Bond Risk Premiums with Machine Learning

Daniele Bianchi
Queen Mary, University of London

Matthias Büchner
University of Warwick

Andrea Tamoni
Rutgers Business School

We show that machine learning methods, in particular, extreme trees and neural networks (NNs), provide strong statistical evidence in favor of bond return predictability. NN forecasts based on macroeconomic and yield information translate into economic gains that are larger than those obtained using yields alone. Interestingly, the nature of unspanned factors changes along the yield curve: stock- and labor-market-related variables are more relevant for short-term maturities, whereas output and income variables matter more for longer maturities. Finally, NN forecasts correlate with proxies for time-varying risk aversion and uncertainty, lending support to models featuring both channels. (*JEL* C38, C45, C53, E43, G12, G17)

Received XXXX XX, XXXX; editorial decision XXXX XX, XXXX by Editor XXXXXXXXXXXX.

We thank the editor Stijn Van Nieuwerburgh and two anonymous referees. We are grateful to Marcus Buckmann, Hui Chen, Serdar Dinc, Winston Dou, Darrell Duffie, Chulwoo Han, Otto Van Hemert, Marcin Kacperczyk, Bryan Kelly, Zhaogang Song, Ansgar Walther, and Marcin Zamojski and participants at the 2020 AFA in San Diego, the 2019 SFS Cavalcade North America at Carnegie Mellon University, the 2019 USC Dornsife INET Panel Data Conference, the 2019 Georgia State FinTech Conference, the 2019 FMA Consortium on Factor Investing in Cambridge, the conference on "New Developments in Factor Investing" at Imperial College, the workshop "Modelling with Big Data and Machine Learning" at the Bank of England, the workshop "Predicting Asset Returns" in Örebro, the 13th Imperial conference on Advances in the Analysis of Hedge Fund Strategies at Imperial College, the 29th EC^2 Conference on Big Data Econometrics, the "Alternative Risk Premia" conference at Imperial College, the 2019 FMA European Conference in Glasgow and research seminars at Durham University Business School, Lancaster University Management School, and Warwick Business School for useful comments and suggestions. We kindly thank Paul Whelan and Marco Giacoletti for sharing their proxy of bond risk premia with us. We thank the Centre for Scientific Computing at the University of Warwick for support with the supercomputing clusters. Send correspondence to Andrea Tamoni, andrea.tamoni.research@gmail.com.

doi:10.1093/rfs/Sample Advance Access publication XXXXXX

The recent advancements in the fields of econometrics, statistics, and computer science have spurred interest in dimensionality reduction and model selection techniques, as well as predictive models with complex features, such as sparsity and nonlinearity, both in finance and in economics.[1] Over the last two decades, however, the use of such methods in the financial economics literature has been mostly limited to data compression techniques, such as principal component and latent factor analysis.[2] A likely explanation for the slow adoption of advances in statistical learning is that these methods are not suitable for structural analysis and parameter inference (see Mullainathan and Spiess, 2017). Indeed, the primary focus of machine learning is prediction, that is, to produce the best out-of-sample forecast of a quantity of interest based on a potentially large conditioning information set. The suitability of machine learning methodologies for predictive analysis makes them particularly attractive in the context of financial asset return predictability and risk premium measurement (e.g., Gu, Kelly, and Xiu, 2018). As a matter of fact, while many problems in economics rely on the identification of primitive underlying shocks and structural parameters, the quantification of time variation in expected returns is essentially a forecasting problem. This practical view complements the theory-driven approach, which often provides the building blocks for the empirical analysis of financial markets. Modeling the predictable variation in Treasury bond returns, which is the focus of this paper, provides a case in point. Forecasting excess bond returns requires a careful approximation of the a priori unknown mapping between the investors' information set and excess bond returns (e.g., Duffee, 2013, pp. 391–2). In this paper, we employ machine learning methods to revisit the debate on the presence of predictable variation in bond returns. We work with two traditional frameworks; one that exploits information in the yield curve only, as in Cochrane and Piazzesi (2005), and one that also uses information from a data set of hundreds of macroeconomic indicators as in Ludvigson and Ng (2009). The research design follows the structure outlined in Gu et al.

[1] See, for example, Rapach et al. (2013), Kelly and Pruitt (2013; 2015), Freyberger, Neuhierl, and Weber (2017), Giannone et al. (2017), Giglio and Xiu (2017), Heaton, Polson, and Witte (2017), Kozak, Nagel, and Santosh (2017), Messmer (2017), Fuster et al. (2018), Gu, Kelly, and Xiu (2018), Kelly, Pruitt, and Su (2018), Rossi (2018), Sirignano et al. (2018), Chen, Pelger, and Zhu (2019), Feng, Giglio, and Xiu (2019a), Feng, Polson, and Xu (2019b), and Huang and Shi (2019).

[2] In economics, the initial idea of data compression techniques can be traced back to Burns and Mitchell (1946), who argue for a business-cycle indicator that is common across macroeconomic time series. This idea was formally modeled by Geweke (1977) and Sargent and Sims (1977). Since then, principal component analysis and factor analysis have been widely adopted by researchers in financial economics for forecasting problems involving many predictors (see, among others, Stock and Watson (2006; 2002a; 2002b), Forni and Reichlin (1996, 1998), Bai and Ng (2003, 2006, 2008), De Mol et al. (2008), and Boivin and Ng (2006)).

(2018), whereby a comparison of different machine learning techniques is based on their out-of-sample predictive performance. Methodologically, we consider a variety of machine learning techniques to forecast excess Treasury bond returns across different maturities including partial least squares, penalized linear regressions, boosted regression trees, random forests, extremely randomized regression trees, and shallow and deep neural networks (NNs). All of these methods fall under the heading of "supervised learning" in the computer science literature.[3] Although not exhaustive, this list covers the vast majority of modern statistical learning techniques (e.g., Friedman et al., 2001). We also employ more classical dimensionality reduction techniques, such as principal component analysis (PCA), which arguably represents an almost universal approach to regression-based forecasting of Treasury bond returns. Our contribution to the bond return predictability literature is threefold. First, within each empirical application (i.e., yields-only or yields plus macroeconomic variables), we show that nonlinear machine learning methods, such as extreme trees and NNs, are useful to detect predictable variations in bond excess returns, as indicated by out-of-sample predictive R^2s that are significantly higher than those obtained by data compression techniques (e.g., linear combinations of forward rates, as in Cochrane and Piazzesi (2005), and factors extracted from macroeconomic variables, as in Ludvigson and Ng (2009)) and penalized regression techniques. Importantly, a battery of asset allocation exercises confirms that the deviations from the Expectations Hypothesis documented in this paper are economically large. In this regard, our paper contributes to the debate on the statistical evidence supporting bond return predictability (e.g., Fama and Bliss (1987), Campbell and Shiller (1991) and Cochrane and Piazzesi (2005), for applications with yields-only) or absence thereof (e.g., Thornton and Valente, 2012). Second, zooming in on nonlinear methods, we document that using information from macroeconomic and financial variables improves the predictive accuracy of forecasts based only on (potentially nonlinear transformations of) the yield curve. Indeed, the best-performing NN that exploits macroeconomic and term structure information attains out-of-sample R^2s that are about 10 percentage points larger (for maturities ranging from 2 to 10 years) than the best-performing NN that employs yields only. Similarly, we document that employing the NN forecasts based on macroeconomic and yield information produces significantly higher certainty equivalent return values than those implied by the NN forecasts based only on yield

[3] In "supervised" statistical learning, mapping between the quantity of interest y and the predictors $\boldsymbol{x}$ is learned by using information on the joint distribution. Unsupervised learning (e.g., PCA) instead does not explicitly condition on the quantity of interest y to summarize the information content in $\boldsymbol{x}$.

curve information. In this respect, our paper contributes to the debate on whether there is macroeconomic variation not spanned by bond yields that helps forecast excess bond returns (see, e.g., Joslin et al., 2014 for evidence in favor of unspanned macroeconomic information, and Bauer and Hamilton, 2018; Bauer and Rudebusch, 2017 for a critical analysis of such evidence, along with a discussion of econometric issues plaguing the "spanning" linear regressions). On one hand, our analysis reinforces the evidence in favor of unspanned macroeconomic information useful to forecast excess bond returns (e.g., Cooper and Priestley, 2009; Ludvigson and Ng, 2009; Duffee, 2011b; Joslin et al., 2014; Cieslak and Povala, 2015; Coroneo et al., 2016; Gargano et al., 2019). On the other hand, our evidence is novel in three respects. First, we continue to find support for unspanned macroeconomic risk even after accounting for potential nonlinearities in interest rates. Second, we find that it is important to account for nonlinearities within macroeconomic categories in order to detect information useful for predicting excess bond returns above and beyond the yield curve. Finally, we document substantial heterogeneity in the relative importance of macroeconomic and financial variables across bond maturities: variables pertaining to the stock and labor markets are more important for short-term maturity bonds, whereas variables pertaining to orders and inventories, and output and income are more relevant for variation in long-term bonds. Thus, the type and nature of unspanned factors may depend on bond maturity. Our third contribution concerns the economic properties of the forecasts implied by deep NNs. First, to provide insight on the origins of the improvements in out-of-sample predictability, we investigate the ability of NNs to forecast the first three principal components of the term structure: level, slope, and curvature. We show that when using yields only as predictors, the NNs improve the forecast of the level of the term structure. However, when both macroeconomic and financial information are used, in addition to yields, we find that the factors extracted from the NNs contribute to the ability to predict the level of the yield curve, as well as the slope. This is consistent with the idea that the slope of the yield curve is related to the state of the economy, and a NN is able to extract the relevant information from the large set of macroeconomic variables used. Next, we document that NN forecasts are countercyclical and mostly related to variables that proxy for macroeconomic uncertainty and time-varying risk aversion. Thus, our results support models that feature both time variation in risk prices and in time-varying risk as in, for example, Bekaert et al. (2009) and Creal and Wu (2018). However, our statistical measure of expected bond returns contrasts with recent survey-based measures like the one proposed by Buraschi et al. (2019). Their measure is mostly related to financial (specifically, bond) volatility. In the context of machine learning in asset pricing,

we document three novel facts.[4] First, our result that extreme trees and NNs constitute the best-performing methods even in the case when only information in the term structure is used to forecast bond returns (i.e., in a low-dimensional setting) is new and provides evidence that the gain from nonlinear machine learning methods is not relegated to a big data context. Second, we show that an economically driven choice of the network structure may perform on par with more data-driven network architectures. More specifically, when macroeconomic data are included as potential bond return predictors, we find that the out-of-sample predictive R^2 increases almost monotonically when we move from shallow specifications (one hidden layer) to deeper networks (up to three hidden layers). However, we also find that economic priors about the role of variables may improve the performance of the network. In particular, grouping variables within economic categories and then training a shallow network within each group—a network structure that we dub "group ensembling"—attains a performance that is on par with the best-performing deep NN where no economic priors are utilized.[5] Thus, the depth of the network and the economic priors used to design the network (e.g., grouping within categories) interact with one another, a result that is new to the empirical finance literature. Third, the fact that the group-ensembled network outperforms more complex and agnostic specifications, is important since it highlights what type of nonlinearities are important from an economic perspective: Is it the interaction of many variables (across categories) or rather a higher polynomial of the same variable (within a category)? Since our group-ensembled network switches off interactions across categories, our analysis shows that it is the nonlinearity *within* a group that is ultimately relevant for the performance of the network. In this respect, our results for Treasury bond returns echo those in Gu et al. (2018) and Chen et al. (2019) for the equity market: the success of NNs lies in their ability to exploit the nonlinear mapping between returns and the predictors. However, whereas Chen et al. (2019) emphasize the importance of identifying the relevant interaction between firm characteristics for equity returns, we document that, in the bond market, the interaction across economic categories matters to a lesser extent than the interaction within a

[4] The literature on machine learning and asset pricing is rapidly growing (cf. footnote 1). Except for Huang and Shi (2019), none of these papers explores machine learning methods to forecast excess bond returns.

[5] In this respect, our paper is mostly related to Huang and Shi (2019). They use an adaptive group-lasso linear regression and cluster economic variables. They show that a linear combination of such clusters significantly predicts excess bond returns. Similar to them, we explore economically motivated structures based on an ex ante clustering. Different from them, we focus on nonlinear methodologies. In fact, our results show that linear sparse regression methods, such as lasso and elastic net, substantially underperform, statistically and economically, both shallow and deep nonlinear methods, such as extreme trees and neural networks.

category. Thus, different types of network structures may be needed for different asset markets.

1. Motivating Framework

In this section, we provide a motivation for the use of machine learning to predict excess Treasury bond returns. The discussion is framed within the context of regression approaches for forecasting treasury yields. We start with the accounting identity of Campbell and Shiller (1991). We consider a zero-coupon bond with maturity $t+n$ and a payoff of one dollar. We denote its (log) price and (continuously compounded) yield at time t by $p_t^{(n)}$ and $y_t^{(n)} = -\frac{1}{n} p_t^{(n)}$, respectively. The superscript refers to the bond's remaining maturity. The (log) excess return to the n-year bond from t to $t+1$, when its remaining maturity is $n-1$, is denoted by $xr_{t+1}^{(n)} = p_{t+1}^{(n-1)} - p_t^{(n)} - y_t^{(1)}$. Then, it is possible to express the log returns to bonds as

$$xr_{t+1}^{(n)} = -(n-1)\left(y_{t+1}^{(n-1)} - y_t^{(n)}\right) + \left(y_t^{(n)} - y_t^{(1)}\right). \tag{1}$$

The identity states that (after controlling for the slope $y_t^{(n)} - y_t^{(1)}$) any variable that forecasts the change in the bond yield from t to $t+1$, that is, $\left(y_{t+1}^{(n-1)} - y_t^{(n)}\right)$, also forecasts the log returns to bonds. Assuming that the investors' information set at time t can be summarized by a latent k-dimensional state vector $\boldsymbol{x}_t$, and exploiting the identity $y_t^{(n)} = \frac{1}{n}\sum_{j=0}^{n-1} E_t\left(y_{t+j}^{(1)} \,|\, \boldsymbol{x}_t\right) + \frac{1}{n}\sum_{j=0}^{n-1} E_t\left(xr_{t+j+1}^{(n-j)} \,|\, \boldsymbol{x}_t\right)$, we can write

$$\boldsymbol{y}_t = f(\boldsymbol{x}_t; N),$$

where we stack time-t yields on bonds with different maturities in a vector $\boldsymbol{y}_t$, and the maturities of the bonds in the vector N. Combining the equation above with Equation (1), we obtain

$$E_t\left[xr_{t+1}^{(n)}\right] = g(\boldsymbol{x}_t; N)\ , \tag{2}$$

for some function $g(\boldsymbol{x}_t; N)$. Every term structure model reduces to a specific mapping between yields and state variables. In the simplest case, yields are linear affine functions of the state variables: $\boldsymbol{y}_t = A + B\boldsymbol{x}_t$. The linearity of $f(\cdot)$, together with a dimensionality reduction of the space of yields, gives rise to principal component regression (PCR) where the quantity of interest (excess bond returns) is regressed onto principal components $\boldsymbol{x}_t$ (see chap. 3.5 in Friedman et al., 2001):

$$E_t\left[xr_{t+1}^{(n)}\right] = \hat{\alpha} + \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}_t \qquad \text{with} \qquad \boldsymbol{x}_t = \boldsymbol{W}\boldsymbol{y}_t + b, \tag{3}$$

where the columns of $\boldsymbol{W}$ form an orthogonal basis for directions of greatest variance, and b captures the average "reconstruction error" or bias. Practically, the linear predictive system outlined in Equation (3) represents a two-step procedure where researchers extract the latent factors $\boldsymbol{x}_t$, and then learn the regression coefficients $\hat{\theta}=\left(\hat{\alpha},\hat{\boldsymbol{\beta}}^{\top}\right)$ by minimizing a loss function that depends on the residual sum of squares. In addition to this yields-only specification, researchers have often evaluated the role of macroeconomic variables as an important driver of bond returns. This leads to an augmented predictive regression:

