Institute

Working Paper 128-2022 | **July 2022**

Credit Factor Investing with Machine Learning techniques

Confidence must be earned

Amundi ASSET MANAGEMENT

Credit Factor Investing with Machine Learning techniques

Amina CHERIEF
Quantitative Research, Amundi Institute
amina.cherief@amundi.com

Mohamed BEN SLIMANE
Quantitative Research, Amundi Institute
mohamed.benslimane@amundi.com

Jean-Marie DUMAS
Head of Fixed Income Solutions
jean.dumas@amundi.com

Hamza FREDJ, CFA
Head of Fixed-Income Investment Lab
hamza.fredj@amundi.com

Abstract

The most common models to assess asset returns are a linear combination of risk factors. We have employed tree-based machine learning algorithms to capture nonlinearities and detect interaction effects among risk factors in the EUR and USD credit space. We have built a nonlinear credit pricing model and compared it to our baseline linear credit pricing model using error metrics on training and testing sets and during different periods. In-sample error metrics revealed the benefit of tree-based regressions. Then, we analysed the explanatory and predictive power measure by factor category and by period in order to evaluate the contribution of each factor in the explanation and prediction of credit excess returns. We found value in adding alternative factors to a traditional factor model and point out which of them prevail across different time horizons and during market crisis periods. Finally, tree-based regressions methods assisted us in improving our understanding of prices through the interaction between features and between each feature and the output of the model.

Keywords: Credit, Value, Size, Momentum, Duration, Liquidity, ESG, Machine learning models, Tree-based regressions, SHAP values, Feature importance.

JEL classification: G12, C41, C51, C52

Acknowledgement

The authors are grateful to Takaya Sekine, Lauren Stagnol, Thierry Roncalli and Edmond Lezmi from Amundi Institute and Jeremy Jouvance and Florian Vincent from Amundi Technology for their helpful comments.

About the authors

Amina CHERIEF

Amina Cherief joined the Quantitative Research team of Amundi in December 2020. Prior to that, Amina was a Cross-Asset Financial Engineer at SG CIB in New York (2019-2020), a Financial Engineer at Natixis AM (2017-2019) and an intern in Amundi's Quantitative Research dedicated to development of a multi-factor risk and performance analysis tool (2017). Her areas of research cover factor investing, sustainable investing and AI-ML in both fixed-income and equities. Amina holds a Master Degree (with honors) in Risk and Asset Management from Paris-Saclay University.

Mohamed BEN SLIMANE

Mohamed Ben Slimane is Head of Fixed-Income Quantitative Research in the Amundi Quantitative Research team. Prior to his current position, he was Senior Quantitative Analyst (2015-2020), Head of Credit and Counterparty Risks (2013-2015) and Head of Regulatory Risk for Fixed-Income and Structured Funds (2010-2013). Before joining Amundi, he was Head of Regulatory Risk at Sociéte Générale AM (2005-2010), an IT Consultant with Altran (2002-2005) and a Software Engineer at Sungard Systems (2001-2002). He started his working career with Cap Gemini in 1999 as a Software Engineer.

Mohamed holds an Engineering Degree from the École Nationale de l'Aviation Civile (1999). He is Certified International Investment Analyst CIIA (2010) and FRM Charterholder (2013).

Jean-Marie DUMAS

Jean-Maris Dumas is Head of Fixed Income Solutions since June 2013, a 11-people investment team. As such he covers hybrid debt management, income-oriented credit funds, total return funds and ALM solutions. He is also in charge of structuring and promoting innovative funds (private debt related solutions, green finance investments). Previously, he held various positions within Crédit Agricole Asset Management, from Fixed Income and Balanced quantitative analyst in the structured products department (2002-2005) to Credit & High Yield portfolio manager (2005-2007), Head of Managed Structured Products (2007-2009) and Head of Core+ Aggregate Bond Management (2010-2013) at Amundi. Jean-Marie Dumas holds a Master of Science (MSc) in Statistics from the ENSAI (Rennes) with a specialization in Financial Risk Management.

Hamza FREDJ

Hamza is Head of the Fixed-Income Investment Lab since March 2022. Prior to that, he worked as a Portfolio Manager within the Fixed Income Solutions team of Amundi since 2010, first as a structured funds PM, then as a Credit PM. He began his career in 2006 as a Senior Front Office Consultant at Murex.

Hamza holds an engineer's degree from Ecole Supérieure d'Electricité and a master's degree in Financial Markets from Dauphine University. He is also a CFA and a Financial Data Professional (FDP) Charterholder.

1 Introduction

Explaining portfolio returns through *Fama-French-Carhart* models supposes linearity between returns and risk factors. It makes it possible to simply identify relationships as a linear combination of factors, while in reality, these relationships could be much more complex. In support of this argument, Figures 18 and 19 of Appendix A.1 show quantile regression coefficients for each decile of our credit risk factors to the credit excess returns of the market for EUR and USD-denominated bonds over the 2011-2021 period. A first observation stands out: each factor curve corroborates both figures in terms of factor exposures to the credit excess returns and displays for most of them a bell-shaped curve.

These models may be refined to address more specific problems. In this sense, studies by Bonne *et al.* (2021), Dixon and Polson (2020), Gu *et al.* (2020), and Simonian *et al.* (2019) demonstrate that tree-based regressions and deep neural networks are particularly attractive in capturing nonlinearities and detecting interaction effects among variables in the context of financial asset returns thanks to their ability to fit flexible functional forms and deal with a large amount of structured and unstructured data. We sought in our study to address the best known tree-based regressions, namely Random Forest (Brieman, 2001) and Gradient Boosting regressions (Friedman, 2001) to identify these nonlinearities in the credit space. Accordingly, we use the investment grade corporate bond universe to explain the credit excess returns of the market with our daily traditional and alternative factors returns specified in Ben Slimane *et al.* (2018). To be in line with our previous research, we added ESG returns in our model (Ben Slimane *et al.*, 2019), therefore becoming a 7-Factor model.

Following the popularity of machine learning algorithms over the past decades, there is a need to understand and interpret models for the sake of identifying relationships. We therefore took an interest in using a method to interpret machine learning models called TreeSHAP (Lundberg *et al.*, 2019) that explains not only the relationships but the dependence and interaction effects between variables.

We conducted our study as follows: in the first part, we introduced the risk factors that we use to explain the EUR and USD market excess returns analysed following Ben Slimane *et al.* (2018). Next, in a second section, we defined the tree-based algorithms adopted in our research. In the third part, we built the nonlinear credit pricing model and compared it to the linear credit pricing model using error metrics on training and testing sets and different time periods to assess the method's robustness. In-sample error metrics reveal the power-fullness of tree-based regressions. Then, we sought to analyse the explanatory and predictive power measures by factors category and by period to evaluate the contribution of each of them in explaining and predicting market excess returns. We, therefore, found value in adding alternative factors to a traditional factorial model. Finally, in the last section, we attempted to apply the TreeSHAP method to evaluate which factors prevail across different time horizons and during market crisis periods.

2 Factor models in investment grade corporate bond investing

This section aims to briefly define the risk factors specified in Ben Slimane *et al.* (2018) with the ultimate goal of determining the sensitivity of the credit market excess returns to these variables. More precisely, we turn on defining traditional factors for the corporate bond universe that refers to the measure of the DTS (duration time spread) risk, duration risk, and liquidity risk. Then, we focus on the alternative risk factors in the corporate universe illustrated by the momentum, the value, and the size risk factors. In the last section, we aim to define the Environmental, Social, and Governance risk that have outperformed the market since 2014 in the EUR Investment Grade corporate bond universe (Ben Slimane *et al.*, 2019).

2.1 Traditional risk factors

We determine the risk factors affecting credit excess returns bond's through either their own specific characteristics or across the bond market's behavior and structural conditions. For this purpose, we have selected three components frequently quoted by academics in a traditional bond risk factor bucket. We face a paradox because there are various ways to measure these risks, as some of them are estimated over more than one metric. For instance, we could have chosen a set of factors per currency, per interest rate curves, or credit rating. In our analysis, we propose a single factor per premium, to give an overview of how the risks have been priced over time.

The first source of corporate bond risk is assigned to the capacity of a borrower to honor its debt engagement. This premium has been initiated by Fama and French (1993) through the default risk factor (also called DEF) and is defined as the spread of corporate bond yields over sovereign bonds yield. For our study, we measure this factor with the duration-time-spread (DTS) corresponding to the estimation of the sensitivity of the price of a bond to spread movement on relative terms (Ben Dor *et al.*, 2007). The second factor is the duration, which finds its roots in the consequences of interest rate fluctuations in the macro-economic environment. When interest rates rise, then yields on outstanding bonds fall compared to newly issued bonds resulting in the payment of higher coupons. The duration risk premium is represented by the modified duration (MD) that evaluates the bond price sensitivity to a yield movement on absolute terms. Finally, the last premium refers to the ease of trading or the liquidity premium that has been extensively studied during the 2000s. Several researchers explain this dimension as the main source of risk in the corporate market (Huang and Huang, 2012) or as a specific systematic risk factor for others across the cross section of returns (Chen *et al.*, 2007; Bao *et al.*, 2011; De Jong and Driessen, 2012). Recently, Roncalli *et al.* (2021) studied the liquidity of a bond from the point of view of the costs implied in trading corporate bonds and sovereign bonds. Thus, the cost is a function of the bid-ask spread and the market impact estimated, among others, over the duration-time-spread and the outstanding based participation rate for the corporate market. For the needs of this study, the illiquidity premium (LTP) is approximated by the bid-ask spread itself computed over the liquidity score of Ben Slimane and De Jong (2017)[1].

[1] The liquidity score is built according to the Bloomberg Barclays Multiverse Bond Index and the bid-ask spread are coming from Bloomberg.

Thus, the total return of a corporate bond can be explained by the following equation:

$$R_i(t) = \alpha(t) - MD_i(t) \times R^I(t) - DTS_i(t) \times R^S(t) + LTP_i(t) \times R^L(t) + \epsilon_i(t) \quad (1)$$

where $R_i(t)$ is defined as the credit excess return of a bond i at t and $\alpha(t)$ is a constant. $R^I(t), R^S(t)$ and $R^L(t)$ are the return components for respectively the interest rate movements, credit spread variation and change in liquidity[2].

2.2 Diversifying with the Alternative risk factors in the corporate bond universe

Houweling and Van Zundert (2014) introduce alternative credit risk factors with the objective of building an insight into the equity pricing model of Fama-French-Carhart using corporate bond data. The authors have replicated the low risk, momentum, and value risk premia in the corporate bond market space. Similarly, shortly afterward, Bektic *et al.* (2019) created four risk premia for corporate bonds based on the scores of equity stocks (size, value, profitability, and investment). Israel (2018) introduced the idea that risk factors for corporate bonds (low risk, momentum, and value) can be fitted using equity and bonds data. For our study, we focus on the momentum, size, and value factors that are listed by Ben Slimane *et al.* (2018). The data are given by selecting the bonds denominated in EUR from the Intercontinental Exchange Bank of America Merrill Lynch (ICE BofAML)[3] Large Cap (Investment Grade) on a daily basis from January 2003 to December 2021. These alternative risk factors are defined as follows:

Table 1: Alternative risk factors

Factor	Definition
Momentum	Six-month trailing bond returns
Size	Total Debt value of the issuing firm
Value	Excess log spread over peers

Source: Amundi Institute Quantitative Research

[2]For more information, the details of this formula are available in Ben Slimane *et al.* (2018).

[3]Source ICE Data Indices, LLC ("ICE DATA"), is used with permission. ICE DATA, its affiliates and their respective third-party suppliers disclaim any and all warranties and representations, express and/or implied, including any warranties of merchantability or fitness for a particular purpose or use, including the indices, index data and any data included in, related to, or derived therefrom. Neither ICE DATA, its affiliates nor their respective third-party suppliers shall be subject to any damages or liability with respect to the adequacy, accuracy, timeliness or completeness of the indices or the index data or any component thereof, and the indices and index data and all components thereof are provided on an "as is" basis and your use it your own risk. ICE DATA, its affiliates and their respective third-party suppliers do not sponsor, endorse, or recommend AMUNDI, or any of its products or services.

2.3 Adding ESG to the risk factors

Ben Slimane *et al.* (2018) concluded their analysis by questioning the relevance of ESG as a risk premium in the corporate bond market. The topic was intensively discussed for the stock market (Bennani *et al.*, 2018,b; Drei *et al.*, 2019; Lepetit *et al.*, 2021) and has been then studied in the corporate bond market in Ben Slimane *et al.* (2019). An important point arising from this paper is that the behaviour of ESG in the EUR denominated Investment Grade universe is in line with the conclusion of Bennani *et al.* (2018) in the European equity market. Between 2010 and 2013, the alpha best-in-class versus worst-in-class is negative and turned positive after 2014. The second important point emerging from this paper is that ESG is increasingly correlated with credit risk, which is undeniable since rating agencies use extra-financial signals when building their default models. ESG is vitical from our point of view as we hypothesize that it is a risk premium along with the risk factors previously mentioned to determine the sensitivity of these variables to the market. In the case of a linear regression model the collinearity between the predictors is an issue.

Ben Slimane *et al.* (2020) extend the previous study to the period before and during the COVID-19. This study confirms the strong sensitivity of the ESG to the credit excess returns of the market from January 2019 until February 2020. The ESG variable is ranked among the main factors explaining the market even though its position has gradually declined during the COVID-19 crisis but this fact does not call into question the importance of integrating ESG into the factor investment universe. On the contrary, ESG acted as a hedge in credit portfolios. In the next chapter, we describe the new methodology used to assess and evaluate the exposition of the above-listed risk premia to the credit excess returns of the market. In the next section, we focus on the definition of decision trees that are by nature immune to the embedment of collinearity between the predictors.

3 A bird's view of tree-based regressions

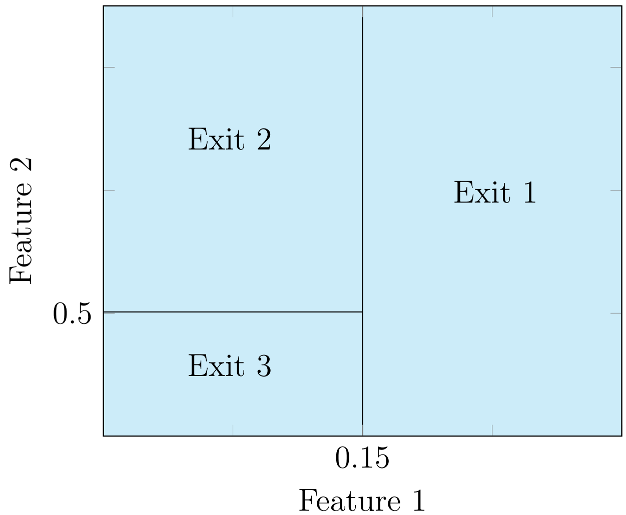
Tree-based methods are powerful for predictive modeling, given their nature to fit simple models in a set of rectangles partitioning the predictors (see Figure 1). Brieman *et al.* (1984) introduced the CART model as the acronym of the book title *"Classification and Regression Trees"* for regression or classification forecasting problems. The CART algorithm lays out the premise for an algorithm such as Random Forest, Gradient Boost, or AdaBoost.

As mentioned in Section 1, we intend to use these methods to rank the importance of our risk factors on the credit excess returns. These algorithms are interesting given their nature with regards to the non-linear effects between features. In this section, we focus on the definition of these tree models, and more particularly on regression trees.

3.1 Decision trees regression

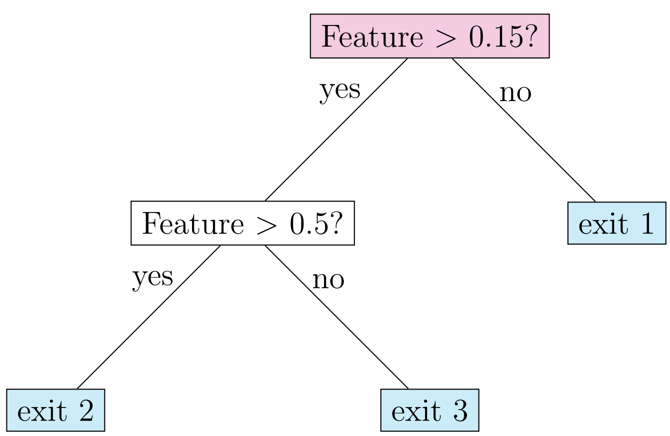
As the name suggests, a decision tree is a tree composed of nodes to which tests are associated and two branches linking these nodes. These branches make the decision to answer *"yes"* or *"no"* to the *"backward"* test. A decision tree is built according to a top-down approach. Each tree is at least composed of a *"root node"* relying upon the top node of the tree and *"leaf nodes"* corresponding to the back end of the tree. In general, the tree has intermediate nodes between

Figure 1: Two-dimensional feature space partition obtained with recursive binary splitting



the root and the leaf called *"internal nodes"*. Figure 2 shows the typical decision tree flow for a regression. In this example, to simplify the problem, we consider only one feature as input data that can be either numerical or categorical. The root node corresponds to the box integrating the first question *"Is Feature above 0.15?"*, then a first exit or leaf node is observed (to the right) by answering *"no"* to the question because there is no more question to ask or enough points to separate the data. On the left (the answer *"yes"*), the tree establishes a second question *"Is the Feature above 0.5?"*. This node is the intermediate node. Finally, the answers *"exit 2"* and *"exit 3"* are considered as the end of the tree (the last leaf nodes) as we cannot split these two nodes in smaller groups. The leaf node *"exit$_i$"* contains the average output observations falling in that node.

Figure 2: Decision Tree



The question that also arises from Figure 2 is how to select the numbers (0.5 or 0.15) in the nodes. For regression purposes, the selection generally refers to the mean squared errors or the reduction in variance method. In this sense, we calculate the sum of squared residuals in each region for different thresholds and we select the threshold having the smallest sum of squared errors using the training set. Throughout the paper, we split our dataset into a training and a testing set in order to evaluate the performance of our models and thus, improving the accuracy of the training set and controlling over fitting. The models are trained with the training set and we use several way of splitting the data set (see Figure 3 for the cross-validation methods used in this paper). Therefore, the reduction in variance method is described as follows:

Algorithm 1 Reduction in variance method (for one-dimensional input data x)

1: **Input**: Training data $\{(x_i, y_i)\}_{i=1}^{N}$
2: Sort x values from the lowest to the highest values
3: **for** $t \leftarrow 0$ to $N-1$ **do**
4: Estimate the threshold as the average between two data:
$$\tau = \frac{x_t + x_{t+1}}{2}$$
5: Calculate the sum of squared residuals of x_i values being above or equal to the threshold and their associated y_i:
$$R_{\{(x_i,y_i)|x_i \geq \tau\}} = \sum_{\{(x_i,y_i)|x_i \geq \tau\}} \|\mathbf{y_i} - \mu_{(\mathbf{x_i},\mathbf{y_i})|\mathbf{x_i} \geq \tau}\|^2$$
6: Calculate the sum of squared residuals of x_i values being below the threshold with their associated y_i:
$$R_{\{(x_i,y_i)|x_i < \tau\}} = \sum_{\{(x_i,y_i)|x_i < \tau\}} \|\mathbf{y_i} - \mu_{(\mathbf{x_i},\mathbf{y_i})|\mathbf{x_i} < \tau}\|^2$$
7: The residual R^t is computed as the sum of the sum of squared residuals of each branch:
$$R^t = R_{\{(x_i,y_i)|x_i \geq \tau\}} + R_{\{(x_i,y_i)|x_i < \tau\}}$$
8: **end for**
9: The candidate having the lowest sum of squared residuals is then selected:
Output: $\tau_F = \operatorname{argmin}_t R$

where y_i represents the target variable of the observation i and N is the corresponding number of observations i in each branch. $\mu_{(x_i,y_i)}$ corresponds to the average of the target variable y in each region and τ_F represents the final selected threshold.

