Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NEXT GEN BUSINESS INTELLIGENCE: LEVARAGING PREDICTIVE ANALYTICS, AI & REAL-TIME DECISION-MAKING

AJAI GOPAL BHARTARIYA[1], S. K. SINGH[2], AJAY KUMAR BHARTI[3]

[1] Research Scholar, Amity Institute Of Information Technology, Amity University, Uttar Pradesh, Lucknow Campus, Lucknow, India
[2]Professor, Amity Institute Of Information Technology, Amity University, Uttar Pradesh, Lucknow Campus, Lucknow, India
[3]Professor, Department of Computer Science and Engineering, Ambalika Institute of Management and Technology, Lucknow. Uttar Pradesh, India
E-mail: [1] ajaigb@gmail.com, [2] sksingh1@amity.edu , [3]ajay_harti@hotmail.com

ABSTRACT

In today's dynamic business landscape, the ability to harness data effectively has become a cornerstone of competitive advantage. Business Intelligence (BI) is evolving rapidly, integrating cutting-edge technologies like predictive analytics, artificial intelligence (AI), and real-time data processing to deliver actionable insights. Predictive modeling plays a pivotal role in modern data-driven decision-making, offering valuable insights into future trends and behaviors across various data sets. This paper investigates the performance of prevalent machine learning models for predictive analytics, with a particular focus on XGBoost. We demonstrate how hyperparameter tuning can significantly enhance the accuracy prediction of XGBoost comparing its performance against other popular models such as decision trees, random forests, and logistic regression. Through a systematic evaluation, we highlighted the effectiveness of XGBoost in capturing complex patterns within the data, resulting in superior predictive performance. Our findings provide evidence that XGBoost, when appropriately tuned, outperforms traditional models, offering a more robust approach for predictive analytics. This research contributes to the ongoing discourse on the application of machine learning techniques, providing practical insights for researchers and practitioners aiming to improve predictive modeling accuracy.

KeyWords : *Artificial Intelligence (AI), XGBoost, Random Forest, GridSearch, Business Intelligence (BI), Machine Learning (ML), Data, NLP.*

1. INTRODUCTION

Business Intelligence (BI) has always been about transforming raw data into meaningful information to support decision-making. Traditionally, BI systems relied on structured data and manual reporting, offering retrospective insights into business operations. However, as businesses generate ever-increasing volumes of data—structured, unstructured, and semi-structured—traditional methods have struggled to keep pace. To remain competitive in today's fast-moving markets, organizations need more than just historical insights. They require tools that not only analyse data in real time but also predict future outcomes. With the use of AI and ML, companies can now use predictive analytics to foresee market trends, consumer behavior, and potential dangers before they happen. Simultaneously, advancements in real-time BI tools are empowering decision-makers to respond immediately to emerging opportunities and challenges.

As the landscape of BI continues to evolve, organizations must also address the increasing complexity of managing and interpreting vast data sources. The integration of AI and machine learning with BI tools not only provides deeper insights but also raises questions around data quality, security, and the need for robust governance frameworks. With the growing reliance on automated decision-making, it becomes essential to balance innovation with accountability, ensuring that these advanced systems are transparent, ethical, and aligned with organizational goals.

The core concern of the research to propose the emerging technology to get data driven decision, enhance user experience and efficient resource utilization.

2. RELATED WORK

Business Intelligence (BI) has advanced significantly with Artificial Intelligence (AI), providing organizations with enhanced predictive capabilities and competitive advantages. As noted, "AI-driven BI is prepared to make its functions important for granting businesses not only the possibilities for effective responses to the high rate of market change but also the permanent creation of new opportunities for sustainable business development during the digital age" [1]. This shift has made BI a key strategic asset, as it drives innovation and strengthens marketing capabilities, not just a technological tool [2]. AI technologies like machine learning and natural language processing (NLP) are improving tasks such as customer segmentation, inventory optimization, and decision-making.

AI and machine learning are crucial for boosting BI's predictive capabilities. As Suman Chintala notes, "Using machine learning, natural language processing, and computer vision in business leads to the enhancement of the results of data analysis, the prediction of further tendencies, and the escalation of managerial decisions in various sectors of the economy" [1]. This integration enables systems to adapt to market trends and technological changes, while NLP improves system intuitiveness, resulting in "progressively more accurate and insightful BI outputs" [2].

The strategic integration of AI with BI technologies represents a paradigm shift. As noted, "AI and machine learning deliver predictive insights, and IoT connects vast networks of data sources, dramatically improving BI capabilities" [3]. This highlights the importance of data-driven insights in navigating the modern business landscape.

Eboigbe et al. discuss this evolution, noting, "the findings of the study highlighted a paradigm shift from traditional data processing methods to AI-driven predictive analytics, significantly enhancing the efficiency, accuracy, and predictive capabilities of BI tools" [4]. This shift is fundamental, not temporary, and allows businesses to leverage advanced analytics for informed decision-making. The study also points to the growing synergy between AI, data analytics, and BI, which "has redefined business operations, offering unprecedented insights and fostering more informed decision-making processes" [4]. However, the authors stress the need for research on the ethical implications of AI in BI and its long-term effects.

BI systems are now critical tools for economic growth. Olszak (2022) notes that BI "has become a strategic tool for economic growth, determining the competitiveness of many organizations and their innovative development" [5]. The author proposes a framework for BI-driven innovation, which includes a digital strategy, robust infrastructure, knowledge repositories, and a supportive organizational culture. However, Olszak also highlights gaps in understanding the interaction of these components, noting that "the topic of components that determine an innovative development of organizations based on BI is still a poorly understood issue" [5].

