HEC MONTRÉAL

Greenium Estimation using Machine Learning Algorithms

par

Nihel Seghaier

Michèle Breton
HEC Montréal
Directrice de recherche

Sciences de la gestion
(Spécialisation M.Sc. Financial Engineering)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

June 2022

ACKNOWLEDGEMENTS

I am profoundly indebted to Prof. Michèle Breton who has given me her invaluable support, time and constant supervision, as well as for imparting her knowledge and expertise in this project.

I would like to convey my heartfelt gratitude to my beloved husband Ghazi, who has supported and helped me at both the technical and personal levels.

I would like to express my deepest gratitude to my parents, who have encouraged me throughout the process of my thesis completion, even from afar.

Finally, I am thankful to all my family, friends, and colleagues at Deloitte who have supported me throughout my thesis.

Abstract

Green bonds are becoming increasingly popular in the fixed-income market as they strive to address a variety of environmental challenges. It was discovered that these securities differ in pricing from brown bonds in both the primary and secondary markets, and that they generally include a premium for their green labelling named "greenium". Nevertheless, the existence of the greenium is yet to be determined and evidence on its value is still mixed. In this thesis, we use Machine Learning models to propose a framework allowing the detection and estimation of the greenium in the primary market. We start by training the Machine Learning algorithms to predict brown bonds yields using their key characteristics. Then, we investigate the performance of the selected models using the K-Fold cross-validation technique, the Mean Square Error, as well as R^2. We show that our chosen models perform considerably well in predicting the brown bonds yields, and they also generate stable results when challenged with previously unseen data. In the second part of this thesis, we aim to extend our work by computing the yield of each brown bond that has similar characteristics to its green counterpart. Hence, we apply the trained models to the green bonds database. We examine the residuals for respectively the brown and green bonds databases and we observe that their mean shifts from zero to a positive value when dealing with green bonds. This confirms the existence of greenium and allows us to estimate its value, which was discovered to be approximately 30 bps.

Keywords: Fixed-Income market, Green bonds, Bond yields, Bond premium, Greenium, Machine Learning algorithms.

Résumé

Les obligations vertes deviennent de plus en plus populaires sur le marché des titres à revenu fixe car elles visent à résoudre divers problèmes environnementaux. Il a été découvert que le prix de ces titres diffère de celui des obligations brunes, tant sur le marché primaire que sur le marché secondaire, et qu'ils incluent généralement une prime pour leur étiquette verte, appelée "greenium". Néanmoins, l'existence du greenium reste à déterminer et les preuves de sa valeur sont encore mitigées. Dans cette thèse, nous utilisons des modèles d'apprentissage automatique pour proposer un cadre permettant la détection et l'estimation du greenium dans le marché primaire. Nous commençons par entraîner les algorithmes d'apprentissage automatique à prédire les rendements des obligations brunes en utilisant leurs caractéristiques principales. Ensuite, nous étudions la performance des modèles sélectionnés en utilisant la technique de validation croisée K-Fold, l'Erreur Quadratique Moyenne, ainsi que R^2. Nous montrons que les modèles que nous avons choisis sont très performants pour prédire les rendements des obligations brunes, et qu'ils génèrent également des résultats stables lorsqu'ils sont confrontés à des données préalablement non vues. Dans la deuxième partie de cette thèse, nous cherchons à étendre notre travail en calculant le rendement de chaque obligation brune qui a des caractéristiques similaires à son homologue verte. Nous appliquons donc les modèles entraînés à la base de données des obligations vertes. Nous examinons les résidus pour les bases de données d'obligations brunes et vertes respectivement et nous observons que leur moyenne passe de zéro à une valeur positive lorsqu'il s'agit d'obligations vertes. Cela confirme l'existence du greenium et nous permet d'estimer sa valeur, qui s'est révélée être d'environ 30 bps.

Mots-clés: Marché des titres à revenu fixe, Obligations vertes, Rendement des obligations, Prime des obligations, Greenium, Algorithmes d'apprentissage automatique.

Contents

List of Abbreviations

AI	Artificial Intelligence
bps	Basis Points
CBI	Climate Bond Initiative
CDOR	Canadian Interest Rate Benchmark
CF	Cash-Flow
CPI	Consumer Price Index
ESG	Environmental Social Governance
FISD	Fixed Income Securities Database
GBP	Green Bond Principles
GSS	Green, Social, and Sustainability
ICMA	International Capital Markets Association
LIBOR	London Interbank Offer Rate
LOOCV	Leave-One-Out Cross-Validation
MARS	Multivariate Adaptive Regression Spline
ML	Machine Learning
MSE	Mean Square Error
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PLS	Partial Least Squares
SVM	Support Vector Machine
SVR	Support Vector Regression

YTM	Yield-to-Maturity

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The global fixed-income markets represent the largest subset of financial markets in terms of number of issuances and market capitalization. In these, the bond market is particularly significant. Bonds are fixed-income securities that are particularly attractive to investors because they provide a stable income, and because they help mitigate exposure to volatile instruments, such as equities. Some of these instruments, known as *green* bonds, are currently being used as Environmental, Social, and Corporate Governance (ESG) investments, providing capital for companies that score high on environmental and societal responsibilities. Since their introduction in 2008, green bonds have emerged as an important financial tool in addressing environmental issues. Their recognition in the capital markets across investors has been steadily increasing, showing an almost exponential rise in terms of annual issuance (see Figure 2.3.1).

Green bonds are discussed in a broad range of the academic literature, and have become a timely topic in many areas, such as finance, business, economics, law and environment. In the financial literature, academics are interested in the pricing of these securities in the primary and secondary markets, and more precisely whether green bonds are priced differently from their non-green counterparts. Several studies have been conducted to investigate the presence of a price premium, which is referred to as *greenium*.

So far, evidence on the existence, magnitude and sign of the greenium is mixed. One way to quantify the greenium is by subtracting the brown bond yield from its green counterpart. The yield indicates the overall return expected by an investor, and, unlike the bond price, is not affected by the currency. In this thesis, we take this approach and propose a greenium estimation framework based on yield comparisons. In the sequel, the term *brown* bonds refer to bonds that do not qualify as green bonds.

Most of the recent studies use a matching approach to estimate the greenium; this ap-

proach involves considering either a hypothetical or an existing brown bond with similar characteristics. In this case, brown bond yields are determined using closed-form formulas for their intrinsic values (see Section 2.2.2.3). However, these theoretical yields differ from actual market yields as they exclude any bond premiums, such as liquidity or credit risk premiums. Furthermore, the matching approach involves the extraction of multiple interest rates at the different issuance dates for each bond, which is a lengthy procedure.

The aim of this thesis is to propose an alternative approach to address the issue of the existence and importance of the greenium. Instead of matching pairs of bonds, we employ Machine Learning (ML) to relate the actual yield of bonds to their key features. We first train the ML models to predict the yield of a brown bond, given its features. The pre-trained algorithms are then applied to a database of green bonds, in order to determine the yield of their brown counterparts (that is, presenting the same features). Finally, we use the difference in the predicted and actual yields in the green bonds database in order to simultaneously detect and estimate the greenium.

1.2 Project Framework

1.2.1 Objectives

Green bonds are presumed to have an additional yield premium. This yield difference has been investigated by researchers and has been found to be either i) positive, due to the willingness of bondholders to invest in environmentally friendly securities, ii) negative, due to their novelty and consequent riskiness, or iii) zero, because they do not present a significant difference from other bonds.

The aim of this thesis is to develop a framework that provides an estimation of the greenium using ML algorithms. We focus on the use of Linear Regression, Random Forest, Support Vector Machine (SVM), Partial Least Square Regression (PLS) and Multivariate Adaptive Regression Spline (MARS) models.

In this thesis, all bonds that do not meet the characteristics of green bonds (see Section 2.3) are identified as *brown*. Our proposed framework starts by training the above-mentioned ML algorithms on a dataset of brown bonds, such that they learn how to predict the yield of any bond once given its features. We then investigate the performance of these algorithms using various evaluation metrics. Finally, we select the best models and apply them to a dataset of green bonds. Our goal is to estimate the greenium by comparing the actual yields of green bonds to the model-predicted yields, which correspond to those of comparable brown counterparts.

1.2.2 Literature Review

Greenium estimation

The finance literature investigates the prospects of green bonds investments in two directions. The first stream of the literature is primarily concerned with the pricing of green bonds, in both the primary and secondary markets, with the existence and sign of the greenium, as well as with the impact of green considerations on market players. The second stream in finance literature focuses on the value implications for green bond issuers, and more specifically on the economic and environmental effects of green issues.

Three different results are reported in the first stream of the finance literature, depending on the data selected: i) positive greenium, ii) negative greenium, and iii) no greenium. Overall, the matching method is the most commonly used methodology to estimate the greenium. To successfully carry out the matching, researchers either i) create hypothetical conventional bonds with the same characteristics as their green counterparts, or ii) consider comparable conventional bonds that already exist in the market and set a threshold for the maximum difference accepted.

Several empirical and theoretical studies establish that bondholders are willing to pay a premium (positive greenium) for climate-friendly bonds. Baker et al. [2018] propose an asset-pricing framework that incorporates investors' preferences into a theoretical model in which the corporate behaviour is primarily decided exogenously, regardless of market sentiment. Using green and brown US bond data, the authors find that securities with higher environmental ratings offer lower expected returns, which they view as an indication of the existence of a positive greenium. Several other studies find evidence consistent with a positive greenium, including Zerbib [2019], who examines a sample of green bonds issued in the US market and their matched brown equivalents, and Ehlers and Packer [2017], who compare the yields of green bonds and their brown equivalents in a sample of 21 green bonds issued between 2014 and 2017.

More recently, Kapraun et al. [2021] also investigate the extent of the pricing differences between green and brown bonds, using a larger sample of over 1,500 green bonds issued globally and considering both primary and secondary market data. The authors find that a small selection of government or large corporations green bonds do trade at a premium at the time of issuance. However, they show that, in the secondary market, the green premium only applies to government bonds. Fatica et al. [2021] analyze a selection of 1,397 green bonds, with a primary focus on determining if green bonds that are issued by financial institutions generate a greater greenium than green bonds issued by other industries. Their findings suggest that, while green bonds issued by governments are subject to a premium, green bonds issued by financial institutions are not.

Contrasting the research that supports the existence of a positive greenium, numerous other

studies show the opposite, the rationale being that green bonds, a new type of instrument, may be perceived as risky or less appealing by investors, thus demanding higher yields compared to their brown equivalents. Karpf and Mandel [2017] investigate the yields of 1,880 green US municipal bonds and discover that green bonds trade at a higher yield in the secondary market than their corresponding brown bonds with similar attributes. Using a small sample of 89 green bonds, Bachelet et al. [2019] obtain that these have higher yields than their similar brown equivalents, the difference reflecting either the quality of the issuer or the green labeling.

Another stream of empirical research contends that there is no difference in yields between green and brown bonds with similar characteristics, as both instruments are equivalent at issuance, suggesting that greenium does not exist. Using a sample of 640 matched pairs of green and brown US municipal bonds, Larcker and Watts [2019] find that the greenium is essentially zero, but because they only consider one type of green bonds – that is, US bonds issued by municipalities - their findings may be limited in their generalizability. A similar finding is reported by Reed et al. [2019], who allude to the lack of investor trust in green bonds' environmental impact, underlining the difficulties in tracking whether green bonds are actually green.

Machine Learning

Thanks to their ability to reduce prediction errors, finance applications of ML methods have recently received a lot of attention (see, e.g., Culkin and Das [2017], De Spiegeleer et al. [2018] and Ghoddusi et al. [2019]). Henrique et al. [2019] provide a comprehensive analysis of the most prominent research published in the last two decades on the ML application in financial market prediction. Their review of the literature clearly demonstrates that various ML methods, such as artificial neural networks, SVM, and random forest, have been applied in multiple forecasts of financial markets and that they were found to have better performance than traditional linear models in some of the applications.

Mishra and Padhy [2019] use the support vector regression (SVR) algorithm to forecast stock prices in recent studies and demonstrate that the model's anticipated prices almost match the observed market prices. As a result, they contend that their proposed framework can be used to efficiently build a portfolio. Rasekhschaffe and Jones [2019] show that ML algorithms outperform linear models as an effective portfolio management framework. Ma et al. [2021] evaluate random forest, SVR, as well as deep learning models in a portfolio management application, where decisions are based on predicted stock returns and find that the random forest approach generates better results than other models in terms of accuracy. In the same direction, Mishra et al. [2021] apply a hybrid regression model incorporating a combination of a selection operator and least absolute shrinkage, learning-based optimization, and SVR to select stocks in a portfolio.

Furthermore, ML techniques have been commonly employed for credit rating forecasting

and are effective for assessing financial risk. Golbayani et al. [2020] use four ML models (random forest, bagged decision tree, multiple layer perceptron, and SVM) to predict corporate credit ratings. Their findings suggest that random forest and bagged decision tree outperform SVM and multiple layer perceptron when applied to three sectors of stock data, namely healthcare, energy, and finance. Moscatelli et al. [2020] find that ML models outperform traditional statistical models in forecasting corporate default risk, particularly when dealing with insufficient data.

ML algorithms have also been used to predict yields and returns. Using big data and ML, Bali et al. [2020] examine the predictability of stock and bond returns. Kirczenow et al. [2018] investigate the use of ML in deriving characteristics from historical market corporate bond yields by constructing a hypothetical illiquid fixed income market and learning the characteristics of the lacking yield from historical data of the securities exchanged in the chosen liquid market. Nunes et al. [2019] propose a forecast of the European yield curve using two models, multivariate linear regression and multilayer perceptron, at five different prediction horizons ranging from the next day to 20 days in the future. Kim et al. [2021] examine nine different forecasting techniques, including state-of-the-art ML models, for predicting corporate bond yield spreads, and analyze their performance on out-of-sample outputs using two distinct forecast horizons. Barboza et al. [2017] show that forecasting accuracy is considerably enhanced by ML models, compared to discriminant analysis and logistic regression. Kim and Jung [2019] find that ML models outperform the traditional least squares method in forecasting winning bids.