However, in practice, the input data might have more than one predictor increasing the number of candidates for the root and intermediate nodes split. The feature selection in each branch is thus elaborated according to the same scheme. We perform Algorithm 1 for each feature x, and we select the feature having the lowest residuals given the optimal value τ_F of the threshold. Finally, the tree stops growing when it does not improve the sum of squared residuals or usually called the entropy.

Decision trees have a disadvantage that causes the irrelevance of this model when implemented and results in over-fitting of the training data set. In other words, they are inaccurate when trying to predict the dependent variable using another subset of data. There are several ways to decrease this effect, such as pruning trees which replaces useless split or redundant subtree with a leaf node. Another method introduces features randomness set at each branch, while the last method formulates a decision based on weak learners. In the following sections, we present some well-known tree algorithms applying these methods to reduce over-fitting. According to Zhou (2009), the following programs are part of ensemble learning that combines the property of basic models to build a forecasting framework. These techniques find their root cause with the *"Wisdom of crowds theory"*, which asserts that aggregating information from a group of individuals to state a decision leads to better results than doing it from any member of the group separately (Galton, 1907). In a statistical framework, the above assertion means that instead of training one model to solve a regression problem giving a good performance, it is often more relevant to generate an ensemble of classifiers having less quality good results and weight the prediction group.

3.2 Random Forest regression

Brieman (2001) defines Random Forest as an aggregation of tree predictors where each tree is built independently from each other with randomly selected data and such that all the trees in the forest have the same distribution. These tree-based regressions combine Bootstrapping-Aggregation (BaGGing), and feature randomness for several trees to prevent correlation issues among decision trees and reduce the variance bias. The author highlights the robustness of this model against outliers or noise and the higher accuracy. Random Forests are detailed in the following algorithm:

Algorithm 2 Random Forest for regression

1: **Input**: Training data $\{(x_i, y_i)\}_{i=1}^{N}$
2: **for** $k \leftarrow 1$ to K **do**
3: *Bootstrapping*: Create a dataset B_k of size N^* by randomly selecting data from the entire sample. A data can be selected more than once.
4: Develop the decision tree k with B_k considering a random subset of j features at each node.
5: Follow Algorithm 1 to choose the feature to use among the j features at each node.
6: **end for**
7: *Aggregating*: Average the prediction of each decision tree as follows:
$$\hat{y}_t = \frac{1}{K}\sum_{k=1}^{K} f_k(B^j)$$
8: **Output**: $\hat{y}_t$

where K represents the number of trees. The accuracy of the Random Forest is computed over the Out-Of-Bag Estimator. The Out-Of-Bag prediction of an input x_i is measured as the average prediction of the decision tree B_k when B_k does not include xi. We estimate the Out-Of-Bag-error by comparing the output of the predictor versus the estimated dependent variable.

3.3 Gradient Boosting regression

This type of regression relies on another dimension of regression trees known as *"Tree Boost"*. The objective is to sequentially aggregate models of a training dataset and progressively correct the weights. This method allows the creation of new trees by fitting the residuals also called the *"pseudo responses"* to the x independent variables using the gradient descent method until a new tree does not bring any added value. Gradient Boosting is interesting because it focuses on the amplitudes of the errors. The new tree is built based on the error made by the previous trees where the first predicted values is defined as the average number of the x values for the first tree. The algorithm is described by Friedman (2001) and is defined as follows:

Algorithm 3 Gradient Boosting for regression

Input: Training data $\{(x_i, y_i)\}_{i=1}^{N}$
Initialize $F_0(x)$ as follows:
$$F_0(x) = \operatorname{argmin}_\theta \sum_{i=1}^{N} \mathcal{L}(y_i, \theta)$$
for $k \leftarrow 1$ to K **do**
    Compute the residuals r by solving the gradient for each observations i of the tree k:
$$r_{i,k} = -\left[\frac{\partial \mathcal{L}(y_i, F_{k-1}(x_i))}{\partial F_{k-1}(x_i)}\right] \text{ for } i \in (1, \ldots, N)$$
    Fit a regression tree to r_k values using Algorithm 1 for the variable selection.
    Create leaves $R_{l,k}$ for each leaf $l \in (1, \ldots, L_k)$ in the tree k.
    Compute the predicted values $\theta_{l,k}$:
$$\theta_{l,k} = \operatorname{argmin}_\theta \sum_{x_i \in R_{i,l}} \mathcal{L}(y_i, F_{k-1}(x_i) + \theta)$$
    Update $F_k(x)$[a]
end for
Output: $F_K(x)$

[a]where $F_k(x) = F_{k-1}(x) + \psi \sum_{l=1}^{Lk} \widehat{\theta}_{l,k} I(x \in R_{l,k})$

where N is the number of observations in the training set, K is the total number of trees and ψ is defined as the learning rate.[4] $\mathcal{L}(y_i, \theta)$ is the differential loss function[5] where y are the observed values and θ are the predicted values. Trees algorithms may be complex given their number of parameters to adjust. We defined in Appendix A.2.1 on page 42 our selection for the hyper-parameters of Random Forest and Gradient Boosting algorithms. In machine learning, hyper-parameters are *"tuning parameters"* that allows to control the learning process (Probst *et al.*, 2019).

3.4 Catching non-linearities in a linear model

In the context of ensemble learning, we identify a major interest to combine regression models. For instance, we can forecast the predictor with a linear model and then predict the linear regression error with a non-linear regression. Bonne *et al.* (2021), for example, finds similarities between stock returns and factor exposure with non-linear models for the component that was not identified by the linear factor model. Zhang *et al.* (2019) capture linearity in their models across a lasso regression and attends to predict the model error with the Random Forest regression. According to the authors, this method aims to predict the dependent variable better than Random Forest regressions themselves. Tree-based regression may suffer from overfitting. We believe that mixing linear and non-linear methods will add stability and robustness in the model. Thus, the algorithm used in the study is the Enhanced Random Forest regression and it is described as follows:

[4]The learning rate is between 0 and 1. This value is important to avoid the variance errors, it allows to reduce the prediction of each tree and finally improves the accuracy when the value is small.

[5]Gradient Boosting is designed to cover a large set of loss functions, each adapted to a particular problem. For our study, we use the least-square loss function.

Algorithm 4 Enhanced Random Forest regression

1: **Input**: Training data $\{(x_i, y_i)\}_{i=1}^N$
2: Fit a Lasso regression and compute the model error as follows:

$$\hat{\beta}(\lambda) = \text{argmin}_\beta \, ||\mathbb{Y} - X\beta||^2 + \lambda||\beta||_1$$

$$\hat{\mathbb{Y}}_{lasso} = \alpha + \sum_{j=1}^{J} X_j \hat{\beta}_j(\lambda)$$

$$e_{lasso} = \mathbb{Y} - \hat{\mathbb{Y}}_{lasso}$$

3: **for** $k \leftarrow 1$ to K **do**
4: *Bootstrapping*: Create a dataset B_k of size N^* by randomly selecting data from the entire sample considering the dependent variable as the error of the lasso model. A data can be selected more than once.
5: Develop the decision tree k with B_k considering a random subset of j features at each node.
6: Follow Algorithm 1 to choose the feature to use among the j features at each node.
7: **end for**
8: *Aggregating*: Average the prediction of each decision tree as follows:

$$\hat{e_{rf}} = \frac{1}{K} \sum_{k=1}^{K} f_k(B^j)$$

9: **Output**: $\hat{\mathbb{Y}}_{lasso} + \hat{e}_{rf}$

This regression may seem complex given the number of parameters to optimize. Indeed, we seek to calibrate the hyper-parameters of the Random Forest described in Appendix A.2.1 plus the λ penalty parameter of the lasso regression. Within this context, we create an object combining these parameters and build a recursive generator that runs regressions using the training data set for each combination. Finally, we select the model minimizing the mean squared error on the test set. In the next section, we compare the performance of these models against the ordinary least square (OLS) regression. The objective is to improve the explanatory power of the linear regression with the non-linear effects.

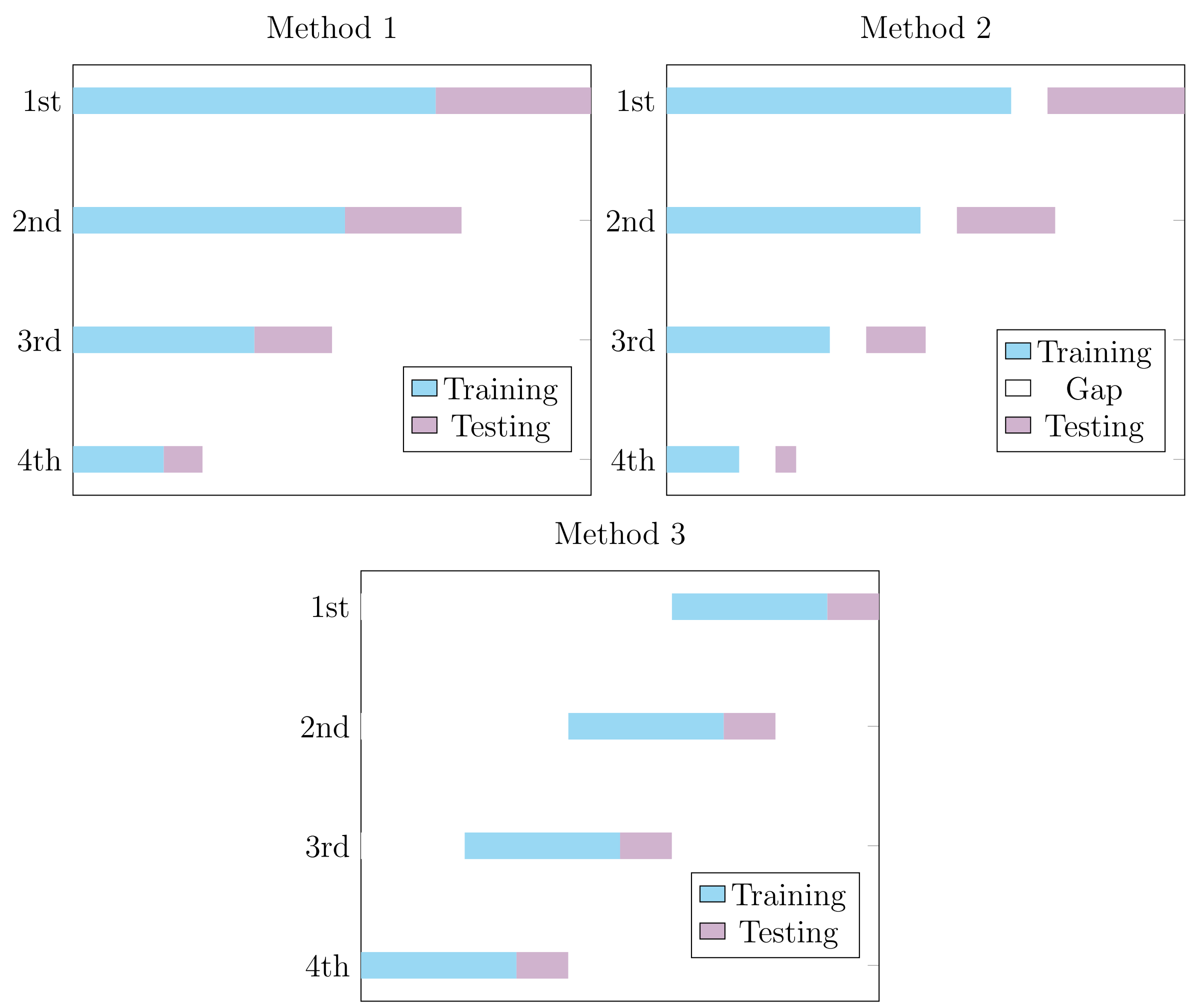
4 Building the ensemble factor models

In this section, we build the non-linear credit factor model given the data described in Section 2. As a first step, we aim to evaluate the performance of the non-linear models versus the linear one and thus, we confirm the added-value of tree-based regressions. Consequently, we compare the errors and adjusted R-squared (R^2_{Adj}) of the OLS, Random Forest, Gradient Boosting, and the Enhanced Random Forest regressions described in Section 3.4.

4.1 Comparing the performance of the non-linear vs. linear models

We analyze the error of the models based on several ways of splitting the dataset (described in Figure 3). As mentioned previously, we split our dataset into a training and a testing set in order to evaluate the performance of our models.

Figure 3: Cross-validation methods for time series

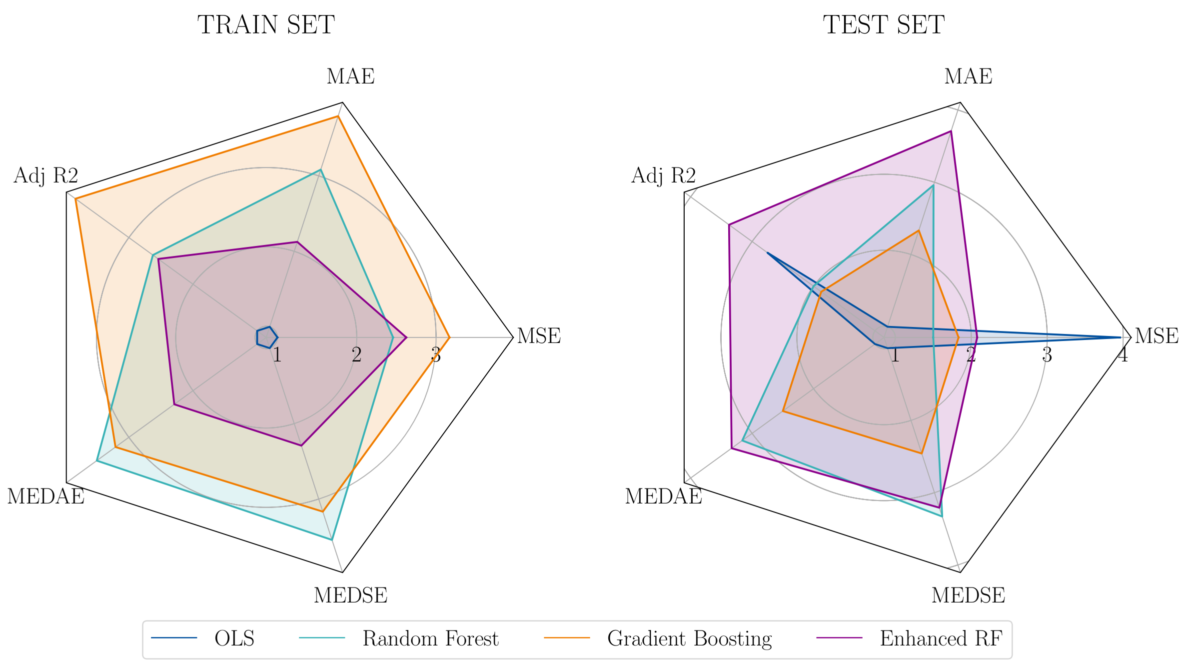


Source: Amundi Institute Quantitative Research

Therefore, with *Method 1*, we adopt a time-series split, which means that the training set retains the oldest data and the test set follows the training set by incorporating data from the most recent years. The first iteration considers the whole period and the last iteration 25% of the data set considering 4 iterations or 20% of the dataset considering 5 iterations. For *Method 2*, we apply the same method incorporating a gap between the training and the test sets of about 6 months. In *Method 3*, we use similar size of data with different time horizons. We split the data with the time series method. In each block, assigning the first part of the data to the training set and the last one to the test set.

First, we focus our study on the entire sample period (2011-2021) using daily data. The objective is against regress credit excess returns to the risk factors[6]. We consider 4 then 5 iterations per method and split the time series into different proportions of the test set (25%, 30%, 35%, and 40%). We determine which regression best explains and predicts the market in terms of excess returns among those described in Section 3 and according to a set of metrics detailed in Appendix A.2.2. We select these 5 metrics to be consistent with our intention to display robustness in the models. We launch all the algorithms in *Mlflow* (Zaharia *et al.*, 2020), part of the *Alto Studio platform*. The framework is described in Appendix A.3. We perform 96 runs for each model for a total of 2304 runs. For each method of Figure 3, the error metrics are computed for each iteration and then averaged to allow us to obtain robust statistics.

Figure 4: Average rank for bonds denominated in EUR (2011-2021)

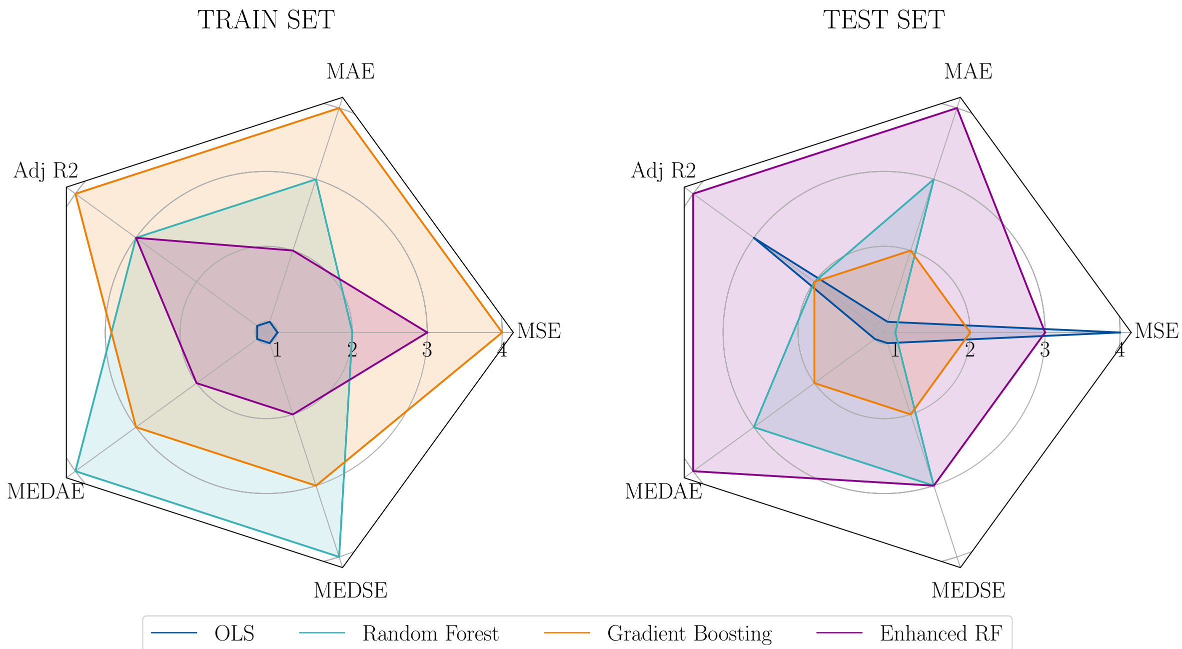


Source: Amundi Institute Quantitative Research

We aggregate our results by ranking the statistics of each model and showing the average rank in Figure 4 on the training set and the test set for EUR-denominated bonds. Similarly, Figure 5

[6]The risk factors are the following: duration, DTS, liquidity, value, size, and momentum.