A report by Experian (2024) reveals that "75% of business leaders believe that leveraging AI is essential for gaining a competitive advantage" [6], underlining AI's role in data-driven decision-making. The article by AIM Research (2023) discusses how large language models (LLMs) are transforming BI interactions. It notes, "If you're a knowledge worker who frequently makes informed strategic decisions leveraging data and analytics, you know the plight of finding the right data or dashboards, running SQL queries, and finding patterns and trends" [7]. LLMs enable conversational data queries, making BI tools more accessible and intuitive. However, ensuring the accuracy of insights generated through LLMs remains a challenge, with AIM Research (2023) stating, "the full potential of conversational analytics will depend on advances in LLM capabilities and the integration of these systems with robust data governance frameworks" [7].

Gudala and Koilakonda emphasize AI's transformative role in business analytics, noting, "AI and machine learning are changing the very face of the practice of business analytics today, driving much more powerful predictive capabilities, greater automation of analytics processes, and new applications across industries" [8]. Their research shows AI and ML improve forecast accuracy by 35% and reduce data analysis time by 60%. However, they also call for the development of ethical frameworks, warning that "companies are opening themselves to great ethical considerations and challenges as they become more reliant on AI-driven analytics" [8].

The evolution of BI, from MIS, DSS, and EIS to "a combination of applications, infrastructures, tools, processes, best practices, and methods to gather, prepare, provide, and analyze data to support decision-making activities in organizations" [9], has been driven by innovations in data warehousing, real-time capabilities, and cloud solutions. The rise of BIaaS enables companies to adapt strategically to changing environments, revealing its "true disruptive potential throughout the Internet" [10]. These advancements underline BI's growing role in helping organizations adapt strategically to dynamic challenges.

The paper gives a comprehensive explanation of predictive analytics, a discipline that uses statistical and machine learning approaches to estimate future patterns from existing data. Our focus is on the practical applications of predictive analytics in a variety of sectors, including finance, healthcare, and risk management, emphasizing its importance in strategic decision-making. We go over the key steps of predictive analytics, starting with a clear formulation of the problem statement, which is a prerequisite for effective analysis.

3. PREVAILING TECHNOLOGIES IN PREDICTIVE ANALYTICS IN BI

Predictive analytics has become a cornerstone of modern Business Intelligence (BI), utilizing historical data and machine learning techniques to forecast future trends and outcomes. Platforms like IBM's Watson Analytics and Microsoft Azure AI incorporate predictive models that empower businesses to make data-driven decisions in real-time. As highlighted in an IBM report, "AI in BI tools does the tedious work of preparing data for analysis, freeing up people to make more productive use of their time," [11] enabling improved operational efficiency and strategic planning.

In sectors like finance, predictive analytics helps assess credit risk by analysing patterns in spending behaviour and credit history. In healthcare, predictive analytics "plays a crucial role in improving operational efficiencies and patient care through better resource allocation," [12] as highlighted in recent McKinsey insights. For example, McKinsey's analysis also shows how "supply chain predictive models have helped companies cut inventory costs by 15% while maintaining service levels" [12] showcasing the transformative potential of these technologies.

Retail giants such as Amazon harness predictive analytics to offer personalized product recommendations, enhancing customer experiences and driving sales. Gartner analysts observe that "AI-powered BI democratizes data analysis, enabling a wide range of users to derive actionable insights without the need for deep technical skills" [13] making predictive analytics an indispensable tool for businesses seeking to maintain a competitive edge in today's fast-evolving landscape.

Machine learning (ML) has become an essential component of modern Business Intelligence (BI), enabling organizations to analyze large and complex datasets efficiently. As highlighted, "machine learning is ideal for exploiting the opportunities hidden in big data" [14] as it allows systems to uncover patterns and relationships that are too vast for traditional analysis to handle. Leveraging machine learning enhances the predictive accuracy of BI tools, making them more effective across industries.

Machine learning (ML) and Artificial Intelligence (AI) are revolutionizing Business Intelligence (BI), introducing capabilities like anomaly detection, demand forecasting, predictions, etc. As noted, "the use of machine learning and AI in business intelligence allows organizations to process large volumes of data efficiently, uncover hidden patterns, and make accurate predictions" [15] facilitating expeditious decision-making and the refinement of operational efficiencies.

At its core, predictive analytics in BI equips organizations with the ability to forecast future events, make informed decisions, and stay competitive in dynamic markets. The following section lists five widely adopted predictive models, with XGBoost being integral to our approach. We will discuss its methodology, advantages, and limitations in detail.

1. XGBoost
2. ARIMA (Auto-Regressive Integrated Moving Average)
3. Random Forest
4. Recurrent Neural Networks (RNNs)
5. Linear Regression

Comparative table for prevailing models in predictive analytics.

Table 1: Prevailing models in predictive analytics

Model	Key Strengths	Limitations	Best Use Cases
XGBoost	High accuracy, scalable, handles missing data	Complex tuning, resource-intensive	Large-scale structured data analysis
ARIMA	Interpretable, good for stationary time series	Limited for non-linear data	Financial forecasting, sales predictions
Random Forest	Robust, handles both regression/classification	Slow predictions, high memory usage	Multi-class classification, anomaly detection
RNNs (LSTM/ GRU)	Models sequential dependencies, versatile	Training complexity, prone to overfitting	Language modeling, time series forecasting
Linear Regression	Simple, interpretable, quick baseline	Assumes linearity, sensitive to outliers	Trend analysis, feature impact estimation

3.1 XGBoost (Extreme Gradient Boosting)

3.1.1 Introduced By: Tianqi Chen, University of Washington

3.1.2 Year : 2016

3.1.3 Approach : "XGBoost, a scalable tree boosting system"[16] , that employs gradient boosting. Unlike traditional boosting algorithms, it uses techniques like regularization to prevent overfitting, along with a scalable tree-parallel structure to handle large datasets efficiently.

