Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Bottom-up leading macroeconomic indicators: An application to non-financial corporate defaults using machine learning

Tyler Pike, Horacio Sapriza, and Tom Zimmermann

2019-070

Please cite this paper as:
Pike, Tyler, Horacio Sapriza, and Tom Zimmermann (2019). "Bottom-up leading macroeconomic indicators: An application to non-financial corporate defaults using machine learning," Finance and Economics Discussion Series 2019-070. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2019.070.

Bottom-up leading macroeconomic indicators: An application to non-financial corporate defaults using machine learning*

Tyler Pike
Federal Reserve Board

Horacio Sapriza
Federal Reserve Board

Tom Zimmermann
University of Cologne

August 30, 2019

Abstract

This paper constructs a leading macroeconomic indicator from microeconomic data using recent machine learning techniques. Using tree-based methods, we estimate probabilities of default for publicly traded non-financial firms in the United States. We then use the cross-section of out-of-sample predicted default probabilities to construct a leading indicator of non-financial corporate health. The index predicts real economic outcomes such as GDP growth and employment up to eight quarters ahead. Impulse responses validate the interpretation of the index as a measure of financial stress.

Keywords: Early Warning Indicators, Corporate Defaults, Machine Learning, Economic Activity
JEL Classification: C53, E32, G33

*The views presented here are solely those of the authors and do not necessarily represent those of the Federal Reserve System.

1 Introduction

Accurate forecasts of macroeconomic activity are central to the decisions of both private firms and public policymakers, yet forecasting aggregate outcomes such as real Gross Domestic Product has been proven to be notoriously difficult (e.g. Faust and Wright (2009)) using aggregate time series data alone. While a vast amount of granular individual or firm-level information could potentially aide forecasting, it is unclear how to best incorporate information from surveys, income statements or stock returns into forecasts of aggregate activity.

In this study, we build a bottom-up indicator of corporate health based on detailed firm-level information and a trained machine-learning model that predicts firm-level default up to eight quarters ahead. We find that our index effectively forecasts future real macroeconomic activity, both in- and out-of-sample, better than well known and widely used aggregate measures of financial conditions, financial stress, and investor sentiment. In addition, our index can be interpreted as a measure of corporate health, where negative shocks to the index are associated with depressed future macroeconomic activity. Overall, our study shows how a large amount of granular information can effectively be used to improve forecasts for macroeconomic aggregates, when such granular information is first summarized by machine learning methods.

In the first part of the study, we use income statement, balance sheet, and stock market variables to estimate the probability of default of each publicly listed non-financial US firm. We compare several modeling approaches, from traditional logistic regressions to more recent approaches such as random forests and gradient boosted trees. We find that the random forest model, an approach that averages predictions from several regression trees (a method that can capture non-linearities and interactions between predictor variables quite well) has the highest accuracy to predict firm defaults out-of-sample. Unlike the widely used logistic regression approach, the random forest model performance is also very stable for predicting defaults at different horizons.

The estimated model gives rise to a cross-sectional probability distribution of defaults that we map into an index of non-financial corporate health using different moments of the estimated distribution. We consider information not only from central tendency moments, but also from higher-order moments of the distribution. In particular, our index is a combination of the mean, variance and skewness of the cross-sectional distribution of estimated default probabilities. We

find that the index correlates well with NBER-dated recessions in the data.

Applying our index to macroeconomic forecasting, we find that it accurately forecasts future real macroeconomic activity. We test its forecasting performance both in- and out-of-sample, and against a set of other measures of financial conditions, financial stress, and investor sentiment widely used in the literature, including the Kansas City Fed FSI, the St. Louis Fed FSI, the Chicago Fed NFCI, the Goldman Sachs FCI, and the excess bond premium (EBP). In a final exercise, we show that our index of corporate health is indeed a reasonable measure of future macroeconomic health: The impulse response of measures of real macroeconomic activity (such as GDP and employment) to shocks to the index is negative.

Our paper connects to two strands of the literature. First, a large literature tries to predict adverse corporate events such as defaults at the firm-level. Early studies started with the analysis of simple ratios and yielded corporate health measures such as the Altman (1968) Z-score, Ohlson (1980) O-score. Methodologically, the literature has gone through a few distinct phases, moving from logistic regressions to predict defaults (Martin (1977)), to neural networks in the early 1990s (Odom and Sharda (1990), Wilson and Sharda (1994), and for a thorough literature survey see Atiya (2001)), to support vector machines in the early 2000s (Härdle et al. (2007), Shin et al. (2005), and see Ravi Kumar and Ravi (2007) for a survey of literature at the time), to tree-based methods more recently. Jones et al. (2015) conduct a thorough review and evaluation of classification techniques used to predict bankruptcy and credit downgrades, and find that tree methods, such as tree-based gradient boosting and random forest, are the best predictors of firm credit changes. The first part of our study is in the spirit of Jones et al. (2015), but while these authors explore a larger number of binary classifier techniques, we focus on five algorithms that are representative of the different waves in the default prediction literature. Moreover, while Jones et al. (2015) only compare algorithms based on their performance in a single cross-section analysis, we evaluate algorithms *recursively* through *time* (1989 through 2018) and across *several forecasting horizons* (1 through 8 quarters ahead), leading to a more rigorous comparison of techniques.

Second, another body of the literature considers the implications of corporate default risk for macroeconomic activity. For instance, several authors have analyzed the implications of corporate distress and the probability of default for explaining risk premiums in stock and bond returns (see for example Fama and French (1996) or Merton (1973)). However, much evidence suggests

that deteriorating firm health, as measured as the probability of default, results in higher than average stock return volatility and lower excess returns (see Campbell et al. (2008) or Chava and Purnanandam (2010)). In addition, recent studies have taken an interest in exploring the impacts of corporate bond market crises on macroeconomic activity, finding that broad corporate debt defaults historically do not have the same real effects as banking sector crises (see Giesecke et al. (2012)). Unlike these studies, we concretely link our improvements in measuring firm defaults to future macroeconomic conditions.

The rest of the paper is organized as follows. Section 2 describes the data used in this study, Section 4 describes the index construction from bottom-up machine learning models, Section 5 evaluates the forecasting abilities of the constructed index and investigates the reaction of real macroeconomic outcomes to structural shocks to the index. Section 6 concludes.

2 Data

2.1 Firm-level data

We estimate the probability of firm-level default using three sets of historical data that track non-financial defaults and bankruptcies. Prior to 1987:Q1 we use the UCLA-LoPucki bankruptcy research database and NYU's Altman Default Database. Post 1987:Q1 we use the Mergent corporate FISD daily feed. A default may occur at the time a company declares bankruptcy in court filings, or if it simply ceases to pay interest due on its corporate securities (i.e., bonds). All three databases denote a default event as the filing date when a company enters Chapter 7 or Chapter 11 court proceedings. These databases combined provide default coverage from 1984:Q1 through 2018:Q1[1]. We restrict our analysis to defaults of publicly traded companies, as our models require inputs from firm balance sheet data.

Figure 1, panel (a), shows that the total number of defaults covered in this time period is 407, an average of 12 defaults per year. The maximum number of defaults in a given quarter occurs in 2001:Q2, approximately 30 defaults. Our panel data tracks all publicly traded firms in US stock markets. Panel (b) of Figure 1 shows that the number of public firms in the US increases steadily

[1]We have additionally conducted all analyses using only the Mergent data starting in 1987, and we find that all of our presented results are robust.

until 1998, and declines steadily afterwards. The number of firms in our dataset never drops below 6000, and as a result, the percent of firms in our set that are labeled as defaulting is less than one-percent for any given quarter. While 407 defaults is not the population of defaults during this time period, Figure 1 suggests that there are sufficiently many defaults to clearly discern business cycles in the data, with peaks occurring in 2001 and 2008.

For our default models, we use a set of 26 firm-level variables. As it is not the goal of this paper to identify a *new* set of variables for predicting non-financial defaults, we use variables previously identified in the corporate finance literature. Our explanatory variables are: Tobin's Q (Tobin and Brainard (1976)), the Kaplan-Zingales Index (Kaplan and Zingales (1997)), firm age and the HP index (Hadlock and Pierce (2010)), a dividend dummy variable and the Whited-Wu index (Whited and Wu (2005)), Altman's Z-score (Altman (1968)), measures of net income, total liabilities, cash, stock price, market equity (Campbell, Hilscher and Szilagyi (2008)), measures of working capital, current ratio, asset turnover, return on equity, EBIT cover, and CAPEX (Jones, Johnstone, and Wilson (2015)), as well as excess returns, average excess returns, 3-digit sic industry dummy variable, industry sales, and industry sales growth. We provide a full description and summary statistics of all variables in tables 3 and 4 in the Appendix A. All balance sheet and income statement data are from Compustat, while stock market data for the excess return variables are from CRSP.

2.2 Macroeconomic data

When evaluating our index of non-financial corporate health, we will relate it to several measures of real macroeconomic activity: Nonfarm payroll employment and the U3 unemployment rate, both from the US Bureau Labor Statistics, an index of Industrial Production from the Federal Reserve Board (accessible online via the FRED website) and Real Gross Domestic Product from the US Bureau of Economic Analysis. To compare the forecasting performance of our index to various financial variables and financial conditions and stress indicators, we use two sets of variables: The first set includes the three-month minus ten-year Treasury bond yield spread, with both yields calculated according to Dahlquist and Svensson (1996), and Moody's seasoned Baa ten-year corporate bond yield minus the ten-year Treasury bond yield spread. The second set of measures

includes well-known financial conditions and financial stress indexes, specifically the Chicago Fed NFCI, the St. Louis FSI, the Kansas City Fed FSI, and the EBP. All financial data was accessed through FRED, save for the term spread and the EBP which were provided by the Federal Reserve Board.

3 Firm-level probability of default

We assume that the probability of a firm default event between time t and $t+h$ takes the functional form:

$$P(I_{i,t,t+h} = 1) = f(X_{i,t}; \theta) \tag{1}$$

where $I_{i,t,t+h}$ is the indicator function that is 1 if firm i defaults between times t and $t+h$, f is a linear or non-linear function of the firm-level variables contained in $X_{i,t}$, and θ is a vector of model parameters that needs to be estimated.

The classical approach to estimating equation (1) is via a logistic regression. However, there are two primary drawbacks to doing the classic approach in this setting. First, covariates typically enter linearly into a logistic regression, imposing a linear structure on what may be a non-linear relationship. Non-selectively including many non-linearities and interactions, on the other hand, easily leads to a model that fits too closely to the estimation data and that does not generalize well, leading for instance to poor out-of-sample forecasts. Second, logistic regressions are not very robust to outliers and to sparse data. These two drawbacks may be acceptable in some settings and may be traded off against a very interpretable model, but it is clear that the determinants of firm health arise as several non-linear relationships, and as we are working with firm-level data, there are unavoidable data restrictions and outliers.

