Computers and Artificial Intelligence; Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.177192
Published by: Zeus Press

Machine Learning in Green Bond Pricing

Haokai Huang*
College of Finance and Statistics, Hunan University, Changsha 410013, China
**Corresponding author: Haokai Huang.*

Abstract

The pricing of green bonds has emerged as a critical issue of international scholarly and policy interest amid the rapid expansion of the global green bond market. Existing empirical research, however, remains constrained by limited sample coverage and superficial model comparisons, leaving substantial gaps in the understanding of the key determinants of green bond yields. This study examines a comprehensive sample of 7,062 Chinese green bonds issued between 2016 and 2024, employing three machine learning algorithms—Random Forest (RF), XGBoost, and Support Vector Machine (SVM)—to predict green bond coupon rates. The empirical results indicate that XGBoost achieves the strongest predictive performance, followed closely by Random Forest, whereas SVM exhibits comparatively weaker performance. Feature importance analysis reveals that macroeconomic factors—particularly year-on-year GDP growth rate, year-on-year M2 growth rate, and year-on-year CPI growth rate—constitute the most critical determinants of green bond pricing. Extensive robustness tests, encompassing alternative train-test split ratios and cross-validation settings, confirm that these findings remain stable across different experimental configurations. This study offers timely theoretical support and practical guidance for investors and regulatory authorities in green bond market decision-making.

Keywords

green bond pricing, machine learning, Random Forest, XGBoost

1. Introduction

The escalating urgency of global climate change has catalysed unprecedented policy attention toward green finance as a pivotal instrument for financing sustainable development and the low-carbon transition. Green bonds, as a cornerstone of the green finance architecture, have established themselves as essential financing vehicles for environmental protection, renewable energy, and climate adaptation projects [1]. According to data from the Climate Bonds Initiative, the global green bond market has experienced sustained exponential growth, and China has consolidated its position as one of the world's largest green bond issuance markets, accounting for a substantial share of global cumulative issuance [1]. In this context, the scientific and accurate pricing of green bonds has become a research topic of considerable theoretical and practical significance.

The academic literature on bond pricing has traditionally relied upon two foundational approaches: structural models, which link default to the issuer's asset value process [2], and reduced-form models, which treat default as an exogenous jump process governed by intensity-based hazard rates [3]. While these frameworks have provided invaluable insights into the determinants of credit spreads and yield dynamics, they encounter significant limitations when confronted with high-dimensional, nonlinear data structures characteristic of modern fixed-income markets [4]. Conventional linear econometric methods, such as ordinary least squares regression, often fail to adequately capture the complex interactions and nonlinear dependencies among bond characteristics, issuer attributes, and macroeconomic conditions that collectively shape green bond pricing [5].

The green bond literature has additionally grappled with the empirical question of whether green bonds command a pricing premium relative to conventional bonds—a phenomenon often termed the "greenium." Flammer [6] documents that corporate green bonds are issued at a slight premium, reflecting investor preferences for environmental sustainability. Wang and Liu [7] further demonstrate that green bond credit spreads are influenced by a heterogeneous set of factors spanning bond-specific characteristics, issuer creditworthiness, and external certification. Despite these advances, the existing body of knowledge exhibits notable limitations in sample representativeness and methodological sophistication. Prior studies have frequently employed relatively small or geographically narrow samples, and the comparative evaluation of alternative predictive methodologies remains underdeveloped [8,9].

In recent years, machine learning techniques have been increasingly deployed in financial research, capitalising on their capacity for nonlinear function approximation and high-dimensional pattern recognition [10]. Kocaarslan and Soytas [11] demonstrate that machine learning models, including ensemble methods, can deliver superior predictive accuracy in forecasting green bond market performance relative to traditional statistical approaches. Similarly, Ma and Zheng [12] apply machine learning algorithms to analyse the relationship between bond ratings and issuance costs, underscoring the methodological advantages of data-driven techniques in fixed-income pricing applications. Machine learning algorithms can autonomously learn complex, latent patterns in financial data, effectively capturing multidimensional determinants of bond pricing that may elude conventional econometric specifications [4,13]. Nevertheless, the application of machine learning to Chinese green bond pricing remains in its nascent stages, with existing studies offering limited model comparisons and insufficient interrogation of the relative importance of micro-level versus macro-level pricing determinants [14,15].

This paper seeks to address the following research gaps in the extant literature. First, existing studies on Chinese green bond pricing have typically relied upon relatively modest sample sizes and narrow variable sets, potentially compromising the external validity of their findings. Second, the comparative performance of alternative machine learning methodologies—particularly ensemble methods versus kernel-based approaches—has not been systematically evaluated in the context of Chinese green bond markets. Third, the relative contributions of bond-specific, issuer-specific, financial, and macroeconomic factors to green bond pricing remain insufficiently disentangled.

The principal contributions of this paper are fourfold:

(1) This study assembles a large-scale sample of 7,062 Chinese green bonds spanning the period from 2016 to 2024, encompassing four comprehensive dimensions: bond characteristics, issuer characteristics, financial indicators, and macroeconomic variables. This extensive data architecture provides a substantially more comprehensive empirical foundation than previous studies in this domain.

(2) This paper conducts a rigorous, systematic comparison of the predictive performance of three mainstream machine learning models—Random Forest, XGBoost, and SVM—for green bond coupon rate prediction. By evaluating these models across multiple performance metrics, including R^2, RMSE, and MAE, this study offers a more complete and nuanced model evaluation framework for the field.

(3) Through feature importance analysis derived from tree-based ensemble models, this study reveals that macroeconomic factors constitute the core determinants of green bond pricing in the Chinese market. This finding provides novel empirical evidence on the primacy of systematic risk factors over idiosyncratic characteristics in shaping green bond yields.

(4) This paper undertakes an extensive battery of robustness tests, including alternative train-test split ratios (80:20, 70:30, 60:40, and 50:50), k-fold cross-validation, feature exclusion tests, and non-macro feature-only specifications, collectively confirming the reliability, stability, and generalisability of the research conclusions.

2. Research Methodology

2.1 Data Source and Description

The research sample comprises 7,062 green bonds issued in the Chinese onshore green bond market over the period from January 2016 to December 2024. This time window is deliberately chosen to span the complete development trajectory of China's green bond market, which was formally inaugurated with the issuance of dedicated green bond guidelines by the People's Bank of China and the National Development and Reform Commission in 2015–2016. The sample period thus captures the market's nascent phase, its rapid expansion through 2020–2021, and its subsequent maturation, ensuring that the empirical analysis reflects the full heterogeneity of issuance conditions across different macroeconomic environments.

Primary bond-level data are sourced from the Wind Financial Terminal, the most widely used financial data provider in Chinese capital markets. Macroeconomic variables, including year-on-year GDP growth rate, year-on-year CPI growth rate, and year-on-year M2 money supply growth rate, are obtained from the People's Bank of China and the National Bureau of Statistics. The sample encompasses a diverse range of bond types, including medium-term notes and financial bonds, corporate bonds, enterprise bonds, and local government bonds, thereby ensuring broad representativeness across the Chinese green bond market's structural composition.

The raw sample is subjected to a systematic data cleaning protocol. Observations with missing values for key variables—specifically, bond coupon rate, bond maturity, entity rating, bond rating, and issuer financial indicators—are excluded from the analysis. Following the application of these selection criteria, 4,650 valid observations are retained for model estimation. The substantial reduction from the initial 7,062 bonds to the final estimation sample reflects the prevalence of incomplete financial disclosures among certain issuer categories, particularly non-listed enterprises and private placement issuers. All continuous variables are winsorised at the 1st and 99th percentiles to mitigate the influence of extreme outliers on model estimation.