Finally, Ryll and Seidens [2019] use meta-analysis to support that ML applications outperform stochastic models in the finance field in general.

1.2.3 Main Contributions

The main contributions of this thesis to the Fixed-Income market research are the following:

- Investigation of the capability of ML models in predicting bond yields in the primary market, given their fundamental features.
- Development of an Artificial Intelligence (AI)-based framework for estimating multiple types of bond premiums, such as liquidity or risk premiums.
- Investigation of the existence of a green bond premium (or greenium) and estimation of its value using various ML models.

1.3 Outline

This thesis is structured as follows.

Chapter 2 recalls fundamental concepts. The first part of the chapter is devoted to the characteristics of the Fixed-Income market, and more specifically of bonds, their features and their valuation, as well as the definition and characterization of green bonds. The second part is dedicated to ML algorithms, including data processing, description of the supervised learning methods used in our work, and of various performance metrics.

Chapter 3 presents the data preprocessing, where we start by explaining the steps followed to collect and clean the brown and green bonds data, and we provide a walkthrough of both datasets where we explain the different features that we used and added to the data. Then, we proceed with providing a visualization of the data, as well an explanation of the transformation step.

Chapter 4 highlights the models implementation, results and the greenium estimation. We report on the implementation of the various ML models on the brown bonds database, and we illustrate their performances using multiple evaluation metrics. We then apply each ML model to the green bonds features in order to generate the predicted yields of corresponding brown bonds. We conclude this chapter by inquiring into the ML models residuals for the brown and green bonds datasets and we provide an estimation of the greenium.

Chapter 5 is a short conclusion, in which we present a summary of our results, as well as some possible extensions for our work.

Chapter 2

Fundamental Notions

This chapter recalls the basic concepts related to green bonds and Machine Learning. The first section introduces the basics of the fixed-income market. We give an overview of the market size and the terminology often used by practitioners, as well as the different types of risks faced in this market. The second section details the specifics applying to the bond market and the way bonds are priced. The third section of this chapter is dedicated to green bonds and the *Green Bonds Principles*, that is, the criteria that a bond needs to satisfy in order to obtain the green labeling. The last section is dedicated to ML algorithms. We briefly outline the main steps in developing such models and we introduce from a mathematical perspective the most commonly used supervised ML algorithms.

2.1 The Fixed-Income Market

The fixed-income market, also commonly referred to as the debt securities market, consists of instruments that pay investors fixed interest or dividend payments, in the form of coupons, until the maturity date. Typically, the payments are made at a predefined frequency while the principal is repaid to the investor at the maturity date.

Fixed-income securities include publicly traded securities, such as commercial paper, notes, and bonds, as well as non-publicly traded loans. The most famous fixed-income securities are bonds, which are usually classified according to the type of issuer (governments, municipalities or corporations).

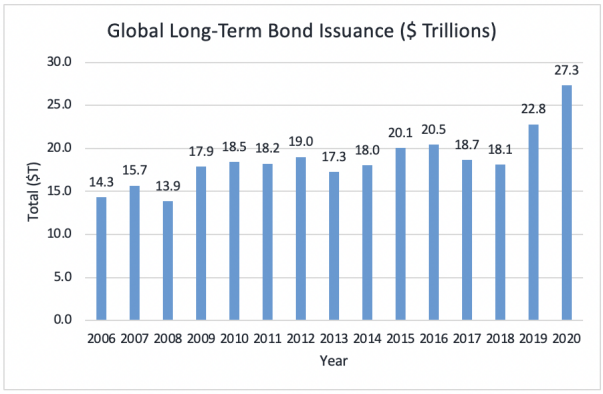
Debt securities are generally seen as less risky than equity investments. Unlike equities, where payments vary depending on some underlying criteria, the payments (coupon and principal) of an instrument in the fixed-income market are known in advance. As a result, their potential returns are often lower.

2.1.1 Size

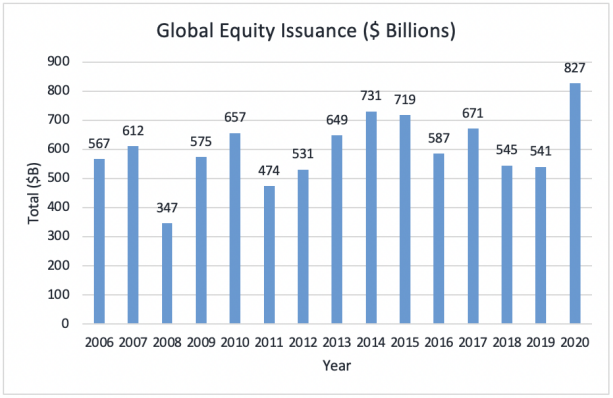
The global fixed-income market represents the largest subset of financial markets in terms of number of issuances and market capitalization. Although they typically receive less attention than equity markets, fixed-income markets are more than three times their size. According to the Institute of International Finance, the size of the global debt market reached USD 253 trillion in the third quarter of 2019, which represents a 322% global debt-to-GDP ratio (Source: CFA Institute [2020]).

In 2020, the US long-term fixed income issuance reached $12.2 trillion, which represents a 48.1% increase from the previous year, whereas equity issuance in the US market, including common and preferred shares, totaled $390 billion in 2020, a 71% increase over the previous year.

At a global level, the bond markets' outstanding value increased by 16.5% to reach $123.5 trillion in 2020, while global long-term bond issuance rose by 19.9% to $27.3 trillion (see Figure 2.1.1a). In comparison, the global equity market capitalization increased by 18.2% year-over-year to $105.8 trillion in 2020, where the global equity issuance reached $826.8 billion (see Figure 2.1.1b).



(a) Global Long-Term Bond Issuance



(b) Global Equity Issuance

Figure 2.1.1: Comparison between the Fixed-Income and the Equity markets in terms of Global Issuance from 2006 to 2020 (Source: SIFMA).

2.1.2 Terminology

The following recalls some of the terminology used in the context of financial instruments of the fixed-income market:

- Issuer: the entity that borrows money from investors by issuing the debt security, and is due to pay interest and repay principal at the maturity date.

- Holder: the investor who buys the debt security from the issuer.
- Principal (also known as maturity value, face value or par value): the amount borrowed by the issuer that must be reimbursed to the lender upon maturity, this amount is also used as the reference for the determination of the interest payments.
- Coupon rate: the rate of interest that the issuer must pay, expressed as a percentage of the principal.
- Coupon dates: the dates at which the issuer will make coupon payments, based on a predefined interval of time also known as frequency. Generally, the frequency is semiannually, but it can also be annually or monthly.
- Maturity: the date at which the debt security matures, and the issuer must return the principal to the investor.
- Issue price: the price at which the debt security is traded at issuance.
- Indenture: the contract that states all of the terms of the debt security.

2.1.3 Sources of Risk

Risks associated with fixed-income securities include, but are not limited to, the following:

- Interest rate risk: As interest rate increases, fixed-income securities lose values, and hence their price decrease. Changes in interest rates are the major drive of changes and volatility in bonds prices.
- Inflation risk: Similarly, fixed payments imply a change in purchasing power when the rate of inflation changes, giving rise to inflation risk.
- Credit risk: Credit risk, also known as business risk or financial risk, refers to the likelihood that an issuer would fail to meet its debt obligations, as could be the case, for instance, in the corporate bond market.
- Liquidity risk: This risk is prompted by the scarcity of some instruments; it represents the likelihood that an investor cannot find a buyer to divest his fixed-income asset.

2.2 Bonds

Bonds are fixed-income instruments used by governments, companies, and municipalities to finance their debts, projects and operations. These securities are attractive to investors because they i) provide a predictable income, ii) help offset exposure to volatile instruments such as equities, and iii) allow to preserve capital while investing.

2.2.1 Categories and flavours

Bonds are traded in the fixed-income market in many different varieties according to the terms agreed upon in the indenture. These instruments vary according to the issuer type, the coupon payment type, the maturity type, and multiple other attributes. *Plain-vanilla* bonds refer to the basic setup with respect to coupon and maturity payments.

Definition 2.2.1 (Plain-vanilla bond). *A plain-vanilla bond is a fixed-income indebtedness security, wherein issuer (debtor) owes the holder (creditor) a debt, and is required to reimburse by paying interest (the coupon) and the face value upon maturity, according to the contract terms. Interest payments are typically made at regular intervals of time called payment frequency (semiannual, annual, sometimes monthly).*

Many variations exist around the plain-vanilla bond. Appendix A describes the various classifications of bonds, their possible special features, and the corresponding terminology.

2.2.2 Valuation

In the following section, we present the most commonly used approach to determine the value of a bond. We then explain the bond's yield, which is widely used to characterize the value of a bond. We conclude by describing bonds' credit rating and the various scores that can be attributed to this class of instruments.

2.2.2.1 The Discounted Cash Flow Approach

Definition 2.2.2 (Bond price). *The fair value of a bond is the present value of all expected future cash flows (CFs) that the bond will generate. This value is obtained by discounting the bond's expected CFs to the present date using the appropriate discount rate.*

The value of a bond depends on multiple of its characteristic features, such as its maturity, the creditworthiness of its issuer, and its coupon rate at issuance compared with current interest rates.

Under the discounted cash flow approach, the value of a bond is assimilated to the present value of an investment opportunity with deterministic future returns. Assuming a discrete discount rate, the value P of a plain-vanilla bond is then given by

$$\begin{aligned} P &= \left(\sum_{n=1}^{N} \frac{C}{(1+i)^n} \right) + \frac{F}{(1+i)^N}, \\ &= C \left(\frac{1-(1+i)^{-N}}{i} \right) + F\,(1+i)^{-N}. \end{aligned} \tag{2.2.1}$$

where

F: the face value, often normalized to \$1000,

$C = c \times F$: the periodic coupon payment, with c the periodic coupon rate,

N: the number of coupon payments,

n: the periodic coupon dates index,

i: the periodic discount rate.

Valuation becomes more complex when a bond has one or more embedded options. One possible approach is to add the value of the embedded option(s) to that of the plain-vanilla bond given in Equation 2.2.1. Since embedded options are generally contingent claims, their value can be obtained using an appropriate numerical or analytical evaluation method. However, it is often the case with complex bonds that option and bond values are not additive.

2.2.2.2 Accrued Interest

Equation 2.2.1 allows to compute the value of a bond at issuance or at a given coupon date. At any intermediate date, the value of a bond must be adjusted for *accued interest*. Bond ownership can be transferred between investors at any point during the life of the bond; if a bond is sold between two coupon dates, accrued interest accounts for the fact that the seller owns a part of the next coupon.

Definition 2.2.3 (Accrued interest). *The accrued interest on a bond refers to the interest that has accumulated but not yet been paid since the principal investment, or since the most recent coupon payment.*

Accrued interests are generally computed based on an agreed-upon day-count convention, which is a standardized methodology for calculating the number of days between two coupon dates, using the following linear approximation:

$$I_A = t \times F \times c_A. \tag{2.2.2}$$

where

I_A: the accrued interest,

t: the elapsed period, expressed in years, using the agreed upon day-count convention,

c_A: the annualized coupon rate.

The accrued interest creates two different quotes for bond prices, which leads to two different terms used in financial markets, that is, the *clean* and the *dirty* prices.

Definition 2.2.4 (Clean price). *The clean price of a bond refers to the price that does not include any previous or current accrued interest.*

Definition 2.2.5 (Dirty price). *The dirty price of a bond is the price that takes into account the accrued interest, which is obtained by summing up the clean price and the accrued interest.*

The value of the bond, and the actual price at which it is traded, correspond to the dirty price.

The fair value (eventually adjusted for accrued interest) is a way to assess bonds having different coupons, maturities and/or face values. The discount rate used by an investor to assess the value of a bond according to the discounted cash flow approach is normally equal to the return that the investor can secure on comparable investments (e.g. in terms of maturity, liquidity, or credit risk). At a given rate, Equation 2.2.1 allows the investor to compare the fair value of a given bond to its market-quoted price. Equation 2.2.1 can also be inverted to determine the discount rate that would equate the fair value of a bond to its market-quoted price, leading to the concept of yield.

2.2.2.3 Yield

Definition 2.2.6 (Yield). *The yield is a metric used to assess common stocks, preferred stocks, convertible stocks, and fixed income instruments, including bonds. The yield is a measure of the ex-ante return received by the security holder.*

In the case of bonds, Equation 2.2.1 defines what is called the yield-to-maturity.

Yield-to-maturity: The *yield-to-maturity* (YTM), also known as book yield or redemption yield, is an estimate of the total return expected to be earned by an investor under the following assumptions:

1. The bondholder keeps the security until the maturity date,
2. The issuer respects all the coupon and capital payments schedule,
3. The bondholder is able to reinvest all interest payments at the YTM and earn the benefit of compounded returns.

The YTM is obtained by solving Equation 2.2.1 for i, using the market price for P. It is usually expressed as an annual rate. The YTM is a useful metric to compare bonds with different prices, coupons, face values and maturities. Clearly, for a fixed coupon schedule and face value, the YTM of a bond is inversely related to its market price. Equation 2.2.1 also allows to relate the coupon rate of bonds to their YTM according to the relative value of their market price w.r.t. their face value.

