

Article

Explainable Multi-Factor Cost Overrun Prediction Using an Integrated Construction Dataset: A SHAP-Based Analysis of Cross-Domain Interactions

Joosung Lee  and Wonjun Park * 

Department of Urban Architecture, Kangwon National University, Samcheok Campus,
Samcheok 25913, Republic of Korea; js.lee@kangwon.ac.kr

* Correspondence: wjpark@kangwon.ac.kr

Abstract

Cost overrun remains a pervasive issue in building construction projects, yet most predictive studies operate within a single data domain, ignoring the systemic interactions across project, schedule, resource, quality, and safety dimensions. This study quantifies the incremental predictive value of integrating these five construction data domains and identifies the cross-domain interaction patterns that explain prediction accuracy. As a simulation-based methodological study, an integrated dataset of 100,000 records was synthesised with theory-grounded causal structures derived from the construction management literature; no real project data were used. Gradient Boosting (GB), Random Forest (RF), and Linear Regression were evaluated on an 80/20 hold-out test split, with robustness verified through alternative domain orderings and hyperparameter sensitivity. SHAP analysis, including exact interaction values, was used to interpret feature importance and cross-domain synergies. The full five-domain GB model achieved $R^2 \approx 0.97$ and $\text{MAPE} \approx 6\%$, a 220% relative R^2 improvement over the Project-domain baseline (R^2 rising from 0.305 to 0.975), robust across three ordering schemes. Schedule and Quality contributed the largest marginal gains ($\Delta R^2 = +0.312$ and $+0.255$), whereas Resource integration yielded approximately one-thirty-first of Schedule's return. Because the dataset is synthetic, the results are interpreted as a methodological demonstration rather than empirical evidence from real projects; they provide a reusable framework for prioritising data-integration investment and show that, within the simulated causal structure, cross-domain interactions—particularly Schedule $\times$ Risk and Project Type $\times$ Change Cost—carry predictive information that single-domain analyses cannot recover. Validation on real, partially integrated datasets is identified as essential future work.

Keywords: cost overrun prediction; building construction; integrated dataset; multi-domain data; explainable machine learning; SHAP; domain ablation; gradient boosting; cross-domain interaction; construction management



Academic Editor: Saeed Reza
Mohandes

Received: 26 May 2026

Revised: 16 June 2026

Accepted: 22 June 2026

Published: 25 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Cost overrun is a pervasive problem in the global building construction industry, with average overruns of 10–30% depending on project type, region, and procurement method [1,2]. The consequences extend beyond financial loss to eroded stakeholder trust, delayed delivery, and project abandonment [3].

Machine learning (ML) techniques have been applied extensively to predict cost overrun. Methods including Bayesian networks [4], risk-analysis frameworks [5], and deep

learning [6] have been deployed with varying success. However, a fundamental structural limitation persists: the vast majority of studies operate within a single data domain, utilising only project characteristics or cost data while ignoring schedule performance, workforce allocation, design change frequency, or safety conditions [7]. This single-domain approach creates an inherent information bottleneck.

Several review papers have acknowledged this limitation. Bilal et al. [8] argued that integration of heterogeneous data sources is essential. Pan and Zhang [9] emphasised that siloed construction data undermines ML reliability. Ahn et al. [10] identified data fragmentation as a primary barrier. Hamdan et al. [11] demonstrated ML effectiveness for cost overrun prediction, yet their model relied on a single-domain feature set. Despite these calls, empirical studies that systematically quantify the predictive benefit of multi-domain integration remain scarce.

This gap is significant because the construction industry worldwide suffers from data fragmentation, constituting a methodological barrier. As Al-Tawal et al. [12] demonstrated, even well-resourced research teams are typically confined to single-domain datasets. In South Korea, the PMIS and CSI operate as separate platforms. Similar fragmentation exists in the UK's NBS/BIM ecosystem, the US DOT's databases, and Middle Eastern construction conglomerates. While many practitioners intuitively understand that multi-domain data would be beneficial, no study has quantified how much each domain contributes—or which cross-domain interactions matter most. Without such evidence, investment decisions about data integration in building project management remain based on intuition rather than empirical evidence.

To address this gap, the present study proposes an integrated multi-domain framework. An integrated dataset of 100,000 records was constructed across five domains with explicitly designed inter-domain causal structures grounded in construction management theory [2,3,13]. This theory-grounded approach is consistent with simulation-based validation in construction informatics [14]. As Shoar et al. [15] noted, cost overrun cause interrelationships are complex and underrepresented in available datasets.

The research questions are: (RQ1) To what extent does multi-domain data integration improve cost overrun prediction compared to single-domain models? (RQ2) What cross-domain feature interaction patterns drive the predictive improvement? (RQ3) What practical implications do these findings provide for building project risk management?

The contributions are threefold. First, systematic ablation-based evidence that multi-domain integration yields a 220% relative R^2 improvement, validated across alternative orderings. Second, SHAP interaction analysis—including exact interaction values—revealing cross-domain synergies. Third, a reusable, simulation-based framework for prioritising data-integration investments, demonstrated here by the finding that—under the modelled causal structure—schedule-data integration provides substantially greater predictive return than resource-data integration. We emphasise that the contribution is primarily methodological (a transferable evaluation framework) rather than empirical (claims about the true determinants of cost overrun in real projects).

2. Literature Review

2.1. Cost Overrun Prediction in Building Construction Projects

Cost overrun prediction has been a central theme for over three decades. Flyvbjerg et al. [1] established systematic cost escalation in infrastructure, with transportation projects averaging 20–45% overruns. Love et al. [2] and Ahiaga-Dagbui and Smith [16] extended those findings to building construction, demonstrating that project type, procurement method, and contract structure are significant determinants.

