

Construction Cost Estimation Using Machine Learning

¹Mr. Jaydeep Madhukar Jadhav, ²Dr. J. M. Shinde*

¹PG Student, ²Head of Civil Engineering Department

^{1,2}Ashokrao Mane Group of Institutions Vathar tarf Vadgaon, Maharashtra, India

*Corresponding Author: Mr. Jaydeep Madhukar Jadhav

Abstract

The estimation of construction costs is a very important subject in the project planning, budgeting, and financial decision making of a civil engineering project. The conventional methods of estimation, however, are mainly based on the information of the historical cost, manual calculation and intuition, which tend to create inaccuracies as construction activity is very dynamic and complicated. To overcome this obstacle, this research paper suggests a machine learning-based architecture of automated and precise building cost forecasting. A detailed dataset as a compilation of 10 000 records of construction projects was gathered using various sources such as records of contractors, government rate schedules as well as estimates of real-estate projects. Data cleaning, data imputation, feature scaling, and correlation-based feature selection were the techniques used to preprocess the dataset to enhance the quality of data and the performance of the model. Several regression models were run such as Linear regression, Decision tree, Random forest, Gradient Boosting and XGBoost as a MultiOutputRegressor to predict at once the quantity of materials, costs of materials and the total cost of construction. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2) were used to estimate the performance of the model. The outcome of the experiment showed that XGBoost, had the highest predictive power with a corresponding R^2 of approximately 0.9999, compared to the random forest (0.9998), among others. Also, the trained model was implemented in the form of a web application written in Flask with the help of which one can enter the parameters of the project and predict the costs in real time. These results indicate that ensemble machine learning models can also significantly increase accuracy and efficiency of construction cost estimation and offer a viable decision-support tool to engineers, contractors, and project planners.

Keywords: Construction Cost Estimation, Machine Learning, Multi-Output Regression, XGBoost, Ensemble Learning, Predictive Modeling, Civil Engineering Applications.

1 Introduction

Construction cost estimation can be considered an essential part of construction project management because it has a direct influence on the planning, budgeting, and financial performance of a construction project. Estimation of construction cost has always been a complex task because of its dynamic nature. The ever-changing nature of construction projects can be explained by various factors, including the fluctuation of materials, inflation rates, labor costs, and environmental conditions. All these factors have a significant influence on construction costs, and it has been observed that the estimated cost of a construction project may differ from the actual cost incurred (Rayabharapuet al., 2025). Conventional cost estimation techniques have limitations in estimating construction costs because they mostly depend on historical data and deterministic calculations. The conventional cost estimation techniques do not have the ability to handle the complex relationships between various cost-influencing factors (Shi & Li, 2010). The limitations of conventional cost estimation techniques have a direct influence on the estimation errors of construction cost. The errors associated with the conventional cost estimation techniques need to be reduced to a minimum to save time for preparation.

Recent years have seen the application of advanced computational techniques and machine learning algorithms to improve traditional and software-based construction cost estimation methods (Ali et al., 2022; Ameya, 2025; G.C.Sarode, 2020).. Predictive

models have been developed for the purpose of improving the accuracy and efficiency of cost predictions in construction projects. Enhanced Support Vector Machines (SVM) optimized by Particle Swarm Optimization (PSO) have provided a promising performance with an average prediction deviation as low as 1.57% (Zhang & Song, 2022). Back Propagation (BP) neural networks that minimize mean square errors through error gradient descent mechanisms have forecasted construction costs with high accuracy successfully utilized (Wang, 2018). The combination of Building Information Modeling (BIM) with Elman Neural Networks (ENN) increases prediction accuracy by using digital and visual data from intelligent building models, achieving a root mean squared error (RMSE) below 75 and a determination coefficient above 0.95 (Zhang & Mo, 2023). Also, Bird Swarm Algorithm-optimized Random Forest algorithms have better predictive performance in tasks related to predicting construction costs where the maximum relative error was reported as low as 1.24% (Zheng et al., n.d.). Neural networks, case-based reasoning, adaptive neuro-fuzzy inference systems, and extreme gradient boosting are examples of artificial intelligence (AI) methods that have been widely investigated to overcome challenges presented by classical estimation approaches. These techniques help to accomplish two objectives. First, they are able to identify nonlinear relationships between different variables, and second, they can handle large quantities of data from the construction industry. Research has shown that AI-based models are capable of producing greater accuracy in predictions ascribed to traditional methods, such as earned value management based approaches (Yalcin et al., 2024). In addition, building information modelling (BIM) based cost estimation methods produce more accurate and reliable results than traditional methods, such as just using square footage calculations. BIM-based cost estimation methods capture more of the variables associated with the project (e.g., the value of the material used and the structural components) than do traditional approaches (Chandra & Yuliana, 2024; Yang et al., 2022). Recent studies have incorporated quantum-computing techniques to increase efficiency and accuracy among traditional machine learning models, such as Random Forest, for estimating costs of construction (Lakshana et al., 2024). The existence of these technologies indicates that data accumulations can be used to improve cost estimating systems in the construction sector.

Despite these advances, certain limitations have been identified in the existing research and practical applications. For instance, most of the existing models have used a single machine learning algorithm, which may be efficient for a particular situation but may not be able to generalize well for various construction problems. The construction data may have heterogeneous variables, non-linear relationships, and variable data distributions. Therefore, it may be difficult for a predictive model to achieve high accuracy for all these scenarios. Furthermore, certain studies have focused on the improvement of individual algorithms but have not considered the hybridization of multiple predictive models, which could be useful for better efficiency. The absence of a hybrid framework that can be used to combine the efficiency of multiple algorithms is a significant gap in the research of construction cost estimation. Furthermore, very little attention has been paid to the development of a system that can not only reduce errors but also improve the efficiency of decision-making for construction managers.