Definition 2.2.7 (Bond at par). *A par bond refers to a bond that has a market price equal to its face value, which means that its coupon rate is equal to its YTM.*

Definition 2.2.8 (Bond at discount / premium). *A discount (resp. premium) bond is a bond that is traded at a market price that is lower (resp. higher) than its face value, which means that it offers a coupon rate that is lower (resp. higher) than its YTM.*

The YTM is the most commonly used metric to assess the rate of returns of bonds and will be the yield used in the sequel to compare green and brown bonds. Other types of yields are also used to characterize bonds under different assumptions about how long the investor expects to hold the security.

Coupon yield: also known as coupon rate, the *coupon yield* is the amount of income interest that investors can expect to receive as long as they hold the bond. It represents the percentage of the yearly interest rate paid by the bond with respect to its face value and is obtained by

$$\text{Coupon yield} = \frac{c_A}{F}. \tag{2.2.3}$$

Current yield: also known as running yield, the *current yield* is the annual coupon payment divided by the current price of the bond. This measure evaluates the yield of the bond at the current moment, rather than reflecting the total return over the life of the bond. Thus, the current yield represents the return an investor would expect to earn by purchasing the bond and holding it for one year, which is different from the actual return the investor would get by purchasing and holding the bond until maturity. The current yield is computed by

$$\text{Current yield} = \frac{c_A}{P}. \tag{2.2.4}$$

Yield-to-call, Yield-to-worst: these measures are used to assess bonds having uncertain maturity dates, for instance when they include embedded call or conversion options. The yield-to-call is computed assuming that the call option will be exercised. The yield-to-worst corresponds to the lowest possible yield for bonds with multiple call options.

Yield measures allow the comparison of bonds with different contractual characteristics; these measures are computed under the assumption that all contractual payments will be made by the issuer. However, credit risk is an important component of the quality of a bond, which is reflected in its value or market price. For that reason, yields can differ according to the credit quality of the issuer.

Definition 2.2.9 (Yield spread). *The yield spread is the difference between the yields on two different investments, usually of different credit qualities but similar maturities. This difference is most often expressed in basis points (bps).*

Yields can also differ according to the maturity of securities, as represented by the *yield curve.*

Definition 2.2.10 (Yield Curve). *The yield curve, also known as term structure of interest rates, is a graph that shows how the yields on debt securities fluctuate depending on the remaining time to maturity. The horizontal x-axis of the graph is typically a time line of months or years, and the vertical y-axis represents the annualized YTM.*

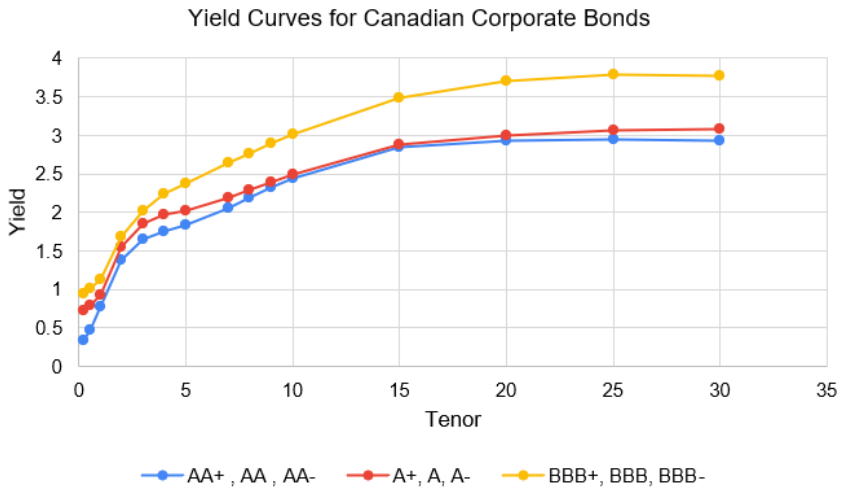


Figure 2.2.1: Yield Curves for AA^+, AA, AA^- (blue), A^+, A, A^- (red) and BBB^+, BBB, BBB^- (yellow) Canadian Corporate bonds as of 12/31/2021 (Source: Bloomberg).

A properly constructed yield curve should be built from a group of instruments with varying remaining times to maturity, with all YTMs calculated as of the same point in time. Furthermore, all securities included in the computation of the yield curve must have comparable credit ratings in order to eliminate the credit risk differentials effect. Figure 2.2.1 is an example of yield curves of Canadian corporate bonds for three different ratings categories (see Section 2.2.2.5).

The shape and slope of the yield curve are thought to be related to changes in investors' expectations for the economy, and there are three well-known shapes of yield curves:

- Normal curve: an upward sloping curve, meaning that the YTMs increase as time to maturity increases. A positive slope reflects investors' expectations of future economic growth.
- Inverted curve: a downward sloping curve, meaning that short-term interest rates are higher than long-term ones. A negative slope corresponds to periods of economic decline, where investors expect returns to decrease in the future.
- Flat curve: when YTMs are similar across all maturities. A few mid-term maturities may have slightly higher YTMs, resulting in a minor hump along the flat curve. A yield curve that is flat or humped often indicates an uncertain economic situation.

2.2.2.4 Bonds Premiums

It is important to note that a bond's theoretical price and yield satisfying Equation 2.2.1 may be different from the actual quoted price and yield available in the market. This is mainly due to the existence of premiums in a bond's market price and yield, such as:

- Liquidity premium: a compensation that aims to encourage investors to invest in instruments that are illiquid, i.e. cannot be easily and efficiently converted into cash.
- Risk premium: a compensation that aims to encourage investors to tolerate instrument's extra risk over that of a risk-free asset, that is, the risk that the issuer will fail to meet its debt obligations.

Other premia may appear to compensate the investor for other sources or risk, as mentioned in Section 2.1.3.

2.2.2.5 Bond ratings

As mentioned in Section 2.1.3, bonds are associated with different types of risks, including credit risk. A number of rating agencies are in the business of providing bond ratings, which is an evaluation of the creditworthiness of their issuers.

Definition 2.2.11 (Rating Agency). *A rating agency evaluates a company's or government agency's financial strength and capability to satisfy its debt payments, then designates it by a letter grade that represents the investor's trust towards that firm or government, as well as the likelihood that the debt will be repaid.*

The US Securities and Exchange Commission recognises three major credit rating agencies as Nationally Recognized Statistical Rating Organizations: Standard and Poor's, Moody's Investor Services, and Fitch Group. Rating agencies assess bond's risk through a top-down forecasting of broad economic conditions, an in-depth bottom-up analysis of the instruments' features, and statistical estimates of the firm's default probability.

Bond ratings are denoted by letters ranging from "AAA" (the highest grade) to "D" (the lowest grade). Multiple rating agencies employ the same letter grades but differentiate themselves by using different combinations of upper- and lower-case letters and modifiers. Bonds are classified into two types depending on their corresponding credit rating (see Table 2.2.1)

1. Investment grade bonds: these are the bonds with higher ratings and they are believed to be more secure and stable instruments. These products are linked to publicly traded firms and governmental institutions with positive outlooks. Standard and Poor's rates investment grade bonds as "AAA" to "BBB-," while Moody's rates them as "Aaa" to "Baa3".
2. Junk bonds: these are bonds with lower ratings. These bonds are considered riskier investments and may be interesting for some investors because of their higher yields. However, junk bonds may present liquidity problems and may default, leaving investors with nothing. They typically have Standard and Poor's ratings ranging from "BB+" to"D", or from "Baa1" to "C" for Moody's.

Moody's	Standard & Poor's	Type	Risk
Aaa	AAA	Investment	Lowest Risk
Aa	AA	Investment	Low Risk
A	A	Investment	Low Risk
Baa	BBB	Investment	Medium Risk
Ba/B	BB/B	Junk	High Risk
Caa/Ca/C	CCC/CC/C	Junk	Highest Risk
C	D	Junk	In Default

Table 2.2.1: Bond rating, type and risk level for Moody's and Standard and Poor's (Source: Investopedia).

2.3 Green Bonds

Environmental, social, and corporate governance (ESG) refer to the three major factors in measuring the sustainability and societal impact of an investment in a company or business. Socially conscious investors use these three criteria to evaluate companies in which they might want to invest. Green bonds are a distinguished element of ESG and socially responsible investing, as they enable investors to finance green projects.

Definition 2.3.1 (Green Bond). *A green bond, also referred to as climate bond, is a financial instrument that has the same characteristics and specifications as a normal bond (see Section 2.2), with however an aim to finance green projects.*

Green bonds fund projects that promote sustainable energy, environmental protection, sustainable agriculture, fisheries, and forestry, marine and terrestrial ecosystem preservation, green transportation, clean water, and sustainable water management. They are also intended to promote sustainability and the development of ecologically friendly products, as well as climate change mitigation.

In response to a demand from a group of Swedish pension funds looking to engage in climate-conscious investment, the World Bank issued its first green bond in November 2008. It was considered to be the first of its kind globally, and served as a model for today's green bond market, allowing investors to finance green technologies while earning a profit.

The attributes of green bonds can essentially be classified into four categories, as follows:

- Green use of proceeds bond: a conventional financial obligation with recourse to the issuer for which the earnings are traced and verified by the issuer through a structured internal process related to the issuer's borrowing and investing activities.
- Green revenue bond: a conventional financial obligation with non recourse to the issuer in which the bond's credit risk is to the promised revenue, fees, taxes, etc, and the earnings of the bond are used to fund related or unrelated green projects. The earnings

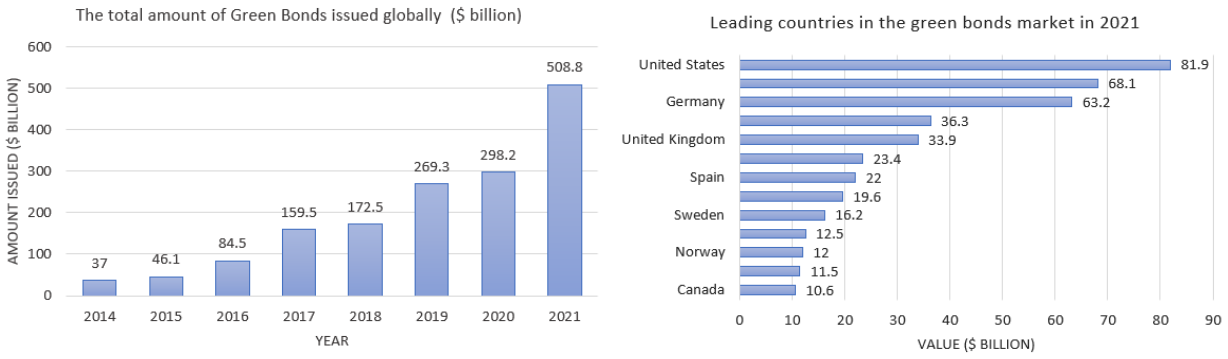
are traced and verified by the issuer through a structured internal process related to the issuer's borrowing and investing activities.

- Green project bond: is linked to one or more green projects in which the owner has direct risk exposure.
- Green securitized bond: a bond backed up by a single or multiple projects, such as covered bonds, asset-backed securities, or other forms. The cash flows from the assets securing the bonds are typically the first source of repayment.

2.3.1 Green Bonds Market Size

The growth of green bonds in the capital markets has been explosive and these are increasingly attracting attention from investors. The Climate Bond Initiative (CBI) claims that green bonds experienced a 49% growth rate in the five years preceding 2021, with an annual issuance that could exceed $1 Trillion by 2023.

The CBI announced that issuance volumes of Green, Social, and Sustainability (GSS) debt reached nearly half a trillion dollars ($496.1 Billion) during the two initial quarters of 2021, illustrating a 59% increase over 2020, with green bond issuance doubling to $227 Billion. Furthermore, the CBI determined a cumulative GSS issuance of $2.1 Trillion by the end of 2021, including cumulative green debt issuance of $1.3 Trillion. The CBI stated that green bond issuance reached $508.8 Billion in 2021, and estimated that the annual green bond issuance would reach $1 Trillion by 2023 (see Figure 2.3.1a). In Europe, a noticeable increase was highlighted in October 2021, when the European Union issued approximately $14 Billion in bonds, making it the biggest transaction ever. These funds will be used to finance projects such as a research platform for energy conversion in Belgium and wind power infrastructure in Lithuania.



(a) Global Green Bonds Issuance

(b) Green Bonds value by Country

Figure 2.3.1: Global Green Bonds Issuance from 2014 to 2021, and value of the Green Bonds market in leading countries as of 2021 in Billion U.S. Dollars (Source: CBI).

2.3.2 Green Bonds Principles

The International Capital Markets Association (ICMA) established the Green Bonds Principles (GBP) as voluntary guidelines, allowing to facilitate the growth of the green bond market (see the GBP document in Association et al. [2018]). The publication of these principles aims to promote transparency, disclosure, as well as integrity, and to provide guidance to issuers on the main attributes to satisfy when issuing green bonds.

The GBP recommend that issuers clearly and transparently communicate their use of proceeds categories to investors to help them make decisions regarding the bond's consistency with their investment strategy, so that investors are better positioned to assess the environmental and/or social impact of their investments.

The GBP comprise the four following elements:

1. Use of proceeds: Issuers ought to divulge the projects eligible for green investment. The GBP recommends to clearly declare and quantify environmental advantages, and/or assess them wherever possible. The GBP include a non-exhaustive list of officially approved green projects, such as, renewable energy, energy efficiency, pollution prevention and control, conservation of terrestrial and aquatic biodiversity, etc.

2. Project evaluation and selection process: Issuers have to explain the decision-making behind the evaluation of projects eligibility, including the sort of projects they are intended to finance, the criteria for evaluating the ecological benefits, and the environmental influence they anticipate the projects to provide. A third-party review can supplement the processes for project evaluation and selection.

3. Management of proceeds: The net earnings have to be transferred to a sub-portfolio or else traced by the issuer and documented by a structured internal process related to the issuer's borrowing and investing activities. The GBP recommends that issuers disclose to investors the sorts of provisional investments intended for the remaining unassigned earnings.