2.2 Variable Selection and Description

This paper selects explanatory variables from four theoretically grounded dimensions: bond characteristics, issuer characteristics, financial indicators, and macroeconomic factors. The dependent variable is the bond coupon rate at issuance, expressed as a percentage. This measure captures the all-in cost of debt at the time of issuance and serves as the primary pricing metric of interest. Table 1 provides detailed definitions and descriptive statistics for each variable.

Table 1: Variable Definitions and Descriptive Statistics

Variable	Definition	Mean	Std. Dev.	Min	Max
Coupon Rate	Bond coupon rate at issuance (%)	2.12	0.49	0.91	3.83
Bond Type	1=Local Govt, 2=Corp, 3=MTN/Fin, 4=Ent.	2.82	0.67	1.00	4.00
Bond Issue Scale	Ln of total issuance amount	1.52	1.59	-6.91	6.55
Bond Maturity	Years from issuance to maturity	5.88	6.23	0.03	30.02
Entity Rating	Numerical rating (higher=better)	3.37	1.89	1.00	6.00
Bond Rating	Numerical rating (higher=better)	3.58	1.82	1.00	6.00
Company Type	1=SOE, 2=Private, 3=Foreign	1.91	0.70	1.00	3.00
Listed Status	1=Listed, 0=Non-listed	0.14	0.35	0.00	1.00
Third-party Cert.	1=Certified, 0=Non-certified	0.34	0.47	0.00	1.00
Asset-Liab. Ratio	Total liabilities / Total assets (%)	46.16	35.53	0.00	154.42
Asset Scale	Ln of total assets	24.86	2.07	17.71	31.52
Return on Assets	Net profit / Total assets (%)	3.51	7.14	-97.91	60.84
Gross Profit Margin	(Revenue-COGS)/Revenue (%)	17.45	22.37	-212.61	99.99
Net Profit Margin	Net profit / Revenue (%)	7.53	18.37	-268.63	99.99
Return on Equity	Net profit / Equity (%)	3.63	12.13	-411.08	245.44
Current Ratio	Current assets / Current liab.	2.48	3.56	0.00	73.57
Int. Coverage Ratio	EBIT / Interest expense	4.22	20.25	-987.36	494.43
Altman Z-score	Financial distress indicator	1.67	2.73	-19.98	34.55
GDP Growth (YoY)	Year-on-year GDP growth (%)	4.00	4.54	-6.80	18.90
CPI Growth (YoY)	Year-on-year CPI growth (%)	1.70	1.54	-0.80	5.40

Variable	Definition	Mean	Std. Dev.	Min	Max
M2 Growth (YoY)	Year-on-year M2 growth (%)	9.49	1.89	0.00	19.20
Public Offering	1=Public, 0=Private	0.61	0.49	0.00	1.00
Embedded Option	1=With option, 0=Without	0.27	0.45	0.00	1.00
Govt. Subsidy	1=With subsidy, 0=Without	0.10	0.30	0.00	1.00

The selection of variables is guided by both theoretical considerations from the fixed-income pricing literature and empirical findings from prior green bond studies. Bond characteristics—including bond type, issue scale, maturity, entity rating, and bond rating—capture the structural and contractual features that directly influence yield determination. Issuer characteristics, such as company type, listed status, and third-party certification, control for ownership structure, information transparency, and the signalling value of external green verification [9]. Financial indicators, encompassing profitability, leverage, liquidity, and solvency measures, proxy for the issuer's fundamental creditworthiness and default risk [13]. Macroeconomic factors—GDP growth, CPI inflation, and M2 money supply growth—capture the systematic risk environment within which green bonds are priced, reflecting the opportunity cost of capital, inflation expectations, and monetary policy conditions [5,15].

2.3 Methodology

This paper employs three machine learning models—Random Forest, XGBoost, and Support Vector Machine—for green bond coupon rate prediction. The dependent variable in all specifications is the coupon rate at issuance. Model performance is evaluated using three complementary metrics: the coefficient of determination (R^2), which measures the proportion of variance explained; Root Mean Square Error (RMSE), which penalises larger prediction errors; and Mean Absolute Error (MAE), which provides an interpretable measure of average prediction deviation.

Prior to model estimation, the complete dataset is partitioned into training and test sets using an 80:20 split ratio, stratified by bond type to preserve the distributional characteristics of the original sample. All models are trained exclusively on the training set, with hyperparameter optimisation performed via grid search combined with 5-fold cross-validation. The optimal hyperparameters are subsequently applied to the held-out test set to generate out-of-sample predictions and compute the final evaluation metrics. This procedure ensures that the reported performance statistics reflect genuine out-of-sample predictive accuracy rather than in-sample overfitting.

(1) Random Forest

Random Forest (RF) is an ensemble learning algorithm founded upon the bootstrap aggregating (Bagging) framework proposed by Breiman [16]. It constructs an ensemble of B de-correlated decision trees through Bootstrap sampling—random sampling with replacement from the training data—and aggregates individual tree predictions via arithmetic averaging. A critical mechanism controlling overfitting is feature subsampling: at each split point in each tree, only a random subset of m features is considered when searching for the optimal partition; the subset size m is treated as a tunable parameter relative to the total number of features p, thereby reducing correlation among constituent trees and improving generalisation performance. For a given input vector x, the Random Forest prediction is computed as:

$$\hat{y}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x) \tag{1}$$

where B is the number of trees and $T_b(x)$ is the prediction of the b-th decision tree. Random Forest additionally provides an internal estimate of prediction error through out-of-bag (OOB) observations—approximately one-third of training samples not selected in each bootstrap replicate—which can be used for model validation without requiring a separate hold-out set. The combination of bootstrap aggregation, random feature selection at each split, and ensemble averaging makes RF particularly well-suited for processing high-dimensional financial data with complex nonlinear interactions and heterogeneous variable importance.

(2) XGBoost

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm based on the gradient boosting framework that has demonstrated state-of-the-art performance across a wide range of machine learning applications [17]. Unlike Random Forest, which trains trees in parallel, XGBoost sequentially adds new trees to fit the negative gradients (residuals) of the loss function from previous iterations. A key algorithmic innovation is the use of a second-order Taylor expansion of the loss function to approximate the optimisation objective, enabling faster convergence and more accurate tree structure learning. The objective function comprises a differentiable loss term and a regularisation term that penalises model complexity:

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{2}$$

where $l(y_i, \hat{y}_i)$ is the loss function and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ is the regularisation term used to control model complexity, with T being the number of leaf nodes and w_j the leaf weights. The regularisation parameters γ (minimum loss reduction required for a further partition) and λ (L2 penalty on leaf weights) jointly constrain model complexity, mitigating overfitting. Additional regularisation mechanisms include shrinkage (learning rate $\eta < 1$), which scales the contribution of each new tree, and column (feature) subsampling, which introduces randomness analogous to Random Forest's feature selection. These design elements collectively render XGBoost highly effective for predictive modelling in financial applications where generalisation to unseen data is paramount.