To overcome these limitations, this study proposes the design and implementation of a multi-regressor construction cost estimation system that integrates several machine learning algorithms and optimization techniques. The proposed method intends to use the strengths of different regression models so as to improve predictive accuracy and, therefore, enhance the robustness of cost estimation. By combining multiple predictive models within one framework, it can better capture complex relationships among variables related to costs and minimize errors in estimations. The multiregressor

approach also allows for comparative evaluation as well as ensemble learning strategies which increase generalization capability over various types of construction projects.

The main goals of this research are to create an intelligent system for predicting construction costs based on a multi-regressor framework, assess how well different machine learning algorithms work for cost estimation, and study how effective the integrated method is in lowering prediction errors. The study also aims to make the cost estimation process more efficient, cut down on unnecessary computations, and offer a more dependable decision-support tool for construction professionals.

The key contributions of this research include the development of an integrated multi-regressor framework for construction cost estimation, the comparative evaluation of different machine learning algorithms within a unified system, and the demonstration of improved prediction accuracy compared with traditional single-model approaches. By addressing the limitations of existing estimation techniques and introducing a hybrid predictive framework, this study contributes to the advancement of intelligent cost estimation systems that support more accurate budgeting, improved project planning, and better decision-making in construction management.

2 Literature Review

Before starting the research work, a thorough literature review as done to understand the currently used techniques. A number of techniques for construction cost estimation were studied and analysed the important ones are discussed in this chapter.

Zheng et al., (2023) presented a prediction model of construction cost based on the Random Forest which is optimized with Bird Swarm Algorithm (BSA). The traditional models were prone to limited levels of accuracy and inability to deal with complex construction data. To solve this problem, BSA was combined with Random Forest to maximize the features and prediction ability. The hybrid method enhanced forecasting of projects at the enterprise level. The findings indicated high accuracy in prediction using conventional algorithms as results were better than the conventional algorithms with a maximum relative error of 1.24% .

Magdum et al., (2017) developed a cost prediction model that predicts costs by utilizing neural networks and multilayer perceptrons (MLP) with reference to six major features of materials. The conventional models had difficulties in revealing nonlinear association between construction data. To counter this shortcoming, deep learning structures were used to learn complicated cost forms. The way it was done was by comparison of network structures and functions of activation in improving performance. The findings demonstrated that, MLP was better than neural networks in training data whereas the generalization was performed better in the neural networks with ELU activation function being the best.

Xu et al., (2023) suggested a hybrid building cost prediction model which employs t-SNE dimensionality reduction and enhanced grey correlation algorithm. The classical methods were not able to process high dimensional data and extract meaningful features. The model used t-SNE to reduce the complexity of features and enhance feature correlation analysis to solve this problem. This composite structure increased the speed of computation and accuracy of prediction. The findings showed that the accuracy of prediction had risen by 5.1 percent, and the efficiency with which it is computed had risen by 12.75 percent, over the conventional cost estimation methods.

Ye et al., (2021) designed a model based on Particle Swarm Optimization (PSO) and a Back Propagation (BP) neural network that would be used to forecast the cost. Convergence and local minima were common problems associated with traditional BP networks. These issues were solved by optimizing network weights with the help of PSO and enhancing the efficiency of learning. The optimization contributed to the improved

stability and the training of the neural network. The findings indicated that convergence speed and accuracy of the forecasting were significantly much better than the traditional BP neural network models.

Tayefeh Hashemi et al., (2020) suggested a systematic review method to examine machine learning models in estimation of construction costs in the last 30 years. The past research did not have systematic knowledge of various predictive methods. In order to fill this gap, the authors classified some of the models and assessed their strengths and weaknesses. The paper has highlighted the relevance of hybrid machine learning methods in dealing with uncertainty. The findings suggested the use of integrated and hybrid models in the establishment of construction where cost estimation could be improved in complex projects.

Zhang et al., (2024) suggested a model of cost predictions combining Building Information Modeling (BIM) and Elman Neural Networks (ENN). The classic estimation techniques did not have the capability of handling dynamic information about projects. To address this issue, digital data of BIM were integrated with the ability of ENN to operate on time-sequential data. This integration resulted in improved prediction of the system. The findings revealed that more than 95% of predictions were correct, and the BIM-ENN framework was able to process data associated to time-dependent construction.

Jirait et al., (2024) suggested the use of machine learning on a GUI-based machine learning system that estimates the cost of construction of houses. The conventional estimation applications were not user friendly and did not have easily accessible interfaces. To address this problem, the researchers have created a graphical user interface that is combined with machine learning models. The system estimated cost effectively using various features like number of floors and location. The findings showed good prediction results and the future opportunities of implementing machine learning models and user-friendly applications.

Park and Yun (2023) suggested a schematic cost prediction model that is based on deep learning by utilizing the BIM capabilities, floor perimeter, and wall area. Earlier methods of estimation were based on past information about the project and hence only on little information the prediction was done thus leading to a wrong estimation. In order to overcome this drawback, not only design features but also building features were taken into consideration in the model. This incorporation enhanced the project description capacity in the model. The findings revealed much higher early stage prediction cost accuracy as opposed to methods with restricted features.

Yaseen et al., (2018) forecasted the cost of construction and the duration of the project using a combination of genetic algorithm and artificial neural networks and support vectors machines. The researchers solved the issue that the traditional model of prediction usually demonstrated poor accuracy because of poor selection of features and poor performance at training. In order to solve this problem, genetic algorithms were used to optimize the features, which improved the learning performance of ANN and SVM types of models. The findings revealed that the hybrid model increased the quality of training and the accuracy in prediction and it was more effective in estimating the cost of constructing and the duration of a project.

