MANTECH Publications

International Journal of Advance Civil Engineering and Technology
Volume 11, Issue 1, January-June 2026
ISSN: 3107-7374 (Online)

Predictive Analytics for Construction Cost Management Using Machine Learning Techniques

Kadambari Ashok Nikam[1], Dr. Satish Waysal[2]

PG Scholar[1], Research Guide & Assistant Professor[2]

Department of Civil Engineering (Construction & Management)[1], Department of Civil Engineering[2]

MVPS's Karmaveer Adv. Baburao Ganpatrao Thakare College of Engineering

Email ID: *nikamkadambari@gmail.com[1], waysal.satish@kbtcoe.org[2]*

ABSTRACT

Construction projects frequently encounter considerable obstacles in achieving cost efficiency due to intricate variables, fluctuating site conditions, and unreliable forecasting techniques. This research examines the use of predictive analytics combined with sophisticated machine learning methods to enhance construction cost management. By utilizing historical project data, essential cost factors such as material costs, labor hours, changes in project scope, and scheduling delays are identified and scrutinized. Machine learning algorithms, including Random Forest, Gradient Boosting, and Support Vector Regression, are trained and assessed for their predictive precision. The suggested methodology shows a significant enhancement in the reliability of cost estimations and supports real-time, data-driven decision-making. The incorporation of intelligent systems into construction management practices presents a transformative approach to minimizing budget overruns, optimizing resource distribution, and improving project outcomes. This study offers important insights into the creation of scalable, automated cost management solutions for the construction sector.

KEYWORDS: *Construction cost prediction, predictive analytics, machine learning, cost management, data-driven decision-making, project budgeting, resource optimization.*

INTRODUCTION

The construction sector is a fundamental pillar of economic and social development, yet it remains vulnerable to persistent issues related to cost overruns and budget inaccuracies. As construction projects become increasingly complex and resource-intensive, the limitations of traditional cost estimation practices have become more apparent. Inaccurate forecasts often lead to inefficient resource allocation, financial disputes, and reduced project performance.

Conventional estimation methods are typically based on expert opinion and historical averages, which are often subjective and unable to adapt to changing project conditions. These approaches struggle to incorporate the combined effects of multiple cost drivers such as labor productivity, material price volatility, project duration, and site-specific risks.

The emergence of predictive analytics using machine learning offers new opportunities to improve construction cost estimation. ML models can analyze extensive datasets, uncover hidden patterns, and generate forecasts that evolve with incoming data. When combined with dynamic management control, these models enable continuous cost monitoring and early identification of financial risks.

This study develops a machine learning-based framework for construction cost estimation and dynamic cost control. The framework focuses on predicting total construction cost, monthly expenditure distribution, and budget categorization. The objective is to enhance forecasting reliability, support proactive cost management, and contribute to improved decision-making in construction project planning and execution.

1. Objectives of study

The primary objective of this study is to develop a machine learning–based predictive analytics framework for efficient construction cost management, enabling accurate forecasting and data-driven financial decision-making throughout the construction project lifecycle.

The specific objectives of the study are to:

1. Develop and implement a machine learning model capable of accurately predicting total construction cost using key project attributes such as building area, number of floors,

material quality, labor cost, equipment cost, and project duration.

2. Estimate month-wise construction expenditure to provide a phased financial profile that supports effective cash flow planning and timely allocation of resources.
3. Design a budget categorization mechanism that classifies construction projects into predefined cost tiers (low, medium, and high) based on predicted cost outcomes.
4. Incorporate contextual and project-specific features, including project type, location characteristics, and time overrun probability, to enhance the robustness and adaptability of the prediction model.
5. Evaluate the predictive performance of the proposed framework using appropriate statistical and machine learning evaluation metrics, and compare the results with baseline estimation methods to validate its effectiveness in real-world construction scenarios.

LITERATURE REVIEW

1. Overview of Construction Cost Estimation Research

Construction cost estimation has long been recognized as a critical component of project planning and control. Early studies relied primarily on deterministic models, expert judgment, and statistical regression techniques. While these approaches provided baseline estimates, they often failed to address uncertainty, nonlinearity, and the dynamic nature of construction projects. As project complexity increased and large volumes of data became available, researchers began exploring intelligent and data-driven methods to improve estimation accuracy and decision-making reliability.