4. Reporting: Issuers should report on the specific investments made with the proceeds of green bonds at least once a year, through newsletters, website updates, or filed financial reports, detailing each individual project and amounts allocated, as well as the anticipated long-term benefits for the environment. Investors are constantly interested in the disclosure of the concrete environmental impact as a prominent evaluation metric that holds issuers accountable for the accomplished environmentally sustainable impact, and also as a mechanism to estimate their own investment performance in terms of sustainability.

2.4 Machine Learning Algorithms

Machine learning (ML) is a sub-category of artificial intelligence that relies on mathematical and statistical approaches to give computers the ability to "learn" from data. All types of digital information can be used as data for ML. By detecting patterns in this data, the algorithms learn and improve their performance in accomplishing a specific task, such as prediction, optimization, or forecasting.

There are five main steps in developing a ML model.

1. The first step consists of collecting relevant data in sufficient quantity and quality, and avoiding any bias in its representativeness. This data must be carefully prepared, organised and cleaned before use. Indeed, some attributes are irrelevant, others need to be modified in order to be understood by the algorithm, and some features have missing information. For this matter, several techniques such as data visualization, data transformation or normalisation are used.

2. The second step consists of feature extraction, also known as feature engineering, in which attributes can be combined with each other to create new ones that are relevant and effective for training the model. Subsequently, the dataset is divided into a training and a testing set, which will be later used to estimate the performance of the model.

3. The third step is to select an algorithm to run on the training data set. The choice of algorithm to be used depends on the type and volume of training data and the type of problem to be solved.

4. The fourth step is the actual training of the algorithm. This is an iterative process that involves running the algorithm, then examining the results, and applying the right adjustments to improve the performance of the model.

5. The fifth and final step is to deploy the model on a new dataset, different from the training one (the testing set). The efficiency of the model on this new dataset is investigated using various evaluation metrics.

Traditional ML approaches are classified into three broad categories based on the nature of the learning system.

- Supervised learning, where for each observation, there is an associated response measurement, also known as target variable. This setting seeks to fit a model that relates the target to the predictors, with the aim of accurately predicting the target for future observations (prediction) or better understanding the relationship between the target and the predictors (inference). Supervised learning methods include many regression-based approaches.

- Unsupervised learning applies to a more challenging situation in which there exist no target variable associated to each observation.

- Reinforcement learning is a behavioural ML model that is similar to supervised learning except that the algorithm is not trained on sample data. This model learns as it goes through its output, using trial and error.

In this work, we will use supervised learning ML models, where the target is to evaluate the value of green and brown bonds, given their salient features. The next sections describe the various supervised models used in this thesis, as well as the mathematical framework incorporated in each model, namely linear regression, decision tree, support vector machine (SVM), partial least squares regression (PLS), and multivariate adaptive regression spline (MARS).

2.4.1 Linear Regression

Linear regression (Zou et al. [2003]) is a statistical method for modelling the relationship between a quantitative dependent variable and one or more independent variables (also known as response or predictors). It is considered the simplest ML model, and has various extensions such as multiple linear regression (Eberly [2007]) and polynomial regression (Ostertagová [2012]). Simple linear regression is used when there is only one independent variable; whereas multiple linear regression is used when there are more than one.

The simple linear regression is given by the following:

$$\hat{y} = \beta_0 + \beta_1 x. \tag{2.4.1}$$

where

$\hat{y}$ is a prediction of the dependent variable y,

β_0 is the intercept,

β_1 is the slope,

x is the independent variable.

The multiple linear regression is given by the following:

$$\begin{aligned} \hat{y} &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \\ &= \beta_0 + \sum_{k=1}^{p} \beta_k x_k. \end{aligned} \tag{2.4.2}$$

where

$\beta_1, \ldots, \beta_p$ are the slopes for the independent variables $x_1, \ldots, x_p$,

$x_1, \ldots, x_p$ are the independent variables,

p is the number of independent variables included in the model.

The least-squares technique is the most commonly used for fitting the linear regression model. This consists of obtaining the best-fitting line by minimizing the sum of the squares of the residuals, i.e. the difference between the observed values obtained from the data, and the predicted values provided by the model.

2.4.2 Decision Tree

Decision tree is a predictive supervised learning method, first introduced by Belson [1959], that can be used for classification and regression. The goal is to build a model that predicts the value of a target variable using simple decision rules derived from data features. It employs a decision tree (as a predictive model) to progress from observations about an item (represented by branches) to conclusions about the item's target value (represented in the leaves). This model can be thought of as a piecewise constant approximation, and it is considered among the most popular ML algorithms thanks to its comprehensibility and simplicity.

Some key terms used in the decision tree model are:

- Root: the node present at the beginning of a decision tree, from which the population starts dividing according to various features.
- Decision nodes: the sub-nodes obtained after splitting nodes in the previous level.
- Leaf nodes: also known as terminal nodes, these are nodes where further splitting is not possible.
- Branch: a subsection of the decision tree consisting of a succession of multiple nodes.
- Pruning: the process of cutting down some sub-nodes of a decision tree.

The relationship between the dependent variable and the features is given by the following formula:

$$\hat{y} = \sum_{n=1}^{N} l_n \mathbb{1}_{\{x \in R_n\}}. \tag{2.4.3}$$

where

$R_1, \dots R_N$ are the leaf nodes,

$l_1, \dots l_N$ are the averages of all training observations located in each of the leaf nodes $R_1, \dots R_N$,

N is the number of leaf nodes,

$\mathbb{1}$ is the identity function.

Various extensions of the decision tree approach exist, such as bagging and random forests. These extensions involve producing multiple trees that are combined to yield a single consensus prediction.

Bagging was first introduced by Breiman [1996] and it consists of bootstrapping multiple times the training dataset in order to obtain different datasets, thus different predictions for the target. The final estimate is then obtained by averaging all the predictions.

Random forests were developed by Breiman [2001] and it also entails constructing multiple decision trees from bootstrapped datasets of the original training data and randomly selecting a subset of independent variables at each step of the decision tree. The final estimate is then obtained by choosing the mode of each decision tree's predictions.

2.4.3 Support Vector Machine (SVM)

SVM is a supervised learning model developed by Vapnik [1999] that analyzes data for classification problems. The equivalent model treating a regression problem is known as Support Vector Regression (SVR). This model's central idea consists of finding the maximum marginal hyperplane (MMH) in a multidimensional space that best divides the dataset into classes. For this purpose, SVM iteratively generates optimal hyperplanes, which are used to minimize a predefined error.

Some key terms used when dealing with SVM are:

- Support vectors: the data points, which are closest to the hyperplane, and which will be used to define the separating lines.
- Hyperplane: a decision plan that divides between the groups of data that belong to distinct classes. Intuitively, the further the data points are from the hyperplane, the more accurate the model is.
- Margin: the distance between the hyperplane and the nearest data point from the distinct classes. This is calculated as the perpendicular distance between the hyperplane and the nearest support vectors. A good margin refers to a relatively large distance between classes.

Generally, the SVM algorithm uses the Hinge loss function to maximize the margin between the data points and the hyperplane. The Hinge loss function is defined by

$$H(x, y, f(x)) = max(0, 1 - y \times f(x)). \tag{2.4.4}$$

The SVM classifier is computed by minimizing a cost function that takes the form

$$\frac{1}{n}\sum_{i=1}^{n} H(x_i, y_i, w) + \lambda||w||^2. \tag{2.4.5}$$

where:

$x_1, \ldots, x_n$ are the independent variables of the training set,

$y_1, \ldots, y_n$ are the dependent variables of the training set,

n is the size of the training set,

λ is a regularization parameter used to weight the margin maximization vs. the loss.

w is the margin to be optimized.

In addition to performing linear predictions, SVM can efficiently perform a non-linear prediction using what is called the kernel trick, which consists of implicitly mapping the inputs into high-dimensional feature spaces. The resulting algorithm is formally similar to the original linear one, except that each dot product given in the cost function (Equation 2.4.5) is replaced by a nonlinear kernel function, such as homogeneous polynomial, complex polynomial, or Gaussian radial basis function. The algorithm can then fit the maximum-margin hyperplane in a transformed feature space.

2.4.4 Partial Least Squares Regression (PLS)

PLS was first introduced by Wold [1966]. It is a predictive technique that combines features from Principal Component Analysis (PCA) and multiple regression. PLS is an alternative to Ordinary Least Squares (OLS)-based methods, and it is especially useful when the set of dependent variables is highly correlated or when the number of independent variables exceeds the number of observations.

At the core of PLS regression is a dimension reduction technique that assumes a latent decomposition of the target and predictors matrices, by projecting them into a new space. The general matrix decomposition in the PLS model takes the following form:

$$X = TP^T + E. \tag{2.4.6}$$

$$Y = UQ^T + F. \tag{2.4.7}$$

where

X is the predictors matrix,

Y is the response matrix,

T and U are, respectively, projections matrices of X and Y, also known by scores of X and Y,

P and Q are, respectively, orthogonal loading matrices,

E and F are, respectively, error term matrices assumed to be independent and identically distributed random Normal variables.

The latter matrix decomposition is obtained through successive optimization problems, in which one seeks to find the projection matrix U that has the maximum covariance with the projection matrix T. Once the projection matrix U is determined, the PLS model uses this decomposition to find the predictions of the target variables by applying Equation 2.4.7.

2.4.5 Multivariate Adaptive Regression Spline (MARS)

MARS was first introduced by Friedman [1991]. It is a non-parametric model that extends linear models by automatically modeling nonlinearities and interactions between variables. While generalized linear models and generalized additive models assume that the coefficients of the independent variables are constant across all values of a predictor, the MARS algorithm explicitly recognises that this is not always the case.

The MARS algorithm creates predictions of the following form:

$$\hat{y} = \sum_{i=1}^{k} \alpha_i B_i(x). \tag{2.4.8}$$

where:

$\alpha_1, \dots \alpha_k$ are constant coefficients,

$B_1, \dots B_k$ are basis functions.

Each basis function can take one of the following three forms:

1. A constant, which will only be the case for the intercept.
2. A Hinge function under either of the following two forms:

$$H(x) = \begin{cases} max(0, \text{constant} - x). \\ max(0, x - \text{constant}). \end{cases} \tag{2.4.9}$$

3. A combination of two or more Hinge functions that can model the interactions between two or more variables.

The MARS algorithm constructs a model in two stages, a forward pass and a backward pass. It begins with a model that consists solely of the intercept term equaling the mean of the response values. Then, it evaluates each independent variable in order to find the basis function pair consisting of opposing sides of a mirrored Hinge function that produces the greatest improved performance in model error. The process is repeated until the algorithm reaches a predefined limit of terms or the error improvement reaches a predefined threshold.

Chapter 3

Data Preprocessing

In ML, data preprocessing refers to the technique of preparing, cleaning and organizing raw data in order to build and train ML models. This is an important step that assists the extraction of meaningful insights from the data, and helps improve its quality. It is generally divided into four steps: i) data quality assessment, ii) data cleaning, iii) data transformation, and iv) data reduction. This chapter describes the data sources and the data sets resulting from the data preprocessing step. In order to estimate the greenium, two sources of data are needed, containing prices and features of brown and green bonds respectively.

3.1 Data Collection and Cleaning

We collected the data from two data providers with bonds issued no less than 2007-01-02. The brown bonds database was obtained from the Mergent Fixed Income Securities Database (FISD), which is widely used for empirical research on the corporate bonds market. The FISD data base contains characteristics of all publicly-traded U.S. bonds. The green bonds database was collected from the Bloomberg terminal. Each green bond provided in the Bloomberg list has obtained the green labelling based on the GBP.

3.1.1 Brown Bonds Data

Table 3.1.1 describes the output files generated from the FISD database.

File Outputs			
Name	**Description**	**Number of Features**	**Size**
Issue	lists basic characteristics of each issue in the database	24	514,625
Issuer	contains information on the issuer's industry, current financial status and corporate parent	2	16,056
Rating	lists the Standard & Poor's ratings for each issue in the database	2	70,499
Coupon_Info	lists the initial interest rate and interest payment frequency for all issues in the database	3	514,623
Foreign_Currency	lists the issuing currency, par amount and exchange rate as of the issuance date for non-U.S. dollar denominated securities in the database	2	6,599
Treasury	lists characteristics of each issue such as the auction date, reopening information on bids, yields, prices, ratio's and the tail	2	8,806

Table 3.1.1: The files generated from FISD, including their description, number of features and size.

After investigating the size of each output file, we eliminate the files i) Issuer, ii) Foreign_Currency and iii) Treasury because they contain a limited number of bonds and because the relevance of this information is not expected to warrant the important reduction in the data. We then merge the i) Issue, ii) Coupon_Info and iii) Rating files, obtaining a dataset of 70,499 observations.

As part of the data preprocessing, we eliminate the non-vanilla bonds because they differ in terms of yield and price computation. Hence, we search for bonds that have embedded options, that is, redeemable, putable, perpetual, exchangeable, fungible and preferred bonds, as well as bonds with variable or zero coupons and we eliminate them from the dataset.

We compute the "Time to Maturity" for each remaining bond, using its "Offering Date" and "Maturity Date".

We then investigate the missing information contained in the remaining dataset. Thus, we search for bonds with missing yields, prices, coupons, ratings and time to maturities, which we deem the most crucial information for our purpose, and we remove them from our selection.

Finally, we check for the existence of outliers in the dataset and we conclude that the fi-

nal data obtained does not contain outliers or noisy information due to data collection errors.

