Olli Aaron Kiverä

On Issuance: What Matters, What Doesn’t and Where Green Fits

School of Accounting & Finance
Finance
On Issuance: What Matters, What Doesn’t and Where Green Fits

Vaasa 2026

UNIVERSITY OF VAASA
School of Accounting & Finance
Author: Olli Aaron Kiverä
Title of the thesis: On Issuance: What Matters, What Doesn't and Where Green Fits
Degree: Master of Science in Economics and Business Administration
Degree Programme: Finance
Supervisor: Klaus Grobys
Year: 2026 **Pages: 99**

ABSTRACT:

Green bond pricing and the greenium remains unsettled, and most evidence comes from secondary markets where liquidity and trading frictions complicate interpretation. This thesis focuses on the primary market and asks two practical questions: whether the green label is associated with an issuance G-spread differential once standard spread determinants are controlled for, and whether machine-learning models improve out-of-sample prediction. The dataset consists of corporate bonds issued between 2023 and 2025 and totals 3654 bonds. Issuance G-spread is used as the dependent variable and controls include bond characteristics and market conditions, such as broad credit spreads and volatility proxies. Issuance G-spreads are estimated with a nested OLS ladder to track how the green coefficient behaves as control blocks are introduced. Predictive performance is benchmarked across OLS, LASSO, and Random Forest and evaluated on a held-out test set, across 100 repeated random splits and in an out-of-time setting that trains on earlier years and tests on a later period. The results show that green labelled issues are associated with a spread difference of about +16.4 bps at the representative spread level. Random Forest provides a large improvement in predictive accuracy relative to linear benchmarks. On the held-out test set, RMSE decreases from 0.251 to 0.152 and test R^2 increases from 0.858 to 0.948. In out-of-time evaluation, performance falls for all models but Random Forest remains most accurate. State splits indicate that implied green differential is larger in tight conditions at 28–31 bps and smaller and less precise in stressed regimes at 9–10 bps. Issuance spreads are dominated by conventional drivers such as credit quality, deal structure, and issuance-time market conditions. The green label adds secondary but sometimes economically meaningful information, and the magnitude is regime dependent.

KEYWORDS: Corporate Bond, Greenium, OLS, Random Forest, Machine Learning, Primary Market, Issuance

Contents

Figures

Tables

1 Introduction

1.1 Purpose of the Study

In this thesis the objective is to investigate what drives corporate bond pricing at issuance and if machine learning models can predict the issuance G-spreads better than linear econometric models. The focus is on primary bond market, which is where new bonds are issued and the initial yield is set through underwriting and bookbuilding processes. The issuance G-Spread is the difference between issue yield and benchmark rate, it measures the cost of debt financing and reflects how investors evaluate for example risk, liquidity, and market conditions that are available when the bond is priced (Benveniste & Spindt, 1989; Cornelli & Goldreich, 2003). Understanding this process is important because it affects the issuer's cost of capital and gives an insight on how information is transmitted in financial markets.

Historically, the fixed-income market (that bonds are part of) has changed due to environmental and social aspects becoming more important and as such are more included in capital allocation. For this reason, this thesis studies the green bond market, which has grown rapidly since the European Investment Bank issued the first Climate Awareness Bond back in 2007. The market for green bonds is also supported by policy frameworks like the Paris Agreement and the European Union Green Deal, and by investor demand for sustainable instruments (Baker et al., 2018).

While a lot of studies have focused on the pricing dynamics of the green bond market, they are still unclear. Some studies show that green bonds trade at a premium, which is referred to as a "greenium", when they are compared with similar non-green bonds (Fatica et al., 2021; Flammer, 2021). These studies show that investors may accept lower yields for bonds that are labelled as green, perhaps due to environmental benefits or transparency. Other studies don't find a greenium when credit risk and liquidity among other factors are controlled for (Larcker & Watts, 2020; Pástor et al., 2021). Most prior research focuses on secondary-market pricing, but the primary market, where information asymmetries might be more impactful, are less researched.

There are a lot of variables that determine issuance G-spreads, and those are further affected by both risk factors and informational elements. Information asymmetry theory suggests that issuers may have private information which the investors do not. To fight this, investors commonly demand compensation for this kind of uncertainty, which results in a higher spread if information is incomplete or not available (Akerlof, 1970; Myers & Majluf, 1984).

Signaling theory then argues that certification, external reviews, or adhering to standards can act as credible signals that reduce required yields (Spence, 1973; Ross, 1977). In the case of green bonds, there are green bond standards that must be fulfilled for a bond to be labelled “green”. Investor-preference models suggest that investors gain non-financial utility from holding environmentally positive assets and then accept lower returns (Heinkel et al., 2001; Pástor et al., 2021). This shows that bond pricing can depend not only on financial risk, but also on information disclosure and preferences.

In the bond literature, traditional econometric models such as ordinary least squares (OLS) are commonly used, but G-spreads can be affected by nonlinearities and interaction effects between bond characteristics (Hastie et al., 2009). Due to this a machine learning model, Random Forest, is incorporated. The motivation for this is simple, recent studies have shown that machine learning models outperform traditional models in predicting asset returns, credit spreads, and default probabilities (Varian, 2014).

The main objective for this thesis is to identify the main determinants of issuance stage bond G-spreads and to examine if the “green label” affects pricing of bonds. After this, I introduce a machine learning model to test if it can improve predictive performance and if its diagnostics provide additional information. By combining the beforementioned theories and modern modelling of data, the aim is to improve understanding of how green label and advanced quantitative techniques can influence the pricing of bonds in the primary market.

In this study, green bonds are classified as green if they are given a “green” tag by LSEG. LSEG assigns green bond tags based on issuer labelling in public documents, review of the issuer’s sustainability framework and verification of the use of proceeds to ensure that the classification is justified (London Stock Exchange Group, 2024). Below are the specific inclusion and exclusion criteria.

Inclusions	**Exclusions**
Currency: USD	Financial sector and non-USD issues
Issuer Sector: non-financial corporates	Observations with missing credit rating
Timing: 5 Jan 2023 – 21 Aug 2025	Rare SIC categories with 2 or less observations
Credit rating: all rating levels included	Total excluded: 27 observations
Callability: only callable	

1.2 Research Problem and Hypotheses

As the research on corporate bonds and their pricing has concentrated on secondary-market yields and often analyzes spreads after issuance, the primary market has received less attention. In the green bond context empirical evidence is mixed: some studies report a “greenium” (Bachelet, et al., 2019), while others report no significant differences once standard determinants are controlled for (Larcker & Watts, 2020; Flammer, 2021). More recently, Caramichael and Rapp (2024) researched green bonds in the primary market and document that green bonds price at lower issuance G-spreads and link this advantage to demand pressure at issuance (Caramichael & Rapp, 2024).

What remains less understood is whether the green label carries incremental information in more recent issuance windows and if the green label contributes beyond standard determinants out-of-sample and out-of-time. This thesis addresses these gaps by studying USD-denominated non-financial corporate bonds issued between 5 January

2023 and 21 August 2025, using issuance G-spreads to focus on primary-market pricing. This thesis further incorporates 3 complementary methods: a nested OLS “ladder”, LASSO regression, and a Random Forest model. Doing so allows to track how the estimated green coefficient evolves as controls are introduced and if any green label signal improves out-of-sample and out-of-time predictions, when more complex interactions and nonlinearities are allowed.

Specifically, this thesis addresses the following research questions:

Does a Random Forest model predict issuance G-spreads more accurately than a nested OLS ladder out-of-sample?

Which variables, including bond characteristics, credit quality, sector composition, and issuance-time market conditions, are most important in explaining issuance G-spreads according to model-based importance measures?

Does the inclusion of a green label variable provide explanatory and predictive content, and is the green label associated with a statistically significant G-spread differential once conventional determinants are controlled for?

Based on these questions, three hypotheses are formulated:

H1: A Random Forest model predicts corporate bond issuance G-spreads more accurately than ordinary least squares out-of-sample.

H2: Credit quality, maturity, issue size, and issuance-time market conditions are among the most important determinants of issuance G-spreads.

H3: The green label is associated with a statistically significant conditional issuance G-spread differential after controlling for bond characteristics, credit quality, sector composition, and issuance-time market conditions.

1.3 Contribution

This thesis contributes to the green bond literature by focusing explicitly on primary market issuance pricing. Building on the results of Caramichael and Rapp (2024), who showed that green bonds price at lower issuance spreads, the aim is to show if a green bond differential is also present in a later US market sample.

Methodologically the contribution is separating two questions that are often conflated in the greenium debate: conditional association and incremental predictability. The OLS ladder adds control blocks stepwise to show transparently how the green coefficient changes as bond characteristics, credit quality and issuance-time market conditions are introduced. At the same time the Random Forest benchmark tests if the green label increases out-of-sample predictive accuracy, once the same determinants are added to the model. To show that results are robust, the models are also evaluated out-of-time. Heterogeneity is also addressed by dissecting the green bond issuances based on credit ratings, market states (tight vs. stressed).

Last, the aim is to offer a unified interpretation that helps understand the mixed findings in the greenium debate. Basically, do the standard determinants explain most variation, or does the green label add value after the fundamentals are already included.

1.4 Structure of the Thesis

Chapter 1 motivates the research problem, defines the research questions and hypotheses. Chapter 2 sets out the theoretical mechanisms relevant for issuance pricing and motivates the use of machine-learning models in this setting. Chapter 3 reviews the empirical evidence on corporate bond spreads, green bond premia, and predictive modelling in finance, and positions the study relative to existing work. Chapter 4 describes the empirical framework: the OLS ladder used for conditional inference on the green label and the Random Forest model used for prediction. Chapter 5 describes the data sources, sample construction, and variable definitions. Chapter 6 presents the results from approaches and compares what the models imply. Chapter 7 concludes and discusses practical implications, limitations, and directions for future research.

2 Theoretical Background

This section includes the theoretical background for green and non-green bonds and their pricing. It also includes the background for machine learning models, the Random Forest. Since green bonds are simply just labelled corporate bonds, the issuance pricing of normal corporate bonds is the starting point. Following this, green bonds are introduced.

2.1 Bonds

Corporate bonds are debt instruments that are issued by companies to raise capital. They provide regular interest payments, which is known as the coupon, and then the principal is repaid at maturity. This offers issuers a reliable funding source and investors gain anticipated returns (Fabozzi, 2012). Although considered stable and fixed-income, bonds carry risks such as inflation, interest rate changes and issuers creditworthiness. Hull (2018) shows that bonds are diverse and come in different forms such as treasury, municipal, and corporate bonds, furthermore different types of bonds offer different risk and return profiles. To start off, bond fundamentals, the risks associated, pricing mechanism, and the central role of credit rating agencies are examined.

2.1.1 Key Characteristics

Bonds are financial instruments with defined characteristics that determine how they are structured, priced, and perceived by investors. These characteristics define the rights and obligations of issuers and investors (bondholders) and form the bond's risk-return profile. The par value is the amount repaid at maturity and provides the basis for calculating interest. The coupon rate is the annual interest rate as a percentage of par value, and it can be fixed or floating. Fixed coupons offer stability for investors, whereas floating coupons adjust to a benchmark like SOFR periodically (Hull, 2018). Bonds maturity date is when the principal is repaid and it can range from under one year (short-term) to over

10 years (long-term). Bodie et al. (2018) note that longer maturities often include higher risks, such as inflation and interest rate risks, but they come with higher yields.

Bonds also have additional features like embedded options, such as puttable or callable bonds, which permit early repayment under strict conditions. Issuers can call back callable bonds, benefiting them and puttable bonds offer some protection for investors by allowing resale back to the issuer at a fixed price (Fabozzi, 2012). Bondholders with convertible bonds have the possibility to convert their bonds into shares. As can be seen, bonds come with a diverse set of features which serve different types of investors making them a flexible financial instrument.

2.1.2 Risks of Bonds

Bonds are often considered as safe or stable investments, they still are subject to several risks that can influence their value and appeal. As bonds are very sensitive to interest rate fluctuations, the most visible is interest rate risk. When interest rates increase, existing bond prices tend to decline, and this effect is worse for bonds with long maturities (Hull, 2018). Credit risk is another major risk that concerns corporate bonds because they depend on issuers financial health. Corporate bonds also involve a higher default risk, which makes the role of credit rating such as Moody's crucial (Fabozzi, 2012).

When prices rise fixed coupon payments lose purchasing power. This is known as inflation risk, which is especially important to consider for longer rated bonds (Tuckman & Serrat, 2011). If markets become illiquid, liquidity risk becomes a major concern that forces investors to sell at discount if bonds must be sold off quickly. Liquidity risk is especially prominent with less frequently traded bonds or lesser-known issuers (Hull, 2018). Callable bonds add further risks such as reinvestment risk.

2.1.3 Bond Pricing

Bond pricing is easiest to describe through the price-yield relationship. For a given set of predetermined cash flows, an increase in the required yield lowers the present value of those cash flows reducing the bond's price. Conversely, if the required yield falls, the present value rises (Tuckman & Serrat, 2012). In this sense yields summarize the discount rate investors use to calculate future cash flows into a market price. The most commonly used yield metric is the yield to maturity (YTM), which is defined by Fabozzi (2012) as the single constant discount rate that equates the present value of the bond's predetermined cash flows to its current market price.

In fixed-income markets yields are useful but incomplete as the sole basis of comparison. Yield does not only reflect issuer-specific risks but also the prevailing level of risk-free interest rates and term structure at any given point. The underlying benchmark curve can shift over time and differs across maturities, yields alone do not isolate compensation that investors demand for holding bonds. Spreads are defined as the yield premium of a bond over a determined reference rate or curve (Fabozzi, 2012). They capture the required excess compensation investors demand for friction and risk that the benchmark does not carry. For corporate bonds the most important risks are credit and liquidity risks.

In this thesis the specific spread used is the G-spread at issuance. G-spread is the bond's yield premium in basis points (bps) over a government yield curve of similar maturity. It measures how much additional yield investors demand on top of the government curve for holding the corporate bond. The G-spread is widely used and easy to interpret making it practical when comparing bonds with different maturities and are issued in different conditions.

There are several alternative spread measures that are commonly used. Z-spread is different from the G-spread as it is defined as a constant spread added to each point of the benchmark curve in a way that the discounted cashflows match the current price

(Tuckman & Serrat, 2012). I-spreads use interest rate swap curves rather than government curves.

2.1.4 Primary vs Secondary Market Pricing

Corporate bonds are priced in two distinct settings: the primary market where new issues are sold to investors and the secondary market, where already existing bonds trade. Bond prices reflect the same underlying building blocks, but the pricing process differs. For this reason issuance pricing should be treated as its own objective, which is the focus of this thesis.

In the primary market, pricing is the outcome of a placement process. The issuer appoints a syndicate of banks to arrange the transaction, market it to investors and build an order book. During bookbuilding, investors indicate demand at different spread levels and the price is adjusted as the demand develops. The final issuance spread is set so that the deal can be placed in full at intended size and maturity. This differs from "trading" drastically, because the underwriters manage information and the end goal is a successful placement (Dick-Nielsen, 2021).

After issuance the bond's price is determined through trading, and this is known as the secondary market. Secondary-market prices react to new information continuously and they are also affected by liquidity conditions. Transaction costs and dealer intermediation affect the observed prices, especially when it comes to corporate bonds because trading can be intermittent and bid-ask spreads are economically meaningful (Edward et al., 2007; Bao et al., 2011).

The main difference between the two markets is that at issuance secondary trading frictions do not exist. At issuance the bond has no trading history and investors cannot observe where the bond might trade once it hits the market. This creates an uncertainty component in primary-market pricing and understandably, investors may demand additional compensation for committing their capital to a new unknown security. This idea is consistent with evidence that newly issued corporate bonds tend to offer slightly better terms than comparable bonds that are already trading (Cai et al., 2007).

For these reasons issuance pricing should not be treated as just secondary-market pricing but "one step earlier". The mechanisms that affect issuance pricing differ from the forces that drive secondary prices.

2.1.5 Underwriting, Bookbuilding and the New-issue Concession

Issuance G-spreads are set through a process that is run by underwriters. As explained before, demand is gathered and this is then translated into pricing guidance and then the final G-spread is arrived at. The process typically starts with an initial price talk (IPT), after which the order book is opened, and revisions are made until demand is high enough and issuers agree with the spread. In other words, underwriters are the intermediaries between issuers and investors. If demand is low, new-issue concessions (discussed later) can be offered to increase it and vice versa, if the orderbook is oversubscribed spreads tend to be tighter (Benveniste & Spindt, 1989; Cornelli & Goldreich, 2003). Thus, the process of bookbuilding is akin to walking a tightrope, issuers cost of capital and the likelihood of successful placement must be balanced or the bond may become illiquid.

The core mechanism that produces information for the issuers and underwriters is the bookbuilding process. Investors indicate demand at different spread levels and this information is used to either widen or tighten them (ICMA, 2018). Investors may try not to show their indications and underwriters on the other hand have incentives to manage the entire process in a way that the bond is issued and performs well in the aftermarket. This is why primary market pricing is quite commonly discussed in terms of information revelation.

New-issue concession or new-issue premium is the extra yield that a new bond offers in comparison to similar bonds already in the market. It works as a placement tool, if demand is uncertain the issuer needs to offer better terms to make sure the orderbook is filled (Hillebrand, 2021). The concession also indicates placement risk from the underwriter's side. If the book is weak it could be priced poorly or trading badly in the aftermarket, which can become a reputational problem for underwriters. For this reason, underwriters commonly offer new-issue premia, which tends to be small in normal markets. However, in highly volatile markets this cushion is larger. Boyarchenko (2022) showed this clearly for March 2020: issuance G-spreads were more than 50 bps higher than comparable secondary market bonds. The logic is simple, if uncertainty is high the gap between primary and secondary pricing increases even more.

2.2 Green Bonds and the Greenium

Green bonds are bonds that are earmarked to finance projects that offer environmental benefits, such as renewable energy, energy efficiency etc. (Tang & Zhang, 2020). Green bonds are structured the same way as any other bonds and they offer coupons, have maturities and seniorities. The difference comes from the use of proceeds, which must be allocated only to projects offering environmental benefits.

The first green bond was issued by the European Investment Bank in 2007, since then the market has grown quickly. This reflects demand from institutional and retail investors who may have ESG mandates or targets for climate or environmental exposure. On the issuer side, green bonds are a way to communicate environmental strategy in a concrete way. To support the growing green bond market frameworks such as the Green Bond Principles (GBP), developed by the International Capital Market Association (ICMA) have increased transparency and standardization which reinforces the market's credibility. EU has also introduced the European Green Bond Label (EuGB) which sets requirements around disclosures, external review and the use of proceeds (European Union, 2023).

The pricing concept that follows from this is the "greenium", which is the main topic of this thesis. Greenium is used to describe a pricing advantage for green bonds, meaning that investors may accept a lower yield compared to a similar non-green bond. At issuance this means a tighter G-spread. In this thesis the dependent variable is the issuance G-spread, so greenium at issuance means a lower issue G-spread for a green bond. In short, if two bonds are comparable but the green one is priced tighter, that differential is the greenium.

