Comparative analysis of machine learning models for Indian bond market predictions: novel insights with XGBoost.

Vigneshwaraan S
Department of Artificial Intelligence and Data Science
Chennai Institute of Technology
Chennai India
Vigneshwaraans.aids2022@citchennai.net

T Shankar Sai
Department of Artificial Intelligence and Data Science
Chennai Institute of Technology
Chennai, India
Shankarsait.aids2022@citchennai.net

Karthika K
Department of Artificial Intelligence and Data Science
Chennai Institute of Technology
Chennai, India
karthikak@citchennai.net

Abstract —Indian bond markets are at the core of the overall financial system of the country. The overall macro-economic stability of the country depends on the bond markets. In this study, the focus will be to explore the use of machine learning techniques for the prediction of bond prices in the context of the Indian economy. Specifically, the focus will be to explore the use of machine learning techniques for the prediction of 3-year and 10-year bond prices. For 3-year and 10-year bond prices, the study will explore the use of simple linear regression, ARIMA and deep learning technique, specifically the DLSTM technique. To take this study to the next level, the focus will be to explore the use of the XGBoost technique for the prediction of bond prices. This adds a new dimension to the field of bond price prediction. While most studies focus on the prediction of bond yields, this study will focus on the prediction of bond prices. This will be a major breakthrough in the field of bond price prediction. The accuracy of the techniques are evaluated by using the MAPE, Mean Absolute Error, Mean Squared Error, and R-squared techniques. The data was collected for the period 2013 to 2024. The results will show the efficacy of the XGBoost technique for the efficient prediction of bond prices.

Keywords Used: Indian Bond Market, Bond Price Prediction, Machine Learning, Deep Learning, XGBoost, ARIMA, Linear Regression, DLSTM, Time Series Forecasting.

1. INTRODUCTION

1.1 Financial Market Overview:

The financial market is the backbone of the economy because it facilitates financial deals in stocks, bonds, currencies, and derivatives. It helps channelize capital to organizations, governments, and people, which in turn empowers investments, savings, and growth. Among the many financial instruments, the bond market holds an important position in stabilizing economies and financing public and private initiatives. In particular, this market provides a means of financing developmental projects for governments while being a low-risk avenue for investors through investment.

The G-Sec and corporate bond markets in India are growing. Such markets, being an important price determiner in the economy for interest rates, have become an appropriate indicator of the cost incurred by the government and others while borrowing. The Indian government bond market is one of the biggest in Asia, and it is considered a safe heaven for the investors. This is basically because of the relatively lower risk involved with government securities. Moreover, government bonds are considered benchmarks to ascertain the creditworthiness of private companies.

1.2 Indian Government/Treasury Bonds:

Government Securities (G-Secs) contributes a large part in the Indian bond market and it has small amount of risk as they are covered by the Indian government. G-Secs are issued by the Reserve Bank of India on behalf of the Indian government and their availability is also with different maturity periods. Treasury Bills mature in a year or less and long-term bonds may exceed 40 years. G-Secs are dominating the Indian bond market and hence very important for understanding the dynamics of the economy [5].

The yield curve of government bonds can be said to indicate bond yields in relation to their maturities. It is very vital in providing insight into the general state of the economy. The normal upward-sloping yield curve reflects that more returns in long-term bonds compensating the risks associated with longer investment horizons. Conversely, an inverted yield curve usually indicates concern among investors over the slowing down of economic growth.

The Indian government bond market is an integral part of the setting of expectations of the investors about future interest rates. As such, it is being closely monitored by institutional investors, financial analysts, and policymakers [5].

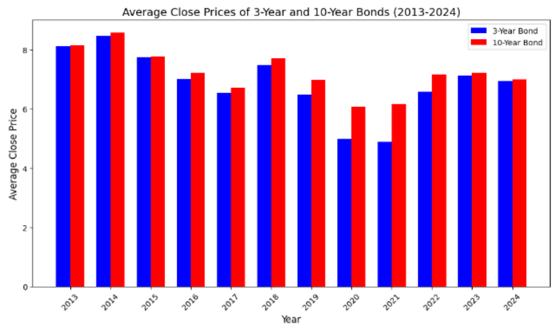

Figure 1. Average Close Price of Indian bond Market (3yr and 10yr).

Inherently, complexity is involved with the process of predicting the bond prices, which makes the usage of advanced Machine Learning and Deep Learning techniques inevitable. At this juncture, we depend on these financial tools to improve the accuracy of our forecast and enlighten us on the behavior of the Indian bond market.

1.3 Motivation of Research and Contributions:

Machine Learning (ML) and Deep Learning (DL) have proven to be promising tools in financial forecasting in an unstable financial market. Statistical tools have failed to deliver satisfactory results in financial forecasting because of the non-linear nature of financial data. This study is based on the application of existing research pertinent to predicting Indian bond yields using Linear Regression, ARIMA, and DLSTM methods.

As an additional contribution, the gradient-boosting technique XGBoost [1],[4] will be applied for the first time to the field of predicting Indian bond prices. Contrary to prior studies which focused on bond yield predictions, the present paper will emphasize the crucial topic of predicting bond prices, which is of enormous importance to every investor and business forecaster. The results will prove the superiority of XGBoost in terms of predictive power and stability, comparing it to traditional as well as deep learning techniques.