Recent literature indicates a clear paradigm shift from traditional cost estimation approaches toward machine learning (ML), artificial intelligence (AI), and predictive analytics. These methods leverage historical project data to identify complex relationships among cost drivers and continuously improve prediction performance.

2. Machine Learning and AI-Based Cost Prediction Models

Several studies have demonstrated the effectiveness of machine learning techniques in construction cost prediction. Yanfen Zhang et al. (2024) proposed an intelligent building cost prediction model by integrating Building Information Modeling (BIM) with an Elman neural network, achieving an accuracy of 95.83%. Their work highlights the potential of combining digital modeling with neural networks to support intelligent cost optimization and enterprise-

level digital transformation.

Similarly, Yanqin Wang et al. (2023) developed a recurrent neural network (RNN)-based cost prediction model that significantly reduced root mean square error (RMSE), demonstrating improved prediction accuracy despite limitations related to the absence of a comprehensive cost index system. Deep learning approaches such as convolutional neural networks (CNNs) have also been explored, with Xiaojuan Xue et al. (2020) applying CNNs to expressway cost estimation by incorporating bridge and tunnel factors, thereby enhancing early-stage financial planning.

Ensemble and hybrid models have shown superior performance in capturing nonlinear cost relationships. Chakraborty et al. (2020) identified a hybrid light gradient boosting and natural gradient boosting model as the most accurate among six ML algorithms tested. Likewise, Chien-Hsun Huang et al. (2020) combined Random Forest and linear regression models to improve BIM-based labor cost estimation accuracy using CRISP-DM methodology.

3. BIM, Big Data, and Digitalization in Cost Management

The integration of BIM, big data, and AI has emerged as a major research trend in construction cost management. Apeesada Sompolgrunk et al. (2024) examined strategic alignment between BIM and big data, identifying organizational and technological domains necessary for effective integration. Majed Alzara et al. (2023) demonstrated that BIM-based 5D models combined with genetic algorithms can reduce project time and cost by approximately 20%.

Big data–driven frameworks have also been explored beyond construction-specific contexts. Weisi Chen et al. (2023) discussed real-time analytics architectures and AI integration, emphasizing the importance of scalable infrastructure and streaming data for predictive decision-making. However, studies such as Nikolay Garyaev et al. (2019) noted that despite the availability of BIM, IoT, and cloud computing, construction data often remain underutilized for optimization purposes.

4. Time-Series Forecasting and Dynamic Cost Control

Time-dependent cost forecasting has gained increasing attention due to its importance in cash

flow management. Alberto De Marco et al. (2024) introduced a Holt–Winters time-series framework to capture trend and seasonality effects in project cost estimation under uncertain economic conditions. Tolga İnan et al. (2022) applied Long Short-Term Memory (LSTM) networks to forecast project costs and demonstrated superior performance compared to traditional Earned Value Management (EVM) indicators.

Julian Mæhlena et al. (2024) further highlighted the role of data-driven uncertainty analysis in reducing cost overruns, identifying correlations between project size and unit cost deviations. These studies emphasize the importance of dynamic, time-aware models rather than static, one-time cost predictions.

5. Risk, Uncertainty, and Cost Overrun Mitigation

Cost overruns remain a persistent issue in construction projects. Abroon Qazi et al. (2021) applied Monte Carlo simulation to prioritize sustainability-related risks, identifying labor productivity and scope definition as major contributors to cost escalation. Lisandra Seecharan et al. (2024) reviewed AI-based tools for mitigating cost overruns and highlighted limitations related to scalability, real-time data integration, and domain-specific knowledge.

Zhao Zeng et al. (2024) proposed a three-stage resource allocation method combining machine learning, fuzzy logic, and auction theory to predict both costs and delays, resulting in improved performance for large-scale projects. These studies demonstrate that cost estimation accuracy is closely linked to risk modeling and uncertainty management.

