# APPLICATIONS OF MACHINE LEARNING IN FORECASTING RECESSIONS: BOOSTING UNITED STATES AND JAPAN

Jonathan B. Ma
Advisor: Nobuhiro Kiyotaki

Submitted to Princeton University
Department of Economics
In Partial Fulfillment of the Requirements for the A.B. Degree

April 2015

## 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 parsimonious principle, that simpler models often outperform more complex models.

# Acknowledgements

Firstly, I want to thank my advisor Nobuhiro Kiyotaki for his mentorship, wisdom and willingess to meet with me so often. Thank you for the timely and thoughtful feedback and helping me find a topic I was passionate about. Thank you for being so understanding while setting a high standard of expectation. You are the best thesis advisor a student can ask for. All errors are my own.

Secondly, I want to thank Elizabeth Bogan for her introductory to macroeconomics course which motivated my interest in forecasting recessions and confirmed my desire to major in economics. I also want to thank Macro Del Negro, Chris Sims, Ulrich Muller, Hank Farber, Greg Kaplan, Uwe Reinhardt, Mark Aguiar, Will Dobbie and Alan Blinder for teaching and walking me through the fascinating world of economics and answering my numerous questions. Thank you to Alan Blinder for the helpful discussion and for being an incredible human being. Thank you to Han Liu and Barbara Engelhardt for introducing me to the world of machine learning and motivating the topic. Thank you Mark Watson for meeting with me in person multiple times and the thoughtful discussion and helpful guidance. Additionally I wanted to thank Serena Ng and Travis Berge for helpful discussion. A special thank you to Travis Berge for providing me his dataset and the tremendous amount he has given me.

Thank you to Dean Rebecca Graves-Bayazitoglu and Kristin Frasier for the treats and accomodation. Special shout out to Mary Kemler for the coffee.

Special thanks to my roommate Shafin Fattah for the encouraging words of advice, the late night Wawa runs, and proofreading my paper. You are like the big brother I never had. Thank you Luke Li for the helpful discussion.

Thank you to Andy Qin for the encouragement, proofreading the paper, and the helpful discussion. Thank you to Andy Lee for checking in, proofreading, and the helpful discussion. Thank you to Justin Lu for all the rides and reminding me what's important in life.

Last but not least, thank you mom, dad, Janice and Gina. These last few months were perhaps the toughest time of our lives. Thank you Gina for sticking by my side and being there for me. I am extremely lucky to call you my girlfriend. I dedicate my senior thesis to my mother. Mom, you are my role model and I love you.

To my mom:

# Pledge

This paper represents my own work in accordance with University regulations. I authorize Princeton University to lend or reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Jonathan B. Ma

# Contents

# List of Figures

# List of Tables

# 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 parsimonious 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–to a strong learner that can classify significantly better. For example, our initial weak learner could be the classification rule that the current state of an economy is in 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. as 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 parsimonious principle, that simpler is often better and that parsimony is especially important for forecasting recessions because even if a model incorporates even a slight amount of noise, the consequences could be inaccurate forecasts.

The paper is organized as follows. We first review the literature on predicting recessions in the United States and Japan as well as methods used for predicting recessions. We then detail the methods used in forecasting recessions in our paper, notably explaining how boosting works. Then we discuss the dataset used for both Japan and United States as well as our evaluation criteria for comparing different classification models. We then look at the in-sample and out-of-sample forecasting performance of boosting in the United States to evaluate boosting's performance. We then turn our attention to Japan and run a similar empirical anlaysis. We conclude with discussion and further work that can be done.

# 2 Related Work

## 2.1 Predicting U.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.

## 2.2 Predicting Japan Recessions

Bernard and Gerlach (1998) summarizes how well the term structure 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 pseudo $R^2$ 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. Majority of the forecasting literature in Japan has focused on the predictive power of 1 or a few variables, thus boosting a large dataset and understanding the predictive merits of such a model will be novel and a contribution to the literature.

## 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.

### 2.3.1 Logit and Probit

Probit or logit is the gold standard 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. Kauppi and Saikkonen (2008) extends the simple probit model to a dynamic probit model that incorporates lags of explanatory variables and recessionary dummies. 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.

### 2.3.2 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 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.

## 3.1 Logit

Logit assumes the log-odds ratio as a function of $x_t$

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

$$P(Y_t = 1|x_{t-h-1}) = \frac{exp(f(x_{t-h-1}, \theta)}{1 + exp(f(x_{t-h-1}, \theta)} \tag{2}$$

where $\theta$ 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 can calculate $\hat{\beta}$ and hence the probability predictions of $Y_t$ using maximum likelihood estimate or software such as R's glm package and specifying a bernoulli loss function and using logit as the link.

## 3.2 Boosting

In this section, we explain how AdaBoost, the algorithm that set off 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 Ng (2014); Berge (2014).

### 3.2.1 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) proves 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.

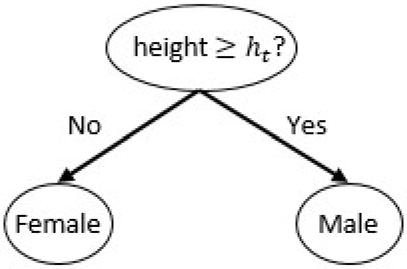
**Algorithm 1** Discrete AdaBoost Schapire (1999)

```
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 $sign(F_m(x))$
```

$M$ is the total number of iterations, $T$ is the total number of observations, $\epsilon$ 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 $\theta$ 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 1

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



Notes: 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.

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:

$$w_t^{(m+1)} = \frac{w_t^{(m)}}{Z_m} \begin{cases} exp(-\alpha_m) & y_t = f_m(x_t|\theta) \\ exp(\alpha_m) & y_t \neq f_m(x_t|\theta) \end{cases} \tag{3}$$

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 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 intitial 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

$$y_t = \begin{cases} 1 & HWI < -0.44 \\ -1 & HWI \geq -0.44 \end{cases} \tag{4}$$

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 ($\epsilon$) of 0.167 calculated according to 2(a) as the initial weak learner classifies 2 of the 12 months incorrectly so $2 * w_1 = 2 * 1/12 = 2/12 = 0.167$. 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:

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

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 $\alpha_m$ 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.

### 3.2.2 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:

$$
\begin{aligned}
E[y|x_1, x_2, ..., x_M] &= \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_M(x_M) & (5)\\
&= \beta_0 + \sum_{m=1}^{M} f_m(x_m) & (6)
\end{aligned}
$$

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:

$$F_M(x) = \sum_{m=1}^{M} \rho_m f_m(x, \theta_m) \tag{7}$$

Where

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

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

$$\hat{F}(x) = argmin_{F(x)} E[L(y, F(x))] \tag{8}$$

$$L(y, F(x)) = exp(-yF(x)) \tag{9}$$

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 (8) 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
        $u_m \equiv \{u_{m,t}\}_{t=1,..,T} = -\frac{\partial L(y_t,F)}{\partial F}|_{F=\hat{F}_{m-1}(x_t)}, t = 1, ..T.$
        Fit each weak learner $f^k$ for $k = 1, ..., K$ to the current negative gradient vector $u_m$
    (b) Let $\hat{f}_m^k$ be the best fit of $u_m$ among the $K$ weak learners.
    (c) 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 equation (9) 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 quantitative 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 information criteria and Berge (2014) uses Schwarz information criterion 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. $\rho$ 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

$$f^k(x_{t,k}) = c^L 1(x_{t,k} \leq \tau) + c^R 1(x_{t,k} > \tau) \tag{10}$$

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 $\tau$. $c^L$ 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

$$I_k^2 = \frac{1}{M}\sum_{m=1}^{M} i_m^2 1(id(x^m) = k) \tag{11}$$

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[1].

[1] For instance, Berge (2014) uses the R package mboost

# 4 Data and Evaluation

We use recent-vintage data which include the most up-to-date data that include revisions. A fundamental criticism of using recent-vintage data is that doing so is not realistic of forecasting in practice, as forecasters do not have the recent-vintage data but instead use real-time data that is subject to revision. Koenig et al. (2003) discusses that using recent-vintage data may exaggerate the forecasting performance relative to real-time forecasting. In light of this criticism, we proceed cautiously with our forecast models.

## 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. 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 1. The Conference Board Leading Indicators and the modifications made can be found in the appendix.

Table 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.

## 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.[2] The Organisation for Economic Co-operation and Development (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

[2] More on the computer routine can be found in Bry and Boschman (1971)

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.[3]

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 business survey indicators. Description of the variables from the small dataset can be found in Table 2. We include additional information about the small dataset and the transformations of the small dataset in the appendix.

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

Table 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 |

| 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

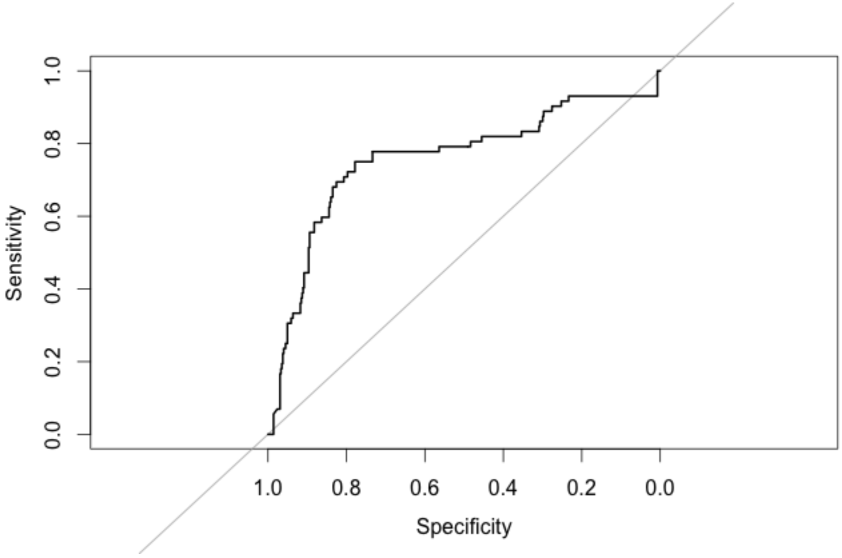
5 recessions during the same time period. Japan has been in a recession, as defined by the OECD, 41% of the time during this period. In our training set, 1975-01-01 to 1995-08-01, there are 5 recessions. In our testing set from 1995-08-01 to 2014-06-0, there are 6 recessions.[4]

## 4.3 Evaluation Criteria

A very popular evaluation metric from biostatistic, the Receiver Operating Characteristic has been making its way into economics. Liu and Moench (2014) and Berge (2014) use Receiver Operating Characteristic (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 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.

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

Figure 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):

$$t = \frac{AUC_1 - AUC_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2}} \tag{12}$$

$AUC_1$ and $AUC_2$ refers to the area under the curve for model 1 and 2 respectively and $\sigma_1^2$ and $\sigma_2^2$ refer to the variance of model 1 and model 2 respectively and $r$ is the correlation between $AUC_1$ and $AUC_2$ .[5] We use the R package pROC by Robin et al. (2011) to calculate the ROC curve, the AUC, the t-statistic and p-value.

[5] For more on ROC and AUC: Liu and Moench (2014)

# 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 x 4 lags = 520 predictors. In forecating the small dataset, we do not include lags as we aim to have a simpler model. 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 push the analysis further 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.

## 5.1 In-Sample Results

### 5.1.1 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 $\rho$ as 0.01.

### 5.1.2 Forecasting Performance

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

Table 3: U.S. Forecast Performance: In-Sample.

| 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 |

Notes: 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.

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. 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.

Out of curiosity, we set $M$ or the number of iterations for boosting to an extremely high number such as 5000 and boost the large and small dataset. We recover results with AUC for forecasting 3, 6 and 12 months ahead close to 1.0 or near perfect classification ability (results are omitted for concision). The result is unsuprising as Ng (2014) notes that $M$ is a stopping rule to prevent overfitting, and thus setting $M$ to an extremely high number may increase in-sample fit but not necessarily out-of-sample fit. 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.

### 5.1.3 Variable Selection

Variables selected from boosting using the large dataset are shown in Table 4. For the left most column, $(h, h+d)$ represents the forecast horizon and the number of lags and the right most column indicates the lag order in which the variable appears. For example, for horizon $(3, 6)$ the 1 year T-bond - Federal Funds Spread (subsequency referred to as 1 year - FF spread) has lags at horizon 4, 5 and 6. $I_k^2$ as discussed in the methods section describe the relative importance of variables. Thus, higher $I_k^2$ values mean higher importance of variable $k$ and $I_k^2$ value of 0 means that the variable $k$ was not selected by boosting in the $M$ iterations. We only include variables in Table 4 with $I_k^2 > 3$ to set a threshold of relative importance.

Table 4: U.S. Boosting Top Variables Chosen by Cross-Validation: Large Dataset, In-Sample

| $(h, h+d)$ | Variable Name | $I_k^2$ | $lag$ | | | |
|---|---|---|---|---|---|---|
| (3,6) | 1 Year T-bond - Federal Funds Spread | 31.79 | 4 | 5 | 6 | |
| | NAPM new orders | 8.82 | 4 | | | |
| | Emp. on Nonfarm Pay Goods Producing | 4.24 | 4 | | | |
| | 3 Month T-bill - Federal Funds Spread | 4.18 | 4 | | | |
| | AAA - Federal Funds Spread | 3.92 | | | 6 | |
| | Emp. On Nonfarm Pay Manufacturing | 3.76 | 4 | | | |
| | 6 month T-bill - Fedaral Funds Spread | 3.37 | 4 | | | |
| | Help Wanted/Employement Ratio | 3.02 | 4 | | | |
| (6.9) | 1 Year T-bond -Federal Funds Spread | 41.39 | 7 | 8 | 9 | |
| | AAA - Federal Funds Spread | 11.77 | | 8 | 9 | |
| | 5 Year T-bond - Federal Funds Spread | 5.42 | | | 9 | |
| | NAPM Commodity Prices Index | 3.49 | | 8 | | |
| | 5 year T-Bond - Federal Funds Spread | 3.04 | 7 | | | |
| (12,16) | 5 year T-Bond - Federal Funds Spread | 39.92 | 13 | 14 | 15 | 16 |
| | 10 year T-Bond - Federal Funds Spread | 9.19 | 13 | | | |
| | NAPM Inventories index | 8.07 | | 14 | 15 | |
| | Purchasing Manager's Index | 3.28 | | | | 16 |

NOTES: Forecasts for period t are based predictors at lag $t-h-1, t-h-2, ..., t-h-d$. The column $I_k^2$ is an indicator of importance. The last column $lag \in [h+1, h+d]$ denotes that lag at which the corresponding predictor is chosen. We only include $I_k^2 > 3$ or variables that are above a certain threshold.