Furthermore, as mentioned in Section 2.2.2, the term structure of interest rates may be a significant determinant in the evaluation of a bond's value. Since the FISD database contains bonds traded in the US market, we incorporate the H.15 release from the Federal Reserve System [2022]. The H.15 release contains daily US interest rates from 01-02-1962 for 11 different tenors (1 month to 30 years). We use this information to associate with each bond of the dataset an interest rate corresponding to its issue date and time to maturity, using linear interpolation between the reported tenors.
Finally, we obtain a brown bond dataset of size 52,563 containing the features described in Table 3.1.2.

Feature	**Description**
Issue ID	A Mergent-generated number unique to each issue used to link each issue's data among the other tables.
Coupon frequency	Code indicating how often coupon payments will be made.
Coupon	The current applicable annual interest rate that the bond's issuer is obligated to pay the bondholders in annual percentage.
Day count basis	Code indicating the day count basis that is agreed upon.
Prospectus issuer name	The name of the issuer as in the prospectus.
Issuer CUSIP	A unique code assigned to the issuer by the Committee on Uniform Securities Identification Procedures.
Issue CUSIP	A unique code assigned to the issue by the Committee on Uniform Securities Identification Procedures.
Issue name	Issue type description as taken from the prospectus.
Maturity date	The issue maturity date.
Time to maturity	The time to maturity of the corresponding bond.
Interest Rate	The interest rate, as of the offering date of the bond, that corresponds to its time to maturity.
Security level	Indicates if the security is a secured, senior or subordinated issue of the issuer.
Offering amount	The volume of debt initially issued in thousand dollars (K$).
Offering date	The date the issue was originally offered.
Offering price	The price in dollars at which the issue was originally sold to investors.
Offering yield	YTM at the time of issuance, based on the coupon and any discount or premium to par value at the time of the sale in annual percentage.
Delivery date	The date the issue was or will be initially delivered by the issuer of the security.
Principal amount	The face or par value of the bond in dollars.
Rating	The Standard & Poor's rating assigned to each bond.

ISIN	The International Securities Identification Number associated with the issue.

Table 3.1.2: Description of features extracted from the FISD dataset, and included in the final brown bonds dataset.

3.1.2 Green Bonds Data

Green bonds still rarely exist in the market, and obtaining a database that contains this type of instruments has been a noteworthy challenge in this project. We searched in different websites such as the Climate Bonds Initiative, Environmental Finance, as well as Bloomberg, in order to collect as much data as possible. Finally, we decided to use only the data provided by Bloomberg, in which there is an identification of green labelling for each bond through an examination if they satisfy the GBP mentioned in Section 2.3.2. We obtain a total of 4,848 green bonds from Bloomberg. For further information about the green bonds labelling on the Bloomberg Terminal, we refer the reader to the BNEF Bloomberg Terminal Guide [2015].

Furthermore, the FISD database, from which we extract the brown bond data, only contains bonds that are traded in the US market. Hence, we extend this constraint into our green bonds database in order to avoid any data or currency mismatch. After removing non-US bonds, the size of the green bond sample reduces to 340. Finally, as for brown bonds, we also limit the sample to vanilla green bonds and we delete all instruments with embedded options, further reducing the size of the green bonds dataset to 82.

Table 3.1.3 shows the length of the the green bonds dataset extracted from Bloomberg, after removing non-US traded, and non-vanilla bonds.

	Green bonds data	**After removing non-US bonds**	**After removing non-vanilla bonds**
Size	4,848	340	82

Table 3.1.3: Description of the size of the green bonds database extracted from Bloomberg after applying the filter on the country and the embedded options.

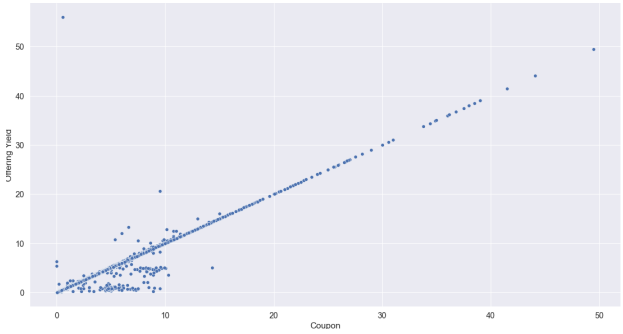
Finally, we search for the identified green bonds in the FISD database using their CUSIP and/or ISIN, and we extract them in a separate database in order to i) obtain green bonds data that have all the needed features mentioned in Table 3.1.2, and ii) ensure that the FISD database contains only brown bonds.

3.2 Data Visualization

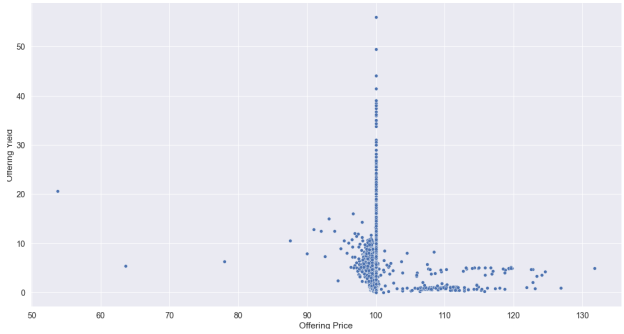
The data quality assessment process consists of examining carefully the data in order to check its quality, its relevance aligned with the project's objective, as well as its consistency. One way to assess the data is to visualize it in order to gain a better understanding of the relationship between the different features. In this section, we focus on gaining a more concrete perspective about the brown and green bonds databases by displaying the relationships between the different variables and the target, as well as the distribution of some specific features.

3.2.1 Brown Bonds

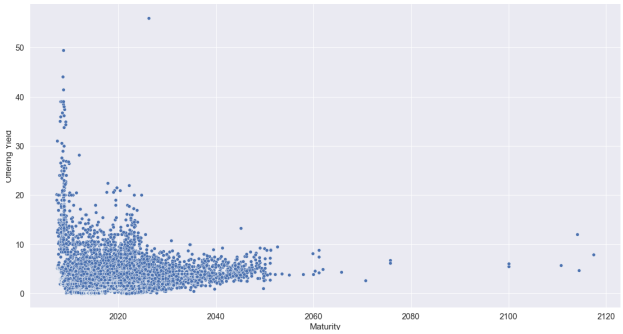
Figure 3.2.1 shows the relationship between the feature of interest in our research, namely the offering yield, and the other characteristics of the bonds, specifically the coupon, offering price, maturity, as well as the credit rating of the bond.



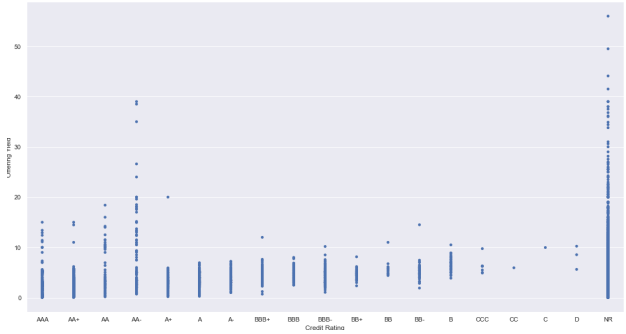
(a) Relationship between the offering yield and the coupon.



(b) Relationship between the offering yield and the offering price.



(c) Relationship between the offering yield and the maturity.



(d) Relationship between the offering yield and the credit rating.

Figure 3.2.1: Visualization of the brown bonds relationship between the offering yield and a) the coupon, b) the offering price, c) the maturity date, and d) the credit rating.

Figure 3.2.1b shows that the majority of the bonds have an offering price of $100, implying

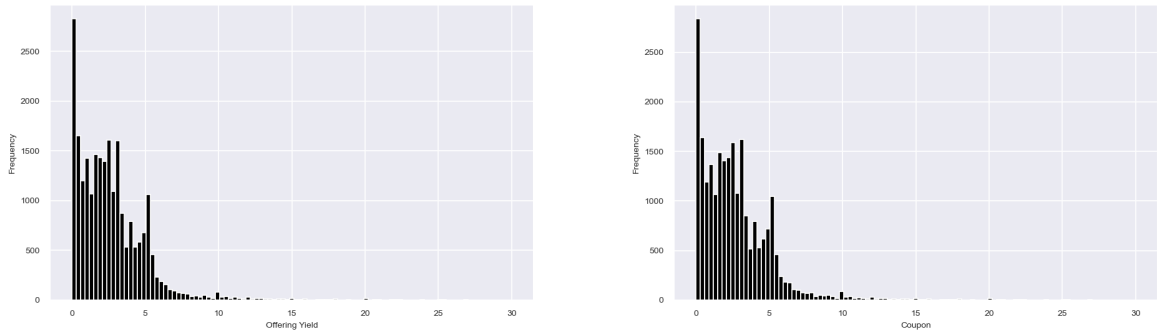
that the majority of bonds are sold at par in the primary market. We also see that the offering price ranges between $63 and $133, with no obvious outliers in the brown bonds dataset.

Figure 3.2.1a depicts a predominant linear relationship between the offering yield and the bond coupon. As indicated in Section 2.2.2.3, this is a reasonable expected result since most of the bonds in the dataset are offered at par.

Furthermore, Figure 3.2.1c shows that the majority of bonds mature between 2015 and 2060, with only a few bonds with a maturity date reaching 2120. As a result, we believe that we have well-distributed data between the various maturity dates, which is considered necessary in the ML framework in order to capture the impact of features.

Figure 3.2.1d depicts the distribution of the issuer's credit rating as determined by Standard & Poor's, and the offering yield. We notice that our selection contains bonds with a fairly even distribution of credit ratings, where each credit rating is represented by a large number of bonds. Again, this is considered positive in the ML framework.

Figure 3.2.2 displays the data distribution of the most important features, in order to investigate the existence of a proportionally distributed data and further investigate the existence of outliers.



(a) Histogram of the offering yield.

(b) Histogram of the coupon.

Figure 3.2.2: Visualization of the histograms of the brown bonds a) offering yield, and b) coupon.

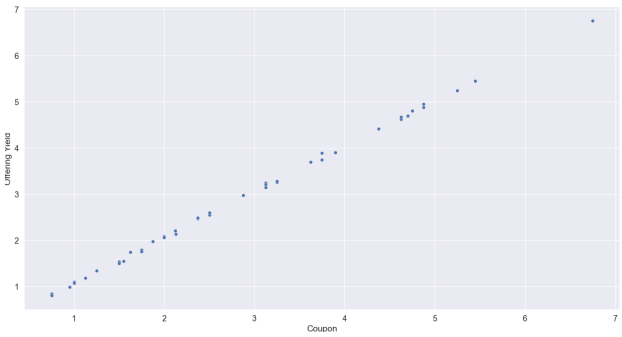
Figure 3.2.2a displays the histogram of the offering yield in our brown bonds selection, and shows well distributed data, with a yield ranging between 0% and 13%, with only a few bonds reaching a yield of 20%.

This same observation can be made from Figure 3.2.2b displaying the coupon histogram,

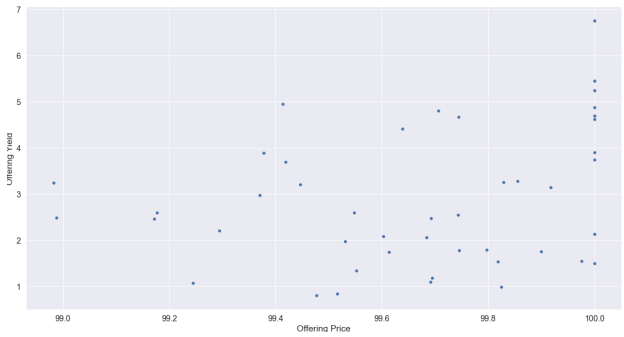
where the distribution ranges between 0% and 13%. As outlined above, we expect similar histograms for the offering yield and coupon, because we are considering primary market data, where most of the bonds are offered at par.

3.2.2 Green Bonds

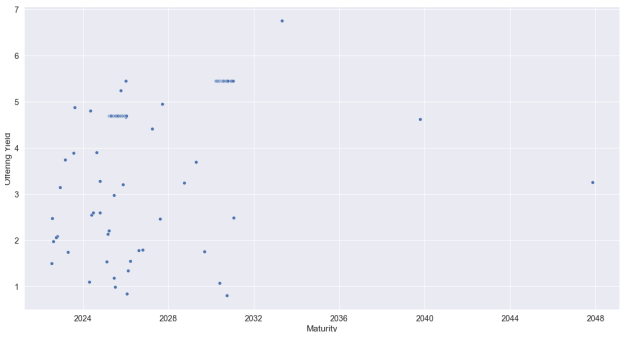
As mentioned in Section 3.1.2, the data on green bonds is scarce and our sample is of limited size. Nevertheless we proceeded with the visualization of our final green bonds database in order to observe the relationship between the different features and the target variable.



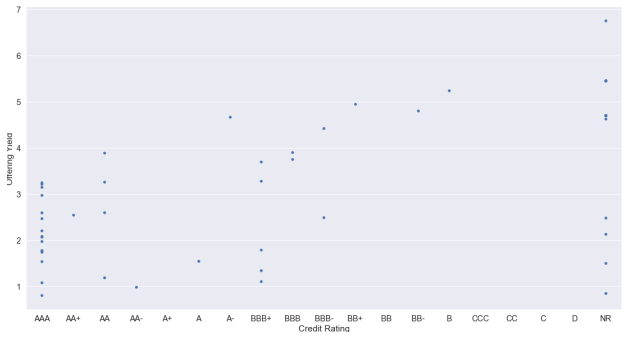
(a) Relationship between the offering yield and the coupon.



(b) Relationship between the offering yield and the offering price.



(c) Relationship between the offering yield and the maturity.



(d) Relationship between the offering yield and the credit rating.