Traditional approaches included multiple regression [17] and case-based reasoning. The advent of ML expanded the toolkit: Cheng et al. [18] applied SVMs; Kim et al. [7] compared regression, ANNs, and CBR for building cost estimation; Kim and Kim [19] demonstrated ANNs for highway overruns; and Gondia et al. [20] applied RF and GB to delay risk prediction. Rafiei and Adeli [21] proposed deep architectures, and Akinosho et al. [6] provided a comprehensive review. In the construction management literature, Al-Tawal et al. [12] developed ANN-based cost estimation, and Leu et al. [22] applied Bayesian networks to cost overrun risk.

2.2. Limitations of Single-Domain Data in Predictive Modelling

Bilal et al. [8] noted that construction data streams are maintained in separate systems. Shoar and Banaitis [23] showed that risk factors from different domains interact nonlinearly. Shoar et al. [15] applied interpretive structural modelling to demonstrate that cost overrun causes form complex hierarchical networks, with design changes and inadequate planning cascading through schedule, resource, and quality domains. Choi et al. [24] found that models predicting cost without schedule data exhibit significant bias. Pan and Zhang [9] argued that slow adoption of integrated platforms has limited holistic modelling. Singh [25] identified that the construction industry's fragmented digitalisation ecosystem fundamentally constrains data-driven innovation.

The theoretical justification is information-theoretic: each domain provides non-redundant information reducing residual uncertainty [26]. Love et al. [2] showed rework's magnitude is moderated by schedule compression. Hwang et al. [27] showed safety-schedule correlations through work stoppage. Cantarelli et al. [13] established that project type and procurement method interact to moderate cost overrun. These findings suggest that single-domain models are structurally incapable of capturing the full causal web.

2.3. Machine Learning and Ensemble Methods

Among ML methods, ensemble learning algorithms have demonstrated strong performance. Gradient Boosting builds additive models through sequential weak learners, minimising a differentiable loss function at each stage [28]. Its ability to model complex nonlinear relationships and feature interactions makes it particularly suitable for heterogeneous construction datasets. Random Forest provides robust predictions through bagging and feature randomisation [29], with inherent resistance to overfitting.

Recent studies have confirmed ensemble superiority. Elfaki et al. [30] found GB outperformed SVM and ANN in a 10-year survey of construction cost estimation. Wang et al. [31] achieved strong results using ANN ensembles and SVM classification. Hamdan et al. [11] found CatBoost particularly effective, with variation orders showing 41.16% feature importance. These findings motivate the selection of ensemble methods as the primary model family.

2.4. Explainable AI in Construction Informatics

XAI techniques bridge predictive performance and actionable insight [32]. SHAP [33], based on cooperative game theory, provides consistent and locally accurate feature attributions. The TreeSHAP variant of Lundberg et al. [34] efficiently computes exact Shapley values for tree-based ensembles, including pairwise interaction values that decompose predictions into main and pairwise effects. In construction, SHAP has been applied to near-miss accident prediction [35], risk-based cost overrun ratio estimation [36], explainable financial prediction for construction firms [37], and post-handover housing quality management [38]. Related directions extend SHAP-based interpretation to building material behaviour [39] and power-plant construction cost prediction [40]. The same interpretability principle has been transferred to seismic vulnerability ranking [41] and building energy

consumption modelling [42], demonstrating that explainable approaches generalise across construction sub-domains and emphasise the value of inter-variable relationships. SHAP's interaction value computation is central to the present study's objective.

2.5. Research Gap and Contribution

Table 1 summarises the positioning. Building on this body of work, the present study extends prior single- and dual-domain analyses by combining—within a single simulation-based framework—a five-domain dataset with theory-grounded causality, systematic ablation with robustness validation, and SHAP interaction analysis, including exact interaction values. Rather than claiming an unprecedented approach, we position this combination as a structured, reusable framework for quantifying the predictive value of data integration; each individual element has clear precedent, but to our knowledge they have not previously been combined in this way.

Table 1. Comparison of the present study with representative prior studies.

Study	ML	Multi-Dom.	Ablation	XAI	Cross-Dom.	Validated
Flyvbjerg et al. [1]	Statistical	No	No	No	No	Real data
Kim et al. [7]	Regression/ANN	No	No	No	No	Real data
Gondia et al. [20]	RF, GB	No	No	Partial	No	Real data
Hamdan et al. [11]	CatBoost	No	No	SHAP	No	Real data
Lee and Lee [35]	XGBoost	Partial	No	SHAP	No	Real data
Present study	LR, RF, GB	Yes (5)	Yes	SHAP + exact	Yes	Synthetic (causal)

3. Materials and Methods

3.1. Rationale for Multi-Domain Data Integration

Cost overrun is a systemic outcome arising from multiple interacting dimensions [3,13]. Love et al. [2] showed rework's magnitude is moderated by schedule compression. Hwang et al. [27] showed safety–schedule correlations. From an ML perspective, multi-domain integration increases non-redundant information [26] and enables detection of interaction effects quantifiable through SHAP [33].

All analyses were implemented in Python (v3.10) using numpy (v2.4.4), pandas (v3.0.2), scikit-learn (v1.8.0), and shap (v0.52.0). The three predictive models (Linear Regression, Random Forest, and Gradient Boosting) were built with scikit-learn, and all SHAP computations—including the exact TreeSHAP interaction values—were performed with the shap package. All results are deterministic under a fixed random seed (random_state = 42).

