Beckmann, Lars; Debener, Jörn; Kriebel, Johannes

Understanding the determinants of bond excess returns using explainable AI

Journal of Business Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Beckmann, Lars; Debener, Jörn; Kriebel, Johannes (2023) : Understanding the determinants of bond excess returns using explainable AI, Journal of Business Economics, ISSN 1861-8928, Springer, Berlin, Heidelberg, Vol. 93, Iss. 9, pp. 1553-1590, https://doi.org/10.1007/s11573-023-01149-5

Journal of Business Economics (2023) 93:1553–1590
https://doi.org/10.1007/s11573-023-01149-5

ORIGINAL PAPER

Understanding the determinants of bond excess returns using explainable AI

Lars Beckmann[1] · Jörn Debener[1] · Johannes Kriebel[1]

Accepted: 25 February 2023 / Published online: 30 May 2023

Abstract

Recent empirical evidence indicates that bond excess returns can be predicted using machine learning models. However, although the predictive power of machine learning models is intriguing, they typically lack transparency. This paper introduces the state-of-the-art explainable artificial intelligence technique SHapley Additive exPlanations (SHAP) to open the black box of these models. Our analysis identifies the key determinants that drive the predictions of bond excess returns produced by machine learning models and recognizes how these determinants relate to bond excess returns. This approach facilitates an economic interpretation of the predictions of bond excess returns made by machine learning models and contributes to a thorough understanding of the determinants of bond excess returns, which is critical for the decisions of market participants and the evaluation of economic theories.

Keywords Asset pricing · Bond excess returns · Machine learning · Explainable artificial intelligence

JEL Classification C40 · G11 · G12 · G17 · E44

We thank Wolfgang Breuer (the editor), two anonymous referees, Heiner Beckmeyer, and Andreas Pfingsten for providing us with very helpful comments and suggestions.

✉ Johannes Kriebel
johannes.kriebel@wiwi.uni-muenster.de

Lars Beckmann
lars.beckmann@wiwi.uni-muenster.de

Jörn Debener
joern.debener@wiwi.uni-muenster.de

[1] University of Münster, Münster, Germany

1 Introduction

What drives bond excess returns has been the subject of extensive academic research over the past few decades. Recent studies have documented the ability of machine learning models to substantially predict bond excess returns (Bianchi et al. 2021a, b; Huang and Shi 2023; Fan et al. 2022). However, although the predictive performance of machine learning models in this context is intriguing, their lack of transparency presents a major problem, making it unclear which determinants contribute in what way to the predictions produced. Consequently, an economic interpretation of the relationship between determinants and bond excess returns is difficult. This is a concern because, for the monetary policy of central banks and the portfolio management of investors, it is of high importance to understand why and under which economic conditions long-term bonds exhibit excess returns (Kessler and Scherer 2009; Ludvigson and Ng 2009; Bauer and Hamilton 2018).

Our work connects to two important strands in the literature. One strand examines the determinants of bond excess returns based on linear prediction models. Several studies document the usefulness of information from the yield curve for predicting bond excess returns (e.g., Fama and Bliss 1987; Campbell and Shiller 1991; Cochrane and Piazzesi 2005). Furthermore, many studies provide evidence of the additional predictive power of macroeconomic variables related to employment and production (Ludvigson and Ng 2009), inflation (Wright 2011; Joslin et al. 2014), the output gap (Cooper and Priestley 2009), and growth and real interest (Coroneo et al. 2016).

The second strand uses machine learning models to predict bond excess returns. Machine learning models deliver much stronger performances than linear prediction models in realistic out-of-sample settings (Bianchi et al. 2021a, b). This is because these models can reflect non-linear relationships between bond excess returns and their determinants and can include many more input variables. Recent important contributions to this field include Bianchi et al. (2021a, b), Huang and Shi (2023), and Fan et al. (2022). However, the black-box nature of machine learning models means that it remains unclear what determinants enable the strong predictive performance of these models and how those determinants relate to bond excess returns.

To better understand the determinants of bond excess returns, we apply a three-step empirical approach based on explainable artificial intelligence. In particular, we use SHapley Additive exPlanations (SHAP) to open the black box of machine learning models with a strong performance in the prediction of bond excess returns. SHAP is a state-of-the-art explainable artificial intelligence technique that uses concepts from game theory to identify the contribution of individual variables to the prediction of a machine learning model (Lundberg and Lee 2017). In the first step of our empirical approach, we predict bond excess returns across different maturities for the U.S. bond market using machine learning models in a realistic out-of-sample setting that adapts to new information every month. In the second step, we uncover the most important determinants

of U.S. bond excess returns in machine learning models. In the third step, we examine the direction in which these determinants are related to U.S. bond excess returns, information that is critical for a thorough economic understanding. We then apply this three-step approach to examine the determinants of bond excess returns in the German bond market, enabling a comparison with the determinants in the U.S. context.

Our empirical approach reveals that information from the yield curve, especially the slope of the yield curve, is a key determinant of U.S. bond excess returns. With respect to the functional relationship, a steeper slope of the yield curve predicts higher bond excess returns. Accordingly, we provide empirical evidence for consumption-based theoretical asset pricing models explaining the predictability of bond excess returns based on the slope of the yield curve (e.g., Gabaix 2012). Beyond information from the yield curve, we find that macroeconomic variables—especially variables related to the housing market—drive predictions of U.S. bond excess returns. Specifically, a weaker housing market predicts higher bond excess returns. These findings add to the relatively scarce literature on the importance of the housing market for asset prices (Piazzesi et al. 2007; Huang and Shi 2023). Turning to differences across bond markets, our study provides empirical evidence that the slope of the yield curve is also an important determinant of German bond excess returns. However, in contrast to the U.S., the local housing market does not seem to provide additional explanatory power beyond the information from the yield curve for German bond excess returns. We suspect that the German housing market poses less risk to German bonds than the U.S. housing market poses to U.S. bonds.

We contribute to the literature in three important ways. First, we contribute to a deeper understanding of the determinants of bond excess returns by examining the key determinants of bond excess returns in machine learning models and investigating the nature of the relationship those determinants have with bond excess returns. Second, we highlight differences in the determinants of bond excess returns across countries by contrasting two central bond markets, drawing attention to this scarcely studied research area. Third, from a methodological perspective, we contribute to the broader asset pricing literature by presenting an empirical approach based on explainable artificial intelligence that is suitable for opening up the black box of machine learning models to predict the returns of not only bonds but also other assets, including stocks and options.

Our findings have important implications for practitioners and academics. Investors can benefit from our results by better understanding which factors determine bond portfolio returns. Central banks can gain a better understanding of the price dynamics of long-term bonds, which play an important role in the transmission of monetary policy. For future empirical asset pricing research, our study suggests the application of explainable artificial intelligence, especially SHAP, to better understand the predictions of machine learning models.

The remainder of the paper is structured as follows. Section 2 provides an overview of the related literature and derives hypotheses. Section 3 presents the data upon which our analysis is based. Section 4 describes our methodology. Section 5 presents our results regarding the determinants of bond excess returns, and Sect. 6 concludes.

2 Literature review and hypotheses

2.1 Research on the determinants of bond excess returns

Bond excess returns have long been the subject of academic research. From a theoretical perspective, the pure expectation hypothesis posits that investors expect the same return on long-term bonds as on short-term bonds when the bonds are held for the same period (McCallum 1975; Campbell 1995). This implicitly assumes that the term structure of yields is determined entirely by expectations of future yields. According to this hypothesis, there should be no systematic difference between holding period returns on long-term bonds and short-term bonds. A weaker form of the pure expectation hypothesis proposes that investors may expect a higher return on long-term bonds than short-term bonds, but this difference is constant and does not change over time (Campbell 1995). Therefore, any deviations between these returns should be completely random and unpredictable (Cochrane and Piazzesi 2005).

While the expectation hypothesis has long been the most common theory on bond excess returns, there is substantial empirical evidence against it, at least for in-sample settings. Several studies find that bond excess returns in the U.S. and international markets can be predicted to some extent using information from the yield curve) (e.g., Fama and Bliss 1987; Campbell and Shiller 1991; Cochrane and Piazzesi 2005; Kessler and Scherer 2009; Sekkel 2011). In particular, the first three principal components of yields over different maturities—reflecting the level, the slope, and the curvature of the yield curve—and a linear combination of forward rates, commonly known as the Cochrane–Piazzesi factor, provide insight into bond excess returns (Litterman and Scheinkman 1991; Cochrane and Piazzesi 2005).

Researchers have developed different theoretical asset pricing models that can explain the predictive power of yields for bond excess returns. Among the most notable studies, Gabaix (2012) proposes a consumption-based disaster model that incorporates inflation jumps in rare consumption disasters. Because long-term bonds are more sensitive to inflation jumps, investors demand a higher risk premium for these bonds, producing a positive slope of the yield curve without a direct link to higher expected yields in the future. In this model, the size of the expected inflation jumps varies over time. When investors expect particularly large inflation jumps, the slope of the yield curve is particularly steep. Because the slope of the yield curve is assumed to be mean-reverting, a particularly steep slope of the yield curve predicts falling yields—that is, rising prices—for long-term bonds, which translates into positive bond excess returns. However, where the model introduced by Gabaix (2012) is based on consumption disasters, other studies suggest habit formation (Wachter 2006) and long-run risk models (Bansal and Shaliastovich 2013) to explain the predictive power of the yield curve.[1]

The theoretical view on predicting bond excess returns indicates that all risks relevant to bondholders—like the risk of inflation jumps in the model proposed by

[1] We refer the reader to Cochrane (2017) for an overview of different state-of-the-art theoretical asset pricing models used to explain the predictability of returns from assets such as stocks and bonds.

Gabaix (2012)—should be incorporated into current bond prices by investors. This idea is reflected in the spanning hypothesis, which posits that all information relevant for predicting bond excess returns is spanned by the yield curve (Bauer and Hamilton 2018). Consequently, any other variable potentially relevant for predicting bond excess returns should contain no predictive power beyond information from the yield curve.

Whether the spanning hypothesis holds is subject to extensive and ongoing empirical debate. Ludvigson and Ng (2009) apply dynamic factor analysis to a large number of macroeconomic indicators to investigate the spanning hypothesis for the U.S. market. They find that macroeconomic indicators substantially increase the predictability of bond excess returns and that indicators related to employment and production are most useful for predictions. This evidence against the spanning hypothesis aligns with Wright (2011) and Joslin et al. (2014), studies indicating that inflation risk is unspanned by the yield curve and can explain risk premia in the U.S. bond market and other international bond markets. Cooper and Priestley (2009) instead focus on the macroeconomic output gap, finding that it has predictive power for U.S. bond excess returns. However, Bauer and Hamilton (2018) argue in favor of the spanning hypothesis, criticizing prior methodological approaches. Furthermore, Bauer and Rudebusch (2017) provide empirical evidence in favor of the spanning hypothesis for the U.S. market.

Until recently, most studies investigating the predictability of bond excess returns and their determinants have relied on linear regressions (e.g., Cochrane and Piazzesi 2005; Ludvigson and Ng 2009; Sekkel 2011; Ioannidis and Ka 2021). Furthermore, most of these studies have focused on the in-sample predictability of returns. This is problematic for two reasons. First, in-sample analyses only consider the information available at a single point in time and are based on expost knowledge rather than the knowledge available at the time of the investment decision. Hence, in-sample analyses do not reflect a realistic decision setting. Second, the in-sample performance of predictive models usually correlates poorly with the more realistic out-of-sample performance of these models (Inoue and Kilian 2004; Thornton and Valente 2012). Recent evidence shows that in an out-of-sample setting, linear predictive regressions based on information from the yield curve (Thornton and Valente 2012; Bianchi et al. 2021a, b) and based on both information from the yield curve and macroeconomic information (Bianchi et al. 2021a, b) have almost no predictive power for U.S. bond excess returns. This is contrary to the findings described above and highlights the need for further empirical analyses of the determinants of bond excess returns.

Empirical research has addressed these methodological drawbacks most recently by using machine learning methods instead of linear regressions to predict asset returns and by applying these methods to out-of-sample rather than in-sample settings. Machine learning can be understood as an approach in which 'computer algorithms (...) infer meaningful patterns from a dataset' (Bartram et al. 2021). Applying machine learning methods to bond excess return prediction enables the use of a large set of variables and allows for non-linear relationships between these variables and returns. In this context, Bianchi et al. (2021a, b) use several machine learning methods and find that neural networks fed with yield data and macroeconomic data together can predict bond excess returns in the U.S., thereby presenting empirical

evidence against the spanning hypothesis. Huang and Shi (2023) apply regularized regressions to predict U.S. bond excess returns and also argue against the spanning hypothesis, demonstrating that some macroeconomic variables have significant predictive power and are, therefore, not spanned by the yield curve. Fan et al. (2022) use neural networks to predict U.S. bond excess returns and find that they can be predicted to a substantial extent based on macroeconomic data.