Greenium cannot be identified from simple “green” versus “non-green” averages without carefully planned controls. This is because issuance G-spreads are sensitive to other factors such as bond characteristics and market conditions. Even within academic literature the evidence on a greenium is mixed across samples, time periods and continents. For example, ECB’s work on green bond pricing notes that results depend on how comparability and liquidity are handled (ECB, 2022). Due to this, it is important to distinguish between a greenium that is real and one that is only apparent. A real greenium would mean that investors price the green label and its framework. An apparent greenium would emerge when green bonds differ from their comparison group in practical ways, for example in credit rating, issue size and market conditions.

2.3 Information Asymmetry and Signaling

At issuance the issuer and syndicate tend to know more about the security than outside investors. This includes soft information about the issuer’s current situation, motivation for issuing and how the deal is expected to trade. Investors are limited to the public information, but they can still face uncertainty about the quality of the bond and the pricing after issuance. This imbalance of information is known as information asymmetry (Akerlof, 1970; Stiglitz, 2000).

In finance, investors typically demand compensation for uncertainty, meaning that issuance G-spreads typically widen due to asymmetric information. This resembles the classic “lemons” logic: when quality can’t be verified the price has to reflect the risk of adverse selection (Akerlof, 1970). Even when underwriters try to reduce this uncertainty, it cannot be fully removed because the bond has not established trading history at all, and the demand is still just developing (Benveniste & Spindt, 1989).

For green bonds there is an added layer of uncertainty. Beyond the standard credit and liquidity, investors must be certain that the issuer will use the proceeds as described and that financed projects are genuinely green and not subject to greenwashing. Greenwashing is not always about intentional deception, it can also reflect weak reporting, unclear criteria and other factors that make environmental claims difficult to verify (Delmas & Burbano, 2011). This means that green label can controversially increase uncertainty, if the issuer's commitments or monitoring are not credible.

To decrease this uncertainty, signaling theory suggests that credible signals can resolve problems around informational uncertainty. Spence (1973) shows that the idea is separation: if high-quality issuers can send a signal at a lower cost than low-quality ones, then that signal helps investors distinguish between credible and non-credible ones. When the signal is easy to copy, it does not carry much weight (Spence, 1973). In finance a similar logic can be applied, when firms use actions such as allowing external monitoring or issuing claims that can be scrutinized, they can reduce adverse selection (Leland & Pyle, 1977).

In the context of green bonds, the "signal" is the green label and credibility around it. Green bonds are issued under the GBP, which emphasizes clear use of proceeds, management of proceeds and reporting (ICMA, 2021). By giving issuers and investors a shared framework, they make green issues easier to compare and track, which in turn reduces information gaps between issuers and investors. This standardization has resulted in wider participation, especially from institutional investors (Tang & Zhang, 2020). Following the GBP also benefits issuers reputation, because it is a visible commitment to standard practices and disclosure.

2.4 From Theory to Prediction

To move from theory to the testable predictions the green coefficient and greenium are discussed in this subchapter.

Under investor-preference models some investors are willing to accept a lower expected return in exchange for holding securities that satisfy some non-financial preferences or mandate constraints (Heinkel et al., 2001; Hong & Kacperczyk, 2009). In bond markets this implies that holding risk, cash-flow and bond characteristics constant, green bonds can be issued at tighter G-spreads than comparable non-green ones. The core prediction from investor-preference therefore is a negative green coefficient. Another prediction under these models is state dependence, the green discount should be larger when demand for green assets is stronger (Pástor et al., 2021; Pedersen et al., 2021). In this thesis direct demand is not observed, therefore this is evaluated indirectly. If a negative green coefficient emerges only after controlling for fundamentals and is more pronounced in specific market states (e.g., lower volatility or easier funding conditions), this pattern would be associated with preference driven pricing.

Under signaling models issuing a green bond can serve as a signal of a firm's environmental commitment or lower long-run transition risk, but this signal is only informative when it is credible (Spence, 1973). The prediction is not that green bonds are cheaper, but that any G-spread advantage should be stronger for bonds with credible verification (Bachelet et al., 2019). In other words, the sign is expected to be negative conditional on credibility. Green classification relies on LSEG's green label in this thesis, and the analysis does not split bonds by certification or second-party opinions. As a result signaling predictions are only assessed at a high level: if the green coefficient is weak or unstable in the full sample, signaling theory suggests that the sample averages over heterogenous green credibility and dilutes the effect.

Third explanation is that green pricing differences arise mechanically because green and non-green bonds differ in characteristics such as maturity, ratings, issue size etc. or on unobservable traits. The tested prediction is attenuation: the green coefficient should shrink in size as controls are added (Tang & Zhang, 2020; Zerbib, 2019). This is directly addressed with the OLS ladder, if the green coefficient is large in the univariate model but collapses towards zero as controls are added, the evidence supports composition/selection effects rather than a pure label effect. However, if the coefficient remains stable after all controls are introduced, the result is less consistent with selection and composition.

2.5 Machine Learning in Financial Modelling

Since this thesis relies on a machine learning (ML) algorithm to study the determinants and G-spread prediction of corporate (green) bonds at issuance, it only makes sense to examine machine learning as a whole in detail. This chapter explains what ML is, digs into history of ML, its usage and then the methods used in this thesis.

2.5.1 Definitions of Machine Learning

Machine learning is a rapidly developing field of computational algorithms that aim to mimic human intelligence by learning from their surroundings and experiences. One of the hallmark features of ML is its ability to learn from training data or a past experience and then to solve a given task. ML models are an essential part of modern commercial and scientific environments due to the emergence of big data and where probability predictions play a key role. ML techniques have been used in multiple different fields, such as pattern recognition, engineering, finance, and even medical applications (El Naqa & Murphy, 2015).

Zhou (2021) emphasizes that one of the most important objectives of ML is to develop algorithms that learn by using data to build models. Simply put, the algorithm is fed with data after which the model makes predictions on new observations. Mohri et al. (2018) claim that ML can be defined as computational methods that make accurate predictions based on learned experiences from data. Alzubi et al. (2018) say that ML tries to improve actions of computers from experiences, and this leads to obtaining higher accuracy. They further define that the measure of accuracy is the number of times the chosen actions lead to the right outcomes.

2.5.2 History of Machine Learning

The background of ML goes all the way back to the 1950's and 1960's, when a team led by Alan Turing developed the "Turing test" (Turing, 1950) to see if a computer could think like a human being. The first paper on artificial intelligence was written by McCarthy et al. (1955) and following this paper, a conference was held in Darthmouth in 1956. Perhaps the first ML algorithm was developed by Rosenblatt in 1958, when he developed the "Percepton" which was based on biological neurons. Alzubi et al. (2018) claim that Arthur Samuel popularized the entire "Machine Learning" term in 1959.

ML entered a new age in the 1970's, coined as the "knowledge age" (Zhou, 2021). During the 1970s a massive amount of different algorithms and systems were developed across multiple different fields. Moving on to 1986, the first journal solely dedicated to ML was created in 1986, creatively the journal was named "Machine Learning". Since then, ML has become a bread-and-butter research field, that works independently and not only as a tool to solve a problem from a different field. The "modern" ML can be rooted to the 1990s, when a shift towards statistical learning techniques emerged alongside neural networks. In the world of today, companies and organizations among others have collected huge amounts of data that affect multiple aspects of our lives, ML methods are in an important position to process and use this data.

2.5.3 Machine Learning in General

Simply put, ML always contains six components no matter which algorithm is used. First phase in the whole ML process is the collection and preparing of data that is usable for the algorithm as an input. Essentially this is cleaning and pre-processing of the data to make it structured, usable, and relevant for the model.

Once the data is prepared, the data comprises of multiple features, of which not all are relevant for the specific learning process. This differs from task to task. Naturally, the next step is then to remove those features that are irrelevant and only select those that are relevant.

As different ML algorithms are suited for different tasks and problems, selecting the best suited one is the next step. This is perhaps the most critical step, as choosing the wrong model may not yield good results. Following the model choice are the next two steps that are embedded in modern ML models already: selection of models and parameters and training. Training means splitting the data into a train and test sets, because the model needs to be trained on data so that it can make accurate predictions.

After these steps, the final step is to evaluate its performance. This step involves the test subset created in an earlier step, in short the model is tested on unseen data to evaluate how it performs and whether performance parameters are suited for the specific problem (Alzubi et al., 2018).

Mahesh (2020) decomposed ML into multiple different areas based on the type of model and its objectives.

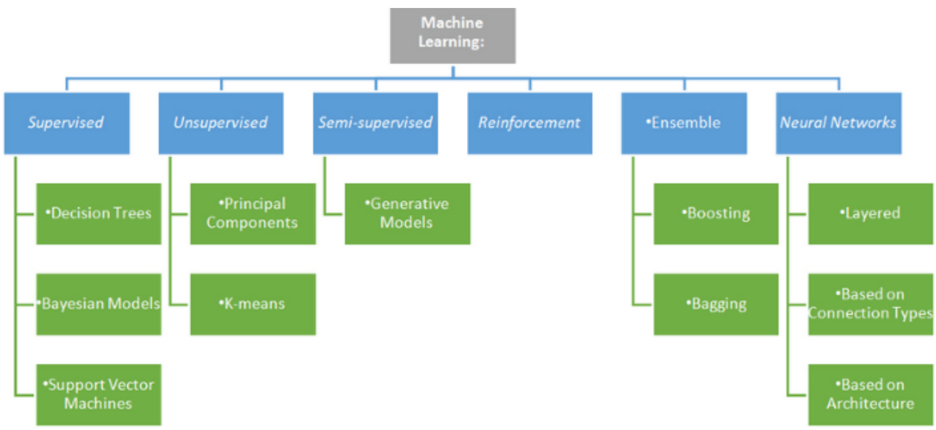

Figure 1. Decomposition of ML branches and models (Mahesh, 2020)

In this thesis the focus is on ensemble and supervised learning techniques, as the objective is to predict a predetermined objective. To effectively use a supervised ML algorithm, the data is split into training and test splits, where the training dataset includes the output variable that I am trying to predict. As the name suggests, during training the algorithm will attempt to learn patterns from the training dataset, after this the patterns are tested on the unseen test split for prediction (Mahesh, 2020). This thesis more specifically uses a random forest (RF) algorithm which is included in the ensemble learning models. RF is an extension of Classification and Regression Trees (CART), and as the name suggests it includes multiple CARTs to create a "forest", simply put the CARTs are put together by using bootstrap aggregating.

2.5.4 Random Forest for Regression

In this chapter the focus is on providing background information on regression and tree-based models. First, it discusses decision trees in general and concludes with the random forest that incorporates multiple decision trees.

Linear models have certain limitations, for this thesis the most important one is that it constricts the relationships between variables to be linear, which I believe may not be the case when using real-world data. Decision trees are able to catch non-linear dependencies by creating a decision function, furthermore this does not require significant feature engineering. Wilkinson (1992) showed that decision trees can handle datasets with categorical variables, which is important for this thesis as some of the predictors are categorical.

There are multiple ways to build a decision tree. Boehmke and Greenwell (2019) stated that the most common and recognized way is the CART algorithm. This type of a decision tree comprises nodes that form a "rooted tree". What this means is that the tree has direction and a root node that has no incoming branches. All of the other nodes have specifically one incoming branch and are called internal or test nodes. Terminal nodes are called leaf nodes.

Borrowing from Maimon and Rokach (2005) each node has a logical statement, in the form of 2 [$xi \leq t$], where xi is a value of the ith feature variable, t is a threshold that was chosen during fitting of the decision tree. The next nodes are then evaluated until a leaf node is reached.

CART on the other hand uses a different procedure. At every node the task is to identify the optimal feature (x_i) that can split the remaining data into two regions (R1 and R2) in a way that the overall error between actual response (y_i) and the constant being predicted (c_i) is minimized. In regression the goal is to reach the minimum total sum of squared errors (SSE):

$$SSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2$$

As explained by Boehmke and Greenwell (2019), once the optimal feature and split combination is found, the data is split into two regions and the splitting is repeated on both of these regions. This process continues until either the tree becomes too complex, or a stopping criterion such as reaching the maximum depth is met.

Decision trees can be used to solve regression and classification tasks, but relying on just one decision tree can be unreliable. One of the main issues with decision trees is that they are susceptible to overfitting the data, meaning that the model learns patterns in the training subset too closely leading to inaccurate predictions on the unseen data.

The issue of overfitting a single decision tree is easily fixed by using a Random Forest, because Random Forest trains multiple individual decision trees (hence the forest) and then computes the Random Forest prediction, which is the average prediction from the forest. Each tree in this ensemble is trained on a randomized split of the data via a bootstrap aggregating. Breiman (2001) showed that this method reduces the model variance, limiting overfitting and increasing the random forest's predictive accuracy. Two decades later, James et al. (2021) confirmed these findings and showed that random forests increase robustness and performance of decision trees.

2.5.5 Bootstrap Aggregating

A bootstrapped sample is created by drawing observations at random from the entire dataset with replacement. This means that any observation that gets picked can show up again later in the same sample. It is important to note that each bootstrapped sample is exactly same in size as the original dataset (Efron & Tibshirani, 1986). Figure 4 illustrates this: every bootstrapped sample contains exactly 12 observations, which matches the original dataset. Because the sampling is redone from the same pool of data, the bootstrap samples usually have similar overall distribution to the original data.

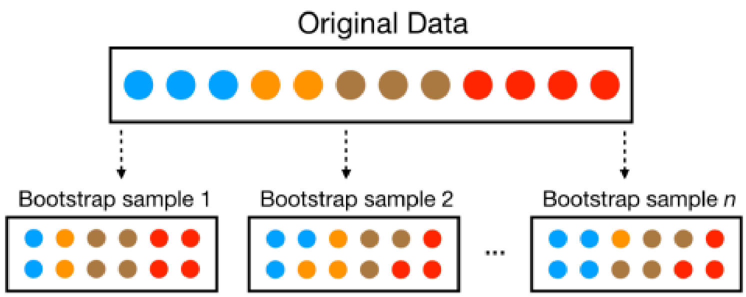

Figure 2. Example of bootstrapping process (Boehmke & Greenwell, 2019)

Because bootstrap samples are selected with replacement, the same observation can appear more than once in any given resample. As a rule of thumb, each bootstrap contains about two-thirds of the unique observations from the original dataset. The observations that are left out of a particular resample are called out-of-bag (OOB). Boehmke and Greenwell (2019) show that in practice, it is possible to fit the model on the bootstrap samples and then evaluate it on the OOB observations. This provides a “built-in” validation set and it is widely used for example in random forests.

Bootstrap aggregating provides a general approach for building prediction models. It trains many models and combines their outputs into a single prediction. This method draws *b* number of bootstrapped samples of the original training set. Then a base learner which is a regression or classification algorithm is applied to all of the bootstrapped samples. Finally, the resulting predictions are averaged:

$$f_{\text{bag}} = f_1(x) + f_2(x) + \cdots + f_b(x)$$

Where: b = The number of bootstrap copies, x = The record for which a prediction is required and f, (x) = the predictions from the individual base learners.

As explained before, bootstrap aggregating reduces the variance by averaging multiple predictions, Breiman (1996) states that this reduces the risk of overfitting the entire model.

2.5.6 Random Forests

As discussed before, the process of aggregating the predictions results in enhanced predictive performance. However, bootstrap aggregating does not come without its issues. One of the biggest problems with using only bootstrap aggregating is that it can lead to trees correlating with one another, limiting the usefulness of reduced variance. Bagged decision trees add randomness by fitting each tree on a different bootstrapped resample of the training data, but this alone may not decorrelate the trees enough.

Random Forests are an upgrade to simply bootstrapping decision trees. The logic of Random Forest is to choose a different subset of data and features it uses in each split. Doing this, the decision tree “forest” is more varied. Boehmke and Greenwell (2019) theorize that this reduces the correlation between individual trees compared to only bootstrap aggregating decision trees, resulting in significant improvement in predictive accuracy. Figure 5 shows the random forest algorithm in action.

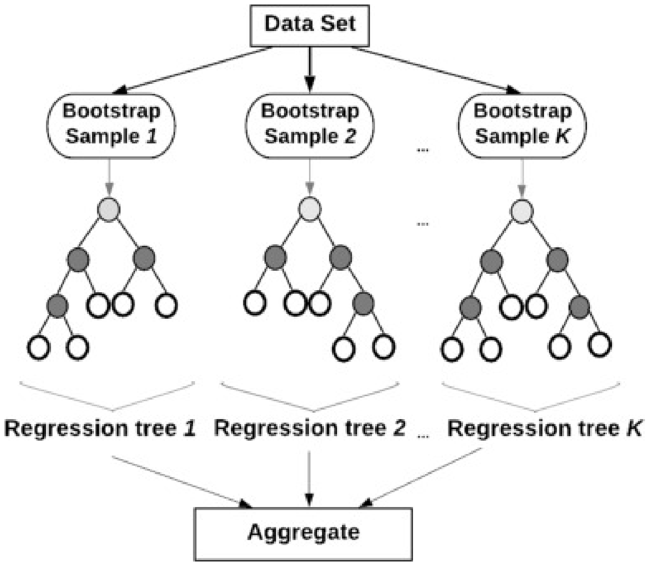

Figure 3. Random Forest model (Makariou et al., 2021)

2.5.6.1 Tuning Parameters of a Random Forest

Random forests are a very commonly used ML model, partly because they offer great performance with very minimal configuration. Probst et al. (2018) showed that random forests have the lowest amount of variability in prediction accuracy between the baseline configuration and a fully tuned version. Essentially, multiple hyperparameters can be tuned, but this is often not necessary for random forests. Regardless of the fact that the model performs well without tuning, Probst et al. (2018) further note that the following hyperparameters should still be tuned:

Number of trees

Number of trees in a forest is important for minimizing the error rates. As a starting point it is common to start with a number of trees that is ten times larger than the number of features used in the model. Notably, the time it takes to run the model (e.g., computational time) increases drastically with the number of trees. However, increasing the number of trees has a positive effect on the robustness and stability of error estimates and even variable importance metrics, so it typically is worth the extra time.

Number of Features to consider at splits (mtry)

This hyperparameter controls how much randomness is introduced when selecting variables at each split. It has a central role in balancing the trade-off between increased predictive performance and low tree correlation. The default first value for mtry is often $p/3$ where p is the number of predictors. In this thesis best results were obtained with mtry = 8. If there are only a few relevant predictors or the data is noisy, a larger value could be more effective because it increases the likelihood of selecting features with more predictive power. Vice versa if there are many relevant predictors one should try using a lower value of mtry.

Complexity of each tree

The result of constructing a random forest from individual decision trees is that many random forest implementations (such as R packages: randomForest and ranger) provide hyperparameters that allow the user to fully control the depth and complexity of individual trees. These typically include node size, maximum depth, maximum number of leaf nodes and minimum node size required for splitting any further. The latter is considered the most common hyperparameter to control tree complexity. In this thesis complexity is controlled by tuning the minimum node size. Goldstein et al. (2011) found that most implementations use five as the default value for minimum nodes as it yields good results overall.