Figure 5: Median rank for bonds denominated in EUR (2011-2021)

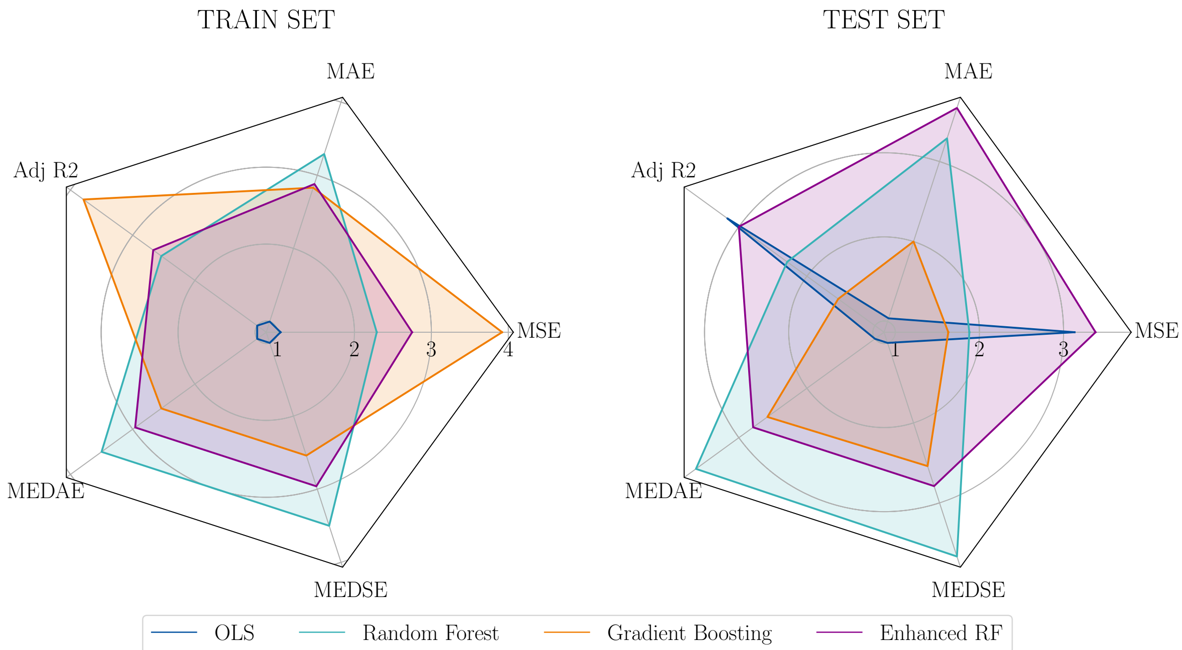


Source: Amundi Institute Quantitative Research

exhibits the median rank of each model's statistics. The first observation from these figures is that Gradient Boosting regression have better prediction in the training set but does not perform well out-of-sample. If we compare its metrics with the statistics of the other regressions, we may think that this regression is overfitting, but it is important to note that we are considering rank. On the other hand, the Random Forest algorithm, the second tree-based regression used in our analysis, is ranked second in-sample and oscillates between second and third place out-of-sample. The second observation that can be made is that the OLS regression is ranked last for all metrics in the training set and is the weakest out-of-sample for 3 over 5 metrics. In terms of predictive power, the OLS ranks second out of sample and first for the mean squared error. Finally, we notice that the Enhanced Random Forest regression is the best performing out-of-sample in terms of predictive power and error metrics for 4 over 5 metrics that we retained during the 2011-2021 period and performs well whatever the period used in the analysis (see Appendix A.4.1).

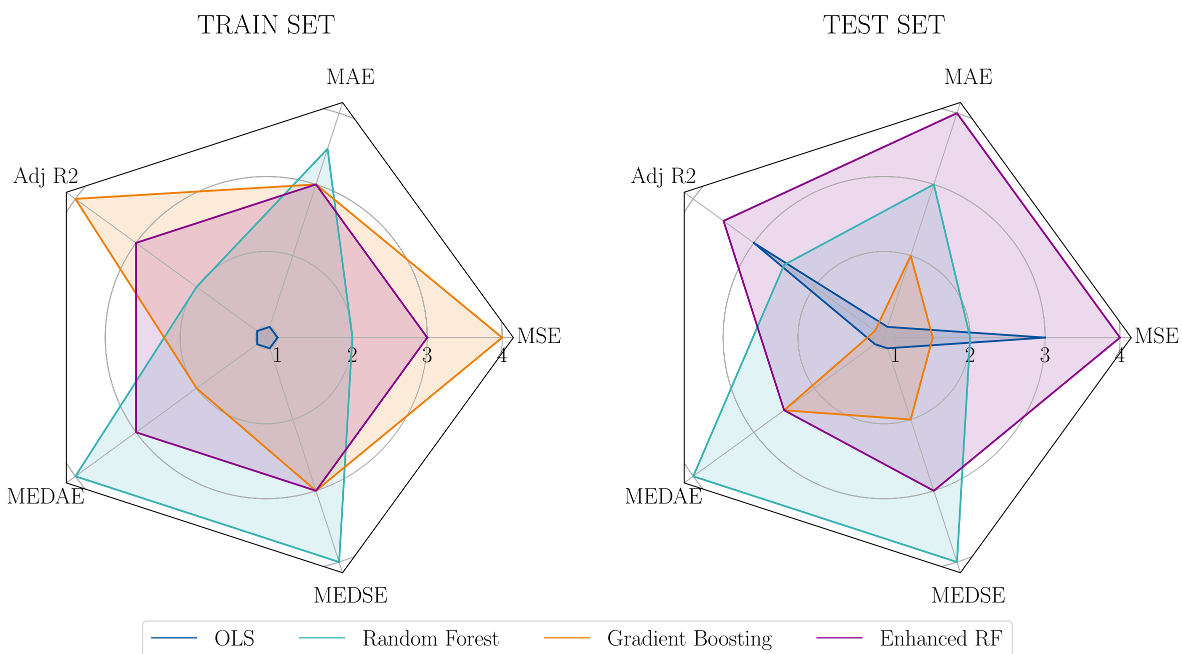
For USD-denominated bonds, the difference in gap narrows on the train set for the three regressions using a nonlinear process regardless of the time period used (see Figures 6, 7 and 24, 25, 26, 27 in Appendix A.4.2). Moreover, we note that the rank of the linear regression is always at the lowest. When focusing on the out-of-sample dataset, we can make two observations. First, for the entire considered period, the best-ranked regressions are those using the Random Forest algorithm. Also for the OLS regression is still performing well for the MSE and R^2_{Adj} metrics (see Figures 6 and 7). The Gradient Boosting regression is performing less compared to the others. Then, if we focus on the two sub-periods: 2011-2014 and 2015-2021, we observe a switch in the best ranked regressions. The Enhanced Random Forest and OLS are ranked first in

Figure 6: Average rank for bonds denominated in USD (2011-2021)



Source: Amundi Institute Quantitative Research

Figure 7: Median rank for bonds denominated in USD (2011-2021)



Source: Amundi Institute Quantitative Research

this configuration.

To conclude, regardless of the dataset, when we consider in-sample data, nonlinear models and in particular Gradient Boosting perform well for all measures. Therefore, tree-based regressions are good candidates to explain past returns. However, the nonlinear algorithms fit too much to the in-sample data, which leads to lower out-of-sample performance and better scores for the linear models. A model mixing these two approaches seems more appropriate for predicting the market excess returns. In this study, we considered traditional and alternative factors to compare the error of the models. Next, we seek to analyse the R^2_{Adj} measures by factors category and by period to evaluate their respective contribution in explaining credit excess returns of the market.

4.2 Factor model evaluation

In this section, we seek to evaluate the contribution of the factors in the credit pricing model. Ben Slimane *et al.* (2018) decomposed the dataset by traditional factors and alternative factors presented in Section 2 to evaluate which factors category contributes the most to the returns of the bonds belonging to the BofA Merrill Lynch Global Large Cap Corporate Index by periods. In this analysis, we conduct a different approach where we estimate the contribution to the predictive of each factor group compared to the 3-Factors model (3-F) the traditional factors.

4.2.1 Explanatory and predictive power analysis for the factors denominated in EUR

We start our analysis with the factors denominated in EUR. We still evaluate the R^2_{Adj} with the cross-validation methods, this time considering 4 iterations implying 4 metrics that we average. The final statistic is computed as an average of the R^2_{Adj} for test sizes 25%, 30%, 35%, and 40%. We first consider the traditional factors to explain the market excess returns, then integrate alternative factors into the model which becomes a 6-Factors model (6-F). We finally incorporate the ESG returns to form a 7-Factors model (7-F).

Table 2 shows the contribution of the R^2_{Adj} when we add first the alternative factors, then the later factors and ESG returns to the 3-F model considering the train and test set for the 2011-2021 period. We notice that adding alternative factors to the 3-F model brings higher prediction when explaining the credit excess returns. By considering each regression alone, we remark a slightly higher contribution for the OLS regression compared to others, even though Gradient Boosting regression displays good improvement in fitting power for methods 2 and 3 when considering the train set. If we focus on the test set, we clearly notice the added value of alternative factors on the period where the minimal value in the table is +28.67%. However, when adding the ESG returns to the 6-F model, we note that ESG seems to be poorly captured by the R^2_{Adj} compared to the 6-F model on both the train and the test sets. Ben Slimane *et al.* (2019) decomposed the dataset into sub-periods to confirm the contribution of the ESG from 2015 onwards in explaining market returns. We, therefore, find it interesting to perform the same study focusing on the contribution of the alternative factors and ESG over these two sub-periods.

Thus, between 2011 and 2014, the contribution of alternative factors is relatively small compared to the numbers observed in the previous table on the train set (see Table 3). However, when we focus on the predictive aspect, it is interesting to note that the contributions spectacularly improve for all regressions on the test set. Then, Table 4 displays the figures for the period

Table 2: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for EUR data (2011-2021)

Split method	Model	TRAIN SET		TEST SET	
		3F→6F	3F→7F	3F→6F	3F→7F
Method 1	OLS	12.57	12.82	36.72	34.52
	RF	8.95	8.64	30.63	29.93
	GB	8.38	8.86	29.63	26.16
	ERF	4.97	5.74	36.41	36.70
Method 2	OLS	12.22	12.46	38.09	35.18
	RF	8.57	8.35	30.16	29.85
	GB	14.42	13.82	28.67	26.83
	ERF	4.73	5.38	40.04	39.99
Method 3	OLS	19.81	21.84	89.81	87.09
	RF	8.59	9.33	69.68	71.46
	GB	14.42	14.29	40.20	43.21
	ERF	7.71	7.67	71.29	72.73

The risk factors of the 3-F model are the following: DTS, duration and liquidity factors.
The risk factors of the 6-F model include those of the 3-F model plus size, momentum and value factors.
The risk factors of the 7-F model includes those of the 6-F model plus ESG.
Source: Amundi Institute Quantitative Research

Table 3: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for EUR data (2011-2014)

Split method	Model	TRAIN SET		TEST SET	
		3F→6F	3F→7F	3F→6F	3F→7F
Method 1	OLS	−0.47	4.36	55.51	53.70
	RF	2.82	2.68	25.06	27.78
	GB	1.39	3.37	24.84	24.22
	ERF	5.42	6.79	30.55	30.00
Method 2	OLS	0.54	4.16	63.69	61.70
	RF	2.85	2.73	30.87	32.46
	GB	3.10	3.89	21.34	24.09
	ERF	6.80	7.24	32.94	34.53
Method 3	OLS	1.71	6.66	50.11	36.01
	RF	2.43	3.44	51.67	50.54
	GB	1.20	1.78	89.76	83.54
	ERF	−0.34	1.90	62.31	59.92

Source: Amundi Institute Quantitative Research

Table 4: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for EUR data (2015-2021)

Split method	Model	TRAIN SET		TEST SET	
		3F→6F	3F→7F	3F→6F	3F→7F
Method 1	OLS	27.44	32.92	12.04	10.09
	RF	8.07	9.17	18.21	17.82
	GB	17.96	16.03	18.56	20.43
	ERF	9.02	8.16	25.67	25.42
Method 2	OLS	28.15	33.86	12.26	10.09
	RF	9.13	9.06	18.59	18.28
	GB	17.51	16.55	18.56	21.42
	ERF	8.57	7.25	25.86	25.21
Method 3	OLS	15.55	19.08	15.86	15.76
	RF	5.74	6.47	22.58	22.02
	GB	11.90	12.71	32.95	35.09
	ERF	6.97	7.12	27.13	26.51

Source: Amundi Institute Quantitative Research

2015-2021. We distinguish several effects. First, we notice that the variation in R^2_{Adj}s have clearly increased compared to the 2011-2014 period for all regressions, it can be assumed that the credit excess returns of the market are explained by the alternative factors during this period. Second, for OLS regression, we observe a substantial enhancement of the R^2_{Adj} when adding alternative factors in the model for the train set which is not reflected in the test set compared to other regressions.

For ESG, the contribution to the R^2_{Adj} is also low in this period. If we focus on the improvement in the predictive power of ESG by breaking it down by sub-pillars (Table 5), looking first at the train set, we distinguish a marginal added value in predictive power for the social pillar compared to the other sub-pillars as well as from ESG itself. We also assess a higher contribution of the ESG pillar in the R^2_{Adj} on the test set compared to the sub-pillars contributions.

Table 5: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for EUR data (2015-2021)

Split method	Model	TRAIN SET 3F→7F			TEST SET 3F→7F		
		E	S	G	E	S	G
Method 1	OLS	29.94	32.12	28.58	9.57	9.63	13.15
	RF	8.26	9.75	9.50	17.59	17.49	17.38
	GB	18.29	16.97	17.80	18.82	21.40	20.40
	ERF	8.07	7.87	7.76	24.82	24.54	24.84
Method 2	OLS	30.77	32.99	29.31	9.60	9.21	13.54
	RF	8.57	9.81	8.54	17.60	18.60	18.12
	GB	18.35	18.71	18.75	19.45	21.18	20.09
	ERF	8.59	7.25	7.86	25.17	24.45	24.94
Method 3	OLS	16.52	18.50	18.24	15.31	11.86	17.44
	RF	5.61	6.80	6.34	20.31	20.69	21.41
	GB	12.92	13.14	12.52	30.40	30.45	31.58
	ERF	6.75	7.12	6.48	25.60	23.57	22.41

The risk factors of the 3-F model are the following: DTS, duration and liquidity factors.
The risk factors of the 6-F model include those of the 3-F model plus size, momentum and value factors.
The risk factors of the 7-F model includes those of the 6-F model plus ESG sub-pillars.
Source: Amundi Institute Quantitative Research

4.2.2 Explanatory and predictive power analysis for the factors denominated in USD

We perform the same analysis for the USD factors against the credit excess returns in USD. In general, alternative factors and ESG contribute less to credit excess returns of the market over the 2011-2021 period. Traditional factors alone can explain a large part of the market excess returns. We notice that the predictive power is higher for returns of USD-denominated bonds than for returns of EUR-denominated bonds. On the other hand, we can observe that the 7-F contributes more to the explanation of the model than the 6-F. This result is noticed both on the train and the test sets, whatever the period (see Tables 8, 9, and 10 in Appendix A.5).

Table 11 in Appendix A.5 shows the improvement in predictive power from the addition of ESG sub-pillars to the traditional factors. Each column represents one pillar for the 2015-2021 period. We cannot draw the same conclusions as for the EUR-denominated bonds since ESG prevails in terms of additional predictive power compared to its sub-pillars. However, when comparing the contributions of the three sub-pillars, we observe that governance contributes marginally more than the other sub-pillars on the train set.

To conclude, alternative factors bring a real added value to explain the credit returns in excess whatever the currency or the period. value, momentum and size factors have been key drivers of credit excess returns of the market since 2015. For ESG, we cannot draw the same conclusions. But the question is whether it is a transitory factor and whether the ESG contribution would not appear succinctly over short periods? This is what we try to determine in Section 5.2.2. In the next section, we seek to determine the exposure of credit excess returns of the market to factors. However, before that, we evaluate the collinearity of the factors that can be a major obstacle to model interpretation.

4.2.3 Dealing with multi-collinearity in the models

Table 6 shows the variance inflation factor (VIF) of the EUR factors. The VIF aims to detect multi-collinearity between variables computed on the linear model. Multi-collinearity does not affect how well the model fits but can be a huge problem for linear models as it can lead to a misunderstanding of the model. We observe that the VIF associated with DTS and value are inflated by a factor between 4 and 8. The reason is the substantial correlation between these variables. We find similar figures in Table 7 for the USD factors. Our value score is built upon a spread model. We consider a valuable bond if it has a high discrepancy between the log difference of the market spread and a theoretical spread (see Ben Slimane *et al.* (2018)). The DTS is a product of a spread and the duration, the DTS of our value factor is elevated by a factor of 3.0 of the market DTS, which could explain such results.

According to O'Brien (2007), the fact that VIF values are below 7 suggests a moderate probability of multi-collinearity in our data. But authors such as Johnston *et al.* (2018) consider that VIF values greater than or equal to 2.5 demonstrate considerable collinearity. In the next studies, we will focus our analysis on the 2015-2021 period which exhibits the highest VIF factors for DTS and particularly for the EUR factors which displays high VIF values. This suggests that analyzing this period might indicate an erroneous significance of the variables because of multi-collinearity. Another consequence is the excessive sensitivity of the coefficients impacted by a minor change in the model. Consequently, we decided not to rely on the OLS coefficients in the following sections. The next step consists in capturing the feature importance of the Random Forest and Gradient Boosting regressions to deduct the accuracy and the prediction interpretability.

Table 6: Variance Inflation Factors for the explanatory variables over the analysis periods (EUR)

VIFs	2011-2021	2011-2014	2015-2021
DTS	**5.889**	**5.442**	**7.214**
$Duration$	1.538	1.569	1.601
$Liquidity$	1.943	1.629	2.212
$Value$	**5.299**	**6.380**	**4.983**
$Size$	1.461	3.379	1.253
$Momentum$	1.094	1.148	1.108
ESG	1.426	1.463	1.765

Source: Amundi Institute Quantitative Research

Table 7: Variance Inflation Factors for the explanatory variables over the analysis periods (USD)

VIFs	2011-2021	2011-2014	2015-2021
DTS	**5.428**	**4.066**	**6.570**
Duration	1.607	1.571	1.801
Liquidity	2.149	1.650	2.639
Value	**4.190**	**3.563**	**5.259**
Size	1.419	1.822	1.625
Momentum	1.134	1.135	1.227
ESG	1.520	1.551	1.596

Source: Amundi Institute Quantitative Research

5 Model interpretability

As mentioned in Section 1, our objective is to explain the tree-based regression models with feature importance. In this section, we first present different methods, and we interpret feature importance through our market insight.

5.1 Feature importance methods

One common approach is a heuristic method introduced by Brieman (2001) for Random Forest algorithms and is related to an accuracy-based importance. The variable importance is built according to the prediction accuracy of shuffled out-of-bag samples of data for each variable j and each tree. Another method consists of summing in each tree the accumulation of the decrease in *"impurity"* (mean squared error)[7] that we used for the variable selection and normalized it. The final feature importance is measured as the average normalized variable importance per feature for all the trees in the forest. The higher the number of appearances in trees with a high score, the higher the importance. This assumption carries two limitations. On the one hand, this statistic is computed on the training set and does not consider the test set, which is crucial to measure the prediction accuracy and it is inaccurate in case of over-fitting. On the other hand, this statistic can be biased for high cardinality[8] features.

We have tested two supplementary methods to determine the importance of the risk factors estimated with the test set. Permutation importance allows evaluation with a test of how important a pre-defined model is. But it cannot conclude about the robustness of the model prediction. The principle is to permute the values of a single feature, then estimate the tree regression model to measure the *new* model's accuracy. The feature importance is the difference between the initial accuracy score and the drop in all-inclusive accuracy generated by the permutation. When two features are collinear, the permutation method becomes a real problem because the outputs are misleading (Hooker and Mentch, 2019), given the lower consideration of highly important variables if these features are dependent.

Lundberg and Su-In (2017) propose the SHAP values for SHapley Additive exPlanations as

[7] The decrease in *"impurity"* is defined in Equation 2 of Appendix A.2.1.

