3

Applications of Machine Learning in Forecasting Regressions: Boosting United States and Japan

Jonathan B. Ma

Princeton University (Class of 2015)

Applications of Machine Learning in Forecasting Regressions: Boosting United States and Japan

Abstract

Does applying machine learning on large datasets yield accurate recession forecasts? This paper applies boosting, considered one the best off-the-shelf classifiers in machine learning, to forecasting recessions in the United States and Japan. Instead of forecasting recessions with one or a few predictors, we utilize large macroeconomic datasets and use boosting to select the most predictive variables and perform prediction. We investigate if a large predictor set, specifically the 132 monthly predictors from Stock and Watson [2005], combined with boosting can forecast recessions better than the best logit model in the United States. We then look ouside of the United States to see if a similarly large predictor set in Japan predicts recessions better than the best logit model. We find that while boosting outperforms the best logit model in-sample, boosting actually performs worse than the best logit model in the United States and Japan out-of-sample. By carefully selecting a smaller dataset that consists of leading indicators, we are able to boost a small dataset that performs better than boosting the large dataset. Our general finding reiterates the parsimony principle, that simpler models often outperform more complex models.

CLASSIFICATION:

KEYWORDS:

3.1 Introduction

In today's world where economists have access to big data, we are curious if more data can lead to more accurate recession forecasts. While more data should theoretically allow for more accurate forecasting models, Occam's Razor and the parsimony principle suggests that simpler models often better explain phenomenon than more complicated models. Thus, we look to settle this debate by applying machine learning on large datasets to forecasting business cycle turning points. We are motivated to study business cycle turning points and recessions because better understanding how to forecast recessions may provide policymakers information ahead of time to mitigate the severity of recessions through early adoption of expansionary policy.

We are specifically interested in the machine learning algorithm boosting, considered the best off-the-shelf classifier. Boosting has applications from spam detection to page ranking for search engines and has only recently been making its way into economics. The power of boosting lies in its ability to combine weak learners – rules of thumb that can predict only slightly better than chance – into a strong learner that can classify significantly better. For example, our initial weak learner classifies the current state of an economy as a recession if unemployment rate is above 10% , otherwise classify the state of the economy as not in a recession. Boosting initially weighs all observations equally and observations that were classified incorrectly (e.g the months where there is a recession when unemployment rate is below 10%, thus violating our initial weak learner) are weighted even more heavily. Additional rules of thumbs (e.g. classify economy as a recession if consumer confidence index is below 50, otherwise not in a recession) are introduced and aim to correct for these previous misclassifications. The output of boosting combines these rough rule of thumbs into a highly accurate prediction rule.

Ng [2014] introduces boosting as a way to forecast recessions and applies boosting to predict U.S. recessions 3, 6, and 12 months in advance. Ng [2014] finds that boosting provides valuable information on which variables are the strongest predictors at specific time horizons. Ng also finds that boosting allows for the composition of predictor sets to change over time. We hypothesize that allowing the composition of predictors to change will allow boosting to take into account the changing nature of business cycles and the fact that no two business cycles are alike. Ng [2014] admits boosting is far from perfect for analyzing recessions as the model presented in the paper misses recessions and produces false positives. However, Ng does not evaluate how much worse (or better) boosting does relative to other methods that forecast recessions. In our paper, we fill in this gap and benchmark boosting's ability.

Boosting has been shown to be effective at variable selection [Zeng, 2014] and to forecast macroeconomic variables such as industrial output [Buchen and Wohlrabe, 2011], but not a lot of work has been done on investigating boosting's ability to forecast recessions and whether or not there are gains from incorporating big datasets in recesion forecasting. Berge [2014] finds that a non-linear and linear specification of boosting in a small dataset of 20 macroeconomic variables outperforms the best logit model in forecasting recessions. His finding motivates us to try boosting a even larger dataset in the United States and to investigate boosting's merits or drawbacks in forecasting recessions outside of the United States. Limited work has been done on applying boosting to large datasets outside of the U.S. because gathering these datasets are not readily available and constructing these datasets are tedious. The one exception is Wohlrabe and Buchen [2014] who apply boosting to forecasting macroeconomic variables in Germany and the Euroarea, though not looking specifically at recessions like we do. We turn our attention to Japan because there is a wealth of macroeconomic data since the 1970s and because of the prevalence of recessions since the 1970s.

While boosting outperforms the best logit model in-sample for the U.S., we find that forecasting with boosting out-of-sample using the Stock and Watson [2006] 132 predictor set actually performs worse than the best logit model in the United States at the 3, 6 and 12 month horizon. Similarly, we also find that boosting a large dataset in Japan that mimics that of the Stock and Watson [2005] dataset predicts well in-sample but worse out-of-sample compared to the best logit model. We find that the main reason why boosting does relatively worse out-of-sample than the best logit model is that boosting tends to overfit on the training data thereby incorporating weak predictors and diluting the predictive power of the strongest predictors. Our finding reiterates the parsimony principle, that simpler is often better and that parsimony is especially important for forecasting recessions because if a model incorporates even a slight amount of noise, the consequences could be inaccurate forecasts. Boosting may perform poorly forecasting recessions because of the small number of observations of distinct recessions but we hesitate to extrapolate this conclusion to all other forecasting exercises.

The paper is organized as follows. Section 2 reviews the literature on predicting recessions in the United States and Japan as well as methods used for predicting recessions. Section 3 details the methods used in forecasting recessions in our paper, notably explaining how boosting works. Section 4 discusses the dataset used for both Japan and United States as well as our evaluation criteria for comparing different classification models. Section 5 explores the in-sample and out-of-sample forecasting performance of boosting in the United States to evaluate boosting's performance. Section 6 applies the empirical analysis in the previous section but to Japan. Section 7 concludes with discussion and further work that can be done.

3.2 Related Work

3.2.1 Predicting U.S's Recessions

Recession dating in the U.S. starts with Burns and Mitchell [1946] and the tedious and manual search of leading indicators and coincident indicators by looking at the co-movement of economic variables. These indicators were used to inform the National Bureau of Economic Research's (NBER) understanding of business cycles and ultimately used to date recessions and expansions and were influential in forecasting recessions. The NBER Business Cycle Dating Committee officially dates business cycle turning points in the United States. Stock and Watson [1989] revised the indexes that Burns and Mitchell [1946] constructed and created an index of Leading Economic Indicators (LEI), Coincident Economic Indicators (CEI) and a Experimental Recession Index (XRI) that were very similar to the NBER's index of leading and coincident indicators. We note that our research is fundamentally looking at forecasting recession given the NBER's classification of the economy as a recession ($Y = 1$) or expansion ($Y = 0$) and thus treat the NBER turning point dates as a gold standard.

Much of the attempts to forecast recessions rely on one or a few predictors. The term structure of treasury yields–defined as the 10 year Treasury Bond - 3 month Treasury Bill Spread–is shown to have strong predictive power of U.S. recessions up to eight quarters or two years into the future [Estrella and Hardouvelis, 1991, Estrella and Mishkin, 1998]. Additionally, stock prices [Estrella and Mishkin, 1998], Stock and Watson [1989]'s index of the Leading Economic Indicator (LEI), the credit market [Levanon et al., 2011], and sentiment [Christiansen et al., 2013] are predictive of U.S. recessions. Liu and Moench [2014] finds that balances in broker-dealer margin accounts can improve recession predictions. Ng [2014] also finds that employment and interest rate measures are also strongly predictive and that AAA Corporate Bond - Federal Funds spread, a measurement of credit and liquidity risk, are the strongest predictor 3 and 6 months ahead whereas the 5 year Treasury Bond - Federal Funds spread is the most predictive 12 months in advance. Ng [2014] finds that the predictive power of spreads are often recession specific.

Ng and Wright [2013] argue that forecasting recessions are inherently difficult because of evidence that business cycles facts in the U.S. have changed in the last 2 decades. Ng and Wright describe how business cycles of the 1970s and 1980s were due to supply side shocks whereas the recessions in 1990-1992, 2001, and 2008-2009 originated from the financial sector. The authors conclude that the ability of predictors to predict recessions are episodic. The finding by Ng motivates our study of boosting because of boostings ability to select different variables over time and thus we hypothesize that boosting should outperform univariate models which rely only on one predictor. Estrella and Mishkin [1998], however, warns of "overfitting" and that including more predictors may hurt rather than help forecasting out-of-sample. Hence, Estrella and Mishkin [1998] advocates for predicting with just simple financial indicators like interest rates, spreads, stock prices and money aggregates. We seek to understand if boosting suffers from "overfitting" or if boosting can perform better than incorporating just simple financial indicators.

3.2.2 Predicting Japan's Recessions

Bernard and Gerlach [1998] summarizes how well the term structure of interest rates predicts recessions in Japan, Belgium, Canada, France, Germany, Netherlands, UK, and the U.S. The predictive power of the term structure is found to helpful in all countries except Japan as predicting recessions in Japan with the term structure has the lowest R^2 of all the countries. Bernard notes that this is likely because of tight regulation of Japanese financial markets earlier on, which limited the role of market expectations in determining interest rates or because Japan had fewer and shallower recessions than the other countries. Hirata and Ueda [1998], however, finds that during the period between the free interest rate movements and the publiciation of the paper in 1998, the yield spread did in fact predict recessions in Japan and that other financial variables such as monetary aggregates, stock price also seem to predict recessions. Hirata and Ueda [1998] are cautious of their findings due to the limited sample size. Hasegawa and Fukuta [2011] updates the literature by exploring the predictive power of the yield spreads before and after the structural break in Japanese growth in 1996. The authors find that while the yield spreads are predictive of recessions prior to the growth break in 1996, after 1996 the yield spread is not predictive. The majority of the forecasting literature in Japan has focused on the predictive power of 1 or a few variables, thus we explore boosting a large dataset and understanding the predictive merits of such a model.

3.2.3 Forecasting Recession Methods

Stock and Watson [1994] surveys 49 univariate forecasting models and other pooling forecasting methods of 215 U.S. macroeconomic time series variables from 1959 - 1996, finding that the autoregression does the best. However, this "horse race" paper does not cover how well these different methods forecast recessions. We review different methods used specifically in forecasting recessions.