6. Research Gaps Identified

Although existing studies demonstrate the effectiveness of ML and AI techniques in construction cost estimation, several research gaps remain:

1. Most models focus on predicting total project cost, with limited attention to month-wise expenditure forecasting and dynamic cash flow planning.
2. Real-time adaptability and feedback-driven cost control mechanisms are often absent.
3. Budget categorization and decision-support features are rarely integrated into prediction frameworks.
4. Many studies rely on limited datasets or project-specific models, reducing generalizability.

5. There is a lack of holistic frameworks that combine prediction, monitoring, classification and managerial decision support.

The reviewed literature confirms the growing adoption of machine learning, deep learning, BIM, and big data technologies in construction cost estimation. While significant progress has been made in improving prediction accuracy, existing approaches remain fragmented and largely static. This study addresses these limitations by proposing an integrated predictive analytics framework that estimates total cost, forecasts monthly expenditures, and classifies budget levels, thereby supporting dynamic cost management and data-driven decision-making throughout the construction lifecycle.

RESEARCH METHODOLOGY

1. Problem Definition

Construction projects frequently suffer from cost overruns and inefficient financial management due to limitations in traditional cost estimation approaches. Conventional methods are largely static, dependent on expert judgment and historical averages, and fail to accommodate the dynamic, nonlinear, and multifactorial nature of modern construction environments. Moreover, existing approaches rarely provide detailed month-wise cost forecasts or automated budget classification, which are essential for effective cash flow management and proactive decision-making.

To address these limitations, this study proposes a machine learning–based predictive analytics framework capable of estimating total construction cost, forecasting monthly expenditure, and categorizing projects into budget tiers using historical project data.

2. Data Collection and Dataset Description

The proposed framework is developed using a structured dataset derived from historical building construction projects. Each data record consists of key project attributes, including building area, number of floors, material quality, labor cost, material cost, equipment cost, project duration, time overrun probability, project type, and location type. The dataset supports both regression and classification tasks, with target outputs defined as total construction cost, monthly cost estimates, and budget category (low, medium, or high).

3. Data Preprocessing and Feature Engineering

To ensure data quality and model reliability, a comprehensive preprocessing pipeline was applied. Missing values were handled using appropriate statistical imputation techniques, while numerical features were normalized or standardized to eliminate scale disparities. Categorical variables such as project type, location type, and material quality were transformed into numerical representations using encoding techniques.

Outliers were identified and treated using interquartile range (IQR) analysis and Z-score methods to prevent distortion of model learning. Feature engineering was employed to enhance predictive capability by generating interaction features, cost ratios, and derived indicators such as cost per unit area and complexity-related attributes. These engineered features enabled the models to capture hidden patterns and improve generalization performance.

4. Model Development

Supervised machine learning techniques were employed for predictive modeling. Random Forest and Gradient Boosting algorithms were selected due to their robustness, ability to handle nonlinear relationships, and strong performance on heterogeneous datasets.

Regression models were developed to predict total construction cost and monthly expenditure profiles, while a classification model was implemented to categorize projects into predefined budget tiers. Hyperparameter tuning was conducted using grid search combined with k-fold cross-validation to optimize model performance and prevent overfitting.

5. Model Evaluation

The performance of regression models was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics to assess prediction accuracy and variance explanation. Classification performance was measured using accuracy, precision, recall, F1-score, and confusion matrix analysis. These evaluation metrics provided a comprehensive assessment of model reliability and predictive effectiveness.

6. System Architecture and Deployment

To enhance real-world applicability, the trained models were deployed within a lightweight

web-based system developed using the Flask framework. The application allows users to input project-specific parameters and obtain real-time predictions for total cost, monthly expenditure, and budget category. Visualization components such as cost trend graphs and expenditure distributions were integrated to facilitate intuitive interpretation and informed decision-making.

7. Summary of Methodology

The proposed methodology integrates data preprocessing, feature engineering, ensemble machine learning models, and system deployment into an end-to-end predictive analytics framework. By combining robust predictive modeling with a user-accessible interface, the framework supports accurate construction cost estimation, phased financial planning, and effective budget control, thereby addressing key limitations of traditional cost management practices.

DATA COLLECTION AND ANALYSIS

The proposed predictive construction cost estimation framework is developed using a structured dataset obtained from historical building construction projects. The dataset was curated to capture the key technical, financial, and project-related parameters that significantly influence construction cost and budget allocation.

