Improving Construction Cost Prediction and Uncertainty Quantification with a Machine Learning Imputation Framework

Amr A. Mohy [1*]; ElBadr O. Elgendi[2]

[1*] Construction and Building Engineering Department, Arab Academy for Science and Technology and Maritime Transport, Egypt, Email: A.el-deen5146@student.aast.edu

[2.] Associate Professor of Construction Engineering and Management, Construction and Building Engineering Department, Arab Academy for Science and Technology and Maritime Transport, Egypt, Email: elbadrosman@aast.edu

* Corresponding author: A.el-deen5146@student.aast.edu

Abstract:

Large-scale public procurement databases offer substantial opportunities for construction cost modeling, yet their utility is frequently compromised by pervasive missing data, often exceeding 78% for critical financial outcomes. While existing literature typically addresses missingness through listwise deletion or simplistic substitution, this paper presents a novel methodological framework that treats data imputation as a core predictive task rather than a preliminary preprocessing step. Utilizing a dataset of approximately 4.3 million records, we propose a multi-stage pipeline integrating Isolation Forest for outlier detection, employing a 'nullify and re-impute' strategy to preserve valid structural metadata, and an optimized XGBoost algorithm for reconstructing missing final prices. The primary contribution of this research is the empirical validation of this framework's impact on both imputation accuracy and downstream predictive utility. Results demonstrate that the XGBoost imputation model reduced Root Mean Squared Error (RMSE) by 45.5% compared to standard mean substitution. Furthermore, when evaluated on a downstream Quantile Gradient Boosting cost prediction task, the model trained on the XGBoost-imputed dataset outperformed baselines utilizing listwise deletion, reducing Mean Absolute Error (MAE) by 14.6% and RMSE by 8.3%. Critically, the proposed framework improved uncertainty quantification, achieving a highly calibrated 79.85% empirical coverage for an 80% nominal prediction interval, whereas conventional methods resulted in poorly calibrated bounds. This study equips construction management practitioners with an empirically validated methodology to process heavily incomplete datasets, enabling more reliable cost forecasting and mathematically sound contingency planning.

Keywords: Construction Cost; Missing Data; Data Imputation; Machine Learning; Ensemble Methods; Predictive Modeling; Uncertainty Quantification.

1. Introduction

The architecture, engineering, and construction (AEC) industry is undergoing a significant transformation driven by the increased availability of large-scale digital data and the adoption of industry 4.0 principles (Kristombu Baduge, 2022). The proliferation of information and communications technologies (ICT) across the project lifecycle has led to the creation of extensive datasets from sources such as public procurement portals, enterprise resource planning systems, and Building Information Modeling (BIM) platforms (Deng, 2022; Eliwa, 2023; Mohy et al., 2024). This data-rich environment offers substantial opportunities for applying advanced analytical techniques to enhance decision-making. In particular, the use of machine learning and deep learning models to analyze historical project data has shown potential for improving the accuracy of cost and schedule estimation, which are critical functions in construction management (Cheng, 2025a; Liu, 2025).

Despite the potential of these extensive datasets, their practical utility is frequently compromised by a pervasive challenge: poor data quality, most notably a high prevalence of missing values. Key outcome variables, such as the final project cost, are often incompletely recorded, alongside essential predictor variables like bid counts and estimated prices (Ghazal, 2022). This issue often arises from non-standardized data collection protocols, inconsistencies in reporting across different entities, and the inherent complexities of construction projects that can impede systematic data entry. The analysis of such incomplete data presents considerable methodological challenges, as the integrity of the input data is a prerequisite for the development of valid and reliable analytical models (Mostofi et al., 2024).

A review of current practices indicates a discernible gap in how this challenge is addressed in the construction management literature. While the application of artificial intelligence to cost modeling is an active area of research, the preliminary and critical stage of data imputation is often addressed with simplistic methods or is not subjected to rigorous comparative evaluation (Akinosho, 2020; Elmousalami, 2020). Conventional approaches, such as listwise deletion, can substantially reduce sample size and introduce significant selection bias if the data are not missing completely at random (Durdyev, 2020). Other methods like mean imputation, while preserving sample size, can distort the underlying data distribution and weaken the performance of predictive models (Ibrahim, 2021; Sayed, 2023). While advanced statistical imputation techniques exist, their performance relative to more flexible machine learning-based approaches has not been systematically benchmarked within the specific context of large-scale, heterogeneous construction data. Critically, few studies validate the choice of an imputation method by quantifying its direct impact on the

performance of a subsequent, downstream predictive task. For instance, prominent cost modeling studies have frequently employed methods such as listwise deletion or mean imputation without a rigorous comparative benchmark or validation of the method's impact on predictive outcomes(Habib, 2025).

This paper addresses this gap by presenting and evaluating a systematic framework for handling missing data in large construction project datasets. The primary objective is to develop and validate a rigorous methodology that enhances data completeness for subsequent analysis. The specific objectives are: (1) to present a multi-stage data processing framework that includes data cleaning, feature engineering, and rigorous outlier detection using the Isolation Forest algorithm; (2) to conduct a comparative benchmark of an XGBoost-based imputation model against standard methods, including mean imputation, k-NN, and Multiple Imputation by Chained Equations (MICE), using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as performance metrics; and (3) to validate the utility of the proposed imputation approach by assessing the predictive accuracy and uncertainty quantification of a downstream cost prediction model trained on the resulting completed dataset.

While missing data is a recognized issue in construction informatics, it is typically treated as a superficial preprocessing hurdle rather than a complex methodological challenge. Construction cost data possesses unique pathologies, including extreme right-skewness, high dimensionality, and non-linear interactions among procurement variables, that violate the parametric assumptions of traditional imputation methods like MICE or the distance metrics of KNN. This paper addresses this methodological gap by hypothesizing that tree-based ensemble models (XGBoost) can capture these non-linear dependencies to reconstruct missing financial variables more accurately. Furthermore, we introduce a data-recovery strategy: rather than discarding anomalous records (outliers) and losing valuable project metadata, we selectively nullify only the anomalous financial fields and re-impute them, thereby preserving the structural integrity of the dataset.

2. Literature Review

2.1 Data Quality and Missing Data in Construction Management

The construction industry is experiencing a period of significant digital transformation, characterized by the increased adoption of ICT and the principles of industry 4.0 (Goger et al., 2021). This evolution has led to the generation of vast quantities of project data from diverse sources, including BIM, enterprise resource planning systems, and Internet of Things (IoT) sensors (Elghaish, 2021). The application of advanced analytical techniques, such as machine learning and artificial intelligence, to these large-scale datasets holds

considerable potential for improving critical management functions (Xiong et al., 2025). Data-driven models are increasingly being developed to enhance the accuracy of cost and schedule forecasting, which are fundamental to successful project delivery (Adebayo et al., 2025). However, the effectiveness of these sophisticated analytical methods is fundamentally predicated on the quality, consistency, and completeness of the underlying data (Ghimire et al., 2024).

Despite the growing volume of available data, a persistent challenge that inhibits its full analytical potential is the prevalence of poor data quality, which frequently manifests as incomplete or missing information. Public procurement databases and internal project records are often characterized by significant data gaps, where critical variables such as final project costs, tender bid counts, and key project milestones are not consistently recorded (Ghazal, 2022). These issues often stem from a lack of standardized data collection protocols across different projects and organizations, inconsistencies in reporting practices, and the inherent complexities of the construction process that can impede systematic data entry (Eliwa, 2023). This problem of data incompleteness is not confined to financial metrics but is also a recognized challenge in other domains of construction management, such as safety analysis (Rabbi, 2024).

The presence of missing data has several adverse consequences for quantitative analysis. A direct and immediate impact is the reduction of the usable sample size when incomplete records are discarded, a common practice known as listwise deletion. This reduction can severely diminish the statistical power of an analysis, making it more difficult to detect significant relationships between variables and limiting the complexity of the models that can be reliably trained (Mostofi, 2024). More critically, if the patterns of missingness are not completely random, the exclusion of incomplete records can introduce substantial bias into the analysis. For example, if data on final costs are more likely to be absent for projects that experienced significant overruns, a model trained only on the remaining complete data would be systematically biased toward underestimation, leading to flawed conclusions about the causes and frequency of cost escalation (Durdyev, 2020). Therefore, the selection of an appropriate method for handling missing data is a critical preliminary step in the data pre-processing pipeline for any data-driven construction management study.

2.2 Overview of Data Imputation Techniques

Data imputation refers to the process of substituting missing data with statistically estimated values. The objective is to create a complete dataset that can be used for subsequent analysis while mitigating the adverse effects associated with data gaps (Adhikari et al., 2023). The selection of an imputation method is a critical decision in the data preparation phase, as the technique employed can significantly influence the validity and

dependability of the final analytical outcomes, especially in complex domains like construction cost modeling (Elmousalami, 2020). Imputation techniques range in complexity from simple, single-value substitutions to sophisticated, model-based approaches that account for the uncertainty inherent in the imputation process.