(3) Support Vector Machine

Support Vector Machine (SVM) is a supervised learning method grounded in statistical learning theory and the principle of structural risk minimisation [18]. For regression tasks, the variant known as Support Vector Regression (SVR) seeks to identify a function $f(x) = w^{\top}\phi(x) + b$ that deviates from the observed target values by at most ε for all training samples, while simultaneously maintaining maximal flatness by minimising the Euclidean norm of the weight vector. The primal optimisation problem is formulated as:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{3}$$

subject to $y_i - w^{\top}\phi(x_i) - b \leq \varepsilon + \xi_i$, $w^{\top}\phi(x_i) + b - y_i \leq \varepsilon + \xi_i^*$, and $\xi_i, \xi_i^* \geq 0$, where $\phi(\cdot)$ is the kernel function mapping features to a high-dimensional space and C is the penalty parameter. The slack variables ξ_i and ξ_i^* accommodate violations of the ε-insensitive tube around the regression function, while the penalty parameter C controls the trade-off between empirical error and model complexity. In this study, the Radial Basis Function (RBF) kernel, defined as $K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2)$, is employed as the kernel function, as it can implicitly map input features into an infinite-dimensional space and capture complex nonlinear relationships without explicit specification of the mapping $\phi(\cdot)$. The hyperparameters C, γ, and ε are jointly optimised through grid search with cross-validation.

3. Results and Discussion

3.1 Descriptive Statistical Analysis

This paper presents a comprehensive descriptive statistical analysis of the sample comprising 7,062 green bonds. Figure 1 illustrates the distribution of bond types, revealing that medium-term notes and financial bonds collectively constitute the largest share at 71.8%, followed by corporate bonds at 18.1%, enterprise bonds at 6.3%, and local government bonds at 3.8%. This composition is indicative of the structural feature of the Chinese green bond market, wherein financial institutions and large corporate entities serve as the primary issuers, reflecting the market's reliance on established financial intermediaries to channel capital toward environmentally sustainable projects.

From the perspective of issuer listing status (Figure 2), non-listed firms account for 86.2% of issuers, whereas listed firms represent 13.8%. This pronounced dominance of non-listed issuers underscores the role of green bonds in providing financing access to entities that face constraints in raising equity capital through public markets. With respect to third-party certification (Figure 3), 34.3% of bonds in the sample carry independent third-party certification, while 65.7% do not. The relatively modest certification rate highlights the voluntary nature of green bond verification in the Chinese market during the sample period and raises questions regarding the credibility and standardisation of green labelling.

Figure 1: Bond Type Distribution

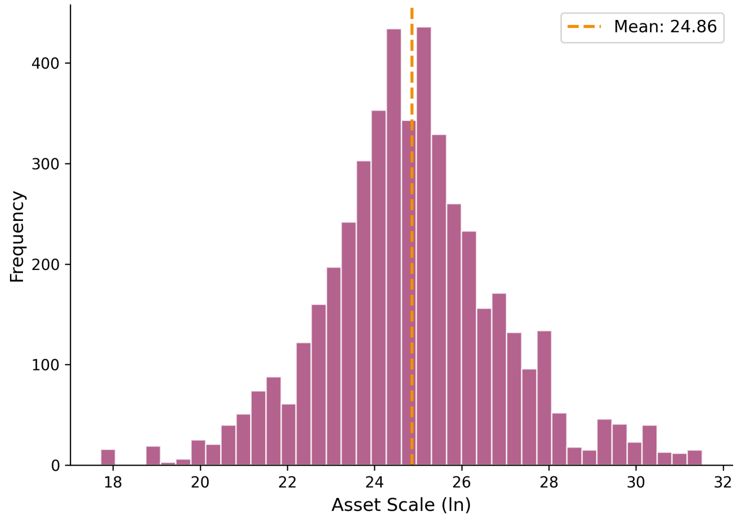


Figure 2: Listing Status Distribution

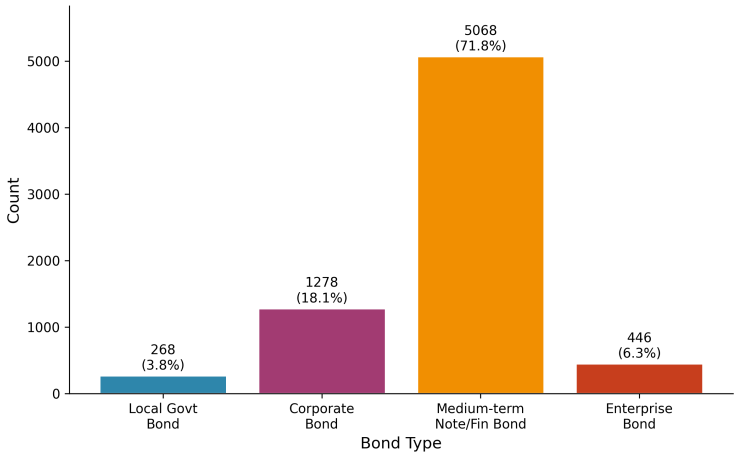
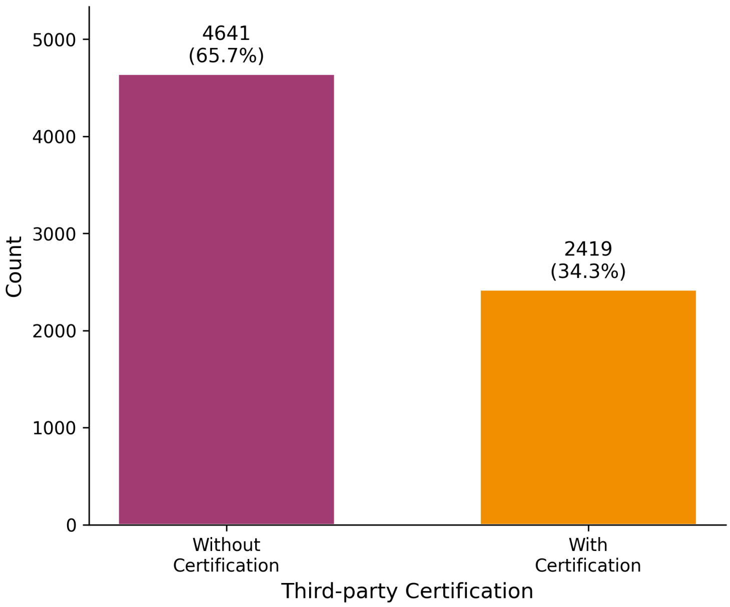


Figure 3: Third-party Certification Distribution



The distributional properties of key financial indicators are summarised in Figures 4 and 5. The mean debt-to-asset ratio of sample issuers stands at 46.16%, which is broadly consistent with the capital structure of firms in capital-intensive industries that typically dominate green bond issuance. The logarithm of total assets (Figure 5) exhibits an approximately normal distribution with a mean of 24.86, suggesting that the sample encompasses a representative cross-section of issuer sizes and mitigating concerns regarding size-related selection bias.