$$E_t\left[xr_{t+1}^{(n)}\right]=\hat{\alpha}+\hat{\boldsymbol{\beta}}^{\top}\boldsymbol{x}_t+\hat{\boldsymbol{\gamma}}^{\top}\boldsymbol{F}_t \tag{4}$$

where $\boldsymbol{F}_t\subset f_t$ and f_t is an $r\times 1$ vector of latent common factors extracted from a $T\times N$ panel of macroeconomic data with elements $m_{it}, i=1,\ldots,N$, $t=1,\ldots,T$, and $r\ll N$. This is the framework originally proposed by Ludvigson and Ng (2009). Equations (3) and (4) constitute two important applications of unsupervised data compression for bond forecasting. Another very popular set of reduced-form term structure models consists of Gaussian linear-quadratic models (e.g., Ahn, Dittmar, and Gallant, 2002). In this case, the relation between yields and state variables is given by $\boldsymbol{y}_t=A+B\boldsymbol{x}_t+\boldsymbol{x}_t'C\boldsymbol{x}_t$. Note that in this case the mapping between yields and factors is nonlinear in the state variables, meaning that standard linear PCA represents a mere approximation and does not necessarily give consistent estimates of the true underlying quadratic factor (see Schölkopf et al., 1998). Nonlinearities are also featured in reduced-form term structure models with regime switches (e.g., Dai et al., 2007), in shadow rate models (Black, 1995; Wu and Xia, 2016), and in the model by Feldhutter et al. (2016) where the price of a bond is a time-varying weighted average of bond prices in artificial, affine Gaussian economies. Interestingly, nonlinearity between bond yields and factors also emerges naturally in structural models with habit formation. For example, Buraschi and Jiltsov (2007) use habit formation preferences as a source of time varying market price of risk in fixed income models. In their model, yields are nonlinear in the state variables, and depend on the habit stock and the factors affecting the monetary aggregate. For completeness, Internet Appendix A provides a simple habit formation economy that leads to bond yields being a linear-quadratic function of macroeconomic variables like consumption growth, expected inflation, and habit. Motivated by this literature, we investigate the possibility that a more precise measurement of bond risk premiums can be obtained by using nonlinear transformations of the data, an avenue that has also been advocated by Stock and Watson (2002b, p. 154) within the context of forecasting macroeconomic time series. Differently from the Gaussian linear-quadratic models, we do not

postulate a specific functional form connecting bond yields and state variables; instead, we use various statistical techniques, such as trees and networks, to learn about it. Besides being agnostic about the functional form between excess bond returns and macroeconomic and financial variables, the use of machine learning techniques has two additional advantages relative to the principal component regressions in Equations (3) and (4). First, the implementation of regression-based forecasts of excess bond returns using principal components as outlined in Equations (3) and (4) typically implies that no direct use of the response variable (i.e., the excess bond returns) is made to learn about the state variables $\boldsymbol{x}_t$ and $\boldsymbol{F}_t$. This is not surprising as data compression methods, such as PCA, are a form of "unsupervised learning." However, Equation (2) suggests that excess bond returns play the implicit role of a conditioning argument; namely, one should be able to tailor the extraction of hidden latent states $\boldsymbol{x}_t$ to the response variable $xr_{t+1}^{(n)}$. In this respect, "supervised learning" algorithms, such as the lasso, elastic net, partial least squares, regression trees, random forests, and NNs, that explicitly condition on the response variables to summarize the information in the predictors, may arguably prove useful to overcoming the limitations of standard data compression methods.[6] Second, traditional PCA and factor analysis (FA) are based on the assumption that all variables could bring useful information for the prediction of future excess bond returns, although the impact of some of them could be small. However, PCA or FA does not guarantee that by simply adding any number of predictors, we can be sure that the extracted factors will provide an optimal summary. Boivin and Ng (2006) formalize this argument by providing evidence that the structure of the common components is sensitive to the input variables, and that more data does not always mean there will be more sensible estimates. In this respect, one may want to "select" the variables that actually matter for forecasting excess bond returns. Penalized regressions, such as lasso and elastic net, as well as NNs exploit the entire span of the input variables without imposing that they all carry useful information for determining excess bond returns. The existing literature on bond return predictability has vastly ignored the potential capability of machine learning techniques to address the issue of nonlinearity and variable regularization. Arguably, this comes at the expense of not fully capturing the extent to which yields and

[6] Ludvigson and Ng (2009, p. 5034) acknowledge that "factors that are pervasive for the panel of data [input] need not be important for predicting [the output]" and propose a three-step forecasting procedure where a subset of principal components extracted from a large panel of macroeconomic variables is selected according to the information criteria before running the bond return forecasting regressions. In line with this intuition, we provide evidence that supervised learning methodologies, such as NNs, are useful to exploit the information in predictors other than yields, and to improve the out-of-sample forecast of bond returns.

macroeconomic variables are relevant for the measurement of expected excess bond returns. This is the focus of our paper.

2. Research Design

In this section, we outline the research design for the empirical analysis. We start with a description of the data, along with the specific applications. We then review the methodologies implemented in the main empirical analysis. We conclude with a short discussion of our estimation strategy.

2.1 Data and empirical applications

The empirical analysis is based on two main benchmark applications. The first application concerns the forecasting of future bond excess returns based on the cross-section of yields as originally proposed by Cochrane and Piazzesi (2005). We use the novel zero-coupon Treasury yield curve data set constructed by Liu and Wu (2019). This data set allows us to study bond returns with maturity of more than 5 years (the longest maturity in the Fama-Bliss data set). This is important since long maturity yields contain substantial extra predictive power over and above the first five yields (Le and Singleton, 2013, discuss the importance of using long-maturity bond yields in assessing the dynamic properties of risk premiums in Treasury markets). Moreover, Liu and Wu (2019) construct the zero-coupon curve using a nonparametric kernel-smoothing method that does not discard Treasury bills, which is instead the case for the parametric approach adopted by Gurkaynak et al. (2007). This is important since Liu and Wu (2019) find that securities at the short end of the yield curve contain important information in disciplining the overall behavior of the curve. Using the Liu and Wu (2019) yield curve data set, we then construct forward rates and excess bond returns as described in Section 1. We focus on bonds with maturities up to 10 years. Since the U.S. Treasury started issuing 10-year notes in September 1971, this also defines the start of our sample period. We do not use bonds with longer maturities for two reasons. First, the Treasury began issuing 20-year bonds in July 1981, and 30-year bonds in November 1985. This would force us to start the analysis later, reducing further the out-of-sample period since training the NNs

requires a sufficient amount of data.[7] Second, the issuance of long-maturity Treasury notes and bonds occurs at irregular intervals; to compensate for the lack of observations at long-maturities, the Liu and Wu (2019) method pulls information for the 20- and 30-year bonds from maturities that are 10 years (or more) away. The second application consists of forecasting future bond excess returns based on both forward rates and a large panel of macroeconomic variables as proposed by Ludvigson and Ng (2009). We consider a balanced panel of $N = 128$ monthly macroeconomic and financial variables. McCracken and Ng (2015) provide a detailed description of how variables are collected and constructed. The series were selected to represent broad categories of macroeconomic time series: real output and income, employment and hours, real retail, manufacturing and sales data, international trade, consumer spending, housing starts, inventories and inventory sales ratios, orders and unfilled orders, compensation and labor costs, capacity utilization measures, price indexes, interest rates and interest rate spreads, stock market indicators, and foreign exchange measures. This data set has been widely used in the literature (e.g., Ludvigson and Ng, 2009; Stock and Watson, 2006, 2002b), and permits comparison with previous studies.

2.2 Forecasting methods

2.2.1 Principal component regressions and partial least squares. The first method we employ is a linear, dimensionality-reduction technique known as principal component regressions (PCRs). Undoubtedly, PCRs constitute the most common method used to forecast interest rates and Treasury bond returns. In the classical implementation of PCRs, the target variable is discarded when extracting the latent factors. Thus, we also consider an alternative data compression methodology called partial least squares (PLS). Unlike PCR, with PLS the common components of the predictors are derived by conditioning on the joint distribution of the target variable and the regressors. Internet Appendix E.1 provides additional details on PLS, while contrasting this method to PCRs and penalized regressions, which we discuss next.

2.2.2 Penalized regressions: Ridge, lasso, and elastic net. Confronted with a large set of predictors, a popular strategy is to

[7] Contrary to typical machine learning applications, such as image recognition, common signal-to-noise ratios in financial data are low, exacerbating the need for sufficient data. Hence, we delay the start of the out-of-sample period so far that we have at least a handful of observations per weight to be estimated in our smallest NN (i.e., a single hidden layer with three nodes). For larger network architectures in which the number of parameters exceeds the number of available observations, regularization methods are used to ensure satisfactory training.

impose sparsity/shrinkage in the set of regressors via a penalty term. The idea is that by selecting a subset of variables with the highest predictive power out of a large set of predictors, and discarding the least relevant ones, one can mitigate in-sample overfitting and improve the out-of-sample performance of the linear model. In its general form, a penalized regression entails adding a penalty term on top of the ordinary least squares (OLS) objective function $\mathcal{L}_{OLS}(\boldsymbol{\theta}) = \frac{1}{t}\sum_{\tau=1}^{t-1}\left(xr_{\tau+1}^{(n)} - \alpha - \boldsymbol{\beta}^\top \boldsymbol{y}_\tau\right)^2$ with $\boldsymbol{\theta} = \left(\alpha, \boldsymbol{\beta}^\top\right)$:

$$\mathcal{L}(\boldsymbol{\theta};\cdot) = \underbrace{\mathcal{L}_{OLS}(\boldsymbol{\theta})}_{\text{Loss Function}} + \underbrace{\phi(\boldsymbol{\beta};\cdot)}_{\text{Penalty Term}} \quad . \tag{5}$$

Depending on the functional form of the penalty term, the regression coefficients can be regularized and shrunk towards zero (as in ridge), exactly set to zero (as in lasso), or a combination of the two (as in elastic net). In Appendix E.2, we describe each method in detail. Penalized regressions still do not account for nonlinear relations. To address this issue, we consider a third class of nonlinear methods: "shallow learners," such as regression trees, and more deep structures, such as neural networks.

2.2.3 Regression trees. Regression trees are based on a partition of the input space into a set of "rectangles." Then, a simple linear model is fit to each rectangle. Figure 1 displays an example of a binary partition (panel a) and the corresponding regression tree (panel b). Regression trees are conceptually simple, yet powerful, and therefore highly popular in the machine learning literature. In addition to a standard regression tree methodology, we consider extensions that employ an ensemble of individual trees, like "random forests' (Breiman, 2001), and, furthermore, take into account the randomness in the predictors' partition process, like "extremely randomized trees" (Geurts et al., 2006). Appendix E.3 provides additional technical details on the estimation of regression trees (and their extensions).

2.2.4 Neural networks. Neural networks (NNs) represent a widespread class of supervised learning methods. We focus on traditional "feed-forward" networks or multilayer perceptrons (MLP). Throughout the paper, we follow Feng et al. (2018) and adopt the convention of counting only the hidden layers, without including the output layer. For details on the estimation of NNs, see Appendix E.4. Next, we provide a high-level description of the NNs structure we consider in our empirical applications.

Neural networks: yields-only. When we forecast bond returns using only information from the term structure of interest rates, we use

variants of the NN architecture depicted in Figure 2. That is, we use a classical MLP. An MLP consists of at least three layers of nodes (the case displayed in the figure): an input layer, a hidden layer, and an output layer. The result is a powerful learning method that can approximate virtually any continuous function with compact support (Cybenko, 1989; Diaconis and Shahshahani, 1984; Hornik, Stinchcombe, and White, 1989; Kolmogorov, 1957). In the empirical application, we study the predictive accuracy of this classical network as we vary the number of hidden layers, as well as the number of nodes per layer.

Neural networks: macro plus yields. When we forecast bond returns using macroeconomic variables in addition to information from the term structure, we consider three alternative specifications that extend the typical MLP structure by taking into account the economic structure of the input data, as well as the nature of the forecasting problem. The first specification, displayed in panel (a) of Figure 3, can be thought of as a "hybrid" modeling framework in the sense that forward rates are simply included as an additional predictor in the output layer ("fwd rates direct"). This structure simulates the idea of Ludvigson and Ng (2009) in which the latent factors $\boldsymbol{F}_t$ are extracted from a large cross-section of macroeconomic variables and a linear combination of forward rates is included as proposed by Cochrane and Piazzesi (2005). We label this structure where forward rates have been preprocessed a *hybrid* neural network. The second specification, displayed in panel (b) of Figure 3, ensembles two separate networks at the output layer level: one network is trained for the forward rates ("fwd rates net") and one for the macroeconomic variables. In contrast to the specification in panel (a), this specification allows for a nonlinear transformation of forward rates. The third specification, displayed in panel (c) of Figure 3, entails a collection of networks, one for each group of macroeconomic variables, which are trained in parallel and ensembled at the output layer level. The groups of macroeconomic variables are constructed following the classification provided by McCracken and Ng (2016).[8] This latter specification, which we call *group ensembling* switches off the (nonlinear) interactions across groups of macroeconomic variables, which are instead present in the specifications of panels (a) and (b). To the best of our knowledge, these specifications have not been proposed before in the empirical asset pricing literature within the context of nonlinear predictive methods. We study the predictive accuracy of these networks

[8] Specifically, we group 128 predictors into eight categories: (1) output (16 series); (2) labor market (31 series); (3) housing sector (10 series); (4) orders and inventories (10 series); (5) money and credit (14 series); (6) bond and FX and interest rates or financial (22 series); (7) prices or price indices (16 series); and (8) stock market (5 series). Four series in our sample could not be matched to McCracken and Ng (2016) and are left unclassified.

as we vary the number of hidden layers and the number of nodes per layer.

2.3 Estimation strategy

Following common machine learning practice, we split the data into three subsamples: a training set used to train the model, a validation set used to evaluate the estimated model on an independent data set, and a testing set, which represents the out-of-sample period in a typical forecasting exercise.[9] We follow Gu et al. (2018) and, conditional on the estimates from the training set, we produce forecasting errors over the validation sample. We then use the prediction errors over the validation sample to iteratively search the hyperparameters that optimize the objective function. Thus, the validation sample represents the part of data that is used to provide an unbiased evaluation of a model fit. It is trivial to see that predictions in the validation set are not out-of-sample as they are used to tune the model hyperparameters. The third subsample, or the testing sample, contains observations that are not used for estimation or tuning. This third subsample is known as the "out-of-sample" period and can be used to test the predictive performance on observations yet unseen by the machine learning model.[10] There are a variety of splitting schemes that could be considered but the trade-off between the size of the training and validation samples is ultimately an empirical question (for a comprehensive survey of cross-validation procedures for model selection, see Arlot et al., 2010). We keep the fraction of data used for training and validation fixed at 85% and 15% of the in-sample data, respectively. The training and the validation samples are consequential. In this respect, we do not cross-validate by randomly selecting independent subsets of data to preserve the time-series dependence of both the predictors and the target variables. Forecasts are produced recursively by using an expanding window procedure; that is, we reestimate a given model at each time t and produce out-of-sample forecasts for 1-year holding period excess returns. Figure 4 provides a visual representation of the sample splitting scheme we adopt in the empirical analysis. Notice that for some of the methodologies, validation is not required. For instance, neither standard linear regressions nor PCA require a pseudo out-of-sample period to validate the estimates. In these cases, we adopt a traditional separation between in-sample versus

[9] See Chen et al. (2017) for an in-depth discussion of model fragility (i.e., the tendency of a model to overfit the data in-sample) at the expense of poor out-of-sample performance.