Sampling

As discussed before, bootstrapping is the common sampling scheme used with random forests, by default all the observations are sampled with replacement. There is however an option to change the sample size and decide whether sampling occurs with or without replacement. Boehmke and Greenwell (2019) demonstrated that if there are only a few dominating features, decreasing the sample size and altering the sampling scheme can reduce correlation between trees.

Grid search is a common approach for hyperparameter tuning, it involves assigning pre-determined values or ranges to the selected tuning parameters. Hsu et al. (2003) demonstrate that this method then trains new models for all parameter combinations using k-fold cross-validation to reduce errors. Then the final models are evaluated based on error metrics and the best is chosen. In this thesis the chosen implementation (ranger) performs grid search on its own and automatically selects the best performing model.

3 Literature Review

When it comes to a natural starting point for the literature on primary market pricing, it is the bookbuilding model that was developed by Benveniste and Spindt in 1989. In their study they examined how investment banks set offer prices and allocate securities in new issues in scenarios where investors differ in the information they possess. The main point of the paper is that underwriters strategically use both pricing and allocation decisions to encourage informed investors to reveal their true demand. From this perspective, underpricing and allocation discretion are not anomalies or mistakes, but deliberate tools designed to address information asymmetry between issuers and investors. The model also highlights the active role of investors in the pricing process, which shows that their behavior influences directly how securities are offered and distributed. Over time, this has become a cornerstone of the literature on new issue markets and it also continues to inform subsequent research on issuance pricing and investor behavior (Benveniste & Spindt, 1989).

Cornelli and Goldreich (2003) examined how much information the order book contains. They showed that investors` orders provided important signals about demand, which then underwriters use to set prices and decide how to allocate shares. This suggests that pricing and allocation in the primary market are not random but responds to the information that comes from investors. While the paper focuses on equity, the findings are also relevant for bonds, especially when thinking about how investor demand might shape the pricing and allocations in both conventional and green bond markets. (Cornelli & Goldreich, 2003).

Building on earlier work, Gabbi and Sironi (2005) look at what drives pricing in the primary corporate bond market. Focusing on Eurobond issues, they explored how factors like market conditions, issuer characteristics, and the specific features of each bond could affect yield spreads at the time of issuance. The results show that credit risk variables, for example credit rating and maturity, play a central role in explaining primary market spreads, while factors that are related to market timing and liquidity also contribute. The paper offers early empirical support for the view that bond pricing at issuance can reflect both prevailing market and issuer risk conditions. It is also the reason why it has become a reference point for subsequent research on corporate bond primary markets. This paper functions as the main guideline for variable selection in this thesis (Gabbi & Sironi, 2005).

A widely cited contribution to the green bond literature is a paper by Ehlers and Packer in 2017. The study examines the role of certification in the emerging green bond market, which was still relatively small and heterogeneous at the time. Using market-level evidence, Ehlers and Packer document that green bonds tend to be issued at slightly lower yields than comparable conventional bonds. This happens especially when third-party certification is present. The authors argue that certification helps to reduce investor uncertainty about the environmental credibility of the bond. In this way, it lowers required returns at issuance. The whole paper highlights certification as a key institutional feature in green bond markets and establishes a link between credibility mechanisms and issuance pricing. This has shaped much of the subsequent empirical work in this area (Ehlers & Packer, 2017).

Table 1 below summarizes the findings and methods used in the most prominent recent green bond literature.

Table 1. Recent Green Bond Literature

Study	Market	Method	Liquidity Measure	Green definition	Key Finding
Zerbib, 2019	Secondary	Matching + FE panel, yield regressions	Bid-ask spread	GBP alignment	Lower yields for green (-1 to -2)
Bachelet et al., 2019	Secondary	Matching + FE	Bid-ask spread	CBI alignment	Higher yields for green (2-5bps)
Larcker & Watts, 2020	Primary	Matching	Controlled via matching and liquidity tests	Self-labelled (Bloomberg)	No pricing difference
Flammer, 2021	Primary	Within-issuer Matching	Controlled via matching	Bloomberg indicator + third-party	No pricing difference
Caramichael & Rapp, 2024	Primary	FE regressions + matching for robustness	Bond/market controls and FEs	GBP alignment	Lower yields for green (-3 to -8)
Gabbi & Sironi, 2005 (not green)	Primary	Regression	Log Amount issued	N/A	Not applicable

Caramichael and Rapp (2024) study pricing at the time of issuance and argue that a “green discount” shows up when demand for green bonds is strong. They find that green corporate bonds can be issued at lower yield spreads than comparable non-green bonds and link this difference to measures of demand pressure such as oversubscription and index-related demand. For example, more pronounced in euro-denominated deals and in USD deals by foreign issuers). This perspective explains why signs are different across papers: the green coefficient can move with investor base and issuance conditions. The research design they use helps substantially against selection and composition effects as they include extensive fixed effects and controls. Caramichael and Rapp (2024) further report heterogeneity and don’t report just a simple estimate. For this thesis, their results motivate the ladder’s market state block: if the green coefficient changes significantly once issuance-time market conditions are controlled for, that pattern is consistent with demand dependence. However, this thesis does not observe oversubscription, index inclusion, or holdings, the ladder cannot “close the mechanism” in the way Caramichael and Rapp were able to. The resulting green coefficient remains a reduced-form pricing effect for the 2023-2025 regime rather than a direct demand-pressure estimate.

Flammer (2021) studied the primary market issuance and compares green bonds to within issuer non-green bond with very tight matching rules. They find a mean difference of -1.9 bps and a median difference of zero. The methodological contribution here is not only the near-zero estimate, but the identification logic: by matching within issuer the “green issuers are systematically different” argument is neutralized and matching on issuance amount partly explains expected liquidity at issuance. Flammer’s paper is therefore especially strong on issuer-level composition and unobservables. In this thesis the ladder replicates the observable logic (bond characteristics and credit quality enter sequentially), but it does not replicate the within-issuer aspect to the same extent, issuer fixed effects are used, however. If the green coefficient remains positive in this thesis after the ladder, one interpretation is that the effect is not purely due to observable selection.

Larcker & Watts (2020) show that a green effect can disappear quickly once comparisons are made between nearly identical bonds. In the municipal market they compare green and non-green bonds issued by the same issuer on the same day with closely aligned terms and find extremely small pricing differences: 0.44 to 0.46 bps. They also test whether green bonds trade differently after issuance using turnover, number of trades and price dispersion. They find no meaningful liquidity differences in these matched samples. They interpret their results as economically negligible. This paper is central for explaining sign differences because it shows that when controls are tight the label effect can be very small. Larcker & Watts (2020) clarifies what the ladder used in this thesis can and cannot do: it provides transparency on how the coefficient changes with controls, but without tight matching it is more exposed to selection than their approach is. One key difference is that Larcker & Watts studied municipal bonds, not non-financial corporates that this thesis does.

Zerbib (2019) studied greenium estimates in the secondary market and found that they depend heavily on liquidity controls. He builds a synthetic “non-green” comparator and then adds an explicit liquidity adjustment, because even close matching cannot perfectly equalize that. He defines a bid-ask based liquidity proxy which is the green bond's bid-ask spread minus the distance weighted bid-ask spread of a matched conventional bond. The reason the study relies on this kind of liquidity measure is that the dataset does not allow for intraday liquidity indicators and lacks TRACE information. The paper shows that a green coefficient in secondary markets can reflect a liquidity premium unless handled carefully, mixed results can therefore occur when liquidity is handled differently. For this thesis this study is relevant as a contrast: using issuance G-spreads mitigates the most direct trading-liquidity channel, but the ladder cannot fully replicate expected secondary market liquidity. As such separating the preference driven pricing from expected liquidity effects is not fully possible.

Bachelet et al. (2019) make heterogeneity and measurement explicit by pairing green and brown "twins," constructing synthetic matches when needed, and then examining not only yield/return differentials but also liquidity (bid–ask spread) and trading inactivity (a zero-trading-day indicator). They find that green bonds are, on average, more liquid than their matched twins (around 5 bps in their liquidity metric), and they show that the sign of the "premium" depends on issuer type and credibility. This directly supports a reconciliation narrative: the sign can flip when "green" interacts with asymmetric information, credibility, and liquidity in different issuer segments, rather than reflecting a real discount. The ladder addresses these partly, the sector and credit quality blocks reduce composition driven sign flips that come from mixing issuer types. Furthermore, as the thesis uses a single vendor green label and does not split external review or different certifications, it cannot test whether any estimated effects concentrate in "more credible" subsets, due to all bonds having the same credibility label.

MacSkill et al. (2021) conducted a systematic review and showed that disagreement is systematic. They report that a green premium is found in 56% of primary market studies and 70% of secondary market studies, and that the average reported greenium is often in the range of -1 to -9 bps in the secondary market, primary market studies report higher variation. This supports the idea of treating the green coefficient as contingent on market, sample and methodology, and justifies why a ladder is valuable. Rather than presenting just a single "best" specification, the ladder shows which blocks account for the changes in the coefficient.

A key recent contribution to the heterogeneity narrative is Pietsch (2025), who studies euro-area greenium dynamics from 2016–2023 using a matched panel approach. They report that retail investors' demand partly drives the greenium, with the greenium evolving across market conditions such as the post-pandemic inflation shock and subsequent tightening. This result helps to interpret a green coefficient in any single window: the coefficient is not only a function of fundamentals and label definition; it also depends on who the marginal buyer is in that regime. For this thesis, the implication is that a positive green coefficient in 2023–2025 USD issuance can be consistent with a regime where demand pressure is weaker even if earlier periods or other markets produced a negative premium. The ladder's market-state controls can absorb broad risk conditions correlated with demand regimes, but without holdings or bookbuilding data the thesis cannot frame any greenium found as a universal construct, but only representative of the sample.

Moving to the usage of machine learning in bond pricing context, Kim et al. explored in 2021 how modern techniques, specifically machine learning, could be applied to forecasting corporate bond yield spreads. Even though the study focuses on corporate bonds in general rather than green bonds specifically, it offers useful insights into what drives pricing in both primary and secondary markets. They show that machine learning models can predict yield spreads better than traditional econometrics methods. This is because they can pick up on complex interactions between macroeconomic variables, issuer characteristics, and market conditions. They also point out that which factors matter most can change depending on the model used and the time period considered. They show that these data-driven approaches can give us a clearer picture of what influences corporate bond yields. They could be especially useful for marketing green bonds, where investor behavior is still developing and market conditions can be rather unpredictable.

Gu et al. (2020) showed that machine learning can uncover patterns in asset prices that traditional models often miss. They highlight that these methods are particularly good at spotting complex relationships between firm characteristics, macroeconomic conditions, and asset returns. Even though the study is not focused on bonds specifically, it still offers useful lessons for fixed income markets, where investor behavior, market conditions, and issuer traits all interact together. Bianchi et al. (2021) built on this idea in the context of corporate bonds. The authors found that risk premiums fluctuate over time and are shaped by a mix of visible and less obvious factors. Machine learning can detect these patterns more effectively than traditional methods. This can give a clearer picture of what drives bond returns. Taken together, these two studies suggest that data-driven approaches could be a valuable tool for improving both pricing and risk assessment in bond markets.

Recently, machine learning has been used in financial research to complement the standard econometric approaches. Varian (2014) argued that machine learning enables economists to analyze high-dimensional data without prespecifying functional forms. Mullainathan and Spiess (2017) emphasized that predictive modelling allows for more robust generalization when theoretical relationships are uncertain. Chen et al. (2024), found that nonparametric methods identify latent factors that improve interpretability and predictive accuracy, showing that flexible modelling can uncover relationships that traditional econometrics may overlook.

4 Methodology

Issuance G-spreads at launch reflect how issuer risk, contractual bond features, sector composition, and market conditions translate into the risk premia investors demand at pricing relative to a government benchmark. Let y_i denote the issuance spread (G-spread, basis points) of bond i at the pricing date and let X_i be the information set observable at issuance. The task is to characterize the following:

$$y_i = f(X_i) + \varepsilon_i, \mathbb{E}[\, \varepsilon_i \mid X_i \,] = 0$$

where f is the pricing function and ε_i captures pricing noise. Multiple estimators are implemented to characterize f from complementary perspectives: OLS and LASSO benchmarks designed for inference and a Random Forest model designed for prediction (Breiman, 2001; Hastie et al., 2009). Combining inference-oriented and prediction-oriented methods follows the view that linear models support transparent interpretation, while supervised learning can improve prediction when the dataset is complex (Varian, 2014; Mullainathan & Spiess, 2017). To compare the performance of the models they are evaluated on the same out-of-sample (OOS) test set and as an additional robustness check also out-of-time (OOT).

4.1 Target Construction, Timing Alignment, and Preprocessing

G-spreads can have heavy right tails, and extreme values can distort estimation and diagnostics. To combat this the dependent variable is winsorized at the upper 99.5th percentile. This replaces values above the cutoff with the cutoff value but does not remove any observations.

Because maturity artefacts can in the same way distort relationships and fitted models, observations with life over 50 years are excluded. As the topic of interest is the primary market, all market-state variables are aligned to each bond's pricing date. If a series is not observed on the exact pricing date due to, for example, non-trading days, the most recent prior observation is used. Macro variables with lower frequency (e.g., CPI) are recorded from the publication month to the issuance date so that they reflect information that is available at issuance date. These alignments are required so that the controls represent information available at issuance, not after it. After this all at issuance macro variables are aligned to the pricing date. Rare sector categories are collapsed into "Other" as describe in section 5.5. After this the sample is split into training and test sets.

4.2 Performance Metrics and Repeated Evaluation

The performance of the OLS and RF in this thesis are measured by commonly used error metrics; Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R²). RMSE measures the average size of the errors of the regression models, MAE measures the average absolute differences between predicted and actual values and R² shows how much of the variance in the dependent variable is explained by the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

4.3 OLS benchmark: Conditional Inference on the Green Label

4.3.1 Baseline Regression

The baseline OLS model is described below. Let $Green_i \in (0,1)$ be an indicator equal to one if bond i is green-labelled. The baseline model is

$$y_i = \alpha + \beta\, Green_i + X_i'\gamma + u_i$$

where X_i contains issuance-date controls, and β is interpreted as the conditional-spread differential associated with the green label. The model is conditioned on determinants in blocks, added one after another to create a “ladder” structure. This is important because unconditional differences between green and non-green bonds can result from factors other than the green label.

4.4 OLS Ladder

As explained prior, controls are added one after another to the OLS model to create a “ladder”. This is done to track how the estimated green coefficient changes as more controls are introduced. This offers a transparent way to see if a greenium exists and if it survives controls and remains stable. Basically, this offers a way to separate the green label from other factors.

Let i index bond issues and y_i denote the issuance G-spread measured in basis points. The baseline regression is:

$$y_i = \alpha + \beta\,\mathrm{Green}_i + u_i \quad (0)$$

where α is a constant, Green_i is an indicator equal to 1 if issue i is green-labelled. β is the coefficient of interest capturing the association between the green label and G-spreads, and u_i is an error term.

After this bond characteristics are added to the model:

$$y_i = \alpha + \beta\,\mathrm{Green}_i + B_i'\gamma + u_i \quad (1)$$

where B_i is a vector of bond characteristics measured at issuance and $\gamma$is the associated coefficient vector. In this thesis, B_i includes maturity (life), log issue size (log amount issued), and coupon terms (coupon level and coupon frequency).

Credit quality is then introduced for issuer risk:

$$y_i = \alpha + \beta\,\text{Green}_i + B_i'\gamma + C_i'\theta + u_i \quad (2)$$

where C_i denotes credit quality observed at issuance and θ is the associated coefficient vector.

After this sector composition controls are added:

$$y_i = \alpha + \beta\,\text{Green}_i + B_i'\gamma + C_i'\theta + S_i'\psi + u_i \quad (3)$$

where S_i is two-digit industry indicator variable and ψ is the corresponding coefficient vector.

Next, issuance-time conditions aligned to the pricing date are included:

$$y_i = \alpha + \beta\,\text{Green}_i + B_i'\gamma + C_i'\theta + S_i'\psi + M_i'\eta + \lambda L_i + u_i \quad (4)$$

where M_i is a vector of market controls aligned to the issuance date and η is the associated coefficient vector. M_i includes a term slope measure, an interest-rate volatility proxy, an inflation measure, and a broad credit conditions proxy. The scalar L_i is a funding/liquidity proxy aligned to the issuance date, and λ is its coefficient.

And finally, as a robustness check, issuance-year fixed effects are estimated as an alternative to the observed market-state block:

$$y_i = \alpha + \beta\,\text{Green}_i + B_i'\gamma + C_i'\theta + S_i'\psi + \delta_{t(i)} + u_i \quad (5)$$

where $\delta_{t(i)}$ denotes a full set of issuance-year indicators derived from the pricing date, and $t(i)$ maps issue i to its issuance year. Model M5 is treated as a robustness check for model M4, as it shows if the chosen macro variables capture issuance conditions.

Put simply, in M0 β is an unconditional mean difference in G-spreads. In M1 β is conditioned on bond structure. M2 uses credit quality to rule out the possibility that green bonds are disproportionately issued by safer issuers. M3 includes sector and addresses the concern that green bonds are issued only in specific sectors. M4 considers market conditions at the time of issuance, addressing the possibility that green bonds are issued only in favorable conditions.

4.4.1 Inference: Robust Standard Errors and Estimator Choice

Cross-sectional issuance data may exhibit heteroskedasticity, since the G-spread residuals can naturally be more dispersed for some bonds than others across ratings, market environments etc. To account for this the standard errors reported are heteroskedasticity-robust standard errors that use the HC1 correction (White, 1980; MacKinnon & White, 1985). This adjustment relaxes the assumption of constant error variance and makes inference more reliable when residual variance is not constant. Furthermore, as there are multiple issues in the same day, day clustering is also done.

4.5 Diagnostics

Diagnostics are used to make sure the results are interpretable and not being driven by artefacts. In the best performing OLS specification, that is compared with the RF, a Breusch-Pagan test is used to check for heteroskedasticity. Multicollinearity among the continuous covariates is checked using variance inflation factors (VIF). Fixed effects and dummy variables can render VIFs misleading, VIFs are computed on the same covariates without fixed effects.

4.6 Random Forest

In this thesis the Random Forest uses issuance time G-spread as the dependent variable and bond characteristics, market conditions and the green label as independent ones. The actual Random Forest algorithm used is the ranger implementation implemented in R. This implementation follows the classic Random Forest theorem by Breiman (Breiman, 2001), where each tree is built from a bootstrapped resample of the training set.