We seek to understand why boosting performs so well in-sample for the large dataset relative to the best logit model. Variables with the highest relative importance from boosting the large dataset differed by horizon. For the 3 month horizon, variables about the real economy seem to be prevalent such as NAPM new orders, Employment on Nonfarm Payroll, and the Help Wanted Employment Ratio. For the 6 month and 12 month horizon, the most relevant variables are a mix of term and default spreads which matches the literature. As Ng (2014) notes in her analysis, whereas most of the literature uses just one spread in predicting, Table 4 shows how boosting is able to incorporate spread and predictors with different lags and of different combinations.

With this in mind, we hypothesize that in-sample boosting the large dataset performed better because of incorporating different types of term and default spreads and predictors whereas the best in-sample logit model only use 1 spread. For instance, the best logit model at the 3 month horizon in-sample used the 3 month - FF spread to predict whereas boosting the large dataset incorporated information about the real economy on top of different spreads. Whether or not incorporating different combinations of spreads improves out-of-sample forecasting is a question we are interested in investigating in the next section.

The top variables selected by boosting the small and large dataset and logit can be found in Table 3. We found surprising that the variables used by the best logit models and the top variables selected by boosting the large dataset and small dataset mostly differed, likely accounting for the different in performances. One of the exceptions was forecasting 6 months ahead where the logit model and boosting the small dataset shared the 5 year - FF spread as the strongest predictor. As show in in Table 5, boosting the small dataset incorporates other variables such as the NAPM new orders and the S&P 500 on top of the 5 year - FF spread which may explain for superior performance over the best logit model which just uses the 5 year - FF spread.

Table 5: U.S. Boosting Top Variables Chosen by Cross-Validation: Small Dataset, In-Sample

| $h$ | Description | $I_k^2$ |
|---|---|---|
| 3 | NAPM new ordrs | 52.80 |
| | 5 Year T-bond - Federal Funds Spread | 28.37 |
| | S&P 500 | 10.32 |
| | Business Permits Total | 5.42 |
| 6 | 5 Year T-bond - Federal Funds Spread | 57.20 |
| | NAPM new ordrs | 24.40 |
| | S&P 500 | 9.08 |
| | Business Permits Total | 5.83 |
| 12 | 5 Year T-bond - Federal Funds Spread | 83.39 |
| | NAPM new ordrs | 9.15 |

Notes: Forecasts for period t are based on predictors at lag $t-h-1$ The column $I_k^2$ is an indicator of importance. We only include $I_k^2 > 3$ or variables that are above a certain threshold.

## 5.2 Out-Of-Sample Results

### 5.2.1 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 6. 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 6: Rolling Window Out-Of-Sample Forecasts in United States: 12 months

| 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 |

Notes: 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 6, 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

$$\bar{I}_k^2 = \sum_{t=t_1}^{T} \frac{1}{336} I_{k,t} \tag{13}$$

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:

$$freq_k = \frac{1}{336} \sum_{t=t_1}^{T} 1(I_{k,t}^2 > 0) \tag{14}$$

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.

### 5.2.2 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, Figure 4, and Figure 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 7.

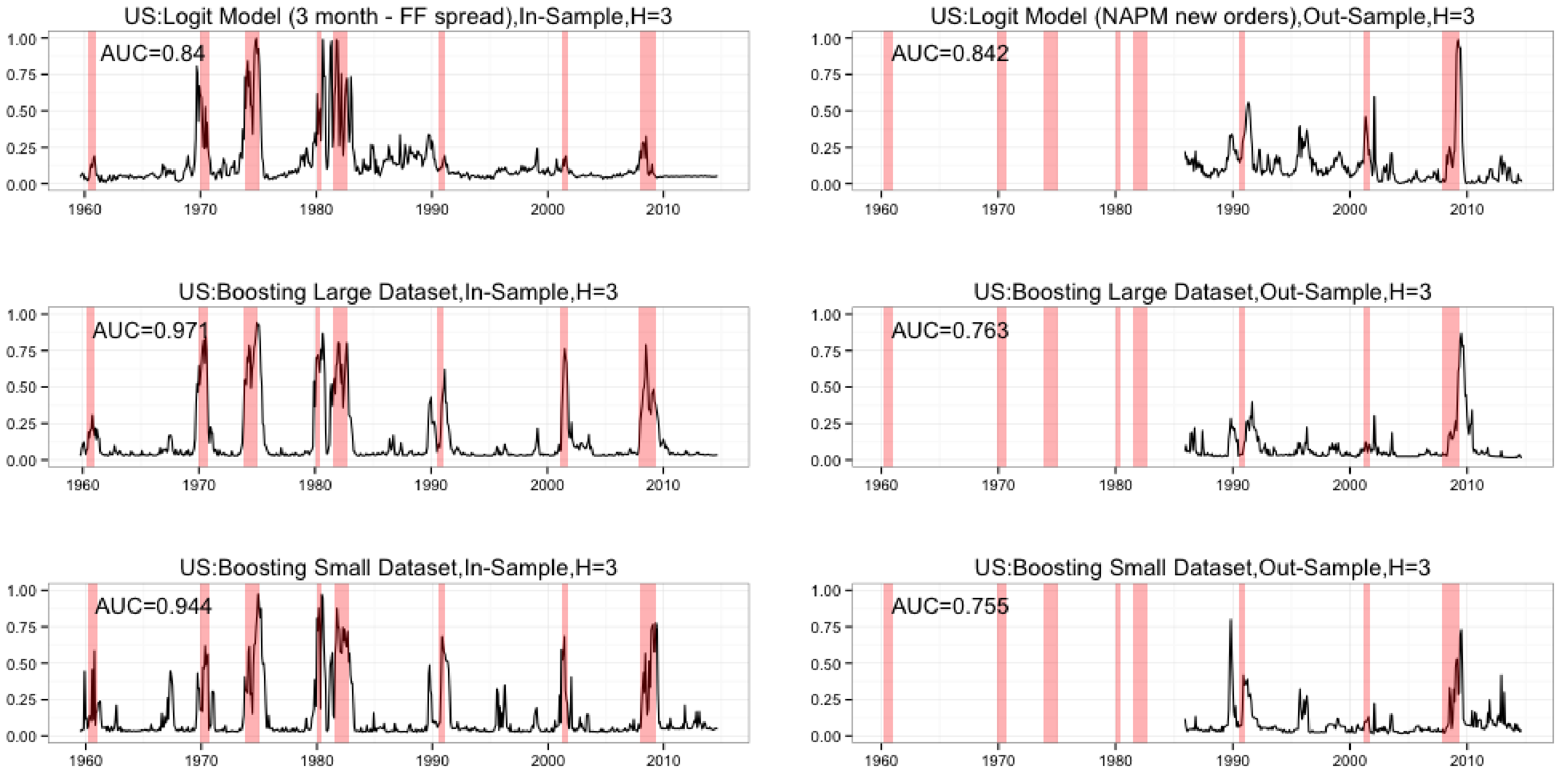
Table 7: U.S. Forecast Performance: Out-Of-Sample

| 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 |

Notes: 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.

Suprisingly, 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.

Figure 3: U.S. Forecasting Recession Performance 3 Months In Advance



US:Logit Model (3 month - FF spread),In-Sample,H=3

US:Logit Model (NAPM new orders),Out-Sample,H=3

US:Boosting Large Dataset,In-Sample,H=3

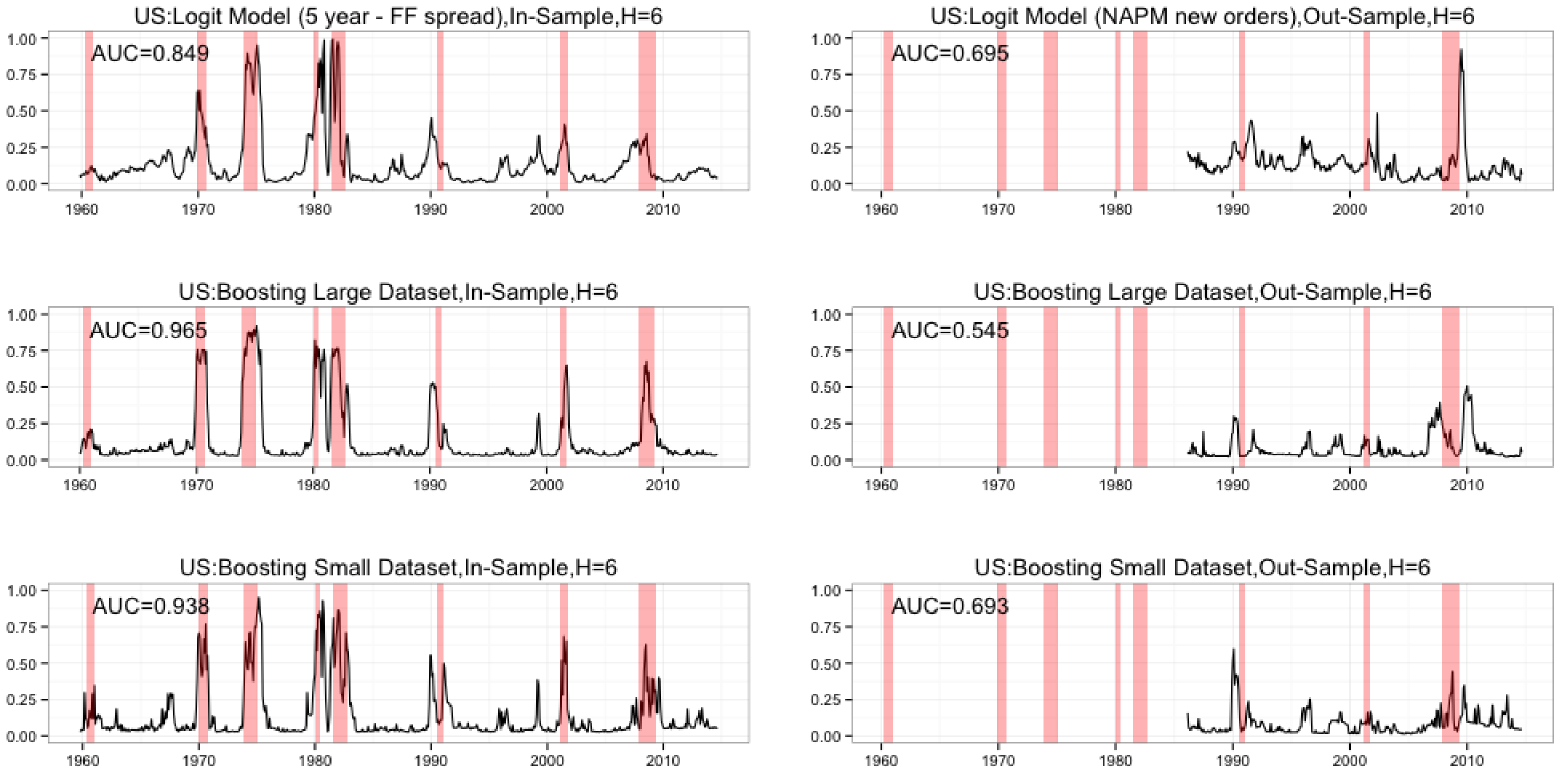
US:Boosting Large Dataset,Out-Sample,H=3

US:Boosting Small Dataset,In-Sample,H=3

US:Boosting Small Dataset,Out-Sample,H=3

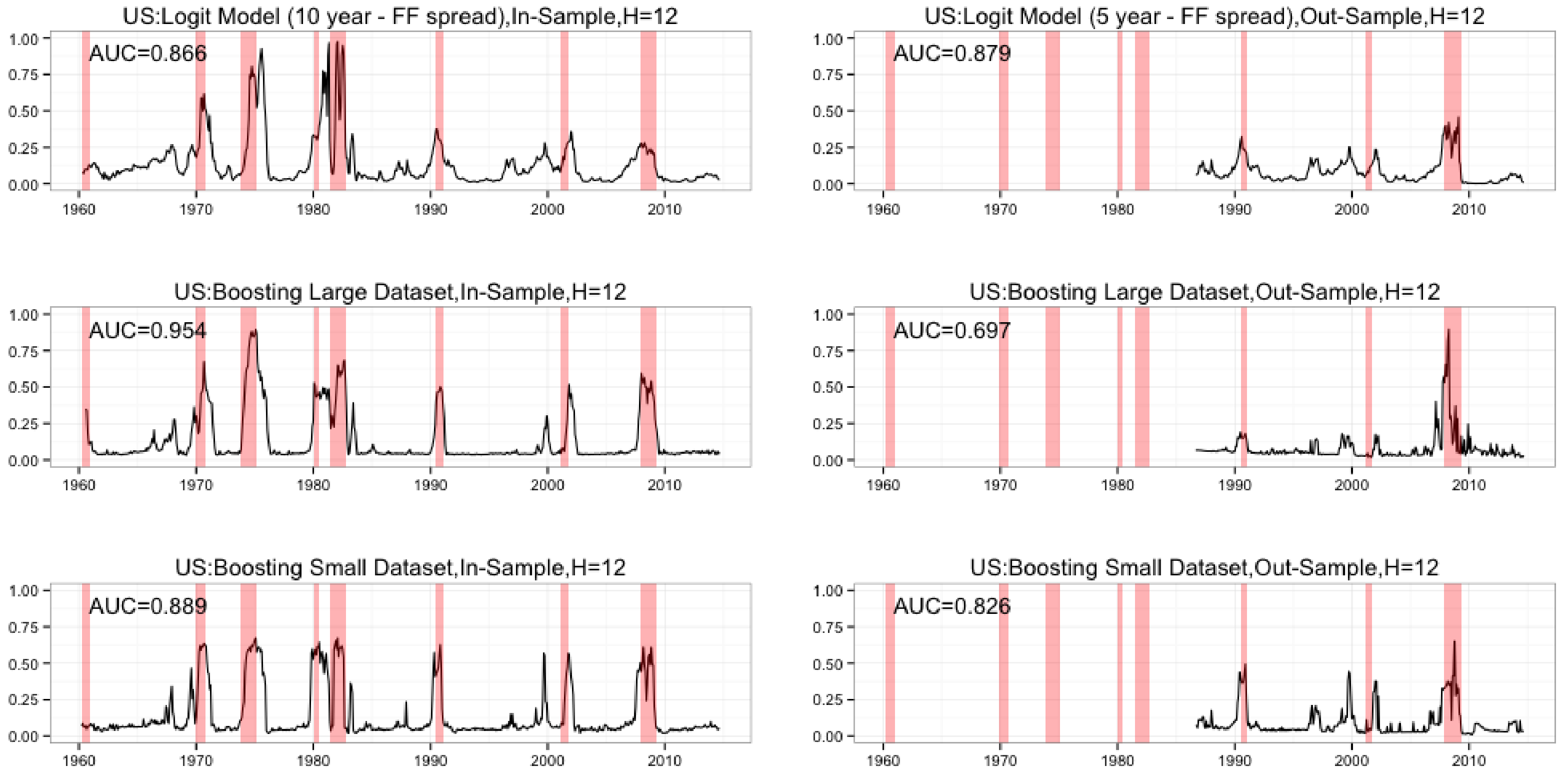
Notes: The left column displays in-sample forecasting performance in the U.S. at the 3 month horizon from 1959-04-01 to 2014-09-01 of the best logit model, boosting model with the large dataset (130 predictors), boosting model with small dataset (10 predictors), in that respective order. The right column 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.

