NBER WORKING PAPER SERIES

GLASS BOX MACHINE LEARNING AND CORPORATE BOND RETURNS

Sebastian Bell
Ali Kakhbod
Martin Lettau
Abdolreza Nazemi

Working Paper 33320
http://www.nber.org/papers/w33320

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2024, Revised July 2025

We thank Hui Chen, Will Cong, Alex Dickerson, Amir Kermani, Leonid Kogan, Ralph Koijen, Eben Lazarus, Dmitry Livdan, Ananth Madhavan, Asaf Manela, Tim McQuade, David Sraer, Richard Stanton, and conference participants at the NBER Summer Institute (Asset Pricing) for their helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Glass Box Machine Learning and Corporate Bond Returns
Sebastian Bell, Ali Kakhbod, Martin Lettau, and Abdolreza Nazemi
NBER Working Paper No. 33320
December 2024, Revised July 2025
JEL No. C45, C55, G11, G12

ABSTRACT

Machine learning methods in asset pricing are often criticized for their black box nature. We study this issue by predicting corporate bond returns using interpretable machine learning on a high-dimensional bond charac-teristics data set. We achieve state-of-the-art performance while maintaining an interpretable model structure, overcoming the accuracy-interpretability trade-off. The estimation uncovers nonlinear relationships and eco-nomically meaningful interactions in bond pricing, notably related to term structure and macroeconomic un-certainty. Subsample analysis reveals stronger sensitivities to these effects for small firms and long-maturity bonds. Finally, we demonstrate how interpretable models enhance transparency in portfolio construction by providing ex ante insights into portfolio composition.

Sebastian Bell
School of Economics and Management,
Karlsruhe Institute of Technology
sebastian.bell@kit.edu

Ali Kakhbod
University of California, Berkeley
Haas School of Business
akakhbod@berkeley.edu

Martin Lettau
University of California, Berkeley
Haas School of Business
and CEPR
and also NBER
lettau@haas.berkeley.edu

Abdolreza Nazemi
Karlsruhe Institute of Technology
nazemi@kit.edu

1 Introduction

Over the past decades, the asset pricing literature has identified various risk factors priced in securities markets. Most of this literature focuses on equity markets, but some recent studies have turned to the corporate bond market, resulting in a factor zoo for bond returns, as summarized by Dickerson et al. (2023a). Recently, machine learning methods have become popular in asset pricing. Such methods can be beneficial for applications in empirical asset pricing for two main reasons: Firstly, high-capacity machine learning models are designed to efficiently approximate complex functional relationships using nonlinearities and interactions between input features. Secondly, machine learning models are equipped to handle large numbers of predictors and high-dimensional datasets, which complements the numerous candidate factors the literature has identified as potential drivers of risk premiums.[1] Hence, machine learning methods can often reduce dimensionality and find a suitable selection of predictors.

However, one of the main problems associated with traditional machine learning methods is their "black box" character: ensembling or layering multiple learning systems makes it difficult to disentangle the contributions of individual input features to the output prediction, which leads to a lack of interpretability. These approaches sacrifice the transparency of simple linear regressions or decision trees for higher model capacity and forecast accuracy. The performance improvements of more complex models hint at the importance of nonlinear model specifications and interactions of predictors. However, how individual input features (or their interactions) affect the prediction remains largely unanswered.[2]

Understanding the reasoning behind predictions is critical for finance research and its practical applications, so we prioritize interpretability in our analysis. We, therefore, turn to "glass box" models that are inherently explainable due to their structure but retain much of the flexibility and scalability of other machine learning methods. Specifically, we employ Explainable Boosting Machine (EBM) models, which are an extension of Generalized Additive Models (GAM) with first-order interactions (see Hastie and Tibshirani (1986) and Nori et al. (2019)). We provide new insights into the relationships underlying return predictions by using the interpretable nature of the EBM model, which allows us to isolate the influence of individual features or first-order combinations of two features and analyze the nature of their impact. Contrary to prior research, which often assumes linearity, we find many nonlinear relationships and interactions that explain corporate bond returns, supplementing findings based on black box models, such as Gu et al. (2020), who observe but do not interpret such effects in their stock market analysis.

The paper makes the following contributions. We add to the return prediction literature by ap-

[1]See, for example, Gu et al. (2020), Bianchi et al. (2021), Feng et al. (2023), Murray et al. (2024), Chen et al. (2024) and Kakhbod et al. (2024).

[2]Rudin (2019) provides several reasons why model-agnostic methods that provide post-hoc explainability to black box models, such as SHAP, may not be suitable to answer this question. Crucially, these methods only provide an approximation rather than an explicit representation of the true prediction process and may be unreliable. While EBMs do not categorically eliminate all challenges of model explainability (Molnar et al., 2020), they are favored for their transparent, separable structure and avoidance of approximation errors.

plying interpretable machine learning to a new dataset of bond price predictors and identifying the most important factors. The performance of our model is superior to linear models (e.g., OLS and LASSO) and on par with other black box state-of-the-art techniques, such as Random Forests and XG-Boost[3], reaching an out-of-sample R^2 of over 12%. Our interpretable model allows us to answer how individual predictors affect corporate bond returns. We find that the most important explanatory variables are measures of macroeconomic uncertainty (see Jurado et al. (2015) and Ludvigson et al. (2021)) and a term structure factor (Fama and French (1993)), followed by firm-level yield spreads. Moreover, the effects of many variables are highly nonlinear. For example, large positive realizations of the term structure factor (i.e., returns of long-term government bonds exceed those of short-term Treasury bills) predict a 100 basis point return increase of corporate bonds. This effect is exclusive to the right tail of the factor distribution and not matched by a corresponding return decrease for large negative values, indicating an asymmetric reaction of corporate bond returns to interest rate change risk.

In addition, we find a negative association between changes in macroeconomic uncertainty and returns, which takes the form of a step function. Its main impact can be tracked to the right tail of the distribution, suggesting an asymmetric response to changes in uncertainty, as the market reacts more strongly to rising uncertainty than to falling uncertainty. Another influential uncertainty factor, financial uncertainty, affects returns through a monotonically decreasing function. The difference in returns resulting from the highest and lowest values of this factor amounts to nearly 130 basis points.

We also uncover heterogeneity for some of these effects between different firms. By sorting the data into subsamples based on characteristics, we find that bonds of small firms and long-maturity bonds are more sensitive to changes in financial uncertainty than big and shorter-maturity ones. We also find stronger return reactions for more risky bonds characterized by higher spreads and lower credit ratings. Our model identifies asymmetric differences that are most pronounced for large increases in uncertainty.

The EBM model can also investigate how input variables *interact* to produce return predictions. Among the most important terms is the interaction of the financial uncertainty factor with bond spreads, which induces an additional return premium associated with falling uncertainty exclusive to high-spread bonds. In comparison, the additional return decrease resulting from the interaction in times of rising financial uncertainty is smaller in absolute value and applies more uniformly across spread values. For bonds with low spreads, the impact of the interaction term is practically independent of the financial uncertainty factor. This suggests that the interaction term mainly represents an outsized price recovery for the most risky bonds as financial uncertainty diminishes.

Further analysis of important interactions reveals a similar pattern for financial uncertainty and the spread-times-rating characteristic. When financial uncertainty falls, the interaction term increases returns disproportionally for low-grade bonds with high spreads over the past six months, while rising uncertainty results in a more homogeneous effect across all bonds.

[3]See Ho (1995) and Chen and Guestrin (2016), respectively.

We show how the benefits of the glass box approach translate to portfolio construction. Based on the interpretable model's return predictions, we build a monthly long-short portfolio with a higher Sharpe ratio than other benchmark portfolios. By analyzing the most important characteristics and interactions regarding their impact on out-of-sample model fit and portfolio Sharpe ratio, we identify our model's most influential functional relationships for portfolio selection. This includes bond spread, yield, and the abovementioned interactions with the financial uncertainty factor. The observed functions allow us to form *ex-ante* expectations regarding the characteristics of the long and short legs. We confirm that these expectations are reflected in the EBM portfolios, which combine strong risk-adjusted performance with a transparent composition.

This paper contributes to the emerging literature applying novel methods in asset pricing. Large parts of the field rely on linear regression models to characterize the relationship between factors and returns and to estimate premiums associated with different risk factors. However, in addition to the traditional asset pricing literature, several recent studies have explored new econometric and machine learning techniques to explain asset returns. Gu et al. (2020) examine future stock returns and find large economic gains for investors using machine learning forecasts. Machine learning models outperform traditional statistical methods, which the authors attribute to allowing nonlinearities and predictor interactions. However, the precise nature of these effects remains unclear due to the complex structure of their most potent models. These include machine learning methods which are commonly applied throughout this literature, such as neural networks (e.g., Gu et al. (2021), Fallahgoul et al. (2024), Chen et al. (2024)), tree-based models (e.g., Lopez-Lira and Roussanov (2020), Bryzgalova et al. (2023), Binsbergen et al. (2023), Cong et al. (2024)) and generative models (e.g., Chen et al. (2022), Fedyk et al. (2024)). Bali et al. (2022a) apply similar techniques to a dataset of corporate bond returns and firm characteristics.

Another popular approach utilizes variants of principal component analysis (PCA) to estimate latent factors in high-dimensional data (e.g., Kelly et al. (2019), Lettau and Pelger (2020), Kelly et al. (2023), Lettau (2023)). Alternative approaches to PCA add regularization to linear models, resulting in regression variants, such as LASSO (Feng et al. (2020)), adaptive group LASSO (Freyberger et al. (2020)), or elastic net (Kozak et al. (2020)).[4] In contrast to the previously discussed papers, we prioritize interpretability in our analysis. We can show when and where specific factors are important and how individual factors affect predictions using glass box models.[5]

Our work is also related to several asset pricing studies investigating bond return predictors. Many studies in the traditional asset pricing literature follow the seminal work of Fama and French (1993). Gebhardt et al. (2005) study the explanatory power of factor loadings and bond characteris-

[4]See Nagel (2021), Kelly and Xiu (2023) and Eisfeldt and Schubert (2024) for excellent reviews of machine learning applications in finance.

[5]There are other methods that aim to make black box models explainable using post-hoc explainability tools (e.g., Cong et al. (2022), Griffin et al. (2023), Bali et al. (2023)). While these approaches rely on approximations of the black box, our model, in contrast, achieves inherent interpretability through its separable structure. Separable models (GAMs and their variants) have also recently been applied in other areas: e.g., in criminal recidivism (Wang et al. (2023)), genetic anomalies (Wagner et al. (2023)), recovery rates (Nazemi and Fabozzi (2024)), and obesity rates (Kiss et al. (2024)). See Molnar (2020) for an excellent review of interpretable methods in machine learning.

tics. Significant pricing effects have also been documented for size, credit risk, value, and momentum (Houweling and van Zundert (2017)), liquidity (Lin et al. (2011)), long-term term reversal (Bali et al. (2021a)), and macroeconomic uncertainty (Bali et al. (2021b)). Dang et al. (2023) additionally emphasize the role of factors related to carry, duration, and equity momentum. Using a Bayesian approach, Dickerson et al. (2023a) identify factors referencing the bond post-earnings announcement drift, inflation risk, and equity-market factors as likely components of the bond pricing kernel. We test the predictive quality of many of these variables by building a large dataset combining bond and firm characteristics with risk factors and macroeconomic variables from bond and equity markets. In addition to the Fama and French (1993) term structure factor, we find that changes in uncertainty rank among the most important predictors in our dataset. These factors are derived from the uncertainty indices defined by Jurado et al. (2015) and Ludvigson et al. (2021), who estimate macroeconomic uncertainty by calculating the conditional volatility in the unforecastable component of sets of economic indicators and aggregating them into a single measure. Using our interpretable method, we discover an asymmetric market reaction to uncertainty changes and heterogeneous effects across subsamples.

The rest of the paper is organized as follows. Section 2 describes the prediction methodology and the machine learning models. Section 3 provides information on the data and pre-processing steps. Section 4 presents the empirical results of the prediction procedure, including a detailed description of the functional relationships associated with the most important predictors found by the glass box model. Section 5 provides information on the benefits of our model for portfolio construction. Section 6 compares the glass box model with a popular post-hoc approach based on Shapley values. Section 7 extends this comparison to simulated data. Section 8 concludes.

2 Methodology

Suppose we want to estimate the relationship between future excess bond returns $r_{l,t+1}$ of firm $l = 1, ..., L$ and information available at time t:

$$r_{l,t+1} = g(\mathbf{z}_{l,t}, \mathbf{z}_t) + e_{l,t+1}, \tag{1}$$

where $\mathbf{z}_{l,t}$ is a vector of bond and firm characteristics for firm l and $\mathbf{z}_t$ is a vector of risk factors and other variables derived from macroeconomic data and bond or stock market characteristics. Let $\mathbf{x}_{l,t} = (\mathbf{z}_{l,t}, \mathbf{z}_t)'$ be the vector of all N explanatory variables, so $x_{i,l,t}$ is variable i of firm l at time t. We follow the assumptions of Jurado et al. (2015) regarding the information set in order to form the most historically accurate estimates of explanatory variables.

We begin by employing traditional methods, such as OLS and LASSO regressions, and more complex "black box" models like Random Forests and XGBoost (described in detail by, e.g., Gu et al. (2020)). These are used as performance benchmarks, as we subsequently focus on interpretable models that allow us to examine the learned relationships between returns and inputs.

2.1 EBM: An interpretable machine learning model

Our goal is to model bond returns using a flexible model that retains the interpretability of the results. The model is based on the Generalized Additive Model (GAM) introduced by Hastie and Tibshirani (1986) that is additive in nonlinear functions of individual explanatory variables. Lou et al. (2013) generalize GAMs by adding nonlinear functions of selected interacted variable pairs:

$$r_{l,t+1} = \beta + \sum_{i=1}^{N} f_i(x_{i,l,t}) + \sum_{i>j} f_{ij}(x_{i,l,t}, x_{j,l,t}) + e_{l,t+1}. \tag{2}$$

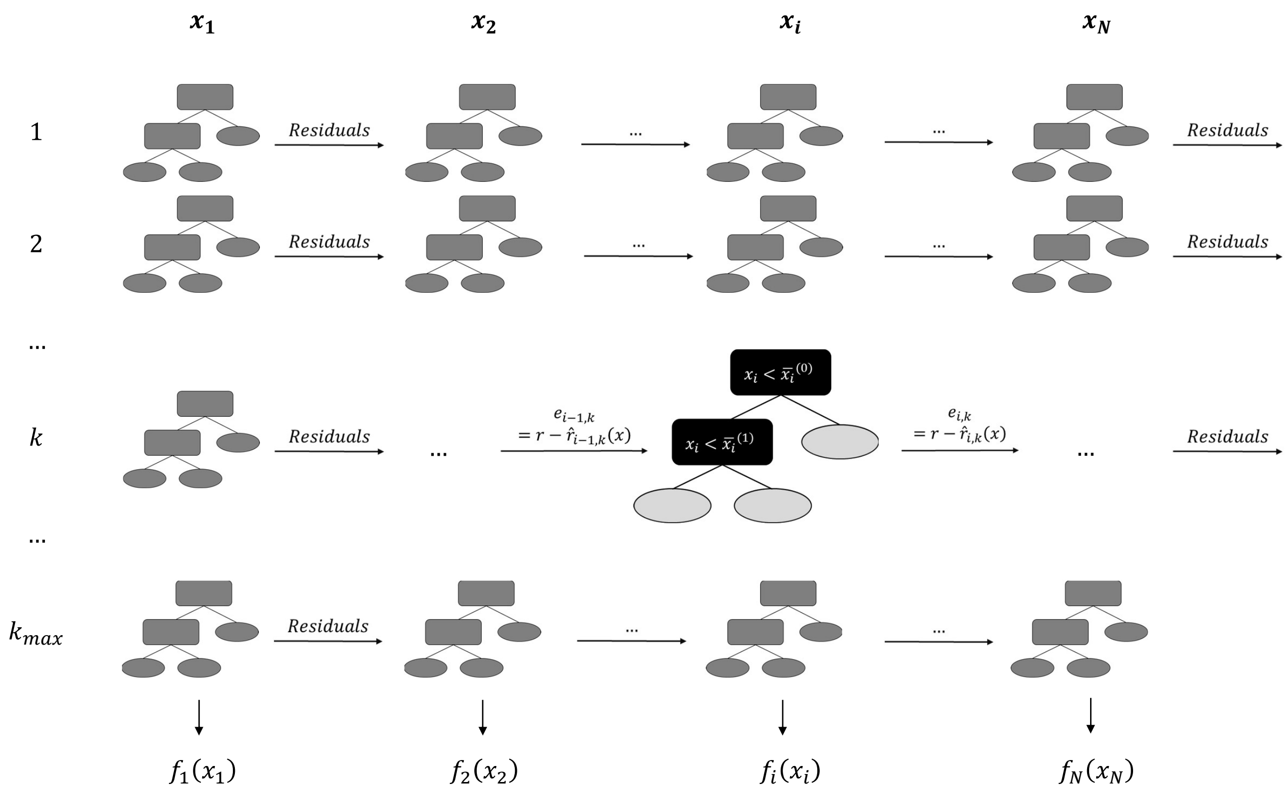
The functions f_i and f_{ij} define unknown univariate and multivariate "shape" functions, respectively, that capture the effects of explanatory variables on bond returns. While GAMs with few regressors and without interaction terms can be estimated using classical econometric methods (e.g., backfitting or maximum likelihood), recent machine learning methods allow for larger and more complex models. To this end, we use the Explainable Boosting Machine (EBM) by Nori et al. (2019) to estimate (2). While this method uses machine learning tools, such as decision trees, bagging, and boosting for the estimation of the f_i and f_{ij} functions, its results remain interpretable. For example, the additive model structure allows us to separate the influence of individual input variables and analyze each predictor independently, as we explain below.

The EBM algorithm works as follows (see Online Appendix O-A.1 for more details). Each f_i and f_{ij} takes the form of a step function over bins and is modeled as an ensemble of decision trees. The objective is to estimate the implied collection of decision trees for $i, j = 1, ..., N$ that minimizes the sum of squares of the residuals $e_{l,t+1}$. Starting from an initial guess of f_i and f_{ij} (usually $f_i = f_{ij} = 0$), the algorithm first cycles through all univariate variables $x_i, i = 1, ..., N$, sequentially, fitting a decision tree, and updating residuals after each tree is built. This process is usually repeated thousands of times and ends after a predefined maximum number of iterations or when performance does not improve significantly for multiple iterations, as defined by an early stopping threshold. Standard errors for each element of f_i are computed using the bootstrapped data from the bagged model. Figure 1 shows a conceptual visualization of this core algorithm.

Following the estimation of the univariate f_i, the EBM algorithm identifies and estimates the most important variable interactions , using a time-efficient method to detect pairs of variables that improve the fit the most (see Online Appendix O-A.2). Shape functions for interaction terms f_{ij} are step functions over a 2-dimensional grid defined over variables i and j and are also modeled as decision trees.

The construction of each decision tree uses familiar ensemble techniques from machine learning. Ensemble methods combine multiple trees to achieve better predictive performance than a single tree. The core idea is that a group of poor models can collectively form a better model. Two primary techniques for creating these ensembles are bagging and boosting. Bagging, short for Bootstrap Aggregation, aims to reduce variance in decision trees. It creates multiple subsets of the original data through random sampling with replacement, then trains a decision tree on each subset. The final

Figure 1: EBM algorithm



Note: This figure visualizes the core algorithm underlying the shape functions of the Explainable Boosting Machine. In each iteration k ($k = 1, ..., k_{max}$), the algorithm cycles through all features x_i ($i = 1, ..., N$), fitting a shallow decision tree and updating residuals $e_{i,k} = r - \hat{r}_{i,k}$ based on the current model predictions $\hat{r}(x) = \beta + \sum_{i'=1}^{N} f_{i'}^{(i,k)}(x_{i'})$ after each tree is built. Finally, all decision trees associated with x_i are combined into the final shape function $f_i = f_i^{(N,k_{max})}$. This algorithm is initially used to generate univariate functions and subsequently extended to generate multivariate functions f_{ij} by cycling through selected pairs of features (x_i, x_j). Additional details can be found in Online Appendix O-A.1.

prediction averages all individual tree predictions, resulting in a more robust model. Boosting, on the other hand, trains models sequentially. Each new model in the sequence focuses on correcting errors made by previous models. When an input is misclassified, its importance increases for subsequent models. Gradient Boosting combines gradient descent optimization with boosting principles. It builds trees one by one, with each new tree attempting to minimize the errors of the previous ensemble.

The full sample is divided into a training sample $\mathcal{T}_1$, a validation sample $\mathcal{T}_2$, and a testing sample $\mathcal{T}_3$. First, the model is estimated for combinations of hyperparameters in the training sample. Given the estimated functions f_i and f_{ij}, the sum of squared residuals is calculated for all combinations of hyperparameters in the validation sample. The combination with the smallest errors in the validation sample is chosen and evaluated in the testing sample.

When predicting new data, the shape functions are lookup tables that provide term contributions associated with input values for univariate features and selected interaction terms. These contributions are measured in the same unit as the target variable and can be interpreted as the local incremental influence of a term on the output value. The overall prediction for an individual data point is finally computed as the sum of the model intercept and the contribution values retrieved from all term functions. The additive nature of the model allows the contributions of individual input variables or combinations to be separated. Therefore, each shape function can be interpreted individually.

2.2 Model evaluation

Since the estimation minimizes the sum of squared residuals of (2), we focus on the R^2 as the main performance measure to evaluate the predictive capabilities of different models. Let $\mathcal{T} \in (\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3)$ be the training, validation, or out-of-sample testing subsamples. The R^2 in sample $\mathcal{T}$ is given by

$$R^2_{\mathcal{T}} = 1 - \frac{\sum_{(l,t)\in\mathcal{T}} e^2_{l,t}}{\sum_{(l,t)\in\mathcal{T}} r^2_{l,t+1}}. \tag{3}$$

The same performance measure is also applied in the corporate bond market by Kelly et al. (2023). It implicitly benchmarks model predictions against a naive forecast of zero for future excess returns by not demeaning in the denominator of our R^2 expression. This avoids overestimating model performance by comparing it to a poor prediction benchmark (Gu et al., 2020).[6] When comparing the EBM estimation of (2) to other estimation methods, we use the R^2 in the out-of-sample testing sample $\mathcal{T}_3$ as an evaluation criterion.

2.3 Effects of individual variables

In most machine learning methods, isolating how a given variable affects the model's prediction is not straightforward. In contrast, (2) has an additive structure so that the effect of a variable

[6] Note that the choice of benchmark does not affect any qualitative conclusions about, for example, the ordering of models during parameter tuning or final performance evaluation.

on the forecast can be seen directly from the shape functions f_i and f_{ij}. We consider two such measures to evaluate a variable's overall importance. First, we use the shape of the f_i functions, which control the marginal impact of changes in x_i while holding all other variables constant. We define the *mean absolute score* S_i as the absolute values of $f_i(x_{i,l,t})$ averaged over the training sample while accounting for sample weights $w_{l,t}$:

$$S(i) = \sum_{(l,t)\in\mathcal{T}_1} w_{l,t}\,|f_i(x_{i,l,t})|\,, \tag{4}$$

$S(i)$ captures the average effect of variable i, taking the distribution over the sample into account, and is therefore similar to the absolute value of a (standardized) coefficient in a linear regression. Hence, the mean absolute scores capture the average "size" of the effect of an individual variable on the prediction of the model. The mean absolute scores of the interaction terms are defined analogously:

$$S(i,j) = \sum_{(l,t)\in\mathcal{T}_1} w_{l,t}\left|f_{ij}(x_{i,l,t}, x_{j,l,t})\right|. \tag{5}$$

An alternative measure of the importance of a variable is its effect on the overall fit. To this end, we compute the change in the R^2 when a variable is removed from the model. We calculate the residuals when the univariate effect of variable i is eliminated from (2),

$$e^{-i}_{l,t+1} = r_{l,t+1} - \left(\beta + \sum_{i'\neq i} f_{i'}(x_{i',l,t}) + \sum_{i\neq j} f_{ij}(x_{i,l,t}, x_{j,l,t})\right), \tag{6}$$

and compute the corresponding R^2_{-i}. The effect of variable i on the explanatory power is the reduction relative to the R^2 of the full model: $\Delta R^2_i = R^2_{-i} - R^2$. For the joint effect of variables i and j, we compute

$$e^{-ij}_{l,t+1} = r_{l,t+1} - \left(\beta + \sum_{i'} f_{i'}(x_{i',l,t}) + \sum_{(i',j')\neq(i,j)} f_{i'j'}(x_{i',l,t}, x_{j',l,t})\right), \tag{7}$$

and the associated $\Delta R^2_{ij} = R^2_{-ij} - R^2$.

ΔR^2_i and ΔR^2_{ij} measure the reduction in the fit when a single term in (2) is removed. To capture the total effect of a variable on the fit of the model, we remove not only the univariate term f_i but also *all* interaction terms that involve x_i:

$$\tilde{e}^{-i}_{l,t+1} = r_{l,t+1} - \left(\beta + \sum_{i'\neq i} f_{i'}(x_{i',l,t}) + \sum_{i'\neq i,j} f_{i'j}(x_{i',l,t}, x_{j,l,t})\right), \tag{8}$$

and define $\Delta\tilde{R}^2_i = \tilde{R}^2_{-i} - R^2$.

3 Data

We combine bond and firm characteristics data with time series data related to many factors and macroeconomic measures to form a novel dataset. Kelly et al. (2023) provide monthly data on duration-adjusted bond returns and characteristics from Trade Reporting and Compliance Engine via Wharton Research Data Services (WRDS/TRACE), which we supplement with equity data from CRSP to impute missing market capitalization data.[7] We construct firm-level value-weighted bond returns and characteristics using the "amount outstanding" of each bond as weight.[8] We combine this dataset with several traded and non-traded bond and equity variables that have seen previous use in the corporate bond literature. Our data set includes 41 firm-level variables, which we merge with 40 market-level variables. We obtain the five Fama and French (2015) factors and the momentum factor from Kenneth French's data library. We add several variables which are available from the authors' websites, including the intermediary capital non-traded risk factor of He et al. (2017), Pástor and Stambaugh (2003)'s liquidity factors, and Jurado et al. (2015)'s macroeconomic uncertainty indices. Data on traded corporate bond factors as well as a bond momentum measure is provided by Dickerson et al. (2023a) via Open Source Bond Asset Pricing.[9] Economic policy uncertainty indices and VIX and CPI data are obtained from FRED. We use lower cases for firm-level variables and upper cases for market-level predictors. Variables that exhibit a trend over the sample period are divided by their monthly mean, and variables censored at zero are transformed using the natural logarithm. Finally, we rescale each feature in our dataset to a mean of zero and a standard deviation of one. A complete overview of the set of characteristics and risk factors is given in Table A.1 of the Appendix.

The dataset includes 81 predictors from July 2002 to August 2020 (218 months) of 1,207 individual firms for a total of 106,265 firm-date observations.[10] Figure A.1 shows the behavior of some bond characteristics over the sample. The number of firms per month, plotted in Panel A, rose from 311 at the start of the sample to 400 before declining to 330 during the financial crisis. Starting in 2009, the number of firms increases steadily to 624 in 2014 followed by a decline to 534 at the end of the sample. The mean rating of all firm-level bonds is BBB-. Since ratings are correlated with firm size, the average rating when firms are weighted by their market cap is higher and close to A. About 56% of the bonds in the sample are investment-grade. The median duration of all bonds in the sample is 5.6 years. However, durations are increasing over the sample as shown in Panel B. The median duration at the end of sample is 6 years compared to 4.75 years at the start. The effect is stronger for the upper part of the distribution as the 90% percentile has almost doubled from 9 years to 16 years. In other words, there are some firms that started to issue bonds with very long maturities in the second half of the sample.

Panels C and D of Figure A.1 plot the monthly median yield and yield spread along with the respective 10% to 90% percentile intervals. Bond yields range between 4% and 8% during the first part of the sample but spike during the financial crisis, when the median bond yield reached almost 9%. As interest rates declined, yields of corporate bonds fell as well and ranged from around 2%

[7]In TRACE-based data, price-related variables are subject to 'microstructure market noise' (MMN), which can be mitigated with implementation gaps for the signals (see, e.g., Bartram et al. (2024)). In Appendix E, we confirm qualitatively similar findings using MMN-adjusted data from Open Source Bond Asset Pricing (Dickerson et al., 2023c).

[8]The raw bond-level data set includes 8,758 individual non-convertible bonds with an average of 7.25 bonds per firm. About one-third of firms have a single bond, and 75% have five bonds or fewer in a given month. Estimation results using the bond-level data are similar to those using firm-level data.

[9]We use this momentum measure because it skips the prior month, consistent with Jostova et al. (2013).

[10]Summary statistics of the variables are in Table O-B.1 in the Online Appendix.

Table 1: Descriptive statistics of excess returns

	Mean	S.D.	5% pct.	95% pct.
Mean of $r_{l,t}$ by firm	0.25	0.22	-0.00	0.64
S.D. of $r_{l,t}$ by firm	2.18	1.08	0.72	4.21
SR of $r_{l,t}$ by firm	0.13	0.10	-0.00	0.31

Note: This table reports means, standard deviations, and the 5% and 95% percentiles of the distribution of moments of excess bond returns. The rows show the distributions of the means, standard deviations, and Sharpe ratios of excess returns by firm. Firms with less than 24 observations in the sample are excluded. The sample is from July 2002 to August 2020.

for investment-grade bonds to 6% for non-investment-grade bonds. Following Israel et al. (2018), we construct yield spreads ("spread") as the duration-matched option-adjusted spread over the Treasury curve. For most of the sample, spreads were between 1% and 6% but reached over 10% during the financial crisis. Note that unlike during the financial crisis, corporate bond yields rose only slightly during the COVID period since higher spreads were offset by lower Treasury rates.

Table 1 reports descriptive statistics of our sample's excess returns of bonds.[11] The first row shows the distribution of means across all bonds with more than 24 observations in the sample. The grand mean is 0.25% per month, but the distribution of mean returns is spread out. The standard deviation is 0.22%, and the 90% confidence interval ranges from 0 to 0.64%. The mean standard deviation is 2.18% with a standard deviation of 1.08%. The average monthly Sharpe ratio is 0.13, or 0.36 annually, but some bonds have monthly Sharpe ratios above 0.3 (0.56 annually).

The distribution of bond returns varies substantially over the sample. On average, bond prices fell at the beginning of the financial crisis in 2008 and COVID in 2020. However, bond returns were high when the economy recovered from these episodes. The median returns were -9.28% and -6.38% in February 2020 and September 2008, respectively. The highest returns occurred in March 2009 (4.50%), December 2008 (4.38%), and March 2020 (4.14%).

4 Empirical results

This section reports the results of the estimation of the EBM model. We first compare its fit to several benchmark models used in the literature. We consider linear specifications (OLS, LASSO) and non-linear machine learning methods (Random Forest and XGBoost). Next, we exploit the interpretable structure of EBM to investigate the roles of individual variables and their interactions. In the main estimation, we weigh firm observations by their market capitalization to reduce the impact of small firms. However, the results for the estimation with equal weights are similar and are reported in section C of the Appendix.

[11]We use the winsorized return variable, but our findings are materially unchanged when using unwinsorized returns. The results are practically identical when we predict the duration-adjusted returns provided via Open Source Bond Asset Pricing or the duration-adjusted returns of Binsbergen et al. (2024).

4.1 Model comparison

The data spans July 2002 to August 2020 (218 months) and is split into three subsamples: training data $\mathcal{T}_{\text{train}}$ to fit models, validation data $\mathcal{T}_{\text{valid}}$ to tune hyperparameters, and out-of-sample test data $\mathcal{T}_{\text{oos}}$ to evaluate out-of-sample performance.[12] A standard split for training and test samples in machine learning is around 80% of the data for training/validation and 20% for testing, which implies a short period of less than two years for the out-of-sample analysis. Instead, we choose a relatively long testing window at the expense of a short training sample: $\mathcal{T}_{\text{train}}$ consists of the first eight years of data (7/2002-6/2010), $\mathcal{T}_{\text{valid}}$ contains data from the following year (7/2010-6/2011), while the remaining data is reserved for out-of-sample testing (7/2011-8/2020). The longer out-of-sample window allows us to assess the estimation results over a full economic cycle, including the COVID period. The drawback is that the estimation is based on fewer observations and is thus more noisy. However, we explore alternative data splits (e.g., a longer training window of 14 years) and rolling windows in the Online Appendices O-C.1 and O-C.2 with similar results. We impute missing values using the cross-sectional median within each month for linear regression and LASSO. We impute the value zero if characteristic values are unavailable for all firms in a month. The predictive performance of all models on subsamples $\mathcal{T}_j$ is summarized in Table 2.

Given its negative R^2_{valid} and R^2_{oos}, the OLS model provides poor predictions out-of-sample and is comfortably outperformed by a naive forecast of 0. Due to the linear model's limitations and our dataset's high dimensionality, such results are expected and broadly in line with other works from the return prediction literature. Performance improves significantly when regularization is introduced via the l_1-penalty in the LASSO specification as the R^2 in the test sample is 7.99%.[13] The nonlinear machine learning models further enhance in-sample and out-of-sample fit. The Random Forest, XGBoost, and EBM models achieve significant predictive performance, reaching R^2 results of around 11% to 13% on the test data. Even though EBM is more restrictive, it yields comparable out-of-sample fits as state-of-the-art machine learning methods. Note that the training sample R^2_{train} of the Random Forest and XGBoost estimations are substantially higher than that of the EBM model. This indicates that Random Forest and XGBoost are prone to overfitting in the training sample, which is penalized in the testing sample.

Table 2: Comparison of R^2's

	OLS	LASSO	R-Forest	XGBoost	EBM
R^2_{train}	31.15%	17.51%	46.93%	42.52%	24.79%
R^2_{valid}	-0.27%	15.94%	15.57%	21.80%	18.54%
R^2_{oos}	-2.34%	7.99%	13.34%	11.97%	12.10%

Note: This table shows R^2's of OLS, LASSO, Random Forest, XGBoost, and EBM estimations in the training (July 2002 – June 2010), validation (July 2010 – June 2011), and testing (July 2011 – August 2020) samples.

[12] An overview of the hyperparameter values used for validation are reported in Table B.1 in the Appendix.

[13] Other regularized linear models, such as Ridge or Elastic Net, yield similar results.

Table 3 reports descriptive statistics and correlations of the models' errors. The mean forecast error of OLS is -0.62, implying that the model systematically overestimates returns. Mean errors of LASSO, Random Forest, and XGBoost are positive, and EBM yields forecasts that are, on average, closest to observed returns. Errors of all models are negatively skewed and leptokurtic. Aside from OLS, the errors are highly correlated, suggesting that the models produce similar forecasts that differ in their precisions.

Table 3: Model errors

	OLS	LASSO	R-Forest	XGBoost	EBM
Panel A: Descriptive Statistics					
Mean	-0.62	0.03	0.08	0.11	0.02
S.D.	2.15	2.12	2.06	2.07	2.08
Skewness	-1.64	-0.89	-0.60	-0.77	-0.96
Kurtosis	18.19	21.76	20.12	20.41	21.82
Panel B: Correlations					
OLS	1.00	0.91	0.88	0.88	0.90
LASSO	0.91	1.00	0.98	0.96	0.99
R-Forest	0.88	0.98	1.00	0.97	0.99
XGBoost	0.88	0.96	0.97	1.00	0.97
EBM	0.90	0.99	0.99	0.97	1.00

Note: This table shows summary statistics (Panel A) and correlations (Panel B) for prediction errors of OLS, LASSO, Random Forest, XGBoost, and EBM estimations in the test sample (July 2011 – August 2020).

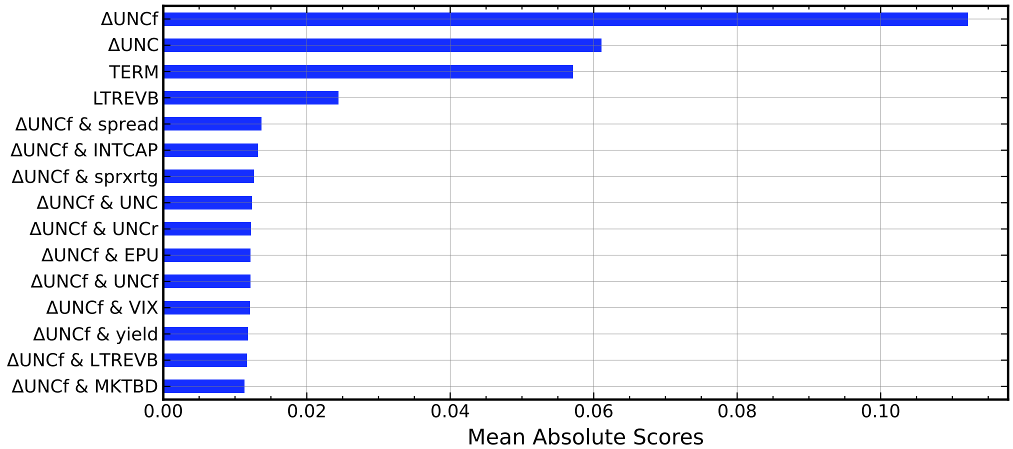
4.2 Predictor importance

Figure 2 shows the most important predictors. Panel A plots the 15 variables with the highest mean absolute scores $S(i)$ and $S(i,j)$ as defined in (4) and (5). Panels B and C plot the marginal effects of excluding a predictor on the R^2. ΔR_i^2 indicates the univariate effect when variable i is excluded and $\Delta\tilde{R}_i^2$ measures the total effect when all interaction effects involving x_i are also removed.

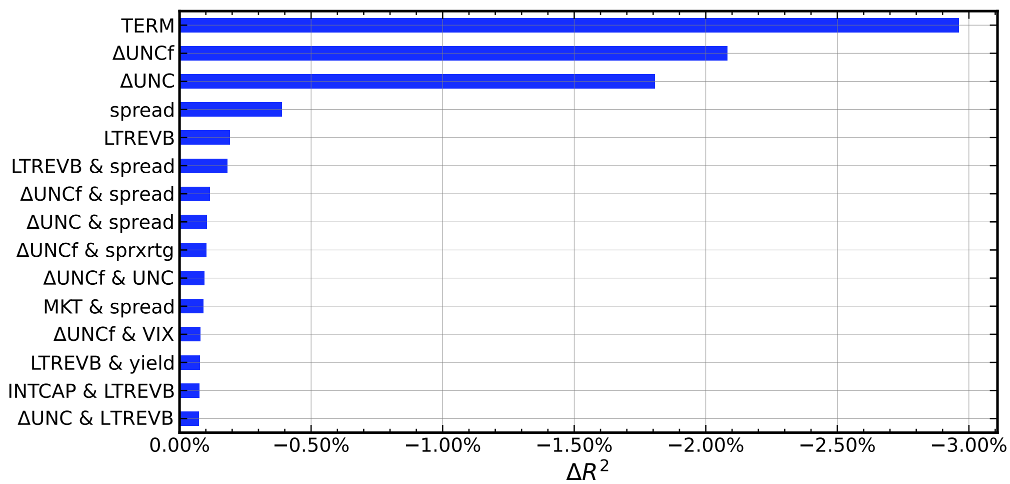
The highest mean absolute scores are associated with macroeconomic and financial uncertainty variables. ΔUNC and ΔUNCf are the first differences of indices constructed by Jurado et al. (2015) and Ludvigson et al. (2021), who estimate macroeconomic and financial uncertainty by calculating the conditional volatility in the unforecastable component of sets of economic indicators aggregated into a single measure. The financial uncertainty measure ΔUNCf is particularly relevant for corporate bond returns. In contrast, the uncertainty index for the real economy (UNCr) has only the 23rd highest score and plays a lesser role than ΔUNCf and ΔUNÇ. The third and fourth most important

Figure 2: Importance of predictors

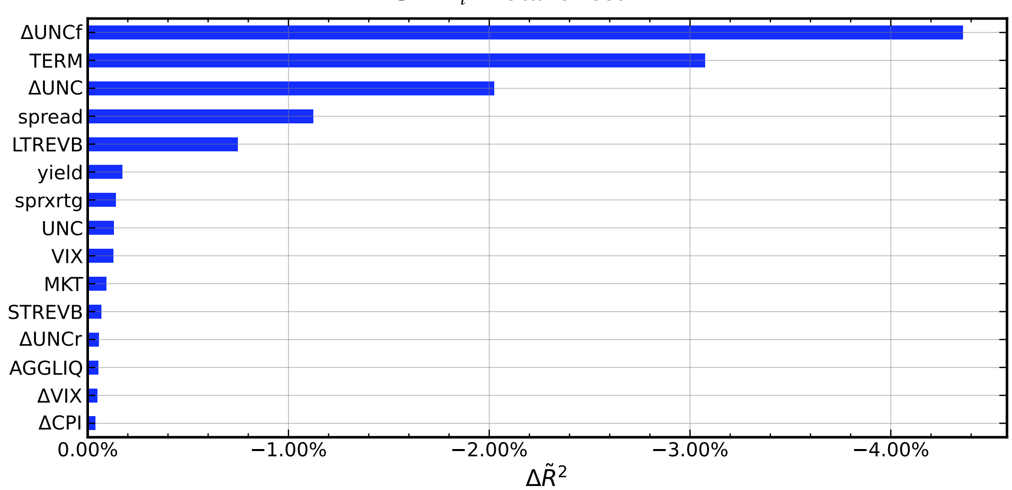
A: Mean Absolute Scores: $S(i)$ and $S(i, j)$



B: ΔR_i^2: Univariate effect



C: $\Delta \widetilde{R}_i^2$: Total effect



Note: This figure shows the most important predictors for the EBM model. In Panel A, importance values are mean absolute scores calculated on the training data (the native EBM importance measure). In Panel B, importance values are calculated as reductions in R^2 when removing a term, i.e., an individual shape function, from the fitted model. In Panel C, importance values are calculated as reductions in R^2 when removing all terms associated with a variable from the fitted model.

variables are the term structure factor TERM and long-term bond reversals LTREVB[14]. While the four most important predictors are univariate, they are followed by interaction terms, such as the interaction of ΔUNCf and the bond yield spread ('ΔUNCf & spread'). The high importance of uncertainty factors is underlined by the multitude of interaction terms within the rankings that contain different uncertainty factors. These findings align with Bali et al. (2021b), who find significant return premiums associated with macroeconomic uncertainty. This observation also mirrors prior evidence that macroeconomic climate and the business cycle are highly relevant determinants of bond risk premiums and returns. Both theoretical predictions, such as general equilibrium model-based estimates by Gomes and Schmid (2021), and empirical findings, e.g., by Elkamhi et al. (2024), suggest that macroeconomic fluctuations possess significant explanatory power for bond prices. Note that the EBM model implies that *changes* in uncertainty are more relevant for future returns than their levels. We analyze the role of uncertainty in more detail in section 4.4.

Recall that the mean absolute score measures the average "size" of the effect of a variable on bond return. Panels B and C show the importance of variables in the model's fit as measured by the R^2. Panel B plots ΔR_i^2 and ΔR_{ij}^2, which indicate the reduction in the R^2 when variable x_i or the interaction effect of x_i and x_j are removed from the model. The pattern is similar to that in Panel A. The three most relevant variables are TERM, ΔUNCf, and ΔUNC. Excluding TERM from the model has a substantial effect on the fit of the model as the R^2 is reduced by 2.97% (12.10% to 9.13%), which is about one-quarter of the R^2 of the full model. For ΔUNCf and ΔUNC, the ΔR_i^2 are -2.08% and -1.81%, respectively. The fourth most important variable is the firm-level bond yield spread. Several interaction terms involving uncertainty and the yield spread also have meaningful effects.

Panel C shows the effect on R^2's when interaction effects are removed from the model. Even though TERM was associated with the largest univariate decrease in the R^2, ΔUNCf has a larger effect when interaction terms are excluded: -4.36% vs. -3.08%. The reason is that interaction terms involving ΔUNCf are also relevant for the fit. Similarly, the univariate ΔR_i^2 of the yield spread is only -0.39%, but since interaction effects are important, the total $\Delta \widetilde{R}_i^2$ is -1.12%.