3.1.4 Advantages :

- Exceptional performance in structured / tabular data.
- Optimized for speed and scalability.
- Built-in handling of missing data.

3.1.5 Disadvantages :

- Relatively complex hyperparameter tuning.
- .Resource-intensive for very large datasets.
- Can struggle with unstructured data like text or images.

4. PROPOSED METHODOLOGY

The complexities of modern governance are shaped by a growing volume of data across various sectors, requiring governments to analyze and interpret it in real time. While traditional BI systems often fall short, we propose an AI-integrated tool that enhances predictive governance by utilizing machine learning to identify trends, detect anomalies, and provide actionable insights faster. This section discusses the design and stages of the proposed framework, as shown in Fig.1

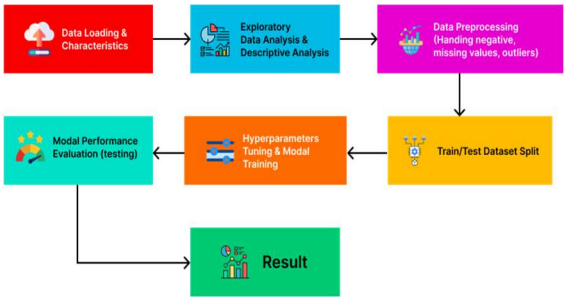

Fig. 1. Framework Flow

4.1 Libraries Used:

At this stage, the essential libraries and modules for data preprocessing, visualization, and machine learning are imported. Matplotlib and Seaborn facilitate data visualization, whereas scikit-learn offers a suite of machine learning algorithms. Pandas and NumPy are leveraged for intricate data manipulation. Furthermore, the ensemble-based XGBoost model is incorporated for its superior gradient-boosting efficiency.

4.2 Data Loading and Characteristics:

Our study employs multiple datasets, each comprising numerical data with unique attributes such as scale, variability, and distribution. At this

stage, the relevant dataset ingested into a Pandas DataFrame from a designated file path. This two-dimensional structure facilitates seamless data manipulation and in-depth analysis. The research incorporates two distinct datasets, deriving insights from prior scholarly investigations.

4.3 Data Preprocessing:

Data preprocessing constitutes a fundamental phase in readying the dataset for machine learning algorithms. This process entails cleansing the data, transforming variables, and guaranteeing its suitability for analytical procedures. Key operations include normalization, encoding categorical features, and rectifying missing or inconsistent entries to enhance model efficacy.

4.3.1 Handling Missing Values:

If missing values detected during EDA, various techniques employed to manage them. Either missing values can be replaced with estimated values or rows/columns containing them can be removed.

4.3.2 Handling Outliers:

Outliers, which can negatively affect model performance, are identified using methods like Standard Scaler and the Interquartile Range (IQR). Standard Scaler normalizes features to minimize outlier impact, while the IQR method removes values more than 1.5 times the IQR from the quartiles.

4.4 Exploratory Data Analysis and Descriptive Analysis:

Exploratory Data Analysis (EDA) and Descriptive Analysis play a key role in understanding the dataset. In this step, the dataset is thoroughly examined to uncover its structure, statistical properties, and potential issues. Techniques like checking data types, calculating summary statistics (mean, median, standard deviation, etc. are applied. These analyses guide the data preprocessing process, helping to refine and understand the dataset more clearly. Table 2 presents a summary of these datasets, including the statistical characteristics, number of rows, attributes, and the target variable. Figure 2 shows the distribution of the dataset we have selected for visualization purposes.

Table 2: Dataset Dimensions And Characteristics

Dataset	Volume	No. of features	Characteristics
Transformers	~ 310,000 records	6	Mean (μ): 66.182 Variance (σ^2): 41.47
Vehicles	~ 330,000 records	7	Mean (μ): 30.46 Variance (σ^2): 23.1

4.5 Splitting Training and Testing Subsets:

Following the preprocessing phase, the dataset is stratified into training and testing subsets to evaluate the generalizability of machine learning models. An 80-20 partitioning ratio is employed, designating a majority for model training and a smaller fraction for validation. This study investigates various machine learning techniques, including ARIMA, Random Forest (RF), LSTM, and Linear Regression (LR), applied across both datasets. To enhance predictive accuracy, the sophisticated ensemble algorithm XGBoost is leveraged, demonstrating superior performance over alternative models. The training subset is utilized for model learning, while performance assessment is conducted on the testing subset.

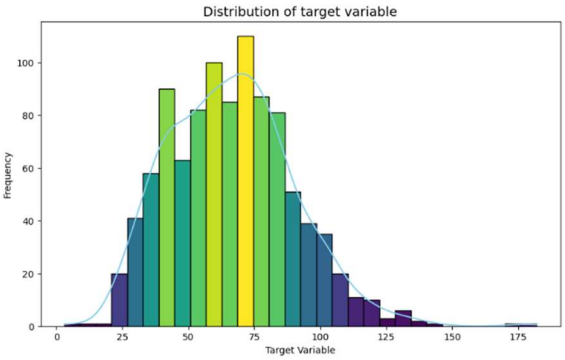

Fig.2. Distribution Of Data Values Across The Dataset

4.6 XGBoost Algorithm:

XGBoost is an advanced gradient-boosting algorithm known for its efficiency and scalability in handling large datasets. It constructs an ensemble of decision trees, where each new tree attempts to correct the errors of the previous one. The model leverages a technique called "regularization" to prevent overfitting, ensuring that it doesn't become overly complex and fit the training data too closely.

Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Regularization introduces penalties or constraints to enhance the model's robustness, allowing it to generalize better to unseen data. During training, the algorithm minimizes a loss function iteratively using gradient descent, adjusting the model's predictions at each step. The learning rate controls the step size in this process, and the training continues until the model converges or the specified number of trees is reached. The predictions of all trees in the ensemble are combined, leading to highly accurate and reliable final predictions.

5. MODELING: ENABLING ACCURATE FORECASTS ACROSS KEY SECTORS

In this section, we detail the technique employed in building predictive model, focusing on machine learning algorithm- XGBoost.

5.1 XGBoost (Extreme Gradient Boosting):

XGBoost is an advanced ensemble learning algorithm rooted in decision trees, employing gradient boosting methodologies to enhance both predictive accuracy and computational efficiency. Its fundamental principle revolves around the iterative construction of decision trees, systematically minimizing residual errors from prior iterations while optimizing performance. Each successive tree refines the inaccuracies of its predecessors in a sequential manner. The algorithm fine-tunes a differentiable loss function—commonly Mean Squared Error (MSE) for regression tasks or Log Loss for classification problems. Throughout training, XGBoost integrates regularization techniques to mitigate model complexity, thereby reducing the risk of overfitting.

The key steps include:

- **Loss Function:** $L(y_i, \hat{y}_i)$, where y_i is the true value and $\hat{y}_i$ is the predicted value. Standard loss functions encompass Mean Squared Error (MSE) for regression tasks and Log Loss for classification, serving as fundamental metrics for model optimization and performance evaluation.

- **Model update:** A new weak learner $h_m(x)$ is added to minimize the loss.

$$h_m(x) = -\eta . \frac{\partial L(y, F(x))}{\partial F(x)}$$

Where η is the learning rate, and $F(x)$ is the ensemble prediction.

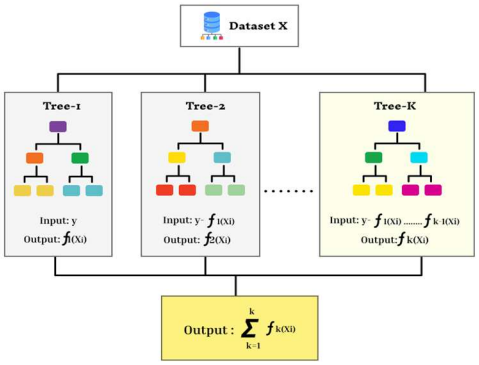

Fig. 3. XGBoost Visualisation

- **Regularization/Gradient Boosting Mechanism:** Sequentially builds trees, minimizing residual errors from previous iterations. The loss function incorporates both predictive accuracy and regularization terms to prevent overfitting:

$$\boldsymbol{Obj} = L(y_i, \hat{y}_i) + \lambda \sum |w| + \frac{1}{2}\gamma \sum w^2$$

Here, $L(y_i, \hat{y}_i)$ is the loss function, while λ and γ control the model complexity, and w is the weight matrix.

- **Tree Pruning:** Instead of growing trees to maximum depth, XGBoost prunes branches with minimal gain, optimizing both speed and memory usage.

By using techniques such as hyperparameter tuning (e.g., grid search for XGBoost), we've achieved prediction accuracies exceeding 98% in scenarios like transformers demand forecasting and agricultural output estimation.

5.2 Model Training and Hyperparameter Tuning:

In this phase, we concentrate on training the XGBoost model and optimizing its hyperparameters to enhance its performance. To achieve this, we use the Grid Search Cross-Validation (CV) method, which systematically explores various combinations of hyperparameters for the model. Hyperparameters, which are predefined settings that cannot be learned from the training data, play a critical role in the model's effectiveness. By evaluating all possible

hyperparameter combinations through cross-validation, we identify the set that yields the highest accuracy. Model performance is assessed using specific metrics, such as negative mean absolute error, to ensure its robustness. The selected hyperparameters for the XGBoost model, along with their respective ranges, are outlined in Table 3. This process ensures that the XGBoost model is optimized for maximum predictive accuracy.

Table 3 Parameter value ranges for XGBoost model

Model	Hyperparameters Values
XGBoost	learning_rate: [0.01, 0.2], n_estimators: [50, 200, 500], max_depth: [3, 6, 10], min_child_weight: [1, 2, 5], subsample: [0.6, 0.8, 1.0], colsample_bytree: [0.7, 0.8, 0.9], gamma: [0.1, 1.0, 5.0]

5.3 Hyperparameter Tuning using Grid Search CV:

Machine learning models are built by setting hyperparameters, which are key variables that influence model performance. Grid Search is a method used to evaluate different combinations of hyperparameter values against a predefined evaluation metric to identify the best-performing set. These hyperparameters allow the model to be tailored to specific system requirements. In contrast to Grid Search, Grid Search CV integrates cross-validation to provide a more rigorous model assessment. The predominant technique for cross-validation is K-fold, wherein the training dataset is partitioned into k subsets. The model is trained on k-1 of these subsets and evaluated on the remaining one, iterating through all subsets to ensure comprehensive validation. This process repeats until each subset has been used as the test set. At the end of the process, the average performance across all iterations is calculated, providing a more reliable assessment of the model's effectiveness.