Figure 4: U.S. Forecasting Recession Performance 6 Months In Advance



Notes: The left column displays in-sample forecasting performance in the U.S. at the 6 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 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.

Figure 5: U.S. Forecasting Recession Performance 12 Months In Advance



Notes: 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.

Figure 3, Figure 4 and Figure 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 the Great Recession is predicted too early.

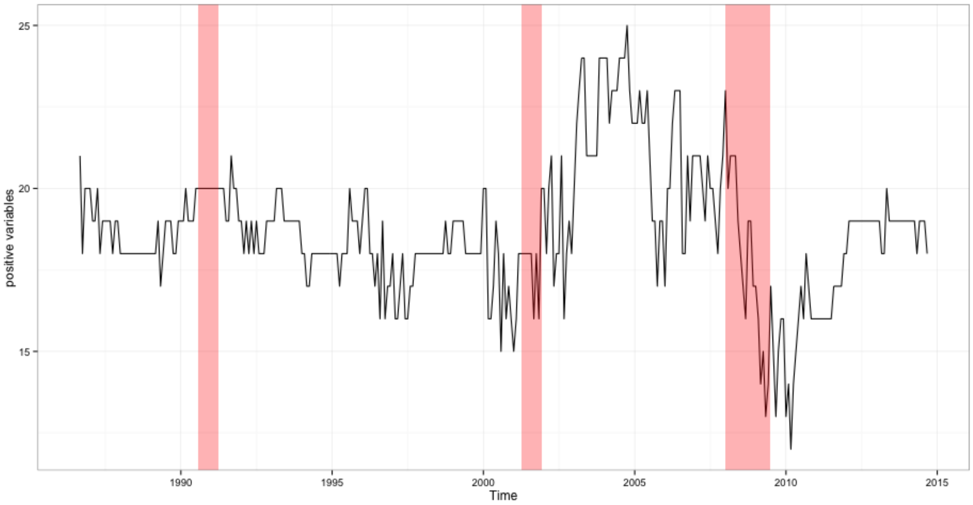
We found surprising that at the 12 month horizon, the best logit model out-of-sample (AUC 0.879) actually performed better than the best in-sample logit model (0.866), leading us to wonder why boosting out-of-sample did not beat boosting in-sample forecasts at the 12 month horizon.

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.

### 5.2.3 Variable Selection

We investigate why boosting with large datasets may actually hurt forecasting performance by investigating the variables chosen by boosting for the large and small dataset. We first turn our attention to the graph of the number of positive variables selected by boosting the large dataset in each of the rolling window subsamples at the 12 month horizon in Figure 6. The 3 month and 6 month equivalent can be found in the appendix. Ng (2014) reasons that Figure 6 illustrates the changing nature of business cycles as the number of predictors selected into the model at each subsample differs over time. There are almost 25 variables selected by boosting to forecast recessions in 2005 but just after the Great Recession in 2010 less than 10 are selected to forecast. One would imagine the ability of rolling windows to allow for boosting to change the amount of predictors selected by boosting at each forecast should help rather than hurt the model's prediction ability. However, we found boosting to underperform the best single variable model and we seek to understand this counterintuitive result.

Figure 6: U.S. Number of Variables Selected by Boosting in Large Dataset: Out-Of-Sample, 12 Month Horizon



Notes: Shaded red bars indicate recessions and black lines indicate number of positive variables selected by boosting to forecast the out-of-sample period 1986-09-01 to 2014-09-01.

We investigate Table 8 and Table 9 which show the variables with the highest average importance $\bar{I}_k^2$ as well as their frequency from boosting in the large dataset and small dataset respectively. Ng (2014) argues that one of the strengths of boosting is its ability to choose many predictors whereas most forecast models use just one predictor. Indeed out-of-sample, boosting the large dataset selects multiple spreads such as AAA spread, 3 month spread, 1 year spread, 5 year spread, 10 year spread in predicting the different horizons. However, boosting's strength in-sample to incorporate many spreads may be boosting's weakness when forecasting out-of-sample. The 5 year spread is the strongest predictor in our logit model for the 12 month horizon with AUC of 0.879. Boosting the large dataset for the same time horizon has an AUC of 0.697 with the 10 year spread, 5 year spread, BAA spread, and AAA spread all sharing similar average relative importance and appearing with equally high frequency. The prevalance of these 4 spreads may be the explanation for why boosting the large dataset underperforms the best logit model: boosting the large dataset dilutes the strength of the strongest predictor. Indeed, when we boost the small dataset which includes only 1 interest rate spread (5 year spread) unlike the large dataset that has 8 types of term and default spreads, we get a substantial increase in AUC from 0.697 (boosting large dataset) to 0.826 (boosting small dataset). When we boost a predictor set that includes just the 5 year spread and NAPM new orders and forecast at the 12 month horizon, we get an AUC of 0.876 which is just shy of the AUC score of the best logit model (result is omitted). However, when more predictors are added such as the the 10 year spread such that we predict with the 5 year spread, NAPM new orders, 10 year spread, the AUC starts to decrease (we again omit these results for concision purposes), suggesting one spread rather than more spreads when boosted has the best forecasting performance.

Table 8: U.S. Boosting Variables By Average Importance: Large Dataset, Out-Of-Sample

| $(h, h+d)$ | Variable Description | $I_k^2$ | Freq | lag | | |
|---|---|---|---|---|---|---|
| (3,6) | AAA - Federal Funds Spread | 14.2 | 0.91 | | | 6 |
| | 3 month T-bill - Fed Funds Spread | 13.95 | 1.05 | 4 | 5 | 6 |
| | NAPM new ordrs | 9.99 | 0.53 | 4 | | |
| | 6 month T-bill - Fed Funds Spread | 5.86 | 0.61 | | | 6 |
| | 1 Year T-bond - Federal Funds Spread | 4.77 | 0.65 | | | 6 |
| | Help wanted/emp | 4.66 | 0.83 | 4 | | |
| | Purchasing Manager's Index | 3.86 | 0.26 | 4 | | |
| | Build Permits:Midwest | 3.37 | 0.28 | 4 | | |
| (6.9) | AAA-Federal Funds Spread | 44.03 | 2.84 | 7 | 8 | 9 |
| | 1 Year T-bond - Federal Funds Spread | 5.89 | 0.66 | 7 | | 9 |
| | 5 Year T-bond - Federal Funds Spread | 5.45 | 0.40 | | | 9 |
| | 6 month T-bill - Federal Funds Spread | 5.03 | 0.42 | 7 | | |
| | 10 Year T-bond - Federal Funds Spread | 3.58 | 0.55 | | | 9 |
| (12,14) | 10 Year T-bond - Federal Funds Spread | 17.23 | 1.11 | 13 | 15 | |
| | 5 Year T-bond - Federal Funds Spread | 15.26 | 1.5 | | 15 | 16 |
| | BAA - Federal Funds Spread | 13.08 | 0.99 | 13 | | 16 |
| | AAA- Federal Funds Spread | 10.05 | 0.7 | 13 | | 16 |
| | Commercial - Fed Funds Spread | 5.25 | 0.69 | | | 16 |

Notes: Forecasts for period t are based on predictors at lag $t-h-1, t-h-2, ..., t-h-d$ where $d$ stands for lag and $h$ stands for forecast horizon. The column $I_k^2$ is an indicator of importance. The last column $lag \in [h+1, h+d]$ denotes that lag at which the corresponding predictor is chosen. We include $\bar{I}_k^2 > 3$. Frequency can exceed 1.0 as we add up the frequency of each individual lag.

Table 9: U.S. Boosting Variables By Average Importance: Small Dataset, Out-Of-Sample

| h | Description | $I_k^2$ | Freq |
|---|---|---|---|
| 3 | 5 Year T-bond - Federal Funds Spread | 39.27 | 1.00 |
| | NAPM new orders | 34.00 | 1.00 |
| | Business Permits Total | 18.16 | 1.00 |
| | S&P 500 | 4.70 | 1.00 |
| 6 | 5 Year T-Bond - Federal Funds Spread | 66.18 | 1.00 |
| | NAPM new orders | 14.94 | 1.00 |
| | Business Permits Total | 8.70 | 0.96 |
| | S&P 500 | 6.31 | 1.00 |
| 12 | 5 Year T-Bond - Federal Funds Spread | 83.49 | 1.00 |
| | Orders: consumer goods | 4.51 | 0.77 |
| | NAPM new orders | 4.23 | 0.91 |
| | Business Permits Total | 3.48 | 0.78 |

Notes: Forecast for period t are based on predictors at lag $t-h-1$ where $h$ represents the forecast horizon. The column $I_k^2$ is an indicator of importance. We only include variables of relative importance $I_k^2 > 3$

We also investigate why boosting with the large dataset outperforms boosting the small dataset at the 3 month horizon. We suspect that boosting on a large dataset in the short run contains more information on the real economy that predictors in the small dataset do not contain. We validate our hypothesis by seeing that some of the top variables selected out-of-sample in Table 8 when boosting on the large dataset consists of variables such as Building Permits, Purchasing Manager Index, Help Wanted Over Employement Ratio which gives greater granularity of the current state of the economy that one wouldn't get with the smaller dataset. Table 8 shows that boosting the small dataset at the 3 month horizon, the 5 year spread, NAPM new orders, Business Permits and the S&P have the highest relative importance but do not include the other real economy variables that the large dataset has. Thus, the gains from reducing a dataset to improve the performance of boosting may be limited in the short-term horizon.

In sum, the biggest gains in AUC seems to come from removing the other interest rate spread which seemed to dilute the strength of the 5 year spread. Boosting in general does significantly better in-sample and significantly worse out-of-sample and we suspect this is due to overfitting in-sample and diluting the strength of the strongest predictors out-of-sample. We are surprised that the best logit model consistently outperforms boosting a large dataset and even a small dataset that includes only predictors that we know are leading indicators. We discover that using a kitchen sink approach and throwing a big predictor set does not nessarily improve forecasting ability and in fact hurts the boosting model's ability to forecast properly in the United States. Carefully selecting a smaller dataset can improve the forecasting performance by removing predictors that predict poorly out-of-sample. Our finding here also echoes the parsimonious principle as a single variable logit model seems to consistently outperform boosted models loaded with many predictors.

# 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.

## 6.1 In-Sample Results

### 6.1.1 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.

### 6.1.2 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 10. 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 10: Japan Forecast Performance: In-Sample

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

Notes: 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.

### 6.1.3 Variable Selection

Since boosting a large and small dataset in Japan has not been explored before, we take a closer look at the variables selected by boosting these datasets.

Table 11 illustrates the in-sample selection of variables in Japan from boosting the large dataset. Unlike the United States, interest rate spreads (10 year Government Bonds - Tokyo Interbank 3 Month Rates) were not selected as strong predictors of recession in Japan in-sample but instead TANKAN business surveys on financial position, lending attitude, business conditions, domestic supply and demand conditions were found to be most predictive. The TANKAN business surveys selected in-sample from the large dataset mostly consist of manufuacturing and electrical machinery companies which suggest the financial position, lending attitude and business conditions of these companies may be leading indicators of the economy. Since the business surveys are meant to assess the current state or short term forecasts of businesses (3 months), we were surprised the TANKAN business surveys forecast well in the longer term forecasting horizons. Thus, we expect the business surveys and our models to be less predictive out-of-sample at the 12 month horizon.

Table 11: Japan Boosting Recession Variables Chosen by Cross-Validation: Large Dataset, In-Sample

| $h$ | Variable Description | $I_k^2$ |
|---|---|---|
| 3 | Financial Position: All, Transportation Machinery, Actual | 35.06 |
| | Financial Position: Large Enterprises, Manuf., Electrical Machinery | 14.17 |
| | Lending Attitude, All Enterprises, Manuf., Electrical Machinery | 6.46 |
| | Business Conditions, Large Enterprises, Manuf., Electrical Machinery | 6.40 |
| | Business Conditions, All Enterprises, Manuf., Electrical Machinery | 6.10 |
| | Business Conditions, Medium Enterprises, Manuf., Petroleum, Coal | 5.71 |
| 6 | Business Conditions, Medium Enterprises, Manuf., Electrical Machinery | 24.01 |
| | Lending Attitude, All Enterprises, Manuf., Electrical Machinery | 13.79 |
| | Financial Position: All, Transportation Machinery, Actual | 13.60 |
| | Business Conditions, Large Enterprises, Manuf., Electrical Machinery | 4.77 |
| | Financial Position, Medium Enterprises, Manuf., Transport. Machinery | 4.45 |
| | Financial Position, Small Enterprises, Manuf., Processed Metals | 4.14 |
| 12 | Business Conditions, Medium Enterprises, Manuf., Electrical Machinery | 24.55 |
| | Domestic Supply/Demand, Medium Enterprises, Manuf., Food, Beverages | 12.92 |
| | Lending Attitude, Small Enterprises, Manuf., Petroleum and Coal | 11.36 |
| | Inventory Level, Small Enterprises, Manuf., Processing | 7.61 |
| | Financial Position, Large Enterprises, Manuf., Pulp and Paper | 5.02 |
| | Lending Attitude, All Enterprises, NonManuf., Real Estate | 4.82 |

Notes: Forecast for period t are based on predictors at lag $t-h-1$ where $h$ represents the forecast horizon. The column $I_k^2$ is an indicator of importance. We only include variables of relative importance $I_k^2 > 4$

In looking at the raw data, we find that interest rate spreads aren't predictive in Japan because interest rate spreads (10 year Government Bonds - Tokyo Interbank 3 Month Rates) are rarely negative or inverted after the mid-1990s because of the nearly 0% Tokyo Interbank Offered Rates (3 months) since the mid-1990s. Both the graphs of interest rate spreads and the interbank offer rates can be found in the appendix. This result supports the finding of Bernard and Gerlach (1998) that interest rate spreads are not strongly predictive of Japanese recessions.

Table 12 illustrates the in-sample selection of variables in Japan from boosting the small dataset. The broad TANKAN business survey results such as business conditions of all enterprsies across all industries are the strongest predictor in the 3 and 6 month horizon. Inventory level of goods of finished goods and merchandise across all enterprises in manufacturing companies is the most predictive in forecasting 12 months ahead in-sample. Once again, interest rate spreads were not included by boosting even though the small dataset.