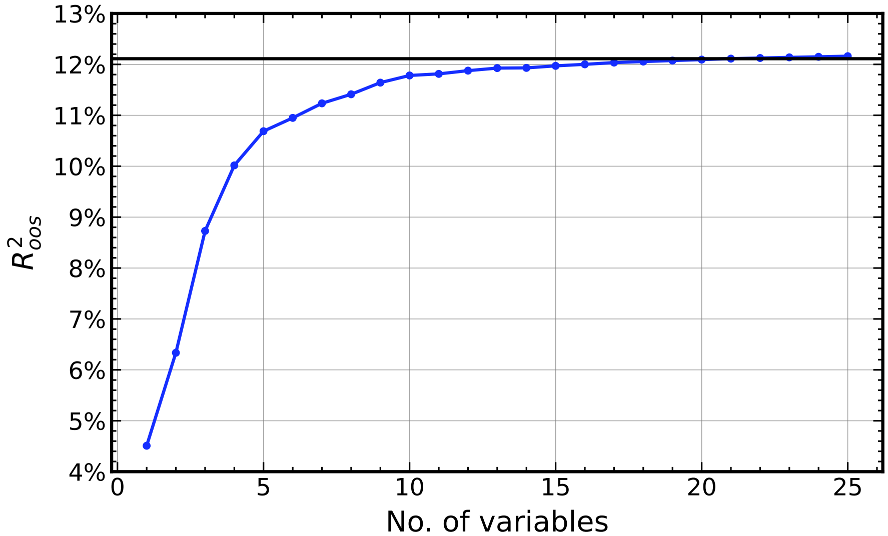
Both measures suggest that a few variables account for a large part of the full model's predictive strength. This is confirmed by the results displayed in Figure 3, where we track R^2_{oos} when adding terms to the model successively ranked by $\Delta \widetilde{R}_i^2$. The R^2 with only the first variable, ΔUNCf, is 4.51%. Adding the next three variables, TERM, ΔUNC, and spread, doubles the R^2_{oos} to 10.01%, while a model with ten variables yields an R^2_{oos} of 11.78%. We conclude that parsimonious EBM models can capture most of the predictability of corporate bond returns with relatively few variables.

4.2.1 Univariate effects

In contrast to previous studies applying machine learning methods to asset pricing, our analysis is not limited to studying their fit. Traditional machine learning methods are subject to a trade-off

[14] TERM is the difference between the holding-period returns of the 20-year and 3-month Treasuries. LTREVB is the bond long-term reversal factor defined by Bali et al. (2021a).

Figure 3: Number of predictors – R^2_{oos}



Note: This figure shows R^2_{oos} when predictors to the model are successively added. Terms are included in order of descending importance given by $\Delta\widetilde{R}^2_i$. The horizontal line shows the R^2_{oos} of the full EBM model with 81 variables.

between accuracy and explainability. Linear models or simple decision trees are easily interpretable but often provide limited or unstable out-of-sample performance due to insufficient capacity or overfitting. More complex models can exhibit stronger performance but usually use high-dimensional interactions in the form of decision trees combined with ensemble techniques, such as bagging (e.g., Random Forests) or boosting (e.g., XGBoost). In neural networks, multiple layers of simple learners and activation functions make it difficult to isolate the contribution of individual features to the final output and write down an easily interpretable function in closed form (see, for example, Gu et al. (2020) and Chen et al. (2024)). Therefore, studies using these techniques rely on simplifications, such as holding other features constant or partial dependence plots, to study variable influences and interactions. As Rudin (2019) argues, the explanations provided by such post-hoc methods are misleading and, therefore, unsuitable for many research questions or investment decisions. Due to the EBM model's separable structure, we can instead study the precise relationships and feature interactions underlying the predictions.

As (2) shows, the univariate contribution of a feature x_i is fully defined by the associated shape function f_i that only depends on x_i. Analogously, the contribution of an interaction (x_i, x_j) is represented by the function $f_{ij}(x_i, x_j)$. Figure 4 plots the univariate shape functions f_i associated with the four most important predictors (by $\Delta\widetilde{R}^2_i$): ΔUNCf, TERM, ΔUNC, and spread. We also plot f_i for sprxrtg, defined as the product of bond spread and credit ratings, since we refer to it in sections 4.2.2 and 5.3.[15] The following section discusses the shapes of f_{ij} functions of interaction terms. Each plot shows the estimated function $f_i(x_i)$ and the associated standard deviations which are obtained

[15]This variable is labeled 'mom6xrtg' in the original data by Kelly et al. (2023), but is actually computed as spread multiplied by rating, with credit ratings converted to a numerical scale of 1 (AAA) to 22 (default). We thank Alex Dickerson for pointing out this definition.

via bagging. At the bottom is the smoothed empirical histogram of x_i. Recall that all variables are standardized. The EBM model specifies each $f_i(x_i)$ as a step function over bins with approximately equal numbers of observations. The standard errors are substantial for some variables, largely due to our choice of a relatively short training sample of eight years. The Online Appendix O-C.1 shows that standard errors are significantly smaller when we extend the training sample to 14 years.

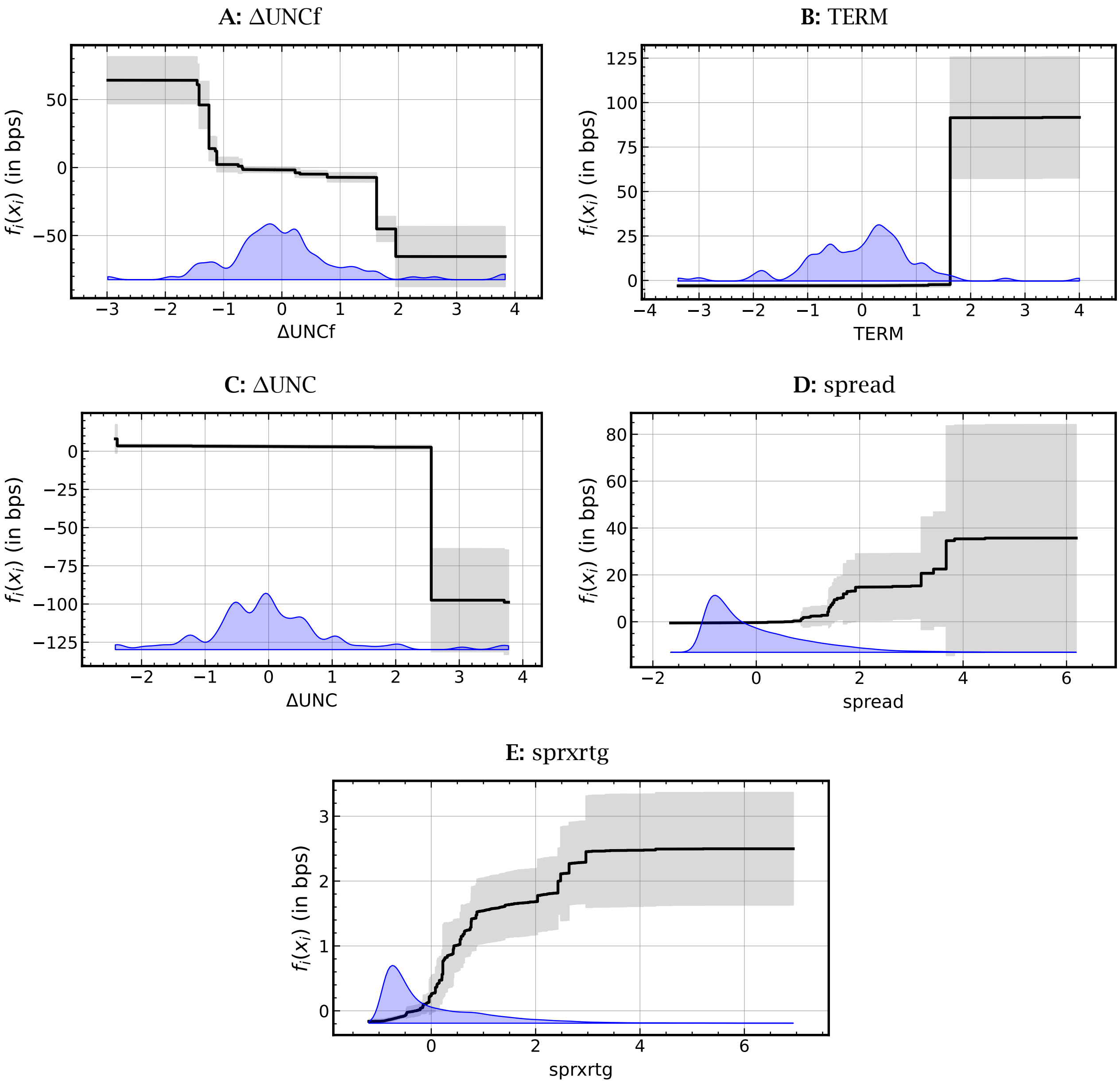
Panel A plots the estimated function $f_i(x_i)$ for the change in financial uncertainty index ΔUNCf. EBM estimates a decreasing but nonlinear effect of ΔUNCf on bond returns. Large increases in financial uncertainty lower the expected returns of corporate bonds by about 66 basis points. On the other hand, substantial declines in ΔUNCf predict higher bond returns by 64 basis points. This indicates a positive (negative) effect of falling (rising) financial uncertainty on future returns. We can also conclude that these effects are only meaningfully present at the tails of the data distribution. At the same time, there is no or only negligible contribution in an inactive region around the distribution's center. Falling financial uncertainty only affects predictions below a threshold of about 1.1 standard deviations below the mean, corresponding to a return premium of up to 64 basis points. On the other end of the ΔUNCf distribution, positive changes in financial uncertainty exceeding a threshold of about 1.6 standard deviations result in a penalty of approximately 65 basis points. Small changes in ΔUNCf have little effect on bond returns. Confidence intervals are small in the middle of the distribution but are larger towards the edges. However, the null that ΔUNCf does not affect bond returns is rejected.

The second most important predictor is the term structure factor TERM, which captures the risk arising from unexpected interest rate changes and is calculated as the difference between the monthly long-term government bond return and the one-month Treasury bill rate. Its shape function, displayed in Panel B, reveals a nonlinear effect of excess returns of Treasury bonds. The factor's contribution to the prediction is negligible for TERM values below a threshold of about 1.6 standard deviations from the mean. TERM is the difference between long and short-term returns of government bonds and is positive when long-term intrest rates fall relative to short-term rates. Since corporate bonds have long maturities (the average duration in our sample is 5.61 years), they have a positive excess return if TERM is high. Why is the effect of TERM close to zero for most of the TERM range and close to 100 bps for the right tail of the distribution? Recall that the shape functions f_i are estimated in the training sample and kept fixed in the testing sample. It turns out that there is little correlation between bond returns and TERM for most values of TERM. However, the three periods with the highest mean bond return in the training sample concur with the three highest TERM observations, as shown in the scatter plot of TERM and returns in Figure A.2. The EBM estimator captures this nonlinear relationship as a step function close to zero for all but very high TERM realizations.[16]

In contrast to financial uncertainty, macroeconomic uncertainty has little effect on corporate bond returns for most of the parameter space. However, for very large increases upwards of about

[16]The scatter plot also shows a non-parametric estimation line, which is flat for most values of TERM and positive for the highest values of TERM.

Figure 4: Univariate functions $f_i(x_i)$



Note: This figure shows the estimated $f_i(x_i)$ functions of ΔUNCf, TERM, ΔUNC, spread, and sprxrtg. The x-axis plots the input variable x_i standardized to a mean of 0 and a standard deviation of 1. At the same time, the y-axis represents the additive contribution to the prediction output through the shape function $f_i(x_i)$ (in basis points). The underlying data distribution is displayed in blue, and uncertainty intervals are shown in gray.

2.5 standard deviations from the mean in ΔUNC, bond returns suffer large losses and fall by about 100 basis points. In the sample, this occurred during the financial crisis in the third quarter of 2008 and during COVID in the first quarter of 2020. The nonlinear step functions estimated by our model suggest that investors react to substantial macroeconomic or financial uncertainty changes, while smaller fluctuations do not result in meaningful return movements. The negative correlation between changes in uncertainty and bond returns implied by this function is consistent with investors becoming more pessimistic about firms' prospects and choosing to forego investment into corporate bonds (possibly in favor of less risky assets such as government bonds) in times of rising uncertainty, resulting in lower corporate bond prices and returns. This aligns with findings by Bali et al. (2021b), who conclude that investors demand compensation for bearing credit risk and macroeconomic risk associated with uncertainty, leading to a significant risk premium in the cross-section of corporate bonds.

The most important firm-level characteristic is the bond spread. Panel D shows a relatively smooth upwards-sloped curve that links higher spread values to positive incremental effects on return predictions. Once again, effects are close to zero for large parts of the data, only starting to affect predictions after spreads deviate from the mean by at least 1.4 standard deviations. Contributions grow substantially to 35 basis points when reaching large, rare spread realizations exceeding the mean by approximately 3.2 standard deviations. This observation is consistent with high-yield bonds offering greater compensation for investors' risk in the form of higher future returns. We investigate the effect of yield spreads on bond returns in more detail below. We note that the confidence intervals of f_i are wide and include zero for most values of x_i. As mentioned above, since the f_i is estimated in the training sample, the limited amount of data in the relatively short training sample of eight years can lead to large standard errors. As a robustness check, we also estimate the model with a training sample of 14 years; see Online Appendix O-C.1 for more details. Figure D.1 shows the estimated shape functions. While the functions are similar to those in the estimation with a training sample of eight years, the standard errors are substantially smaller. For example, f_i of spread increases and statistically differs from zero. The qualitative nature of these relationships also persists across temporal subsamples, while magnitudes and sensitivities to certain predictors vary.[17]

4.2.2 Interaction effects

Gu et al. (2020) emphasize variable interactions as a source of the predictive strength of high-complexity models. In line with this observation, we find multiple interaction terms among the top 15 of the importance ranking (see Figure 2). Notably, many interactions among the most influential terms link the financial uncertainty feature with other factors or characteristics, underlining its relevance for predictions. We present results for two of the most important such terms in this section, describing the interactions of financial uncertainty ΔUNCf with bond spread and the spread-times-rating (sprxrtg) variables. Note, that $f_{ij}(x_i, x_j)$ measure only interaction effects. To calculate

[17] See the Online Appendix O-C.3 for an in-depth analysis of the global financial crisis.

the total effect of a variable, the univariate and interaction effects have to be combined. We first investigate the interaction terms and then discuss the joint effects.

Panel A of Figure 5 illustrates the shape function $f_{ij}(x_i, x_j)$ associated with the interaction of financial uncertainty on the x-axis and bond spread on the y-axis. The heatmap shows a complex interaction of the two variables. The interaction term results in an additional positive contribution of falling uncertainty for high-spread bonds, reaching up to approximately 11 basis points. The additional negative effect of rising uncertainty reducing future returns by up to 4 basis points is present for larger parts of the spread distribution. However, it also exhibits regions of more pronounced impact for bonds with lower spreads. These findings suggest that high-risk bonds (firms) benefit the most from improving macroeconomic conditions, i.e., falling uncertainty, while rising uncertainty affects bonds (firms) more uniformly. [18]

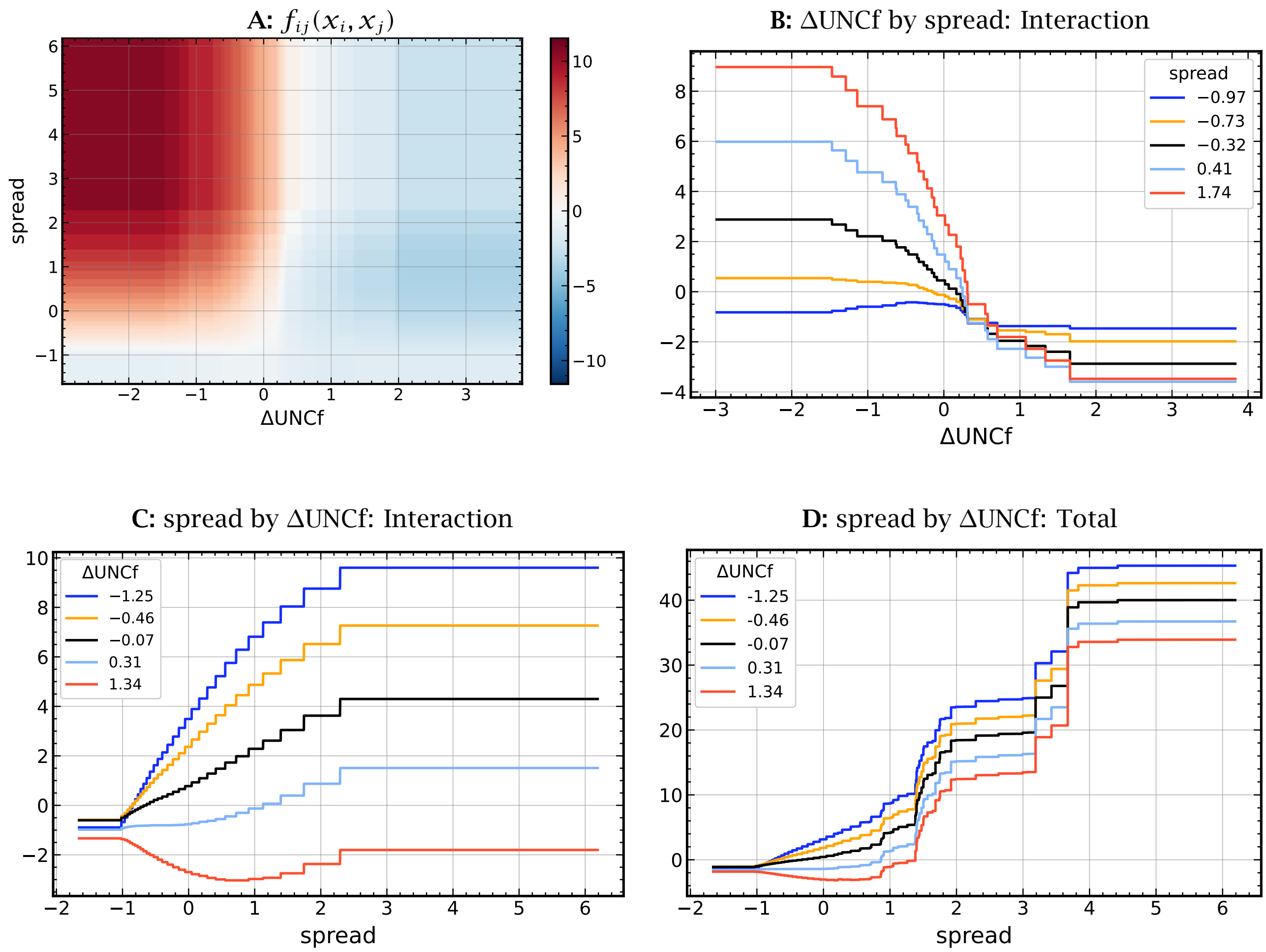
Another way to illustrate interaction effects is to plot x_i conditional on different values of x_j, i.e., $f_{ij}(x_i|x_j)$. Panel B shows the effect of ΔUNCf on the x-axis for five different values of spread, equivalent to taking horizontal cuts of the heatmap in Panel A. This plot suggests that low-spread bonds do not experience large additional effects associated with ΔUNCf but become more sensitive to changes in financial uncertainty as the spread grows. This phenomenon is especially pronounced for falling uncertainty. Panel C presents the opposite case of holding ΔUNCf constant at different levels and observing the resulting contributions as a function of spread (equivalent to taking vertical cuts in the heatmap). The spread function is monotonically increasing for changes in financial uncertainty below or around the mean. This relationship is partially reversed in times of sharply rising uncertainty, as indicated by the downwards-sloped curve around the mean.

As we have shown, the model structure allows us to separate the influence of an input variable x_i into univariate effects $f_i(x_i)$ and multivariate effects $f_{ij}(x_i, x_j)$. We can, therefore, find the aggregate effect of an input dependent on all x_j it interacts with by summing over the corresponding functions $f_i(x_i)$ and $f_{ij}(x_i|x_j)$. The univariate effect of spread, shown in Panel D of Figure 4, is increasing and ranges from -0.5 to 36 bps. Panel D of Figure 5 shows the joint effect when we add the interaction terms $f_{ij}(x_i|x_j)$ to the univariate function $f_i(x_i)$. While the univariate effect is an order of magnitude larger, the interaction terms show significant differences in predictions for bonds with high spreads when financial uncertainty declines or increases.

Figure 6 shows the interaction effect of ΔUNCf and the spread-times-ratings variable sprxrtg. The variable sprxrtg is calculated as the product of bond spread and numerical bond rating. As such, it is correlated with other variables that capture default risk, such as yield, spread, and rating. We investigate the role of sprxrtg in more detail in section 5.2. The shape function $f_i(x_i)$ of sprxrtg increases from -0.2 bps for low values of sprxrtg to 2.5 bps, see Panel E of Figure 4. Hence, the univariate quantitative effect of sprxrtg is substantially smaller than those of ΔUNCf, TERM, ΔUNC, and spread.

[18] As a robustness test, we confirm that similar patterns persist in relation to other measures of macroeconomic conditions when the uncertainty factors are excluded in Online Appendix O-C.4.

Figure 5: Interactions – ΔUNCf & spread



Note: Panel A displays the shape function of the interaction between the financial uncertainty feature and bond spread. The scaled values of ΔUNCf and spread are shown on the x- and the y-axis, respectively. The color-coded axis represents the additive contribution to the prediction output associated with combinations of x and y values in basis points. In Panels B and C, conditional contributions are derived from the shape function by separately holding each variable involved in the interaction constant at different levels. Panel D shows the sum of the univariate and interaction effects of spread conditional on ΔUNCf.

The pattern of the heatmap of the ΔUNCf and sprxrtg interaction is similar to that of ΔUNCf and spread. Therefore, the resulting interaction with the financial uncertainty factor is similar to the previously discussed interaction involving spread: Returns of low-grade high-momentum bonds increase disproportionately when financial uncertainty falls and moderately decreases otherwise. Consequently, the conditional functions resulting from the interaction term also behave similarly to the interaction between financial uncertainty and spread: The impact of changes in financial uncertainty grows as sprxrtg levels increase, with the most pronounced effect present for falling uncertainty. Once again, rising uncertainty affects bonds markedly more uniformly across the sprxrtg spectrum. The overall contribution is mostly independent of ΔUNCf for the lowest sprxrtg values. The contribution of sprxrtg to return predictions conditional on ΔUNCf levels increases in sprxrtg for ΔUNCf values around or below the mean but is mainly decreasing in sprxrtg for high ΔUNCf values. In contrast to spread, the univariate effect of sprxrtg is smaller so that the interaction terms dominate the total effect (Panel D).

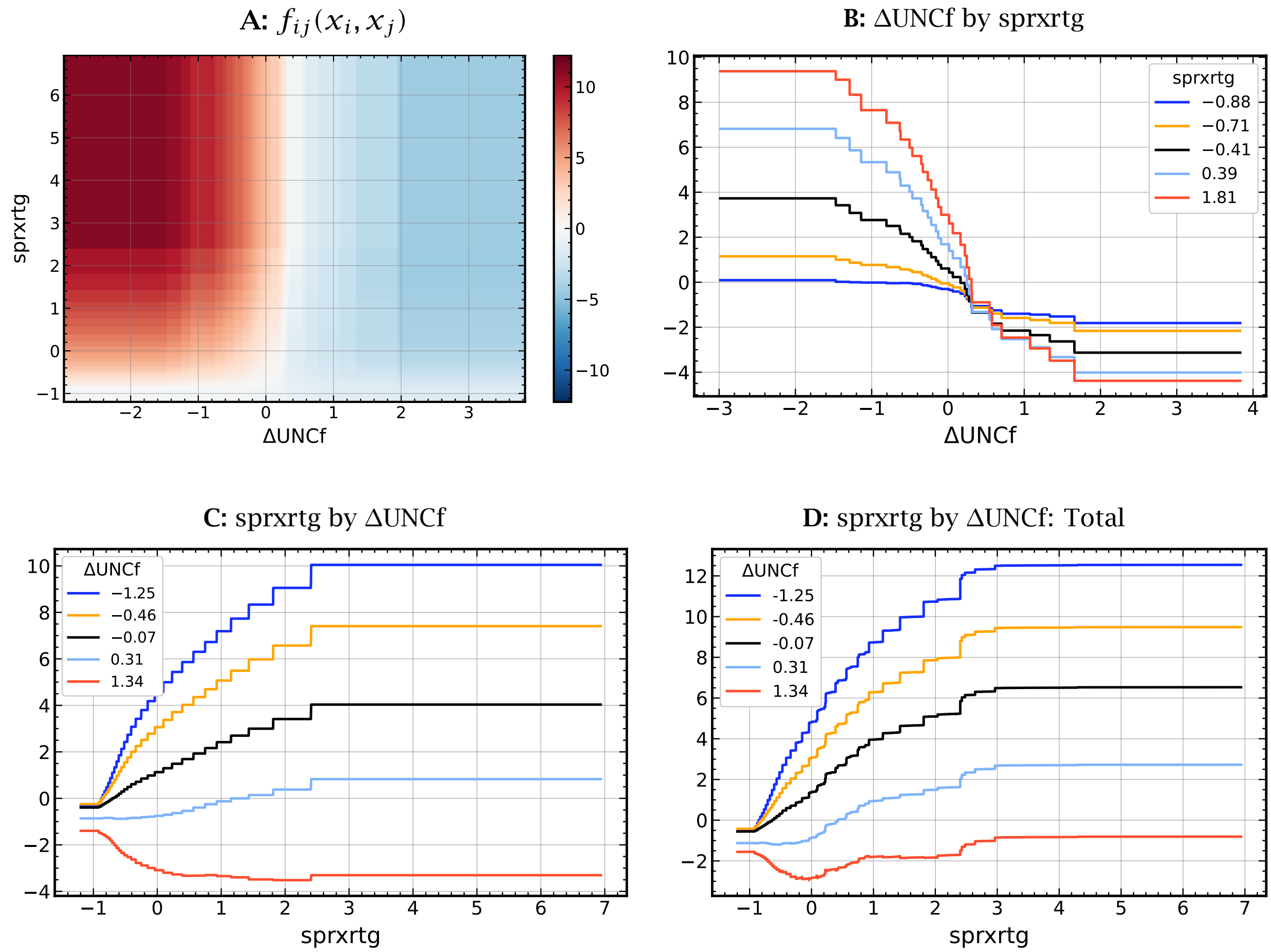
4.3 Interaction effects in returns

The shape functions capture the marginal effect of a single variable or combinations of two variables. In this section, we study how financial uncertainty and bond spreads relate to predicted and realized returns in the sample. We group observations by ΔUNCf (10 bins) and spread (25 bins) and compute the mean predicted and observed returns for each bin and the interaction bins.[19] Let b_{ij} be the bin j of variable x_i and N_{ij} be the number of observations in bin b_{ij}. The mean of returns conditional on x_i being in bin j is given by $\overline{r}_{ij} = (1/N_{ij}) \sum_{l,t} r_{l,t+1}$ if $x_{i,l,t} \in b_{ij}$. To compute the means of predicted returns, $r_{l,t+1}$ is replaced by $\hat{r}_{l,t+1}$. Panel A of Figure 7 plots the means of predicted (orange) and realized (blue) returns by spread bins. Higher spreads are associated with higher observed returns. Since the spread shape function is upward-sloping, the EBM model captures this pattern in the data even when other variables are considered. However, the fit is imperfect as the model understates the effect of spread on bond returns. Panel B shows results for financial uncertainty. High ΔUNCf is associated with low predicted returns, consistent with the negative slope of the ΔUNCf shape function. The model closely matches the negative link between ΔUNCf and the observed return.

Panels C and D show heat maps of the interaction effect of ΔUNCf (x-axis) and spread (y-axis) on realized and predicted returns, respectively. The plot for observed returns shows the importance of the bond spreads and financial uncertainty interaction. Returns are highest for bonds with high spreads when financial uncertainty is decreasing. Conditional on ΔUNCf<0, the effect of the spreads on the returns becomes weaker, and conditional on high spreads, the effect of ΔUNCf also decreases. Moreover, for positive values of ΔUNCf, the interaction terms are weaker than for ΔUNCf<0. The results for predicted returns are in Panel D. The heatmap closely matches that of realized returns, both qualitatively and quantitatively, showing that the EBM model captures the interaction pattern

[19] We choose fewer bins for ΔUNCf than for spread since ΔUNCf is a market-level variable while spread differs by firm.

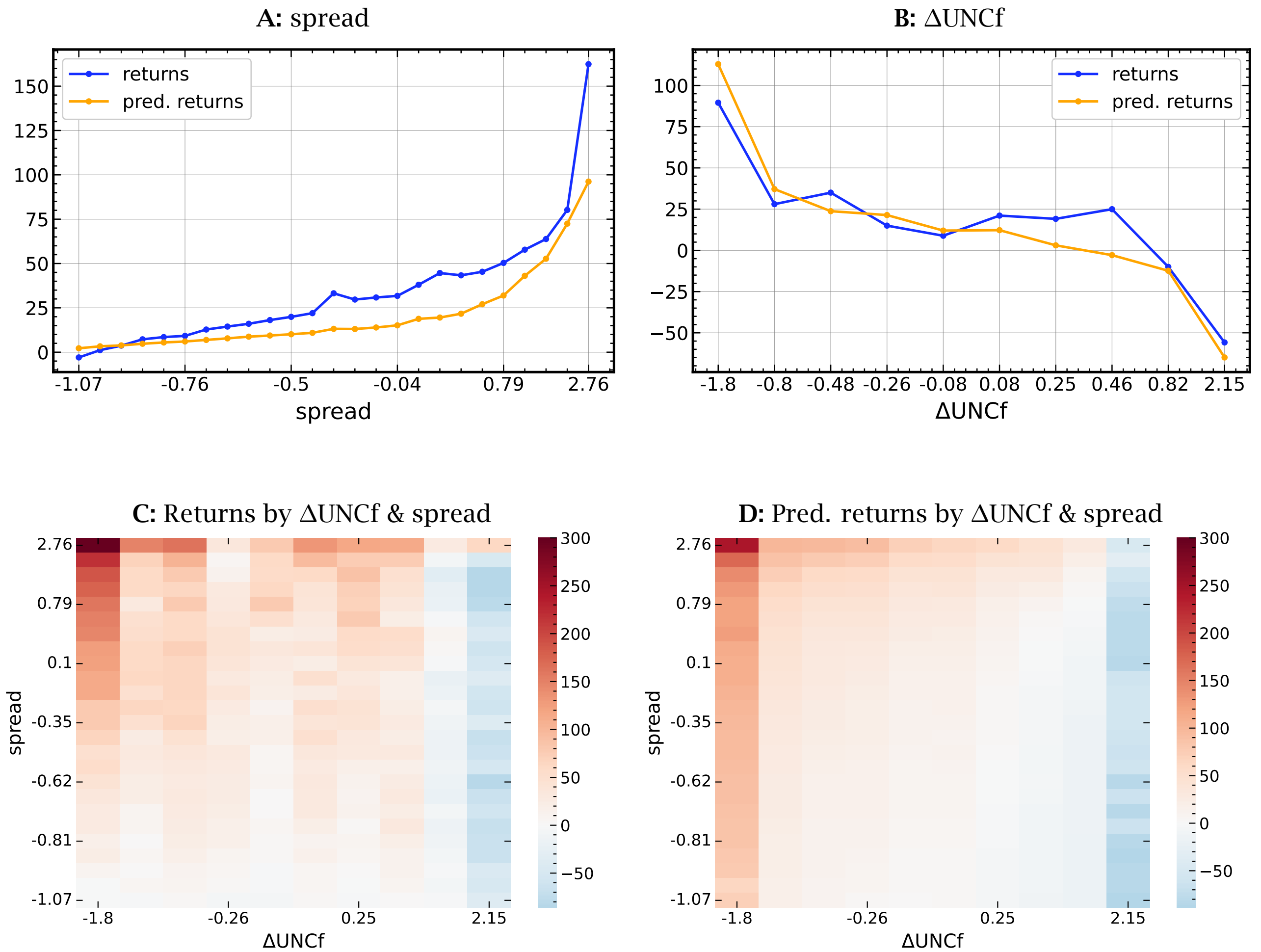
Figure 6: Interactions – ΔUNCf & sprxrtg



Note: Panel A displays the shape function of the interaction between the financial uncertainty feature and bond spread times rating. The scaled values of ΔUNCf and sprxrtg are shown on the x- and the y-axis, respectively. The color-coded axis represents the additive contribution to the prediction output associated with combinations of x and y values in basis points. In Panel B and C, conditional contributions are derived from the shape function by separately holding each variable involved in the interaction constant at different levels. Panel D shows the sum of the univariate and interaction effects of sprxrtg conditional on ΔUNCf.

of financial uncertainty and bond spreads. The ΔUNCf and sprxrtg interaction results, reported in Figure 7, are similar.

Figure 7: Interactions – ΔUNCf & spread and returns



Note: This figure shows returns and predicted returns by ΔUNCf and spread. We group observations by UNCf (10 bins) and spread (25 bins) and compute the means of observed and predicted returns for each bin and the interaction bins. Panels A and B show the means of observed returns (blue) and predicted returns (orange) by spread and ΔUNCf bins. Panel C plots the heatmap of observed returns by ΔUNCf (x-axis) and spread (y-axis) bins. High (low) returns are in red (blue). Panel D shows the same plot for predicted returns.

4.4 Uncertainty and corporate bonds

The previous section showed that uncertainty, particularly changes in financial uncertainty measured by ΔUNCf, is an important state variable for corporate bond returns. This section investigates the effect of ΔUNCf in more detail. First, we focus on the roles of levels vs. changes of uncertainty. Second, we analyze how uncertainty affects different firms.

The EBM estimation suggests that changes in uncertainty are more relevant for returns than levels. Figure 8 plots cross-correlations of mean returns by month, $\bar{r}_t$, with the level of financial uncertainty, UNCf, and first-differences, ΔUNCf. The level of financial uncertainty is only weakly correlated with returns. For example, their contemporaneous correlation is -0.06 and statistically insignificant. The correlations of UNCf and returns beyond 3 months are positive and marginally statistically significant. Given these results, it is not surprising that the level of financial uncertainty does not play a major role in the EBM estimation. The correlation pattern for *changes* in financial uncertainty is different. The cross-correlations $\rho_k = \rho(\Delta\text{UNCf}_t, \bar{r}_{t+k})$ are economically and statistically significant for $k = -1, 0, 1, 2$. Lagged returns are negatively correlated with ΔUNCf ($\rho_{-1} = -0.2$), which could be due to an anticipation effect of investors in the bond market, as described by Jurado et al. (2015). The contemporaneous correlation is $\rho_0 = -0.54$. Most importantly for forecasting models, ΔUNCf is negatively correlated with next month's bond returns as the point estimate of $\rho_1 = -0.5$ with a t-statistic of -8.93.[20] EBM exploits this predictive return correlation of the change in the financial uncertainty index and, as a result, estimates ΔUNCf as an important predictor.

To analyze how uncertainty affects different firms, we split each month of the sample into two subsamples, denoted "low" and "high", according to a firm-level variable, e.g., size or spread.[21] We then estimate the EBM model on the two subsamples and compare the effects of ΔUNCf. Figure 9 shows the estimated shape functions $f_i(x_i)$ from the subsamples based on me, spread, nrtg, sprd2d, sprxrtg, and maturity.[22] For comparison, we also plot $f_i(x_i)$ for the full sample.

Panel A shows the result for firm-size subsamples. The effect of ΔUNCf on large firms is similar to that for the full sample, which is to be expected since the estimation weighs firms by their market cap. The estimated $f_i(x_i)$ for small firms is substantially steeper, indicating that their bond returns react more strongly to changes in financial uncertainty. For large increases in ΔUNCf, expected bond returns of small firms decline by up to 200 basis points, four times larger than the effect for large firms. If ΔUNCf decreases, the effect is twice as large as for large firms.

Next, we consider variables that are related to default risk. Panel B shows $f_i(x_i)$ for firms with low and high bond spreads. We find that the ΔUNCf effect is about twice as large for bonds with high spreads than for firms with low spreads. The difference is even larger for bonds with different credit ratings. Panel C shows that bonds with low credit ratings (high nrtg) react substantially more to changes in financial uncertainty than high-quality bonds (low nrtg). Expected bond returns fall by up to 265 bps when ΔUNCf rises compared to only 50 bps for highly rated bonds. The effects of declining financial uncertainty are 128 bps and 58 bps, respectively. Next, we split the sample by the spread-to-D2D measure sprd2d suggested by Correia et al. (2012) and sprxrtg, see Panels D and E. The patterns are similar to those for bond spreads and ratings, as ΔUNCf affects firms with high sprd2d and sprxrtg substantially more than those with low sprd2d and sprxrtg. Finally, we find that bonds with longer maturities are more sensitive to financial uncertainty than those with shorter maturities (see Panel F). We conclude that bond returns of smaller firms with riskier bonds

[20]We confirm the correlation results using a bivariate VAR with $(\Delta\text{UNCf}_t, \overline{r}_t)$, where $\bar{r}_t$ is the mean return of bonds in month t.

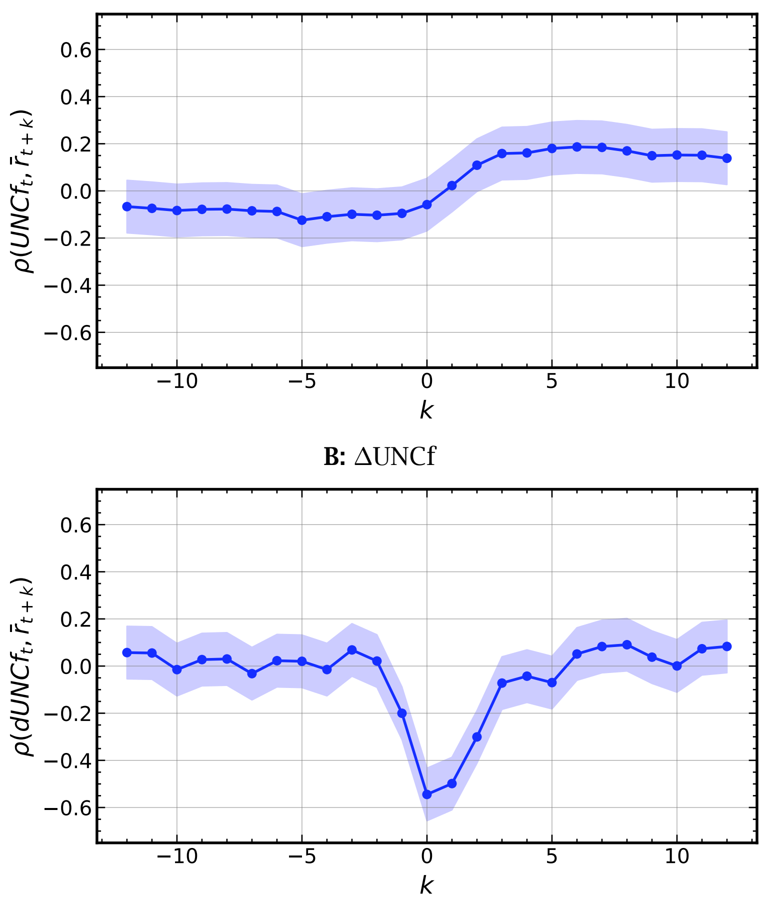
[21]We present a similar analysis for the term structure factor in the Online Appendix O-C.5.

[22]Subsample analysis for distance to default and duration is provided in the Online Appendix O-C.6.

Figure 8: Cross-correlation of financial uncertainty and mean returns

A: UNCf

B: ΔUNCf



Note: This figure plots cross-correlations of mean returns with the level of financial uncertainty, $\rho(\text{UNCf}_t, r_{t+k})$, and first-differences, $\rho(\Delta\text{UNCf}_t, r_{t+k})$ where k is the lead/lag order in months.

of long maturity are most sensitive to changes in financial uncertainty, particularly large increases in financial uncertainty.

5 Portfolios based on the EBM model

Next, we use our interpretable model to construct bond portfolios. Leveraging the interpretable model's return predictions, we form bond portfolios as follows. Each month t of the sample period, we rank firms based on predicted $t+1$ excess returns from the EBM model (2) and create ten decile portfolios.[23] Using this procedure, we construct monthly equal and value-weighted portfolios, in which each firm is weighted by the market value of its bonds, which we obtain from Open Source Bond Asset Pricing. For each decile portfolio, we compute the (excess) return of the buy-and-hold strategy that is long $1 in the portfolio. Finally, we create long-short portfolios by taking long positions in the highest decile and short positions in the lowest decile.

We also consider an alternative investment strategy that exploits time-series forecastability embedded in the EBM model. Recall that the EBM model yields forecasts for bonds of each firm, $\hat{r}_{l,t+1}$. For each month and each portfolio, we compute the predicted return of the portfolio by taking the weighted average of all predicted bond returns $\hat{r}_{l,t+1}$ in the portfolio. We implement a simple investment strategy, EBM-pos, that exploits high-return forecasts while considering that predicted bond returns could be negative. If the predicted excess return of a portfolio is negative, we invest in the risk-free rate instead of the portfolio. Since we use excess returns, the return of the investment strategy in such periods is equal to zero. Hence, this investment strategy coincides with the standard buy-and-hold strategy if expected bond returns are positive but exploits the time series forecastability of the EBM model for market timing.

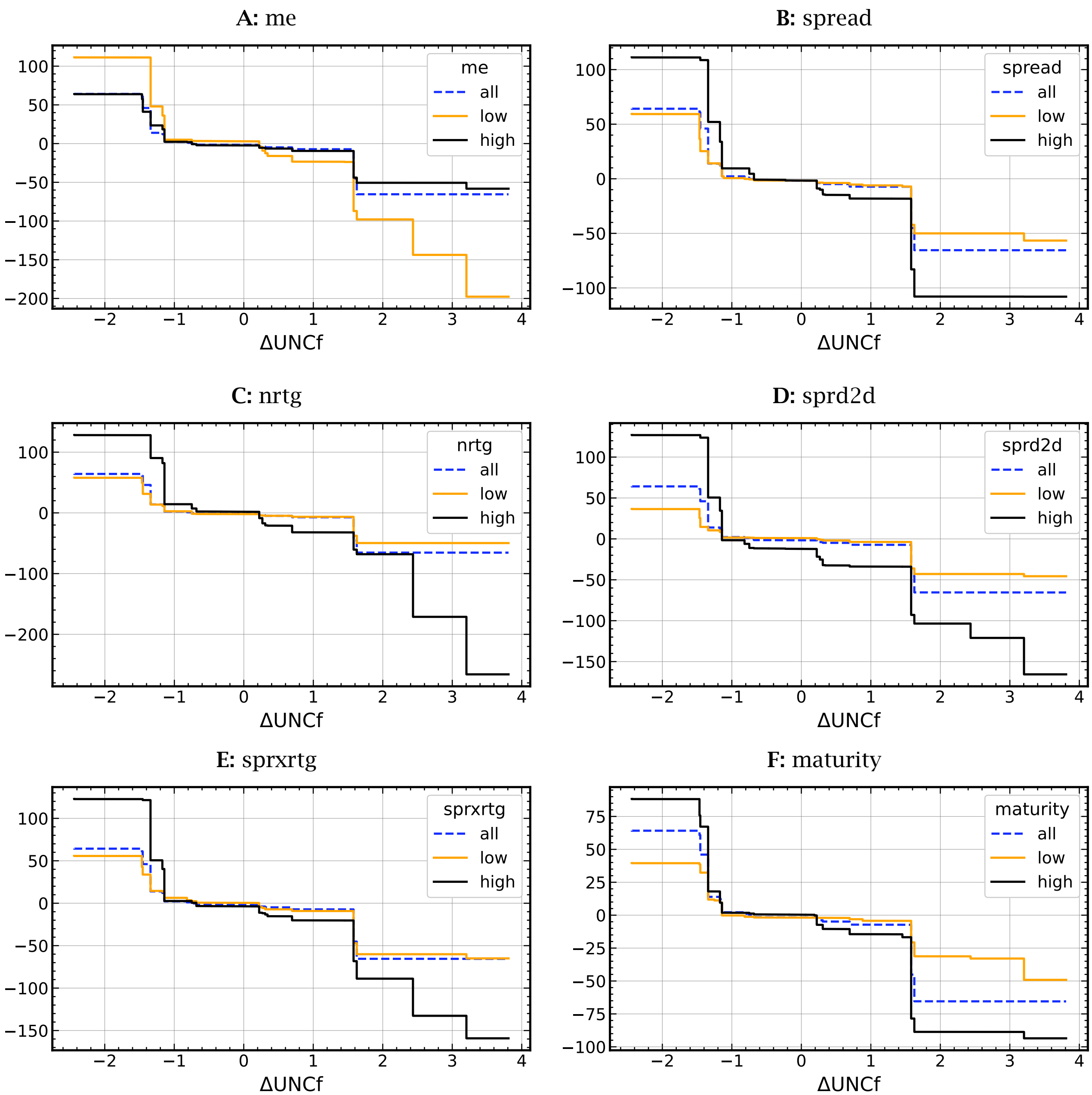
5.1 Portfolio returns

Table 4 reports the decile portfolios' means, standard deviations, and Sharpe ratios in the July 2011 to August 2020 testing sample. Means and Sharpe ratios of the value-weighted portfolios are monotonic from portfolio 1 to portfolio 10 (except the Sharpe ratio of portfolio 7). Moreover, the return spread is economically meaningful. The mean excess return of the first portfolio is negative (-0.06%), while the top portfolio has a mean excess return of 0.61% per month. Thus, the mean return of the 10-1 portfolio is 0.67% per month or 8.04% per year. Relative to the duration-adjusted bond market return, this long-short strategy generates a monthly alpha of 0.55% (t=4.22). Monthly Sharpe ratios range from -0.05 for the first portfolio to 0.23 for the tenth portfolio. The Sharpe ratio of the 10-1 portfolio is 0.38, or 1.32% annually.[24]

[23] As return predictions and characteristics are aggregated on the firm level, we select firms, representing a mean of all their bonds weighted by the amount outstanding.