Logit and Probit

Probit or logit is often used for forecasting recessions as Hirata and Ueda [1998] , Liu and Moench [2014], Bernard and Gerlach [1998], Hasegawa and Fukuta [2011] and many other studies use logit or probit to predict recessions. We do not consider other binary prediction models such as the linear probability model because of the well documented drawbacks highlighted in Horrace and Oaxaca [2006]. Kauppi and Saikkonen [2008] extends the simple probit model to a dynamic probit model that incorporates lags of explanatory variables and recessionary dummies and produced mixed results. Ng [2014] takes a similar approach by lagging explanatory variables in the boosting model used. Because of the ubiquity of logit and probit models, we will use the best logit model as our benchmark against our boosting models. In this paper, we use logit instead of probit due to probit's longer computation time of calculating the Maximum Likelihood Estimate.

Boosting

Boosting in theory makes for a better classifier than logit because boosting can take into account more variables than a simple logit model, account for nonlinearity, and train and forecast in a computationally reasonable time. Bai and Ng [2009] discover that boosting as a means to select predictors from a large dataset can perform quite well. Ng [2014] applies boosting on 132 macro time series to predict recessions in 610 months from 1961-3-01 to 2011-12-01. When forecasting 3 months in advance and using 4 lags, there becomes 532 predictors. When forecasting 3 months in advance and using 12 lags, there are 1596 predictors. Picking by hand which variables to include into the forecasting model would be tedious and computationally intensive, however boosting is able to automatically select which variables are most predictive in its model. While there is no economic theory that goes into a model like boosting, the variables the author found match economic theory and the literature in forecasting recessions. Berge [2014] continues the literature by applying boosting to 20 variables consisting of the slope, level and curve of the yield curve as well as other real economy and financial variables. Berge compares boosting with other models and finds that a non-linear variation of the boosting technique to have the best performance when forecasting recessions. It should be noted that Berge [2014] finds the gains from the boosting models relative to the best univariate models are relatively small. Since Ng [2014] does not benchmark her boosting method on her large dataset like Berge [2014] does, we look to quantify and evaluate how well Ng's boosting method performs. Furthermore, we extend Ng [2014] analysis to Japan to see if the predictive variables in the U.S. are the same in Japan.

3.3 Methodology

We explain in more detail how the boosting algorithm works and how we will apply the method to forecasting recessions in the U.S and Japan. For $t = 1, .., T$, we define $Y_t = 1$ if month t is a recession and $Y_t = 0$ if month t is not in a recession as defined by the NBER for the United States or the OECD for Japan. We define the predictor set as $x_{t-h} = (x_{1,t-h}, ..., x_{K,t-h})'$ where K is the number of predictors and h is the forecasting horizon. When we forecast Y_t we assume to observe x_{t-h-1} as opposed to x_{t-h} because some of the data we use are not released until 2-3 weeks into the month.

3.3.1 Logit

Logit assumes that log odds ratio of Y_t is a random function of x_t.

$$\log \frac{P(Y_t = 1|x_{t-h-1})}{P(Y_t = 0|x_{t-h-1})} = f(x_{t-h-1}, \theta) \tag{3.1}$$

$$P(Y_t = 1|x_{t-h-1}) = \frac{\exp(f(x_{t-h-1}, \theta)}{1 + \exp(f(x_{t-h-1}, \theta)} \tag{3.2}$$

where θ is the model parameters and h denotes forecast horizon, x_{t-h-1} is defined just as before. We assume a linear relationship between the covariates and the outcome variable such that $f(x_{t-h-1}|\theta) = x_{t-h-1}\beta$. We calculate the estimate $\hat{\beta}$ and predict the probability of of Y_t using maximum likelihood estimation or software such as R's glm package and specifying a Bernoulli loss function and using logit as the link.

3.3.2 Boosting

In this section, we explain how AdaBoost Schapire [1999], the algorithm that began the boosting literature, works. We then discuss how AdaBoost can be generalized and solved using gradient descent to create a generalized Gradient Boost technique that we ultimately apply using R's GBM package. The following sections borrow heavily from the explaination of Berge [2014], Ng [2014].

AdaBoost

AdaBoost is an algorithm that combines weak learner classifiers–rules of thumbs that perform only slightly better than random guessing – to form strong learners that classify substantially better than weak learners on their own. Stronger learners are defined as a classifier $f(x)$ that has an error rate $ERROR = E[1(f(x) \neq Y]$ that is arbitrary small such that $P(ERROR < \epsilon) \geq 1 - \delta$ for all $\delta > 0, \epsilon > 0$. Weak learners are defined as having $ERROR$ such that there exists $\gamma > 0$ such that $P(ERROR < 1/2 - \gamma) \geq 1 - \delta$. Schapire [1990] finds that weak learners, regardless of distribution, can be boosted to become a strong learner with high accuracy.

Because we utilize the sign function sign(X) which returns 1 or -1 depending on the sign of X, we define $y_t = 2Y_t - 1$ such that $y_t = 1$ if there is a recession and $y_t = -1$ if there isn't a recession. A sketch of the AdaBoost algorithm is given in Algorithm 1.

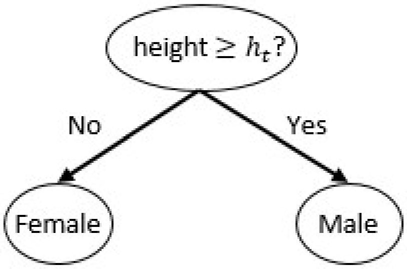
Algorithm 1: Discrete AdaBoost Schapire [1999]

```
1. Initialize observation weights $w_t = 1/T, t = 1, 2, ..., T$ and $F_0(x) = 0$
2. For $m = 1$ to $M$ :
    (a) Find $f_m(x)$ from the set of candidate models to minimize the weighted error:
        $\epsilon_m = \sum_{t=1}^{T} w_t^{(m)} 1(y_t \neq f_m(x_t))$
    (b) If $\epsilon_m < 0.5$, update $F_m(x_t) = F_{m-1}(x_t) + \alpha_m f_m(x_t|\theta)$ and calculate the updated weights:
        $w_t^{(m+1)} = \frac{w_t^{(m)}}{Z_m} exp(-\alpha_m y_t f_m(x_t;\theta)$
        where $Z_m = 2\sqrt{\epsilon_m(1-\epsilon_m)}$ and $\alpha_m = \frac{1}{2}log(\frac{1-\epsilon_m}{\epsilon_m})$
3. Return classifier $sign(F_M(x))$
```

Here, M is the total number of iterations, T is the total number of observations, ϵ is the error rate, $y_t = \{-1, 1\}$, Z_m is a normalizing factor optimally chosen such that $\sum_{n=1}^{N} w_t^{(m+1)} = 1$

At step 1, we must select the weak learner $f_m(x)$ which is a function parameterized by θ and maps features of x into the class labels $y_t = \{-1, 1\}$. For example, $f_m(x)$ could be a decision stump that assigns $y_t = 1$ if $x \geq \theta$ or that predicts a person as a male or female based on their height as seen in Figure 3.1.

Figure 3.1: Example of Decision Stump for Determining Male or Female[1]



Step 2(a) calculates the error rate of each of the weak learner and selects the model that yields the lowest weighted error. Furthermore, AdaBoost specifies in step 2(b) that $\epsilon_m < 0.5$, otherwise $e_m \geq 0.5$ the classification ability is < 0.5 implying the weak learner is as worse than random guessing which would disqualify the learner from being a weak learner as defined earlier.

In step 2(b), the weak learner that minimizes the weighted error from 2(a) is added to the stronger learner. The magic of boosting happens with the reweighting that takes place in 2(b), noting that:

$$w_t^{(m+1)} = \frac{w_t^{(m)}}{Z_m} \begin{cases} exp(-\alpha_m) & y_t = f_m(x_t|\theta) \\ exp(\alpha_m) & y_t \neq f_m(x_t|\theta) \end{cases}$$

Since if $y_t \neq f_m(x_t, \theta)$ then the resulting value must be -1 thus making the weight $exp(-\alpha * -1) = exp(\alpha)$. Note that $exp(\alpha) > 1$ and $exp(-\alpha) < 1$. Thus, cases where the weak learner misclassifies on an observation t, the weight is increased (since $exp(\alpha) > 1$) whereas observations where the

[1] For all practical purposes, say $h_t = 70$ inches and $height \geq 70$ inches would lead to a classification of a person as a male, otherwise, female.

classification is correct the weight is decreased (since $exp(-\alpha) < 1$). Because of the reweighting scheme, the algorithm forces the classifier to train on misclassified observations t by increasing the weight at 2(b).

After M iterations, step 3 returns either -1 or 1 depending on the sign of the strong learner.

Ng [2014] gives a toy example in the appendix of her paper to illustrate how Adaboost would apply to forecasting recessions 3 months ahead in the 12 months in 2001. We briefly summarize below.

At the initial iteration, we assign equal weight $w_1 = 1/12$ for all twelve months in 2001. We also set our weak learners as decision stumps. At the first iteration $m = 1$, the decision stump that minimizes the classification error classifies as follows

$$y_t = \begin{cases} 1 & HWI < -0.44 \\ -1 & HWI \geq -0.44 \end{cases}$$

where HWI stands for the help wanted index and where -0.44 is the threshold that minimizes classification error. We find that the initial weak learner has a error rate (ϵ) of 0.167 calculated according to 2(a) as the initial weak learner classifies 2 of the 12 months incorrectly so $2 \cdot w_1 = 2 \cdot 1/12 = 2/12 = 0.167$. We calculate $\alpha_1 = 0.5log(\frac{1-\epsilon}{\epsilon}) = 0.5log(\frac{1-0.167}{0.167}) = 0.804$. Furthermore the weights are updated in accordance to 2(b) such that the 2 months that were misclassfied now each have weight 0.25 whereas the other 10 months that were classified correctly each have weight of 0.05.

After 5 iterations, an example of a stronger learner or a combination of weak learners is:

$$\begin{aligned} \hat{y} = 0.804 * 1(HWI < -0.44) + 1.098 * 1(NAPM < 49.83) + 0.710 * 1(HWI < -0.1) \\ + 0.783 * 1(SPREAD > -0.622) + 0.575 * 1(NAPM < 47.062) \end{aligned}$$

Where

- $\hat{y}$ represents the classifier returned at step 3 and is the ensemble of the 5 weak learners and predicts a recession if $\hat{y} > 0$ and not a recession if $\hat{y} < 0$.

- HWI stands for the help wanted index
- $NAPM$ is the number of new orders for manufacturers
- $SPREAD$ is the 10 year Treasury Bond - Fed Funds Rate spread
- The weights 0.804, 1.098, 0.710, 0.783, 0.575 represent α_m calculated in Algorithm 1 step 2(b).

Whereas the initial weak learner in the TOY example had a classification error of 0.167, the stronger learner at the end had a classification error of 0, thus classifying perfectly. Also note that variables were selected more than once which is permissible in AdaBoost.