1.4 Research Question:

The study in question is about optimizing the degree of precision in our ability to calculate the prices of the Indian government bonds, the 3-year and the 10-year papers, using more sophisticated machine learning algorithms. In this regard, it is important to note that commonly used conventional techniques such as Linear Regression and ARIMA have always been utilized to get the forecasts in the analysis of time series data, but are less effective in the attempt to capture the more complex, difficult to execute relationships and patterns, particularly in the realm of finance, utilizing the more contemporary and more sophisticated DLSTMs algorithms.

One of the interesting angles employed in this work was the introduction of XGBoost into bond price prediction. One of the advantages of the XGBoost algorithm is that it can efficiently deal with large problems while avoiding overfitting, also enhancing the prediction power. This makes XGBoost a good candidate to predict the movements of the financial market.

The research poses the core research question:

To what extent can XGBoost improve the predictions of bond prices compared to Linear Regression, ARIMA, and DLSTM in the Indian bond market? The method by which the performance of the models will be evaluated is based on using certain parameters such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. This will enable the evaluation process to gain a better insight into the potential that machine learning can bring to the field of bond price forecasting. The end objective is one that can provide valuable predictions that can help guide the decisions by various players in the market.

2. LITERATURE REVIEW

At the heart of the bond market, there is a thread of research extending throughout the course of history. The importance of correctly forecasting bond prices is a challenge for both investors and policymakers, who are quick to adapt to the markets. In the long run, the challenge has been met by exploring numerous forecasting techniques for the bond market's trends, from simple statistical tricks to the latest in machine learning technology.

2.1 Traditional Statistical Techniques:

In the traditional approach, the journey starts with Linear Regression and ARIMA, as they are simple and intuitive. In linear regression, for instance, the relationship between the macro variables and the closing price of the bond can be easily visualized. The catch, however, is the linear relationship between the variables, as the bond market is definitely not linear.

While the classic models provide good grounding for understanding, they don't always capture the nonlinear twists and turns, or the outside shocks, that the modern financial markets throw our way. This has contributed to the growing trend towards machine learning for bond market research, where flexibility and pattern-spotting are key to cutting through the chaos.

2.2 Machine Learning and Deep Learning Methods:

Recently, however, there has been a significant increase in the use of machine learning and deep learning to forecast bond prices. Models such as XGBoost, deep neural networks, and deep long short-term memory models have been promising in taming the chaos in the data.

2.3 Gaps in Existing Research:

However, the existing literature is not without its drawbacks. For instance, the majority of the existing research is based on developed countries, while emerging markets like India are relatively under-represented. Even though it is universally accepted that inflation, exchange rates, interest rates, moving averages, etc., are the major determinants of bond prices, these factors are not utilized to the best possible extent in the existing models. Besides, not enough research is available on the relative efficiency of traditional versus machine learning models in the context of bond pricing, especially in emerging markets.

2.4 Contributions of This Paper:

In order to bridge the gap, this study has incorporated various macro factors, including Inflation, Interest Rates, RBI actions, changes in the currency, and Moving Averages, along with historical bond data on the 3-year and 10-year Government Bonds of India, to increase the precision of the forecast. The aim is to have an extensive and thorough analysis of the effectiveness of Linear Regression, ARIMA, DLSTM, and XGBoost in forecasting the closing prices of the Government Bonds of India by pitting these models against each other and finding out the one that is the most effective. The unique contribution of the study is its focus on the effectiveness of the XGBoost in the bond market.

3. METHODOLOGY

In this section, the data used by the forecasting model for the secondary bond market in India will be explained. The data selection, the variables used in the study, and the findings from the descriptive statistics will be explained in this section.

In this study, the data will be collected from various sources, including nseindia.com and investing.com, as these are the commonly used sources by practitioners in the bond market in India. The data will range from January 1, 2013, to October 31, 2024, and will have 2,946 data points for the 10-year bonds and 2,944 data points for the 3-year bonds. The data points will include Close, Open, High, Low, Change %, Daily Return %, and various other macro factors, including Interest Rates, SMA, and Currency Rates. The data on these was obtained from trusted sources like the Reserve Bank of India (RBI) and the World Bank to increase the accuracy of the model.

The data set revolves around the prediction of the Close price of bonds, which represents the market price of the bond at the end of each trading day. Essentially, in this particular research study, we validate the extent to which state-of-the-art deep learning algorithms, such as Deep Long Short-Term Memory (DLSTM), perform relative to classical machine learning algorithms, including Linear Regression, ARIMA, and XGBoost, in the prediction of bond prices. The studies illustrate that XGBoost outperforms the results relative to the other approaches, making this research distinct from past studies that focused on bond yields and/or spreads.

The study seeks to add more predictive power through using various economic input indicators, such as interest rates, currency fluctuations, and other macroeconomic factors that may impact bond and stock markets, given that both tend to fluctuate simultaneously. Typically, when interest rates go up, bond prices fall, while interest in bonds, perceived to be relatively secure investments in comparison to stocks, increases. Currency fluctuations are also a factor, closely linked to changes in interest rates, and impact bond prices.

With the incorporation of all these economic indicators, the forecasting model tries to ensure an accurate yet complete forecasting of the bond prices in India.

3.1 Data Collection and Expansion:

The data set that the Indian Bond Market deals with is about the bond price holds from January 1, 2013, to October 31, 2024. It is about the important indicators such as government securities, namely, the 3-year and 10-year notes, along with other macroeconomic factors such as interest rates, currency, moving averages, and RBI monetary policy. Figure 1 illustrates the Average Close Price for the Indian Bond Market.