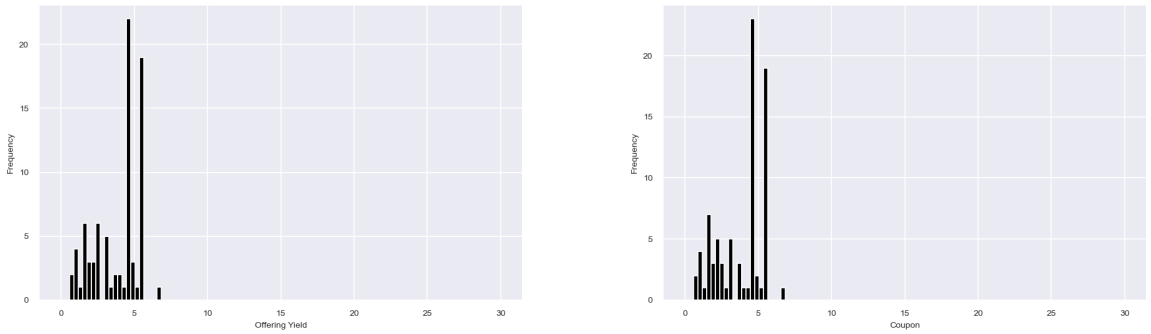
Figure 3.2.3: Visualization of the green bonds relationship between the offering yield and a) the coupon, b) the offering price, c) the maturity date, and d) the credit rating.

Figure 3.2.3a shows a linear relationship between the coupon and the offering yield, which is aligned with our observation in Figure 3.2.1a.

Regarding the offering price, we notice that our green bonds database is limited in terms of offering price variability, as this variable ranges between 99 and 100 as shown in Figure 3.2.3b. This indicates that all the green bonds in our dataset are offered at par, which is often the case when dealing with primary market transactions.

In terms of maturity date, we consider that our green bonds database represents a fair distribution within a wide data range, varying between 2024 and 2030, as depicted in Figure 3.2.3c.

Finally, Figure 3.2.3d depicts the distribution of the issuer's credit rating as determined by Standard & Poor's with regards to the offering yield. We also notice that our selection contains bonds with a well distributed credit ratings, as observed for brown bonds in Figure 3.2.1d.



(a) Histogram of the offering yield.

(b) Histogram of the coupon.

Figure 3.2.4: Visualization of the histograms of the green bonds a) offering yield, and b) coupon.

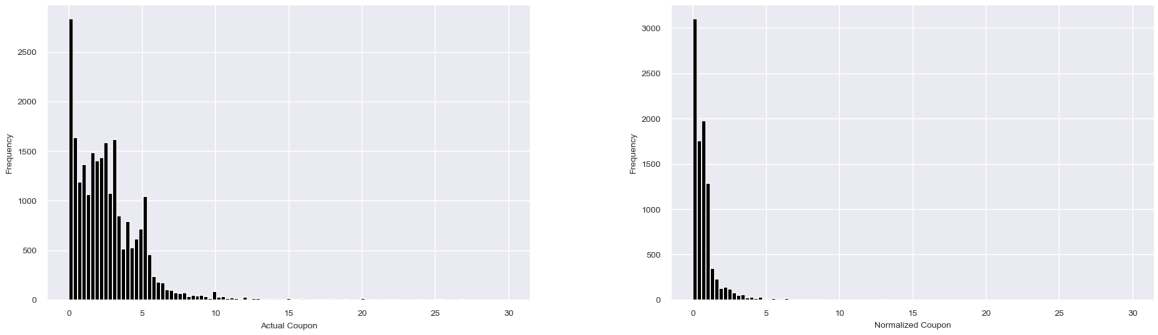
Figure 3.2.4 displays the histograms of the offering yield and coupon in the green bonds database, and shows that unfortunately we do not have well distributed data, due to the scarceness of green bonds data, as mentioned in Section 3.1.2.

3.3 Data Transformation

After cleaning, data transformation consists of performing changes in structure, value or format in order to enhance the performance of the ML algorithms. A first step is selecting the features that are the best predictors of the model. In our framework, we include only the features that have a major impact on the bond yield. Hence, the final database is limited to the following features: coupon, principal amount, offering amount, credit rating, interest rate, and the time to maturity, as well as the target variable, namely the offering yield.

The data transformation process consists of two main procedures: i) data encoding, and ii) data normalization. Our selected features are numerical variables, except for one categorical variable, namely the credit rating. We use the *Label Encoder* embedded in the scikit learn package in Python to transform the credit rating from categorical to numerical values.

We then normalize all the features included in the database by subtracting the mean and then scaling to unit variance using the *StandardScaler*, which is also included in the scikit learn package in Python, with the objective of having a homogeneously scaled database. To illustrate, Figure 3.3.1 shows the impact of the normalization on the coupon variable.



(a) Histogram of the actual coupon.

(b) Histogram of the normalized coupon.

Figure 3.3.1: Visualization of the impact of the normalization process using the histograms of the brown bonds a) actual coupon, and b) normalized coupon.

It is important to note that the target variable, namely the offering yield, is not included in the normalization process. Actually, this variable will be used later in Section 4.3 to estimate the greenium using the difference between the actual offering yield and model-predicted offering yield. Hence, it is important to keep the target variable at its actual scale to get the appropriate estimation of the greenium.

Chapter 4

Results and Analysis

After preparing both the brown and green bonds data, we are now interested in implementing the chosen ML models, namely the linear regression, decision tree, support vector machine (SVM), partial least squares regression (PLS), and multivariate adaptive regression spline (MARS), and assessing their performance on the brown bonds dataset. Then, we apply the pre-trained models on the green bonds data, and we conclude the chapter by inquiring into sequentially the brown and green bonds residuals, in order to provide an estimation of the greenium.

4.1 ML Models Implementation

After preparing the data, we can start the implementation of the ML models presented in Section 2.4. It is important to note that these models have been selected based on some of their advantages that are aligned with the framework of this project.

The linear regression algorithm is an easy to implement and easy to interpret model that performs exceptionally well for linearly separable data, which is clearly observed in the relationship between the coupon and the offering yield in the brown bonds data, as shown in Figure 3.2.1a.

On the other hand, in the decision tree family, we choose to use the random forest algorithm because it incorporates the bagging technique, as explained in Section 2.4.2, creating multiple trees and combining their results to get the final predictions. This model has two main advantages: i) it avoids the overfitting issue, reduces the variance and improves the accuracy; and ii) it is a very stable model that does not allow big deviations in its predictions when facing new observations, because these new data points might impact one tree but they will have minor influence on all the trees.

Concerning the SVR algorithm, it is known for its robustness and stability. Indeed, a minor change in the data does not have a considerable impact on the hyperplanes of the

SVR model, and hence on the predictions. The SVR also has an excellent generalization capability, which reduces the overfitting risk, and increases its accuracy.

While we did perform feature selection in our datasets, as described in Section 3.3, nevertheless, we investigate if an even more intense dimensionality reduction is appropriate, through the use of the PLS model. This algorithm is known for its good performance compared to other models for dealing with multicolinearity problems, thanks to its use of the covariance instead of the variance between dependent and independent variables. It is important to note that this technique may cause overfitting and needs to be applied carefully in order to avoid this issue.

Finally, we consider the use of the MARS algorithm because it has good bias-variance trade-off. In fact, the MARS model offers the flexibility needed to automatically model non-linearity in the variables, thus having fairly low bias, yet its constrained form of basis functions prevents too much flexibility, hence having fairly low variance. Also, it is known to work well for both large and small datasets, and it is easy to understand and interpret.

The general workflow for implementing the ML models follows the diagram shown in Figure 4.1.1.

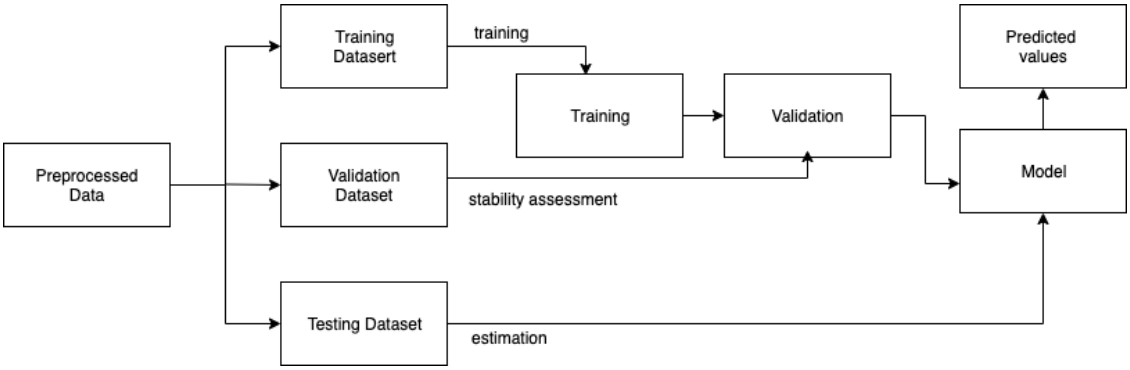


Figure 4.1.1: Diagram of the general workflow for implementing ML algorithms.

We start by dividing the preprocessed brown bonds database into 3 subsets: i) training set, ii) validation set and iii) testing set. For this matter, we use the *train_test_split* function located in the scikit learn package in Python. We devote 75% of the data to be used in the cross-validation process and the remaining 25% for the testing task. We also distinguish the independent variables used in each model, which consist of the five features described in Section 3.3, and the dependent variable, which is the offering yield.

4.2 Model Evaluation

We now describe the validation techniques and evaluation metrics used to evaluate the various ML algorithms applied in this thesis.

4.2.1 Cross-Validation

Cross-Validation (Berrar [2019]) is a statistical method for evaluating and comparing learning algorithms by splitting data into two segments: i) one used to train and learn the model, and ii) one used to validate and assess the performance of the model.

The ability to generalize a model to new data is ultimately what allows us to use ML algorithms to make the desired predictions. Cross-validation aims at overcoming the overfitting problem, which occurs when the model fits perfectly to its training data, but performs poorly on new, previously unseen data, which effectively defeat the learning purpose.

Cross-Validation consists of dividing the available data into two non-overlapping sections, one for training, and one for testing. In this thesis, we use the K-fold cross-validation approach. The K-fold validation approach consists of dividing the data into k equally, or nearly equally-sized segments, also known as *folds*. Then, k iterations of training and validation are performed, with each iteration holding out a different fold of the data for validation, while the remaining $k-1$ folds are used for training. This procedure avoids the dependence of the results on the training/test split chosen, which can result in overfitting to the training data and poor performance on new data.

Prior to splitting data into k folds, it is a common practise to stratify it. Stratification is the process of rearranging data so that each fold is an accurate representation of the whole data set.

The Leave-One-Out cross-validation (LOOCV) is a special case of K-fold cross-validation in which k is the number of instances in the data. In other words, nearly all of the data except for a single observation is used for training in each iteration, and the model is tested on that one single remaining observation. LOOCV is still widely used when available data is scarce, and it results in an accuracy that is known to be almost unbiased, but has a high variance, resulting in unreliable estimates.

4.2.2 Evaluation Metrics

Evaluation metrics are used to investigate the performance of the algorithms in the model evaluation phase, in order to determine the optimal model that gives the best results, and the most accurate predictions.

There exist multiple evaluation metrics to assess the goodness of a supervised ML algorithm.

In this thesis, we use the mean square error (MSE) and the coefficient of determination (R^2), defined precisely below.

1. Mean square error (MSE): this metric is used to quantify the difference between the predicted values obtained from the ML model and the observed values obtained from the data. It is computed as follows:

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}. \tag{4.2.1}$$

   where

   $y_1, \ldots, y_n$ are the observed values,

   $\hat{y}_1, \ldots, \hat{y}_n$ are the predicted values,

   n is the size of the dataset.

2. R Squared: this metric, also known as the coefficient of determination, represents the goodness-of-fit measure for regression algorithms. It expresses the proportion of variance in the outcome that the model can predict based on its features, and it ranges between 0 and 1. The model with best performance has the R Squared that is closest to 1. R Squared is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{4.2.2}$$

   where:

   $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average of the observed values.

4.2.3 Results

We apply the K-fold cross-validation method using $k = 10$ folds in order to assess the stability of the models and their ability to generate consistent results when facing new unseen observations. The results, for the five algorithms selected, applied to the brown bonds dataset, are summarized in Tables 4.2.1 (for the MSE metric) and 4.2.2 (for the R^2 metric). We refer the reader to Appendix B for the stability assessment of the models, the functions embedded in the Python package scikit learn and parameters used to achieve the same results.

ML Model	Train MSE	Validation MSE	Test MSE
Linear regression	0.2937	0.2963	0.1636
Random forest	0.3072	0.3367	0.2617
SVR	0.4912	0.5512	0.9881
PLS	0.2967	0.3953	0.1649
MARS	0.2933	0.2945	0.1632

Table 4.2.1: Performance of the ML models: Train, Validation and Test MSE for each model.

Table 4.2.1 shows that all models perform exceptionally well for all three datasets of train, validation and test in terms of MSE. The low MSE values for all models indicate that the deviation between the actual bonds yields and their predicted values is small[1]. Hence, we can rely on the ML models to predict the yield of a brown bond once given the features described in Section 3.3. Furthermore, the low MSE values reveal that the models have a good "Bias-Variance" trade-off, which means that the bias error resulting from erroneous assumptions in the learning algorithm and the variance error resulting from sensitivity to small fluctuations in the training set are reasonably balanced. Also, we observe that we have the same MSE scaling for the three data subsets, which means that the models are able to avoid the underfitting and overfitting issues.

ML Model	Train R^2	Validation R^2	Test R^2
Linear regression	0.9596	0.9611	0.9810
Random forest	0.9578	0.9555	0.9697
SVR	0.9487	0.9445	0.9042
PLS	0.9592	0.9475	0.9809
MARS	0.9597	0.9614	0.9810

Table 4.2.2: Performance of the ML models: Train, Validation and Test R^2 for each model.