AdaBoost in a nutshell initializes and weighs all observations equally, and through each iteration increases the weight of observations that were classified incorrectly and decreases the weight on correctly classified observations. As AdaBoost adds additional weak learners, these weak learners are forced to focus on the misclassified observations. The output of AdaBoost is a classifier that is boosted by the M weak learners and classifies substantially better than an individual weak learner on its own.

Gradient Boosting

Friedman et al. [2000] draws the connection between AdaBoost and a stage wise additive model with an exponential loss function, turning a seemingly powerful but unfamilar algorithm in AdaBoost into a familiar statistical concept. A generalized additive model can take the form:

$$
\begin{aligned}
E[y|x_1, x_2, ..., x_M] &= \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_M(x_M) && (3.3) \\
&= \beta_0 + \sum_{m=1}^{M} f_m(x_m) && (3.4)
\end{aligned}
$$

Where Y is the outcome variable, $X_1, ..., X_M$ are M different predictors and f_m are unspecified nonparametric functions.

Thus, AdaBoost can be viewed in the lens of an additive model as follows:

$$F_M(x) = \sum_{m=1}^{M} \rho_m f_m(x, \theta_m) \tag{3.5}$$

Where

- $F_M(x)$ is the stronger learner
- $f_m(x)$ is the weak learner at iteration m
- ρ_m is the step-size or the regularization parameter
- M is the number of total iterations
- θ_m is the parameter of the weak learner
- x is the data

AdaBoost is then the solution to the following loss function with exponential loss:

$$\hat{F}(x) = argmin_{F(x)} E[L(y, F(x))] \tag{3.6}$$

$$L(y, F(x)) = exp(-yF(x)) \tag{3.7}$$

Friedman [2001] generalizes AdaBoost to Gradient Boosting to take into account other loss functions besides the exponential loss function such as the bernoulli loss function which Ng [2014] uses in her paper. Furthermore, Friedman [2001] introduces the empirical counterpart to the original AdaBoost to solve for (3.6) using gradient descent. The gradient boosting algorithm is sketched in Algorithm 2 and is adapted to take in the time series data we will be using to forecast recessions.

Algorithm 2: Gradient Boosting Friedman [2001] minimizing $L(y, F)$

```
Input: Choice of loss function $L(y, F)$ , number of iterations $M$, choice of functional form
  for weak learner $f^{(k)}$ for $k = 1, .., K$, shrinkage factor $\rho$, data $(y_t, x_{t,1}...x_{t,K})_{t=1}^{T}$ where there
  are $T$ observations and $K$ number of covariates

  1. Set $F_0$ to the constant that minimizes empirical loss.

  2. For $m = 1, ..., M$

      (a) Compute the negative gradient of the loss function evaluated at the current estimate of
          F which is $\hat{F}_{m-1}$. This produces
          $u_m \equiv \{u_{m,t}\}_{t=1,..,T} = -\frac{\partial L(y_t, F)}{\partial F}|_{F=\hat{F}_{m-1}(x_t)}, t = 1, ..T.$
      (b) Fit each weak learner $f^k$ for $k = 1, ..., K$ to the current negative gradient vector $u_m$
      (c) Let $\hat{f}_m^k$ be the best fit of $u_m$ among the $K$ weak learners.
      (d) Update the estimate of $F$ by adding the weak learner $k$ to the estimate of
          $F_m(x) = F_{m-1}(x) + \rho \hat{f}_m^k$
          Where $\rho$ is a predetermined step size

  3. Return $F_M(x)$
```

We discuss the inputs for Gradient Boosting. The loss function chosen for Gradient Boosting can be an arbitrary loss function. For instance, specifying an exponential loss function as in (**??**) would lead to solving for the original AdaBoost algorithm. However, if we were looking to solve a regression problem where the outcome variable is a cardinal variable $Y_t = (-\infty, \infty)$ the loss function could be the squared loss function. M is important because the total number of iterations M weighs the tradeoff of bias and variance. While Bai and Ng [2009] uses BIC and Berge [2014] uses Schwarz in determining M, we use cross-validation to determine M as Buchen and Wohlrabe [2011] find that using the information criteria tends to lead to more overfitting and cross-validation proves to have more accurate forecasts. ρ is between 0 and 1 and is considered the step size and regularization parameter. To follow the work by Ng [2014] we set $\rho = 0.01$.

At step 1 of Gradient Boosting, we could set our weak learner $f^{(k)}$ as a decision stump. Formally our weak learner as a decision stump would look like

$$f^k(x_{t,k}) = c^L 1(x_{t,k} \leq \tau) + c^R 1(x_{t,k} > \tau) \tag{3.8}$$

where $x_{t,k}$ is the macroeconomic variable at time t of predictor k and belongs to one of two partitions depending on the value of a data dependent threshold τ. c^L and c^R are parameters and are typically the mean of observations in the partition.

Step 2(a) fits each of the K weak learners to the negative gradient of the specified loss function given the current estimate of the strong learner. Step 2(b) searches across all the weak learners to choose the one that most quickly descends the function space. Step 2(c) updates the strong learner with the weak learner with the best fit. Note that at each iteration, we update the current strong learner F_{m-1} at each step and add iteratively the best weak learner f_m but do not update or affect previous weak learners selected in prior iterations.

Ng [2014] states that the relative importance of predictor k can be assessed by how it affects variation in $F_M(x)$. To determine the relative importance of predictor k, Friedman [2001] suggests using

$$I_k^2 = \frac{1}{M} \sum_{m=1}^{M} i_m^2 1(id(x^m) = k) \tag{3.9}$$

where $id(x^m)$ is a function that returns the identity of the predictor chosen at stage $m.I_k^2$ signifies the number of times predictor k is selected over M iterations weighted by predictor k's improvement in squared error as given by i_m^2. $\sum_{k=1}^{K} I_k^2 = 100$. Thus variables with higher I_k^2 signify higher importance of the associated variables and variables not selected at all have 0 importance.

Ridgeway [2007] extends the work of Friedman [2001] and Schapire [1999] and develops a package in R called GBM to implement generalized boosting models. We follow Ng [2014] in using the GBM package in R in our paper for our boosting models though other alternatives exist[2].

[2] For instance, Berge [2014] uses the R package mboost

3.4 Data and Evaluation

We discuss the macroeconomic variables used in the United States and Japan to forecast recessions.

3.4.1 United States

In the United States, we use NBER recession dating which does not use the typical definition of two consecutive quarters of declining GDP to define a recession. Instead the NBER considers many measures of activity such as the real GDP measured on the product and income sides, economy-wide employment, and real income. Further, the NBER also look at indicators that do not cover the entire economy, for instance real sales and the Federal Reserve's index of industrial production (IP). We use NBER based Recession Indicators for the United States from the Period following the Peak through the Trough provided by the Federal Bank Economic Data (FRED) as our formal definition of a recession.

We use two types of datasets in the United States, one large and one small. For the large dataset, we use the same standard data that Stock and Watson [2005] and Ng [2014] use to forecast recessions which are 132 monthly U.S. variables but with a longer time horizon from 1959-02-01 to 2014-09-01. We start with 1959-02-01 and end with 2014-09-01 since that's the period when all 132 variables are available. The 132 monthly variables are split up into 7 groups: Output and Income, Labor Market, Housing, Consumption Orders and Inventories, Money and Credit, Bond and Exchange rate. We drop the "Index of Help-Wanted Advertising in the Newspapers" because the series was discontinued in 1966 and we do not include the lagged NBER recession variable in our predictor set because the NBER typically do not date recessions until 5-21 months later and including lagged recession variables in our model would be unrealistic of forecasting in real-time. Because we omit 2 variables from our predictor set, we use 130 monthly predictors in the U.S. In the appendix, we illustrate any differences between our dataset and the ones used by Stock and Watson [2005] and Ng [2014]. We also include the the appendix description of the large dataset of 130 monthly indicators and the transformations done to achieve stationarity.

For our more parsimonious small dataset, we use 10 predictors from the Conference Board Leading indicators with slight modifications. The small dataset can be found in Table 3.1. The Conference Board Leading Indicators and the modifications made can be found in the appendix.

Table 3.1: U.S. Monthly Indicators: Small Dataset

Description
Average weekly hours, manufacturing
Average weekly initial claims for unemployment insurance
Manufacturers' new orders, consumer goods and materials
ISM® Index of New Orders
Manufacturers' new orders, nondefense capital goods excluding aircraft orders
Building permits, new private housing units
Stock prices, 500 common stocks
Leading Credit Index
Interest rate spread, 10-year Treasury bonds less federal funds
Average consumer expectations for business conditions

There have been 8 recessions between 1959-04-01 and 2014-09-01, and about 14% of the time the U.S. has been in a recession during that period. In our training set, from 1959-04-01 to 1985-08-01, there have been 5 recession. In our out-of-sample from 1985-08-01 to 2014-09-01, there have been 3 recessions.

3.4.2 Japan

The Economic and Social Research Institute (ESRI) Business Cycle Indicators Committee looks at coincident indicators to date business cycle turning points in Japan. One method of dating business cycles is using the Bry-Boschan method that formalizes the rules used by the NBER recession dating commmittee in a computer routine.[3] The Organisation for Economic Co-operation and De-

[3] More on the computer routine can be found in Bry and Boschman [1971]

velopment (OECD) uses this procedure to identify turning points in business cycles in Japan. To be consistent with the United States, we specifically use OECD based Recession Indicators for Japan from the Period following the Peak through the Trough as the recession variable in Japan.

For Japan, we construct a “large” dataset that uses monthly indicators that closely models that of the standard one used by Stock and Watson [2005]and Ng [2014] for the United States. We select 93 macroeconomics variables collected by the Federal Reserve Economic Data (FRED), Global Insight Database, Japan’s Cabinet Office, and the Bank of Japan. The variables are broken down into 10 groups: Export, Import, Trade; Output and Income; Labor Market; Housing; Consumption, Order and Inventories; Money and Credit; Bond and Exchange Rates; Prices; Stock Market; TANKAN Business Surveys. In total we have 436 variables: 93 macroeconomics variables and 343 TANKAN business surveys that begin in 1975-01-01 and end in 2014-06-01.