Liu et al., (2023) came up with a model to predict engineering costs. The research solved this issue because the conventional BP neural networks were usually associated with slow convergence and reduced prediction accuracy because of the inability to optimize parameters. To eliminate this problem, the researchers combined genetic algorithms to optimize the weights and the parameters of the BP neural network, thus enhancing its learning performance and prediction. The findings indicated that the GA-BP hybrid

model had much more superior convergence rate, and forecasting accuracy than the traditional BP neural network models with no evolutionary optimization.

Aung et al., (2023) suggested a machine learning approach to cost overruns prediction in construction projects with the help of an extended list of project parameters. The use of traditional cost estimation approaches could not effectively predict overruns because of the limited chances they have in managing complex and dynamic project variables. To overcome this, machine learning algorithms were used in the analysis of large datasets and defining patterns that affect cost deviations. The suggested information-based strategy enhanced the accuracy of cost estimation and control. It was demonstrated that machine learning models demonstrated a much higher predictive accuracy than traditional estimation techniques allowing to plan better, eliminate risks, and make better financial decisions in construction projects.

Das et al., (2024) suggested a machine learning model based on machine learning, a construction cost prediction methodology of Artificial Neural Network (ANN), Natural Gradient Boosting (NGBoost) and Linear Regression (LR). The conventional methods of estimation were usually poor at estimating the cost of construction projects efficiently in the initial phases of the construction process because the parameters affecting it are complex. In order to overcome this drawback, the NGBoost model was used and compared to ANN and LR in terms of evaluation indices including MAE, MAPE, CE, index of agreement, and R^2 . The findings reflected that NGBoost was better than the rest of the models in terms of $R^2 = 0.992$ (training) and 0.985 (testing) and minimum RMSE values whereas sensitivity analysis established the most significant factor to be formwork (41%).

Lee et al., (2024) suggested a machine learning based approach that could enhance the construction cost prediction with a range of missing value replacements methods and conditional outlier deletion. The missing values and non-uniform data were typical of construction cost datasets that decreased the accuracy and reliability of classical prediction models. To solve this problem, various single and multiple imputation approaches were compared and a verdict was arrived at on the best approach to use in case of missing data. The performance of prediction was then evaluated using machine learning models. The findings showed that median imputation worked optimally on single method, whereas lasso regression imputation had the highest results on multiple methods of imputation.

Draz et al., (2024) suggested a hybrid model of cost estimation where Convolutional Neural Networks (CNN) and Particle Swarm Optimization (PSO) algorithm are used to forecast costs of software using time series. Conventional estimation schemes tended to make erroneous and volatile forecasts and were very manual in the parameter tuning. To overcome these shortcomings, CNN model was combined with PSO so as to allow automatic extraction of features and hyperparametric optimization to enhance robustness and generalization ability. The model was trained and tested on 13 benchmark datasets and tested using a variety of metrics. The findings showed best performance as opposed to the current models in all the assessment criteria.

3. Research Methodology

The construction cost estimation system evolution occurred in four stages, namely data collection and curation, data preprocessing and validation, model development and training, and Flask-based deployment. Other operations like feature selection, hyperparameter optimization, cross-validation and bias management were also added to guarantee the role of robustness, scalability, and generalization capability of the suggested system. Each of the phases is described below.

3.1 Dataset Collection and Curation

In order to build a dependable prediction system a large dataset of 10,000 instances was present together where different sources were used. These sources consisted of tender websites by the government, records of the construction contractors, estimates of the real estate project in the public and market rate sheets of construction materials. The main objective was to make a broad scope of project types, geographical locations, and grade of materials to be covered in order to be able to generalize the model. Some of the aspects that were captured in every data entry were the type of the location (urban, semi-urban and rural), the amount of built-up area in square feet, quality measure of construction elements (cement, steel, bricks, flooring), and the anticipated time within which the project will be completed.

The desired products were estimated values and cost of each construction material and the final cost of construction. This step involved data validation processes to eliminate inconsistencies, duplicate records and unrealistic cost values. Reliability was ensured by statistical checks and domain checking by the available construction cost standards. To reduce the nature of bias in the data, there was a balance of the dataset in terms of project sizes, quality grades, and geographic parameters to ensure the trained model does not incline towards a specific project type or region. This selected set of data served as a robust basis in developing a machine learning model which could predict detailed construction cost estimates in a variety of project situations.

3.2 Data Preprocessing

The dataset was preprocessed before model training in order to enhance quality and compatibility. Ordinal variables were coded in terms of numbers and missing values were done away with by imputing or through exclusion. The feature scaling made the numerical attributes to be normalized and the feature selection eliminated the irrelevant variables. The data was categorized into input variables and target outcomes and divided into 80 percent training and 20 percent testing data.

- **Data Cleaning and Validation**

Before using machine learning methods to analyze the dataset, the collected data was thoroughly checked to make sure that it was of good quality and reliability. It was found that duplicate records, inconsistency of records, and unrealistic cost values were eliminated to avoid misleading predictions. This move assisted in clearing unnecessary mistakes that would have adversely affected the learning of the model. Moreover, the data validation processes were also implemented to ensure that all the construction records had logical values and uniform formats. The cleaning and validation of the dataset enhanced the integrity of the whole dataset such that the model was trained on the correct and significant data.

