction of US stock markets using industry returns," *Empirical Economics*, vol. 52(4), 1451-1480.

Lanne, M. (2002), “Testing the predictability of stock returns”. *Review of Economics and Statistics*, 84, 407–415.

Leung, M., H. Daouk, and A. Chen (2000). “Forecasting stock indices: a comparison of classification and level estimation models”. *International Journal of Forecasting*, 16(2), 173–190.

Liao, Y. (2017), “Machine Learning in Macro-Economic Series Forecasting,” *International Journal of Economics and Finance*; 9(12).

Linton, O. and Whang, Y. (2007). “The quantilogram: With an application to evaluating directional predictability,” *Journal of Econometrics*. 141. 250-282. 10.1016/j.jeconom.2007.01.004.

Lussange, J., Lazarevich, I., Bourgeois-Gironde, S., et al. (2020). “Modelling stock markets by multi-agent reinforcement learning,” *Computational Economics,* https://doi.org/10.1007/s10614-020-10038-w.

Natekin, A. and Knoll, A. (2013), “Gradient boosting machines, a tutorial,” Front. Nerorobot,” https://doi.org/10.3389/fnbot.2013.00021

Qiu, M. and Y. Song (2016). “Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model”. *Plos one,* 11(5)

Rhee, Hongjai. (2021), “A discounted measure of the past stock market returns and its impact on the future returns: a machine learning approach,” Mimeo. https://drive.google.com/file/d/1bDEzFdfzy7J7uNoQ-H4g0J6PvyCqWZop/view?usp=sharing

Rossi, A.G. (2018), “Predicting Stock Market Returns with Machine Learning.”

Soybilgen, B., & Yazgan, E. (2020). “Nowcasting US GDP Using Tree-Based Ensemble Models and Dynamic Factors,” *Computational Economics*, https://doi.org/10.1007/s10614-020-10083-5.

Tashman, L. (2000), “Out-of-sample tests of forecasting accuracy: An analysis and review,” *International Journal of Forecasting*, 16(4), 437–450.

Varian, H. R. (2014), “Big data: New tricks for econometrics,” *Journal of Economic Perspectives*, 28(2), 3–28.

Wang, Yanshan. (2014), “Stock Price Direction Prediction by Directly Using Prices Data: An Empirical Study on the KOSPI and HIS,” Int. J. Bus. Intell. Data Min.. 9. 145-160. 10.1504/IJBIDM.2014.065091.

Yoon, J. (2020). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*. https://doi.org/10.1007/s10614-020-10054-w

Zhong, X., Enke, D. (2019), “Predicting the daily return direction of the stock market using hybrid machine learning algorithms,” *Financial Innovation*, 5, 24.