[24] We observe similar results for the Random Forest and XGBoost models (see Online Appendix O-D.1).

Figure 9: $f_i(x_i)$ – ΔUNCf by subset of firms



Note: This figure compares shape functions associated with ΔUNCf across subsets of firms obtained from monthly sorts on firm size, spread, rating, maturity, sprd2d, sprxrtg, and maturity. Each month, firms are grouped into "low" and "high" subsets according to one of the four variables. The EBM model is then estimated on the two subsets of firms. The dashed blue line plots the shape function for the full sample with all firms. Models are trained on the first 8 years of data, validated in the following year, and tested in the remaining sample.

Panel B shows results for portfolios based on the EBM-pos strategy that invests in the riskfree asset when the predicted portfolio return is negative. Compared to the simple buy-and-hold strategy, mean returns are higher, and volatility is lower for all ten portfolios. As a result, the effect on Sharpe ratios is pronounced. For portfolios one to eight, the Sharpe ratio more than doubles. Moreover, the Sharpe ratios of the long-only portfolios eight to ten are above 0.3 and close to that of the standard buy-and-hold strategy's 10-1 long/short portfolio. These results suggest that market timing based on the time series forecastability of the EBM model boosts returns significantly compared to the simple buy-and-hold strategy. To illustrate this effect, we plot cumulative log returns of portfolio ten based on the two strategies in Panel A of Figure 10. The shaded areas indicate quarters where the predicted portfolio return is negative, so the EBM (blue) and EBM-pos (orange) portfolios differ. The EBM model predicted negative returns for portfolio 10 during the financial crisis and the COVID pandemic. Buy-and-hold returns were negative during those periods, but the EBM-pos strategy anticipated the bear market and exited the bond market. Hence, it avoids the downturns and has stable returns during these periods. The results for equal-weighted portfolios in Panels C and D of Table 4 are similar, which implies that small firms do not drive the results.

Table 4: Returns of EBM portfolios

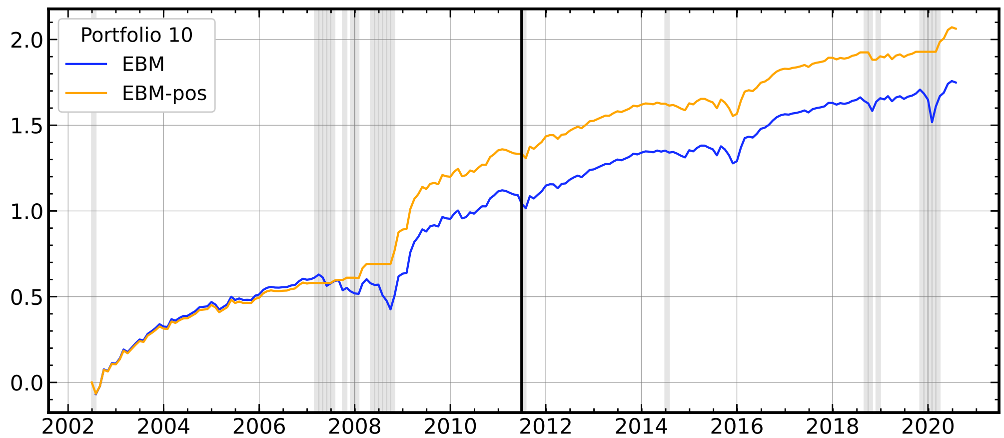
	1	2	3	4	5	6	7	8	9	10	10-1
					Panel A: EBM VW						
Mean	-0.06	0.04	0.09	0.12	0.16	0.21	0.21	0.28	0.37	0.61	0.67
S.D.	1.28	1.22	1.12	1.25	1.37	1.50	1.71	1.92	2.09	2.60	1.75
SR	-0.05	0.03	0.08	0.10	0.12	0.14	0.12	0.14	0.18	0.23	0.38
					Panel B: EBM-pos VW						
Mean	0.04	0.10	0.15	0.20	0.25	0.34	0.37	0.44	0.50	0.68	0.64
S.D.	0.40	0.55	0.62	0.73	0.80	0.96	1.10	1.20	1.46	1.96	1.72
SR	0.10	0.18	0.25	0.28	0.31	0.35	0.34	0.37	0.35	0.35	0.37
					Panel C: EBM EW						
Mean	-0.08	0.02	0.09	0.11	0.15	0.18	0.22	0.30	0.38	0.65	0.73
S.D.	1.27	1.19	1.12	1.22	1.38	1.52	1.70	1.88	1.99	2.67	1.81
SR	-0.07	0.02	0.08	0.09	0.11	0.12	0.13	0.16	0.19	0.24	0.40
					Panel D: EBM-pos EW						
Mean	0.03	0.09	0.15	0.19	0.25	0.32	0.38	0.46	0.52	0.74	0.71
S.D.	0.35	0.46	0.55	0.67	0.75	0.91	1.05	1.18	1.39	1.97	1.77
SR	0.09	0.20	0.27	0.29	0.34	0.36	0.36	0.39	0.37	0.38	0.40

Note: This table compares monthly means, standard deviations, and Sharpe ratios for ten value-weighted portfolios constructed based on predicted return deciles in the testing sample. The last columns show the results of the 10-1 portfolio, which is long in portfolio 10 and short in portfolio 1. EW are equally-weighted portfolios and VW are value-weighted portfolios by bond market value. EBM-pos is the strategy that invests in the riskfree asset when the predicted portfolio return is negative.

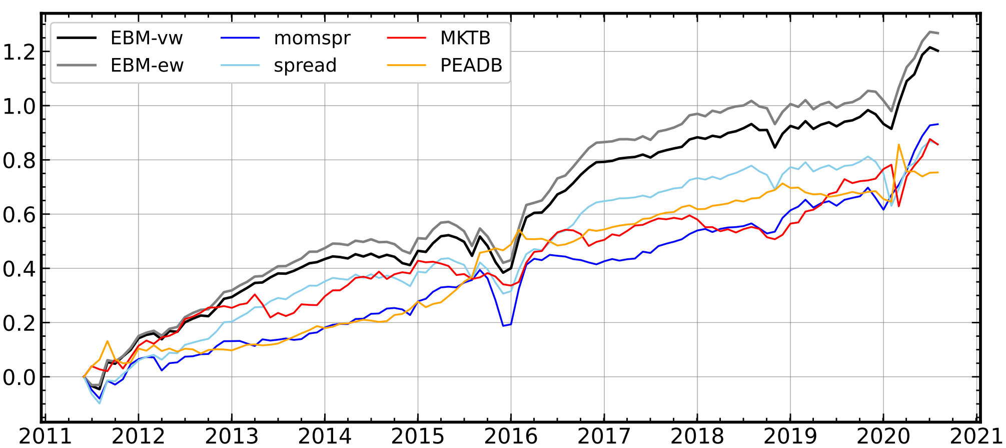
To benchmark its performance, we compare the EBM portfolios to several alternatives. First, we construct quintile portfolios based on all available bond and firm characteristics and create char-

Figure 10: Cumulative returns

A: EBM portfolio 10



B: EBM, characteristics, and factors



Note: This figure shows cumulative log returns of EBM portfolios over the sample period. Panel A plots the value-weighted portfolio 10 of the standard buy-and-hold strategy (EBM) and the strategy that invests in the riskfree asset when the predicted portfolio return is negative (EBM-pos). The shaded areas indicate quarters in which the predicted portfolio return is negative. Panel B plots the cumulative log returns of the 10-1 EBM equal and value-weighted portfolios, the two 10-1 characteristic portfolios with the highest Sharpe ratios (momspr and spread), and the two factors with the highest Sharpe ratios (MKTB and PEADB). Returns are scaled to have an annualized standard deviation of 10%. The black vertical line indicates the start of the testing period (July 2011).

acteristic factors as 5-1 long-short portfolios.[25] Additionally, our data set uses the traded factor portfolios from the bond and stock markets. Table 5 reports results for the ten characteristics-sorted portfolios and factors with the highest Sharpe ratios, respectively, in the combined training and validation samples (July 2002 to June 2011). The results for characteristic portfolios are shown in Panel A. The top three characteristics are momspr[26], spread, and sprxrtg with Sharpe ratios of 0.30, 0.28, and 0.27, respectively. Note that these are also among the most important characteristics in the EBM model; see Panel C of Figure 2 and Online Appendix O-D.2. The Sharpe ratios of the remaining characteristic portfolios range from 0.26 to 0.13. The Sharpe ratios of the factors in Panel B are similar. The bond market factor MKTB has the highest Sharpe ratio of 0.28, followed by PEADB (0.24) and TERM (0.20).[27] Recall that the Sharpe ratios of the EBM portfolios range from 0.38 to 0.40 and are thus considerably higher than those of characteristic portfolios and factors. Panel B of Figure 10 compares the cumulative returns of the value and equal-weighted buy-and-hold EBM 10-1 portfolios with those of momspr, spread, MKTB, and PEADB. The plot shows that the EBM portfolios outperform the other strategies consistently over the sample.

Table 5: Returns of characteristics-sorted portfolios and factors

	Panel A: 5-1 characteristic portfolios									
	momspr	spread	sprxrtg	nime	yield	sprd2d	coupon	voleq	ret	skew
Mean	0.29	0.45	0.39	0.10	0.42	0.26	0.24	0.21	0.22	0.13
S.D.	0.98	1.61	1.44	0.39	1.77	1.23	1.20	1.15	1.52	0.98
SR	0.30	0.28	0.27	0.26	0.24	0.21	0.20	0.18	0.15	0.13
	Panel B: Factors									
	MKTB	PEADB	TERM	CRY	VAL	DRF	HMLB	RMW	LTREVB	SMB
Mean	0.44	0.12	0.62	0.52	0.37	0.40	0.15	0.14	-0.04	-0.24
S.D.	1.58	0.52	3.02	2.58	1.97	2.34	1.62	1.55	0.92	2.44
SR	0.28	0.24	0.20	0.20	0.19	0.17	0.09	0.09	-0.05	-0.10

Note: This table reports monthly means, standard deviations, and Sharpe ratios of characteristics-sorted portfolios and factors. We sort characteristics-sorted portfolios and factors by their Sharpe ratios in the combined training and validation samples. We then report statistics of the ten characteristics-sorted portfolios (Panel A) and factors (Panel B) in the testing sample.

5.2 Properties of EBM portfolios

Having established the performance of the EBM portfolio, we shift our focus to a new question: What insights can the interpretable model provide regarding portfolio composition? In contrast to black box approaches, which only allow for ex-post analysis of selected assets, our interpretable model

[25] We construct quintile instead of decile portfolios because of a limited number of observations for some characteristics.

[26] The variable momspr (spread momentum) is defined as the log of the spread six months earlier minus current log spread.

[27] MKTB is the corporate bond market excess return by Dickerson et al. (2023a), PEADB is the bond earnings announcement drift factor of Nozawa et al. (2023), and TERM is the difference between returns of long and short-term government bonds (Fama and French (1993)).

offers an ex-ante understanding of how portfolios are built. In other words, the EBM model produces transparent rules describing which values of characteristics lead to higher return predictions and, therefore, feature in the portfolios.

To this end, we compute the average characteristics of the bonds in the ten EBM portfolios in each quarter. The results are in Figure 11. Panel A shows the average characteristics in the ten portfolios of the characteristics with the highest scores $S(i)$ in the EBM model: spread, var, and d2d.[28] The pattern is monotonic across portfolios for all three characteristics. Recall from Panel D of Figure 4 that the shape function f_i of spread is increasing so that higher spreads predict higher returns. Consequently, portfolios 1 (10) include bonds with below (above) average spreads. The shape functions of var and d2d are downward sloping, and the pattern across EBM portfolios is reversed. For comparison, Panel B repeats the same plot but for the three characteristics with the lowest scores: turnvol, levop, and salesat. The figure shows that these characteristics do not vary across the EBM portfolios. Panels C and D show characteristics with the highest and lowest Sharpe ratios. Portfolio 10 includes bonds with properties associated with high Sharpe ratios, such as momspr, spread, and sprxrtg. The first portfolio includes bonds with opposite features. In contrast, portfolios do not exhibit monotonic patterns and do not load on characteristics with low Sharpe ratios (e.g., dolvol, retlag, dureq).[29]

The previous results are based on averages across the sample. Next, we study the variation of portfolio characteristics over time. Panel E plots the time series of spread of portfolios 1 and 10. We scale by the mean spread over all bonds in each quarter to eliminate changes in the spread level. The plot shows that portfolio 10 consistently exhibits high spreads that range from 1 to 3. On the other hand, spreads of portfolio 1 are consistently negative. Panel F shows the time variation of salesat.[30] Note that the average salesat is close to zero for all portfolios. The plot shows that this is not only true on average but consistently throughout the sample.

5.3 Impact of interactions

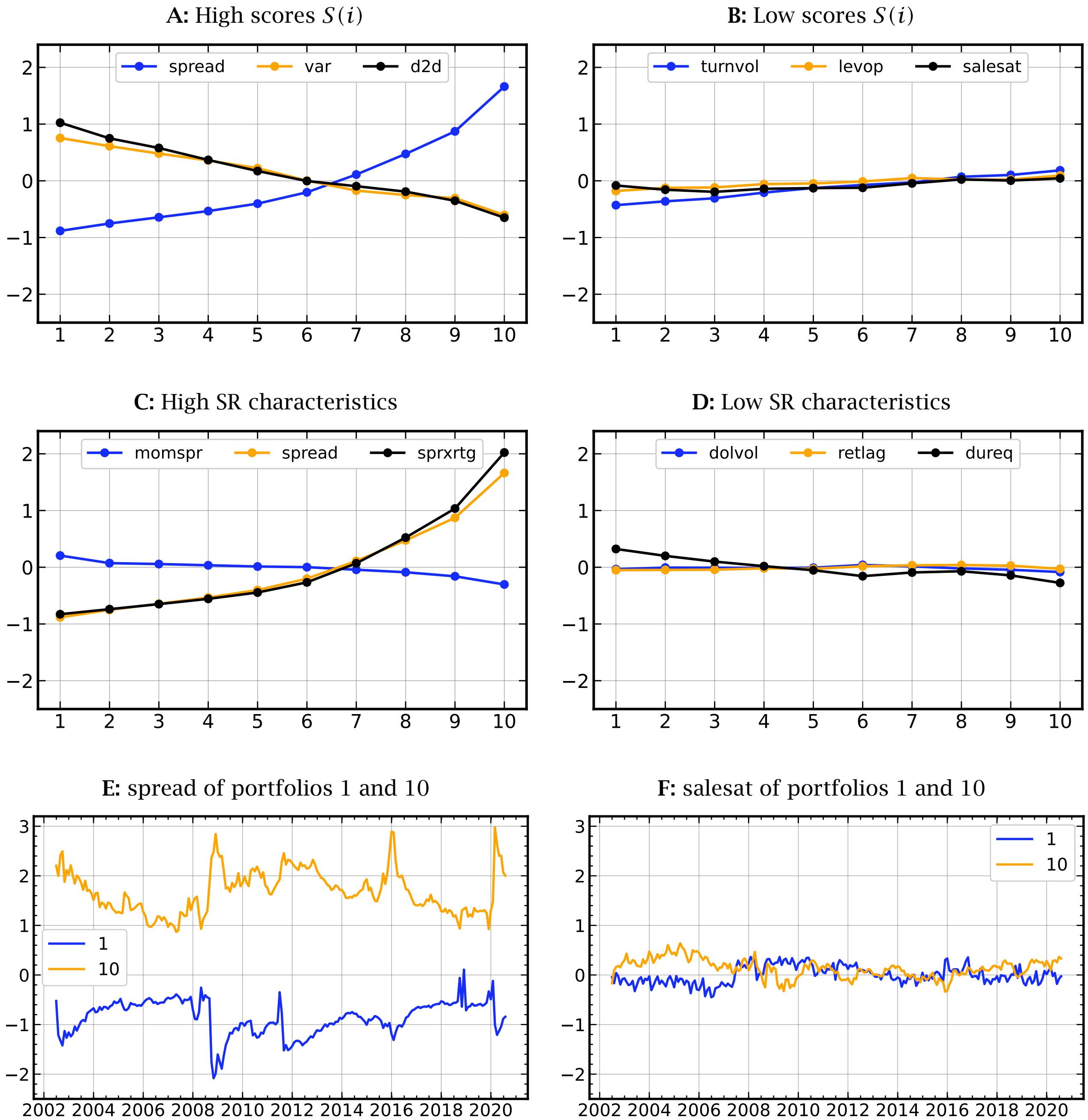
In section 4.2, we showed that interactions of characteristics play an important role in the fit of the EBM model. Next, we study how interaction terms are related to EBM portfolios. We focus on the interactions of financial uncertainty and characteristics whose shape functions are presented in section 4.2.2, spread, and sprxrtg. Based on the observed functions, we can derive implications for the composition of the EBM portfolios. Consider first the interaction between ΔUNCf and bond spread. The estimated function $f_{ij}(x_i, x_j)$ of the ΔUNCf/spread interaction is shown in Figure 5, Panel A. Panel D plots the sum of the univariate and interaction terms of spread for different values of ΔUNCf, that is, $f_{ij}(x_i|x_j)$. The effect of spread on predicted returns varies with ΔUNCf for bonds

[28]var is value-at-risk defined as the 2nd lowest credit excess return over the past 36 months (Dickerson et al. (2023b)) and d2d is the distance-to-default measure suggested by Bharath and Shumway (2008).

[29]dolvol is the trading volume in dollars, retlag is the bond return lagged by one month, and dureq is equity duration.

[30]salesat is asset turnover calculated as sales over total assets.

Figure 11: Characteristics of EBM portfolios



Note: This figure shows characteristics of the ten EBM portfolios. Panels A and B show the average of the three characteristics with the highest and lowest scores $S(i)$ in the testing sample. Panels B and C show the characteristics with the highest and lowest Sharpe ratios. Panels E and F show spread and salesat for portfolios 1 and 10 over time.

with high spreads. When financial uncertainty falls, the spread coefficient is larger than when ΔUNCf is negative. For low spread levels, spread has little effect on predicted returns for any value of ΔUNCf.

This interaction effect is reflected in EBM portfolios. The high-predicted return portfolio 10 includes bonds with high spreads. When financial uncertainty falls, spread is a relatively stronger predictor than when it rises. Hence, portfolio 10 should include bonds with higher spreads when ΔUNCf is negative relative to when ΔUNCf is positive. Panel A of Figure 12 shows the scatter plot of ΔUNCf on the x-axis and spread of the tenth portfolio on the y-axis. The plot also shows the regression line in orange with 95% confidence intervals and the regression coefficient in the top left corner. As expected, there is a negative relationship between financial uncertainty and spread. Panel B shows the same plot for portfolio 1. Based on the estimated shape function, the role of ΔUNCf is small for low-spread bonds, which make up the first portfolio. Therefore, the interaction of ΔUNCf and spread should be much weaker for portfolio 1 than for portfolio 10, which is confirmed in the scatter plot.

Panels C and D show the corresponding plots for the ΔUNCf/sprxrtg interaction. As documented in section 4.2.2, the interaction between ΔUNCf and sprxrtg follows a similar pattern. From the shape function associated with the interaction, we can again derive ex-ante implications for portfolio composition: Return increases are concentrated on high-sprxrtg bonds in times of falling uncertainty. Rising uncertainty leads to return decreases, especially for high values of sprxrtg, as does falling uncertainty for low-sprxrtg bonds. Therefore, we expect a negative correlation between ΔUNCf and sprxrtg in the long portfolio and a positive correlation in the short portfolio, which is consistent with the observed correlations of -0.18 in the long leg and 0.45 in the short leg.

6 Comparison with SHAP

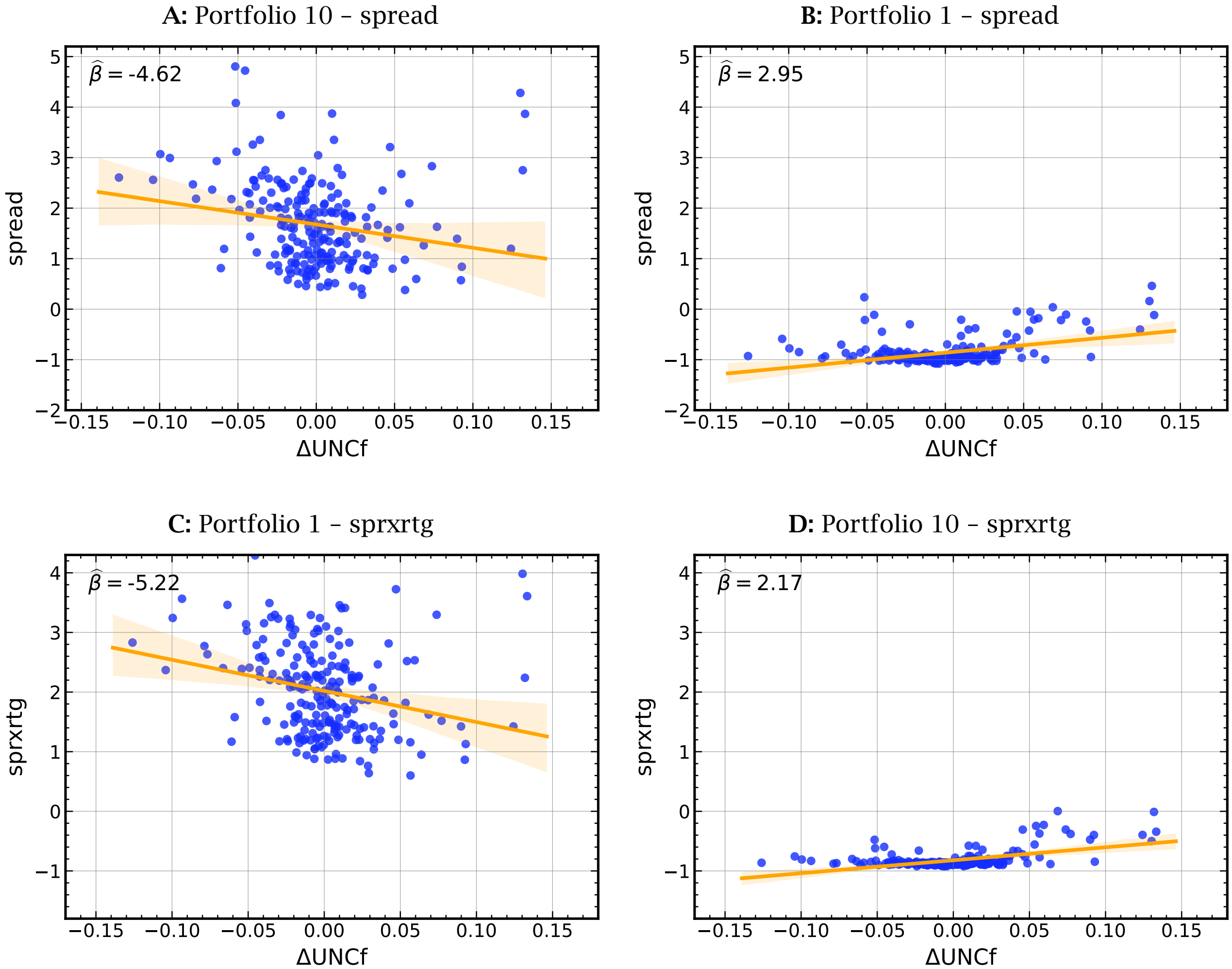
Although black box ML models are inherently uninterpretable, there are several model-agnostic methods that can be used to explain their predictions, see, for example, Molnar (2020). One of the most popular techniques is SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017; Lundberg et al., 2020), which is based on the concept of Shapley values from cooperative game theory (Shapley, 1953). In this section, we compare EBM to SHAP values implied by Random Forest and XGBoost estimations. The SHAP method assigns each feature a set of values that represent its contribution to the predictions of the model. SHAP decomposition is additive and satisfies several desirable properties, such as local accuracy, consistency, and missingness (Lundberg and Lee, 2017). In particular, it ensures that the sum of the SHAP values is equal to the prediction of the model for each observation.

SHAP decomposes the predictions of a black box model $\hat{r}_{l,t+1}$ in the following way:

$$\hat{r}_{l,t+1} = \phi_0 \; + \; \sum_{i=1}^{N} \phi_{i,l,t} \; + \; \sum_{i>j} \phi_{ij,l,t}, \tag{9}$$

where $\phi_{i,l,t}$ and $\phi_{ij,l,t}$ represent the SHAP main and interaction effects, respectively, corresponding

Figure 12: ΔUNCf and characteristics of EBM Portfolios



Note: This figure shows scatter plots of ΔUNCf and spread and sprxrtg of EBM portfolios 1 and 10. The orange line is the regression line with 95% confidence intervals.

to the individual and pairwise Shapley values of features in the model prediction.[31] Note that the SHAP decomposition has a similar form as the EBM model (2). Indeed, if the data-generating process is linear, i.e., $f_i(x_i) = \gamma_i x_i$, and the variables are independent, SHAP values are given by $\phi_i = \gamma_i(x_i - E(x_i))$ (Lundberg and Lee, 2017). In this case, EBM and SHAP are conceptually identical.[32]

While SHAP values are the most popular attribution method in machine learning, they are not without problems (see Rudin (2019) for a critique of the interpretability methods of black-box models). In contrast to the glass box model, SHAP is a post-hoc approach, i.e., an attempt to explain the predictions of the original black box with a simplified model that is interpretable. As such, SHAP values are approximations of the true effects of variables. In general, the quality of the approximation is unknown and it is possible that SHAP values are poor measures of the true variable effects. For example, Slack et al. (2020) show that SHAP values can be unreliable and mask biases in the black box model.

We apply TreeSHAP, a SHAP variant designed for tree-based models, to the Random Forest and XGBoost models described in section 4.1. We compute $\phi_{i,l,t}$ for each feature i and $\phi_{ij,l,t}$ for each pair of features (i, j) for each firm-month observation (l, t) in our sample. Table 6 shows the ten most important variables in the three models. For RF and XGB, we sort the variables by their sum of absolute SHAP values across the observations in the training sample. Although there is some commonality between the models, they differ in several dimensions. The uncertainty variables ΔUNCf and ΔUNC are among the most relevant variables in the three models. However, the remaining variables are different. For example, levmkt is the second most important variable in the Random Forest estimation, but it plays no major role in the other models. Similarly, UNC and VIX are relevant in RF and XGBoost but not in EBM. Furthermore, interaction effects are less important in RF and XGB than in EBM. Although six of the ten most important EBM effects are interaction terms, all RF variables and all but three XGB variables are univariate.

To assess the differences in the estimated effects of a variable i between the models, we calculate correlations as well as relative root mean square errors (RMSE) between SHAP values and EBM shape functions. For example, the relative RMSE of variable i between the EBM and SHAP values is given by

$$\text{RMSE}_i = \frac{\sqrt{\frac{1}{|\mathcal{T}_1|}\sum_{(l,t)\in\mathcal{T}_1}\left(\phi_{i,l,t} - f_i(x_{i,l,t})\right)^2}}{\sigma\left(\phi_{i,l,t}\right) + \sigma\left(f_i(x_{i,l,t})\right)}. \tag{10}$$

To account for the size of the individual effects, we scale the RMSE by the sum of the standard deviations of $\phi_{i,l,t}$ and $f_i(x_{i,l,t})$. The RMSEs of interaction effects are computed analogously. Table 7 presents both correlations (Panel A) and the corresponding RMSEs (Panel B) for the ten most important variables of the EBM model. For univariate effects, the correlations between the EBM shape functions f_i and the SHAP values are generally high, indicating agreement between the models on

[31] See Online Appendix O-E.1 for a detailed explanation of SHAP calculations. To be consistent with the EBM interactions, we use $\phi_{ij,l,t}$ to denote the full interaction effect in the following, i.e. the sum of the symmetric SHAP interaction values assigned to the pairs (i, j) and (j, i).

[32] We thank the referee for pointing out this property.

Table 6: Top 10 variables

	EBM	Random Forest	XGBoost
1	ΔUNCf	ΔUNCf	ΔUNCf
2	ΔUNC	levmkt	spread
3	TERM	ΔUNC	UNC
4	LTREVB	TERM	ret
5	ΔUNCf, spread	UNC	ΔUNC
6	ΔUNCf, INTCAP	spread	ΔUNCf, spread
7	ΔUNCf, sprxrtg	ΔUNCr	ΔUNCf, UNCf
8	ΔUNCf, UNC	LTREVB	VIX
9	ΔUNCf, UNCr	EPUT	ΔUNCf, UNC
10	ΔUNCf, EPU	VIX	MOMB

Note: This table presents the top 10 EBM, Random Forest, and XGBoost effects. Variables are ranked by the mean absolute score for EBM and by the sum of absolute SHAP values across observations in the training sample for Random Forest and XGBoost.

the direction and functional form of the most influential effects. In contrast, interaction effects show considerably lower correlations. Some interactions even yield negative values. These low correlations suggest substantial divergence in the interaction effect structures across models.[33]

The RMSE results in Panel B align with the correlation patterns. For univariate effects, the RMSE of the EBM shape functions f_i and the Random Forest SHAP values $\phi^{\mathrm{RF}}_{i,l,t}$ are smaller than those between EBM and XGBoost and between RF and XGB, respectively. In contrast, the RMSE of the XGBoost model relative to the other two models is substantially larger for important predictors. For interactions, the RMSE of RF SHAP values $\phi^{\mathrm{RF}}_{ij,l,t}$ and the shape functions of EBM are larger than those of the univariate variables.

Figure 13 illustrates these differences for ΔUNCf (Panel A) and spread (Panel B). The figure shows the SHAP values $\phi^{\mathrm{RF}}_{i,l,t}$ (in orange) and $\phi^{\mathrm{XGB}}_{i,l,t}$ (in blue) for all observations in the test sample. For comparison, the figure also plots the EBM shape function f_i (in black). The observations of ΔUNCf and spread are on the respective x-axes. Consider first the plot for ΔUNCf in Panel A. Since ΔUNCf is a market-level variable, each month corresponds to a vertical set of points given by the SHAP values $\phi_{i,l,t}$ of the l firms in month t. For example, the lowest observed ΔUNCf of -3.61 occurred in May 2020. Although $\phi_{i,l,t}$ are SHAP values of the univariate effect of variable i, they can differ for different firms l and l' even when $x_{i,l,t} = x_{i,l',t}$. This variability arises because SHAP values provide local explanations that reflect the marginal contribution of a given feature for each individual observation. As such, they depend not only on the value of that specific feature but also on the values of all other features used in the model. In tree-based models like RF and XGB, interactions across features mean that the same value of $x_{i,l,t}$ can yield different SHAP contributions depending on the combination of other covariates in the feature vector. As a result, SHAP values introduce a degree

[33]We find similar results when calculating correlations across all data points (see Online Appendix O-E.3).

Table 7: Top 10 effects – Comparison of EBM and SHAP

	EBM vs. RF	EBM vs. XGB	RF vs. XGB
	Panel A: Correlations		
ΔUNCf	0.965	0.898	0.905
ΔUNC	0.995	0.999	0.995
TERM	0.981	0.973	0.971
LTREVB	0.971	0.665	0.749
ΔUNCf & spread	0.357	0.415	0.561
ΔUNCf & INTCAP	0.493	0.212	0.499
ΔUNCf & sprxrtg	0.357	—	—
ΔUNCf & UNC	0.491	0.518	0.710
ΔUNCf & UNCr	0.370	−0.033	−0.212
ΔUNCf & EPU	0.083	−0.342	0.168
	Panel B: RMSE		
ΔUNCf	0.146	0.397	0.439
ΔUNC	0.052	0.219	0.207
TERM	0.145	0.472	0.388
LTREVB	0.190	0.869	0.894
ΔUNCf, spread	0.575	0.747	0.722
ΔUNCf, INTCAP	0.512	0.629	0.503
ΔUNCf, sprxrtg	0.686	1.046	1.003
ΔUNCf, UNC	0.518	0.889	0.829
ΔUNCf, UNCr	0.576	0.950	0.991
ΔUNCf, EPU	0.767	0.833	0.821

Note: This table presents correlations (Panel A) and root mean squared errors (Panel B), defined in (10), between EBM outputs $f_i(x_{i,l,t})$ and SHAP decompositions $\phi_{i,l,t}$ for the top 10 EBM effects. Terms are ordered by their EBM importance ΔR^2. Missing values are due to constant values in the XGBoost SHAP explanations.

of heterogeneity or apparent noise, even for identical values of a single feature. For example, the RF and XGB SHAP values in May 2020 range from 45.5 to 78.7, and 114.9 to 133.8, respectively, with means across all l of 65.5 and 124.6. In contrast, the estimated shape function of the EBM $f_i(x_{i,l,t})$ has a unique value for a given observation $x_{i,l,t}$, independent of the values of other features. That is, EBM predictions are based on fixed, global shape functions, so there is no distinction between local and global explanations. For $x_{i,l,t} = -3.62$ in May 2020, $f_i(x_{i,l,t}) = 64.3$, which is close to the average RF effect but significantly smaller than the mean XGB effect. The plot shows that the mean RF SHAP values are close to the EBM shape function for all but the largest ΔUNCf observations.

Panel B shows the same plot for the firm-level variable spread. The estimated effect of spread on future returns in the XGBoost model is substantially larger than for the EBM and Random Forest models. The mean absolute value of the XGB SHAP value is 30.9. Since the variables are normalized, XGB implies that a one-standard deviation increase leads to a 33.3 bps increase in expected bond returns. The mean spread effects of the EBM and RF models are 1.6 and 6.5 bps, respectively.

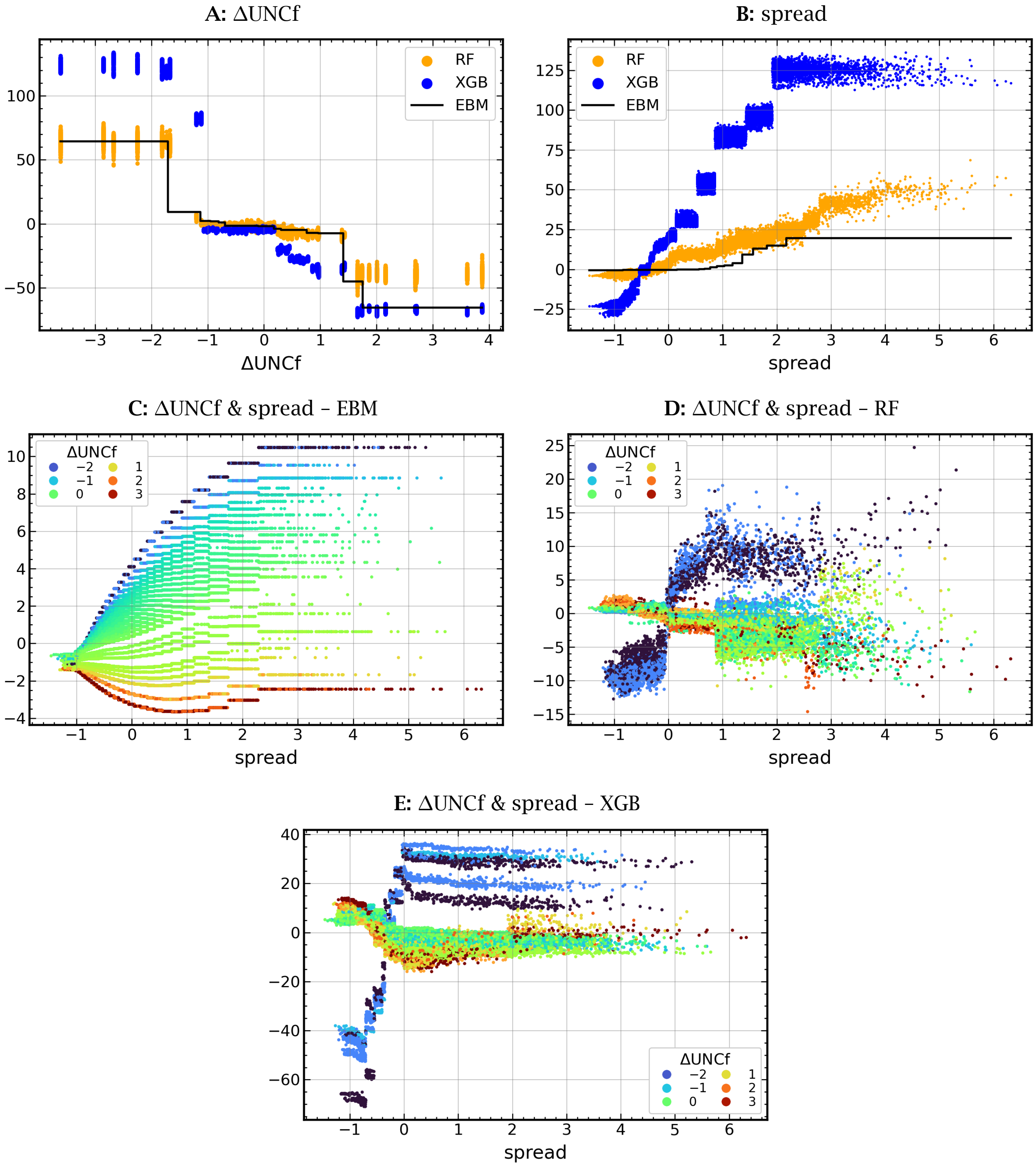
Panels C, D, and E illustrate the $(i, j) =$ (ΔUNCf, spread) interaction effects. Panel C shows a scatter plot of the EBM estimates $f_{ij}(x_{i,l,t}, x_{j,l,t})$ for all observations in the test sample. Observations with different values of ΔUNCf are indicated by different colors (with low (high) ΔUNCf in blue (red) colors). If ΔUNCf is negative, higher spreads have a large positive effect on future returns. However, the magnitude of this effect decreases for higher values of ΔUNCf and reverses for the highest ΔUNCf observations. The SHAP values $\phi_{ij,l,t}$ of the Random Forest and XGBoost models plotted in Panels D and E show a less clear pattern. Although there is a positive slope for negative ΔUNCf observations, the effect of spread on future returns is not monotonic, suggesting that the SHAP values are subject to estimation and/or approximation noise.

6.1 Portfolios

Similarly to the portfolio sorting in section 5, we form monthly decile portfolios on the main effects of EBM and SHAP and record their returns over each month $t + 1$. We also compute the corresponding 10-1 long-short portfolios and test for monotonicity of average returns and Sharpe ratios across deciles. To take advantage of cross-sectional (firm-level) interactions, we relax the restrictions on interaction terms in our main EBM model and refit the model, while allowing it to potentially learn the full set of interactions instead of the default of 40 selected pairs. We provide a more detailed analysis of this model in section 6.2.

Table 8 reports the out-of-sample performance of the best-performing long-short portfolios according to their Sharpe ratio in the training and validation samples. We observe similar magnitudes when comparing the highest Sharpe ratios of the three models, suggesting a similar explanatory power for the three decompositions. Although the Random Forest's long-short Sharpe ratio peak is the highest at 0.37, EBM produces more consistently high returns and Sharpe ratios for more characteristics. Table O-E.2 shows selected portfolios sorted into deciles, which allows us to analyze the

Figure 13: SHAP values – ΔUNCf, spread



Note: This figure compares EBM effects and SHAP values associated with the RF and XGB models. Panels A and B show the shape function $f_i(x_{i,l,t})$ and the corresponding SHAP values $\phi_{i,l,t}$ for ΔUNCf and spread, respectively. Panel C shows the interaction of ΔUNCf and spread by plotting $f_{ij}(x_{i,l,t}, x_{j,l,t})$ for all observations in the test sample. Panels D and E present the corresponding SHAP values $\phi_{ij,l,t}$ for RF and XGB, respectively.

monotonicity of the portfolio relationships. We report the underlying decile portfolios for the best-performing long-short portfolio of each model. In the selected examples, we only observe monotonic patterns for the Random Forest's market equity portfolios, although returns and Sharpe ratios are monotonic in some other EBM shape functions' outputs, e.g. spread.

Next, we extend this approach to the interaction terms. For each pair (i, j) of bond or firm characteristics, we obtain $\phi_{ij,l,t}$ values for each firm-month observation. Once again, we observe monthly returns and Sharpe ratios of each decile bin over the sample period. Analogously, we form decile sorts for all EBM's $f_{i,j}(x_{i,l,t}, x_{j,l,t})$. Note that we use the EBM model with the full set of interactions in order to compare all pairs of characteristics. The highest Sharpe ratios are comparable for EBM and XGBoost, while Random Forest seems to produce interaction portfolios with weaker out-of-sample performance.

Overall, both EBM and SHAP decompositions seem to discover effects of similar order of magnitude but exhibit little agreement in the variables to which these effects are attributed, especially for interactions. Although the discovered patterns are not always monotonic across decile portfolios, taking both univariate effects and interactions into account, EBM most consistently provides high Sharpe ratio portfolios.

6.2 Decompositions of fit

In our main analysis, we find that the most import interactions involve market-level variables. This raises the question whether EBM's nonlinearities and interactions truly capture firm-level cross-sectional predictability and whether firm-level interactions are economically meaningful. We address these questions by comparing the relative magnitude of univariate and interaction effects for all models. As the main model is restricted to 40 interactions by default, we refit the model without any restriction on the number of interactions, allowing it to model the full set of $\frac{81\times80}{2} = 3240$ interaction pairs. We find that the expanded EBM model with all interactions achieves slightly higher out-of-sample performance compared to the main model, reaching an R^2 of 12.41%. At the same time, the EBM portfolio Sharpe ratio on the test data rises substantially to 0.53 (0.58) for value (equal) weights. This indicates that the additional firm-level information contained in the newly found interactions increases coss-sectional predictive power and is economically meaningful. It complements the high Sharpe ratios of firm-level interaction portfolios we observe in section 6.1.

The mean absolute scores $S(i)$ and $S(i, j)$, defined in (4) and (5), of the EBM model measure the average effects of the univariate and interaction terms across firm-month observations. The total score of the model is given by $S = \sum_i S(i) + \sum_{i>j} S(i, j)$. The first column of Table 9 decomposes S into various subsets of univariate and interaction effects. Univariate effects account for 11% of the total EBM score, while interactions account for 89%, indicating that interaction terms are crucial for model fit. Next, we decompose the total score into firm-level effects, which include firm-level univariate variables and interaction terms with only firm-level variables, market-level effects (market variables and interaction terms with only market variables), and interaction terms that

Table 8: Returns of long-short portfolios sorted on EBM and SHAP effects

Panel A: EBM – univariate

	sprd2d	sprxrtg	spread	yield	age	voleq	skew	levmkt	retlag	ret
Mean	0.33	0.51	0.61	0.39	0.10	0.18	0.15	0.09	0.16	0.17
S.D.	1.03	1.71	2.06	1.45	0.44	1.24	1.03	0.66	1.64	1.92
SR	0.32	0.30	0.30	0.27	0.23	0.15	0.14	0.13	0.10	0.09

Panel B: Random Forest – univariate

	me	amtout	nrtg	coupon	dolvol	beme	maturity	offramt	sales	duration
Mean	0.37	0.24	0.33	0.10	0.07	0.04	0.06	0.07	0.03	0.03
S.D.	1.00	0.66	1.07	0.75	0.59	0.51	0.77	1.09	0.74	0.66
SR	0.37	0.35	0.31	0.13	0.12	0.08	0.08	0.06	0.05	0.04

Panel C: XGBoost – univariate

	assets	spread	momspr	nrtg	ret	offramt	atme	sprxrtg	voleq	maturity
Mean	0.15	0.57	0.21	0.20	0.30	0.24	0.15	0.05	0.07	0.04
S.D.	0.50	1.93	1.15	1.13	1.65	1.35	0.91	0.47	1.08	0.68
SR	0.30	0.30	0.19	0.18	0.18	0.18	0.17	0.11	0.07	0.06

Panel D: EBM – interactions

	profitbl, ret	ret, volbond	maturity, ret	duration, ret	ret, var	nime, ret	dureq, ret	ret, salesat	ret, skew	mom, ret
Mean	0.37	0.49	0.41	0.36	0.39	0.35	0.31	0.30	0.30	0.22
S.D.	0.82	1.09	1.01	0.92	1.36	1.27	1.12	1.29	1.45	1.71
SR	0.46	0.45	0.41	0.40	0.28	0.28	0.27	0.23	0.21	0.13

Panel E: Random Forest– interactions

	totaldebt, yield	dolvol, sprd2d	duration, spread	mom, ret	momspr, sales	levop, spread	assets, ret	levmkt, yield	ret, sales	debtebitda, spread
Mean	0.25	0.07	0.09	0.11	0.06	0.04	0.04	0.02	0.02	0.01
S.D.	0.85	0.62	0.80	1.22	0.79	0.57	0.62	0.51	0.94	0.86
SR	0.30	0.11	0.11	0.09	0.07	0.07	0.06	0.04	0.02	0.01

Panel F: XGBoost– interactions

	beme, spread	spread, volbond	atme, spread	d2d, yield	atbe, spread	nime, ret	nime, offramt	offramt, yield	amtout, dolvol	amtout, ret
Mean	0.31	0.38	0.37	0.22	0.18	0.11	0.09	0.16	0.07	0.07
S.D.	0.72	0.96	1.07	1.01	0.81	0.58	0.48	0.97	0.54	0.67
SR	0.43	0.39	0.35	0.22	0.22	0.19	0.19	0.16	0.12	0.10

Note: This table reports monthly means, standard deviations, and Sharpe ratios of 10-1 long-short portfolios sorted on EBM outputs and SHAP values from Random Forest and XGBoost models for univariate characteristics in Panels A to C and interaction variables in Panels D to F. We identify the top ten portfolios by their Sharpe ratios in the combined training and validation samples and report statistics of the ten portfolios in the testing sample.

involve a firm-level and a market-level variable. Firm-level variables account for 17% of the total score, the market-level variables account for 36%, but the largest share of 47% is due to mixed interactions between the firm-level and market-level variables. Finally, we decompose the total score into univariate/interaction-firm/market components. Univariate firm-level and market-level variables account for only 1% and 10% of the score, respectively. The effects of the interaction of the variables at the firm level account for 16%, while the interactions at the market level contribute 26%.