4.6.1 Training, Split and Hyperparameters

To evaluate the out-of-sample (OOS) performance the dataset is partitioned into a training and test splits where 75% of the data belongs to the training set and 25% to the test set. A time-based split is commonly used when the goal is forecasting forward, but in this thesis the objective is cross-sectional pricing that depends on issuance-time conditions, so the split is not time-based. To verify that the OLS-RF performance comparison is not an artefact of one particular split, repeated stratified random splits are used. This is standard practice when evaluating supervised learning (Hastie et al., 2009; Varian, 2014).

There are multiple ways to tune a Random Forest model in addition to manual trial-and-error. Most used ones are Grid Search, Random Search and Bayesian Optimization. Random Search requires the user to define number of trials, and the model randomly tests that number of different combinations. Grid Search on the other hand tries all possible combinations of hyperparameters and their values, which can be quite heavy computationally. Bayesian optimization works differently, in that it tries to constantly improve its performance using past performances. However, if the earlier estimates are noisy, it can overfit random fluctuations. As such, this thesis uses Grid Search as the safest option, despite the heavy computational power required.

Following the suggestions of Probst et al. (2018) the hyperparameters that are tuned are the mtry which is the number of candidate predictions considered at each split and min.node.size which is the required minimum number of observations in a leaf node. The grid evaluates min.node.size in 5,10,20 and 40

Hyperparameter tuning is performed through grid search combined with 5-fold cross-validation within the training set, with folds stratified by the green label.

The grid evaluates min.node.size in 5,10,20 and 40 and a practical set of mtry candidates spanning from low to high feature-subsampling regimes (implemented as evenly spaced candidates between 2 and the full predictor count, in this application the grid spans 2 to 13). The number of trees is fixed at 500 to keep the tuning computationally feasible, as the power required increases linearly. This also ensures that the differences in error metrics used for validation do not result from changes in forest size. After the hyperparameters that minimizes error are selected, the final RF model is trained on the full training data using 1000 trees.

4.6.2 Handling Categorical Sector Identification

Sector composition enters RF as a categorical two-digit industry classification (SIC2). Rare levels can cause two problems: factor levels that appear only in the test split lead to out-of-sample prediction failures, and unordered factors can exceed feasibility limits under unrestricted partition-based splitting. To avoid these issues, any SIC2 category with fewer than 20 observations in the full sample is collapsed into “Other” category.

4.6.3 Permutation Importance and Partial Dependence

Interpreting the RF model focuses on which variables are the most important predictors for issuance G-spread and how much the green label contributes. Variable importance is assessed using permutation importance, where a predictor is randomly permutated in the training data and the effect it has on prediction error. This increase in prediction error measures how much predictive information is attributed to that predictor while holding rest of the model constant (Fisher et al., 2019).

Partial dependence plots (PDP) are used to complement the importance by showing the relationship between selected predictors and predicted G-spreads. PDPs visualize a feature's average impact on the model's prediction which shows how the output changes as that specific feature varies (Hastie et al., 2009).

4.7 LASSO regression

To complement OLS and Random Forest, this thesis also includes the Least Absolute Shrinkage and Selection Operator (LASSO) as an additional predictive benchmark. LASSO is a linear regression model with an L_1 penalty on coefficients, which encourages sparsity by shrinking some coefficients exactly to zero. In practice, this provides a simple transparent benchmark that can improve out-of-sample prediction when the covariate set is high-dimensional or when multicollinearity makes OLS estimates unstable. Unlike Random Forest, LASSO does not model nonlinearities or interactions unless they are explicitly engineered, so it helps separate "gains from regularization" from "gains from nonlinearity."

Formally, given outcome $y_i = \log(\text{G-spread}_i)$and predictor vector x_i, LASSO estimates coefficients $(\alpha, \beta)$by solving

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha,\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where $\lambda \geq 0$ governs the strength of shrinkage. As λ increases coefficients are increasingly shrunk toward zero and some are set exactly to zero. The penalty therefore trades a small increase in bias for a potentially large reduction in variance, which can improve predictive performance out of sample (Tibshirani, 1996; Hastie et al., 2009).

LASSO provides an additional benchmark for this thesis. If LASSO improves over OLS the data benefits from linear shrinkage. If Random Forest improves further, the gains are consistent with nonlinearities and interactions.

4.8 Linking Inference and Prediction: The Green Label Across Models

One of the most important distinctions to make is that OLS (and LASSO) and RF address different empirical questions. OLS and LASSO are mainly an inference tool: they estimate the coefficient on the green label (β), which is the conditional difference in issuance G-spreads between green and non-green bonds. RF on the other hand is mainly a prediction tool: it focuses on out-of-sample accuracy with less assumptions than linear models. It is also used to evaluate the role and importance of the green label by showing if it improves performance and appears as an important factor in the model.

The best way to interpret the findings is to use the approaches together. If the linear models produce a stable negative β across all specifications and the RF results show that the green label ranks as an important feature support for a greenium is found. However, if the green coefficient does not stay stable and negative (e.g., survive controls) or it adds little predictive performance, then the existence of a greenium in this sample cannot be supported. Together these approaches offer a unified assessment if the green label matters for primary-market issuance G-spreads.

5 Data

In this chapter the data used in this thesis is presented. First, the data collection process is explained, including final sample formation and filtering criteria. Then the data limitations encountered are discussed. This is followed by the variable definition table. In this thesis some variables are constructed and aligned to issuance dates, this process is then reported. The chapter concludes with descriptive statistics for both green and non-green samples.

5.1 Data Collection

The data used in this thesis was collected from LSEG using the LSEG DataStream application. The data covers both green and non-green bonds issued between 5 January 2023 and 21 August 2025, all bonds used in this study are USD-denominated US-market non-financial corporate bonds. The sample includes both first-time and subsequent green and non-green bond issuances. To ensure comparability, only bonds with fixed coupon rates are included and all bonds in this dataset are callable. As such, no dummy for callability is needed.

For each bond the dataset includes the issuer name and ticker, issue date, amount issued, coupon rate, amount of coupons per year, Moody’s credit rating at issuance, green label, and sector classification. Furthermore issuance-time macro variables collected include the 10-year U.S. Treasury yield, 2-year U.S. Treasury yield, MOVE index, U.S. CPI, Corporate OAS index and the 3-month financial commercial paper rate and the 3-month Treasury bill rate. Variable construction and transformations are reported later.

Green-labelled bonds are identified using the LSEG DataStream green label classification that each green bond carries. The indicator equals one if the bond is classified as green and zero otherwise. This reduces false positives and is in line with the definition used in prior literature, only bonds that are credibly labelled green are treated as such (Gianfrate & Peri, 2019; Flammer, 2021). LSEG identifies green bonds using Climate Bonds Initiative

(CBI) data for “use of proceeds” and checks that this aligns with the GBP, as such the label is considered credible in this thesis.

5.2 Variable Selection and Construction

The variables used in this thesis are anchored to the Eurobond issuance spread framework in Gabbi and Sironi (2005). They included maturity, coupon structure, issue size, credit quality, sector composition, and issuance-time conditions. This thesis adopts that structure but extends the measurement of issuance conditions: rather than relying only on time dummies to proxy the pricing environment, issuance-time conditions are measured using observed macro variables aligned to the pricing date. Notably, Gabbi and Sironi used the natural logarithm of amount issued as a proxy for liquidity, in this thesis a secondary liquidity proxy is constructed.

The dependent variable is the issuance G-spread measured in basis points. Conceptually G-spread is the difference between the yield offered and a matched government benchmark yield. The central explanatory variable is the green label, this thesis asks whether a greenium exists once the models control for standard spread determinants. In literature green premium is commonly defined as a yield or spread differential between green bonds and conventional ones (Zerbib, 2019; Flammer, 2021). G-spread is modelled as log(g-spread) rather than G-spread in basis points because issuance G-spreads are positive and as such right-skewed. Log-transforming the variable means that the results must be interpreted as approximate percentage effects on G-spreads.

The bond characteristics and market variables that are used in the final dataset are explained below.

Amount Issued

Amount issued is expressed as the natural logarithm of the amount in millions of euros to control for outliers and reduce skewness. Wang et al. (2020) report that issue size increases credibility in the context of green finance. Issue size also directly relates to the bond's liquidity, which is priced. Bao et al. (2011) also link trading frictions and illiquidity to yields and spreads.

Coupon

Coupon is the fixed interest rate attached to each bond. Coupon levels reflect issuer's risk profile and investor demand (Fabozzi, 2012). Hull (2018) notes that higher coupon rates can signal higher perceived risk.

Coupon Amount

Similarly to coupon, increasing the payment frequency can play a role in issuance pricing.

Life

As the life of bond increases, so does the uncertainty: rates can move, credit conditions shift and issuer's situation can weaken (Elton et al., 2001). G-spreads at issuance can differ with maturity, by including maturity this thesis addresses that term structure matters and that the investment horizon is priced.

Credit rating

Credit quality captures expected default risk at issuance. Controlling credit quality is necessary because rating differences can dominate G-spread differences (Gabbi & Sironi, 2005). Issuers with better credit quality may also issue more green bonds, as such credit rating is commonly considered as one of the most important factors to include. In this thesis Moody's credit rating at issuance is transformed into an ordinal numeric scale where higher values indicate worse credit quality. Pietsch & Salakhova (2025) also show that default risk dominates spreads across ratings even when the non-default components change over time.

Sector (SIC2)

Sector is included to capture industry differences related to risk and issuance premia, it also addresses composition effects because green bond may not be issued evenly across sectors. SIC2 categories with fewer than 20 observations are collapsed into “Other” category.

Unlike Gabbi & Sironi (2006), this thesis uses issuance-time market conditions instead of dummy variables. The variables introduced below are included because G-spreads respond to changing market conditions. Collin-Dufresne et al. (2001) show that corporate credit spread changes are highly correlated across bonds, and that there is strong comovement between the bond spreads. Gilchrist and Zakrajšek (2012) identify excess bond premium, that captures the market-wide and time-varying credit risk premium. Furthermore, Brunnermeier and Pedersen (2009) show that funding constraints can increase market illiquidity, which motivates the addition of a funding/liquidity proxy.

Term Slope

Term slope is constructed as the 10-year U.S. Treasury yield minus the 2-year U.S. Treasury yield at the pricing date and then expressed in basis points. This captures the yield-curve slope and macro-state. The slope can also relate to risk appetite.

MOVE index

MOVE captures interest rate uncertainty and volatility conditions. Volatility tends to be associated with wider credit spreads via risk-bearing capacity. MOVE is considered as a rate volatility proxy.

U.S. CPI

This variable captures the inflation environment that can influence yields, policy expectations and risk premia.

OAS

Corporate OAS is considered as a broad credit condition proxy. It captures the market-wide compensation for credit risk at issuance and is intended to absorb the common component related to credit risk premia.

Funding/Liquidity Proxy

A funding/liquidity proxy is used to capture short-term liquidity stress that can spill over to issuance pricing through balance sheet constraints. The proxy is constructed as:

$$L_i = (CP_{3M} - TBill_{3M}) \times 100$$

where CP_{3M} is the 3-month financial commercial paper rate and $TBill_{3M}$ is the 3-month Treasury bill rate. It is then multiplied by 100 to express it in basis points.

Issuance-year fixed effects (year dummies)

Issuance-year fixed effects are used to capture year-level common shocks and shifts in regimes. These are only used in the last OLS model where macro controls are excluded. Below is the variable definition table, reported for full transparency.

Table 2. Variable definitions, sources, units, transformations, and timing alignment

Variable	Unit	Transformation	Timing alignment
G-spread	bps	Winsorized at upper 99.5th %, Log-transformed	At issuance
Green label	0/1	None	Observed at issuance
Life	years	Exclude if > 50 years	at issuance
Issue size	USD	Log (amount issued)	at issuance
Coupon rate	%	None	at issuance
Coupon Amount	ordinal	None	at issuance
Credit Rating	ordinal	AAA =1,..., C = 21	at issuance
Industry classification (SIC2)	categorical	Rare levels grouped "Other"	Fixed security attribute
SLOPE (10-2)	bps	$10Y - 2Y$	Aligned to issuance date
MOVE	index	None	Aligned to issuance date last
U.S. CPI	index	None	Monthly value mapped to issuance month
OAS bps	bps	None	Aligned to issuance date
Funding/liquidity proxy	bps	$(DCPF3M - DTB3) \times 100$	Aligned to issuance date

Notes: when the series is not reported on the exact issuance date (e.g., non-trading days), the most recent prior value is used to avoid look-ahead. CPI is aligned at monthly frequency, so the value corresponds to information available at issuance month, rather than ex-post revisions. For full credit rating transformation see Appendix 3.

5.3 Final Dataset

Before arriving at the final dataset, two cleaning rules are applied (as briefly explained earlier). Observations with implausibly long life are excluded by restricting life to 50 years or less. After this issuance G-spread is winsorized at the upper 99.5th percentile to avoid a handful of extreme values or artefacts influencing the results. Winsorizing changes the values in the extreme tail but does not reduce the number of observations. After these cleaning rules are applied the final sample contains no missing observations. For cases removed at each filtering step, see appendix 1.

Table 3. Sample composition by issuance year and green label

Issue year	Non-green	Green	Total	Green (%)
2023	1031	47	1078	4.36 %
2024	1453	46	1499	3.07 %
2025	1061	16	1077	1.49 %
Total	3545	109	3654	2.98 %

The final dataset contains 3654 bonds of which 109 (2.98%) are labelled as green. Importantly, massively imbalanced samples can be an issue for random forest because the green label is rarely seen by the model, which can distort its prediction. To address this, random forest models are estimated using weights. Each green bond is assigned a weight which is the same as the ratio of non-green to green bonds in the full sample (31). This is done to ensure that prediction errors for green bonds contribute similarly as non-green ones to the Mean Squared Error (MSE). The weighting is only applied to random forests, OLS results are estimated with robust standard errors.

5.4 Descriptive Statistics

The descriptive statistics for the overall and the split green vs. non-green samples are reported. Appendix 2 shows the full sample descriptive statistics for the dependent variable and all the numeric covariates. Table 4 compares green and non-green issues. Without including controls it is evident that green bonds are issued at tighter G-spreads than non-green ones (184 vs 207 bps) and offer slightly lower coupons. Green bonds also tend to be larger deals as seen by the amount issued and have better credit ratings on average.

When green bonds are issued, the MOVE index is slightly higher, indicating more interest rate fluctuation and that the outlook is more uncertain. Likewise, OAS is also higher, meaning that investors demand slightly more compensation for holding corporate bonds overall. On green issuance dates the gap between long- and short-term yields is smaller and the funding/liquidity proxy is lower pointing to somewhat easier funding conditions.

Table 4. Green vs non-green summary statistics with mean-difference tests (bps)

Variable	Green M (SD)	Not Green M (SD)	Green – Not Green	p
G-spread	183.72 (112.51)	206.87 (167.46)	-23.14	0.040*
Life	10.24 (8.01)	11.11 (25.04)	-0.87	0.323
Log Amount Issued	13.18 (0.38)	12.87 (1.97)	0.31	<.001**
Coupon	5.85 (1.15)	6.01 (1.57)	-0.16	0.159
Coupons Amount	2 (0)	2.01 (0.122)	-0.01	0.009**
Credit Rating	8.05 (2.99)	9.54 (3.51)	-1.49	<.001**
Slope BPS	-0.87 (21.91)	-11.94 (23.42)	11.07	<.001**
MOVE	111.08 (14.20)	105.96 (16.59)	5.12	<.001**
US CPI	0.03 (0.01)	0.03 (0.01)	0.00	0.007**
OAS bps	137.51 (24.14)	127.25 (24.45)	10.26	0.001***
Liq./funding Proxy	10.92 (6.90)	12.56 (7.37)	-1.64	0.0158*

** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

6 Empirical Results

This chapter reports the empirical evidence on whether green-labelled corporate bonds are associated with different issuance G-spreads in the USD primary market, and whether any such difference remains once standard determinants of corporate credit spreads are controlled for.

The results are presented using two approaches that complement each other. First is a stepwise OLS ladder that is used to estimate how the effect of green label changes as more controls are introduced. In literature this is common practice when the stability of a coefficient itself is informative (Hachenberg & Schiereck, 2018). After the OLS a Random Forest (RF) model is used to evaluate predictive performance and to report which variables the model relies on. As an additional benchmark LASSO is also used. The best performing OLS model and LASSO are then compared with the RF out-of-sample using multiple training/test splits to show that the results are robust and not due to a "lucky" split. As further robustness check the models are also evaluated out-of-time.

6.1 Evidence on Issuance G-Spreads and Log-Transform

Before moving to the models the distribution of issuance-time G-spreads is inspected. Alongside the issuance G-spreads, some bivariate relationships are also shown. Figure 4 shows that G-spreads in this sample are strongly right-skewed, meaning that most bonds price at relatively low spreads and a smaller amount of bonds with high G-spreads creates a long upper tail even after winsorization. This is consistent with literature where spreads are found to vary widely across credit risk and market conditions, as such tail observations remain economically meaningful and not just noise (Collin-Dufresne et al., 2001).

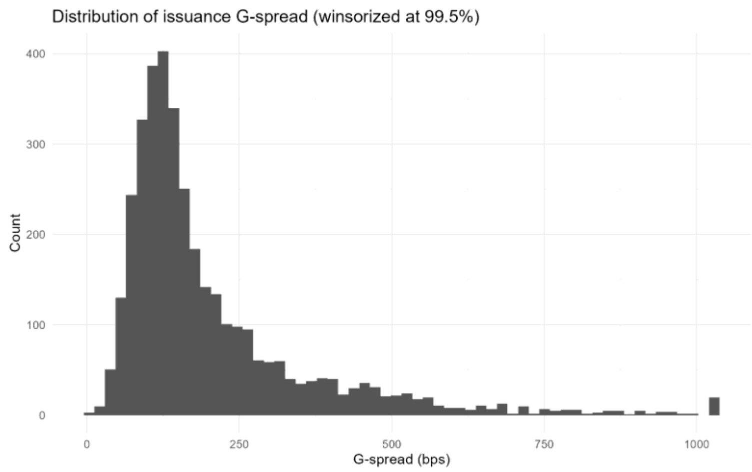

Figure 4. Distribution of issuance G-spread (winsorized).

Figure 5 compares green and non-green bonds without any controls. The plots show that the distributions have substantial overlap and that the G-spread distribution is wider with higher spreads. This hints that variance of G-spreads is not constant across the sample.