We are specifically interested in TANKAN judgment surveys that begin in the 1970s. In the judgement surveys, enterprises are asked questions broken down by business conditions, domestic supply and demand conditions for products and services, inventory level of finished goods and merchandise, employment conditions, financial position, and lending attitude of financial institutions. To give an example, for business conditions, enterprises judge the general business conditions in light of individual profits as “(1) favorable”, “(2) not so favorable”, or “(3) unfavorable”. For financial position, enterprises rank their judgement of the general cash position on account as “(1) easy”, “(2) not so tight”, or “(3) tight”. The TANKAN diffusion indices (DI) are then calculated as the percent share of choice (1) minus the percent share of choice (3). An example is that the business conditions DI is calculated by subtracting the percentage share of enterprises responding “(3) unfavorable” from that of "(1) favorable”. Since the TANKAN business surveys are given every quarter, we use linear approximation to convert the quarterly data into monthly data.[4]

Descriptions of the variables from the large dataset can be found in the appendix as well as the transformations done to achieve stationarity. In addition, we construct a small dataset of 26 predictors consisting of the 14 leading indicators used by the cabinet office as well as 12 broad TANKAN

[4] More information about TANKAN can be found at the Bank of Japan: https://www.boj.or.jp/

business survey indicators. Description of the variables from the small dataset can be found in Table 3.2. We include additional information about the small dataset and the transformations of the small dataset in the appendix.

Table 3.2: Japan Monthly Indicators: Small Dataset

Cabinet Office Leading Indicators
Description
Index of Producer’s Inventory Ratio of Finished Goods: Final Demand Goods
Index of Producer’s Inventory Ratio of Finished Goods: Mining and Manufacturing
New Job offers (Excluding New School Graduates)
New Orders for Machinery at Constant Prices (Excluding Volatile Orders)
Total Floor Area of New Housing Construction Started
Consumer Confidence Index
Nikkei Commodity Price Index (42 items)
Interest Rate Spread (10 Year Gov. Bond - 3 month Interbank Rates)
Newly Issued Government Bonds Yield (10 Years)
Tokyo Interbank Offered Rates(3 Months)
Stock Prices(TOPIX)
Index of Investment Climate (Manufacturing)
Ratio of Operating Profits to Total Assets (Manufacturing)
Sales Forecast D.I. of Small Businesses

Table 3.3: Japan Monthly Indicators: Small Dataset *(continued)*

TANKAN Business Survey
Description
Business Conditions, All Enterprises, All industries, Actual result
Business Conditions, All Enterprises, All industries, Forecast
Inven Lvl of Finished Goods Merchandise, Actual result
Inven Lvl of Finished Goods Merchandise, Manufacturing, Actual result
Domestic Supply & Demand , All Enterprises, Manufacturing, Actual result
Domestic Supply & Demand, All Enterprises, Manufacturing, Forecast
Financial Position, All Enterprises, All industries, Actual result
Financial Position, All Enterprises, Manufacturing, Actual result
Employment Conditions, All Enterprises, All industries, Actual result
Employment Conditions, All Enterprises, All industries, Forecast
Employment Conditions, All Enterprises, Manufacturing, Actual result
Employment Conditions, All Enterprises, Manufacturing, Forecast

Godbout and Lombardi [2012] finds using a Quandt-Andrews breakpoint test that there exists a structural break in Japan in 1991 Q1. To take into account this break in the growth rate in 1991 Q1, we demean rate variables and detrend quantitative variables using 1991-01-01 as the break point. Detailed transformations of each variable from the large and small dataset is included in the appendix.

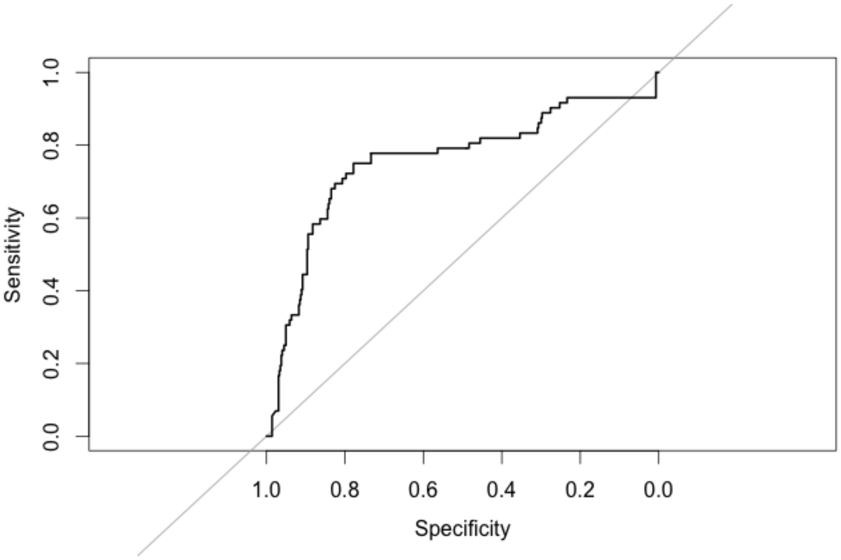
From 1975-01-01 to 2014-06-01, Japan has had 11 recessions whereas the U.S. has had 5 recessions during the same time period. Japan has been in a recession, as defined by the OECD, 41% of the time during this period. In our training set, 1975-01-01 to 1995-08-01, there are 5 recessions. In our testing set from 1995-08-01 to 2014-06-0, there are 6 recessions.[5]

[5] Japan's most recent recession ended in Q4 of 2014

3.4.3 Evaluation Criteria

A very popular evaluation metric from biostatistics, the Receiver Operating Characteristic (ROC) has been making its way into economics.Liu and Moench [2014] and Berge [2014] use ROC to evaluate the performance of different classification models. The ROC curve displays the trade-off between false positives and true positives, plotting the true positives (TP) on the y-axis against the False Positives (FP) on the x-axis. We illustrate an example of the ROC curve in Figure 3.2. As Liu and Moench [2014] explain, a model with 100% accuracy would draw a ROC curve that hugs the top left most corner whereas a model which does as well as random guesses would follow the 45 degree line running from the bottom left to the top right corner.

Figure 3.2: ROC Curve of Forecasting 12 Months Ahead in the United States with a AUC (Area Under Curve) of 0.7584



Area Under the Curve (AUC) of ROC summarizes the classification ability of a model. AUC illustrates how well a classification model discriminate for all possible threshold c where predictions greater than c would be classified as a 1 otherwise as 0. AUC is preferable to other methods such as the Root Mean Squared Error $(\hat{y} - y)^2$ because models with different squared errors may classify exactly same. For example, a model that predicted every recession with 0.51 probability and every non-recessionary period with 0.49 probability would have a significantly worse Root Mean Squared

Error than a model that classified every recession with 1.0 probability and every non-recessinoary period with 0 probability although the models classify precisely the same. Furthermore, AUC does not impose a loss function over the tradeoff between true positives and false positives. AUC lower than 0.5 signals the model has negative predictions and AUC greater than 0.5 signal positive predictions. To determine if a model has a AUC significantly bigger than another model's AUC, we define the following t-statistic following Hanley and McNeil [1983]:

$$t = \frac{AUC_1 - AUC_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2}} \tag{3.10}$$

AUC_1 and AUC_2 refers to the area under the curve for model 1 and 2 respectively and σ_1^2 and σ_2^2 refer to the variance of model 1 and model 2 respectively and r is the correlation between AUC_1 and AUC_2 .[6] We use the R package pROC by Robin et al. [2011] to calculate the ROC curve, the AUC, the t-statistic and p-value.

3.5 Forecast Results in United States

We conduct in-sample and out-sample forecasts of recessions in the United States. We forecast using 3 types of models at the 3 month, 6 month and 12 month horizon: best logit, boosting a large dataset, boosting a small dataset. The large dataset is the 130 predictor set used by Stock and Watson [2005] and the small dataset is the Conference Boarding Leading Indicators. Additional information about both datasets can be found in the appendix. Furthermore, to emulate the setup used by Ng [2014], we allow for dynamics in the large dataset by allowing for lags, specifically 3 lags for the 3 month horizon, 3 lags for the 6 month horizon, and 4 lags for the 12 month horizon. Thus, boosting the large dataset at the 12 month horizon will select from 520 predictors as 130 predictors $\times$ 4 lags $=$ 520 predictors. In forecating the small dataset, we do not include lags as we aim to have a simpler model. In instances when we do add lags, the model tends to perform

[6] For more on ROC and AUC: Liu and Moench [2014]

worse out-of-sample but better in-sample because of overfitting. We find the best logit model at each horizon by systematically going through all 130 predictors from the large dataset.

We recreate the work done in Ng [2014] but extend the analysis by calculating the AUC, a measurement of a model's classification ability, of boosting to determine how much better or worse boosting does relative to the best logit model. Furthermore, we boost a small dataset to evaluate performance against the the best logit model. Our focus is understanding why boosting large predictor sets underperforms or outperforms the best logit model. All results relating to variable selection are included in the appendix.

3.5.1 In-Sample Results

Model Set Up

We perform full in-sample forecast of U.S. recessions from 1959-04-01 to 2014-09-01 and use the entire dataset to estimate our models. We follow the setup used in Ng [2014] for in-sample boosting by using cross-validation to determine the number of iterations M for boosting and set the step size ρ as 0.01.

Forecasting Performance

The forecasting results for the 3, 6 and 12 months horizons can be found in Table 3.4.

Table 3.4: U.S. Forecast Performance: In-Sample.[7]

Model		3 Months Ahead	6 Months Ahead	12 Months Ahead
Best Logit	AUC	0.840	0.849	0.866
	Variable	3 month - FF spread	5 year - FF spread	10 year - FF spread
Boosting	AUC	0.971	0.965	0.954
	T-test 1	-5.34***	-5.08***	-4.59***
	Top Var.	1 Year - FF Spread	1 Year -FF Spread	5 year- FF Spread
	Dataset	Large Dataset	Large Dataset	Large Dataset
Boosting	AUC	0.947	0.943	0.902
	T-test 2	-4.12***	-3.94***	-1.49*
	Top Var.	NAPM new ordrs	5 year - FF Spread	5 year - FF Spread
	Dataset	Small Dataset	Small Dataset	Small Dataset

We find that boosting with the large dataset has the highest in-sample AUC across all horizons, indicating that boosting the large dataset has the strongest classification ability. Furthermore, boosting the large dataset does better than the best logit model across all horizons at the 1% level as indicated by T-test 1 in Table 3.4. Furthermore, boosting the large dataset does better than boosting the small dataset across all horizons. We also note that boosting the small dataset outperforms the best logit model significantly at the 1%, 1%, and 10% level for the 3 month, 6 month and 12 month horizon respectively as indicated by T-test 2 in Table 3.4.

We hestitate to draw any broad generalizations from our in-sample performance as strong in-sample performance does not necessarily mean strong out-of-sample performance. We ultimately care about the out-of-sample performance as out-of-sample forecasting is more applicable and useful for forecasting recessions in practice.