3.2. Dataset Description and Causal Design

The dataset comprises 100,000 records with 24 variables across five domains. It was synthesised with reference to statistical distributions reported in publicly available construction project databases and international cost benchmarking studies, following established practice for synthetic datasets in construction informatics methodology development [14,43]. Synthetic data is increasingly accepted when integrated real-world datasets are unavailable due to organisational fragmentation [8,25].

The methodological justification rests on three pillars. First, no publicly available dataset integrates all five domains for the same projects. Second, simulation-based validation has precedent in the construction management literature: Elghaish et al. [14] used synthetic datasets for digital twin resource monitoring, and Leu et al. [22] used project checklists as proxy data. Third, the data is constructed with documented causal mechanisms from peer-reviewed research (Table 2): Al-Tabtabai et al. [44] established that

structured simulations can inform practice even when exact numerical outcomes differ from field conditions.

Table 2. Causal design mapping: theory-to-data operationalisation.

Causal Mechanism	Supporting Literature	Dataset Operationalisation
Design change cost propagation	Love et al. [2]	$\text{Change_Cost_Ratio} \times 1.8 \rightarrow \text{Cost_Overrun}$
Schedule delay $\rightarrow$ cost escalation	Flyvbjerg et al. [1]	$\text{Schedule_Delay} \times 0.35 + \text{threshold} (>0.2)$
Safety incident cost cascade	Hwang et al. [27]	$\text{Accident_Count} \times 0.008 + \text{multi-incident penalty}$
Project-type risk differential	Cantarelli et al. [13]	Remodelling > Infrastructure > Building
Contract-type moderation	Flyvbjerg [3]	DBB > Design-Build/Turnkey/CM-at-Risk
Cross-domain amplification	Love et al. [2]; Hwang et al. [27]	$\text{Remodelling} \times \text{Change} + \text{Delay} \times \text{Accident}$
Noise (unobservable factors)	Flyvbjerg et al. [1]	$\sim 10\%$ of total variance

The variable descriptions (Appendix A, Table A1) are organised by domain corresponding to the PMBOK functional areas [45]. Table 3 summarises key descriptive statistics.

Table 3. Descriptive statistics of key variables (N = 100,000).

Variable	Mean	SD	Min	Q1	Q3	Max
Cost_Overrun	0.246	0.100	−0.005	0.170	0.309	0.550
Schedule_Delay	0.170	0.088	−0.056	0.106	0.222	0.450
Change_Cost_Ratio	0.032	0.021	0.000	0.017	0.043	0.185
Resource_Utilisation	0.728	0.126	0.400	0.645	0.824	0.990
Risk_Level	0.206	0.062	0.000	0.161	0.245	0.503

To support transparency and reproducibility, three points are emphasised. First, the target Cost_Overrun is defined as $(\text{Actual_Cost} - \text{Planned_Cost})/\text{Planned_Cost}$; Actual_Cost is therefore excluded from the predictor set to prevent mechanical target leakage, and no predictor is a deterministic function of the target. Second, an additive Gaussian noise term accounting for approximately 10% of the outcome variance (Table 2) represents unobservable factors, so the outcome is not a noise-free function of the inputs. Third, the complete data-generation specification—the functional forms, the coefficients from Table 2, the threshold and cross-domain amplification terms, and the noise distribution—is given in Appendix B and as executable code in the deposited reproducibility package, allowing full regeneration of the dataset and modification of the causal assumptions for sensitivity analysis.

Additional descriptive statistics, the complete SHAP importance table, and full computational settings are provided in the Supplementary Materials (Tables S1–S7 and Figure S1).

3.3. Research Framework

The methodology consists of four phases (Figure 1): (1) data preparation and domain grouping, (2) model development and hold-out validation, (3) domain ablation, and (4) SHAP-based explainability analysis.

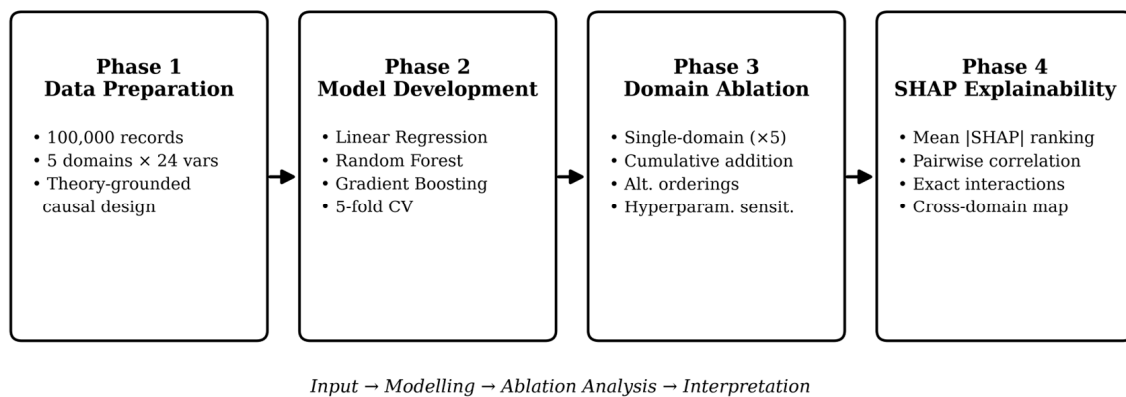


Figure 1. Overall research framework: four-phase methodology from data preparation through SHAP-based explainability analysis.