Although the predictive performance of machine learning models for bond excess returns is intriguing, a central shortcoming of these models is their lack of transparency. As such, it is difficult to identify which variables contribute in what way to the bond excess return predicted by the model. Thereby an economic interpretation of the relationship between the determinants and the bond excess return is hindered, which makes it difficult for decision makers to act based on the outcomes of machine learning models. This calls for techniques that can open the black box of these models, typically referred to as explainable artificial intelligence.

Overall, there is still an extensive academic debate about the determinants of bond excess returns. Specifically, it is unclear which variables drive bond excess returns. Furthermore, most studies have focused on a single market, typically the U.S. market. Therefore, further research is needed to understand which variables are most important for predicting bond excess returns and investigate whether these variables differ between bond markets.

2.2 Hypotheses on the determinants of bond excess returns

To guide our empirical examination, we derive hypotheses on what information is most likely to have predictive power for bond excess returns. In this regard, economic theory suggests that information from the yield curve is highly informative for future bond excess returns (Wachter 2006; Gabaix 2012; Bansal and Shaliastovich 2013). In particular, the described consumption-based disaster model by Gabaix (2012) considers the slope of the yield curve important because it reflects investor expectations of inflation jumps, meaning that a higher expected inflation jump in the case of a consumption disaster leads to a steeper slope of the yield curve. Because the slope is assumed to be mean-reverting, a particularly steep slope predicts a subsequently less steep slope, which is equivalent to increasing prices for long-term bonds, implying positive bond excess returns. Based on this theoretical reasoning—that the slope of the yield curve reflects substantial risk for future long-term bond prices—we hypothesize the following:

H1: The slope of the yield curve is one of the most important determinants of bond excess returns.

However, despite the empirical findings and the theoretical asset pricing models in favor of the high predictive power of information from the yield curve for bond excess returns, there is some empirical evidence that not all relevant macroeconomic risks are reflected in the yield curve (Ludvigson and Ng 2009; Wright 2011; Joslin et al. 2014; Cooper and Priestley 2009; Bianchi et al. 2021a, b; Huang and Shi 2023).

Furthermore, the literature has documented the importance of local risk factors for predicting bond excess returns (Barr and Priestley 2004; Pérignon et al. 2007). However, which macroeconomic determinants have explanatory power beyond information from the yield curve in different local bond markets has not been in the focus of research to date. Intuitively, there are different local macroeconomic risks that are relevant to investors in different local bond markets. For example, high inflation expectations in the Eurozone will more substantially impact European government bonds than U.S. government bonds. Furthermore, the degree to which such risks are reflected in the yield curve can differ between local bond markets. Therefore, it is likely that the macroeconomic determinants that have explanatory power beyond information from the yield curve differ between bond markets. Following this line of reasoning, we hypothesize the following:

H2: The macroeconomic determinants of bond excess returns that have explanatory power beyond information from the yield curve differ between bond markets.

The following two sections describe the data and the methodology used to investigate these two hypotheses.

3 Data

3.1 Yield data and macroeconomic data

Based on the literature described in the previous section, we use two types of information to predict bond excess returns. On the one hand, we predict bond excess returns based on the structure of yields over different maturities, as proposed by Fama and Bliss (1987), Campbell and Shiller (1991), and Cochrane and Piazzesi (2005), among others. On the other hand, we use both yield data and a large set of macroeconomic data to predict bond excess returns (Ludvigson and Ng 2009; Bianchi et al. 2021a, b; Huang and Shi 2023). Our study focuses on two important international bond markets, namely, the U.S. and the German bond market.

For the U.S., we use a monthly data set of the zero-coupon U.S.-Treasury bond yield curve constructed by Liu and Wu (2021), which is available online.[2] This data set contains monthly information on yields for maturities from 1 to 10 years. Our sample period ranges from August 1971 to December 2018. Using these data on the structure of yields, we then calculate the bond excess returns for the U.S. bond market, as illustrated in the following subsection, and forward rates. Furthermore, we conduct a principal component analysis (PCA) of the yield data to summarize the most important information from these data. In particular, we extract the first three principal components of the yield data. Earlier studies showed that the principal components of yields are highly informative and related to the level, slope, and curvature of the yield curve, respectively (e.g., Litterman and Scheinkman 1991; Bauer and Rudebusch 2017).

[2] https://sites.google.com/view/jingcynthiawu/yield-data.

For our macroeconomic data, we use a large panel of 124 monthly macroeconomic variables for the U.S. bond market. This panel, constructed by McCracken and Ng (2016), is also available online.[3] The time series in the panel were grouped to reflect the following eight categories of macroeconomic information: "Output and income" (1), "Labor market" (2), "Housing" (3), "Consumption, orders, and inventories" (4), "Money and credit" (5), "Interest and exchange rates" (6), "Prices" (7), and the "Stock market" (8). This data set has been widely used in previous studies (e.g., Stock and Watson 2002, 2006; Ludvigson and Ng 2009).

For the German bond market, we use monthly data on the yields of German government bonds provided by the Deutsche Bundesbank.[4] These data are also available online.[5] Again, we use the data to calculate bond excess returns and forward rates for the German bond market and conduct a PCA of the yield data to summarize the most important information from the yield curve. A short analysis of the correlation between the first three principal components and proxies for the level, slope, and curvature of the yield curve following Diebold and Li (2006) and Diebold et al. (2006) confirms that, as for the U.S., the first two principal components strongly relate to the level and slope of the yield curve, with the third principal component rather weakly related to the curvature.[6] For our macroeconomic data for the German bond market, we construct a large panel of 67 monthly macroeconomic variables from the Thomson Reuters Eikon database and the Federal Reserve Economic Data (FRED). These macroeconomic variables are selected to match the U.S. variables as closely as possible. As such, we have grouped them into the same eight categories of macroeconomic information previously introduced. Table 3 and Table 4 in the Appendix provide a full description of the macroeconomic variables used for each bond market.

3.2 Bond excess returns

A bond excess return is defined as the return from buying a long-term bond and selling it at a future point in time T less the return from investing in a short-term bond with maturity T and holding it until maturity. Bond excess returns are positive if the returns on long-term bonds exceed the returns on short-term bonds over this period and negative if the returns on long-term bonds are below the returns on short-term bonds.

Using the notation $p_t^{(n)}$ for the (log) price of a zero-coupon bond at time t and a remaining maturity of n, and the notation $y_t^{(n)} = -\frac{1}{n} p_t^{(n)}$ for the (continuously compounded) yield at time t with a remaining maturity of n, the (log) excess return of a n-year bond from t to $t+1$ can be calculated as follows:

[3] https://research.stlouisfed.org/econ/mccracken/fred-databases.

[4] Due to data availability, the sample period for Germany lasts from August 1972 to December 2018.

[5] https://www.bundesbank.de/dynamic/action/de/statistiken/zeitreihen-datenbanken/zeitreihen-datenbank/759778/759778?listId=www_skms_it03a.

[6] For our sample period, the Pearson correlation coefficients for the relationships between the three principal components of the German yield data and the level, slope, and curvature proxies are −99.98%, 84.33%, and −23.04% respectively.

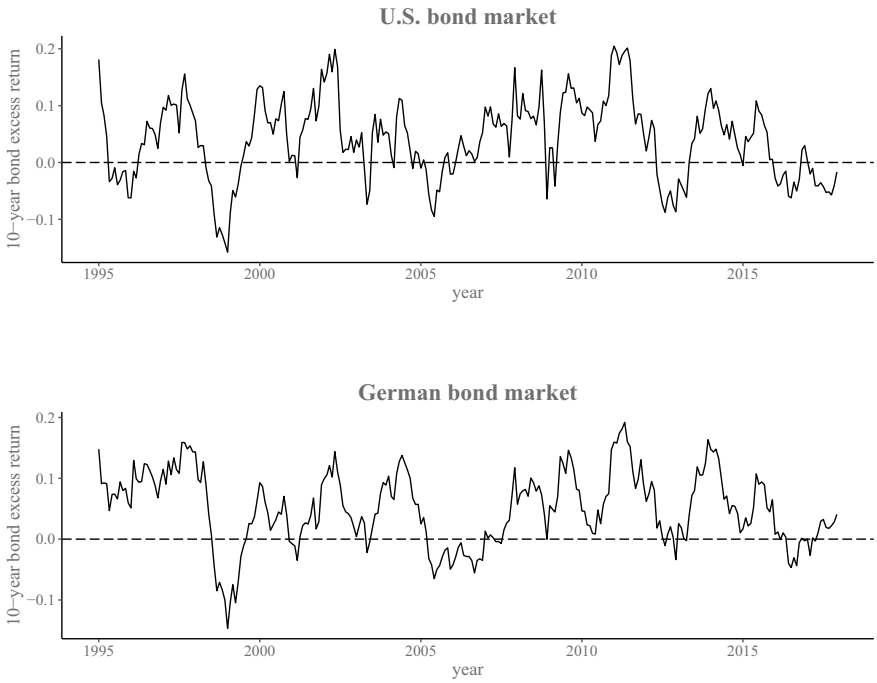

Fig. 1 10-year bond excess returns over time. The upper plot displays the excess returns on 10-year government bonds between January 1995 and December 2017 for the U.S. market, and the bottom plot displays the excess returns on 10-year government bonds between January 1995 and December 2017 for the German market

$$xr_{t+1}^{(n)} = (p_{t+1}^{(n-1)} - p_t^{(n)}) - y_t^{(1)} \tag{1}$$

The return from buying a long-term bond today and selling it after a certain holding period depends on the price of the long-term bond at the end of the holding period. Because this information is unknown at the time of the investment, buying a long-term bond and selling it later is associated with uncertainty. This is reflected in the fact that bond excess returns vary substantially over time (e.g., Ludvigson and Ng 2009), as demonstrated by Fig. 1, which shows the observed excess returns on 10-year government bonds between January 1995 and December 2017—the out-of-sample period in our analysis—for the U.S. and German bond markets.[7] In general, bond excess returns in both bond markets vary from approximately −15% to 20%. This means that sometimes, returns on long-term bonds are higher, and, at other times, returns on short-term bonds are higher. Although bond excess returns in the two bond markets move in a broadly similar direction, this is not the case at every point in time, and bond excess return levels can vary substantially between the two bond markets. The following analyses will identify and compare the determinants of the bond excess returns in these two markets.

[7] Due to the holding period of one year, the last observed bond excess return in our data is from December 2018.

4 Methodology

4.1 Estimation strategy

To predict bond excess returns, it is crucial that we very carefully consider the temporal structure of how information becomes available to decision makers. This consideration is especially important when building machine learning models because technical parameters (called hyperparameters) must be set for these models (see Sect. 4.3 for more detail). Neither the training of machine learning models nor the choice of hyperparameters should be based on ex-post knowledge. Therefore, we split the data available to the decision maker at the respective time into a training, validation, and testing sample using a realistic rolling approach that adapts to newly acquired information every month. This aligns with the recent literature predicting asset returns using machine learning methods (e.g., Gu et al. 2020; Bianchi et al. 2021a, b).

In line with the investment situation of a decision maker, we aim to predict the bond excess returns with different maturities over a holding period of 1 year in every month based on the information available until each respective point. We start the rolling out-of-sample prediction in January 1995 and predict the bond excess return between January 1995 and January 1996. In this step, it is important to be very careful with the information on past bond excess returns that the decision maker could actually have in this situation. Because the most recent bond excess return the decision maker can observe initially is the one between January 1994 and January 1995, the data available for the training and validation sample corresponds to the period August 1971 to January 1994. We follow Bianchi et al. (2021a, b) and use 85% of these data as the training sample and 15% as the validation sample.[8] After predicting the bond excess return between January 1995 and January 1996, we move the rolling window 1 month ahead, build new models based on the training and validation sample that is—in total—1 month longer, and subsequently predict the bond excess return between February 1995 and February 1996. We continue this process until we reach the final time period, which corresponds to the bond excess return between December 2017 and December 2018.

[8] We also conduct analyses with different choices of training and validation sample splits. The results are discussed in Sect. 5.1.