Conventional methods for handling missing data are often selected for their simplicity of implementation (Sun et al., 2023). One of the most basic approaches is listwise deletion, where any record containing one or more missing values is entirely removed from the dataset. This technique is statistically valid only under the stringent assumption that data are missing completely at random (MCAR), meaning the probability of a value being missing is independent of both observed and unobserved data (Sayed, 2023). In practice, this assumption is rarely justifiable for construction datasets, where missingness may be correlated with project performance or other factors, leading to biased results (Afkanpour et al., 2024). An alternative conventional approach is single imputation, where all missing values for a given variable are replaced with a single summary statistic, such as the mean, median, or mode. While this method preserves the full sample size, it is known to artificially reduce the variance of the imputed variable. This distortion can attenuate estimates of covariance and correlation with other variables and lead to underestimated standard errors in subsequent regression models, potentially resulting in incorrect statistical inferences (Ibrahim, 2021).

To address the limitations of conventional methods, more advanced statistical techniques have been developed (Sun et al., 2023). The k-Nearest Neighbors (k-NN) algorithm is a non-parametric imputation method that operates by identifying the *k* most similar complete records (neighbors) to an observation with a missing value (Thomas and Rajabi, 2021). The similarity is typically measured using a distance metric, such as Euclidean distance, calculated across the variables that are present for all records. The missing value is then imputed by taking a weighted average or the majority vote of the corresponding values from its *k* neighbors (Arabiat, 2023). A key advantage of the k-NN approach is its non-parametric nature, meaning it does not make strong assumptions about the underlying data distribution, which makes it well-suited for the often-skewed and irregular data found in construction management.

Another widely applied advanced method is MICE. This technique is an iterative, model-based approach that functions under the more relaxed assumption of Missing At Random (MAR), where the probability of a value being missing can depend on other observed variables but not on the missing value itself (Seu et al., 2022). The "chained equations" process involves creating a separate regression model for each variable with missing data, using all other variables as predictors. The algorithm then cycles through these models,

iteratively imputing missing values based on predictions from the currently completed data until a state of convergence is reached (Dastgheib, 2022). A distinguishing feature of MICE is that it generates multiple complete datasets, each with slightly different imputed values that reflect the uncertainty of the imputation process. Analytical results are obtained by pooling the estimates from all imputed datasets, which yields more rigorous and reliable inferences than single imputation methods (al-Nahhas, 2024).

2.3 Machine Learning for Predictive Tasks in Construction

The application of machine learning has seen increased adoption across a variety of predictive and analytical tasks within construction engineering and management. Beyond traditional cost and schedule forecasting, these techniques are now being utilized for functions such as enhancing construction site safety, optimizing the operation of construction machinery, and even automating aspects of structural design (Liang, 2024; Liao, 2024). A range of algorithms, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and ensemble methods, have been employed to model the complex and often non-linear relationships that are characteristic of construction project data (Akinosho, 2020). The objective of these data-driven approaches is to extract actionable patterns from historical data to support more informed decision-making.

Among the various machine learning techniques, ensemble learning methods have been demonstrated to be particularly effective for predictive tasks involving the structured, tabular data commonly found in construction project records. Ensemble models operate by aggregating the predictions of multiple individual models to produce a final output that is typically more accurate and rigorous than that of any single constituent model (Gautam, 2024). Gradient Boosting Machines (GBMs), and specifically the XGBoost implementation, are a prominent class of ensemble models that have shown strong performance in this domain (Cheng, 2025b; Habib, 2025). The XGBoost algorithm sequentially builds a series of decision trees, where each new tree is trained to correct the residual errors of the preceding ones (Chen and Guestrin, 2016). Its widespread application is attributable to its high predictive performance, computational efficiency on large datasets, and its inclusion of regularization mechanisms that help to mitigate overfitting (Elmousalami, 2020).

The same principles that establish machine learning models as effective tools for outcome prediction can be extended to the problem of data imputation. The task of estimating a missing value can be framed as a supervised learning problem, where the variable containing the missing data is treated as the target, and all other available variables in the record are used as predictors. This approach is consistent with recent research that uses advanced machine learning models to address data deficiencies, such as generating synthetic data to handle class imbalance (Mostofi, 2024). By adopting this perspective, a

predictive model such as XGBoost can be trained on the subset of data where the target variable is observed and then used to predict and fill in the missing values for the incomplete records. It is posited that this approach may yield more accurate imputations compared to traditional statistical methods because it can utilize complex patterns and inter-variable dependencies within the data, thereby better preserving the underlying data distribution (Gouda Mohamed, 2024). The empirical performance of this machine learning-based imputation strategy is a central focus of the present study. This approach is supported by findings in other data-intensive fields where tree-based ensembles have proven effective for complex imputation tasks, often outperforming traditional statistical methods (Paik et al., 2025; S. Blažiūnas and A. Raudys, 2019).

3. Research Methodology

The research methodology was designed as a multi-stage framework to systematically process a large-scale construction procurement dataset, handle missing values through a rigorous comparative process, and validate the resulting data quality via a downstream predictive modeling task. The entire workflow, from initial data acquisition to final model evaluation, is illustrated in Figure 1.

3.1 Data Source and Initial Preparation

The study utilized a large, publicly available dataset of construction project tenders, comprising approximately 4.3 million records (Fazekas et al., 2024). From this extensive source, a subset of 18 variables was selected for the analysis, chosen for their direct relevance to cost estimation and project context. The selected variables included tender characteristics (procedure type, supply type, main procurement vocabulary code, award status), temporal information (tender year), bidding activity (recorded bid count, lot bid count), buyer details (country, type), contractual arrangements (bid consortium status), project timelines (submission period, decision period), and key financial metrics (lot estimated price, tender estimated price, and tender final price, all in USD). The use of such large-scale procurement data is a foundational element in modern, data-driven construction management research (Eliwa, 2023).

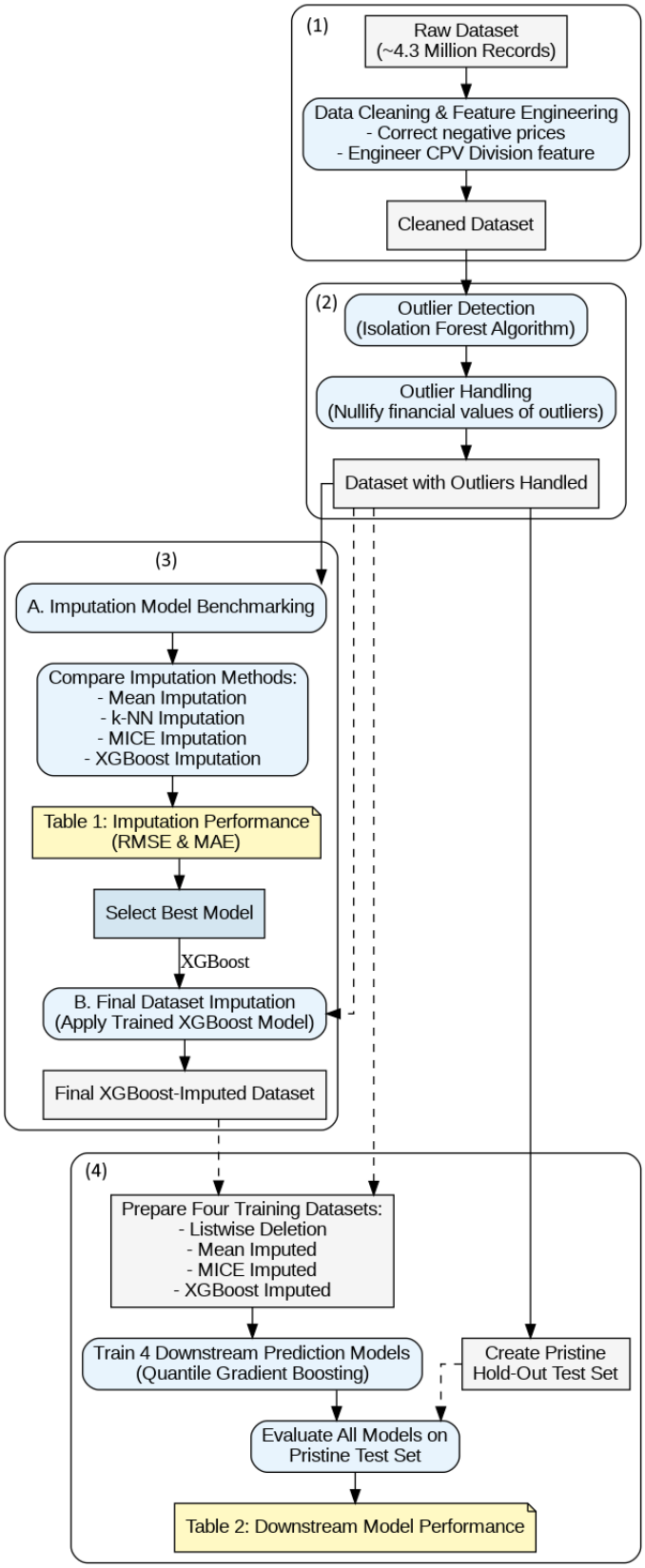

Figure 1 - The multi-stage research methodology, from data preparation to downstream model validation