The literature on default probabilities, and more generally the studies on firm stress-events, have primarily relied upon the logistic regression to estimate default probabilities, although some attempts were made in the 1990s and early 2000s to use methods from the machine learning literature (such as neural networks or support vector machines). Given the historical context, and motivated by recent findings (see e.g. Jones et al. (2015)) on the benefits of using machine learning techniques to estimate equation (1) that may circumvent these pitfalls, we estimate equation (1)

by both logistic regression and several modern machine learning algorithms, including artificial neural networks, support vector machines, random forests, and tree-based gradient boosting machines.

We run two empirical exercises. The first exercise is an in-sample estimation of firm-level defaults within h quarters. The second exercise is a recursive estimation of the defaults, where we estimate the model with data up to some period t and then we estimate the probability of default within the next h quarters. We evaluate the predictions on the *area under the receiver operating curve*, a standard measure to assess the quality of different classification models (see e.g. Drehmann and Juselius (2014)). Our evaluation of these techniques through time and across different forecast horizons is novel to the default probability literature, and further details regarding the exercises and techniques are outlined in appendix B.

When we fit the models with all the data, in-sample, we find that the random forest dominates all other classification techniques. Figure 2 shows the random forest achieves an in-sample AUC greater than 0.99 for all forecast horizons tested and appears to increase as the forecast horizon increases. In comparison, all other techniques, except for the support vector machine, appear to decrease their AUC as the forecast horizon increases. Further, we find that when examining the variables of importance in each model, both the logistic regression and random forest put most weight on measures of leverage, in other words, financial conditions.

When the models are evaluated recursively, out-of-sample, we again find that the random forest dominates all other classification techniques[2]. Figure 2 shows that the random forest achieves an in-sample AUC greater than 0.96 for all the forecast horizons being tested, and a maximum AUC of approximately 0.985. However, as we evaluate the random forest in real time, we find that its AUC peaks at forecast horizons between two and five quarters-ahead.

We find that the random forest algorithm outperforms all other techniques at all forecast horizons, both in- and out-of-sample. While the tree-based gradient boosting machine (a similar algorithm to the random forest) is somewhat worse than the random forest, it outperforms the logistic regression, artificial neural network, and support vector machine algorithms. Therefore, we construct our index using predictions from the random forest algorithm, our best forecasting

[2]The out-of-sample logistic regression exercise used data winsorized at the 5th and 95th percentiles due to the extremely poor performance of the model without the data pre-processing.

model.

4 The index of non-financial corporate sector health

Our index is based on a bottom-up approach. We first estimate firm-level defaults, and then we aggregate the out-of-sample predictions to create an index of non-financial corporate sector health. We discuss each step below.

We use the cross-section of default probabilities predicted with the random forest algorithm to construct our measure of non-financial corporate sector health. We construct the index from three components that reflect different moments of the cross-sectional distribution of predicted defaults: We use asset-weighted observations to derive the weighted mean, so that defaults affecting larger corporations are considered to be more informative about the health of the sector. We allow for the possibility that the shape of the cross-sectional distribution contains additional relevant information, so we also consider the standard deviation and the skewness of the distribution. The current prevailing view in the forecast combination literature is that a simple unweighted averaging of forecasts will outperform more complicated forecast combination techniques (for a textbook treatment of the topic, see Elliott and Timmermann (2016), and for an applied demonstration, see Stock and Watson (2004)). Following the forecast combination literature, we create an aggregate index of corporate health by taking the unweighted average of the sub-indexes, i.e., at each time t, our index is the simple average of the weighted mean, the unweighted standard deviation and the unweighted skewness of the predicted default probabilities distribution.

Starting with the weighted mean, we construct the first component of the index as

$$NFCH^{m}_{t,h} = \frac{1}{n_t} \sum_{i=1}^{n_t} w_{i,t} p_{i,t,h} \tag{2}$$

$$w_{i,t} = \sum_{j=1}^{n_t} \frac{A_{i,t}}{A_{j,t}}$$

where i indexes the firm, t the time, n is the number of firms in the sample at time t, $w_{i,t}$ is the firm's cross-sectional weight determined by total assets, A, and $p_{i,t,h}$ is the firm's probability of default within h quarters from time t.

Figure 3 depicts the four-quarter simple moving average of the $NFCH^m$ constructed to forecast events up to eight-quarters ahead[3]. We use an 8-quarters ahead time horizon because 8 quarters provides an earlier warning sign of adverse business conditions than any other tested time-horizon[4]. The $NFCH^m$ has the desirable quality of rising prior to the 2001 and 2008 recessions. Also, there is a greater increase in the index prior to the 2001 rather than the 2008 recession. This may respond to the fact that the run-up to the 2001 recession was characterized by over-investment in the information technology sector, giving rise to a slew of weak firms entering the sector, while the 2008 recession was brought on by a crisis in the housing and financial sector, not weakness of non-financial corporate firms.

Note that by weighting a firm's probability of default by the firm's share of assets, we largely mute the effects of small firm distress. However, given that small firms may be more capital constrained and therefore less able to weather deteriorating economic conditions, the probability of default of small firms may provide valuable information to forecast real economic activity. Furthermore, while we note that the mean is the most efficient measure of central tendency (for the normal distribution), its does not capture information concerning the dispersion or symmetry of the distribution. However, it is documented that higher-order moments of firm stock returns vary predictably over the business cycle (for example Alles and Kling (1994)), suggesting that moments describing the tails of a distribution, i.e., standard deviation and skew, may lend valuable information regarding the current and future state of economic activity. With these motivations in mind, we next construct an index by calculating the mean of the quarter-over-quarter[5] difference in the unweighted standard deviation and quarter-over-quarter difference in the unweighted skew at time t:

$$NFCH_t^s = \frac{1}{2}(\gamma_t + \sigma_t) \tag{3}$$

where t is the time index, γ is the quarter-over-quarter difference of the cross-sectional skew, and σ is the quarter-over-quarter difference of the cross-sectional standard deviation.

[3]We find our results are generally robust to smoothing or not smoothing the series, we choose to present the smoothed series, as it reduces noise, making business cycles more discernible and maximizes its in- and our-of-sample forecasting ability.

[4]We only present indexes constructed using $h = 8$, and as a constant, it will be dropped from further index notation.

[5]We find our results to be robust to constructing the $NFCH^s$ with the first difference of the moments, however, as with the $NFCH^m$, we find that using the first difference of the moments makes business cycles more discernible and maximizes the forecasting ability of the index.

Figure 4 shows the dispersion index $NFCH^s$. The second index moves procyclically, rising prior to all three recessions, 1991, 2001, and 2008, with a maximum achieved one quarter prior to the 2008 recession. Such behavior suggests that this index may contain useful information for forecasting real economic activity.

To capitalize on the information content of both indexes, we take the average of the two and construct an ensemble index as:

$$NFCH_t = \frac{1}{2}(NFCH_t^m + NFCH_t^s). \tag{4}$$

Figure 5 shows the NFCH, standardized with mean zero and standard deviation one. As shown in the figure, the index experiences large increases prior to recessions, especially for the 2001 and 2008 episodes. When the index rises two standard deviations above the historical mean, there is a recession within two quarters. The fact that the NFCH distinctly rises prior recessions suggests that it can be a valuable tool for measuring stress in the economy. In the next section we investigate the ability of the index to forecast future real economic activity more generally.

5 Applications to real macroeconomic activity

We next validate the use of our bottom-up non-financial corporate health index as a leading indicator of real macroeconomic activity through in- and out-of-sample forecasting exercises considering a number of widely used measures of real economic activity, and by performing several impulse response exercises with these measures following the direct projection approach by Jordà (2005).

5.1 In-sample forecasting ability

To determine the efficacy of our non-financial corporate health index as a leading indicator of economic activity, we run an in-sample forecasting test, as in Gilchrist and Zakrajšek (2012), considering payroll employment, real GDP, industrial production, and the unemployment rate as

measures of economic activity. The forecast specification is:

$$\nabla^h Y_{t+h} = \alpha + \sum_{i=1}^{4} \beta \nabla^1 Y_{t-i} + \delta_1 TS_t + \delta_2 RFF_t + \delta_4 NFCH(t,h) + \epsilon_{t+h} \tag{5}$$

where t indexes time, h is the forecast horizon, $NFCH(t,h)$ is our index of interest, $\nabla^h Y_{t+h} \coloneqq \frac{400}{h+1} \log(\frac{Y_{t+h}}{Y_{t-i}})$, Y is the measure of real economic activity. In addition to lagged growth values of the dependent variable, we control for the stance of monetary policy by including the term spread, TS, between the constant maturity three-month and ten-year Treasury yield, and the real federal funds rate, RFF.

We measure the forecasting accuracy in terms of adjusted R^2. Furthermore, we compare our index to the Excess Bond Premium, EBP, an information rich sentiment or risk appetite indicator constructed in Gilchrist and Zakrajšek (2012), based on the credit spread of non-financial corporate firms, and a financial conditions index, the Chicago Fed $NFCI$ (see Brave and Butters (2011)).

Table 1 displays the in-sample forecast exercise results. The NFCH is statistically significant at the one-percent level in all model specifications. Moreover, the NFCH has a larger adjusted R^2 than any other index when forecasting real GDP and industrial production. All the forecast coefficients associated with the NFCH have the expected sign, i.e., positive for payroll employment, real GDP, and industrial production, and negative for unemployment. Finally, when all four indexes are included in the forecasting exercises, the NFCH remains highly significant, while the Chicago Fed NFCI becomes insignificant across all exercises. The EBP becomes statistically insignificant when forecasting real GDP.

5.2 Out-of-sample forecasting ability

To determine the efficacy of the NFCH as a leading indicator of US economic conditions and assess its potential value for practitioners and policymakers, we supplement our in-sample forecasting tests with an out-of-sample forecasting exercise.

To assess the ability of the NFCH to forecast real macroeconomic activity in real-time, we compare the forecast errors generated by a baseline autoregressive (AR) model and a similar index-augmented AR model. We use a model specification similar to that of our in-sample exercise, but we now drop the term-spread and real federal funds rate as controls, yielding the baseline

model:[6]

$$Y_{t+h} - Y_{t-1} = \alpha + \sum_{i=1}^{4} \beta(Y_{t-i} - Y_{t-i-1}) + \epsilon_{t+h} \tag{6}$$

and the index-augmented model:

$$Y_{t+h} - Y_{t-1} = \alpha + \sum_{i=1}^{4} \beta(Y_{t-i} - Y_{t-i-1}) + \gamma Index + \epsilon_{t+h} \tag{7}$$

We use a four-quarter forecast horizon. All data start in 1989 so that all indexes produce the same number of forecast estimates. The forecasts are built using an expanding, recursive, window, with the first forecast estimation using 5 years of data.

We test the Term Spread, Corporate Bond Spread (Baa - 10-year Treasury yield), Chicago Fed NFCI, EBP, Kansas City Fed FSI, St. Louis Fed FSI, Goldman Sachs FCI, and our NFCH. As the firm-level default models heavily load on measures of corporate leverage (see appendix B), one can think of the NFCH as a function of corporate financing conditions. The Term Spread, Corporate Bond Spread, Chicago Fed NFCI, EBP, Kansas City Fed FSI, St. Louis Fed FSI, and Goldman Sachs FCI, are all indexes that attempt to characterize financing conditions, stress, or investors' sentiment. Hence, we consider it appropriate to compare the NFCH to these indexes.