[10] Note, the out-of-sample period is only "pseudo" out-of-sample in the sense that its observations are available to the researcher at the time of the study. Nevertheless, given that the observations in the out-of-sample period have not been used to train the model itself, it is standard in the machine learning literature (in asset pricing) to refer to the testing sample as the "out-of-sample" period (e.g., see Feng et al., 2018; Gu et al., 2018).

out-of-sample period, where the former consists of both the training data and the validation data. We recursively forecast bond returns in excess of the short-term rate. We focus on 1-year holding period excess returns for comparability with the original settings in Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009). We recursively fit machine learning methods at each time t (i.e., we increase the in-sample period by one monthly observation). This scheme allows us to incorporate the most recent updates from the yield curve, as well as the set of macroeconomic and financial variables.[11] When enlarging the in-sample period, we roll it forward to include the most recent information in a recursive fashion but keep constant the ratio between the training sample and the validation sample. In this respect, we always retain the entire history of the training sample, thus its window size gradually increases. By keeping the proportion of the training and validation sets fixed, the validation sample gradually increases as well. The result is a sequence of performance evaluation measures that correspond to each recursive estimate. Although computationally expensive, this leverages more information for prediction. In each empirical application, the sample spans from 1971:08 to 2018:12. Since our forecasting exercise is iterative, the computational challenge becomes sizable. Thus, we perform all computations on a high-performance computing cluster consisting of 84 nodes with 28 cores each, totaling to more than 2,300 cores. Internet Appendix F provides a complete description of the computational specifications.

2.4 Statistical performance

We compare the forecasts obtained from each methodology to a naive prediction based on the historical mean of excess bond returns. In particular, we calculate the out-of-sample predictive R^2 as suggested by Campbell and Thompson (2007). The R^2_{oos} is akin to the in-sample R^2 and is calculated as

$$R^2_{oos} = 1 - \frac{\sum_{t_0=1}^{T-1} \left(xr_{t+1}^{(n)} - \widehat{xr}_{t+1}^{(n)}(\mathcal{M}_s) \right)^2}{\sum_{t_0=1}^{T-1} \left(xr_{t+1}^{(n)} - \overline{xr}_{t+1}^{(n)} \right)^2}, \tag{6}$$

where $\overline{xr}_{t+1}^{(n)}$ is the prediction error obtained based on the historical mean and $\widehat{xr}_{t+1}^{(n)}(\mathcal{M}_s)$ is the forecast of the excess bond returns for maturity n obtained using model $\mathcal{M}_s$, and t_0 is the date of the first prediction. The first forecast error obtains by comparing the excess

[11] Note that this is different from the implementation of machine learning methods for stock returns, where trading signals from firm characteristics are often updated once per year, which means that retraining of the models could be performed with lower frequency (Gu et al., 2018).

holding period return during the February 1989 through January 1990 period and its forecast made on January 1989. We also build a portfolio-level return forecast from the individual maturity forecasts produced by our models. We construct the forecast of an equally weighted portfolio by $\widehat{xr}_{t+1}^{(EW)} = \frac{1}{6}\sum_{n=2}^{10} \widehat{xr}_{t+1}^{(n)}(\mathcal{M}_s)$. We compute $R^2_{oos,EW}$ by constructing forecast errors using the realized return $xr_{t+1}^{(EW)} = \frac{1}{6}\sum_{n=2}^{10} xr_{t+1}^{(n)}$ and comparing to the historical mean. Testing the null hypothesis, $R^2_{oos} \leq 0$, against the alternative hypothesis, $R^2_{oos} > 0$, is tantamount to testing whether the predictive model has a significantly lower mean squared prediction error (MSPE) than the historical average benchmark forecast. Thus, to test whether R^2_{oos} is significantly greater than zero, we implement the MSPE-adjusted Clark and West (2007) statistic. A limitation of the R^2_{oos} measure is that it does not explicitly account for the risk borne by an investor over the out-of-sample period. To this end, we also calculate realized utility gains for a mean-variance or power utility investor (see Section 5).

3. An Empirical Study of U.S. Treasury Bonds

3.1 Bond return predictability and the yield curve

We start by forecasting the excess returns of Treasury bonds with the yield curve. In this case, the classical specification is given by the principal component regression (PCR) in Equation (3). In words, excess returns are regressed on PCs of the Treasury term structure, that is, $\boldsymbol{x}_t = [PC_{1,t} \ldots, PC_{k,t}]$. We use the first three, five, or ten PCs. The case with ten PCs essentially corresponds to the setting in Cochrane and Piazzesi (2005), where excess returns are regressed on a linear combination of short-rate, $y_t^{(1)}$, and nine forward rates for loans between $t+n-1$ and $t+n$, $f_t^{(n)}, n=2,\ldots,10$. Table 1 displays the out-of-sample R^2_{oos} (and its p-value) for different bond maturities; we also report the R^2_{oos} for an equally weighted portfolio. Panel A of Table 1 displays the results for PCRs and partial least squares (PLS). The first three rows show the predictive performance of PCRs for $k=3,5$, and 10 PCs. The predictive R^2 are negative across different maturities. A parsimonious representation with only three PCs significantly outperforms the specification with five and ten PCs, particularly at long maturities. Further, adding simple forms of nonlinearities, such as squared PCs, worsens performance. Perhaps surprisingly, a linear supervised learning method like PLS does not lead to any improvement relative to PCR. Panel B of Table 1 displays the results from various configurations of the linear penalized regressions. Ridge regression performs poorly out-of-sample with predictive R^2_{oos} that are mostly negative across bond maturities. The second and third rows of panel B show that sparse

modeling improves the forecasting performance of the current term structure relative to ridge: the R^2_{oos} for both the lasso and elastic net are positive for maturities of more than 4 years, as well as for the equally weighted bond portfolio. However, the performance of elastic net is on par to that obtained from a principal component regression with three PCs, which proves to be a tough benchmark. Panel C of Table 1 shows the results for boosted regression trees, random forests, extremely randomized trees, and neural networks. All these methods attain good performance with significantly positive R^2_{oos} across maturities. With respect to trees, the randomization of the feature split locations (i.e., for extreme trees) improves the out-of-sample performance over random forests, particularly for long maturities. Turning to NNs, we observe that a shallow network with a single hidden layer and three nodes (cf. Figure 2) performs on par with the best, deeper network with two hidden layers and seven nodes. Interestingly, further increasing the depth of the network deteriorates its performance. This continues to be the case even when we consider alternative structures, like a NN with three hidden layers and pyramidal node architecture.[12] The last row of panel C presents the results of an interesting case. In their paper, Cochrane and Piazzesi (2005) conclude that lags of forward rates (dated $t-1$ and earlier) contain information about excess returns that is not spanned by month t forward rates. They note that this result is inconsistent with the logic that the time-t term structure contains all information relevant to forecasting future yields and excess returns (cf. Equation (2)). We therefore ask whether the flexibility of a NN can help reconcile the theoretical assumption that yields at time t already incorporate all information about the term structure that is needed to understand bond risk premiums. To this end, the last row in panel C reports the results obtained by feeding the NN with the ten forward rates at time t and lagged forward rates from time $t-11$ to $t-1$. By comparing the last row to the NN with one hidden layer and three nodes, we find no evidence that we can improve on a NN that uses just the month-t forward rates.[13] The evidence in Table 1 confirms that, even in a small dimensional setting using only information in the yield curve, we can improve bond risk premium measurement by acknowledging that (1) the function $g(\boldsymbol{y}_t;N)$ in Equation (2) can be nonlinear, and that (2) such improvement depends on the neural network specification; a shallow NN with one hidden layer performs on par with a network with two layers, but deeper networks worsen the performance. Finally, consistent

[12] In Internet Appendix B, we employ a Diebold and Mariano (1995) pairwise test to compare the predictions from different models.

[13] We report the best specification with lagged forward rates (i.e., a shallow NN with seven nodes). NNs with more layers or a different number of nodes underperform.

with Equation (2), we find that lagged values of the yield curve cannot improve the forecast obtained using just the time-t term structure when we account for nonlinearities.

3.2 Bond return predictability and macroeconomic variables

Next, we consider the setup where information embedded in the yield curve does not necessarily subsume information contained in macro variables. In this case, the classical specification is given by Equation (4), where the factors F_t now have the potential to serve as the model's state vector beyond yields only. To ensure comparability with the literature, we adopt the specification proposed by Ludvigson and Ng (2009), whereby $\boldsymbol{F}_t$ is a subset of the first eight PCs extracted from a large cross-section of macroeconomic variables and $\boldsymbol{x}_t$ represents a linear combination of forward rates as proposed by Cochrane and Piazzesi (2005), aka the CP factor. Panel A in Table 2 displays results from a simple principal component regression with eight PCs, the specification proposed in equation (8) of Ludvigson and Ng (2009) (i.e., $\boldsymbol{F}_t = \left(F_{1t}, F_{1t}^3, F_{3t}, F_{4t}, F_{8t}\right)$), and partial least squares (PLS). Panel B shows the results from two alternative implementations of sparse and regularized linear regressions. In the first implementation ("using CP factor"), we employ the CP factor as an additional regressor; this specification ensures a closer comparability to Ludvigson and Ng (2009). In the second implementation ("using fwd rates directly"), we treat the whole set of forward rates as additional regressors with respect to macroeconomic variables. The results in panels A and B shows that (1) dense modeling, such as data compression techniques and ridge regression, tends to perform poorly out-of-sample and (2) sparse modeling with both regularization and shrinkage (i.e., elastic net regressions), perform well, particularly when restricting the linear combination of forward rates. Comparing panel B in Table 1 to that in Table 2, it seems apparent that there is information beyond the term structure of interest rates that can be used to predict bond returns. Turning to nonlinear machine learning methods, panel C in Table 2 shows the results from the three alternative network specifications discussed in Subsection 2.2.4: (1) a *hybrid* framework in which the forward rates enter linearly as additional predictors in the output layer ("fwd rates direct"; Figure 3, panel (a)); (2) a specification that ensembles one network for the forward rates ("fwd rates net") and one for the macroeconomic variables (Figure 3, panel (b)); and (3) a specification that entails a collection of networks, one for each group of macroeconomic variables (Figure 3, panel (c)). This latter specification, dubbed *group ensembling*, switches off the (nonlinear) interactions across groups of macroeconomic variables, which are present in the hybrid network, as well as in the specification that ensembles

separately forwards and macroeconomic variables. The performance of hybrid networks stands out. Interestingly, and differently from Table 1, increasing the depth of the NN from one to three layers improves its accuracy. However, a careful choice of network structure based on prior economic information exerts a great impact on performance. In particular, a one-layer *group-ensembled* model (see third-to-last row) performs on par with the three-layer *hybrid* NN for short- and medium-term maturities, and attains the highest predictive accuracy for the 7- and 10-year bonds. Interestingly, adding more layers is detrimental to the performance of the group-ensembled NN. Panel C in Table 2 also shows that the performances of boosted regression trees, random forests, and extreme trees improve substantially when using a large panel of macroeconomic information. In fact, the extreme trees performs better than shallow (hybrid and ensembled) NNs but worse than the (best-performing) one-layer NN with group ensembling; see Internet Appendix B for a formal comparison of the predictions from different models using a Diebold and Mariano (1995) pairwise test. The results in Table 2 show that macroeconomic variables carry information that is not contained in the yield curve. When we compare the best-performing (one-layer and three nodes) NN in Table 1 to the one-layer group-ensembled NN, we observe approximately a 10 percentage point increase in R^2_{oos} for each maturity. The results also show that the depth and structure of the network interact with one another: having a separate network for each group of macroeconomic variables compensates for the need of a deep NN when macroeconomic variables are processed together without further classification.[14]

4. Dissecting Predictability

4.1 Bond return predictability in expansions and recessions

We start by investigating whether bond return predictability varies over the economic cycle. To this end, we split the data into recession and expansion periods using the NBER recession indicator. Table 3 shows the R^2_{oos} values computed separately for the recession and expansion subsamples. For yields-only PCA, we recover a classical result: predictability is concentrated in economic recessions (in particular for long maturity bonds) and is absent during expansions. Turning to NNs, we continue to observe R^2_{oos} that are generally higher during recessions than in expansions. However, the difference in R^2_{oos} values decreases with bond maturity. More importantly, a formal test confirms that the bond return prediction from NNs is statistically different from that of

[14] Internet Appendix C.6 demonstrates that our results continue to hold when we allow for macroeconomic information to be released with a delay to investors.

the expectations hypothesis (EH) model, both in expansions and in recessions. In contrast to NNs, the return predictability implied by trees is actually stronger during expansions. A formal test confirms that the predictive accuracy of trees is significantly better than that generated by the EH benchmark only for expansion periods. Finally, a pairwise test (untabulated) confirms that the improvement of NNs over trees is mainly due to the better predictive accuracy of networks in recessionary periods; however, the predictions from trees and NNs are indistinguishable in expansions. Our finding that the predictability of bond returns implied by machine learning methods is not concentrated exclusively in bad times, but is present also in expansions is novel to the literature and contrasts with evidence for equities (Dangl and Halling, 2012; Rapach et al., 2010) and bonds (Gargano et al., 2019). In Internet Appendix C.1, we analyze the models' performance in different periods using the cumulative sum of squared errors and confirm that the out-performance of machine learning-based forecasts versus the EH benchmark is not concentrated in isolated events. Interestingly, however, it is possible to relate, ex post, NN forecasts to specific patterns of the yield curve. Internet Appendix Table C.1 shows that NNs and, in particular, the group-ensembled network that exploits macroeconomic and financial information in addition to interest rates predict high excess bond returns when there is a steep slope in the yield curve (e.g., right after recessions) and when the level of the yield curve is high. Further, these NN predictions (conditional on specific shapes of the yield curve) are highly correlated with realized returns, thus leading to high R^2s.

4.2 Understanding the performance of neural networks: Level, slope, or both?

In this subsection, we provide an heuristic interpretation of the performance of the NNs based on the Campbell and Shiller (1991) accounting identity (Equation (1)). Such identity posits that the forecasts of future yields (or their PCs) using current yields are necessarily also forecasts of expected log returns to bonds. Thus, we investigate the ability of the latent factors extracted by the NN to predict the year-on-year changes in the first three PCs extracted from the cross section of forward rates. $PC_{1,t}, PC_{2,t}$, and $PC_{3,t}$ denote the principal components. The auxiliary forecasting regressions are

$$PC_{i,t+1} - PC_{i,t} = b_0 + \boldsymbol{b}_1^\top \mathcal{P}_t + \boldsymbol{b}_2^\top \boldsymbol{x}_t + \epsilon_{i,t+1} \qquad \text{for} \quad i = 1,2,3,$$

where we stack the first three PCs of the term structure in the vector $\mathcal{P}_t$ and denote by $\boldsymbol{x}_t$ the hidden factors extracted by the NN.[15] Table

[15] Take a shallow network with $L = 1$ hidden layers as an example, that is, $E_t\left[xr_{t+1}^{(n)}\right] = \hat{\alpha}_n + \hat{\boldsymbol{\beta}}_n^\top \boldsymbol{x}_t$, where $\boldsymbol{x}_t = h(\boldsymbol{W}\boldsymbol{y}_t + b)$. The latent factor $\boldsymbol{x}_t$ is extracted at each time t conditional on estimates of the weights $\boldsymbol{W}$ and bias b from the vector of inputs y_t.

4 reports the in-sample R^2 of such predictive regressions. The first row provides the benchmark results based on the sole vector $\mathcal{P}_t$. In support of Duffee (2011a, 2013), we find weak evidence that changes in the first PC (level) are forecastable ($R^2 = 9.28\%$ being the lowest), whereas the slope and curvature are unquestionably forecastable with 21.66% (48.70%) of the variation in slope (curvature) that is predictable. Next, we add to the regression the hidden factors extracted from the two best-performing NNs in Tables 1 and 2: the *NN One-Layer (three nodes)*, when forecasting only with the forward rates, and *NN One-Layer Group Ensem + fwd rate net*, when including also macroeconomic variables. We observe that the factors extracted from NNs that use yields-only (second row in Table 4) contribute substantially to the predictability of the level and curvature. On the other hand, the statistical evidence for slope forecasts—after controlling for the standard three principal components—is weak. We conclude that there is substantial information in the time-t term structure not only about future values of slope but also about the level (and curvature). Standard PCs are not entirely able to extract all the relevant information about the level. A shallow NN is successful in extracting such information about the future level of the curve, an information which leads to excess returns being more predictable out-of-sample. The last row in Table 4 focuses on factors extracted from a NN that exploits macroeconomic variables in addition to forward rates. We observe that the factors extracted from the group-ensembled NN not only contribute to the ability to predict the level of the yield curve but also contribute to the slope. This suggests that the slope of the yield curve is related to the state of the economy, and our NN is able to extract the relevant information from our large set of macroeconomic variables.

