Review of Bagging Performance

Disa Alda Naomi **Prarthna Khemka** **Samita Sridhar Kamath**

Abstract

We investigate Breiman's conclusions in his 1994 'Bagging predictors' paper with the objective of replicating his analyses to verify his results. Our findings agree with those in the original paper that bagging gives substantial gains in accuracy for regression and classification trees and forward variable selection. In fact, we observe that bagging outperforms forward variable selection to a far greater extent than in Breiman's paper. Additionally, we implement extensions to the analyses in the original paper including studying the effect of the number of bootstrap replications on the performance of the bagging methods.

1 Motivation

Leo Breiman proposed bootstrap aggregating (coined "Bagging") as a method to improve the stability of predictions and accuracy of a machine learning algorithm. Bagging is an ensemble method that aggregates the predictions of weak learners either by averaging the output for a regression problem, or through a plurality vote for a classification problem. The weak learners are generated by making bootstrap replicates of the "learning" or training set and fitting models to these bootstrap samples. Through this method, Breiman tries to solve the common problem of data scarcity, since we may not have the luxury of having actual replicates of the learning set. Bagging has also been shown to act as a variance reduction technique that reduces overfitting.

The motivation behind this algorithm was to use bootstrap replicates to improve overall model performance. Having studied and implemented bagging with various machine learning models, we were curious to study the seminal research that first introduced it. In particular, we hope to replicate Breiman's methods and extend his original analysis on the effect of varying number of bootstrap replicates. Breiman notes that bagging optimizes unstable procedures such as neural nets, classification and regression trees, and subset selection in linear regression. In contrast, improvements in performance may not be as apparent in stable procedures like k-nearest neighbors. We also investigate these claims in the following sections.

2 Bagging Classification Trees

Breiman displays how bagging impacts classification tasks through his experiments on Classification Trees. Classification trees are decision trees used to predict categorical outputs.

2.1 Methodology

Bagging was applied to classification trees using five datasets as outlined in the next section.

2.1.1 Datasets

Some of the datasets below are available in the UCI repository [5] cited in the paper. However, a few of the datasets have been modified over the years, and when we were not able to find datasets identical to the original ones used in the paper, we sourced these datasets from other online sources. In some of the datasets, missing values denoted by '?' are replaced by zeroes to allow for fitting the decision trees.

- *Waveform*: The dataset contains simulated 21 predictors and 3 classes each having probability roughly of a third, with 5000 observations. In the paper, Breiman used the Waveform dataset that has the same predictors but only 300 observations, and then generated a balanced dataset with 4500 observations himself. We note that this slight discrepancy in the dataset may influence our results.
- *Heart*: We were unable to find the heart disease dataset from University of California, San Diego Medical Center, as the heart disease dataset that is currently online in UCI repository are sourced from hopsitals in Budapest, Zurich, Basel, Long Beach, and Cleveland. We attempted to replicate the bagging analyses for this dataset, however our results showed significantly different rates than Breiman's analyses with the original dataset. Thus we will omit the results of this analysis in this paper.
- *Breast Cancer*: This dataset contains two class data classifying breast cancer to benign and malignant cases, with 699 cases and 9 variables consisting of cellular characteristics. The dataset is provided by University of Wisconsin Hospitals, Madison.
- *Ionosphere*: This dataset is radar data collected by the Space Physics Group at Johns Hopkins University. There are 351 cases with 34 variables, consisting of 2 attributes for each at 17 pulse numbers. There are two classes (bad and good).
- *Diabetes*: This dataset is collected by the Pima Indians by the National Institute of Diabetes and Digestive and Kidney Diseases. We do not find the appropriate dataset from the UCI repository, however we found the dataset posted on Kaggle (3). The dataset contains 768 observations, with 8 predictors and two classes. As per the guidance in the paper, to generate a balanced dataset, we duplicated the diabetes cases from 286 cases to 536 cases, giving a dataset of 1036 observations.
- *Glass*: This dataset was created by the Central Research Establishment, Home Office Forensic Science Service Aldermaston, Reading, Berkshire. Each case vonsists of 9 chemical measurements on one of six types of glass. There are 214 cases.
- *Soybean*: The dataset contains 307 observations, 35 variables, and 19 classes. The classes are types of soybean diseases, and the predictors consist of the plan characteristics and climactic variables.