Figure 4: Asset-Liability Ratio Distribution

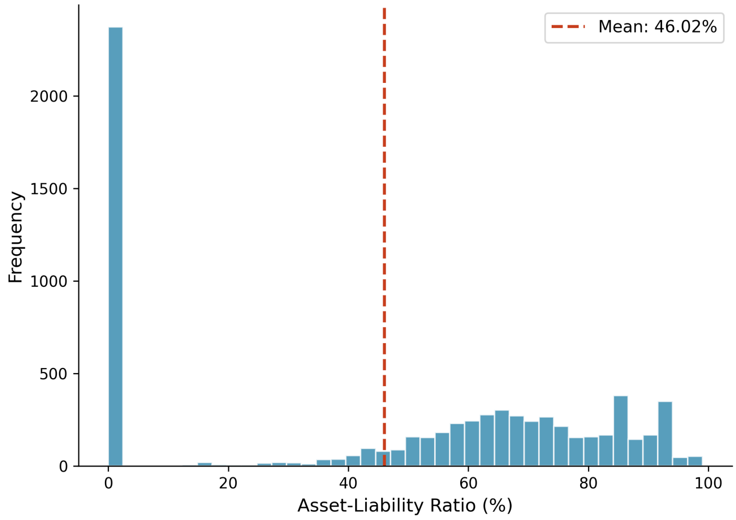
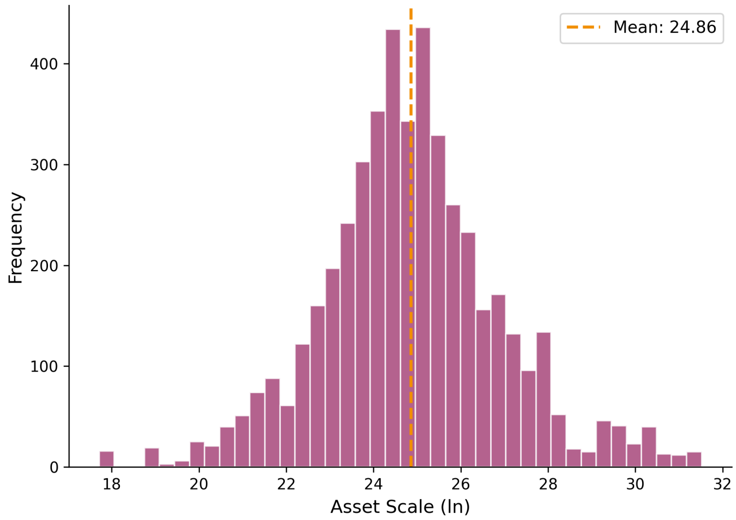


Figure 5: Asset Scale Distribution



The distribution of coupon rates (Figure 6) has a mean of 2.12%, while the logarithm of issuance scale (Figure 7) has a mean of 1.52. These statistics characterise the prevailing interest rate environment and issuance volume dynamics within the sample, capturing the period of relatively accommodative monetary policy conditions during much of the sample window.

The distributions of entity ratings and bond ratings are presented in Figures 8 and 9. Bonds rated at level 6 constitute the modal category, consistent with the well-documented feature that green bond issuers typically exhibit superior credit quality relative to conventional bond issuers. The presence of non-trivial proportions at rating levels 1 and 5 indicates meaningful heterogeneity in the credit structure of the market, which is conducive to examining the differential pricing implications of credit risk.

The macroeconomic indicators exhibit the following distributional properties. The year-on-year GDP growth rate has a mean of 4.00% (Figure 10), and the year-on-year CPI growth rate has a mean of 1.70% (Figure 11), capturing the trajectory of China's macroeconomic performance over the observation period, characterised by a gradual deceleration in headline growth alongside subdued inflationary pressures.