These include reputable financial websites such as investing.com and the RBI. This gives a complete picture. The dataset consists of 23,560 data points. The 3-year and 10-year bonds are included in the analysis. Table 1 provides a brief summary of the data used to train the model.

The dataset includes bond price values and important macroeconomic indicators that could affect bond markets. These macroeconomic indicators are Reverse Repo Rate, Cash Reserve Ratio, and Foreign Exchange Reserves. Incomplete and inconsistent values are not included in the dataset to ensure the accuracy and quality of the data.

Table 1. Summary of Dataset for Bond Price Prediction

Name of the bond	Dataset collection period	Model	Total Data Points	Training Data	Testing Data
3-years bond	01-01-2013 to 31-10-2024	Linear Regression	2944	2355	589
		ARIMA	2944	2355	589
		DLSTM	2944	2355	589
		XGBoost	2944	2355	589
10-years bond	01-01-2013 to 31-10-2024	Linear Regression	2946	2357	589
		ARIMA	2946	2357	589
		DLSTM	2946	2357	589
		XGBoost	2946	2357	589
Total	01-01-2013 to 31-10-2024	All Models	23,560	18,848	4,712

3.2 Data Preprocessing:

The paper focuses on the Indian government bond with maturities of 3 years and 10 years for the period gathered from January 1, 2013, to October 31, 2024, gathered at a daily cadence. Data preprocessing included cleaning, feature engineering, and preparing it ready for the machine learning models. Details on the columns used and their importance with the reason for pre-processing are given in Tables 2 and 3 [2].

Table 2. Features and justification

Feature	Justification
Date	Serves as the time index for aligning bond prices and macroeconomic indicators.
Close	Target variable representing the bond's closing price, a key metric for price prediction.
Open, High, Low	Provide insights into daily price movements and market volatility.
Change %	Captures daily price changes as a percentage, reflecting market dynamics.
Daily Return (%)	Indicates the return on the bond based on daily price fluctuations, critical for financial analysis.
Volatility (7D)	Measures weekly price variability, helping capture short-term market risk.
SMA (7D, 30D)	Simple Moving Averages (7-day and 30-day) smooth price trends, aiding in capturing short- and medium-term trends.
Forward Premia of US$ (1M, 3M, 6M) (%)	Reflect currency risk and interest rate differentials affecting bond pricing.
Reverse Repo Rate (%)	Represents the rate at which the RBI borrows funds, impacting bond yields and prices.
Marginal Standing Facility (MSF) Rate (%)	Indicates short-term borrowing costs for banks, indirectly affecting market liquidity.
Bank Rate (%)	Represents the long-term lending rate from the RBI, influencing economic activity and bond prices.
Base Rate (%)	Minimum lending rate for banks, affecting overall financial conditions.
Treasury Bill Yields (91D, 182D, 364D)	Serve as benchmarks for short-term interest rates, crucial for pricing bonds.
10-Year G-Sec Yield (FBIL) (%)	Provides a benchmark for long-term interest rates, closely linked to bond prices.
Cash Reserve Ratio (CRR) (%)	Indicates the liquidity level in the banking system, influencing bond market dynamics.
Statutory Liquidity Ratio (SLR) (%)	Determines the minimum percentage of assets banks must hold, impacting market liquidity.
Policy Repo Rate (%)	Key monetary policy tool affecting borrowing costs and investment flows.
Foreign Exchange Reserves (US$ Million)	Reflects the country's economic stability, influencing investor sentiment in the bond market.

Table 3: Summary of Preprocessed Dataset

Bond Type	Time Period	Frequency	Features	Dataset Size
3-Year Bond	01-Jan-2013 to 31-Oct-2024	Daily	25 (as detailed above)	2943 rows
10-Year Bond	01-Jan-2013 to 31-Oct-2024	Daily	25 (as detailed above)	2945 rows

3.3 Designing Data with Statistics:

The dataset used was created meticulously for both training and testing purposes. A total of 23,560 data points were processed for training and testing purposes. For training purposes, 80% of the points were used, while 20% were reserved for testing purposes. Here, history of bond prices is used for training; verification of how the model performs is also based on this [2]. This allows for the creation of any desired machine learning models for predicting bond close prices based on its historical price history as well as other factors.

The time period from January 1, 2013, to October 31, 2024, was considered, and the data was processed using a variety of Python libraries to make feature engineering possible. In this regard, the number of data points in the training dataset was considered to be 18,848, while the number was considered to be 4,712 in the testing dataset.

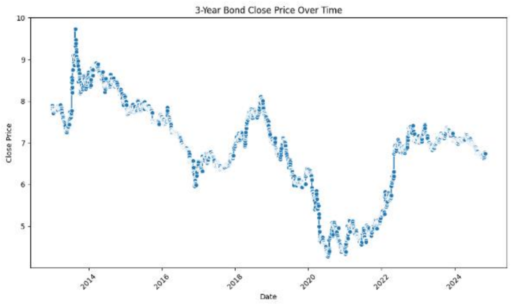

Figure 2. 3-Year Bond Close Price Over Time.