2.1.2 Computations

To assess the performance of bagging for each of the datasets above, we compare the misclassification rates of a single classifier tree and bagged classifiers on randomly sampled test sets using 50 bootstrap replicates, averaged across 100 repetitions for each dataset. We follow the procedure as outlined in the paper:

(i) The dataset was randomly divided into a test set T and learning set L. The test and learning set sizes are chosen according to Table 2: Data Set Summary in Breiman's paper.

(ii) We construct the single classification tree from L with the best-performing hyperparameters chosen from 10-fold CV. As Breiman did not specify what hyperparameters were considered in this process, we tuned the $criterion$, $splitter$, and max_depth hyperparameters used in Python's $sklearn.tree$ models. These three hyperparemeters are among the most commonly used to tweak decision trees. Running the test set T down this tree gives the misclassification rate $e_S(L, T)$ computed with zero-one loss formula.

(iii) The random division of the data is repeated 100 times and the reported $\bar{e_S}$ is the average of $e_S(L, T)$ over the 100 iterations.

(iv) A bootstrap sample $L^{(B)}$ is selected from L, and a tree grown using $L^{(B)}$ and 10-fold cross-validation, with similar hyperparameters as in the previous step. This is repeated 50 times giving 50 tree classifiers $\varphi_1(x), ..., \varphi_{50}(x)$.

(v) If an observation belongs to the set set, i.e. $(j_n, x_n) \in T$ then the estimated class of the observation x_n is the class that is the majority class out of predictions from the models $\varphi_1(x), ..., \varphi_{50}(x)$. The bagging misclassification rate $e_B(L, T)$ is computed as the proportion of times the estimated class of the test set differs from the true class.

(vi) The random division of the data is repeated 100 times and the reported $\bar{e_B}$ is the average of $e_B(L, T)$ over the 100 iterations.

2.2 Results

The results from our computations agree with Breiman's results, however, due to the discrepancies in the datasets outlined above, we observe slightly different results for two of the datasets (Waveform and diabetes). We also note that some small discrepancies in the percentages of misclassification rates can lead to moderate changes in percent decrease, as observed with the results with the glass and ionosphere datasets. The reduction in test set misclassification rates in our analyses range from 15% to 36%.

Table 1: Misclassification Rates (Percent)

Dataset	$\bar{e}_S$	$\bar{e}_B$	Decrease	Breiman's Results
waveform	24.0	16.22	32%	33%
breast cancer	6.0	3.8	36%	30%
ionosphere	11.5	7.4	35%	23%
diabetes	28.5	24.1	15%	20%
glass	33.9	22.4	34%	22%
soybean	12.1	9.1	25%	27%

2.3 Extensions

In the above analyses, 50 bootstrap replicates was used, however Breiman noted that this particular choice of number of bootstrap replicates is neither necessary nor sufficient, yet it is an arbitrary choice as the number "seemed reasonable." Breiman claimed that even with only 10 bootstrap replicates, we are already getting most of the improvements. He concluded that "more than 25 bootstrap replicates is love's labor lost."

We decided to verify his arguments by comparing the misclassification rates by the number of bootstrap replicates to the datasets used in this section. Our findings agree with Breiman's claim that the most substantial improvements are already seen with only 10 bootstrap replicates in all datasets. In a few of the datasets, the misclassification rates fluctuate (decrease, then increase) for larger number of bootstrap replicates, and thus we conclude that increasing the number of bootstrap replicates does not always decrease the misclassifications rates.

Furthermore, Breiman claimed that "more replicates are required with an increasing number of classes." Comparing the number of classes with the misclassification rates, we observe similar improvements in performances in datasets with fewer number of classes (waveform, breast cancer, ionosphere, and diabetes) with the datasets with greater number of classes (glass and soybean). That is, 10 bootstrap replicates already show substantial improvements. We computed the percent decrease in misclassification rates from 10 bootstrap replicates to 100 bootstrap replicates and we find that the improvements in the rates of glass is similar to improvements of heart and ionosphere, and the improvements in soybean is similar to the improvements in diabetes. Thus, we conclude that high number of classes does not require more bootstrap replicates to show substantial improvements, given the datasets considered.