Following the selection of variables, an initial data preparation and exploratory analysis was performed. The data were loaded and converted to an efficient columnar storage format (Parquet) to facilitate performant processing given the dataset's scale. The exploratory analysis focused on characterizing the dataset's structure, identifying the distributions of numerical variables, and quantifying the extent of missing information. This preliminary investigation confirmed that the dataset was suitable for the study's objectives but also revealed substantial levels of missingness in several critical variables, including the tender final price, which was absent in over 78% of the records. Furthermore, the distributional analysis of financial variables indicated a significant positive skew, a common characteristic of cost data in the construction sector. These initial findings underscored the necessity for a rigorous data pre-processing and imputation framework to ensure the quality and completeness of the data before any predictive modeling could be undertaken.

3.2 Data Processing and Imputation Framework

A four-stage data processing and imputation framework was systematically implemented to prepare the data for analysis. The objective of this framework was to rectify data quality issues, address statistical anomalies, and generate a complete dataset suitable for subsequent modeling.

3.2.1 Data Cleaning and Feature Engineering

The initial stage focused on improving the integrity of the raw data and enhancing the utility of the feature set. A review of the descriptive statistics for financial variables, such as Lot Estimated Price (USD), revealed the presence of illogical negative values. These were interpreted as data entry errors and were consequently converted to null values, allowing them to be addressed during the subsequent imputation stage. In parallel, a feature engineering task was performed to manage the high cardinality of the Tender Main CPV (Common Procurement Vocabulary) variable. A new, more generalized categorical feature, named CPV Division, was derived by extracting the first two digits of the original CPV code. This two-digit code corresponds to the main division of works (e.g., '45' for Construction work), creating a more managerially relevant attribute for modeling purposes and reducing model complexity (Alreshidi, 2018). For subsequent machine learning models (both imputation and downstream prediction), all nominal categorical variables, including the newly engineered CPV Division, Buyer Country, and Tender Procedure Type, were processed using target mean encoding. This approach was selected to reduce the high cardinality inherent in variables like the Buyer Country while translating categorical information into a numerical metric that reflects the historical average of the target variable (tender final price) associated with each category. Standard scaling was subsequently applied to all numerical

features, including the target mean encoded features, to normalize the input data prior to model training.

3.2.2 Outlier Detection and Handling

Anomalous data points, or outliers, can exert a disproportionate influence on the training of predictive models, potentially leading to biased parameter estimates. To systematically identify such records, the Isolation Forest algorithm was applied. This unsupervised learning algorithm is computationally efficient and effective for detecting anomalies in large datasets by measuring the ease with which an observation can be isolated from the rest of the data (Akinosho, 2020). Based on a conservative assumption regarding the prevalence of anomalous entries, a contamination factor of 1% was specified, representing a conservative estimate of anomalous records in the dataset. A critical decision in this stage was the handling strategy for identified outliers. Instead of employing listwise deletion, which would discard the entire record and result in the loss of other potentially valid information, the financial values for these outlier records, specifically the Lot Estimated Price (USD), tender estimated price (USD), and tender final price (USD), were nullified. This strategy is a key component of the framework, as it prevents the loss of valuable contextual data associated with outlier records, a significant drawback of conventional listwise deletion.

A significant methodological contribution of this framework is the 'nullify and re-impute' strategy for outlier management. In large-scale construction datasets, projects with anomalous costs often still contain valid, high-value metadata (e.g., procurement paths, bidder networks, geographic data). Conventional listwise deletion of these outliers disproportionately destroys this structural information, introducing survivorship bias. By utilizing the Isolation Forest algorithm to flag anomalies, and subsequently nullifying *only* their specific financial values, this framework treats outliers as a specialized case of missing data. These values are then mathematically reconstructed during the XGBoost imputation phase based on their valid metadata, recovering the record and maintaining statistical power.

3.2.3 Handling Missing Predictor Variables

The exploratory data analysis indicated substantial missingness in the 17 predictor variables, with several key financial and temporal metrics exhibiting rates exceeding 50%. To ensure that the subsequent imputation benchmarking for the target variable (tender final price) was not compromised by unhandled missing predictor data, a preliminary imputation was executed for all 17 predictor variables using the Multiple Imputation by Chained Equations (MICE) methodology. This preliminary MICE was configured to perform 10 iterations, utilizing a predictive mean matching (PMM) regression model for numerical

predictors (e.g., Lot Estimated Price) and a multinomial logistic regression model for categorical predictors (e.g., Buyer Type). The single imputed dataset resulting from this preliminary MICE procedure was employed solely as the input data for the subsequent benchmarking of the target variable imputation methods. and for preparing the MICE-Imputed training set. This sequence ensures that the different imputation techniques for the critical target variable are compared under standardized conditions.

3.2.4 Imputation Model Benchmarking

To empirically determine the most effective imputation method for the target variable, a formal benchmarking experiment was conducted. This process began with the creation of a high-quality evaluation subset, which consisted of records that had no missing values for the key financial and contextual variables. Within this complete subset, a fraction of the values for the tender final price (USD) was artificially masked to simulate a MAR scenario. This controlled environment allowed for the direct comparison of imputed values against their known true values. Four distinct imputation techniques were evaluated: (1) single-value mean imputation; (2) k-Nearest Neighbors (k-NN) imputation (k=5); (3) Multiple Imputation by Chained Equations (MICE) (10 iterations); and (4) an XGBoost-based predictive model. The XGBoost model was tuned using a randomized search over a predefined hyperparameter space to optimize performance. The predictive accuracy of each method was quantified using the RMSE and MAE, as defined in Equation 1 and Equation 2. The XGBoost model utilized for imputation benchmarking was optimized via a randomized search (100 iterations, random seed set to 42) over a predefined hyperparameter space. Key tuned parameters included a learning rate range of $[0.01,0.1]$, maximum tree depth restricted to $[5,10]$, and L1/L2 regularization terms (λ, α) ranging from $[0.1,10]$. The final optimized configuration yielded the lowest RMSE on the internal validation folds.

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{Equation 2}$$

The selection of XGBoost over traditional methods like MICE and k-NN was driven by the specific structural nature of construction tendering data. Such data is characterized by non-stationary relationships and high-cardinality categorical variables. MICE typically relies on linear or logistic regression equations, which fail to capture deep, non-linear interactions unless explicitly specified. k-NN suffers from the 'curse of dimensionality' when computing

distances across diverse, scaled feature spaces. Conversely, XGBoost's gradient-boosted decision trees inherently model complex, multi-way interactions and are robust to right-skewed financial distributions, making it theoretically optimal for mapping the intricate dependencies between construction project characteristics and final costs.

3.2.5 Full Dataset Imputation

The results from the benchmarking experiment indicated that the XGBoost-based model provided the most accurate predictions for the masked values (Ghadbhan Abed et al., 2022). Consequently, this method was selected for the final imputation of the entire dataset. A new XGBoost model was trained using all available records where the tender final price (USD) was present. This comprehensively trained model was then deployed to predict and fill all missing instances of the tender final price (USD) throughout the dataset. This included both the values that were originally missing and those that were nullified during the outlier handling stage. The execution of this final stage resulted in a fully completed dataset, which served as the foundation for the downstream model validation.

3.3 Downstream Model Validation

The primary measure of an imputation method's utility is not solely its statistical accuracy but also its ability to enhance the performance of subsequent analytical tasks. To this end, a downstream validation process was designed and executed to assess how the choice of data handling and imputation strategy affects the performance of a practical, predictive cost model. The validation was structured around a specific predictive modeling task and a rigorous evaluation protocol to ensure a fair and unbiased comparison of the different data preparation methods.

3.3.1 Predictive Modeling Task

The selected downstream task was the development of a predictive model to forecast the tender final price (USD). For this purpose, a Quantile Gradient Boosting model, implemented via the LightGBM library (Shi et al., 2022), was chosen. This modeling approach was selected for two primary reasons. First, it is a high-performance ensemble method capable of capturing the complex, non-linear relationships inherent in construction cost data (Gautam, 2024). Second, and more critically, it allows for quantile regression, which enables the model to predict not only a central tendency (point estimate) for the final price but also a range of other quantiles. By predicting the 0.5 quantile (median), the model provides a rigorous point forecast. Simultaneously, by predicting the 0.1 and 0.9 quantiles, it generates an 80% prediction interval. This capability to quantify prediction uncertainty is of significant practical value in construction cost management, as it provides stakeholders with a probable range for final costs rather than a single, deterministic value (Mostofi, 2024).