Table 12: Japan Boosting Recession Variables Chosen by Cross-Validation: Small Dataset, In-Sample

| $h$ | Description | $I_k^2$ |
|---|---|---|
| 3 | Business Conditions,All Enterprises,All industries,Actual | 25.45 |
| | Ratio of Operating Profits to Total Assets (Manuf.) | 23.59 |
| | Business Conditions,All Enterprises,All industries,Forecast | 11.38 |
| | Inventory Level, All Enterprises,Manuf.,Actual | 9.20 |
| | Nikkei Commodity Price Index (42items) | 6.25 |
| | Employment Conditions,All Enterprises,Manuf.,Actual | 4.68 |
| 6 | Business Conditions,All Enterprises,All industries,Actual | 26.32 |
| | Domestic Supply/Demand,All Enterprises,Manuf.,Forecast | 17.70 |
| | Inventory Level, All Enterprises,Manuf.,Actual | 15.94 |
| | Ratio of Operating Profits to Total Assets (Manuf.) | 9.44 |
| | Financial Position,All Enterprises,All industries,Actual | 5.44 |
| | Consumer Confidence Index | 4.08 |
| 12 | Inventory Level, All Enterprises,Manuf.,Actual | 23.37 |
| | Financial Position,All Enterprises,All industries,Actual | 15.81 |
| | Domestic Supply/Demand ,All Enterprises,Manuf.,Forecast | 15.17 |
| | Employment Conditions,All Enterprises,Manuf.,Forecast | 14.23 |
| | Domestic Supply/Demand ,All Enterprises,Manuf.,Actual | 10.99 |
| | Inventory Level, All Enterprises, All industries, Actual | 7.56 |

Notes: Forecast for period t are based on predictors at lag $t-h-1$ where $h$ represents the forecast horizon. The column $I_k^2$ is an indicator of importance. We only include variables of relative importance $I_k^2 > 4$

Our finding thus far suggest that, like the United States, boosting the large dataset in-sample in Japan far outperforms the small dataset and the logit model. Furthermore we find that business survey results of the TANKAN are the most predictive across all time horizons while interest rate spreads are not.

## 6.2 Out-Of-Sample Results

### 6.2.1 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.

### 6.2.2 Forecast Performance

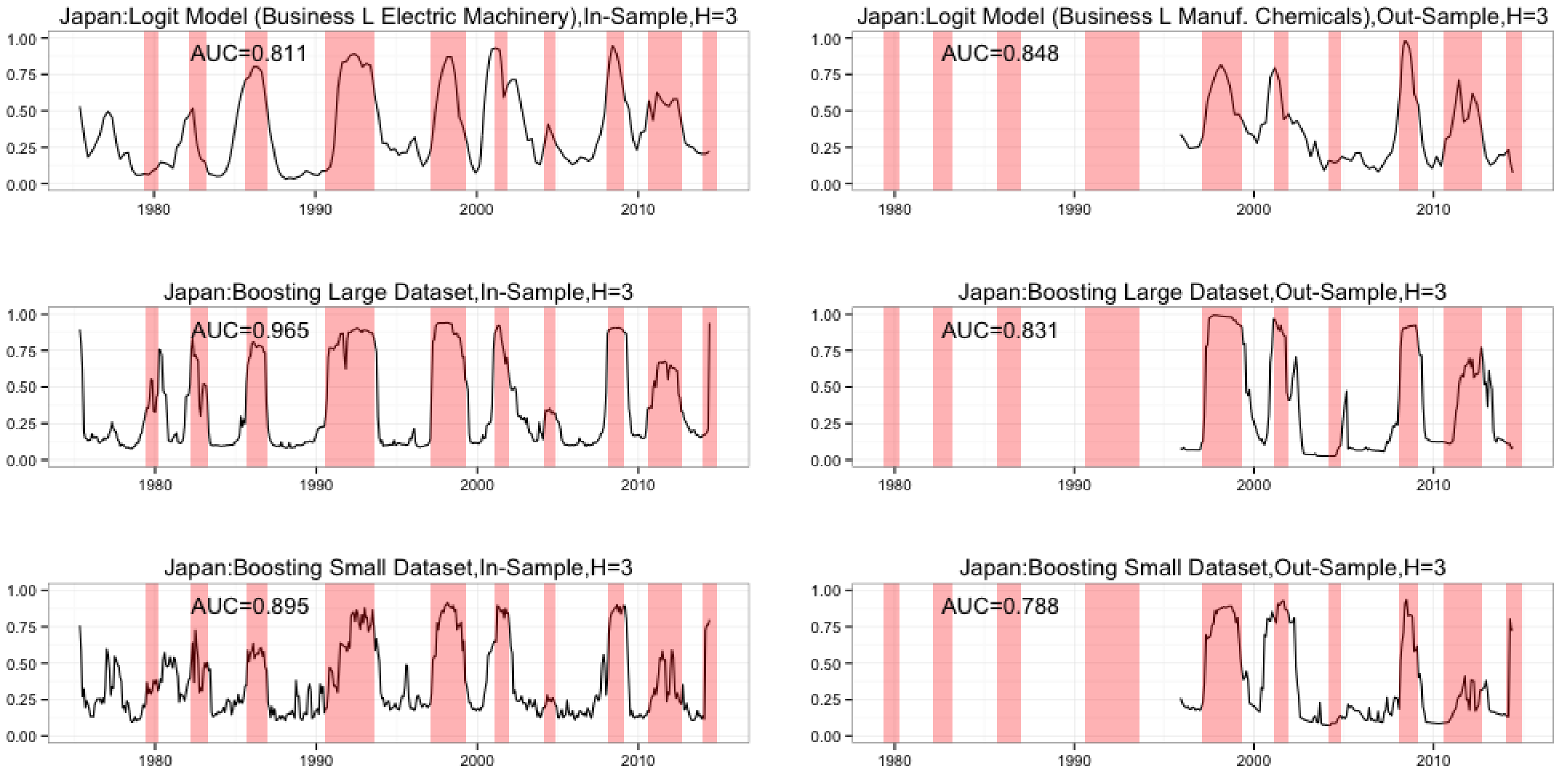
In-sample and out-of-sample forecasts of the best logit model and boosting models for Japan are shown in Figure 7, Figure 8 and Figure 9 for the 3 month, 6 month and 12 month horizon respectively. Out-of-sample performance with T-tests comparing boosting models with the best logit models can be found in Table 13.

Table 13: Japan Forecast Performance: Out-Of-Sample

| 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 |

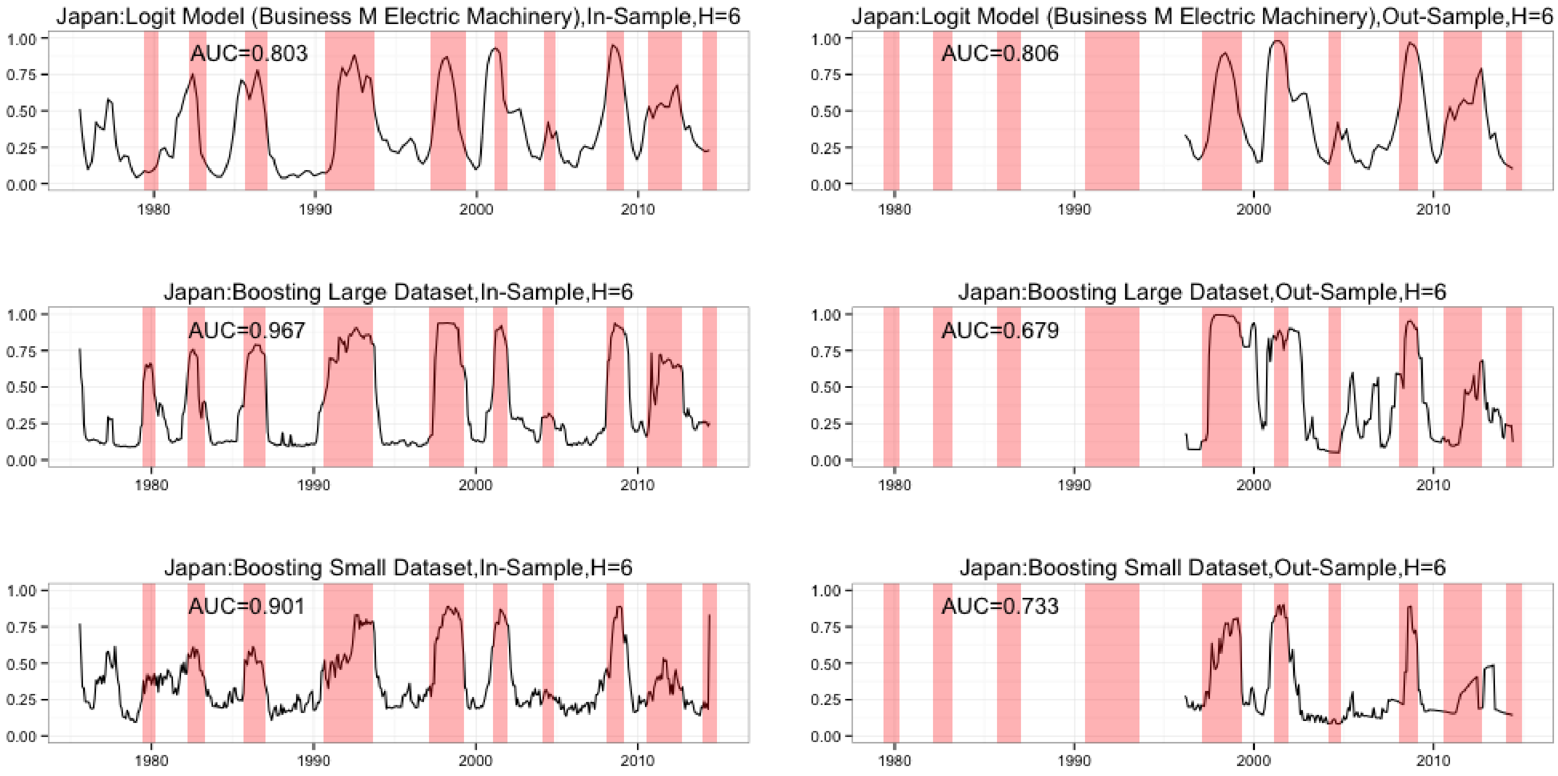
Notes: Rolling windows begin 1975-01-01:1995-08-01 to forecast out-of-sample 1995-09-01:2014-06-01. Small dataset refers to 20 leading variables from Japan's Cabinet Office and business surveys. Large dataset refers to 436 business surveys and macrovariables. T-Test compares the AUC of boosting to the 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 7: Japan Forecasting Recession Performance 3 Months In Advance



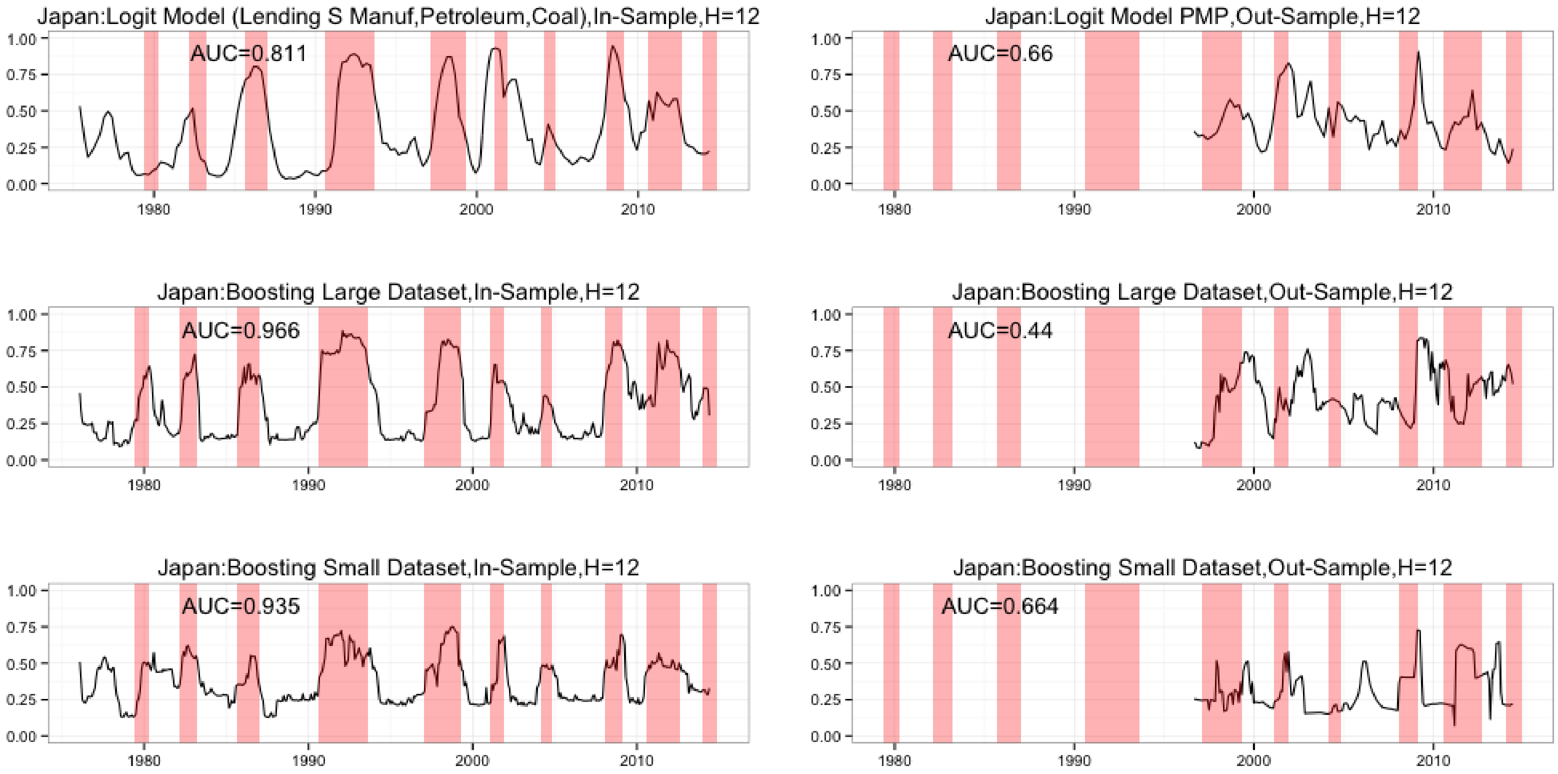
Notes: The left column displays in-sample forecasting performance in Japan at the 3 month horizon from 1975-01-01 to 2014-06-01 of the best logit model, boosting model with the large dataset (436 predictors), boosting model with small dataset (27 predictors), in that respective order. The right column 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.