4.3 Relative importance of macroeconomic variables

In this subsection, we investigate which variables drive the performance of NNs studied in Tables 1 and 2. To this end, we examine the marginal relevance of single variables based on the partial derivative of the target variable, $xr_{t+1}^{(n)}$, with respect to each input, where the gradient is evaluated at the in-sample mean value of the input, that is,

$$\mathbb{E}\left[\frac{\partial}{\partial y_{it}} xr_{t+1}^{(n)} \middle| y_{it} = \overline{y}_i\right], \tag{7}$$

where $\overline{y}_i$ represents the in-sample mean of the input variable i. The partial derivative represents the sensitivity of the output to the ith input evaluated at its sample mean, conditional on the network structure and the average value of the other input variables (see Dimopoulos et al., 1995). Further, we focus on the magnitude of the gradients by taking their absolute values. Internet Appendix C.3 provides additional

discussion about the computational details. Figure 5 shows the relative importance of each input variable based on the gradient in Equation (7). The analysis is carried out for the best-performing NN in Table 2, the NN with one hidden layer where macroeconomic variables and forward rates are modeled separately through ensembling at the output layer level. For ease of exposition, we report only the top 20 most relevant predictors. Panels (a) and (b) display results for the 2- and 10-year bond maturities, respectively. To gain further intuition about the systematic patterns in the drivers of expected excess bond returns, we also calculate the relative importance from the absolute gradients averaged for each class of input variables as labeled in McCracken and Ng (2016). The results, displayed in panels (c) and (d), provide an indication of which economic category dominates. Comparing panels (c) to (d) of Figure 5, we observe that the variables pertaining to inflation, and money and credit are important independent of the maturity considered. However, the results also show that the effect of other classes of predictors is heterogeneous over the term structure. For instance, variables related to the stock and labor market are more important for the short-end of the yield curve, while variables pertaining to the categories output & income and orders & inventories become more relevant for the long-end of the curve. This analysis is important for two reasons. First, it suggests that inflation has a level-like effect on bond yields, whereas variables pertaining to the labor market (order & inventories) are likely to have a slope effect acting mostly on short-term (long-term) bonds, while leaving long-term (short-term) bonds unaffected. This evidence can therefore provide guidance for theoretical models that include macro risk factors as drivers of bond risk premiums by highlighting their permanent or transient nature. Second, our analysis suggests that the use of excess bond returns averaged across maturities is unlikely to flesh out the true impact of macro risk factors on bond risk premiums.[16] In all, our results show that there is information in macroeconomic and financial variables beyond that conveyed by the yield curve, and this information improves the predictions of bond returns (Tables 1 and 2). In addition, the type of unspanned (by the yield curve) information may vary across different bond maturities. To our knowledge, this fact is novel and provides a new angle to revisit a central question in the term structure literature, which is whether yields data contain all the relevant information to predict future bond returns.

[16] Our evidence of a level-like effect of inflation on bond yields is in line with the analysis in Joslin et al. (2014, section VI). Furthermore, within the orders and inventories category, we find that "New Orders for Durable Goods" is of single-most importance for forecasts at the far end of the curve (see panel (b) of Figure 5). Yang (2011) provides theoretical and empirical evidence that is consistent with our finding by showing that the impact of durable consumption growth on the yield curve strengthens with bond maturity.

4.4 Interactions within or across categories?

The finding that a shallow group-ensembled network performs on par with (or better than) a deep, three layer NN that models all macroeconomic and financial variables together is novel to the literature on machine learning and asset prices, and is important for two reasons. First, our results on group-ensembled NNs show that the depth of the network and the economic priors used to design it (e.g., grouping variables that pertain to the same category) interact with one another. In particular, for the application at hand, group ensembling can compensate for the depth of the network. Second, our results highlight what type of nonlinearities are important from an economic perspective: Is it the interaction of many variables (across categories) or a higher polynomial of the same variable (within a category)? Since our group-ensembled network switches off interactions across categories, our results show that nonlinearity within a group drives the outperformance of the network. Of course, this is true only in so far as cross-group network weights are not already small in the fully connected network. Internet Appendix Table C.2 shows that this is indeed not the case. More precisely, we calculate the second-order derivatives of the output with respect to each input conditional on the inputs being in different groups of predictors.[17] We estimate cross-group partial derivatives for both a fully connected network and a network with group ensembling. The results in panel A show that the absolute value of the sum of interaction derivatives in the fully connected network is orders of magnitude larger than the value obtained from a group-ensembled NN (i.e., the cross-group interactions are indeed large in the fully connected network). In panel B of Internet Appendix Table C.2, we provide the within-group second-order partial derivatives of the outputs with respect to the inputs conditional on being in the same group. We find that the magnitude of the within-group effects is similar between the fully connected and group-ensembled NNs. Hence, the performance of the group-ensembled NN is driven by imposing the absence of interactions across categories while allowing for nonlinearity within an economic category.

4.5 Model uncertainty

Faced with multiple NN estimates, the question of how to best exploit ex ante different forecasting specifications immediately arises. In particular, should we rely on a single, ex post, dominant model specification or

[17] That is, we calculate

$$\mathbb{E}\left[\frac{\partial^2}{\partial y_i \partial y_j} x r_{t+1}^{(n)} \Big| y_i \in G_A, y_j \in G_B\right], \tag{8}$$

where G_A and G_B are two nonoverlapping groups of variables defined as in McCracken and Ng (2016), and we sum the absolute value of Equation (8) for each interaction of variables that do not belong to the same group.

should a combination of different forecasts be used to produce a better forecast? From a pure theoretical perspective, unless the best forecasting model can be identified ex ante, forecast combinations may offer some diversification benefits (see Clemen, 1989, for a discussion). However, it may also be the case that a carefully designed validation procedure is able to systematically pick the best out-of-sample model specification. To answer this question, we first compare the best-performing NN within the context of forecasting bond returns with both yields and macroeconomic variables—the *NN One-Layer Group Ensem + fwd rate net* (see Table 2)—against a combined forecast of the form,

$$\hat{xr}_{c,t+1}^{(n)} = \sum_{i=1}^{\mathcal{M}} \omega_{i,t} \cdot \hat{xr}_{i,t+1}^{(n)} \tag{9}$$

where $\hat{xr}_{c,t+1}^{(n)}$ denotes the one-step-ahead combined forecast for maturity n, $\omega_{i,t}$ is the weight assigned to each individual prediction, $\hat{xr}_{i,t+1}^{(n)}$, and $i=1,\ldots,\mathcal{M}$ are the forecasts from the set of NNs $\mathcal{M}$ in Table 2. We choose two representative model combination schemes: (1) an equal weight assigned to each forecast, that is, $\omega_{i,t}=1/\mathcal{M}$, and (2) a linear combination of forecasts based on the validation losses, that is, $\omega_{i,t}=\frac{1/L(e_{i,t}|\theta_i)}{\sum_{i=1}^{\mathcal{M}}(1/L(e_{i,t}|\theta_i))}$, where $L(e_{i,t}|\theta_i)$ is the validation loss obtained from the cross-validation prediction error $e_{i,t}$ given the network hyperparameters θ_i.[18] In addition to an equal-weight and a relative-performance combination scheme we also compare our best-performing NN forecasts against a full-blown cross-validated network. In particular, we expand the set of hyperparameters that are cross-validated and selected every 5 years; that is, we let optimization procedures select not only the dropout rate and the L1/L2 penalties but also the number of hidden layers, the nodes per group of macroeconomic variables, and the nodes in the forward rate network (see Internet Appendix Table F.1 for the details of the hyperparameters).[19] The logic for comparing our best-performing model against two representative forecast combination schemes and a full-blown cross-validated network is to make sure that our results are robust to more flexible and adaptive modeling strategies. Internet Appendix Table C.3 reports the results. Two interesting aspects emerge from the table. First, the group-ensembled NN (see first row) outperforms both forecast combination schemes (second and third rows),

[18] Note that the loss function we use is a simple mean squared error plus a penalty to induce regularization in the weights. This means that the weighting scheme reflects the performance of each model relative to the performance of the average model (e.g., Bates and Granger, 1969; Newbold and Granger, 1974; Stock and Watson, 1998; Elliott and Timmermann, 2004).

[19] We thank an anonymous referee for suggesting this exercise.

the sole exception being at the 2-year maturity. Second, our group-ensemble network specification tends also to perform on par with, or better than, the full-blown cross-validated network for maturities of more than 3 years. Hence, we conclude that the optimal structure of layers and nodes, which is endogenously chosen through an adaptive cross-validation exercise, does not improve (except for the very short-end) on a more parsimonious and economically motivated network structure, like our one-layer group-ensemble NN. We next examine the recursive performance, meaning cross-validation error, of the top performing neural network. In principle, the performance of the *NN One-Layer Group Ensem + fwd rate net* specification could be justified by the fact that such network is consistently chosen through cross-validation and across time. In Internet Appendix Table C.4, we compare how often the four on average best-performing NN structures from Table 2 are selected throughout the out-of-sample period. Two interesting facts emerge. First, our best-performing group-ensembled NN generates the smallest validation error (and thus it would be chosen through cross-validation) for about a half of the out-of-sample period. This could explain the similar performance between the best-performing (group-ensemble) NN in Table 2 and the full-blown cross-validate model, which contains the benchmarking specification in the model set. Second, shallow NNs tend to consistently deliver lower validation errors. This reinforces our result that network depth and structure interact with one another: in fact, a carefully designed network outperforms a deeper, and more data-driven, network structure.

5. Economic Value of Excess Bond Return Forecasts

So far, our analysis concentrated on statistical measures of predictive accuracy. Next, we evaluate whether the apparent gains in predictive accuracy translate into better investment performance relative to the no-predictability alternative. This is important since Thornton and Valente (2012) find that yield-based predictors, when used to guide the investment decisions of an investor with mean-variance preferences, do not lead to higher out-of-sample Sharpe ratios compared with investments based on a no-predictability expectations hypothesis (EH) model. Sarno et al. (2016) reach a similar conclusion. However, the large time variation in expected bond returns that is detectable in real time by machine learning methods naturally calls for revisiting these findings.

5.1 The asset allocation framework

To assess the economic importance of machine learning methods (particularly trees and NNs) in forecasting bond returns, we use a classic portfolio choice problem (Della Corte et al., 2008; Thornton and Valente,

2012). Specifically, we consider an investor who optimally invests in a portfolio comprising $K+1$ bonds: a risk-free one-period bond and K risky n-period bonds. We consider both univariate and multivariate asset allocation exercises. In the univariate case, the investor selects between an n-year bond and the risk-free return based on the expected return implied by a given model. We focus on the results for $n=2$ and $n=10$ years. In the joint asset allocation exercise, the investor selects bonds with maturities of 2 to 10 years, and the risk-free return. We analyze the asset allocation decisions of a mean-variance investor and those of a power utility investor. The discussion of the power utility problem and its solution is in Internet Appendix D.1, and in the remaining discussion we focus on the mean-variance case. At each time t, the decision-maker selects the weights on the risky n-period bonds $\boldsymbol{w}_t = \left[w_t^{(2)} \dots w_t^{(10)}\right]$ to maximize the quadratic utility:

$$\max_{\boldsymbol{w}_t} \; E[R_{p,t+1}] - \frac{\gamma}{2}\mathrm{Var}(R_{p,t+1}),$$

where γ is the risk aversion coefficient of the mean-variance investor, $R_{p,t+1} = 1 + y_t^{(1)} + \boldsymbol{w}_t' \boldsymbol{xr}_{t+1}$ is the gross return on the portfolio, $E[R_{p,t+1}]$ is the sample mean portfolio return, and $\mathrm{Var}(R_{p,t+1})$ is the sample variance portfolio return. Then the solution of the above optimization is $\boldsymbol{w}_{t,s} = \frac{1}{\gamma}\Sigma_{t+1|t}^{-1}\widehat{\boldsymbol{xr}}_{t+1}(\mathcal{M}_s)$, where $\widehat{\boldsymbol{xr}}_{t+1}(\mathcal{M}_s)$ is the vector of bond returns' forecast obtained using model $\mathcal{M}_s$, and $\Sigma_{t+1|t} = Var_t(\boldsymbol{xr}_{t+1} - E_t[\boldsymbol{xr}_{t+1}])$. For the univariate allocation exercise we have: $w_{t,s}^{(n)} = \frac{\widehat{xr}_{t+1}^{(n)}(\mathcal{M}_s)}{\gamma \sigma_{t+1|t}^{(n)}}$ where $\widehat{xr}_{t+1}^{(n)}(\mathcal{M}_s)$ is the bond returns' forecast for maturity n given model $\mathcal{M}_s$, and $\sigma_{t+1|t}^{(n)}$ the diagonal element of $\Sigma_{t+1|t}$ relative to the bond with n-year maturity. To proxy for $\Sigma_{t+1|t}$, we employ a rolling sample variance estimator as in Thornton and Valente (2012): $\widehat{\Sigma}_{t+1|t} = \sum_{l=0}^{\infty} \Omega_{t-l} \odot \epsilon_{t-l}\epsilon_{t-l}'$, where $\epsilon_t = \left[\epsilon_t^{(2)} \dots \epsilon_t^{(10)}\right]'$ are forecast errors, $\Omega_{t-l} = \alpha \exp(-\alpha)\mathbf{1}\mathbf{1}'$ is a symmetric matrix of weights, $\odot$ denotes element-by-element multiplication, and we set the decay rate α to 0.05 (same value as in Thornton and Valente (2012) and within the range of those reported in studies like Fleming et al. (2001)). We also winsorize the weights for each of the n-period bonds to $-1 \le w_t^{(n)} \le 2$ to prevent extreme investments; however, we evaluate the robustness of our results to alternative assumptions about the portfolio weights. Finally, to make our results directly comparable to other studies (e.g., Gargano et al., 2019; Thornton and Valente, 2012), we assume a coefficient of risk aversion of five. Given the Markowitz optimal weights on the risky bonds, we compute the realized utilities. Then, following Fleming et al. (2001), we obtain the certainty equivalent gains (annualized and in percentages) by equating the average utility of the EH model with the

average utility of any of the alternative models. To test whether the certainty equivalent return (CER) values are statistically greater than zero, we use a Diebold and Mariano (1995) test. Specifically, to evaluate the allocation implied by the NN forecasts, we estimate the following regression:

$$u_{t+1,NN} - u_{t+1,EH} = \alpha^{(n)} + \varepsilon_{t+1} \ ,$$

where $u_{t+1,s} = \boldsymbol{w}'_{t,s}\boldsymbol{x}\boldsymbol{r}_{t+1} - \frac{\gamma}{2}\boldsymbol{w}'_{t,s}\Sigma_{t+1}\boldsymbol{w}_{t,s}$ and $s = \{EH, NN\}$; that is, we use the optimal weights together with the realized returns.

5.2 Asset allocation: Results

Table 5 shows the annualized CER values computed relative to the EH model. Positive values indicate that the predictive model performs better than the EH model. We focus on the best predictive models from Tables 1 and 2, the extreme trees and the *NN One-Layer (three nodes)*, when forecasting only with forward rates, or *NN One-Layer Group Ensem + fwd rate net*, when including the macroeconomic variables. With the sole exception of the mean-variance investor selecting the 2-year bond, the remaining CER values for the trees and NNs are significantly higher than those generated by the EH benchmark. The CER values increase with bond maturity, but the highest CER values are found in the multivariate setting, which suggests that the economic gains associated with NNs forecasts are not limited to specific maturities. Interestingly, when the (mean-variance or power utility) investor makes no use of information beyond the term structure of interest rates, then trees deliver CER values that are 0.2%–0.7% greater than those obtained using NNs. However, when the investor also considers information from macro and financial variables, then NNs outperform trees by 0.6% (power utility and 10-year bond) to 1% (multivariate setting). A pairwise test confirms that this improvement of NNs over trees is statistically significant. Furthermore, the results in Table 5 also show that (for the univariate and multivariate allocation, independent of utility) the group-ensemble NN that exploits macroeconomic information produces significantly higher CER values than those implied by the best-performing NN using yields-only. Overall, our machine learning-based forecasts of bond returns provide support for the hypothesis that a (statistically and economically) significant portion of macroeconomic information is not captured by the yield curve, even after accounting for nonlinearity in interest rates.[20] Internet Appendix Table D.1 shows the investment

[20] The p-values based on power utility (panel B) are lower than those reported for mean-variance (panel A). This is because the power utility setting generates less volatile CER series. To address the higher persistence of power utility CER, we compute Newey-West standard errors using a larger truncation parameter equal to 20 lags (see Lazarus et al., 2018). Even in this case, we continue to find statistical support for our conclusions.

performance when we change assumptions about the portfolio weights. In the first scenario (panel A), we restrict the weights on the risky bonds to the interval [0, 0.99] to ensure that the expected utility is finite even with an unbounded return distribution (e.g., Geweke, 2001; Kandel and Stambaugh, 1996). In the second scenario (panel B), we leave the portfolio weights unrestricted and instead restrict the bond returns to fall between −100% and 100% to prevent the expected utility from becoming unbounded (e.g., Johannes et al., 2014). In both cases, our conclusions continue to hold: the CER values of the tree and NN models are generally significantly higher than those generated by the EH benchmark; moreover, NNs outperform trees provided that macroeconomic variables are included in the set of predictors. In summary, we find that a NN that exploits the nonlinearities within groups of macroeconomic variables delivers high predictive accuracy (see Table 2), which, in turn, translates into investment strategies with large economic value (see Table 5).