1. Data Sources

The primary data source consists of historical construction project records collected from completed residential and commercial building projects. These records include both quantitative and categorical attributes relevant to cost estimation and financial planning. In addition, user-provided project inputs are collected dynamically through the system interface to enable real-time prediction.

2. Data Attributes and Description

Each project instance in the dataset is represented by a set of input features and output variables. The selected attributes were identified based on domain relevance and prior studies in construction cost modelling.

Input Features: Building area, Number of floors, Project duration, Material quality/type, Labor cost, Material cost, Equipment cost, Project type, Location type

Target Variables: Total construction cost**,** Monthly cost distribution**,** Budget category (low, medium, high)

These attributes enable both regression-based cost prediction and classification-based budget categorization.

3. Data Collection Workflow

The overall data collection workflow follows a structured process. Initially, project-specific parameters are entered by the user (Project Manager) through the system interface. Simultaneously, historical construction data are retrieved from the centralized database and used for model training and validation. The system integrates both real-time user inputs and historical data to ensure context-aware predictions.

The hierarchical decomposition of this workflow is represented through:

- **Level 0 DFD**: High-level interaction between the user, predictive system, and historical data repository.
- **Level 1 DFD**: Modular breakdown into data input, preprocessing, machine learning, and output modules.
- **Level 2 DFD**: Detailed internal processes of the machine learning module, including feature extraction, model selection, training, and prediction.

4. Data Storage and Management

Two primary data repositories are maintained:

a) **Historical Construction Database** – Stores cleaned and validated historical project data used for model training.

b) **Project Feature Database** – Temporarily stores user-input project parameters during prediction.

Trained models and configuration files are preserved in a Trained Model Repository to enable efficient reuse without repeated training.

5. Data Integrity and Quality Assurance

To ensure reliability, the collected data undergo validation checks for completeness,

consistency, and logical correctness. Erroneous or missing entries are either corrected using statistical techniques or excluded from model training. This quality control process improves model robustness and reduces prediction bias.

6. Summary of Data Collection Process

The data collection process integrates historical construction records with real-time user inputs to support accurate and scalable cost prediction. The structured flow of data, supported by well-defined system architecture and repositories, ensures efficient handling, traceability, and reproducibility of results. This systematic data collection approach forms a strong foundation for subsequent data preprocessing, model training, and performance evaluation stages.

RESULTS AND DISCUSSION

This section presents the performance outcomes of the proposed machine learning–based construction cost prediction framework and discusses their practical implications for construction cost management. The results are analyzed for both regression and classification tasks to evaluate the effectiveness, robustness, and real-world applicability of the developed models.

1. Performance of Cost Prediction Models

The Random Forest–based multi-output regression model demonstrated strong predictive accuracy in estimating both total construction cost and monthly cost expenditure. The model achieved very high R^2 values (0.999 for total cost and 0.993 for monthly cost estimates), indicating that the selected project parameters effectively explain variations in construction costs.

The low Mean Squared Error (MSE) values confirm minimal deviation between actual and predicted values, highlighting the model's ability to capture complex, non-linear relationships among cost-driving factors such as building area, material cost, labor cost, equipment usage, and project duration. These results validate the suitability of ensemble learning techniques for construction cost estimation, where cost behavior is inherently multidimensional and uncertain.

Compared to traditional estimation approaches that rely on static rates and expert judgment, the proposed model provides data-driven adaptability, making it more resilient to variations in project scale, material quality, and site conditions.

2. Interpretation of Regression Results

Visual analysis using actual-versus-predicted scatter plots further reinforces the numerical evaluation. The clustering of data points along the ideal reference line confirms strong agreement between predicted and observed values across both small- and large-scale projects. The limited presence of outliers indicates stable generalization and robustness against noise in real-world construction data.

Importantly, the model's ability to simultaneously predict total cost and monthly expenditure offers a significant advantage over existing models that focus solely on lump-sum estimates. Monthly cost forecasting enables improved cash-flow planning, phased fund allocation, and early detection of potential budget stress, which are critical for effective project financial control.

3. Budget Classification Results and Implications

The Random Forest classification model was employed to categorize projects into budget levels (under-budget, on budget, and over-budget). The model achieved high overall accuracy, with particularly strong recall for the on-budget class. This indicates that the classifier effectively identifies financially stable projects, which is valuable for routine project monitoring.