4.2 Predicting bond excess returns with machine learning

Our analysis uses random forests, extremely randomized trees, and artificial neural networks as machine learning methods to predict bond excess returns.[9] We further benchmark these methods with a linear regression.

In the classical linear regression approach, the target of observation i is modeled as a random variable Y_i, which has a distribution that is conditional on the values of the inputs x_{i1}, x_{i2},..., x_{ip}, where p is the number of inputs. Y_i is assumed to be normally distributed conditional on the input variable values. The value of the response variable Y_i of observation i is then assumed to comprise two components: First, the deterministic term that depends on the values of the inputs x_{i1}, x_{i2},..., x_{ip}; second, the random component ε_i. The relationship between the inputs and the random component and the target variable is assumed to be linear, with the parameters β_0, β_1,..., β_p.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{2}$$

The random variable ε_i is assumed to be independent and identically distributed and to follow a normal distribution. Its expected value is given by $E(\varepsilon_{ii}) = 0$, and it has a variance of $Var(\varepsilon_i) = \sigma^2$. The parameters β_0, β_1,..., β_p can be estimated using the least squares method.

The random forest is a tree-based prediction method introduced by Breiman (2001) that builds an ensemble of classification or regression trees. The classification trees and regression trees that make up a random forest are relatively easy-to-explain machine learning methods for approximating non-linear relationships in a data set and for using these relationships to make predictions about new observations. During the training phase, the training data set is used to build a binary tree structure that divides the data set into subsets. The data can be divided based on one of the input variables—in our setting, for example, a forward rate—being above or below a certain threshold. Each leaf node of the tree that results from these splits corresponds to a subset of the training data, while the internal nodes of the tree correspond to a decision rule. In later prediction steps, new observations are classified using these decision rules. Each new observation traverses the tree according to the decision rules. Then, the prediction is calculated based on the final observations in the leaf nodes. In a regression problem such as the one studied in this setting, the mean of the observed responses is used as the predicted value. The splits in the tree are chosen to reduce the mean squared error MSE in the individual leaf nodes τ

$$MSE(\tau) = \sum_{k \in obs.\, in\, \tau} (y_k - \bar{y}(\tau))^2 \tag{3}$$

where $\bar{y}(\tau)$ is the mean target value in τ. The splits of a node can then be chosen to minimize the overall MSE in the resulting child nodes

$$\max_{s \in poss.\, splits} \Delta(s, \tau) = MSE(\tau) - MSE(\tau_L) - MSE(\tau_R) \tag{4}$$

where τ_L is the left child node, and τ_R is the right child node.

[9] We use the terms artificial neural network and neural network interchangeably.

When creating decision trees, there is a trade-off between a strong fit of the training data and its usefulness for predicting new data. If a very deep tree is constructed, it will fit the training data very well, but it might perform poorly when applied to new data, because it might overfit random characteristics of the training data. A tree should be deep enough to capture the important characteristics of the data but flat enough to be useful for making predictions. Therefore, the size of the tree will usually be restricted by hyperparameter choices (see Sect. 4.3) or by reducing the tree size via pruning (Breiman et al. 1984). For example, depth can be restricted by imposing some penalty on new splits or by establishing a minimum for the number of observations in a final node. An approach commonly used in most current studies deploying tree-based methods is to build ensembles of multiple models.

The use of ensembles to improve classification and regression trees was established by several authors. In 1996, Breiman introduced a method called bagging, the short form for bootstrap aggregation (Breiman 1996), which involves producing a set of decision trees by repeatedly sampling from a data set and building a decision tree for each bootstrap sample. The main advantages of bagging are reducing prediction variance by averaging the outcomes of the ensemble of trees and reducing bias by including a larger variety of possible predictions by using the averages of the predictions of the single trees. Later, Breiman (2001) combined the idea of bagging with ideas of other authors such as random split selection (Dietterich 2000), naming the new method "random forest."

The procedure used to build a random forest can be described as follows. For each decision tree T_k, $k \in 1, 2, ...K$ where K is the number of trees, one draws a bootstrap sample as a subset from the training data set. In each node of the tree T_k, one then draws a random sample of size m from the number of input variables M available for the split selection. Then, the tree is fully developed with no pruning. To make predictions, one determines the leaf node of the trees T_k the observation is categorized into and uses the mean target value of the individual leaf nodes as predictions. One then calculates an aggregated prediction over the ensemble using the mean prediction over the individual trees T_k.

As a second machine learning method, we use extremely randomized trees. Introduced by Geurts et al. (2006), the extremely randomized trees method also uses an ensemble of trees to develop multiple classification or regression trees, with each tree randomly selecting the input variables used to split the data. However, extremely randomized trees differ from random forests in two main ways. First, the trees that form a random forest only use a subset of the data; extremely randomized trees use the entire data set. Second, extremely randomized trees randomly choose the split values of the input variables; random forests choose the split values based on an optimization procedure. The method aims to make the trees even more dissimilar to the trees in a random forest and potentially generate a smoother surface of the non-linear function that machine learning methods aim to approximate.

As a third machine learning method, we use neural networks. Neural networks comprise different layers of neurons. The features enter the model through the first layer, the input layer. The input layer comprises one neuron for each feature in the model. Then, the features are passed on ("fed forward") to one or more hidden layers

of neurons.[10] For fully connected hidden layers, each neuron in the layer is connected to all neurons in the previous layer. The neurons assign weights to the inputs from neurons in previous layers and typically have non-linear activation functions that transform the inputs and determine whether they are passed on to the neurons in the next layer. Finally, the transformed features are fed into an output layer. The weights in the layers are optimized via back-propagation to reduce the prediction error.[11] The great advantage of neural networks is their flexibility, a product of the connected hidden layers that allows them to powerfully model non-linear relationships.

4.3 Hyperparameter search

Several technical choices that can be made when designing and building specific machine learning models can affect prediction quality. These include, for example, the minimum size of nodes in the trees of a random forest. Deciding how these hyperparameters are chosen is a crucial step in any machine learning study.

When searching for the best hyperparameter set for a machine learning model, we generally apply a random search. This approach involves sampling various combinations of hyperparameter values using random distributions. Based on these hyperparameter combinations, we then build different models on the training data and validate those models on a separate partition of the available data, namely, the validation data in every month.[12] We then use the model with the best performance in predicting the bond excess returns in the validation data to make a prediction for the test data. Because we use a rolling training, validation, and test split, different hyperparameter combinations could be selected while the rolling window proceeds. As such, we ensure that only information available to decision makers at the time is used.

For the random forest, we consider as hyperparameters the number of variables used in each node of the trees, the number of trees, the minimum size of terminal nodes, the maximum depth of the trees, and the number of observations considered for building each tree. For the extremely randomized trees, the same hyperparameters are used, with the exception of the number of observations. This is because the full training data set is typically used in these models to build each tree. Furthermore, the number of trees is not a typical hyperparameter for either tree-based method because a larger number is always beneficial when reducing measures such as mean squared error while computational effort increases (Probst and Boulesteix 2017). Consequently, we set the number of trees to the reasonably large and typical value of 1,000 (Probst et al. 2019). For neural networks, hyperparameters related to

[10] Due to the relatively small number of observations in the sample, we use neural networks with one hidden layer.

[11] For a more detailed description of neural networks, see, e.g., Hastie et al. (2009).

[12] Due to restrictions regarding computational power, we deviate from monthly hyperparameter tuning for neural networks and follow Bianchi et al. (2021a, b). Details about their procedure appear in their online Appendix.

the architecture of the neural network, such as the number of neurons within the hidden layer(s), are commonly tuned (e.g., Bergstra and Bengio 2012). Neural networks have recently been found to very successfully predict bond excess returns when a certain network architecture is specified ex-ante (Bianchi et al. 2021a, b). Despite these intriguing results, one typically does not know which network architecture produces optimal results ex-ante. Therefore, we choose as hyperparameters the neurons within the hidden layer, the dropout rate (the proportion of neurons in the hidden layer that is omitted during model training to avoid overfitting), and the penalization parameters that decrease the weight of uninformative predictors. Furthermore, we allow the neural network to process the yield data and the macroeconomic data jointly or separately. The exact hyperparameters used for the three machine learning methods appear in Table 5 in the Appendix.

4.4 Performance measures and statistical testing

To assess the performance of the machine learning and linear benchmark models, we compare their predictions of bond excess returns to a naive prediction of bond excess returns based on the historical mean. This involves calculating the out-of-sample R^2 of the predictions according to Campbell and Thompson (2008) using the following equation:

$$R^2_{oos} = 1 - \frac{\sum_{t=0}^{T-1}(xr_{t+1}^{(n)} - \hat{xr}_{t+1}^{(n)})^2}{\sum_{t=0}^{T-1}(xr_{t+1}^{(n)} - \bar{xr}_{t+1}^{(n)})^2} \tag{5}$$

where T is the number of predicted periods in the test sample, $\bar{xr}_{t+1}^{(n)}$ is the naive historical mean prediction of the bond excess returns with maturity n between t and $t+1$ based on the training and validation sample until $t-1$, and $\hat{xr}_{t+1}^{(n)}$ is the prediction produced by a machine learning model or a linear benchmark model.

In a classical (in-sample) linear regression, R^2 values are necessarily between 0 and 1 because the linear regression is fit to reduce the squared errors in the same data. In this way, the prediction produced by linear regression cannot be less accurate than the mean. However, this study calculates the out-of-sample R^2 values based on the test data, meaning the out-of-sample R^2 from Campbell and Thompson (2008) is not bound to the interval between 0 and 1.

To evaluate whether the out-of-sample R^2 values significantly exceed zero, we use a Clark–West test. This step follows Clark and West (2007), who derive a statistic for the difference in *MSE* between models. This statistic accounts for prediction models being susceptible to overfit noise in-sample when the data provide limited or no true information. In such cases, out-of-sample predictions are usually less accurate than simple averages in *MSE* performance (Clark and West 2007). According to Clark and West (2007), t-statistics rejection regions can provide significance levels.

4.5 Explainable artificial intelligence approach

Our study proposes using explainable artificial intelligence to derive interpretable results concerning what determines bond excess returns and how these determinants relate to bond excess returns. This can allow decision makers to obtain useful information from machine learning models. Following Lundberg and Lee (2017), we use SHAP values for this purpose, which provide insight into how much a certain input variable has contributed to a particular prediction produced by a machine learning model. This is referred to as local explainability. Meanwhile, aggregating the SHAP values for each input variable across all predictions enables global explainability.

To derive the contribution of an input variable for a particular prediction, SHAP calculates the change to a target value upon adding an input variable to a model. However, the input variables previously used in the model affect how much the target value changes when the input variable is added. Consequently, determining the contribution of a specific input variable becomes a challenge. To resolve this problem, SHAP uses concepts from game theory that were developed to share the outcome of a game between players making mutually non-exclusive contributions to that outcome (Shapley 1953). SHAP does this by weighting the contribution of the input variables across the different models in which the variables could be added. Interestingly, the resulting SHAP values sum the difference between the mean prediction of the target variable for the training data and the prediction of the target variable for the test data.

This is a particularly useful property if the approach is compared with traditional variable importance techniques. In SHAP, the contribution of each input variable is directly interpretable with regard to the dimension of the target variable and the individual prediction. Traditional global techniques, such as permutation importance, usually only develop an ordering of the importance of individual input variables.

5 Results

5.1 The predictability of U.S. bond excess returns

In the first step of our analysis, we investigate the predictability of U.S. bond excess returns across maturities ranging from 2 to 10 years using the machine learning methods described in Sect. 4.2. We then compare the predictions of the machine learning models with those produced using a traditional linear regression. This enables us to revisit recent findings from the literature suggesting that machine learning models significantly predict U.S. bond excess returns and substantially outperform linear models (Bianchi et al. 2021a, b).

Table 1 presents the results of our analysis. In the table's first section, we see the predictive power of linear regression and machine learning models based on information from the yield curve. We estimate the linear regression using the first three principal components of the yield data as predictors. As discussed, these are typically associated with the level, the slope, and the curvature of the yield curve.

