International Journal of Mechanical, Industrial and Aerospace Sciences
ISSN: 2517-9950
Vol:14, No:5, 2020

# Forecasting Stock Indexes Using Bayesian Additive Regression Tree

Darren Zou

***Abstract*—**Forecasting the stock market is a very challenging task. Various economic indicators such as GDP, exchange rates, interest rates, and unemployment have a substantial impact on the stock market. Time series models are the traditional methods used to predict stock market changes. In this paper, a machine learning method, Bayesian Additive Regression Tree (BART) is used in predicting stock market indexes based on multiple economic indicators. BART can be used to model heterogeneous treatment effects, and thereby works well when models are misspecified. It also has the capability to handle non-linear main effects and multi-way interactions without much input from financial analysts. In this research, BART is proposed to provide a reliable prediction on day-to-day stock market activities. By comparing the analysis results from BART and with time series method, BART can perform well and has better prediction capability than the traditional methods.

***Keywords*—**Bayesian, Forecast, Stock, BART.

## I. Introduction

THERE are a number of primary economic indicators that can be used to predict the stock market changes. Gross Domestic Product (GDP) is an important indicator, which is defined as the total amount of goods and services produced by a country. As a major economic indicator, it has a significant impact on the stock market. When the economy is growing, it is expected that GDP will increase due to business expansion. A lower GDP number will have a negative effect on the stock market. Unemployment rate is another important indicator that illustrates the strength of the economy. In the U.S., this figure is reported monthly by the Bureau of Labor Statistics. Other important indicators include, but are not limited to the Consumer Price Index for All Urban Consumers, Advance Retail Sales: Retail (Excluding Food Services), Consumer Sentiment (University of Michigan), Effective Federal Funds Rate and 10-Year Breakeven Inflation Rate, etc.

The goal of this research is to provide a reliable statistical method using BART to predict stock indexes, such as Dow Jones Averages, based on multiple economic indicators.

## II. Background of BART Method

Chipman et al. proposed a Bayesian "sum-of-trees" model with Bayesian backfitting MCMC algorithm implemented [1]. One major advantage of BART is the capability to handle non-linear main effects and multiway interactions without much input from researchers [2]. There are two versions of BART proposed by Chipman et al. [1]: BART-cv and BART-default. In BART-cv, the prior hyperparameters were tuned via cross-validation as operational parameters. Different from BART-cv, those parameters were set as default values in BART-default. Chipman implemented BART in his R package BayesTree written by C++. The performance of BART-cv and BART-default were compared to some other methods, including Lasso, Boosting, Neural net and Random Forest. BART-cv has the smallest relative root-mean-square error (RMSE) values than the other methods per Chipman's analyses. The other BART method, BART-default, has the second best performance. Since most other methods, such as neural nets, random forests and gradient boosting rely on cross-validation, it is very impressive that BART-default, which does not need cross-validation, can out-perform those methods.

Kapelner et al. developed another R package bartMachine for similar analysis on BART [3]. Different from BayesTree which was written by C++, bartMachine was written by Java, with some additional features, such as external prediction function, model persistence across sessions, parallelization, native missing data mechanism, built-in cross-validation, and model diagnosis, etc.

Geirsson explored a parallel implementation of BART using the Apache Spark framework [4]. The most significant modification is the serial improvement of the code, which reduces the workload considerably as the scanning of data is minimized.

Lakshminarayanan et al. proposed a new PG sampler for BART [5]. Unlike existing samplers which make local moves, the PG sampler can propose complete trees. Experimental results confirm that PG dramatically increases mixing when the true posterior consists of deep trees or when the data dimensionality is high.

## III. BART Algorithm

The BART model consists of two parts: a sum-of-trees model and prior distributions on the parameters of that model.