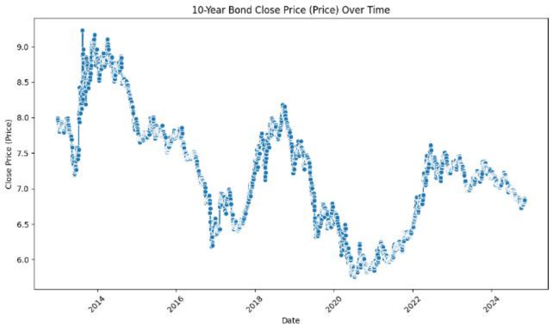

Figure 3. 10-Year Bond Close Price Over Time.

3.4 Models:

The study employed a combination of various machine learning techniques to enhance how well a forecast could be made. They imported a deep LSTM recurrent neural network, the ARIMA model, and simple linear regression to boost the predictive potential of their forecast. The model also introduced the XGBoost algorithm to provide a new perspective on the study.

3.4.1. Linear Regression:

Linear regression is one of the fundamental techniques used by machine learning that was utilized to determine the extent to which each feature was responsible for the output by assigning weights to the numerical features. The general form of the multiple linear regression model can be written as:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + .. + \varepsilon$$

Where,
y = Dependent variable (response variable)
$\beta 0$ = y intercept
$\beta 1$ and $\beta 2$ = slope coefficients
X1 and X2= Independent variables (predictors)
ϵ = Error term

The goal of linear regression is to minimize a cost function called Mean Squared Error, or MSE, which represents the difference between what a model predicts and reality.

For evaluating how good a Linear Regression model has performed, different parameters are utilized, such as MSE, MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), R-squared, etc.

3.4.2. Deep LSTM Neural Network:

Long Short-Term Memory (LSTM) networks are a kind of Recurrent Neural Network (RNN) that is specifically designed to learn and remember both short- and long-range patterns. The use of the LSTM model is particularly popular when trying to make financial time series predictions such as the price of the bond.

A network of LSTMs is constructed from individual units that contain input gates, output gates, as well as forget gates to control flow within it appropriately to ensure that critical information is not lost as it moves from one time step to successive ones. During its forward pass, an LSTM's operations follow a number of equations:

$$f_t = \sigma_g \, (W_f \times x_t + U_f \times h_{t-1} + b_f)$$
$$i_t = \sigma_g \, (W_i \times x_t + U_i \times h_{t-1} + b_i)$$
$$o_t = \sigma_g \, (W_o \times x_t + U_o \times h_{t-1} + b_o)$$
$$c'_t = \sigma_c \, (W_c \times x_t + U_c \times h_{t-1} + b_c)$$
$$c_t = f_t \, . \, c_{t-1} + i_t \, . \, c'_t$$
$$h_t = o_t \, . \, \sigma_c(c_t)$$

Where,
$\sigma_g = Sigmoid$
$\sigma_c = tanh$
. = element wise multiplication
f_t = Forget gate
i_t = Input gate
o_t = Output gate
c_t = Cell state
h_t = Hidden state

The hyperparameters of a different type of RNN, specifically LSTM networks, will include a number of layers, a number of units in each of these layers, learning rate, batch size, and length of sequence in each input. To attain the best results, a process called hyperparameter tuning tunes all these parameters. Table 4 lists all the hyperparameters used in building the DLSTM model when used in determining 3-year and 10-year bond prices, along with a justification of these parameters.

Table 4: Hyperparameters and Justifications (DLSTM)

Hyperparameter	Value	Justification
Sequence Length (SEQ_LENGTH)	90	A 90-day sequence length captures both short-term and medium-term temporal dependencies, which are crucial for financial time series forecasting.
LSTM Units	128 per layer	The number of units is increased to 128 to allow the model to capture complex patterns and dependencies in the time series data.
Dropout Rate	0.3	Dropout technique is used to avoid overfitting by randomly dropping out 30% of the units during training. This technique proves effective for models with a lot of parameters.
Learning Rate	0.0003	A smaller learning rate ensures gradual updates to the model weights, helping achieve convergence without overshooting the optimal point.
Batch Size	32	A smaller batch size provides a balance between training efficiency and convergence stability, particularly in noisy financial data.
Epochs	200	A high number of epochs allows the model to learn effectively over multiple iterations, while early stopping prevents overfitting.
Early Stopping Patience	15	The patience of 15 epochs ensures that the training stops early if no improvement is observed in the validation loss, saving computational resources.
Loss Function	Mean Squared Error	MSE penalizes large errors more heavily, which is crucial for minimizing the deviation between predicted and actual bond prices.
Optimizer	Adam	The Adam optimizer is chosen for its adaptive learning rate and ability to handle sparse gradients, which is beneficial for time series data.

When assessing the performance of the bond price prediction model, the following metrics were used to measure the extent of the performance of the LSTM model: Mean squared error, Mean Absolute Error, Mean Absolute Percentage Error, and the R-squared statistic.

3.4.3. ARIMA

The Auto-Regressive Integrated Moving Average (ARIMA) model is a traditional time series forecasting technique that is used in predicting the value of bonds in the stock market. The ARIMA model is capable of revealing the link between different data values. The ARIMA model can be used when the time series data is stationary.

With regard to the forecast of bond prices, it makes use of both the autoregressive (AR) and the moving average (MA). These are the components of ARIMA, which make use of past values of the series as well as the forecast errors while conducting the forecast of the future values.