Figure 8: Japan Forecasting Recession Performance 6 Months In Advance



Notes: 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 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.

Figure 9: Japan Forecasting Recession Performance 12 Months In Advance



Notes: 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.

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. Surprisingly, 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.

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 9 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 9, 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 7 and Figure 8 where the in-sample fit is nearly perfect for boosting but out-of-sample the performance declines.

### 6.2.3 Variable Selection

We now look to investigate the following three questions:

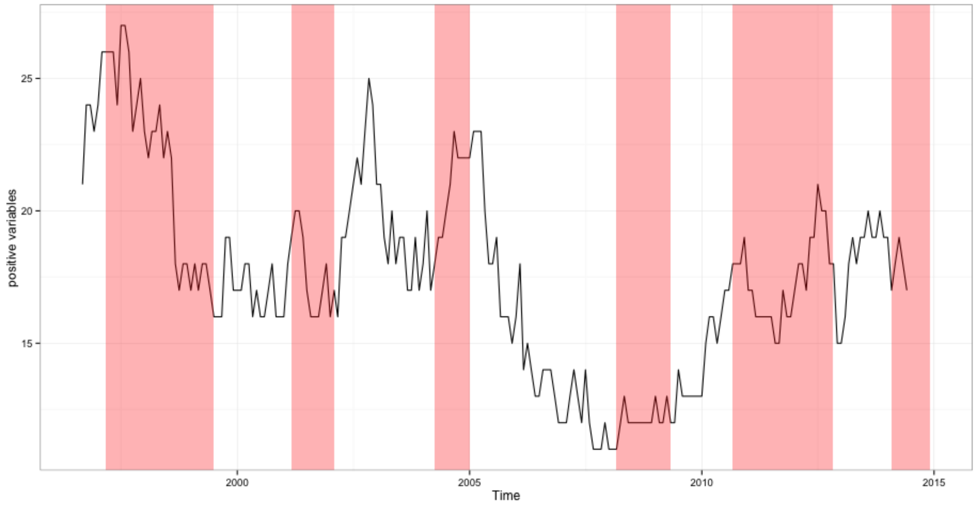
1. Why does boosting the large dataset do worse than the best logit model at the 6 month and 12 month horizon?

2. Why does boosting the small dataset outperform boosting a large dataset at the 6 month and 12 month horizon?

3. Why does boosting the small dataset actually outperforms the best logit model at the 12 month horizon?

First we investigate why boosting the large dataset does so much worse than the best logit model at the 6 month and 12 month horizon. Like we found in the United States, the number of variables selected by boosting at the 12 month horizon in Japan varies overtime as seen in Figure 10. The number of positive variables selected peaks at 26 in 1997-07-01 and falls to 11 in 2007-09-01. The changing nature of the variables predictor sets suggests the changing nature of business cycles. One would think the ability of the predictor set to change overtime would allow for more accurate forecasting performance via boosting but instead, like we found in the United States, a single predictor is able to outperform the boosting model. We seek to understand this counterintuitive result in Japan.

Figure 10: Japan Number of Variables Selected by Boosting in Large Dataset at 12 Month Horizon: Out-Of-Sample,



Note: Shaded red bars indicate recessions and black lines indicate number of positive variables selected by boosting to forecast the period from 1996-09-01 to 2014-06-01.

Table 14 displays the variables selected from boosting the large dataset. We find that variables with the highest average relative importance for each horizon are very different from the variables from the best logit models. As seen in Table 13, the variable selected by the best logit model at the 6 month horizon concerns Business Conditions whereas boosting the large dataset selects a survey about Inventory Levels. At the 12 month horizon, the best logit model selects a variable on Inventory Levels whereas boosting the large dataset selects Lending Attitudes as the most important variable. Our finding suggests that boosting the large dataset may underperform the best logit model simply because boosting does not prioritize the best performing variable in the boosting model. Why logit selects one variable as the strongest predictor and boosting chooses another variable as the strongest predictor may come from how boosting incorporates many different types of classification rules and variables whereas logit can only pick one variable. We hypothesize that boosting tends to fit very well to the rolling subsamples at each iteration, selecting variables with classification rules that work within the rolling sub-sample. However, the same classification rules and variables that worked in the rolling subsample may not work as well outside of the rolling sample. Hence we believe the boosting property to incorporate many variables leads to overfitting in the rolling samples and subsequently inferior forecasting performance out-of-sample compared to that of a simple logit model that uses the same one variable across all rolling subsamples.

Table 14: Japan Boosting Recession Variables By Average Importance: Large Dataset, Out-Of-Sample

| $h$ | Variable Description | $\bar{I}_k^2$ | Freq |
|---|---|---|---|
| 3 | Financial Position: All, Transportation Machinery, Actual | 43.15 | 0.93 |
| | Financial Position, Large Enterprises, Manuf., Transport Machine | 11.40 | 0.83 |
| | Financial Position, Medium Enterprises, Manuf., Transport Machine | 5.50 | 0.83 |
| 6 | Financial Position, Large Enterprises, Manuf., Transport Machinery | 12.43 | 0.64 |
| | Financial Position: All, Transportation Machinery, Actual | 8.12 | 0.50 |
| | Business Conditions, All Enterprises, NonManuf., Retailing | 6.79 | 0.65 |
| | Financial Position, Medium Enterprises, Manuf., Transport Machinery | 6.76 | 0.55 |
| | Business Conditions, Medium Enterprises, Manuf., Electrical Machine | 6.01 | 0.63 |
| | Financial Position, Medium Enterprises, Manuf., Chemicals | 5.97 | 0.37 |
| 12 | Lending Attitude, Small Enterprises, Manuf., Petroleum and Coal | 13.27 | 0.66 |
| | Lending Attitude, Medium Enterprises, Manuf., Chemicals | 12.97 | 0.42 |
| | Inventory Level, All Enterprises, Manuf., Pulp and Paper | 7.15 | 0.63 |
| | Business Conditions, Medium Enterprises, Manuf., Electrical Machine | 6.81 | 0.53 |
| | Inventory Level, All Enterprises, All Industries | 6.71 | 0.50 |
| | Lending Attitude, All Enterprises, NonManuf., Real Estate | 4.31 | 0.54 |

Notes: Forecast for period t are based on predictors at lag $t-h-1$ where $h$ represents the forecast horizon. The column $I_k^2$ is an indicator of importance. We only include variables of relative importance $I_k^2 > 4$

Table 15: Japan Boosting Recession Variables By Average Importance: Small Dataset, Out-Of-Sample

| $h$ | Description | $\bar{I}_k^2$ | Freq |
|---|---|---|---|
| 3 | Financial Position,All Enterprises,Manuf.,Actual result | 25.29 | 0.95 |
| | Inventory Level, All Enterprises,Manuf.,Actual result | 23.48 | 0.98 |
| | Ratio of Operating Profits to Total Assets (Manuf.) | 12.40 | 1.00 |
| | Domestic Supply & Demand, All Enterprises,Manuf.,Forecast | 12.12 | 0.70 |
| | Business Conditions,All Enterprises,All industries,Forecasts | 8.88 | 0.35 |
| | Employment Conditions,All Enterprises,Manuf.,Forecast | 5.53 | 0.95 |
| | Inventory Level, All Enterprises,All industries,Actual result | 5.03 | 0.73 |
| 6 | Inventory Level, All Enterprises,Manuf.,Actual result | 24.40 | 1.00 |
| | Inventory Level, All Enterprises,All industries,Actual result | 19.38 | 1.00 |
| | Domestic Supply & Demand ,All Enterprises,Manuf.,Actual | 14.25 | 0.61 |
| | Financial Position,All Enterprises,Manuf.,Actual result | 7.44 | 0.85 |
| | Employment Conditions,All Enterprises,All industries,Forecast | 6.45 | 0.94 |
| | Domestic Supply & Demand ,All Enterprises,Manuf.,Forecast | 5.28 | 0.79 |
| | Financial Position,All Enterprises,All industries,Actual result | 4.69 | 0.73 |
| | Employment Conditions,All Enterprises,Manuf.,Forecast | 4.69 | 0.90 |
| 12 | Inventory Level, All Enterprises,All industries,Actual result | 30.86 | 0.97 |
| | Inventory Level, All Enterprises,Manuf.,Actual result | 18.41 | 1.00 |
| | Employment Conditions,All Enterprises,All industries,Forecast | 17.47 | 0.83 |
| | Employment Conditions,All Enterprises,Manuf.,Forecast | 15.30 | 0.92 |
| | Financial Position,All Enterprises,Manuf.,Actual result | 8.04 | 0.64 |
| | Financial Position,All Enterprises,All industries,Actual result | 4.97 | 0.71 |

Notes: Forecast for period t are based on predictors at lag $t-h-1$ where $h$ represents the forecast horizon. The column $\bar{I}_k^2$ is an indicator of average importance. We only include variables of average relative importance $\bar{I}_k^2 > 4$.

We now investigate why boosting the small dataset generally leads to a higher AUC than boosting the large dataset. At forecasting at the 12 month horizon, boosting the large dataset has a AUC score of 0.44, thus predicting recessions worse than random chance. However, boosting the small dataset leads to a substantially higher AUC of 0.664 which is even slightly better than the best logit model. We find that the main difference between the two boosting models is that boosting the large dataset selects

Lending Attitude of Medium and Small Enterprises as the top two variables whereas boosting the small dataset selects Inventory Levels of All Enterprises and Inventory Level of All Manufacturing companies as the top two variables. The best logit model at the 12 month horizon uses Inventory Level of Medium Enterprises that Manufacture Electrical Machinery. Thus, the likely reason why boosting the large dataset performs so poorly relative to boosting the small dataset is that the large boosting does not prioritize Inventory Levels as a strong predictor whereas boosting the small dataset does. Once again, we believe boosting the large dataset incorporates noise that fits well at each of its rolling subsamples but do not predict well outside of the rolling subsamples.

We now turn our attention to investigate why boosting the small dataset leads to an AUC slightly better than the best logit model at the 12 month horizon. As noted earlier, the best logit model at the 12 month horizon predicts uses Inventory Level of Medium Enterprises and boosting small datasets focus on Inventory Levels of All Enterprises and Employment Condition Forecasts. It seems that boosting the small dataset performs slightly better because it incorporates a more general measurement of Inventory Level whereas the logit model focuses specifically on Medium Enterprises that Manufacture Electrical Machinery.

We note that TANKAN business surveys forecast relatively poorly at the 12 month horizon as the best logit model that predits with the Inventory Level of Medium Enterprises has a AUC of 0.66 whereas the AUC of the the 6 month horizon logit model has an AUC of 0.80. The result confirms our suspicion that there was overfiting in-sample as the in-sample models at the 12 month horizon had AUC of 0.95 for boosting the large dataset. Furtheremore, we believe predicting recessions beyond the 6 month horizon will be difficult in Japan as interest rate spreads aren't predictive whereas in the U.S. spreads are the strongest predictor at the 12 month horizon.

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 parsimonious principle that models with fewer variables can outperform models that involve many predictors, some that may not be predictive.

# 7 Conclusion

We have contributed to the literature by finding 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. We find that, unsurprisingly, the best logit models in the U.S. predict with term spreads. 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. Additionally, we find that spreads are not predictive of recessions in Japan from 1995 to 2014 and that Japan TANKAN business surveys are the strongest predictors at the 3, 6 month horizon but do poorly at the 12 month horizon. 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. We find our result counterintuitive as we hypothesized that the flexibility to incorporate different predictors shoud allow boosting to predict recessions better than logit models that rely on just one variable. Our explanation is that the parsimonious 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.

Furthermore, an inherent flaw of forecasting excercises done in our paper and often other papers is the use of recent-vintage data. Berge (2014) runs an experiment and uses real-time data instead of recent-vintage data and finds classification ability of the models to drop significantly. Thus, while researchers may find forecast models that outperforms the best logit model, applying the models in practice may be limited due to data revision. An approach to avoid working with revised data is using financial data that isn't subject to revision such as stock market prices and term/default spreads, an approach that Estrella and Mishkin (1998) advocates for.

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. The parsimonious principle continues to ring true even after William of Ockham declared the law of parsimony back in the 14th century as we find simple models predict recessions best.

# 8 Appendix

## 8.1 U.S. Monthly Indicators: Large Dataset

Below is the U.S. Monthly Dataset used by Stock and Watson (2005) and Ng (2014). In the transformation column, $ln$ denotes logarithm transformation, $\triangle ln$ and $\triangle^2 ln$ denote the first and second difference of the logarithm, $lv$ denotes the level of the series, and $\triangle level$ denotes the first difference of the series. The data are available from 1959:01-2014:06. To transform some series to be stationary, real variables are expressed in growth rates, first differences are used for nominal interest rates, and second log differences are used for prices. We folow the transformations used by Stock and Watson (2005).