Figure 6: Coupon Rate Distribution

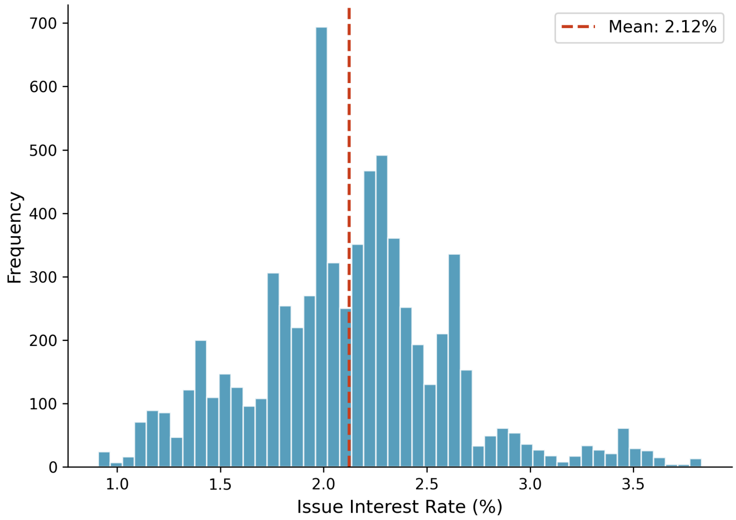


Figure 7: Bond Issue Scale Distribution

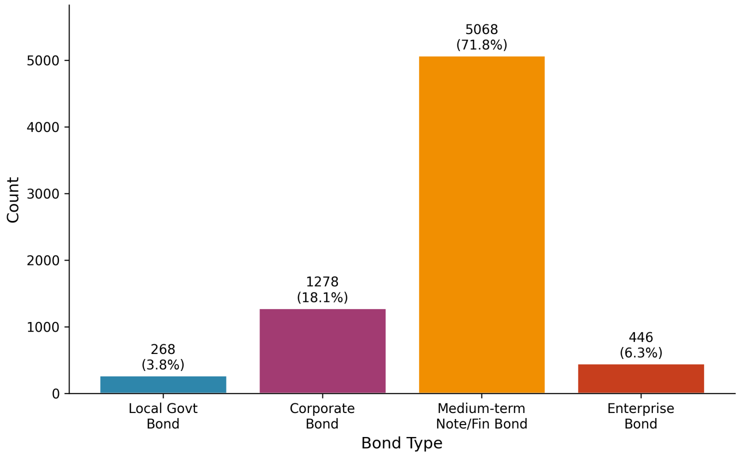


Figure 8: Entity Rating Distribution

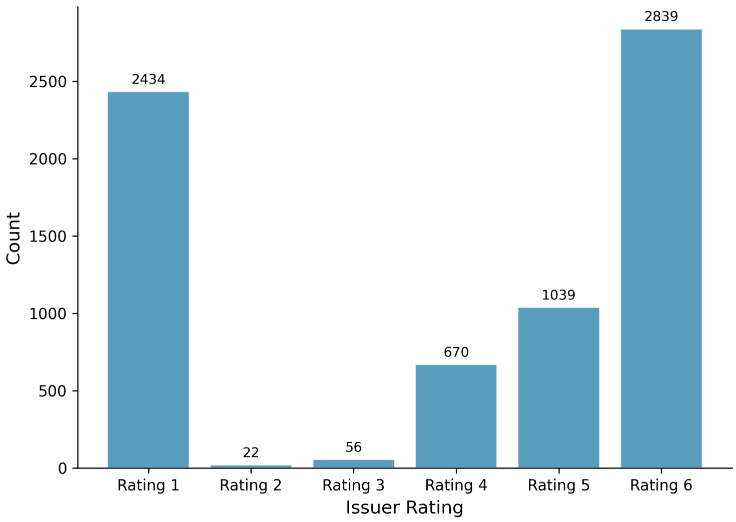


Figure 9: Bond Rating Distribution

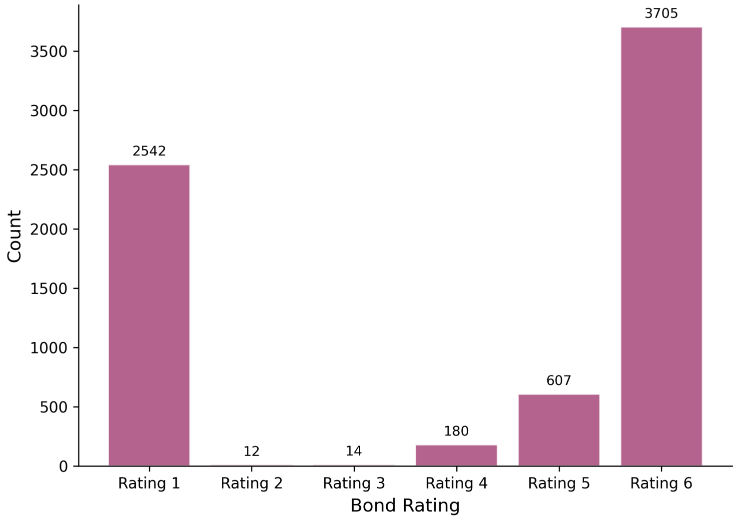


Figure 10: GDP Year-on-Year Growth Rate Distribution

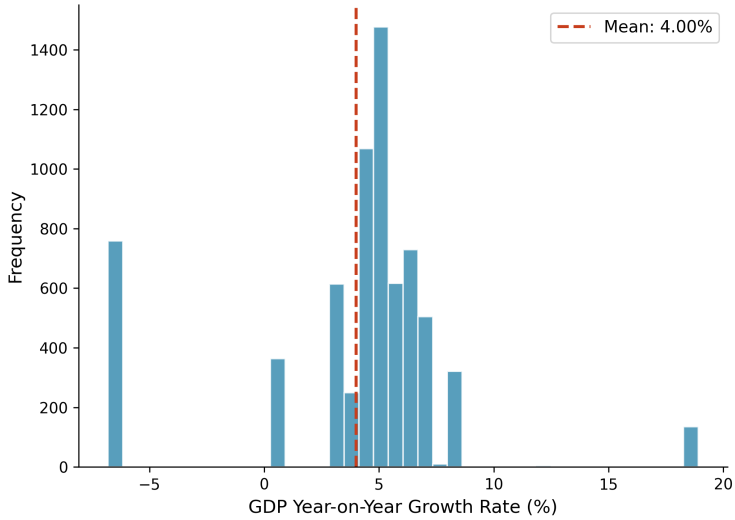
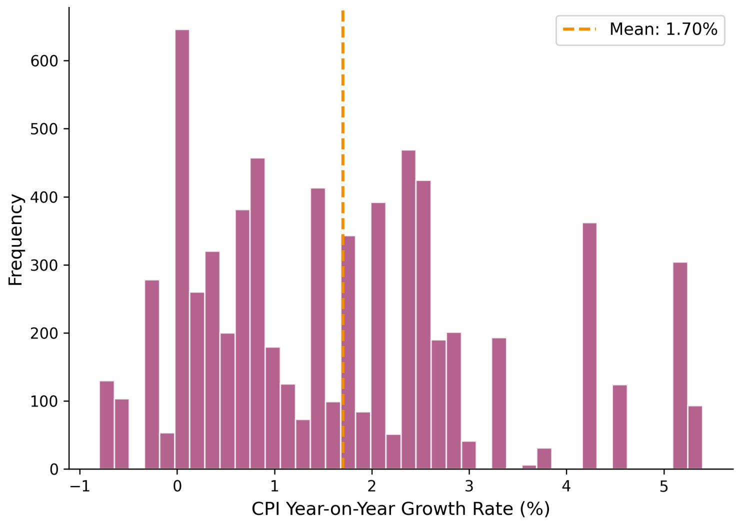


Figure 11: CPI Year-on-Year Growth Rate Distribution



The year-on-year M2 growth rate has a mean of 9.49% (Figure 12), reflecting the broadly accommodative stance of monetary policy throughout the sample period, which has supported liquidity conditions in the bond market. The distribution of government subsidies (Figure 13) indicates that 10.1% of bonds in the sample receive explicit fiscal subsidies, whereas 89.9% do not. This low incidence of direct subsidy suggests that government support operates primarily through indirect channels, such as monetary accommodation and regulatory incentives, rather than direct fiscal transfers.

Figure 12: M2 Year-on-Year Growth Rate Distribution

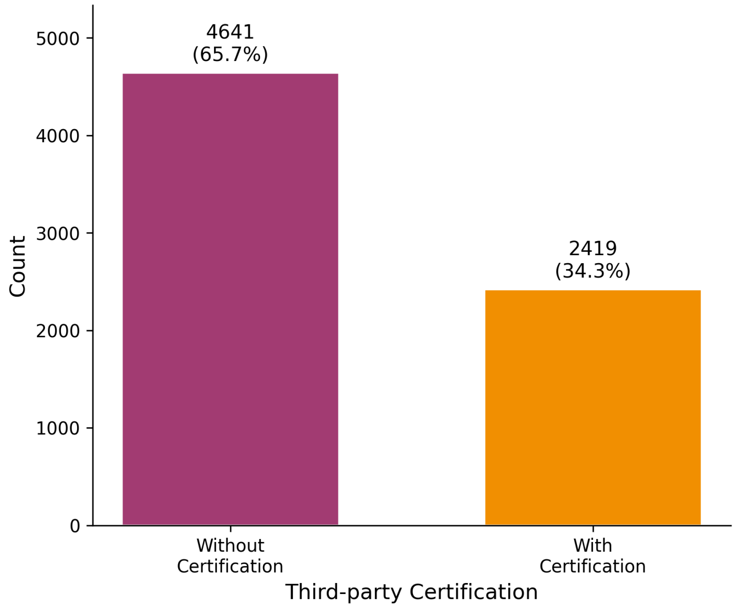
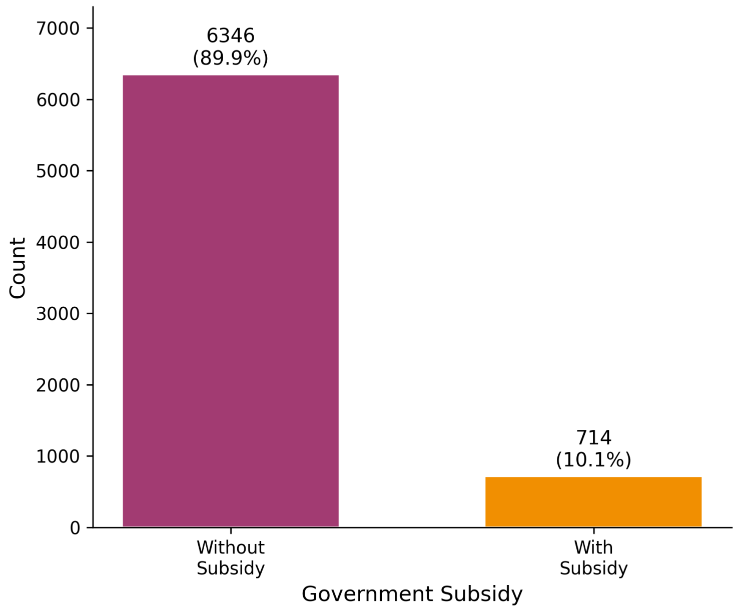