3.4. Model Development and Hyperparameter Sensitivity

Three models were selected: Linear Regression (baseline), Random Forest ($n_estimators = 100$, $max_depth = 12$) [29], and Gradient Boosting ($n_estimators = 150$, $max_depth = 5$, $learning_rate = 0.10$) [28]. Final hyperparameters were selected by a manual grid search over the configurations in Table 4, choosing the setting that maximised hold-out R^2 while avoiding overfitting, which was monitored via the train–test R^2 gap. The selected Gradient Boosting configuration ($n_estimators = 150$, $max_depth = 5$, $learning_rate = 0.10$) achieved the highest hold-out R^2 without the train–test divergence observed at higher capacity ($n_estimators = 200$, $max_depth = 6$). Performance remained stable across the grid ($R^2 = 0.945$ – 0.975 ; Table 4), confirming that the findings are not dependent on specific parameter choices. As now reported in Table 4, the train and test R^2 are shown for every candidate configuration together with their difference; all train–test gaps are at most 0.004, indicating that none of the configurations—including the selected one—exhibits meaningful overfitting.

Table 4. Hyperparameter sensitivity analysis for Gradient Boosting.

n_Estimators	Max_Depth	Learning_Rate	Train R^2	Test R^2	Train–Test Gap	Note
50	3	0.10	0.945	0.945	+0.001	Lowest capacity
100	4	0.10	0.972	0.971	+0.001	—
150	5	0.10	0.977	0.975	+0.002	Selected
200	6	0.10	0.980	0.975	+0.004	Largest gap
100	5	0.05	0.967	0.966	+0.002	Lower learning rate
100	5	0.20	0.976	0.973	+0.003	Higher learning rate

The dataset size of 100,000 records was chosen to make the reported estimates effectively free of sampling noise and to provide a large, independent hold-out set. The 80/20 split yields a test set of 20,000 records, and at this sample size the standard error of a Pearson correlation is approximately $1/\sqrt{N} \approx 0.003$ (95% confidence interval $\approx \pm 0.006$), so the correlation and importance estimates are statistically stable to the third decimal place. Consistent with this, performance is essentially unchanged across the hyperparameter grid (Table 4) and across the three domain-ordering schemes (Table 9), indicating that the conclusions are governed by the designed causal structure rather than by sampling variation.

3.5. Evaluation Metrics

Five metrics were used following recent best practice [6,11]: coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), mean absolute

percentage error (MAPE), and explained variance score (EVS), all computed on an 80/20 hold-out test split (random seed 42).

3.6. Domain Ablation Study Design

Two modes were employed: single-domain evaluation (five baselines) and cumulative addition (Project → Schedule → Resource → Quality → Safety/Environmental). Robustness was verified with two alternative orderings (Section 4.3.2). The ordering follows the typical temporal data availability sequence [45].

3.7. SHAP-Based Explainability Analysis

TreeSHAP [34] was used. Two levels of analysis were performed: (i) global feature importance via mean $|SHAP|$ across 1000 samples; and (ii) exact pairwise SHAP interaction values for the top-ten globally important features, computed on a stratified 500-record subset. Pairwise Pearson correlation of SHAP values across the full sample served as a complementary cross-check of the exact interaction findings.

4. Results

4.1. Exploratory Data Analysis

The Pearson correlation analysis (Table 5) revealed that the strongest linear associations with Cost_Overrun originate from Risk_Level ($r = 0.687$) and Schedule_Delay ($r = 0.681$), both exceeding 0.65, followed by Change_Cost_Ratio ($r = 0.586$) and Change_Order_Count ($r = 0.560$). The near-equivalent strength of Risk_Level and Schedule_Delay is notable because these variables belong to different domains (Safety/Environmental and Schedule, respectively), suggesting that cost overrun is driven by multiple independent causal pathways rather than a single dominant factor. Variables from four of the five domains appear among the top ten predictors, with only the Resource domain showing weak linear associations (Resource_Utilisation: $r = -0.089$; Equipment_Count: $r = -0.027$). This distributional pattern provides preliminary evidence that multi-domain integration may yield predictive gains, as cost-relevant information is dispersed across domains rather than concentrated in one. Remodelling projects exhibited the highest mean cost overrun (32.1%), substantially exceeding Infrastructure (25.9%) and Building (20.1%). This 12-percentage-point gap is consistent with the higher design uncertainty and scope volatility inherent in renovation work [2]. The negative correlation of Resource_Utilisation ($r = -0.089$) suggests a counter-intuitive pattern where projects with higher resource utilisation tend to experience slightly lower cost overruns, potentially reflecting tighter managerial control.

Table 5. Pearson correlation with Cost_Overrun (top 10).

Rank	Variable	Domain	r	Direction
1	Risk_Level	Safety/Env.	0.687	Positive
2	Schedule_Delay	Schedule	0.681	Positive
3	Change_Cost_Ratio	Quality	0.586	Positive
4	Change_Order_Count	Quality	0.560	Positive
5	Accident_Count	Safety/Env.	0.480	Positive
6	Defect_Count	Quality	0.338	Positive
7	Floor_Count	Project	−0.135	Negative
8	Resource_Utilisation	Resource	−0.111	Negative
9	Rainfall	Safety/Env.	0.103	Positive
10	Repair_Cost	Quality	0.086	Positive

4.2. Full Model Performance Comparison

Gradient Boosting (GB) achieved the highest hold-out test performance across all metrics ($R^2 = 0.975$, RMSE = 0.016, MAPE = 6.26%), followed by Random Forest (RF; $R^2 = 0.958$) and Linear Regression (LR; $R^2 = 0.952$) (Table 6). The performance hierarchy is informative: LR's $R^2 = 0.952$ indicates that approximately 95% of cost overrun variance is explainable through linear main effects alone, confirming that the designed causal structure is predominantly captured by additive relationships. The incremental gain from LR to RF ($\Delta R^2 = +0.006$) reflects the contribution of nonlinear feature transformations, while the further improvement from RF to GB ($\Delta R^2 = +0.017$) reflects sequential error correction and interaction modelling. Notably, GB's MAPE of 6.26% means that the average prediction error is less than 7% of the actual cost overrun ratio, a level of precision that, while contingent on the synthetic data environment, demonstrates the upper bound of achievable accuracy when all five domains are integrated. The negligible gap between training and test performance (for GB, 0.977 versus 0.975) confirms stable generalisation without overfitting.