Table 2: Misclassification Rates by No. of Bootstrap Replicates

Dataset	10	25	50	100	No. of Classes	Percent Decrease
waveform	18.1	16.6	16.2	16.2	3	10.5%
breast cancer	4.4	4.0	3.8	3.8	2	14.9%
ionosphere	8.0	7.0	7.4	7.0	2	11.5%
diabetes	25.5	24.6	24.1	24.2	2	5.1%
glass	25.1	23.6	22.4	22.2	6	11.6%
soybean	9.7	9.5	9.1	9.3	19	4.5%

3 Bagging Regression Trees

Breiman demonstrates how bagging affects regression tasks through his experiments on Regression Trees. Regression trees are decision trees used as predictive models to draw conclusions about continuous valued outputs.

3.1 Methodology

3.1.1 Datasets

To replicate Breiman's analysis, we utilize these 5 datasets:

1. **Real Datasets**: We find the datasets through R packages [4], since they are not widely available.
   - *Boston Housing*: 506 cases corresponding to census tracts in the greater Boston area, with response variable as the median housing price and 13 (mostly socio-economic) predictor variables. The dataset Breiman used in the paper had 12 predictor variables and this discrepancy may influence the results.
   - *Ozone*: 366 readings of maximum daily ozone (response variable) at a hot spot in Los Angeles and 11 (mostly meteorological) predictor variables. We perform the same data cleaning process he performed, leaving us with 330 readings. Again, the dataset Breiman used in the paper had 9 predictor variables instead, and this discrepancy may influence the results.
2. **Simulated Datasets**: We generate the datasets following Breiman's outlined simulation structure.
   - *Friedman #1*: Ten independent predictor variables $x_1, ..x_{10}$, each uniformly distributed over [0, 1]. Response is then given by:

     $$y = 10sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon$$

     where ϵ is $N(0, 1)$.
   - *Friedman #2, #3*: Four independent predictor variables with

     $$\#2 : y = (x_1^2 + (x_2 x_3 - (1/x_2 x_4))^2)^1/2 + \epsilon_2$$

     $$\#3 : y = tan^{-1}\left(\frac{x_2 x_3 - (1/x_2 x_4)}{x_1}\right) + \epsilon_3$$

     where x_1, x_2, x_3, x_4 are uniformly distributed over these ranges:

     $$0 \leq x_1 \leq 100$$

     $$40\pi \leq x_2 \leq 560\pi$$

     $$0 \leq x_3 \leq 1$$

     $$1 \leq x_4 \leq 11$$

     The noises are distributed as $N(0, \sigma_2^2), N(0, \sigma_3^2)$ where σ_2, σ_3 were chosen to give 3:1 signal-to-noise ratios.

3.1.2 Computations

We perform the following procedure on the datasets:

(i) Divide real data sets at random into learning set L and test set T using a 95-5 split. For simulated data sets, generate a learning set L of 200 samples and test set T of 1000 samples.

(ii) Grow regression tree using L and 10-fold cross-validation for hyperparameter tuning of $criterion$, $splitter$, and max_depth hyperparameters used in Python's $sklearn.tree$ models.

(iii) Run test set T down this tree to compute mean squared error (MSE) as $e_S(L, T)$

(iv) Generate 25 bootstrap replicates $L^{(B)}$ of L and grow a regression tree using $L^{(B)}$ and 10-fold cross-validation with hyperparameters tuned as before to create 25 tree regressors $\varphi_1(x), ..., \varphi_{25}(x)$.

(v) Average the predictions over all bootstrap replicates

$$\hat{y}_B = av_k \varphi_k(\mathbf{x}_n)$$

and compute MSE between $\hat{y}_B$ and true values as $e_B(L, T)$

(vi) Repeat the procedure 100 times and average the respective errors to obtain a single tree error $\bar{e_S}$ and a bagged error $\bar{e_B}$.

3.2 Results

The results from our replication of Breiman's analysis align with his results and trends. However, due to the differences in our datasets, variability in σ_2, σ_3 choices, and lack of details on which hyperparameters Breiman utilized for his analysis, some results do not match in magnitude, but all match in range (since even small changes in errors can lead to moderate changes in percent decrease).