Table 9: Decompositions of effects

Variables	EBM	RF	XGB
Univariate	0.11	0.34	0.42
Interactions	0.89	0.66	0.58
Firm	0.17	0.17	0.18
Mkt	0.36	0.53	0.54
Firm/mkt	0.47	0.30	0.28
Univariate-firm	0.01	0.09	0.13
Univariate-mkt	0.10	0.25	0.29
Interaction-firm	0.16	0.08	0.05
Interaction-mkt	0.26	0.29	0.25
Interaction-firm/mkt	0.47	0.30	0.28

Note: This table shows the decompositions of the total score of the EBM model S and the total SHAP values ϕ of the RF and XGB models.

Next, we compute a similar decomposition for SHAP values of the Random Forest and XGBoost models. In analogy to mean absolute EBM scores, we define the total SHAP value ϕ as

$$\begin{aligned}
\phi_i &= \sum_{(l,t)\in\mathcal{T}_1} w_{l,t}\,\left|\phi_{i,l,t}\right|, \\
\phi_{i,j} &= \sum_{(l,t)\in\mathcal{T}_1} w_{l,t}\,\left|\,\phi_{ij,l,t}\right|, \\
\phi &= \sum_i \phi_i + \sum_{i>j} \phi_{i,j},
\end{aligned}$$

where $w_{l,t}$ are sample weights. The decompositions of ϕ for RF and XGB are reported in the second and third columns of Table 9. Compared to the EBM model, interaction effects contribute less to the fit of RF and XGB than in the EBM model. In contrast, market-level variables are relatively more important, accounting for more than half of ϕ. Compared to SHAP-based estimates, EBM assigns greater importance to interaction effects, particularly those involving firm-level interactions.

7 Simulations

Next, we compare the EBM, Random Forest, and XGBoost estimation methodologies using a Monte Carlo simulation. We assume that the data are generated by a nonlinear factor model of the form similar to Gu et al. (2020). Following the notation in (1), bond returns are given by

$$r_{l,t+1} \;=\; g(\mathbf{x}_{l,t}) \;+\; \mathbf{z}'_{l,t}\,\mathbf{v}_{t+1} \;+\; \varepsilon_{l,t+1}, \tag{11}$$

where $\mathbf{x}_{l,t} = (\mathbf{z}_{l,t}, y_t)'$. $g(\mathbf{x}_{l,t})$ is given by a nonlinear 4-factor model with two univariate and two interaction factors:[34]

$$g(\mathbf{x}_{l,t}) = \left(z^2_{l,t,1}, y_t, z_{l,t,1} \times y_t,\ \text{sign}\left(z_{l,t,2} \times y_t\right)\right) \boldsymbol{\theta}, \quad \boldsymbol{\theta} = (0.04, 0.01, 0.05, 0.012)'.$$

The observed data include the true factors, as well as additional variables that do not affect bond returns. In addition to Gu et al. (2020), we assume that the data set has "copies" of the true factors. Copies are created by adding independent noise to the true factors with the standard deviation chosen so that their correlations are 0.9. The copies and remaining characteristics represent pure noise and do not affect the DGP. This additional complication allows us to assess how each model handles multicollinearity in practice. Specifically,

1. $\mathbf{z}_{l,t}$ is an $L \times N_z$ matrix of simulated N_z characteristics of L firms. The raw characteristics $c_{l,t,i}$ follow AR(1) processes with different degrees of persistence: $c_{l,t,i} = \rho_i c_{l,t-1,i} + \epsilon_{i,l,t}$, where $\rho_i \sim \mathcal{U}[0.9, 1]$, $\epsilon_{lt,i} \sim \mathcal{N}(0, 1-\rho^2)$. In each t, $z_{l,t,i}$ are rank-normalized $c_{l,t,i}$ values and transformed to the $[-1, 1]$ interval.
2. The first two elements of $\mathbf{z}_{l,t}$ are the true factors. The third and fourth characteristics $z_{l,t,3}, z_{l,t,4}$ are copies of the true factors given by $z_{l,t,3} = z_{l,t,1} + w_{l,t,1}$ and $z_{l,t,4} = z_{l,t,2} + w_{l,t,2}$. $w_{l,t,1}$ and $w_{l,t,2}$ are normally distributed with standard deviations chosen so that the correlations of the copies with the true factors are 0.9.
3. y_t is a persistent univariate time series: $y_t = \rho y_{t-1} + u_t,\ \ u_t \sim \mathcal{N}(0, 1-\rho^2)$ where $\rho = 0.95$.
4. $\mathbf{v}_{t+1} \sim \mathcal{N}(0, \sigma_v^2 \times \mathbf{I}_2)$ is a 2×1 vector of factor shocks.
5. $\varepsilon_{l,t+1} \sim t_5(0, \sigma_u^2)$ is an uncorrelated i.i.d. idiosyncratic error.

We set $L = 500$, $T = 250$, and $N_z = 50$ to match our data set.[35] Finally, σ_v and σ_u are chosen to yield a cross-sectional R^2 of 50% and a time series R^2 of 12%. The following results are based on 100 simulations.

For each simulated data set, we divide the sample into training (50% of the sample), validation (25%), and test subsamples (25%). We fit EBM, Random Forest, and XGBoost models and apply standard hyperparameter tuning following the procedure used in the empirical estimation. We obtain EBM and SHAP decompositions for all data points (l, t), yielding main effects $f_i(x_{i,l,t})$ and $\phi_{i,l,t}$, and

[34] Given the importance of macroeconomic variables in our data set, we add an y_t term to the nonlinear terms in Gu et al. (2020).

[35] Our simulations contain only 500 firms even though the data set includes 1,207 firms to keep the computation time manageable. Simulations with fewer runs suggest that the results do not change significantly when the number of firms is increased beyond 500.

Table 10: Simulation results

	EBM	RF	XGB
	Panel A: R^2		
Train	14.25	12.88	15.90
Valid	10.69	9.42	10.28
Test	9.96	8.49	8.65
	Panel B: Term selection		
z_1	1.00	1.00	1.00
y	0.90	0.98	0.97
(z_1, y)	1.00	1.00	0.99
(z_2, y)	1.00	0.99	0.97
	Panel C: RMSE		
z_1	1.47	1.56	1.50
y	0.18	0.30	0.25
(z_1, y)	0.84	1.05	1.13
(z_2, y)	0.64	0.79	0.77

Note: This table reports aggregate measures for EBM and the two SHAP decompositions over 100 simulation runs. The decompositions are compared based on model fit (R^2), importance rankings, and term-level RMSE.

interaction effects $f_{i,j}(x_{i,l,t}, x_{j,l,t})$ and $\phi_{ij,l,t}$. First, we compare R^2 in the training, validation, and test sample. Second, we assess how precisely the models identify the true factors. To this end, we rank terms by mean absolute EBM scores and mean absolute SHAP values within each simulation run and calculate the share of runs where the correct variables are among the top-5 terms. Finally, we calculate how well each model captures the true effects of each component of the DGP. The RMSE of factor i in the training sample is

$$\text{RMSE}_i = \sqrt{\frac{1}{|\mathcal{T}_1|} \sum_{(l,t)\in\mathcal{T}_1} \left(h_i(\mathbf{x}_{l,t}) - g_i(\mathbf{x}_{l,t})\right)^2},$$

where $h_i(\mathbf{x}_{l,t})$ is either the corresponding EBM shape function or the RF and XGB SHAP values.

The results are reported in Table 10. The estimated R^2 in the test sample in Panel A are slightly lower than the true time series R^2 of 12% and range from 8.49% for Random Forest to 9.96% for EBM, suggesting that the models are able to extract most but not all of the forecastable components of returns. In addition, all three estimation methods almost always select the correct variables that drive the DGP. As shown in Panel B, the lowest detection percentage is 90% for the main effect of EBM of y and 97% for the interaction term (z_2, y) in XGBoost. Finally, Panel C shows that EBM achieves the lowest RMSE for all terms, indicating that it captures the true effects more accurately than both SHAP decompositions. Especially for the two interaction terms, EBM consistently exhibits lower errors.

To illustrate the differences in fit, we plot the estimated shape functions and SHAP values (averaged across simulations) in Figure 14.[36] Panels A and B show the univariate effects of z_1 and y, respectively. The first factor in the DGP is quadratic in z_1, $g_1(z_1) = \theta_1 z_1^2$, and is shown in dashed blue. The estimated shapes of the EBM, RF, and XGBoost are shown in black, gray, and orange, respectively. All three models fit the nonlinear effect reasonably well. The second factor is a linear function of y, $g_2(y) = \theta_2 y$, and is also captured well by the models, although the fit of EBM is slightly better than that of RF and XGB. Panels C to F show the results for the interaction effects involving (z_1, y) and (z_2, y). To illustrate the 2-dimensional function, we plot conditional effects as exemplary cuts through the resulting interaction planes, holding each variable constant at a fixed level. Since $g_3(z_1, y) = \theta_3 z_1 y$, the conditional interaction effects are linear in z_1 and y and shown in dashed blue in Panels A and B. The estimated EBM shape function (in black) is close to the true DGP when z_1 is fixed while the SHAP values of RF and XGB are less steep and thus underestimate the true effect of y on future returns. Since $g_4(z_2, y) = \theta_4 \operatorname{sign}(z_2 y)$, the conditional interaction effects are nonlinear in z_2 and y. The EBM shape function (in black) captures the nonlinearity reasonably well. In contrast, SHAP decompositions tend to underestimate the magnitude of the interaction effects and, in some cases, even assign the wrong direction to them, leading to a worse fit. We conclude that the differences between the three estimation methods are relatively small for univariate effects. Differences are somewhat larger for interaction effects and EBM recovers the true nonlinear effects better than Random Forest and XGBoost.[37]

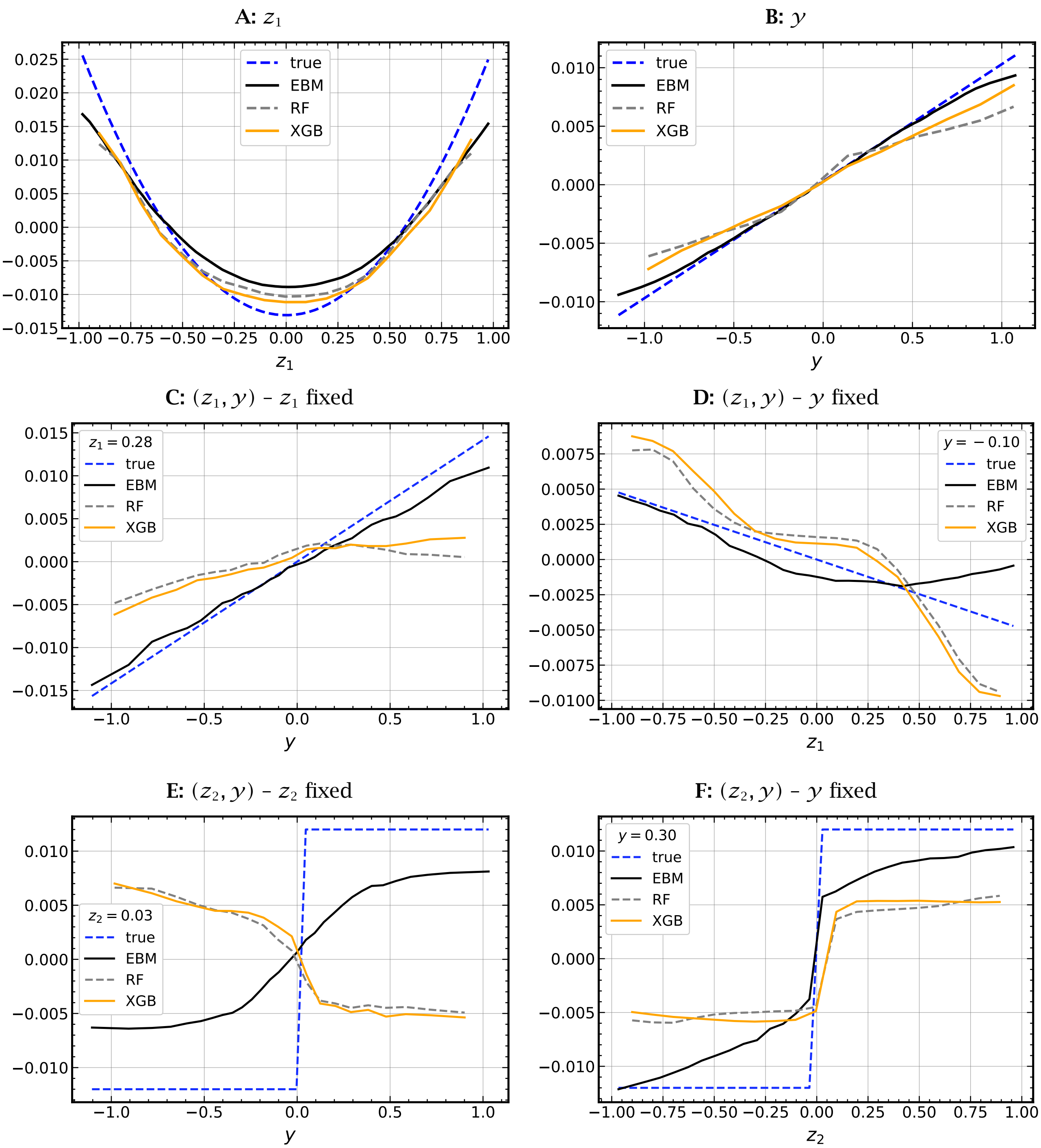
8 Conclusion

Due to the complex structure of asset prices, advanced machine learning algorithms often outperform linear methods in return prediction. However, the ability to process higher complexity typically comes at the expense of interpretability. This greatly limits the utility of many machine-learning models for investors and financial researchers. We show that state-of-the-art predictive performance can be combined with interpretability when inherently explainable glass box methods, such as the Explainable Boosting Machine, are used instead to predict returns in the corporate bond market. In contrast to black box methods, the interpretable model allows us to study the relationships underlying corporate bond returns. We find that nonlinearities and variable interactions are crucial to the fit of our model. The most important market-level predictors are related to financial and macroeconomic uncertainty, and the term structure. The effect of all three variables is nonlinear and especially

[36] We present SHAP estimates as continuous curves, representing bin-wise averages across simulations. However, note that SHAP values remain local explanations, resulting in discrete point estimates for each observation, as shown in Figure 13.

[37] In addition to this simulation case, we consider the two data-generating processes of Gu et al. (2020). The results, which are reported in Appendix O-F.2, confirm that while all models capture both DGPs with similar R^2, EBM identifies the truly influential components more consistently and captures the interaction effects more accurately. The outperformance of EBM over the SHAP decompositions is even more pronounced than in the above case, as indicated by considerably lower errors and higher correlations with the DGP for interaction terms. The corresponding shape functions also illustrate how the SHAP interaction values systematically underestimate the true magnitude of the underlying effects.

Figure 14: Comparison of fitted effects



Note: This figure presents conditional contributions derived from the mean shape functions over 100 simulation runs. Panels A and B show the univariate effects of z_1 and x, respectively. Panels C to F show interaction effects by separately holding each variable involved in the interaction constant at fixed levels. The three model decompositions are compared with the underlying data-generating process based on exemplary cuts through the interactions (z_1, y) (Panel A and B) and (z_2, y) (Panel C and D).

pronounced for large realizations towards the tails of the underlying data distribution (and large declines in financial uncertainty). For more "normal" realizations, the effects are close to zero. In addition, bond characteristics such as spread and interactions involving uncertainty factors possess considerable explanatory power.

When comparing firms based on monthly sorts on market equity, we find that the impact of uncertainty is amplified for the returns of small firms compared to big firms. Further subsample analysis reveals similar disparities between bonds with long and short maturities and between bonds with high and low default risk. Sensitivity to changes in financial uncertainty varies asymmetrically, emphasizing large increases.

Moreover, we show how our approach can add ex-ante explainability to portfolio construction. We find that a long-short portfolio based on model predictions can outperform all benchmarks while being explainable through the transparent asset selection performed by the glass box model. As the final simulation analysis confirms, this method offers a more accurate representation of the interaction structure underlying returns compared to post-hoc explainability techniques.

References

Asness, Clifford S., Andrea Frazzini, and Lasse Heje Pedersen (2019) "Quality minus junk," *Review of Accounting Studies*, 24 (1), 34–112.

Bai, Jennie, Turan G. Bali, and Quan Wen (2016) "Do the distributional characteristics of corporate bonds predict their future returns?" *Working Paper*.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016) "Measuring economic policy uncertainty," *The Quarterly Journal of Economics*, 131 (4), 1593–1636.

Bali, Turan G, Heiner Beckmeyer, Mathis Moerke, and Florian Weigert (2023) "Option return predictability with machine learning and big data," *The Review of Financial Studies*, 36 (9), 3548–3602.

Bali, Turan G., Amit Goyal, Dashan Huang, Fuwei Jiang, and Quan Wen (2022a) "Predicting Corporate Bond Returns: Merton Meets Machine Learning," *Swiss Finance Institute, Research Paper Series 20–110*.

Bali, Turan G., Avanidhar Subrahmanyam, and Quan Wen (2021a) "Long-term reversals in the corporate bond market," *Journal of Financial Economics*, 139 (2), 656–677.

——— (2021b) "The macroeconomic uncertainty premium in the corporate bond market," *Journal of Financial and Quantitative Analysis*, 56 (5), 1653–1678.

Bali, Turan G., Avanidhar Subrahmanyam, and Quan Wen (2022b) "Return-based factors for corporate bonds," *Working paper*.

Bartram, Sönke M., Mark Grinblatt, and Yoshio Nozawa (2024) "Book-to-market, mispricing, and the cross-section of corporate bond returns," *Journal of Financial and Quantitative Analysis (Forthcoming)*.

Bharath, Sreedhar T. and Tyler Shumway (2008) "Forecasting default with the Merton distance to default model," *The Review of Financial Studies*, 21 (3), 1339–1369.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (2021) "Bond risk premiums with machine learning," *The Review of Financial Studies*, 34 (2), 1046–1089.

Binsbergen, Jules van, Xiao Han, and Alejandro Lopez-Lira (2023) "Man versus machine learning: The term structure of earnings expectations and conditional biases," *The Review of Financial Studies*, 36 (6), 2361–2396.

Binsbergen, Jules van, Yoshio Nozawa, and Michael Schwert (2024) "Duration-based valuation of corporate bonds," *The Review of Financial Studies (Forthcoming)* .

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu (2023) "Forest through the trees: Building cross-sections of stock returns," *Journal of Finance*, Forthcoming.

Campbell, John Y. and Glen B. Taksler (2003) "Equity volatility and corporate bond yields," *The Journal of Finance*, 58 (6), 2321–2350.

Carhart, Mark M. (1997) "On persistence in mutual fund performance," *The Journal of Finance*, 52 (1), 57–82.

Chen, Luyang, Markus Pelger, and Jason Zhu (2024) "Deep learning in asset pricing," *Management Science*, 70 (2), 714–750.

Chen, Tianqi and Carlos Guestrin (2016) "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu (2022) "Expected returns and large language models," *Working paper*.

Choi, Jaewon and Yongjun Kim (2018) "Anomalies and market (dis)integration," *Journal of Monetary Economics*, 100, 16–34.

Chung, Kee H., Junbo Wang, and Chunchi Wu (2019) "Volatility and the cross-section of corporate bond returns," *Journal of Financial Economics*, 133 (2), 397–417.

Cong, Lin William, Guanhao Feng, Jingyu He, and Xin He (2024) "Growing the efficient frontier on panel trees," *Journal of Financial Economics*, Forthcoming.

Cong, Lin William, Ke Tang, Jingyuan Wang, and Yang Zhang (2022) "AlphaPortfolio: Direct construction through deep reinforcement learning and interpretable AI," *Working Paper*.

Correia, Maria, Johnny Kang, and Scott Richardson (2018) "Asset volatility," *Review of Accounting Studies*, 23, 37–94.

Correia, Maria, Scott Richardson, and İrem Tuna (2012) "Value investing in credit markets," *Review of Accounting Studies*, 17, 572–609.

Dang, Thuy Duong, Fabian Hollstein, and Marcel Prokopczuk (2023) "Which factors for corporate bond returns?" *Review of Asset Pricing Studies*, 13 (4), 615–652.

Dickerson, Alexander, Christian Julliard, and Philippe Mueller (2023a) "The corporate bond factor zoo," *Working paper*.

Dickerson, Alexander, Philippe Mueller, and Cesare Robotti (2023b) "Priced risk in corporate bonds," *Journal of Financial Economics*, 150 (2), 103707.

Dickerson, Alexander, Cesare Robotti, and Giulio Rossetti (2023c) "Noisy prices and return-based anomalies in corporate bonds," *UNSW Business School Research Paper*.

Eisfeldt, Andrea and Gregor Schubert (2024) "AI and Finance," *National Bureau of Economic Research*.

Elkamhi, Redouane, Chanik Jo, and Yoshio Nozawa (2024) "A one-factor model of corporate bond premia," *Management Science*, 70 (3), 1875–1900.

Fallahgoul, Hasan, Vincentius Franstianto, and Xin Lin (2024) "Asset pricing with neural networks: Significance tests," *Journal of Econometrics*, 238 (1), 105574.

Fama, Eugene F. and Kenneth R. French (1992) "The cross-section of expected stock returns," *The Journal of Finance*, 47 (2), 427–465.

——— (1993) "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33 (1), 3–56.

——— (2015) "A five-factor asset pricing model," *Journal of Financial Economics*, 116 (1), 1–22.

Fedyk, Anastassia, Ali Kakhbod, Peiyao Li, and Ulrike Malmendier (2024) "ChatGPT and perception biases in investments: An experimental study," *Working paper*.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu (2020) "Taming the factor zoo: A test of new factors," *The Journal of Finance*, 75 (3), 1327–1370.

Feng, Guanhao, Jingyu He, Nicholas Polson, and Jianeng Xu (2023) "Deep learning in characteristics-sorted factor models," *Journal of Financial and Quantitative Analysis*, 1–36.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber (2020) "Dissecting characteristics nonparametrically," *The Review of Financial Studies*, 33 (5), 2326–2377.

Gamba, Andrea and Alessio Saretto (2019) "Firm policies and the cross-section of CDS spreads," *WBS Finance Group Research Paper no. 191*.

Gebhardt, William R., Soeren Hvidkjaer, and Bhaskaran Swaminathan (2005) "The cross-section of expected corporate bond returns: Betas or characteristics?" *Journal of Financial Economics*, 75 (1), 85–114.

Gomes, Joao and Lukas Schmid (2021) "Equilibrium asset pricing with leverage and default," *The Journal of Finance*, 76 (2), 977–1018.

Griffin, J. M., N. Hirschey, and S. Kruger (2023) "Do municipal bond dealers give their customers "fair and reasonable" pricing?" *Journal of Finance*, 78 (2), 887–934.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020) "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 33 (5), 2223–2273.

——— (2021) "Autoencoder asset pricing models," *Journal of Econometrics*, 222 (1), 429–450.

Hastie, Trevor and Robert Tibshirani (1986) "Generalized Additive Models," *Statistical Science*, 1 (3), 297 – 310.

He, Zhiguo, Bryan Kelly, and Asaf Manela (2017) "Intermediary asset pricing: New evidence from many asset classes," *Journal of Financial Economics*, 126 (1), 1–35.

Ho, Tin Kam (1995) "Random decision forests," in *IEEE Proceedings of 3rd international conference on document analysis and recognition*, 1, 278–282.

Houweling, Patrick and Jeroen van Zundert (2017) "Factor investing in the corporate bond market," *Financial Analysts Journal*, 73 (2), 100–115.

Israel, Ronen, Diogo Palhares, and Scott A. Richardson (2018) "Common factors in corporate bond returns," *Journal of Investment Management*, 16 (2), 17–46.

Jegadeesh, Narasimhan and Sheridan Titman (1993) "Returns to buying winners and selling sosers: Implications for stock market efficiency," *The Journal of Finance*, 48 (1), 65–91.

Jorion, Philippe and Gaiyan Zhang (2009) "Credit contagion from counterparty risk," *The Journal of Finance*, 64 (5), 2053–2087.

Jostova, Gergana, Stanislava Nikolova, Alexander Philipov, and Christof W. Stahel (2013) "Momentum in corporate bond returns," *The Review of Financial Studies*, 26 (7), 1649–1693.

Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng (2015) "Measuring uncertainty," *American Economic Review*, 105 (3), 1177–1216.

Kakhbod, Ali, Leonid Kogan, Peiyao Li, and Dimitris Papanikolaou (2024) "Measuring Creative Destruction," *SSRN working paper 5008685*.

Kelly, Bryan, Diogo Palhares, and Seth Pruitt (2023) "Modeling corporate bond returns," *The Journal of Finance*, 78 (4), 1967–2008.

Kelly, Bryan, Seth Pruitt, and Yinan Su (2019) "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 134 (3), 501–524.

Kelly, Bryan and Dacheng Xiu (2023) "Financial machine learning," *Foundations and Trends in Finance*, 13 (3-4), 205–363.

Kiss, Orsolya, Fiona C Baker, Robert Palovics, Erin E Dooley, Kelley Pettee Gabriel, and Jason M Nagata (2024) "Using explainable machine learning and fitbit data to investigate predictors of adolescent obesity," *Scientific Reports*, 14 (1), 12563.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2020) "Shrinking the cross-section," *Journal of Financial Economics*, 135 (2), 271–292.

Lettau, Martin (2023) "High-dimensional factor models and the factor zoo," *National Bureau of Economic Research working paper no. 31719*.

Lettau, Martin and Markus Pelger (2020) "Factors that fit the time series and cross-section of stock returns," *The Review of Financial Studies*, 33 (5), 2274–2325.

Lin, Hai, Junbo Wang, and Chunchi Wu (2011) "Liquidity risk and expected corporate bond returns," *Journal of Financial Economics*, 99 (3), 628–650.

Lintner, John (1965) "Security prices, risk, and maximal gains from diversification," *The Journal of Finance*, 20 (4), 587–615.

Lopez-Lira, Alejandro and Nikolai L Roussanov (2020) "Do Common Factors Really Explain the Cross-Section of Stock Returns?" *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker (2013) "Accurate intelligible models with pairwise interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 623–631.

Ludvigson, Sydney C., Sai Ma, and Serena Ng (2021) "Uncertainty and business cycles: exogenous impulse or endogenous response?" *American Economic Journal: Macroeconomics*, 13 (4), 369–410.

Lundberg, Scott M, Gabriel Erion, Hugh Chen et al. (2020) "From local explanations to global understanding with explainable AI for trees," *Nature machine intelligence*, 2 (1), 56–67.

Lundberg, Scott M. and Su-In Lee (2017) "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Advances in Neural Information Processing Systems, 4768–4777.

Molnar, Christoph (2020) *Interpretable machine learning*, Lulu.com.

Molnar, Christoph, Gunnar König, Julia Herbinger et al. (2020) "General pitfalls of model-agnostic interpretation methods for machine learning models," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 39–68.

Murray, Scott, Yusen Xia, and Houping Xiao (2024) "Charting by machines," *Journal of Financial Economics*, 153, 103791.

Nagel, Stefan (2021) *Machine Learning in Asset Pricing*, Princeton University Press: Princeton.

Nazemi, Abdolreza and Frank J Fabozzi (2024) "Interpretable machine learning for creditor recovery rates," *Journal of Banking & Finance*, 164, 107187.

Nori, Harsha, Samuel Jenkins, Paul Koch, and Rich Caruana (2019) "InterpretML: A unified framework for machine learning interpretability," *arXiv:1909.09223*.

Nozawa, Yoshio, Yancheng Qiu, and Yan Xiong (2023) "Disagreement and bond PEAD," *Working Paper, University of Toronto*.

Pástor, Lúboš and Robert F. Stambaugh (2003) "Liquidity risk and expected stock returns," *Journal of Political Economy*, 111 (3), 642–685.

Rudin, Cynthia (2019) "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 1 (5), 206–215.

Shapley, LS (1953) "A VALUE FOR n-PERSON GAMES," *Contributions to the Theory of Games* (28), 307.

Sharpe, William F. (1964) "Capital asset prices: A theory of market equilibrium under conditions of risk," *The Journal of Finance*, 19 (3), 425–442.

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju (2020) "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.

Wagner, Nils, Muhammed H Çelik, Florian R Hölzlwimmer, Christian Mertes, Holger Prokisch, Vicente A Yépez, and Julien Gagneur (2023) “Aberrant splicing prediction across human tissues,” *Nature Genetics*, 55 (5), 861–870.

Wang, Caroline, Bin Han, Bhrij Patel, and Cynthia Rudin (2023) “In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction,” *Journal of Quantitative Criminology*, 39 (2), 519–581.

Appendix

A Tables and figures

Table A.1: Bond/Firm characteristics

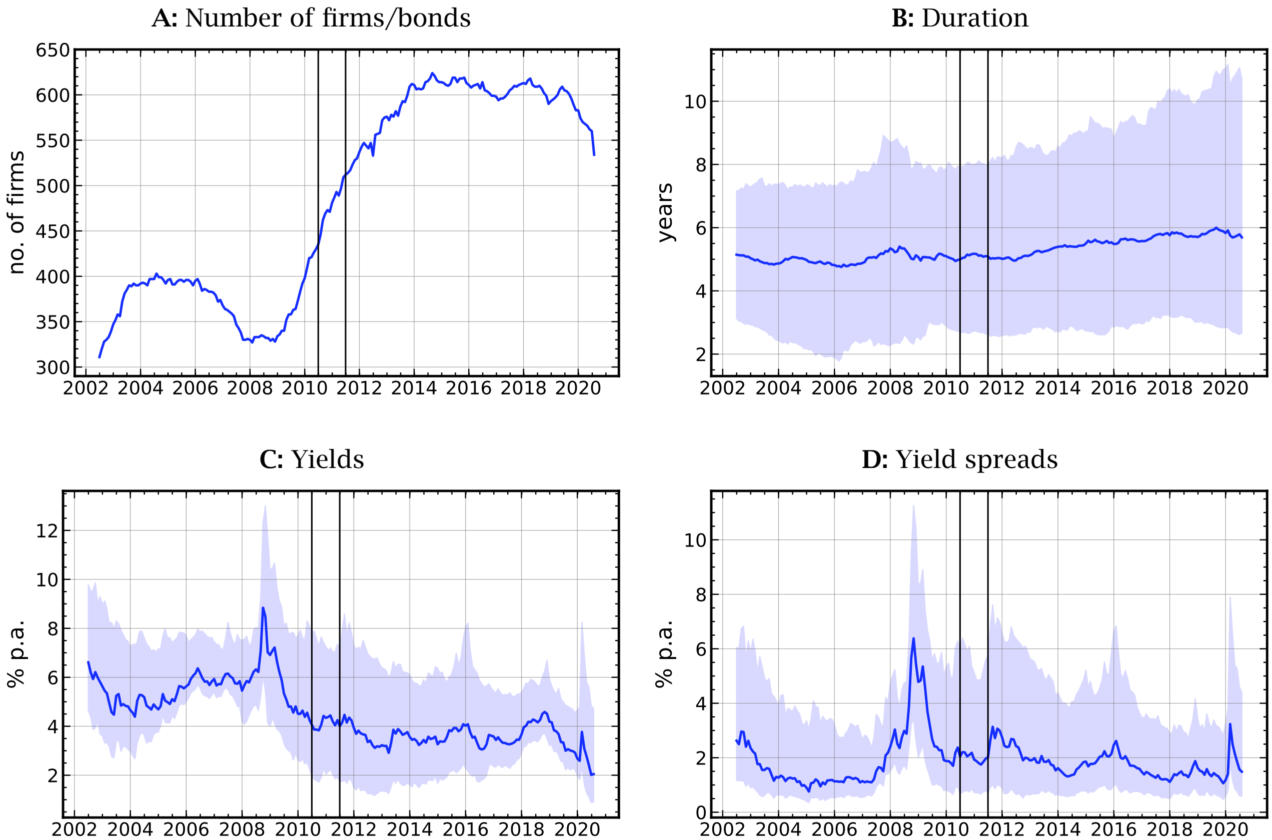
	Abbreviation	Measure	Reference
Firm-level variables			
1.	age	Bond age	Israel et al. (2018)
2.	amtout	Face value	Chung et al. (2019)
3.	assets	Assets	
4.	atbe	Total assets to market value of equity	Asness et al. (2019)
5.	atme	Total assets to book value of equity	Asness et al. (2019)
6.	beme	Book-to-price	Bartram et al. (2024)
7.	chgprof	Profitability change	Asness et al. (2019)
8.	coupon	Coupon	Israel et al. (2018)
9.	d2d	Distance-to-default	Israel et al. (2018)
10.	debtebitda	Debt-to-EBITDA	Kelly et al. (2023)
11.	dolvol	Dollar trading volume	
12.	duration	Duration	Israel et al. (2018)
13.	dureq	Equity Duration	
14.	ebitdasales	EBITDA margin (EBITDA/sales)	
15.	gpat	Profitability	Choi and Kim (2018)
16.	levop	Operating leverage	Gamba and Saretto (2019)
17.	marketlev	Market leverage	Asness et al. (2019)
18.	maturity	Time to maturity (years)	Bali et al. (2022a)
19.	me	Equity market capitalization	Choi and Kim (2018)
20.	mom	Mom. 6m	Gebhardt et al. (2005)
21.	momeq	Mom. 6m equity	Gebhardt et al. (2005)
22.	momindeq	Mom. 6m industry	Jorion and Zhang (2009)
23.	momspr	Mom. 6m log(Spread)	Israel et al. (2018)
24.	nime	Earnings-to-price	Correia et al. (2012)
25.	nrtg	Numerical Rating	Kelly et al. (2023)
26.	offramt	Face value at issue	
27.	ret	Bond excess return in month t (r_t)	
28.	retlag	Bond excess return in month $t-1$ (r_{t-1})	
29.	sales	Sales	
30.	salesat	Asset turnover (sales/assets)	
31.	skew	Bond skewness	Bai et al. (2016)
32.	sprd2d	Spread-to-D2D	Correia et al. (2012)
33.	spread	Spread	Israel et al. (2018)
34.	sprxrtg	Spread x ratings	
35.	totaldebt	Firm total debt	Kelly et al. (2023)
36.	turnvol	Turnover volatility	Correia et al. (2018)
37.	var	Value-at-risk	Dickerson et al. (2023b)
38.	vixbeta	VIX beta	Chung et al. (2019)
39.	volbond	Bond volatility	Bai et al. (2016)
40.	voleq	Equity volatility	Campbell and Taksler (2003)
41.	yield	Bond yield	
Traded corporate bond factors			
42.	CRF	Credit risk factor	Dickerson et al. (2023b)
43.	CRY	Bond carry factor	Houweling and van Zundert (2017)
44.	DEF	Bond default risk factor	Fama and French (1993)
45.	DRF	Downside risk factor	Dickerson et al. (2023b)

Continued on next page

	Abbreviation	Measure	Reference
46.	DUR	Bond duration factor	Dang et al. (2023)
47.	HMLB	Bond book-to-market factor	Bartram et al. (2024)
48.	LTREVB	Bond long-term reversal factor	Bali et al. (2021a)
49.	MKTB	Corporate Bond Market excess return	Dickerson et al. (2023b)
50.	MKTBD	Corporate Bond Market duration adjusted return	Binsbergen et al. (2024)
51.	MOMB	Bond momentum factor	Bali et al. (2022b)
52.	PEADB	Bond earnings announcement drift factor	Nozawa et al. (2023)
53.	STREVB	Bond short-term reversal factor	Bali et al. (2021a)
54.	TERM	Bond term structure risk factor	Fama and French (1993)
55.	VAL	Bond value factor	Houweling and van Zundert (2017)
Traded equity factors			
56.	CMA	Investment factor	Fama and French (2015)
57.	HML	Value factor	Fama and French (1992)
58.	LIQ	Liquidity factor	Pástor and Stambaugh (2003)
59.	MKT	Market excess return	Sharpe (1964), Lintner (1965)
60.	MOM	Momentum factor	Carhart (1997), Jegadeesh and Titman (1993)
61.	RMW	Profitability factor	Fama and French (2015)
62.	SMB	Size factor	Fama and French (1992)
Non-traded corporate bond and equity variables			
63.	AGGLIQ	Level of aggregated stock market liquidity	Pástor and Stambaugh (2003)
64.	CPTL	Intermediary capital non-traded risk factor	He et al. (2017)
65.	INTCAP	Aggregate Capital Ratio of Financial Intermediaries	He et al. (2017)
66.	LIQNT	Liquidity factor	Pástor and Stambaugh (2003)
67.	VIX	CBOE VIX	Chung et al. (2019)
68.	ΔVIX	First difference of VIX	
Macroeconomic variables			
69.	ΔCPI	First difference in the Consumer Price Index	
70.	EPU	Economic Policy Uncertainty Index	Baker et al. (2016)
71.	EPUT	Economic Tax Policy Uncertainty Index	Baker et al. (2016)
72.	ΔEPU	First difference of EPU	
73.	ΔEPUT	First difference of EPUT	
74.	RF	Risk-free rate	
75.	ΔRF	First difference of RF	
76.	UNC	Macroeconomic Uncertainty Index ($h = 1$)	Jurado et al. (2015)
77.	UNCf	Financial Economic Uncertainty Index ($h = 1$)	Jurado et al. (2015)
78.	UNCr	Real Economic Uncertainty Index ($h = 1$)	Jurado et al. (2015)
79.	ΔUNC	First difference of UNC	
80.	ΔUNCf	First difference of UNCf	
81.	ΔUNCr	First difference of UNCr	

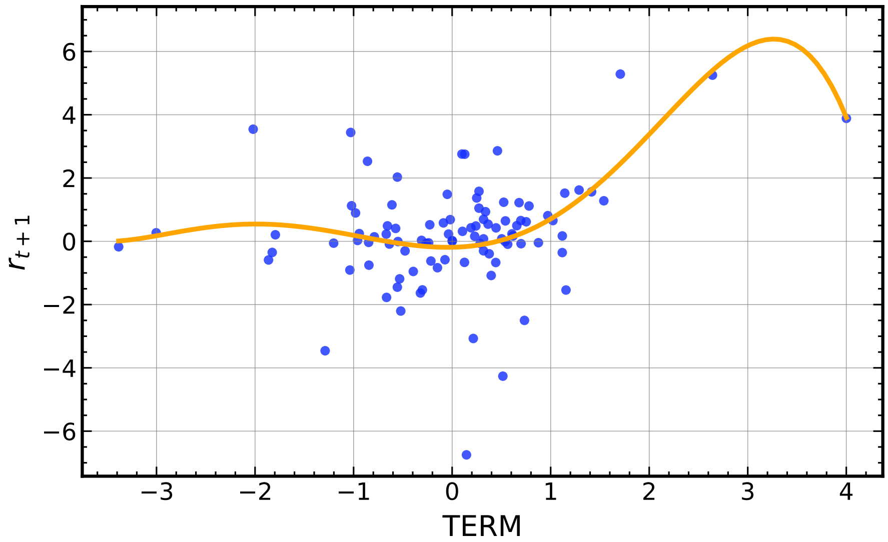
Note: This table presents variable definitions for the set of predictors in our dataset. We differentiate between bond or firm characteristics, traded factors, non-traded factors, and additional macroeconomic predictors. We provide descriptions of each measure with references to previous literature.

Figure A.1: Sample characteristics over time



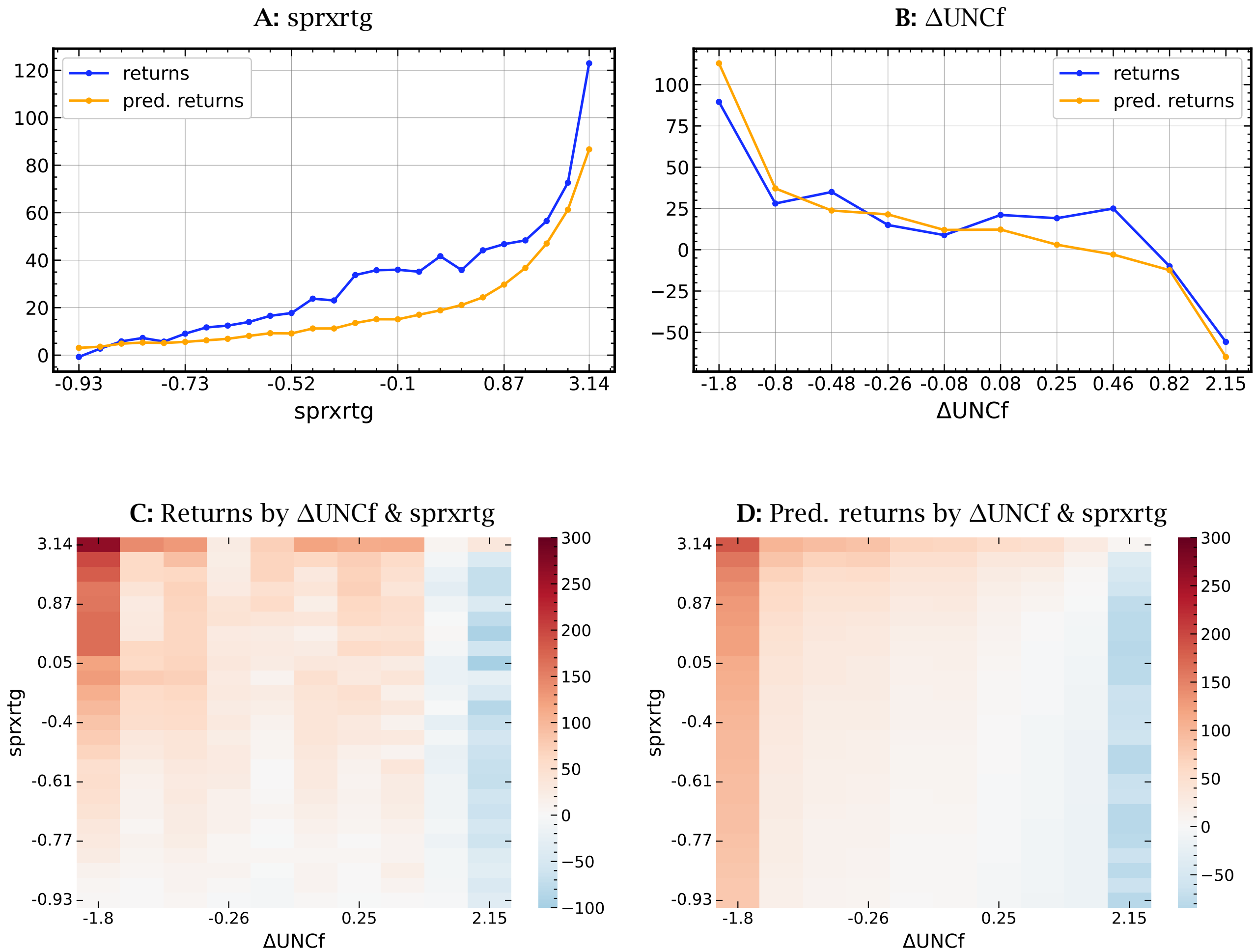
This figure shows the number of firms in the sample and the median duration, yields, and yield spreads for each month of our sample. The shaded areas indicate the 10% to 90% percentile intervals. The vertical lines represent the ends of the training and validation samples.