[8] Too many unique values

a response to the complexity to interpret the feature importance. They propose this measure in order to explain the impact that each variable exhibits on the model's prediction. It is based on the shapley values of the game theory and is described as the average marginal contribution of a feature over every probable coalition. In other words, the goal is to compute the average difference of the contribution of each feature to the prediction and the average prediction for a set of possible combinations of features. The added value of SHAP comes from the representation of SHAP values as an additive attribution method. Lundberg *et al.* (2019) proposed TreeSHAP as an alternative to SHAP values built for tree-based models using conditional expectation instead of the marginal expectation. This method is more likely to attribute importance to a feature that has no-effect on the prediction if this feature is correlated to another feature. Thus, we use the causal inference rules to cut the dependencies between features (Janzing *et al.*, 2019).

5.2 Market interpretation

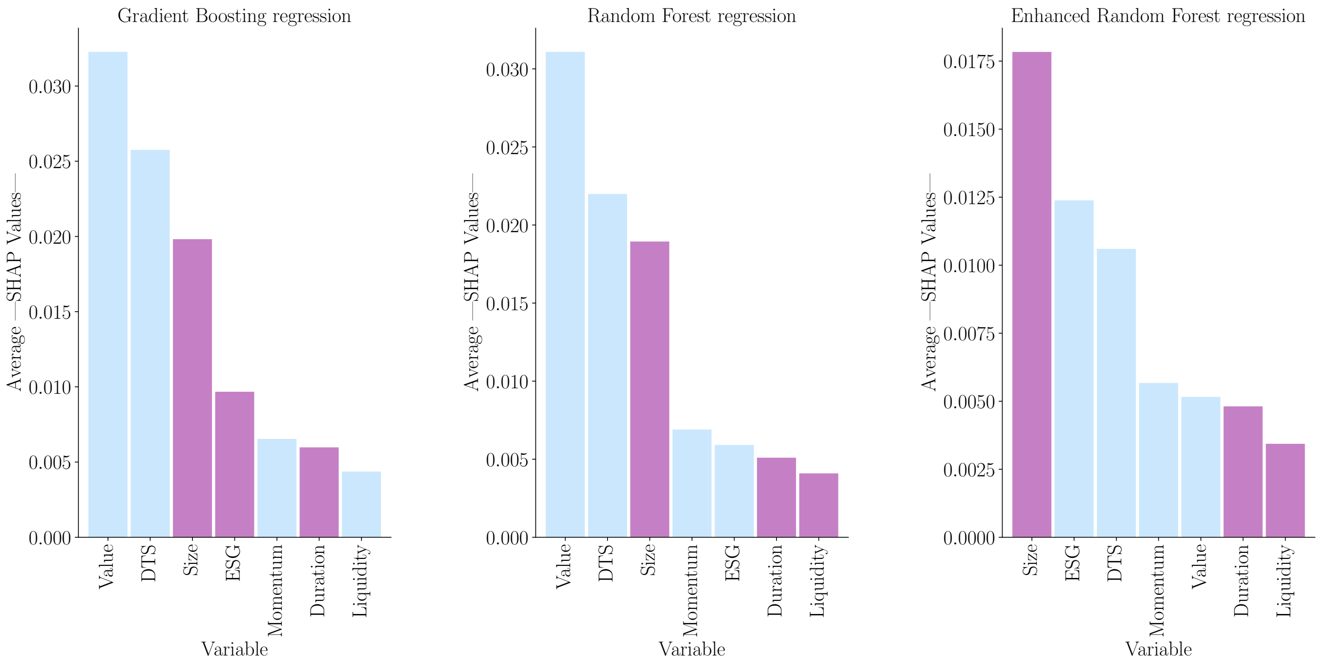
The goal of this section is to introduce the feature importance using the SHAP values that we applied in the context of equity factor investing in a previous research (Lepetit *et al.*, 2021). With this method, we can break down the prediction of the risk factors in order to measure the impact of each variable on the estimated credit excess returns.

5.2.1 Factors exposure

We explain the raw output of the tree-based regression models that we calibrated using the optimized hyper-parameters captured with the cross-validation grid-search. We measure the features importance over the 2015-2021 period. We average the absolute SHAP values to determine the impact of each factor on the credit returns in excess. Figures 8 and 9 respectively show the average absolute values for EUR-denominated and USD-denominated factors for tree-based regressions that we used in our analysis. On the one hand, we observe for most cases that three variables appear in the top 3 important features over the period, namely DTS, value and size. The absolute average SHAP values allow us to estimate the strength of a factor's returns relative to the credit excess returns but not its direction. In these figures, the pink bars indicate whether the returns of the variables are negatively correlated with their SHAP values and blue if they are positively correlated. It reveals that DTS and value factors returns are positively correlated with their SHAP values, while the size factor is negatively correlated.

It is interesting to observe the interaction effect of DTS and value factors on credit excess returns of the market. Figures 28 and 29 in Appendix B.1 show the relationship between DTS returns and its SHAP values for Random Forest algorithm and Gradient Boosting regression. Thus, we observe a linear and increasing relationship between the importance of DTS in the model's prediction and its returns. This means that for periods when the factor returns are positive, it has a positive and moderate impact on the output model, and conversely when DTS returns are strongly negative, it has a negative and a non negligible impact on the market excess returns. Furthermore, the colors indicate the positive (blue) and negative (red) interaction of DTS and value factors. For example, it means that very low standardized returns for DTS (-3.0) and negative standardized returns for value factor (-0.6) impact negatively but in a moderate manner (around -0.2) the credit returns in excess. On the contrary, for very high standardized returns of

Figure 8: average —SHAP values— for EUR factors using 2015-2021 period



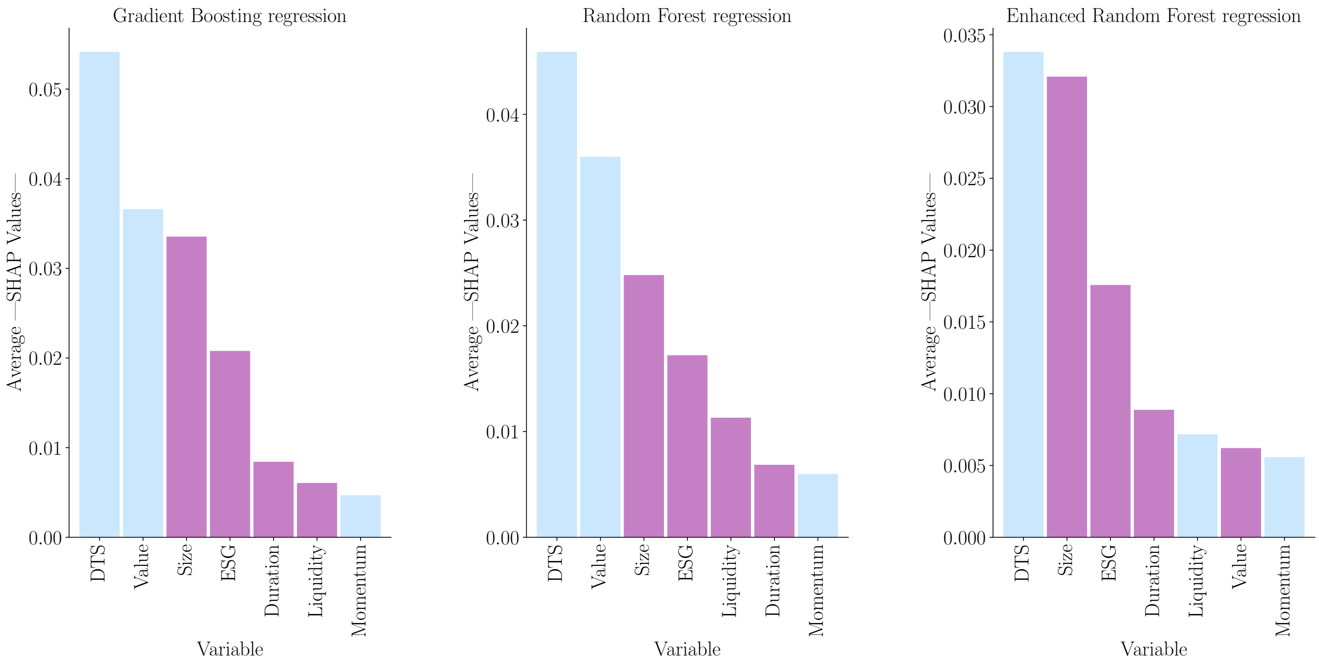
Source: Amundi Institute Quantitative Research

DTS (+3.0) and positive standardized returns of value factor (+0.4), we notice a positive impact on the output model. We distinguish the same relationship between DTS and value factor for USD-denominated bonds (see Figure 32 for Random Forest algorithm and Figure 33 for Gradient Boosting regression in Appendix B.2).

Moreover, Figures 8 and 9 show the feature importance of the Enhanced Random Forest algorithm. As a reminder, these average absolute SHAP values exhibit the exposures that have not been detected by the linear model. Indeed, we explain the error of the lasso regression with the explanatory variables using Random Forest regression. We present the feature importance because we can evaluate that a linear combination of factors is not enough sufficient to explain the credit excess returns of the market. For example, for USD-denominated bonds (see Figure 9), over the period, we observe that the average absolute SHAP values record significant numbers. The three variables appearing as the most important either for EUR-denominated bonds or USD-denominated bonds are DTS, size, and ESG. Some relations have not been recorded mainly for these three features.

Finally, for the USD-denominated bonds, we observe that the ESG returns have a non-negligible impact on the credit excess returns of the market as it is ranked 4th or 3rd in the feature importance. In the next section, we analyse if this variable has any relation with the other top features.

Figure 9: average —SHAP values— for USD factors using 2015-2021 period



Source: Amundi Institute Quantitative Research

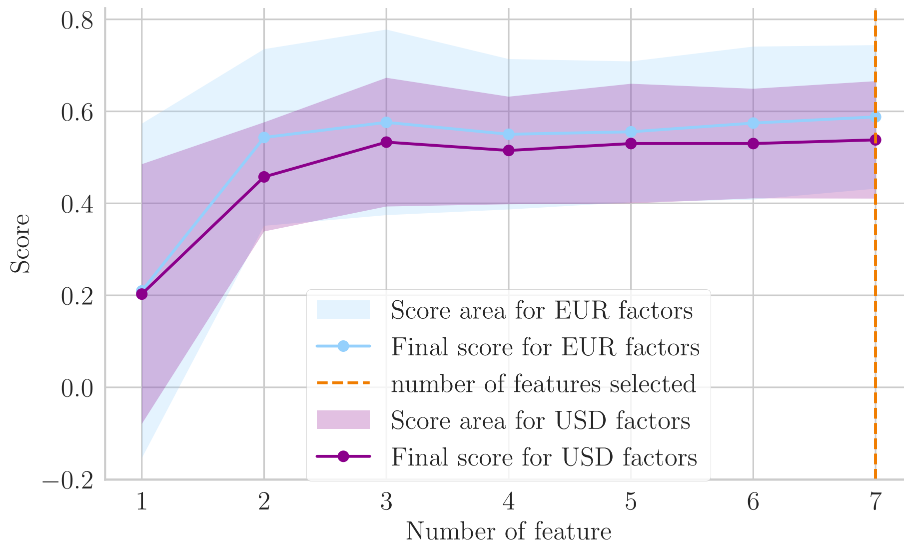
5.2.2 ESG as a transitory factor?

As mentioned in Section 4.2.2, we assess the impact of ESG returns on the credit excess returns of the market. We raised the point that ESG could appear as a transitory explanatory variable. Figure 10 shows the optimal number of selected factors for the Random Forest algorithm based on the 2015-2021 period. We then run the same analysis using Gradient Boosting instead (see Figure 11). This algorithm works as follows: we first eliminate 0 to N features according to the second feature importance method presented in Section 5.1.

In a second step, we compute a cross-validation score[9] for each subset of selected models. Finally, the average of the cross-validation score determines the best-performing model. The shaded area of the figures represents the variability of the cross-validation score from +1 to -1 standard deviation of the accuracy score mean[10]. From these figures, we note that whatever the currency, the number of features selected by recursive feature elimination is 7. From the Random Forest algorithm point of view, based on Figure 10, the optimal number of features is 3 but keeping 7 factors does not decrease the value of the model. For Gradient Boosting regression (see Figure 11), the scores seem to improve, even with the addition of a seventh factor. This result implies that all variables remain important in explaining credit excess returns of the market during the period. There is a value added from including ESG return in our factor model, but the question raised is how to evaluate this contribution.

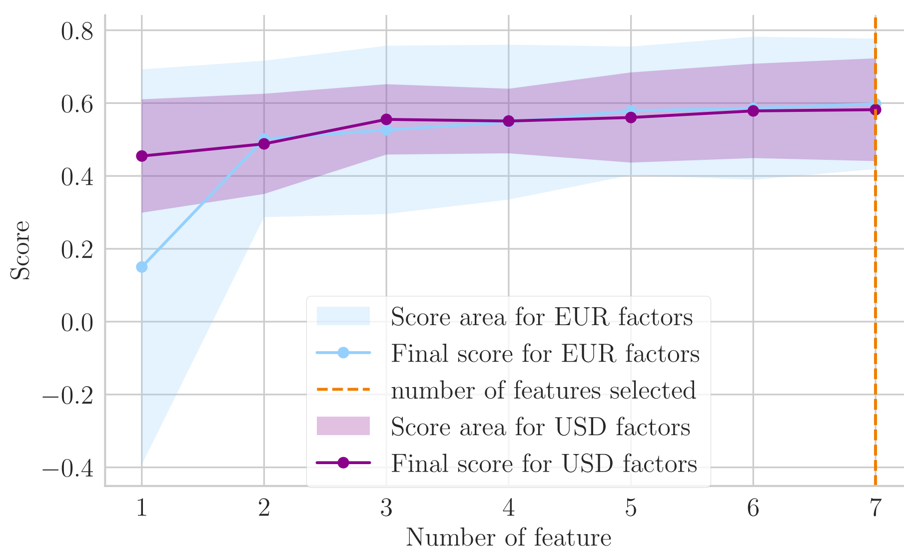
[9]The cross-validated score is computed according to *Method 1* using 4 iterations.
[10]This score being the R-squared.

Figure 10: Optimal number of factors using Random Forest algorithm for EUR and USD factors (2015-2021)



Source: Amundi Institute Quantitative Research

Figure 11: Optimal number of factors using Gradient Boosting algorithm for EUR and USD factors (2015-2021)

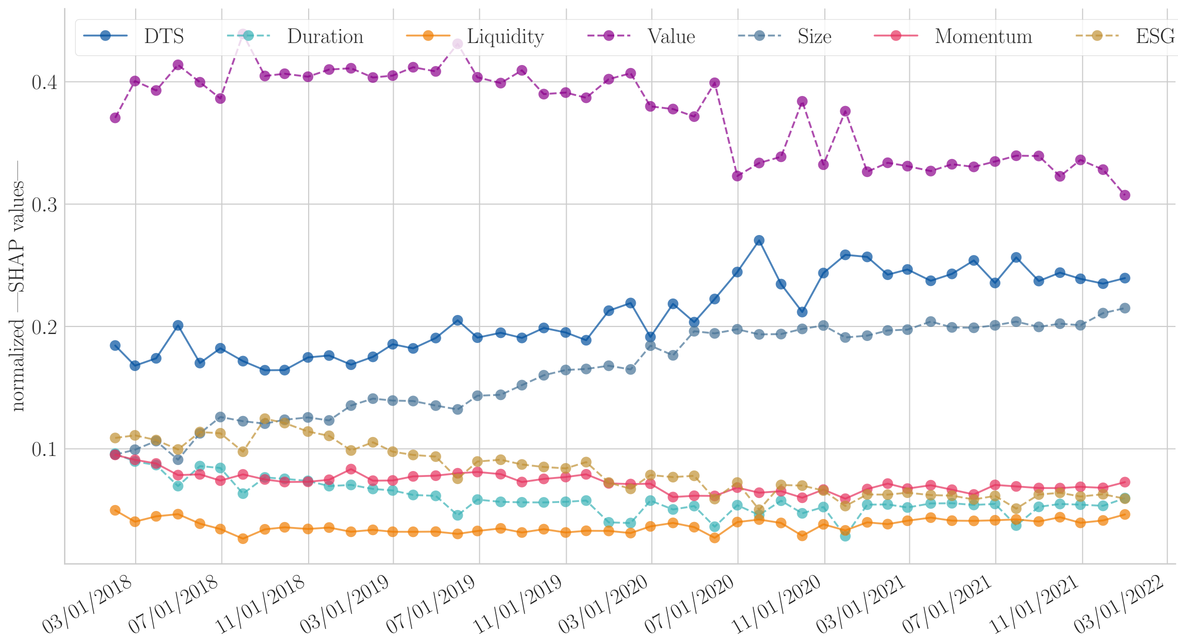


Source: Amundi Institute Quantitative Research

To do so, we recursively run tree-based regressions for the 7-F model over the 2015-2018 period and add the next month's returns each time. First, we show in Figure 12 and Figure 13, the normalized —SHAP values— for the EUR-denominated factors after running the Random

Forest and the Gradient Boosting regressions, respectively. In these figures, we notice that the importance rank of the features is almost similar for both regressions, indicating the robustness and consistency of our results. We can confirm the importance of factors such as value, DTS, and size that we observed in Section 5.2.1 over the period. We also validate the less significant importance of duration, liquidity, and momentum factors.

Figure 12: Normalized average —SHAP values— for EUR factor and Random Forest regression (2018-2021)

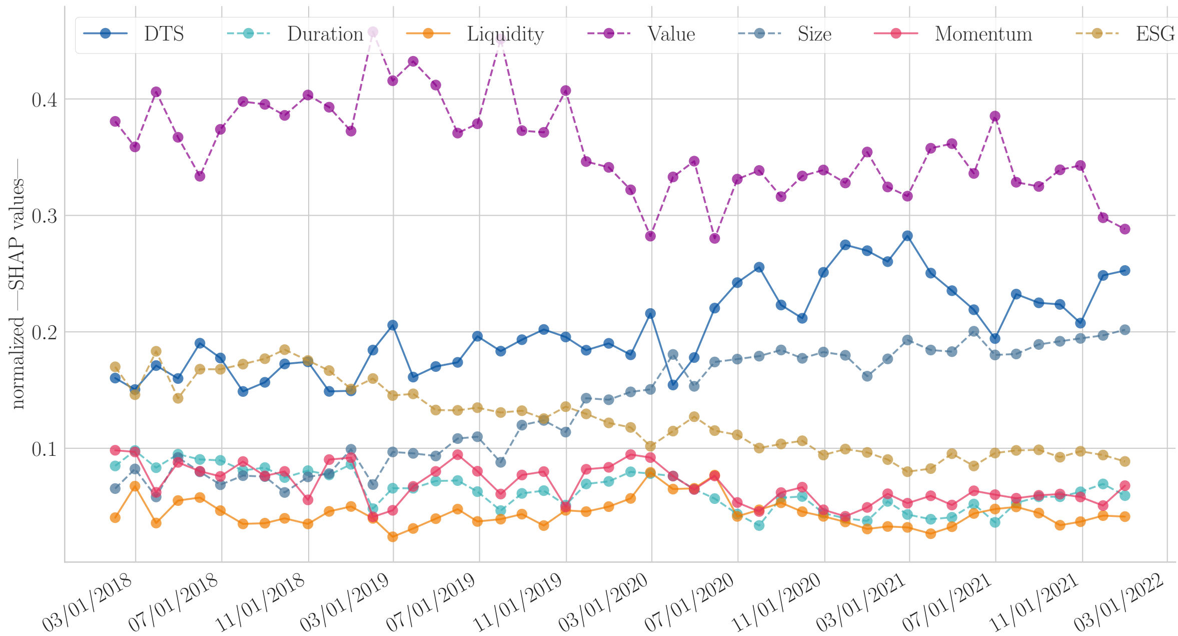


Source: Amundi Institute Quantitative Research

If we now focus on the ESG variable, we find that the importance of this feature is decreasing as the size factor importance increases. In that respect, we confirm the results of Ben Slimane *et al.* (2020) mentioned in Section 2.3, the ESG variable is ranked among the most important market drivers and its position gradually declined during the COVID-19 crisis because it plays a role in hedging investors' concerns. Indeed, we can identify that ESG is as important as DTS and size for Figure 13 or size only for Figure 12 to explain the EUR credit excess returns of the market between 2015 and early 2019. For Gradient Boosting regression, we then detect a decline in significance between 2019 and the early stage of COVID-19 crisis (March 2020), ending with a slight rebound during the COVID-19 crisis (March 2020 to June 2020) and finally quasi-stagnating at a level well below value, DTS and size factors since July 2020. For Random Forest regression, the break is much less gradual, with a detachment between size and ESG starting in 2019.