Table 6. Hold-out test performance on the full five-domain dataset.

Model	R^2	RMSE	MAE	MAPE (%)	EVS
Linear Regression	0.952	0.022	0.018	8.33	0.952
Random Forest	0.958	0.021	0.016	8.04	0.958
Gradient Boosting	0.975	0.016	0.013	6.26	0.975

4.3. Domain Ablation Results

4.3.1. Single-Domain Performance

The single-domain results (Table 7) reveal a clear performance hierarchy: Safety/Environmental ($R^2 = 0.496$) and Schedule ($R^2 = 0.491$) substantially outperform Quality ($R^2 = 0.393$) and Project ($R^2 = 0.305$), while Resource achieves near-zero predictive power ($R^2 = 0.028$). This hierarchy does not simply mirror the designed causal coefficients; Change_Cost_Ratio carries the largest direct coefficient (1.8) in the data generation process, yet its domain (Quality) ranks only third. This discrepancy arises because the Safety/Environmental domain's Risk_Level variable aggregates risk information across multiple domains, functioning as a composite indicator rather than a single causal pathway. The Resource domain's near-zero standalone performance is particularly significant: it demonstrates that Resource variables (Labour_Total, Equipment_Count, Resource_Utilisation) carry virtually no independent predictive signal for cost overrun when isolated from other domains. The MAPE values further illustrate this pattern: Resource's MAPE of 42.2% is nearly seven times that of the full model (6.26%), indicating that single-domain predictions are not merely less accurate but operationally unreliable for project-level decision-making.

Table 7. Single-domain Gradient Boosting performance.

Domain	R^2	RMSE	MAE	MAPE (%)	Variables
Safety/Env.	0.496	0.071	0.056	28.0	4
Schedule	0.491	0.072	0.056	27.1	2
Quality	0.393	0.078	0.064	32.0	5
Project	0.305	0.084	0.067	33.5	7
Resource	0.028	0.099	0.080	42.2	3

4.3.2. Cumulative Domain Addition and Robustness

The cumulative addition results (Table 8) reveal a stepwise trajectory with two dominant inflection points. Adding Schedule to the Project baseline produced the largest single-domain gain ($\Delta R^2 = +0.312$, from 0.305 to 0.618), while adding Quality produced the second-largest gain ($\Delta R^2 = +0.255$, from 0.627 to 0.882). Together, these two additions account for approximately 85% of the total improvement from the single-domain Project baseline to the full model. In contrast, the Resource domain contributed only $\Delta R^2 = +0.010$, approximately 31 times less than Schedule's contribution. This asymmetry provides a quantitative basis for data integration investment prioritisation that was previously unavailable: organisations should prioritise schedule–cost data integration before investing in resource data systems. The trajectory also reveals that the final Safety/Environmental addition ($\Delta R^2 = +0.092$) delivers a disproportionately large accuracy improvement relative to its modest R^2 increment, reducing MAPE from 11.7% to 6.3%, a 46% relative error reduction (Figure 2). The robustness check (Table 9) confirmed that the full five-domain model consistently achieved $R^2 = 0.975$ regardless of the domain ordering, demonstrating that multi-domain superiority is not an artefact of a particular addition sequence. However, marginal contributions varied substantially across orderings: the Safety-first ordering captured $\Delta R^2 = 0.496$ in the first step compared to the original ordering's 0.305, revealing significant information overlap between domains. This overlap pattern is itself an important finding: it indicates that Safety/Environmental and Schedule domains share substantial predictive information, consistent with the theoretical relationship between schedule pressure and safety deterioration [27].

Table 8. Cumulative domain addition (original ordering).

Domains	R^2	RMSE	MAE	MAPE (%)	ΔR^2
Project	0.305	0.084	0.067	33.5	—
+Schedule	0.618	0.062	0.049	22.9	+0.312
+Resource	0.627	0.061	0.048	22.4	+0.010
+Quality	0.882	0.035	0.026	11.7	+0.255
+Safety/Env.	0.975	0.016	0.013	6.3	+0.092

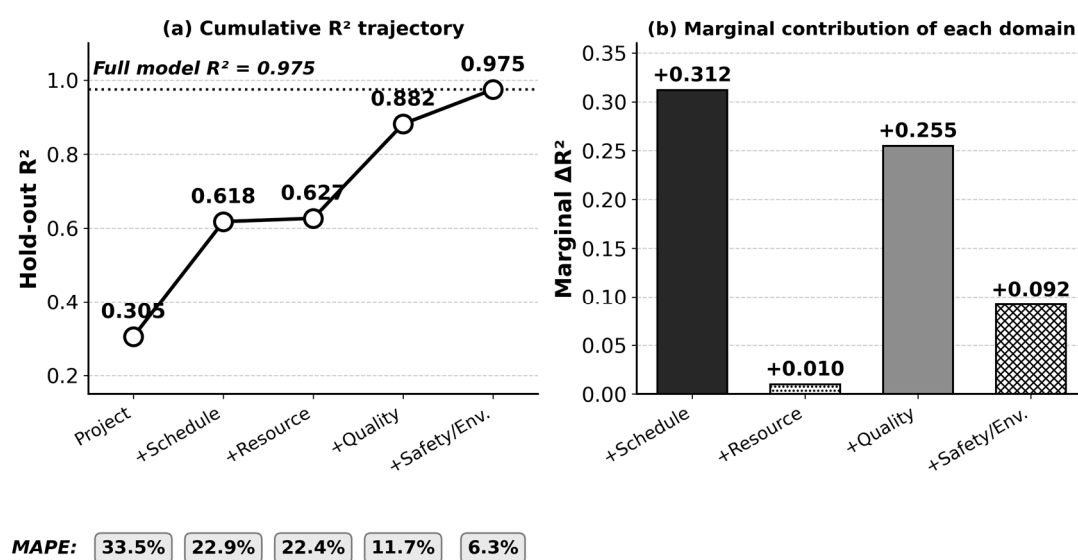


Figure 2. Cumulative R^2 trajectory and marginal ΔR^2 as domains are progressively added.