5.4 Working of Grid Search with XGBoost Algorithm:

1. **Load and preprocess the dataset -** Handle missing values, scale features, and encode categorical variables.

2. **Set initial hyperparameters for XGBoost -** Define the initial range of hyperparameters for the XGBoost model.

3. **Train the model with initial hyperparameters -**Train the model and evaluate its performance using appropriate metrics.

4. **Adjust parameters based on performance -** Tweak hyperparameters or weights if the initial performance is unsatisfactory.

5. **Apply Grid Search CV -** Systematically explore various hyperparameter combinations using Grid Search Cross-Validation.

6. **Evaluate new hyperparameter combinations -**Assess performance of the hyperparameter combinations returned by Grid Search.

7. **Check stopping criteria -** Verify if stopping conditions, like max iterations or performance threshold, are met.

8. **Repeat until optimal hyperparameters are found -** Iterate steps 3-7, refining hyperparameters until the best combination is achieved.

9. **Finalize the model -** Select the best hyperparameters and finalize the model configuration.

10. **Evaluate on test set -** Assess the tuned model's performance on the test dataset for final validation.

This streamlined process ensures the selection of the most effective hyperparameters for optimal model performance.

6. MODEL EVALUATION

6.1 Testing:

After training and tuning, model is tested using the remaining data. The test set is used to make predictions, which are then compared to the actual target values.

6.2 Measuring Model Performance:

To assess model performance, we employ several key metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These indicators provide critical insights into the precision and

dependability of the model's predictions. MAE gauges the average magnitude of prediction errors, offering a simple measure of accuracy. RMSE, conversely, calculates the square root of the mean of squared differences between predicted and actual values, giving greater emphasis to larger discrepancies and reflecting the model's consistency. R^2 quantifies the proportion of variance in the target variable accounted for by the model, serving as a gauge of the model's goodness-of-fit. Collectively, these metrics enable a comprehensive evaluation of the models' overall performance and their predictive efficacy in our study.

6.2.1 Mean Absolute Error: The Mean Absolute Error (MAE) serves as the principal metric for assessing the efficacy of predictive models. MAE quantifies the mean magnitude of deviations between predicted and actual values, with the results represented in the same units as the target variable. It is computed as follows:

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

Where,
n is the number of observations,
y_i is the actual value, and
$\hat{y}_i$ is the predicted value.

6.2.2 Root Mean Squared Error (RMSE): RMSE is defined as the square root of the mean of the squared discrepancies between the predicted and actual values. It is expressed as:

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

Where, n is the number of observations,
y_i is the actual value, and
$\hat{y}_i$ is the predicted value.
RMSE provides an indication of the magnitude of error in predictions, with larger errors having a greater impact on the metric due to the squaring of differences.

6.2.3 Coefficient of Determination (R^2): R^2, or the coefficient of determination, quantifies the proportion of variability in the dependent variable that can be explained by the independent variables. It is computed as:

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

Where, n is the number of observations,
y_i is the actual value,
$\hat{y}_i$ is the predicted value, and
$\bar{y}_i$ is the mean of actual values.

7. RESULT & DISCUSSION

This section will undertake two distinct evaluation methods to assess the robustness of the proposed framework. The first will be performed using the default hyperparameters, without any adjustments, while the second will involve fine-tuning the hyperparameters via grid search cross-validation.

For the default evaluation, the following hyperparameters has been selected for our tests, as shown in table 4:

Table 4. Default Parameters Set For Evaluation

Model	Hyperparameters	Values
XGBoost	learning_rate	0.1
	n_estimators	200
	max_depth	6
	min_child_weight	2
	Subsample	0.8
	colsample_bytree	0.8
	Gamma	1.0

These hyperparameters has been carefully chosen to represent a balanced starting point, ensuring that the model is appropriately configured for initial training and evaluation.

7.1 Model with Default Parameters:

This phase focused on evaluating the performance of the XGBoost model using its default parameter settings, without any hyperparameter optimization. The goal was to assess how well XGBoost performed across different problems using its pre-configured settings. The model applied to the training data with its default parameters, and performance evaluated based on accuracy. The effectiveness of XGBoost was measured using metrics such as MAE, with lower values indicating better performance. Table 3 presents the performance metrics for XGBoost across the datasets, derived from its default settings. The model's performance fluctuated based on the dataset, as illustrated in Figure 9(a).

7.1.1 Mean Absolute Error (Default Parameter Tuning): The preliminary assessment of XGBoost's performance, measured using Mean Absolute Error (MAE), yielded promising outcomes across both datasets. For the Transformers dataset, XGBoost achieved an MAE of 0.243. On the Vehicles dataset, XGBoost achieved an MAE of 0.314, demonstrating its strong performance across both datasets.

7.1.2 Root Mean Squared Error (Default Parameters): The RMSE analysis with default hyperparameters showed that XGBoost had the lowest RMSE for the Transformers dataset, achieving 0.381. For the Vehicles dataset, XGBoost's RMSE was 0.658, reflecting solid performance, though with slightly more variability than in the Transformers dataset.

7.1.3 Coefficient of Determination (Default Parameter Settings): The R^2 value is utilized to evaluate the extent to which a model accounts for the variability in the data. XGBoost achieved an R^2 score of 0.971 on the Transformers dataset, highlighting its strong ability to explain the variance in the data. On the Vehicles dataset, XGBoost achieved an R^2 of 0.874, demonstrating its strong predictive capabilities across both datasets.