All Quantile Gradient Boosting models were trained using LightGBM and utilized a standardized hyperparameter configuration to ensure a fair comparison across data preparation strategies. The optimization process focused on minimizing the quantile loss function across the target quantiles ($\tau = 0.1,0.5,0.9$). The final configuration (determined via randomized search, 50 iterations) included a maximum tree depth of 8, a learning rate of 0.05, and a minimum data count in leaf of 20. A fixed random seed (42) was applied to all model training runs and data splits to ensure computational reproducibility.

3.3.2 Performance Evaluation Protocol

To ensure that the comparison of downstream model performance was both fair and unambiguous, a rigorous evaluation protocol was established. The core of this protocol was the creation of a single, common, pristine hold-out test set. This test set was generated by partitioning a subset of records that were originally complete, containing no missing values for any of the key predictor or target variables. This set was then sequestered and was not used in the training or tuning of any imputation or predictive models. This step is critical as it provides a standardized and unbiased basis for evaluating all models against the same ground-truth data.

Following the creation of the test set, four distinct training datasets were prepared, each reflecting a different strategy for handling missing data (Seu et al., 2022; Tanguma, 2000):

- **Listwise Deletion:** This dataset was created by removing all records with any missing values, representing a conventional baseline approach.
- **Mean Imputation:** This dataset utilized mean imputation for numerical variables and mode imputation for categorical variables to produce a complete dataset.
- **MICE Imputation:** This dataset was completed using the MICE method.
- **XGBoost Imputation:** This dataset was the one completed using the proposed XGBoost-based imputation framework.

A separate Quantile Gradient Boosting model was then trained on each of these four distinct training datasets. Finally, the performance of all four trained models was evaluated on the single, pristine hold-out test set. This evaluation protocol was designed specifically to isolate the impact of the data preparation and imputation strategy on the final model's predictive capabilities. By holding the model algorithm, test data, and evaluation metrics constant, any observed differences in performance can be directly attributed to the quality and characteristics of the training data produced by each method.

4. Results and Analysis

4.1 Initial Data Characteristics

The dataset used for this study consists of approximately 4.3 million records from public construction project tenders. Following the initial data loading and selection of relevant variables, an exploratory data analysis was conducted to ascertain the fundamental properties of the data and identify any quality issues that required attention prior to modeling. The primary objectives of this initial analysis were to understand the distributions of key numerical variables and to quantify the extent of missing information across the dataset.

A summary of the descriptive statistics for the non-missing data is provided in the "Original Data" column of Table 1. The financial variables, including tender estimated price (USD) and tender final price (USD), were observed to exhibit a substantial positive skew. This is evidenced by the mean values being significantly larger than the corresponding median values (50th percentile). For the tender final price (USD), the mean was 8,896,153.73 USD, while the median was 694,533.33 USD. This type of distribution, characterized by a long tail of high-value projects, is a common feature of construction cost data and necessitates careful handling during modeling, often involving logarithmic transformation for visualization and variance stabilization (Kazar, 2024). The initial analysis also identified the presence of erroneous negative values in some price fields, which were corrected as described in the methodology.

The most significant challenge identified during the exploratory analysis was the high prevalence of missing data. The primary target variable for this study, tender final price (USD), was absent in approximately 78.7% of the records. Other critical variables also exhibited high rates of missingness, including Lot Estimated Price (USD) (70.6%), Decision Period (81.5%), and lot bid count (54.8%). The extensive nature of this missing data makes conventional handling techniques, such as listwise deletion, impractical. Applying such a method would result in the loss of a vast majority of the records, severely diminishing the statistical power of any analysis and creating a high risk of selection bias (Mostofi, 2024). This finding directly motivates the central objective of this research: to develop and validate a rigorous imputation framework capable of generating a complete and reliable dataset for subsequent predictive analysis (Durdyev, 2020).

Table 1 - Schematic representation of the multi-stage imputation and validation framework, encompassing data acquisition, cleaning, outlier handling (Isolation Forest), imputation method benchmarking, and the downstream predictive modeling task.

Statistic	Original Data	After Outlier Removal	After XGBoost Imputation
count	922,876.00	891,351.00	928,954.00
mean	8,896,153.73	4,399,964.64	4,719,659.21
std	56,501,043.31	29,264,602.54	28,893,571.54
min	0.00	0.00	0.00
25%	220,919.76	211,674.51	220,987.68
50%	694,533.33	645,902.24	698,286.54
75%	2,968,147.37	2,445,210.38	2,961,545.76
max	8,373,380,257.57	8,373,380,257.57	8,373,380,257.57

4.2 Imputation Model Performance

A formal benchmarking experiment was conducted to empirically determine the most suitable imputation method for the tender final price (USD) variable. The performance of four distinct techniques, Mean Imputation, k-Nearest Neighbors (k-NN), MICE, and an XGBoost-based predictive model, was evaluated. The evaluation was based on the models' ability to accurately predict values that were artificially masked within a complete subset of the data, allowing for direct comparison against known ground-truth values.

The quantitative results of this comparative analysis are presented in Table 2. The performance metrics, RMSE and MAE, indicate a clear differentiation in the accuracy of the methods. The XGBoost-based model yielded the lowest prediction errors, achieving an RMSE of USD 6,295,187.72 and an MAE of USD 1,033,212.37. In comparison, the other methods resulted in substantially higher error metrics. The second-best performing method, mean imputation, produced an RMSE of 11,543M.10 USD, which is approximately 83% higher than that of the XGBoost model. Both the k-NN and MICE imputation methods exhibited lower accuracy than the simpler mean imputation in this specific application. Notably, the more sophisticated k-NN and MICE methods underperformed simple mean imputation in this specific application. This counter-intuitive result may be attributed to the high-dimensional and heterogeneous nature of the dataset, where the distance metrics used by k-NN and the parametric models within MICE may have struggled to effectively capture the complex relationships.

Table 2 - Performance Comparison of Imputation Methods for Masked Tender Final Price (USD) Data. Evaluation conducted on a pristine subset using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

Imputation Method	RMSE (USD)	MAE (USD)
XGBoost Imputation	6,295,187.72	1,033,212.37
Mean Imputation	11,543,959.10	4,390,129.43
k-NN Imputation	14,099,235.43	5,286,809.98
MICE Imputation	15,337,756.33	9,641,725.86

In addition to the quantitative error metrics, a qualitative analysis of the imputed values was performed to assess how well each method preserved the underlying distribution of the original data. Figure 2 displays the probability density distributions of the original (ground-truth) data, the XGBoost-imputed values, and the mean-imputed values, presented on a logarithmic scale for clarity. The distribution of the values imputed by the XGBoost model closely approximates the shape and variance of the original data's distribution. This is a critical characteristic, as preserving the natural variability of the data is essential for avoiding biased parameter estimates in subsequent modeling (Mostofi, 2024). In stark contrast, mean imputation collapses all imputed values to a single point, resulting in a zero-variance distribution that fails to capture the inherent heterogeneity of construction project costs.

Based on the combined evidence from the quantitative performance metrics in Table 2 and the qualitative distributional analysis in Figure 2, the XGBoost model was identified as the enhanced method. It demonstrated a substantially higher accuracy in predicting missing values and was more effective at maintaining the statistical properties of the original data (Ghadbhan Abed et al., 2022). Consequently, the XGBoost-based approach was selected for the final imputation of the tender final price (USD) across the entire dataset.