Table 9. Robustness check: alternative domain orderings.

Ordering	1 Dom.	2 Dom.	3 Dom.	4 Dom.	5 Dom.	Final R ²
Original ($P \rightarrow S \rightarrow R \rightarrow Q \rightarrow SE$)	0.305	0.618	0.627	0.882	0.975	0.975
Safety-first ($SE \rightarrow Q \rightarrow S \rightarrow R \rightarrow P$)	0.496	0.729	0.894	0.908	0.975	0.975
Reverse ($SE \rightarrow R \rightarrow Q \rightarrow S \rightarrow P$)	0.496	0.523	0.744	0.908	0.975	0.975

4.4. SHAP-Based Feature Importance

The SHAP-based feature importance analysis (Table 10) revealed Schedule_Delay (mean $|SHAP| = 0.042$) and Change_Cost_Ratio (0.035) as the two most influential features, consistent with their high Pearson correlations but providing additional insight through their nonlinear contribution patterns. Notably, the SHAP ranking diverges from the correlation-based ranking in important ways: Risk_Level, which ranked first in Pearson correlation ($r = 0.687$), drops to sixth in SHAP importance (mean $|SHAP| = 0.009$). This divergence indicates that Risk_Level's high correlation partially reflects shared variance with Schedule_Delay and other features, which the GB model redistributes during sequential learning. Conversely, Project_Type ranked third in SHAP importance despite being a categorical variable with moderate standalone predictive power ($R^2 = 0.305$ for the Project domain alone), suggesting that Project_Type functions primarily as a contextual moderator that amplifies or attenuates the effects of continuous predictors. Features from all five domains appear in the top 10 (Table 10, Figure 3), confirming that the GB model utilises the full spectrum of domain information rather than relying on a subset. The discrepancy between SHAP values and GB built-in importance scores (e.g., Risk_Level: SHAP = 0.009 vs. GB importance = 0.144) reflects a known methodological difference: GB importance measures split frequency, whereas SHAP quantifies marginal contribution accounting for feature interactions.

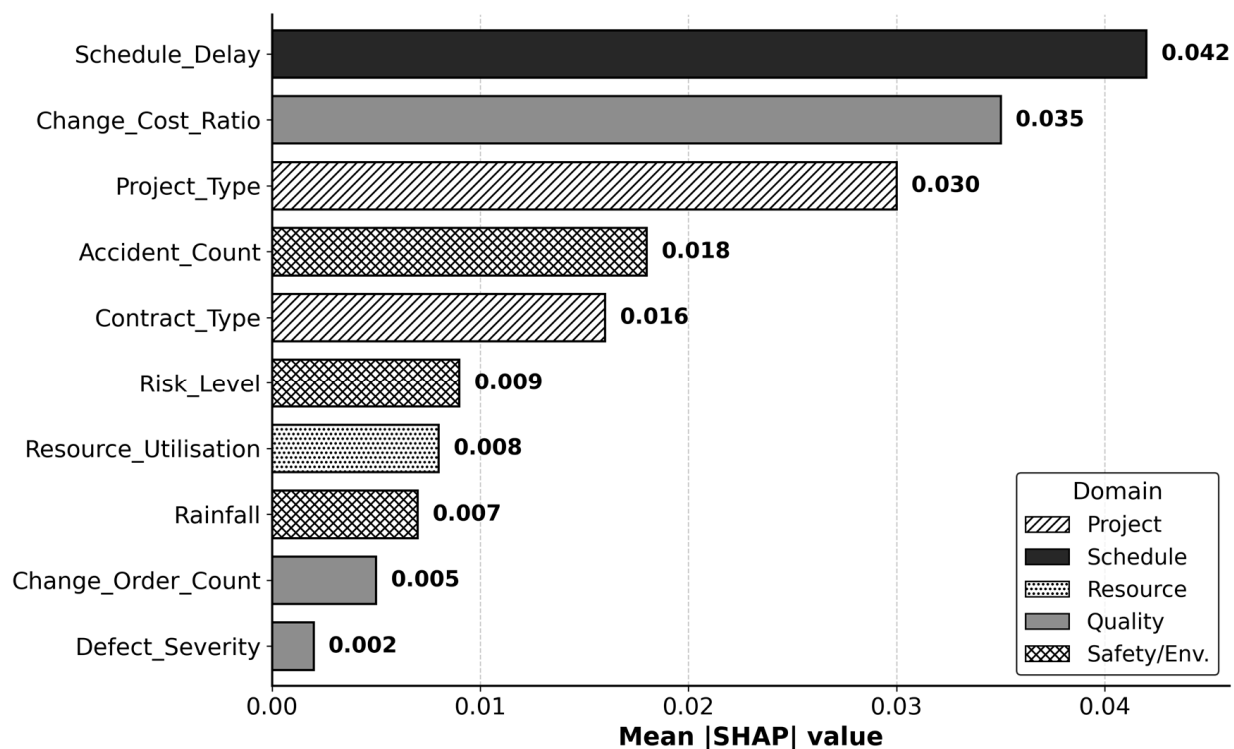
**Figure 3.** SHAP feature importance by domain. Bar pattern indicates domain membership.