However, the comparatively weaker performance for minority classes (under-budget and over-budget) suggests class imbalance within the dataset. This highlights an important insight: while machine-learning models can effectively support budget categorization, balanced and representative datasets are essential to improve sensitivity toward cost deviation risks. Addressing this limitation through resampling techniques or cost-sensitive learning can further enhance early-warning capabilities in future implementations.

Despite this limitation, the classification results still demonstrate the model's potential as a decision-support tool, enabling stakeholders to identify financially critical projects at early

planning stages.

4. Practical Significance for Construction Cost Management

The combined regression–classification framework provides a holistic solution for construction cost management by integrating prediction accuracy with interpretability and usability. The inclusion of contextual variables such as project type, location type, and time overrun probability significantly improves the model's adaptability across diverse construction scenarios.

The deployment of the model through a Flask-based graphical user interface further enhances its practical relevance. By allowing users to input project parameters and receive real-time predictions, the system bridges the gap between advanced analytics and on-site decision-making. This transforms cost estimation from a static planning activity into a dynamic, data-driven process.

5. Comparison with Existing Studies

Unlike many prior studies that focus exclusively on total cost prediction, the proposed framework uniquely integrates multi-output regression, budget classification, and real-time deployment within a unified system. The high predictive performance achieved in this study compares favorably with existing machine learning–based cost estimation models, while offering additional value through monthly cost forecasting and budget categorization.

6. Findings

- Ensemble machine learning models effectively capture non-linear cost relationships in construction projects.
- Simultaneous prediction of total and monthly costs enhances financial planning accuracy.
- Budget classification provides early financial risk indicators for stakeholders.
- Integration with a web-based interface significantly improves usability and real-world adoption.

CONCLUSIONS

This study presents an effective machine learning–driven framework for construction cost management, demonstrating the strong potential of data-driven predictive analytics in

addressing long-standing challenges of cost uncertainty and budget overruns in the construction industry. By employing ensemble learning techniques—specifically Random Forest and Gradient Boosting algorithms—the proposed system accurately predicts total construction cost, monthly expenditure patterns, and budget categorization using historical project data.

The high predictive performance achieved by the regression models confirms their capability to capture complex, non-linear relationships among critical project parameters such as building area, material quality, labor and equipment costs, and project duration. The inclusion of monthly cost estimation provides an added practical advantage by enabling improved cash-flow planning and phased financial control, which are often overlooked in conventional cost estimation approaches.

Furthermore, the budget classification component enhances strategic decision-making by offering early insights into potential budget deviations. The integrated framework surpasses traditional estimation methods by reducing dependency on manual judgment, minimizing human error, and improving adaptability across varying project conditions. The successful deployment of the model through a user-friendly interface further emphasizes its real-world applicability and usability for construction professionals.

While the framework demonstrates strong performance, its effectiveness is influenced by data quality and feature availability, highlighting the importance of comprehensive and updated datasets for large-scale adoption. Despite these limitations, the findings clearly indicate that machine learning-based cost estimation systems can significantly enhance planning accuracy, financial transparency, and resource optimization in construction projects.

Overall, this research contributes to the growing body of knowledge on intelligent construction management systems and establishes a scalable foundation for future integration with advanced technologies such as BIM, GIS, and real-time data analytics. The proposed approach represents a meaningful step toward smarter, more resilient, and data-driven cost management practices aligned with the demands of modern construction engineering.