**Group 1: Output and Income**

| Variable | Description | Transformation |
|---|---|---|
| YPR | Personal Income (AR, Bil. Chain 2000 $) | $\triangle ln$ |
| IPS10 | Industrial Production Index - Total Index | $\triangle ln$ |
| IPS11 | Industrial Production Index - Products, Total | $\triangle ln$ |
| IPS299 | Industrial Production Index - Final Products | $\triangle ln$ |
| IPS12 | Industrial Production Index - Consumer Goods | $\triangle ln$ |
| IPS13 | Industrial Production Index - Durable Consumer Goods | $\triangle ln$ |
| IPS18 | Industrial Production Index - Nondurable Consumer Goods | $\triangle ln$ |
| IPS25 | Industrial Production Index - Business Equipment | $\triangle ln$ |
| IPS32 | Industrial Production Index - Materials | $\triangle ln$ |
| IPS34 | Industrial Production Index - Durable Goods Materials | $\triangle ln$ |
| IPS38 | Industrial Production Index - Nondurable Goods Materials | $\triangle ln$ |
| IPS43 | Industrial Production Index - Manufacturing (Sic) | $\triangle ln$ |
| IPS307 | Industrial Production Index - Residential Utilities | $\triangle ln$ |
| IPS306 | Industrial Production Index - Fuels | $\triangle ln$ |
| PMP | NAPM Manufacturing Production Index (Percent) | $level$ |
| UTL11 | Capacity Utilization (SIC-Mfg) (TCB) | $\triangle level$ |

**Group 2: Labor Market**

| Variable | Description | Transformation |
|---|---|---|
| LHEL | Index Of Help-Wanted Advertising In Newspapers (1967=100;Sa) | $\triangle level$ |
| LHELX | Employment: Ratio; Help-Wanted Ads:No. Unemployed Clf | $\triangle level$ |
| LHEM | Civilian Labor Force: Employed, Total (Thous.,Sa) | $\triangle ln$ |
| LHNAG | Civilian Labor Force: Employed, Nonagric.Industries (Thous.,Sa) | $\triangle ln$ |
| LHUR | Unemployment Rate: All Workers, 16 Years & | $\triangle level$ |
| LHU680 | Unemploy.By Duration: Average(Mean)Duration In Weeks (Sa) | $\triangle level$ |
| LHU5 | Unemploy.By Duration: Persons Unempl.Less Than 5 Wks (Thous.,Sa) | $\triangle ln$ |
| LHU14 | Unemploy.By Duration: Persons Unempl.5 To 14 Wks (Thous.,Sa) | $\triangle ln$ |
| LHU15 | Unemploy.By Duration: Persons Unempl.15 Wks + (Thous.,Sa) | $\triangle ln$ |
| LHU26 | Unemploy.By Duration: Persons Unempl.15 To 26 Wks (Thous.,Sa) | $\triangle ln$ |
| LHU27 | Unemploy.By Duration: Persons Unempl.27 Wks + (Thous,Sa) | $\triangle ln$ |
| CLAIMUII | Average Weekly Initial Claims, Unemploy. Insurance (Thous.) | $\triangle ln$ |
| CES002 | Employees On Nonfarm Payrolls: Total Private | $\triangle ln$ |
| CES003 | Employees On Nonfarm Payrolls - Goods-Producing | $\triangle ln$ |
| CES006 | Employees On Nonfarm Payrolls - Mining | $\triangle ln$ |

**Group 2: Labor Market**

| Variable | Description | Transformation |
|---|---|---|
| CES011 | Employees On Nonfarm Payrolls - Construction | $\triangle ln$ |
| CES015 | Employees On Nonfarm Payrolls - Manufacturing | $\triangle ln$ |
| CES017 | Employees On Nonfarm Payrolls - Durable Goods | $\triangle ln$ |
| CES033 | Employees On Nonfarm Payrolls - Nondurable Goods | $\triangle ln$ |
| CES046 | Employees On Nonfarm Payrolls - Service-Providing | $\triangle ln$ |
| CES048 | Employees On Nonfarm Payrolls - Trade, Transportation, And Utilities | $\triangle ln$ |
| CES049 | Employees On Nonfarm Payrolls - Wholesale Trade. | $\triangle ln$ |
| CES053 | Employees On Nonfarm Payrolls - Retail Trade | $\triangle ln$ |
| CES088 | Employees On Nonfarm Payrolls - Financial Activities | $\triangle ln$ |
| CES140 | Employees On Nonfarm Payrolls - Government | $\triangle ln$ |
| A0M048 | Employee Hours In Nonag. Establishments (AR, Bil. Hours) | $\triangle ln$ |
| CES151 | Avg Weekly Hrs of Prod or Nonsup Workers On Private Nonfarm Payrolls - Goods-Producing | $level$ |
| CES155 | Avg Weekly Hrs of Prod or Nonsup Workers On Private Nonfarm Payrolls - Mfg Overtime Hours | $\triangle level$ |
| A0M001 | Average Weekly Hours, Mfg. (Hours) | $level$ |
| PMEMP | Napm Employment Index (Percent) | $level$ |
| CES275 | Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Goods-Producing | $\triangle^2 ln$ |
| CES277 | Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Construction | $\triangle^2 ln$ |
| CES278 | Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Manufacturing | $\triangle^2 ln$ |

**Group 3: Housing**

| Variable | Description | Transformation |
|---|---|---|
| HSFR | Housing Starts:Nonfarm(1947-58);Total Farm&Nonfarm(1959-)(Thous.,Saar) | $ln$ |
| HSNE | Housing Starts:Northeast (Thous.U.)S.A. | $ln$ |
| HSMW | Housing Starts:Midwest(Thous.U.)S.A. | $ln$ |
| HSSON | Housing Starts:South (Thous.U.)S.A. | $ln$ |
| HSWST | Housing Starts:West (Thous.U.)S.A. | $ln$ |
| HSBNE | Housing Authorized: Total New Priv Housing Units (Thous.,Saar) | $ln$ |
| HSBMW | Houses Authorized By Build. Permits:Midwest(Thou.U.)S.A. | $ln$ |
| HSBSOU | Houses Authorized By Build. Permits:South(Thou.U.)S.A. | $ln$ |
| HSBWST | Houses Authorized By Build. Permits:West(Thou.U.)S.A. | $ln$ |

**Group 4: Consumption Orders and Inventories**

| Variable | Description | Transformation |
|---|---|---|
| PMI | Purchasing Managers' Index (Sa) | $level$ |
| PMNO | Napm New Orders Index (Percent) | $level$ |
| PMDEL | Napm Vendor Deliveries Index (Percent) | $level$ |
| PMNV | Napm Inventories Index (Percent) | $level$ |
| A1M008 | Mfrs' New Orders, Consumer Goods And Materials (Mil. $) | $\triangle ln$ |
| A0M007 | Mfrs' New Orders, Durable Goods Industries (Bil. Chain 2000 $ ) | $\triangle ln$ |
| A0M027 | Mfrs' New Orders, Nondefense Capital Goods (Mil. Chain 1982 $) | $\triangle ln$ |
| A1M092 | Mfrs' Unfilled Orders, Durable Goods Indus. (Bil. Chain 2000 $) | $\triangle ln$ |
| A0M070 | Manufacturing And Trade Inventories (Bil. Chain 2000 $) | $\triangle ln$ |
| A0M077 | Ratio, Mfg. And Trade Inventories To Sales (Based On Chain 2000 $) | $\triangle level$ |
| CONS-R | Real Personal Consumption Expenditures (AC) (Bill $) pi031 / gmdc | $\triangle ln$ |
| MTQ | Manufacturing And Trade Sales (Mil. Chain 1996 $) | $\triangle ln$ |
| A0M059 | Sales Of Retail Stores (Mil. Chain 2000 $) | $\triangle ln$ |
| HHSNTN | U. Of Mich. Index Of Consumer Expectations | $\triangle level$ |

**Group 5: Money and Credit**

| Variable | Description | Transformation |
|---|---|---|
| FM1 | Money Stock: M1(Curr,Trav.Cks,Dem Dep,Other Ck'able Dep)(Bil$,S | $\triangle^2 ln$ |
| FM2 | Money Stock:M2(M1+O'nite Rps,Euro$,G/P&B/D & Mmmfs&Sav& Sm Time Dep(Bil$,Sa) | $\triangle^2 ln$ |
| FMSCU | Money Stock: Currency held by the public (Bil$,Sa) | $\triangle^2 ln$ |
| FM2-R | Money Supply: Real M2, fm2 / gmdc (AC) | $\triangle ln$ |
| FMFBA | Monetary Base, Adj For Reserve Requirement Changes(Mil$,Sa) | $\triangle^2 ln$ |
| FMRRA | Depository Inst Reserves:Total, Adj For Reserve Req Chgs(Mil$,Sa) | $\triangle^2 ln$ |
| FMRNBA | Depository Inst Reserves:Nonborrowed,Adj Res Req Chgs(Mil$,Sa) | $\triangle^2 ln$ |
| FCLNBW | Commercial & Industrial Loans Outstanding + NonFin Comm. Paper (Mil$, SA) | $\triangle^2 ln$ |
| FCLBMC | Wkly Rp Lg Com'l Banks:Net Change Com'l & Indus Loans(Bil$,Saar) | $level$ |
| CCINRV | Consumer Credit Outstanding - Nonrevolving(G19) | $\triangle^2 ln$ |
| CCIPY | Ratio, Consumer Installment Credit To Personal Income (Pct.) | $\triangle level$ |

**Group 6:Bond and Exchange Rates**

| Variable | Description | Transformation |
|---|---|---|
| FYFF | Interest Rate: Federal Funds (Effective) (% Per Annum,Nsa) | $\triangle level$ |
| CP90 | Commercial Paper Rate | $\triangle level$ |
| FYGM3 | Interest Rate: U.S.Treasury Bills,Sec Mkt,3-Mo.(% Per Ann,Nsa) | $\triangle level$ |
| FYGM6 | Interest Rate: U.S.Treasury Bills,Sec Mkt,6-Mo.(% Per Ann,Nsa) | $\triangle level$ |
| FYGT1 | Interest Rate: U.S.Treasury Const Maturities,1-Yr.(% Per Ann,Nsa) | $\triangle level$ |
| FYGT5 | Interest Rate: U.S.Treasury Const Maturities,5-Yr.(% Per Ann,Nsa) | $\triangle level$ |
| FYGT10 | Interest Rate: U.S.Treasury Const Maturities,10-Yr.(% Per Ann,Nsa) | $\triangle level$ |
| FYAAAC | Bond Yield: Moody's Aaa Corporate (% Per Annum) | $\triangle level$ |
| FYBAAC | Bond Yield: Moody's Baa Corporate (% Per Annum) | $\triangle level$ |
| SCP90F | Commerical Paper- Federal Funds spread | $level$ |
| SFYGM3 | 3 month Treasury Bill - Federal Funds spread | $level$ |
| SFYGM6 | 6 month Treasury Bill -Federal Funds spread | $level$ |
| SFYGT1 | 1 year Treasury Bond -Federal Funds spread | $level$ |
| SFYGT5 | 5 year Treasury Bond - Federal Funds spread | $level$ |
| SFYGT10 | 10 year Treasury Bond -Federal Funds spread | $level$ |
| SFYAAAC | AAA Corporate - Federal Funds spread | $level$ |
| SFYBAAC | BAA Corporate -Federal Funds spread | $level$ |
| EXRU.S | Ex rate: avg | $\triangle ln$ |
| EXRSW | Ex rate: Switz | $\triangle ln$ |
| EXRJAN | Ex rate: Japan | $\triangle ln$ |
| EXRUK | Ex rate: UK | $\triangle ln$ |
| EXRCAN | EX rate: Canada | $\triangle ln$ |

**Group 7: Prices**

| Variable | Description | Transformation |
|---|---|---|
| PWFSA | Producer Price Index: Finished Goods (82=100,Sa) | $\triangle^2 ln$ |
| PWFCSA | Producer Price Index: Finished Consumer Goods (82=100,Sa) | $\triangle^2 ln$ |
| PWIMSA | Producer Price Index:Intermed Mat.Supplies & Components(82=100,Sa) | $\triangle^2 ln$ |
| PWCMSA | Producer Price Index: Crude Materials (82=100,Sa) | $\triangle^2 ln$ |
| PSCCOM | Spot market price index: bls & crb: all commodities(1967=100) | $\triangle^2 ln$ |
| PW102 | Producer Price Index: Nonferrous Materials (1982=100, Nsa) | $\triangle^2 ln$ |
| PMCP | Napm Commodity Prices Index (Percent) | $level$ |
| PUNEW | Cpi-U: All Items (82-84=100,Sa) | $\triangle^2 ln$ |
| PU83 | Cpi-U: Apparel & Upkeep (82-84=100,Sa) | $\triangle^2 ln$ |
| PU84 | Cpi-U: Transportation (82-84=100,Sa) | $\triangle^2 ln$ |
| PU85 | Cpi-U: Medical Care (82-84=100,Sa) | $\triangle^2 ln$ |
| PUC | Cpi-U: Commodities (82-84=100,Sa) | $\triangle^2 ln$ |
| PUCD | Cpi-U: Durables (82-84=100,Sa) | $\triangle^2 ln$ |
| PUS | Cpi-U: Services (82-84=100,Sa) | $\triangle^2 ln$ |
| PUXF | Cpi-U: All Items Less Food (82-84=100,Sa) | $\triangle^2 ln$ |
| PUXHS | Cpi-U: All Items Less Shelter (82-84=100,Sa) | $\triangle^2 ln$ |
| PUXM | Cpi-U: All Items Less Midical Care (82-84=100,Sa) | $\triangle^2 ln$ |
| GMDC | Pce, Impl Pr Defl:Pce (2000=100) (AC) (BEA) | $\triangle^2 ln$ |
| GMDCD | Pce, Impl Pr Defl:Pce; Durables (2000=100) (AC) (BEA) | $\triangle^2 ln$ |
| GMDCN | Pce, Impl Pr Defl:Pce; Nondurables (2000=100) (AC) (BEA) | $\triangle^2 ln$ |
| GMDCS | Pce, Impl Pr Defl:Pce; Services (2000=100) (AC) (BEA) | $\triangle^2 ln$ |

**Group 8: Stock Market**

| Variable | Description | Transformation |
|---|---|---|
| FSPCOM | S&P's Common Stock Price Index: Composite (1941-43=10) | $\triangle ln$ |
| FSPIN | S&P's Common Stock Price Index: & Industrials (1941-43=10) | $\triangle ln$ |
| FSDXP | S&P's Composite Common Stock: Dividend Yield (% Per Annum) | $\triangle level$ |
| FSPXE | S&P's Composite Common Stock: &Price-Earnings Ratio (%,Nsa) | $\triangle ln$ |

## 8.2 U.S. Monthly Indicators: Small Dataset

The small dataset for U.S. Monthly predictors used consists of the 10 leading indicators from the Conference Board. Since the Leading Credit Index™ is a fairly new metric that does not start as early as our dataset, we use Consumer Credit Outstanding - Nonrevolving as a substitute. We substitute the 5 year Treasury bonds less federal funds for the 10 year treasury bonds less federal funds as we find the former much more predictive for recessions across the 3, 6 and 12 month horizons. All transformations done follow that of the Stock and Watson (2005).

| Conference Board Leading Indicators |
|---|
| 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 |

## 8.3 Japan Monthly Indicators: Large Dataset

We demean rate variables (e.g interest rates, unemployment rate) and detrend quantity variables (index, prices) separating by pre-1991-01-01 variables and post 1991-01-01 variables. We take the level differences of detrended and demeaned variables to remove stationarity. We include the transformation in the table below. Data sources include Global Insight Database, Japan Cabinet Office, Bank of Japan, and FRED.