Figure A.2: TERM and returns in the training sample



This figure shows a scatter plot of TERM in period t and mean bond returns in $t + 1$ in the training sample. A non-parametric regression line is drawn in orange.

Figure A.3: Interactions – ΔUNCf & sprxrtg and returns



This figure shows returns and predicted returns by ΔUNCf and sprxrtg. We group observations by ΔUNCf (10 bins) and sprxrtg (25 bins) and compute the means of observed and predicted returns for each bin and the interaction bins. Panels A and B show the means of observed returns (blue) and predicted returns (orange) by sprxrtg and ΔUNCf bins. Panel C plots the heatmap of observed returns by ΔUNCf (x-axis) and sprxrtg (y-axis) bins. High (low) returns are in red (blue). Panel D shows the same plot for predicted returns.

B Hyperparameter tuning

We provide more detailed descriptions of the EBM algorithm and its hyperparameters.[38] The EBM uses tree-based gradient boosting to optimize a given objective. This objective is (root) mean squared error in a regression setting. The core estimation algorithm is based on a cyclical procedure, which cycles through all input features, fitting shallow decision trees based only on one feature and updating residuals after each tree. One cycle through all features constitutes an iteration. Iterations are added until either the maximum number of iterations (*max_rounds*) is reached or the early stopping criterion is met (defined via *early_stopping_tolerance* and *early_stopping_rounds*). There have been several modifications and additions to this algorithm over the years. Unfortunately, there is no up-to-date paper or other comprehensive explanation for its current state, so the following is an overview based on public information provided by the package's authors in the code, in the documentation, or via GitHub.

- All continuous variables are first discretized into roughly equal-sized bins. The number of bins is equal to a predefined number (*max_bins*), except when insufficient data populate that number of bins. The bins mitigate the effect of outliers and speed up the tree estimation because the set of possible candidate splits is reduced to the set of boundaries between bins.
- Before the regular boosting rounds, a number (*smoothing_rounds*) of highly regularized precursor boosting rounds are executed where tree splits are randomly chosen. The idea is to establish the basic shape function for each feature before moving to more aggressive gain-based boosting.
- The process of how individual trees are fit is regularized by hyperparameters which define the minimum number of samples allowed in a leaf (*min_samples_leaf*) and the maximum number of leaves allowed in each tree (*max_leaves*). Trees are also stopped from growing based on the hessian within leaves resulting from a potential split (*min_hessian*).
- The learning rate (*learning_rate*) describes the weight of an additional tree added to the model. Small learning rates are used so that residuals are only updated by small amounts in each boosting step, and the order of features does not matter.
- In addition to the cyclic boosting rounds described above, the EBM also uses greedy rounds where features are visited in order of calculated gain.[39] The proportion of the number of boosting steps in greedy rounds vs. cycling rounds is a hyperparameter (*greedy_ratio*).[40] Another parameter (*cyclic_progress*) describes the proportion of cyclic rounds where updates, i.e., calculated trees, are applied to the model. Cyclic rounds where updates are not applied will still recalculate gains.
- After all univariate functions are fit, pairs of features are ranked using the heuristic FAST algorithm. FAST estimates the possible gain from an interaction without fully constructing the associated shape function. A certain number (*interactions*) of the highest-ranked pairs is then chosen, and the associated shape functions are estimated using an analogous procedure to the univariate case. The main difference is that each tree can now incorporate both features of a given pair instead of only a single feature.
- The model also incorporates multiple layers of bagging. Outer bagging creates a certain number (*outer_bags*) of randomly drawn subsets of the dataset that are of size (1-*validation_size*). An individual model is then

[38] We make use of the InterpretML implementation of the Explainable Boosting Machine by Nori et al. (2019).

[39] Gain has to be calculated for all features during cyclic rounds in order to grow trees. Due to the low learning rate, gain calculations remain relatively valid for multiple update steps and can be used for greedy boosting afterward.

[40] Example: For a data set with 100 features and a default greedy ratio of 1.5, a cycling round will consist of 100 boosting steps followed by a greedy round with 150 boosting steps.

fit on each of the outer bags. The uncertainty intervals shown on the graphs are standard deviations of estimates per bin across each bag. The final model is the mean across all bags.

- Additionally, inner bagging can be applied to each step of the estimation procedure. Cycling through all features generates a tree for each inner bag (*inner_bags*). The averaged update is applied to the feature before moving to the next feature.

Table B.1 reports the values of the hyperparameters that are used in the training and validation samples for the LASSO, Random Forest, XGBoost, and EBM models.

Table B.1: Hyperparameters

Model	LASSO	Random Forest	XGBoost	EBM
Parameters	$\alpha \in (10^{-3}, 10^{-1})$	$\#Trees = 100$ *Min Samples Leaf* $\in \{1, 2, 4\}$ *Min Samples Split* $\in \{2, 4, 8\}$ *Max Features* $\in \{0.1, 0.25, 0.5, 0.75, 1\}$ *Max Depth* $\in \{2, 4, 8\}$	$\#Trees \in \{10, 50, 100\}$ *Max Depth* $\in \{5, 6, 7, 8, 9\}$ *Learning Rate* $\in \{0.01, 0.1, 0.3, 0.5\}$ $Subsample \in \{0.5, 0.75, 1\}$ $Lambda \in \{1, 2, 4, 8\}$	*Early Stopping Rounds* $\in \{10, 20\}$ *Early Stopping Tolerance* $\in \{0.01, 0.005\}$ *Smoothing Rounds* $= 25$ *Interaction Smoothing Rounds* $\in \{250, 500\}$ *Greedy Ratio* $\in \{1.5, 5\}$ *Learning Rate* $= 0.00025$ *Max Leaves* $= 2$ *Max Rounds* $= 500$ *Min Hessian* $= 0.00001$ *Min Samples Leaves* $\in \{2, 20\}$ $Interactions \in \{30, 40\}$ *Inner Bags* $\in \{0, 5\}$ *Outer Bags* $= 10$ *Validation Size* $\in \{0.5, 0.75\}$

This table shows the hyperparameters used in the training and validation samples for the LASSO, Random Forest, XGBoost, and EBM models.

C Estimation with equal weights

Our main model weights data points based on market equity. As Table C.1 shows, model performance decreases slightly when using equal weights instead. The ordering of the models is similar, with the linear regression still performing poorly out-of-sample, followed by LASSO with a positive R^2 and the other models which reach R^2_{oos} values greater than 11%. Figure C.1 displays predictor rankings according to the different importance measures for the equal-weight EBM model. We confirm that the previously documented importance of uncertainty factors, the term structure factor, and bond spread is robust to the choice of weighting scheme.

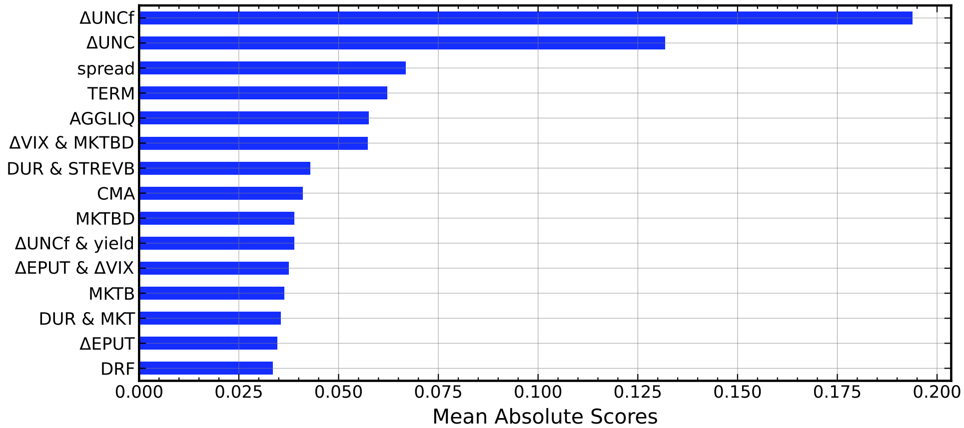
Table C.1: Comparison of R^2's – EW

	OLS	LASSO	R-Forest	XGBoost	EBM
R^2_{train}	35.28%	19.38%	34.75%	42.57%	39.57%
R^2_{valid}	-27.54%	12.09%	12.54%	16.87%	18.01%
R^2_{oos}	-22.12%	5.77%	10.16%	11.41%	11.38%

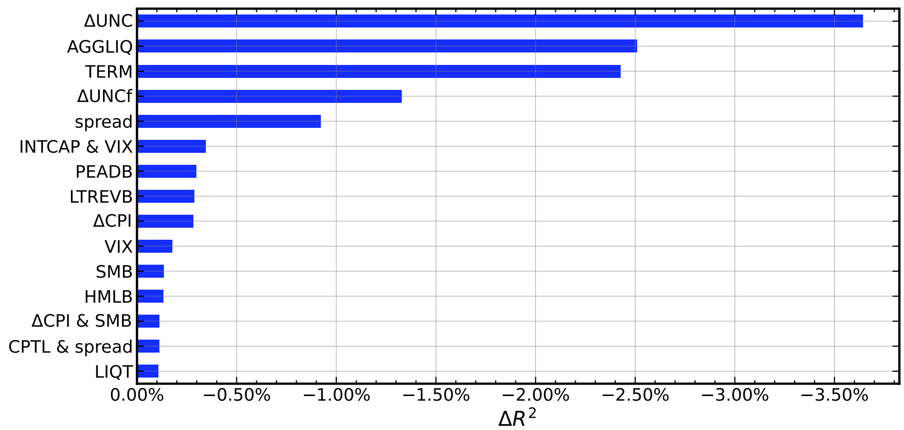
This table shows R^2's of OLS, LASSO, Random Forest, XGBoost, and EBM estimations in the training (July 2002 – June 2010), validation (July 2010 – June 2012), and testing (July 2012 – August 2020) samples when all observations are weighted equally.

Figure C.1: Importance of predictors – EW

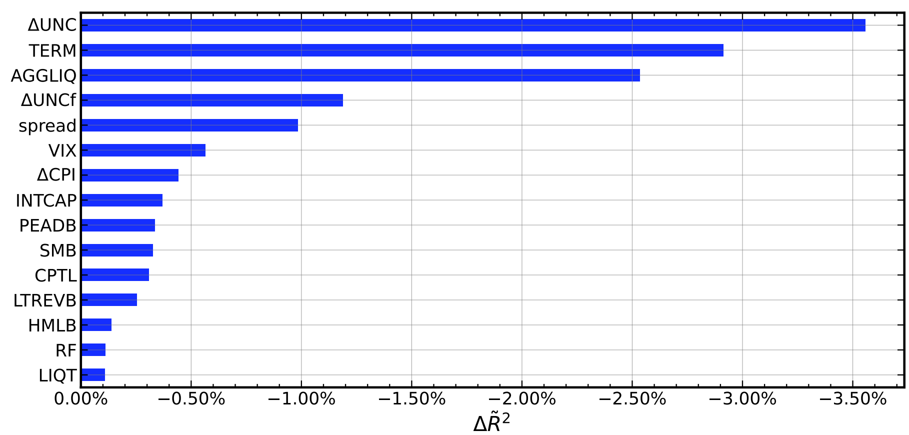
A: Mean Absolute Scores: $S(i)$ and $S(i,j)$



B: ΔR_i^2: Univariate effect



C: $\Delta \widetilde{R}_i^2$: Total effect

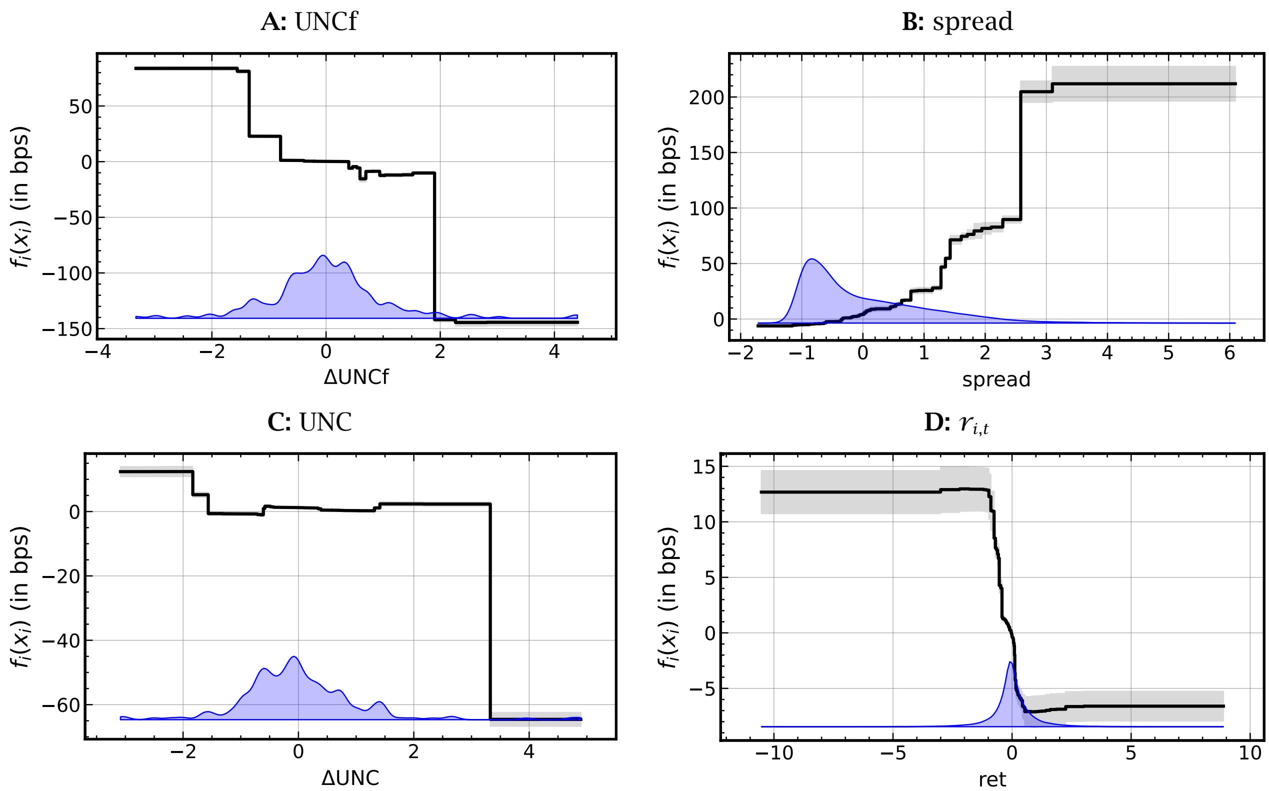


This figure shows the most important predictors for the EBM model when all observations are weighted equally. In Panel A, importance values are mean absolute scores calculated on the training data (the native EBM importance measure). In Panel B, importance values are calculated as reductions in R^2 when removing a term, i.e., an individual shape function, from the fitted model. In Panel C, importance values are calculated as reductions in R^2 when removing all terms associated with a variable from the fitted model.

D Estimation with 14-year training sample

The benchmark estimation uses a training sample of eight years. For robustness, we also analyze another model resulting from a longer training period of 14 years (7/2002–6/2016). With a validation window of one year (7/2016–6/2017), we are left with a little more than three years of data for testing (7/2017–8/2020). We again weight samples by market equity and adjust the set of parameters to account for the longer training period. Figure D.1 shows the shape functions of some of the most important out-of-sample predictors. Further results are reported in Appendix O-C.1.

Figure D.1: Univariate functions $f_i(x_i)$ – 14-year training sample



This figure shows the contributions to return predictions of the three most important out-of-sample predictors in the estimation with a 14-year training sample. We also present the functions associated with lagged returns because we refer to them when constructing EBM portfolios. The x-axis describes the input variable standardized to a mean of 0 and a standard deviation of 1, while the y-axis represents the additive contribution to the prediction output.

E Estimation with MMN-adjusted data

We show the robustness of our results to market microstructure noise (MMN), which describes serially correlated measurement errors in TRACE-based bond prices that could result in look-ahead bias (see, e.g., Bartram et al. (2024)). To remove the influence of MMN, we follow the recommendations by Dickerson et al. (2023c) and the documentation available on Open Source Bond Asset Pricing. We drop the sprxrtg and sprd2d measures and replace other price-related variables in our dataset with their MMN-adjusted versions obtained from Open Source Bond Asset Pricing. Namely, this concerns bond duration, returns, spread, spread momentum, and yield. In line with Dickerson et al. (2023c), we do not use the MMN-adjusted returns as month-ahead ex ante bond returns, but only as return signals.

Using the adjusted data, EBM reaches R^2 values of 23.29%, 16.15%, and 11.04% on the training, validation, and testing sample, respectively. Note that the MMN adjustments result in reduced data availability for price-related characteristics, which might be partially responsible for the reduction in predictive value.

We proceed by reproducing our main results on the adjusted dataset: Figure E.1 reports the observed predictor importances. While the univariate effect associated with spread is now slightly less important, the interaction of ΔUNCf & spread remains among the most important interaction terms, according to both mean absolute score and ΔR^2. In total, spread remains among the top 5 most important features, ranked no. 5 in terms of total R^2 importance (compared to no. 4 on the unadjusted data). Additionally, the adjusted spread and yield measures remain the two most important characteristics in our dataset.

Furthermore, Figure E.2 confirms that the main shape functions are not qualitatively affected by the adjustments. The average contribution of spread decreases in terms of magnitude, but we still observe a monotonically increasing function. Compared to our main model, standard errors are smaller, allowing us to deduce the nature of the spread effect with high confidence.[41] The interaction of ΔUNCf & spread is illustrated in Figure E.2. We find large agreement with the effect from our main model, both qualitatively and quantitatively (see Figure 5 in the paper). The interaction heatmap and the resulting conditional effects seem highly robust to the MMN adjustments (Panel A-C). The main difference compared to the results on the original data is the relative size of the univariate spread effect compared to the effect through the interaction term. As a result, the aggregate effect of spread is now somewhat smaller and more differentiated between ΔUNCf values, i.e. more dependent on ΔUNCf realizations (Panel D). The effect is still monotonically increasing for most ΔUNCf values.

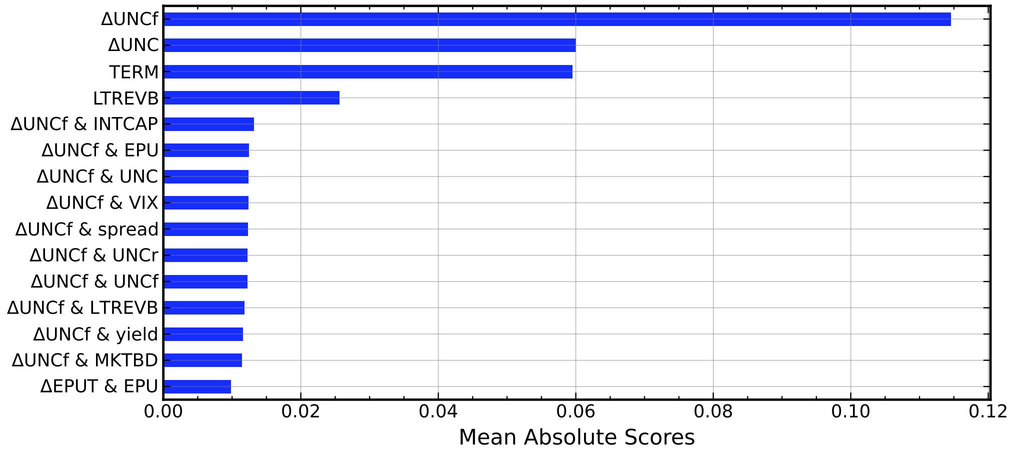
Our portfolio results are also qualitatively unchanged: In Table E.1, we report the updated benchmark portfolios derived from characteristics sorts. Spread, yield, and momspr remain among the top 10 benchmark characteristic portfolios, albeit with slightly reduced returns and Sharpe ratios. As Table E.2 shows, decile portfolio sorted on EBM predictions continue to exhibit monotonic patterns in returns and Sharpe ratios. The EBM-based portfolio strategies still outperform all benchmarks and particularly the "pos" strategies, which exploit time-series forecastability, remain highly profitable. The value-weighted EBM portfolio's alpha in excess of the duration-adjusted bond market return is still significant and measures 0.51% (t=3.53).

Overall, we conclude that our main results are robust to MMN adjustments. The effects of spread and other characteristics are economically meaningful and not merely artifacts of market microstructure noise.

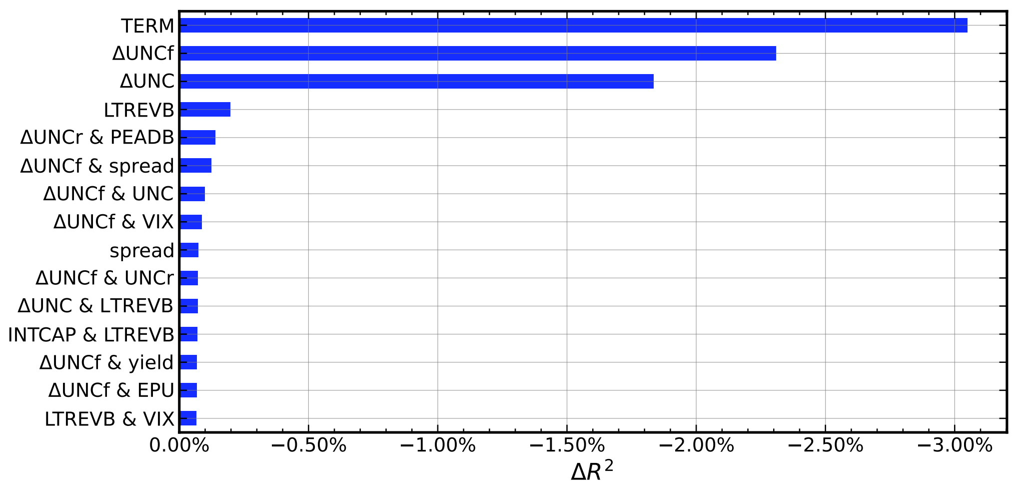
[41] The MMN adjustments also produce some outliers in the spread data. We do not remove these as EBM is robust to outliers, but truncate the spread axes of the shape function plots at the 0.1% level for legibility.

Figure E.1: Importance of predictors

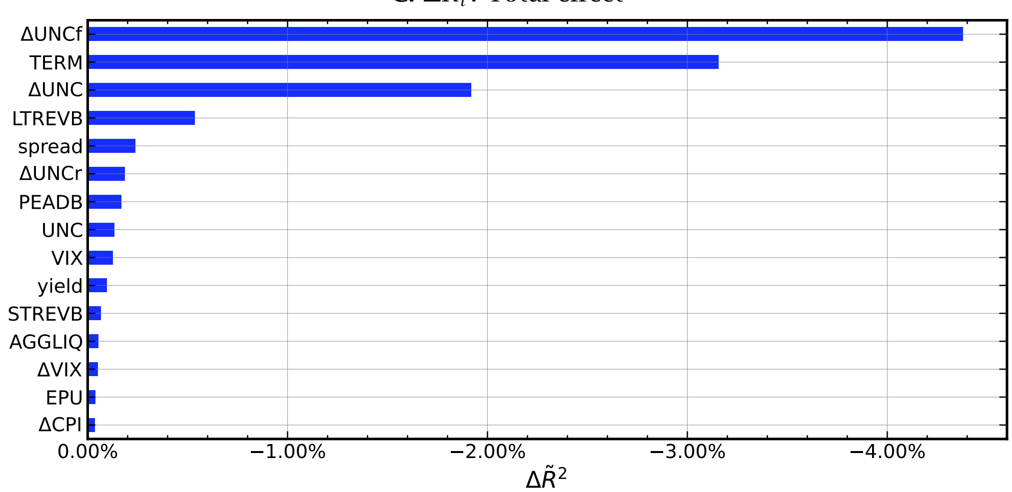
A: Mean Absolute Scores: $S(i)$ and $S(i,j)$



B: ΔR_i^2: Univariate effect

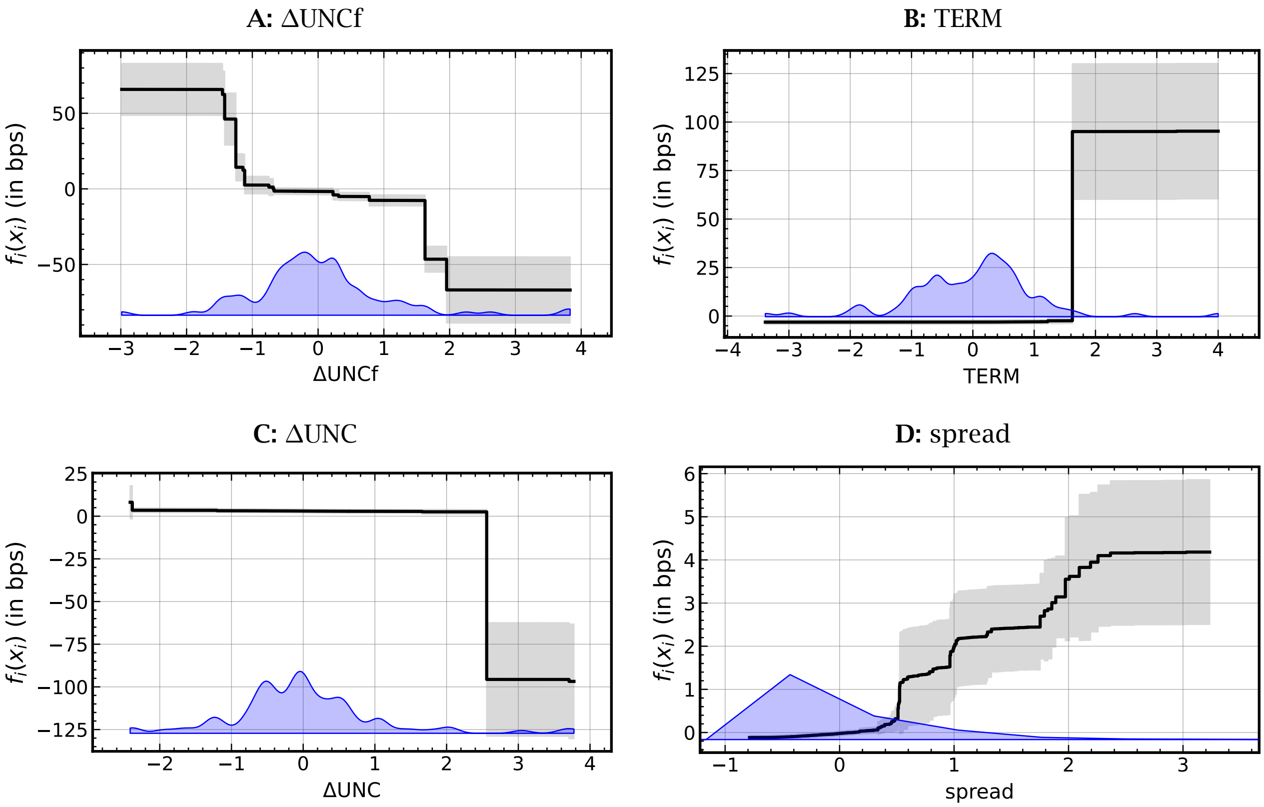


C: $\Delta\tilde{R}_i^2$: Total effect



This figure shows the most important predictors for the EBM model. In Panel A, importance values are mean absolute scores calculated on the training data (the native EBM importance measure). In Panel B, importance values are calculated as reductions in R^2 when removing a term, i.e., an individual shape function, from the fitted model. In Panel C, importance values are calculated as reductions in R^2 when removing all terms associated with a variable from the fitted model.

Figure E.2: Univariate functions $f_i(x_i)$



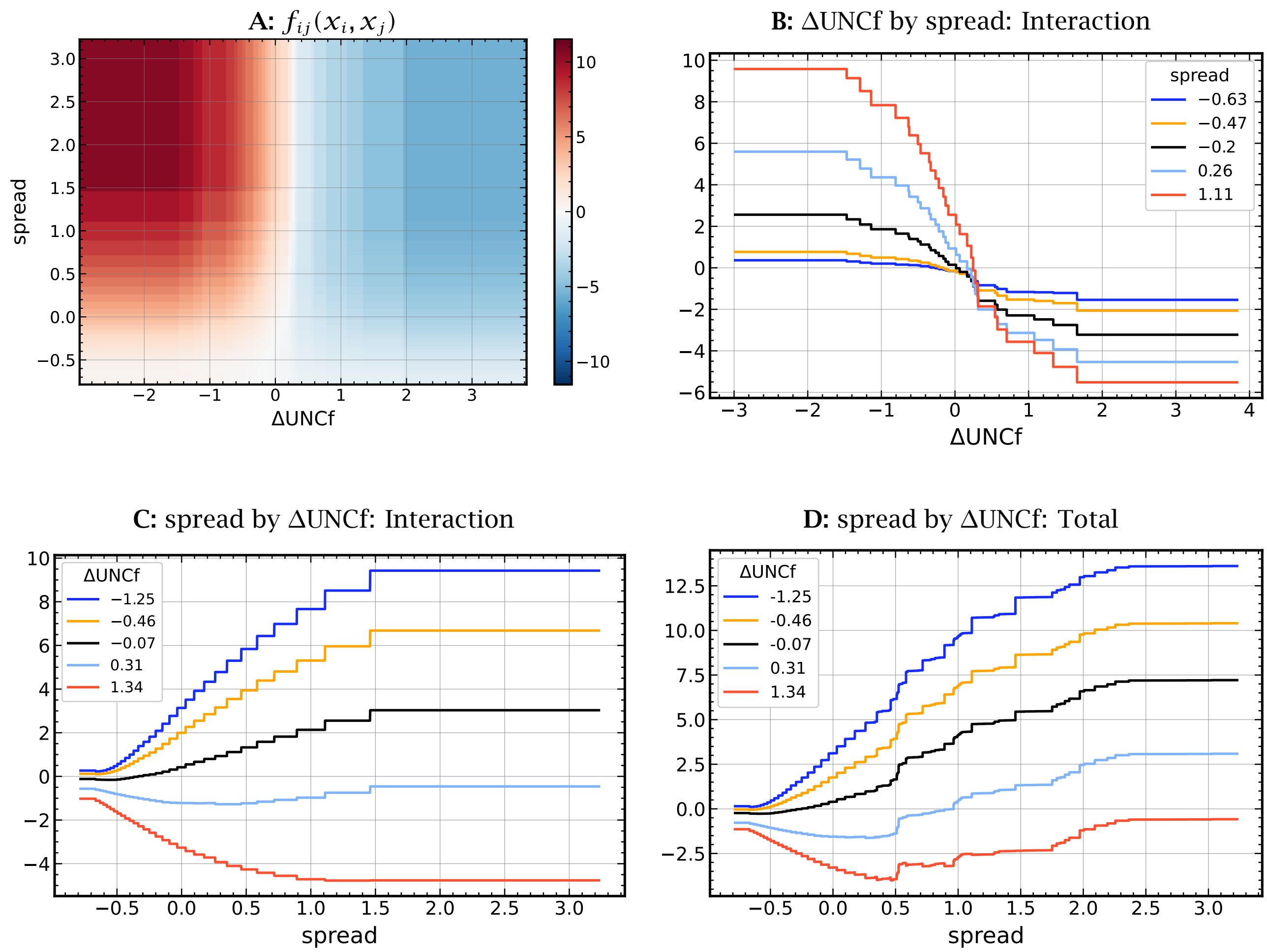
This figure shows the estimated $f_i(x_i)$ functions of ΔUNCf, TERM, ΔUNC, spread. The x-axis plots the input variable x_i standardized to a mean of 0 and a standard deviation of 1. At the same time, the y-axis represents the additive contribution to the prediction output through the shape function $f_i(x_i)$ (in basis points). The underlying data distribution is displayed in blue, and uncertainty intervals are shown in gray. For better readability, the spread axis is truncated at the 0.1% level.

Table E.1: Returns of characteristics-sorted portfolios and factors

	Panel A: 5-1 characteristic portfolios									
	spread	nime	sales	nrtg	yield	levmkt	coupon	momspr	voleq	skew
Mean	0.41	0.10	0.20	0.28	0.39	0.16	0.24	0.20	0.21	0.13
S.D.	1.61	0.39	0.82	1.14	1.72	0.72	1.20	1.02	1.16	0.97
SR	0.26	0.25	0.25	0.24	0.23	0.22	0.20	0.19	0.18	0.13

This table reports monthly means, standard deviations, and Sharpe ratios of characteristics-sorted portfolios. We identify the top ten characteristics-sorted portfolios by their Sharpe ratios in the combined training and validation samples and report their statistics in the testing sample.

Figure E.3: Interactions – ΔUNCf & spread



Panel A displays the shape function of the interaction between the financial uncertainty feature and bond spread. The scaled values of ΔUNCf and spread are shown on the x- and the y-axis, respectively. The color-coded axis represents the additive contribution to the prediction output associated with combinations of x and y values in basis points. In Panels B and C, conditional contributions are derived from the shape function by separately holding each variable involved in the interaction constant at different levels. Panel D shows the sum of the univariate and interaction effects of spread conditional on ΔUNCf. For better readability, the spread axes are truncated at the 0.1% level.

Table E.2: Returns of EBM portfolios

	1	2	3	4	5	6	7	8	9	10	10-1
					Panel A: EBM VW						
Mean	0.02	0.08	0.08	0.11	0.16	0.18	0.21	0.22	0.37	0.67	0.65
S.D.	1.33	1.47	1.27	1.34	1.41	1.52	1.59	1.70	1.95	2.82	1.95
SR	0.02	0.05	0.06	0.08	0.11	0.12	0.13	0.13	0.19	0.24	0.33
					Panel B: EBM-pos VW						
Mean	0.10	0.14	0.15	0.19	0.24	0.30	0.34	0.35	0.52	0.77	0.66
S.D.	0.54	0.63	0.63	0.69	0.76	0.91	0.98	1.10	1.32	1.78	1.47
SR	0.19	0.22	0.24	0.27	0.31	0.33	0.35	0.32	0.39	0.43	0.45
					Panel C: EBM EW						
Mean	0.01	0.08	0.08	0.11	0.16	0.20	0.24	0.25	0.37	0.52	0.50
S.D.	1.37	1.44	1.26	1.38	1.41	1.53	1.64	1.67	1.88	2.40	1.57
SR	0.01	0.06	0.06	0.08	0.11	0.13	0.14	0.15	0.20	0.21	0.32
					Panel D: EBM-pos EW						
Mean	0.11	0.15	0.16	0.20	0.24	0.31	0.37	0.37	0.51	0.73	0.62
S.D.	0.49	0.56	0.57	0.66	0.75	0.91	0.99	1.08	1.23	1.72	1.41
SR	0.21	0.27	0.27	0.30	0.32	0.35	0.37	0.35	0.41	0.42	0.44

This table compares monthly means, standard deviations, and Sharpe ratios for ten value-weighted portfolios constructed based on predicted return deciles in the testing sample. The last columns show the results of the 10-1 portfolio, which is long in portfolio 10 and short in portfolio 1. EW are equally-weighted portfolios and VW are value-weighted portfolios by bond market value. EBM-pos is the strategy that invests in the riskfree asset when the predicted portfolio return is negative.

Online Appendix:

Glass Box Machine Learning and Corporate Bond Returns

by

S. Bell, A. Kakhbod, M. Lettau, and A. Nazemi

O-A Additional material for Section 2

O-A.1 Algorithms

We provide stylized algorithmic descriptions of the estimation procedure used to produce a singular EBM model. Algorithm 1 illustrates how shape functions f_i are generated for univariate inputs x_i. Algorithm 2 returns multivariate shape functions $f_{i,j}$ for interactions of inputs x_i and x_j. These functions are used in the different stages of Algorithm 3, which returns the final model.

Algorithm 1 Univariate Shape Functions

function GETSHAPEFUNCTIONS
    Input Dataset of length L with predictors x_i and returns r, parameters p, learning rate α
    Output Shape functions f_i
    $f_i \leftarrow 0$
    for $k = 1$ to k_{max} **do**
        if not `early stopping criterion` **then**
            for all x_i **do**
                $\mathcal{T} \leftarrow \left\{x_i^{(l)}, r^{(l)} - \sum_i f_i\left(x_i^{(l)}\right)\right\}_{l=1}^{L}$ ▷ Update Residuals
                Learn tree $t : x_i \to y$ using parameters p on $\mathcal{T}$
                $f_i \leftarrow f_i + \alpha t$
            end for
        end if
    end for
    return f_i
end function

Algorithm 2 Multivariate Shape Functions

function GETSHAPEFUNCTIONS2D
    Input Dataset of length L with predictors x_i and returns r, parameters p, learning rate α
        univariate model F, set of interaction pairs S
    Output Shape functions $f_{i,j}$
    $f_{i,j} \leftarrow 0$
    for $k = 1$ to k_{max} **do**
        if not `early stopping criterion` **then**
            for all $s \in S$ **do**
                $(i, j) \leftarrow s$
                $\mathcal{T} \leftarrow \left\{x_i^{(l)}, x_j^{(l)}, r^{(l)} - F(x) - \sum_{i,j} f_{i,j}\left(x_i^{(l)}, x_j^{(l)}\right)\right\}_{l=1}^{L}$
                Learn tree $t : (x_i, x_j) \to y$ with parameters p on $\mathcal{T}$
                $f_{i,j} \leftarrow f_{i,j} + \alpha t$
            end for
        end if
    end for
    return $f_{i,j}$
end function

O-A.2 Interaction detection

A significant improvement, that allows Explainable Boosting Machines to outperform traditional generalized additive models, is the inclusion of some first-order interaction terms (Lou et al. (2013)). This means that for certain index values (i, j), combinations of features x_i and x_j are considered

Algorithm 3 Model Algorithm

function MODEL
  Input Dataset D of length L with predictors x_i and returns r, parameters p, learning rate α,
  Output Model instance F
  $F \leftarrow 0$
  $A \leftarrow \text{GETSHAPEFUNCTIONS}(D, p, \alpha)$
  for all f_i in A **do**
    $F \leftarrow F + f_i$
  end for
  $C \leftarrow \emptyset$
  for all x_i, x_j **do** ▷ Apply FAST algorithm
    $C \leftarrow C \cup \text{EVALUATEINTERACTION}(x_i, x_j)$
  end for
  $S \leftarrow \text{TOPRANKEDPAIRS}(C)$
  $A \leftarrow \text{GETSHAPEFUNCTIONS2D}(D, p, \alpha, F, S)$
  for all $f_{i,j}$ in A **do**
    $F \leftarrow F + f_{i,j}$
  end for
  return F
end function

and mapped to suitable output values, defining a function f_{ij}. As an exhaustive search through all possible combinations is inefficient or infeasible for high-dimensional datasets, an EBM model learns a predefined number of interaction functions automatically detected based on their predictive value via the time-efficient FAST algorithm. Conceptually, the FAST algorithm works by adding interaction terms to the model one by one using a greedy procedure. In each iteration, the pair of features that maximizes fit on the residuals of the best previous model is chosen. Importantly, the FAST algorithm ranks all possible pairs of features without computing the full-complexity shape functions associated with their interactions. Instead, a simple model for f_{ij} is built using a single cut in each variable x_i and x_j. This results in four quadrants in the (x_i, x_j) plane for which the mean target values are calculated, defining a simple predictor $P_{i,j}$ that is evaluated on the training sample (of size L) using the residual sum of squares:

$$RSS = \sum_{l=1}^{L} \left(y^{(l)} - P_{i,j}(x_i^{(l)}, x_j^{(l)}) \right)^2 . \tag{O-A.1}$$

The computational cost of the required calculations can be optimized so that all possible cuts can be efficiently compared. The cuts that minimize the residual sum of squares are chosen, and the corresponding RSS value is used to define the strength of an interaction. Further efficiency gains are achieved by not requiring refitting the full model after adding a new interaction. Instead, univariate functions are fitted first and left unchanged for the interaction detection stage. This means the training algorithm finds the best GAM before fitting interactions on residuals. Shape functions for interaction terms are estimated using the boosting procedure described in Section 2, with trees now using both features of an interaction term.

O-B Summary statistics of all variables

Table O-B.1: Summary statistics

Variable	Count	Mean	S.D.	25%	50%	75%
age	106265	3.017	2.017	1.572	2.712	4.063
amtout	106265	4.077	8.643	0.550	1.387	3.719
assets	89903	71.867	233.527	6.155	15.494	40.578
atbe	86665	6.997	81.650	2.082	2.812	4.724
atme	89903	3.096	9.496	0.805	1.463	2.685
beme	86665	0.670	1.980	0.288	0.501	0.816
chgprof	83741	-0.015	0.114	-0.052	-0.006	0.030
coupon	106265	5.614	1.829	4.250	5.534	6.875
d2d	100054	7.409	5.228	3.731	6.520	10.130
debtebitda	89335	3.704	4.891	1.675	2.804	4.590
dolvol	106265	76.353	164.865	20.045	40.566	80.169
duration	105958	5.606	2.563	3.938	5.293	6.856
dureq	85329	16.078	8.854	14.931	16.433	17.484
ebitdasale	89825	0.150	9.323	0.122	0.205	0.328
levmkt	80604	1.589	4.743	1.009	1.200	1.593
levop	89710	5.933	10.674	1.958	3.737	6.874
maturity	106265	7.865	5.031	4.648	6.673	9.502
me	106265	26.442	55.779	3.379	10.315	25.973
mom	96658	0.013	0.051	-0.003	0.010	0.029
momeq	89752	0.049	0.222	-0.067	0.046	0.156
momeqind	89752	-1.053	142.660	-0.115	-0.012	0.087
momspr	97023	0.037	0.368	-0.139	0.052	0.237
nime	89899	0.036	0.223	0.030	0.052	0.076
nrtg	106265	9.305	3.196	7.000	9.000	11.251
offramt	106265	0.738	0.500	0.426	0.600	0.876
profitbl	89727	0.235	0.183	0.102	0.200	0.322
ret	104841	0.002	0.023	-0.005	0.002	0.010
retlag	103626	0.002	0.023	-0.005	0.002	0.010
sales	89899	20.321	36.085	3.677	8.935	20.259
salesat	89899	0.763	0.697	0.311	0.587	0.988
skew	88427	0.037	0.710	-0.278	0.072	0.404
sprd2d	99763	39.366	29.999	17.543	29.529	54.407
spread	105958	2.371	1.824	1.020	1.741	3.270
sprxrtg	105958	26.135	26.454	7.407	15.241	37.643
totaldebt	89903	18.764	66.234	1.916	4.280	10.404
turnvol	88518	0.110	0.133	0.037	0.073	0.135
var	88427	-0.031	0.024	-0.039	-0.023	-0.015
vixbeta	88427	-0.017	0.050	-0.035	-0.011	0.005
volbond	88427	0.020	0.013	0.011	0.016	0.025
voleq	89894	0.019	0.014	0.011	0.016	0.023
yield	105958	4.593	2.007	3.085	4.407	5.832
AGGLIQ	106265	-0.017	0.062	-0.043	-0.012	0.019
CMA	106265	-0.031	1.483	-1.000	-0.050	0.910

Continued on next page

Variable	Count	Mean	S.D.	25%	50%	75%
ΔCPI	106265	0.158	0.096	0.105	0.164	0.207
CPTL	106265	0.000	0.064	-0.031	0.006	0.041
CRF	106265	0.129	2.533	-0.911	0.184	1.297
CRY	106265	0.568	2.780	-0.461	0.478	1.537
DEF	106265	0.043	1.996	-0.790	0.050	0.890
DRF	106265	0.521	2.576	-0.708	0.531	1.577
DUR	106265	0.345	2.297	-0.955	0.433	1.711
EPU	106265	4.595	0.468	4.254	4.513	4.864
ΔEPU	106265	0.970	38.948	-15.330	-0.620	15.090
EPUT	106265	4.605	0.580	4.223	4.558	4.975
ΔEPUT	106265	0.850	45.886	-18.659	0.077	17.397
HML	106265	-0.279	2.649	-1.810	-0.310	1.180
HMLB	106265	0.349	2.208	-0.381	0.135	1.060
INTCAP	106265	0.062	0.015	0.050	0.061	0.075
LIQNT	106265	0.008	0.054	-0.022	0.009	0.035
LIQT	106265	0.277	3.600	-1.580	0.293	2.498
LTREVB	106265	0.073	1.163	-0.337	0.017	0.466
MKT	106265	0.894	4.205	-1.220	1.350	3.330
MKTB	106265	0.459	1.709	-0.373	0.499	1.119
MKTBD	106265	0.194	1.630	-0.279	0.290	0.783
MOM	106265	0.155	4.223	-1.590	0.300	2.280
MOMB	106265	-0.125	2.277	-1.050	0.048	1.091
PEADB	106265	0.136	0.519	-0.022	0.102	0.254
RF	106265	0.090	0.117	0.000	0.020	0.150
ΔRF	106265	-0.000	0.020	0.000	0.000	0.010
RMW	106265	0.174	1.800	-0.800	0.250	1.200
SMB	106265	0.016	2.467	-1.710	0.190	1.680
STREVB	106265	0.194	1.929	-0.833	0.171	0.863
TERM	106265	0.528	3.154	-1.500	0.400	2.480
UNC	106265	-0.445	0.164	-0.539	-0.481	-0.427
ΔUNC	106265	0.002	0.025	-0.008	-0.001	0.007
UNCf	106265	-0.146	0.192	-0.282	-0.219	0.016
ΔUNCf	106265	0.000	0.036	-0.018	-0.001	0.015
UNCr	106265	-0.470	0.158	-0.537	-0.515	-0.471
ΔUNCr	106265	0.003	0.027	-0.008	0.001	0.009
VAL	106265	0.356	1.908	-0.409	0.397	1.088
VIX	106265	2.849	0.352	2.600	2.766	3.014
ΔVIX	106265	-0.015	4.848	-1.970	-0.320	1.000

This table presents summary statistics for the full set of bond and firm characteristics and risk factors. Bond characteristics are aggregated on the firm level using a weighted average of a firm's outstanding bonds with the amount outstanding as weights. Variables representing dollar amounts are reported in billions of dollars. Bond returns, traded factor returns, and interest rates are reported in percent.