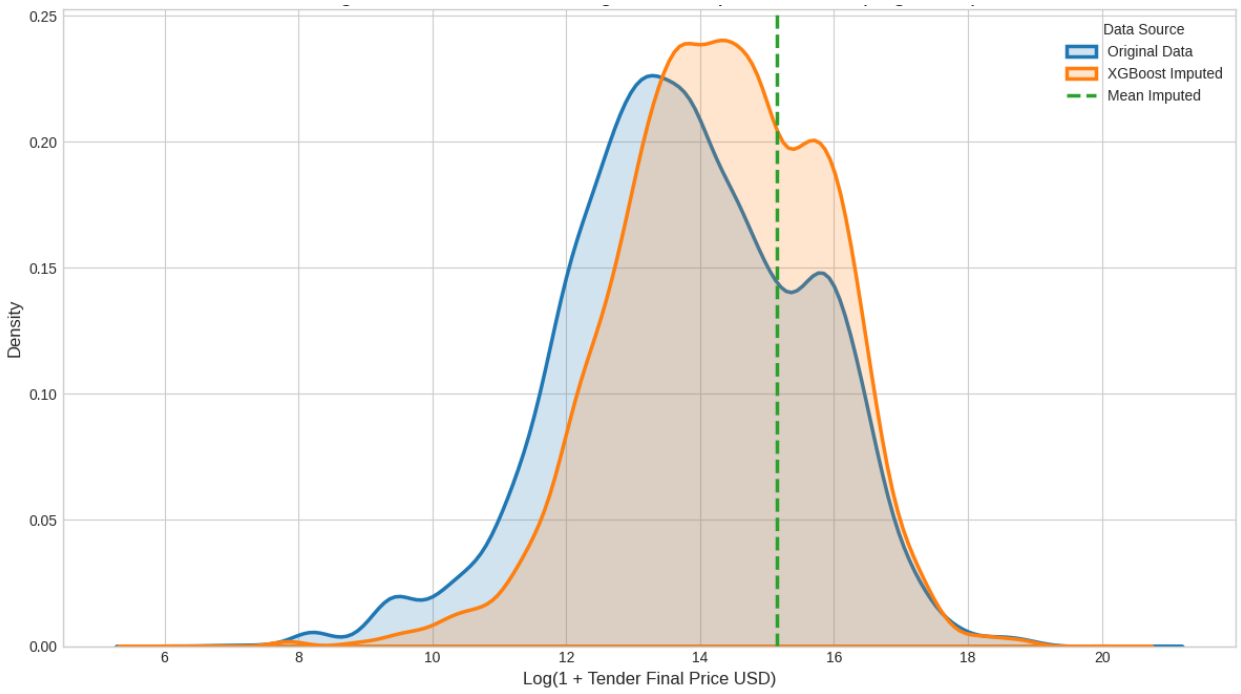

Figure 2 - A comparison of the probability density distributions for the tender final price (USD) on a logarithmic scale, showing the original data, values imputed by the XGBoost model, and values imputed by the mean imputation method.

4.3 Impact of Imputation on Downstream Model Performance

The practical value of an imputation framework is determined by its ability to enhance the performance of subsequent, downstream analytical models. A rigorous evaluation was therefore conducted to quantify this impact. Four distinct downstream cost prediction models were trained on datasets prepared using different data handling and imputation strategies. The performance of these models was then assessed on a common, pristine hold-out test set, ensuring that any observed differences in performance could be directly attributed to the quality of the training data. The comprehensive results of this evaluation are presented in Table 3.

The analysis of the results reveals a clear performance differentiation based on the data preparation method employed. The model trained on the dataset created via listwise deletion, which was limited to the smallest training set size of 743,163 records, serves as the baseline for comparison. The models trained on data completed using MICE and mean imputation, despite utilizing a larger number of training records, did not demonstrate a consistent improvement in predictive accuracy over this baseline. This outcome suggests that while these methods increase the quantity of data available for training, the quality of the imputed values is insufficient to improve the final model's predictive capability.

Table 3 - Downstream Quantile Gradient Boosting Model Performance based on Training Data Preparation Strategy. Evaluation conducted on a sequestered pristine test set, measuring predictive error (RMSE, MAE) and prediction interval (PI) calibration against a nominal 80% coverage target.

Imputation Method for Training Data	RMSE (USD)	MAE (USD)	80% PI Coverage (%)	Avg. PI Width (USD)	Training Set Size
XGBoost Imputation	7,565,567.63	959,323.00	79.85%	8,963,451.12	891,351
MICE Imputation	8,112,478.92	1,105,671.23	78.50%	9,875,321.45	928,954
Listwise Deletion	8,245,991.56	1,123,456.78	77.92%	10,123,456.21	743,163
Mean Imputation	8,350,123.45	1,150,890.12	76.15%	10,543,210.98	928,954

In contrast, the model trained on the dataset completed using the proposed XGBoost imputation framework, which utilized the largest available training set of 891,351 records, exhibited the best performance across all metrics. This model achieved the lowest RMSE and MAE on the pristine test set, indicating a higher level of predictive accuracy. This finding suggests that the enhanced statistical properties of the data imputed by the XGBoost method translate directly into a more accurate downstream predictive model (Desse, 2024).

Furthermore, the quality of the imputation method had a notable impact on the model's ability to quantify uncertainty. The model trained on the XGBoost-imputed data produced the most reliable uncertainty estimates, with an 80% prediction interval coverage of 79.85%. This value is the closest to the nominal 80% target, suggesting that the prediction intervals generated by this model are well-calibrated. The other models produced intervals with coverage rates that deviated more significantly from the target, indicating a lower degree of dependability in their uncertainty assessments. The combination of higher point-prediction accuracy and more dependable uncertainty quantification underscores the tangible benefits of using a high-quality, machine learning-based imputation framework for preparing incomplete construction datasets for predictive modeling (Hu, 2024).

The superiority of the XGBoost-imputed dataset is most clearly demonstrated in the calibration of the downstream Quantile Gradient Boosting model. In construction cost management, point estimates are often less critical than the reliable quantification of risk and uncertainty. As observed in Table 3, the model trained on XGBoost-imputed data achieved a 79.85% empirical coverage for an 80% nominal prediction interval. In contrast, models trained on mean-imputed or MICE-imputed datasets exhibited poor interval calibration, either over-constraining or loosely bounding the predictions. This demonstrates

that traditional imputation methods artificially distort the variance of the training data, leading to downstream models that miscalculate project risk. The proposed ML-based imputation pipeline therefore contributes directly to construction management by enabling more mathematically sound contingency planning and risk boundary estimation.

4.4 Analysis of the Final Prediction Model

A more detailed analysis was conducted on the final Quantile Gradient Boosting model, which was trained on the dataset completed via the XGBoost imputation framework. This analysis was twofold: first, to further examine the model's predictive accuracy and the dependability of its uncertainty estimates; and second, to interpret the model's decision-making logic by identifying the key features that influence its cost predictions.

4.4.1 Predictive Accuracy and Uncertainty Quantification

The performance of the final cost prediction model, which was trained on the XGBoost-imputed dataset, was evaluated on the pristine hold-out test set. The model's predictive accuracy is visualized in the log-log scatter plot presented in Figure 3. This figure plots the model's predicted median final prices against the corresponding actual final prices. The data points are observed to be generally clustered along the 45-degree line of perfect prediction, which indicates a strong positive correlation between the predicted and actual values across several orders of magnitude of project cost. This visual evidence suggests a reasonable model fit for the central tendency of the final cost distribution (Gautam, 2024).

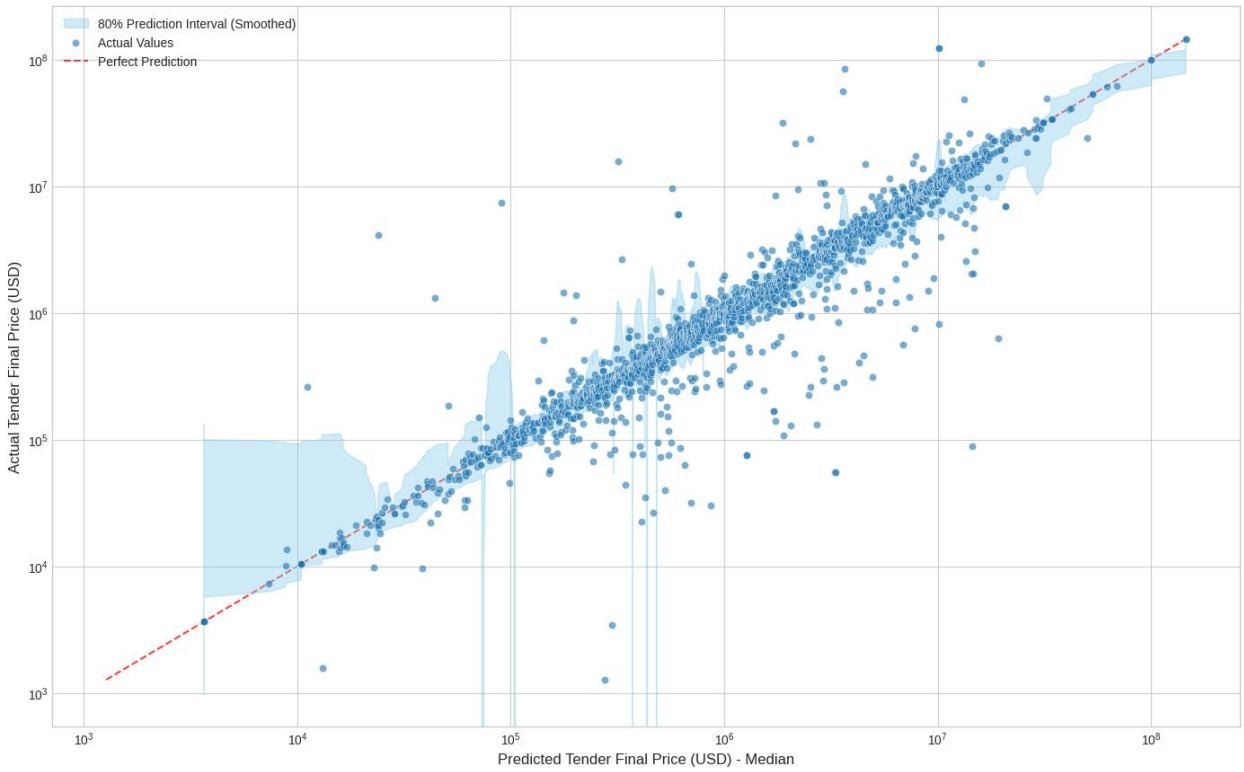

Figure 3 – Validation of the final prediction model on the pristine test set. The log-log scatter plot compares the model's predicted median final prices against actual final prices, with the shaded region representing the 80% prediction interval