Table 1 Prediction of U.S. bond excess returns

	Models	R^2_{oos}								
		$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(6)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(8)}_{t+1}$	$xr^{(9)}_{t+1}$	$xr^{(10)}_{t+1}$
Yield only	Linear regr. (3 PCs)	−53.49%	−44.78%	−39.71%	−30.05%	−31.68%	−26.33%	−27.76%	−24.29%	−15.98%
	Random forest	−30.41%	−34.19%	−35.98%	−38.06%	−36.69%	−37.50%	−36.57%	−35.76%	−36.70%
	Extr. randomized trees	−18.18%	−23.36%	−28.45%	−30.50%	−30.21%	−29.84%	−28.78%	−28.37%	−27.57%
	Neural network	2.78%	1.12%	0.36%	−1.28%	−0.84%	−2.85%	−3.21%	−3.80%	−6.05%
Incl. macro	Linear regr. (8 PCs)	−63.72%	−43.14%	−36.70%	−27.12%	−24.67%	−20.03%	−19.59%	−17.69%	−9.68%
	Random forest	−9.44%	2.99%**	7.40%**	8.08%**	7.90%**	11.11%**	10.86%**	11.54%**	14.25%**
	Extr. randomized trees	−51.73%	−18.45%	−3.56%	4.72%**	7.98%**	13.41%**	13.36%**	17.82%***	20.50%***
	Neural network	5.72%*	8.00%**	12.00%**	12.89%***	14.84%***	13.52%***	12.77%**	13.12%**	11.31%**

This table reports out-of-sample R^2_{oos} values as in Campbell and Thompson (2008) from predicting U.S. bond excess returns across different maturities with a linear regression and machine learning models using yield data only and using yield data and macroeconomic data together. p-values for the null hypothesis $R^2_{oos} \leq 0$ are calculated following Clark and West (2007). *, **, and *** denote significance at the 10%, 5%, and 1% levels

The linear approach produces negative R^2_{oos} values, indicating that the predictions are substantially less accurate than naive predictions based on the historical mean of U.S. bond excess returns. We subsequently turn to a linear approach based on yield data and macroeconomic data. Specifically, we estimate a linear regression using the first eight principal components across the 1-year spot rate, the forward rates, and all 124 macro variables for the U.S. described in Sect. 3.1 and presented in Table 3 in the Appendix. Again, the linear approach cannot generate positive R^2_{oos} values. Therefore, even the addition of macroeconomic variables seems not to enable a linear regression to significantly predict bond excess returns in a realistic out-of-sample setting.

Turning to machine learning methods, we first investigate the predictive power of a random forest model, an extremely randomized trees model, and a neural network model based solely on yield data. While the random forest and the extremely randomized trees model produce negative R^2_{oos} values across all maturities when using the first three principal components of the yield data as predictors, the neural network produces positive but statistically insignificant R^2_{oos} values in the short term and negative R^2_{oos} values in the long term. Considered alongside the results produced by the linear approach, this indicates that predicting U.S. bond excess returns using only yield data is difficult regardless of the choice of predictive method.

However, combining yield and macroeconomic data substantially changes the capacity of machine learning models to predict U.S. bond excess returns. The lower part of Table 1 displays positive R^2_{oos} values for all three machine learning models. While the neural network model produces statistically significant positive values across all maturities, the extremely randomized trees model produces statistically significant positive values from maturities of five years onward and the random forest model from maturities of three years onward. For long maturities, the tree-based models are superior to the neural networks, explaining up to 20.5% and 14.3% of the variation in U.S. ten-year bond excess returns. These positive R^2_{oos} values are statistically significant. Furthermore, our findings concerning the predictability of U.S. bond excess returns using machine learning methods and the performance ranking

of the models are generally robust against different choices of training and validation sample splits and where the mean squared error is adopted as a measure of prediction accuracy (see Tables 6 and 7 in the Appendix for further details).

Taken together, our results indicate that machine learning models can significantly predict U.S. bond excess returns by using both yield and macroeconomic data, a finding that challenges the idea that the yield curve reflects all information relevant to bond excess return predictions (the spanning hypothesis). This means that the results broadly align with a recent study by Bianchi et al. (2021b), demonstrating that machine learning models outperform the traditional linear approach in terms of predicting excess returns on U.S. bonds. However, closer consideration of the results shows that our findings differ from previous findings in two ways. First, we observe a lower predictive accuracy for neural networks than Bianchi et al. (2021b). This is because we have adopted a flexible approach to constructing the neural networks. Because we cannot know ex ante what the optimal architecture for the neural network is, our approach allows the neural network to choose the optimal number of neurons in the hidden layer and the joint or separate processing of yield and macro data as part of its hyperparameter tuning (see Table 5 in the Appendix for further details).[13] Second, the predictive performances of the tree-based models also differ from Bianchi et al. (2021b). Again, this is due to differences in the methodological approach. For instance, we use the first three principal components of the yield data as input data for the tree-based models rather than using the yield data directly, because the principal components have been shown to provide insights into bond excess returns and have useful interpretations. Furthermore, our hyperparameter tuning deviates from the approach of Bianchi et al. (2021b) (see Table 5 in the Appendix and Sect. 4.3 for further details on our hyperparameter tuning).

The results regarding the predictive performance of the machine learning models are certainly intriguing. However, several important questions remain unanswered: Why do the models predict what they predict? That is, what are the most important determinants of bond excess returns in these models? How exactly do the determinants relate to bond excess returns? Do key determinants differ between bond markets? To answer these questions, our analysis proceeds with the use of SHAP, an explainable artificial intelligence technique that allows us to open the black box of machine learning models.

5.2 The determinants of U.S. bond excess returns

This section uses explainable artificial intelligence to address the central shortcoming of machine learning models, namely, the lack of transparency. This enables us to better understand which determinants drive bond excess returns. For this analysis, we focus on the machine learning model with the best predictive performance,

[13] Using the specific neural network architecture chosen in Bianchi et al. (2021b) based on one layer with 32 neurons for the macroeconomic variables and two separate layers with three neurons each for the yield data, we find comparably strong results, with R^2_{oos} values of $3.40\%^{*}$, $11.05\%^{***}$, $14.78\%^{***}$, $17.07\%^{**}$, $18.01\%^{***}$, $19.56\%^{***}$, $19.93\%^{***}$, $19.94\%^{***}$, and $20.37\%^{***}$.

namely, the extremely randomized trees model for the 10-year U.S. bond excess return. We calculate the mean absolute SHAP values for the model's input variables to examine the average absolute impact of each input variable on the model output. Then, we aggregate the macroeconomic variables to the eight macroeconomic categories according to McCracken and Ng (2016).

Figure 2 presents the results of our analysis. The x-axis shows the mean absolute SHAP values of the first three principal components of the yield data and the aggregated mean absolute SHAP values of the eight macroeconomic categories. We present the principal components and the macroeconomic categories on the y-axis in descending order of relative importance. This means listing the more influential determinants of excess bond returns at the top and the less influential ones at the bottom.

In line with the explanations of theoretical asset pricing models (Wachter 2006; Gabaix 2012; Bansal and Shaliastovich 2013), we can see that information from the yield curve is very important for predicting bond excess returns. According to those models, the yield curve, and especially the slope of the yield curve, captures information such as inflation expectations, making it helpful for predicting bond excess returns. Indeed, we observe that the second principal component of the yield data, which is typically associated with the slope of the yield curve, most impacts the model prediction. This provides evidence in favor of our first hypothesis, namely, that the slope of the yield curve is among the most important determinants of bond excess returns. Furthermore, the first principal component of the yield data, which is typically associated with the level of the yield curve, and—to a somewhat lesser extent—the third principal component of the yield data, which is typically associated with the curvature of the yield curve, also have a considerable impact on the model prediction.

Moving beyond yield curve information, macroeconomic variables related to specific categories also importantly contribute to predictions of U.S. bond excess returns. Interestingly, variables related to the macroeconomic category "Housing," on average, have a particularly high mean absolute impact on the model output. This suggests that housing market information is important for U.S. bond excess returns but does not appear to be fully included in U.S. yield data. These findings add to the relatively scarce literature on the importance of the housing market for asset prices. Most notably in this regard, Piazzesi et al. (2007) develop a consumption-based asset pricing model that explicitly includes housing as a consumption good. In that model, investors favor assets that hedge against negative housing consumption shocks and require excess returns on assets that correlate positively with housing consumption. The authors show that stocks and bonds indeed correlate positively with housing consumption, meaning investors require excess returns on these assets. Our findings provide evidence in favor of the model and align with more recent empirical evidence (Huang and Shi 2023) indicating that variables related to the housing market have predictive power for the excess returns on U.S. bonds. Consideration of other macroeconomic categories reveals that variables related to interest and exchange rates and the labor market also contain information useful for predicting U.S. bond excess returns that is not spanned by the yield curve. Meanwhile,

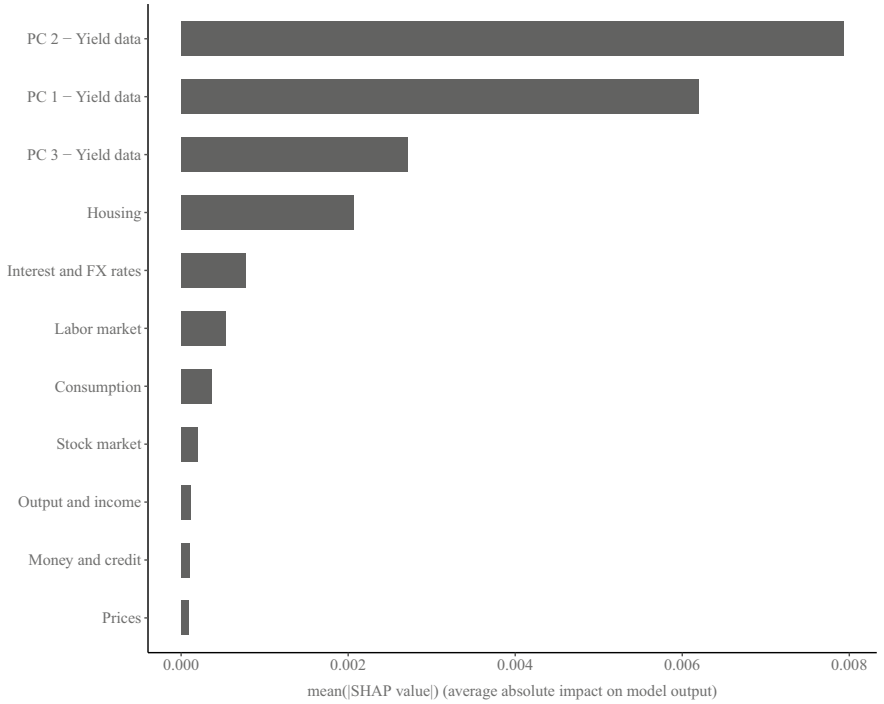

Fig. 2 Importance of 10-year U.S. bond excess return determinants. This figure displays mean absolute SHAP values for the first three principal components of the U.S. yield data and for the macroeconomic variables described in Sect. 3.1, aggregated to the eight macroeconomic categories as in McCracken and Ng (2016). The SHAP values presented are obtained from an extremely randomized trees model predicting 10-year U.S. bond excess returns

the other macroeconomic categories included in this plot, on average, demonstrate rather small impacts on the prediction of bond excess returns.

To test the robustness of our results, we repeat the analysis for the random forest and the neural network. The results appear in Fig. 6 in the Appendix and generally confirm our findings, with all the machine learning models characterized by the high level of importance of several determinants, namely, the first two principal components of the yield data and the housing variables.

After gaining a better understanding of the key determinants of U.S. bond excess returns, we now investigate whether the identified key determinants remain static or change over time. We divide the full sample period into three subperiods of roughly similar length including a crisis period from 2000 to 2009 covering the dotcom bubble and the global financial crisis, a pre-crisis period from 1995 to 1999, and a post-crisis period from 2010 to 2017. For these subperiods, we again calculate mean absolute SHAP values following the procedure previously described.