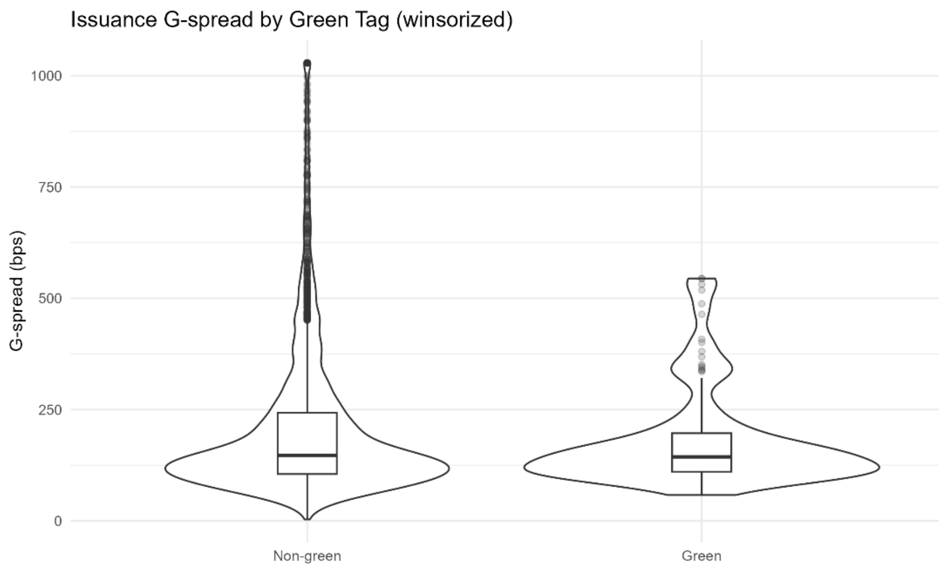

Figure 5. Distribution of issuance G-spread green vs. non-green (winsorized).

Figure 6 shows how issuance G-spreads vary across key covariates. The main takeaway is dispersion, spreads are much more variable at higher levels. In short, this reinforces what is already evident: G-spreads are right-tailed and show variance that depends on the level of DV.

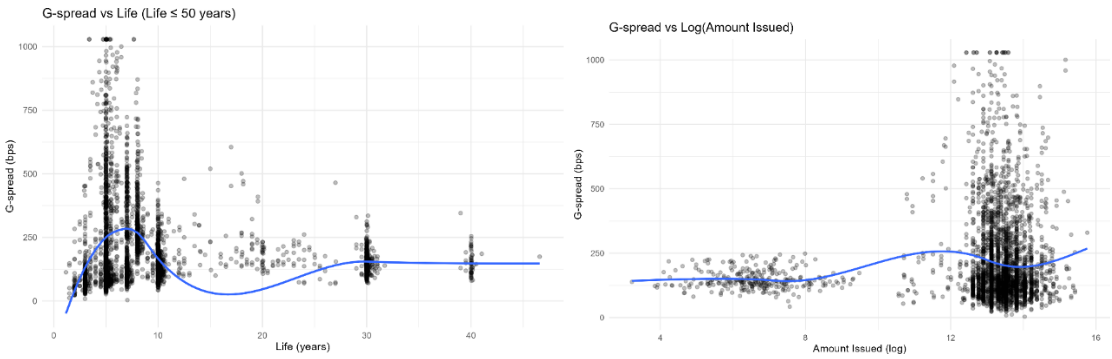

Figure 6. Issuance G-spread vs. Life and Amount Issued (log)

These descriptives motivate log transforming the dependent variable. The distribution is clearly right-tailed, and the scatterplots show that variance increases with the level of the G-spread. A log-transform is the standard response to this. It compresses the upper tail, reduces the influence of extreme observations. Wooldridge (2020) explicitly notes that using a log functional form for the dependent variable is a common fix when heteroscedasticity rises with the level of DV.

Following this, the OLS and RF use log-transformed version of G-spread as the dependent variable, which is not an uncommon practice (Gilchrist & Zakrajšek, 2012). This is done because logging the DV makes error variance more stable and this improves fit and inference for regressions (Wooldridge, 2020). Once the DV is logged, coefficients are interpreted as follows: for a one-unit increase in a regressor the expected spread changes by approximately 100 × β percent, holding other covariates fixed. To express this effect in basis points, the implied proportional change is back-transformed as $\exp(\beta\Delta x) - 1$ and then multiplied by a representative baseline spread level S_0, so $\Delta\text{bps} = S_0(\exp(\beta\Delta x) - 1)$ (for a dummy regressor, $\Delta x = 1$); in this thesis S_0 is set to 147.5 bps.

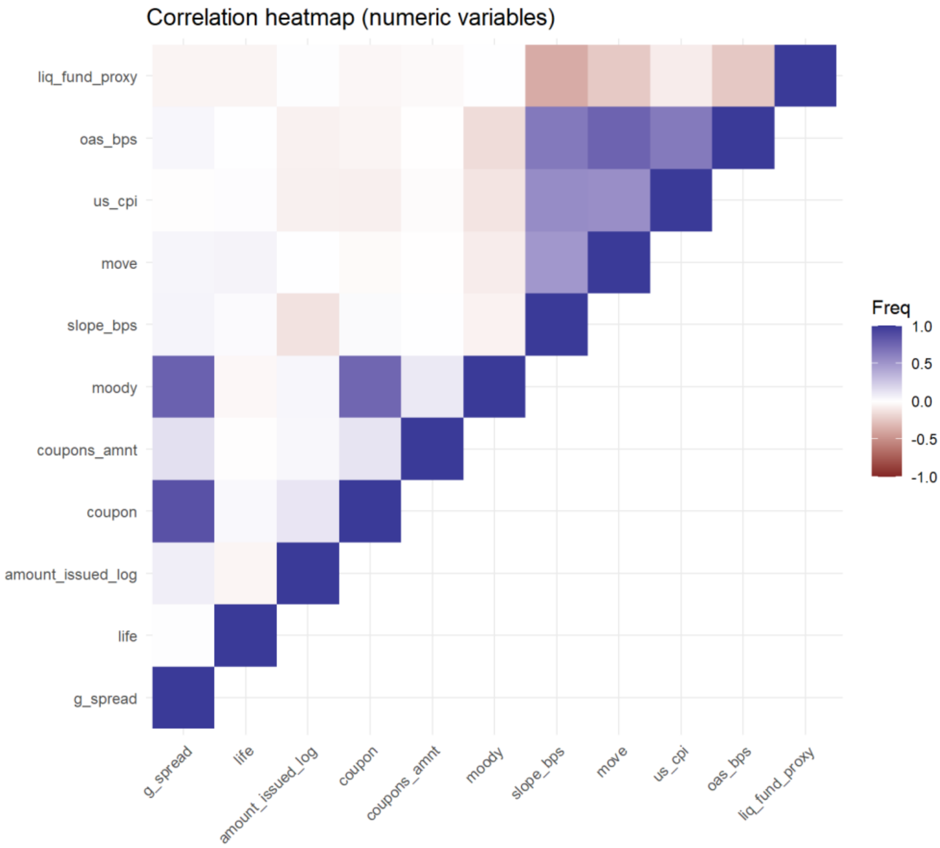

Figure 7. Correlation heatmap of key continuous predictors

Figure 7 reports pairwise correlations between the continuous predictors (and the DV for reference). The strongest correlations with issuance G-spreads are observed for coupon and credit rating, which is consistent with the view that terms and default risk dominate (Gabbi & Sironi, 2005); Elton et al., 2001). The heatmap also shows that several issuance-time market proxies covary, which is expected when multiple variables are used to capture market conditions.

6.2 OLS Ladder Results: The Green Coefficient Across Specifications

This section reports how the associations between green label and issuance G-spreads changes as common determinants are introduced in blocks. Elton et al. (2001) and Gilchrist & Zakrajšek (2012) showed that cross-sectional spreads are related to contractual terms and credit risk and the greenium debate raises the question if green issuance differs systematically from non-green one.

Table 5. OLS ladder estimates with HC1 robust standard errors (M0 to M5)

	Model 0	Model 1	Model 2	Model 3	Model 4	Model 5
Green Label	−0.020	0.044	0.164***	0.143***	0.105***	0.114***
	(0.067)	(0.028)	(0.028)	(0.027)	(0.027)	(0.029)
Life		0.003***	0.011***	0.011***	0.011***	0.011***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Amount Issued		−0.020***	−0.009**	−0.010	−0.005	−0.002
(log)		(0.004)	(0.003)	(0.005)	(0.004)	(0.004)
Coupon		0.363***	0.199***	0.190***	0.175***	0.172***
		(0.009)	(0.013)	(0.012)	(0.012)	(0.012)
Coupons		0.035	0.009	0.005	0.004	0.021
Amount		(0.043)	(0.036)	(0.036)	(0.029)	(0.028)
Credit Rat.			0.099***	0.100***	0.110***	0.109***
			(0.006)	(0.006)	(0.006)	(0.006)
Slope bps					−0.000	
					(0.000)	
MOVE					−0.000	
					(0.001)	
US CPI					−0.979	
					(1.171)	
OAS bps					0.006***	
					(0.000)	
Liquidity Proxy					0.001	
					(0.001)	
FE 2024						−0.232***
						(0.016)
FE 2025						−0.260***
						(0.019)
N	3654	3654	3654	3654	3654	3654
R^2	0.000	0.719	0.826	0.837	0.875	0.864
R^2 Adj.	0.000	0.718	0.826	0.834	0.873	0.861

** p < 0.05, ** p < 0.01, *** p < 0.001.*

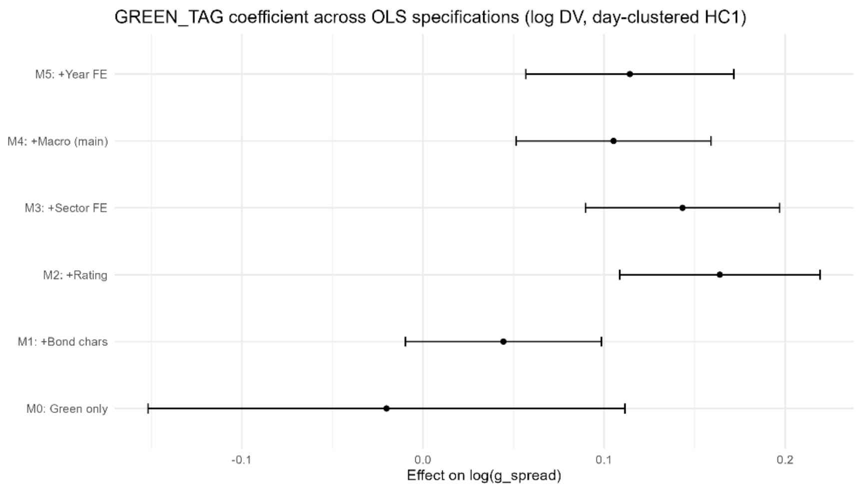

Figure 8. Green Label coefficient across specifications

Table 5 reports the nested OLS ladder with log-transformed G-spread as the outcome variable. The point of the ladder is simple, it shows how the estimated association between the green label and G-spread changes as the typical control blocks are introduced. The first model (model 0) is the unconditional one, which shows if the green label without controls determines G-spreads in a meaningful way. In model 0 the green coefficient is small and not statistically significant (β = -0.020, SE = 0.067), showing that without controls there is no clear difference in issuance spreads between green and non-green bonds in this sample.

Adding bond characteristics turns the green coefficient positive, but it remains statistically insignificant (β = 0.044, SE = 0.028). This shows that unconditional comparisons between green and non-green bonds are not very informative, because spreads are linked to bond characteristics. Key shift occurs once credit rating is added to the mix (model 2). Once credit is added the green coefficient increases drastically and becomes strongly statistically significant (β = 0.164, SE = 0.028, $p < .001$). This corresponds to approximately an 18 % difference in issuance G-spreads for green labelled bonds, while holding bond characteristics and credit rating constant. Adding sector controls in model 3 does not change the estimate much (β = 0.143, SE = 0.027, p = < .001).

Once issuance time market controls are included (model 4) the green coefficient declines but stays highly significant (β = 0.105, SE = 0.027, p < .001), which implies approximately 11 % higher G-spreads for green-labelled issues in this sample. Switching from macro conditions to year-fixed effects produces very similar estimate (β = 0.114, SE = 0.029, p < 001), suggesting that the macro controls capture the issuance-time conditions relatively well.

The ladder shows that the sign and strength of the green label is very control dependent. With no conditions it does not robustly differ from zero yet becomes consistently and strongly positive once credit quality is accounted for. This aligns with prior literature, where it is found that the green premia are control dependent (Flammer, 2021; Hachenberg & Schiereck, 2018).

Model fit also evolves as controls are introduced. Explanatory power jumps massively when bond characteristics are added. Once credit quality is controlled another jump occurs to R^2 = 0.826 in model 2. Interestingly sector controls do not produce a large increase, in model 3 R^2 = 0.837. The best overall fit is produced by model 4 that adds market-wide conditions into the model, reaching R^2 = 0.875. Model 4 is therefore chosen for out-of-sample (OOS) comparison with Random Forest and LASSO.

The ladder makes two points clear: most of the explanatory power comes from standard bond characteristics and credit quality. Conditional on those, the green label is associated with a statistically significant and economically meaningful spread premium in this sample.

6.3 Preferred OLS Specification: Coefficient Estimates and Model Fit

Table 6. OLS coefficients (M4) with robust standard errors

Coefficient	Estimate	SE	t	p
Green Label	0.105	0.027	3.837	< .001***
Life	0.011	0.001	21.129	< .001***
Log Amount Issued	-0.005	0.004	-1.140	0.254
Coupon	0.175	0.012	14.761	< .001***
Coupons Amount	0.004	0.029	0.128	0.898
Credit Rating	0.110	0.006	19.647	< .001***
Slope bps	0.000	0.000	-0.713	0.476
MOVE	0.00	0.001	-0.355	0.723
US CPI	-0.098	0.117	-0.835	0.404
OAS bps	0.006	0.000	13.102	<. 001***
Liq./funding Proxy	-0.146	0.198	-0.736	0.462

** p < 0.05, ** p < 0.01, *** p < 0.001.*

In model M4 the coefficient on the green label is β = 0.105 (SE = 0.027, t = 3.84, p < .001) as reported earlier. Because the dependent variable is log(G-spread) this implies an expected spread differential of $e^{0.105} - 1 \approx$ 11.1% Using the thesis baseline S_0 = 147.5bps, this corresponds to approximately +16.4 bps, with a 95% confidence interval of 7.8, 25.4 bps

Table 6 shows the coefficients and they are analyzed in more detail. Life is also strongly related to G-spreads at issuance β = 0.011 (SE = 0.001, t = 21.13, p < .001), which corresponds to approximately a 1.1 % increase in G-spreads for each additional year of life (about 1.6 bps per year at S_0). Coupon is also positive and significant (β = 0.175, SE = 0.012, t = 14.76, p < .001) suggesting that higher coupon is associated with significantly higher G-spreads, approximately 28.2 bps increase for each 1% increase in coupon. Credit quality, unsurprisingly, is one of the dominant estimators. Credit rating has a large positive coefficient (β = 0.110, SE = 0.006, t = 19.65, p <.001), showing that as credit rating worsens by 1 notch the required spread increases by 11 %, which corresponds to 17.2 bps per notch at S_0 = 147.5 bps. This is consistent with prior literature (Gabbi & Sironi, 2005). Overall, these bond characteristics are consistent with bond literature as they showcase that as credit rating worsens the spreads investors demand increases.

Similarly, as the heatmap showed, credit rating and coupon were correlated, indicating that lower-rated issuers typically must give higher coupons.

The green label is interesting in this sample, because typically less credible issuers must offer higher spreads. Yet in this sample the descriptive statistics showed that green bonds enjoy slightly higher credit ratings, but the label is still associated with higher G-spreads. This suggests that green labelled bonds even when they are issued by more credible issuers may have to offer a higher concession at issuance.

Of the remaining controls only OAS in bps is positive and has a significant effect (β = 0.006, SE = 0.000, t = 13.10, p < .001). Suggesting that market-wide credit conditions at issuance shift the G-spreads meaningfully (e.g., a 10 bps increase in OAS corresponds to $e^{0.006\times10} - 1 \approx$ 6.2% higher G-spreads, and a 100 bps increase to $e^{0.006\times100} - 1 \approx$ 82% higher spreads). The rest of controls do not show a statistically significant effect.

6.4 Regression Diagnostics and Robustness Checks

Diagnostics are reported for two practical reasons. First, because cross-sectional issuance G-spreads are rarely homoscedastic: pricing errors tend to be worse for issues that offer higher spreads and they can vary between bond types. Heteroscedasticity invalidates conventional standard errors; inference is based on HC1 standard errors that are clustered by issuance date. Breusch-Pagan test is used to report heteroscedasticity (Breusch & Pagan, 1979). Variance inflation factors (VIFs) are also reported to check that multicollinearity is not severe enough to make coefficient estimates unstable.

Residual vs. fitted and scale-location plots are also reported to show that the log-transformation reduces variance that is level-dependent and that there are less systematic patterns in residuals.

6.4.1 Heteroskedasticity

Table 7. Breusch–Pagan test for heteroskedasticity (M4)

Test	Statistic	df	p
Breusch-Pagan test	$\chi^2 = 99.764$	11	0.009

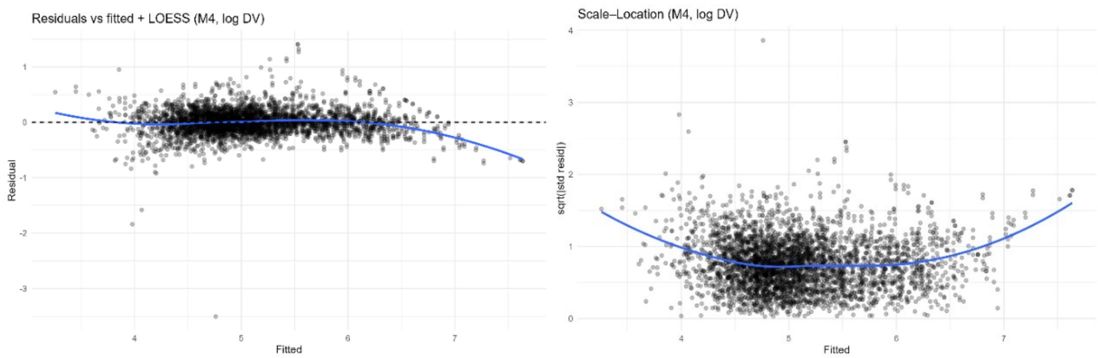

Figure 9. OLS residuals versus fitted values and Scale-location

Breusch-Pagan test shows that the residual variance is not constant, $\chi2$ (11) = 99.76, p < .001, rejecting homoscedasticity. This means that the model's errors are systematically larger for some parts of the sample. For this reason HC1 standard errors are used. Standard errors are also clustered by issue date to avoid treating same day observations as independent. This is standard practice when observations share common time-related shocks (Campbell & Taksler, 2003).

Figure 9 shows that after log-transforming the DV, the residuals plot and scale-location show a flatter pattern of variance, around the center of the fitted range. Some curvature still remains, but log-transform greatly reduces the extremes. This makes the data closer to what linear models assume.

6.4.2 Multicollinearity

Table 8. Variance inflation factors (VIF) for M4 predictors

Variable	VIF
OAS bps	3.648
Credit Rating	2.648
MOVE	2.542
Coupon	2.482
Slope bps	2.114
US CPI	1.922
Liq./funding Proxy	1.218
Life	1.118
Log Amount Issued	1.063
Coupons Amount	1.018
Green Label	1.016

VIFs are reported in table 8 above. Overall, the VIFs do not point to severe multicollinearity in the core regressors. As a common rule of thumb values between 1 and 2 indicate no issue with multicollinearity, and values between 2 and 5 indicate mild multicollinearity. In this case mild multicollinearity is present, but as it is not extreme (VIF > 5) the results can be interpreted with confidence (O'Brien, 2007; Wooldridge, 2016).