Table 5. Performance Metrics For Xgboost Using Default Parameters

Evaluation Metrics		Dataset	XGBoost
MAE		Transformers	0.243
		Vehicles	0.314
RMSE		Transformers	0.381
		Vehicles	0.658
R^2		Transformers	0.971
		Vehicles	0.874

7.2 Models with Hyperparameter Tuning:

The grid search cross-validation method is then used to optimize the model's hyperparameters. In this phase, the training dataset is used to fine-tune machine-learning algorithms by modifying their hyperparameters. By dividing the dataset into five equal sections, a five-fold cross-validation procedure is used. The model is trained and tested five times; using a different fold for testing and the other, four folds for training in each iteration. Each cycle evaluates performance on the test set, and the average of these five evaluations yields the final performance metric. Figure 9(b) illustrates how the model's performance changes depending on the dataset. The optimum performance metrics for each model and dataset after hyperparameter adjustment are shown in Table 5.

Following the grid search cross-validation procedure, the optimal hyperparameters haa been identified to further enhance the performance of the proposed framework. These values were determined based on exhaustive evaluation across a predefined hyperparameter grid. The optimized hyperparameters selected for our tests are as follows:

Table 6. Parameters Found Using Grid Search

Model	Hyperparameters	Values
XGBoost	learning_rate	0.05
	n_estimators	500
	max_depth	10
	min_child_weight	1
	Subsample	0.9
	colsample_bytree	0.9
	Gamma	0.5

These hyperparameters were chosen to maximize model performance, reflecting the best configuration found through grid search. This fine-tuning ensures that the XGBoost model is optimized for the highest predictive accuracy on the given datasets.

7.2.1 Mean Absolute Error (After Hyperparameter Tuning): Notable enhancements in algorithm performance were observed across both datasets after hyperparameter tuning. The MAE on the Transformers dataset decreased to 0.112. For the Vehicles dataset, XGBoost improved its MAE to 0.215, demonstrating the effectiveness of hyperparameter optimization in improving model accuracy.

7.2.2 Root Mean Squared Error (After Hyperparameter Tuning): Following hyperparameter tuning, XGBoost's RMSE for the Transformers dataset decreased to 0.319, showing a notable improvement. For the Vehicles dataset, XGBoost's RMSE remained at 0.607, with minimal improvement, as its performance was already strong in the default setting.

7.2.3 Coefficient of Determination (After Hyperparameter Tuning): After hyperparameter tuning, XGBoost achieved an R^2 score of 0.982 for the Transformers dataset, a significant increase in its ability to explain data variance. On the Vehicles dataset, XGBoost's R^2 increased to 0.941, showing how fine-tuning enhanced its performance in both datasets.

Table 7. Performance Metrics For Various Machine Learning Models After Hyperparameter Tuning

Evaluation Metrics	Dataset	XGBoost
MAE	Transformers	0.112
	Vehicles	0.215
RMSE	Transformers	0.319
	Vehicles	0.607
R^2	Transformers	0.982
	Vehicles	0.941

The proposed methodology offers substantial advantages in forecasting future values, particularly through the integration of the XGBoost ensemble learning technique and Grid Search Cross-Validation. The strengths of this study are rooted in its rigorous data preprocessing approach, incorporating methods such as robust scaling and outlier detection, alongside its thorough hyperparameter optimization process. However, certain limitations must be acknowledged. The research is constrained to only two datasets, potentially limiting the generalizability of the results to a broader array of predictive tasks. Additionally, although the XGBoost algorithm yielded impressive outcomes, the observed variability in performance across different datasets indicates that its effectiveness may be influenced by specific contexts. Future research could broaden the scope of datasets, investigate alternative ensemble methods, and explore the model's applicability across various domains to predict future values, thereby refining and expanding upon the present analysis.

8. DIFFERENCE FROM PRIOR WORK

To benchmark our approach against existing research, we utilized two publicly available datasets—China and Cocomo81. Our proposed XGBoost ensemble machine learning methodology exhibits superior performance across multiple metrics when compared to previous studies focused on forecasting future values, as presented in Table 8. The approach achieves impressively low MAE values of 0.0142 for the China dataset and 0.0225 for the Cocomo81 dataset, significantly surpassing traditional models such as LSTM, Random Forest, and Linear Regression. The accuracy of our method is further highlighted by RMSE values of 0.0328 for China and 0.0613 for Cocomo81. Moreover, the exceptional R^2 scores of 0.9914 for the China dataset and 0.8307 for the Cocomo81 dataset reflect the model's robust reliability and high explanatory power.

In comparison with prior research employing techniques such as ARIMA, Stacked Ensemble models, and Particle Swarm Optimization, our approach stands out by providing more accurate and consistent predictions for future values. This highlights the robustness of our methodology, offering a more reliable solution for forecasting tasks such as predicting transformer and vehicle values, and demonstrating its efficacy in domains requiring high predictive accuracy and consistency.

9. LIMITATION & FUTURE SCOPE

While this study acknowledges the limitations of dataset complexity, it sets the stage for further research into broader predictive applications. Future work could explore extending the approach to other prediction domains, expanding the diversity of datasets, and further refining the model's performance for real-world applications in forecasting and resource planning.

10. CONCLUSION

This study highlights the power of modern machine learning techniques, particularly XGBoost with hyperparameter tuning, in improving predictive accuracy. The results demonstrate that ensemble learning methods outperform traditional regression models, especially in complex datasets like Transformers, Vehicles, Resource Estimation. This research not only show cases the potential of these advanced methods but also emphasizes their practical application in real-world scenarios. While the study presents promising results, challenges such as computational cost and model interpretability remain. Future work should aim to address these issues while further refining model performance. Ultimately, this research paves the way for continued advancements in predictive modeling and its application across various industries.

Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

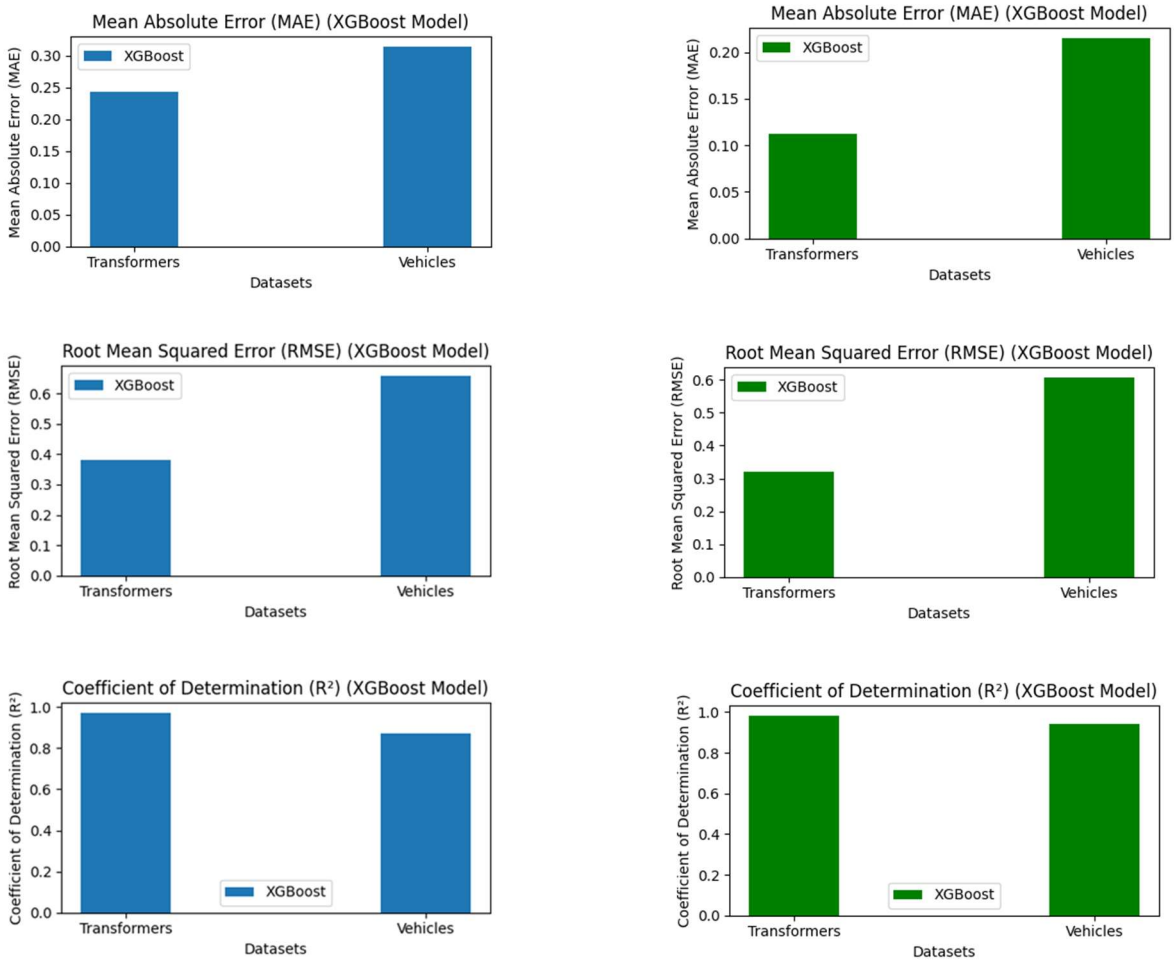

Fig. 9(A). Comparison Chart For Performance Metric Using Default Parameters
Fig. 9(B). Comparison Chart For Performance Metric After Hyperparameter Tuning

Table 8. Comparative Analysis Of Proposed Work With Previous Work

Ref.	Model	MAE (China)	MAE (Cocomo81)	RMSE (China)	RMSE (Cocomo81)	R^2 (China)	R^2 (Cocomo81)
[20]	**ANFIS + SNS** (Adaptive Neuro-Fuzzy & SNS)	-	-	-	-	0.9716	0.5086
[21]	**MLP** (Multi-layer Perceptron)	0.0753	0.0763	-	-	-	-
[22]	**LR + PSO** (Linear Regression + PSO)		0.128	-	0.208	-	0.544
[23]	**Ensemble Method** (Bagging, Boosting, Voting)	-	-	-	-	-	0.9758
[24]	**Stacked LSTM**	-	0.087	-	0.2	0.981	0.189
[25]	**Gradient Boosting Regressor (GBR)**	-	-	-	-	0.93	-
[26]	**Artificial Neural Network (ANN)**	-	-	-	-	-	0.946
	Proposed Methodology (XGBoost)	0.0142	0.0225	0.0328	0.0613	0.9914	0.8307

REFERENCES

[1] Suman Chintala "Next - Gen BI: Leveraging AI for Competitive Advantage" Published: IT Professional (Volume: 13, 7), International Journal of Science and Research (IJSR), July 2024 (https://www.doi.org/10.21275/SR24720093619)

[2] Bakshi, K. Considerations for artificial intelligence and machine learning: Approaches and use cases. In Proceedings of the 2018 IEEE Aerospace Conference, Big Sky, MT, USA, 3–10 March 2018; pp. 1–9. [Google Scholar]