Figure 3 presents the results of this analysis. We see that over time, the second principal component of the yield data consistently has the largest impact and the first principal component of the yield data consistently has the second-largest impact on the model prediction. Focusing on the importance of different

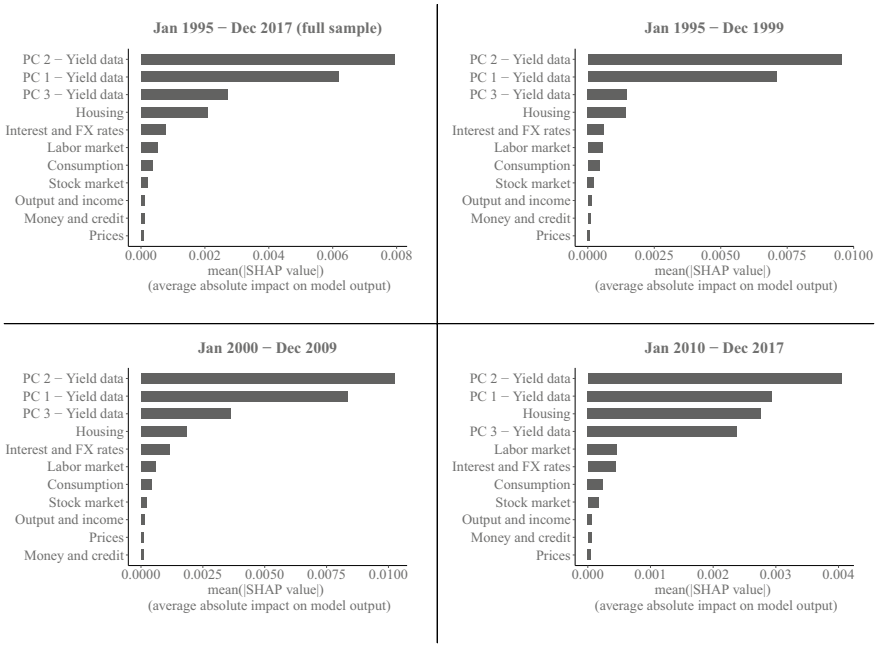

Fig. 3 Importance of 10-year U.S. bond excess return determinants over time. This figure displays mean absolute SHAP values for the first three principal components of the U.S. yield data and for the macroeconomic variables described in Sect. 3.1, aggregated to the eight macroeconomic categories as in McCracken and Ng (2016). The SHAP values presented are obtained from an extremely randomized trees model predicting 10-year U.S. bond excess returns

macroeconomic categories, we find that the "Housing" category has become more important for the prediction of U.S. bond excess returns over the sample period. Interestingly, the relative importance of that category compared to other categories of macroeconomic variables increases particularly strongly in the subperiod after the U.S. subprime mortgage crisis of 2007–2008. This hints towards bond investors paying more attention to the housing market after the crisis. The observation that the importance of the determinants of excess bond returns changes to some degree over time offers a possible explanation for studies focusing on earlier periods (e.g., Ludvigson and Ng 2009) not identifying the housing market as an important determinant of U.S. bond excess returns.

We can conclude that information from the yield curve, especially the slope of the yield curve, is important for predicting U.S. bond excess returns. Beyond information from the yield curve, macroeconomic information related to the housing market is particularly important for predicting U.S. bond excess returns. Moreover, we have observed that the importance of variables related to the housing market has increased substantially over time. While this identifies the key determinants of U.S. bond excess returns, it remains unclear how exactly these determinants relate to bond excess returns. This relationship is important

for an economic interpretation of machine learning predictions and, therefore, of considerable interest to decision makers such as investors and central banks.

5.3 The relationship between U.S. bond excess returns and their key determinants

To understand how the identified key determinants relate to bond excess returns, we further investigate the calculated SHAP values using a different visualization. Figure 4 shows the SHAP values in the form of a sina plot. Compared to the previous figure, here, the SHAP values no longer appear aggregated into macroeconomic categories. Instead, they are shown for the individual input variables. Visualizing the SHAP values with a sina plot provides considerable useful information. First, the plot shows the importance of individual input variables by ranking them in descending order. Furthermore, the plot shows the impact of a particular observation of an input variable, because the horizontal position indicates whether the effect of that input variable's observation is associated with a lower (negative SHAP values) or higher (positive SHAP values) bond excess return prediction. The plot also shows

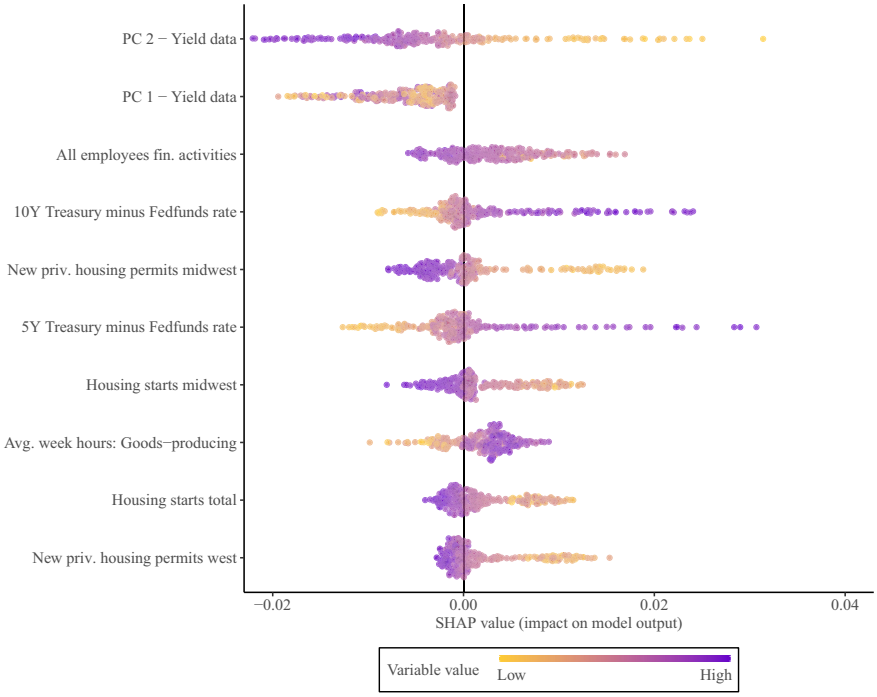

Fig. 4 Relationship between 10-year U.S. bond excess returns and their predictors. This figure displays the SHAP values for the ten variables that are most important for predicting 10-year U.S. bond excess returns. The SHAP values presented are obtained from an extremely randomized trees model predicting 10-year U.S. bond excess returns

the original value of the observation of the input variable, with the color indicating whether that input variable value is low (yellow) or high (violet) for that observation. With the information provided by the plot, it is possible to derive relationships between input variables and bond excess return predictions. For example, a positive relationship between an input variable and bond excess returns can be identified if low input variable values (yellow) lead to lower predictions (negative SHAP values) and high input variable values (violet) lead to higher predictions (positive SHAP values). However, no statistical significance can be inferred from the plot.

Notably, we find a positive relationship between the slope of the yield curve and bond excess return predictions, indicating that a steeper slope of the yield curve leads to higher bond excess return predictions. This is because the second principal component of the yield data is strongly negatively correlated with the slope of the yield curve. In turn, this means that high values for the second principal component of the yield data (low values of the yield curve slope) lead to lower bond excess return predictions, and low values of the second principal component of the yield data (high values of the yield curve slope) lead to higher bond excess return predictions. This aligns strongly with consumption-based asset pricing models. For example, Gabaix (2012) argues that higher inflation expectations are reflected in a steeper slope of the yield curve. Because the slope of the yield curve is mean-reverting, an increase in the slope predicts a subsequent decrease and, therefore, higher bond excess returns.[14]

Interestingly, based on their high relative importance, we also observe that the variables "5Y Treasury rate minus Fedfunds rate" and "10Y Treasury rate minus Fedfunds rate" from the macroeconomic category "interest and exchange rates" seemingly offer further explanatory power for U.S. bond excess returns by using a different reference point (in this case the Fedfunds rate) to calculate the slope of the yield curve. Consistent with the previous finding, the plot also shows a positive relationship between the variables and the bond excess return predictions, with a steeper slope associated with higher bond excess return predictions.

Visualizing the SHAP values in this way enables us to further analyze the relationship between the important housing market variables identified earlier and the predictions for bond excess returns in the same way. In line with our previous analysis, we see four variables related to the housing market among the prediction model's ten most important input variables. The four variables "Permits for new private housing midwest," "Housing starts midwest," "Housing starts total," and "Permits for new private housing west" all indicate the same impact direction on bond excess returns, with a negative relationship observed between the number of new construction starts or permits and the predicted bond excess returns. That is, a lower number of new construction starts or permits leads to higher U.S. bond excess return predictions. Again, this finding aligns strongly with the theoretical asset pricing model described by Piazzesi et al. (2007), who show that the consumption of a housing

[14] A decrease in the slope of the yield curve means that the prices of long-term bonds increase relative to the prices of short-term bonds, which translates into increasing bond excess returns.

good correlates positively with stock and bond prices and, therefore, predicts excess returns.

Overall, implementing explainable artificial intelligence not only identifies the key determinants of bond excess returns but also provides insight into how these determinants relate to bond excess returns. For example, knowing how macroeconomic information, such as housing variables, relate to U.S. bond excess returns gives us a better overall understanding of what drives bond excess returns, which is critical for decision makers, who can act based on these insights. For U.S. monetary policy, therefore, the housing market should be closely monitored because negative developments in that domain indicate increasing excess returns on long-term bonds. Investors should also incorporate this information into their analyses of bond markets and corresponding investment strategies.

5.4 The determinants of German bond excess returns

Having analyzed the determinants of the predictions of U.S. bond excess returns produced by machine learning models, we now investigate whether these determinants differ between bond markets. This involves consideration of another highly important bond market, namely, the German bond market. Although the realized bond excess returns for the German and the U.S. markets exhibit substantial co-movement (see Fig. 1), empirical studies have documented the important role of local factors in predicting bond excess returns (Barr and Priestley 2004; Pérignon et al. 2007). However, whether these local factors differ between bond markets remains unclear, demanding empirical investigation.

As for the U.S. market, we begin by assessing the predictions of German bond excess returns produced by the traditional linear regression and the selected machine learning models. The results appear in Table 2. Again, we see that a linear regression based on yield data cannot successfully predict bond excess returns, with all R^2_{oos} values negative. We also see that, again, adding macroeconomic information

Table 2 Prediction of German bond excess returns

	Models	R^2_{oos}								
		$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(6)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(8)}_{t+1}$	$xr^{(9)}_{t+1}$	$xr^{(10)}_{t+1}$
Yield only	Linear regr. (3 PCs)	−66.05%	−65.80%	−63.92%	−61.86%	−59.74%	−57.67%	−55.62%	−53.32%	−51.16%
	Random forest	−15.06%	−26.15%	−35.42%	−44.17%	−47.57%	−53.33%	−56.62%	−55.92%	−55.59%
	Extr. randomized trees	−8.16%	−18.07%	−24.79%	−29.56%	−35.01%	−41.15%	−45.28%	−46.01%	−46.19%
	Neural network	3.31%*	2.43%	1.16%	−0.30%	−1.49%	−2.45%	−3.00%	−3.35%	−3.52%
Incl. macro	Linear regr. (8 PCs)	−41.83%	−32.73%	−28.48%	−26.35%	−25.39%	−25.00%	−24.96%	−25.34%	−25.78%
	Random forest	-14.88%	−13.51%	−9.33%	−3.58%	0.61%	4.44%	5.43%*	7.14%*	9.06%**
	Extr. randomized trees	−30.49%	−21.78%	−14.29%	−7.05%	−4.76%	−0.30%	3.94%*	6.14%**	9.64%**
	Neural network	3.66%	3.18%	1.83%	0.77%	−0.13%	−0.77%	−1.43%	−1.62%	−1.75%

This table reports out-of-sample R^2_{oos} values as in Campbell and Thompson (2008) from predicting German bond excess returns across different maturities with a linear regression and machine learning models using yield data only and using yield data and macroeconomic data together. p-values for the null hypothesis $R^2_{oos} \leq 0$ are calculated following Clark and West (2007). *, **, and *** denote significance at the 10%, 5%, and 1% levels

through a PCA incorporating yield and macroeconomic data does not produce positive R^2_{oos} values either. Furthermore, as for the U.S. market, machine learning models cannot predict bond excess returns solely based on yield data either. However, incorporating macroeconomic data enables the random forest and the extremely randomized trees model to obtain significantly positive predictions of German bond excess returns for longer maturities. For German 10-year government bonds, both tree-based models predict close to 10% of the variation in excess returns in an out-of-sample setting. Given the difficulty of the task and the lower amount of available macroeconomic variables than in the U.S., we interpret this as a strong result.