**Group 0: Export, Import, Trade**

| Description | Variable | Transformation |
|---|---|---|
| Net Trade: Value Goods for Japan | NTVGJ | Detrend, $\triangle level$ |
| Imports: Value Goods for Japan | IVGJ | Detrend, $\triangle level$ |
| Exports: Value Goods for Japan | EVGFJ | Detrend, $\triangle level$ |
| Trade Imports, Raw Materials, Billions of Japanese Yen | JPNVT0060 | Detrend, $\triangle level$ |

**Group 1: Output and Income**

| Description | Variable | Transformation |
|---|---|---|
| New Job offers (Excluding New School Graduates) | NEWJOB | Detrend, $\triangle level$ |
| COMMERCIAL SALES VALUE RETAIL SALES, NSA | STRNSJ | Detrend, $\triangle level$ |
| COMMERCIAL SALES VALUE TOTAL, NSA | JPNRS0076 | Detrend, $\triangle level$ |
| COMMERCIAL SALES VALUE WHOLESALE, NSA | JPNRS0077 | Detrend, $\triangle level$ |
| Manufacturing Production Capacity Index , NSA | JCAPMNSJ | Detrend, $\triangle level$ |
| Production Index: Mining, Manufacturing, Electricity and Gas , SA | JPNQJ0134 | Detrend, $\triangle level$ |
| Industrial Production Index Construction Materials , SA | JQICNJ | Detrend, $\triangle level$ |
| Industrial Production Index Construction Materials Capital Goods, SA | JQIIJ | Detrend, $\triangle level$ |
| Industrial Production Index(2010) Total, SA | JQITOTAL | Detrend, $\triangle level$ |
| Industrial Production Index(2010) Mining & Manufacturing, SA | JQIJ | Detrend, $\triangle level$ |
| Industrial Production Index(2010) Manufacturing, SA | JQIMJ | Detrend, $\triangle level$ |
| Industrial Production- Synthetic Rubbers, NSA | JQISRNSJ | Detrend, $\triangle level$ |

**Group 1: Output and Income**

| Description | Variable | Transformation |
|---|---|---|
| Index Of Producers' Shipments Mining & Manufacturing, SA | JSJ | Detrend, $\triangle level$ |
| Index Of Producers' Inventories Mining & Manufacturing, SA | JINVJ | Detrend, $\triangle level$ |
| Index Of Inventory To Shipments Mining & Manufacturing, SA | JINV%SJ | Detrend, $\triangle level$ |
| Index Of Capacity Utilization Manufacturing, SA | JRKMJ | Detrend, $\triangle level$ |
| Industry Indices of Wholesale and Retail Trade, SA | JPNTI0015 | Detrend, $\triangle level$ |
| Industry Indices of Real Estate, SA | JPNTI0029 | Detrend, $\triangle level$ |
| Indices of Information and Communications, SA | JPNTI0005 | Detrend, $\triangle level$ |
| Indices of Finance and Insurance, SA | JPNTI0021 | Detrend, $\triangle level$ |
| Indices of Electricity, Gas, Heat Supply and Water, SA | JPNTI0003 | Detrend, $\triangle level$ |
| Disposable Income Wage Earners Households, NSA | YDHWNSJ | Detrend, $\triangle level$ |
| Production Value, Electronic Circuit Boards, NSA | JPNIP0009 | Detrend, $\triangle level$ |
| Production Value Semi-Conducters, NSA | QISCNSJ | Detrend, $\triangle level$ |
| Retail Sales Major Department Stores , NSA | STRDPNSJ | Detrend, $\triangle level$ |
| Retail Sales Major Stores , NSA | STRMNSJ | Detrend, $\triangle level$ |
| Savings Wage Earners' Households , NSA | SAVHWNSJ | Detrend, $\triangle level$ |
| Total Income Wage Earners' Households NSA | YHWNSJ | Detrend, $\triangle level$ |
| Wages And Salaries Wage Earners' Households , NSA | WSHWNSJ | Detrend, $\triangle level$ |

**Group 2: Labor Market**

| Description | Variable | Transformation |
|---|---|---|
| EMPLOYED AGRICULTURE, NSA | EAGNSJ | Detrend, $\triangle level$ |
| EMPLOYED CONSTRUCTION, NSA | ECONSJ | Detrend, $\triangle level$ |
| EMPLOYED ELECTRICITY, GAS, HEAT SUPPLY & WATER, NSA | EPUNSJ | Demean, $\triangle level$ |
| EMPLOYED FISHERIES, NSA | EFISNSJ | Detrend, $\triangle level$ |
| EMPLOYED FORESTRY, NSA | EFORNSJ | Detrend, $\triangle level$ |
| EMPLOYED MANUFACTURING, NSA | EMNSJ | Detrend, $\triangle level$ |
| EMPLOYED MINING & QUARRYING OF STONE AND GRAVEL, NSA | EMINSJ | Detrend, $\triangle level$ |
| EMPLOYMENT NEW JOB OPENINGS, SA | EJONJ | Detrend, $\triangle level$ |
| LABOUR FORCE EMPLOYEES, SA | LFEJ | Detrend, $\triangle level$ |
| SELF EMPLOYED WORKERS, NSA | ESEWNSJ | Detrend, $\triangle level$ |
| UNEMPLOYMENT RATE, SA | RUJ | Detrend, $\triangle level$ |
| Monthly Earnings: Manufacturing | MEMFJ | Detrend, $\triangle level$ |
| Monthly Overtime Hours: Manufacturing | MOHMFJ | Detrend, $\triangle level$ |

**Group 3: Housing**

| Description | Variable | Transformation |
|---|---|---|
| Total Floor Area of New Housing Construction Started | NEWHOUSE | Detrend, $\triangle level$ |
| Building Construction, Grand Total, NSA | JPNCSG0004 | Detrend, $\triangle level$ |
| Quick Estimate of Construction Investment, Public Sector, Building, Non-housing | JPNQH0059 | Detrend, $\triangle level$ |
| BUILDING CONSTRUCTION STARTED FLOOR SPACE FOR DWELLINGS | BCSTDNSJ | Detrend, $\triangle level$ |
| BUILDING CONSTRUCTION STARTED NON DWELLING VALUE | BCSTNDNSJ | Detrend, $\triangle level$ |
| Housing Construction Started, SA | JPNQH0037 | Detrend, $\triangle level$ |
| New Construction Starts of Dwellings, Total, NSA | JPNHSS0001 | Detrend, $\triangle level$ |
| New Construction Starts of Dwellings, Total Private | JPNHSS0003 | Detrend, $\triangle level$ |
| New Construction Starts of Dwellings, Total Public | JPNHSS0005 | Detrend, $\triangle level$ |

**Group 4: Consumption, Order and Inventories**

| Description | Variable | Transformation |
|---|---|---|
| Index of Producer's Inventory Ratio of Finished Goods (Final Demand Goods) | IPIRFG | Demean, $\triangle level$ |
| Index of Producer's Inventory Ratio of Finished Goods | IPIRFGMM | Demean, $\triangle level$ |
| Index of Investment Climate (Manufacturing) | INVESTCLIM | Detrend, $\triangle level$ |
| Consumer Confidence Index | CONCONF | Demean, $\triangle level$ |
| Sales Forecast D.I. of Small Businesses | DISB | Demean, $\triangle level$ |
| Ratio of Operating Profits to Total Assets (Manufacturing) | OPTA | Detrend, $\triangle level$ |
| New Orders for Machinery at Constant Prices (Excluding Volatile Orders) | NEWORD | Detrend, $\triangle level$ |
| Wholesale Trade and Retail Sales, big-Scale Retail Store Sales Value, Supermarkets | JPNRS0139 | Detrend, $\triangle level$ |

**Group 5: Money and Credit**

| Description | Variable | Transformation |
|---|---|---|
| The monthly number of corporate bankruptcies | BANKRUPT | Detrend, $\triangle level$ |
| Total Reserves excluding Gold | TOTALRJ | Detrend, $\triangle level$ |
| M2 | M2J | Detrend, $\triangle level$ |
| M1 | M1J | Detrend, $\triangle level$ |

**Group 6: Bond and Exchange rates**

| Description | Variable | Transformation |
|---|---|---|
| Interest Rate Spread (Lending Rate - Deposit Rate) | INTSPREAD | Demean, $\triangle level$ |
| Interest Rates, Government Securities, Treasury Bills | IRGSTBJ | Demean, $\triangle level$ |
| Newly Issued Government Bonds Yield (10 Years) | GB10 | Detrend, $\triangle level$ |
| Tokyo Interbank Offered Rates(3 Months) | IBORATE | Demean, $\triangle level$ |
| Real Narrow Effective Exchange Rate for Japan | RNJPBIS | Detrend, $\triangle level$ |
| Immediate Rates: Less than 24 Hours: Central Bank Rates | CBRL24 | Demean, $\triangle level$ |

**Group 7: Prices**

| Description | Variable | Transformation |
|---|---|---|
| Nikkei Commodity Price Index (42items) | NIKKEICOM | Detrend, $\triangle level$ |
| Domestic Corporate Goods Price Index (2010=100) All Commodities | PPIZNSJ | Detrend, $\triangle level$ |
| Consumer Price Index: Total, All Items for | JAPCPI | Detrend, $\triangle level$ |

**Group 8: Stock Market**

| Description | Variable | Transformation |
|---|---|---|
| Stock Prices (TOPIX) | STOCKPRIC | Detrend, $\triangle level$ |
| Tokyo Stock Exchange, Average Stock Yield, NSA | JPNIR0004 | Detrend, $\triangle level$ |
| Total Share Prices for All Shares for | TSPASJ | Detrend, $\triangle level$ |

**Group 9: TANKAN Business Surveys**

Enterprises surveyed are broken down by Large Enterprises, medium-sized enterprises and small enterprises. Enterprises with capital exceeding 1 billion year or more are considered big, between 100 million yen and 1 billion year is considerd medium, and between 100 million yen and 20 million yen is considered small. Enterprises are categorized as manufacturing and non-manufacturing. Manufacturing is divided into 17 categories and nonmanufacturing into 14 categories. The 14 manufacturing categories include: textiles, lumber and wood products, pulp and paper, chemicals, petroleum and coal products, ceramics stone and clay, iron and steel, nonferrous steel, food and beverages, processed metals, general-purpose & production & business oriented machinery, electrical machinery, transportation machinery, and other manufacturing. The nonmanufacturing categories include: construction, real estate & goods rental & leasing, wholesaling & retailing, transport & postal activities, information communication, electric & gas utilities, services for businesses, services for individuals, accomdations, mining and quarrying of stone and gravel. All the variables in the following table have data from 1975 Q1 to 2014 Q4 and were downloaded from the Bank of Japan. We use linear approximation to convert the quarterly data into monthly data by setting Q1 data as January, Q2 data as April, Q3 data as July, and Q4 data as October and then approximating the diffusion indexes in between those months. For brevity purposes, we do not list the individual variable names of 343 variables but instead provide the number of variables included for each type of TANKAN business survey question. For instance, we have 57 variables or types of surveys on Business Conditions.

| Description | Number of Variables | Transformation |
|---|---|---|
| Business Conditions | 57 | *level* |
| Inventory Level of Finished Goods & Merchandise | 32 | *level* |
| Domestic Supply & Demand Conditions | 32 | *level* |
| Financial Position | 74 | *level* |
| Employment Conditions | 74 | *level* |
| Lending Attitude | 74 | *level* |

## 8.4 Japan Monthly Indicator: Small Dataset

The small dataset consists of 14 Japan cabinet office leading indicators (Japan Cabinet Office) and 12 TANKAN business survey diffusion indices (Bank of Japan). The two datasets are specified below.

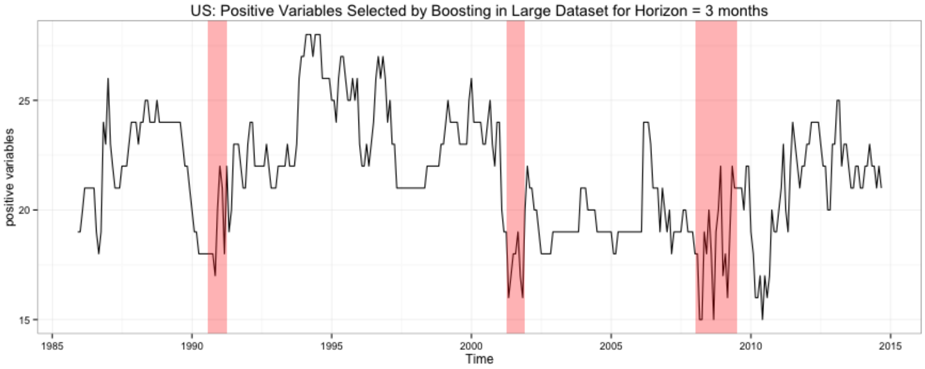
| Cabinet Office Leading Indicators | | |
|---|---|---|
| Description | Variable | Transformation |
| Index of Producer's Inventory Ratio of Finished Goods: Final Demand Goods | IPIRFG | Demean, $\triangle level$ |
| Index of Producer's Inventory Ratio of Finished Goods: Mining and Manufacturing | IPIRFGMM | Demean, $\triangle level$ |
| New Job offers (Excluding New School Graduates) | NEWJOB | Detrend, $\triangle level$ |
| New Orders for Machinery at Constant Prices (Excluding Volatile Orders) | NEWORD | Detrend, $\triangle level$ |
| Total Floor Area of New Housing Construction Started | NEWHOUSE | Detrend, $\triangle level$ |
| Consumer Confidence Index | CONCONF | Demean, $\triangle level$ |
| Nikkei Commodity Price Index (42items) | NIKKEICOM | Detrend, $\triangle level$ |
| Interest Rate Spread | INTSPREAD | Demean, $\triangle level$ |
| Newly Issued Government Bonds Yield (10 Years) | GB10 | Detrend, $\triangle level$ |
| Tokyo Interbank Offered Rates(3 Months) | IBORATE | Demean, $\triangle level$ |
| Stock Prices(TOPIX) | STOCKPRIC | Detrend, $\triangle level$ |
| Index of Investment Climate (Manufacturing) | INVESTCLIM | Detrend, $\triangle level$ |
| Ratio of Operating Profits to Total Assets (Manufacturing) | OPTA | Detrend, $\triangle level$ |
| Sales Forecast D.I. of Small Businesses | DISB | Demean, $\triangle level$ |