O-C Additional material for Section 4

O-C.1 Estimation with 14-year training sample

Our main model uses the first eight years of data for training followed by one year of validation data, with the entire second half of the sample reserved for testing. The long testing window allows for reliable and stable estimation for out-of-sample tests. We also analyze another model resulting from a longer training period of 14 years for robustness. With a validation window of one year, we are left with a little more than three years worth of data for testing. We again weight samples by market equity and adjust the set of parameters to account for the longer training period.

Performance measures for the resulting model are reported in Table O-C.1. The longer training window combined with the adjusted parameter set allows the EBM R^2 performance to increase to 16.7% on the shorter test set. The linear regression's test performance is on par with the nonlinear models in this case, but this result is unstable, as indicated by the poor validation performance and our observations from the other data splits.

Table O-C.1: Comparison of R^2's

	OLS	LASSO	R-Forest	XGBoost	EBM
R^2_{train}	25.06%	0.87%	43.35%	25.85%	30.32%
R^2_{valid}	−26.62%	7.69%	4.34%	2.53%	14.52%
R^2_{oos}	15.94%	0.06%	17.52%	16.26%	16.71%

This table shows R^2's of OLS, LASSO, Random Forest, XGBoost, and EBM estimations in the training (July 2002 - June 2016), validation (July 2016 - June 2017), and testing (July 2017 - August 2020) samples.

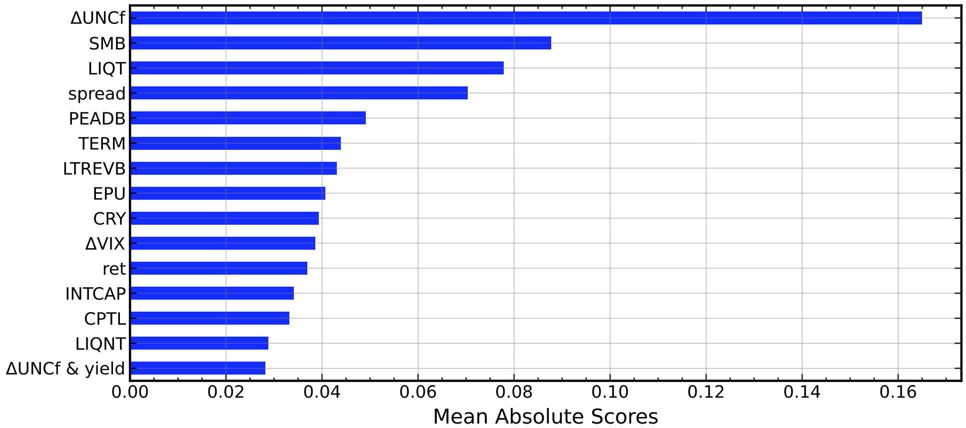
We present importances of predictors for the model in Figure O-C.1 using the in-sample measure as well as out-of-sample R^2 importance. Overall, the most important predictors are similar to the previously presented model. Uncertainty factors are at the top of both rankings, while spread again ranks as the most important characteristic. The term structure factor features less prominently but is still among the top six (seven) predictors in-sample (out-of-sample).

Consistent with the results for these predictors presented in Section 4, we observe a positive effect on returns associated with falling uncertainty and an adverse impact of rising uncertainty. These relationships are represented by step functions that emphasize the tails of the uncertainty distribution. These relationships again indicate an asymmetric response to changes in uncertainty, with a stronger market reaction to large increases. Additionally, we confirm the robustness of the monotonically increasing function we observe for the spread contribution to return predictions.

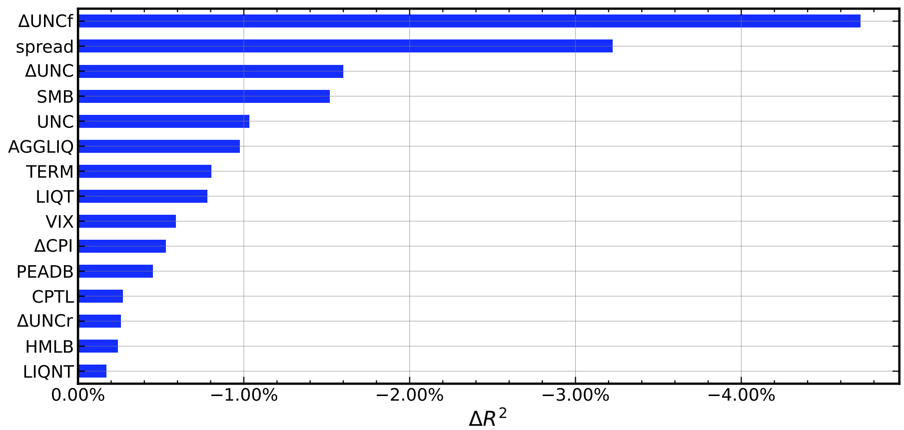
We construct portfolios based on return predictions analogously to the main model: In each month of the sample period, we sort bonds into deciles based on predicted future returns and observe the return of each decile portfolio. Using this procedure, we create equal-weighted and

Figure O-C.1: Importance of predictors

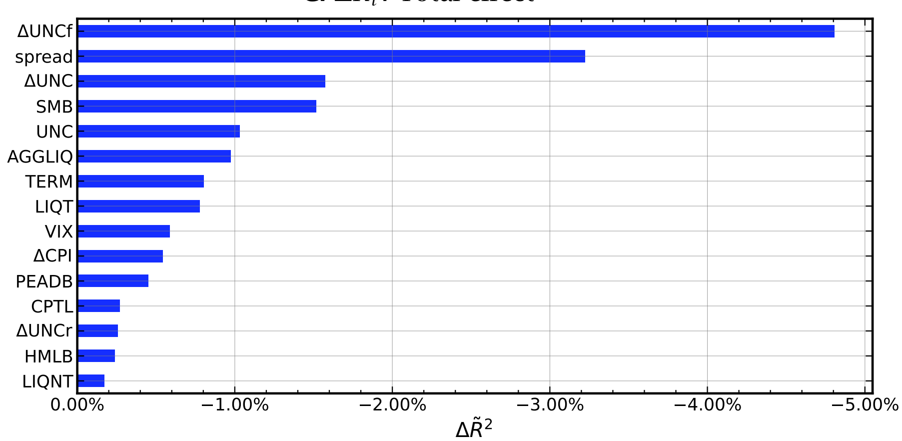
A: Mean Absolute Scores: $S(i)$ and $S(i,j)$



B: ΔR_i^2: Univariate effect



C: $\Delta \widetilde{R}_i^2$: Total effect



This figure shows the most important predictors for the EBM model resulting from the 14-year training window. In Panel A, importance values are mean absolute scores calculated on the training data (the native EBM importance measure). In Panel B, importance values are calculated as reductions in R^2 when removing a term, i.e., an individual shape function, from the fitted model. In Panel C, importance values are calculated as reductions in R^2 when removing all terms associated with a variable from the fitted model.

value-weighted portfolios, where each firm is weighted by the total amount outstanding of its bonds. Finally, we create long-short portfolios where we go long the highest decile and short the lowest decile. Cumulative returns of the standard value-weight portfolio and the EBM-pos portfolio, which exploits time-series forecastability, are displayed in Figure O-C.2, Panel A.

We report descriptive statistics, namely mean return, standard deviation, and Sharpe ratios for the decile portfolios and the high-minus-low (HML) long-short portfolios generated by the different trading strategies and weighting schemes in Table O-C.2. The value-weighted long-short portfolio has a monthly mean return of 0.93%, a monthly standard deviation of 1.39%, and a Sharpe ratio of 0.67.

Table O-C.2: Returns of EBM portfolios

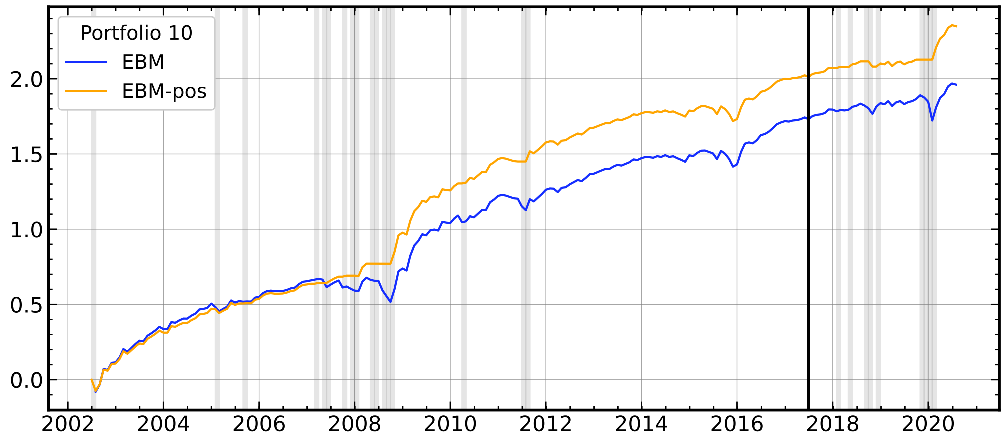
	1	2	3	4	5	6	7	8	9	10	10-1
					Panel A: EBM VW						
Mean	-0.35	-0.01	-0.03	0.06	0.03	0.09	0.17	0.19	0.29	0.58	0.93
S.D.	2.32	1.67	1.97	1.80	1.91	1.98	2.09	2.11	2.47	3.08	1.39
SR	-0.15	-0.00	-0.01	0.03	0.01	0.05	0.08	0.09	0.12	0.19	0.67
					Panel B: EBM-pos VW						
Mean	0.07	0.15	0.20	0.26	0.26	0.31	0.35	0.43	0.54	0.88	0.82
S.D.	0.32	0.41	0.50	0.58	0.72	0.77	0.99	1.06	1.34	2.16	2.01
SR	0.22	0.36	0.39	0.45	0.37	0.39	0.36	0.40	0.40	0.41	0.41
					Panel C: EBM EW						
Mean	-0.35	-0.03	-0.05	0.03	0.00	0.06	0.14	0.18	0.26	0.62	0.96
S.D.	2.32	1.61	2.00	1.79	2.05	2.07	2.16	2.34	2.51	3.16	1.40
SR	-0.15	-0.02	-0.02	0.02	0.00	0.03	0.07	0.08	0.10	0.20	0.69
					Panel D: EBM-pos EW						
Mean	0.06	0.14	0.19	0.26	0.27	0.30	0.35	0.46	0.50	0.94	0.88
S.D.	0.30	0.38	0.48	0.54	0.69	0.75	0.96	1.05	1.24	2.19	2.05
SR	0.19	0.37	0.41	0.47	0.40	0.40	0.36	0.44	0.41	0.43	0.43

This table compares monthly means, standard deviations, and Sharpe ratios for ten value-weighted portfolios constructed based on predicted return deciles in the testing sample. The last columns show the results of the 10-1 portfolio, which is long in portfolio 10 and short in portfolio 1. EW are equally-weighted portfolios and VW are value-weighted portfolios by bond market value. EBM-pos is the strategy that invests in the riskfree asset when the predicted portfolio return is negative.

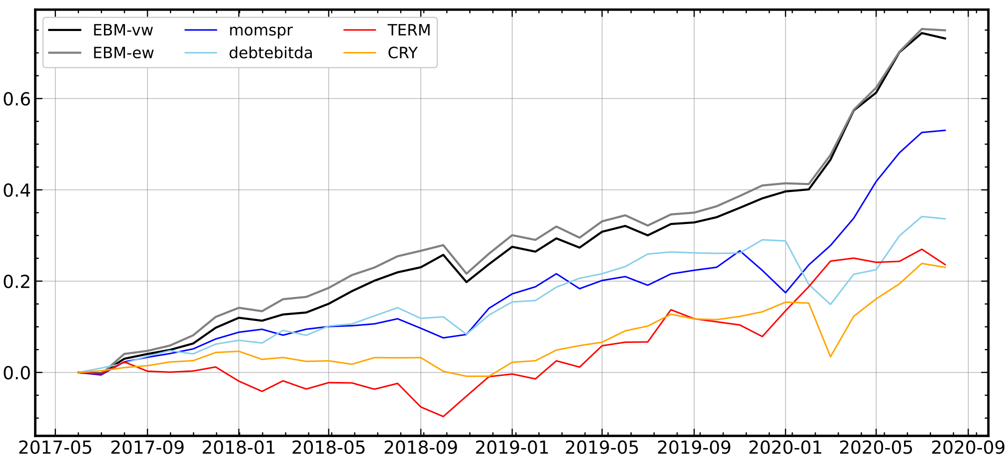
We compare this EBM portfolio to several benchmarks: We construct portfolios from monthly characteristics sorts and use traded factor portfolios from the bond market. Table O-C.3 reports test results for the ten best-performing (in terms of Sharpe ratio on the combined train and validation sample) characteristics portfolios and traded benchmark portfolios. Panel B of Figure O-C.2 compares the long-short EBM portfolio and the best-performing benchmarks over the testing window. These results indicate that the long-short EBM portfolio outperforms the benchmarks from characteristics sorts and factor strategies.

Figure O-C.2: Cumulative returns

A: EBM portfolio 10



B: EBM, characteristics, and factors



This figure shows cumulative log returns of EBM portfolios over the sample period. Panel A plots the value-weighted portfolio 10 of the standard buy-and-hold strategy (EBM) and the strategy that invests in the riskfree asset when the predicted portfolio return is negative (EBM-pos). The shaded areas indicate quarters in which the predicted portfolio return is negative. Panel B plots the cumulative log returns of the 10-1 EBM equal and value-weighted portfolios, the two 10-1 characteristic portfolios with the highest Sharpe ratios (momspr and debtebitda), and the two factors with the highest Sharpe ratios (TERM and CRY). Returns are scaled to have an annualized standard deviation of 10%. The black vertical line indicates the start of the testing period (July 2017).

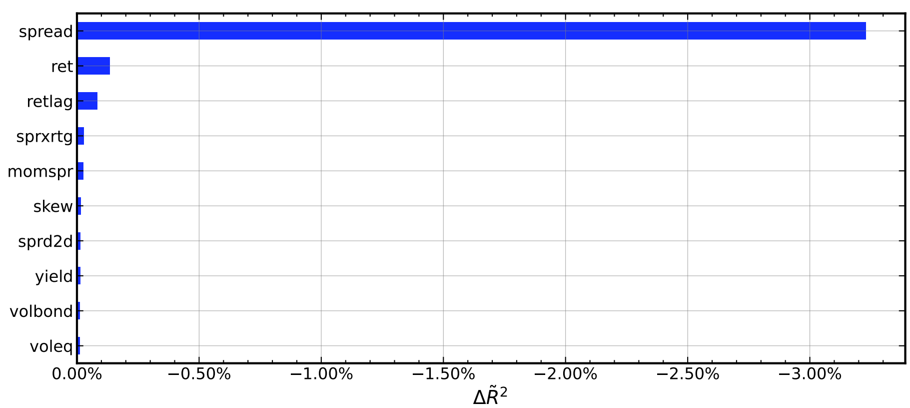
Table O-C.3: Returns of characteristics-sorted portfolios and factors

	Panel A: 5-1 characteristic portfolios									
	momspr	debtebitda	spread	yield	sprxrtg	voleq	ret	nime	coupon	sprd2d
Mean	0.42	0.12	0.30	0.29	0.24	0.16	0.20	0.04	0.10	0.01
S.D.	0.87	0.39	1.88	1.98	1.67	1.22	2.02	0.44	1.57	1.46
SR	0.49	0.31	0.16	0.15	0.14	0.13	0.10	0.08	0.06	0.01
	Panel B: Factors									
	TERM	CRY	MKT	MKTB	DUR	DRF	VAL	PEADB	HMLB	MKTBD
Mean	0.71	0.68	1.16	0.45	0.55	0.51	0.48	0.08	0.16	0.13
S.D.	3.11	2.97	5.20	2.06	2.76	2.58	2.48	0.72	1.54	2.26
SR	0.23	0.23	0.22	0.22	0.20	0.20	0.19	0.12	0.10	0.06

This table reports monthly means, standard deviations, and Sharpe ratios of characteristics-sorted portfolios and factors. We sort characteristics-sorted portfolios and factors by their Sharpe ratios in the combined training and validation samples. We also report statistics of the ten characteristics-sorted portfolios (Panel A) and factors (Panel B) in the testing sample.

Impact of Characteristics. We examine the importance of bond and firm characteristics for asset selection based on the previously presented R^2 importance. An overview is provided in Figure O-C.3. Spread is by far the most important predictor among characteristics, followed by two lagged return variables, $r_{i,t-1}$ and $r_{i,t}$. We focus on these top three, recalling the corresponding shape functions from Figure D.1.

Figure O-C.3: Importance of characteristics



This figure shows the most important characteristics in terms of their impact on out-of-sample predictions based on an EBM trained on the first 14 years of data. Importance values are calculated as reductions in R^2 when removing a term, i.e., an individual shape function, from the fitted model, yielding term-level importances.

From these shape functions, we can draw conclusions on the bonds selected during the portfolio construction: Based on the high return premium associated with high spreads (Figure D.1 Panel B), we expect high-spread bonds to feature in the long leg of the portfolio. In contrast, spreads should be lower for shorted bonds.

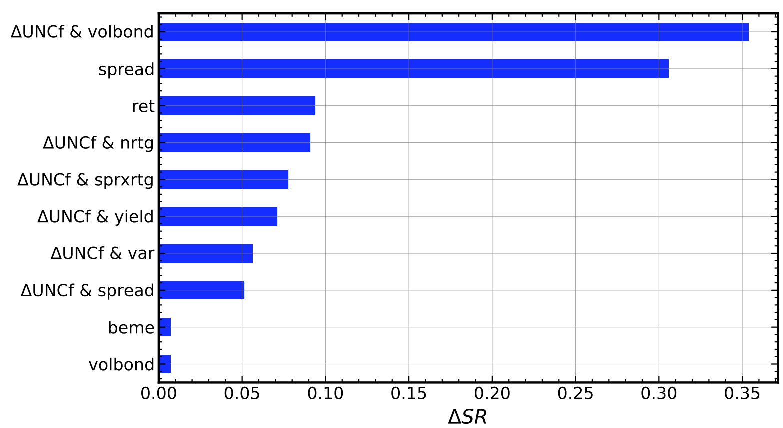
Table O-C.4: Portfolio properties

	spread	ret ($r_{i,t}$)
Long	5.51*	−41 bp
Short	1.02	67 bp*

This table compares mean characteristics value for the long and the short leg of the constructed EBM portfolio. Asterisks highlight the value we expect to be higher based on a characteristic's shape function.

Impact of Interactions. We also analyze the importance of predictors regarding their direct impact on the Sharpe ratio. As for the R^2 measure, we drop individual terms from the prediction equation and observe the drop in the Sharpe ratio of the resulting portfolio. Results are reported in Figure O-C.4. In line with the main model, we find that interactions of the financial uncertainty factor with bond characteristics frequently appear among the most important predictors in their contribution to the Sharpe ratio. In addition to the univariate characteristics spread and $r_{i,t}$, which are discussed in the context of the most important characteristics, we present the two highest-ranked interactions, ΔUNCf and volatility and ΔUNCf and value at risk. Figure O-C.5 shows the functions associated with these interactions.

Figure O-C.4: Importance in terms of SR



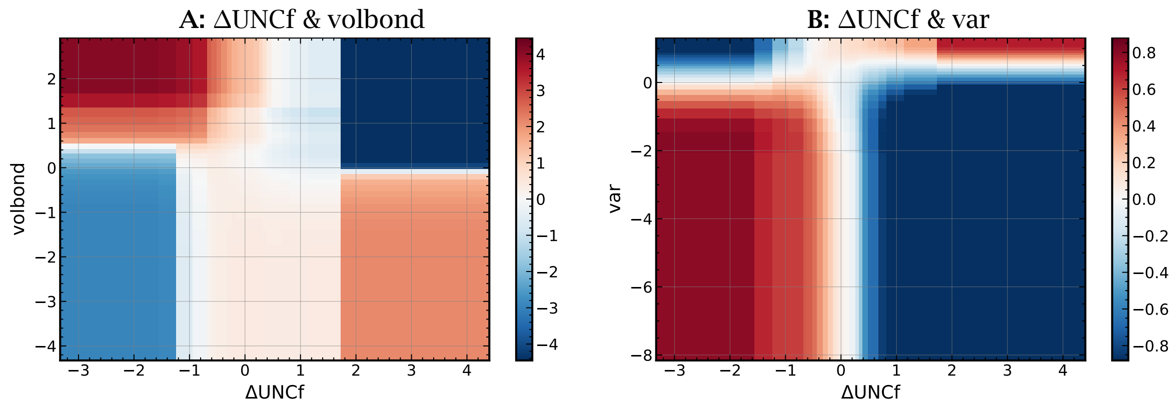
This figure shows the most important terms as measured by their impact on the Sharpe ratio of the EBM portfolio. Importance values are calculated as reductions in the Sharpe ratio when removing a term, i.e., an individual shape function, from the fitted model.

We derive implications for the portfolio composition from these terms: The most important one is the interaction between ΔUNCf and volatility, defined as bond return volatility over the past 24 months (see Kelly et al. (2023)). The function (Figure O-C.5 Panel A) suggests that bonds with high volatility are subject to a large penalty on future returns in times of rising uncertainty. In contrast, low-volatility bonds experience a comparatively small increase. The opposite occurs for falling uncertainty, which leads to a moderate return increase (decrease) for bonds with high (low) volatility. Thus, in the long leg of the EBM portfolio, high-volatility bonds should be picked in times of falling uncertainty, and low-volatility bonds should be picked in times of rising uncertainty. We expect this relationship to be reversed in the short leg of the portfolio, with high (low)-volatility bonds being selected in times of rising (falling) uncertainty.

Analysis of the correlation between ΔUNCf and volatility confirms that these expectations hold for the constructed portfolio: We measure a correlation coefficient of −0.30 in the long leg and of 0.28 in the short leg.

The reverse case can be observed for the interaction between ΔUNCf and value-at-risk[42], shown in Panel B of Figure O-C.5. Return increases are concentrated on high-var bonds in times of rising uncertainty and low-var bonds in times of falling uncertainty. Rising uncertainty also leads to return decreases for low-var bonds, as does falling uncertainty for high-var bonds. Therefore, we expect a positive correlation between ΔUNCf and var in the long portfolio and a negative correlation in the short portfolio. The correlations of 0.22 in the long leg and −0.41 in the short leg are consistent with this pattern.

Figure O-C.5: Interactions $f_{i,j}(x_i, x_j)$



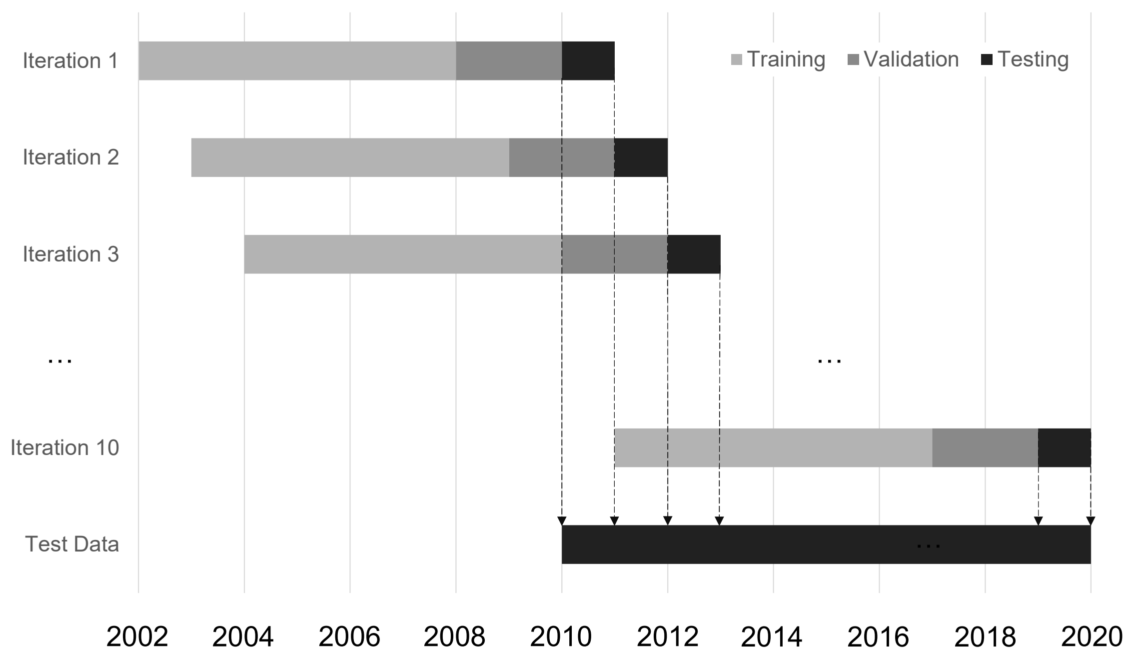
This figure displays shape functions of the most important interaction terms as measured by their impact on the Sharpe ratio. The respective features' scaled values are shown on the x and the y-axis. The color-coded score represents the additive contribution to the prediction output associated with x and y value combinations.

[42]Value-at-risk (var) is defined as the second-lowest credit excess return over the past 36 months.

O-C.2 Estimation in rolling window

Additionally, models are trained on rolling windows of data. We initially used the first six years of our dataset for training, the following two years for validation, and the remaining ten years until June 2020 for testing. The inputs for the EBM models are normalized with the 8-year training data when applying the rolling windows approach. As the EBM estimation procedure is fundamentally based on decision trees, which are robust to scaling, this leaves the results practically unchanged but allows for easier comparisons between windows. A performance comparison for all models is provided in Table O-C.5.

Figure O-C.6: Rolling windows



This figure visualizes the splitting scheme used to train machine learning models. Predictions for the full test dataset containing 10 years of data are produced year-by-year based on rolling training and validation data windows. The sample period begins in July 2002 and ends in June 2020.

Performance comparison Overall, relative model performance is comparable to the results from our static data split: The linear regression performs poorly out-of-sample, while the LASSO model substantially improves performance, with R^2 reaching almost 5% in the rolling windows estimation. The more complex models improve this result and produce predictions of similar quality. All handily outperform linear models and a constant benchmark prediction of 0 with R^2_{oos} values around 8 to 9%. The EBM again provides an interpretable model at a small cost in terms of performance.

Aggregated importances Analogous to the R^2 importance calculations presented in Section 4, we repeat these calculations for each of the ten models trained, validated, and tested on rolling data

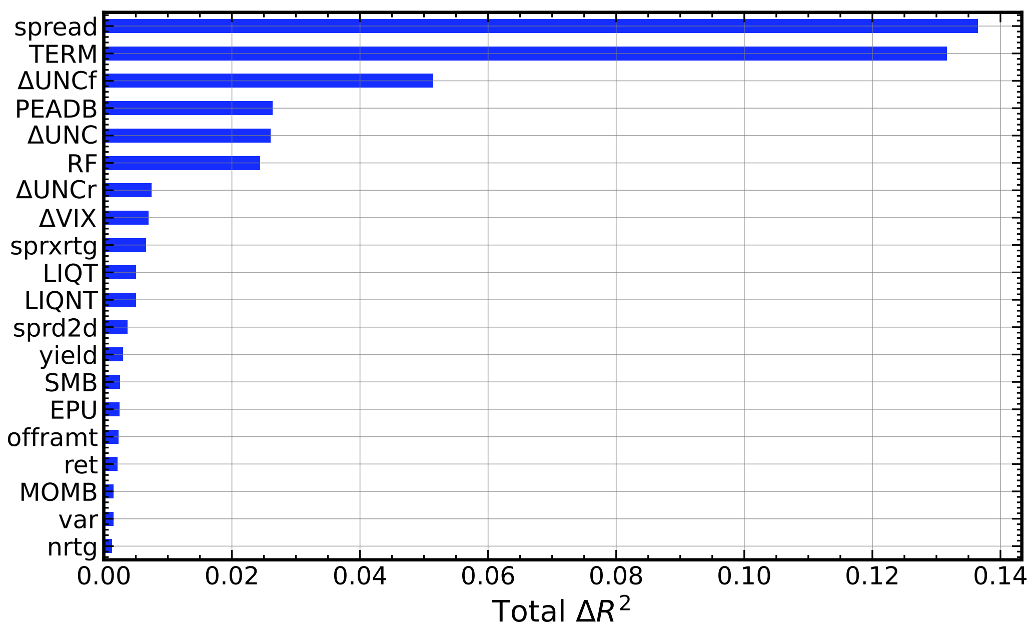
Table O-C.5: Out-of-sample R^2_{oos}

Model	OLS	Lasso	R-Forest	XGBoost	EBM
R^2_{oos}	-206.75	5.04	7.77	9.56	7.46

This table compares machine learning models based on out-of-sample performance measured by R^2 (reported in percent). Models are fit using a rolling windows approach, with the total testing data covering the period from July 2010 to June 2020.

windows. We sum the importances of all terms across windows. These aggregated results are reported in Figure O-C.7. The aggregated measure confirms the previously documented importance of the term structure factor TERM, bond spread, and uncertainty factors such as ΔUNC and ΔUNCf.

Figure O-C.7: Importance across all windows



This figure shows the top 20 predictors in terms of their importance for out-of-sample predictions. Importances are calculated as reductions in R^2 when removing a term from the fitted model and aggregated over ten EBM models trained on different rolling windows.

O-C.3 Subsample: Global financial crisis

We examine how the influence of features on return predictions differs in times of macroeconomic turmoil by focusing on the global financial crisis (GFC). We restrict our dataset to July 2007 to June 2009 and fit an EBM model on training data from the first 18 months. We then divide the remaining six months of data evenly into validation and test data. We normalize the predictor variables using the mean and standard deviation of the 8-year training data to ensure comparability of the units of the output function plots. The model achieves an in-sample R^2 of 50.71%, combined with impressive out-of-sample performance that produces an R^2 value of 48.14%. The corresponding importances from the in-sample and the out-of-sample measures are displayed in Figure O-C.8.

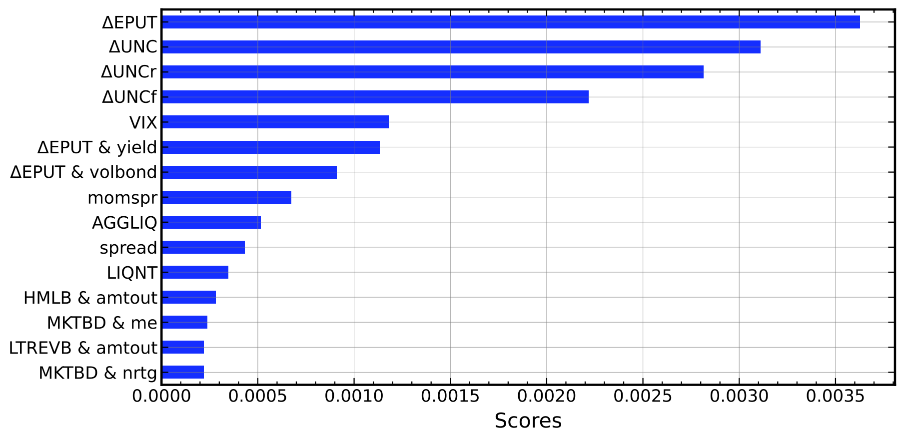
Compared to the main model, we observe an even stronger emphasis on predictors related to macroeconomic uncertainty in the GFC ranking. Factors capturing different facets of uncertainty frequently appear in both rankings as univariate terms and as part of interactions. This most prominently includes the first difference in the financial uncertainty index (UNCf) and the first difference in the economic tax policy uncertainty index (EPUT), which underlines the role of policymakers in influencing markets during a crisis.

Additionally, we can use the shape functions discovered by the crisis model to better understand underlying changes compared to model outputs covering the entire sample period. Figure O-C.9 displays the functions associated with the four main predictors of the 8-year model. We find a similarly shaped function for the term structure factor in times of crisis, although its impact is considerably reduced compared to the main model. Its maximum contribution to returns is about five times smaller, consistent with its much lower ranking in terms of importance. These observations indicate that the influence of term structure is displaced by other effects during the crisis that are accounted for by other variables.

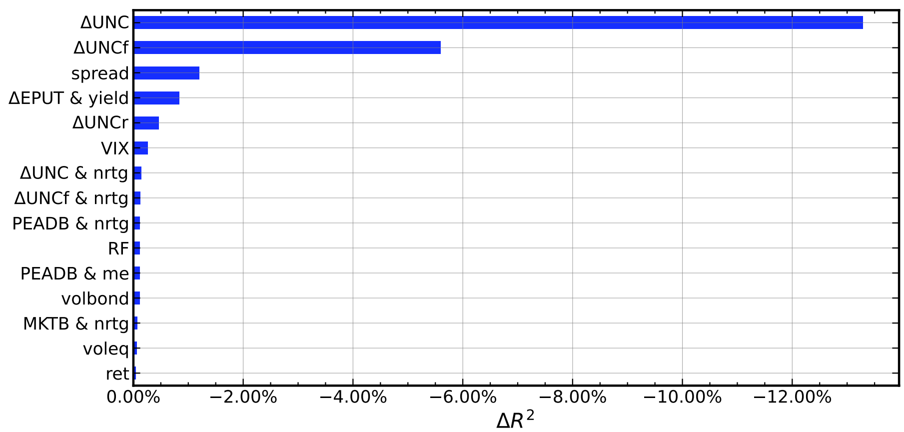
We also find a qualitative resemblance between the shape functions for the two uncertainty factors ΔUNC and ΔUNCf: We again observe a negative trend, where falling uncertainty is associated with a negligible or positive effect and rising uncertainty has a strong negative effect on future return. However, the distribution of the underlying ΔUNC and ΔUNCf data in the training period is noticeably flatter during the crisis, emphasizing realizations at the distribution's tails. The data and the resulting contributions curve for ΔUNCf are also shifted rightwards, indicating that increases in financial uncertainty outweigh decreases in the crisis period compared to the overall ΔUNCf data, which is roughly centered around zero. While the functions themselves are shaped similarly, the main difference is that effects set in earlier for the crisis model towards both ends of the ΔUNCf distribution, i.e., for smaller absolute changes in the ΔUNCf index. In the overall model, effects only occur after changes in financial uncertainty deviate from the mean by about 1 to 1.5 standard deviations at either side of the distribution's center. In contrast, the area where effects are small is roughly half as big in the crisis model, suggesting heightened sensitivity of returns to uncertainty movements during the financial crisis. Finally, the GFC analysis confirms the robustness of the positive risk compensation offered by high-spread bonds. This effect is similar in size to the main model for most of the data distribution but is significantly larger for the highest spreads, indicating increased risk compensation during the crisis.

Figure O-C.8: Importance of predictors

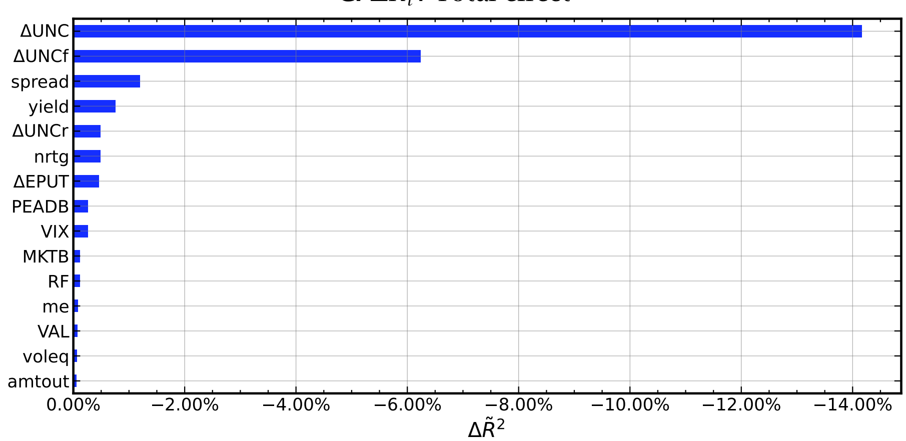
A: Mean Absolute Scores: $S(i)$ and $S(i,j)$



B: ΔR_i^2: Univariate effect

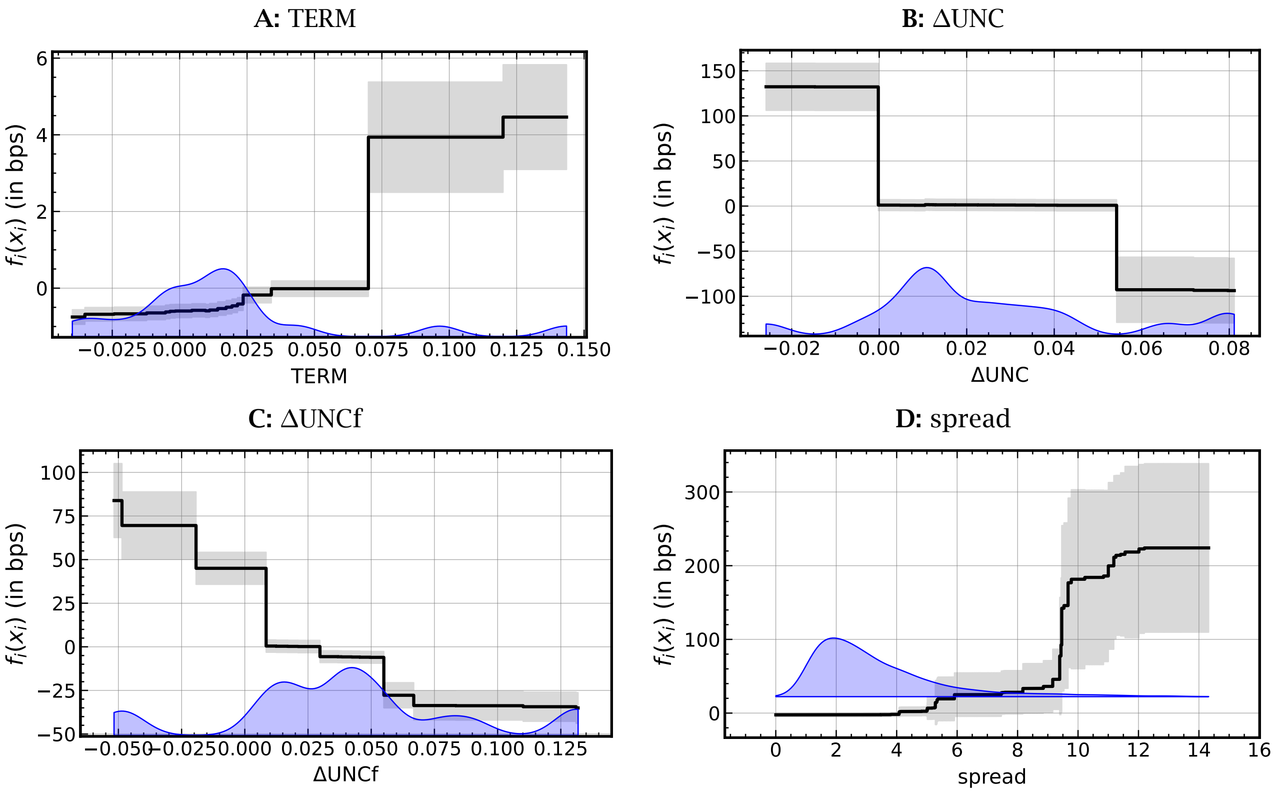


C: $\Delta\widetilde{R}_i^2$: Total effect



This figure shows the most important predictors for the GFC model. In Panel A, importance values are mean absolute scores calculated on the training data (the native EBM importance measure). In Panel B, importance values are calculated as reductions in R^2 when removing a term, i.e. an individual shape function, from the fitted model. In Panel C, importance values are calculated as reductions in R^2 when removing all terms associated with a variable from the fitted model.

Figure O-C.9: Univariate functions $f_i(x_i)$



This figure shows the contributions to return predictions of the four most important out-of-sample predictors from the main model during the global financial crisis. The x-axis describes the input variable standardized to a mean of 0 and a standard deviation of 1, while the y-axis represents the additive contribution to the prediction output.

O-C.4 Omission of Jurado et al. (2015) uncertainty factors

In the main analysis, we follow Jurado et al. (2015) and adopt a historical-data perspective with respect to macroeconomic variables required to construct their uncertainty indices. They argue that relying solely on real-time data is likely overly restrictive as a proxy for the information set available to forecasters. For robustness, we re-estimate the models in the most restrictive case where no information related to these uncertainty factors is used. Consistent with Figure 3, out-of-sample R^2 drops. However, as Table O-C.6 shows, this affects all models uniformly. Machine learning models continue to outperform linear methods, with EBM performance similar to black box models.