[7] Full in-sample forecasts are from 1959-04-01:2014-09-01. Small dataset consists of the Conference Board Leading Indicators. Large dataset refers to the 132 predictors from Stock and Watson [2005]. For T-Test 1, $H_0 : AUC_{logit} = AUC_{boosting-large}$ $H_a : AUC_{logit} < AUC_{boosting-large}$. For T-Test 2, $H_0 : AUC_{logit} = AUC_{boosting-small}$, $H_a : AUC_{logit} < AUC_{boosting-small}$. ***,**, and * represent significance at the 1,5, and 10% confidence levels respectively. Top var. stands for top variable with highest relative importance I_k^2. Variable from the best logit model also included.

3.5.2 Out-Of-Sample Results

Model Set Up

We now turn our attention to out-of-sample forecast performance. For out-of-sample forecasts for boosting and our logit model, we use rolling window estimates to produce forecasts starting 1985-09-01 and ending in 2014-09-01. For example, for forecasting 12 months ahead, we use data from 1959-09-01 to 1985-08-01 to estimate our logit or boosting model and then forecast 13 months later (12 months + 1 months as our predictors are defined as x_{t-h-1} with the observation Y_t) and produce a recession probabiltity for 1986-09-01. Then we increment our window by 1 month and estimate our model from 1959-10-01 to 1985-09-01 and then make a forecast for 1986-10-01 and so on and so forth. An example of using rolling windows to forecast out-of-sample in the United States can be found in Table 3.5. We make $348-h$ rolling forecasts for horizon h and with a fixed rolling window size of 311 months. To find the best out-of-sample logit model, we systematically go through all 130 predictors from the large dataset and perform rolling window forecasts and select the variable that yields the highest AUC to be our best logit model for that time horizon.

Table 3.5: Rolling Window Out-Of-Sample Forecasts in United States: 12 months[8]

Rolling Subsample	Rolling Subsample Window	Forecast Period
1	1959-09-01 to 1985-08-01	1986-09-01
2	1959-10-01 to 1985-09-01	1986-10-01
3	1959-11-01 to 1985-10-01	1986-11-01
...	..	...
334	1987-07-01 to 2013-06-01	2014-07-01
335	1987-08-01 to 2013-07-01	2014-08-01
336	1987-09-01 to 2013-08-01	2014-09-01

[8] We illustrate forecasting 12 months ahead in the United States. When we forecast y_t, we observe x_{t-h-1} where t is period of observation and h is the forecast horizon. Thus, when forecasting 12 months ahead, we in practice forecast 13 months ahead.

Following the convention used by Ng [2014] , we use $\bar{I}_k^2$ or the average importance of each of the variables across all the forecasts. Continuing the example used in Table 3.5, for forecasting 12 months in advance, we forecast results from 1986-09-01 to 2014-09-01 which consists of 336 months. Thus, the average relative importance of each variable over the 336 rolling subsamples is

$$\bar{I}_k^2 = \sum_{t=t_1}^{T} \frac{1}{336} I_{k,t} \tag{3.11}$$

where $I_{k,t}$ is the relative importance at time t of predictor k and t_1 in our example would be first rolling estimation which is 1986-09-01 and T would be our last period of the last forecast so for our example 2014-09-01. We also calculate the average frequency of variable j being selected in the rolling estimation defined as the following for 336 rolling sub-samples:

$$freq_k = \frac{1}{336} \sum_{t=t_1}^{T} 1(I_{k,t}^2 > 0) \tag{3.12}$$

To select the number of iterations M for boosting at each rolling sub-sample, we use 5 fold cross-validation on the first rolling sub-sample and use the optimal number of iterations returned by cross-validation for the rest of the rolling sub-samples. Hence, we use cross-validation once to determine M for the entirety of the rolling forecasts. While we could repeatedly use cross validation to find the number of iterations for each of rolling subsamples, doing so we found computationally unreasonable. We find the average optimal number of iterations used is about 400 iterations.

Forecasting Performance

We first turn our attention to the forecasting performances of each model before investigating the variables selected by boosting. The in-sample and out-of-sample forecasting results for the the forecast horizons 3, 6 and 12 months can be found in Figure 3.3, Figure 3.4, and Figure 3.5. Out-of-sample forecasting performance for horizons 3, 6 and 12 months with T-tests comparing the AUC of the boosting models against the AUC of the best logit model can be found in Table 3.6.

Table 3.6: U.S. Forecast Performance: Out-Of-Sample[9]

Model		3 Months Ahead	6 Months Ahead	12 Months Ahead
Best Logit	AUC	0.842	0.695	0.879
	Variable	NAPM new orders	NAPM new orders	5 year - FF spread
Boosting	AUC	0.763	0.545	0.697
	T-test 1	2.16**	1.68*	3.23***
	Top Var.	AAA- FF Spread	AAA- FF Spread	10year - FF spread
	Dataset	Large Dataset	Large Dataset	Large Dataset
Boosting	AUC	0.755	0.693	0.826
	T-test 2	2.55***	0.029	2.15**
	Top Var.	5 year - FF Spread	5 year - FF Spread	5 year - FF Spread
	Dataset	Small Dataset	Small Dataset	Small Dataset

Surprisingly, the best logit model outperforms both boosting the large and small dataset across all forecast horizons. The best logit model has a higher AUC than the AUC from boosting the large dataset significant at 5%, 10% and 1% for the 3 month, 6 month and 12 month horizon respectively. The best logit model has a higher AUC score than boosting the small dataset significant at 10% and 5% for the 3 month and 12 month horizon respectively. The best logit model did not perform significantly better than boosting the small dataset at the 6 month horizon. Another surprising result was that boosting the large dataset performs the worse of all three methods and even performs worse than boosting a smaller predictor set except for the 3 months horizon. We discuss why we think boosting the large dataset at the 3 month horizon outperforms boosting the small dataset in our discussion about variable selection in the appendix.

[9] Rolling windows begins 1959-04-01:1985-08-01 to forecast out-of-sample 1985-09-01:2014-06-01. Large dataset refers to the 132 predictor from Stock and Watson [2005]. Small dataset consists of the Conference Board Leading Indicators. For T-Test 1, $H_0 : AUC_{logit} = AUC_{boosting-large}$ $H_a : AUC_{logit} > AUC_{boosting-large}$. For T-Test 2, $H_0 : AUC_{logit} = AUC_{boosting-small}$, $H_a : AUC_{logit} > AUC_{boosting-small}$. ***,**, and * represent significance at the 1,5, and 10% confidence levels respectively. Top var. stands for top variable with highest average relative importance $\bar{I}_k^2$. Variable from the best logit model also included.

Figure 3.3: U.S. Forecasting Recession Performance 3 Months In Advance[10]

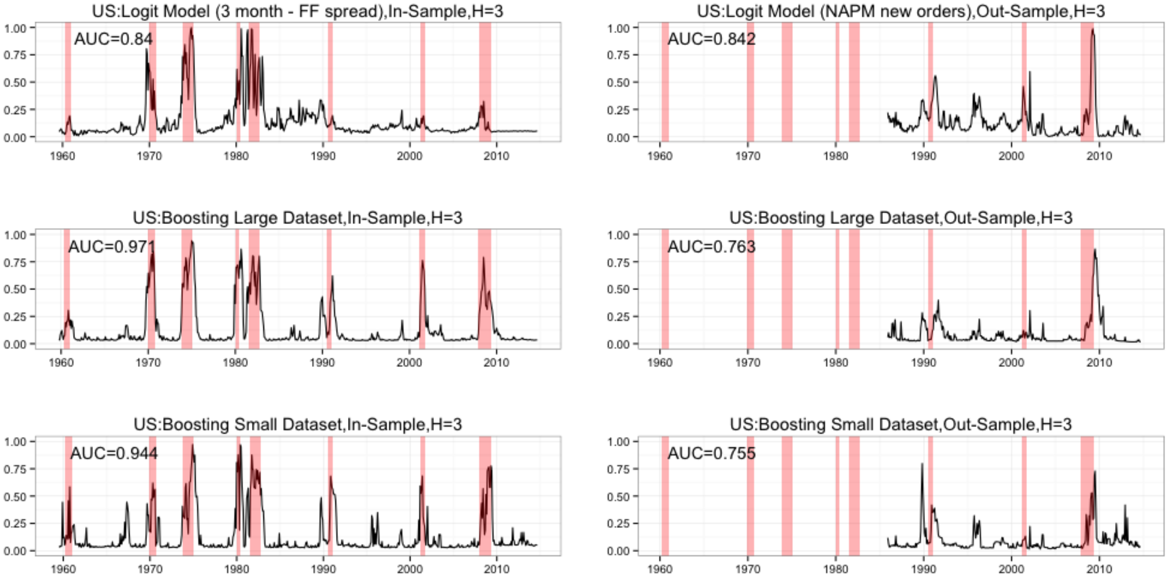
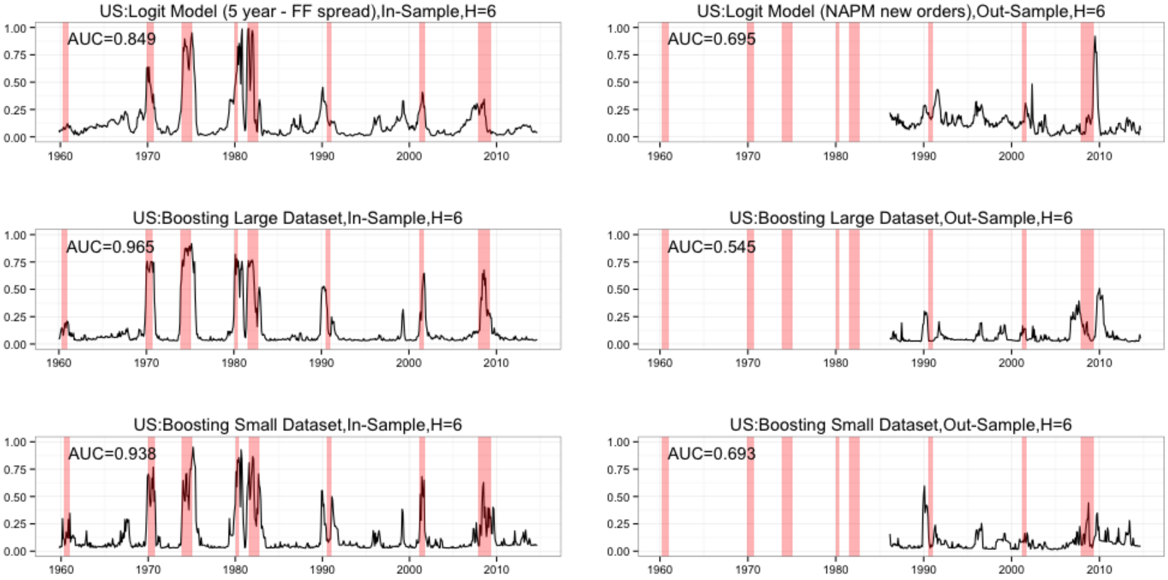


Figure 3.4: U.S. Forecasting Recession Performance 6 Months In Advance[11]



[10] The left column displays in-sample forecasting performance in the U.S. at the 3 month horizon from 1959-04-01 to 2014-09-01 of the best logit model, boosting model with the large dataset (130 predictors), boosting model with small dataset (10 predictors), in that respective order. The right column mirrors the left column, but displays out-of-sample performance of forecasting recessions 3 months in advance from 1985-12-01 to 2014-09-01. Shaded red bars indicate recessions and black lines indicate predicted probability of recession. AUC is a measurement of model classification ability.