Each index's ability to improve forecasts of real macroeconomic activity is measured by constructing the ratio of forecast errors such that the RMSE of equation (7) is divided by the RMSE of equation (6). Therefore, a forecast error ratio less than one is interpreted as indicating that an index improves forecasting ability. To further evaluate the difference between the two sets of errors, we also compute the Diebold and Mariano (1995) forecast error statistic and present the tests' p-values. Additionally, given the sample bias present in testing nested models, we also include significance levels according to Clark and West (2007).

Table 2 presents the out-of-sample forecasting results. Ratios less than one (indicating the tested index improves forecasting ability) are in bold, while Diebold-Mariano p-values significant at the ten-percent confidence level (indicating the index augmented model produces smaller absolute forecast errors) are starred. It is clear that the NFCH has a forecast-error ratio of less than one for all measures of real macroeconomic activity. That is, the NFCH helps to forecast

[6]We drop the term-spread and real federal funds rate because we want to use the term-spread on its own as an indicator of future real activity, and several of the financial conditions indexes that we evaluate in the exercise are built using both the term-spread and real federal funds rate.

all tested measures of real macroeconomic activity out-of-sample. Note as well that the NFCH especially improves the Payroll Employment forecasts by approximately six-percent, with the improvement being statistically significant at the ten-percent level according to the Diebold-Mariano test, and at the five-percent level according to the Clark-West test. The EBP improves the forecast-error ratio for every measure of real macroeconomic activity, but less so than the NFCH for industrial production, and the EBP's forecast improvement for payroll employment is not statistically significant. The Term Spread has a forecast error ratio less than one for the unemployment rate and payroll employment, but it does not improve our ability to forecast GDP or industrial production. The only other indexes with a forecast error ratio less than one are the Kansas City Fed and St. Louis Fed FSIs, but these only help forecast the unemployment rate.

Overall, we find that the NFCH effectively forecasts payroll employment, industrial production, GDP, and the unemployment rate, out-of-sample. Finally, we test several well-known indicators of financial conditions, stress, and investor sentiment, and we find that the only indicator that effectively forecasts all the measures of real macroeconomic activity that the NFCH can forecast, is the EBP.

5.3 Impulse responses

To further evaluate the macroeconomic implications of a given rise or fall in the $NFCH$ and the effects of real economy shocks on the $NFCH$, we construct impulse response functions (IRFs) via the local projections method outlined in Jordà (2005). Impulse responses are estimated via direct local projection:

$$Y_{t+h} - Y_{t-1} = \alpha_h + \sum_{i=1}^{8} \mathbf{M}_{t-i}\overline{\rho}_i \tag{8}$$

where t indexes time, h is the horizon index, and $\mathbf{M}$ is a matrix of real GDP, industrial production, unemployment rate, real federal funds rate, and the NFCH[7]. Note that every column in $\mathbf{M}$ has been standardized to ease the interpretation of the impulse response functions.

First we simulate a positive shock to the $NFCH$ and review its effects on the four measures of economic activity used in the previous exercises. Figure 6 shows the impulse responses to a one standard deviation shock to the $NFCH$.

[7]We use annualized quarter-over-quarter changes of real GDP and industrial production

An increase in the $NFCH$ (a decrease in non-financial corporate health) leads to a decrease in the real federal funds rate. This result is intuitive, as it suggest that as firms' health deteriorates, a recession or slowdown may become more likely, and policy-makers may decrease interest rates to stimulate activity. A similar effect is visible with Real GDP growth. Within eight quarters after an increase to the $NFCH$, real GDP growth decreases by one standard deviation, and the decline is statistically significant at the ten-percent level. The unemployment rate increases by approximately one standard deviation, which is statistically significant at the ten-percent level. Lastly, industrial production growth falls, reaches its minimum of approximately -1.25 standard deviations six quarters after a shock to the $NFCH$, and then begins to return to its historical mean.

The results of the impulse response functions suggest that the NFCH is procyclical, and that it is an effective leading indicator of stress in the real economy.

Reversing our exercise, we simulate a positive shock to each of the individual measures of real economic activity and investigate their effects on the $NFCH$. Figure 7 shows the impulse response functions of the $NFCH$ to a positive one standard deviation shock to measures of real economic activity.

If there is a positive shock to interest rates, then borrowing costs increase and consumers and businesses spend less, so corporate health tends to weaken. This is consistent with the impulse response function, which shows that an increase in interest rates increases the $NFCH$ growth rate by approximately 2 standard deviations. For the rest of the variables, real GDP growth, the unemployment rate, and the industrial production growth rate, the NFCH remains relatively unchanged, suggesting that it takes a broad economic shock to deteriorate overall firms' health. That is, a weaker labor market alone will not deteriorate firm health by a statistically significant margin, and a large increase in GDP growth alone may not act as a powerful enough tail wind to significantly improve the health of non-financial corporate firms. Rather, it may take several factors together to generate a persistent change in overall non-financial corporate health.

6 Conclusion

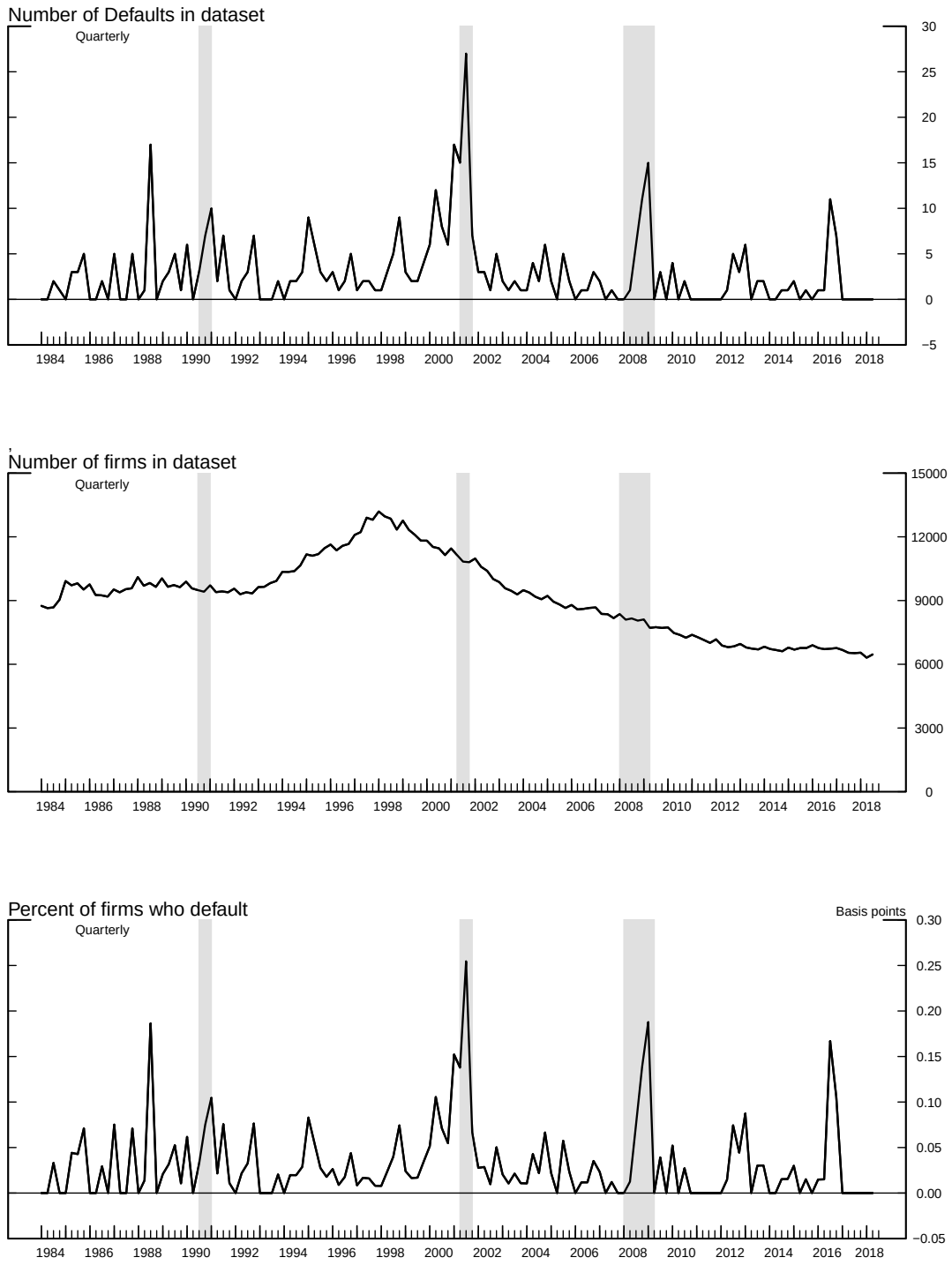
In this paper we built a bottom-up indicator of non-financial corporate financial stress using machine learning techniques. We find that these techniques more accurately measure the probability

of non-financial firm default than traditional models, and we identify moments of the distribution of predicted default probabilities to build an aggregate early warning indicator of corporate stress and macroeconomic conditions.

There are two primary contributions of our study. First, we addressed the various weaknesses of the logistic regression in predicting non-financial firm defaults, both its restrictive linear structure and inability to adapt to sparse data and outliers, by identifying a number of nonstandard statistical techniques that accurately predict future non-financial firm defaults. Using a set of standard balance sheet and income statement variables, we find the random forest to be the most accurate technique, both in and out-of-sample.