In the next step, we analyze the key determinants for the predictions of German bond excess returns produced by machine learning models. As for our previous analyses, we use SHAP to specifically investigate the predictions of the best-performing model, namely, the extremely randomized trees model. The results appear in Fig. 5. The x-axis of the plot again shows the respective mean absolute SHAP values of the first three principal components of the German yield data and the aggregated mean absolute SHAP values for the eight macroeconomic categories used in the analysis for the U.S. The aggregated SHAP values of the different principal components and

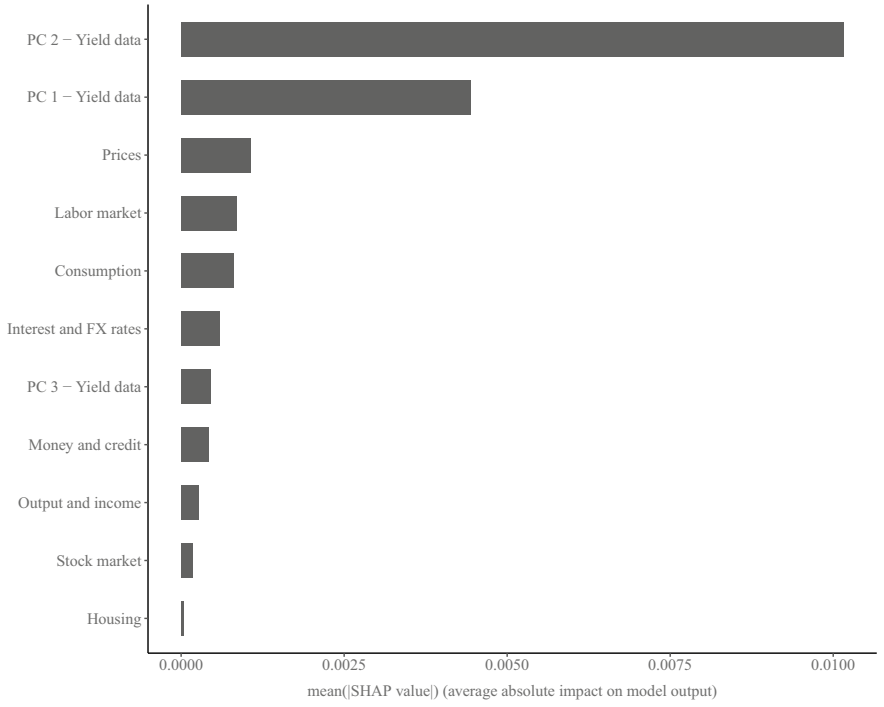

Fig. 5 Importance of 10-year German bond excess return determinants. This figure displays the mean absolute SHAP values for the first three principal components of the German yield data and for the macroeconomic variables described in Sect. 3.1, aggregated to the eight macroeconomic categories as in McCracken and Ng (2016). The SHAP values presented are obtained from an extremely randomized trees model predicting 10-year German bond excess returns

macroeconomic categories appear in descending order according to their relative importance on the y-axis.

In line with the results for the U.S. bond market, we find that, for the German bond market, the second principal component of the yield data is the most important and the first principal component of the yield data is the second most important determinant of bond excess returns. As for the U.S., the first two principal components closely relate to the level and the slope of the German yield curve. This finding provides further evidence in favor of our first hypothesis, which asserts that the slope of the yield curve is an important determinant of bond excess returns. Turning to the relationship between the second principal component of the yield curve and the excess returns on 10-year German bonds, we again find that a steeper slope of the yield curve leads to higher bond excess return predictions (see Fig. 7 in the Appendix), confirming our previous results.

Regarding additional macroeconomic variables, again, some of these variables contain information pertaining to bond excess returns that is not spanned by the yield curve. Variables from the macroeconomic categories "Prices" and "Labor market" are, on average, the most important additional macroeconomic variables for predictions of German bond excess returns. Variables from the macroeconomic categories "Consumption" and "Interest and exchange rates" also, on average, contribute to the machine learning model's predictions. Interestingly, variables from the macroeconomic category "Housing," the most important macroeconomic determinants for U.S. bond excess returns, have nearly no explanatory power on average beyond information from the yield curve for German bond excess returns. This result is consistent with our second hypothesis, which asserts that the macroeconomic determinants of bond excess returns that have explanatory power beyond information from the yield curve differ between bond markets.

To reconcile this finding with economic intuition, we consider two possible explanations. First, it is possible that the risks associated with the local housing market for local bond prices differ between the two markets. Second, it is possible that the risks associated with the local housing market for local bond prices are already reflected in the local yield curve to differing degrees. Based on our findings, we conjecture that the U.S. housing market includes substantial risks for U.S. bonds that are not already fully reflected in the yield curve. Notably, the existence of such risks for U.S. bonds stemming from the local U.S. housing market was particularly evident during the subprime mortgage crisis. For Germany, such risks from the housing market were not observed in the past. Therefore, we suspect that the German housing market poses less risk to German bonds than the U.S. housing market poses to U.S. bonds. However, it is theoretically possible that substantial risks for German bonds from the local housing market exist and are already largely reflected in the yield curve. We leave further analysis of this issue for future research.

6 Conclusion

There is an ongoing debate about the determinants of bond excess returns. A thorough understanding of these determinants and their relationship with bond excess returns is important for decision makers, such as investors and central banks. Recent studies show that machine learning models can predict bond excess returns and clearly outperform linear models. However, a central shortcoming of these machine learning models is their lack of transparency, which complicates the economic interpretations of the relationship between the predicted bond excess returns and their determinants. To address this issue, we have used the state-of-the-art explainable artificial intelligence technique SHAP to open the black box of these machine learning models.

We contribute to the literature by providing a deeper understanding of the determinants of bond excess returns. Specifically, we have identified the key determinants in machine learning models and revealed the relationship between these key determinants and the bond excess returns. By comparing the U.S. and German contexts, we have highlighted differences between the determinants in the two bond markets. Methodologically, we contribute an empirical approach based on explainable artificial intelligence that is suited to opening up black-box machine learning models to predict the returns on assets in general, not only bonds.

Our results have important implications for practitioners and academics. Investors can better understand which factors determine their bond portfolio's returns, and central banks can better understand the excess returns that investors demand of long-term bonds. This is important for the transmission of monetary policy because central banks can only control the short-term interest rates directly. For researchers investigating empirical asset pricing, our study encourages the use of explainable artificial intelligence to increase the transparency of the predictions of machine learning models.

Appendix

Table 3 Macroeconomic variables U.S.

Variable	Description	Tr.	Source
Group 1: Output and income			
CUMFNS	Capacity utilization: Manufacturing	2	FRED
INDPRO	IP Index	5	FRED
IPB51222s	IP: Residential utilities	5	FRED
IPBUSEQ	IP: Business equipment	5	FRED
IPCONGD	IP: Consumer goods	5	FRED
IPDCONGD	IP: Durable consumer goods	5	FRED
IPDMAT	IP: Durable materials	5	FRED
IPFINAL	IP: Final products (market group)	5	FRED
IPFPNSS	IP: Final products and nonindustrial supplies	5	FRED
IPFUELS	IP: Fuels	5	FRED
IPMANSICS	IP: Manufacturing (SIC)	5	FRED
IPMAT	IP: Materials	5	FRED
IPNCONGD	IP: Nondurable consumer goods	5	FRED
IPNMAT	IP: Nondurable materials	5	FRED
RPI	Real personal income	5	FRED
W875RX1	Real personal income ex transfer receipts	5	FRED
Group 2: Labor market			
AWHMAN	Avg weekly hours: Manufacturing	1	FRED
AWOTMAN	Avg weekly overtime hours: Manufacturing	2	FRED
CE16OV	Civilian employment	5	FRED
CES0600000007	Avg weekly hours: Goods-Producing	1	FRED
CES0600000008	Avg hourly earnings: Goods-producing	6	FRED
CES1021000001	All employees: Mining and logging: Mining	5	FRED
CES2000000008	Avg hourly earnings: Construction	6	FRED
CES3000000008	Avg hourly earnings: Manufacturing	6	FRED
CLAIMSx	Initial claims	5	FRED
CLF16OV	Civilian labor force	5	FRED
DMANEMP	All Employees: Durable goods	5	FRED
HWI	Help-wanted index for United States	2	FRED
HWIURATIO	Ratio of help wanted/no. unemployed	2	FRED
MANEMP	All employees: Manufacturing	5	FRED
NDMANEMP	All employees: Nondurable goods	5	FRED
PAYEMS	All employees: Total nonfarm	5	FRED
SRVPRD	All employees: Service-providing industries	5	FRED
UEMP15OV	Civilians unemployed - 15 weeks & over	5	FRED
UEMP15T26	Civilians unemployed for 15–26 weeks	5	FRED
UEMP27OV	Civilians unemployed for 27 weeks & over	5	FRED
UEMP5TO14	Civilians unemployed for 5–14 weeks	5	FRED
UEMPLT5	Civilians unemployed - less than 5 weeks	5	FRED

Table 3 (continued)

Variable	Description	Tr.	Source
UEMPMEAN	Average duration of unemployment (weeks)	2	FRED
UNRATE	Civilian unemployment rate	2	FRED
USCONS	All employees: Construction	5	FRED
USFIRE	All employees: Financial activities	5	FRED
USGOOD	All employees: Goods-producing industries	5	FRED
USGOVT	All employees: Government	5	FRED
USTPU	All employees: Trade, transportation & utilities	5	FRED
USTRADE	All employees: Retail trade	5	FRED
USWTRADE	All employees: Wholesale trade	5	FRED
Group 3: Housing			
HOUST	Housing starts: Total new privately owned	4	FRED
HOUSTMW	Housing starts, midwest	4	FRED
HOUSTNE	Housing starts, northeast	4	FRED
HOUSTS	Housing starts, south	4	FRED
HOUSTW	Housing starts, west	4	FRED
PERMIT	New private housing permits (SAAR)	4	FRED
PERMITMW	New private housing permits, midwest (SAAR)	4	FRED
PERMITNE	New private housing permits, northeast (SAAR)	4	FRED
PERMITS	New private housing permits, south (SAAR)	4	FRED
PERMITW	New private housing permits, west (SAAR)	4	FRED
Group 4: Consumption, orders, and inventories			
AMDMNOx	New orders for durable goods	5	FRED
AMDMUOx	Unfilled orders for durable goods	5	FRED
ANDENOx	New orders for nondefense capital goods	5	FRED
BUSINVx	Total business inventories	5	FRED
CMRMTSPLx	Real manu. and trade industries sales	5	FRED
DPCERA3M086SBEA	Real personal consumption expenditures	5	FRED
ISRATIOx	Total business: Inventories to sales ratio	2	FRED
RETAILx	Retail and food services sales	5	FRED
Group 5: Money and credit			
BOGMBASE	Monetary base	6	FRED
BUSLOANS	Commercial and industrial loans	6	FRED
CONSPI	Nonrevolving consumer credit to personal income	2	FRED
DTCOLNVHFNM	Consumer motor vehicle loans outstanding	6	FRED
DTCTHFNM	Total consumer loans and leases outstanding	6	FRED
INVEST	Securities in bank credit at all commercial banks	6	FRED
M1SL	M1 money stock	6	FRED
M2REAL	Real M2 money stock	5	FRED
M2SL	M2 money stock	6	FRED
NONBORRES	Reserves of depository institutions	7	FRED
NONREVSL	Total nonrevolving credit	6	FRED
REALLN	Real estate loans at all commercial banks	6	FRED
TOTRESNS	Total reserves of depository institutions	6	FRED

Table 3 (continued)

Variable	Description	Tr.	Source
Group 6: Interest and exchange rates			
AAA	Moody's seasoned Aaa corporate bond yield	2	FRED
AAAFFM	Moody's Aaa corporate bond minus FEDFUNDS	1	FRED
BAA	Moody's seasoned Baa corporate bond yield	2	FRED
BAAFFM	Moody's Baa corporate bond minus FEDFUNDS	1	FRED
COMPAPFFx	3-Month commercial paper minus FEDFUNDS	1	FRED
CP3Mx	3-Month AA financial commercial paper rate	2	FRED
EXCAUSx	Canada / U.S. foreign exchange rate	5	FRED
EXJPUSx	Japan / U.S. foreign exchange rate	5	FRED
EXSZUSx	Switzerland / U.S. foreign exchange rate	5	FRED
EXUSUKx	U.S. / U.K. foreign exchange rate	5	FRED
FEDFUNDS	Effective federal funds rate	2	FRED
GS1	1-Year Treasury Rate	2	FRED
GS10	10-Year Treasury Rate	2	FRED
GS5	5-YearTreasury Rate	2	FRED
T10YFFM	10-Year Treasury C minus FEDFUNDS	1	FRED
T1YFFM	1-Year Treasury C minus FEDFUNDS	1	FRED
T5YFFM	5-Year Treasury C minus FEDFUNDS	1	FRED
TB3MS	3-Month Treasury Bill	2	FRED
TB3SMFFM	3-Month Treasury C minus FEDFUNDS	1	FRED
TB6MS	6-Month Treasury Bill	2	FRED
TB6SMFFM	6-Month Treasury C minus FEDFUNDS	1	FRED
Group 7: Prices			
CPIAPPSL	CPI: Apparel	6	FRED
CPIAUCSL	CPI: All items	6	FRED
CPIMEDSL	CPI: Medical care	6	FRED
CPITRNSL	CPI: Transportation	6	FRED
CPIULFSL	CPI: All items less food	6	FRED
CUSR0000SA0L2	CPI: All items less shelter	6	FRED
CUSR0000SA0L5	CPI: All items less medical care	6	FRED
CUSR0000SAC	CPI: Commodities	6	FRED
CUSR0000SAD	CPI: Durables	6	FRED
CUSR0000SAS	CPI: Services	6	FRED
DDURRG3M086SBEA	Personal cons. expend.: Durable goods	6	FRED
DNDGRG3M086SBEA	Personal cons. expend.: Nondurable goods	6	FRED
DSERRG3M086SBEA	Personal cons. expend.: Services	6	FRED
OILPRICEx	Crude oil, spliced WTI and cushing	6	FRED
PCEPI	Personal cons. expend.: Chain index	6	FRED
PPICMM	PPI: Metals and metal products	6	FRED
WPSFD49207	PPI: Finished goods	6	FRED
WPSFD49502	PPI: Finished consumer goods	6	FRED
WPSID61	PPI: Intermediate materials	6	FRED