### *A. Priors and Likelihood*

There are three components of prior distribution in the BART model:

(1) The tree structure, denoted as Tj, where j is the number of trees in the model.
(2) The leaf parameters in a tree, denoted as Mj.
(3) The variation parameter $\sigma^2$.

$$p((T1;M1); \dots ; (Tm;Mm); \sigma) = [\prod_j \; p(Tj, Mj) \quad p(\sigma) \quad (1)$$

$$=[\prod_j \; p(Tj \vee Mj) \quad p(Tj)\, p(\sigma) \quad (2)$$

Darren Zou is with the Montclair Kimberly Academy, United States (e-mail: dzou2021@mka.org).

and,

$$p(Mj|Tj)\prod_j p(\mu ij \vee Tj) \quad (3)$$

where, $\mu ij \in Mj$

The tree components (Tj ;Mj) and σ are independent to each other. The terminal node parameters μij in all trees are also independent.

The prior of $\mu_{ij}|T_j \sim N(\mu_\mu; \sigma_\mu^2)$ and the prior of $\sigma^2 \sim IG(\gamma/2, \gamma\lambda/2)$, where IG( γ/2, γλ/2) is the inverse gamma distribution with shape parameter of γ/2 and the scale parameter of γλ/2.

The prior distribution of tree structure of $P(T_j)$ is composed of three different components.

1. The probability of a node will split at depth $d = \alpha / (1+d)^\beta$, where d is an integer greater or equal to 0. At the root of the tree, d is equal to 0. α controls the probability of a node would split, which follows a uniform distribution from 0 to 1. The larger α, the more likely for a node to be split. β controls the number of terminal nodes in a tree. When the value of β increases, the probability to split goes down.
2. Since there is usually more than one covariate in the model, only one covariate will be checked for the possibility to be split on an internal node of a tree. The probability to pick one specific covariate follows the uniform distribution by default.
3. Each covariate has a list of values in the dataset. Once a covariate is selected for an internal node, which value will be used as the cutoff point in an internal node to split will also follow the uniform distribution.

In addition, the number of trees m also needs to be determined. Chipman et al. suggested the default of m to be 200 [1]. According to their experience, the predictive performance of BART improves dramatically when m is increased from 1. At some point, the predictive performance levels off and starts to go down gradually when m continues to be increased. They suggest not to choose a very small m. In the proposed method, the default of m=200 is used.

### *B. Hyperparameters*

There are six hyperparameters, α, β, $\mu_\mu$, $\sigma_\mu$, γ, and λ, need to be set in BART for continuous endpoints. α and β are usually set to their default values of 0.95 and 2, respectively, which provide balanced penalizing effect for the probability of node splitting [1]. γ is also set to its default value of 3. λ is set to the value such that $P(\sigma^2 < s^2; \gamma; \lambda)=0.9$.

In default, the last two hyperparameters (i.e. $\mu_\mu$, $\sigma_\mu$) are set to appropriate values so that the prior distribution of BART will be a non-informative prior. In this case, $E[Y|X] \sim N(m\,\mu_\mu; m\,\sigma_\mu)$ assigns high probability to the interval (min(Y), max(Y)), where m is the number of trees. In order to calculate the posterior distribution easily, Y is transformed to become Y_tilda = (Y-(max(Y)+min(Y))/2)/(max(Y)-min(Y)). When min(Y)=-0.5 and max(Y)=0.5, Y_tilda will fall in the range of (-0.5, 0.5). As a result, $\mu_\mu$ will be set as 0 and the value of $\sigma_\mu$ will be calculated from $\sigma_\mu=0.5/(\gamma*\sqrt{m})$. The value of γ will be set to its default values of 3, as suggested by Vincent et al. [2]. In order to set the value of λ, a multiple linear regression by using the outcome as the dependent variable and all selected predictors as covariates will be run first to obtain the estimated variance of residuals. The selected value of λ will make the probability of $\sigma^2 < s^2$, i.e. $P(\sigma^2< s^2, \gamma, \lambda)$, equal to 0.9.