6. Economic Drivers of Bond Return Predictability and Portfolio Performance

In this section, we investigate whether our forecasts of excess bond returns are consistent with explanations based on time-varying risk premiums. We then examine the economic drivers of bond return predictability and portfolio performance.

6.1 Cyclical pattern of expected excess bond returns

We start by investigating the cyclicality of our forecasts of excess bond returns. Indeed, standard asset pricing models featuring habit persistence like Wachter (2006) suggest that bond risk premiums are countercyclical. Panels (a) and (b) of Figure 6 plot the forecast of 10-year bond returns obtained from the best-performing NNs against the industrial production index growth. The results are similar for alternative maturities. We report the prediction based on yields only (panel (a)), as well as the prediction obtained by adding macroeconomic variables to forward rates (panel (b)). In panels (c) and (d), we overlay our forecasts with the realized 10-year excess bond returns series.[21] Independent of the set of predictors employed, panels (a) and (b) of Figure 6 reveal that the bond risk premium obtained from NNs displays a clear countercyclical pattern. In particular, the contemporaneous correlations between forecasts of the 10-year excess bond returns and

[21] Relative to yields-only, the addition of macroeconomic variables leads to (1) NN forecasts that are higher in the recession of 2007–2009 and (2) better predictive performance. In Internet Appendix C.1, and, in particular, panels (b) and (d) of Internet Appendix Figure C.1, we examine the predictive accuracy of NNs throughout our sample period.

industrial production is −12.4% (p-value of .07) when only information in the term structure is used (panel (a)). This correlation almost doubles to −24.6% (p-value of .01) when we add macroeconomic variables to forward rates (panel (b)).[22] Thus, using macroeconomic variables greatly improves the estimates of the risk premium. This prima facie evidence suggests that our forecasts may be consistent with the fact that investors want to be compensated for bearing recession-related risks. To the extent that our forecasts of excess bond returns reflect time-varying risk premiums, we would also expect higher Sharpe ratios in recessions. To this end, Table 6 reports, for the 2- and 10-year bond maturities, the Sharpe ratios computed separately for recession and expansion periods. We find that, across all maturities and forecasting models, the Sharpe ratios are substantially higher during recessions than in expansions.

6.2 Economic drivers of expected excess bond returns

Having established that our forecasts of excess bond returns, and the associated Sharpe ratios, move countercyclically, we next investigate whether these forecasts are linked to key drivers of bond risk premiums suggested by asset pricing theory and previous evidence. In particular, we regress the forecasts of 10-year excess bond returns obtained from the best-performing NNs in Tables 1 and 2 on a set of structural risk factors that arise in equilibrium models and generate time-varying bond risk premiums. Each row in Table 7 corresponds to a different specification. Motivated by the literature on the role of disagreement in asset prices (e.g., Buraschi and Jiltsov, 2007; Dumas et al., 2009), we examine the role played by differences in beliefs for the dynamics of excess bond returns. Row 1 in Table 7 presents the results. We proxy for real disagreement (DiB(g)) and nominal disagreement (DiB(π)) using the interquartile range of four-quarter-ahead forecasts of gross domestic product (GDP) and consumer prices (CPI), respectively, obtained from the Survey of Professional Forecasters (SPF). We investigate the link between time-varying risk aversion and excess bond returns in rows 2 and 3 of Table 7. Asset pricing models featuring habit persistence suggest that risk premiums should be higher during recessions due to a reduced surplus consumption ratio. Following Wachter (2006), we proxy for risk aversion using (the negative of) a weighted average of 10 years of quarterly consumption growth rates (dubbed −Surplus). We also employ the new measure of time-varying risk aversion proposed by Bekaert et al. (2019) (dubbed RAbex). This risk aversion measure is calculated from observable financial information at high frequencies. We next examine the role played by economic growth and inflation uncertainty, UnC(g)

[22] The correlation p-values are computed using Newey and West (1987) standard errors with 12 lags.

and UnC(π), for expected bond returns in row 4. This link can be motivated by long-run risk models like Bansal and Shaliastovich (2013) or by habit-models that allow for time variation in quantities of risk like Creal and Wu (2018).[23] Finally, we examine the link between bond volatility and our forecasts of excess bond returns in row 5 of Table 7. To assess this link, we employ two proxies: (1) the intramonth sum of squared yield changes (returns) on a constant maturity 10-year zero-coupon bond (denoted as $\sigma(n)$); and (2) the 1-month implied 10-year maturity bond risk-neutral volatility published by the CME (denoted as $TYVIX$).[24] Several conclusions emerge from the results in Table 7. First, the link between structural risk factors and realized returns is generally weak. The sole factor that is statistically linked to realized bond returns is the risk neutral volatility (panel A, row 5). Our forecasts of excess bond returns paint a completely different picture. Independent of the set of predictors we use, we find a strongly positive coefficient on uncertainty about economic growth, but not on inflation uncertainty (panels B and C, row 4). We also find strong support for the prediction of equilibrium models based on habit preferences (panels B and C, row 2). Adding macroeconomic information strengthens this conclusion: in this case, the slope coefficient on the risk aversion measure proposed by Bekaert et al. (2019) is also positive and statistically significant (panel C, row 3). Finally, in line with Duffee (2002), we find only a weak link between expected bond returns and bond volatility (panels B and C, row 5; the link is marginally significant in panel B but the R^2 is small). There are minor differences between panels B and C. In particular, the addition of macroeconomic variables leads to a positive and statistically significant slope coefficient on nominal disagreement (panel C, row 1). However, in a horse race only (habit) risk aversion and macroeconomic uncertainty continue to stay significant, leading to a large R^2 of about 25%. Instead, (nominal) disagreement is driven out (panel C, row vi). This is also the case in panel B. Comparing row 6 in panels B and C to panel A, it is apparent that using the measure of excess bond returns implied by NNs instead of realized returns leads to stronger support of the predictions of equilibrium models.[25] In Internet Appendix D.3, we discuss the relation between the realized utility obtained in the portfolio

[23] To proxy for uncertainty, we adapt the procedure of Bansal and Shaliastovich (2013). In the first step, we use our SPF on consensus expectation of four-quarter GDP growth and inflation and fit a bivariate VAR(1). In a second step, we compute a GARCH(1,1) process on the VAR residuals to estimate the conditional variance of expected real growth and inflation.

[24] `http://www.cboe.com/products/vix-index-volatility/volatility-on-interest-rates/cboe-cbot-10-year-u-s-treasury-note-volatility-index-tyvix`. Accessed February 10, 2020.

[25] A saturated regression that includes all variables simultaneously leads to the same conclusion: only habit-based risk aversion and macroeconomic uncertainty remain

analysis in Section 5 and the same structural risk factors presented in this section. The results in Internet Appendix Tables D.2 and D.3 show that the relation between utility gains from our portfolio analysis is the strongest with risk aversion and time-varying uncertainty. The evidence in Table 7 for bond returns forecasts, and that in Internet Appendix Tables D.2 and D.3 for realized utility, confirm that the variation in expected bond returns implied by a NNs can be understood in terms of time variation in risk prices and time-varying (macroeconomic) risk. Overall, our results support models that feature both channels, such as Bekaert et al. (2009) and Creal and Wu (2018). Our evidence stands in stark contrast to the recent finding of Buraschi et al. (2019). They find that the quantity of risk as measured by bond volatility has a strong role, whereas habit-based risk aversion matters little. Thus, our statistical measure of bond risk premiums likely captures a potentially different channel component from the subjective bond risk premiums of Buraschi et al. (2019). We investigate this point further in the next subsection.

6.3 Statistical versus subjective forecasts

Table 8 reports the correlations between our forecasts of 10-year excess bond returns obtained from the best-performing tree and NNs in Tables 1 and 2 with three recent proxies for risk premiums that rely on interest rates forecasts as surveyed by the Blue Chip Financial Forecasts (BCFF): (1) the measure of subjective bond risk premiums (EBR^*) proposed by Buraschi et al. (2019) based on aggregation of expectations of (the top decile of) professional forecasters; (2) the Piazzesi et al. (2015) consensus measure of subjective bond risk premiums constructed as the difference between subjective and VAR interest-rate expectations, $E_t^*\left[i_{t+h}^{(n-h)}\right] - E_t\left[i_{t+h}^{(n-h)}\right]$; and (3) the forecasts by Giacoletti et al. (2016) based on a learning rule that updates beliefs using the history of bond yields and disagreement among forecasters.[26] The results in Table 8 show that, across all bond maturities and model specifications, the correlation between our forecasts and the subjective bond risk premiums of Buraschi et al. (2019) is small, and not statistically significant. This is in line with our previous analysis of economic drivers of bond return predictability (Table 7): our forecasts are associated with proxies for time-varying risk

significant. The R^2 from the saturated regressions are 12%, 25.51%, and 25.60%, adding little explanatory power to the specification (vi) in panels A, B, and C.

[26] We measure $E_t^*\left[i_{t+h}^{(n-h)}\right]$ using the median survey forecast of $i_{t+h}^{(n-h)}$ for the 10-year Treasury bond from the Survey of Professional Forecasters. Importantly, Piazzesi et al. (2015) find that "median forecasts from the SPF are similar to those from the Bluechip survey." The statistical forecasts follow Piazzesi et al. (2015): we compute the forecasts by directly running the OLS regression on the system $Y_{t+h} = \mu + \phi Y_t + \varepsilon_{t+h}$, so that we can compute the h-horizon forecast simply as $\mu + \phi Y_t$. The vector of interest rates Y includes the 1-, 2-, 3-, 4-, 5-, 7-, and 10-year maturities.

aversion and macroeconomic uncertainty, whereas Buraschi et al. (2019) find a strong link between the quantity of risk channel (as proxied by bond volatility) and their proxy for bond risk premiums. We instead find a strong and positive association between our forecasts and the Piazzesi et al. (2015) measure of bond risk premiums, in particular when yields-only based forecasts are considered. This is perhaps not surprising given the evidence in Buraschi et al. (2019): the consensus is not a sufficient statistic for the cross-section of expectations so that aggregation of subjective bond risk premiums for each single contributor (as in EBR^*) may differ from measures that rely on the simple arithmetic average of the cross-section of forecasters (as in Piazzesi et al., 2015). However, we note that adding macroeconomic variables weakens the correlation between our forecasts and subjective measures based on consensus. Finally, the correlation between our forecasts and those of Giacoletti et al. (2016) are quite large and mostly significant. This effect is generally stronger for long maturity bonds and when yields-only based forecasts are considered. Overall, dynamic learning effects could account for some of our findings of bond return predictability.

7. Conclusion

In this paper, we evaluate the benefits of using machine learning methods for understanding bond price fluctuations. Three main findings emerge from our analysis. First, we show that nonlinear machine learning techniques, such as extreme trees and neural networks, detect predictable variations in bond returns that are statistically large; importantly, the forecasts implied by these methods translate into similarly large out-of-sample economic gains. Second, we document that employing the NN forecasts based on macroeconomic and yield information produces significantly higher certainty equivalent return values than those implied by the NN forecasts based on yields-only variables, thus providing support for information that is unspanned by (potentially nonlinear transformations of) the yield curve and yet useful to forecast bond returns. We also provide evidence of a significant heterogeneity in the relative importance of macroeconomic variables across bond maturities. Hence, the type and nature of unspanned factors may depend on the bond maturity. Finally, we document that NN forecasts are countercyclical and mostly related to variables that proxy for macroeconomic uncertainty and time-varying risk aversion. Our results provide support for models that feature both time variation in risk prices and in time-varying risk as in, for example, Bekaert et al. (2009) and Creal and Wu (2018). However, our statistical measure of expected bond returns contrasts with recent survey-based measures like the one proposed by Buraschi et al. (2019) that is mostly related to

financial (specifically, bond) volatility. From a pure machine learning perspective in the context of asset pricing, we make three contributions. First, we find that NNs perform well even when, in the context of bond return regressions, we employ just yield-based variables (i.e., in a low-dimensional setting). This finding emphasizes that the success of NNs is largely due to their ability to capture complex nonlinearities in the data. Second, we document that nonlinearities within macroeconomic categories (output, inflation, labor market, etc.) are more important than interactions across categories. Finally, we document that a carefully chosen structure of the network (like group ensembling) may compensate for the depth of the network. Overall machine learning methods that dispense with the linearity assumption in the return-predicting function may prove useful to improve our empirical understanding of asset price movements.

References

Ahn, D., R. Dittmar, and A. Gallant. 2002. Quadratic Term Structure Models: Theory and Evidence. *Review of Financial Studies* 15:243–288.

Arlot, S., A. Celisse, et al. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys* 4:40–79.

Bai, J., and S. Ng. 2003. Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70:191–221.

Bai, J., and S. Ng. 2006. Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica* 74:1133–1150.

Bai, J., and S. Ng. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146:304–317.

Bansal, R., and I. Shaliastovich. 2013. A Long-Run Risks Explanation of Predictability Puzzles in Bond and Currency Markets. *The Review of Financial Studies* 26:1–33.

Bates, J. M., and C. W. Granger. 1969. The combination of forecasts. *Journal of the Operational Research Society* 20:451–468.

Bauer, M. D., and J. D. Hamilton. 2018. Robust Bond Risk Premia. *The Review of Financial Studies* 31:399–448.

Bauer, M. D., and G. D. Rudebusch. 2017. Resolving the Spanning Puzzle in Macro-Finance Term Structure Models. *Review of Finance* 21:511–553.

Bekaert, G., E. Engstrom, and Y. Xing. 2009. Risk, uncertainty, and asset prices. *Journal of Financial Economics* 91:59–82.

Bekaert, G., E. C. Engstrom, and N. R. Xu. 2019. The Time Variation in Risk Appetite and Uncertainty. NBER Working Papers 25673, National Bureau of Economic Research, Inc. URL `https://ideas.repec.org/p/nbr/nberwo/25673.html`.

Black, F. 1995. Interest Rates as Options. *The Journal of Finance* 50:1371–1376.

Boivin, J., and S. Ng. 2006. Are more data always better for factor analysis? *Journal of Econometrics* 132:169–194.

Breiman, L. 2001. Random forests. *Machine learning* 45:5–32.

Buraschi, A., and A. Jiltsov. 2007. Habit Formation and Macroeconomic Models of the Term Structure of Interest Rates. *The Journal of Finance* 62:3009–3063.

Buraschi, A., I. Piatti, and P. Whelan. 2019. Subjective Bond Risk Premia and Belief Aggregation. Tech. rep.

Burns, A. F., and W. C. Mitchell. 1946. *Measuring Business Cycles*. National Bureau of Economic Research, Inc. URL `https://EconPapers.repec.org/RePEc:nbr:nberbk:burn46-1`.

Campbell, J. Y., and R. J. Shiller. 1991. Yield Spreads and Interest Rate Movements: A Bird's Eye View. *Review of Economic Studies* 58:495–514.

Campbell, J. Y., and S. B. Thompson. 2007. Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies* 21:1509–1531.