- **Handling Missing Values (Imputation)**

Statistical imputation methods were used to fill areas that had no values so that the dataset would be complete. The methods of mean and median imputation were used based on the data distribution. The missing values in a feature are replaced by a mean of the available values in a feature using mean imputation. This method should be used in case data distribution is comparatively normal and will not have extreme values.

i. **Mean Imputation Formula:**

$$X_i = \frac{1}{n} \sum_{j=1}^n X_j$$

X_i is the missing value which has been replaced and X_j is the available values and nnn is the number of observations.

In cases where there were outliers or skewed values in the dataset, median imputation was taken since median is not influenced by the extreme values.

ii. Median Imputation:

$$X_i = \text{Median}(X_1, X_2, \dots, X_n)$$

• **Categorical Data Encoding (Label Encoding)**

Other variables in the dataset were categorical like the type of material (Basic, Medium, Premium). Given that machine learning models need numerical values to be fed, these categorical values were transformed into numerical values with the help of label encoding. Under this approach, all categories were given a distinct numeric data, and the algorithm was able to work with them.

• **Feature Scaling (Normalization)**

The data were numerical variables with various ranges including, built-up area and project duration. Normalization was done to ensure every feature makes a contribution during model training. Min-max normalization method brought the values between 0 and 1 in order to enhance the performance of the models and training stability. Min-Max Normalization Formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

with X being the original value, X_{min} is the minimum value and X_{max} is the maximum value of the feature.

• **Feature Selection (Correlation Analysis)**

The process of feature selection was conducted with the aim of finding out the most significant variables that have an impact on construction cost. The relationship between the input variables and the target variable was measured by correlation analysis. Data variables that had weak or overlap correlation were eliminated to make the data simpler and enhance the efficiency of the model. Correlation Coefficient Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where r is the correlation between variables x and y .

• **Dataset Splitting (Train-Test Split)**

The data was preprocessed and separated into input features and target outputs. To test the predictive capacity of the model, the data was divided into training and testing set. The model was constructed using the training set and performance evaluated using the testing set. Train-Test Split Representation: $D = D_{train} \cup D_{test}$

where $|D_{train}| = 0.8D$, $|D_{test}| = 0.2D$

This implies that it trained 80% of the data and tested 20 percent of the data in order to ensure a good assessment of the model.

3.3 Model Development and Training

In the case of predictive modeling, we used several machine learning algorithms as a multi-target regression task MultiOutputRegressor wrapper to implement the algorithm. The models that were evaluated were Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor, Decision Tree Regressor and Linear Regression. These algorithms were chosen because they have been effective in regression based

prediction activities as well as the fact that they are in a position to identify nonlinear relationships among complex data.

In order to improve the performance of the models, hyper parameter tuning was carried out with grid search and randomized search methods. The important parameters including number of trees, learning rate, max depth and regularization terms have been optimized to enhance the accuracy of prediction. Furthermore, the cross-validation with k-fold was used in training so that the training would be robust and the overfitting can be avoided. The method enabled the model to be trained and evaluated on several smaller sets of the data, which enhanced reliability and generalization.

Regarded metrics of regression performance, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and coefficient of determination (R^2 score) were used to model evaluation. Such metrics comparatively measured prediction accuracy in various models on a quantitative basis. The XGBoost Regressor was the most successful in the performance assessment, as it showed the lowest MAE and RMSE values and the highest R^2 value. Another factor that was observed is that this model did not vary in terms of predictability and confidence in the cross-validation, meaning it was very stable and robust in most partitions of the data.

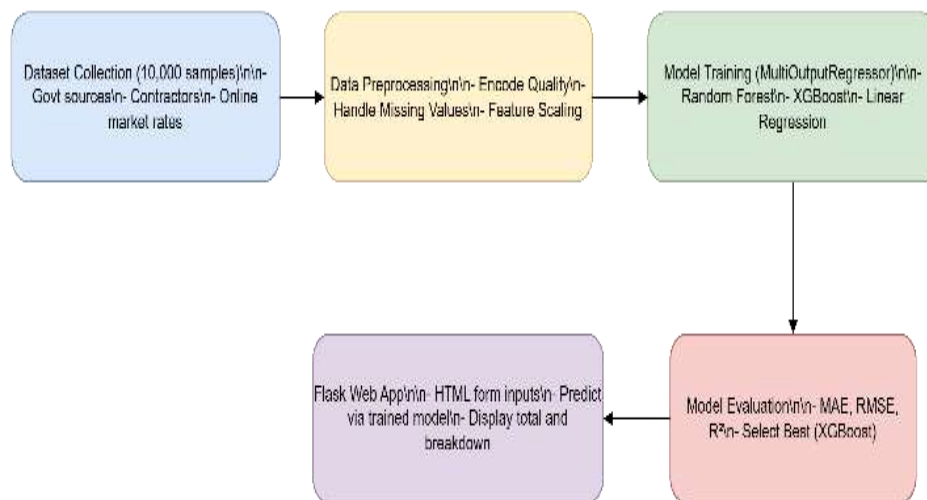
3.4 Deployment via Flask Application

A web-based interface based on the Flask framework was built to make the developed system available to the non-technical people like civil engineers, contractors, and project planners. The application is secure and gives the user the option to enter the construction project parameters with easy dropdown menus and text entries. The variables that can be specified by users include the category of the project location, built-up area and material quality levels.

When the application is submitted, the application sends the input to the trained XGBoost model to produce the predictions of the material quantities, materials costs separately and the construction cost in general. The outcomes are presented in real time via web interface, allowing users to do quick cost evaluation of various construction situations.