Table 3: Mean Squared Test Set Error

Data set	$\bar{e_S}$	$\bar{e_B}$	Decrease	Breiman's Results
Boston Housing	22.1	11.6	54%	39%
Ozone	22.6	17.5	34%	22%
Friedman #1	11.4	5.7	50%	46%
Friedman #2	79,925	58,747	26%	30%
Friedman #3	0.0784	0.0496	37%	38%

3.3 Extensions

In our analysis above, we use 25 bootstrap replicates, but as stated earlier in the classification extension section, Breiman claims that you see most improvement even with 10 replicates. In fact, he further claims that he proposed less number of replicates for regression (25) than the classification problem (50) because when response is numerical, "his sense is that you need fewer replicates". We decided to verify his claims by implementing our computation procedure for bootstrap values of 10, 25, 50, and 100.

Table 4: Mean Squared Test Set Error by No. of Bootstrap Replicates

Data set	10	25	50	100
Boston Housing	12.1	11.6	11.2	11.2
Ozone	18.9	17.5	17.5	17.6
Friedman #1	6.2	5.7	5.5	5.4
Friedman #2	61,765	58,747	57,754	57,214
Friedman #3	0.0529	0.0496	0.0486	0.0482

We agree with Breiman's first claim that you see the most improvement with 10 replicates itself and choosing smaller number of replications is reasonable. For his second claim that you need more replicates for classification than regression, we found that in both problems, 10 replicates were more than sufficient to observe significant improvement.

4 Forward Variable Selection

Breiman illustrates the effects of bagging in the context of forward variable selection. Forward selection is a type of stepwise linear regression that begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model.

4.1 Methodology

We follow Breiman's simulation structure to generate data. In particular, the simulated data are drawn from the model.

$$y = \sum_m \beta_m x_m + \epsilon$$

where ϵ is $N(0, 1)$. The number of variables $M = 30$ and sample size is 60. The x_m are drawn from a mean-zero joint normal distribution with $EX_iX_j = \rho^{|i-j|}$ and at each iteration, is selected from a uniform distribution on $[0, 1]$.

Three sets of coefficients are generated, wherein each set of coefficients consists of three clusters centered at $m = 5, 15, 25$, respectively. Each cluster is of the form where k is the cluster center, and

$$\beta_m = c[(h - |m - k|)^+]^2, m = 1, ..., 30$$

$h = 1, 3, 5$ for the first, second and third set of coefficients, respectively. The normalizing constant c is further chosen as described in the paper. We do the following for each set of coefficients:

(i) Draw data L from the simulation structure.

(ii) Perform forward selection on the data. The metric used in the forward selection procedure is mean squared error (MSE). The mean-squared prediction error of each model is computed giving $e_1^{(S)}, ..., e_M^{(S)}$.

(iii) Fifty bootstrap replicates $L^{(B)}$ of L are generated. For each one of these, perform forward selection to construct $M = 30$ m-predictor models. Then, average the predictions over all bootstrap replicates for every m-predictor model. The prediction errors $e_1^{(B)}, ..., e_M^{(B)}$ for this sequence are computed.

Breiman replicates the following procedure 250 times in his paper but after a couple of trials we found that the results stabilize after 5 runs if we employ 5-fold cross validation (CV) while performing the forward selection. We chose this strategy instead as it was a more computationally efficient method and also allowed us to report CV prediction error. The computed CV mean-squared errors are averaged over the 5 repetitions to give two sequences $\bar{e}_m^{(S)}$, $\bar{e}_m^{(B)}$ for each of the three set of coefficients.

We also repeated this exercise with 25 and 10 bootstrap replicates to see if varying the number bootstrap replicates changed our findings.

4.2 Results

On the whole, our results support Breiman's conclusion that bagging performs better than subset selection. The range of our prediction errors are very similar to the ones obtained by Breiman but are lower overall. Unsurprisingly and in line with Breiman's results, bagging outperforms subset selection, which can be an unstable procedure. However, we find that in our case, bagging performs substantially better than vanilla subset selection even in cases where the number of predictors m is high.