In summary, the ARIMA model combines the various components to create the temporal process and produce the predictions.:

$$y_t = \beta + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + .. + \alpha_p y_{t-p}$$

Where,
y_t = the next timestamp value which will be forecasted,
β = the intercept term or a constant
α_i = the coefficient of lagged observation
y_{t-i} = the lagged observation
p = the number of lag orders

3.4.4. XGBoost

XGBoost, or Extreme Gradient Boosting, is a swift machine learning method, which can be commonly used for regression. It functions through several decision trees, and each tree works to correct the mistakes of preceding trees in the list, thereby promoting prediction accuracy [1],[4].

To forecast the Close prices of the bonds, XGBoost was utilized for predicting the relationship between the historical metrics of the bond such as Close, Open, High, Low, etc., and other important financial indicators. It improves the accuracy of its predictions by optimizing its objective function:

$$L(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

The loss function that we aim to optimize is the Mean Squared Error (MSE), and Ω(f) represents the regularizing factor in order to prevent over-complexity.

Table 5: Hyperparameters and Justifications (XGBOOST)

Hyperparameter	Value	Justification
Objective	reg:squarederror	Specifies the regression task, minimizing the squared error for continuous target variables like bond prices.
n_estimators	200	A moderate number of estimators ensures a balance between performance and computational cost.
learning_rate	0.1	A learning rate of 0.1 provides a balance between convergence speed and avoiding overfitting.
max_depth	7	Allows the model to capture moderately complex patterns in the data without excessive overfitting.
subsample	0.8	Uses 80% of the training data for each tree, reducing overfitting and improving generalization.
colsample_bytree	0.9	Utilizes 90% of the features for each tree, ensuring diverse trees and capturing important features.
random_state	42	Ensures reproducibility of results by fixing the random seed.

3.5 Evaluation Metrics:

With this piece of work, we seek to determine the level of accuracy that the models in machine learning (ML) and deep learning (DL) achieve in terms of forecasting the prices of the bonds using a set of essential evaluation criteria: Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2).

3.5.1. Mean Squared Error (MSE)

- Definition: MSE refers to the mean of all squared differences between what is predicted by a model and what really happens.

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

  Where:
    - MSE = mean squared error
    - N = number of data points
    - y_i = observed values
    - $\hat{y}_i$ = predicted values
- Purpose: MSE shows how similar or dissimilar predictions of a model are to real values. A smaller MSE means a better model since its predictions are close to exact values. The squaring of errors distinguishes larger discrepancies.

3.5.2. Mean Absolute Error (MAE):

- MAE informs you how close your predictions have been to the real numbers. It provides you with a simple yet understandable level of accuracy of your model. This feature proves useful when you need to understand how well various models perform on different datasets or scales. The smaller the MAE of your model, the more accurate your predictions.

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

  Where:
    - MAE = mean absolute error
    - N = number of observations
    - y_i = actual value
    - $\hat{y}_l$ = predicted value

3.5.3. Mean Absolute Percentage Error (MAPE):

- MAPE indicates just how off-target these predictions are as a percentage of the actual values, essentially giving a quick idea of just how accurate a prediction or prediction method is. In other words: a smaller number means better predictive accuracy.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_l - y_i}{y_i}\right| \times 100$$

  Where:
    - MAPE = mean absolute percentage error
    - N = number of observations
    - y_i = actual value
    - $\hat{y}_l$ = predicted value

3.5.4. R-squared (R^2):

- R^2 or the coefficient of determination describes the proportion of the variation in the dependent variable explained by the model with its independents. If the R^2 is close to 1, then most of the variance is explained by the model; if it is close to 0, little is explained.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

  Where:
    - $\bar{y}$ = mean of actual value
    - y_i = actual value
    - $\hat{y}_i$ = predicted value

4. RESULTS AND DISCUSSIONS

The aim of this study is to examine how the various algorithms compare in the prediction of India's bond prices, specifically with regard to accuracy and performance. Four different algorithms are utilized here: Linear Regression, ARIMA (Autoregressive Integrated Moving Average), DLSTM (Deep Long Short-Term Memory) Neural Networks, and XGBoost (Extreme Gradient Boosting). The performance of each, again specifically with regard to accuracy, was evaluated using the usual error metrics of MAE, MAPE, MSE, and R^2.

All of the algorithms are able to provide forecasts on India's bond prices, but not all to the same level of accuracy.

In particular, Linear Regression and ARIMA are more likely to catch the overall trends of the bond prices' movements but struggle to handle the increased complexity of the data. On the other hand, DLSTM and XGBoost are more effective in handling nonlinear and time-dependent relationships.

The XGBoost model excelled in all aspects in all parameters. XGBoost had the minimum MAE, MAPE, and MSE values, as well as the highest R^2 value. This indicates that XGBoost is a better choice to deal with the complex nonlinear relationship in bond markets.

The model used in this research is a predictive model that uses both financial and macroeconomic indicators to generate a complete forecast of bond prices.

4.1 MODEL ACCURACY:

In this section, we are going to discuss how these models are performing in predicting bond prices for 3-year and 10-year bonds. We are going to use Linear Regression, ARIMA, DLSTM, and XGBoost as our models to be compared. We use MAPE, MAE, MSE, and R-squared as evaluation metrics to compare the models. These indicators provide information on how well each model can effectively predict bond prices and how tightly their predictions track the actual prices, reflecting each model's ability to explain the fluctuations in price movements.

Table 6. Results of Linear Regression, ARIMA, Deep LSTM and XGBoost models for 3-year and 10-year government bond price prediction.