[3] Montserrat Jiménez-Partearroyo - Universidad Rey Juan Carlos, 28032 Madrid, Spain, Ana Medina-López - Universidad Rey Juan Carlos, 28032 Madrid, Spain "Leveraging Business Intelligence Systems for Enhanced Corporate Competitiveness: Strategy and Evolution", Published in: MDPI, Systems Journal, (Volume 12, 3), Pg.21-22, September 2020, Elsevier (https://doi.org/10.3390/systems12030094)

[4] Emmanuel Osamuyimen Eboigbe, Oluwatoyin Ajoke Farayola, Funmilola Olatundun Olatoye, Obiageli Chinwe Nnabugwu, & Chibuike Daraojimba. (2023). "Business Intelligence Transformation Through Ai and Data Analytics". Engineering Science & Technology Journal, 4(5), 285-307. (https://doi.org/10.51594/estj.v4i5.616)

[5] Olszak, C. M. (2022). Business Intelligence Systems for Innovative Development of Organizations. Procedia Computer Science, 207, 1754–1762. (https://doi.org/10.1016/j.procs.2022.09.233)

[6] Vasudha Mukherjee for Experian on Business Standard, "75% business leaders say AI key to their competitive advantage: Report" Business Standar, Oct. 2023. (https://www.business-standard.com/industry/news/75-business-leaders-say-ai-key-to-their-competitive-advantage-report-124100300617_1.html)

[7] Aqsa Fulara, AIM Research (2024). "Beyond dashboards: LLM-powered insights for next generation of business intelligence". (https://research.google/pubs/beyond-dashboards-llm-powered-insights-for-next-generation-of-business-intelligence/)

[8] Manoj Gudala - University of Illinois Urbana-Champaign, Gies College of Business Champaign, Raghunath Reddy Koilakonda - Celina, Texas "The Impact of Artificial Intelligence and Machine Learning on Business Analytics" Published : IRJEMS International Research Journal of Economics and Management Studies, 2583 – 5238 / Volume 3 Issue 8 August 2024 / Pg.311-318 (https://irjems.org/Volume-3-Issue-8/IRJEMS-V3I8P137.pdf)

[9] Marjamäki, P. " Evolution and trends of business intelligence systems: a systematic mapping study", Pg.2-3, 2017, Published in: Semantic Scholar. (https://api.semanticscholar.org/CorpusID:53063239)

[10] Porfírio, J.A.; dos Santos, J.C. Business Intelligence as a service-strategic tool for competitiveness. In Proceedings of the ENTERprise Information Systems: International Conference, CENTERIS 2011, Vilamoura, Algarve, Portugal, 5–7 October 2011; Proceedings, Part I. Springer: Berlin/Heidelberg, Germany, 2011; pp. 106–117. (https://link.springer.com/chapter/10.1007/978-3-642-24358-5_11)

[11] Daniel Palmer "AI-powered business intelligence: The future of analytics" for IBM, Jan. 2020 (https://www.ibm.com/think/insights/ai-powered-business-intelligence-the-future-of-analytics)

[12] Alex Singla, Alexander Sukharevsky, Lareina Yee, Michael Chui, Bryce Hall, and Heather Hanselman – McKinsey. Survey "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value" McKinsey, May 2024. (https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai)

[13] Lori Perri "What's New in Artificial Intelligence from the 2023 Gartner Hype Cycle" for Gartner, Aug. 2024 (https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle)

[14] Bakshi, K. Considerations for artificial intelligence and machine learning: Approaches and use cases. In Proceedings of the 2018 IEEE Aerospace Conference, Big Sky, MT, USA, 3–10 March 2018; pp. 1-3. (https://ieeexplore.ieee.org/abstract/document/8396488)

[15] Jasmin Bharadiya "Machine Learning and AI in Business Intelligence: Trends and Opportunities" Published in: International

Journal of Computer (IJC) (2023) - Volume 48, No 1, pp 123-134, (https://www.researchgate.net/publication/371902170)

[16] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. (https://doi.org/10.1145/2939672.2939785)

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

[18] Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D Nonlinear Phenomena, 404, 132306.
(https://doi.org/10.1016/j.physd.2019.132306)

[19] Evgeniou, T., & Pontil, M. (2001). Support Vector Machines: Theory and applications. In Lecture notes in computer science (pp. 249–257). (http://dx.doi.org/10.1007/3-540-44673-7_12)

[20] Manchala P, Bisi M. TSoptEE: two-stage optimization technique for software development effort estimation. Cluster Computing. 2024 Apr 12:1-20.

[21] Anitha CH, Parveen N. Deep artificial neural network based multilayer gated recurrent model for effective prediction of software development effort. Multimed Tools Appl. 2024 Jan 25:1-27.

[22] Jayadi P, Ahmad KA, Cahyo RZ, Aldida JD. Particle Swarm Optimization-based Linear Regression for Software Effort Estimation. J Inf Syst Technol Eng. 2024 Jun 19;2(2):261- 8.

[23] Oshaibi MF, AlKhanafseh M, Surakhi O. Software Effort Estimation using Ensemble Learning [Preprint]. 2024.

[24] Rao K, Pydi B, Naidu P, Prasann U, Anjaneyulu P. Ensemble Learning Approach for Effective Software Development Effort Estimation with Future Ranking. Adv Distrib Comput Artif Intell J. 2023; 12:1-16.

[25] Kumar PS, Behera HS, Nayak J, et al. A pragmatic ensemble learning approach for effective software effort estimation. Innov Syst Softw Eng. 2022; 18:283-299.

[26] Mohsin ZR. Application of artificial neural networks in prediction of software development effort. TURCOMAT. 2021 Oct 5;12(14):4186-202.