### *C. Posterior Distribution and Prediction*

After all hyperparameters are set and all prior distributions and m are determined, the posterior distribution will be generated by a Gibbs sampler with Metropolis-Hasting method embedded using Bayesian backfitting MCMC algorithm [6].

The residual response of each tree is defined as:

$$R_{-j} = Y - \sum_{t \neq j} \binom{n}{k} Tt(X) \quad (4)$$

As the first step, the posterior distribution of T1 will be obtained from $R_{-1}$ and $\sigma^2$, i.e.: $T_1|R_{-1}, \sigma^2$.

The proposed tree structure may or may not be accepted via a Metropolis-Hastings step. The proposed tree structure can either grow or prone or change or swap. Chipman et al. [1] propose the probabilities as 0.25 for growing a terminal node, 0.25 for pruning a pair of terminal nodes, 0.40 for changing a nonterminal rule, and 0.10 for swapping a rule between parent and child. The sampling from the posterior distribution of tree structure $T_1$ does not depend on the leaf parameter $M_1$ because they can be integrated out.

As the next step, the posterior distribution of $M_1$ will be generated using the posterior formula derived in appendix: $M1|T1, R_{-1}, \sigma^2$. Similarly, the posterior distributions of other tree structures Tj and leaf parameters Mj are generated sequentially in the following order for up to m trees.

$T_2|R_{-2}; \sigma^2$
$M_2|T_2;R_{-2}; \sigma^2$
$T_3|R_{-3}; \sigma^2$
$M_3|T_3;R_{-3}; \sigma^2$
……
$T_m|R_{-m}; \sigma^2$
$M_m|T_m;R_{-m}; \sigma^2$
$\sigma^2 | T_1, M_1, \ldots, T_m, M_m, \epsilon$

All above steps represent a single Gibbs iteration. According to Kapelner et al. [3] the default setting in R package bartMachine is 250, and no more than 1,000 iterations are needed as "burn-in". In addition, the results will converge after 200 iterations according to trace plots. As a result, the first 1000 iterations are discarded in the proposed method.

After the model based on proposed method is converged, the full set of trees will be obtained. The terminal node μ's can be summed up to get the predicted values of Y. A large number of random draws from the posterior distribution will provide estimation for the response.

## IV. Proposed BART Method

Dow Jones Averages and economic data, including Real

GDP, Unemployment Rate, Consumer Price Index for All Urban Consumers (CPI), Advance Retail Sales: Retail (Excluding Food Services), Consumer Sentiment (University of Michigan), Effective Federal Funds Rate and 10-Year Breakeven Inflation Rate, are downloaded from [7]. Since not all data are available before the year 2003, only data available on or after January 1, 2003 are used in the analysis. Not all economic indicators are measured on a daily basis. For example, CPI and the unemployment rate, etc., are measured monthly and GDP is only measured quarterly. When any economic indicators are missing on a specific day, the value reported on an earlier day would be carried over. If any stock indicator is still missing on a day, that day would be excluded from the analysis. The days of the weekend are also excluded from the analysis since the stock market was closed. As a result, 4186 days between January 1, 2003 and January 10, 2020 are used in the analysis.

Half of the available days are used as the training set to build the BART model. For the remaining days, Dow Jones Averages will be predicted using economic indicators based on the model built on training set. The predicted stock index will be compared to the true stock index to determine the accuracy of prediction.

BART with 200 trees and seven covariates listed above are used in the analysis. Metropolis-Hastings method is used to run 5000 iterations with the first 1000 burn-in sets excluded to obtain the posterior distribution.

Based on the posterior distribution, 200 random draws are taken. Those 200 random draws are combined using Rubin's rule to get the final prediction.

The time series analysis with the same covariates is also performed to predict the stock index so that the accuracy of the proposed BART method and time series method can be compared.