Methods	Metrics	MAPE	MAE	MSE	R squared
Linear regression	3 yrs	0.0108	0.0762	0.0092	0.7008
	10 yrs	0.0114	0.0817	0.0077	0.7743
Arima	3 yrs	0.0199	0.1391	0.0321	-
	10 yrs	0.0362	0.2551	0.0968	-
DLSTM	3 yrs	0.0090	0.0633	0.0078	0.7556
	10 yrs	0.0062	0.0450	0.0035	0.8894
XGBoost	3 yrs	0.0018	0.0753	0.0101	0.9890
	10 yrs	0.0013	0.0398	0.0026	0.9938

The XG Boost model emerged as a winner by producing the lowest MAPE values for both the 3-year and 10-year bonds, which were 0.0018 and 0.0013 respectively. Therefore, it was clear that XG Boost was the most accurate model compared to other models for this particular problem, and its MAPE values were much lower compared to other models, especially for the 10-year bond Linear Regression and ARIMA models.

Another model, DLSTM, was also able to perform well, especially when dealing with the 10-Years Bond, when it surpassed Linear Regression and ARIMA by a good margin with a MAPE of 0.0062. This captures its ability to learn temporal dependencies in the price movements for bonds, although it was not able to achieve XGBoost's accuracy level.

Overall, the MAPE comparison reinforces XGBoost as the strongest model in predicting bond prices, according to the study.

4.2 Results Visualization and Discussion:

To show our forecasting capability, we have created a graph visualization and table showing the predictions of each model for 3-year and 10-year bond prices.

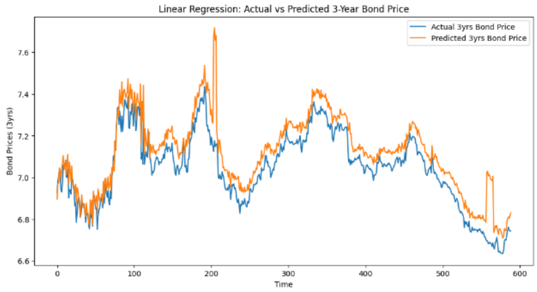

Figure 4. Predicted vs Actual Bond Prices for Linear Regression (3-Year)

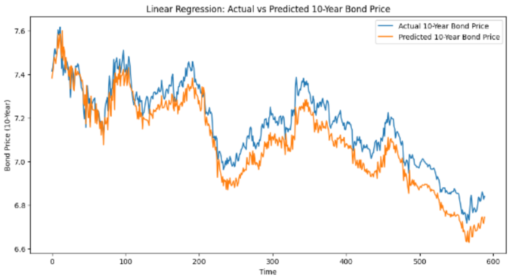

Figure 5. Predicted vs Actual Bond Prices for Linear Regression (10-Year)

Figure 4 shows the Linear Regression Model in action for the 3 years and 10 years bond prices in Figure 5. As can be observed from the figures, Linear Regression finds it hard to interpret or understand the volatility and complexity of the patterns in the prices of the bonds. It gives an idea of the direction of the trends; however, as can be seen from the graph, its forecasts tend to differ significantly from the actual prices of bonds, especially for the 10-year bonds.

- The model works moderately well for the 3-year bond but does not pick up all the finer wiggles, and it has a resultant MAPE of 0.0108.
- For example, for a 10-year bond, the knowledge of mismatch from true values becomes more visible with a relatively higher MAPE of 0.0114.

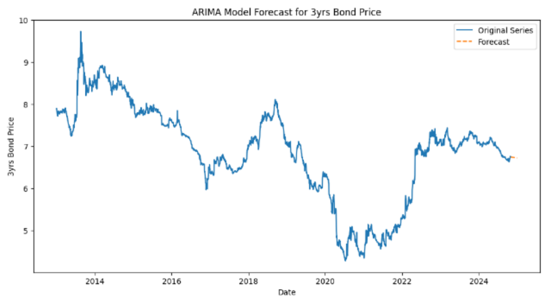

Figure 6. Predicted vs Actual Bond Prices for ARIMA (3-Year)

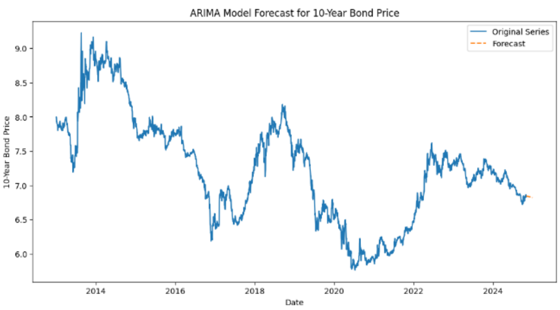

Figure 7. Predicted vs Actual Bond Prices for ARIMA (10-Year)

Figure 6 and Figure 7 present the ARIMA forecasts for the prices of 3-year and 10-year bonds, respectively. The ARIMA technique models time-series data by analyzing past trends to forecast future trends. The technique, however, is not suited for data with non-linear movements and dynamic relationships, like the prices in the market.

- For instance, in the case of the 3-year bond, whilst ARIMA picks up most of the changes in upward and downward movements, larger movements are missed, as demonstrated by a larger MAPE of 0.0199.
- The performance of the 10-Year Bond is worse, with an unusually higher MAPE of 0.0362. The orange prediction lines deviate more in the graphs from the actual prices, especially in periods of volatility.