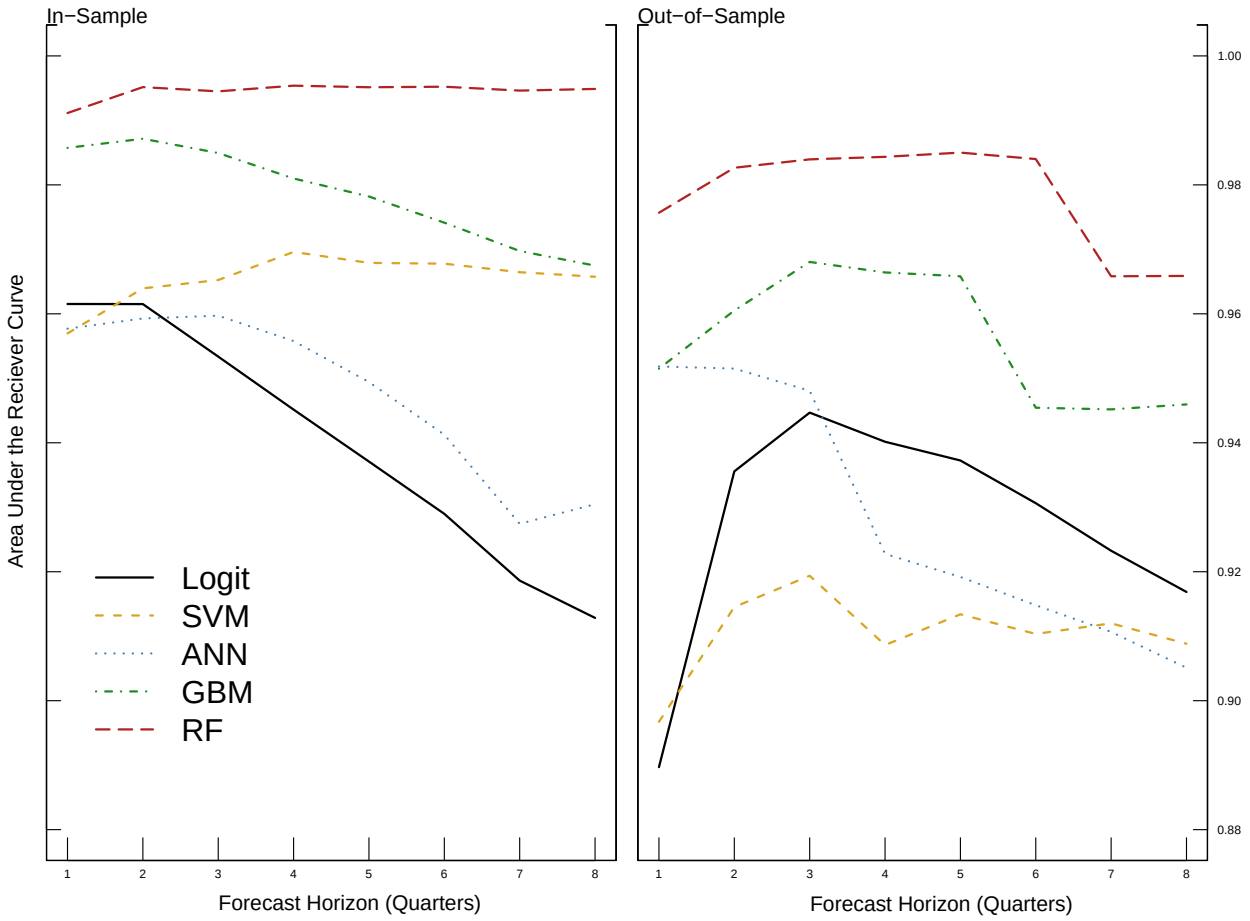
Second, we leveraged the improved accuracy in measuring default probabilities by constructing an information rich macroeconomic index. The macroeconomic index is an aggregation of micro-level default probabilities, specifically utilizing the first three moments of the firm default probability distribution. The indicator is highly effective in predicting future industrial production, payroll employment, real GDP, and the unemployment rate, outperforming the Chicago NFCI and Excess Bond Premium, both in- and out-of-sample. We further validate our index as being an accurate and useful measure of US non-financial corporate health by conducting direct projection impulse response exercises, where positive shocks to the non-financial corporate stress index elicit future downturns in macroeconomic conditions.

Figure 1: *Firm-level data over time*



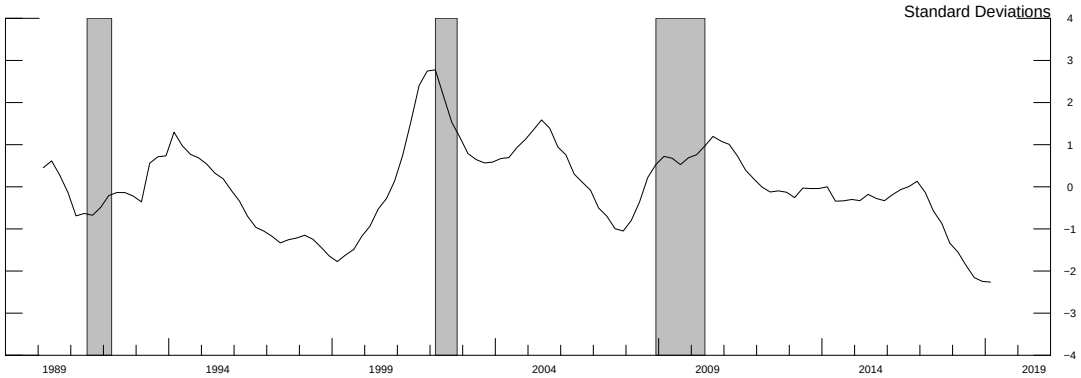
All data is quarterly. A firm is considered to have defaulted when it files for chapter 7 or chapter 11 bankruptcy. All bankruptcy filing dates come from a combination of the UCLA-LoPucki bankruptcy research database, NYU's Altman default database, and Mergent's corporate FISD daily feed. The universe of firms is the CRSP and COMPUSTAT set of US public firms.

Figure 2: *AUC by Machine Learning and Logistic Algorithms*



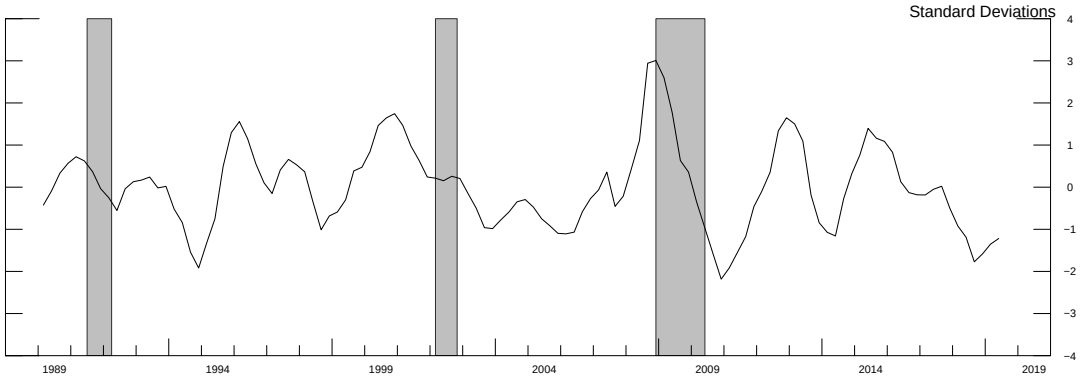
The area under the receiver curve (AUC) across 1- through 8-quarter ahead horizons, obtained with the machine learning and logistic algorithms. A score of one denotes a perfect classifier. The out-of-sample logistic regression was trained on winsorized data, at the 5th and 95th percentiles, due to its extremely poor performance using the uncleaned data.

Figure 3: *Level index, $NFCH^m$, derived from predicted corporate default probability distribution*



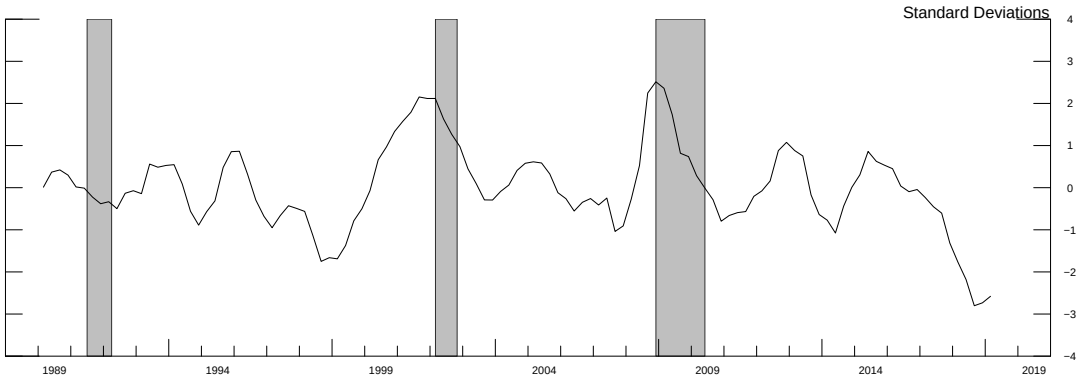
The weighted mean index, $NFCH^m$, is the 4-quarter moving average of the weighed mean of the probabilities of default within 8-quarters, predicted by the RF algorithm. The probability of default of a given firm is weighted by the firm's share of the assets of all firms in the sample. The index has been standardized to have mean 0 and standard deviation 1.

Figure 4: *Dispersion index, $NFCH^s$, derived from predicted corporate default probability distribution*



The Dispersion Index, $NFCH^s$, is average of the QoQ change in skewness and standard deviation of the probabilities of default within 8-quarters, predicted by the RF algorithm. The index has been normalized to have mean 0 and standard deviation 1.

Figure 5: *Ensemble index, NFCH, constructed as the average of $NFCH^m$ and $NFCH^s$*



The ensemble index, NFCH, is the unweighted average the $NFCH^m$ and $NFCH^s$. The index has been normalized to have mean 0 and standard deviation 1.

Table 1: *In-Sample Forecast Results*

Payroll Employment

EBP	-0.830^{***}			-0.612^{***}		-0.388^{***}
	(0.159)			(0.131)		(0.144)
NFCI		-0.851^{***}			-0.606^{***}	-0.033
		(0.211)			(0.204)	(0.169)
NFCD			-0.693^{***}	-0.491^{***}	-0.517^{***}	-0.369^{***}
			(0.178)	(0.169)	(0.142)	(0.117)
Observations	100	100	100	100	100	100
Adjusted R^2	0.692	0.668	0.681	0.765	0.749	0.802

Industrial Production

EBP	-2.944^{***}			-1.888^{***}		-1.164^{***}
	(0.707)			(0.347)		(0.439)
NFCI		-2.595^{**}			-1.639^{**}	0.414
		(1.002)			(0.687)	(0.665)
NFCD			-2.926^{***}	-2.373^{***}	-2.496^{***}	-1.914^{***}
			(0.880)	(0.740)	(0.590))	(0.499)
Observations	100	100	100	100	100	100
Adjusted R^2	0.395	0.347	0.519	0.603	0.586	0.678

Real GDP

EBP	-0.850^{***}			-0.477^{***}		-0.153
	(0.192)			(0.159)		(0.169)
NFCI		-0.940^{***}			-0.562^{***}	-0.064
		(0.212)			(0.194)	(0.225)
NFCD			-0.989^{***}	-0.820^{***}	-0.803^{***}	-0.644^{***}
			(0.246)	(0.248)	(0.213)	(0.176)
Observations	100	100	100	100	100	100
Adjusted R^2	0.347	0.369	0.501	0.551	0.565	0.613