Figure 1 shows Breiman's findings that beyond a certain point, bagged predictors have a larger prediction error than the unbagged. His rationale for this is that linear regression using all or close to all variables is a fairly stable procedure, Hence, he claims that bagging does not improve prediction accuracy over subset selection for high values of m. However, we find that bagging consistently outperforms subset selection for all values of m.

Additionally, for each set of coefficients, we observe a decrease in prediction error as we approach the true number of non-zero coefficients followed up an increase in error as we begin to overfit the model.

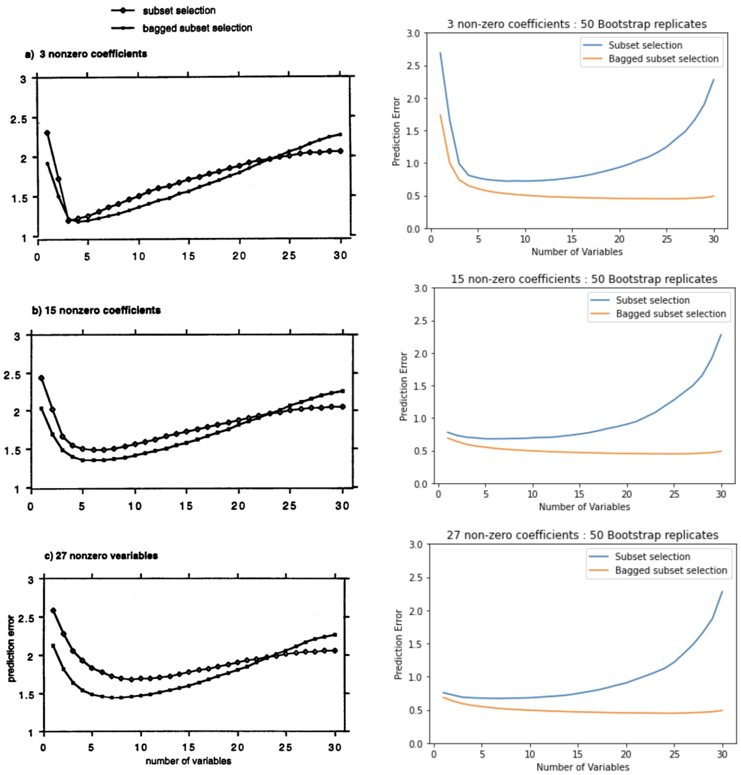

Figure 1: Breiman's results (left) compared to our results (right)

This is the canonical illustration of the bias-variance trade-off and we observe that the increase in error due to overfitting is vastly greater in the case of subset selection as compared to bagging where the increase in error is very slight. This coincides with what we know about bagging as a variance reduction technique.

Repeating the exercise with 25 and 10 bootstrap replicates did not change our findings in any way. This lines up with Breiman's intuition that fewer bootstrap replicates are necessary when y is numerical. These results are shown in the appendix.

4.3 Bagging Nearest Neighbor Classifiers

We repeat the analyses of bagging classification trees with nearest neighbor classifiers on all of the datasets described in Section 2 except for the soybean data, as Breiman pointed out the variables were categorical.

4.4 Methodology

The datasets used are identical to the datasets used in Section 2.1.1., and the computations carried out here are similar to the computations in Section 2.1.2. We used the same procedure to randomly divide the datasets into learning and test sets with 100 bootstrap replicates, and 100 iterations in each run. As per Breiman's guideline in the paper, we used a Euclidean metric that is standardized by the standard deviation of each predictors over the learning set.

4.5 Results

Our results confirm the findings in the paper that bagging with nearest neighbor is more accurate than the single trees in Section 2 for all but one of the datasets, but bagging with classification trees are more accurate than bagging with nearest neighbor in all but one of the datasets.

Table 5: Misclassification Rates (Percent)

Dataset	$\bar{e_s}$	$\bar{e_b}$
waveform	22.7	22.7
breast cancer	4.6	4.6
ionosphere	47.7	36.8
diabetes	27.6	27.7
glass	28.2	28.4

Our findings show that bagging with nearest neighbors provide very similar performance with nearest neighbor models without bagging, further confirming Breiman's findings. However, we note that our results for the ionosphere dataset is actually improved with bagging.