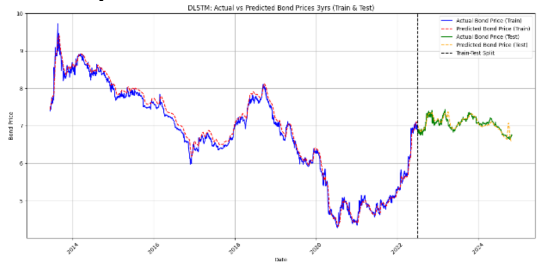

Figure 8. Predicted vs Actual Bond Prices for DLSTM (3-Year)

Figure 9. Predicted vs Actual Bond Prices for DLSTM (10-Year)

Figure 8 and Figure 9 graphically represent the performance of the model when predicting bond prices for the 3 years and 10 years maturity, respectively. The DLSTM model, which stands for Deep Long Short-Term Memory, excels in the detection of long-range dependencies and trends, which makes it performance-enhancing when it comes to detecting the dynamic changes of the bond prices.

- In the case of the 3-year bond, the DLSTM shows improved accuracy compared to Linear Regression and ARIMA, with a lower MAPE of 0.0090. In this case, the orange line of the predicted values closely follows the blue line of the actual values but fails to attain the level of precision achieved by XGBoost.
- For the 10-year bond, the accuracy level is significantly improved using DLSTM, achieving a MAPE of 0.0062. It can be observed that predictions have a good alignment with the actual prices, except for a slight deviation during the volatile period.

Figure 10. Predicted vs Actual Bond Prices for XGBoost (3-Year)

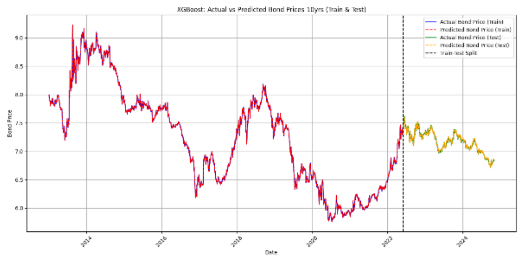

Figure 11. Predicted vs Actual Bond Prices for XGBoost (10-Year)

Figure 10 and Figure 11 show the XGBoost prediction results for the 3-year and 10-year bond prices. XGBoost is a robust gradient boosting algorithm that is very popular for its potential to handle complex, non-linear relationships between features and achieve elevated levels of prediction accuracy.

- In the case of the 3-year bond, the model shows outstanding results, as evidenced by the fact that the predicted prices follow the actual prices closely. This can be seen from a MAPE of 0.0018.
- Again, the performance occurs in the 10-year bond case, where the MAPE is 0.0013. Once more, the forecasts track the actual values nearly perfectly. It again emphasizes the capabilities of the XGBoost model to estimate the movements of the bonds.

4.3 Actual vs Predicted Values:

In this section, we create a table containing a brief summary where we compare the actual bond prices with their predicted values based on whether they are 3-year or 10-year bond values. The table will show all four models, namely Linear Regression, ARIMA, DLSTM, and XGBoost, comparing their predictions with the actual values.

Table 7. Actual vs Predicted Bond Prices (3yr):

Date (3yr)	Actual Price (3yr)	Predicted Price (3yr) - Linear Regression	Predicted Price (3yr) - ARIMA	Predicted Price (3yr) - DLSTM	Predicted Price (3yr) - XGBoost
01-06-2022	6.925	6.999790	7.062807	6.987325	6.937465
12-07-2022	6.887	6.961380	7.024051	6.948983	6.899397
25-08-2022	6.915	6.989682	7.052608	6.977235	6.927447
07-10-2022	7.338	7.417250	7.484026	7.404042	7.351208
22-11-2022	7.104	7.180723	7.245370	7.167936	7.116787
02-01-2023	7.146	7.223177	7.288205	7.210314	7.158863
10-02-2023	7.188	7.265630	7.331041	7.252692	7.200938
27-03-2023	7.125	7.201950	7.266788	7.189125	7.137825
15-05-2023	6.925	6.999790	7.062807	6.987325	6.937465
23-06-2023	7.010	7.085708	7.149499	7.073090	7.022618
04-08-2023	7.161	7.238339	7.303504	7.225449	7.173890
18-09-2023	7.179	7.256533	7.321862	7.243611	7.191922
01-11-2023	7.319	7.398045	7.464648	7.384871	7.332174
14-12-2023	7.144	7.221155	7.286166	7.208296	7.156859
29-01-2024	7.043	7.119064	7.183156	7.106387	7.055677
12-03-2024	7.039	7.115021	7.179076	7.102351	7.051670
30-04-2024	7.188	7.265630	7.331041	7.252692	7.200938
13-06-2024	6.993	7.068524	7.132161	7.055937	7.005587
26-07-2024	6.839	6.912861	6.975096	6.900551	6.851310

Table 8. Actual vs Predicted Bond Prices (10yr):