[11] The left column displays in-sample forecasting performance in the U.S. at the 6 month horizon from 1959-04-01 to

Figure 3.5: U.S. Forecasting Recession Performance 12 Months In Advance[12]

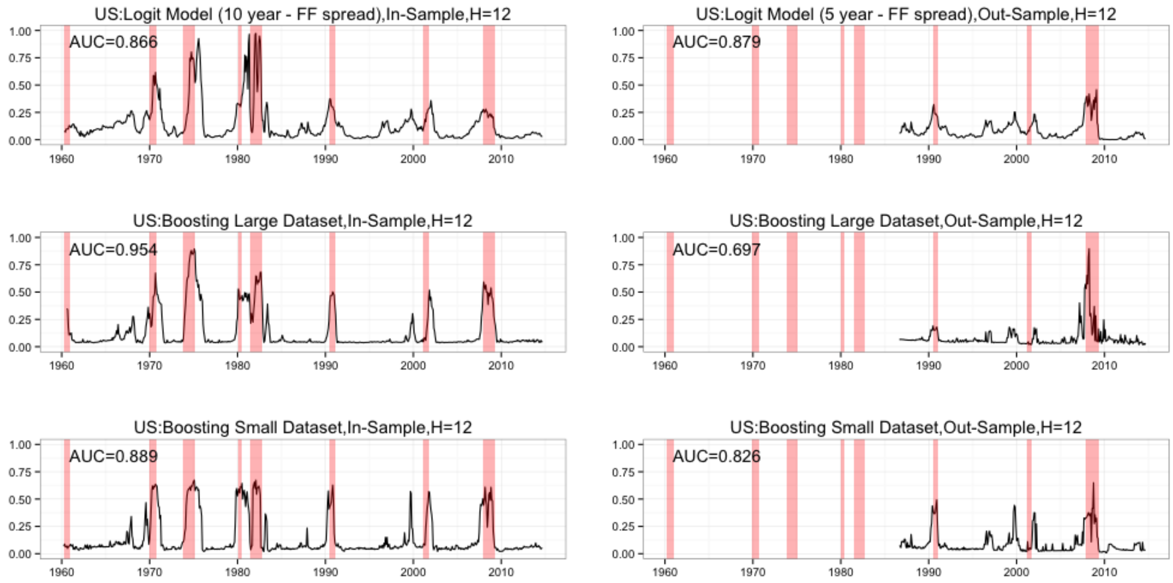


Figure 3.3, Figure 3.4 and Figure 3.5 illustrate how boosting the large dataset seems to perfectly forecast the state of the economy in-sample but out-of-sample predictions for the same time horizon and same dataset performs significantly worse. For example, boosting the large dataset in-sample 3 months ahead almost perfectly predicts the three most recent recessions. However, out-of-sample boosting the large dataset at the 3 month horizon seems to predict the three most recent recessions too late. The out-of-sample logit model seems to perform much better as there are noticeable signals in forecasting the 2001 and 2008-2009 recessions, although suffers from the same problem as boosting the large dataset and forecasting the 1990-1991 recession too late. Similarly in-sample boosting of the large dataset at the 12 month horizon perfectly predicts all three recent recessions whereas out-of-sample, the 2001 recession is predicted too early and too late and

2014-09-01 of the best logit model, boosting model with the large dataset (130 predictors), boosting model with small dataset (10 predictors), in that respective order. The right column mirrors the left column, but displays out-of-sample performance of forecasting recessions 6 months in advance from 1986-03-01 to 2014-09-01. Shaded red bars indicate recessions and black lines indicate predicted probability of recession. AUC is a measurement of model classification ability.

[12] The left column displays in-sample forecasting performance in the U.S. at the 12 month horizon from 1959-04-01 to 2014-09-01 of the best logit model, boosting model with the large dataset (130 predictors), boosting model with small dataset (10 predictors), in that respective order. The right column mirrors the left column, but displays out-of-sample performance of forecasting recessions 12 months in advance from 1986-09-01 to 2014-09-01. Shaded red bars indicate recessions and black lines indicate predicted probability of recession. AUC is a measurement of model classification ability.

the Great Recession is predicted too early.

We found surprising that at the 12 month horizon, the best logit model out-of-sample (AUC 0.879) actually performed better than the best in-sample logit model (0.866), leading us to wonder why boosting out-of-sample did not beat boosting in-sample forecasts at the 12 month horizon.

So far, our finding broadly suggests that boosting may be working too hard and overfitting in-sample or on the training model and thus predicting poorly out-of-sample, whereas models that are more parsimonious like boosting a smaller dataset or predicting using a single variable is able to forecast fairly well. Our finding also suggests that the approach by Ng [2014] when benchmarked with a logit model can be greatly improved. For instance in forecasting 12 months ahead, by reducing the predictor set from 520 (130 predictors x 4 lags) to 20 variables in the small dataset, we were able to get much stronger predictions though not better than the best logit model.

3.6 Forecast Results in Japan

While we have found that boosting a large dataset does not lead to superior forecasting of recessions in the United States for the 3, 6 and 12 month horizon, we do not know if the same applies to outside of the United States. Thus, we extend our analysis to Japan and explore boosting's performance on a large dataset of Japanese macroeconomic variables similar to the dataset used by Stock and Watson [2006] and a small dataset that includes leading indicators used by Japan's cabinet office. More specifics about both datasets can be found in the appendix. We conduct in-sample and out-sample forecasting in Japan, evaluating how well boosting with a large dataset with 436 predictors and a small dataset with 26 predictors does compared to the best logit model. We do not include lags in our large or small dataset like we did for the large dataset in the United States as we find including lags lead to inferior performance in general. We find the best logit model at each horizon by systematically going through all 436 predictors. All results relating to variable selection are included in the appendix.

3.6.1 In-Sample Results

Model Set Up

We perform full in-sample forecast of recessions in Japan from 1979-01-01 to 2014-06-01 and use the entire dataset to train our boosting models as well as to forecast. We use the same model specifications used for in-sample boosting in the United States for Japan.

Forecast Performance

In-sample performance of boosting the small and large datasets significantly outperforms the best logit logit model at the 1% level across all horizons as shown in Table 3.7. We find an almost identical pattern in Japan and in the United States, finding that boosting the large dataset significantly outperforms the best logit model and boosting the small dataset. We also find that boosting the smaller dataset also outperforms the best logit model at each horizon. Boosting the large datasets in-sample have AUC of at least 0.95 in all 3 horizons, which makes us suspicious that the model is overfitting. Out of curiosity, we set M the number of iterations in boosting to a really high number of iterations and find that the AUC of boosting the large dataset approaches 1.0 or perfect classification ability. This experiment against highlights boosting's ability to overfit in-sample. Conversely, when we set M the number of iterations very low, we get a significantly diminished AUC. We remind the reader that we use cross-validation to select the optimal number of iteration M that balances between maximizing AUC in-sample and out-of-sample.

Table 3.7: Japan Forecast Performance: In-Sample[13]

Model		3 Months Ahead	6 Months Ahead	12 Months Ahead
Best Logit	AUC	0.811	0.803	0.736
	Variable	Business Cond. Large Enterprise Electric Machine	Business Cond. Medium Enterprise Electric Machine	Lending Attitude Small Enterprise Manuf,Petrol,Coal
Boosting	AUC	0.967	0.9666	0.9577
	T-test 1	-6.64***	-6.94***	-6.14***
	Top Var.	Financial Position All Enterprise Transp. Machine	Business Cond. Medium Enterprise Electric Machine	Business Cond. Medium Enterprise Electric Machine
	Dataset	Large Dataset	Large Dataset	Large Dataset
Boosting	AUC	0.904	0.9038	0.9307
	T-test 2	-3.24***	-3.84***	-4.83***
	Top Var.	Business Cond. All Enterprises All industries	Business Cond. All Enterprises All industries	Inventory Level All Enterprises Manuf.
	Dataset	Small Dataset	Small Dataset	Small Dataset

[13] Full in-sample forecasts 1978-01-01:2014-06-01. Small dataset refers to 20 leading indicators from Japan's Cabinet Office and business surveys. Large dataset refers to 436 business surveys and macrovariables. T-Test compares the AUC of boosting to the AUC of the best logit model. For T-Test 1, $H_0 : AUC_{logit} = AUC_{boosting-large}$ $H_a : AUC_{logit} < AUC_{boosting-large}$. For T-Test 2, $H_0 : AUC_{logit} = AUC_{boosting-small}$, $H_a : AUC_{logit} < AUC_{boosting-small}$. ***,**, and * represent significance at the 1,5, and 10% confidence levels respectively. Top var. stands for top variable with highest relative importance I_k^2. We also include the variable from the best logit model.

3.6.2 Out-Of-Sample Results

Model Set Up

We continue our analysis to determine whether or not boosting a large dataset will improve recession forecast performance against the benchmark. We use rolling window estimators (method detailed in out-of-sample results for the United States) starting with 1978-02-01 to 1995-08-01 as our initial window and 1995-09-01 + h months as our initial forecast period and 2014-06-01 as our last forecast period. We make $225 - h$ months of forecasts for horizon h with rolling window size of 210 months. To find the best out-of-sample logit model, we systematically go through all 436 predictors from the large dataset and perform rolling window forecasts and select the variable that yields the highest AUC to be our best logit model.