Figure 13: Government Subsidy Distribution



3.2 Machine Learning Model Results

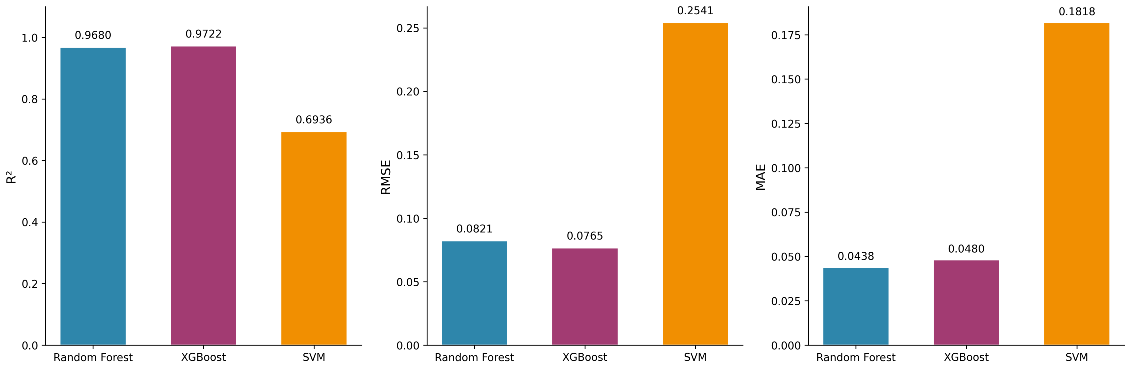
This paper employs Random Forest, XGBoost, and SVM to predict green bond coupon rates. The performance evaluation results are reported in Table 2. XGBoost achieves the strongest predictive performance, with an R^2 of 0.9722, an RMSE of 0.0765, and an MAE of 0.0480. Random Forest follows closely, achieving an R^2 of 0.9680, an RMSE of 0.0821, and an MAE of 0.0438. SVM exhibits markedly weaker performance, recording an R^2 of 0.6936, an RMSE of 0.2541, and an MAE of 0.1818. As illustrated in Figure 14, XGBoost and Random Forest substantially outperform SVM across both R^2 and RMSE metrics, indicating that ensemble tree-based methods possess considerable advantages in modelling the pricing dynamics of green bonds.

Table 2: Machine Learning Model Performance Comparison

Model	R^2	RMSE	MAE
Random Forest	0.9680	0.0821	0.0438
XGBoost	0.9722	0.0765	0.0480
SVM	0.6936	0.2541	0.1818

The pronounced performance differential between tree-based ensemble methods and SVM can be attributed to several factors rooted in model architecture and the nature of the data. First, Random Forest and XGBoost are inherently well-suited to tabular financial data characterised by heterogeneous feature scales, non-linear relationships, and potential interaction effects among predictor variables. The recursive partitioning mechanism of tree-based models captures complex non-linearities and high-order interactions without requiring explicit specification of functional form, which is particularly advantageous in bond pricing where the mapping from macroeconomic conditions to coupon rates involves non-linear discount rate channels. Second, the ensemble structure of Random Forest (bagging) and XGBoost (gradient boosting) mitigates overfitting through variance reduction (Random Forest) and systematic bias correction (XGBoost), yielding superior generalisation relative to single-model approaches. In contrast, although the SVM specification uses an RBF kernel, its predictive performance is more sensitive to feature scaling, kernel-parameter selection, and sample heterogeneity, which appears to limit its ability to capture the intricate relationship between macro-financial predictors and green bond coupon rates in this setting. Third, the high dimensionality of the feature space relative to the complexity of the underlying data-generating process favours models capable of implicit feature selection; both Random Forest and XGBoost naturally rank and utilise the most informative predictors, whereas SVM treats all features uniformly, rendering it more susceptible to the curse of dimensionality.

Figure 14: Model Performance Comparison



3.3 Feature Importance Analysis

As shown in Figure 15, feature importance analysis based on the Random Forest and XGBoost models reveals that macroeconomic factors exert a dominant influence on green bond coupon rate determination. The year-on-year GDP growth rate emerges as the single most influential predictor, accounting for 70.78% of total feature importance in the Random Forest model and 65.34% in the XGBoost model. This is followed by the year-on-year M2 growth rate (17.91% in Random Forest; 12.70% in XGBoost) and the year-on-year CPI growth rate (8.82% in Random Forest; 10.90% in XGBoost). Collectively, these three macroeconomic indicators account for over 95% of aggregate feature importance, whereas bond-specific characteristics, issuer attributes, and financial indicators each contribute less than 2%.

Figure 15 (a):Random Forest Feature Importance Analysis

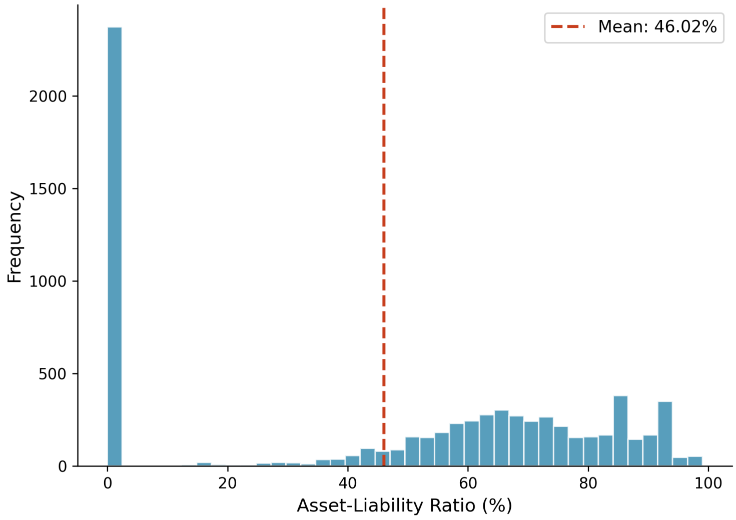
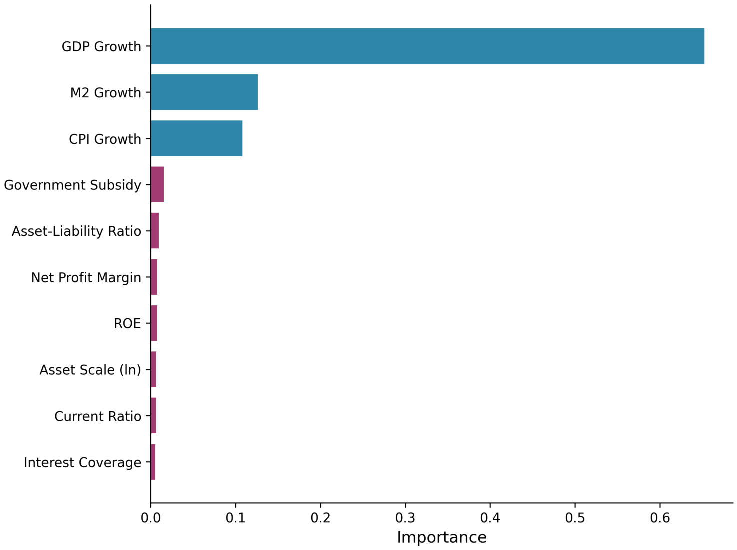


Figure 15(b): XGBoost Feature Importance Analysis



The overwhelming dominance of GDP growth in driving green bond pricing can be rationalised through multiple financial channels. From a theoretical standpoint, GDP growth serves as a broad proxy for the risk-free rate and the discount rate environment: higher GDP growth is typically associated with elevated inflationary expectations and tighter monetary policy, which translate into higher benchmark interest rates and, consequently, higher coupon rates on newly issued bonds. Furthermore, GDP growth captures aggregate demand conditions that influence the opportunity cost of capital, thereby affecting the pricing of fixed-income instruments through the term structure of interest rates. The substantial importance of M2 growth reflects the monetary policy transmission mechanism; the People's Bank of China's liquidity management directly affects interbank funding costs and benchmark yields, which propagate into corporate and financial bond pricing. The non-negligible contribution of CPI growth underscores the role of inflation expectations in nominal rate determination, consistent with the Fisher equation and the inflation-risk premium embedded in fixed-income securities.