Next, we present in Figure 14 and Figure 15 the normalized average —SHAP values— for the Random Forest and Gradient Boosting regressions, respectively, over the same period for USD-denominated bonds. We notice here that the most important factors are DTS, value, ESG and size. duration, liquidity, and momentum are in the last positions similarly to the of EUR-denominated bonds. We identify in the figures that the size and ESG importance lines cross at

Figure 13: Normalized average —SHAP values— for EUR factor and Gradient Boosting (2018-2021)

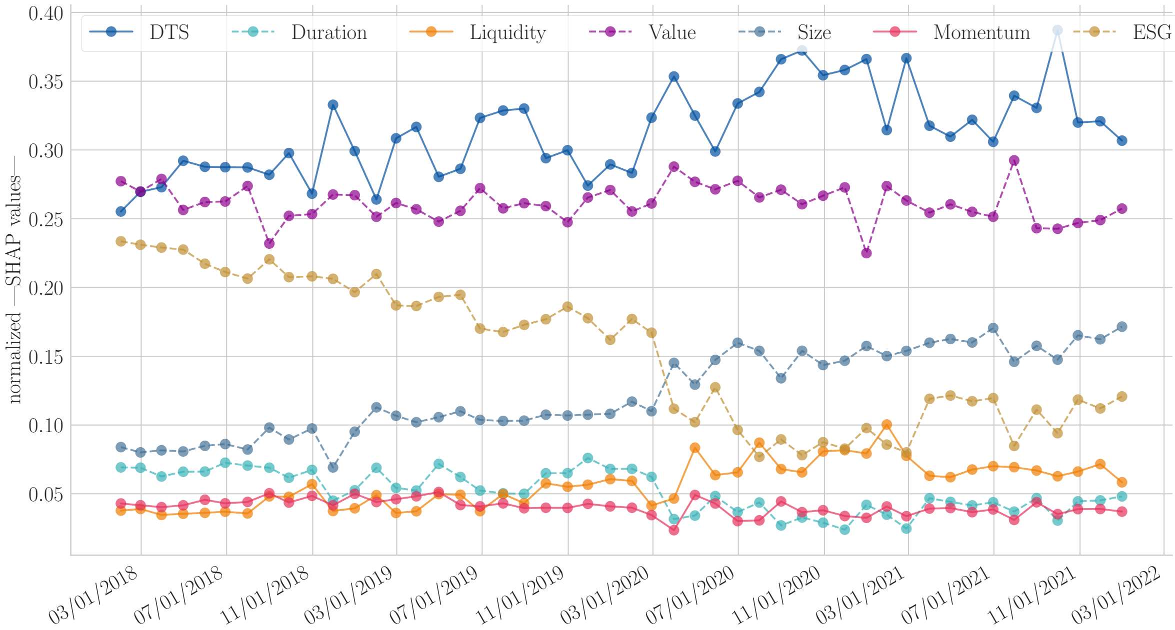


Source: Amundi Institute Quantitative Research

the beginning of the COVID-19 crisis. As was the case for EUR-denominated bonds, ESG loses importance to the benefit of the size factor. ESG is in the top 3 most important factors before the COVID-19 crisis, then falls below the top-ranked variables.

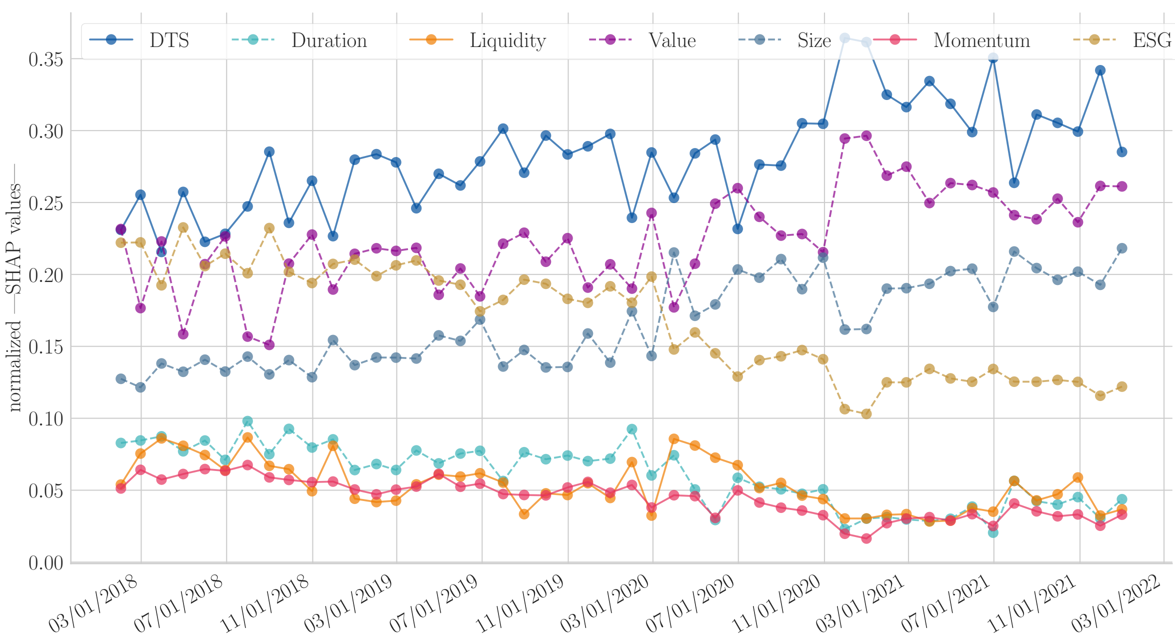
Therefore, we want to understand whether there is a relationship between the size and ESG factors, as the importance of the ESG variable in explaining market excess returns has been gradually replaced by the size factor. To this end, we show in Figure 30 and Figure 31 in Appendix B.1 the dependence between the returns of the size factor in EUR and its SHAP values, for the Random Forest and the Gradient Boosting regressions considering the global period 2015-2021. The relationship shown by the curved figures is non-linear, and the gradient function decreases along the curves. In reality, high positive returns have a negative impact on the output variable, or that large total debt value of the issuing firm explain the credit excess returns of the market over the period. Moreover, in these figures, we show the interaction between ESG and size returns through the colors, suggesting no empirical evidence of any relationship between the variables. We conduct the same study in Figure 34 and Figure 35 in Appendix B.2 for the USD factors on the same period and as previously, for the Random Forest algorithm and the Gradient Boosting regression leading to same conclusion.

Figure 14: Normalized average —SHAP values— for USD factor and Random Forest regression (2018-2021)



Source: Amundi Institute Quantitative Research

Figure 15: Normalized average —SHAP values— for USD factor and Gradient Boosting (2018-2021)



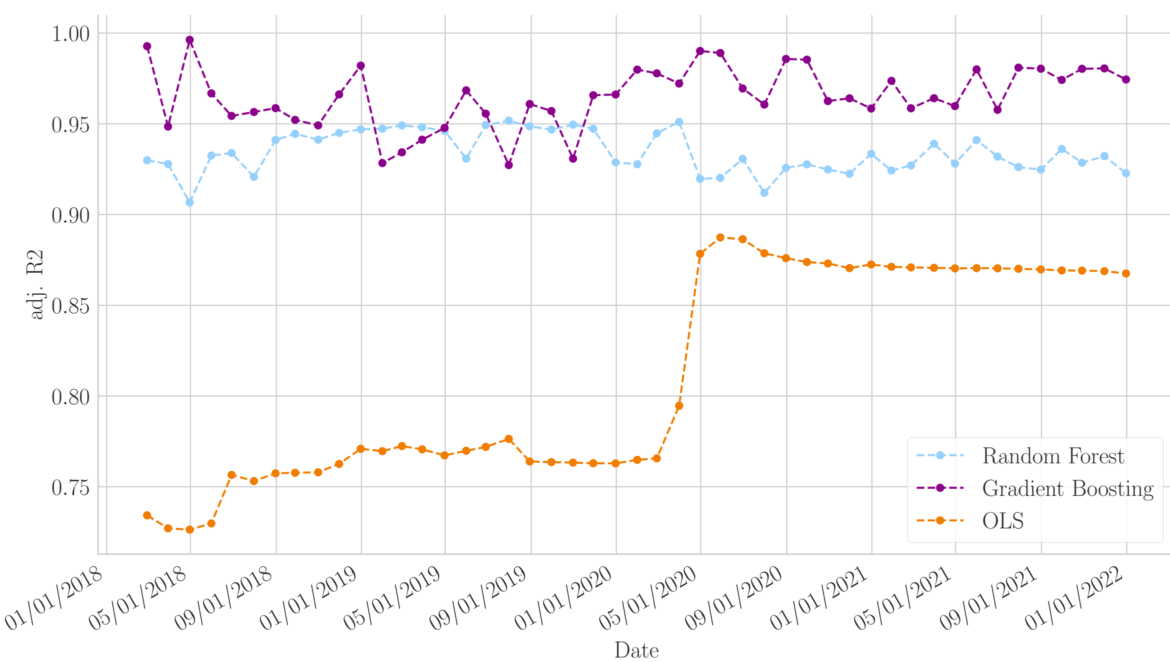
Source: Amundi Institute Quantitative Research

The importance of ESG varies with significantly amplified movements over the period similar to the size factor. We identify a breakout phase in early 2020. Despite this decline in importance, we cannot neglect the importance of this variable in the explanation of the credit excess returns of the market. Indeed, ESG explains market movements and remains higher significance rankness than traditional factors such as liquidity or alternative factor as the momentum factor. We also find that the Random Forest algorithm separates the factors into two distinct groups, regardless of currency, whereas the importance rank of the factors in the Gradient Boosting regression is more homogeneous.

5.2.3 Focus on the COVID-19 crisis

The objective of this sub-section is to analyze the explanatory power of tree-based regressions relative to the OLS regression during turbulent periods. It is an interesting exercise to quantify the fitness of the models during periods of market stress, but also to measure factor rotations and hedge behaviors during crisis.

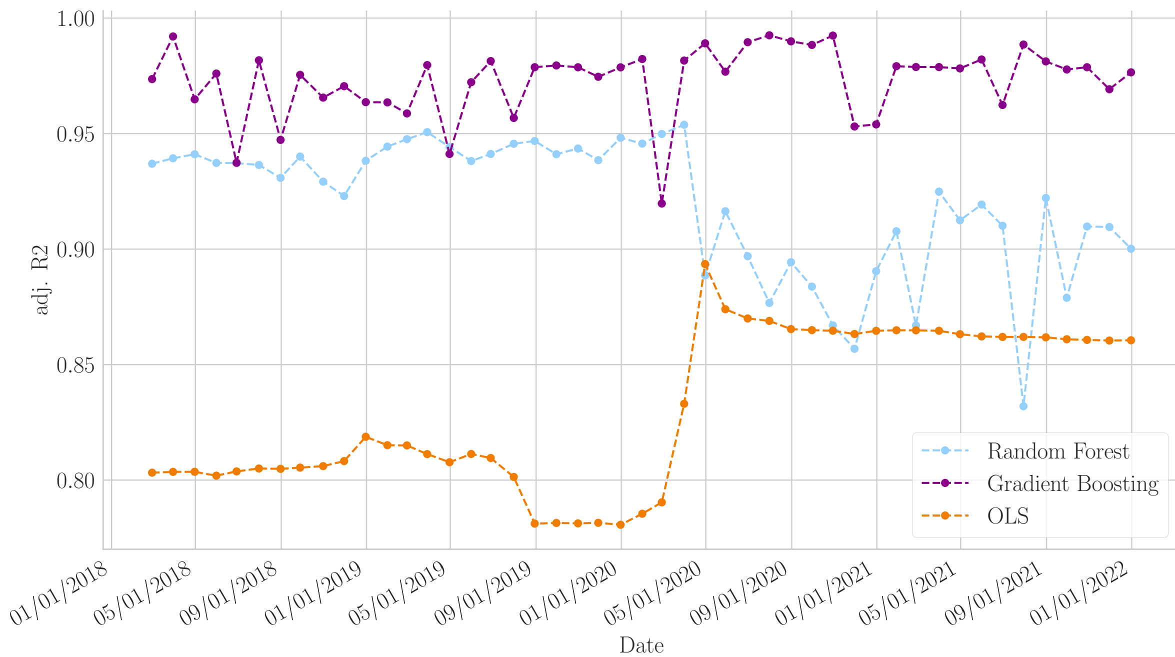
Figure 16: Rolling R^2_{Adj} for bonds denominated in EUR (2018-2021)



Source: Amundi Institute Quantitative Research

Figure 16 displays the rolling R^2_{Adj} of the 7-F model for the OLS regression and the tree-based regressions for EUR-denominated bonds. We perform a first regression from January 2015 to January 2018 and we add the following month's returns to the next's regressions. We observe a phenomenon in the explanatory power of OLS for EUR and USD-denominated bonds (see Figure 17) where the addition of returns attributed to the beginning of the COVID-19 crisis increases the explanatory power of the model by almost +112bps for the bonds denominated in

Figure 17: Rolling R^2_{Adj} for bonds denominated in USD (2018-2021)



Source: Amundi Institute Quantitative Research

EUR currency and +103bps for the bonds denominated in USD currency. This result is very interesting and shows the instability of the R^2_{Adj} during regime changes on the returns for the OLS case. We observe such deviations because the amplitude of the returns variation were extreme during the crisis, both for the variable we seek to explain or for the explanatory variables. The factor returns were positively or negatively correlated linearly to the market.

Concerning the Random Forest regression, we observe two disjoint events: for the EUR denominated bonds, the R^2_{Adj} seems stable over the whole period. We distinguish a slight drop in explanatory power at the beginning of the COVID-19 crisis but this is only temporary since the R^2_{Adj} automatically re-stabilises after July 2020. However, for the USD-denominated bonds, we discern two separate periods: first, we have a stability of the explanatory power from the beginning of the period until March 2020. Then, we notice volatile R^2_{Adj}, leading to disappointing results. Moreover, for the Gradient Boosting regression, we perceive a variation of the R^2_{Adj} around the average R^2_{Adj} of the period being 96.52% for bonds denominated in EUR and 97.30% for the bonds denominated in USD. We observe a higher amplitude of the variation for the latter but they seem quite stable compared to the other regressions. Empirically, we assume that there should be no extensive instability of the R^2_{Adj}. Consequently, Gradient Boosting seems to be an excellent candidate for factor analysis whatever the returns regime.

Let us now get back to Figure 13 and Figure 15 in order to study the impact of explanatory features on the explanation of the output variable at the beginning of the crisis by focusing on Gradient Boosting regressions results. We find that features such as value, DTS, size, and ESG for the EUR-denominated bonds and value, DTS, size, and liquidity for USD-denominated factors participated more than other features in explaining the credit excess returns during this period. In the research of Ben Slimane *et al.* (2020), we noticed a drastic shift from the 3rd rank to last as of the beginning of the COVID-19 crisis for the ESG variable while tree-based regressions do not neglect the importance of this feature. We observed also in the study that the VIF factor of ESG variable has increased strongly from these dates which could mislead its effect on the credit excess returns of the market with the linear model as indicated in the Section 4.2.3. Furthermore, focusing on a longer sample for analysis, we find results that are consistent with the previous study based on EUR-denominated bonds and on monthly returns, since the top three exposures to the market were value, DTS, and Size factors from 2014 to April 2020.

We see a major contribution of working with tree-based regressions for factor analysis which better fits our interpretation of what explains credit markets. This does not negate completely the exposures to credit excess returns of the market that we found in previous studies with linear models. Indeed, the higher contributor are still the same, which may be reassuring. This is a good way to complete the analysis on linear models as we found in Figure 8 and Figure 9 components that has not been recorded by the linear models. Moreover, it is a good way to evaluate the interaction between a feature and credit excess returns in the past as relationships were not always linear (see Appendix B.1 and Appendix B.2), especially for the size factor.

6 Conclusion

We notice a significant benefit in working with tree-based regressions for factor analysis. The results remain consistent with our studies but add robustness. In the previous study of factor investing and ESG (Ben Slimane *et al.*, 2020), we applied a Lasso regression process to determine the impact of ESG on the credit excesss return of the investment grade market. We noticed a drastic shift from the third rank to the last one after the beginning of the COVID-19 crisis for the ESG variable, while the tree regressions do not neglect the importance of this feature. We also observed that the VIF factor of the ESG variable increased strongly from these dates, which could mislead its effect on credit excess returns within the linear model.

Another way to prove the necessity and reliability of tree-based regressions is their insight into the error term of linear models. Accordingly, we detected nonlinear relationships between the error term of the linear model and credit risk factors. Indeed, the Enhanced Random Forest regression captured relationships from Lasso model errors, so it appears to be an appropriate way to complement the linear analyses.

Tree-based regressions methods assist us in improving our understanding of prices through the interaction assessment, not only between features but between each feature and the output of the model. As a result, we confirmed that DTS and value factors are the first drivers of excess returns in the EUR or USD markets over the 2015-2021 period, consistent with Ben Slimane *et al.* (2019). We were also able to identify the interaction effects between these two variables that show strong similarities in the impact of the output variable. Furthermore, in line with the study of Ben Slimane *et al.* (2020), we included ESG returns in our factor model and noticed that ESG the peak of importance was reached before the COVID-19 crisis, in line with previous study results.

As a next step, we aim to build a strategy using tree-based regressions in a cross-section instead of a time series in the context of improving the prediction of the spread model. We have chosen to take advantage of the tree-based regressions because of their better prediction accuracy than linear models, even if they are much more difficult to interpret than the latter. However, we show that using linear regressions leads to underfitting in the model by increasing the bias but may also lead to misinterpretation or inaccuracy within the model in the event of multi-collinearity or transformation of dummy variables, for example.

References

Bao, J., Pan, J., and Wang, J. (2011), The Illiquidity of Corporate Bonds, *Journal of Finance*, 66(3), pp. 911-946.

Bektic, D., Wenzler, J-S., Wegener, M., Schiereck, D., and Spielmann, T. (2019), Extending Fama-French to Corporate Bond Markets, *The Journal Of Portfolio Management*, 4(3), pp. 141-158.

Bennani, L., Le Guenedal, T., Lepetit, F., Ly, L., Mortier, V., and Sekine, T. (2018), The Alpha and Beta of ESG Investing, *Amundi Working Paper*, 76.

Bennani, L., Le Guenedal, T., Lepetit, F., Ly, L., Mortier, V., Roncalli, T., and Sekine, T. (2018), How ESG Investing Has Impacted the Asset Pricing in the Equity Market, *Available at SSRN*: http://ssrn.com/abstract=3316862.

Ben Dor, A., Dynkin, L., Hyman, J., Houweling, P., van Leeuwen, E., and Penninga, O. (2007), Duration Times Spread, *Journal of Portfolio Management*, 33(2), pp. 77-100.