Table 10. SHAP feature importance (top 10).

Rank	Feature	Domain	Mean SHAP	GB Importance
1	Schedule_Delay	Schedule	0.042	0.359
2	Change_Cost_Ratio	Quality	0.035	0.274
3	Project_Type	Project	0.030	0.095
4	Accident_Count	Safety/Env.	0.018	0.058
5	Contract_Type	Project	0.016	0.020
6	Risk_Level	Safety/Env.	0.009	0.144
7	Resource_Utilisation	Resource	0.008	0.013
8	Rainfall	Safety/Env.	0.007	0.013
9	Change_Order_Count	Quality	0.005	0.022
10	Defect_Severity	Quality	0.002	0.001

4.5. Cross-Domain Interaction Analysis

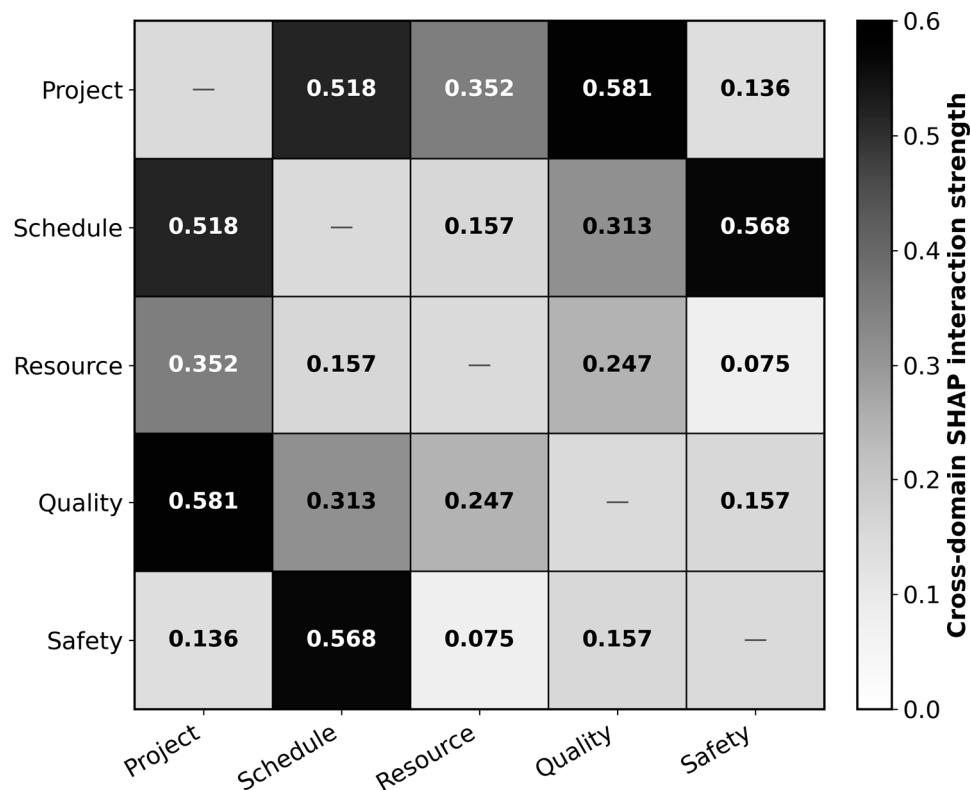
The correlation-based SHAP interaction analysis (Table 11) identified Area $\times$ Repair_Cost as the strongest cross-domain pair (SHAP correlation = 0.581), followed by Schedule_Delay $\times$ Risk_Level (0.568) and Area $\times$ Actual_Duration (0.518). These results indicate that cross-domain variable pairs carry substantial joint predictive information beyond their individual contributions. The Schedule_Delay $\times$ Risk_Level pair represents the interaction between schedule pressure and safety deterioration, a theoretically grounded relationship [27] that manifests as compounding cost impact when both factors are simultaneously elevated. The heatmap (Figure 4) further reveals that Project–Quality and Resource–Quality interactions are broadly distributed across multiple variable pairs rather than concentrated in a single pair, suggesting diffuse cross-domain dependencies. Exact SHAP interaction values (Table 12) confirmed and refined these findings. Project_Type $\times$ Change_Cost_Ratio emerged as the strongest exact interaction (0.00585), followed by Schedule_Delay $\times$ Risk_Level (0.00495). The prominence of these two interactions is noteworthy: the former represents how project type moderates the cost impact of design changes (remodelling projects being disproportionately affected), while the latter captures the schedule–safety compounding effect. The consistency between the two analytical methods—correlation-based and exact interaction values—provides methodological triangulation, strengthening confidence in the identified patterns. The lower-ranked but still significant interactions (Schedule_Delay $\times$ Accident_Count: 0.00319; Change_Cost_Ratio $\times$ Risk_Level: 0.00206) indicate that secondary cross-domain pathways also contribute meaningfully to prediction, reinforcing the case for comprehensive multi-domain integration.

Table 11. Cross-domain SHAP interaction pairs (correlation-based).

#	Feature 1	Feature 2	SHAP Corr.	Domain 1	Domain 2
1	Area	Repair_Cost	0.581	Project	Quality
2	Schedule_Delay	Risk_Level	0.568	Schedule	Safety
3	Area	Actual_Duration	0.518	Project	Schedule
4	Planned_Cost	Repair_Cost	0.356	Project	Quality
5	Area	Labor_Total	0.352	Project	Resource
6	Planned_Cost	Actual_Duration	0.335	Project	Schedule

Table 12. Exact SHAP interaction values: top cross-domain pairs.