REFERENCES

1. Yanfen Zhang et al, “Intelligent building construction cost optimization and prediction by integrating BIM and elman neural network” Heliyon 2024.
2. Temitope Omotayo et al., “Generative AI for BIM-based Digital Construction Cost Management: A Qualitative Sentiment Analysis Approach” IEOM 2024
3. ZHAO ZENG et al., “Cost Control Management of Construction Projects Based on Fuzzy Logic and Auction Theory” IEEE ACCSEE 2024. VOLUME 12,
4. Apeesada Sompolgrunk et al., “Strategic alignment of BIM and big data through systematic analysis and model development” Automation in Construction, ELSEVIER 2024.
5. Yanfen Zhang et al., “Intelligent building construction cost optimization and prediction by integrating BIM and elman neural network” Heliyon 2024. https://doi.org/10.1016/j.heliyon.2024.e37525
6. Alberto De Marcoa et al., “Time series-based Project Cost Forecasting Framework” ScienceDirect, ELSEVIER 2024.
7. Julian Mæhlena et al., “Reducing cost overruns through data-driven methods used in uncertainty analyses” ScienceDirect, ELSEVIER 2024.
8. Lisandra Seecharan et al., “Artificial Intelligence (Ai)Tools for Cost Overruns on Construction Projects” International Journal of Communication Networks and Information Security 2024.
9. Kudirat Ayinla et al., “The impact of artificial intelligence on construction costing practice” Abdullahi Saka 2023.
10. WEISI CHEN et al., “Real-Time Analytics: Concepts, Architectures, and ML/AI Considerations” IEEE ACCESS 2023. VOLUME 11.
11. Yanqin Wang et la., “Research on Construction Project Cost Prediction Model Based on Recurrent Neural Network” SHS Web of Conferences, 2023. https://doi.org/10.1051/shsconf/202317002009
12. MAJED ALZARA et al., “Building a Genetic Algorithm-Based and BIM-Based 5D Time and Cost Optimization Model” IEEE ACCSEE, 2023 VOLUME 11,
13. Shanaka Kristombu Baduge et al., “Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications” Automation in Construction, ELSEVIER 2022. https://doi.org/10.1016/j.autcon.2022.104440

14. Tolga İnan et al., “A Machine Learning Study to Enhance Project Cost Forecasting” ScienceDirect, ELSEVIER 2022.
15. RAKESH GUPTA et al., “AEHO: Apriori-Based Optimized Model for Building Construction to Time-Cost Tradeoff Modeling” IEEE ACCESS, 2022.
16. Abroon Qazi et al., “Prioritizing risks in sustainable construction projects using a risk matrix-based Monte Carlo Simulation approach” Sustainable Cities and Society, ELSEVIER 2021.
17. Haytham H. Elmousalami et al, “Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review” Haytham Elmousalami 2021.
18. Frank Bodendo et al., “A machine learning approach to estimate product costs in the early product design phase: a use case from the automotive industry” ScienceDirect, ELSEVIER 2021.
19. Debaditya Chakraborty et al., “A novel construction cost prediction model using hybrid natural and light gradient boosting” Advanced Engineering Informatics, ELSEVIER 2020. https://doi.org/10.1016/j.aei.2020.101201
20. XIAOJUAN XUE et al., “Expressway Project Cost Estimation with a Convolutional Neural Network Model” IEEE ACCESS 2020. VOLUME 8,
21. Sanaz Tayefeh Hashemi et al ., “Cost estimation and prediction in construction projects: a systematic review on machine learning techniques” SN Applied Sciences 2020. https://doi.org/10.1007/s42452-020-03497-1
22. Chien-Hsun Huang et al., “Predicting BIM labor cost with random forest and simple linear regression” Automation in Construction, ELSEVIER 2020. https://doi.org/10.1016/j.autcon.2020.103280
23. Sanaz Tayefeh Hashemi et al., “Cost estimation and prediction in construction projects: a systematic review on machine learning techniques” SN Applied Sciences 2020. https://doi.org/10.1007/s42452-020-03497-1
24. Sidharan Gurappa Manakoji et al., “STUDY OF TIME AND COST MANAGEMENTS IN HIGHWAY PROJECTS” IJCRT 2020. Volume 8,
25. Jasmine Ngo et al., “Big Data and Predictive Analytics in the Construction Industry Applications, Status Quo, and Potential in Singapore’s Construction Industry” Construction Research Congress 2020
26. Structural Engineering Dep et al., “Estimation and prediction of construction cost index using neural networks, time series, and regression” Alexandria Engineering

Journal ELSEVIER 2019.

27. Nikolay Garyaev et al., “Big data technology in construction” EDP Sciences 2019.
28. Dr. V. K. Divakar et la., “Factors Affecting Effective Implementation of Cost Management Process in Construction Industry” International Research Journal of Engineering and Technology (IRJET) 2018.Volume: 05