Table 3 (continued)

Variable	Description	Tr.	Source
WPSID62	PPI: Crude materials	6	FRED
Group 8: Stock market			
S &P 500	S &P's Common stock price index: Composite	5	FRED
S &P div yield	S& P's composite common stock: Dividend yield	2	FRED
S &P PE ratio	S& P's composite common stock: PE ratio	5	FRED
S &P: indust	S& P's common stock price index: Industrials	5	FRED
VIXCLSx	VIX	1	FRED

This table displays the U.S. macroeconomic variables used to predict U.S. bond excess returns. It contains the variable descriptions, the transformation of the variables ("Tr."), and their data source. The variables are transformed and grouped based on McCracken and Ng (2016). The transformation code denotes the following transformation for time series x: (1): no transformation; (2): Δx_t; (3): $\Delta^2 x_t$; (4): $log(x_t)$; (5): $\Delta log(x_t)$; (6): $\Delta^2 log(x_t)$; (7): $\Delta(x_t/x_{t-1} - 1.0)$

Table 4 Macroeconomic variables Germany

Variable	Description	Tr.	Source
Group 1: Output and income			
aDECEXPB/A	Merchandise exports, stand	5	Reuters
aDECIMPB/A	Merchandise imports, stand	5	Reuters
aDECVISB/A	Visible trade balance, stand	1	Reuters
aDEEXPGDSB/A	Exports total	5	Reuters
aDEIMPGDSB/A	Imports total	5	Reuters
DEUPRCNTO01GPSAM	Production: Construction: Total construction	1	FRED
DEUPRINTO01GYSAM	Production: Total industry excl. construction	1	FRED
DEUPRMNIG01IXOBSAM	Production: Manufacturing: Intermediate goods	1	FRED
DEUPRMNTO01GYSAM	Production: Total manufacturing	1	FRED
DEUPRMNVG01IXOBSAM	Production: Manufacturing: Investment goods	1	FRED
DEUPROCONMISMEI	Production of total construction	1	FRED
DEUPROINDMISMEI	Production of total industry	1	FRED
DEUPROMANMISMEI	Production in total manufacturing	1	FRED
DEUXTEXVA01CXMLM	International trade: Exports: Value (goods)	5	FRED
DEUXTIMVA01CXMLM	International trade: Imports: Value (goods)	5	FRED
DEUXTNTVA01CXMLM	International trade: Net trade: Value (goods)	2	FRED
XTEXVA01DEM667S	Exports: Value goods	5	FRED
XTIMVA01DEM667S	Imports: Value goods	5	FRED
XTNTVA01DEM667S	Net trade: Value goods	2	FRED
Group 2: Labor market			
aDECUNPO	Unempl. level, stand	4	Reuters
aDECUNPPO/A	Unempl. level, % MOM, stand., chg P/P	1	Reuters
aDECUNPYO	Unempl. level, % YOY, stand., chg Y/Y	1	Reuters
aDECUNPYQ/A	Unempl. rate, YOY, stand	1	Reuters
aDECVACPO/A	Job vacancies, % MOM, stand., chg P/P	1	Reuters
aDECVACYO/A	Job vacancies, % YOY, stand., chg Y/Y	1	Reuters
aDEWGSLHRBS.1D/C	Wages & sal, hrly basis - prod sct, 2005=100	1	Reuters
aDEWSTOT	Wages and salaries, overall economy, monthly	5	Reuters
LMJVTTUVDEM647S	Total unfilled job vacancies	5	FRED
LMUNRLTTDEM647S	Registered unempl. level	5	FRED
LMUNRRTTDEM156S	Registered unempl. rate	2	FRED
Group 3: Housing			
aDEBPERMITI/A	Building permits, non-residential, industrial	5	Reuters
aDEBPERMITR/A	Building permits, residential	5	Reuters
aDEBPERMNR/A	Building permits, non-residential, total	5	Reuters
Group 4: Consumption, orders, and inventories			
aDENCAR	New passenger car registrations	5	Reuters
aDENOMFG/CA	New orders, manufacturing industry	1	Reuters
DEUSARTMISMEI	Total retail trade	1	FRED
DEUSLRTTO02IXOBSAM	Sales: Retail trade: Total value	1	FRED
DEUSLWHTO02IXOBSAM	Sales: Wholesale trade: Total value	1	FRED

Table 4 (continued)

Variable	Description	Tr.	Source
Group 5: Money and credit			
aDEBANKLPA	Lending to enterprises & individuals	5	Reuters
aDEDEBTD	Central government debt	5	Reuters
aDEM2BC/A	Money supply - M2	6	Reuters
aDEM3ABC/A	Money supply - M3	6	Reuters
TRESEGDEM052N	Total reserves excluding gold	5	FRED
Group 6: Interest and exchange rates			
aDEBISNXNR	BIS, nominal narrow effective exch. rate index	1	Reuters
aDEBISRXNR	BIS, real narrow effective exch. rate index	1	Reuters
CCUSMA02DEM618N	National currency to US Dollar exch. rate	5	FRED
CCUSSP01DEM650N	US Dollar to national currency spot exch. rate	5	FRED
IR3TIB01DEM156N	3-month or 90-day rates and yields	2	FRED
IRLTLT01DEM156N	Long-term government bond yields: 10-year	2	FRED
IRSTCI01DEM156N	Immediate rates: Less than 24 h	2	FRED
NNDEBIS	Narrow effective exch. rate for Germany	1	FRED
Group 7: Prices			
aDECPPIE/CA	PPI, stand	1	Reuters
aDECPPIPE/A	PPI, % MOM, stand., chg P/P	1	Reuters
aDECPPIYF	PPI, % YOY, stand., chg Y/Y	1	Reuters
aDEEXP	Export prices	1	Reuters
aDEIMP	Import prices	1	Reuters
aDEWPI	Wholesale prices	1	Reuters
CPALTT01DEM659N	CPI: Total all items	1	FRED
CPGDFD02DEM657N	CPI: Total food excluding restaurants	1	FRED
DEUCP040500GYM	CPI: Housing, water, electricity, gas, oth. fuels	1	FRED
DEUCPIALLMINMEI	CPI of all items in Germany	1	FRED
DEUCPICORMINMEI	CPI: All items excluding food and energy	1	FRED
DEUCPIENGMINMEI	CPI: Energy	1	FRED
DEUCPIFODMINMEI	CPI: Food	1	FRED
DEUCPIHOUMINMEI	CPI: Housing	1	FRED
DEUPPDMMINMEI	Domestic producer prices index: Manufacturing	1	FRED
Group 8: Stock market			
SPASTT01DEM657N	Total share prices for all shares for Germany	1	FRED

This table displays the macroeconomic variables for Germany used to predict German bond excess returns. It contains the variable descriptions, the transformation of the variables ("Tr."), and their data source. The variables are transformed and grouped based on McCracken and Ng (2016). The transformation code denotes the following transformation for time series x: (1): no transformation; (2): Δx_t; (3): $\Delta^2 x_t$; (4): $log(x_t)$; (5): $\Delta log(x_t)$; (6): $\Delta^2 log(x_t)$; (7): $\Delta(x_t/x_{t-1} - 1.0)$

Table 5 Hyperparameter tuning. This table reports the hyperparameter sets for the three machine learning methods we use to predict bond excess returns

Machine learning method	Hyperparameter set
Random Forest	Number of variables ∈ {10%, 15%, 25%, 33.33%, 40%, 50%}
	Number of trees ∈ {1000}
	Minimum terminal node size ∈ {1, 3, 5}
	Maximum depth of the trees ∈ {4, 6, 8, 10}
	Sample fraction ∈ {10%, 15%, 20%, 25%, 30%, 40%, 50%}
Extreme Trees	Number of variables ∈ {15%, 25%, 33.33%, 40%, 50%}
	Number of trees ∈ {1000}
	Minimum terminal node size ∈ {1, 3, 5, 7, 10}
	Maximum depth of the trees ∈ {4, 6, 8, 10}
	Sample fraction ∈ {100%}
Neural network	Our neural network design is generally based on Bianchi et al. (2021a, b)[a]. For a detailed description of the neural network settings, we refer to the online Appendix of Bianchi et al. (2021a)
	Dropout rate ∈ {30%, 40%, 50%}
	Penalization parameter ∈ {0.0005, 0.001, 0.0015}
	Setting "Yield data"
	Number of neurons ∈ {[3], [2], [1]}
	Setting "Yield and macro data"
	Separate processing
	Number of neurons (yield data) ∈ {[3], [2], [1]}
	Number of neurons (macro data) ∈ {[64], [32], [16]}
	Joint processing
	Number of neurons ∈ {[64], [32], [16]}

[a]We deviate from this approach by tuning the neurons in the hidden layers as described in Sect. 4.3, by averaging the prediction using the ten best models out of 20, rather than 100 estimated models due to computational constraints and by allowing the neural work to process the yield and macroeconomic data in separate groups or jointly

Table 6 Prediction of U.S. bond excess returns: different training & validation splits

	Models	R^2_{oos}								
		$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(6)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(8)}_{t+1}$	$xr^{(9)}_{t+1}$	$xr^{(10)}_{t+1}$
80 / 20	Random forest	−10.50%	−0.86%	−0.99%	1.85%*	3.22%*	4.95%*	5.26%*	6.82%**	9.91%**
	Extr. randomized trees	−50.03%	−20.79%	−11.03%	−3.28%	2.23%*	4.43%*	5.72%**	8.90%**	11.28%**
	Neural network	−3.70%	6.79%**	8.77%**	10.28%**	11.69%**	10.97%**	10.73%**	10.69%**	9.07%**
90 / 10	Random forest	−15.25%	−5.78%	−7.14%	−4.35%	−0.94%	−1.12%	1.39%*	2.96%*	7.40%*
	Extr. randomized trees	−46.60%	−21.97%	−12.41%	−5.03%	−0.94%	3.75%**	6.18%**	10.89%**	15.02%**
	Neural network	−0.21%	2.29%*	6.39%**	8.35%**	9.68%**	8.19%**	8.01%**	8.25%**	7.43%*

This table reports out-of-sample R^2_{oos} values as in Campbell and Thompson (2008) from predicting U.S. bond excess returns across different maturities with machine learning models using yield data and macroeconomic data together. The first section of the table contains R^2_{oos} values based on a 80% / 20%-split. The second section of the table contains R^2_{oos} values based on a 90% / 10%-split. p-values for the null hypothesis $R^2_{oos} \leq 0$ are calculated following Clark and West (2007). *, **, and *** denote significance at the 10%, 5%, and 1% levels

Table 7 Prediction of U.S. bond excess returns: OOS-MSE as an alternative to measure forecasting accuracy