6.5 Heterogeneity

In the baseline specification (M4) the coefficient of the green label is β = 0.105 (SE = 0.027, t = 3.84, p < .001), implying a +11.1% higher issuance G-spread for green-labelled bonds, conditional on the full control set. This subsection shows whether that spread differential is uniform, or if it varies in a way that that would be consistent with state-dependent pricing and segmentation. Variation would not be surprising as G-spreads are known to be shaped by credit risk and market conditions (Collin-Dufresne et al., 2001). Likewise, the greenium itself is documented to be heterogeneous across bonds and over time rather than a single constant premium. (Zerbib, 2019; Fatica et al., 2021).

Two complementary checks are used. M4 is re-estimated with subsamples (rating buckets, maturity and tight vs. stressed regimes). Then interaction models on the full sample are used to test whether green differs significantly across segmentations.

6.5.1 Heterogeneity by ratings

Table 9. Rating bucket estimates (green coefficient)

			95% Confidence Interval				
Bucket	β	SE	Lower	Upper	t	p	n
Low	0.061	0.055	-0.047	0.170	1.11	0.266	1038
Mid	0.054	0.050	-0.044	0.153	1.09	0.277	1206
High	0.078	0.031	0.018	0.139	2.53	0.011*	1410

** p < 0.05, ** p < 0.01, *** p < 0.001.*

Table 9 reports that the green coefficient remains positive in every rating bucket. For high quality issues the green estimate is β = 0.078 (SE = 0.031, t = 2.53, p = .011), with a 95% CI (0.018, 0.139). This corresponds to an 8.2% higher expected G-spread for green issues. Using the chosen representative spread level S_0 = 147.5 bps that translates to approximately 12 bps 95% CI (2.7, 20.6) for green bonds in the highest quality tercile.

In the mid bucket, the estimate is smaller and statistically weak (β = 0.055, SE = 0.050, t = 1.09, p = .277). Interestingly, the CI crosses to slightly negative values 95% CI (-0.044, 0.153). Transforming back to bps, on the representative spread this correspond to approximately 8.3 bps with 95% CI (-6.3, 24.4). In the low quality bucket the estimate is similar to the mid bucket, β = 0.061, SE = 0.055, t = 1.11, p = 0.266) and 95% CI (-0.047, 0.170). These imply 9.3 bps (-6.8, 27.3) higher spreads at the representative level; however, these estimates are not statistically significant.

Table 10. OLS interaction model: Green × Rating bucket

Term	β	SE	t	p
Intercept	2.983	0.096	31.10	< .001***
Green label (high)	0.112	0.034	3.29	.001**
Rating bucket: Mid	0.329	0.019	17.04	< .001***
Rating bucket: Low	0.779	0.043	18.09	< .001***
Green × Mid	−0.043	0.061	−0.70	.482
Green × Low	−0.124	0.052	−2.40	.016*

** p < 0.05, ** p < 0.01, *** p < 0.001.*

Table 10 tests the same idea inside a pooled model, which allows the green effect to differ by rating bucket. The green coefficient in the baseline bucket (High quality) is β = 0.112, SE = 0.034, t = 3.29, p = .001. The interaction terms are negative, meaning the green differential is smaller outside the safest tercile: for Mid, the interaction is β = −0.043, SE = 0.061, t = −0.703, p = .482 and for Low quality β = −0.124, SE = 0.052, t = −2.40, p = .016.

In the subsample estimates the green coefficient remains positive in each rating bucket, but the estimates for the mid and low buckets are imprecise and their confidence intervals cross zero. In the pooled interaction specification, the baseline (high quality) green effect is clearly positive, while the interaction terms are negative. This implies that the green differential becomes much smaller for riskier issues and can turn slightly negative in the lowest quality tercile. Among the highest quality issuers, the green effect is clearly positive and economically meaningful, while for riskier issues the confidence intervals widen and cross across zero. A natural interpretation is that when credit risk is low, bonds are closer to substitutes and any pricing difference linked to the label is easier to detect. Once credit risk is high, the green component becomes harder to separate from the larger dispersion in G-spreads.

6.5.2 Heterogeneity by Maturity

Table 11 splits the sample into terciles based on life (short, medium, long) and re-estimates the M4 model within each bucket. The green coefficient is clearly positive in shorter maturities and becomes small and insignificant at longer lives.

Table 11. Maturity bucket estimates (green coefficient)

Bucket	β	SE	95% Confidence Interval Lower	Upper	t	p	n
Short	0.213	0.056	0.103	0.323	3.80	< .001***	1218
Mid	0.111	0.046	0.020	0.202	2.39	0.017*	1315
Long	-0.002	0.041	-0.082	0.079	-0.04	0.970	1121

** p < 0.05, ** p < 0.01, *** p < 0.001.*

In the Short bucket, the green coefficient is β = 0.213, SE = 0.056, t = 3.80, p < .001, 95% CI (0.103, 0.323). This corresponds to a 23.7% spread differential (95% CI [10.9%, 38.2%]). At a representative spread level of S_0 = 147.5 bps, that is approximately +35.0 bps (95% CI [16.1, 56.4]). In the Medium bucket, the estimate remains positive but is smaller: β = 0.111, SE = 0.046, t = 2.39, p = .017, 95% CI [0.020, 0.202]. This implies 11.8% (95% CI [2.0%, 22.4%]), or about 17.4 bps at S_0 (95% CI [3.0, 33.0]). In the Long bucket, the estimate is essentially zero and imprecise: β = −0.002, SE = 0.041, t = −0.04, p = .970, 95% CI (−0.082, 0.079). In percent terms that is −0.2% (95% CI [−7.9%, 8.2%]), and roughly −0.3 bps at S_0 with interval of −11.6 to 12.4 bps.

Table 12. OLS interaction model: Green × maturity bucket

Term	β	SE	t	p
Intercept	2.077	0.089	23.36	< .001***
Green label (short)	0.142	0.048	2.95	.003**
Maturity bucket: Mid	0.131	0.012	10.50	< .001***
Maturity bucket: Long	0.249	0.012	19.99	< .001***
Green × Mid	0.006	0.063	0.088	.929
Green × Long	-0.122	0.061	−2.01	.045*

** p < 0.05, ** p < 0.01, *** p < 0.001.*

Using the interaction model, the green effect in the short bucket is β = 0.142, SE = 0.048, t = 2.95, p = .003. The key results is the long interaction term: β = −0.122, SE = 0.061, t = -2.01, p = .045. Showing that the green differential is statistically significantly smaller in long maturities compared to short maturities.

These results show that any conditional green effect that exists in the sample is concentrated in the short and medium maturities. The effect fades out at longer maturities where duration and uncertainty related factors become more pronounced (Duffee, 1998; Collin-Dufresne et al., 2001).

6.5.3 Heterogeneity in Tight vs. Stressed Markets

To test whether the conditional green spread differential depends on broader market conditions, the sample is split into "tight" and "stressed" regimes using two standard issuance-time proxies: (i) OAS, which captures overall credit-risk pricing, and (ii) MOVE, capturing interest-rate uncertainty and volatility. Tight refers to below-median OAS/MOVE, while stressed refers to above-median OAS/MOVE.

Table 13. Green coefficient by OAS state

			95% Confidence Interval				
Group	β	SE	Lower	Upper	t	p	n
Stressed	0.060	0.035	-0.009	0.129	1.70	0.089	1823
Tight	0.174	0.038	0.101	0.249	4.63	< .001***	1831

** p < 0.05, ** p < 0.01, *** p < 0.001.*

Table 14. Green coefficient by MOVE state

			95% Confidence Interval				
Group	β	SE	Lower	Upper	t	p	n
Stressed	0.064	0.034	-0.003	0.130	1.88	0.060	1816
Tight	0.188	0.038	0.113	0.263	4.92	< .001***	1838

** p < 0.05, ** p < 0.01, *** p < 0.001.*

Tables 13 and 14 show the same pattern across both proxies. Under tight regimes the estimated green effect is large and positive. Using OAS as the regime proxy the tight estimate is β = 0.175, SE = 0.038, 95% CI (0.101, 0.249), p < .001. This corresponds to roughly 19.1% higher G-spreads for green-labelled issues, which transforms to approximately 28 bps at the S_0 level. Using MOVE as the proxy yields nearly identical results. β = 0.188, SE = 0.038, 95% CI (0.113, 0.263), p < .001, which is approximately 20.7% and 31 bps at S_0 level.

Under stressed regimes, the estimated green effect is noticeably smaller and statistically weaker. Using OAS as the regime proxy the stressed estimate is β = 0.060 (SE = 0.035, 95% CI: -0.009 to 0.129, p = 0.090). This corresponds to roughly 6.2% higher spreads for green-labelled issues and translates to about 9 bps at the S_0 level, with a 95% CI of roughly -1.4 to 20.4 bps. Using MOVE as the proxy yields a very similar stressed-regime estimate: β = 0.064 (SE = 0.034, 95% CI: -0.003 to 0.130, p = 0.060), which corresponds to approximately 6.6% and about 10 bps at the S_0 level, with a 95% CI of roughly -0.4 to 20.5 bps.

Table 15. OLS interaction model: Green × Tight market states

	Term	β	SE	t	p
OAS State	Green (baseline = stressed)	0.074	0.035	2.12	.034*
	Green × Tight (low OAS)	0.060	0.052	1.16	.246
MOVE State	Green (baseline = stressed)	0.084	0.035	2.41	.016*
	Green × Tight (low OAS)	0.091	0.048	1.88	.060

** p < 0.05, ** p < 0.01, *** p < 0.001.*

In the pooled models the Green × Tight term is positive under both proxies, indicating a larger green differential in tight conditions. Notably, the estimates are not statistically significant at α = .05.

The results are consistent with the idea that when markets are stressed, G-spreads are largely driven by economy-wide repricing of credit and liquidity risk rather than by other cross-sectional factors. A large share of the variation in these conditions reflects shifts in default-risk compensation and time-varying credit risk premia that move together across issuers (Gilchrist & Zakrajšek, 2012). In that setting, determinants that can matter in quieter periods are harder to separate from the dominant component of G-spreads, consistent with evidence that much of the variation is accounted for by systematic factors when spreads move sharply (Collin-Dufresne, Goldstein, & Martin, 2001).

Under stress markets also tend to feature a pronounced liquidity dimension. Trading becomes more costly, immediacy is priced more highly, and dealers' and investors' balance-sheet capacity tightens, all of which raises compensation for holding illiquid credit and increases dispersion, especially in parts of the distribution where uncertainty is already high (Longstaff et al., 2005; Brunnermeier & Pedersen, 2009). Empirically, corporate bond illiquidity is related to transaction costs and spread levels, and liquidity deteriorations are particularly visible in turbulent periods, which makes precise attribution of small “label” effects more difficult (Bao et al., 2011; Dick-Nielsen et al., 2012).

This is also consistent with how the green bond literature interprets pricing differentials. The standard view is that any green premium (or discount) reflects investor demand, segmentation, and clientele effects rather than a mechanical change in expected default risk (Baker et al., 2018; Zerbib, 2019). If that demand channel varies over time, it is most likely to be detectable when aggregate spread movements are smaller and dispersion is lower (Ehlers & Packer, 2017). Recent work documenting that the greenium is not constant and depends on market conditions and market structure is consistent with this interpretation (Pietsch & Salakhova, 2024). Overall, the state splits are naturally read as: whatever green differential is present shows up primarily in normal conditions, while in stressed regimes the pricing of credit risk, liquidity, and intermediary constraints dominates issuance G-spreads, and the green estimate becomes smaller and less precisely estimated (Gilchrist & Zakrajšek, 2012; Longstaff et al., 2005; Brunnermeier & Pedersen, 2009).

6.6 Predictive Performance: OLS vs. LASSO vs. Random Forest

To compare the RF, the best performing OLS model (model 4) and LASSO, they are evaluated on the same OOS set. This is done by first splitting the data into 75/25 train-test split to train the models on the same data, after which the models are tested on the split that was “held-out” from the training. This shows how the models perform in a realistic setting, where data is not seen before. Because the models are estimated with log G-spread as the DV, MAE and RMSE are reported in log points. Log-point errors are interpreted as multiplicative errors in the spread level: if the prediction error is $e = \log(\hat{S}) - \log(S)$, then $\hat{S} = S\, e^{e}$. A MAE of m corresponds to error of approximately $100(e^{m} -$

$1)$ % (and similarly for RMSE). The log-scale errors are reported because they match the estimation scale.

Table 16. Test-set performance comparison for OLS and Random Forest

Model	RMSE	MAE	R^2
OLS (M4)	0.251	0.181	0.858
LASSO	0.251	0.180	0.858
Random Forest	0.152	0.099	0.948

Table 16 presents the test-set results for OLS (M4), LASSO, and RF on the same OOS split. The result is clear: RF is substantially more accurate than the linear benchmarks. OLS (M4) achieves RMSE = 0.251 and MAE = 0.180 with test R^2 = 0.858. LASSO performs almost identically on this split (RMSE = 0.251, MAE = 0.180, R^2 = 0.858), suggesting that regularization adds little predictive value beyond the baseline linear specification given this covariate set. RF, on the other hand, reduces errors to RMSE = 0.152 and MAE = 0.099 and increases test R^2 to 0.948. Relative to OLS, this corresponds to approximately a 39% reduction in RMSE and a 45% reduction in MAE. As the error metrics complement each other, the outcome is reassuring: the gains do not appear to be driven only by a small number of extremes.

To be sure that these results do not result from a single "lucky" split, performance is measured across 100 repeated splits. Table 17 summarizes these results. Across the 100 repetitions, OLS has mean RMSE = 0.279 (SD = 0.082), mean MAE = 0.180 (SD = 0.005) and mean test R^2 = 0.829 (SD = 0.077). LASSO performs very similarly with mean RMSE = 0.283 (SD = 0.080), mean MAE = 0.183 (SD = 0.008) and mean test R^2 = 0.826 (SD = 0.075). RF remains clearly better on average with mean RMSE = 0.152 (SD = 0.017), mean MAE = 0.099 (SD = 0.004) and mean test R^2 = 0.947 (SD = 0.010). Overall, the repeated-split results confirm that the predictive advantage of RF is not a one-off outcome, while LASSO does not materially improve on the baseline linear model. This result is consistent with Makariou et al. (2021) who found that Random Forest performs better in the catastrophe bond markets. Overall, these results align with the consensus that machine learning approaches perform better than traditional models in financial research (Kim et al., 2021).

Table 17. Repeated split summary

Model	RMSE M (SD)	MAE M (SD)	R^2 M (SD)
OLS (M4)	0.279 (0.082)	0.1801 (0.005)	0.829 (0.077)
LASSO	0.283 (0.080)	0.1831 (0.008)	0.826 (0.075)
Random Forest	0.152 (0.017)	0.0990 (0.005)	0.947 (0.010)

Table 18. Mean differences repeated splits

			95 % Confidence Interval		
	Comparison	Mean difference	Lower	Upper	p
RMSE	RF – OLS	-0.127	-0.186	-0.068	< .001***
	LASSO – OLS	0.004	-0.002	0.010	0.201
MAE	RF – OLS	-0.081	-0.084	-0.078	< .001***
	LASSO – OLS	0.003	-0.002	0.008	0.228
R^2	RF – OLS	0.119	0.064	0.174	< .001***
	LASSO – OLS	-0.002	-0.007	0.001	0.173

** p < 0.05, ** p < 0.01, *** p < 0.001.*

To assess whether performance differences are systematic, paired differences in evaluation metrics are computed across the 100 splits. RF significantly outperforms OLS: the mean paired difference in RMSE (RF − OLS) is −0.127 with a 95% CI (−0.186, −0.068), $p < .001$ and the mean paired difference in MAE is −0.081 with a 95% CI (−0.084, −0.078), $p < .001$. Likewise, predictive fit improves: the mean paired difference in test R^2 is 0.119 with a 95% CI (0.064, 0.174), $p < .001$. In contrast, LASSO does not differ meaningfully from OLS across splits. Taken together, the evidence suggests that the main performance gains come from RF's ability to capture nonlinearities and interactions rather than from linear regularization alone.

6.6.1 Out-of-time evaluation

Out-of-time (OOT) evaluation is included as a robustness check, because when market data is time-ordered and regimes shift, randomly mixing observations can make a model look more stable than it in reality is. In forecasting, a standard recommendation is to validate models OOT so each test period is predicted using only information available earlier in time (Hyndman & Athanosopoulos, 2021). In finance this is even more important, because dependence on overlapping information creates leakage; OOT evaluations are recommended to get a more credible estimate of the predictive performance (López de Prado, 2018; Bergmeir et al., 2018). In this thesis, the OOT train window is 2023-2024 and test window is 2025.

Table 19. Out-of-time test performance

Model	RMSE	MAE	R^2
OLS	0.281	0.203	0.855
LASSO	0.323	0.234	0.846
RF	0.229	0.163	0.895

Table 20. Out-of-time performance in bps

Model	RMSE	MAE	R^2
OLS	47.9	33.2	0.855
LASSO	56.2	38.9	0.846
RF	38.0	26.1	0.895

Table 19 shows the test performances for all three models OOT. Relative to the random splits OOS results, RF suffers drastic performance deterioration which is expected due to restricting leakage.

In the OOT window, RF remains as the best performing model, but with higher errors: RMSE = 0.229, MAE = 0.163 and R^2 = 0.895. The OLS model (M4) also weakens OOT (RMSE = 0.281, MAE = 0.203, R^2 =0.855), while LASSO performs the worst (RMSE = 0.323, MAE = 0.234, R^2 = 0.846). Table 19 shows the same performances but in bps at the S_0 level for interpretability.

Overall, the OOT results show that machine learning methods can generalize better than linear benchmarks in this setting. However, the performance deterioration is harsher for them when evaluated OOT, suggesting leakage when evaluation is done purely OOS, not OOT.

6.7 Random Forest Tuning and Model Diagnostics

Hyperparameters are chosen using cross-validation within the training sample as explained previously in the methods section. Therefore, the selected configuration reflects performance that is stable across multiple resamples and not resulting from a specific "lucky" partition (Stone, 1974). Cawley and Talbot (2010) note that this is important because tuning itself is a form of model selection: if the model is continuously recalibrated on limited data the reported accuracy would be biased. To see why these hyperparameters were chosen to be tuned, see methods.

Table 21 shows the tuning grid results. The best configuration is mtry = 8 and min.node.size = 5, with cross-validated RMSE = 0.168, MAE = 0.105, and R^2 = .934. The grid shows two important features: performance is not drive by a single "lucky" configuration. This is shown by multiple "neighboring" configurations delivering very similar accuracies, suggesting that the tuning results are not sensitive to small changes. The other important note is that increasing minimum node size worsens fit consistently, in practice this shows that issuance G-spreads are not linearly formed and they benefit from allowing more nonlinear structure.