The architecture diagram of the flow of the project is as shown below in Figure 1 below:

Fig.1 Architecture Diagram of the System



To define what the system architecture entails, one can start with collection of data sets, in which a diverse collection of 10,000 construction cost records was put together by a number of sources such as government rate schedules, contractor cost sheets and market price databases. This guaranteed that the dataset reflected realistic changes in prices of materials, regions and quality categories.

After collecting the data, the dataset was preprocessed and validated to provide structural consistency and remove noise or biased entries. The method of feature engineering and selection was applied to determine the most applicable variables that had an impact on construction costs. The derived data was then incorporated in the training of the model whereby the multi-regression algorithms were adopted in a multi-output model.

In the evaluation of the models, testing of the trained models was done on the predetermined evaluation measures and cross validation processes. XGBoost model was the most accurate and strong model and hence the final model to be deployed. To examine scalability and generalization the model was run on a variety of different subsets of the data that represented a range of project sizes and location categories. The accuracy of the predictions between these subsets was consistent and this implied a high generalization ability.

The last phase was deployment via the Flask web application whereby users can enter the project specifications and get immediate cost estimates. The system produces a comprehensive output in terms of material quantities, cost involved and the cost of construction estimate. Such a deployment will make sure that the predictive system is fully applicable by the builders, contractors, civil engineers, and project planners to efficiently budget and make decisions made in construction projects.

3.5 Evaluation matrix

- **Mean Absolute Error (MAE)**

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

- **Mean Square Error (MSE)**

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

- **Root Mean Squared Error (RMSE)**

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

- **Mean Absolute Percentage Error (MAPE)**

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

- **Coefficient of Determination (R^2 Score)**

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

- **Adjusted R^2 Score**

$$Adjusted R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1} \right)$$

- **Explained Variance Score**

$$EVS = 1 - \frac{Var(y - \hat{y})}{Var(y)}$$

4. Results and Discussion

The dataset of 10,000 records of construction projects underwent a thorough exploratory analysis and predictive modeling to determine the usefulness of machine learning algorithms in estimating construction costs. The data had variables that were SquareFeet, CementQuantity, SteelQuantity, CementCost, SteelCost, and TotalCost, and categorical variables of project location and material grade. Exploratory analysis and model training were carried out to learn about the correlation between these factors and how they predict.

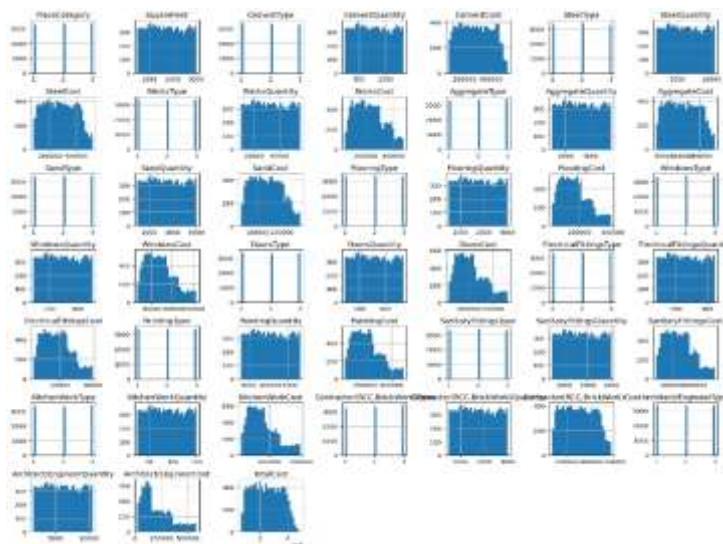
4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to see the distribution and correlation among the variables to train the machine learning models. The features in the dataset were analyzed using univariate analysis to visualize the distribution of a feature and identify anomalies.

Distribution of Place category variable showed a relatively equal representation of construction projects in three categories of; Expensive, Medium, and Economical zone. Balanced representation of such kind guarantees that the trained models are not biased to a particular regional cost pattern.

The positively skewed distribution was observed in histograms of numerical variables, including SquareFeet, CementCost, SteelCost, and TotalCost, and this is natural in the construction datasets because there is a smaller number of large-scale projects add greatly to the upper limits of costs. The findings of the univariate analysis are demonstrated in Figure 2.

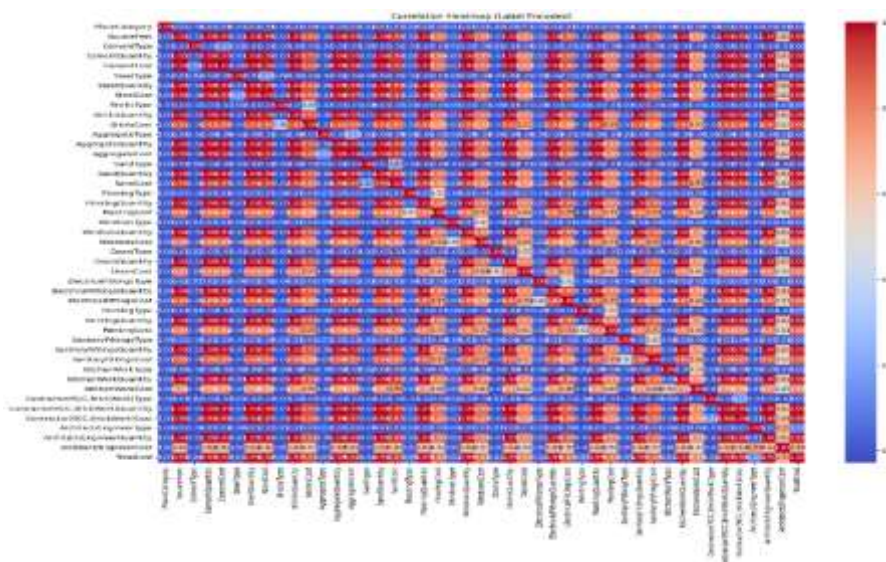
Fig.2 Univariate Analysis of Dataset



After the univariate analysis, a correlation table was created to test the relation among the variables. Figure 3 below is a correlation heatmap, which indicated that there existed strong positive relationships between the amount of materials and the costs of the materials. The Steel Quantity exhibited a good correlation with SteelCost and

Cement Quantity with CementCost. These relationships affirm that the data base is realistic in construction pricing patterns.