The near-complete insignificance of micro-level factors—issuer characteristics, financial ratios, and bond-specific attributes—carries important implications for understanding the Chinese green bond market. This pattern may reflect several market-specific features. First, the predominance of state-owned enterprises and large financial institutions among green bond issuers (as evidenced by the high credit ratings documented in Section 3.1) implies a compressed cross-sectional dispersion in issuer credit quality, reducing the marginal information content of firm-level financial metrics for pricing purposes. Second, pervasive implicit government guarantees in the Chinese bond market may attenuate the sensitivity of coupon rates to individual issuer fundamentals, as market participants price sovereign backing rather than idiosyncratic credit risk [19]. Third, information asymmetries in the green bond market—particularly the absence of standardised environmental impact metrics and the voluntary nature of third-party certification—may render issuer-level disclosures less informative for pricing than macroeconomic signals that are uniformly observable to all market participants. These findings contrast with studies on conventional corporate bond markets in developed economies, where firm-specific creditworthiness indicators typically play a more prominent role in yield determination [6,11].

3.4 Robustness Tests

To ensure the reliability and generalisability of the principal findings, this paper conducts a battery of robustness tests encompassing four distinct validation strategies.

(1) Alternative train-test split ratios. The baseline 80:20 train-test split is systematically varied to 70:30, 60:40, and 50:50. The results, presented in Figure 16, demonstrate that Random Forest R^2 values remain stable in the range of 0.9650 to 0.9727, while XGBoost R^2 values range from 0.9627 to 0.9726. Although SVM performance remains comparatively modest, it likewise exhibits stability across partitioning schemes. The insensitivity of model performance to data splitting provides strong evidence that the reported results are not artefacts of sample-specific random variation and that the models generalise robustly across different training set sizes.

(2) K-fold cross-validation. Five-fold cross-validation is performed on the full sample to assess out-of-sample stability. The results indicate that Random Forest achieves a mean R^2 of 0.9734 with a standard deviation of 0.0030; XGBoost achieves a mean R^2 of 0.9700 with a standard deviation of 0.0067; and SVM achieves a mean R^2 of 0.7013 with a standard deviation of 0.0126 (Figure 16). The small standard deviations relative to the mean performance levels—coefficients of variation of 0.31%, 0.69%, and 1.80% for Random Forest, XGBoost, and SVM respectively—confirm that all three models exhibit high stability, with the ensemble methods demonstrating particularly low cross-fold variability. These results provide statistical assurance that the model rankings and performance magnitudes are reproducible.

(3) Exclusion of the dominant feature. Following the identification of year-on-year GDP growth rate as the most important predictor, this feature is removed from the feature set and the models are retrained. Random Forest and XGBoost achieve R^2 values of 0.9590 and 0.9638, respectively, representing only marginal declines of approximately 0.93 and 0.86 percentage points from their respective baseline levels. This finding demonstrates that the models do not exhibit excessive dependence on a single dominant predictor and possess robust generalisation capacity. The modest performance degradation is attributable to the presence of multicollinearity among macroeconomic variables: GDP growth, M2 growth, and CPI growth are economically interrelated, and the remaining macroeconomic features partially absorb the information content previously captured by GDP growth. This result is consistent with the theoretical expectation that macroeconomic indicators share common underlying drivers of aggregate economic activity.

(4) Non-macro feature-only models. When the models are restricted to bond characteristics, issuer characteristics, and financial indicators—excluding all macroeconomic variables—Random Forest and XGBoost R^2 values decline to 0.4310 and 0.4794, respectively. While this represents a substantial reduction in explanatory power, the retained R^2 values of approximately 43–48% indicate that micro-level factors collectively possess meaningful, albeit secondary, explanatory capacity for green bond pricing. This residual explanatory power likely derives from credit rating information and bond-specific structural features (e.g., issuance scale, maturity, and bond type) that capture idiosyncratic risk components orthogonal to macroeconomic variation. The stark contrast between the macro-only and non-macro-only specifications reinforces the principal conclusion that macroeconomic conditions constitute the primary pricing driver, while micro-level characteristics provide incremental but economically modest contributions.

Figure 16(a): Different Train-Test Split Ratios

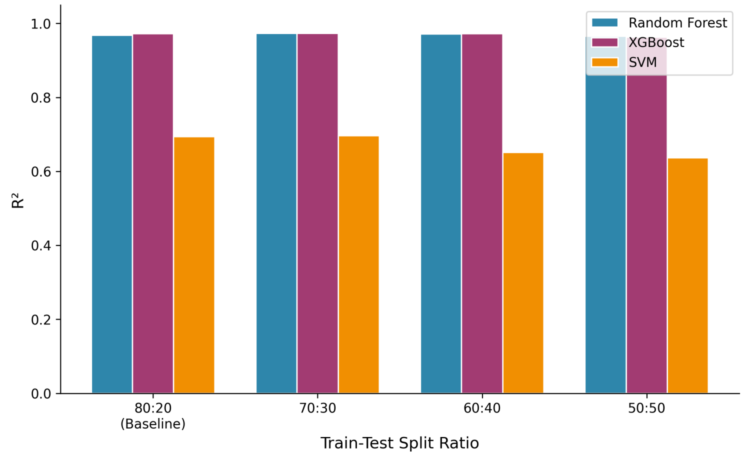
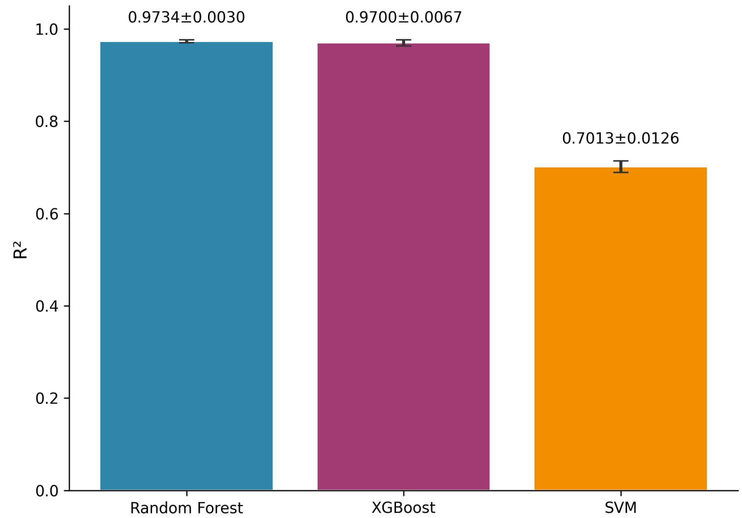


Figure 16(b): 5-Fold Cross-Validation



4. Conclusion

This paper examines 7,062 Chinese green bonds issued between 2016 and 2024, employing three machine learning models—Random Forest, XGBoost, and Support Vector Machine (SVM)—to investigate the determinants of green bond coupon rates. The analysis yields the following key findings.