## V. Results

In the test set, which has 2186 days, the difference between the true Dow Jones Stock Index and the prediction from the time series technique is calculated for each day. The difference is summarized in Table I. The mean difference is 56.115 points with the standard deviation of 1030.650. Similarly, the difference between the true Dow Jones Stock Index and BART prediction is calculated for each day. The difference is summarized in Table II. The mean difference is as small as -0.284 points and the standard deviation is only 242.163. Comparing time series analysis, BART can provide a much more accurate prediction on stock changes.

TABLE I
Summary of Differences between True Dow Jones Stock Index and Prediction from Time Series Technique

| Number of Days | Mean | Median | Std | Min | Max |
|---|---|---|---|---|---|
| 2186 | 56.115 | 52.3633 | 1030.650 | -3499.62 | 3695.73 |

TABLE II
Summary of Differences between True Dow Jones Stock Index and BART Prediction

| Number of Days | Mean | Median | Std | Min | Max |
|---|---|---|---|---|---|
| 2186 | -0.284 | -9.35221 | 242.163 | -1385.78 | 1456.92 |

The true stock index and BART prediction, along with 95% CI, are plotted in Fig. 1. According to the figure, the true stock indexes are always within 95% CI even when confidence interval is very narrow.

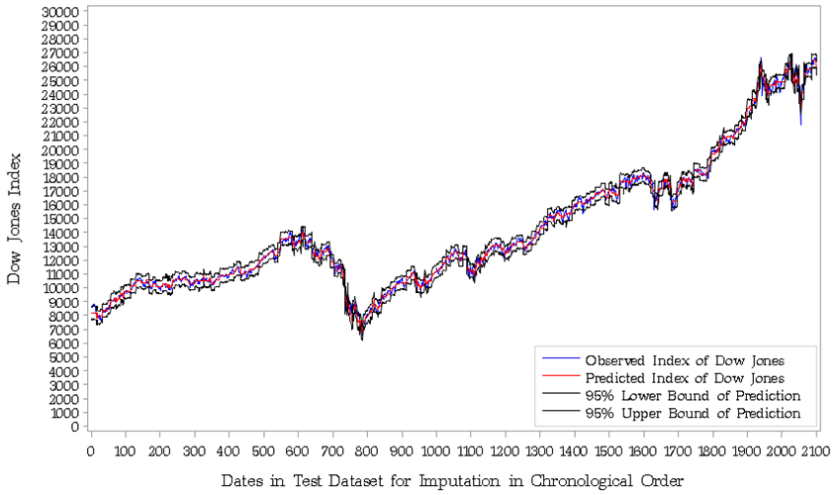


Fig. 1 Forecasting Performance in Test Set Using BART

## VI. Conclusion

BART is a powerful method to analyze financial data. One of advantages of the proposed BART method is that it can predict stock changes accurately based on economic indicators. The obtained analysis results proved that the BART algorithm is more accurate than the traditional time series technique in stock prediction.

BART can handle many covariates at the same time. However, only seven covariates are used in the above analysis due to the availability of data. As a future research interest, the proposed BART method with more available data included

can be applied.

## REFERENCES

[1] Chipman et al., (2010), BART: Bayesian additive regression trees. Annuals of Applied Statistics. Volume 4, Number 1 (2010), 266-298.
[2] Tan et al., (2019), Bayesian additive regression trees and the General BART model. Statistics in Medicine. Volume38, Issue25, 10 November 2019, Pages 5048-5069
[3] Kapelner et al., (2016), bartMachine: Machine Learning with Bayesian Additive Regression Trees. Journal of Statistical Software. 10.18637/jss.v070.i04
[4] Geirsson et al., (2017), Parallel Bayesian Additive Regression Trees, using Apache Spark. Computer Science.
[5] Lakshminarayanan et al., (2016), Particle Gibbs for Bayesian Additive Regression Trees. Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: WCP
[6] Hastie et al., (2000) Bayesian Backfitting, Statistical Science, Volume 15, Number 3, 196-223.
[7] Fred Economic Data, "Economic Research", https://fred.stlouisfed.org, accessed on 17 Apr 2020. volume 38