Ben Slimane, M., and de Jong, M. (2017), Bond Liquidity Scores, *Journal of Fixed Income*, 27(1), pp. 77-82.

Ben Slimane, M., de Jong, M., Dumas, J-M., Fredj, H., Sekine, T., and Srb M. (2018), Traditional and Alternative Factors in Investment Grade Corporate Bond Investors, *Amundi Working Paper*, 78.

Ben Slimane, M., Brard, E., Le Guenedal, T., Roncalli, T., and Sekine, T. (2019), ESG Investing and Fixed Income: It's Time to Cross the Rubicon, *Available at SSRN*: http://ssrn.com/abstract=3683477.

Ben Slimane, M., Dumas, J-M. and Sekine, T. (2020), Factor Investing and ESG in the Corporate Bond Market Before and During the COVID-19 Crisis, *Amundi Discussion Paper*, 46.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984), Classification and regression trees, *Chapman and Hall/CRC*.

Breiman, L. (1996), Bagging Predictors, *Machine Learning*, 26(1), pp. 123-140.

Breiman, L. (2001), Random Forests, *Machine Learning*, 45(1), pp. 5-32.

Bonne, G., Wang, J. and Zhang, H. (2021), Machine Learning Factors: Capturing Nonlinearities in Linear Factor Models, *MSCI Research*.

Chen, L., Lesmond, D.A and Wei, J. (2007), Corporate Yield Spreads and Bond Liquidity, *Journal of Finance*, 62(1), pp. 119-149.

de Jong, F. and Driessen, J. (2012), Liquidity Risk Premia in Corporate Bond Markets, *Quarterly Journal of Finance*, 2(2), pp. 1-34.

Diettrich, T. (2000), An Experimental Comparison of Three Methods For Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, *Machine Learning*, 40(2), pp. 139-157.

Dixon, M. and Polson, N (2020), Deep Fundamental Factor Models, *SIAM Journal on Financial Mathematics*, 11(3), pp. SC26-SC37.

Drei, A., Le Guenedal, T., Lepetit, F., Ly, L., Mortier, V., Roncalli, T., and Sekine, T. (2019), ESG Investing in Recent Years: New Insights from Old Challenges, *Available at SSRN*: http://ssrn.com/abstract=3683469.

Fama, E.F, and French, K.R (1993), Common Risk Factors in the Returns on Stocks and Bonds, *Journal of Financial Economics*, 33(1), pp. 3-56.

Friedman, J.H (2001), Greedy Function Approximation: A Gradient Boosting Machine, *Ann. Statist.*, 29(5), pp. 1189-1232.

Galton, F. (1907), Vox Populi, *Nature*, 75(1949), pp. 450-451.

Gu, S., Kelly, B. and Xiu, D. (2020), Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies*, 33(5), pp. 2223-2273.

Hastie, T. (2009), The Elements of Statistical Learning: Data Mining, Inference and Prediction, *Springer, 2 edition*, pp. 363-592.

Hooker, G. and Mentch, L. (2019), Please Stop Permuting Features: An Explanation and Alternatives, *arXiv:1905.03151.*

Houweling, P., and Van Zundert, J., (2014), Factor Investing in the Corporate Bond Market, *Robeco Quantitative Strategies Working paper.*

Huang, J.Z, and Huang, M, (2012), How Much of the Corporate-Treasury Yield Spread is Due to Credit Risk?, *Review of Asset Pricing Studies*, 2(2), pp. 153-202.

Israel, R., Palhares, D. and Richardson, S.A, (2018), Common Factors in Corporate Bond and Bond Fund Returns, *Journal of Investment Management*, 16(2), pp. 17-46.

Janzing, D., Minorics, L., and Blöbaum, P. (2019), Feature Relevance Quantification in Explainable AI: A Causality Problem, *arXiv:1910.13413.*

Johnston, R., Jones, K., and Manley, D. (2018), Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour, *Qual Quant.*, 52(4), pp. 1957-1976.

Lepetit, F., Le Guenedal, T., Ben Slimane, M., Cherief, A., Mortier, V., Sekine, T. and Stagnol, L. (2021), The Recent Performance of ESG Investing, the Covid-19 Catalyst and the Biden Effect, *Available at SSRN*: http://ssrn.com/abstract=3946483.

Lundberg, S M. and Su-In, L. (2017), A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems 30 (NIPS 2017).*

LUNDBERG, S M., ERION, G G. and SU-IN, L. (2019), Consistent Individualized Feature Attribution for Tree Ensembles, *arXiv:1802.03888*.

O'BRIEN, R.M. (2007), A Caution Regarding Rules of Thumb for Variance Inflation Factors, *Quality & Quantity*, 41(5), pp. 673-690.

PROBST, P., BOULESTEIX, A-L. (2017), To Tune or Not To Tune the Number of Trees in Random Forest?, *arXiv:1705.05654[stat.ML]*.

PROBST, P., BISCHL, B., and BOULESTEIX, A-L. (2019), Tunability: Importance of Hyperparameters of Machine Learning Algorithms, *arXiv:1802.09596v3[stat.ML]*.

RONCALLI, T., CHERIEF, A., KARRAY-MEZIOU, F. and REGNAULT, M. (2021), Liquidity Stress Testing in Asset Management Part 2. Modeling the Asset Liquidity Risk, *Available at SSRN*: http://ssrn.com/abstract=3849496.

SIMONIAN, J., WU, C., ITANO, D. and NARAYANAM, V. (2019), A Machine Learning Approach to Risk Factors: A Case Study Using the Fama-French-Carhart Model, *The Journal of Financial Data Science*, 4(2).

ZAHARIA, M., CHEN, A., DAVIDSON, A., GHODSI, A., ANN HONG, S., KONWINSKI, A., MURCHING, S., NYKODYM, T., OGILVIE, P., PARKHE, M., XIE, F., ZUMAR, C. and DATABRICKS INC. (2020), Accelerating the Machine Learning Lifecycle with MLflow, *Databricks Inc.*.

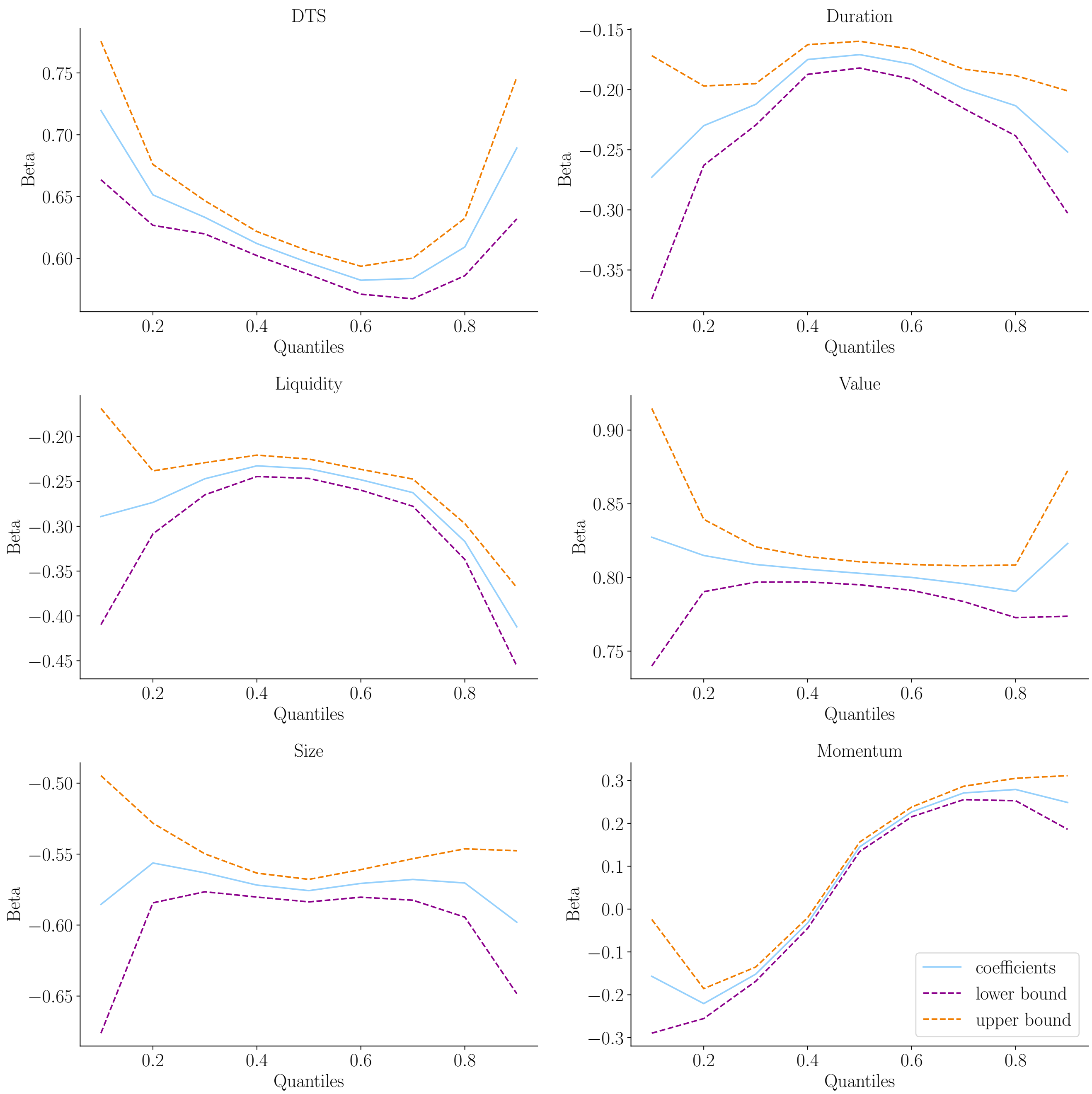
ZHANG, H., NETTLETON, D. and ZHU, Z. (2019), Regression-Enhanced Random Forests, *arXiv:1904.10416*.

ZHOU, ZH. (2009), Ensemble Learning, *Encyclopedia of Biometrics. Springer US*, pp. 270-273.

A Appendix

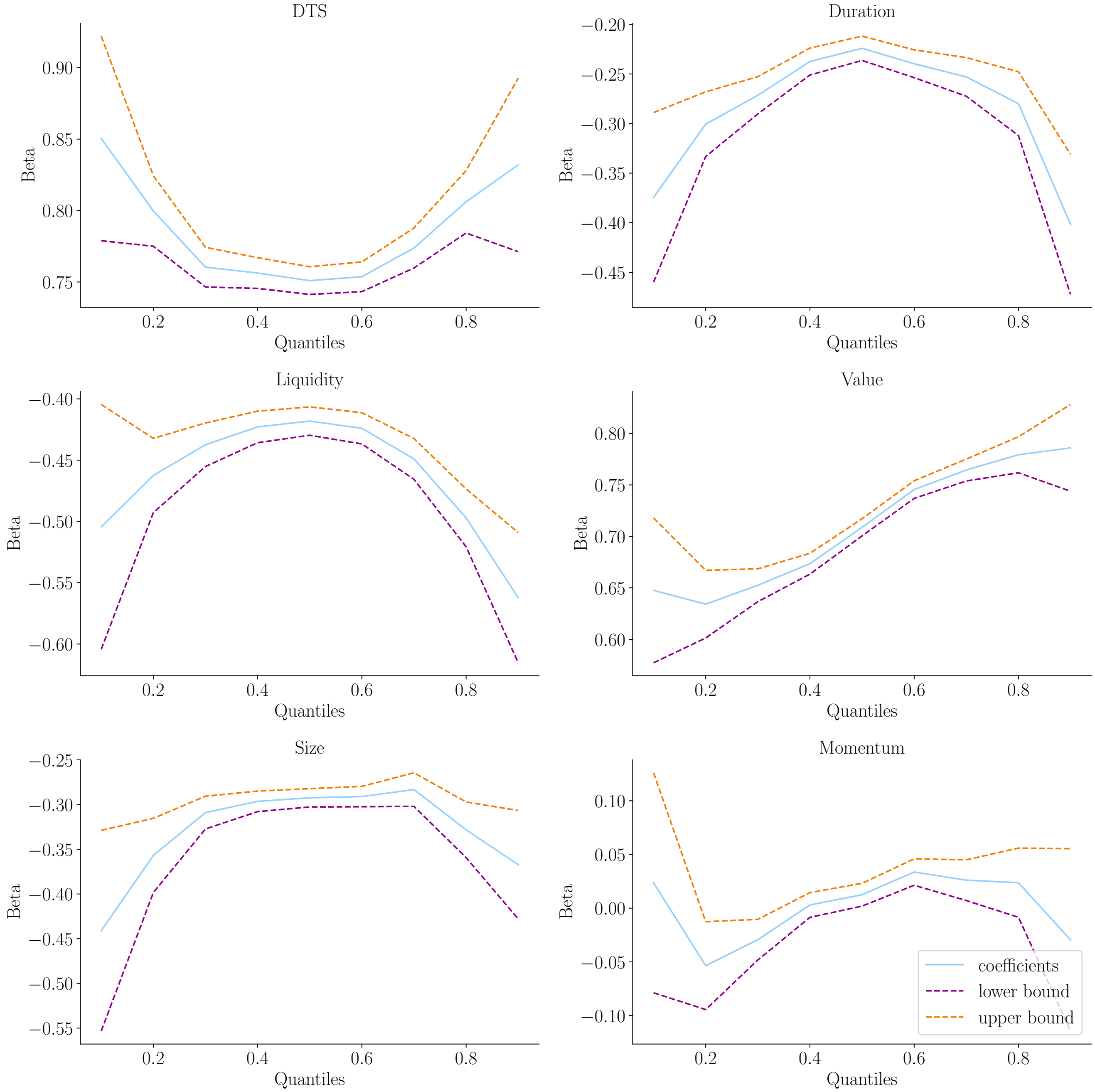
A.1 Motivation

Figure 18: Quantile regressions coefficients for bonds denominated in EUR (2011-2021)



Source: Amundi Institute Quantitative Research

Figure 19: Quantile regressions coefficients for bonds denominated in USD (2011-2021)



Source: Amundi Institute Quantitative Research

A.2 Statistical results

A.2.1 Customizing the hyper parameters

The objective is to adjust the hyper-parameters to improve the performance of the prediction but also to avoid over-fitting. As mentioned in Section 3.3, the tree-based algorithm can be complex given the number of parameters to set before performing the regression. We choose a cross-validated grid search to find the optimal parameters. In this section, we comment on the parameters that we seek to tune and justify our choice.

Find the optimal number of trees There is a debate on whether or not we should consider the number of trees as a parameter or a large and constant number for Random Forest algorithms. Probst and Boulesteix (2017) attempt to answer this question for classification and regression problems involving for the latter a mixed result where only two over four error scores decreased as the number of trees has increased for Random Forest regressions. Before that, Brieman (2001) proved that *"the generalization error converges to a limit"* when rising the number of trees but did not quantify it. As far as Gradient Boosting is concerned, it is much more complex since the trees are dependent, implying that the risk of over-fitting is high after training a large number of trees. That's why Friedman (2001) has introduced the regularization parameter that we also set as a hyper-parameter, that we set between the 0.001 and 0.15 values. For our study, we seek to vary the number of trees from 100 to 1000 for Random Forest algorithm and between 100 to 500 for Gradient Boosting algorithm.

Maximum number of depth in a tree It is important to note that the higher the degree of the tree, the better the data information capture. But, a large depth in the tree also means a high variance, which can lead to overfitting. In the study, we have set the maximum depth in a tree between 3 and 7.

Maximum number of features in a node We use different metrics to find the optimal maximum number of features to consider at each node in order to split the data. According to Hastie *et al.* (2009), the recommendable number of features is p/3 for a tree-based regression with p corresponding to the total number of features. Brieman (2001) suggested to use $\log 2\,(M+1)$ where M is the number of features in his original paper. Finally, we allowed the algorithm to select $J-1$ features at the most, where J is the number of features because selecting all the features would increase the correlation between the trees.

Minimum number of data in a node We can also choose hyper-parameters such as the minimum number of data to consider in a node. Thus, if the minimum number of data to split is not reached, then the tree can be pruned, and the node becomes a leaf node. Moreover, in our case, a node is split only if the impurity decrease is above a constant value. The weighted impurity decrease is described as follows:

$$WID_k = \frac{N_k}{N} \cdot \left(R_k - \frac{N_k^{y|x \geq \tau}}{N_k} \cdot R_k^{y|x \geq \tau} - \frac{N_k^{y|x < \tau}}{N_k} \cdot R_k^{y|x < \tau} \right) \tag{2}$$

where N represents the total number of data, N_k is the number of observations in the k node, $N_k^{y|x<\tau}$ is the number of observations in the first branch, and $N_k^{y|x\geq\tau}$ is the number of observations in the second branch where R_k is defined as the impurity computed in Algorithm 1.