Fig.3 Correlation Matrix



Also, the TotalCost variable showed high correlation with other important variables of prediction like, SquareFeet, SteelCost and CementCost, which indicated that these variables are predominant in overall construction spending.

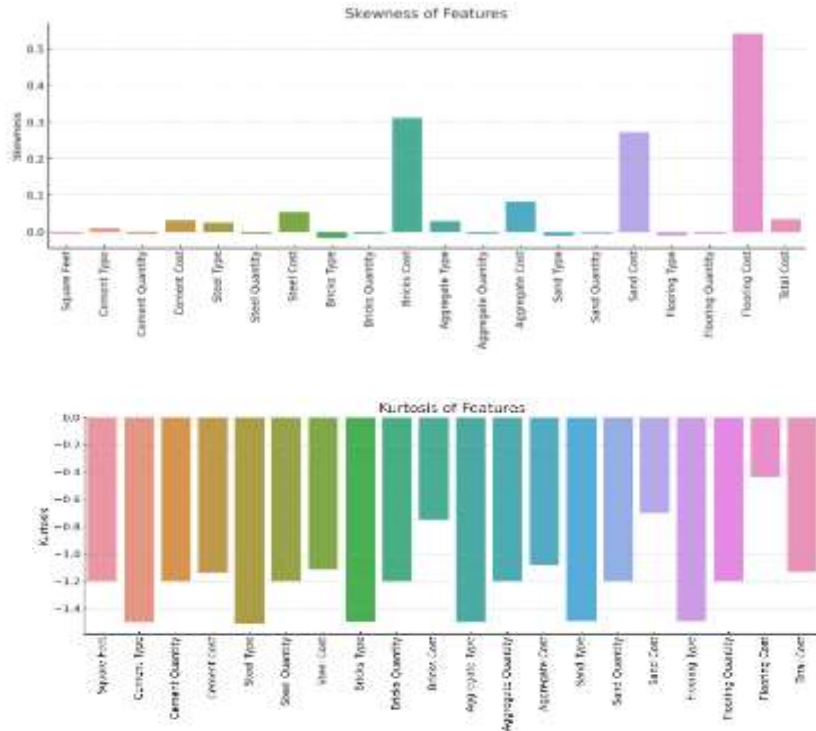
Multivariate relationships between material quantities and costs were again confirmed using pair plots and KDE-based scatter matrix. An example is the correlation between CementQuantity and CementCost, which was proportional to scaling, and SteelQuantity and TotalCost, which were evidently upward trends. The curated data were observed to have good internal consistency and logical organization.

In order to examine more the distribution property of the data, skewness and kurtosis were determined among key numerical variables. The skewness is used to measure the asymmetry of a distribution, and kurtosis to enumerate the extent of tail-heaviness in comparison to a normal distribution.

It was found that the skewness of Steel Cost, CementCost and Total Cost were moderate and in the positive direction. At the same time, the values of kurtosis of the majority of the cost related variables were less than 3, which implies that the data follows a platykurtic distribution with light tails, and no extreme values as the normal distribution would otherwise suggest.

Random Forest and XGBoost were also non-parametric tree-based models that were resistant to irregularities in distribution. However, Linear Regression was a slightly less accurate ($R^2 = -0.983$) model since it is sensitive to skewness and outliers and assumes linearity and homoscedasticity.

Fig.4 Skewness and Kurtosis Plot



Random Forest and XGBoost were also non-parametric tree-based models that were resistant to irregularities in distribution. However, Linear Regression was a slightly less accurate ($R^2 = -0.983$) model since it is sensitive to skewness and outliers and assumes linearity and homoscedasticity.

Even though normalizations like log-scaling or BoxCox might be needed to further normalize the features, they should not have been needed because the ensemble models worked on the raw data. Table 1 demonstrates the dataset features and their impact on the model training.

Table. 1 Impact on Model Training and Related Observations

Metric	Interpretation	Observation in Dataset	Impact on Model Training	
Skewness	Measures asymmetry of data distribution	Moderate right skew in TotalCost, CementCost, SteelCost	Tree-based models unaffected; Linear Regression performance degraded due to skew sensitivity	
Kurtosis		Measures tailedness; >3 = peaked, <3 = flat	Values $< 3 \rightarrow$ platykurtic distribution (light tails)	Minimal impact on tree-based models; further optimization via transformation not necessary

Linear Regression	Assumes linearity and normality	Underperformed with $R^2 \approx 0.983$	Affected by skew/outliers due to parametric nature
Ensemble Models	Robust to non-normality and skewed features	Delivered strong performance even with skewed data	No need for transformation (e.g., Box-Cox, log scaling) due to inherent resilience

These findings underscore that the tree-based ensemble models are resistant to skewed distributions than the parametric models like Linear Regression which assumes both linearity and homoscedasticity.

4.2 Model Training And Evaluation

Training and evaluation are conducted on the model to identify defects that need fixing. Model Training And Evaluation The model is trained and evaluated to identify the defects requiring correction.