(1) Among the three candidate models, XGBoost delivers the strongest predictive performance in green bond coupon rate forecasting, achieving an R^2 of 0.9722, followed closely by Random Forest ($R^2 = 0.9680$), while SVM exhibits substantially weaker results ($R^2 = 0.6936$). The superiority of gradient-boosted and bagged tree ensembles over the kernel-based SVM specification is consistent with the documented advantages of tree-based methods in capturing non-linear relationships and high-order feature interactions inherent in financial tabular data.

(2) Macroeconomic factors emerge as the dominant drivers of green bond pricing, with the year-on-year GDP growth rate, year-on-year M2 growth rate, and year-on-year CPI growth rate collectively accounting for over 95% of aggregate feature importance, whereas micro-level factors—including issuer characteristics, financial indicators, and bond-specific attributes—contribute less than 2%. This finding suggests that green bond coupon rates are predominantly determined by the broader macroeconomic and monetary policy environment rather than by issuer-specific idiosyncratic risk.

(3) The Chinese green bond market is structurally dominated by medium-term notes and financial bonds (71.8%) and corporate bonds (18.1%), with non-listed entities serving as the primary issuers (86.2%), reflecting the market's orientation toward institutional and quasi-sovereign financing vehicles.

(4) Comprehensive robustness tests—including alternative train-test splits, five-fold cross-validation, exclusion of the dominant feature, and non-macro-only specifications—confirm that the principal findings are statistically stable and not sensitive to modelling choices or data partitioning strategies.

Theoretical implications. The results contribute to the emerging literature on machine learning applications in sustainable finance by providing systematic evidence on the relative performance of alternative algorithms for green bond pricing. The finding that macroeconomic variables overwhelmingly dominate micro-level predictors extends the debate on information content in bond markets to the green finance domain and suggests that the pricing of green bonds in China operates through mechanisms more closely aligned with sovereign interest rate dynamics than with firm-level credit risk differentiation. This aligns with the theoretical prediction that in markets characterised by concentrated issuance among high-credit-quality entities, aggregate discount rate factors subsume idiosyncratic risk premia.

Practical and policy implications. Based on the foregoing conclusions, the following policy recommendations are proposed. First, regulatory authorities should further strengthen the green bond information disclosure framework by mandating standardised reporting of environmental impact metrics, use-of-proceeds details, and post-issuance verification. Enhanced transparency would augment the informational content of micro-level features, potentially improving pricing efficiency and reducing information asymmetry between issuers and investors. Second, policymakers should strengthen the coordination between macroeconomic stabilisation policy and green finance objectives. Given the outsized influence of monetary conditions (M2 growth) and aggregate economic performance (GDP growth) on green bond pricing, synchronised macro-prudential and green finance policies can optimise the issuance environment and reduce the cost of capital for sustainable projects. Third, financial institutions and asset managers are encouraged to integrate advanced machine learning methodologies—particularly gradient-boosted tree models—into their fixed-income valuation and risk management infrastructure. The demonstrated predictive accuracy of XGBoost in this context suggests tangible benefits for algorithmic pricing, portfolio construction, and market-making activities in the green bond space. Fourth, the standardisation and wider adoption of third-party green bond certification should be promoted to enhance market credibility and reduce greenwashing risk, with a view toward aligning Chinese certification practices with international standards such as the Green Bond Principles.

Limitations and future research. Several limitations of this study merit acknowledgment and suggest avenues for future inquiry. First, the sample is confined to the Chinese green bond market over the period 2016–2024; whether the findings generalise to other jurisdictional contexts—particularly developed markets with different issuer compositions and regulatory frameworks—remains an open question. Second, the analysis employs a static feature set; the incorporation of dynamic temporal features, rolling window estimation, or regime-switching specifications could capture time-varying pricing relationships and structural breaks associated with major policy shifts. Third, while the present study focuses on coupon rate prediction, future research could extend the machine learning framework to the prediction of secondary market yields, credit spreads, and liquidity premia, thereby providing a more comprehensive characterisation of green bond price formation. Fourth, the investigation of non-linear feature importance through techniques such as SHAP (SHapley Additive exPlanations) values or partial dependence plots could yield richer insights into the directional and interaction effects of individual predictors. Fifth, the development of deep learning architectures tailored to structured bond data—such as tabular neural networks or attention-based models—represents a promising direction for potentially improving upon the already high predictive accuracy documented herein.

References

[1] Zhang, L. and Zhao, Q. (2024) 'Current Development Status and Policy Recommendations for China's Green Bond Market', Securities Market Herald, (3), pp. 12-20. (In Chinese)

[2] Merton, R. C. (1974) 'On the Pricing of Corporate Debt: The Risk Structure of Interest Rates', Journal of Finance, 29(2), pp. 449-470.

[3] Duffie, D. and Singleton, K. J. (1999) 'Modeling Term Structures of Defaultable Bonds', Review of Financial Studies, 12(4), pp. 687-720.

[4] Zhou, J. and Li, M. (2024) 'Research on Green Bond Pricing Mechanism and Influencing Factors', Journal of Financial Research, (5), pp. 112-128. (In Chinese)

[5] Chen, S. and Zhao, Y. (2023) 'Research on the Impact of Macroeconomic Factors on Bond Yields', Economic Perspectives, (12), pp. 105-118. (In Chinese)

[6] Flammer, C. (2021) 'Corporate green bonds', Journal of Financial Economics, 142(2), pp. 499-516.

[7] Wang, X. and Liu, F. (2024) 'Empirical Analysis of Factors Affecting Green Bond Credit Spreads', Finance and Trade Economics, (6), pp. 88-101. (In Chinese)

[8] Wu, G. and Shang, J. (2025) 'Research on Green Bond Issuance Cost Advantage and Its Motivations', Bond, (10), pp. 45-56. (In Chinese)

[9] Huang, L. and Zhou, M. (2025) 'Research on the Information Effect of Third-party Certification for Green Bonds', Securities Market Herald, (1), pp. 23-35. (In Chinese)

[10] Li, X. and Wang, D. (2024) 'Application of Machine Learning in Financial Risk Management', Statistical Research, (8), pp. 56-68. (In Chinese)

[11] Kocaarslan, B. and Soytas, U. (2023) 'The role of major markets in predicting the U.S. municipal green bond market performance: New evidence from machine learning models', Technological Forecasting and Social Change, 196, p. 122820.

[12] Ma, W. and Zheng, X. (2024) 'Machine Learning Analysis of the Relationship Between Bond Rating and Issuance Rate', Quantitative and Technical Economics Research, (9), pp. 156-172. (In Chinese)

[13] Zhao, P. and Sun, L. (2024) 'Enterprise Bond Default Prediction Based on Ensemble Learning', Journal of Management Sciences in China, (4), pp. 78-92. (In Chinese)

[14] Lin, X. and Huang, W. (2025) 'Green Finance Policy and Green Bond Market Development', Financial Forum, (2), pp. 34-47. (In Chinese)

[15] Sun, J. and Li, T. (2023) 'Impact of China's Monetary Policy on the Bond Market', International Finance Research, (5), pp. 45-58. (In Chinese)

[16] Breiman, L. (2001) 'Random forests', Machine Learning, 45(1), pp. 5-32.

[17] Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

[18] Vapnik, V. N. (1995) The Nature of Statistical Learning Theory. New York: Springer.

[19] Walker, T., Zhang, X., Zhang, A. and Wang, Y. (2021) 'Fact or fiction: Implicit government guarantees in China's corporate bond market', Journal of International Money and Finance, 116, p. 102414.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.