Chen, H., W. Dou, and L. Kogan. 2017. Measuring the'Dark Matter'in Asset Pricing Models. *Working Paper* .

Chen, L., M. Pelger, and J. Zhu. 2019. Deep learning in asset pricing. *Working Paper* .

Cieslak, A., and P. Povala. 2015. Expected Returns in Treasury Bonds. *Review of Financial Studies* 28:2859–2901.

Clark, T. E., and K. D. West. 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics* 138:291–311.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting* 5:559–583.

Cochrane, J. H., and M. Piazzesi. 2005. Bond Risk Premia. *American Economic Review* 95:138–160.

Cooper, I., and R. Priestley. 2009. Time-Varying Risk Premiums and the Output Gap. *The Review of Financial Studies* 22:2801–2833.

Coroneo, L., D. Giannone, and M. Modugno. 2016. Unspanned Macroeconomic Factors in the Yield Curve. *Journal of Business & Economic Statistics* 34:472–485.

Creal, D. D., and J. C. Wu. 2018. Bond Risk Premia in Consumption-based Models. NBER Working Papers 22183, National Bureau of Economic Research, Inc. URL `https://ideas.repec.org/p/nbr/nberwo/22183.html`.

Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* 2:303–314.

Dai, Q., K. J. Singleton, and W. Yang. 2007. Regime Shifts in a Dynamic Term Structure Model of U.S. Treasury Bond Yields. *Review of Financial Studies* 20:1669–1706.

Dangl, T., and M. Halling. 2012. Predictive regressions with time-varying coefficients. *Journal of Financial Economics* 106:157–181.

De Mol, C., D. Giannone, and L. Reichlin. 2008. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics* 146:318–328.

Della Corte, P., L. Sarno, and D. Thornton. 2008. The expectation hypothesis of the term structure of very short-term rates: Statistical tests and economic value. *Journal of Financial Economics* 89:158–174.

Diaconis, P., and M. Shahshahani. 1984. On Nonlinear Functions of Linear Combinations. *SIAM Journal on Scientific and Statistical Computing* 5:175–191.

Diebold, F. X., and R. S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business & economic statistics* 20:134–144.

Dimopoulos, Y., P. Bourret, and S. Lek. 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters* 2:1–4.

Duffee, G. 2011a. Forecasting with the term structure: The role of no-arbitrage restrictions. Economics Working Paper Archive 576, The Johns Hopkins University,Department of Economics. URL `https://ideas.repec.org/p/jhu/papers/576.html`.

Duffee, G. 2013. *Forecasting Interest Rates*, vol. 2 of *Handbook of Economic Forecasting*, chap. 0, pp. 385–426. Elsevier.

Duffee, G. R. 2002. Term Premia and Interest Rate Forecasts in Affine Models. *The Journal of Finance* 57:405–443.

Duffee, G. R. 2011b. Information in (and not in) the Term Structure. *Review of Financial Studies* 24:2895–2934.

Dumas, B., A. Kurshev, and R. Uppal. 2009. Equilibrium Portfolio Strategies in the Presence of Sentiment Risk and Excess Volatility. *Journal of Finance* 64:579–629.

Elliott, G., and A. Timmermann. 2004. Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics* 122:47–79.

Fama, E. F., and R. R. Bliss. 1987. The information in long-maturity forward rates. *The American Economic Review* pp. 680–692.

Feldhutter, P., C. Heyerdahl-Larsen, and P. Illeditsch. 2016. Risk Premia and Volatilities in a Nonlinear Term Structure Model. *Review of Finance* 22:337–380.

Feng, G., S. Giglio, and D. Xiu. 2019a. Taming the Factor Zoo: A Test of New Factors. NBER Working Papers 25481, National Bureau of Economic Research, Inc. URL `https://ideas.repec.org/p/nbr/nberwo/25481.html`.

Feng, G., J. He, and N. G. Polson. 2018. Deep Learning for Predicting Asset Returns. *arXiv preprint arXiv:1804.09314* .

Feng, G., N. Polson, and J. Xu. 2019b. Deep Learning Alpha. Chicago Booth Research Paper 23527, Chicago Booth.

Fleming, J., C. Kirby, and B. Ostdiek. 2001. The Economic Value of Volatility Timing. *Journal of Finance* 56:329–352.

Forni, M., and L. Reichlin. 1996. Dynamic Common Factors in Large Cross-Sections. *Empirical Economics* 21:27–42.

Forni, M., and L. Reichlin. 1998. Let's Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics. *The Review of Economic Studies* 65:453–473.

Freyberger, J., A. Neuhierl, and M. Weber. 2017. Dissecting Characteristics Nonparametrically. CESifo Working Paper Series 6391, CESifo Group Munich. URL `https://ideas.repec.org/p/ces/ceswps/\_6391.html`.

Friedman, J., T. Hastie, and R. Tibshirani. 2001. *The elements of statistical learning*, vol. 1. Springer series in statistics New York, NY, USA:.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. 2018. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (November 6, 2018)* .

Gargano, A., D. Pettenuzzo, and A. Timmermann. 2019. Bond Return Predictability: Economic Value and Links to the Macroeconomy. *Management Science* 65:508–540.

Geurts, P., D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. *Machine learning* 63:3–42.

Geweke, J. 1977. *The Dynamic Factor Analysis of Economic Time Series*. D.J Aigner and A.S. Goldberger, eds. (North-Holland, Amsterdam).

Geweke, J. 2001. A note on some limitations of CRRA utility. *Economics Letters* 71:341 – 345.

Giacoletti, M., K. Laursen, and K. J. Singleton. 2016. Learning, dispersion of beliefs, and risk premiums in an arbitrage-free term structure model. *Working Paper* .

Giannone, D., M. Lenza, and G. Primiceri. 2017. Economic predictions with big data: The illusion of sparsity. *Working Paper* .

Giglio, S., and D. Xiu. 2017. Inference on Risk Premia in the Presence of Omitted Factors. NBER Working Papers 23527, National Bureau of Economic Research, Inc. URL `https://ideas.repec.org/p/nbr/nberwo/23527.html`.

Gu, S., B. T. Kelly, and D. Xiu. 2018. Empirical Asset Pricing via Machine Learning. Chicago Booth Research Paper 18-04, Chicago Booth.

Gurkaynak, R. S., B. Sack, and J. H. Wright. 2007. The U.S. Treasury yield curve: 1961 to the present. *Journal of Monetary Economics* 54:2291–2304.

Harvey, D., S. Leybourne, and P. Newbold. 1997. Testing the equality of prediction mean squared errors. *International Journal of forecasting* 13:281–291.

Heaton, J. B., N. G. Polson, and J. H. Witte. 2017. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry* 33:3–12.

Hornik, K., M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. *Neural networks* 2:359–366.

Huang, J.-Z., and Z. Shi. 2019. Determinants of Bond Risk Premia: A Machine-Learning-Based Resolution of the Spanning Controversy. Working Papers 2019-12, Penn State.

Johannes, M., A. Korteweg, and N. Polson. 2014. Sequential Learning, Predictability, and Optimal Portfolio Returns. *Journal of Finance* 69:611–644.

Joslin, S., M. Priebsch, and K. J. Singleton. 2014. Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks. *The Journal of Finance* 69:1197–1233.

Kandel, S., and R. F. Stambaugh. 1996. On the Predictability of Stock Returns: An Asset-Allocation Perspective. *The Journal of Finance* 51:385–424.

Kelly, B., and S. Pruitt. 2013. Market Expectations in the Cross-Section of Present Values. *The Journal of Finance* 68:1721–1756.

Kelly, B., and S. Pruitt. 2015. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics* 186:294–316.

Kelly, B., S. Pruitt, and Y. Su. 2018. Characteristics Are Covariances: A Unified Model of Risk and Return. NBER Working Papers 24540, National Bureau of Economic Research, Inc. URL `https://ideas.repec.org/p/nbr/nberwo/24540.html`.

Kolmogorov, A. K. 1957. On the Representation of Continuous Functions of Several Variables by Superposition of Continuous Functions of One Variable and Addition. *Doklady Akademii Nauk SSSR* 114:369–373.

Kozak, S., S. Nagel, and S. Santosh. 2017. Shrinking the Cross Section. NBER Working Papers 24070, National Bureau of Economic Research, Inc. URL `https://ideas.repec.org/p/nbr/nberwo/24070.html`.

Lazarus, E., D. J. Lewis, J. H. Stock, and M. W. Watson. 2018. HAR Inference: Recommendations for Practice. *Journal of Business & Economic Statistics* 36:541–559.

Le, A., and K. J. Singleton. 2013. The Structure of Risks in Equilibrium Affine Models of Bond Yields. Working paper, Stanford Business School WP.

Liu, Y., and J. C. Wu. 2019. Reconstructing the Yield Curve. NBER Working Paper 24070, University of Notre Dame.

Ludvigson, S. C., and S. Ng. 2009. Macro Factors in Bond Risk Premia. *Review of Financial Studies* 22:5027–5067.

McCracken, M. W., and S. Ng. 2015. FRED-MD: A Monthly Database for Macroeconomic Research. Working Papers 2015-12, Federal Reserve Bank of St. Louis. URL `https://ideas.repec.org/p/fip/fedlwp/2015-012.html`.

McCracken, M. W., and S. Ng. 2016. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics* 34:574–589.

Messmer, M. 2017. Deep Learning and the Cross-Section of Expected Returns .

Mullainathan, S., and J. Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31:87–106.

Newbold, P., and C. W. Granger. 1974. Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Statistical Society: Series A (General)* 137:131–146.

Newey, W. K., and K. D. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55:703–708.

Piazzesi, M., J. Salomao, and M. Schneider. 2015. Trend and cycle in bond premia. Tech. rep.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies* 23:821–862.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2013. International Stock Return Predictability: What Is the Role of the United States? *Journal of Finance* 68:1633–1662.

Rossi, A. 2018. Predicting Stock Market Returns with Machine Learning. Tech. rep., Working paper.

Sargent, T., and C. Sims. 1977. Business cycle modeling without pretending to have too much a priori economic theory. Working Papers 55, Federal Reserve Bank of Minneapolis. URL `https://EconPapers.repec.org/RePEc:fip:fedmwp:55`.

Sarno, L., P. Schneider, and C. Wagner. 2016. The economic value of predicting bond risk premia. *Journal of Empirical Finance* 37:247–267.

Schölkopf, B., A. Smola, and K.-R. Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10:1299–1319.

Sirignano, J. A., A. Sadhwani, and K. Giesecke. 2018. Deep Learning for Mortgage Risk. Economics working paper archive, Stanford Working Paper.

Stock, J. H., and M. Watson. 2006. Forecasting with Many Predictors. 1st ed., chap. 10, pp. 515–554. Elsevier.

Stock, J. H., and M. W. Watson. 1998. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. Tech. rep., National Bureau of Economic Research.

Stock, J. H., and M. W. Watson. 2002a. Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association* 97:1167–1179.

Stock, J. H., and M. W. Watson. 2002b. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics* 20:147–162.

Thornton, D. L., and G. Valente. 2012. Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *The Review of Financial Studies* 25:3141–3168.

Wachter, J. A. 2006. A consumption-based model of the term structure of interest rates. *Journal of Financial Economics* 79:365–399.

Wu, J. C., and F. D. Xia. 2016. Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound. *Journal of Money, Credit and Banking* 48:253–291.

Yang, W. 2011. Long-run risk in durable consumption. *Journal of Financial Economics* 102:45–61.

Table 1: **Forecasting annual holding-period returns with forward rates**

Models	R^2_{oos}						p-value						$R2_{oos}$ EW	p-value EW
	$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(10)}_{t+1}$	$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(10)}_{t+1}$	$xr^{(EW)}_{t+1}$	$xr^{(EW)}_{t+1}$
A. PCA and PLS														
PCA (10 components)	-53.0%	-36.2%	-27.5%	-20.3%	-12.4%	-0.7%							-17.7%	
PCA (5 components)	-54.9%	-38.8%	-30.4%	-20.3%	-15.3%	-3.1%							-20.0%	
PCA (3 components)	-17.6%	-10.2%	-5.3%	1.0%	2.9%	10.2%				0.052	0.028	0.008	1.0%	0.036
PCA-squared (5 components)	-55.0%	-42.1%	-32.5%	-22.6%	-16.4%	-5.5%							-22.4%	
PCA-squared (3 components)	-46.8%	-38.8%	-30.6%	-21.0%	-17.0%	-8.9%							-22.5%	
Partial least squares (5 components)	-56.3%	-40.5%	-33.4%	-25.7%	-18.6%	-9.5%							-24.8%	
Partial least squares (3 components)	-57.9%	-40.7%	-31.9%	-22.9%	-14.6%	-1.8%							-20.2%	
B. Penalized linear regressions														
Ridge	-40.1%	-25.8%	-18.9%	-11.4%	-6.0%	4.6%						0.011	-9.9%	
Lasso	-11.5%	-8.1%	-2.5%	0.2%	2.3%	9.3%				0.072	0.044	0.010	2.3%	0.041
Elastic net	-10.7%	-8.4%	-5.1%	0.4%	0.3%	7.0%				0.064	0.060	0.018	0.7%	0.052
C. Regression trees and neural networks														
Gradient boosted tree	4.9%	4.9%	7.0%	10.8%	7.1%	11.2%	0.002	0.007	0.004	0.001	0.003	0.000	8.9%	0.003
Random forest	12.2%	13.1%	15.2%	17.4%	14.9%	16.0%	0.000	0.000	0.000	0.000	0.000	0.000	15.9%	0.000
Extreme tree	3.9%	10.1%	14.3%	16.9%	19.7%	24.6%	0.014	0.009	0.004	0.001	0.001	0.000	17.9%	0.001
NN - 1 layer (3 nodes)	12.7%	16.4%	19.5%	21.6%	23.1%	26.4%	0.028	0.014	0.007	0.003	0.002	0.001	23.0%	0.002
NN - 1 layer (5 nodes)	7.0%	10.7%	14.6%	17.2%	18.9%	22.9%	0.023	0.013	0.008	0.004	0.003	0.002	18.2%	0.003
NN - 1 layer (7 nodes)	-3.2%	3.0%	7.1%	12.1%	13.6%	17.1%		0.026	0.009	0.003	0.004	0.004	11.5%	0.004
NN - 2 layer (3 nodes each)	7.4%	10.8%	12.9%	14.7%	16.3%	19.3%	0.053	0.031	0.019	0.012	0.008	0.004	16.4%	0.009
NN - 2 layer (5 nodes each)	8.9%	13.0%	16.0%	18.5%	20.4%	23.6%	0.027	0.013	0.007	0.004	0.003	0.002	20.2%	0.003
NN - 2 layer (7 nodes each)	9.0%	15.2%	18.1%	21.2%	23.7%	27.4%	0.013	0.004	0.002	0.001	0.001	0.001	23.1%	0.001
NN - 3 layer (3 nodes each)	3.4%	8.9%	10.9%	11.2%	12.7%	15.1%	0.095	0.040	0.026	0.017	0.012	0.006	12.9%	0.013
NN - 3 layer (5 nodes each)	3.6%	7.4%	8.9%	10.6%	11.6%	13.7%	0.118	0.070	0.057	0.037	0.028	0.018	11.7%	0.031
NN - 3 layer (7 nodes each)	3.9%	6.7%	8.3%	8.9%	11.1%	13.2%	0.088	0.061	0.043	0.031	0.023	0.014	10.8%	0.025
NN - 3 layer (5,4,3 nodes each)	-0.9%	2.8%	5.6%	6.5%	8.3%	10.6%		0.139	0.079	0.061	0.040	0.020	8.0%	0.044
NN - 1 layer (7 nodes), lagged inputs: $t-1:t-11$	6.4%	14.4%	17.0%	19.5%	20.3%	24.4%	0.028	0.009	0.006	0.003	0.001	0.001	20.6%	0.002

This table reports the out-of-sample R^2_{oos} obtained using forward rates to predict annual excess bond returns for different maturities and across methodologies. To compute the out-of-sample R^2_{oos} we compare the forecasts obtained from each methodology to the expectation hypothesis (i.e., to the prediction based on the historical mean). In addition to the R^2_{oos} we report the p-value for the null hypothesis $R^2_{oos} \leq 0$ calculated as in Clark and West (2007). Notice that we report a p-value only when the R^2_{oos} is positive. The out-of-sample prediction errors are obtained by a recursive forecast that starts in January 1990. The sample period is from 1971:08-2018:12.