Forecast Performance

In-sample and out-of-sample forecasts of the best logit model and boosting models for Japan are shown in Figure 3.6, Figure 3.7 and Figure 3.8 for the 3 month, 6 month and 12 month horizon respectively. Out-of-sample performance with T-tests comparing boosting models with the best logit models can be found in Table 3.8.

Table 3.8: Japan Forecast Performance: Out-Of-Sample[14]

Model		3 Months Ahead	6 Months Ahead	12 Months Ahead
Best Logit	AUC	0.848	0.8055	0.660
	Variable	Business Cond.	Business Cond.	Inventory Level
		Large Enterprises	Med. Enterprises	Med. Enterprises
		Manuf. Chemicals	Elect. Machine	Elect. Machine.
Boosting	AUC	0.8309	0.678	0.439
	T-test 1	0.970	4.05***	3.86***
	Top Var.	Financial Position	Inventory Level	Lending Attitude
		All Enterprises	All Enterprises	Small Enterprises
		Transp. Machine	Manuf	Manuf,Petrol,Coal
	Dataset	Large Dataset	Large Dataset	Large Dataset
Boosting	AUC	0.788	0.733	0.664
	T-test 2	2.33***	2.40***	-0.0845
	Top Var.	Financial Position	Inventory Level	Inventory Level
		All Enterprises	All Enterprise	All Enterprises
		Manuf.	Manuf.	All industries
	Dataset	Small Dataset	Small Dataset	Small Dataset

[14] Rolling windows begin 1975-01-01:1995-08-01 to forecast out-of-sample 1995-09-01:2014-06-01. Small dataset refers to 20 leading variables from Japan's Cabinet Office and business surveys. Large dataset refers to 436 business surveys and macrovariables. T-Test compares the AUC of boosting to the AUC of the best logit model. For T-Test 1, $H_0 : AUC_{logit} = AUC_{boosting-large}$ $H_a : AUC_{logit} > AUC_{boosting-large}$. For T-Test 2, $H_0 : AUC_{logit} = AUC_{boosting-small}$, $H_a : AUC_{logit} > AUC_{boosting-small}$. ***,**, and * represent significance at the 1,5, and 10% confidence levels respectively. Top var. stands for top variable with highest average relative importance $\bar{I}_k^2$. We also include the variable from the best logit model.

Figure 3.6: Japan Forecasting Recession Performance 3 Months In Advance[15]

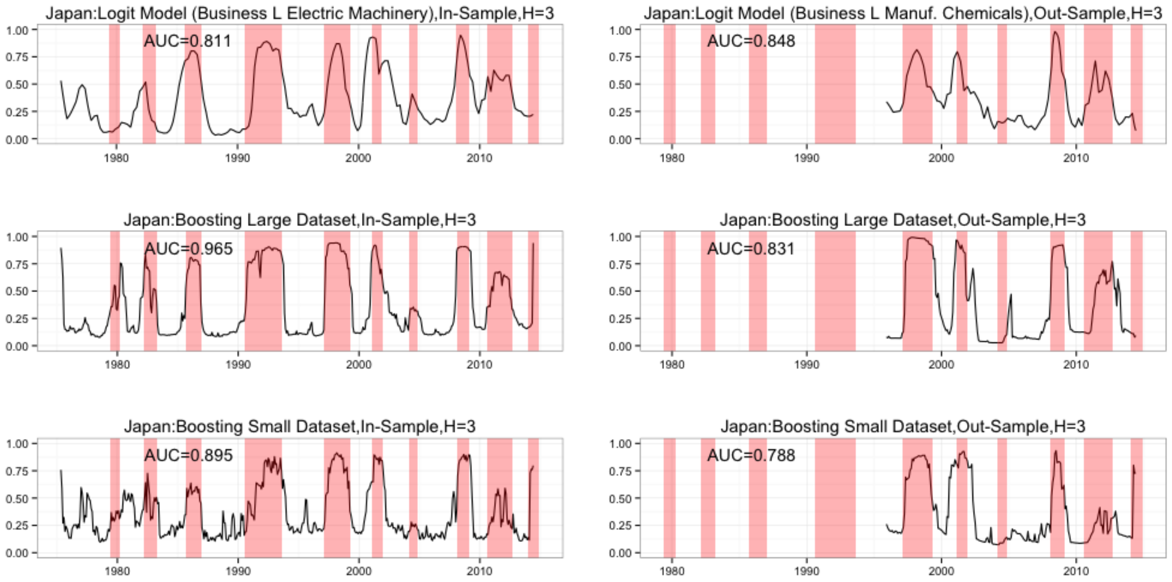
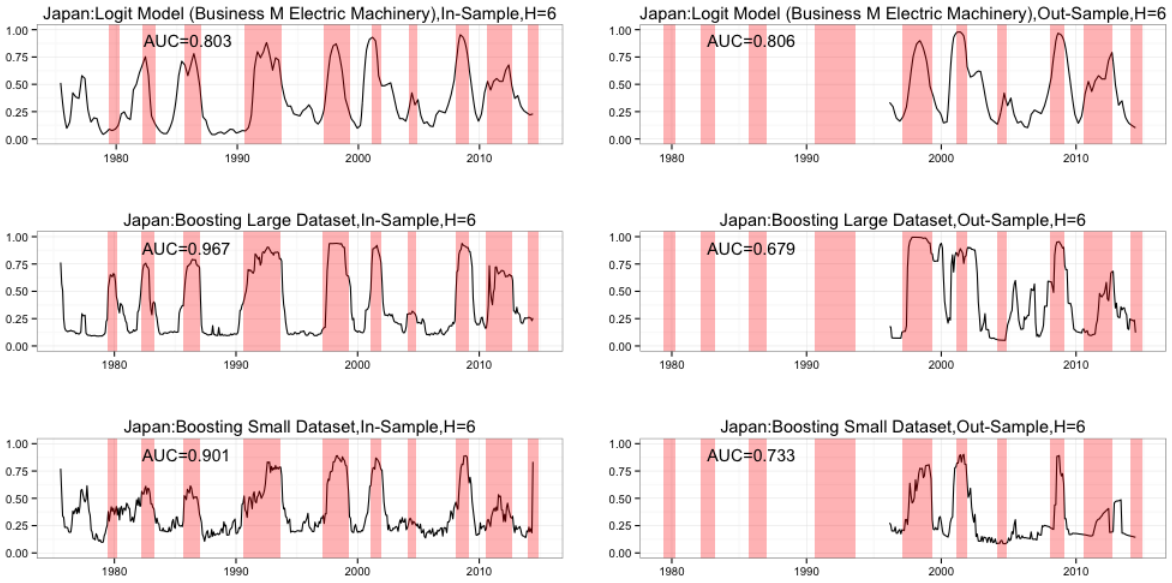


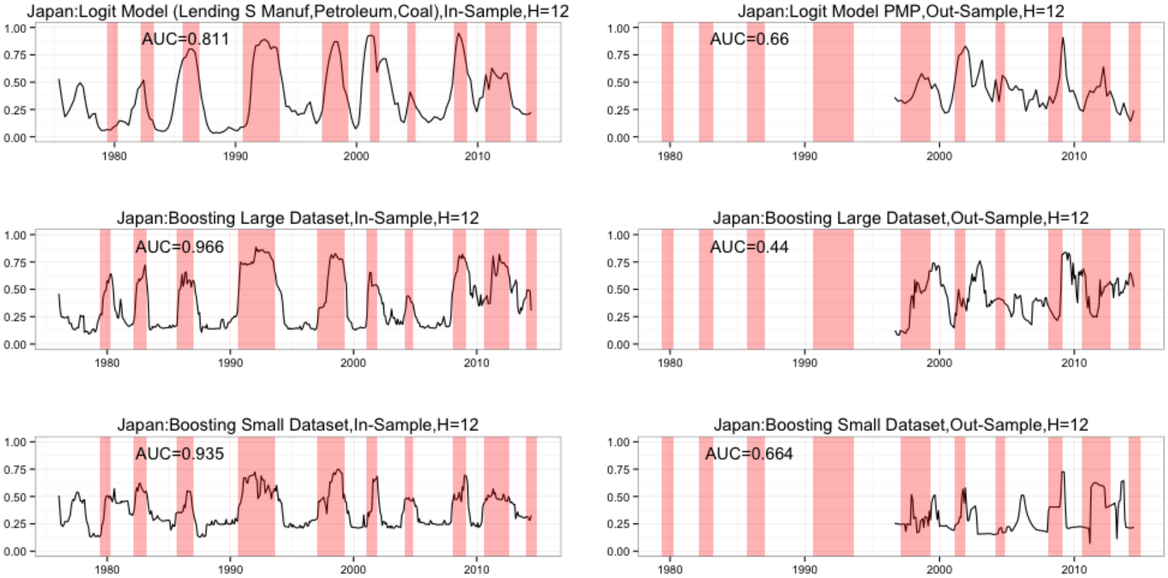
Figure 3.7: Japan Forecasting Recession Performance 6 Months In Advance[16]



[15] The left column displays in-sample forecasting performance in Japan at the 3 month horizon from 1975-01-01 to 2014-06-01 of the best logit model, boosting model with the large dataset (436 predictors), boosting model with small dataset (27 predictors), in that respective order. The right column mirrors the left column, but displays out-of-sample performance of forecasting recessions 3 months in advance from 1995-12-01 to 2014-06-01. Shaded red bars indicate recessions and black lines indicate predicted probability of recession. AUC is a measurement of model classification ability.

[16] The left column displays in-sample forecasting performance in Japan at the 6 month horizon from 1975-01-01 to 2014-06-01 of the best logit model, boosting model with the large dataset (436 predictors), boosting model with small

Figure 3.8: Japan Forecasting Recession Performance 12 Months In Advance[17]



We find that the best logit model outperforms boosting the large dataset across all horizons and significantly better at 1% level for both the 6 month and 12 month horizon. The AUC of the best logit model is not significantly greater than the AUC of boosting the large dataset at the 3 month horizon.

We also find that the best logit model outperforms boosting the small dataset at the 1% significance level at the 3 month and 6 month horizon. Boosting the small dataset actually slightly outperforms the best logit model at the 12 month horizon. We also find that boosting the small dataset outperforms boosting with the large datset out-of-sample at the 6 month and 12 month horizon but not the 3 month horizon. We look to investigate in the next section why boosting the small dataset actually outperforms the best logit model at the 12 month horizon.

dataset (27 predictors), in that respective order. The right column mirrors the left column, but displays out-of-sample performance of forecasting recessions 6 months in advance from 1996-03-01 to 2014-06-01. Shaded red bars indicate recessions and black lines indicate predicted probability of recession. AUC is a measurement of model classification ability.