Five regression models were trained using the data after exploring the data:

- Random Forest
- Gradient Boosting
- Decision Tree
- Linear Regression
- XGBoost

The algorithms were used to model many output variables (quantity of material, cost of material and the overall cost) using Multi Output Regressor framework.

It was divided into 80 percent training data and 20 percent testing data, and the mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2 score) were used to determine the performance of each model. Table 2 gives the performance obtained through training on the various algorithms.

Table.2 Performance Evaluation Metrics

Model	MAE	RMSE	R^2 Score
Random Forest	Low	Low	~0.9998
Gradient Boosting	Lower	Lower	~0.9997
Decision Tree	Higher	Higher	~0.9996
Linear Regression	Highest	Highest	~0.983
XGBoost	Lowest	Lowest	~0.9999

XGBoost had the lowest predictive performance, the lowest values of MAE and RMSE, and the highest value of R^2 of around 0.9999 as compared to the other models. Random Forest also showed good performance where the R^2 was near 0.9998, which means that

the ensemble models were able to capture nonlinear correlations between the construction parameters and the cost.

Nonetheless, even with such large values of R^2 , there is a risk of overfitting, especially since tree-based models are capable of learning complex patterns that can highly resemble the training data. In order to alleviate this problem, the models were tested on invisible test data and hyperparameters under control. Also, the ensemble models like the Random Forest and XGBoost automatically minimize overfitting by methods like the bootstrap aggregation and gradient boosting regularization.

The result of linear Regression gave relatively low accuracy with R^2 = around 0.983 implying that all the linear relationships cannot be used to model the complicated interactions between the construction variables.

- **Error Analysis**

More detailed analysis of the errors in making predictions indicated that the majority of errors were made in large construction projects as small percentage changes in the amount of materials used brought about more significant changes in the absolute cost. Ensemble models had relatively low values of RMSE, which means that they are predictable even in high-cost situations. The residual analysis revealed that the errors in prediction were lived randomly and there was no intense systematic bias and this proved the reliability of the trained models.

- **Comparison with Existing Studies**

The results achieved are in agreement with the findings of other researchers on building cost prediction through machine learning. Indicatively, research by Yaseen et al. and Liu et al. has shown that ensemble learning algorithms like the Random Forest and gradient models are superior to the classical regression models in construction cost prediction exercises. Just like these works, the current research also validates that nonlinear machine learning models offer better representation of complicated interactions among construction parameters compared to the linear model.

- **Model Validation**

To validate the robustness of the trained models, the evaluation metrics were computed on the test dataset rather than the training data. The minimal difference between training and testing errors indicates that the models generalized well to unseen data. Future work may incorporate additional validation techniques such as k-fold cross-validation and statistical significance testing to further strengthen model reliability.

The comparative performance of all models was also visualized using bar charts across MAE, RMSE, and R^2 metrics. These visual comparisons clearly demonstrated the superiority of XGBoost across all evaluation measures.

The screenshots of the final web application developed for construction cost prediction are presented in Figure 5, and 6. These figures illustrate the user interface of the deployed Flask-based application, including the input parameter window and the generated prediction results. The application allows users to enter relevant construction parameters, after which the trained machine learning model processes the inputs and provides the estimated material quantities and total construction cost.

Fig. 5 Application for Construction Cost Estimation



Fig. 6 Input window and Results

5. Conclusion

This study introduced a machine learning based model to predict construction costs using a big dataset of 10000 construction project records. The study solved the shortcomings of conventional cost estimating methods that tend to use manual computations and trends of the past that might not be in a good position to represent intricate relationships among the variables of construction. A set of regression algorithms were tested using systematic preprocessing, exploratory data analysis, and model training, such as, Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost.

As it was shown by the experimental results, ensemble learning models, especially XGBoost, had a better predictive performance with the lowest error scores and the largest coefficient of determination ($R^2 = 0.9999$). These findings suggest that machine learning models have the capacity to generate nonlinear relationships among construction parameters including quantities of materials, costs, and size of a project. Moreover, the created Flask-based web application provided an opportunity to practically apply the suggested system and put project parameters into the system and get immediate cost estimates using a convenient interface.

On the whole, it is possible to note that the study confirms that machine learning methods can drastically enhance the precision and effectiveness of cost estimation of construction. The suggested system will offer an effective decision-support system to the engineers, contractors and project planners. The future studies can be aimed at uniting real-time price data on the market with explainable AI techniques to make the models more transparent and flexible.

Acknowledgement

References (Style name: PP Heading)