Unemployment

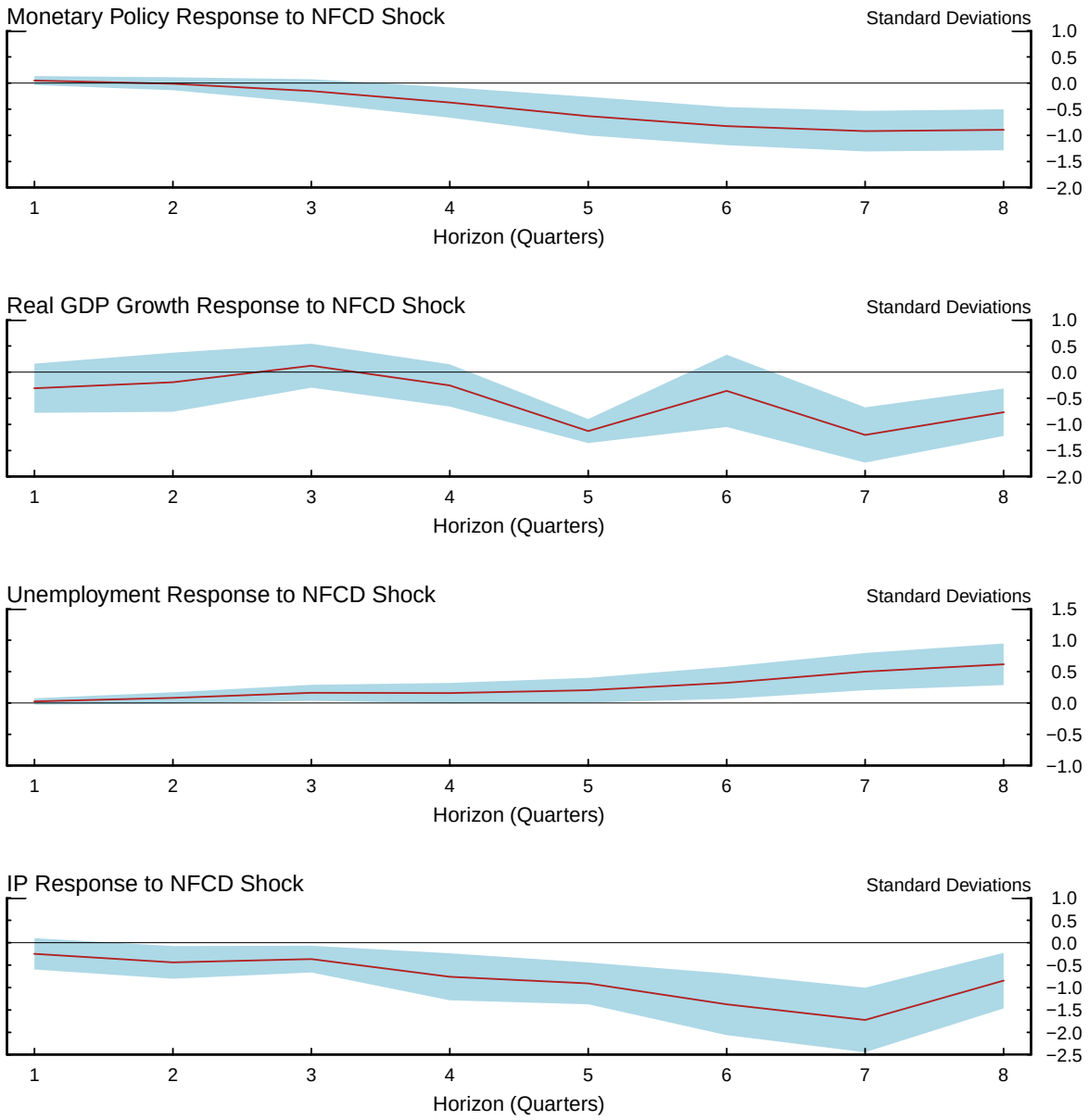
EBP	8.191^{***}			5.467^{***}		3.578^{***}
	(0.985)			(0.844)		(1.185)
NFCI		8.752^{***}			5.356^{***}	-1.517
		(2.385)			(1.813)	(1.615)
NFCD			7.651^{***}	5.723^{**}	6.072^{***}	4.264^{***}
			(2.082)	(2.261)	(1.579)	(0.900)
Observations	100	100	100	100	100	100
Adjusted R^2	0.587	0.548	0.624	0.697	0.673	0.767

Table 2: *Out-of-Sample Forecast Results*

Real Activity	Index	Forecast Error Ratio	Diebold Mariano P-value	Clark-West P-value
Payroll Employment	Kansas FSI	1.004	0.518	
Payroll Employment	Chicago NFCI	1.05	0.648	
Payroll Employment	St. Louis FSI	1.031	0.694	
Payroll Employment	Goldman Sachs FCI	1.047	0.995	
Payroll Employment	Term Spread	**0.986**	0.240	0.10
Payroll Employment	Corp. Spread	1.053	0.952	
Payroll Employment	EBP	**0.935**	0.104	0.05
Payroll Employment	NFCD	**0.944**	0.084*	0.05
GDP	Kansas FSI	1.016	0.595	
GDP	Chicago NFCI	1.01	0.548	
GDP	St. Louis FSI	1.07	0.874	
GDP	Goldman Sachs FCI	1.058	0.952	
GDP	Term Spread	1.015	0.938	
GDP	Corp. Spread	1.015	0.722	
GDP	EBP	**0.956**	0.171	0.10
GDP	NFCD	**0.982**	0.298	0.10
Industrial Production	Kansas FSI	1.075	0.788	
Industrial Production	Chicago NFCI	1.105	0.785	
Industrial Production	St. Louis FSI	1.05	0.904	
Industrial Production	Goldman Sachs FCI	1.049	0.969	
Industrial Production	Term Spread	1.012	0.820	
Industrial Production	Corp. Spread	1.042	0.799	
Industrial Production	EBP	**0.997**	0.483	0.05
Industrial Production	NFCD	**0.947**	0.155	0.10
Unemployment Rate	Kansas FSI	**0.976**	0.383	
Unemployment Rate	Chicago NFCI	1.003	0.512	
Unemployment Rate	St. Louis FSI	**0.966**	0.208	
Unemployment Rate	Goldman Sachs FCI	1.038	0.998	
Unemployment Rate	Term Spread	**0.994**	0.366	0.05
Unemployment Rate	Corp. Spread	1.045	0.911	
Unemployment Rate	EBP	**0.922**	0.057*	0.05
Unemployment Rate	NFCD	**0.972**	0.145	0.05

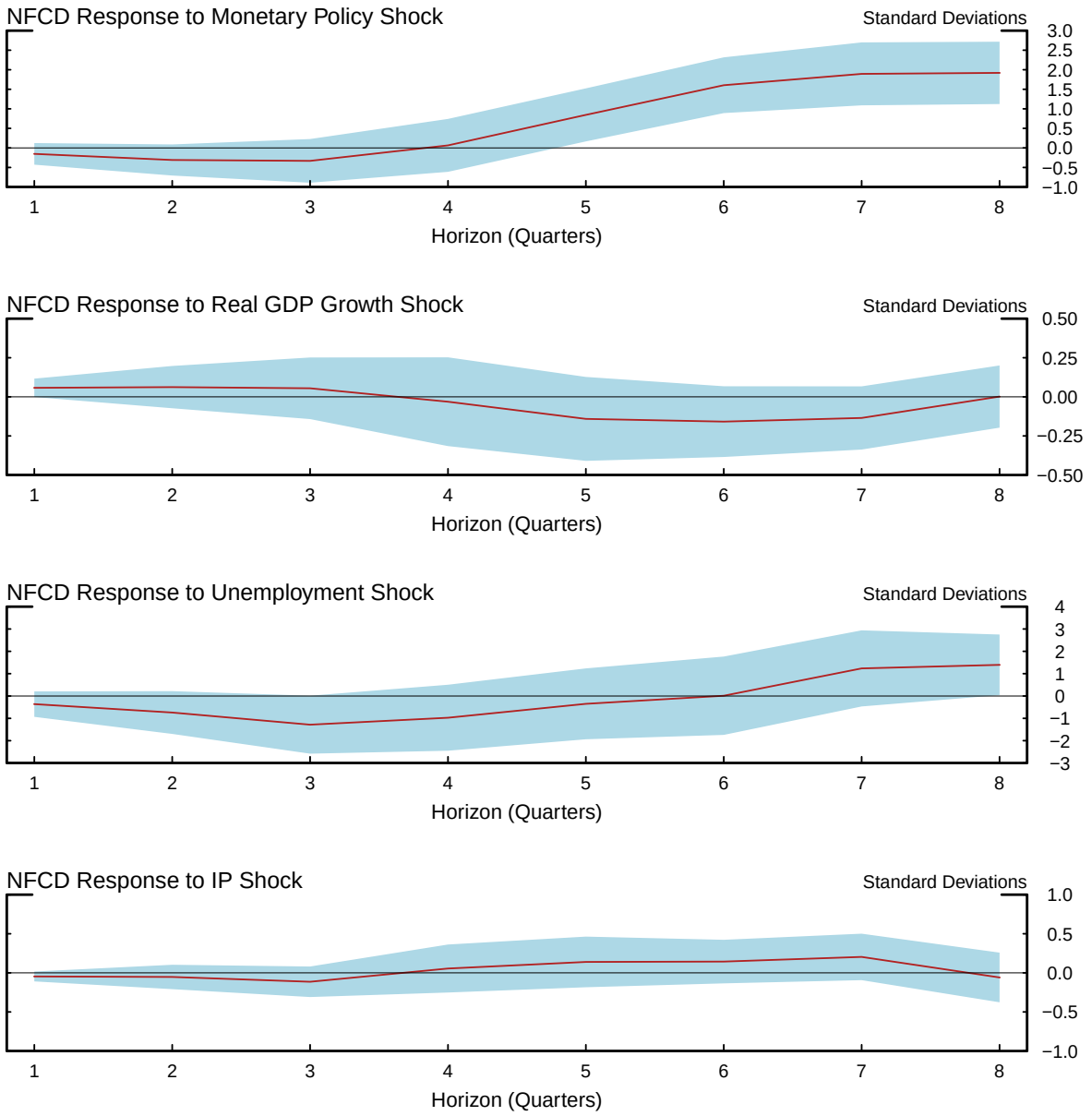
The Forecast Error Ratio (column 3) is defined as the root mean squared error (RMSE) of the index-augmented model (equation 7) divided by the baseline AR (equation 6) RMSE. Ratios less than one, indicating the index improves forecasting ability, are bolded. Diebold-Mariano (1995) P-values (Column 4) statistically significant at the ten-percent level (suggesting the index-augemented RMSE is less than the baseline RMSE) are denoted by an astrics. Clark-West (2007) P-values (Column 5) are non-normally distributed and only identified at the 5 and 10% significance level.

Figure 6: *Impulse Response Functions: Real activity responses to NFCH shocks*



Jordà (2005) style direct projection Impulse Response Functions. All inputs have been standardized with mean zero and standard deviation one. Each IRF is a measure of the response of real economic activity to a one standard deviation shock to the NFCH index. Each regression contains eight lags of each of the four response variables and the NFCH index. Real GDP and industrial production are both quarter-over-quarter differences, and monetary policy refers to the real federal funds rate.