[17] The left column displays in-sample forecasting performance in Japan at the 12 month horizon from 1975-01-01 to 2014-06-01 of the best logit model, boosting model with the large dataset (436 predictors), boosting model with small dataset (27 predictors), in that respective order. The right column mirrors the left column, but displays out-of-sample performance of forecasting recessions 12 months in advance from 1996-09-01 to 2014-06-01. Shaded red bars indicate recessions and black lines indicate predicted probability of recession. AUC is a measurement of model classification ability.

Boosting the small dataset having a higher AUC than the best logit model at the 12 month horizon giving us hope that boosting in a smart or well thought out manner may outperform the best current single variable models. However, we must note that this gain is relatively small and is not significant even at the 10% level. Furthermore, Figure 3.8 shows that boosting the small dataset while having a higher AUC at the 12 month horizon misses the 2004 recession competely whereas the logit model at the 12 month horizon provides some sort of signal in 2004. Also, the boosted small dataset at the 12 month horizon mistakens an expansionary period as a recession between 2004 and 2008. Though boosting the smaller dataset has a small AUC advantage over the best logit model at the 12 month horizon, the boosting model is still imperfect and not much better than the best logit model.

We also consistently see a sharp decline in AUC from in-sample boosting model performance to out-of-sample boosting performance, suggesting overfitting in-sample like in the case of the United States. In Figure 3.8, boosting the large dataset in-sample produces noticeably high and distinct warnings for each of the 6 most recent recessions with an AUC of 0.966, whereas boosting the large dataset out-of-sample misses recessions, produces many false positives, and actually negatively predicts recessions with an AUC of 0.44. We can see that during the recession just after 2010, the probability of recession actually decreased. We see a similar pattern at the 3 month and 6 month horizon in Figure 3.6 and Figure 3.7 where the in-sample fit is nearly perfect for boosting but out-of-sample the performance declines.

Our finding in Japan supports our hypothesis that boosting overfits in-sample or in the training data and thus performs worse out-of-sample. Furthermore, we find that the best logit model outperforms the best boosting model for the most part, and that exceptions to this rule include boosting models that do not perform significantly better than best logit model. We also are able to find a smaller dataset that is a subset of the large dataset and find that this carefully crafted dataset significantly outperforms boosting the large dataset in the same time horizon. Our finding once again illustrates the parsimony principle that models with fewer variables can outperform models that involve many predictors, some that may not be predictive.

3.7 Conclusion

The main purpose of our paper was to evaluate if combining novel machine learning techniques such as boosting with large datasets could serve as a better alternative forecasting method than the single variable models that are popular in forecasting recessions. We have found that boosting large datasets cannot significantly beat the best single variable models. We find that the approach taken by Ng [2014] in using boosting to variable select and predict using a large dataset leads to poor forecasting performance in the United States relative to the best logit model. Furthermore, we compile a dataset similar to the one used in Stock and Watson [2005] for Japan and find that boosting a large dataset in Japan leads to inferior performance versus the best logit models. Our key contribution is finding that both in the United States and Japan, the best logit model mostly outperforms both boosting a large dataset and boosting a smaller dataset with just the leading indicators.

Furthermore, we find that boosting on smaller datasets give us gains in classification ability though not improving upon the best logit model except for one case. Our explanation is that the parsimony principle holds true when forecasting recessions, that often times a single variable predicts better than a group of predictors as incorporating large amounts of predictors may lead to overfitting the training model and poor out-of-sample performance. Thus using a kitchen sink approach and loading up a boosting model will not necessarily lead to superior forecasting performance. Taking the approach of Berge [2014] and carefully selecting only leading indicators to boost may lead to substantial gains and possible improvements over the best logit models.

While we find that boosting does not forecast recessionary binary variable better than the benchmark, Buchen and Wohlrabe [2011] finds that boosting is a strong competitor to forecasting gold standards such as dynamic factor models in forecasting quantiative variables like U.S. industrial production. Furthermore, Wohlrabe and Buchen [2014] find that boosting generally outperforms the autoregressive benchmarks when forecasting macroeconomic variables. Perhaps boosting does poorly in forecasting recessions because of the small number of observations of distinct recessions. While we have a relatively large time series set with 665 months in the U.S. and 427 months in Japan, the number of recessions in those months are 8 and 11 for the U.S. and Japan respectively.

General machine learning classification problems such as spam filtering usually have thousands of different spam and not spam emails to train on, whereas 8 to 11 cases for boosting to learn about recessions is very small and not likely sufficient to train our model. The limited amount of data we have on recessions likely limits our ability to forecast recessions well. Further work can be done to simulate recessions to extend our time series and then seeing at which point boosting outperforms the best logit model, if boosting ever does.

Bibliography

Jushan Bai and Serena Ng. Boosting diffusion indices. *Journal of Applied Econometrics*, 24:607–629, 2009. ISSN 08837252. doi: 10.1002/jae.1063.

Travis Berge. Predicting Recessions with Leading Indicators : Model Averaging and Selection Over the Business Cycle. (April 2013), 2014.

Henri Bernard and Stefan Gerlach. Does the term structure predict recessions? The international evidence. *International Journal of Finance & Economics*, 3:195–215, 1998. ISSN 1076-9307. doi: 10.1002/(SICI)1099-1158(199807)3:3<195::AID-IJFE81>3.0.CO;2-M.

Gerhard Bry and Charlotte Boschman. Cyclical Analysis of Time Series : Selected Procedures and Computer Programs. *NBER Technical Paper*, 20:13, 1971.

Teresa Buchen and Klaus Wohlrabe. Forecasting with many predictors: Is boosting a viable alternative? *Economics Letters*, 113(1):16–18, oct 2011. ISSN 01651765. doi: 10.1016/j.econlet.2011.05.040. URL http://www.sciencedirect.com/science/article/pii/S0165176511002175.

Arthur F. Burns and Wesley C. Mitchell. *Measuring business cycles*, volume I. 1946. ISBN 087014085X. URL http://econpapers.repec.org/bookchap/nbrnberbk/burn46-1.htm.

Charlotte Christiansen, Jonas Nygaard Eriksen, and Stig Vinther Møller. Forecasting US recessions: The role of sentiment, 2013. ISSN 03784266.

Arturo Estrella and Gikas A Hardouvelis. The Term Structure as a Predictor of Real Economic

Activity. *The Journal of Finance*, 46:555–576, 1991. ISSN 0022-1082. doi: 10.1111/j.1540-6261.1991.tb02674.x. URL http://www.jstor.org/stable/2328836.

Arturo Estrella and Frederic S. Mishkin. Predicting U.S. Recessions: Financial Variables as Leading Indicators, 1998. ISSN 0034-6535.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting, 2000. ISSN 00905364.

J.H. Friedman. GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE. *Statistics*, 29(5):1189–1232, 2001.

Claudia Godbout and Marco J. Lombardi. Short-Term Forecasting of the Japanese Economy Using Factor Models. 2012. URL http://ideas.repec.org/p/bca/bocawp/12-7.html.

J A Hanley and B J McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983. ISSN 0033-8419. doi: 10.1148/radiology.148.3.6878708.

Masashi Hasegawa and Yuichi Fukuta. An empirical analysis of information in the yield spread on future recessions in Japan. *Applied Economics*, 43(15):1865–1881, 2011. ISSN 0003-6846. doi: 10.1080/00036840902780136.

Hideaki Hirata and Kazuo Ueda. The Yield Spread as a Predictor of Japanese Recessions. 1998.

William C. Horrace and Ronald L. Oaxaca. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327, 2006. ISSN 01651765. doi: 10.1016/j.econlet.2005.08.024.

Heikki Kauppi and Pentti Saikkonen. Predicting U.S. Recessions with Dynamic Binary Response Models. *The Review of Economics and Statistics*, 90:777–791, 2008. ISSN 0034-6535. doi: 10.1162/rest.90.4.777. URL http://ideas.repec.org/a/tpr/restat/v90y2008i4p777-791.html.

Gad Levanon, Jean-Claude Manini, Ataman Ozyildirim, Brian Schaitkin, and Jennelyn Tanchua. Using a Leading Credit Index to Predict Turning Points in the U.S. Business Cycle. *Conference Board*, 11(05), 2011.

Weiling Liu and Emanuel Moench. What Predicts US Recessions? *Federal Reserve Bank of New York Staff Reports*, 691(September), 2014.

Serena Ng. Boosting recessions. *Canadian Journal of Economics*, 47(1):1–34, 2014. ISSN 15405982. doi: 10.1111/caje.12070.

Serena Ng and Jonathan H. Wright. Facts and Challenges from the Great Recession for Forecasting and Macroeconomic Modeling. *Journal of Economic Literature*, 51:1120–1154, 2013. ISSN 0022-0515. doi: 10.1257/jel.51.4.1120.

Greg Ridgeway. Generalized Boosted Models : A guide to the gbm package. *Compute*, 1:1–12, 2007. ISSN 14679752. doi: 10.1111/j.1467-9752.1996.tb00390.x. URL http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12:77, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-77.

Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990. ISSN 08856125. doi: 10.1007/BF00116037.

Robert E. Schapire. A brief introduction to boosting. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2, pages 1401–1406, 1999. ISBN 3540440119. doi: citeulike-article-id:765005.

James H Stock and Mark W Watson. New Indexes of Coincident and Leading Economic Indicators.

NBER Macroeconomics Annual, 4:351–394, 1989. ISSN 08893365. doi: 10.2139/ssrn.227144. URL http://www.jstor.org/stable/3584985.

James H Stock and Mark W Watson. A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series. *Methods*, No. 6607:1–44, 1994. ISSN 03043932. doi: 10.2139/ssrn.226329. URL http://www.nber.org/papers/w6607.

James H Stock and Mark W Watson. Implications of Dynamic Factor Models for VAR Analysis. *NBER Working Paper Series*, 11467:1–67, 2005. doi: 10.2139/ssrn.755703.

James H. Stock and Mark W. Watson. Forecasting with Many Predictors. *Handbook of Economic Forecasting*, 1(05):515–554, 2006. ISSN 15740706. doi: 10.1016/S1574-0706(05)01010-4.

Klaus Wohlrabe and Teresa Buchen. Assessing the macroeconomic forecasting performance of boosting: Evidence for the United States, the Euro area and Germany. *Journal of Forecasting*, 33(May):231–242, 2014. ISSN 1099131X. doi: 10.1002/for.2293.

Jing Zeng. Forecasting Aggregates with Disaggregate Variables: Does boosting help to select the most informative predictors? *Conference Paper*, 2014.