A.2.2 Statistics to measure the error of the models

Let N be the number of observations and J the number of features of the model. We make use of the following statistics:

- MAE is the Mean Absolute Error:

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}$$

- MSE is the Mean Squared Error:

$$MSE = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}$$

- MEDAE is the Median Absolute Error:

$$MEDAE = med\left\{|y_i - \hat{y}_i|\right\}$$

- MEDSE is the Median Squared Error:

$$MEDSE = med\left\{(y_i - \hat{y}_i)^2\right\}$$

- SSR is the Sum of the Squares of Residuals:

$$SSR = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- SST is the Sum of the Squares total:

$$SST = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

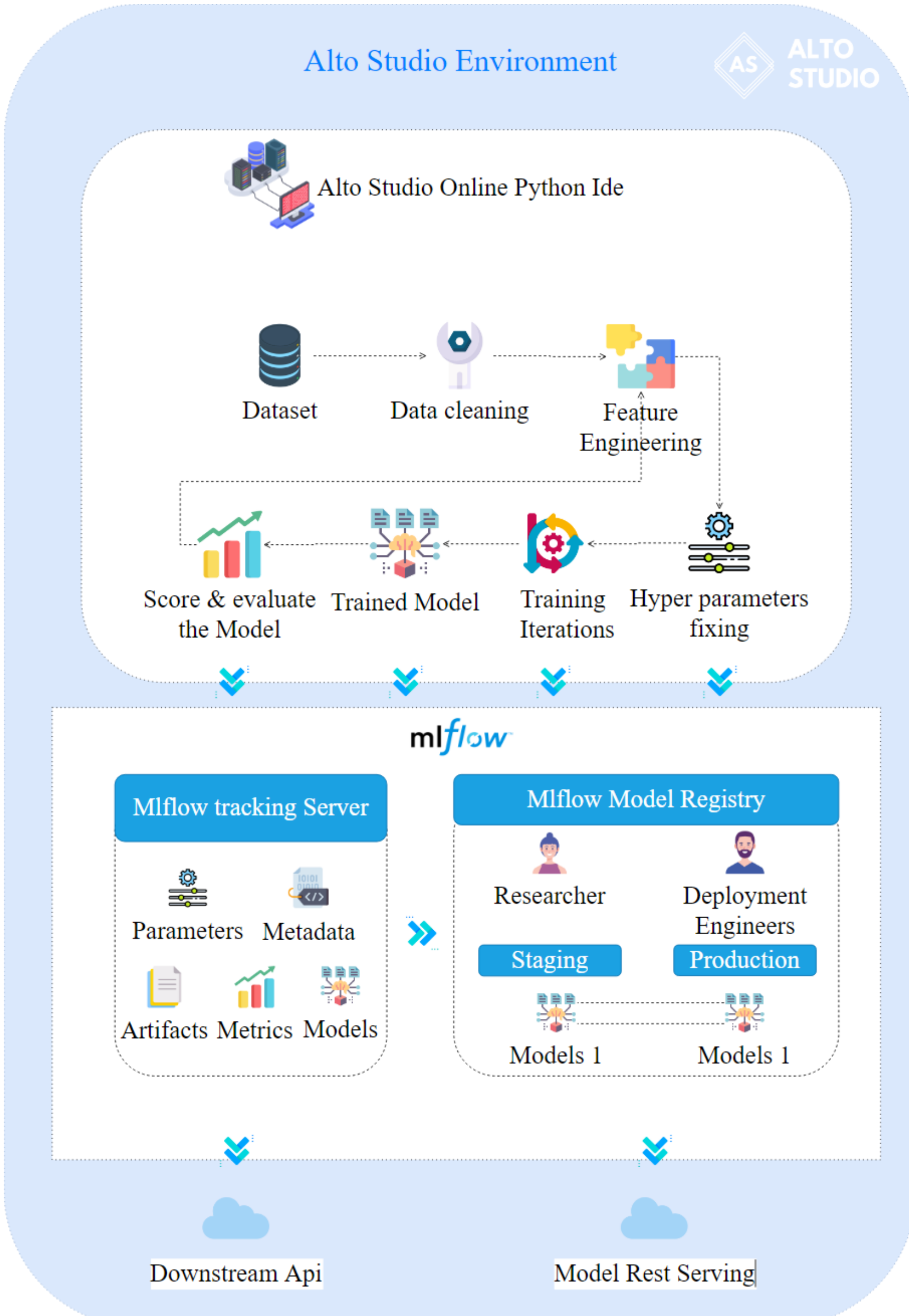
- R^2 is the coefficient of determination:

$$R^2 = 1 - \frac{SSR}{SST}$$

- R^2_{Adj} is the adjusted R^2:

$$R^2_{Adj} = 1 - \left(1 - R^2\right) \times \frac{(N-1)}{(N-J-1)}$$

A.3 Alto Studio Framework



Source: Le Lab - Amundi Technology and Amundi Institute Quantitative Research

A.4 Comparing the performance of the non-linear vs. linear models

A.4.1 Bonds denominated in EUR currency

Figure 20: Average rank for bonds denominated in EUR (2011-2014)

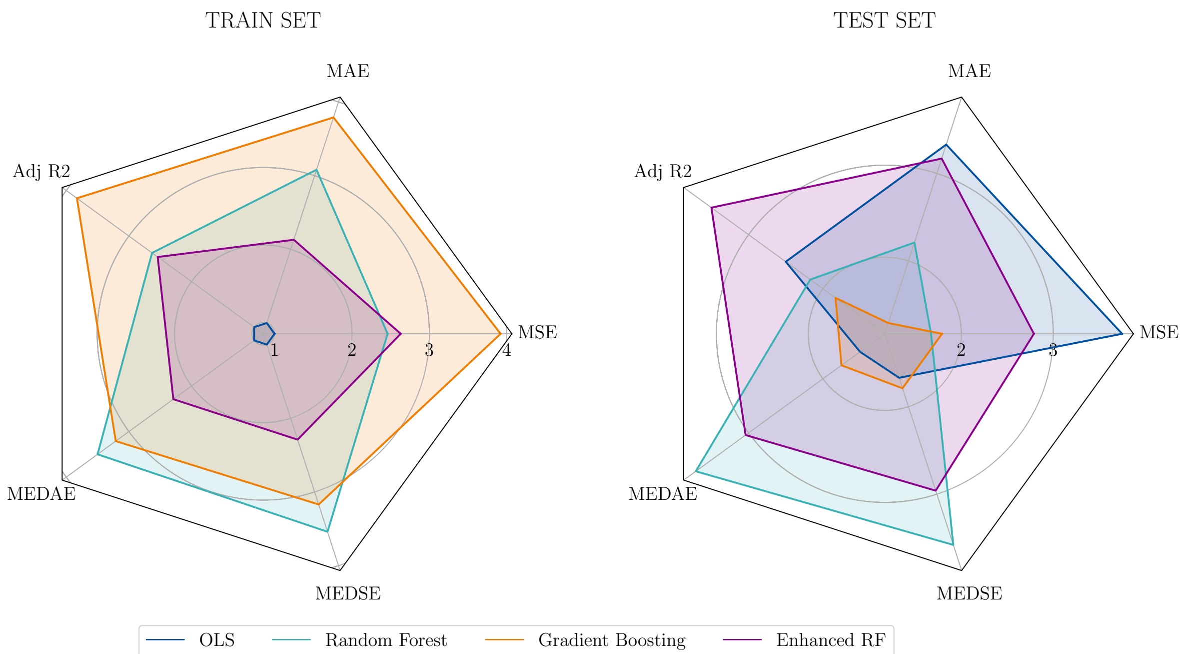


Figure 21: Median rank for bonds denominated in EUR (2011-2014)

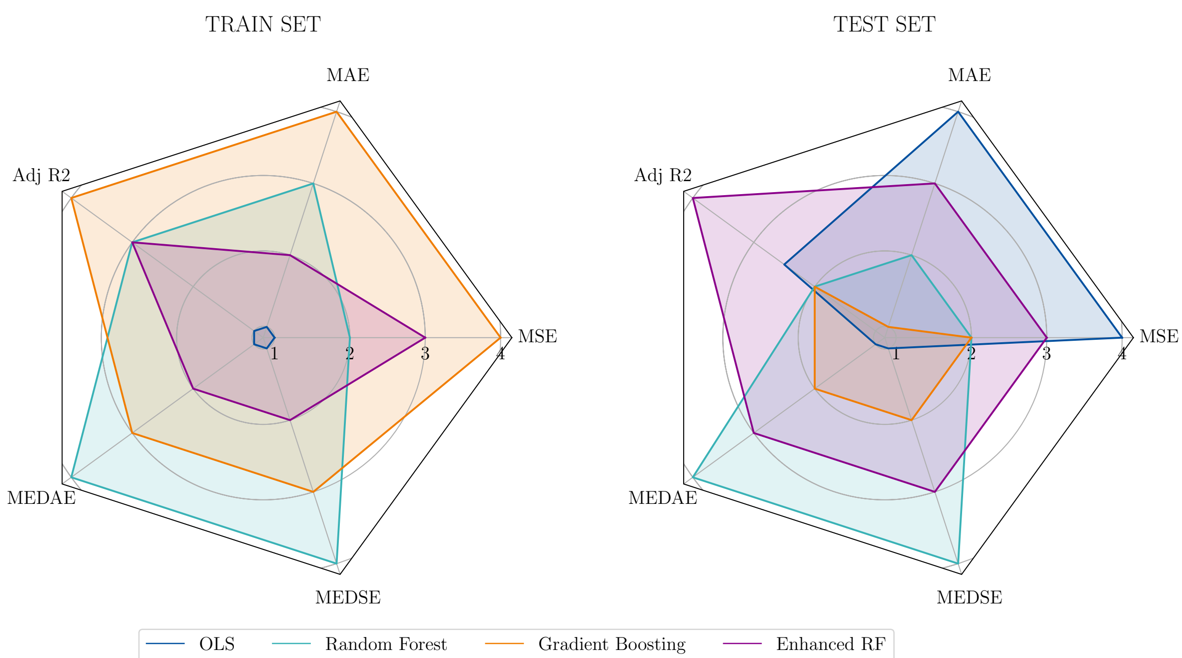


Figure 22: Average rank for bonds denominated in EUR (2015-2021)

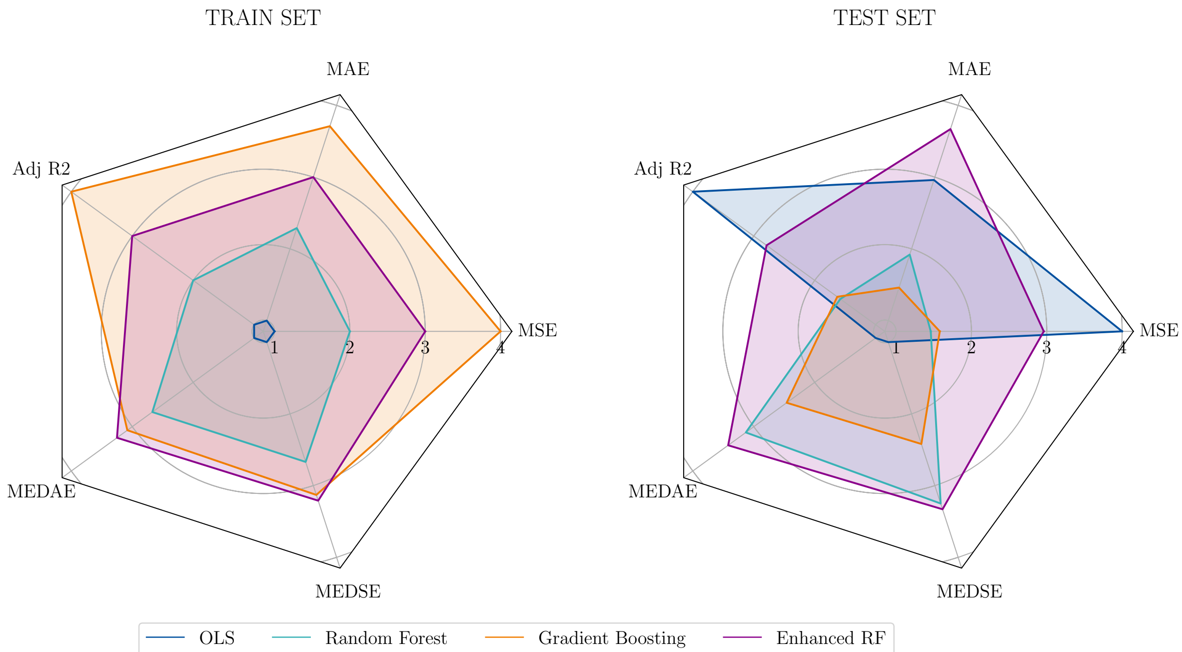
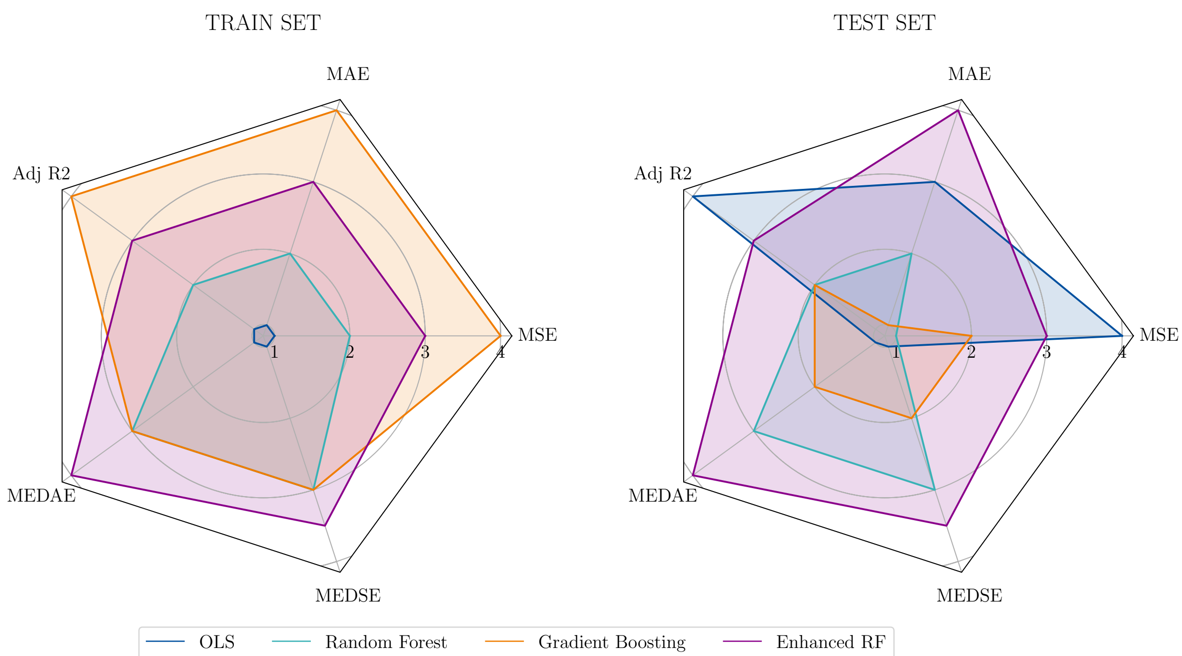


Figure 23: Median rank for bonds denominated in EUR (2015-2021)



Source: Amundi Institute Quantitative Research

A.4.2 Bonds denominated in USD currency

Figure 24: Average rank for bonds denominated in USD (2011-2014)

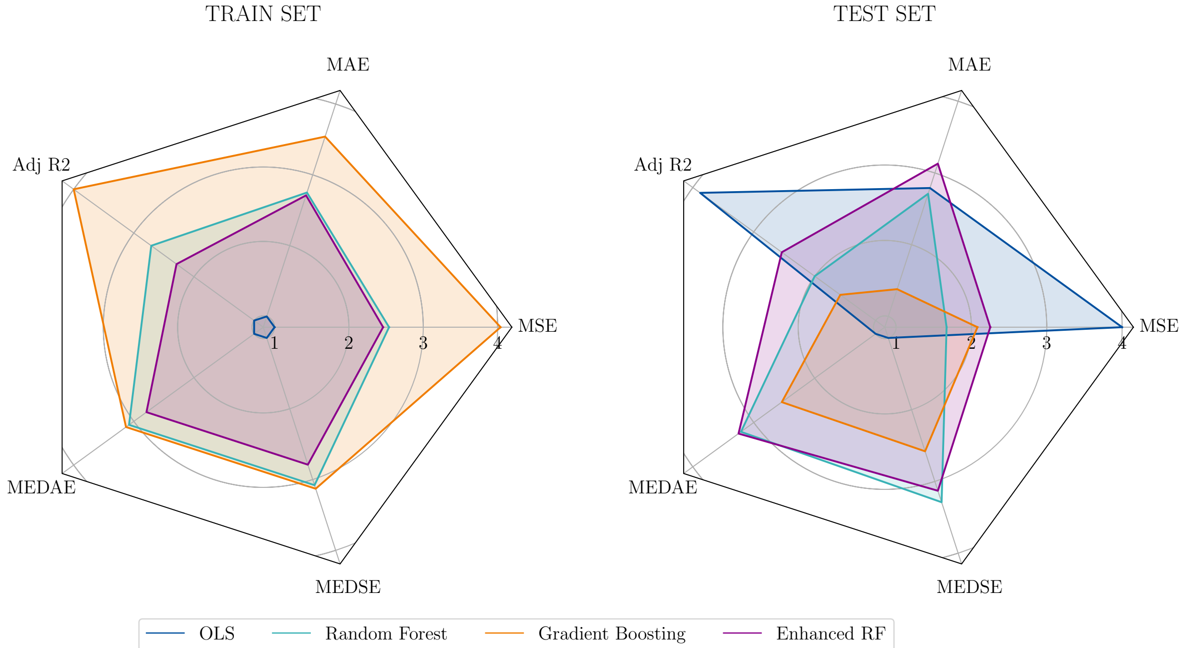


Figure 25: Median rank for bonds denominated in USD (2011-2014)

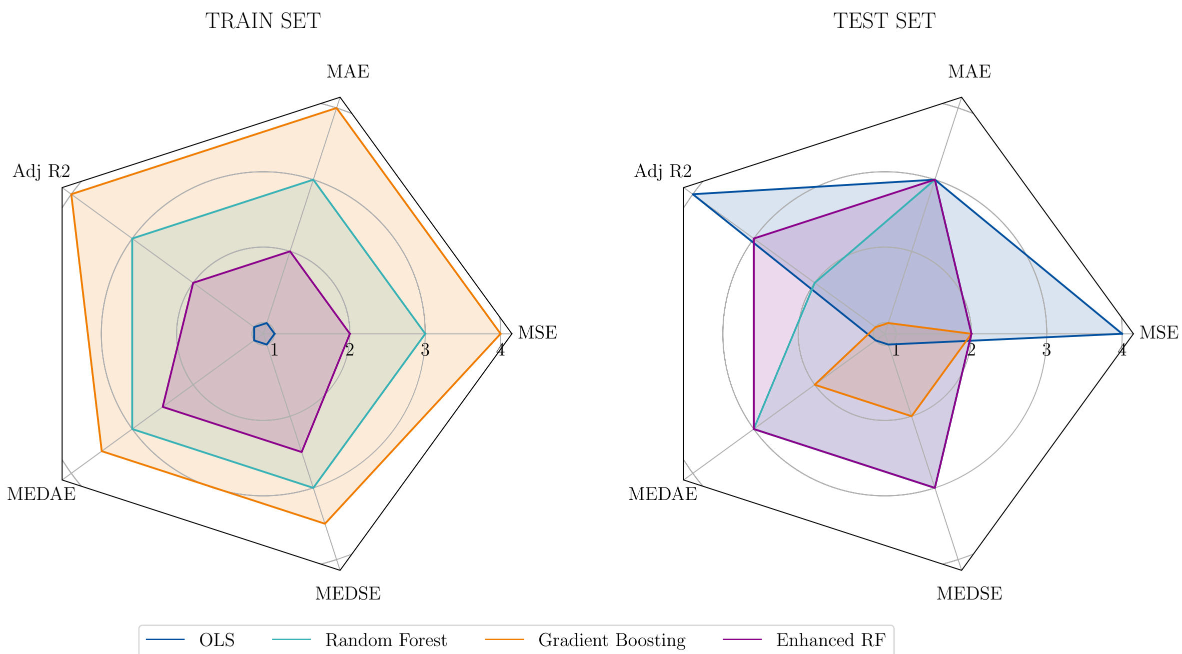


Figure 26: Average rank for bonds denominated in USD (2015-2021)