Date (10yr)	Actual Price (10yr)	Predicted Price (10yr) - Linear Regression	Predicted Price (10yr) - ARIMA	Predicted Price (10yr) - DLSTM	Predicted Price (10yr) - XGBoost
01-06-2022	7.414	7.498520	7.682387	7.459967	7.423638
12-07-2022	7.391	7.475257	7.658554	7.436824	7.400608
25-08-2022	7.292	7.375129	7.555970	7.337210	7.301480
07-10-2022	7.460	7.545044	7.730052	7.506252	7.469698
22-11-2022	7.285	7.368049	7.548717	7.330167	7.294471
02-01-2023	7.340	7.423676	7.605708	7.385508	7.349542
10-02-2023	7.363	7.446938	7.629541	7.408651	7.372572
27-03-2023	7.312	7.395357	7.576694	7.357334	7.321506
15-05-2023	7.001	7.080811	7.254436	7.044406	7.010101
23-06-2023	7.074	7.154644	7.330079	7.117859	7.083196
04-08-2023	7.193	7.275000	7.453387	7.237597	7.202351
18-09-2023	7.149	7.230499	7.407794	7.193324	7.158294
01-11-2023	7.358	7.441881	7.624360	7.403620	7.367565
14-12-2023	7.194	7.276012	7.454423	7.238603	7.203352
29-01-2024	7.171	7.252749	7.430590	7.215460	7.180322
12-03-2024	7.026	7.106096	7.280341	7.069561	7.035134
30-04-2024	7.195	7.277023	7.455459	7.239609	7.204354
13-06-2024	6.986	7.065640	7.238893	7.029313	6.995082
26-07-2024	6.940	7.019116	7.191228	6.983028	6.949022

Tables 7 and 8 show the actual prices of bonds daily, along with the predictions of the actual prices for both 3-year and 10-year bonds based on the predictions of the four different models: Linear Regression, ARIMA, DLSTM, and XGBoost. The explanation will discuss the major differences in terms of the actual and predicted prices of bonds based on the different models.

5. Conclusion

Implications: This study delves into the potential of the application of machine learning and deep learning in the prediction of bond prices in the Indian market. The article pits Linear Regression, ARIMA, DLSTM, and XGB to expose the potential it holds in predicting bond price movements, which is a topic that touches on an issue of global concern to investors.

Findings and Novelty: The findings have proven that XGBoost is the best performing algorithm for both the 10-year and 3-year bonds. In the case of the 10-year bond, XGBoost recorded a MAPE score of 0.0013 and an R-squared score of 0.9938, outperforming other models. This was replicated in the case of the 3-year bond, where XGBoost recorded a MAPE score of 0.0018 and an R-squared score of 0.9890.

Limitations: The research has a few lacunas. First and foremost, though the results look promising, the time frame of the experiment from 2013 to 2024 could have been extended still. A real-time dataset and the responses from the sentiment analysis of news articles could have helped the prediction accuracy. Moreover, the experiment being based on the Indian bond market, the overall accuracy on a different nation/market still has to be evaluated.

Future Research: This piece provides a good foundation for future research into bond market forecasting using machine learning approaches. Research into this topic could involve the use of a combination of machine learning and other types of data, including sentiment, global economic indicators, and live market data. Broadening the scope could provide a better perspective on the use of machine learning techniques in finance as a whole.

References

[1] Authors: Hassan Oukhouya, Hamza Kadiri, Elhimdi Khalid, Raby Guerbaz. (2024) "Forecasting International Stock Market Trends: XGBoost, LSTM, LSTM-XGBoost, and Backtesting XGBoost Models" https://www.researchgate.net/publication/376951516_Forecasting_International_Stock_Market_Trends_XGBoost_LSTM_LSTM-XGBoost_and_Backtesting_XGBoost_Models.

[2] Author: Salisu Garba Abdullahi (2024) "Boosting Nigeria's Bond Market: Evidence from Macroeconomic Perspective"

https://www.researchgate.net/publication/386555653_Boosting_Nigeria%27s_Bond_Market_Evidence_from_Macroeconomic_Perspective.

[3] Authors: Dharika Kapil and Kannan Yamini (2024). "Machine Learning Applications in Predicting Bond Market Trends: Explore How Machine Learning Algorithms Can Be Used to Predict Movements in the Fixed Bond Market". DOI:10.20944/preprints202404.1667.v1 .

[4] Authors: Yan Zhang, Lin Chen (2021). "A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method". https://doi.org/10.4236/tel.2021.112019.

[5] Authors: Wisniewski, T. P. (2016). "Is there a link between politics and stock returns? A literature survey. International Review of Financial Analysis". https://doi.org/10.2139/ssrn.2690852.

[6] Authors: Bhaskar Nandi, Subrata Jana, Krishna Pada Das (2024). "Machine learning-based approaches for financial market prediction: A comprehensive review". https://doi.org/10.59400/jam.v1i2.134.

[7] Authors: Robert Verner, Michal Tkáč sr., Michal Tkáč jr. (2021). "Improving Quality of Long-Term Bond Price Prediction Using Artificial Neural Networks". https://doi.org/10.12776/QIP.V25I1.1532.

[8] Authors: Bhaskar Nandi, Subrata Jana, Krishna Pada Das (2024). "Machine learning-based approaches for financial market prediction: A comprehensive review". https://doi.org/10.59400/jam.v1i2.134.

[9] Authors: Paul, M., & Reddy, K. S. (2022). "US QE and the Indian bond market. Journal of Quantitative Economics". https://doi.org/10.1007/s40953-021-00257-9.

[10] Authors: Wahidin, D., Akimov, A., & Roca, E. (2021). "The impact of bond market development on economic growth before and after the global financial crisis: Evidence from developed and developing countries. International Review of Financial Analysis". https://doi.org/10.1016/j.irfa.2021.101865.