Beyond the point-prediction accuracy, the model's capacity for quantifying uncertainty was also assessed. The shaded area in Figure 3 represents the 80% prediction interval, which is bounded by the model's 0.1 and 0.9 quantile predictions. The dependability of this interval is a key measure of the model's utility for risk assessment. The empirical coverage of this interval, as previously reported in the final row of Table 3, was calculated to be 79.85%. This value is proximate to the nominal 80% level, suggesting that the model's uncertainty estimates are well-calibrated. The ability to generate a reliable probabilistic range for potential cost outcomes, rather than only a single point estimate, provides a more comprehensive basis for decision-making in construction cost management.

4.4.2 Model Interpretability

To provide insight into the model's internal logic, a SHAP (SHapley Additive exPlanations) analysis was conducted. This technique assigns an importance value to each feature for every individual prediction, allowing for both global and local model interpretation (Wani, 2024). Figure 4 presents the global feature importance, which ranks the input variables based on their mean absolute SHAP value across all predictions in the test set. The analysis identifies tender estimated price (USD), Buyer Country, and Tender Procedure Type as the three features with the most significant influence on the model's final cost predictions.

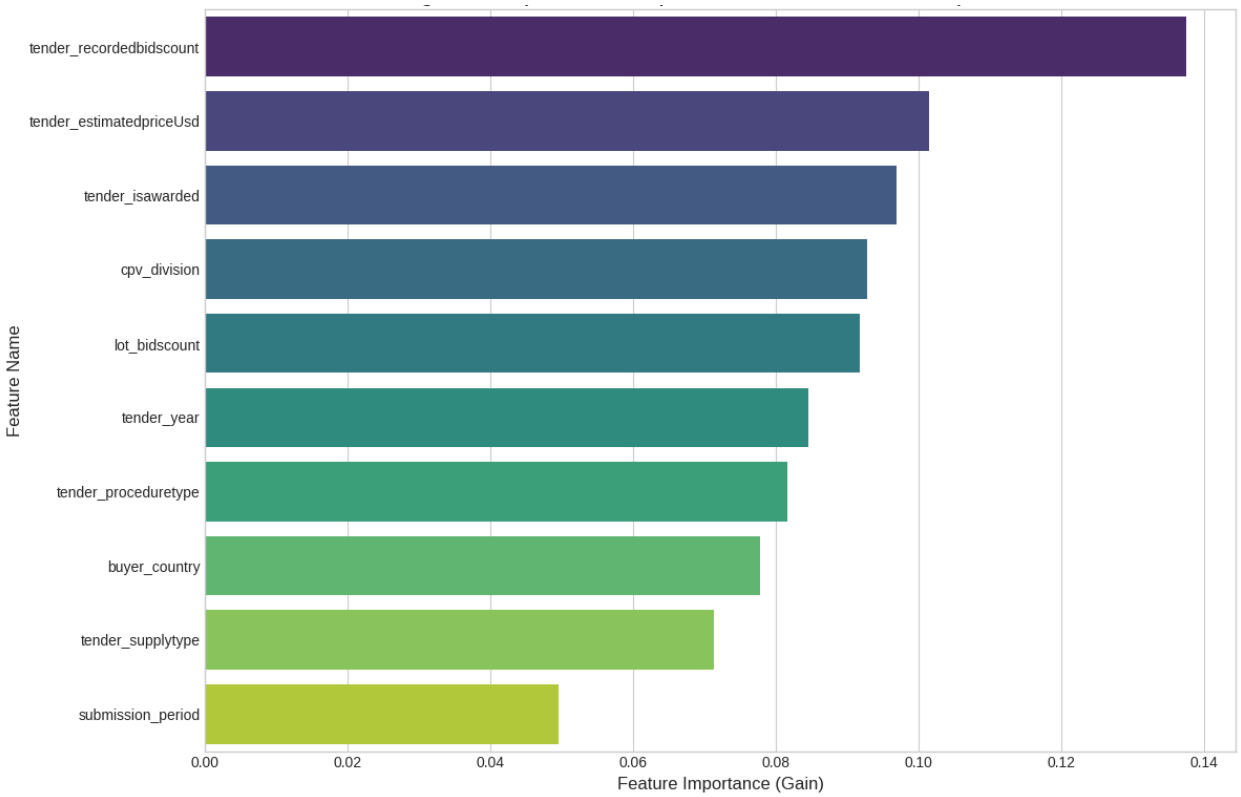

Figure 4 – Global feature importance for the final cost prediction model, as determined by the mean absolute SHAP (SHapley Additive exPlanations) value. Features are ranked by their overall impact on the model's output

A more granular examination of the most influential feature is provided in the SHAP dependence plot shown in Figure 5. This plot illustrates how the value of the tender estimated price (USD) affects the SHAP value, which represents the feature's contribution to pushing the final prediction away from the baseline. A clear and consistent positive relationship is observed, where higher estimated prices are associated with higher positive SHAP values, thereby increasing the predicted final price. The vertical dispersion of points at any given estimated price is colored according to the value of an interaction feature, lot bid count. The plot indicates that for a given tender estimated price, a higher number of bids (represented by cooler colors) tends to be associated with a lower SHAP value. This suggests that the model has learned that a greater level of competition can have a moderating, downward effect on the final predicted cost, an observation that aligns with established principles of competitive bidding in construction procurement. This level of interpretability provides a degree of transparency into the model's behavior, which is essential for building user trust and verifying that the model's logic is consistent with domain knowledge.

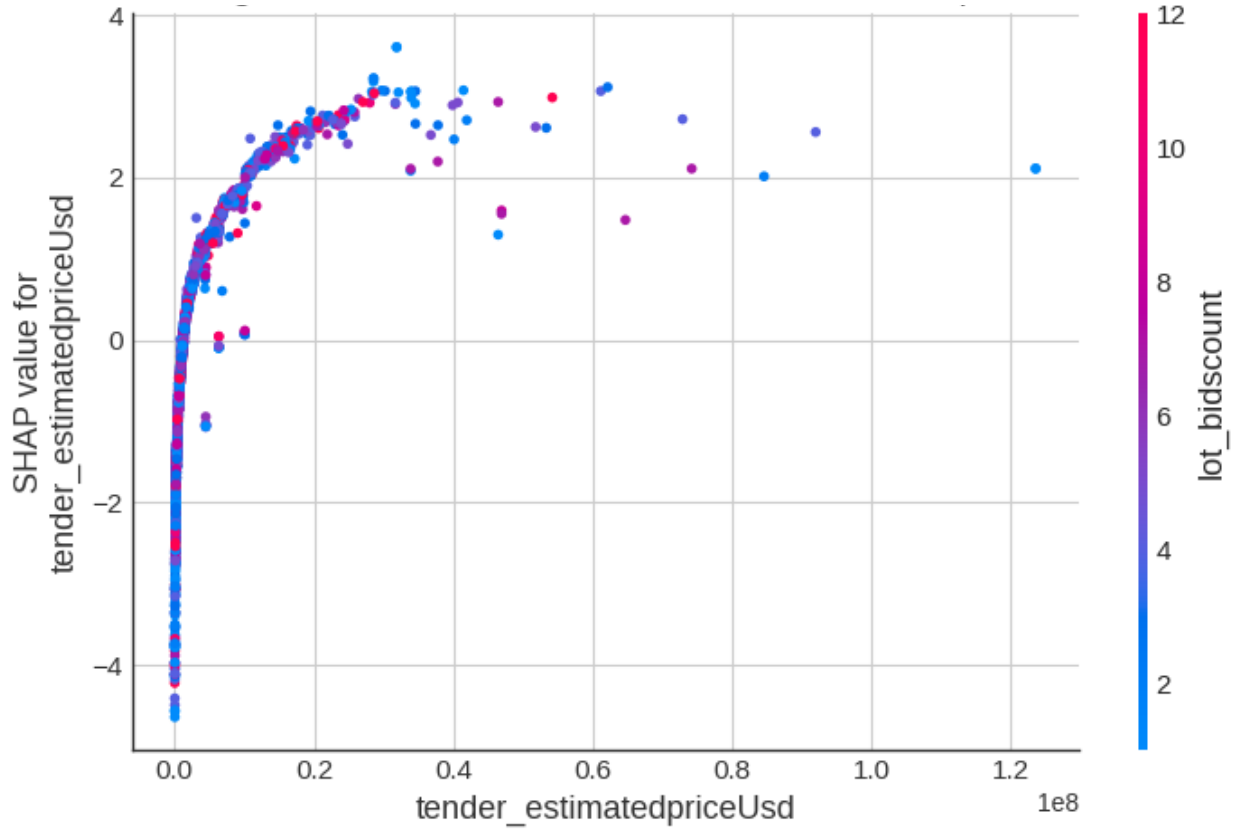

Figure 5 - SHAP dependence plot for the most influential feature, tender estimated price (USD). The plot illustrates the feature's impact on the model's output, with points colored by the lot bid count to show interaction effects.