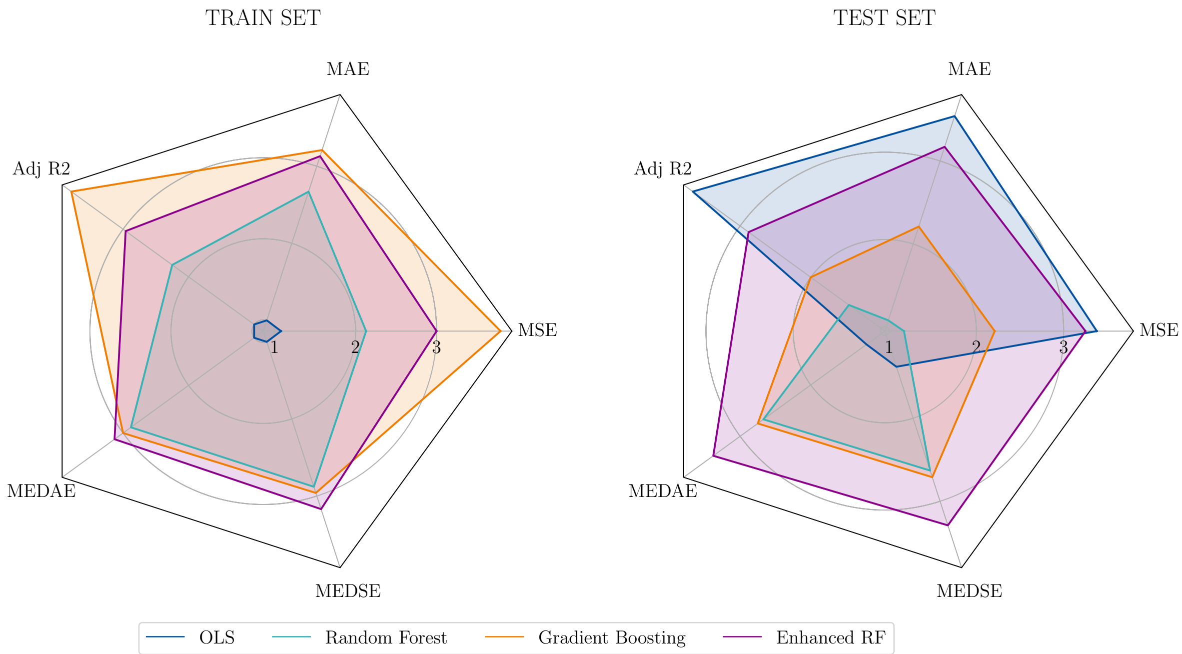
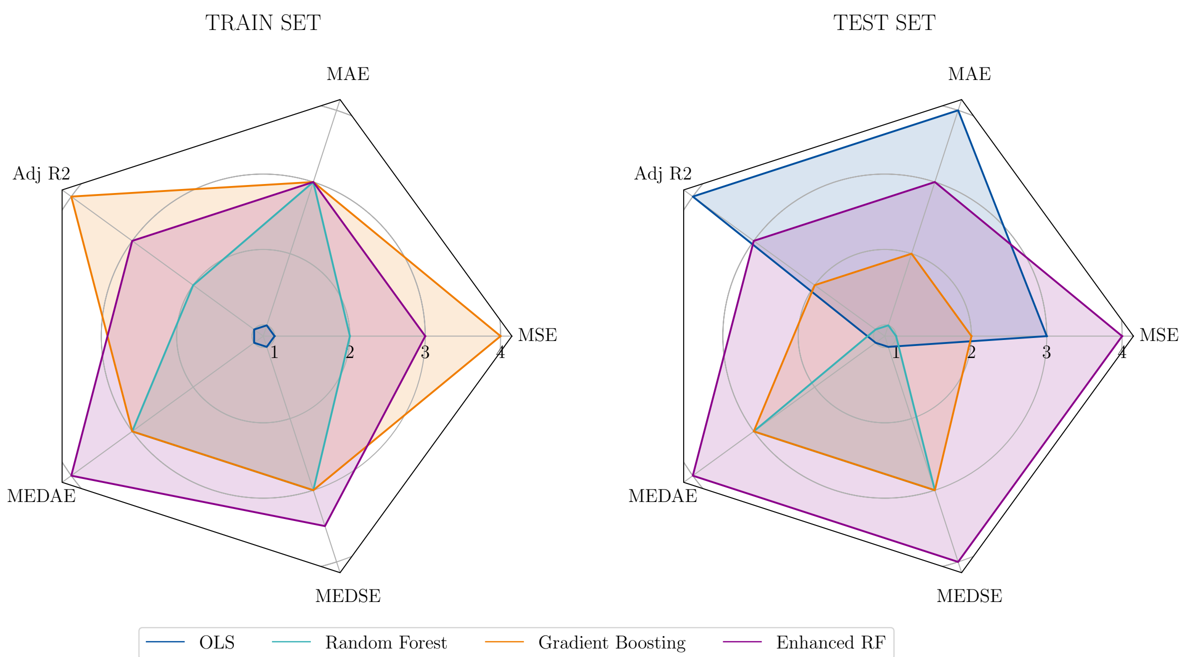


Figure 27: Median rank for bonds denominated in USD (2015-2021)



Source: Amundi Institute Quantitative Research

A.5 Arithmetic improvement in $AdjR^2$ for USD factors

Table 8: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for USD data (2011-2021)

Split method	Model	TRAIN SET		TEST SET	
		3F→6F[11]	3F→7F[12]	3F→6F	3F→7F
Method 1	OLS	10.44	15.00	8.32	15.23
	RF	3.44	3.32	8.16	12.59
	GB	2.78	4.77	9.54	13.48
	ERF	4.61	5.48	17.17	21.38
Method 2	OLS	10.13	14.46	7.25	16.17
	RF	3.39	3.69	9.73	13.21
	GB	5.36	5.40	14.91	18.87
	ERF	3.65	5.67	19.71	24.16
Method 3	OLS	8.38	13.52	7.73	18.86
	RF	5.14	6.75	14.08	13.70
	GB	9.81	11.72	9.91	14.22
	ERF	9.64	12.13	14.20	13.70

Source: Amundi Institute Quantitative Research

Table 9: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for USD data (2011-2014)

Split method	Model	TRAIN SET		TEST SET	
		3F→6F	3F→7F	3F→6F	3F→7F
Method 1	OLS	11.72	14.80	10.97	24.75
	RF	2.98	3.04	4.87	8.67
	GB	6.34	7.23	6.75	12.22
	ERF	3.55	2.97	12.64	17.24
Method 2	OLS	12.51	15.66	15.54	29.15
	RF	2.02	2.78	5.45	10.53
	GB	8.38	9.65	3.51	6.72
	ERF	3.53	3.95	15.12	19.14
Method 3	OLS	6.52	11.93	20.84	33.11
	RF	3.05	4.41	8.79	17.62
	GB	13.49	14.26	18.76	32.57
	ERF	1.31	3.92	14.33	23.23

Source: Amundi Institute Quantitative Research

Table 10: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for USD data (2015-2021)

Split method	Model	TRAIN SET		TEST SET	
		3F→6F	3F→7F	3F→6F	3F→7F
Method 1	OLS	8.23	13.59	7.09	12.23
	RF	5.96	6.66	0.48	1.86
	GB	15.02	16.20	11.03	12.23
	ERF	3.76	5.39	20.72	21.03
Method 2	OLS	8.45	14.00	6.94	14.23
	RF	8.33	8.40	2.43	1.43
	GB	16.79	17.48	10.40	12.30
	ERF	3.08	2.72	25.07	24.54
Method 3	OLS	9.75	14.15	14.10	12.69
	RF	3.71	6.74	10.97	6.55
	GB	7.14	8.79	14.05	1.33
	ERF	5.87	6.21	19.30	16.02

Source: Amundi Institute Quantitative Research

Table 11: Arithmetic improvement in R^2_{Adj} for the train and test set in percent for USD data (2015-2021)

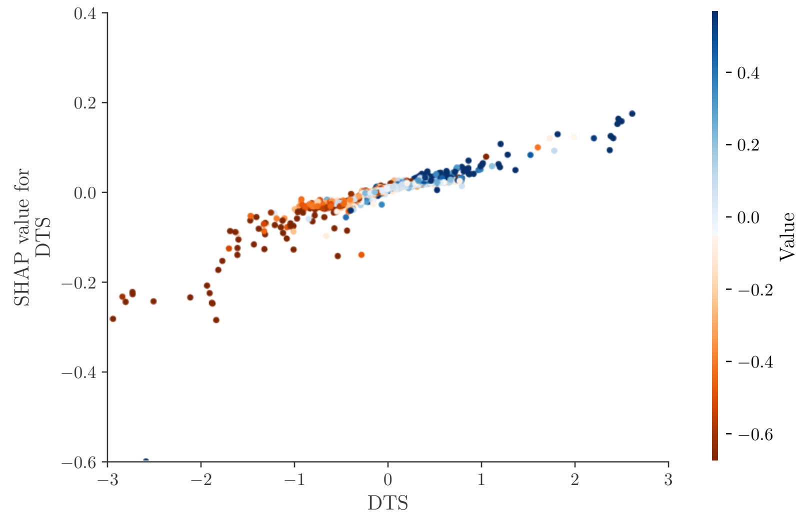
Split method	Model	TRAIN SET 3F→7F			TEST SET 3F→7F		
		E	S	G	E	S	G
Method 1	OLS	10.92	8.38	9.82	8.01	6.88	7.05
	RF	6.31	6.56	6.51	2.03	0.79	−1.42
	GB	15.55	14.71	15.75	9.89	8.78	7.12
	ERF	4.65	4.81	5.72	20.48	20.39	24.52
Method 2	OLS	11.21	8.63	10.19	8.12	6.78	7.34
	RF	7.43	7.55	7.65	0.43	0.38	−0.46
	GB	17.64	17.14	18.35	12.49	10.22	7.53
	ERF	3.70	3.65	4.50	23.96	24.27	25.52
Method 3	OLS	12.08	10.28	12.17	18.57	14.46	16.06
	RF	6.62	5.37	6.55	8.43	9.86	13.39
	GB	8.09	7.62	8.87	14.01	10.09	12.86
	ERF	6.07	6.13	5.65	20.16	21.18	22.10

Source: Amundi Institute Quantitative Research

B Market Interpretation

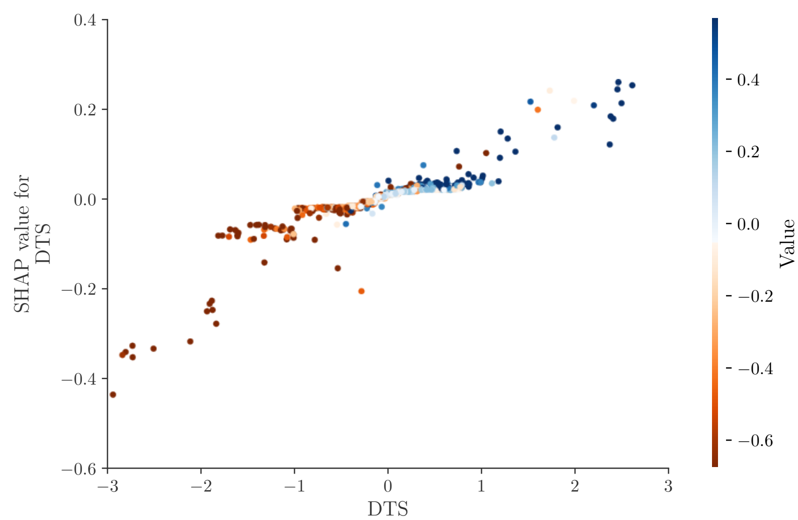
B.1 Dependence figures for EUR factors

Figure 28: Dependence to the output variable for the EUR-denominated DTS factor using Random Forest algorithm (2015-2021)



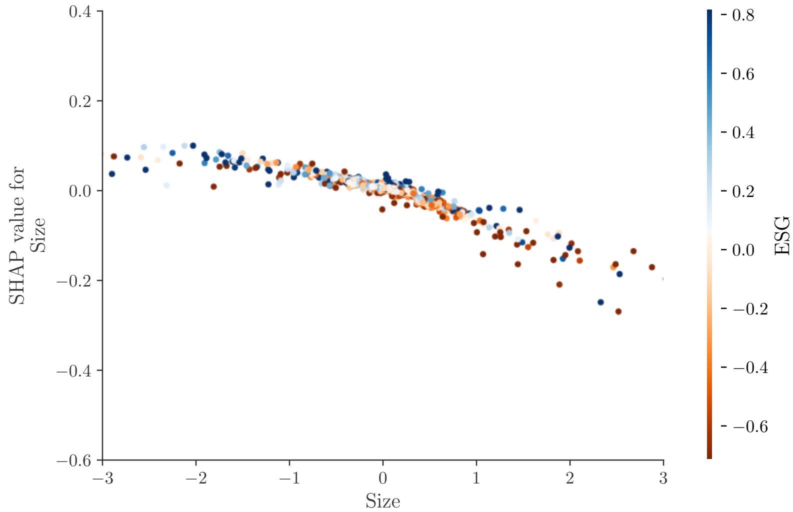
Source: Amundi Institute Quantitative Research

Figure 29: Dependence to the output variable for the EUR-denominated DTS factor using Gradient Boosting algorithm (2015-2021)



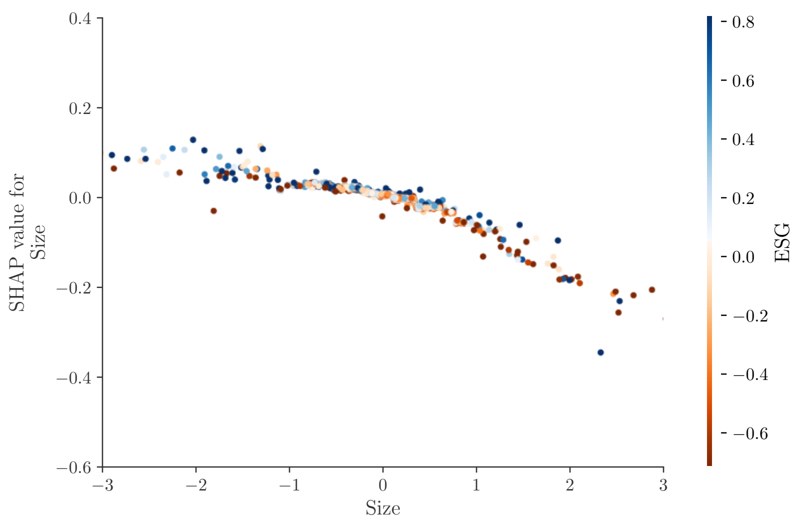
Source: Amundi Institute Quantitative Research

Figure 30: Dependence to the output variable for the EUR-denominated Size factor using Random Forest algorithm (2015-2021)



Source: Amundi Institute Quantitative Research

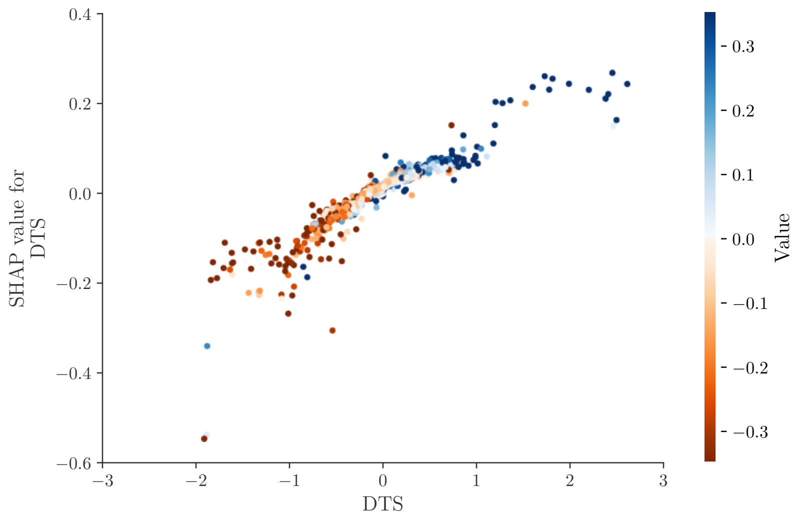
Figure 31: Dependence to the output variable for the EUR-denominated Size factor using Gradient Boosting algorithm (2015-2021)



Source: Amundi Institute Quantitative Research

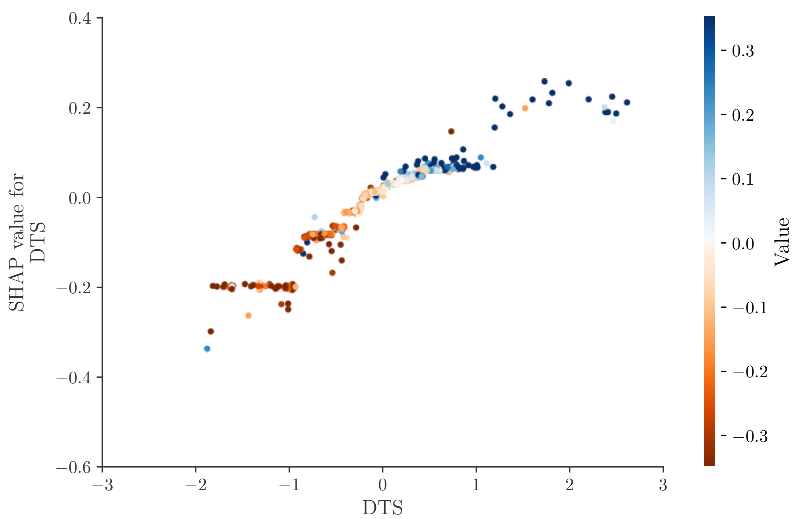
B.2 Dependence figures for USD factors

Figure 32: Dependence to the output variable for the USD-denominated DTS factor using Random Forest algorithm (2015-2021)



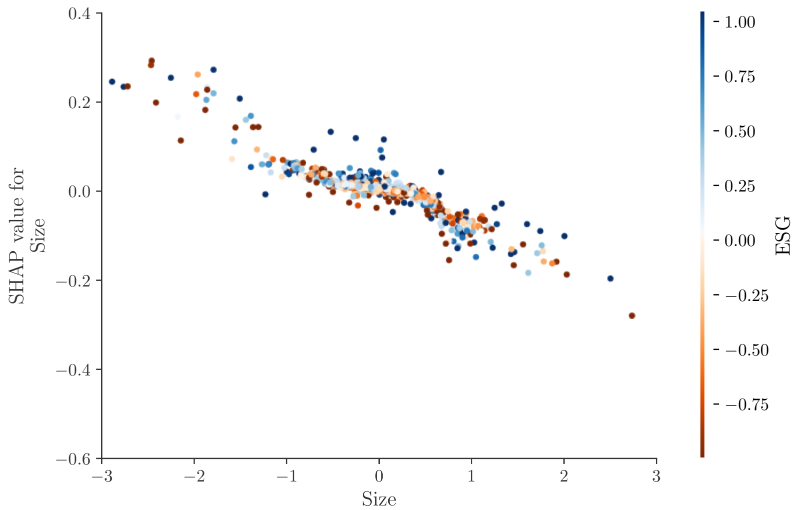
Source: Amundi Institute Quantitative Research

Figure 33: Dependence to the output variable for the USD-denominated DTS factor using Gradient Boosting algorithm (2015-2021)



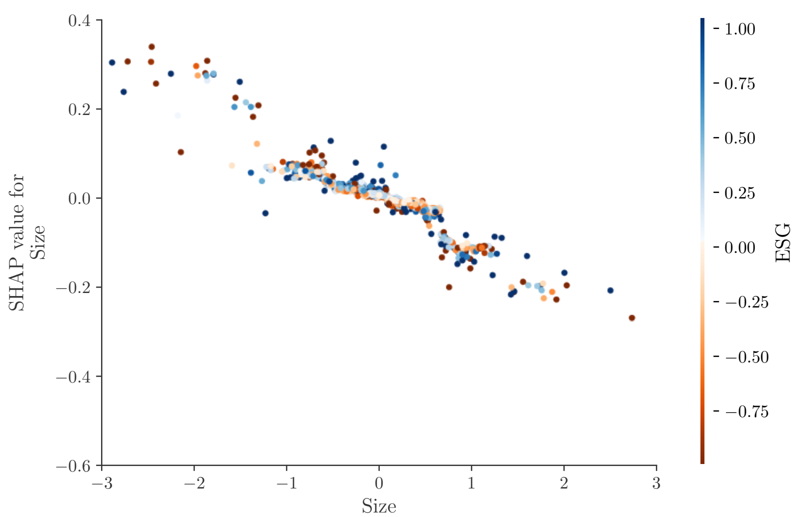
Source: Amundi Institute Quantitative Research

Figure 34: Dependence to the output variable for the USD-denominated Size factor using Random Forest algorithm (2015-2021)



Source: Amundi Institute Quantitative Research

Figure 35: Dependence to the output variable for the USD-denominated Size factor using Gradient Boosting algorithm (2015-2021)



Source: Amundi Institute Quantitative Research

Chief Editors

Pascal BLANQUÉ
Chairman Amundi Institute

Vincent MORTIER
Group Chief Investment Officer

Amundi ASSET MANAGEMENT | Institute

Working Paper
July 2022

Important Information

Date of first use: 1st July 2022.

Document issued by Amundi Asset Management, "société par actions simplifiée"- SAS with a capital of €1,143,615,555 - Portfolio manager regulated by the AMF under number GP04000036 – Head office: 91-93 boulevard Pasteur – 75015 Paris – France – 437 574 452 RCS Paris – www.amundi.com

Photo credit: iStock by Getty Images - monsitj