Cawley and Talbot (2010) emphasize that the best way to protect against overfitting is simply testing OOS and reporting final model performance on that held-out test data. In the present study this would show up as large deviations between cross-validated training performance and test performance or as instability between train-test splits. However, evaluation here shows that the RF's advantage over OLS is not just a "one-off" outcome, but persistent performance gain.

Table 21. Full cross-validation grid results for RF hyperparameters

mtry	min.node.size	RMSE	MAE	R^2
2	5	0.193	0.126	0.921
4	5	0.171	0.108	0.933
5	5	0.168	0.106	0.934
7	5	0.168	0.105	0.934
8	5	0.168	0.105	0.934
10	5	0.169	0.106	0.933
11	5	0.170	0.106	0.932
13	5	0.172	0.108	0.931
2	10	0.200	0.133	0.916
4	10	0.176	0.112	0.929
5	10	0.173	0.110	0.931
7	10	0.171	0.108	0.932
8	10	0.171	0.109	0.931
10	10	0.173	0.109	0.930
11	10	0.174	0.109	0.929
13	10	0.176	0.111	0.927
2	20	0.213	0.143	0.907
4	20	0.185	0.121	0.922
5	20	0.181	0.117	0.925
7	20	0.179	0.115	0.926
8	20	0.178	0.115	0.926
10	20	0.180	0.116	0.924
11	20	0.181	0.116	0.923
13	20	0.183	0.117	0.921
2	40	0.230	0.156	0.893
4	40	0.197	0.132	0.913
5	40	0.192	0.128	0.916
7	40	0.189	0.124	0.917
8	40	0.189	0.124	0.917
10	40	0.190	0.124	0.916
11	40	0.190	0.125	0.915
13	40	0.192	0.126	0.913

To check that performance is uniform across the G-spread distribution, diagnostic plots complement RMSE and MAE. These are shown in Figures 10 and 11. Evidently the errors are much more pronounced when spreads get larger. Figures 10 and 11 use bps for ease of interpretation, these bps estimates are back transformed from the logarithmic estimates.

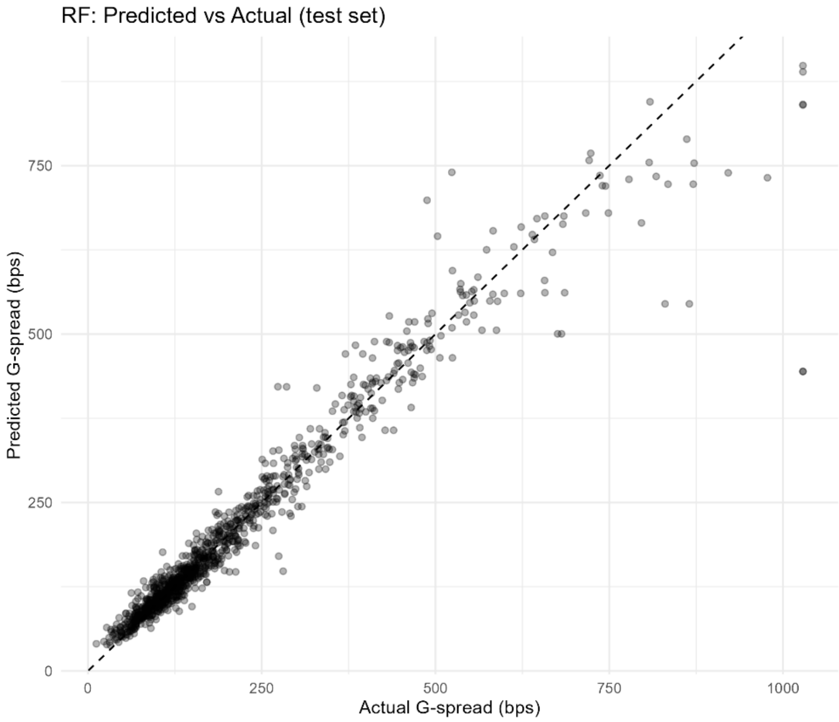

Figure 10. Random Forest predicted versus actual issuance G-spreads.

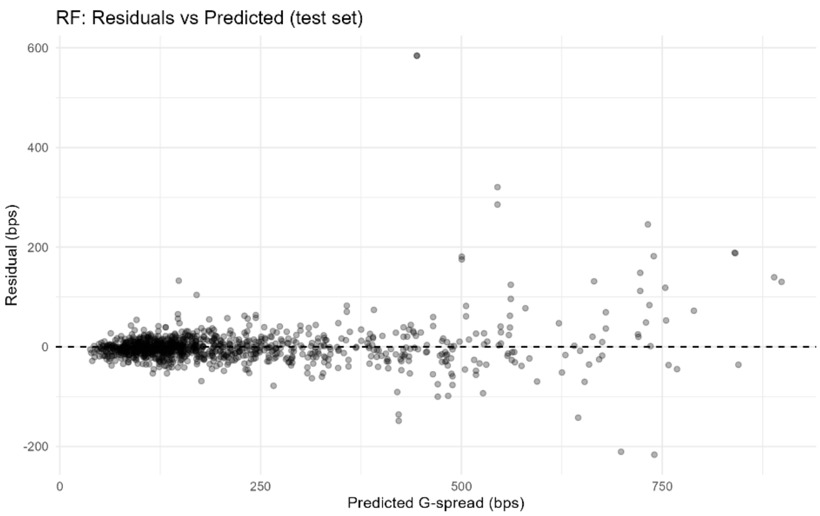

Figure 11. Random Forest residuals versus predicted values.

To summarize this heterogeneity more compactly, the test sample is grouped into maturity and credit-quality buckets and mean absolute error (MAE) is computed within each group. For the bucket based on Moody's credit rating, the sample is split into the 25th, 50th, 75th and 100th percentiles, the bonds with credit rating in the bottom 25 %

are labelled "Worst" and those with credit ratings above 75 % the "Best". For figures 12 and 13 MAE is back transformed from log(G-spread) to bps.

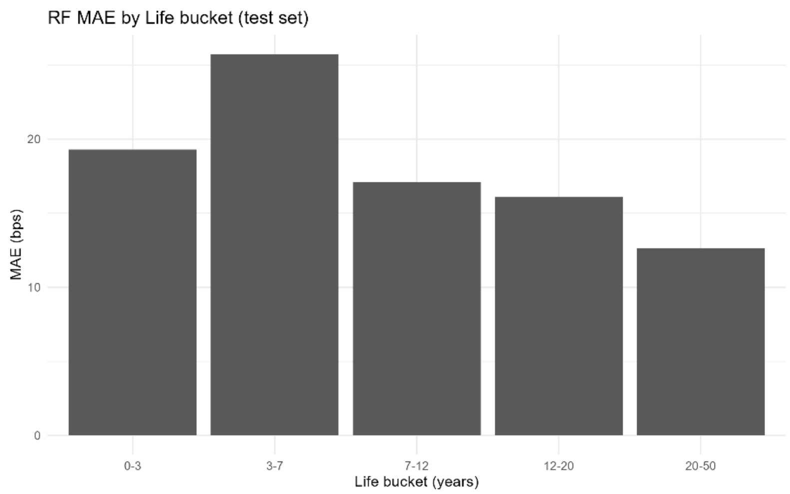

Figure 12. Random Forest MAE by life bucket (test set)

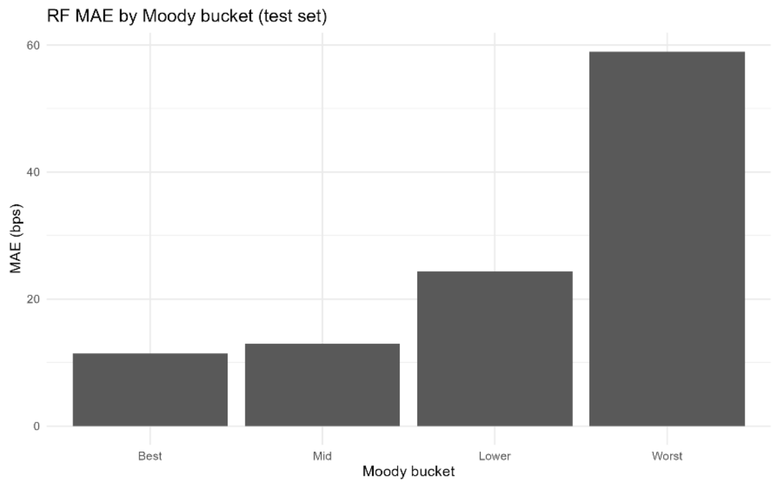

Figure 13. Random Forest MAE by Credit Rating bucket (test set)

Figure 12 shows that MAE varies across maturity ranges, with the largest errors in the 3–7-year bucket and the smallest errors in the longest maturities. Figure 13 shows that errors are low in the "Best" and "Mid" buckets, increase in the "Lower" bucket, and are substantially larger in the "Worst" bucket. These indicate that the model's largest prediction errors are concentrated among the lowest credit-quality issues where investors demand larger spreads. This is not unsurprising, prior evidence shows that credit spreads become more volatile as credit quality drops making them inherently harder to predict (Heinke, 2006; Chung et al., 2019).

6.8 Random Forest Drivers, Nonlinear Effects, and Prediction Errors

To show what information the fitted RF uses, variable importance factors and partial dependence plots (PDPs) are reported. Altmann et al. (2010) explains that importance factors measure the drop in predictive performance when the values of a single variable are randomly permutated. The larger the drop, the more the model relies on that variable.

Table 22. Random Forest variable importance factors

Variable	Importance
Credit Rating	0.3508
Coupon	0.3069
Green Label	0.1271
Life	0.0364
OAS in BPS	0.0277
Amount Issued	0.0096
Slope in BPS	0.0084
Sector	0.0076
U.S. CPI	0.0074
MOVE	0.0060
Liq./funding Proxy	0.0055
Coupons Amount	0.0039

Table 22 presents variable importance results. The picture is extremely clear; credit rating and coupon dominate the RF prediction with a clear gap to every other variable. This ranking is consistent with prior studies that show credit risk and pricing terms are the central correlates of G-spread levels (Collin-Dufresne et al., 2001; Gabbi & Sironi, 2005). After these, the green label provides the most predictive importance of the remaining variables.

Two caveats are very important for interpretation. As the green bonds were assigned a weight of 31 (to account for the sample imbalances for RF), the label should be considered more as a robustness-oriented adjustment. It makes the importance score more comparable despite the large sample size differences. The second is that variable importance is not the same as OLS coefficient. Importance asks if the variable improves prediction given everything else in the model, while OLS estimates a conditional mean difference holding other covariates fixed. In short, variable importance are not economic effect sizes and they should not be interpreted as such.

Another important note is that importance values produced by ranger are not measured in basis points, they reflect the change in loss/accuracy criterion and should be interpreted as relative magnitudes.

Beyond average accuracy, PDPs are used to show key nonlinearities. PDPs show how the model's average predicted outcome changes as one variable changes (Hooker, 2007). The PDPs are used to support the conclusion from variable importance metrics, RFs predictive gains are not due to some exotic variables or opaque interactions alone. Most of the predictive power comes from the same fundamentals that drive linear models: contractual terms and credit quality first and then maturity and market conditions and green label in the middle.

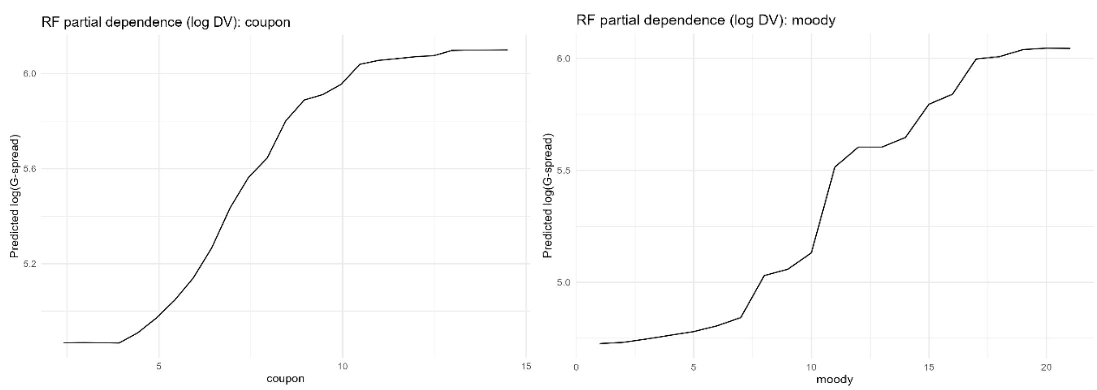

Figure 14. Partial Dependence Plots for Coupon and Credit Rating

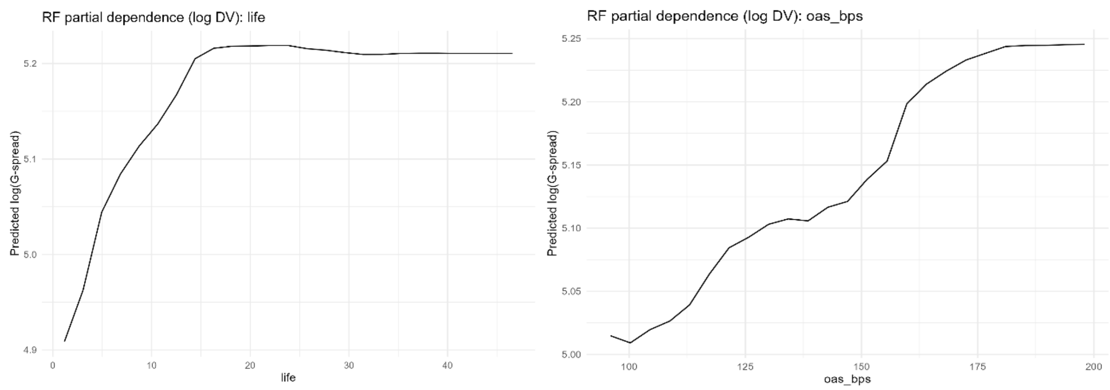

Figure 15. Partial Dependence Plots for Life and OAS

6.9 Summary of Results

In this subchapter the main findings of this thesis are discussed with focus on addressing the hypotheses formulated in chapter 1. The results are compared to earlier literature on green bonds, issuance G-spreads and machine learning in finance.

Hypothesis H1 stated that a Random Forest (RF) model would predict corporate bond issuance G-spreads more accurately than OLS out-of-sample (OOS). The evidence from this study supports H1 very clearly. On the OOS test set, RF outperforms OLS (M4) on all metrics: RMSE declines from 0.251 to 0.152, MAE declines from 0.180 to 0.099 and test R^2 rises from 0.858 to 0.948. LASSO is included as an additional predictive benchmark but performs almost identically to OLS on this split (RMSE = 0.251; MAE = 0.180; R^2 = 0.858), indicating that the main predictive gains come from RF's ability to capture non-linearities and interactions. This aligns with prior literature that models able to capture these often generalize better (Heger et al., 2024; Kim et al., 2021). Cabrol et al. (2024) similarly show that tree-based ML approaches and neural networks consistently outperform linear regression benchmarks on prediction tasks when using the same covariates.

Evaluating performance across 100 repeated splits further solidifies this result. Across partitions, RF reports substantially lower error metrics and higher test R^2 on average.

While LASSO remains very close to OLS. Overall, it is safe to say that ML offers a very valuable method to complement traditional econometrics.

Hypothesis H2 predicted that credit quality, maturity, issue size and issuance time market conditions are among the most important determinants of issuance spreads. This is supported by both the OLS ladder and RF. RF variable importance rankings showed that credit rating and coupon dominate and there is a clear drop to the next best variable, green label. Collin-Dufresne et al. (2001) alongside Gabbi and Sironi (2005) found that credit risk and contractual terms are first-order correlates of corporate credit spreads and other forces explain smaller amounts of variation. Longstaff et al. (2005) similarly showed evidence from the CDS markets that default risk and non-default components both are important for spreads, but credit quality is “foundational” for spread levels. Non-trivial contribution of issuance-time conditions such as OAS also fits the view that broad credit conditions at time of issuance shifts corporate bond spreads (Gilchrist & Zakrajšek, 2012).

One peculiarity is worth explaining further. For OLS the coupon explained the most variance of all predictors. On the RF side however, credit quality overtook coupon as the most important variable. One possibility is that because spread formation is nonlinear in credit risk (Heger et al.,2024) and interacts with other predictors, RF is able to identify this whereas OLS compresses these effects. Similarly to Gu et al. (2020) the finding is therefore that RF can “reshuffle” variable importance due to nonlinearities.

Hypothesis H3 stated that green label is associated with a statistically significant conditional spread difference after controls are introduced. The OLS result supports H3 in the sense of statistical inference. In the model 4 (M4), the green label showed positive conditional association between the label itself and issuance G-spreads in this sample. In literature the sign and magnitude of a greenium is not settled and the estimates vary across samples and strategies. For example, some work reports a modest negative greenium (Zerbib, 2019), while others find that effects are small, unstable and report that overall green bonds provide higher yields than matched non-greens (Bachelet et al., 2019). The difference is that previous studies investigated greenium in the secondary markets where trading frictions play a role. This study is therefore able to extend this further to the primary markets; in this sample the green label is associated with higher spreads while on average enjoying better credit ratings.

In economic terms, the baseline M4 estimate implies a green-labelled issuance differential of about 16.4 bps at the representative spread level of S_0 =147.5 bps. Comparing this with typical primary-market concessions, this is economically meaningful: evidence from U.S. corporate bond primary markets shows that the "issuance premium" is typically positive but modest in normal times and rise sharply in stress periods (Siani, 2022). The state splits in this thesis align with that broader primary-market pattern. In tight conditions the implied green differential rises to roughly 28–31 bps, while in stressed conditions it shrinks to about 9–10 bps and becomes statistically weaker. This supports the idea that any label-related difference is most detectable when aggregate spread movements and dispersion are lower. Whereas in stressed regimes the dominant repricing of credit and liquidity risk makes it harder to isolate smaller effects. This interpretation is consistent with primary-market green bond evidence that issuance pricing differentials depend on demand conditions and vary over time. This includes work linking issuance outcomes to oversubscription and demand pressure (Caramichael & Rapp, 2024), and with evidence that green premia are time-varying rather than constant (D'Amico et al., 2023).

Evidence from the RF suggests that the green label is not the main engine driving predictive accuracy, but its effects are not negligible either. Putting the RF importance rankings together, it is important to remember that permutation importance can be conservative when predictors are correlated (Gregorutti et al., 2016), which can be the case with financial data. This study is no different in that sense, as shown by the correlation heatmap (figure 7). As such, the ranking should be interpreted descriptively.