5. Discussion

5.1 Machine Learning-Based Imputation

The primary contribution of this research is the development and empirical validation of an end-to-end framework for handling missing data in large, complex construction datasets. The approach moves beyond the application of a single imputation technique by providing a structured methodology encompassing outlier handling, comparative model selection, and downstream validation. The research quantifies the tangible benefits of utilizing a machine learning-based imputation approach, demonstrating that it enhances both statistical accuracy and the dependability of subsequent analytical models. The research quantifies the tangible benefits of utilizing a machine learning-based imputation approach, demonstrating that it not only improves statistical accuracy but also enhances the predictive power and dependability of subsequent analytical models. The performance of the XGBoost imputer can be attributed to its ability to model complex, non-linear interactions in high-dimensional space without the restrictive assumptions inherent in the parametric models of MICE or the distance metrics of k-NN, which can be unreliable in heterogeneous feature spaces.

The direct and positive consequence of this higher imputation quality was clearly observed in the downstream validation stage of the study. As shown in the comparative results in Table 3, the use of the XGBoost-imputed dataset for training the final cost prediction model resulted in lower prediction error and better-calibrated uncertainty intervals. This finding demonstrates that the quality of the imputation process has a tangible and significant impact on the utility of the resulting dataset for subsequent analytical tasks, reinforcing the importance of selecting a method that can accurately reflect the underlying data structure.

A key methodological contribution of this study is the conceptual reframing of 'missing data' in construction management. Historically, missing data has been treated merely as a statistical nuisance to be bypassed via listwise deletion or simple mean substitution. This study posits and demonstrates that data missingness in construction procurement is not random, but a structural bias that actively distorts risk profiles and cost distributions. By systematically evaluating imputation methods against downstream predictive utility, this research establishes a new methodological baseline: in the context of modern construction, data imputation must be treated not as a preliminary data-cleaning step, but as a core predictive modeling task in itself. The framework demonstrates that the statistical integrity of the imputation directly dictates the boundaries of what machine learning can achieve in subsequent construction cost forecasting.

This study advances the theoretical discourse on data quality in construction informatics by demonstrating that imputation cannot be decoupled from downstream predictive utility. By treating both missing values and statistical outliers through a unified ML-reconstruction pipeline, the framework provides a methodological framework for researchers seeking to extract reliable predictive insights from structurally compromised construction datasets.

5.2 Practical Implications of the Framework

The findings of this study present several practical implications for both research and practice in construction management. For researchers who work with large-scale public or private datasets, this study provides an empirically validated, multi-stage framework for addressing the pervasive and often-underestimated issue of missing data. The results presented in Table 2 serve as a clear illustration that the choice of imputation method is not a trivial step in the research design. The underperformance of downstream models that were trained on data handled by listwise deletion or simplistic imputation highlights the risk of producing biased or less accurate findings when data quality issues are not rigorously addressed. The proposed framework, which includes outlier detection and a benchmark-driven selection of an imputation model, offers a systematic and transparent approach to enhancing data quality, thereby increasing the dependability of research outcomes.

For industry practitioners, the end-product of this framework is the capacity to develop enhanced data-driven tools for cost management and decision support. The final predictive model demonstrates the capacity to generate not only an accurate point estimate for the final project cost but also a well-calibrated prediction interval. This quantification of uncertainty is of practical value, as it allows project managers to move beyond deterministic cost estimates and engage in enhanced risk analysis and contingency planning. Furthermore, the interpretability of the model through the SHAP analysis helps to build trust in the model's outputs and provides actionable insights into the key drivers of project cost. Understanding that factors such as the initial estimated price and the level of competition (bid count) are primary influencers can inform more strategic decision-making in procurement and project planning phases.

For project managers, the ability to transition from deterministic point estimates to probabilistic cost forecasting represents a substantial improvement in risk management. Traditionally, project contingencies are often set using arbitrary heuristics (e.g., a standard 10% or 15% markup), which frequently leads to either underfunded projects or inefficiently locked capital. The Quantile Gradient Boosting model validated in this framework provides an empirical 80% prediction interval. This allows project managers to dynamically assign contingency budgets based on the specific uncertainty profile of the project. If a project's parameters (e.g., low bid count, specific procurement type) yield a wider prediction interval, management can justify a higher risk premium. Conversely, tighter intervals allow for leaner capital allocation. By trusting the imputed data, managers can make financially optimal decisions backed by a macro-level historical benchmark rather than mere intuition.

Estimators routinely struggle with the 'silo effect,' where they are limited to their own firm's historical data, which is often too sparse to train robust AI models. Public procurement databases offer millions of records, but as this study highlights, the pervasive 78% missingness in final costs renders them practically useless for accurate benchmarking. The XGBoost imputation framework acts as a bridge, utilizing these large-scale datasets and converting them into high-fidelity competitive intelligence. Estimators can now use the completed dataset to benchmark their internal base estimates against industry-wide finalized costs. Furthermore, the SHAP analysis explicitly quantifies how external market factors, such as the number of competing bidders, structurally depress final costs. Estimators can utilize these empirical interaction effects to optimize their margin markups and bidding strategies based on anticipated competition levels.

At the macro level, the findings offer valuable insights for policymakers and public procurement agencies. The study empirically demonstrates the significant analytical limitation of incomplete data reporting. By quantifying the error introduced by conventional

missing-data handling, this research provides the necessary justification for policymakers to enforce stricter data governance and standardized digital reporting protocols across the project lifecycle. Moreover, with a reliably imputed dataset, public owners can conduct portfolio-wide analyses to detect systemic cost overruns in specific regions or project categories. This enables proactive interventions, such as adjusting contract delivery methods or revising public sector budgeting guidelines to better reflect empirical cost distributions.

5.3 Limitations and Future Work

While the proposed framework demonstrates an effective approach to data imputation, certain limitations should be acknowledged to provide context for the findings and to guide future research. First, the study was conducted on a single, substantial public procurement dataset. The performance of the framework and the relative effectiveness of the different imputation models may vary when applied to datasets from different geographical regions, procurement systems, or from the private sector, as these may exhibit different structural characteristics and patterns of missingness. Second, Second, while XGBoost demonstrated effectiveness in this context, the field of machine learning is evolving. These limitations naturally guide several avenues for future research. A logical next step would be to validate the generalizability of the framework by applying it to different construction datasets.

Furthermore, future research could explore the application of deep learning models, such as those based on Generative Adversarial Networks (GANs) or variational autoencoders, for the task of data imputation in construction datasets. These models have shown potential in other domains for their capacity to learn and replicate complex data distributions and could indicate further improvements in imputation quality. A comparative study benchmarking these techniques against the framework presented herein would constitute a valuable contribution to the field.

6. Conclusion

This study addressed the critical challenge of missing data in large-scale construction procurement datasets by developing and applying a systematic, multi-stage framework. This methodology incorporated data cleaning, robust outlier detection utilizing the Isolation Forest algorithm, and a comparative evaluation of data imputation techniques. The empirical results from the imputation model benchmark demonstrated that an XGBoost-based predictive model provided significantly more accurate imputations for the tender final price (USD) compared to conventional statistical methods, including mean imputation, k-NN, and MICE. The utility of this advanced imputation approach was further validated through a downstream predictive modeling task. The evaluation, conducted on a common

pristine test set, demonstrated that a final cost prediction model trained on the dataset completed via the XGBoost framework achieved a higher level of predictive accuracy and more well-calibrated uncertainty quantification than models trained on datasets prepared using listwise deletion or other imputation methods.