Figure 7: *Impulse Response Functions: NFCH response to real activity shocks*



Jordà (2005) style direct projection Impulse Response Functions. All inputs have been standardized with mean zero and standard deviation one. Each IRF is the response of the ensemble NFCH index to a one standard deviation shock to some measure of real economic activity. Each regression contains eight lags of each of the four response variables and the NFCH index. Real GDP and industrial production are both quarter-over-quarter differences, and monetary policy refers to the real federal funds rates.

References

Alles, L. A. and J. L. Kling (1994). Regularities in the variation of skewness in asset returns. *Journal of Financial Research 17*(3), 427–438.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance 23*(4), 589–609.

Atiya, A. F. (2001, July). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks 12*(4), 929–935.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2008). In search of distress risk. *The Journal of Finance 63*(6), 2899–2939.

Chava, S. and A. Purnanandam (2010, 01). Is Default Risk Negatively Related to Stock Returns? *The Review of Financial Studies 23*(6), 2523–2559.

Clark, T. E. and K. D. West (2007, May). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*(1), 291–311.

Dahlquist, M. and L. E. O. Svensson (1996). Estimating the term structure of interest rates for monetary policy analysis. *The Scandinavian Journal of Economics 98*(2), 163–183.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*(3), 253–263.

Drehmann, M. and M. Juselius (2014). Evaluating early warning indicators of banking crises: Satisfying policy requirements. *International Journal of Forecasting 30*(3), 759–780.

Elliott, G. and A. Timmermann (2016). *Economic Forecasting* (1 ed.). Princeton University Press.

Fama, E. F. and K. R. French (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance 51*(1), 55–84.

Faust, J. and J. H. Wright (2009). Comparing greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business & Economic Statistics 27*(4), 468–479.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis 38*(4), 367–378.

Giesecke, K., F. A. Longstaff, S. Schaefer, and I. Strebulaev (2012, February). Macroeconomic effects of corporate default crises: A long-term perspective. Working Paper 17854, National Bureau of Economic Research.

Gilchrist, S. and E. Zakrajšek (2012). Credit spreads and business cycle fluctuations. *American Economic Review 102*(4), 1692–1720.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. Springer series in statistics. Springer.

Härdle, W., R. Moro, and D. Schäfer (2007, June). Estimating Probabilities of Default With Support Vector Machines. SFB 649 Discussion Papers SFB649DP2007-035, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.

Jones, S., D. Johnstone, and R. Wilson (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking & Finance 56*, 72–85.

Jordà, O. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review 95*(1), 161–182.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. *Journal of banking & finance 1*(3), 249–276.

Merton, R. (1973). An intertemporal capital asset pricing model. *Econometrica 41*(5), 867–87.

Odom, M. D. and R. Sharda (1990, June). A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on Neural Networks*, pp. 163–168 vol.2.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131.

Ravi Kumar, P. and V. Ravi (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. *European Journal of Operational Research 180*(1), 1–28.

Scholkopf, B., , C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik (1997, Nov). Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing 45*(11), 2758–2765.

Shin, K., T. S. Lee, and H. jung Kim (2005). An application of support vector machines in bankruptcy prediction model. *Expert Syst. Appl. 28*, 127–135.

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*(6), 405–430.

Wilson, R. L. and R. Sharda (1994, June). Bankruptcy prediction using neural networks. *Decis. Support Syst. 11*(5), 545–557.

A Additional Data Tables

Table 3: *Descriptive statistics for predictor variables*

Variable	mean	sd	min	max	range
EXRET	0.00	0.99	-4.22	4.51	8.72
EXRETAVG	0.01	0.99	-4.12	3.22	7.33
INDSALES	-0.00	1.00	-0.99	5.35	6.34
INDSALESGR	-0.00	1.00	-7.40	12.01	19.41
AGE	10.79	7.12	0.00	31.00	31.00
NITA	0.03	0.96	-28.42	6.85	35.27
TLTA	0.01	0.99	-2.22	27.36	29.58
NIMTA	0.02	0.97	-12.09	5.05	17.14
TLMTA	0.01	0.99	-1.84	3.41	5.24
CASHMTA	-0.01	0.99	-1.00	7.27	8.27
PRICE	0.03	0.98	-6.17	0.98	7.15
MB	-0.02	0.98	-1.46	5.84	7.31
NIMTAAVG	0.02	0.97	-10.34	3.42	13.77
Q	-0.02	0.96	-1.07	61.62	62.69
DIVDUMMY	0.29	0.46	0.00	1.00	1.00
KZ	-0.01	0.98	-10.55	46.11	56.66
HP	0.02	1.00	-3.14	2.61	5.75
WW	0.01	1.01	-1.96	13.76	15.72
Z	-0.02	0.96	-36.18	19.96	56.14
WCTA	0.01	0.96	-72.68	1.25	73.93
CURRENTRATIO	-0.01	0.97	-1.19	13.57	14.76
ASSETTURNOVER	0.01	1.01	-0.46	9.24	9.69
ROE	0.01	0.96	-16.12	20.97	37.09
EBITCOVER	0.01	1.00	-10.59	12.05	22.64
CAPEXTA	0.00	0.99	-7.15	9.22	16.37

Two lags of each variable are used and all missing observations have been removed. All continuous variables have been standardized with mean zero and variance one.

Table 4: *Predictor variable construction*

Variable	Description	Formula (CRSP and Compustat Mnuemonics)
EXRET	Excess Returns	log((1+ret)/100) - log(1+sprtrn)
EXRETAVG	Average Excess Returns	$\frac{(1-\theta)}{1-\theta^{12}}\phi_i(EXRET) + \sum_{i=2}^{12}\theta^{i-1}\phi_i(EXRET)$; $\theta = 2^{(\frac{-1}{3})}$; ϕ is lag operator
INDSALES	Industry Sales	Total sales of SIC 3-digit group
INDSALESGR	Industry Sales Growth	Average QoQ sales growth in SIC 3-digit group
AGE	Firm Age	(datadate - min(datadate))/365
NITA	Net Income / Adjusted Total Assets	$\frac{niq}{atq+.1(me-be)}$; me = abs(prccq)cshoq; be = χ + .1(me - χ); χ = seqq + txditcq - pstkq
TLTA	Total Liabilities/Total Assets	$\frac{ltq+mibq}{atq+.1*(me-be)}$
NIMTA	Net Income / Adjusted Total Assets	$\frac{niq}{(me+ltq+mibq}$
TLMTA	Total Liabilities / Adjusted Total Assets	$\frac{ltq+mibq}{(me+ltq+mibq}$
CASHMTA	Cash / Adjusted Total Assets	$\frac{cheq}{(me+ltq+mibq}$
PRICE	Price	log(min(abs(prccq), 15))
MB	Market Equity/ Book Equity	me/be
NIMTAAVG	Net Income / (Average Adjusted Total Assets)	$\frac{1-\theta^3}{1-\theta^{12}}\phi_1(NIMTA)+$ $+\theta^3\phi_2(NIMTA) + \theta^6\phi_3(NIMTA) + \theta^9\phi_4(MINTA)$
Q	Tobin's Q	(atq + cshoq(prccq) - ceqq - txdbq)/atq
DIVDUMMY	Dividend Dummy	1 if dividends are paid at time t
KZ	Kaplan-Zingales Index	-1.002*CASHFLOW/atq + 0.283*Q+3.319TOTDEBT/(TOTDEBT+ seqq)- -39.368*DIVIDENDS/atq-1.315*cheq/atq; CASHFLOW =ibq + dpq; TOTDEBT = dlcq+dlttq
HP	HP Index	0.737*log(pmin(atq,4500)) + 0.043*$log(pmin(atq, 4500))^2$- -0.040*pmin(AGE,37)
WW	Whited-Wu Index	-0.091*(ibq + dpq)/atq + 0.062*(DIVDUMMY) + 0.021*dlttq/atq
Z	Altman Z-score	1.2*(actq - lctq)/atq + 1.4*req/atq + 3.3*(niq + xintq + txtq)/atq + + 0.6*(cshoq*prccq)/ltq + 0.999*saleq/atq)
WCTA	Working Capital / Total Assets	(actq - lctq)/actq
CURRENTRATIO	Current Ratio	actq/lctq
ASSETTURNOVER	Asset Turnover	saleq
ROE	Return on Equity	niq/ceqq
EBITCOVER	Ebit Cover	ifelse(xintq == 0, EBIT/.01, EBIT/xintq); EBIT = niq + xintq + txtq
CAPEXTA	Capex/ Total Assets	capxq/atq

Note that we calculate all upper-case variables, while all lower-case variables are raw CRSP and Compustat variables.

B Classification Techniques and Evaluation

B.1 Classification techniques

We estimate equation (1) by various machine learning algorithms that we describe below.

We test a classical single layer artificial neural network trained with and stochastic gradient descent via standard backward propagation, emblematic of the networks used in the second wave of default prediction literature. Further, we train a support vector machine with Gaussian radial basis function kernel, emblematic of the third wave of default prediction literature (for a discussion of this technique, see Scholkopf et al. (1997)). These two techniques are well rooted in statistics, economic, and default prediction literature, so we focus our discussion on the more recently developed, and to-be-proven more effective, tree-methods.

Both the tree-based gradient boosting machine and random forest are augmentations of the simple decision tree. A single decision tree is constructed as follows: starting with all observations, the method finds the variable and associated threshold value that best splits the observations in two groups, as measured by some objective loss function, such as the Gini index or absolute error. Each group is then split again into two groups by the same processes, however it is important to note that the splitting predictor variable does not have to be the same. This processes continues until each group has only one observation in it, or until some predefined number of splits is achieved.

A random forest is simply a collection of several decision trees, returning the mean estimation of the single trees (Breiman (2001)). The random forest is outlined in algorithm 1.
Conversely, a tree-based gradient boosting machine is initialized with a single decision tree, then decision tree m is trained on the residuals of decision tree $m-1$. The final output of the gradient boosting machine is the mean probability estimate of all M trees (Friedman (2002)). The tree-based gradient boosting machine is outlined in algorithm 2.

Algorithm 1: In-sample construction of Random Forest

for $b = 1$ *to* B **do**

Draw a bootstrap sample Z^* of size N from the training data;

Grow a decision tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached;

while $n > n_{min}$ **do**

Select m variables at random from the p variables.;

Pick the best variable/split point among the m.;

Split the node into two daughter nodes.;

end

end

Output the ensemble of trees $T_{b_i}^{B}$;

Algorithm 2: In-sample construction of tree-based stochastic gradient boosting machine

Choose loss function $\Psi(y, f)$, learning rate λ, and tree depth L with five-fold cross validation;

Instantiate simple decision tree $f(x)^{(0)}$;

for *iteration m = 1 ... K* **do**

Compute the gradient $\tilde{y}_i = -(\frac{\partial \Psi(y_i, f(x_i)^{m-1})}{\partial f(x_i)^{m-1}})$ for all observations i;

Sample from training data without replacement ;

Train a tree model $h_i^{(m)}$ of depth L on the random subset using the gradient as the outcome ;

Update the model $f_i^{(m)} = f_i^{(m-1)} + \lambda h_i^{(m)}$;

end

Instantiate trained model $f^{(K)}$;

B.2 Classification exercises and evaluation

We conduct estimation exercises in two stages. The first stage is an in-sample fitting exercise of various classification algorithms on the entire data set. This exercise mainly serves to illustrate which variables different algorithms consider important in making predictions for firm defaults. The second stage is running the classification techniques recursively out-of-sample to get predictions that could be used in real-time.