The main question however is, why is the greenium seemingly not a penalty but a gain in this sample. There are a few possible explanations for this. First is that the label can be correlated with deal features and dynamics that are hard to observe and as such be omitted from the analysis (Gilchrist & Zakrajšek, 2012). Another possibility relates to segmentation and liquidity. Perhaps green bonds do not receive cheaper financing in every segment and the sign of the “green” flips with market conditions or investor demand. Karpf & Mandel (2018) studied U.S. municipals and found that historically green bonds were penalized and that premium later turned positive. Finally, as shown by Zerbib (2019) and Flammer (2021) the green coefficient is sensitive to controls, meaning that selection and composition can have an effect in an observed effect. Flammer (2021) further notes that insufficient controls can mechanically generate an apparent greenium even when one “does not actually exist”.

Taken together the results support all three hypotheses in their appropriate sense. RF predicts issuance spreads more accurately than OLS and LASSO out-of-sample. The dominant predictive and explanatory drivers are conventional determinants. The green label is also statistically significant in the best configuration, but its predictive contribution is secondary to standard determinants. Last, it is important to note that the results must be interpreted with caution due to limitations discussed in the next chapter.

7 Conclusion

This thesis studied whether USD-denominated corporate green bonds price differently at issuance than comparable non-green bonds, and whether issuance G-Spreads can be predicted reliably using information available at the time of issuance. The key takeaway is that most of the variance in the issuance spreads is explained by standard determinants such as credit quality, coupon, maturity characteristics, and issuance-time market conditions, while the green label offers some predictive power. As expected, a large part of the cross-sectional variation in spreads remains tied to broader risk and liquidity conditions that are difficult to capture perfectly with a finite set of proxies. However, by combining OLS "ladder" design with a RF model, this thesis shows that issuance spreads can be predicted with reasonable accuracy.

For practical implications the results suggest that primary market pricing remains anchored in standard credit and deal fundamentals. The label adds secondary information that can matter at the margin and drastically changes depending on market conditions. Below are some actionable bullets for issuers, underwriters and investors suggested.

Issuers

- Treat green as an adjustment, not the main pricing lever. In this sample, issuance spreads are mainly driven by conventional determinants, so a green framework should not substitute for any fundamentals.
- Plan for state dependence. The green effect is not stable across market conditions, expected pricing should be stress-tested under different regimes. The greenium or green "penalty" is not constant.
- The label matters more when ESG demand is high and when credibility is perceived as high.

Underwriters

- Keep the pricing anchor in standard drivers. Models should continue to rely mainly on conventional G-spread drivers and not treat the green label as a primary input.
- Regime check should be explicit in practice. Guidance should be conditioned on the credit-market state, since label-related differentials are easier to detect when aggregate spread movements are smaller.

Investors

- Do not assume that green implies cheaper entry at issuance. In this sample, green bonds can price at higher G-spreads even when issuers are, on average, slightly higher quality.
- Unstable regime dependent differentials should be expected. In stressed regimes, credit and liquidity drive pricing, but this can change in tight market conditions.

Interpreting the green label results remains constrained because green issuance is a choice that can be correlated with issuer-level traits and decisions that are difficult to observe in the dataset. Even with a rich set of controls a cross-sectional strategy cannot guarantee that a green coefficient isolates a pricing effect and not issuer selection or unobserved features. Larcker and Watts (2020) highlight this and emphasize that estimates are sensitive to design and identification strategies.

Data availability is a second limitation and considers liquidity and funding measurements especially. Because standard corporate bond liquidity measures often rely on transaction data (e.g., TRACE-based trading costs, turnover), prior research frequently faces a trade-off between sample scale and “first-best” liquidity measurement (Edwards et al., 2007; Dick-Nielsen et al., 2012). Furthermore, demand proxies, different green indicators and debut green issuer indicators are necessary to show the mechanisms at play. As these were not available for this thesis, they are listed as important limitations.

Sample size is also constrained by access and download limits rather than by choices, which matters because power and stability are particularly important for effects that might be economically small. A larger sample would improve precision for the green-label coefficient and make subgroup analyses (e.g., within-rating or within-sector heterogeneity) less sensitive to a small number of observations, which is a standard concern.

Even though the RF is tuned and evaluated using a clean separation of training and test data, the reported performance still reflects the chosen evaluation methods, and different split designs can yield different results. The machine learning literature treats this as a feature of model assessment rather than a defect of any single model. Careful validation design is therefore central for credible performance reporting (Kohavi, 1995; Arlot & Celisse, 2010).

Future research could improve on this study and make the interpretation of the "green" coefficient more convincing. The most direct improvement is to focus more on within-issuer variation. An issuer fixed-effects model would remove all time-invariant issuer differences and force the green estimate to come from comparisons within the same issuer. Another option is within-issuer matching: for each green bond a conventional bond from the same issuer issued within ±90 days. This would not only treat the comparison within issuer, but also within similar market conditions by default. A related design is to treat a firm's first green issuance as an event. Comparing pricing between the first green issuance and another nearby non-green issue (and ideally with similar issuers that never issue green) would help to separate the "who issues green" from "what happens when issuer starts issuing green". Matching green bonds to non-greens is not a completely new approach, similar has been done in prior research (Flammer, 2021; Zerbib, 2019).

Another extension is to treat the green label as more heterogeneous. If the mechanism it operates from is partly signaling or credibility, the effect should be stronger when the signal is more credible and harder to fake. Using a sample with green labels from different sources would allow for comparisons between signals that vary in strength. On the market conditions side, the liquidity proxy in this thesis was a pragmatic solution. With TRACE type data it would be possible to measure liquidity directly and test if the estimated differential is compensation for liquidity related factors.

Finally, regime dependence should be treated as a core feature. Recent evidence ties the green pricing to demand pressure at issuance, which makes it reasonable to include variables for oversubscription and index inclusion (Caramichael & Rapp, 2024). And because sustainable bond markets are still evolving quickly, recent policy syntheses provide a reason to expect that pricing differences may not be stable across time or across different settings (OECD, 2025).

Appendix 1 – Data Appendix

Step	Filter	Removed	Remaining	Rationale
0	Raw downloaded sample		3806	Initial extraction.
1	Exclude missing credit rating	122	3684	Credit risk is core control. Required for compatibility.
2	Exclude rare sectors (≤2 obs)	27	3657	Avoid unstable sector effects.
3	Exclude >50 life	3	3654	Outliers with improbably long life.
4	Final sample		3654	

Appendix 2 – Full Sample Descriptive Statistics

	G-Spread	Life	Log Amount Issued	Coupon	Coupons Amount	Credit Rating	Slope BPS	MOVE	US CPI	OAS bps	Liq./funding Proxy
Mean	205.92	10.54	12.88	6.00	2.01	9.49	-11.60	106.10	0.03	127.57	12.51
SD	165.79	8.72	1.94	1.56	0.12	3.50	23.46	16.54	0.01	24.50	7.35
p25	105.73	5.02	13.02	5.00	2.00	7.00	-34.00	94.30	0.03	110.00	7.00
p75	241.55	10.03	13.71	6.50	2.00	12.00	5.00	114.90	0.03	150.00	18.00
Skewness	2.29	1.79	-2.83	1.52	14.17	0.40	-0.35	1.50	1.78	0.88	-0.04
Kurtosis (Fisher)	6.20	2.07	7.31	3.22	252.11	-0.11	-0.98	5.10	3.60	-0.29	0.82
n	3654	3654	3654	3654	3654	3654	3654	3654	3654	3654	3654

Note: Kurtosis is reported as excess kurtosis. Coupons amount is a discrete coupon frequency variable, so its kurtosis reflects a highly concentrated mass at 2 rather than continuous tail behaviour.

8 Appendix 3 – Moody’s Credit Rating Transformation

Moody’s Rating	Transformed
Aaa	1
Aa1	2
Aa2	3
Aa3	4
A1	5
A2	6
A3	7
Baa1	8
Baa2	9
Baa3	10
Ba1	11
Ba2	12
Ba3	13
B1	14
B2	15
B3	16
Caa1	17
Caa2	18
Caa3	19
Ca	20
C	21

9 References

Akerlof, G. A. (1970). The Market for “Lemons”: Quality Uncertainty and the Market Mechanism. *The Quarterly journal of economics, 84*(3), 488-500. https://doi.org/10.2307/1879431

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*(10), 1340-1347. https://doi.org/10.1093/bioinformatics/btq134

Alzubi, Jafar & Nayyar, Anand & Kumar, Akshi. (2018). Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series. 1142. 012012.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79. https://doi.org/10.1214/09-SS054

Bachelet, M. J., Becchetti, L., & Manfredonia, S. (2019). The Green Bonds Premium Puzzle: The Role of Issuer Characteristics and Third-Party Verification. *Sustainability, 11*(4), 1098. https://doi.org/10.3390/su11041098

Baker, M., Bergstresser, D., Serafeim, G., & Wurgler, J. (2018). Financing the Response to Climate Change: The Pricing and Ownership of U.S. Green Bonds. *NBER Working Paper Series,* 25194. https://doi.org/10.3386/w25194

Bank for International Settlements. (2023). *Debt securities statistics*. BIS Data Portal. Retrieved November 15, 2025, from https://data.bis.org/topics/DSS

BAO, J., PAN, J., & WANG, J. (2011). The Illiquidity of Corporate Bonds. *The Journal of finance (New York), 66*(3), 911-946. https://doi.org/10.1111/j.1540-6261.2011.01655.x

Benveniste, L. M., & Spindt, P. A. (1989). How investment bankers determine the offer price and allocation of new issues. *Journal of financial economics, 24*(2), 343-361. https://doi.org/10.1016/0304-405X(89)90051-2

Bianchi, D., Büchner, M., Hoogteijling, T., & Tamoni, A. (2021). Bond Risk Premiums with Machine Learning. *The Review of financial studies, 34*(2), 1046-1103. https://doi.org/10.1093/rfs/hhaa062

Boehmke, B. C., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. Chapman and Hall/CRC.

Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica, 47*(5), 1287-1294. https://doi.org/10.2307/1911963

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market Liquidity and Funding Liquidity. *The Review of financial studies, 22*(6), 2201-2238. https://doi.org/10.1093/rfs/hhn098

Cabrol, A., Drobetz, W., Otto, T., & Puhan, T. (2024). Predicting Corporate Bond Illiquidity via Machine Learning. *Financial analysts journal, 80*(3), 103-127. https://doi.org/10.1080/0015198X.2024.2350952

Caramichael, J., & Rapp, A. C. (2024). *The green corporate bond issuance premium*. *Journal of Banking & Finance, 162*, 107126. https://doi.org/10.1016/j.jbankfin.2024.107126

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*(70), 2079–2107.

CFA Institute. (2025). *Fixed-income bond valuation: Prices and yields*. CFA Institute.

Chen, L., Pelger, M., & Zhu, J. (2024). Deep Learning in Asset Pricing. *Management science, 70*(2), 714-750. https://doi.org/10.1287/mnsc.2023.4695

Collin-Dufresne, P., Goldstein, R. S., & Martin, J. S. (2001). The determinants of credit spread changes. *The Journal of Finance, 56*(6), 2177–2207.

Cornelli, F., & Goldreich, D. (2003). Bookbuilding: How Informative Is the Order Book? *The Journal of finance (New York), 58*(4), 1415-1443. https://doi.org/10.1111/1540-6261.00572

Dick-Nielsen, J., Feldhütter, P., & Lando, D. (2012). Corporate bond liquidity before and after the onset of the subprime crisis. *Journal of financial economics, 103*(3), 471-492. https://doi.org/10.1016/j.jfineco.2011.10.009

EDWARDS, A. K., HARRIS, L. E., & PIWOWAR, M. S. (2007). Corporate Bond Market Transaction Costs and Transparency. *The Journal of finance (New York), 62*(3), 1421-1451. https://doi.org/10.1111/j.1540-6261.2007.01240.x

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Ehlers, T., & Packer, F. (2017). Green bond finance and certification. *BIS Quarterly Review* (September), 89–104.

Elton, E. J., Gruber, M. J., Agrawal, D., & Mann, C. (2001). Explaining the Rate Spread on Corporate Bonds. *The Journal of finance (New York), 56*(1), 247-277. https://doi.org/10.1111/0022-1082.00324

European Central Bank. (2023). *Climate-related financial disclosures of the ECB's non-monetary policy portfolios*.

Fatica, S., Panzica, R., & Rancan, M. (2021). The pricing of green bonds: Are financial institutions special? *Journal of financial stability, 54*, 100873. https://doi.org/10.1016/j.jfs.2021.100873

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.

Flammer, C. (2021). Corporate green bonds. *Journal of financial economics, 142*(2), 499-516. https://doi.org/10.1016/j.jfineco.2021.01.010

Friewald, N., Jankowitsch, R., & Subrahmanyam, M. G. (2012). Illiquidity or credit deterioration: A study of liquidity in the US corporate bond market during financial crises. *Journal of financial economics, 105*(1), 18-36. https://doi.org/10.1016/j.jfineco.2012.02.001

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gabbi, G., & Sironi †, A. (2005). Which factors affect corporate bonds pricing? Empirical evidence from eurobonds primary market spreads. *The European journal of finance, 11*(1), 59-74. https://doi.org/10.1080/1351847032000143422

Gianfrate, G., & Peri, M. (2019). The green advantage: Exploring the convenience of issuing green bonds. *Journal of cleaner production, 219*, 127-135. https://doi.org/10.1016/j.jclepro.2019.02.022

Giglio, S., Kelly, B., & Stroebel, J. (2021). Climate finance. *Annual Review of Financial Economics, 13*, 15–36. https://doi.org/10.1146/annurev-financial-102620-103311

Gilchrist, S., & Zakrajšek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *The American economic review, 102*(4), 1692-1720. https://doi.org/10.1257/aer.102.4.1692

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of financial studies, 33*(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009

Hachenberg, B., & Schiereck, D. (2018). Are green bonds priced differently from conventional bonds? *Journal of asset management, 19*(6), 371-383. https://doi.org/10.1057/s41260-018-0088-5

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Heger, J., Min, A., & Zagst, R. (2024). Analyzing credit spread changes using explainable artificial intelligence. *International review of financial analysis, 94*, 103315. https://doi.org/10.1016/j.irfa.2024.103315

Heinkel, R., Kraus, A., & Zechner, J. (2001). The Effect of Green Investment on Corporate Behavior. *Journal of financial and quantitative analysis, 36*(4),4 31-449. https://doi.org/10.2307/2676219

Hooker, G. (2007). Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of computational and graphical statistics, 16*(3), 709-732. https://doi.org/10.1198/106186007X237892

Huang, J., & Huang, M. (2012). How Much of the Corporate-Treasury Yield Spread Is Due to Credit Risk? *Review of asset pricing studies, 2*(2), 153-202. https://doi.org/10.1093/rapstu/ras011

International Capital Market Association. (2022). *Green Bond Principles: Voluntary Process Guidelines for Issuing Green Bonds* (June 2022). ICMA.

Kapraun, J., Latino, C., Scheins, C., & Schlag, C. (2021, April). (In)-Credibly green: Which bonds trade at a green bond premium? SSRN

Karpf, A., & Mandel, A. (2017, February 24). Does it pay to be green? *SSRN.* http://dx.doi.org/10.2139/ssrn.2923484

Kim, J., Kim, D. H., & Jung, H. (2021). Applications of machine learning for corporate bond yield spread forecasting. *The North American journal of economics and finance, 58*, 101540. https://doi.org/10.1016/j.najef.2021.101540

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*

Larcker, D. F., & Watts, E. M. (2020). Where's the greenium? *Journal of accounting & economics, 69*(2-3), 101312. https://doi.org/10.1016/j.jacceco.2020.101312

Leland, H. E., & Pyle, D. H. (1977). Informational Asymmetries, Financial Structure, and Financial Intermediation. *The Journal of finance (New York), 32*(2), 371. https://doi.org/10.2307/2326770

Mithal, S., Longstaff, F. A., & Neis, E. (2005). Corporate Yield Spreads: Default Risk or Liquidity? New Evidence from the Credit Default Swap Market. *Journal of Finance, 60*(5), 2213-2253. https://doi.org/10.1111/j.1540-6261.2005.00797.x

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics, 29*(3), 305-325. https://doi.org/10.1016/0304-4076(85)90158-7

Mahesh, B. (2020). *Machine learning algorithms—A review*. *International Journal of Science and Research (IJSR), 9*(1), 381–386.

Maimon, O., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook*. Springer.

Moritz, B., & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. SSRN. http://dx.doi.org/10.2139/ssrn.2740751

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *The Journal of economic perspectives, 31*(2), 87-106. https://doi.org/10.1257/jep.31.2.87

Myers, S. C., & Majluf, N. S. (1984). Corporate financing and investment decisions when firms have information that investors do not have. *Journal of financial economics, 13*(2), 187-221. https://doi.org/10.1016/0304-405X(84)90023-0

Organisation for Economic Co-operation and Development. (2025). *Sustainable bonds: Trends and policy recommendations*. OECD Publishing. https://doi.org/10.1787/26726c68-en

Organisation for Economic Co-operation and Development. (2025). Key issues in the sustainable bond markets. In *Sustainable bonds: Trends and policy recommendations* (pp. 28–40). OECD Publishing. https://doi.org/10.1787/26726c68-en

Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2021). Sustainable investing in equilibrium. *Journal of Financial Economics, 142*(2), 550-571. https://doi.org/10.1016/j.jfineco.2020.12.011

Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2021). Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics, 142*(2), 572-597. https://doi.org/10.1016/j.jfineco.2020.11.001

Pietsch, A., & Salakhova, D. (2025). Pricing of green bonds: Greenium dynamics and the role of retail investors. *Journal of international financial markets, institutions & money, 104*, 102211. https://doi.org/10.1016/j.intfin.2025.102211

Probst, P., Bischl, B., & Boulesteix, A. (2018). *Tunability: Importance of Hyperparameters of Machine Learning Algorithms*. https://doi.org/10.48550/arxiv.1802.09596

Ross, S. A. (1977). The Determination of Financial Structure: The Incentive-Signalling Approach. *Bell Journal of Economics, 8*(1), 23-40. https://doi.org/10.2307/3003485

Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative finance, 19*(9), 1449-1459. https://doi.org/10.1080/14697688.2019.1622295

Spence, M. (1973). Job Market Signaling. *The Quarterly journal of economics, 87*(3), 355-374. https://doi.org/10.2307/1882010

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36 (2),111– 133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Tang, D. Y., & Zhang, Y. (2020). Do shareholders benefit from green bonds? *Journal of Corporate Finance, 61*, 101427. https://doi.org/10.1016/j.jcorpfin.2018.12.001.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica, 48*(4), 817–838. https://doi.org/10.2307/1912934

Wilkinson, L. (1992). Tree Structured Data Analysis: AID, CHAID and CART. Sawtooth/SYSTAT Joint Software Conference, Su Valley, July 1992.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.

Zerbib, O. D. (2019). The effect of pro-environmental preferences on bond prices: Evidence from green bonds. *Journal of banking & finance, 98*, 39–60. https://doi.org/10.1016/j.jbankfin.2018.10.012