In the paper, Breiman argues that given N cases in the learning set, the probability that the nth case is selected 0,1,2,.. times is approximately Poisson distributed with $\lambda = 1$ for large N. Thus the probability that the nth case is selected at least once is $1 - (1/e) \approx .632$. If there are N_B bootstrap repetitions in a 2-class problem, then a test case may change classification only if its nearest neighbor in the learning set is not in the bootstrap sample in at least half of the N_B replications, which happens with probability that the number of heads in N_B coin flips (with probability .632 of heads) is less than $.5N_B$. We conclude that since nearest neighbor classification methods is more stable than other models such as trees and neural networks, bagging may not result in substantial gain in performance.

5 Conclusion

Our replications of Breiman's analysis are successful in that we were largely able to confirm his results using his methods. Our findings agree that bagging improves accuracy in the case of more unstable methods such as regression and classification trees, and subset selection. For subset selection, we in fact observe far better performance with bagged predictors than Breiman. In the case of bagging nearest neighbors, we observe either a slight degradation or no improvement in performance (barring the ionosphere dataset). Some of the datasets seem to have changed since Breiman's paper was published, which could explain some differences in our findings.

We also investigate Breiman's intuition that fewer bootstrap replicates are required when y is numerical, while the converse is true for a categorical y with increasing number of classes. We find that indeed in the case of bagging regression trees and forward variable selection, as low as 10 bootstrap replicates show results similar to 50 replicates. However, contrary to Breiman's intuition, we find that this results holds true even for bagging classification trees.

References

[1] Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

[2] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

[3] Learning, U. C. I. M. (2016, October 6). Pima Indians Diabetes Database. Kaggle. Retrieved March 23, 2023, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[4] Leisch F, Dimitriadou E (2021). mlbench: Machine Learning Benchmark Problems. R package version 2.1-3.

[5] UCI Machine Learning Repository: Data Sets. (n.d.). Retrieved March 13, 2023, from https://archive.ics.uci.edu/ml/datasets.php

Appendix

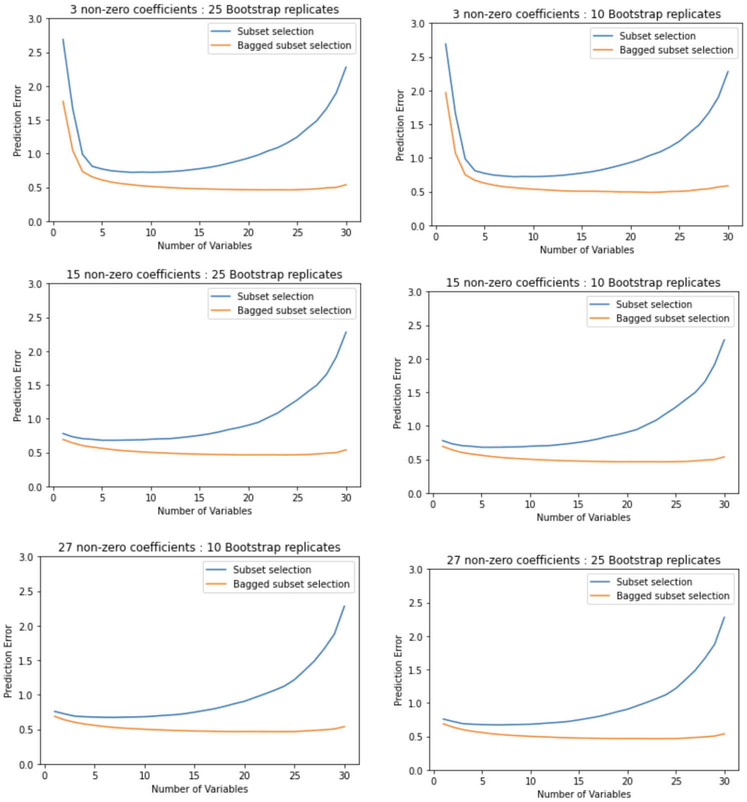

Figure 2: Forward subset selection analysis with $B = 25$ and $B = 10$ bootstrap replicates