During the first stage, the in-sample fit and testing, our data begins in 1985:Q1 and ends in 2018:Q3. Since defaults are rare events, representing less than one-percent of the cross-sectional observations each quarter, we mitigate the effect of this severe class imbalance by implementing a down-sample procedure. Specifically, rather than using all the data for training the models, we use all of the incidents of default and then sample from the set of non-defaulted firms to ensure a 90-10 split between the events. An alternative to this approach is to specify a high cost for missclassifying firms that default.

In the second stage, the out-of-sample fit and testing data begins in 1985:Q1, and we begin our estimation in 1989:Q1, giving our first iteration 20 months of training data. The dataset then grows to include the new observations at time t, as an expanding window. This recursive forecasting process is illustrated in the appendix. Additionally note, that as in the out-of-sample exercise, we down sample to a 90-10 split as we did in the in-sample exercise.

We train our models using monthly data. The balance sheet and income statement data are quarterly and are carried forward until new data is available. Both default dates and equity data are available at the daily frequency, and are updated monthly in our exercises.

To assess the accuracy of our various model specifications, we use a standard measure of binary-classification techniques, the area under the receiver operating characteristic curve (AUC) (see Drehmann and Juselius (2014)). As we estimate predicted probabilities of defaults, bounded between 0 and 1, we need a cutoff value c for which predicted values above are classified as a default prediction. Given the specific cutoff value c, one can calculate the proportion of true positives and false positives (that is, the number of firms which are predicted to default and do not, and vice versa) that occur. The choice between a high or low value of c implies a trade-off between the proportion of true and false positives. The receiver operating characteristic curve

(ROC) attempts to visualize this trade-off by plotting the true positive rate against the false positive rate for various values of c between 0 and 1. The AUC is the definite integral of the ROC. An AUC of 1 indicates a perfect classifier, while an AUC of 0.5 suggests the same accuracy as tossing a fair coin to assign classifications. Our choice of preferred models are driven by which algorithms maximize the AUC.

B.3 In-sample estimation results

In-sample we estimate the probability of default within h-quarters, such that $h \in [1, 8]$. At each time-horizon we fit a logistic regression, support vector machine, artificial neural network, random forest, and tree-based gradient boosting machine. All model parameters are chosen using five-fold cross validation, maximizing the accuracy of each model. In parameter tuning process, accuracy is measured as the proportion of predicted classes that agree with observed classes.

Figure 2 depicts the AUCs of the fitted models across the one through eight quarter time horizons. The random forest dominates all other models across all but one time horizon, as measured by AUC. The tree-based gradient boosting machine dominates all methods, except the random forest, at all time horizons. Both the support vector machine and artificial neural network are dominated by the logistic regression at the one-quarter time horizon. However the support vector machine then dominates both the logistic regression and neural network after the one-quarter horizon. The neural network dominates the logit after the two-quarter horizon. Note that the ability of the classifiers is exactly what one would expect based on the evolution of default forecasting literature. Further, the logistic regression decreases monotonically, suggesting that the linear functional form may poorly approximate the nonlinear relationships as they develop overtime. In contrast, the best classifier, the random forest, increases monotonically, and achieves an AUC greater than 0.99 at all time horizons. Random forest are well known for their ability to capture non-linear relationships and robustness to outlier and missing data (for a greater review of random forest, see Hastie et al. (2009)).

A consequence of the logistic regression's linear structure is its inability to capture the nonlinear relationship between variables and their lags in a mean reverting system (unless quadratic terms are specifically imposed via a basis expansion technique). As Figure 3 shows, nine of the ten

most important variables used by the random forest are really three variables with two associated lags each[8]. Conversely, the logistic regression[9] only ranks two pairs of variables and their first lag as being in the top ten most important variables. However, both methods do flag, TLMTA, total liabilities over market equity plus total liabilities (a measure of leverage), as being the most important variable. In fact, more than half of the most important variables flagged by the random forest are a measure of leverage, and the other variable EBIT cover is a measure of a firm's ability to pay its interest expenses. While these specific variables may be important motivation in a micro-driven structural model of corporate stress, that is beyond the scope of this paper.

Despite the dominance of the random forest across all forecasting horizons, there are serious disadvantages to drawing conclusions based on only in-sample model fits when the true design of the project is to predict defaults in real-time. For example, the machine learning techniques may be overfitting in-sample, potentially leading to increased out-of-sample bias (Hastie et al. (2009)). For this reason, we test all models out-of-sample.

B.4 Out-of-sample estimation results

Out of sample, the random forest once again dominates all tested algorithms at every time horizon, as measured by AUC. In fact, the random forest has an average AUC greater than 0.97.
Figure 2 presents the AUC's of the random forest, tree-based gradient boosted machine, support vector machine, artificial neural network, and logit. The tree-based gradient boosting machine again dominates all other classifiers, excluding the random forest. Both tree-based methods can be seen to increase in efficacy after the one-quarter horizon, sustain a stable AUC from at least the three-quarters through the five-quarters horizons, then decrease to another stable region from

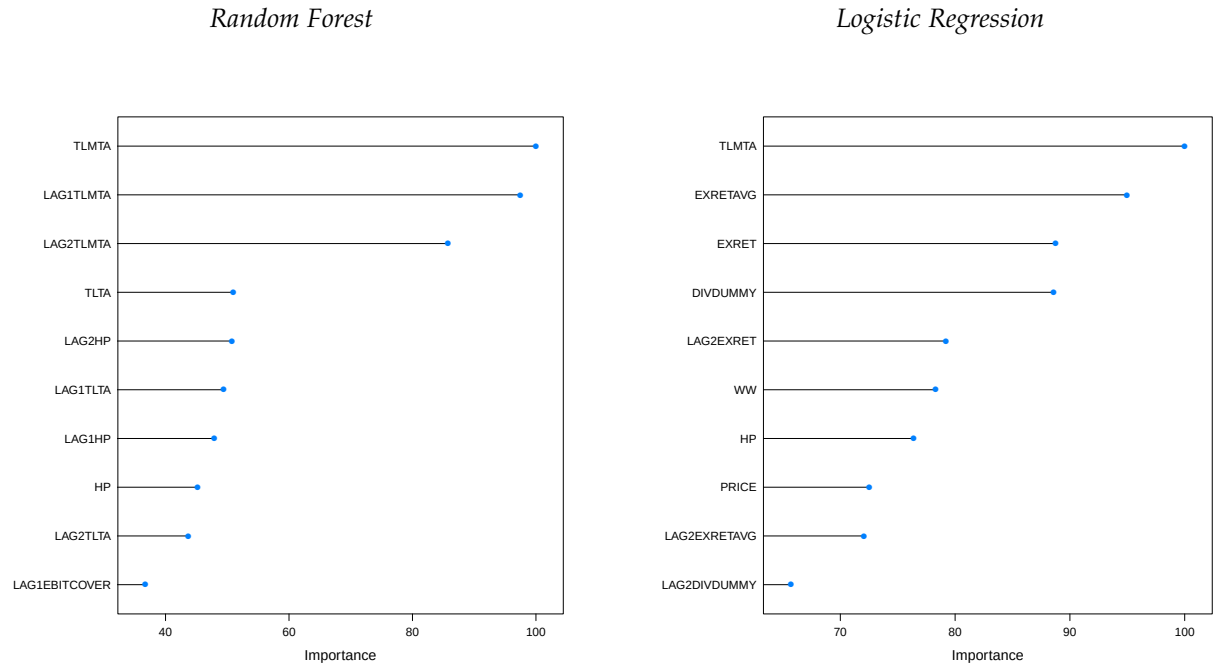
[8]Put forth by Breiman et al. (1984), a variable's relative importance in a single decision tree is given by its squared relevance

$$R_\lambda^2 = \sum_{n=1}^{N-1} i_n^2 I(v(t) = \lambda)$$

where λ is the variable of interest, $N-1$ is the number of splits inside the tree, n is a node in the tree, and i^2 is the maximal marginal improvement in accuracy. The marginal improvement i^2 is calculated by partitioning the region associated with node n by each variable in turn, setting the regional response value to some constant, and measuring the change in accuracy from the tree before generating the node. The indicator function $I(v(t) = \lambda)$ is present to say that only i^2 generated by the variable λ are counted when calculating λ's relative importance. A variables' relative importance in a random forest is the variable's average R^2 across all trees in the forest.

[9]A variable's importance, as measured by a logistic regression, is the absolute value of its coefficient. Because all variables have been previously standardized, comparing the magnitude of betas is straightforward.

Figure 8: *Variable Importance by Machine Learning Algorithm*



The top ten most important variables used in estimating the probability of firm default by random forest and logistic regression. Please see footnotes (3) and (4) for construction details. Note that the relative importance of each variable is scaled such that the most important variable has a relevance of 100.

either six or seven-quarters to the eight-quarters horizon. Conversely, the neural network's AUC monotonically decreases from a maximum of 0.94 to a minimum of approximately 0.91. Further evidence of the logit's weakness in this exercise is the fact that the logit first needed to have the data winsorized at the 5 and 95% levels before producing the AUC's presented by Figure 2. When the data is not winsorized, the logit's one-quarter ahead AUC is 0.83, not 0.87. Additionally, while the general shape of the AUC curve is the same, at the six-quarter ahead horizon, the AUC sporadically drops to approximately 0.9, only to rise back up to .93 at seven-quarters ahead and then monotonically decrease. This marked instability further underscores the logit's susceptibility to data constraints and its own rigid linear structure. Given that the random forest dominates across all time horizons, its predicted probabilities are used to construct our aggregate indexes of the macroeconomy.

C In-sample forecasting results for sub-indices

C.1 (In-sample) Forecasting Exercises

To determine the efficacy of this index as a leading indicator of US non-financial corporate health, and therefore a leading indicator of real economic activity, we run an in-sample forecasting test, as in Gilchrist and Zakrajšek (2012), with payroll employment, real GDP, industrial production, and the unemployment rate. The forecast specification is:

$$\nabla^h Y_{t+h} = \alpha + \sum_{i=1}^{4} \beta \nabla^1 Y_{t-i} + \delta_1 TS_t + \delta_2 RFF_t + \delta_4 X(t,h) + \epsilon_{t+h} \tag{9}$$

where t is the time index, h is the forecast horizon, $X(t,h)$ is our index of interest, $\nabla^h Y_{t+h} \coloneqq \frac{400}{h+1}\log(\frac{Y_{t+h}}{Y_{t-i}})$, Y is the measure of real economic activity. In addition to lagged growth values of the dependent variable, we control for the stance of monetary policy by including the term spread, TS, between the constant maturity three-month and ten-year Treasury yield, and the real federal funds rate, RFF.