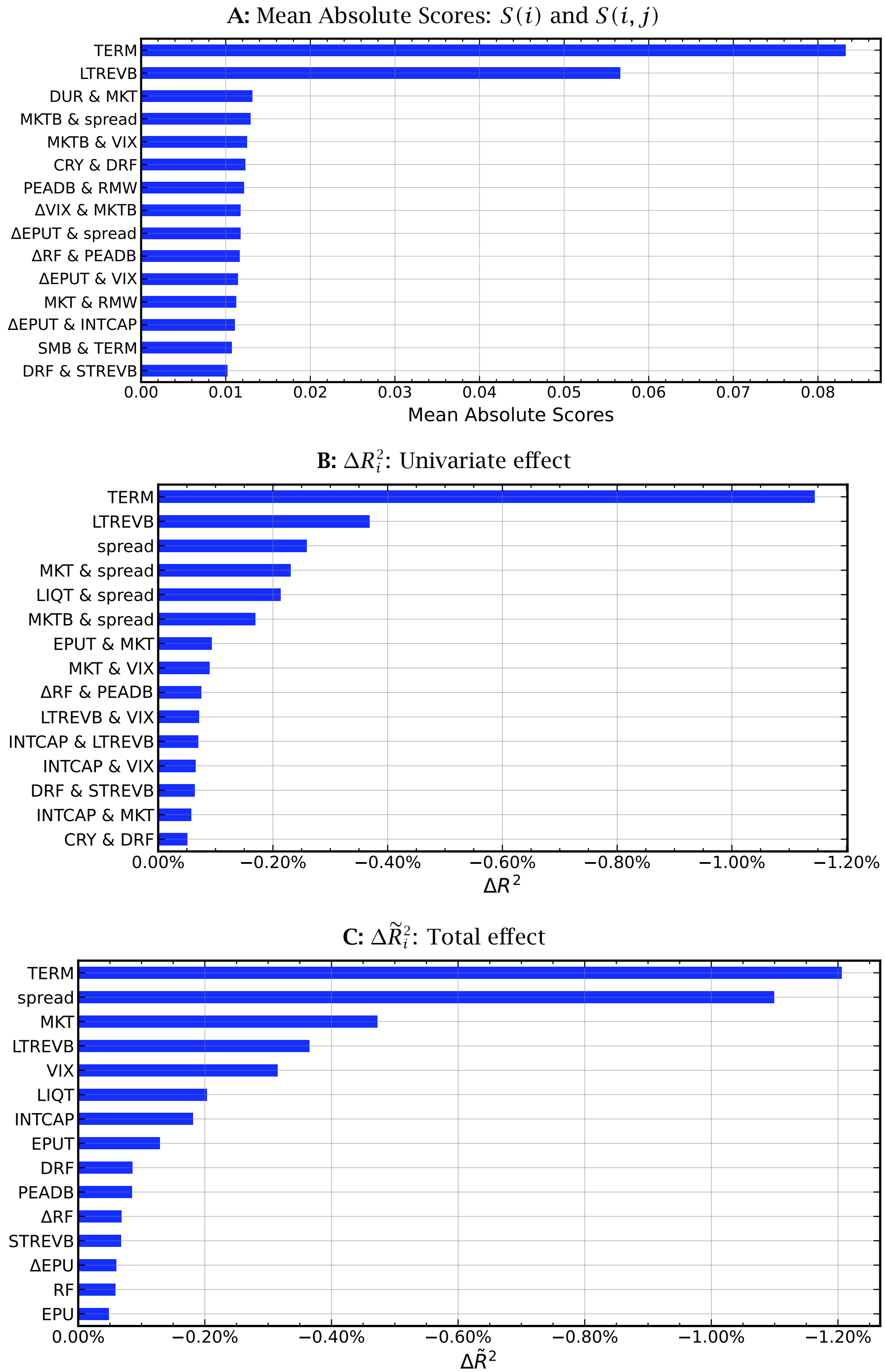
After retraining the EBM model using the original hyperparameter set, we confirm the importance of macroeconomic variables, lead by the term structure factor, and the importance of spread as the most influential characteristic (see Figure O-C.10). Additionally, this analysis confirms that interactions remain highly relevant to bond return predictions. We find very similar shape functions for TERM and spread compared to the main model (see Panels A and B of Figure O-C.11). In the absence of the Jurado et al. (2015) uncertainty factors, other variables gain relevance whose shape functions resemble the previously documented behavior. In particular, terms related to market factors and the Baker et al. (2016) uncertainty measures feature prominently in the rankings. Corresponding univariate and interaction shape functions are displayed in Panels C to F of Figure O-C.11. They confirm that future corporate bond returns are adversely affected by large increases in uncertainty and large downward market movements, complementing our main analysis. The interaction terms closely mirror the previously documented interaction of ΔUNCf and spread, indicating that the identified effect (outsized recovery of high-spread bonds when macroeconomic conditions improve) generalizes beyond the Jurado et al. (2015) uncertainty factors.

Table O-C.6: Comparison of R^2's

	OLS	LASSO	R-Forest	XGBoost	EBM
R^2_{train}	27.56%	18.29%	36.08%	18.62%	26.57%
R^2_{valid}	4.27%	9.47%	11.59%	4.25%	12.47%
R^2_{oos}	-13.33%	3.42%	6.02%	6.83%	6.76%

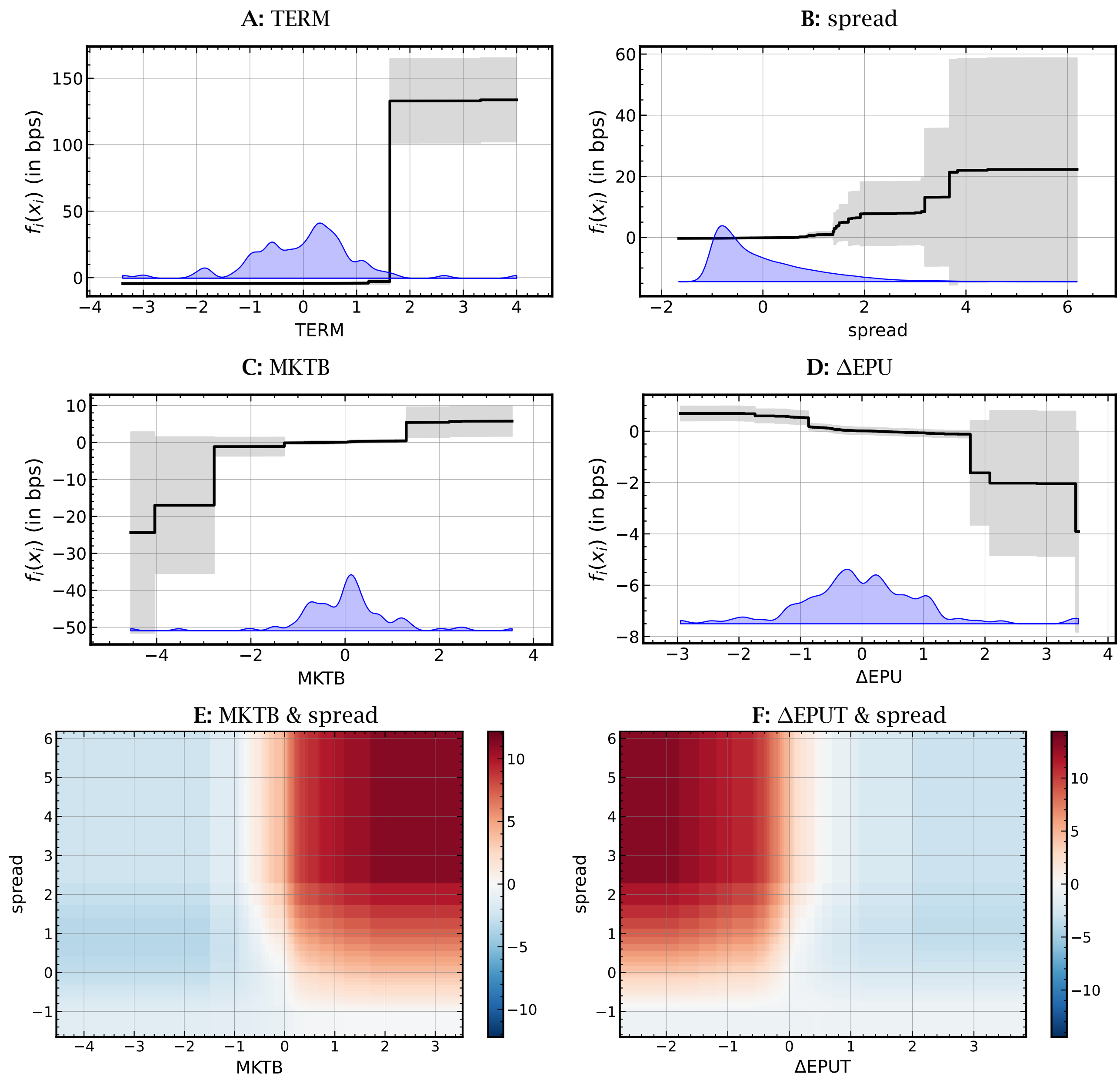
Note: This table shows R^2's of OLS, LASSO, Random Forest, XGBoost, and EBM estimations in the training, validation, and testing samples when refitting models without the Jurado et al. (2015) uncertainty factors.

Figure O-C.10: Importance of predictors



Note: This figure shows the most important predictors for the EBM estimated without the Jurado et al. (2015) uncertainty factors. In Panel A, importance values are mean absolute scores calculated on the training data (the native EBM importance measure). In Panel B, importance values are calculated as reductions in R^2 when removing a term, i.e., an individual shape function, from the fitted model. In Panel C, importance values are calculated as reductions in R^2 when removing all terms associated with a variable from the fitted model.

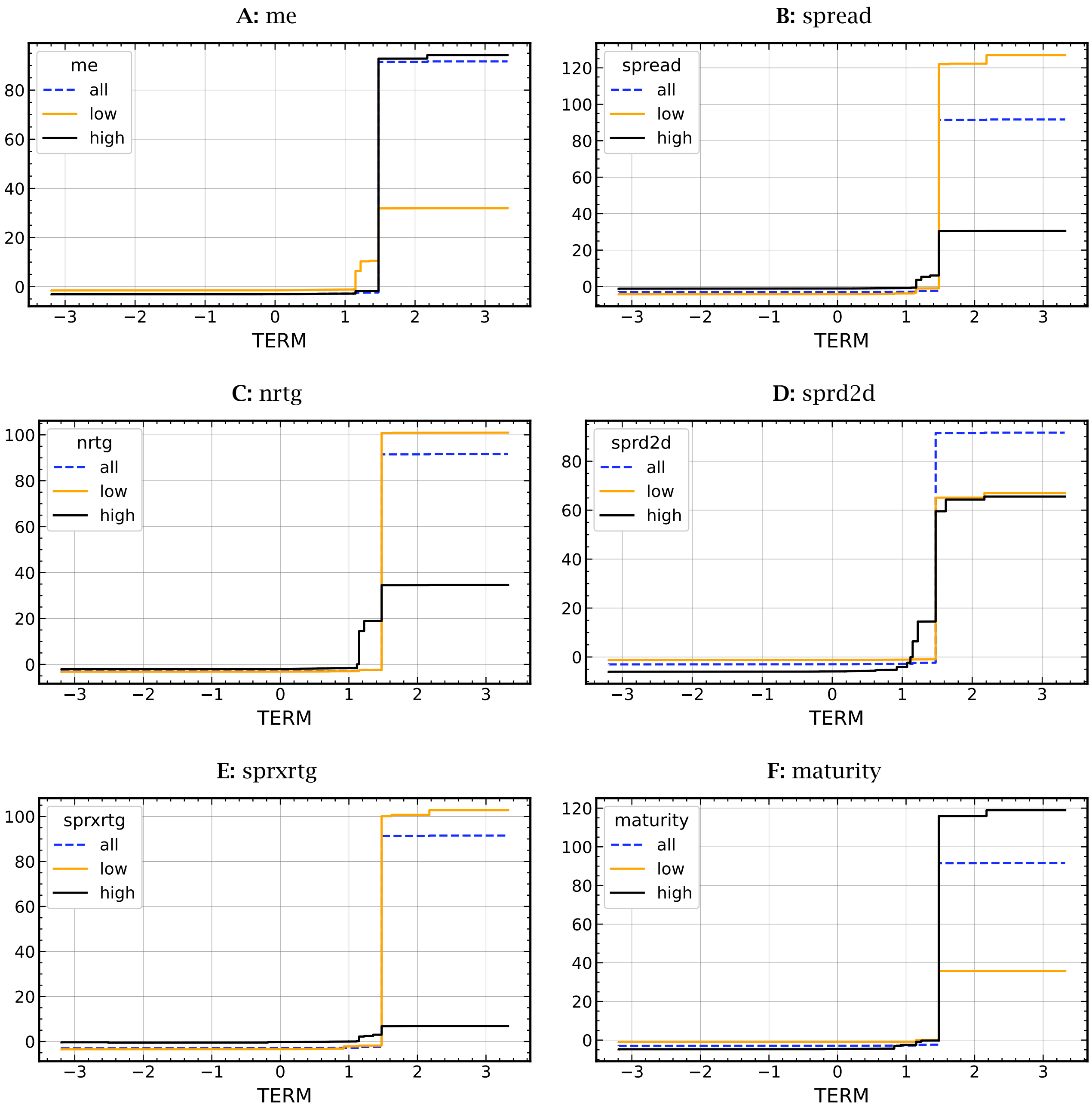
Figure O-C.11: Shape functions after refitting



This figure shows shape functions for the EBM estimated without the Jurado et al. (2015) uncertainty factors. The x-axis describes the input variable standardized to a mean of 0 and a standard deviation of 1, while the y-axis represents the additive contribution to the prediction output.

O-C.5 The role of TERM in more detail

Figure O-C.12: $f_i(x_i)$ – TERM by subset of firms



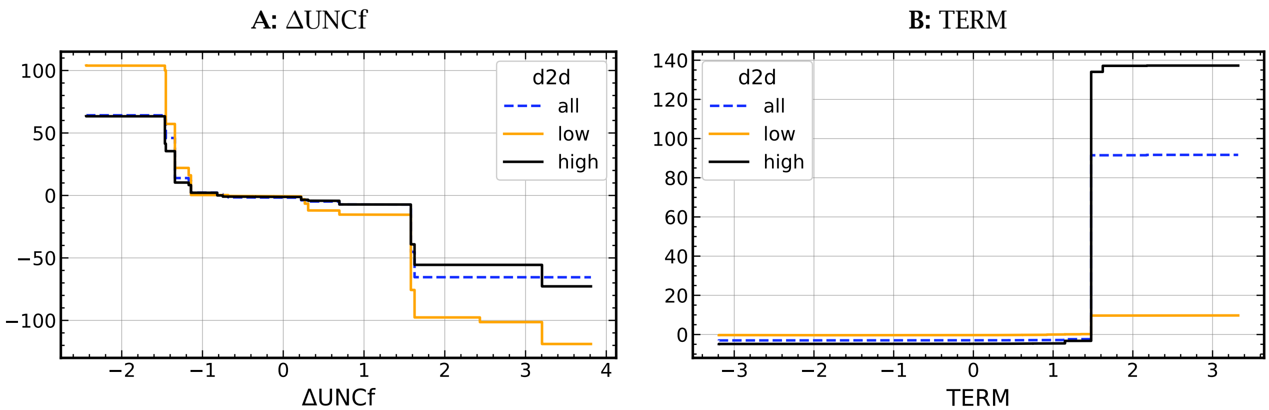
This figure compares shape functions associated with ΔUNCf across subsets of firms obtained from monthly sorts on firm size, spread, rating, maturity, sprd2d, sprxrtg, and maturity. Each month, firms are grouped into "low" and "high" subsets according to one of the four variables. The EBM model is then estimated on the two subsets of firms. The dashed blue line plots the shape function for the full sample with all firms. Models are trained on the first 8 years of data, validated on the following year, and tested in the remaining sample.

O-C.6 Additional firm subsamples

We explore heterogeneity in the observed effects with respect to additional characteristics.

Distance to default. We sort firms into monthly subsets based on distance-to-default (d2d). The distance-to-default measure as described by Bharath and Shumway (2008) is derived from a structural model of default risk, where low (high) distance-to-default implies a low (high) probability of default. As Figure O-C.13 illustrates, the previously described positive (negative) effect of rising (falling) financial uncertainty on bond returns seems to be more pronounced for firms that are deemed closer to default, with effects for low-d2d firms reaching roughly twice the magnitude of effects for high-d2d firms. Conversely, financially stable high-d2d firms are more sensitive to realizations of the term structure factor, with the corresponding effect exceeding the low-d2d effect by an order of magnitude. This is intuitive as the bonds of creditworthy firms are more aligned with risk-free government bonds and are more impacted by term structure risk than credit risk, while the opposite is true for low-grade bonds.

Figure O-C.13: $f_i(x_i)$ by subsamples – d2d



This figure compares shape functions associated with ΔUNCf and TERM across separate subsets of bonds obtained form monthly sorts on bond distance to default. Models are trained on the first 8 years worth of data and validated on the following year.

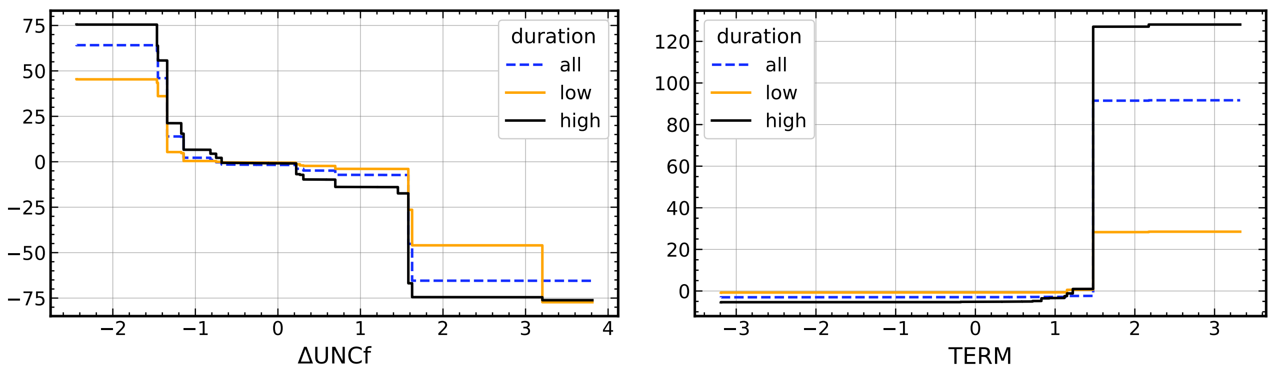
Duration. Conducting a similar data split based on the duration of the bond yields further subsample results: The respective contributions of changes in financial uncertainty and the term structure factor to the prediction of the return are shown in Figure O-C.14. When comparing both shape functions for ΔUNCf, we observe that low-duration firms are consistently less affected by changes in financial uncertainty than high-duration firms. The contribution associated with ΔUNCf values in the tails is about twice as high for high-duration firms, indicating that the sensitivity of bonds to uncertainty fluctuations is correlated with the sensitivity to interest rate changes. This disparity is even more pronounced for the function associated with TERM, where the maximum contribution is about four times that of high-duration firms. This result is expected as TERM represents interest rate change risk and should be closely aligned with duration.

Overall, the duration-based analysis indicates that a firm's exposure to uncertainty movements and term structure risk strongly correlates with its bonds' sensitivity to interest rate changes, with investors reacting to increases in macroeconomic uncertainty by disproportionately divesting from high-duration bonds.

Figure O-C.14: $f_i(x_i)$ by subsamples – duration

A: ΔUNCf

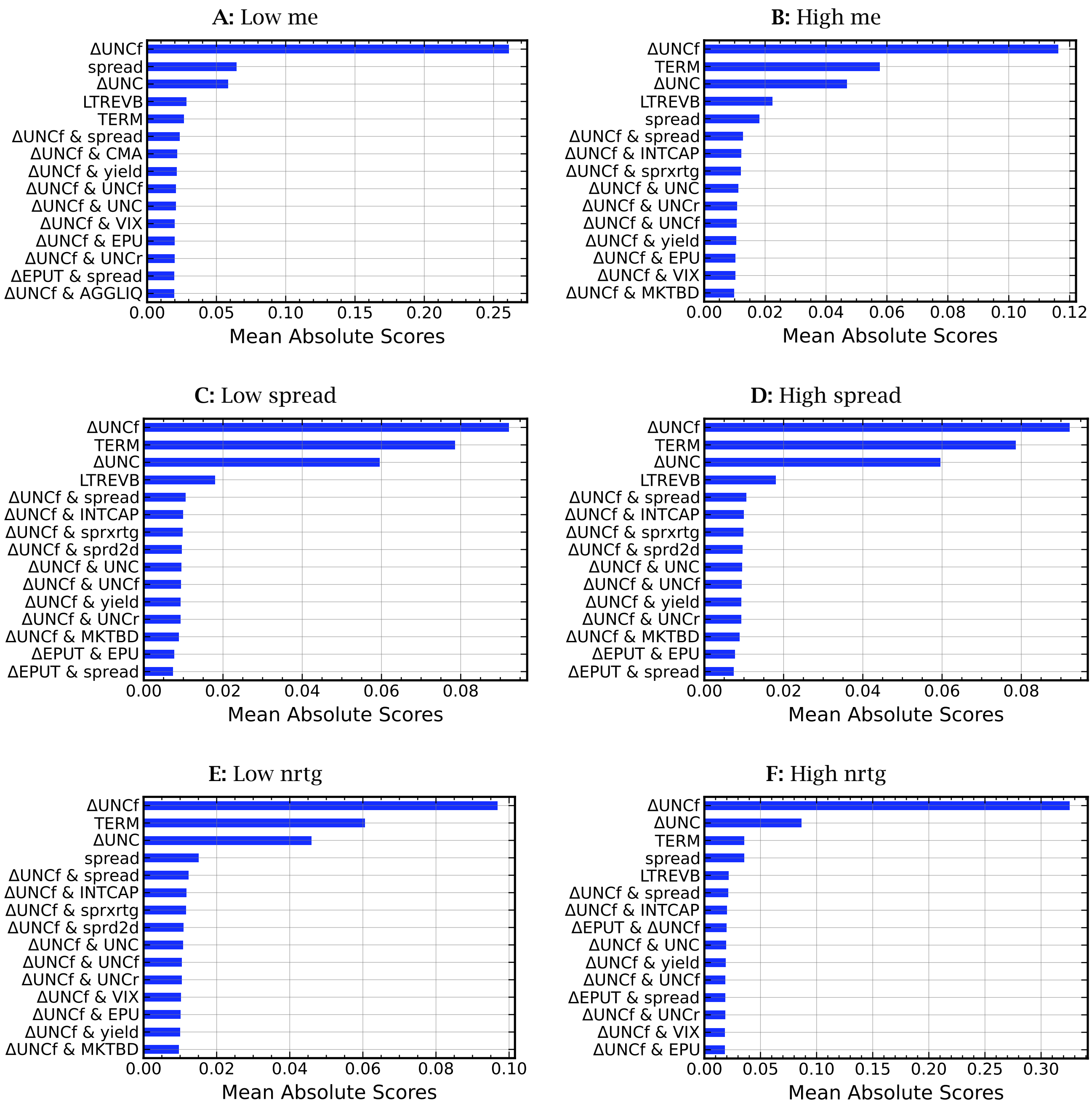
B: TERM



This figure compares shape functions associated with ΔUNCf and TERM across separate subsets of bonds obtained from monthly sorts on bond duration. Models are trained on the first 8 years of data and validated on the following year.

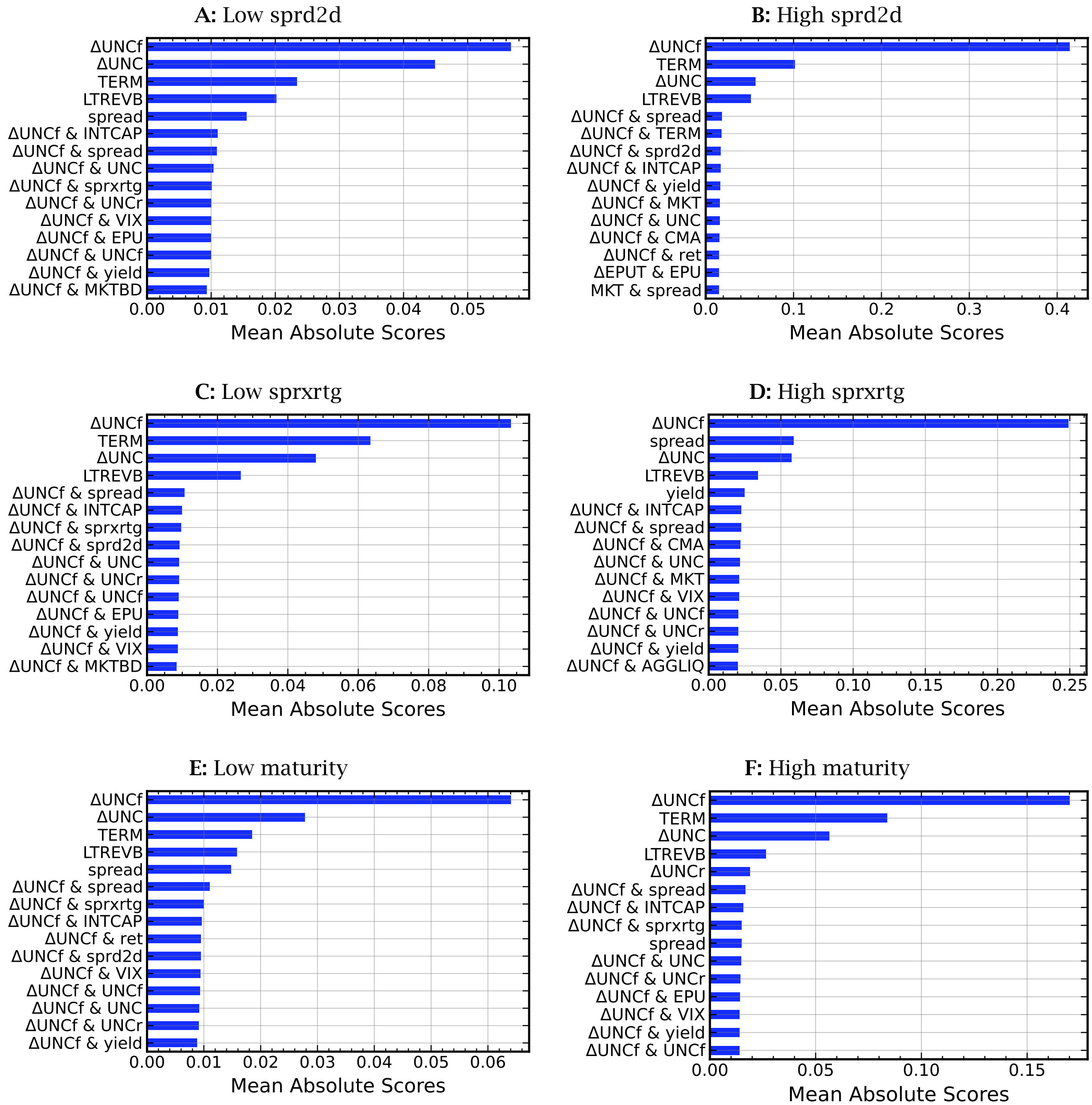
O-C.7 Subsample importances

Figure O-C.15: Mean Absolute Scores: $S(i)$ and $S(i, j)$



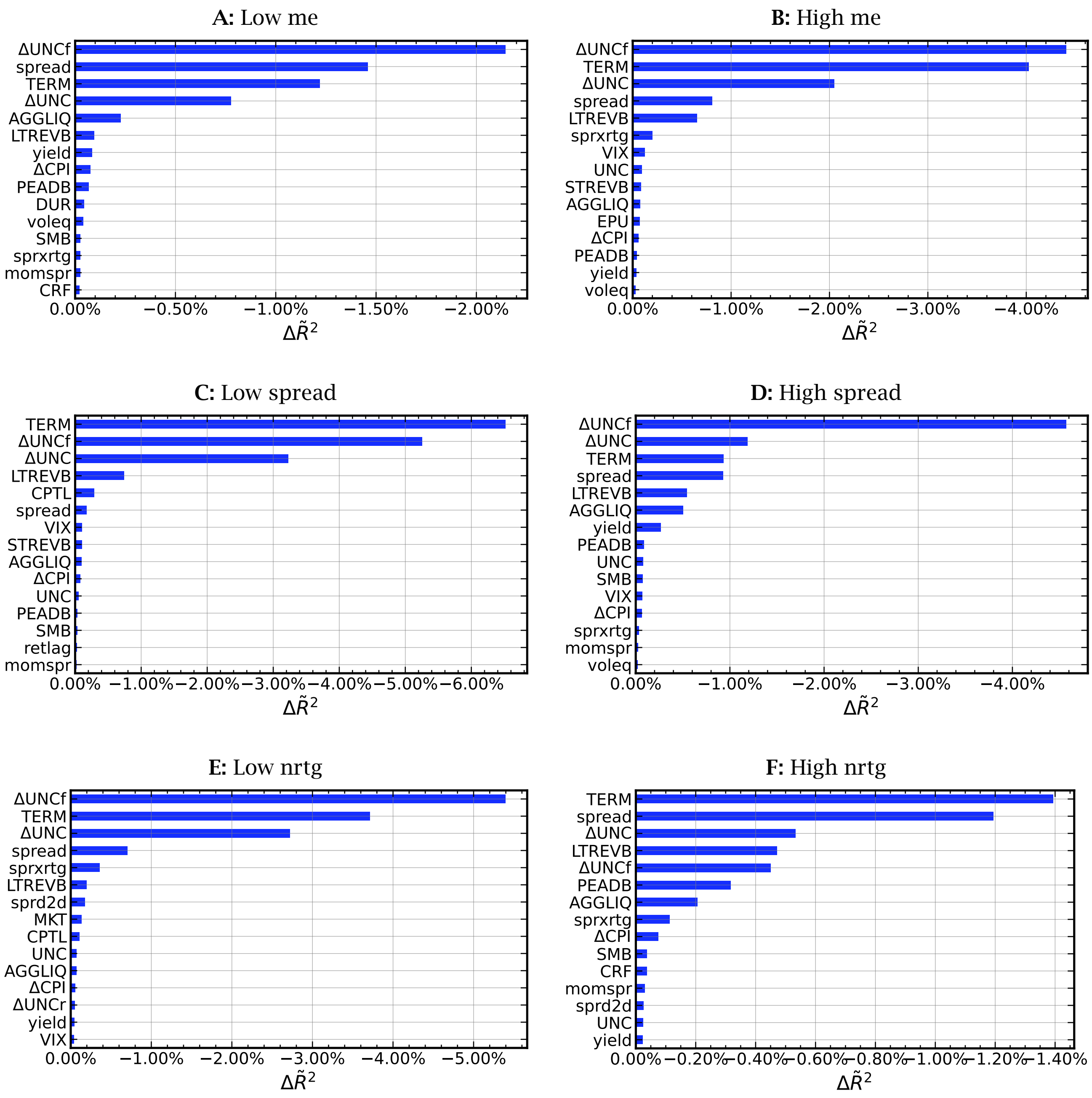
This figure presents mean absolute score importances for models based on characteristics-sorted subsamples (market cap, bond spread, and numerical rating).

Figure O-C.16: Mean Absolute Scores: $S(i)$ and $S(i,j)$



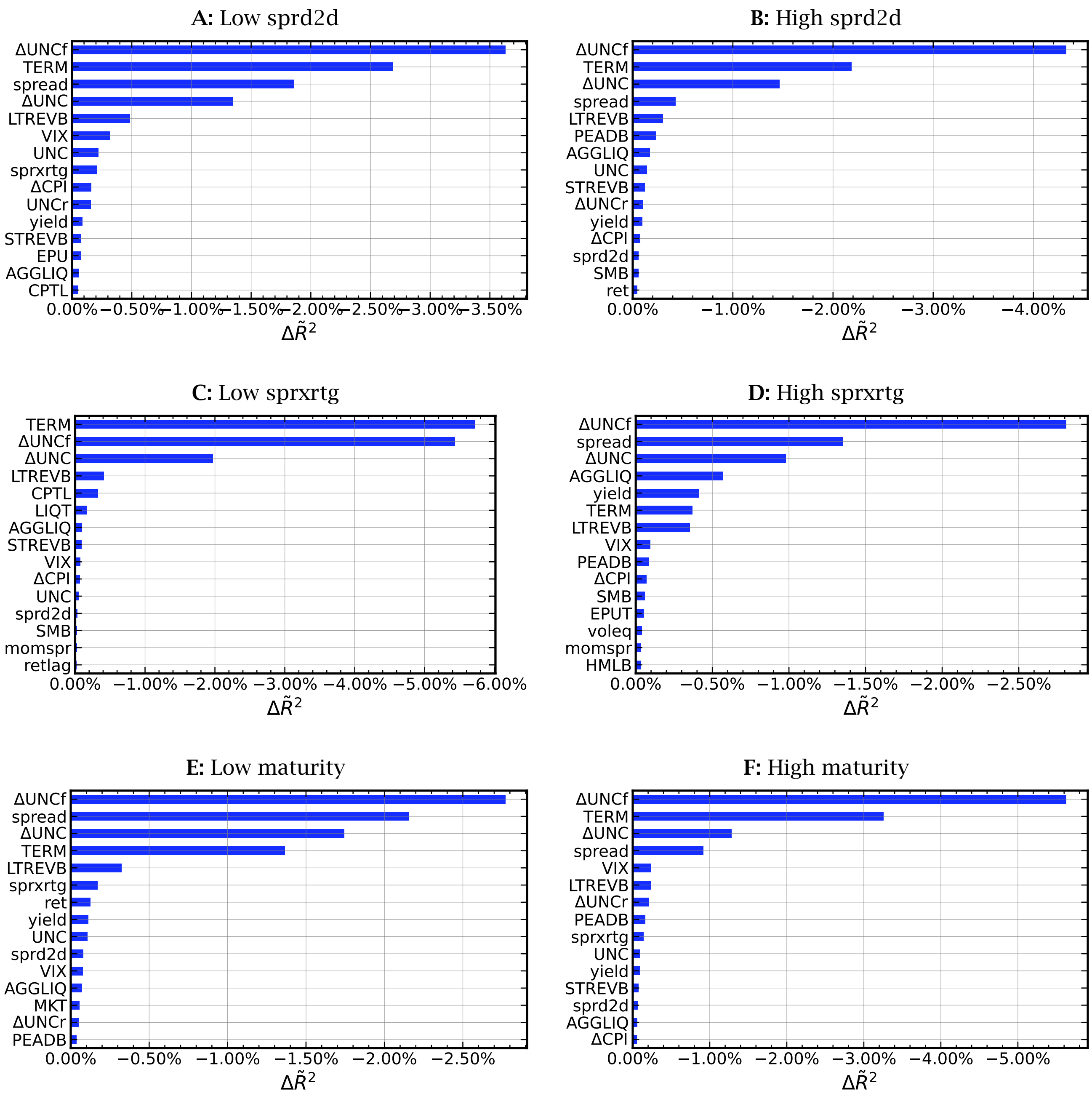
This figure presents mean absolute score importances for models based on characteristics-sorted subsamples (spread/d2d, spread times rating, and maturity).

Figure O-C.17: $\Delta\tilde{R}^2_i$ – Total effect



This figure presents total effects on R^2_{oos} of predictors for models based on characteristics-sorted subsamples (market cap, bond spread, and numerical rating).

Figure O-C.18: $\Delta\tilde{R}_i^2$ – Total effect



This figure presents total effects on R^2_{oos} of predictors for models based on characteristics-sorted subsamples (spread/d2d, spread times rating, and maturity).

O-D Additional material for Section 5

O-D.1 Portfolio returns for other models

We also construct bond portfolios based on the Random Forest (RF) and the XGBoost (XGB) model return predictions. We rank firms based on predicted future excess returns and create monthly decile portfolios (equal-weighted and value-weighted) as well as corresponding long-short portfolios by taking long positions in the highest decile and short positions in the lowest decile. We also consider the alternative investment strategy that invests in the riskfree asset if the predicted excess return of a portfolio is negative, exploiting time-series forecastability. Tables O-D.1 and O-D.2 report mean returns, standard deviations, and Sharpe ratios of the decile portfolios and the 10-1 long/short portfolio in the testing sample of July 2011 to August 2020 for the RF and XGB model, respectively.

Table O-D.1: Returns of Random Forest portfolios

	1	2	3	4	5	6	7	8	9	10	10-1
					Panel A: RF VW						
Mean	-0.11	-0.05	0.01	0.10	0.15	0.23	0.27	0.32	0.45	0.68	0.79
S.D.	1.82	1.66	1.49	1.46	1.40	1.43	1.52	1.54	1.77	2.25	2.01
SR	-0.06	-0.03	0.01	0.07	0.11	0.16	0.18	0.21	0.25	0.30	0.39
					Panel B: RF-pos VW						
Mean	0.07	0.11	0.15	0.21	0.26	0.30	0.34	0.40	0.51	0.68	0.61
S.D.	0.48	0.55	0.63	0.68	0.78	0.89	1.10	1.18	1.41	2.16	1.92
SR	0.14	0.19	0.24	0.30	0.33	0.34	0.31	0.34	0.36	0.31	0.32
					Panel C: RF EW						
Mean	-0.14	-0.07	0.01	0.09	0.13	0.24	0.27	0.33	0.45	0.73	0.87
S.D.	1.87	1.69	1.53	1.49	1.42	1.48	1.57	1.55	1.74	2.25	2.05
SR	-0.08	-0.04	0.00	0.06	0.09	0.16	0.17	0.21	0.26	0.32	0.42
					Panel D: RF-pos EW						
Mean	0.05	0.10	0.15	0.20	0.25	0.31	0.35	0.40	0.53	0.73	0.68
S.D.	0.43	0.48	0.56	0.62	0.73	0.86	1.08	1.17	1.34	2.15	1.94
SR	0.12	0.21	0.27	0.32	0.35	0.36	0.32	0.35	0.39	0.34	0.35

This table compares monthly means, standard deviations, and Sharpe ratios for ten value-weighted portfolios constructed based on predicted return deciles in the testing sample. The last columns shows results of the 10-1 portfolio that is long in portfolio 10 and short in portfolio 1. EW are equally-weighted portfolios and VW are value-weighted portfolios by bond market value. RF-pos is the strategy that invests in the riskfree asset when the predicted portfolio return is negative.

Observed return patterns are similar to the EBM portfolios, with monotonically increasing mean returns and Sharpe ratios along the buy-and-hold decile portfolios. RF-pos and XGB-pos substantially increase returns and Sharpe ratios, indicating that both models enable profitable market timing. Overall, RF portfolios exhibit similar returns and Sharpe ratios across all weightings and strategies. The XGB predictions offer a higher mean return and Sharpe ratio for the equal-weighted buy-and-hold portfolio (driven by lower returns in the lowest decile) and perform similarly otherwise.

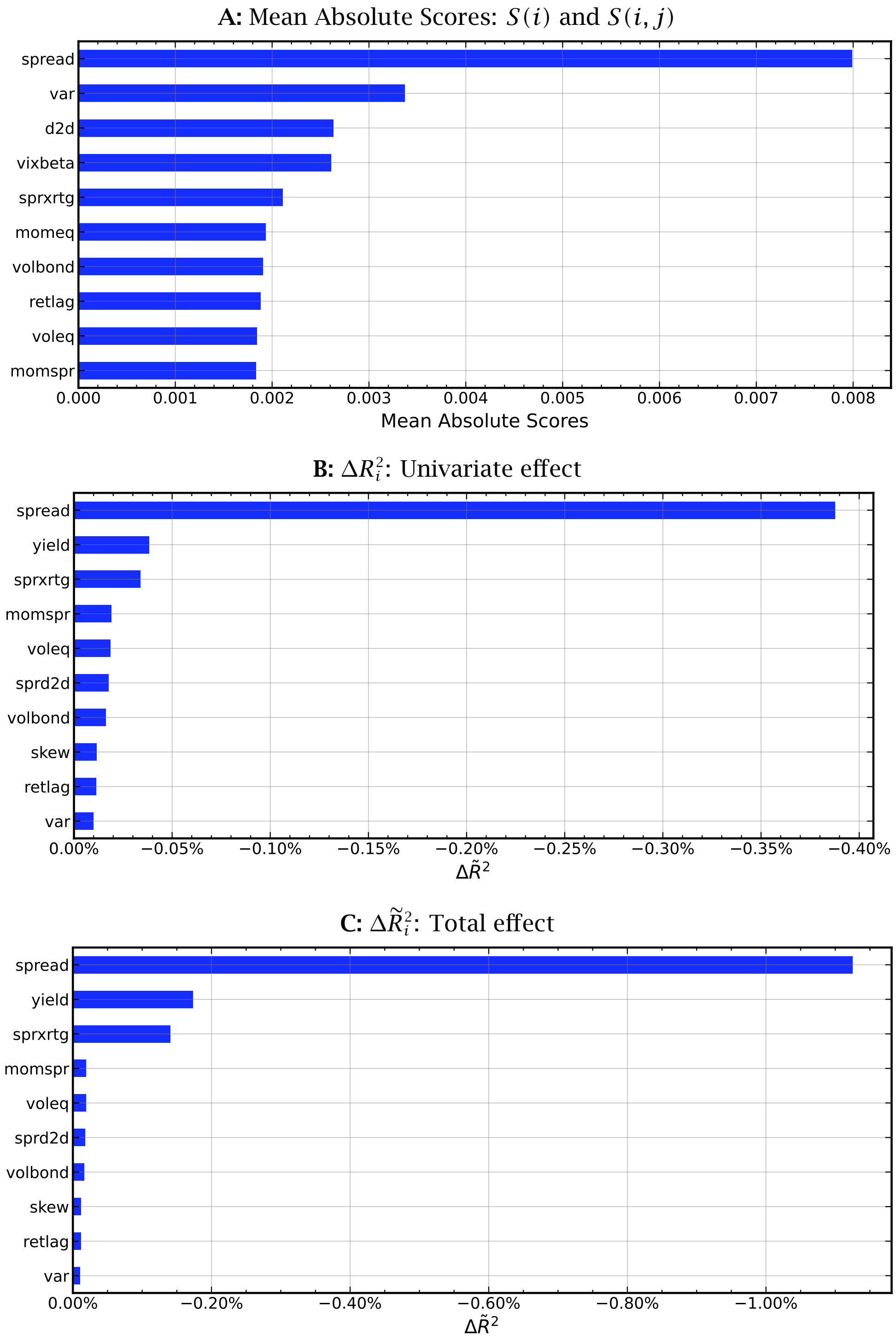
Table O-D.2: Returns of XGBoost portfolios

	1	2	3	4	5	6	7	8	9	10	10-1
					Panel A: XGB VW						
Mean	-0.19	-0.07	-0.02	0.10	0.13	0.19	0.29	0.34	0.46	0.65	0.84
S.D.	1.48	1.56	1.69	1.42	1.45	1.52	1.60	1.64	1.68	2.11	1.71
SR	-0.13	-0.05	-0.01	0.07	0.09	0.12	0.18	0.21	0.27	0.31	0.49
					Panel B: XGB-pos VW						
Mean	0.04	0.09	0.14	0.17	0.20	0.25	0.33	0.41	0.49	0.71	0.67
S.D.	0.27	0.43	0.52	0.65	0.70	0.93	1.06	1.28	1.49	1.99	1.85
SR	0.15	0.22	0.26	0.25	0.29	0.27	0.31	0.32	0.33	0.35	0.36
					Panel C: XGB EW						
Mean	-0.22	-0.06	-0.01	0.09	0.14	0.20	0.29	0.36	0.49	0.72	0.94
S.D.	1.56	1.56	1.64	1.44	1.54	1.55	1.60	1.65	1.68	2.15	1.76
SR	-0.14	-0.04	-0.00	0.06	0.09	0.13	0.18	0.22	0.29	0.33	0.53
					Panel D: XGB-pos EW						
Mean	0.02	0.09	0.13	0.16	0.21	0.23	0.33	0.43	0.50	0.75	0.72
S.D.	0.22	0.38	0.47	0.60	0.69	0.92	1.03	1.23	1.47	2.06	1.96
SR	0.11	0.24	0.28	0.26	0.31	0.25	0.32	0.35	0.34	0.36	0.37

This table compares monthly means, standard deviations, and Sharpe ratios for ten value-weighted portfolios constructed based on predicted return deciles in the testing sample. The last columns show the results of the 10-1 portfolio, which is long in portfolio 10 and short in portfolio 1. EW are equally-weighted portfolios and VW are value-weighted portfolios by bond market value. XGB-pos is the strategy that invests in the riskfree asset when the predicted portfolio return is negative.

O-D.2 Importance of characteristics

Figure O-D.1: Importance of characteristics

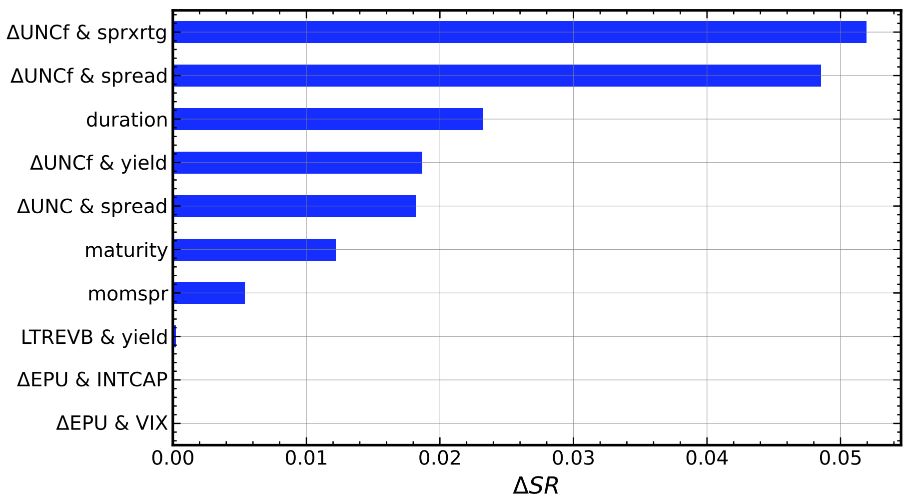


This figure shows the most important firm and bond characteristics for the EBM model. In Panel A, importance values are mean absolute scores calculated on the training data (the native EBM importance measure). In Panel B, importance values are calculated as reductions in R^2 when removing a term, i.e. an individual shape function, from the fitted model. In Panel C, importance values are calculated as reductions in R^2 when removing all terms associated with a variable from the fitted model.