Table 2: **Forecasting annual holding-period returns with forward rates and macroeconomic variables**

Models	R^2_{oos}						p-value						$R2_{oos}$ EW	p-value EW
	$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(10)}_{t+1}$	$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(10)}_{t+1}$	$xr^{(EW)}_{t+1}$	$xr^{(EW)}_{t+1}$
A. PCA and PLS														
PCA - first 8 PC'S	-9.8%	-2.9%	0.3%	3.0%	3.3%	4.5%			0.004	0.004	0.003	0.002	1.8%	0.003
PCA as in Ludvigson and Ng (2009)	-3.4%	0.2%	1.6%	1.6%	-1.4%	-4.7%		0.007	0.006	0.007			-1.3%	
PLS - 8 components	-40.7%	-19.7%	-12.0%	-8.2%	-2.7%	3.4%						0.000	-6.4%	
B. Penalized linear regressions														
Ridge (using CP factor)	-45.3%	-23.6%	-16.7%	-13.2%	-3.1%	5.3%						0.000	-5.6%	
Lasso (using CP factor)	6.4%	11.2%	12.9%	14.4%	19.6%	23.7%	0.008	0.004	0.002	0.002	0.001	0.000 21.0%	0.001	
Elastic net (using CP factor)	6.4%	11.0%	14.3% 15.7%	21.7%	29.1%	0.007	0.004	0.002	0.001	0.001	0.000	22.0%	0.001	
Ridge (using fwd rates directly)	-52.2%	-28.7%	-22.7%	-18.3%	-13.1%	-3.5%							-15.4%	
Lasso (using fwd rates directly)	11.0%	12.0%	12.3%	16.4%	19.9%	23.6%	0.003	0.003	0.002	0.006	0.007	0.004	20.7%	0.004
Elastic net (using fwd rates directly)	10.2%	14.2%	16.0%	13.2%	19.9%	23.6%	0.005	0.003	0.003	0.003	0.002	0.002	21.0%	0.002
C. Regression trees and neural networks														
Gradient boosted tree	13.1%	15.9%	18.1%	23.8% 22.5%	25.5%	0.004	0.006	0.005	0.003	0.004	0.001	26.2%	0.002	
Random forest	26.7%	21.5%	22.0%	24.4%	20.0%	25.0%	0.003	0.003	0.002	0.001	0.004	0.002	26.6%	0.002
Extreme tree	23.0%	23.4%	22.3%	23.7%	29.9%	29.6%	0.004	0.005	0.005	0.004	0.001	0.003	29.2%	0.002
NN 1 layer (32 nodes), fwd rates direct	6.0%	13.6%	17.8%	22.0% 22.5%	26.5%	0.003	0.001	0.000	0.000	0.000	0.000	22.4%	0.000	
NN 2 layer (32, 16 nodes), fwd rates direct	16.6%	21.5%	24.9%	28.0%	29.4%	32.3%	0.001	0.000	0.000	0.000	0.000	0.000	29.3%	0.000
NN 3 layer (32, 16, 8 nodes), fwd rates direct	24.8%	26.3%	29.7%	32.0%	31.7%	33.7%	0.000	0.000	0.000	0.000	0.000	0.000	32.5%	0.000
NN 1 layer (32 nodes), fwd rates net (1 layer: 3 nodes)	8.4%	19.0%	23.8%	25.6%	27.2%	29.5%	0.003	0.001	0.001	0.001	0.001	0.000	27.2%	0.001
NN 2 layer (32,16, nodes), fwd rates net (1 layer: 3 nodes)	12.1%	15.7%	20.0%	23.5%	25.4%	28.1%	0.008	0.004	0.002	0.001	0.001	0.001 25.1%	0.001	
NN 3 layer (32,16, 8 nodes), fwd rates net (1 layer: 3 nodes)	7.6%	16.3%	20.2%	23.7%	25.0%	28.1%	0.014	0.005	0.004	0.002	0.002	0.001 25.0%	0.002	
NN 1 layer group ensem (1 node per group), fwd rates direct	12.6%	17.3%	21.6%	24.2%	25.9%	29.6%	0.002	0.001	0.001	0.000	0.000	0.000	25.9%	0.000
NN 1 layer group ensem (1 node per group), fwd rates net (1 layer: 3 nodes)	20.0%	25.6%	29.5%	31.2%	33.6%	36.3%	0.002	0.001	0.000	0.000	0.000	0.000	34.0%	0.000
NN 2 layer group ensem (2,1 nodes per group / hidden layer), fwd rates net (2 layer: 3 nodes)	17.3%	23.6%	27.8%	29.8%	31.0%	33.0%	0.006	0.001	0.001	0.000	0.000	0.000	31.6%	0.000
NN 3 layer group ensem (3, 2, 1 nodes per group / hidden layer), fwd rates net (3 layer: 3 nodes)	13.6%	20.0%	23.7%	26.1%	27.5%	30.7%	0.011	0.005	0.003	0.002	0.002	0.001	27.9%	0.002

This table reports the out-of-sample R^2_{oos} obtained using forward rates and a large panel of macroeconomic variables to predict annual excess bond returns for different maturities. To compute the out-of-sample R^2_{oos} we compare the forecasts obtained from each methodology to the expectation hypothesis (i.e., prediction based on the historical mean). In addition to the R^2_{oos} we report the p-value for the null hypothesis $R^2_{oos} \leq 0$ calculated as in Clark and West (2007). Notice that we report a p-value only when the R^2_{oos} is positive. The out-of-sample prediction errors are obtained by a recursive forecast that starts in January 1990. The sample period is from 1971:08-2018:12. Penalized regressions are estimated including macro-economic variables plus either raw forward rates or a linear combination of forward rates as introduced by Cochrane and Piazzesi (2005) (CP). Similarly, neural networks are either estimated adding the CP factor as an additional regressor in the output layer ("fwd rates direct") or estimated using a separate network for forward rates and ensembling both macro and forward rates networks in the output layer ("fwd rates net").

Table 3
Forecasting performances in expansions and recessions

	Forward rates		Fwd rates + Macro	
	R^2_{oos}		R^2_{oos}	
	Exp	Rec	Exp	Rec
PCA (10-year maturity)	7.69	**34.88**	-1.08	-41.19
PCA (2-year maturity)	-15.27%	-30.04%	-8.30%	**22.98**
Extreme tree (10-year maturity)	**26.80**	2.85%	**33.38**	-8.74
Extreme tree (2-year maturity)	**7.19**	-14.18	**25.08**	11.63%
Neural net (10-year maturity)	**26.28**	**27.33**	**35.70**	**42.54**
Neural net (2-year maturity)	**9.42**	**30.31**	**17.34**	**34.23**

This table reports the out-of-sample performances, measured by R^2_{oos} (in percentage), separately for expansions (Exp) and recessions (Rec) as defined by the NBER recession index. For ease of exposition, we report the results for the principal component regression with three PCs and for the best-performing nonlinear methodologies, that is extreme trees and the *NN 1 layer (3 nodes)* – when forecasting with only the forward rates – and *NN 1 layer group ensem + fwd rate net* , when including also macroeconomic variables (see Tables 1 and 2 for reference). We focus on the prediction exercise with two- and ten-year maturity bonds. Values in boldface indicate that the predictive accuracy of a given model is better than that obtained from the EH benchmark at the 5% level (p-value calculated as in Clark and West (2007)). The out-of-sample predictions are obtained by a recursive forecast that starts in January 1990. The sample period is from 1971:08 to 2018:12.

Table 4
Ex post diagnostics based on principal components forecasts

	Level	Slope	Curvature
PCA	9.28	21.66	48.70
Neural net (fwd rates only)	36.67	22.05	70.52
Neural net (fwd rates + macro)	30.98	30.91	65.43

This table reports the in-sample R^2 (in percentage) of a predictive regression where the dependent variable is the year-on-year change in the first three principal components extracted from the term structure of interest rates. The first row reports results when the independent variables are the lagged first three principal components. The second and third rows display the in-sample R^2 when, in addition to the first three principal components, we include the factors extracted from the best-performing neural networks obtained using either forward rates only or forward rates plus a large set of macroeconomic variables (see Tables 1 and 2 for reference).

Table 5: **Economic significance of bond predictability**

A. Mean-variance utility

	2-year maturity			10-year maturity			All		
	Fwd rates	Fwd + Macro	Δ	Fwd rates	Fwd + Macro	Δ	Fwd rates	Fwd + Macro	Δ
Neural net	-0.044	-0.002	0.042	**2.622**	**4.194**	**1.571**	**3.555**	**5.015**	**1.461**
p-value	(0.552)	(0.981)	(0.422)	(0.002)	(0.000)	(0.000)	(0.017)	(0.050)	(0.000)
Extreme tree	-0.151	-0.022	0.128	**2.881**	**2.955**	0.074	3.266	**3.961**	0.695
p-value	(0.464)	(0.675)	(0.355)	(0.001)	(0.000)	(0.864)	(0.078)	(0.078)	(0.112)
Δ	0.106	0.020		-0.258	**1.239**		0.289	**1.054**	
p-value	(0.421)	(0.680)		(0.510)	(0.001)		(0.722)	(0.024)	

B. Power utility

	2-year maturity			10-year maturity			All		
	Fwd rates	Fwd + Macro	Δ	Fwd rates	Fwd + Macro	Δ	Fwd rates	Fwd + Macro	Δ
Neural net	**0.056**	**0.111**	**0.054**	**2.714**	**3.077**	**0.363**	**3.152**	**4.829**	**1.678**
p-value	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.020)	(0.000)	(0.000)	(0.000)
Extreme tree	0.022	**0.092**	**0.072**	**3.301**	**2.511**	**-0.795**	**3.831**	**3.943**	0.113
p-value	(0.486)	(0.002)	(0.000)	(0.000)	(0.000)	(0.021)	(0.000)	(0.000)	(0.835)
Δ	0.034	0.017		**-0.591**.	**0.567**		**-0.680**	**0.886**	
p-value	(0.212)	(0.267)		(0.023)	(0.024)		(0.030)	(0.029)	

This table reports the annualized certainty equivalent values (in %) for portfolio decisions based on the out-of-sample forecasts of bond excess returns for an investor with either mean-variance (panel A) or power utility (panel B) and a coefficient of risk aversion equal to five. The table reports two asset allocation exercises. In the univariate asset allocation case, the investor selects either the 2- or the 10-year bond, along with the 1-year short-rate. In the multivariate case, the investor selects bonds across the six maturities, 2 to 5, 7, and 10 years. The asset allocation decision is based on the predictions implied by either the best-performing regression tree specification,that is, extreme tree, or the best performing neural network, namely, the *NN 1 layer (3 nodes)*, when forecasting with only the forward rates, and *NN 1 layer group ensem + fwd rate net*, when also including macroeconomic variables (see Tables 1 and 2 for reference). The row Δ reports the value added by NN relative to extreme tree within each application ("Fwd rates" and "Fwd + Macro"). The column Δ reports the value added by "Fwd+Macro" relative to "Fwd rates" within each model (NN and extreme tree). The models are benchmarked against the expectation hypothesis. The out-of-sample predictions are obtained by a recursive forecast that starts in January 1990. The sample period is from 1971:08 to 2018:12. Statistical significance is based on a one-sided Diebold and Mariano (1995) test as extended by Harvey et al. (1997) to account for autocorrelation in the forecasting errors. We flag in bold those values that are statistically significant at the 5% confidence level.

Table 6
Sharpe ratios in expansions and recessions

	Forward rates		Fwd rates + Macro	
	Exp	Rec	Exp	Rec
PCA (10-year maturity)	0.087	1.364	-0.118	0.190
PCA (2-year maturity)	0.037	1.524	0.403	1.356
Extreme tree (10-year maturity)	0.261	0.521	0.491	0.555
Extreme tree (2-year maturity)	0.253	1.688	0.740	1.541
Neural net (10-year maturity)	0.506	1.769	0.749	1.707
Neural net (2-year maturity)	1.077	2.384	1.093	2.098

This table reports the out-of-sample annualized Sharpe ratio, separately for expansions (Exp) and recessions (Rec) as defined by the NBER recession index. We report the results for the benchmarking regression that employ the first three principal components of the yield curve, and for the best-performing nonlinear methodologies, that is, extreme trees and the *NN 1 layer (3 nodes)*, when forecasting with only the forward rates, and *NN 1 layer group ensem + fwd rate net*, when including also macroeconomic variables (see Tables 1 and 2 for reference). We focus on the prediction of 2- and 10-year bonds. The out-of-sample predictions are obtained by a recursive forecast that starts in January 1990. The sample period is from 1971:08 to 2018:12.

Table 7
Drivers of bond risk premiums

A. 10-year realized bond excess returns

	$DiB(g)$	$DiB(\pi)$	$-Surplus$	$RAbex$	$UnC(g)$	$UnC(\pi)$	$TYVIX$	$\sigma_B^{(n)}$	$R^2(\%)$
(i)	-0.19	0.29							6.43
	(0.41)	(0.19)							
(ii)			0.01						0.25
			(0.90)						
(iii)				0.01					0.25
				(0.81)					
(iv)					-0.14	0.28			5.82
					(0.42)	(0.14)			
(v)							**0.25**	-0.22	6.36
							(0.02)	(0.38)	
(vi)		0.27	0.04	-0.03	0.09				7.01
		(0.08)	(0.66)	(0.67)	(0.14)				

B: 10-year expected bond excess returns (fwd rates)

	$DiB(g)$	$DiB(\pi)$	$-Surplus$	$RAbex$	$UnC(g)$	$UnC(\pi)$	$TYVIX$	$\sigma_B^{(n)}$	$R^2(\%)$
(i)	0.02	0.31							3.06
	(0.96)	(0.22)							
(ii)			**0.35**						9.85
			(0.01)						
(iii)				0.15					1.58
				(0.21)					
(iv)					**0.40**	-0.11			10.34
					(0.00)	(0.72)			
(v)							-0.11	**0.74**	3.30
							(0.53)	(0.04)	
(vi)		0.01	**0.43**	0.04	**0.28**				23.69
		(0.95)	(0.00)	(0.77)	(0.00)				

C. 10-year expected bond excess returns (fwd rates + macro)

	$DiB(g)$	$DiB(\pi)$	$-Surplus$	$RAbex$	$UnC(g)$	$UnC(\pi)$	$TYVIX$	$\sigma_B^{(n)}$	$R^2(\%)$
(i)	-0.35	**0.55**							8.22
	(0.43)	(0.01)							
(ii)			**0.32**						11.80
			(0.02)						
(iii)				**0.27**					6.06
				(0.02)					
(iv)					**0.38**	-0.15			8.49
					(0.01)	(0.71)			
(v)							0.05	0.59	5.44
							(0.80)	(0.14)	
(vi)		0.19	**0.35**	0.15	0.21				25.22
		(0.35)	(0.00)	(0.14)	(0.09)				

This table reports the regression estimates of realized (panel A) and expected (panel B and C) bond excess returns on 10-year bonds on a set of structural determinants of risk premiums (see Section 6.2 for details). The expected bond return (dependent variable) is based on the predictions implied by the best-performing neural network, namely the *NN 1 layer (3 nodes)*, when forecasting with only the forward rates, and *NN 1 layer group ensem + fwd rate net*, when including also macroeconomic variables (see Tables 1 and 2 for reference). We standardize both left- and right-hand side variables, so that a 1-standard deviation change in the right hand variables implies a β-standard deviation in the dependent variable. We report the regression estimates and Newey-West p-values. Bold font indicates significance at the 5% level. The out-of-sample predictions are obtained by a recursive forecast that starts in January 1990. The sample period is from 1971:08 to 2018:12.