Table 4.2.2 reports empirical results similar to the ones observed in Table 4.2.1. Indeed, we observe that all models have a value of R^2 close to 1, for the three data subsets, namely train, validation and test, indicating a high "goodness of fit" for the models. Moreover, we see that the models generate R^2 values of similar range for the train, validation and test sets indicating that we can successfully overcome the problems of underfitting and overfitting.

Overall, we conclude that all five models show very positive empirical results in terms of prediction error, bias-variance trade-off, as well as goodness of fit. In terms of models comparison, we observe that the MARS model generates the best results in terms of MSE and

[1]It is important to note that the MSE values displayed in Table 4.2.1 use the YTM in percentage, which entails that the decimal MSE values are actually divided by 10,000. For example, the Test MSE for the linear regression model in decimals is 0.000016.

R^2, with a test MSE of 0.1639 and R^2 of 0.9810. Notice that the linear regression and PLS models show similar performance that differ by slight differences in terms of MSE and R^2 evaluation metrics. The Random Forest and SVR models have the lowest performance in terms of test MSE and R^2 with their test R^2 being respectively 0.9697 and 0.9042.

Based on the above results, we select the linear regression, PLS and MARS trained models for the next steps of our project, namely the analysis on the green bonds database and the greenium estimation.

4.3 Greenium Estimation

At this point, we trained the ML models to predict any brown bond yield once given its features, namely the coupon, the principal amount, the offering amount, the credit rating, the term structure of interest rates, and the time to maturity.

As explained in Section 2.2.2.4, the market price of a bond may differ from its fair or theoretical price, which depends on the coupon, face value, and discount rate, because of various premia that are not accounted for in Equation 2.2.1. The observed yields in the brown bonds data set are computed from the actual prices on the primary market, and consequently do include these premia. Accordingly, contrary to what is often done in the greenium literature, where theoretical prices are used to match green bonds to their brown counterparts, our ML approach allows to account for premia that could be present in brown bonds prices.

More precisely, the offering amount may be related to the liquidity premium, the credit rating to the credit risk premium, and the term structure of interest rates to the inflation and interest rate premia. The high explanatory power of our ML models indicates that we are able to capture a large portion of the determinants of a brown bond price. This means that we are able to predict the yield (or, equivalently, the price) of a brown bond with the exact same characteristics as each of the green bonds in our green dataset.

In this section, we test our ML models on the green bonds database in order to obtain, for each green bond, the yield of its brown counterpart. Accordingly, the greenium is defined as follows:

$$\mathcal{G} = YTM_G - YTM_B, \tag{4.3.1}$$

where

$\mathcal{G}$: is the greenium,

YTM_G: is the offering yield-to-maturity of the green bond,

YTM_B: is the offering yield-to-maturity of its brown counterpart.

As the greenium is based on the difference between yields, we are interested in investigating the following residuals for respectively the test brown dataset and the green dataset:

$$\mathcal{E}_J = YTM_{Obs,J} - YTM_{Pred,J} \tag{4.3.2}$$

where

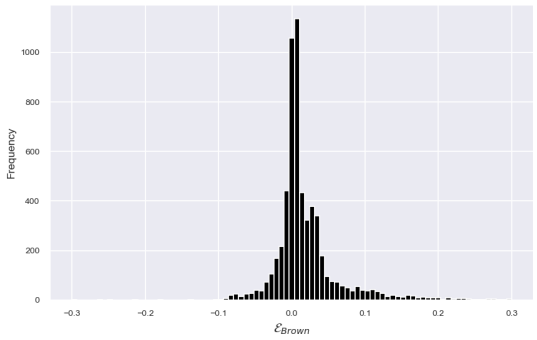
$J \in B, G$ represents either the brown or green bonds dataset,

$\mathcal{E}_J$: is the vector of residuals for dataset J,

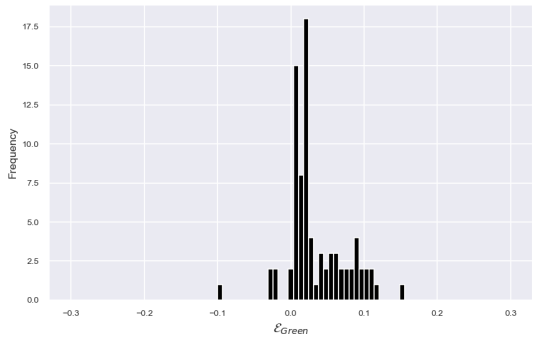
YTM_{Obs}: is the observed offering yield,

YTM_{Pred}: is the predicted offering yield obtained from the ML model.

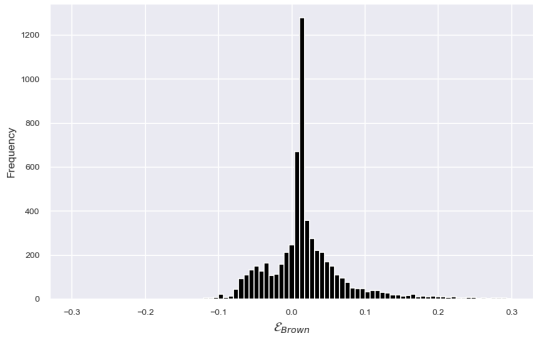
Figure 4.3.0 displays the distribution of residuals $\mathcal{E}_B$ and $\mathcal{E}_G$ for the linear regression, PLS and MARS models, respectively. A first observation is that, for all models, the brown bonds residuals are small, ranging from -0.15 to 0.23 (see the left panel of Figure 4.3.0). As expected, the brown residuals for all ML models are centered around zero, which entails that overall the models predictions are unbiased. As for the green bonds residuals, we notice their irregular shape and asymmetry, which may be caused by the paucity of the green bonds data (see the right panel of Figure 4.3.0). Nevertheless, we observe that the range of green residuals is comparable to that of the brown ones, with values between -0.15 and 0.19. Finally, the mode of the distribution seems to have shifted from zero (brown residuals) to a positive value (green residuals) for all the ML models.



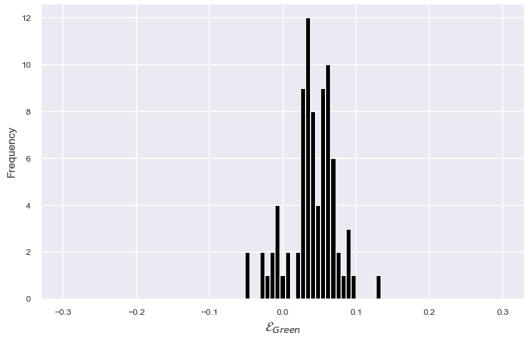
(a) Histogram of the brown bonds residuals for the Linear Regression model.



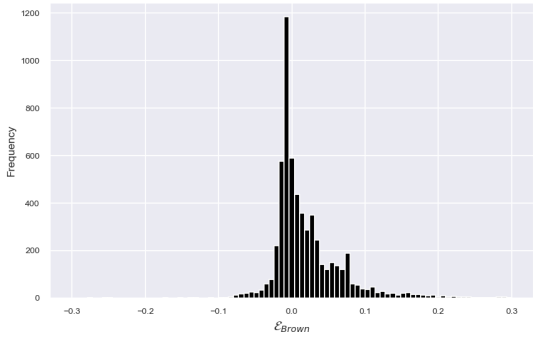
(b) Histogram of the green bonds residuals for the Linear Regression model.



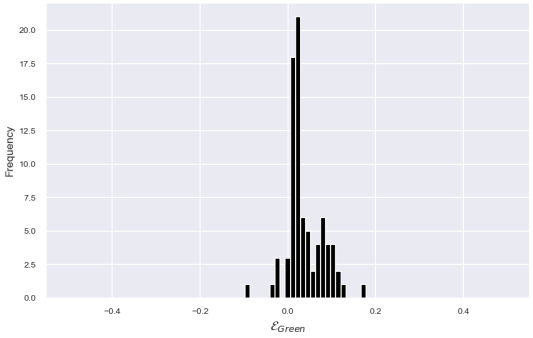
(c) Histogram of the brown bonds residuals for the PLS model.



(d) Histogram of the green bonds residuals for the PLS model.



(e) Histogram of the brown bonds residuals for the MARS model.



(f) Histogram of the green bonds residuals for the MARS model.

Figure 4.3.0: Visualization of the histograms of each ML model residuals obtained when applied on brown bonds (left), and green bonds (right).

Using Equations 4.3.1-4.3.2, the green residuals satisfy

$$\begin{aligned} \mathcal{E}_G &= YTM_{Obs,G} - YTM_{Pred,G} \\ &= YTM_G - YTM_{Pred,B} \\ &= YTM_G - [YTM_{Obs,B} - \mathcal{E}_B] \\ &= \mathcal{G} + \mathcal{E}_B. \end{aligned} \tag{4.3.3}$$

As such, we can use Equation 4.3.3 to estimate the greenium using the brown and green bonds residuals from each ML model, using

$$\mathcal{G} = \mathcal{E}_G - \mathcal{E}_B. \tag{4.3.4}$$

We compute for each dataset the average of the residuals obtained from Equation 4.3.2,

$$m_J = \mathbb{E}[\mathcal{E}_J] = \frac{\sum_{k=1}^{n} \mathcal{E}_{J,k}}{n}. \tag{4.3.5}$$

where

J is either B or G,

n: is the size of the dataset,

$\mathcal{E}_{J,k}$ is the k^{th} residual from dataset J.

According to the greenium computation displayed in Equation 4.3.4, as well as the mean of the brown and green bonds residuals provided in Equation 4.3.5, we obtain the following greenium estimation

$$\begin{aligned} \hat{\mathcal{G}} &= \mathbb{E}[\mathcal{G}] \\ &= \mathbb{E}[\mathcal{E}_G - \mathcal{E}_B] \\ &= m_G - m_B. \end{aligned} \tag{4.3.6}$$

Table 4.3.1 provides the results and the greenium estimation obtained from the three ML models.

ML Model	$\mathbf{m_B}$	$\mathbf{m_G}$	$\hat{\mathcal{G}}$ **(decimal)**	$\hat{\mathcal{G}}$ **(bps)**
Linear regression	0.000677	0.003445	0.002768	27.68
PLS	0.000596	0.004015	0.003419	34.19
MARS	0.000467	0.003904	0.003437	34.37

Table 4.3.1: Mean Residuals of the ML models for the brown and green bonds datasets, and the estimated greenium in respectively decimals and bps.

As illustrated in the left panel of Figure 4.3.0, the residuals (prediction errors) in the brown bonds dataset have a distribution with a mean close to 0 for all ML models, with m_B varying from 4.67 bps to 6.77 bps. The residuals from the green bonds dataset are positive on average, with m_G varying from 34.45 bps to 40.15 bps, an increase by a factor of ~ 10 compared to m_B for all ML models.

We estimate the greenium using Equation 4.3.6 for the three ML models in both a decimal and a bps scale. We obtain an estimate in the same range of around ~ 30 bps. We find that our proposed framework is stable and is independent of the chosen ML model, generating approximately the same estimation.

Finally, we run a statistical t-test test in order to determine if there is a significant difference between the means of the green and brown groups, which means that we are testing

the statistical significance of the obtained greenium estimate. Our test is based on the following hypotheses:

$$\begin{cases} H0 : \text{The two groups have identical expected values.} \\ H1 : \text{The two groups have different expected values.} \end{cases}$$

This system of hypotheses is equivalent to the following:

$$\begin{cases} H0 : \text{The difference is not statistically significant.} \\ H1 : \text{The difference is statistically significant.} \end{cases}$$

We use the *ttest_ind* function embedded in the stats package in python to run the t-test and we summarize the results obtained for the different ML models in Table 4.3.2.

ML Model	Linear regression	PLS	MARS
p-value	1.03×10^{-6}	7.08×10^{-10}	2.95×10^{-8}

Table 4.3.2: The p-value of the t-test for the different ML models.

We observe that all models have p-values < 0.05, which entails that we reject the null hypothesis $H0$ and that the detected difference between the mean of the green and brown bonds residuals, i.e. the greenium estimate, is statistically significant.

Conclusion

Summary

In this thesis, we developed a framework to detect and estimate the greenium embedded in green bonds yields. For this purpose, we used a ML-based approach to predict the yield of a plain-vanilla brown bond, given its key features. We then fed the features of a sample of green bonds to the trained ML models in order to derive the yield of each brown counterpart. Finally, we detected and estimated the greenium by computing the differences between the green bonds data residuals and the brown data residuals.

In the case of brown bonds data, we found that the selected ML models exhibited a good performance in terms of MSE and R^2, and they also generate stable cross-validation results when tested on new unseen observations. Furthermore, we observed small residual values centered around zero, for all ML models.

In the case of green bonds, the availability of data satisfying the applied filters is limited. Still, we were able to detect an obvious shift in the residuals' distribution, resulting in a positive average value for all ML models and we confirmed the statistical significance of this shift through running a t-test. We concluded on the existence of the greenium, and we obtained an estimate of its value, which is around 30 bps in the US market.

Possible Extensions

We now suggest some possible extensions of our work. Firstly, one can further investigate the models performance by understanding its relationship with bond's characteristics (e.g. maturity) and its stability over time. One can also try to explain why complex models do not perform better than linear regression.

Secondly, one could apply our proposed framework to datasets of bonds traded in markets other than the U.S., and investigate if there are any differences in the existence, sign and importance of greenium, and in investors' attitudes toward these instruments, depending on where the green bonds are exchanged.