O-D.3 Importances for Sharpe ratio

We analyze the importance of predictors in terms of their direct impact on the Sharpe ratio. Following a method analogous to the out-of-sample R^2 measure, we drop individual terms from the prediction equation and observe the drop in the Sharpe ratio of the resulting portfolio, with results reported in Figure O-D.2. Unsurprisingly, the most important terms consist of characteristics as well as interactions including characteristics. Notably, the highest-ranked interaction terms link uncertainty factors to bond characteristics. The Sharpe ratio analysis underlines the importance of the previously discussed interaction terms.

Figure O-D.2: Importance in terms of SR



This figure shows the most important terms as measured by their impact on the Sharpe ratio of the EBM portfolio. Importance values are calculated as reductions in the Sharpe ratio when removing a term, i.e. an individual shape function, from the fitted model. Values are scaled so that the highest-ranked item has an importance of 1.

O-E Additional material for Section 6

O-E.1 SHAP Computation

SHAP (SHapley Additive exPlanations) is a popular model-agnostic interpretability method grounded in cooperative game theory, particularly the concept of Shapley values (Shapley, 1953). In this framework, the prediction of a model for a particular input is viewed as the outcome of a cooperative game, where each feature contributes to the final prediction. The goal is to fairly distribute the model's output among the input features based on their contributions.

Formally, for a given prediction $\hat{r}_{l,t+1}$, SHAP decomposes the prediction into a sum of additive contributions:

$$\hat{r}_{l,t+1} = \phi_0 + \sum_{i=1}^{N} \phi_i^{(l,t)} + \sum_{i=1}^{N} \sum_{j=1}^{N} \phi_{ij}^{(l,t)},$$

where ϕ_0 is the expected prediction (i.e., the mean prediction over the dataset), $\phi_i^{(l,t)}$ represents the main effect (individual contribution) of feature i for instance (l,t), and $\phi_{ij}^{(l,t)}$ and $\phi_{ji}^{(l,t)}$ each capture half of the interaction effect between features i and j.

The Shapley value for a feature is computed by averaging its marginal contribution across all possible subsets of features. Let $F = \{1, \dots, N\}$ be the set of all feature indices and $S \subseteq F \setminus \{i\}$ be a subset of the feature set excluding feature i. The marginal contribution of feature i is defined as the difference in the model prediction when i is included in S versus when it is excluded. The Shapley value ϕ_i is then:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} \left[f(S \cup \{i\}) - f(S) \right],$$

where $f(S)$ is the expected model prediction when only features in S are used. This approach ensures that SHAP values satisfy several desirable properties, including local accuracy (the explanation sums to the model prediction), missingness (features missing in the model have zero contribution), and consistency (if a model changes such that the marginal contribution of a feature increases, its Shapley value does not decrease).

In addition to main effects, SHAP can also compute pairwise feature interaction effects. The SHAP interaction value ϕ_{ij} between features i and j ($i \neq j$) is defined as:

$$\phi_{ij} = \frac{1}{2} \sum_{S \subseteq F \setminus \{i,j\}} \frac{|S|!(N - |S| - 2)!}{(N-1)!} \left[f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S) \right].$$

For tree-based models such as Random Forests or XGBoost, SHAP provides specific implementations (TreeSHAP) that make use of the tree structure to compute these values more efficiently (Lundberg et al., 2020). In our analysis, we use this variant to quantify the univariate and interaction effects of features on model predictions.

O-E.2 Comparison with SHAP: Predictor importances

Recall the EBM importance measure defined in Equations 4 and 5. Analogously, we can calculate mean absolute SHAP values S_{SHAP} in the following way, once again using only the training sample $\mathcal{T}_1$ and accounting for sample weights:

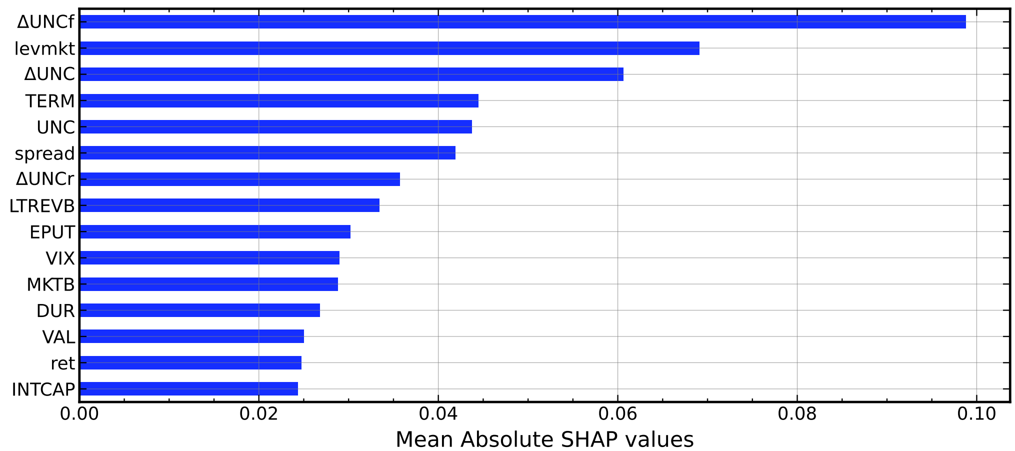
$$S_{SHAP}(i) = \sum_{(l,t)\in\mathcal{T}_1} w_{l,t} \left|\phi_i\left(x_{i,l,t}\right)\right|, \tag{O-E.1}$$

$$S_{SHAP}(i,j) = \sum_{(l,t)\in\mathcal{T}_1} 2\, w_{l,t} \left|\phi_{ij}\left(x_{i,l,t}, x_{j,l,t}\right)\right| \quad \text{for } i < j, \tag{O-E.2}$$

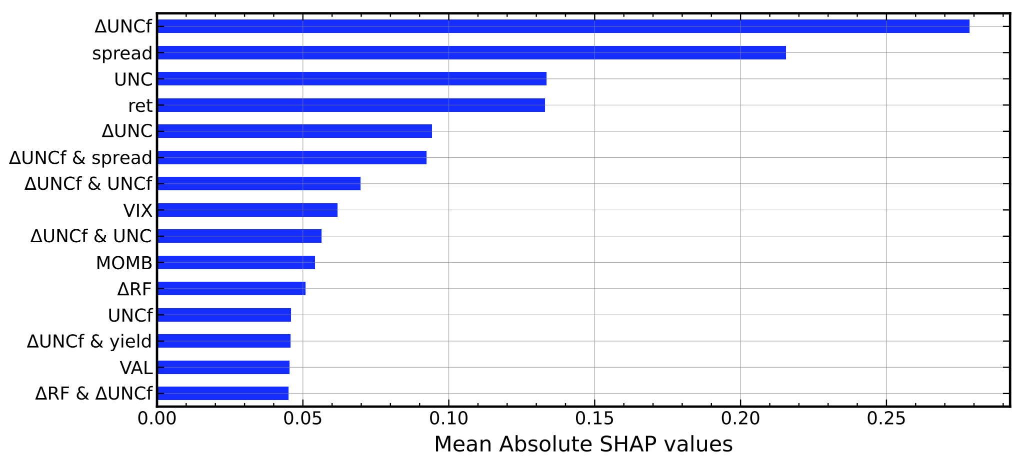
where $\phi_i(x_{i,l,t})$ and $\phi_{ij}(x_{i,l,t}, x_{j,l,t})$ represent the SHAP value for sample (l,t) that is associated with feature i and interaction (i,j), respectively. Figure O-E.1 shows the resulting importance rankings for SHAP. Compared to the EBM ranking (Figure 2 Panel A), we find overlap with respect to some of EBM's most important univariate effects, e.g. ΔUNCf, ΔUNC, TERM, and spread. There are other predictors which exclusively appear in either the Random Forest or the XGBoost ranking (i.e. there is disagreement with EBM and between the two SHAP decompositions), such as levmkt or MOMB. XGBoost and EBM agree on the most important interaction (ΔUNCf & spread), but, overall, SHAP assigns lower importance to interactions compared to the multitude of such terms in the EBM ranking. This disparity is especially pronounced for the Random Forest decomposition, where only univariate effects appear in the top 15. This highlights the ability of EBM to detect interactions and is consistent with the differences regarding the treatment of interaction effects we observe in the other comparisons.

Figure O-E.1: SHAP importance of predictors

A: Random Forest: Mean Absolute SHAP values



B: XGBoost: Mean Absolute SHAP values



This figure shows the most important predictors for the Random Forest (Panel A) and the XGBoost model (Panel B). Importance values are mean absolute SHAP values calculated on the training data (analogous to the native EBM importance measure).

O-E.3 Comparison with SHAP: Correlations

Table O-E.1: Correlations: Top 10 main effects and interactions

	EBM vs RF	EBM vs XGB	RF vs XGB
	Panel A: Main effects		
ΔUNCf	0.97	0.90	0.93
ΔUNC	1.00	1.00	1.00
TERM	0.99	0.90	0.92
LTREVB	0.98	0.62	0.68
ΔUNCr	1.00	1.00	0.99
spread	0.82	0.75	0.95
EPUT	0.91	0.88	0.79
EPU	0.68	0.97	0.61
MKT	0.98	0.69	0.64
ΔCPI	0.76	0.68	0.74
	Panel B: Interactions		
ΔUNCf & spread	0.18	0.28	0.60
ΔUNCf & INTCAP	0.30	0.09	0.41
ΔUNCf & sprxrtg	0.27	—	—
ΔUNCf & UNC	0.52	0.54	0.76
ΔUNCf & UNCr	0.24	−0.05	−0.19
ΔUNCf & EPU	0.25	−0.26	−0.06
ΔUNCf & UNCf	0.06	−0.01	0.74
ΔUNCf & VIX	−0.08	−0.25	0.33
ΔUNCf & yield	0.42	0.37	0.88
ΔUNCf & LTREVB	−0.08	−0.04	0.11

This table presents correlations between EBM outputs $f_i(x_{i,l,t})$ and SHAP decompositions $\phi_{i,l,t}$ for the top 10 EBM univariate effects (Panel A) and the top 10 EBM interactions (Panel B). Correlations are calculated across all firm-month observations. Missing values are due to constant values in the XGBoost SHAP explanations. Main effects and interactions are ordered by their EBM importance $S(i)$ and $S(i,j)$, respectively.

O-E.4 Comparison with SHAP: Portfolios

Table O-E.2: Returns of portfolios sorted on EBM/SHAP main effects

	1	2	3	4	5	6	7	8	9	10
				Panel A: EBM portfolios – sprd2d						
Mean	0.14	0.13	0.07	0.09	0.12	0.14	0.19	0.21	0.31	0.47
S.D.	1.39	1.04	1.02	1.25	1.35	1.58	1.81	2.09	2.05	2.15
SR	0.10	0.13	0.07	0.08	0.09	0.09	0.10	0.10	0.15	0.22
				Panel B: Random Forest portfolios – me						
Mean	0.46	0.37	0.22	0.20	0.20	0.14	0.11	0.11	0.10	0.09
S.D.	1.90	1.81	1.74	1.66	1.61	1.52	1.39	1.37	1.36	1.24
SR	0.24	0.21	0.13	0.12	0.13	0.09	0.08	0.08	0.07	0.07
				Panel C: XGBoost portfolios – assets						
Mean	0.24	0.11	0.28	0.26	0.27	0.18	0.23	0.09	0.09	0.08
S.D.	1.73	1.64	1.62	1.71	1.44	1.84	1.51	1.20	1.31	1.69
SR	0.14	0.07	0.17	0.15	0.19	0.10	0.15	0.08	0.07	0.05

This table reports monthly means, standard deviations, and Sharpe ratios of decile portfolios sorted on EBM outputs (Panel A) and SHAP values from Random Forest (Panel B) and XGBoost models (Panel C) for selected univariate characteristics.

Table O-E.3: Returns of portfolios sorted on EBM/SHAP interactions

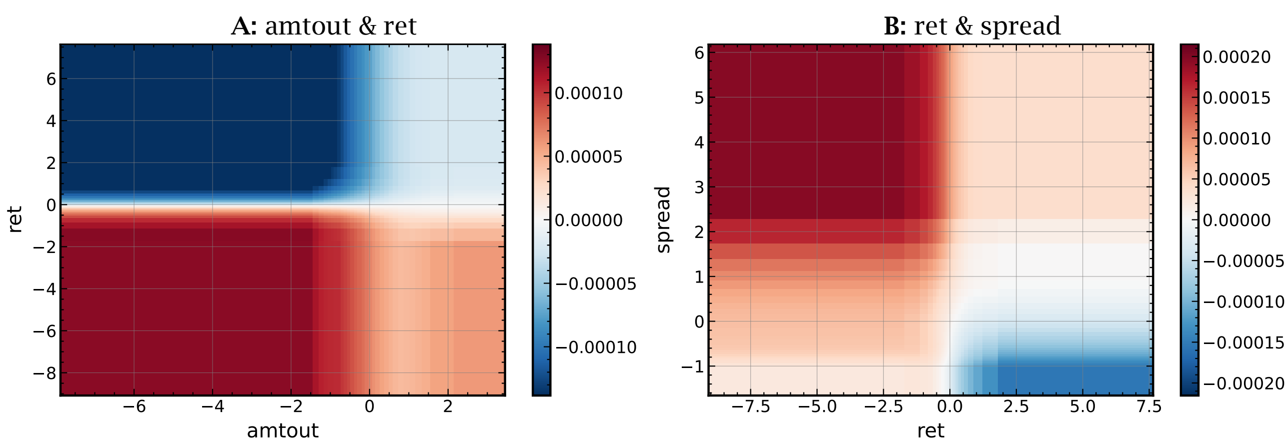
	1	2	3	4	5	6	7	8	9	10
				Panel A: EBM portfolios – profitbl & ret						
Mean	0.03	0.11	0.15	0.14	0.12	0.17	0.20	0.17	0.23	0.40
S.D.	1.45	1.36	1.31	1.45	1.50	1.46	1.62	1.68	1.67	1.74
SR	0.02	0.08	0.12	0.10	0.08	0.12	0.12	0.10	0.14	0.23
				Panel B: Random Forest portfolios – totadebt & yield						
Mean	0.14	0.17	0.15	0.18	0.11	0.07	0.15	0.18	0.21	0.39
S.D.	1.56	1.69	1.43	1.30	1.39	1.49	1.54	1.51	1.50	1.84
SR	0.09	0.10	0.11	0.14	0.08	0.05	0.10	0.12	0.14	0.21
				Panel C: XGBoost portfolios – beme & spread						
Mean	0.20	0.14	0.17	0.07	0.10	0.04	0.15	0.24	0.27	0.51
S.D.	1.91	1.56	1.30	1.51	1.24	1.44	1.30	1.53	1.73	2.13
SR	0.11	0.09	0.13	0.05	0.08	0.03	0.11	0.16	0.15	0.24

This table reports monthly means, standard deviations, and Sharpe ratios of decile portfolios sorted on EBM outputs (Panel A) and SHAP values from Random Forest (Panel B) and XGBoost models (Panel C) for selected interactions.

O-E.5 Comparison with SHAP: Firm-level Interactions

Figure O-E.2 presents EBM shape functions for the two most important firm-level interactions in terms of mean absolute scores. These functions exhibit meaningful variation along both axes, providing evidence that firm-level interactions do not merely represent linear effects in an individual characteristic. Together with the main results of section 6.2, this indicates that the EBM model successfully captures multiple economically relevant firm-level interactions that highlight the importance of nonlinear effects in the cross-section of corporate bond returns.

Figure O-E.2: Firm-level Interactions

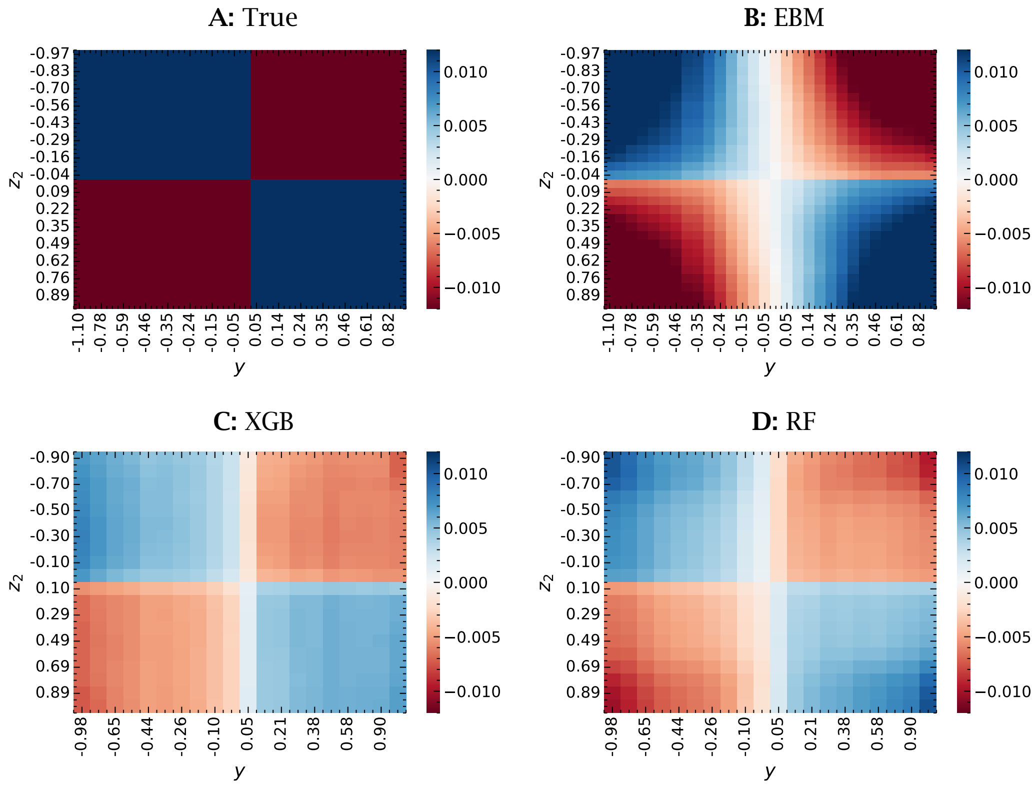


This figure shows the two most important firm-level interactions interms of mean absolute scores, amtout & ret (Panel A) and ret & spread (Panel B).

O-F Additional material for Section 7

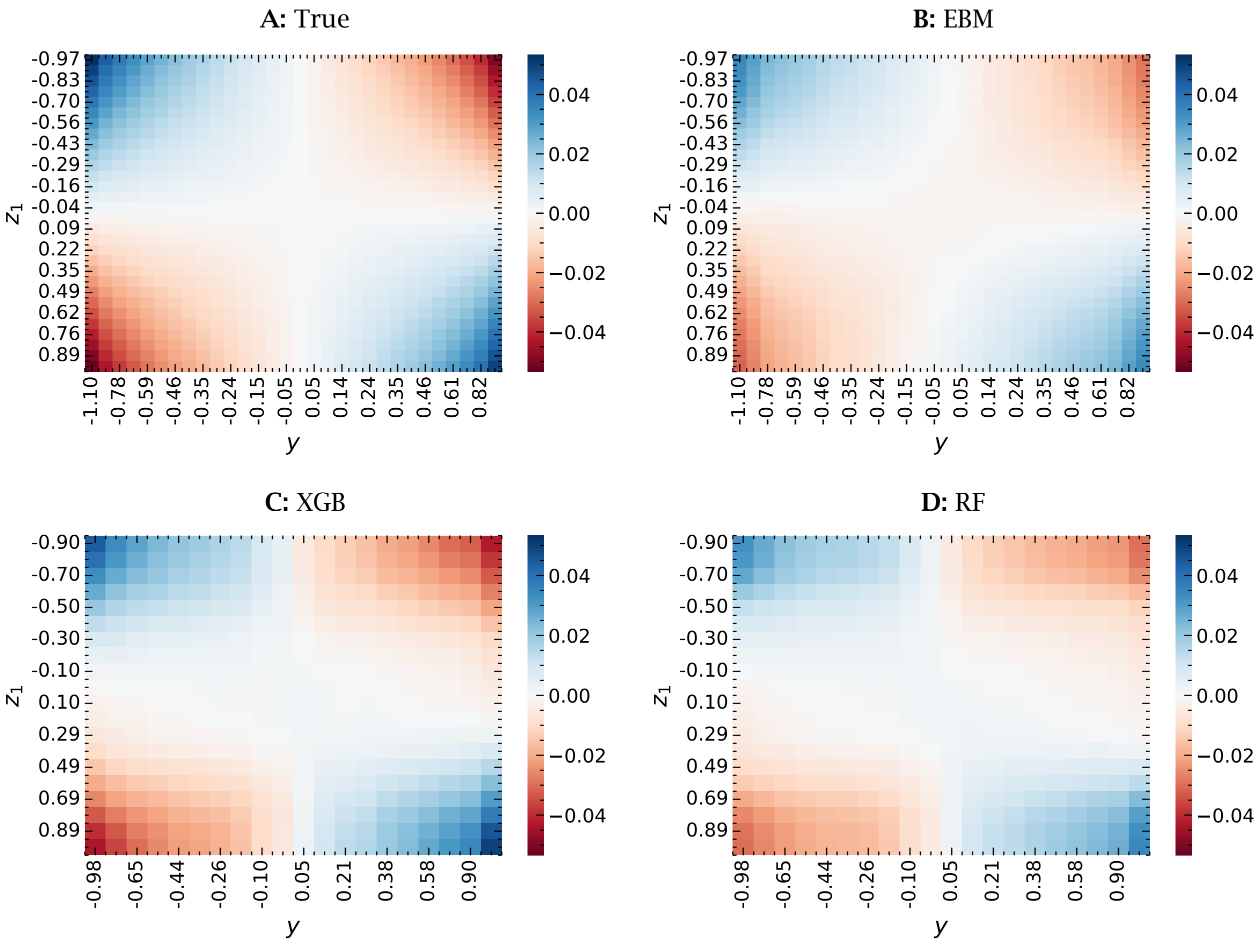
O-F.1 Shape functions

Figure O-F.1: Interaction z_2 & y



This figure displays mean shape functions over 100 simulation runs for the interaction z_2 & y. The output of the data-generating process (Panel A) is compared with the corresponding estimates from EBM (Panel B) and the corresponding SHAP values from the XGBoost (Panel C) and Random Forest models (Panel D).

Figure O-F.2: Interaction z_1 & y



This figure displays mean shape functions over 100 simulation runs for the interaction z_1 & y. The output of the data-generating process (Panel A) is compared with the corresponding estimates from EBM (Panel B) and the corresponding SHAP values from the XGBoost (Panel C) and Random Forest models (Panel D).

O-F.2 Alternative DGPs

In addition to the simulation in Section 7, we implement both the linear and the nonlinear data-generating process of Gu et al. (2020):

a) **Linear Case:** $g^\star(z_{i,t}) = (c_{i1,t},\ c_{i2,t},\ c_{i3,t} \times x_t)\,\theta_0, \quad \theta_0 = (0.02, 0.02, 0.02)'$.

b) **Nonlinear Case:** $g^\star(z_{i,t}) = (c_{i1,t}^2,\ c_{i1,t} \times c_{i2,t},\ \mathrm{sgn}(c_{i3,t} \times x_t))\,\theta_1, \quad \theta_1 = (0.04, 0.03, 0.012)'$.

These DGPs are calibrated so that the cross-sectional R^2 is 50% and the time-series R^2 is 5%.

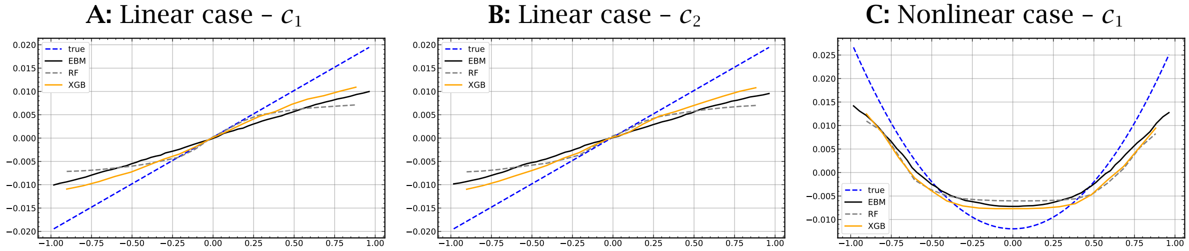
Table O-F.1: Summary: Linear case

	EBM	RF	XGB
		R^2	
Train	7.13	5.64	6.63
Valid	3.31	3.67	3.99
Test	3.49	3.40	3.58
		Term selection	
c_1	1.00	1.00	1.00
c_2	1.00	1.00	0.98
c_3 & x	0.64	0.06	0.06
		RMSE	
c_1	0.54	0.63	0.43
c_2	0.55	0.64	0.47
c_3 & x	0.46	0.62	0.64

This table reports aggregate measures for EBM and the two SHAP decompositions over 50 simulation runs. The decompositions are compared based on model fit (R^2), importance rankings, and term-level deviations from the data-generating process as measured by RMSE.

Figure O-F.3: Univariate functions (Gu et al. (2020) cases)

A: Linear case – c_1 **B:** Linear case – c_2 **C:** Nonlinear case – c_1



This figure presents univariate shape functions for the linear (Panel A and B) and the nonlinear Gu et al. (2020) case (Panel C) together with the true effects from the data-generating process. All results are averaged over 50 simulation runs.

Table O-F.2: Summary: Nonlinear case

	EBM	RF	XGB
		R^2	
Train	6.69	5.58	7.11
Valid	3.23	3.10	3.31
Test	2.96	2.82	2.72
		Term selection	
c_1	1.00	1.00	1.00
c_1 & c_2	0.86	0.76	0.78
c_3 & x	0.94	0.30	0.40
		RMSE	
c_1	1.44	1.49	1.43
c_1 & c_2	0.67	0.82	0.78
c_3 & x	0.71	1.13	1.08

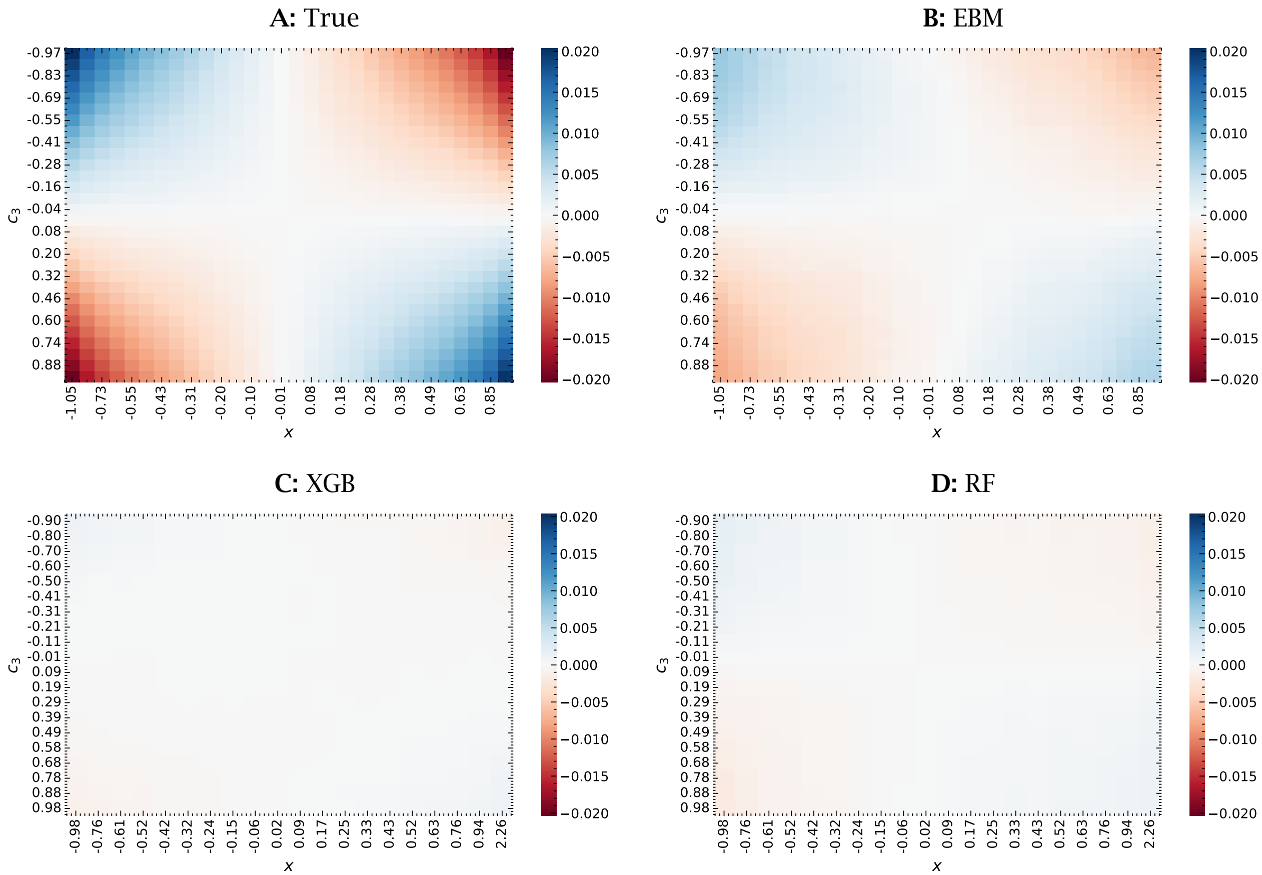
This table reports aggregate measures for EBM and the two SHAP decompositions over 50 simulation runs. The decompositions are compared based on model fit (R^2), importance rankings, and term-level deviations from the data-generating process as measured by RMSE.

Table O-F.3: Average Correlations (Gu et al. (2020) cases)

	EBM vs RF	EBM vs XGB	RF vs XGB	RF vs True	XGB vs True	EBM vs True
			Panel A: Linear case			
c_1	0.96	0.99	0.97	0.95	0.98	1.00
c_2	0.96	0.99	0.98	0.95	0.98	1.00
c_3 & x	0.80	0.66	0.64	0.65	0.61	0.79
			Panel B: Nonlinear case			
c_1	0.95	0.98	0.96	0.85	0.91	0.92
c_1 & c_2	0.74	0.79	0.87	0.80	0.83	0.91
c_3 & x	0.64	0.73	0.65	0.50	0.65	0.84

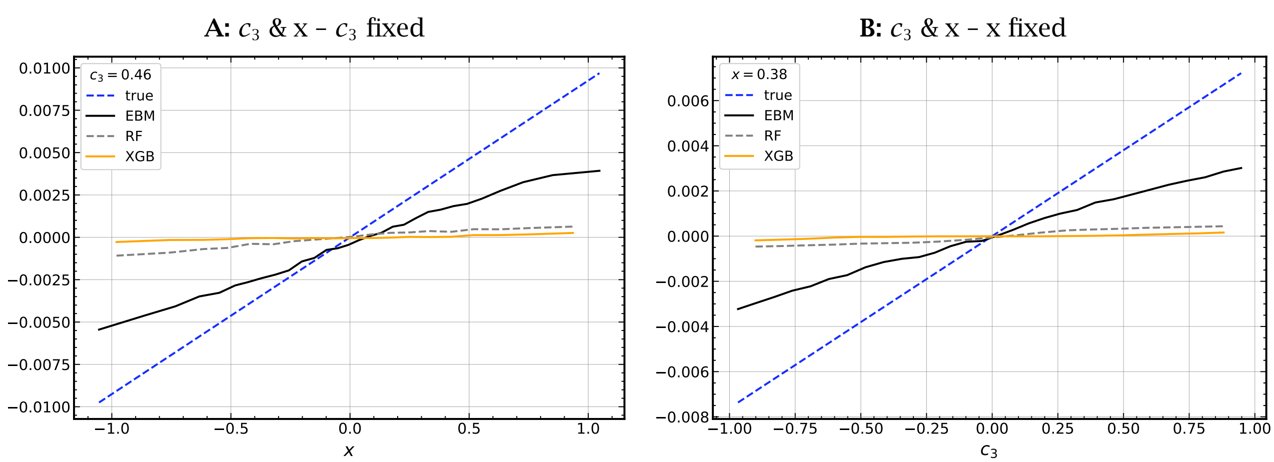
This table presents term-level correlations of EBM and SHAP decompositions with each other (columns 1-3) and with the true data-generating process (columns 4-6) for the linear (Panel A) and the nonlinear case (Panel B). Correlations are calculated across all data points within each simulation run and averaged over 50 runs.

Figure O-F.4: Interaction c_3 & x (Linear case)



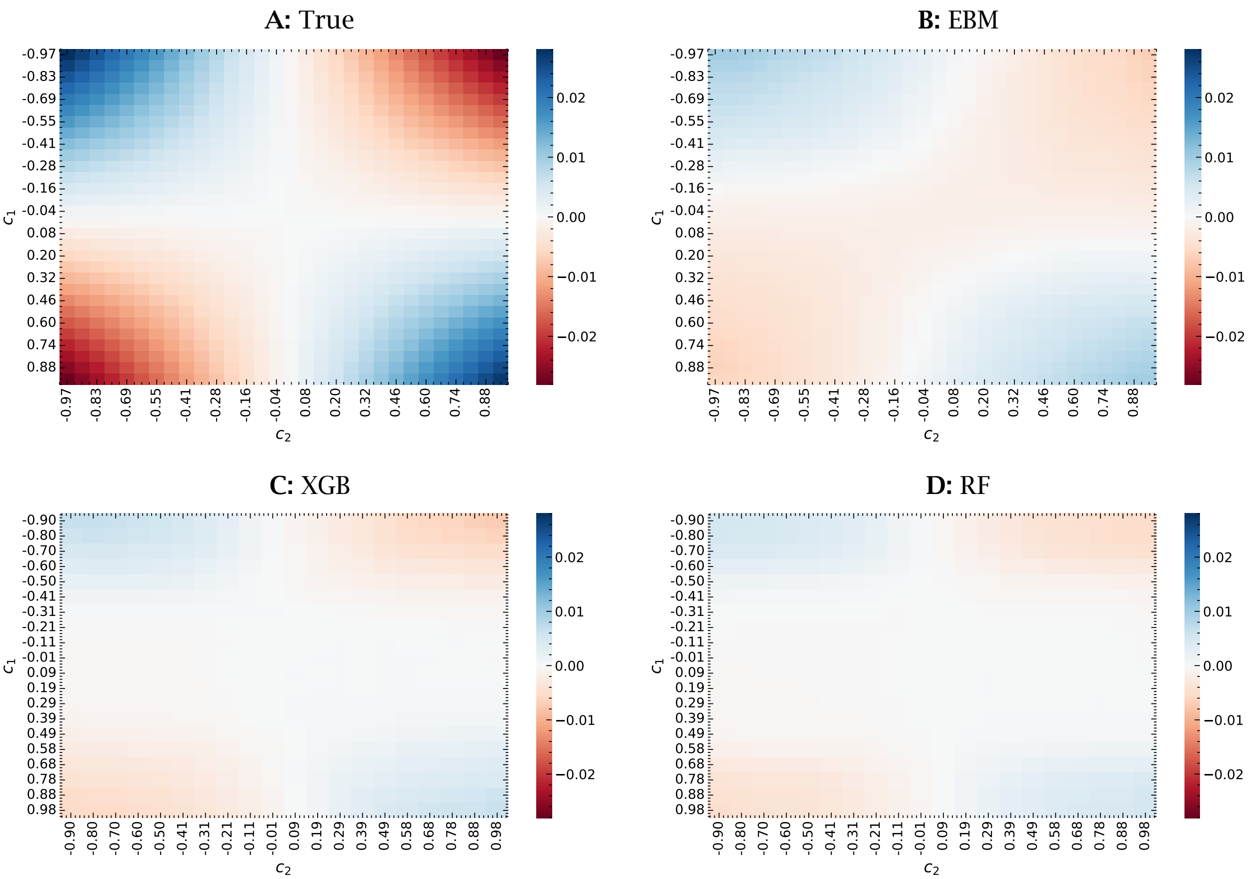
This figure displays mean shape functions over 50 simulation runs for the interaction c_3 & x in the linear Gu et al. (2020) case. The output of the data-generating process (Panel A) is compared with the corresponding estimates from EBM (Panel B) and the corresponding SHAP values from the XGBoost (Panel C) and Random Forest models (Panel D).

Figure O-F.5: Interaction cuts (Linear case)



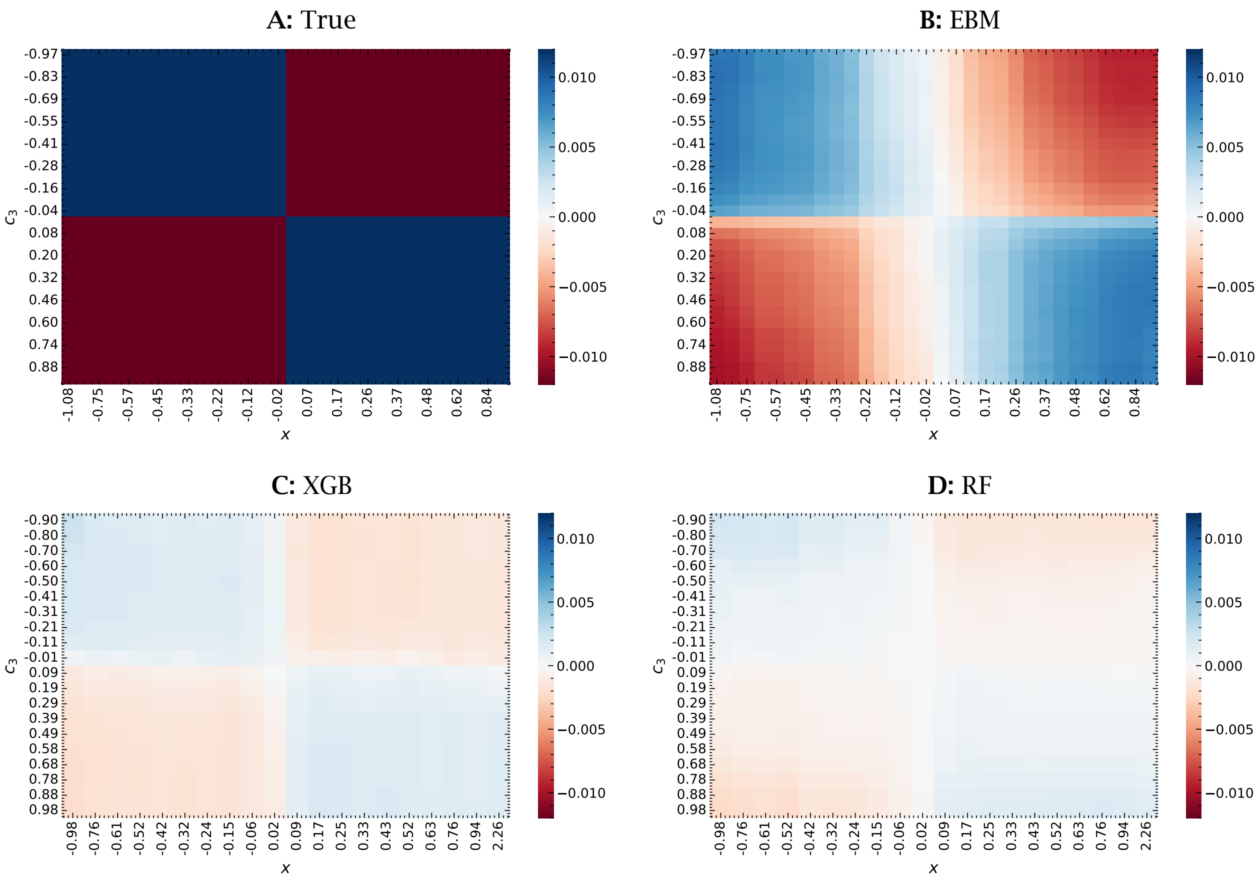
This figure presents conditional contributions derived from the mean shape functions over 50 simulations runs by separately holding each variable involved in the interaction constant at fixed levels. The three model decompositions are compared with the underlying data-generating process based on exemplary cuts through the interaction c_3 & x of the linear Gu et al. (2020) case.

Figure O-F.6: Interaction c_1 & c_2 (Nonlinear case)



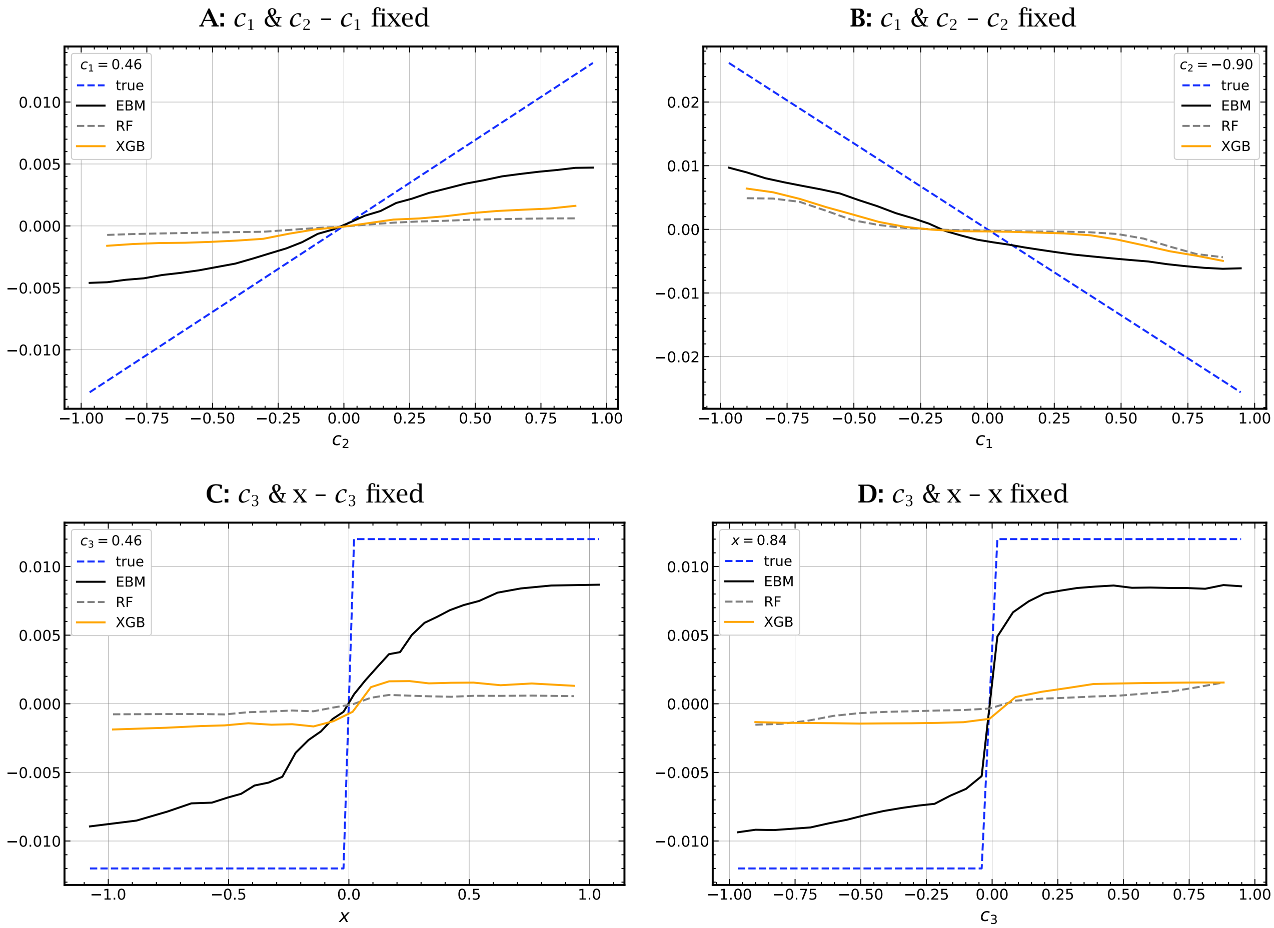
This figure displays mean shape functions over 50 simulation runs for the interaction c_1 & c_2 in the nonlinear Gu et al. (2020) case. The output of the data-generating process (Panel A) is compared with the corresponding estimates from EBM (Panel B) and the corresponding SHAP values from the XGBoost (Panel C) and Random Forest models (Panel D).

Figure O-F.7: Interaction c_3 & x (Nonlinear case)



This figure displays mean shape functions over 50 simulation runs for the interaction c_3 & x in the nonlinear Gu et al. (2020) case. The output of the data-generating process (Panel A) is compared with the corresponding estimates from EBM (Panel B) and the corresponding SHAP values from the XGBoost (Panel C) and Random Forest models (Panel D).

Figure O-F.8: Interaction cuts (Nonlinear case)



This figure presents conditional contributions derived from the mean shape functions over 50 simulations runs by separately holding each variable involved in the interaction constant at fixed levels. The three model decompositions are compared with the underlying data-generating process based on exemplary cuts through the interactions c_1 & c_2 and c_3 & x of the nonlinear Gu et al. (2020) case.