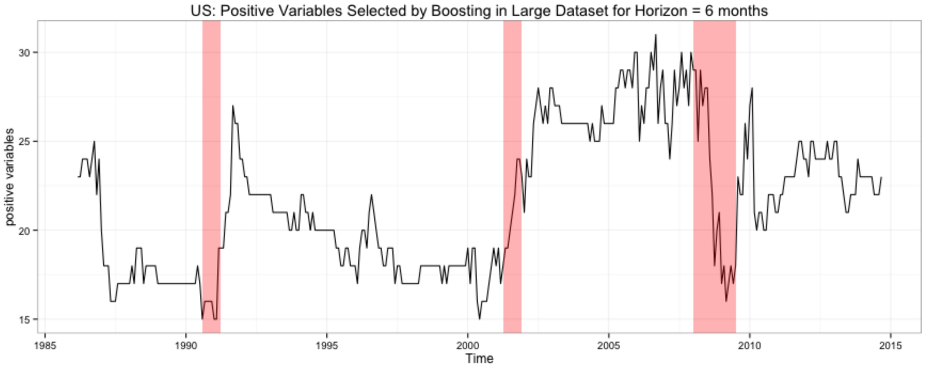
| TANKAN Business Survey | | |
|---|---|---|
| Description | Variable Name | Transformation |
| Business Conditions, All Enterprises, All industries, Actual result | BC_ALL_A | *level* |
| Business Conditions, All Enterprises, All industries, Forecast | BC_ALL_F | *level* |
| Inven Lvl of Finished Goods Merchandise, Actual result | IL_ALL_A | *level* |
| Inven Lvl of Finished Goods Merchandise, Manufacturing, Actual result | IL_MAN_A | *level* |
| Domestic S&D Conditions, All Enterprises, Manufacturing, Actual result | DS_MAN_A | *level* |
| Domestic S&D Conditions, All Enterprises, Manufacturing, Forecast | DS_MAN_F | *level* |
| Financial Position, All Enterprises, All industries, Actual result | FIN_ALL_A | *level* |
| Financial Position, All Enterprises, Manufacturing, Actual result | FIN_MAN_A | *level* |
| Employment Conditions, All Enterprises, All industries, Actual result | EMP_ALL_A | *level* |
| Employment Conditions, All Enterprises, All industries, Forecast | EMP_ALL_F | *level* |
| Employment Conditions, All Enterprises, Manufacturing, Actual result | EMP_MAN_A | *level* |
| Employment Conditions, All Enterprises, Manufacturing, Forecast | EMP_MAN_F | *level* |

## 8.5 How Author Handles U.S. Monthly Data Compared to Ng (2014)

| Variable Description | Action |
|---|---|
| Index of Help-Wanted Advertising in the Newspapers | Drop because series Discontinued |
| Help-Wanted Advertisements:Number of Unemployed | Impute using mean to fill in missing value from 2009:9-2014:9 |
| Employee Hours In Nonag. Establishments | Impule using na.approx to fill in missing values from 1959:2-1964:2 |
| United States; Effective Exchange Rate (Merm) | Combine with Real Broad Effective Exchange Rate for United States |
| Foreign Exchange Rate: Switzerland | Replace with Average of Daily Rates for Switzerland (1957:1-2014:12) |
| Foreign Exchange Rate: Japan | Replace with Average of Daily Rates for (1957:1-2014:1) |
| Spot Market Price Index: BLS & CRB | Replace with Weekly Spot Market Price then average |
| Weekly Repayment Loan Commecial Banks | Impute with mean of 1988-02-01 and 1987-12-01 |

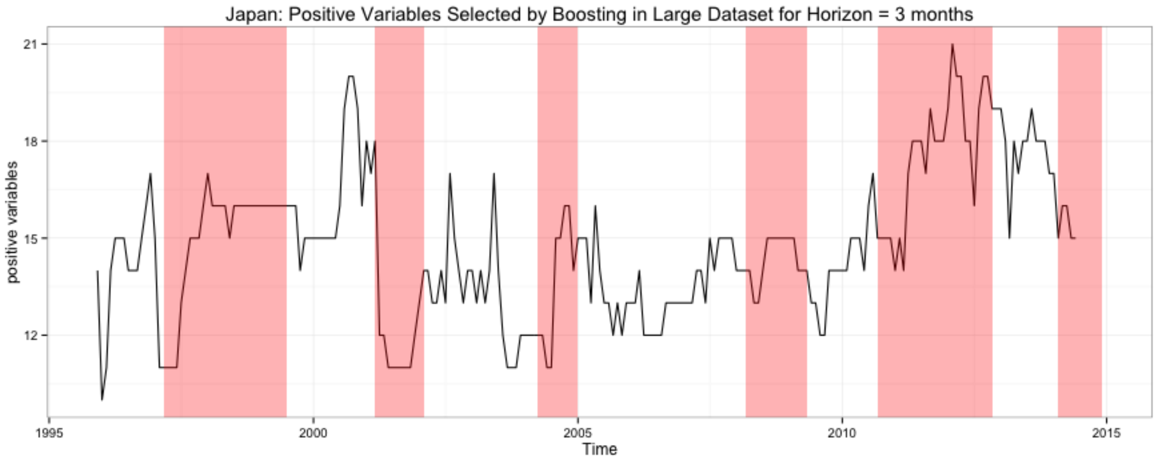
## 8.6 U.S. Number of Positive Variables Selected by Boosting: Out-Of-Sample for 3 month and 6 month

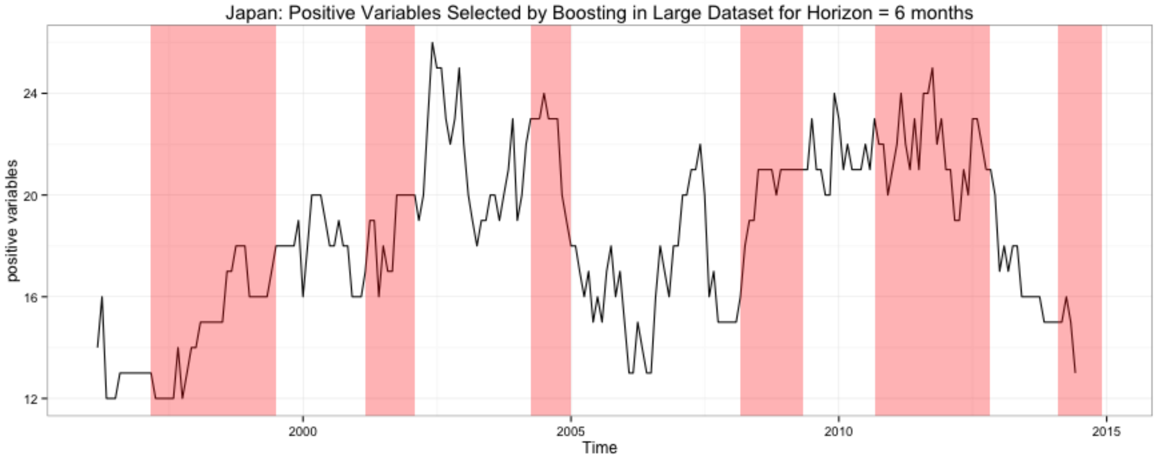

US: Positive Variables Selected by Boosting in Large Dataset for Horizon = 3 months


US: Positive Variables Selected by Boosting in Large Dataset for Horizon = 6 months

We include number of positive variables selected by out-of-sample boosting large datasets for the 3 month and 6 month horizon as we only include the graph of the 12 month horizon in the text.

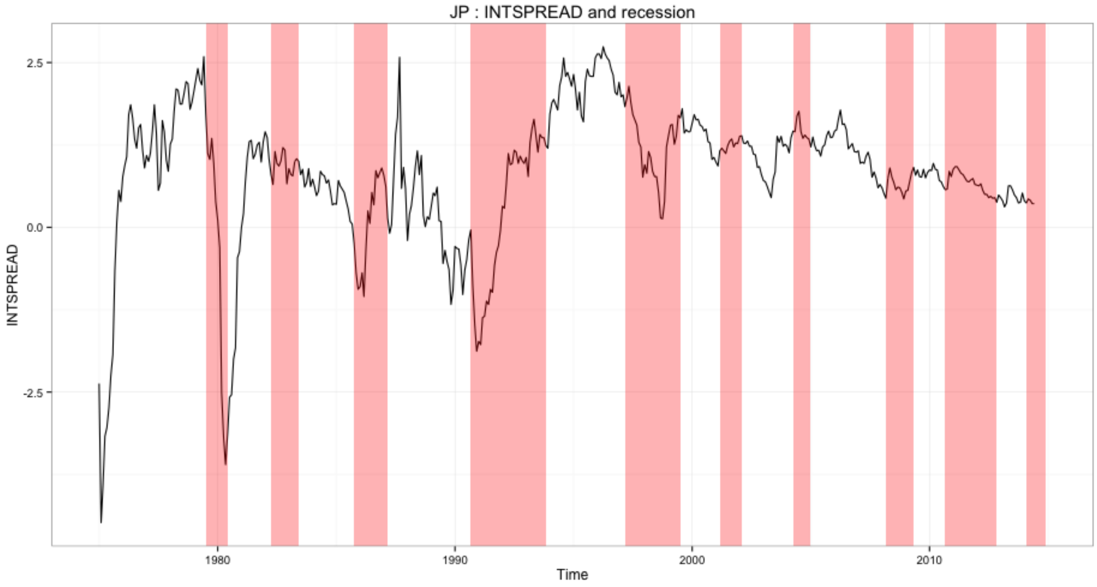
## 8.7 Japan Number of Positive Variables Selected by Boosting: Out-Of-Sample for 3 month and 6 month

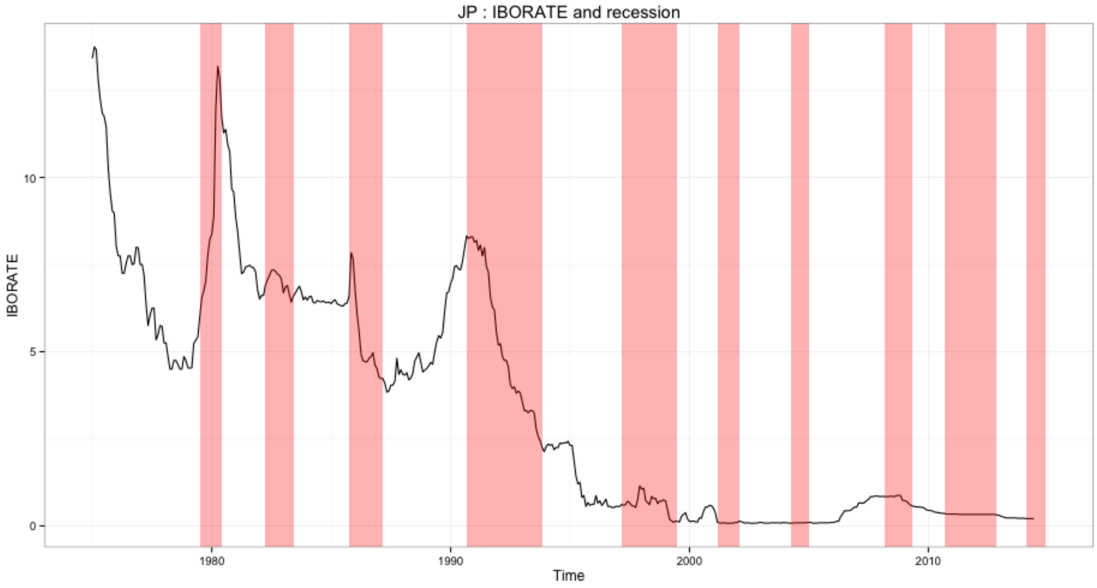




We include number of positive variables selected by out-of-sample boosting large datasets for the 3 month and 6 month horizon as we only include the graph of the 12 month horizon in the text.

## 8.8 Japan Graphs: Raw Data





Notes: Shaded red bars represent recessions. INTSPREAD is the interest rate spread in Japan, specifically the 10 year Government Bonds - Tokyo Interbank 3 Month Rates. IBORATE is the Tokyo Interbank Offered Rates (3 months).

## 8.9 R Code

All code used in this paper can be found at the following link: `https://github.com/jonbma/BoostingRecessions`

# References

Bai, J. and S. Ng (2009). Boosting diffusion indices. *Journal of Applied Econometrics 24*, 607–629.

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

Bernard, H. and S. Gerlach (1998). Does the term structure predict recessions? The international evidence. *International Journal of Finance & Economics 3*, 195–215.

Bry, G. and C. Boschman (1971). Cyclical Analysis of Time Series : Selected Procedures and Computer Programs. *NBER Technical Paper 20*, 13.

Buchen, T. and K. Wohlrabe (2011, October). Forecasting with many predictors: Is boosting a viable alternative? *Economics Letters 113*(1), 16–18.

Burns, A. F. and W. C. Mitchell (1946). *Measuring business cycles*, Volume I.

Christiansen, C., J. N. Eriksen, and S. V. Mø ller (2013). Forecasting US recessions: The role of sentiment.

Estrella, A. and G. A. Hardouvelis (1991). The Term Structure as a Predictor of Real Economic Activity. *The Journal of Finance 46*, 555–576.

Estrella, A. and F. S. Mishkin (1998). Predicting U.S. Recessions: Financial Variables as Leading Indicators.

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

Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting.

Godbout, C. and M. J. Lombardi (2012). Short-Term Forecasting of the Japanese Economy Using Factor Models.

Hanley, J. A. and B. J. McNeil (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology 148*(3), 839–843.

Hasegawa, M. and Y. Fukuta (2011). An empirical analysis of information in the yield spread on future recessions in Japan. *Applied Economics 43*(15), 1865–1881.

Hirata, H. and K. Ueda (1998). The Yield Spread as a Predictor of Japanese Recessions.

Kauppi, H. and P. Saikkonen (2008). Predicting U.S. Recessions with Dynamic Binary Response Models. *The Review of Economics and Statistics 90*, 777–791.

Koenig, E. F., S. Dolmas, and J. Piger (2003). The Use and Abuse of Real-Time Data in Economic Forecasting.

Levanon, G., J.-C. Manini, A. Ozyildirim, B. Schaitkin, and J. Tanchua (2011). Using a Leading Credit Index to Predict Turning Points in the U.S. Business Cycle. *Conference Board 11*(05).

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

Ng, S. (2014). Boosting recessions. *Canadian Journal of Economics 47*(1), 1–34.

Ng, S. and J. H. Wright (2013). Facts and Challenges from the Great Recession for Forecasting and Macroeconomic Modeling. *Journal of Economic Literature 51*, 1120–1154.

Ridgeway, G. (2007). Generalized Boosted Models : A guide to the gbm package. *Compute 1*, 1–12.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics 12*, 77.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*(2), 197–227.

Schapire, R. E. (1999). A brief introduction to boosting. In *IJCAI International Joint Conference on Artificial Intelligence*, Volume 2, pp. 1401–1406.

Stock, J. H. and M. W. Watson (1989). New Indexes of Coincident and Leading Economic Indicators. *NBER Macroeconomics Annual 4*, 351–394.

Stock, J. H. and M. W. Watson (1994). A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series. *Methods No. 6607*, 1–44.

Stock, J. H. and M. W. Watson (2005). Implications of Dynamic Factor Models for VAR Analysis. *NBER Working Paper Series 11467*, 1–67.

Stock, J. H. and M. W. Watson (2006). Forecasting with Many Predictors. *Handbook of Economic Forecasting 1*(05), 515–554.

Wohlrabe, K. and T. Buchen (2014). Assessing the macroeconomic forecasting performance of boosting: Evidence for the United States, the Euro area and Germany. *Journal of Forecasting 33*(May), 231–242.

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