Another important extension would be to include other bond categories, that may have embedded options, such as callable, putable, convertible bonds. These bonds are commonly excluded from most empirical investigations because of the difficulty in estimating their fair values. However, a machine learning approach may be able to predict the prices of exotic bonds without having to directly estimate the value of embedded options, by relating their observed prices to the contract specifications. This approach could be used to investigate the changes in greenium according to multiple categories of green bonds, but could also be applied to many other empirical studies where only plain-vanilla bonds have been used up to now.

Finally, our work could be extended by employing more advanced financial ML models. Combining many forms of green bonds data, such as data from different countries and different bond categories would most probably result in a massive amount of data, with a large number of features, which would require advanced, deep learning-based approaches.

Appendix A

Bond categories and terminology

Firstly, bonds can be classified into three main categories depending on the issuer type:

1. Government bonds: issued by governments, also referred to as sovereign debt, these are further categorized according to their maturities:

   - Bills: sovereign bonds with a maturity of one year or less,
   - Notes: sovereign bonds issued with 1 to 10 years to maturity,
   - Bonds: sovereign bonds issued with more than 10 years to maturity.

   The term *treasuries* is also used to refer to government bonds.

2. Municipal bonds: issued by municipalities, some of them offering tax-free coupon income.

3. Corporate bonds: issued by firms, often providing them more flexible terms and lower interest rates than bank loans.

Secondly, there are four main categories of bonds that differ according to the coupon payment type:

1. Zero-Coupon bonds: these bonds do not pay any coupon. Issued at a substantial discount to par value (see Definition 2.2.7), they will generate a return once the bondholder receives the full principal upon maturity.

2. Fixed-rate bonds: these bonds have constant coupon payments throughout their life until the maturity date.

3. Floating-rate bonds: these bonds have variable coupon payments that are linked to a certain reference interest rate agreed upon in the indenture, such as LIBOR, CPI, or CDOR. At each coupon date, the interest payment is determined according to the changes in the reference rate.

4. Inflation-indexed bonds: these bonds have principal amount and interest payments that are indexed to inflation, so that investors are protected from inflation risk.

Thirdly, bonds can be divided into four categories based on the type of maturity:

1. Term bonds: these have a fixed finite maturity date predetermined in the indenture, on which the issuer will have to pay the principal amount to the bondholder.

2. Perpetual bonds: also known as *perps* or *consols*, these bonds don't have a maturity date. Hence, bondholders will never get their principal back, however they will be paid forever – in perpetuity – interest payments at every coupon date.

3. Methuselah bonds: these bonds have a maturity of at least 50 years. They are named after Methuselah, the oldest person whose age is mentioned in the Hebrew Bible.

4. Serial bonds: these bonds have a set of maturity dates. They are structured such that they mature in steps, where a portion of the principal is paid at specific dates throughout the bond's life. Each maturation segment in the serial bond is issued concurrently, and the repayment schedule is mentioned in the indenture.

Finally, there are different attributes and specifications that may be contained in the indenture, and that lead additional types of bonds, such as the following:

1. Convertible bonds: these are securities that have features of normal bonds, yet they can be exchanged into an agreed upon number of common stocks or equity shares. The conversion happens at a predetermined conversion ratio that determines the number of shares obtained upon conversion. Typically, the conversion occurs at the discretion of the bondholder, at predefined possible times during the bond's life.

2. Exchangeable bonds: very similar to convertible bonds, these securities consist of a straight bond with an embedded option that allows the bondholder to exchange it for a stock of a company other than the issuer (usually a subsidiary or company in which the issuer owns a stake) at some predetermined dates and under prescribed conversion features.

3. Callable bonds: also known as redeemable bonds, these bonds have an embedded call option, providing the issuer with the right, but not the obligation, to redeem it before the maturity date. If interest rates fall, the issuer may decide to call the bond and re-borrow at a lower interest rate. Callable bonds compensate investors for this risk by typically providing a higher coupon rate.

4. Putable bonds: also known as retractable bonds, these bonds have an embedded put option, providing the holder with the right, but not the obligation, to sell the bond back to the issuer and demand an early repayment of the principal at a predetermined price, on predetermined dates.

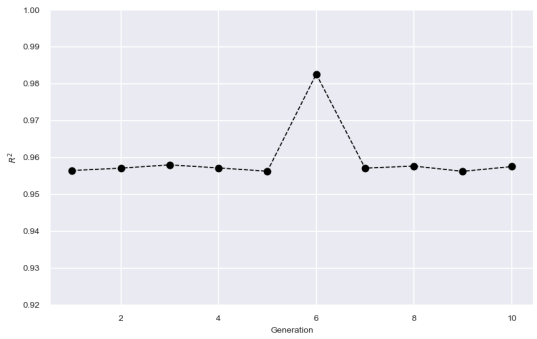
To conclude, note that there are no strict standards or specifications on bond issuance and the above features can be combined; for instance a company can issue a bond that has a fixed-rate coupon, no maturity, and two embedded options (call and put), giving rise to a corporate fixed-rate perpetual callable putable bond.

Appendix B

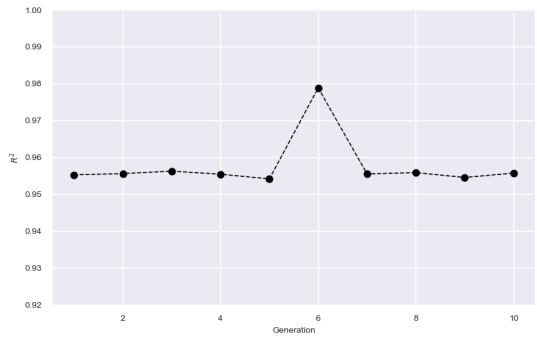
ML Models Implementation

B.1 ML Models Stability Assessment

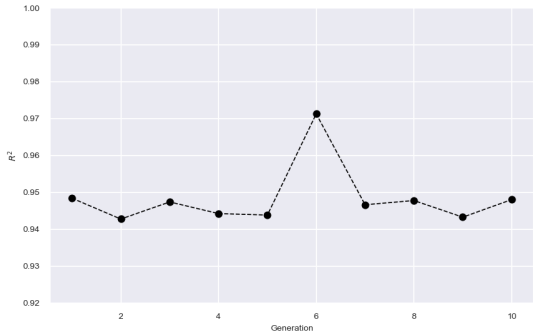
Figure B.1.0 displays the performance of the models, in terms of R^2, across the 10 validation datasets derived in the cross-validation step. We observe that overall all models keep a consistent performance as the R^2 remains roughly the same throughout the different datasets, indicating the stability of the models when facing new unseen observations. Furthermore, we note that the models maintain their great performance throughout the validation datasets, as the R^2 ranges approximately between 0.94 and 0.98 which is very close to 1.



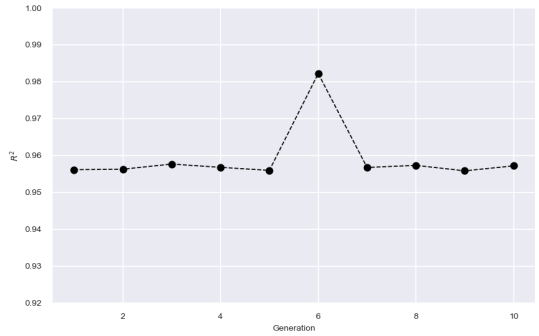
(a) Stability of the Linear Regression model.



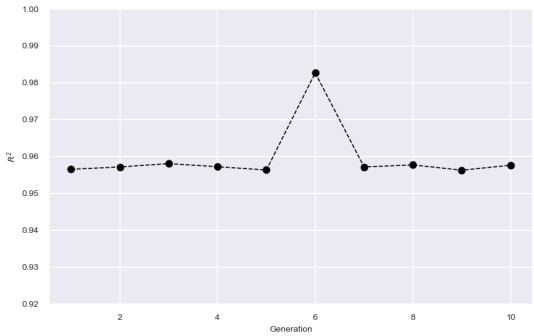
(b) Stability of the Random Forest model.



(c) Stability of the SVR model.



(d) Stability of the PLS model.



(e) Stability of the MARS model.

Figure B.1.0: Stability assessment of the ML models using K-Fold Cross-Validation method with $k = 10$.

B.2 Functions and Parameters of the ML Models

ML model	Function	Parameters
Linear regression	LinearRegression	-
Random forest	RandomForestRegressor	max_depth=4, random_state=50
SVR	SVR	kernel=rbf
PLS	PLSRegression	n_components=3
MARS	Earth	max_degree=2, penalty=1.0, minspan_alpha = 0.01, endspan_alpha = 0.01, endspan=5

Bibliography

I. C. M. Association et al. Green bond principles: voluntary process guidelines for issuing green bonds. *International Capital Market Association, Zürich*, 2018.

M. J. Bachelet, L. Becchetti, and S. Manfredonia. The green bonds premium puzzle: The role of issuer characteristics and third-party verification. *Sustainability*, 11(4):1098, 2019.

M. Baker, D. Bergstresser, G. Serafeim, and J. Wurgler. Financing the response to climate change: The pricing and ownership of us green bonds. Technical report, National Bureau of Economic Research, 2018.

T. G. Bali, A. Goyal, D. Huang, F. Jiang, and Q. Wen. Different strokes: Return predictability across stocks and bonds with machine learning and big data. *Georgetown McDonough School of Business Research Paper*, (3686164):20–110, 2020.

F. Barboza, H. Kimura, and E. Altman. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417, 2017.

W. A. Belson. Matching and prediction on the principle of biological classification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8(2):65–75, 1959.

D. Berrar. Cross-validation., 2019.

BNEF Bloomberg Terminal Guide. Guide to green bonds on the bloomberg terminal, 2015. URL https://data.bloomberglp.com/bnef/sites/4/2015/09/BNEF_Green-Bonds-Terminal-Guide_H2-2015-update.pdf.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

CFA Institute. Institute of international finance global debt monitor, 2020. URL https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/fixed-income-markets-issuance-trading-funding.

R. Culkin and S. R. Das. Machine learning in finance: the case of deep learning for option pricing. *Journal of Investment Management*, 15(4):92–100, 2017.

J. De Spiegeleer, D. B. Madan, S. Reyners, and W. Schoutens. Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. *Quantitative Finance*, 18(10): 1635–1643, 2018.

L. E. Eberly. Multiple linear regression. *Topics in Biostatistics*, pages 165–187, 2007.

T. Ehlers and F. Packer. Green bond finance and certification. *BIS Quarterly Review September*, 2017.

S. Fatica, R. Panzica, and M. Rancan. The pricing of green bonds: are financial institutions special? *Journal of Financial Stability*, 54:100873, 2021.

Federal Reserve System. Selected interest rates (daily) - h.15, 2022. URL https://www.federalreserve.gov/releases/h15/.

J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, 19(1): 1–67, 1991.

H. Ghoddusi, G. G. Creamer, and N. Rafizadeh. Machine learning in energy economics and finance: A review. *Energy Economics*, 81:709–727, 2019.

P. Golbayani, I. Florescu, and R. Chatterjee. A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *The North American Journal of Economics and Finance*, 54:101251, 2020.

B. M. Henrique, V. A. Sobreiro, and H. Kimura. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124: 226–251, 2019.

J. Kapraun, C. Latino, C. Scheins, and C. Schlag. (in)-credibly green: which bonds trade at a green bond premium? In *Proceedings of Paris December 2019 Finance Meeting EUROFIDAI-ESSEC*, 2021.

A. Karpf and A. Mandel. Does it pay to be green? *Available at SSRN 2923484*, 2017.

J.-M. Kim and H. Jung. Predicting bid prices by using machine learning methods. *Applied Economics*, 51(19):2011–2018, 2019.

J.-M. Kim, D. H. Kim, and H. Jung. Applications of machine learning for corporate bond yield spread forecasting. *The North American Journal of Economics and Finance*, 58: 101540, 2021.

G. Kirczenow, M. Hashemi, A. Fathi, and M. Davison. Machine learning for yield curve feature extraction: Application to illiquid corporate bonds. *arXiv preprint arXiv:1812.01102*, 2018.

D. F. Larcker and E. M. Watts. Where's the greenium? *Rock Center for Corporate Governance at Stanford University Working Paper*, (239):19–14, 2019.

Y. Ma, R. Han, and W. Wang. Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, 165:113973, 2021.

S. Mishra and S. Padhy. An efficient portfolio construction model using stock price predicted by support vector regression. *The North American Journal of Economics and Finance*, 50 (C), 2019.

S. Mishra, S. Padhy, S. N. Mishra, and S. N. Misra. A novel lasso–tlbo–svr hybrid model for an efficient portfolio construction. *The North American Journal of Economics and Finance*, 55:101350, 2021.

M. Moscatelli, F. Parlapiano, S. Narizzano, and G. Viggiano. Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161:113567, 2020.

M. Nunes, E. Gerding, F. McGroarty, and M. Niranjan. A comparison of multitask and single task learning with artificial neural networks for yield curve forecasting. *Expert Systems with Applications*, 119:362–375, 2019.

E. Ostertagová. Modelling using polynomial regression. *Procedia Engineering*, 48:500–506, 2012.

K. C. Rasekhschaffe and R. C. Jones. Machine learning for stock selection. *Financial Analysts Journal*, 75(3):70–88, 2019.

P. Reed, T. Cort, and L. Yonavjak. Data-driven green bond ratings as a market catalyst. *The Journal of Investing*, 28(2):66–76, 2019.

L. Ryll and S. Seidens. Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey. *arXiv preprint arXiv:1906.07786*, 2019.

V. N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

H. Wold. Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420, 1966.

O. D. Zerbib. The effect of pro-environmental preferences on bond prices: Evidence from green bonds. *Journal of Banking & Finance*, 98:39–60, 2019.

K. H. Zou, K. Tuncali, and S. G. Silverman. Correlation and simple linear regression. *Radiology*, 227(3):617–628, 2003.