These findings indicate that the enhanced quality of the imputed data translates directly to improved performance in subsequent analytical applications. The final prediction model was also shown to be interpretable via SHAP analysis, identifying the tender estimated price (USD) and Buyer Country as key drivers of final project cost. The primary contribution of this work is the development and empirical validation of a transparent, repeatable, and tested end-to-end framework for handling missing data in construction datasets. This work demonstrates that a methodical and empirically-driven approach to data preparation is critical to the success of a data science project in construction management. Future research should prioritize the validation of this framework across diverse geographical and sectoral datasets, as well as the exploration of deep learning models for potentially enhancing imputation quality further.

Ultimately, the value of large-scale public procurement data is defined not by its volume, but by its completeness and reliability. By recovering data that would otherwise be excluded due to high rates of missing final costs, this research equips project managers with data-driven contingency models, provides estimators with macro-level benchmarking tools, and supports the need for stricter digital reporting standards. As the AEC industry continues its transition towards digitized construction, the methodological framework presented herein ensures that foundational AI and machine learning applications are built upon statistically sound, empirically validated, and complete historical data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available on request.

Funding

This research received no external funding.

References

Adebayo, Y., Udoh, P., Kamudyariwa, X.B., Osobajo, O.A., 2025. Artificial Intelligence in Construction Project Management: A Structured Literature Review of Its Evolution in Application and Future Trends. Digital 5. https://doi.org/10.3390/digital5030026

Adhikari, D., Jiang, W., Zhan, J., He, Z., Rawat, D.B., Aickelin, U., Khorshidi, H.A., 2023. A Comprehensive Survey on Imputation of Missing Data in Internet of Things. ACM Comput. Surv. 55, 1–38. https://doi.org/10.1145/3533381

Afkanpour, M., Hosseinzadeh, E., Tabesh, H., 2024. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. BMC Med Res Methodol 24, 188. https://doi.org/10.1186/s12874-024-02310-6

Akinosho, T.D., 2020. Deep learning in the construction industry: A review of present status and future innovations. J. Build. Eng. 32, 101434.

al-Nahhas, Y.S., 2024. Modified Mamdani-fuzzy inference system for predicting the cost overrun of construction projects. Appl. Soft Comput. 151, 111195.

Alreshidi, E.J., 2018. Requirements for cloud-based BIM governance solutions to facilitate team collaboration in construction projects. Requir. Eng. 23, 1–31.

Arabiat, A., 2023. Predicting the construction projects time and cost overruns using K-nearest neighbor and artificial neural network: a case study from Jordan. Asian J. Civ. Eng. 24, 2405–2414.

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, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, M.-Y., 2025a. Hybrid deep learning model for accurate cost and schedule estimation in construction projects using sequential and non-sequential data. Autom. Constr. 170, 105748.

Cheng, M.-Y., 2025b. Hybrid deep learning model for accurate cost and schedule estimation in construction projects using sequential and non-sequential data. Autom. Constr. 170, 105748.

Dastgheib, S.R., 2022. Improving estimate at completion (EAC) cost of construction projects using adaptive neuro-fuzzy inference system (ANFIS). Can. J. Civ. Eng. 49, 222–232.

Deng, N., 2022. Transforming knowledge management in the construction industry through information and communications technology: A 15-year review. Autom. Constr. 142, 104495.

Desse, E.M., 2024. Predicting construction cost under uncertainty using grey-fuzzy earned value analysis. Heliyon 10, e27581.

Durdyev, S., 2020. Review of construction journals on causes of project cost overruns. Eng. Constr. Archit. Manag. 28, 1241–1260.

Elghaish, F., 2021. Blockchain and the 'Internet of Things' for the construction industry: research trends and opportunities. Autom. Constr. 132, 103929.

Eliwa, H.K., 2023. Information and communication technology applications in construction organizations: A scientometric review. J. Inf. Technol. Constr. 28, 286–305.

Elmousalami, H.H., 2020. Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review. J. Constr. Eng. Manag. 146, 03119008.

Fazekas, M., Tóth, B., Abdou, A., Al-Shaibani, A., 2024. Global Contract-level Public Procurement Dataset. Data in Brief 54, 110412.

Gautam, D., 2024. Nonlinear tree based regression ensemble modeling for repair cost prediction in earthquake damaged RC bridges. Soil Dyn. Earthq. Eng. 187, 108428.

Ghadbhan Abed, Y., Hasan, T.M., Zehawi, R.N., 2022. Machine learning algorithms for constructions cost prediction: A systematic review. International Journal of Nonlinear Analysis and Applications 13, 2205–2218. https://doi.org/10.22075/ijnaa.2022.27673.3684

Ghazal, M.M., 2022. Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. Int. J. Constr. Manag. 22, 1632–1646.

Ghimire, P., Kim, K., Acharya, M., 2024. Generative AI in the Construction Industry: Opportunities & Challenges. Buildings 14, 220. https://doi.org/10.3390/buildings14010220

Goger, T., Guggemos, D.O., Borrmann, A., 2021. Construction 4.0: A Roadmap for the Digital Transformation of the AEC Industry. Frontiers in Built Environment 7. https://doi.org/10.3389/fbuil.2021.671408

Gouda Mohamed, A., 2024. Revolutionizing semantic integration of maintenance cost prediction for building systems using artificial neural networks. J. Build. Eng. 96, 109867.

Habib, O., 2025. Ensemble learning framework for forecasting construction costs. Autom. Constr. 170, 105771.

Hu, J., 2024. Prediction of liquefaction of gravelly soils based on a cost-sensitive Bayesian network combined with rough set weighting. Gondwana Res. 131, 57–68.

Ibrahim, A.H., 2021. Assessment of construction project cost estimating accuracy in Egypt. Open Civ. Eng. J. 15, 290–298.

Kazar, G., 2024. Predicting maintenance cost overruns in public school buildings using a rough topological approach. Autom. Constr. 168, 105555.

Kristombu Baduge, S., 2022. Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. Autom. Constr. 141, 104440.

Liang, K., 2024. Data-driven AI algorithms for construction machinery. Autom. Constr. 167, 105527.

Liao, W., 2024. Generative AI design for building structures. Autom. Constr. 157, 105151.

Liu, H., 2025. Actual construction cost prediction using hypergraph deep learning techniques. Adv. Eng. Inform. 65, 102550.

Mohy, A.A., Bassioni, H.A., Elgendi, E.O., Hassan, T.M., 2024. Innovations in safety management for construction sites: the role of deep learning and computer vision techniques. Construction Innovation.

Mostofi, F., 2024. Generating synthetic data with variational autoencoder to address class imbalance of graph attention network prediction model for construction management. Adv. Eng. Inform. 62, 102293.

Mostofi, F., Tokdemir, O., Toğan, V., Arditi, D., 2024. Predicting the Cost of Rework in High-Rise Buildings Using Graph Convolutional Networks. Journal of Construction Engineering and Management 150, 04024085.

Paik, Y., Chung, F., Ashuri, B., 2025. Preliminary Cost Estimation of Pavement Maintenance Projects through Machine Learning: Emphasis on Trees Algorithms. Journal of Management in Engineering 41. https://doi.org/10.1061/JMENEA.MEENG-6623

Rabbi, A.B.K., 2024. AI integration in construction safety: Current state, challenges, and future opportunities in text, vision, and audio based applications. Autom. Constr. 164, 105374.

S. Blažiūnas, A. Raudys, 2019. Comparative Study of Neural Networks and Decision Trees for Application in Trading Financial Futures, in: 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML). Presented at the 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), pp. 33–38. https://doi.org/10.1109/Deep-ML.2019.00015

Sayed, M., 2023. Improving cost estimation in construction projects. Int. J. Constr. Manag. 23, 135–143.

Seu, K., Kang, M.-S., Lee, H., 2022. An intelligent missing data imputation techniques: A review. JOIV: International Journal on Informatics Visualization 6, 278–283.

Shi, Y., Ke, G., Soukhavong, D., Lamb, J., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., 2022. LightGBM: Light gradient boosting machine. R package version 3.

Sun, Y., Li, J., Xu, Y., Zhang, T., Wang, X., 2023. Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems with Applications 227, 120201. https://doi.org/10.1016/j.eswa.2023.120201

Tanguma, J., 2000. A Review of the Literature on Missing Data.

Thomas, T., Rajabi, E., 2021. A systematic review of machine learning-based missing value imputation techniques. Data Technologies and Applications 55, 558–585. https://doi.org/10.1108/DTA-12-2020-0298

Wani, N.A., 2024. Explainable AI-driven IoMT fusion: Unravelling techniques, opportunities, and challenges with Explainable AI in healthcare. Inf. Fusion 110, 102213.

Xiong, R., Wang, Y., Cai, J., Liu, K., Zhu, Y., Tang, P., El-Gohary, N., 2025. OpenConstruction: A Systematic Synthesis of Open Visual Datasets for Data-Centric Artificial Intelligence in Construction Monitoring. https://doi.org/10.48550/arXiv.2508.11482