- [1] Zheng, Z., Zhou, L., Zhou, L. "Construction Cost Prediction System Based on Random Forest Optimized by the Bird Swarm Algorithm," *Mathematical Biosciences and Engineering*, 20(8), pp. 15044–15074, 2023. <https://doi.org/10.3934/mbe.2023674>
- [2] Magdum, S. K., Adamuthe, A. C. "Construction Cost Prediction Using Neural Networks," *ICTACT Journal on Soft Computing*, 8(1), pp. 1549–1556, 2017.
- [3] Xu, Y., Cao, S. "Building Engineering Cost Prediction Model Based on TSNE and Improved Grey Correlation Algorithm," *Procedia Computer Science*, 228, pp. 957–965, 2023. [https://doi.org/\[Insert DOI if available\]](https://doi.org/[Insert DOI if available])
- [4] Ye, D. "An Algorithm for Construction Project Cost Forecast Based on Particle Swarm Optimization-Guided BP Neural Network," *Scientific Programming*, 2021, Article ID 4309495. [https://doi.org/\[Insert DOI if available\]](https://doi.org/[Insert DOI if available])
- [5] Hashemi, S. T., Ebadati, O. M., Kaur, H. "Cost Estimation and Prediction in Construction Projects: A Systematic Review on Machine Learning Techniques," *SN Applied Sciences*, 2, Art. 1703, 2020. [https://doi.org/\[Insert DOI if available\]](https://doi.org/[Insert DOI if available])
- [6] Zhang, Y., Mo, H. "Intelligent Building Construction Cost Optimization and Prediction by Integrating BIM and Elman Neural Network," *Heliyon*, 10(5), p. e37525, 2024. [https://doi.org/\[Insert DOI if available\]](https://doi.org/[Insert DOI if available])
- [7] Jirait, S., Rede, S. "House Construction Cost Prediction," *International Journal of Novel Research and Development*, 9(5), pp. h311–h314, 2024.
- [8] Park, D., Yun, S. "Construction Cost Prediction Using Deep Learning with BIM Properties in the Schematic Design Phase," *Applied Sciences*, 13(12), Art. 7207, 2023. [https://doi.org/\[Insert DOI if available\]](https://doi.org/[Insert DOI if available])
- [9] Yaseen, M., Shaker, A., Altaee, S. "Hybrid GA-ANN-SVM Model for Construction Cost and Time Estimation," *International Journal of Engineering and Technology*, 7(3.20), pp. 1142–1147, 2018.
- [10] Liu, C., Li, Q., Wang, C. "Engineering Cost Estimation Model Based on GA-Optimized BP Neural Network," *SHS Web of Conferences*, 170, p. 02009, 2023. [https://doi.org/\[Insert DOI if available\]](https://doi.org/[Insert DOI if available])
- [11] Aung, Theingi, Sui Reng Liana, Arkar Htet, and Amiya Bhaumik. "Using machine learning to predict cost overruns in construction projects." *Journal of Technology Innovations and Energy* 2, no. 2 (2023): 1-7.
- [12] Das, Pobithra, Abul Kashem, Imrul Hasan, and Mominul Islam. "A comparative study of machine learning models for construction costs prediction with natural gradient boosting algorithm and SHAP analysis." *Asian Journal of Civil Engineering* 25, no. 4 (2024): 3301-3316.
- [13] Lee, Haneul, and Seokheon Yun. "Strategies for imputing missing values and removing outliers in the dataset for machine learning-based construction cost prediction." *Buildings* 14, no. 4 (2024): 933.
- [14] Draz, Moatasem M., Osama Emam, and Safaa M. Azzam. "Software cost estimation predication using a convolutional neural network and particle swarm optimization algorithm." *Scientific Reports* 14, no. 1 (2024): 13129.
- [15] Zhang, X., Song, J. "A Whole Process Cost Prediction System for Construction Projects Based on Improved Support Vector Machines," 2022. <https://doi.org/10.46300/9106.2022.16.34>
- [16] Wang, X. "Forecasting Construction Project Cost Based on BP Neural Network," 2018. <https://doi.org/10.1109/ICMTMA.2018.00109>
- [17] Zhang, Y., Mo, H. "Intelligent Building Construction Cost Prediction Based on BIM and Elman Neural Network," 2023. <https://doi.org/10.21203/rs.3.rs-3226303/v1>
- [18] Zheng, Z., Zhou, L., Zhou, L. "Construction Cost Prediction System Based on Random Forest Optimized by the Bird Swarm Algorithm," *Mathematical Biosciences*

- and Engineering, [online] Available at: <https://doi.org/10.3934/mbe.2023674> [Accessed: 14 July 2025].
- [19] "Research on Intelligent Prediction of Engineering Cost Based on Artificial Intelligence," 2023. <https://doi.org/10.25236/ijndes.2023.070110>
- [20] Lakshana, J., Madhumitha, L., Dharshini, S. L. P., Kumar, T. R. "Cost Prediction for Home Construction Using Quantum Computing," 2024. <https://doi.org/10.4018/979-8-3693-3601-4.ch006>
- [21] Rayabharapu, V. K., Rao, K. D., Punitha, S., Abbas, S. H., Sivaranjani, L. "Enhancing Construction Project Cost Predictions using Machine Learning," 2025. <https://doi.org/10.2139/ssrn.5080704>
- [22] Shi, H., Li, W. "A Web-Based Integrated System for Construction Project Cost Prediction," 2010. https://doi.org/10.1007/978-3-642-05173-9_5
- [23] Ali, Z. H., Burhan, A. M., Kassim, M., Al-Khafaji, Z. H. K. "Developing an Integrative Data Intelligence Model for Construction Cost Estimation," 2022. <https://doi.org/10.1155/2022/4285328>
- [24] Ameya, F. M. "Evaluation of the Ongoing Use of Approximate Construction Cost Estimation," 2025. <https://doi.org/10.20944/preprints202501.1900.v1>
- [25] Sarode, G. C., E. C. S. "To Study Cost Prediction Analysis of Construction Project Using ANN Model and SVM by MATLAB," 2020.
- [26] Chandra, A., Yuliana, C. "Estimation of Cost Budget Using BIM in School Construction," 2024. <https://doi.org/10.20527/crc.v8i5.13421>
- [27] Yang, S.-W., Moon, S., Jang, H., Choo, S. Y., Kim, S.-A. "Parametric and BIM-Based Cost Estimation for Construction Projects," 2022. <https://doi.org/10.3390/app12199553>
- [28] Yalçın, G., Bayram, S., Çitakoğlu, H. "Evaluation of Earned Value Management-Based Cost Estimation," *Buildings*, 14(12), 372, 2024. <https://doi.org/10.3390/buildings14123772>