	Models	MSE_{oos}								
		$xr^{(2)}_{t+1}$	$xr^{(3)}_{t+1}$	$xr^{(4)}_{t+1}$	$xr^{(5)}_{t+1}$	$xr^{(6)}_{t+1}$	$xr^{(7)}_{t+1}$	$xr^{(8)}_{t+1}$	$xr^{(9)}_{t+1}$	$xr^{(10)}_{t+1}$
Yield only	Linear regr. (3 PCs)	0.000193	0.000687	0.001312	0.002007	0.002878	0.003753	0.004764	0.005719	0.006574
	Random forest	0.000164	0.000637	0.001277	0.002131	0.002987	0.004085	0.005093	0.006247	0.007749
	Extr. randomized trees	0.000149	0.000586	0.001206	0.002014	0.002846	0.003857	0.004802	0.005907	0.007231
	Neural network	0.000122	0.000470	0.000935	0.001563	0.002204	0.003055	0.003849	0.004776	0.006011
Incl. macro	Linear regr. (3 PCs)	0.000206	0.000680	0.001283	0.001962	0.002725	0.003565	0.004460	0.005416	0.006217
	Random forest	0.000138	0.000461	0.000869	0.001419	0.002013	0.002641	0.003324	0.004070	0.004861
	Extr. randomized trees	0.000191	0.000562	0.000972	0.001470	0.002011	0.002572	0.003231	0.003781	0.004506
	Neural network	0.000119	0.000437	0.000826	0.001344	0.001861	0.002569	0.003253	0.003998	0.005027

This table reports out-of-sample mean squared errors from predicting U.S. bond excess returns across different maturities with a linear regression and machine learning models using yield data only and using yield data and macroeconomic data together

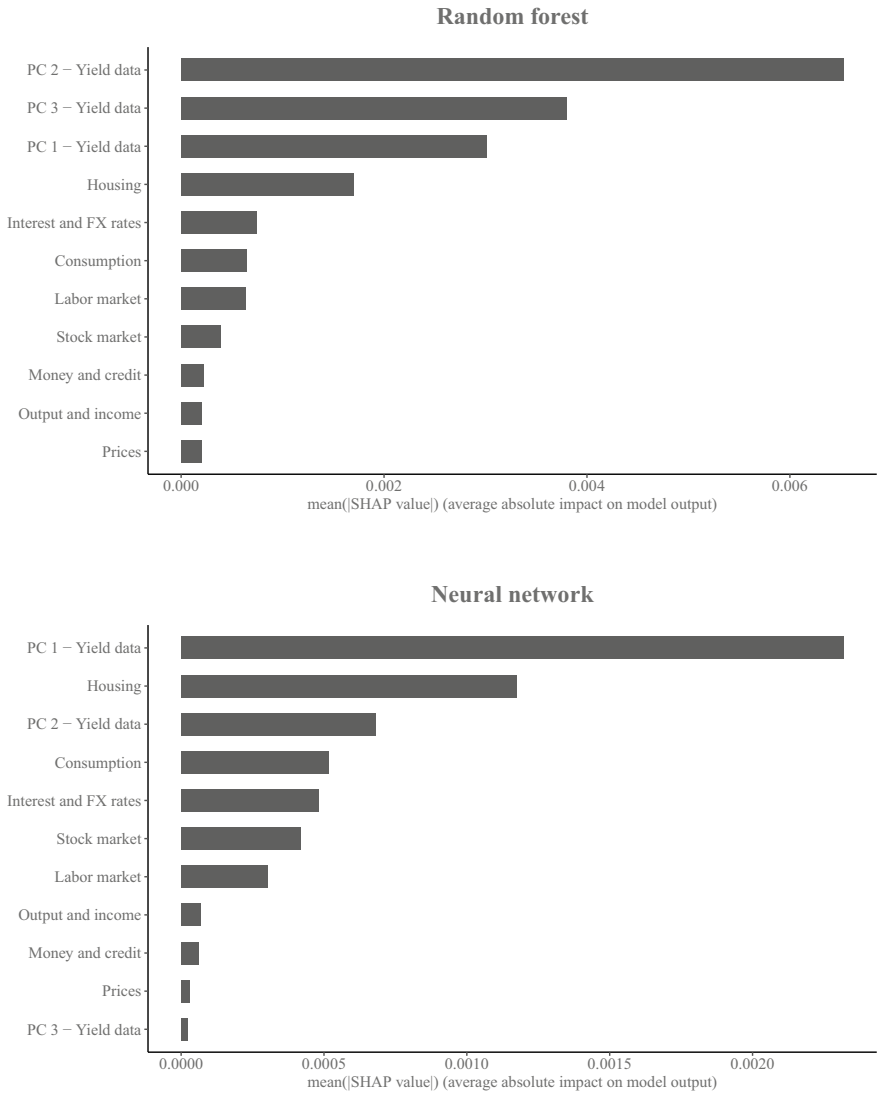

Fig. 6 Importance of 10-year U.S. bond excess return determinants: Random forest and neural network. This figure displays mean absolute SHAP values for the first three principal components of the U.S. yield data and for the macroeconomic variables described in Sect. 3.1, aggregated to eight macroeconomic categories as in McCracken and Ng (2016). The SHAP values presented are obtained from a random forest model (upper plot) and a neural network model (bottom plot) predicting 10-year U.S. bond excess returns

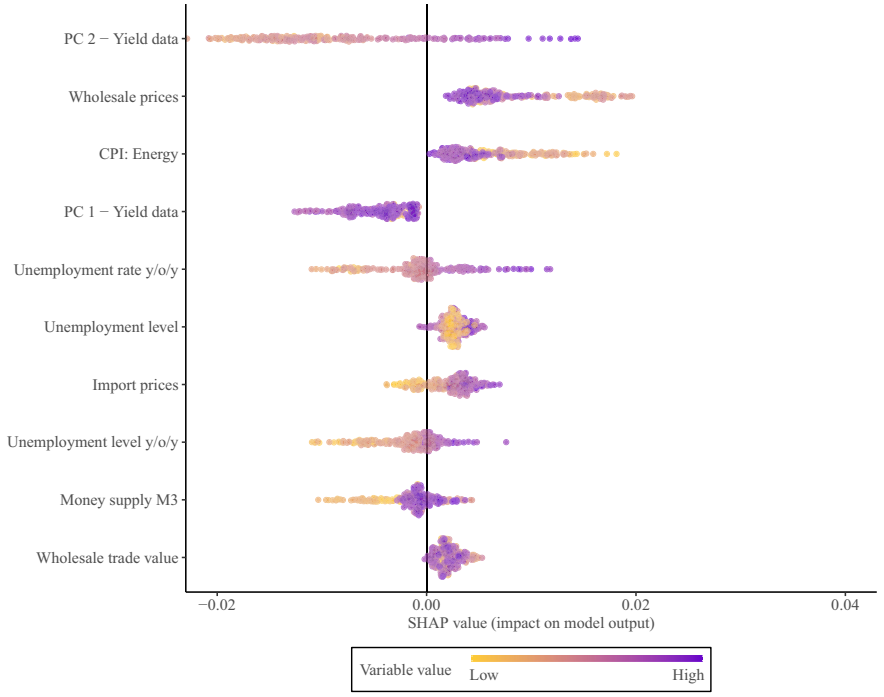

Fig. 7 Relationship between 10-year German bond excess returns and their predictors. This figure displays the SHAP values for the ten variables that are most important for predicting 10-year German bond excess returns. The SHAP values presented are obtained from an extremely randomized trees model predicting 10-year German bond excess returns

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availibility The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Bansal R, Shaliastovich I (2013) A long-run risks explanation of predictability puzzles in bond and currency markets. Rev Financ Stud 26(1):1–33

Barr DG, Priestley R (2004) Expected returns, risk and the integration of international bond markets. J Int Money Financ 23(1):71–97

Bartram SM, Lohre H, Pope PF, Ranganathan A (2021) Navigating the factor zoo around the world: An institutional investor perspective. J Bus Econ 91(5):655–703

Bauer MD, Hamilton JD (2018) Robust bond risk premia. Rev Financ Stud 31(2):399–448

Bauer MD, Rudebusch GD (2017) Resolving the spanning puzzle in macro-finance term structure models. Rev Financ 21(2):511–553

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(2):281–305

Bianchi D, Büchner M, Tamoni A (2021a) Bond risk premiums with machine learning. Rev Financ Stud 34(2):1046–1089

Bianchi D, Büchner M, Hoogteijling T, Tamoni A (2021b) Corrigendum: bond risk premiums with machine learning. Rev Financ Stud 34(2):1090–1103

Breiman L (1996) Bagging predictors. Mach Learn. 24(2):123–140

Breiman L (2001) Random forests. Mach Learn. 45(1):5–32

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth and Brooks/Cole Monterey, New York

Campbell JY (1995) Some lessons from the yield curve. J Econ Perspect 9(3):129–152

Campbell JY, Shiller RJ (1991) Yield spreads and interest rate movements: a bird's eye view. Rev Econ Stud 58(3):495–514

Campbell JY, Thompson SB (2008) Predicting excess stock returns out of sample: can anything beat the historical average? Rev Financ Stud 21(4):1509–1531

Clark TE, West KD (2007) Approximately normal tests for equal predictive accuracy in nested models. J Econom 138(1):291–311

Cochrane JH (2017) Macro-finance. Rev Financ 21(3):945–985

Cochrane JH, Piazzesi M (2005) Bond risk premia. Am Econ Rev 95(1):138–160

Cooper I, Priestley R (2009) Time-varying risk premiums and the output gap. Rev Financ Stud 22(7):2801–2833

Coroneo L, Giannone D, Modugno M (2016) Unspanned macroeconomic factors in the yield curve. J Bus Econ Stat 34(3):472–485

Diebold FX, Li C (2006) Forecasting the term structure of government bond yields. J Econom 130(2):337–364

Diebold FX, Rudebusch GD, Aruoba SB (2006) The macroeconomy and the yield curve: a dynamic latent factor approach. J Econom 131(1–2):309–338

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Mach Learn 40(2):139–157

Fama EF, Bliss RR (1987) The information in long-maturity forward rates. Am Econ Rev 77(4):680–692

Fan Y, Feng G, Fulop A, Li J (2022) Real-time macro information and bond return predictability: a weighted group deep learning approach. Working Paper

Gabaix X (2012) Variable rare disasters: an exactly solved framework for ten puzzles in macro-finance. Q J Econ 127(2):645–700

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63(1):3–42

Gu S, Kelly B, Xiu D (2020) Empirical asset pricing via machine learning. Rev Financ Stud 33(5):2223–2273

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, New York

Huang JZ, Shi Z (2023) Machine-learning-based return predictors and the spanning controversy in macro-finance. Manag Sci 69(3):1780–1804

Inoue A, Kilian L (2004) In-sample or out-of-sample tests of predictability: which one should we use? Econometr Rev 23(4):371–402

Ioannidis C, Ka K (2021) Economic policy uncertainty and bond risk premia. J Money, Credit, Bank 53(6):1479–1522

Joslin S, Priebsch M, Singleton KJ (2014) Risk premiums in dynamic term structure models with unspanned macro risks. J Financ 69(3):1197–1233
Kessler S, Scherer B (2009) Varying risk premia in international bond markets. J Bank Financ 33(8):1361–1375
Litterman R, Scheinkman J (1991) Common factors affecting bond returns. J Fixed Income 1(1):54–61
Liu Y, Wu JC (2021) Reconstructing the yield curve. J Financ Econ 142(3):1395–1425
Ludvigson SC, Ng S (2009) Macro factors in bond risk premia. Rev Financ Stud 22(12):5027–5067
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems, pp 4768–4777
McCallum JS (1975) The expected holding period return, uncertainty and the term structure of interest rates. J Financ 30(2):307–323
McCracken MW, Ng S (2016) Fred-md: a monthly database for macroeconomic research. J Bus Econ Stat 34(4):574–589
Pérignon C, Smith DR, Villa C (2007) Why common factors in international bond returns are not so common. J Int Money Financ 26(2):284–304
Piazzesi M, Schneider M, Tuzel S (2007) Housing, consumption and asset pricing. J Financ Econ 83(3):531–569
Probst P, Boulesteix AL (2017) To tune or not to tune the number of trees in random forest. J Mach Learn Res 18(1):6673–6690
Probst P, Wright MN, Boulesteix AL (2019) Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data mining and knowledge discovery. 9(3):1301
Sekkel R (2011) International evidence on bond risk premia. J Bank Financ 35(1):174–181
Shapley LS (1953) A value for n-person games. Princeton University Press, Princeton, pp 307–318
Stock JH, Watson MW (2002) Macroeconomic forecasting using diffusion indexes. J Bus Econ Stat 20(2):147–162
Stock JH, Watson M (2006) Forecasting with many predictors. Handb Econ Forecast 1:515–554
Thornton DL, Valente G (2012) Out-of-sample predictions of bond excess returns and forward rates: an asset allocation perspective. Rev Financ Stud 25(10):3141–3168
Wachter JA (2006) A consumption-based model of the term structure of interest rates. J Financ Econ 79(2):365–399
Wright JH (2011) Term premia and inflation uncertainty: empirical evidence from an international panel dataset. Am Econ Rev 101(4):1514–34