We measure the forecasting accuracy of our indexes in terms of an adjusted R^2. Further, we compare our index to the Excess Bond Premium, EBP, an information rich sentiment or risk appetite indicator constructed in Gilchrest and Zakrajsek (2012), based on the credit spread of non-financial corporate firms, and a financial conditions index, the Chicago $NFCI$ (see Brave and Butters (2011)). We run the forecasting exercise with each indicator individually, then in pairs with our index, and with all four indexes at the same time. We evaluate the merits of each index by comparing adjusted R^2's and levels of statistical significance.

C.2 Weighted Mean Index

First, we construct an index of non-financial corporate defaults (NFCH) by calculating the weighted mean of firm-level probabilities of default between time t and $t+h$:

$$NFCH^m_{t,h} = \frac{1}{n_t}\sum_{i=1}^{n_t} w_{i,t} p_{i,t} \tag{10}$$

$$w_{i,t} = \sum_{j=1}^{n_t} \frac{A_{i,t}}{A_{j,t}}$$

where i is the firm index, t is the time index, n is the number of firms in the sample at time t, $w_{i,t}$ is the firm's cross-sectional weight determined by assets, A, and $p_{i,t,h}$ is the firm's probability of default within h quarters from time t.

Table 5 shows the in-sample forecast exercise results using the $NFCH^m$ as $I(t,h)$ in equation (2). The $NFCH^m$ is statistically significant at the one-percent level any time it is either alone or with one other leading indicator. Further, the $NFCH^m$ is significant at the five-percent level when in a forecasting regression with all four other leading indicators. In all exercises that include all three indicators, the Chicago NFCI is insignificant, while the Excess Bond Premium is made insignificant in forecasting real GDP. Lastly, the sign of the indicator is as expected, negative, meaning that as the probability of firm default increases, future real economic activity decreases.

C.3 Unweighted Dispersion Index

Second, we note that by weighting a firm's probability of default by the firm's share of assets, we largely mute the effects of small firm distress. However, given that small firms may be more capital constrained and therefore less able to weather deteriorating economic conditions, the probability of small firm default may provide valuable information to forecast real economic activity. Further, while we note that the mean is the most efficient measure of central tendency, its does not capture information concerning a distributions dispersion. However, it is documented that higher-order moments of firm stock returns vary predictably over the business cycle (for example, see Alles and Kling (1994)), suggesting that moments describing the tails of a distribution, standard deviation and skew, may lend valuable information regarding the current and future state of economic activity. Given these motivations, we next construct an index by calculating the mean of the quarter-over-quarter[10] difference in the unweighted standard deviation and quarter-over-quarter

[10]We find our results to be robust to constructing the $NFCH^s$ with differenced moments, however, as with the $NFCH^m$, we find that differencing the moments makes business cycles more discernible and maximizes the index's forecasting abilities.

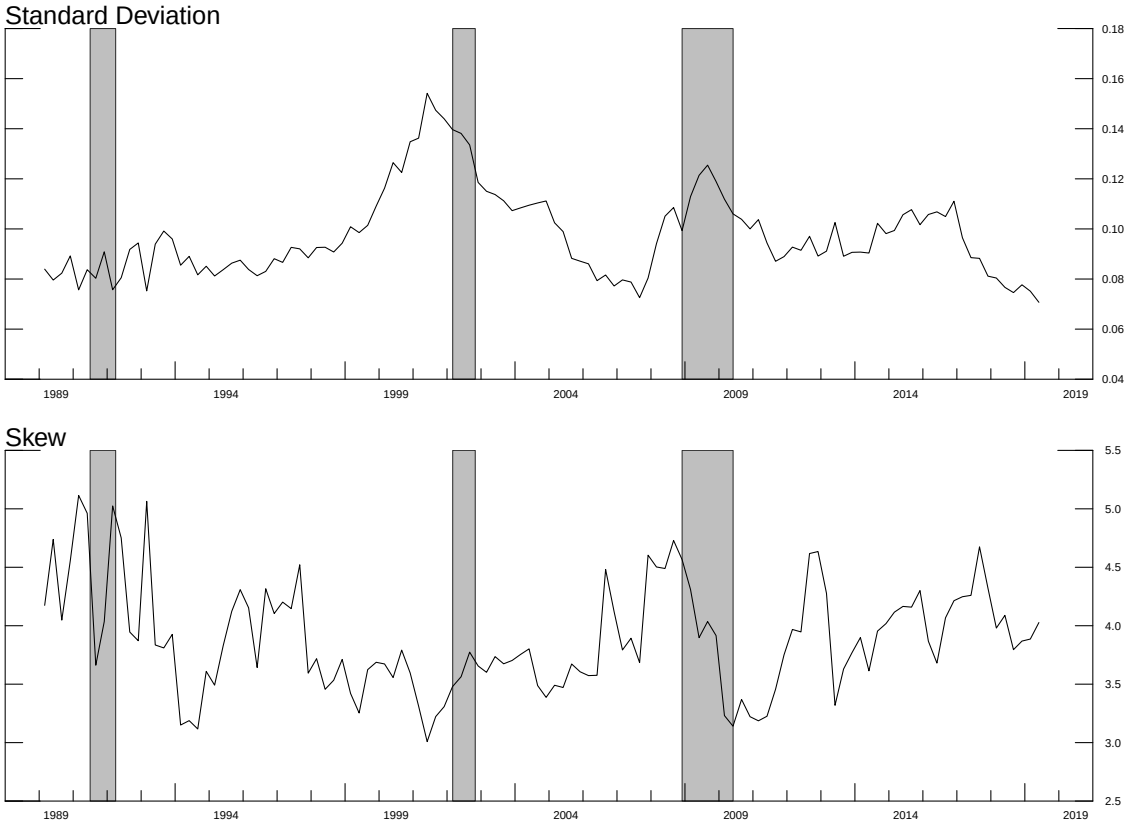
difference in the unweighted skew at time t:

$$NFCH_t^s = \frac{1}{2}(\gamma_t + \sigma_t) \tag{11}$$

where t is the time index, γ is the cross-sectional skew, and σ is the cross-sectional standard deviation.

Figure 9 shows the unweighted standard deviation and skew. It is notable that the standard deviation appears countercyclical, rising before and through the 2001 and 2008 recession while falling in economic recoveries. However, the skew does not appear cyclical, but rather characterized by relative calms punctuated by rapid increases and decreases. The skew spikes two years before the 2001 recession, and three years before the 2008 recession. These pre-recession spikes may suggest that the skew is a good leading indicator of recession, however, it does also rise during recessions. Table 6 presents the results of the in-sample forecasting exercises using equation (2) as the $NFCH^s$. The $NFCH^s$ is statistically significant in all model specifications, except when forecasting the unemployment rate with EBP. The NFCH coefficients are all the expected direction for a pro-cyclical index. However, the adjusted R^2s of the exercises using only one indicator suggest that the skew and standard deviation are less informative than other indexes.

Figure 9: *Unweighted Moments of the Corporate Default Probability Distribution*



The second and third moments generated by the cross-section of unweighted probabilities of firm default within eight-quarters from time t, generated by the RF algorithm.

Table 5: *In-Sample Forecast Results using the Weighted Mean Subindex*

Payroll Employment

EBP	−0.830***			−0.720***		−0.342**
	(0.159)			(0.173)		(0.164)
NFCI		−0.851***			−0.786***	−0.151
		(0.211)			(0.223)	(0.180)
NFCD			−0.514***	−0.262***	−0.434***	−0.237***
			(0.103)	(0.088)	(0.101)	(0.089)
Observations	100	100	100	100	100	100
Adjusted R^2	0.692	0.668	0.593	0.708	0.725	0.777

Industrial Production

EBP	−2.944***			−2.292***		−1.123**
	(0.707)			(0.671)		(0.514)
NFCI		−2.595**			−2.219**	0.196
		(1.002)			(1.018)	(0.724)
NFCD			−2.140***	−1.405***	−1.808***	−1.112***
			(0.335)	(0.344)	(0.350)	(0.305)
Observations	100	100	100	100	100	100
Adjusted R^2	0.395	0.347	0.337	0.459	0.473	0.590

Real GDP

EBP	−0.850***			−0.641***		−0.156
	(0.192)			(0.223)		(0.222)
NFCI		−0.940***			−0.770***	−0.126
		(0.212)			(0.248)	(0.235)
NFCD			−0.664***	−0.402***	−0.464***	−0.264*
			(0.151)	(0.144)	(0.159)	(0.150)
Observations	100	100	100	100	100	100
Adjusted R^2	0.347	0.369	0.301	0.387	0.434	0.516

Unemployment

EBP	8.191***			6.733***		3.458**
	(0.985)			(0.949)		(1.333)
NFCI		8.752***			7.728***	−0.091
		(2.385)			(2.191)	(1.781)
NFCD			5.606***	3.349***	4.680***	2.498**
			(0.939)	(0.776)	(1.068)	(0.958)
Observations	100	100	100	100	100	100
Adjusted R^2	0.587	0.548	0.505	0.621	0.630	0.729

Table 6: *In-Sample Forecast Results using Unweighted Dispersion Measures Subindex*

Payroll Employment

EBP	−0.830***			−0.744***		−0.543***
	(0.159)			(0.135)		(0.135)
NFCI		−0.851***			−0.687***	0.168
		(0.211)			(0.188)	(0.169)
NFCD			−0.563***	−0.443**	−0.327**	−0.362***
			(0.199)	(0.180)	(0.143)	(0.110)
Observations	100	100	100	100	100	100
Adjusted R^2	0.692	0.668	0.610	0.751	0.693	0.798

Industrial Production

EBP	−2.944***			−2.606***		−1.801***
	(0.707)			(0.532)		(0.448)
NFCI		−2.595**			−2.024***	1.071*
		(1.002)			(0.733)	(0.585)
NFCD			−2.074**	−1.734**	−1.556**	−1.488***
			(0.886)	(0.853)	(0.749)	(0.546)
Observations	100	100	100	100	100	100
Adjusted R^2	0.395	0.347	0.331	0.515	0.435	0.636

Real GDP

EBP	−0.850***			−0.755***		−0.351
	(0.192)			(0.174)		(0.216)
NFCI		−0.940***			−0.769***	0.110
		(0.212)			(0.183)	(0.304)
NFCD			−0.683**	−0.586**	−0.491**	−0.474**
			(0.300)	(0.267)	(0.245)	(0.190)
Observations	100	100	100	100	100	100
Adjusted R^2	0.347	0.369	0.308	0.464	0.443	0.572

Unemployment

EBP	8.191***			7.270***		4.914***
	(0.985)			(1.079)		(1.237)
NFCI		8.752***			7.100***	−2.995
		(2.385)			(1.673)	(1.807)
NFCD			5.176**	3.763	3.061***	3.484***
			(2.594)	(2.334)	(0.950)	(0.955)
Observations	100	100	100	100	100	100
Adjusted R^2	0.587	0.548	0.483	0.635	0.569	0.739