Table 8
Statistical versus subjective forecasts of bond risk premiums

A. Forecasting with forward rates

	10-year maturity		
	EBR^*_{10y}	SUBJ_BRP	GLS
Extreme tree	-7.5%	**49.5%**	**52.3%**
	(0.45)	(0.00)	(0.00)
NN - 1 layer (3 nodes)	3.3%	**56.3%**	**59.5%**
	(0.75)	(0.00)	(0.00)

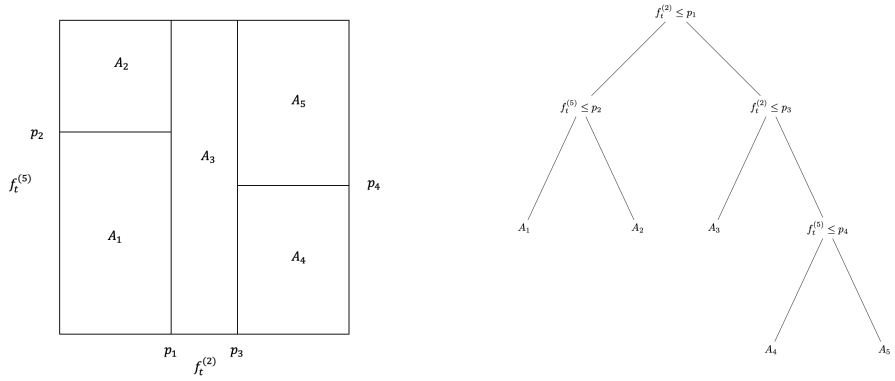
	2-year maturity		
	EBR^*_{2y}	SUBJ_BRP	GLS
Extreme tree	-18.7%	**42.7%**	**63.8%**
	(0.17)	(0.00)	(0.00)
NN - 1 layer (3 nodes)	4.1%	**53.8%**	**50.3%**
	(0.78)	(0.00)	(0.00)

B. Forecasting with forward rates and macro variables

	10-year maturity		
	EBR^*_{10y}	SUBJ_BRP	GLS
Extreme tree	-1.6%	**40.9%**	**48.3%**
	(0.88)	(0.00)	(0.00)
NN - 1 layer group ensem + fwd rate net	13.4%	**38.0%**	**47.2%**
	(0.26)	(0.00)	(0.00)

	2-year maturity		
	EBR^*_{2y}	SUBJ_BRP	GLS
Extreme tree	3.5%	21.3%	20.0%
	(0.80)	(0.13)	(0.16)
NN - 1 layer group ensem + fwd rate net	6.4%	**27.5%**	**30.3%**
	(0.61)	(0.03)	(0.03)

This table reports the correlation between our machine learning implied forecasts and existing measures of bond risk premiums based on subjective forecasts or on asset pricing models with learning dynamics. Panel A: shows the results for the forecasts generated using only the forward rates, whereas panel B: shows the results for the forecasts generated using both forward rates and a large panel of macroeconomic variables. Correlations are computed with respect to the subjective bond risk premiums in Buraschi et al. (2019) (EBR^*_{10y} and EBR^*_{2y}), the subjective risk premiums measure proposed by Piazzesi et al. (2015) (SUBJ_BRP), and the out-of-sample bond returns forecasts in Giacoletti et al. (2016) (GLS). In parentheses we report Newey-West p-values with 12 lags. Bold font indicates significance at the 5% level. The out-of-sample predictions are obtained by a recursive forecast that starts in January 1990. The sample period is from 1971:08 to 2018:12.



(a) Recursive binary splitting (b) Partition tree

Figure 1
Example of a regression tree
This figure shows an example of a regression tree for a predictive regression with a univariate target variable, for example, the holding period excess return of a 1-year treasury bond, and two predictors, for example, the 2-year and the 5-year forward rates, which we label $f_t^{(2)}$ and $f_t^{(5)}$. The left panel shows the partition of the two-dimensional regression space by recursive splitting. The right panel shows the corresponding regression tree.

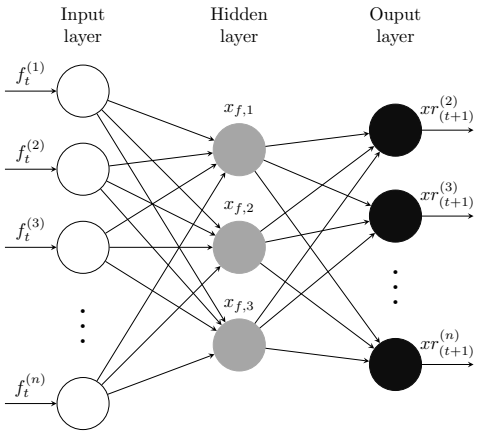
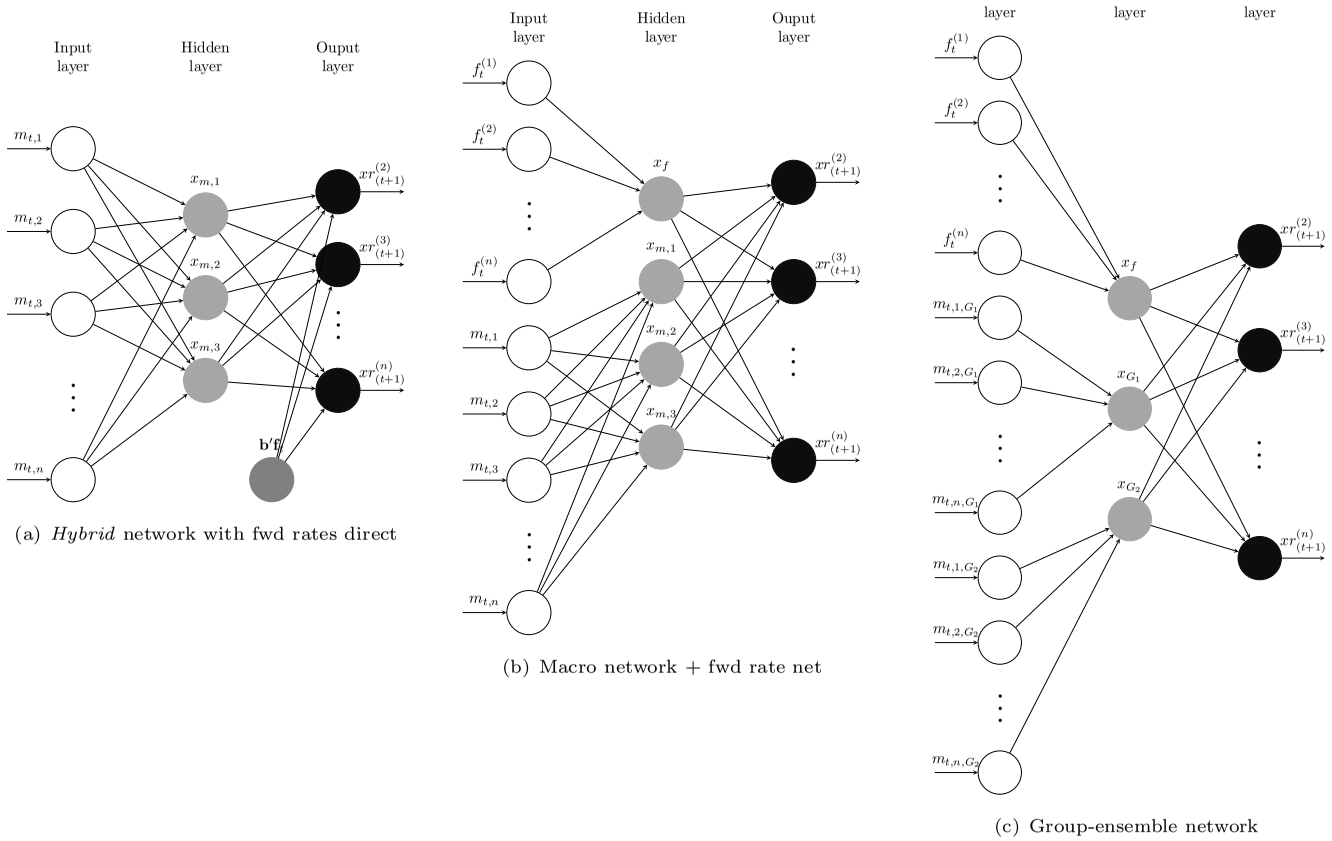


Figure 2
Examples of a neural network with only forward rates
This figure shows a neural network with one hidden layer when forecasts are based only on forward rates as in, for example, Cochrane and Piazzesi (2005).



(a) *Hybrid* network with fwd rates direct

(b) Macro network + fwd rate net

(c) Group-ensemble network

Figure 3: **Examples of neural networks with forward rates and macro variables**
This figure shows examples of the network structures used in the paper. The left panel shows the neural network with a linear combination of forward rates, $b'f_t$, that is included as an exogenous regressors (in the paper such specification is called macro + fwd rates direct). This structure simulate the idea of Ludvigson and Ng (2009) in which the latent macro factors are extracted from a large cross-section of macroeconomic variables and forward rates are included as a linear combination as proposed by Cochrane and Piazzesi (2005). The center panel displays a network structure whereby macro variables, $m_{t,i}$, and forward rates, $f_t^{(n)}$, define two separate groups (in the paper such specification is called macro + fwd rates net). The right panel shows the group-ensemble network whereby groups of macroeconomic variables, m_{t,i,G_1} and , and forward rates, $f^{(n)}$, define a separate network. The collection of networks is then ensembled at the output

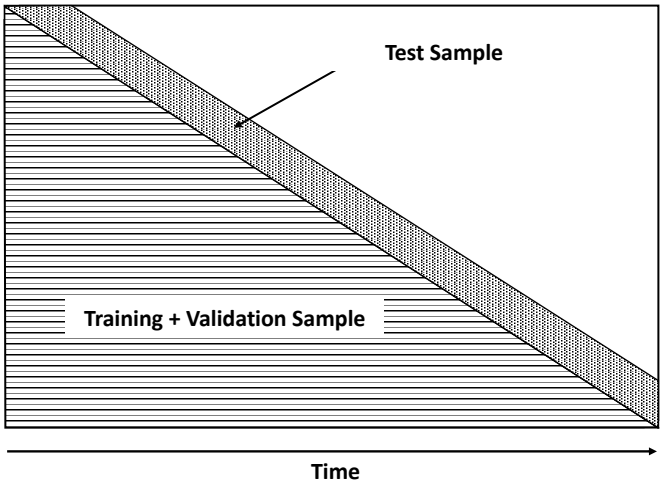
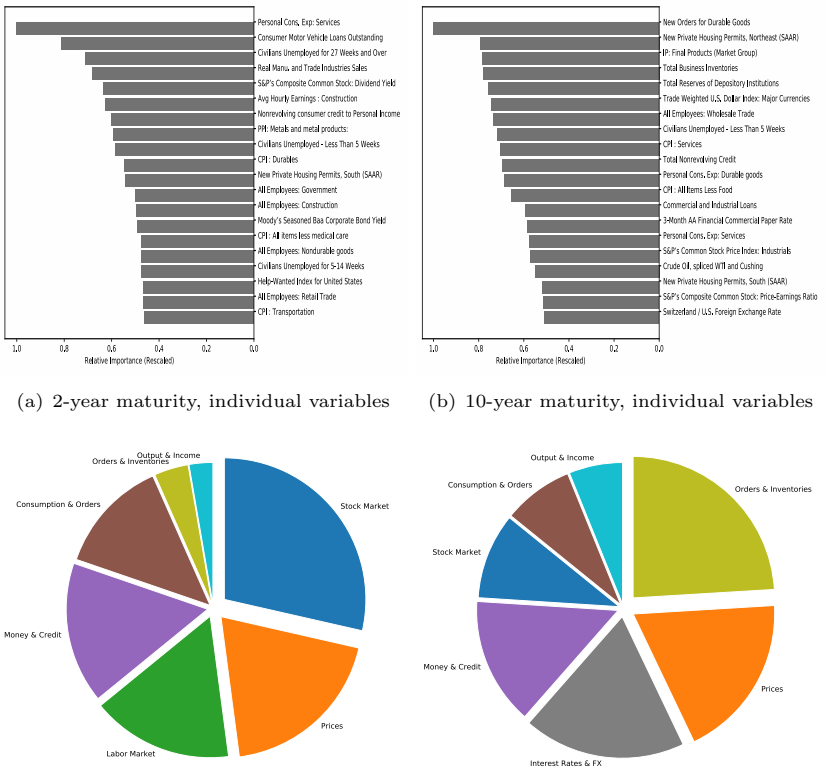


Figure 4
Sample splitting strategy
This figure shows the sample splitting used for cross-validation of the hyperparameters of the penalized regressions, that is, lasso, elastic net, ridge, and the neural networks. The forecasting exercise involves an expanding window that starts in January 1990. The full sample period is from 1971:08 to 2018:12. The training sample consists of the first 85% of the observations, and the validation sample consists of the final 15% of observations. The training and the validation samples are consequential and not randomly selected in order to preserve the time-series dependence. The testing sample consists of observations on 1-year holding period excess Treasury bond returns.



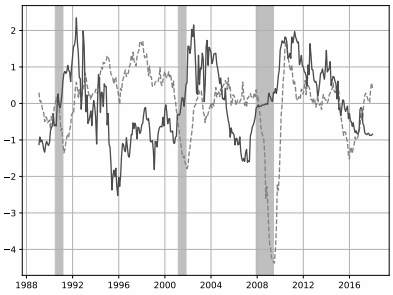
(a) 2-year maturity, individual variables

(b) 10-year maturity, individual variables

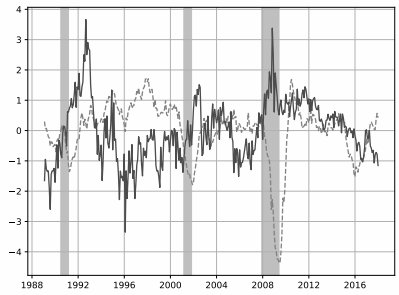
(c) 2-year maturity, groups of variables

(d) 10-year maturity, groups of variables

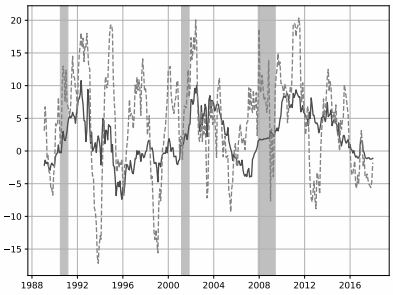
Figure 5
Relative importance of macroeconomic variables
This figure shows the relative importance of different macro economic variables use to forecast bond excess returns. Panels (a) and (b) show results for individual variables, and panels (c) and (d) present results for groups of macro variables. The groups are labeled following McCracken and Ng (2016). The relative importance of an input variable is computed based on the absolute value of the gradient of the network outputs with respect to the input variable. The gradient is evaluated at the in-sample mean of the input variable. The gradient-at-the-mean is calculated for each time t of the recursive forecasting exercise and then averaged over the out-of-sample period. For the grouped results in panels (c) and (d) the relative importance is averaged within groups.



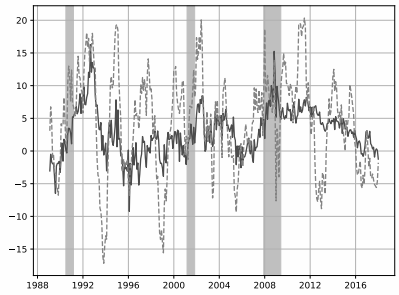
(a) Fwd only: Expected return versus IP



(b) Fwd + Macro: expected return versus IP



(c) Fwd only: Realized versus expected return



(d) Fwd + Macro: Realized versus expected return

Figure 6
Bond excess returns, model-implied risk premia, and economic growth
Panels (a) and (b) plot the model-implied expected bond excess returns for the 10-year maturity (solid lines) against the annual growth rate of industrial production (IP) in the US (dashed line). Panels (c) and (d) display the time series of annual realized (dashed line) and expected (solid line) ten-year bond excess returns (in percentage terms). We report the two best performing forecasts from the neural nets, that is the *NN 1 layer (3 nodes)*, when forecasting with only the forward rates, and *NN 1 layer group ensem + fwd rate net*, when including also macroeconomic variables (see Tables 1 and 2 for reference). The left panels report the results for the expected bond excess returns obtained by using the forward rates, whereas the right panels report the results for the expected bond excess returns obtained by using a large set of macroeconomic variables in addition to the forward rates.