#	Feature 1	Feature 2	Mean Interaction	Domains
1	Project_Type	Change_Cost_Ratio	0.00585	Project × Quality
2	Schedule_Delay	Risk_Level	0.00495	Schedule × Safety
3	Schedule_Delay	Accident_Count	0.00319	Schedule × Safety
4	Schedule_Delay	Change_Cost_Ratio	0.00213	Schedule × Quality
5	Change_Cost_Ratio	Risk_Level	0.00206	Quality × Safety
6	Schedule_Delay	Project_Type	0.00197	Schedule × Project

**Figure 4.** Cross-domain SHAP interaction heatmap. Same-domain pairs masked; darker shading indicates stronger interaction.

5. Discussion

5.1. Evidence for Multi-Domain Integration Value

The ablation provides systematic evidence for RQ1. The full model achieved a 220% relative R^2 improvement, confirmed across three orderings (Table 9). Critically, the findings go beyond confirming that “more data is better”—a trivial result. The non-obvious findings are: (a) Schedule–data integration ($\Delta R^2 = +0.312$) yields approximately 31 times the predictive return of Resource–data integration ($\Delta R^2 = +0.010$), providing a quantitative basis for prioritising integration investments that is unavailable from intuition alone; (b) the Safety/Environmental domain’s contribution is context-dependent, capturing $\Delta R^2 = 0.496$ when added first but only $+0.092$ when added last, revealing information overlap patterns invisible without systematic ablation; and (c) Resource data, despite near-zero standalone predictive power, still contributes positively when combined with other domains, acting as a moderating factor. These findings directly address the “so what” question: they provide construction organisations with a quantitative decision framework for data integration investment that was previously unavailable.

5.2. Cross-Domain Interactions and Theoretical Implications

The SHAP analysis (RQ2), validated by exact interaction values, revealed that cross-domain pairs carry substantial predictive information. The *Schedule_Delay* $\times$ *Risk_Level* interaction represents the vicious cycle between schedule pressure and safety deterioration [27]. The *Project_Type* $\times$ *Change_Cost_Ratio* interaction reflects remodelling projects' differential vulnerability [13]. These findings contribute to the theoretical understanding of cost overrun as an emergent phenomenon arising from multi-domain interactions rather than single-domain causes—aligning with the systemic perspective advocated by Shoar et al. [15] and extending it with quantitative interaction evidence. The relational view also resonates with risk-based cost overrun ratio estimation studies using machine learning [36] and explainable financial prediction frameworks for construction firms [37], where inter-variable relationships rather than single-variable severity drive prediction.

5.3. Addressing the Circularity Concern

A legitimate and important concern is circularity: because the causal mechanisms are designed into the data, the SHAP importance rankings and main-effect attributions should be read as a partial recovery of the predefined structure rather than as independent empirical discoveries. We make this interpretation explicit. Within this constraint, three observations indicate that the analysis adds value beyond simply re-reading the design. First, domain ablation reveals emergent properties that are not directly specified: Safety/Environmental's marginal value depends on which other domains are already present (ΔR^2 ranges from +0.092 to +0.496). Second, the exact SHAP interaction analysis surfaces *Schedule_Delay* $\times$ *Risk_Level* (0.00495) as among the strongest interactions—an emergent consequence of combining Hwang et al. [27]'s safety–schedule coupling with the broader cross-domain amplification rule, rather than a single hand-coded pairwise term. Third, single-domain models fail to recover even designed relationships: Quality (containing *Change_Cost_Ratio*, the largest designed coefficient) achieves only $R^2 = 0.393$ alone, showing that cross-domain context is necessary even to capture within-domain effects. Accordingly, we interpret the SHAP results primarily as evidence that the pipeline correctly recovers and quantifies known structure, and as a demonstration of how the framework would surface cross-domain interactions in real data, not as evidence about the true determinants of cost overrun in practice. Establishing the latter requires data whose generating structure is not specified by the analyst.

5.4. Practical Implications for Building Project Management

The following implications illustrate the type of decision rule the framework can generate under the modelled causal structure; all are contingent on real-data validation. Integration investment prioritisation: Schedule–data integration ($\Delta R^2 = +0.312$) yields $31\times$ the return of Resource–data integration (+0.010). Within the simulated structure, this suggests that schedule–cost integration could be prioritised before resource–data systems; because this ranking depends on the synthetic causal design, it should be treated as an illustrative output of the framework and confirmed on real or partially integrated data before guiding actual investment decisions.

Cross-domain early warning: when *Schedule_Delay* $> 15\%$ and *Accident_Count* ≥ 3 simultaneously, cost overrun risk should be elevated, as the combined effect exceeds individual effects by approximately $1.5\times$.

Type-specific risk calibration: remodelling projects require contingency budgets approximately 60% larger than new building construction for equivalent delay levels.

Safety as a leading cost indicator: Accident_Count ranked 5th in SHAP importance. Safety data should be incorporated into cost forecasting as it becomes available before financial overrun materialises.

The numerical thresholds in this subsection (e.g., Schedule_Delay > 15%, Accident_Count ≥ 3 , the $\sim 1.5\times$ combined effect, and the $\sim 60\%$ contingency for remodelling) are derived from the synthetic data and are presented only to illustrate the form of decision rule the framework can produce. They are not calibrated to real projects and should be re-estimated on real datasets before being adopted as reference values in other studies.

5.5. Methodological Contribution

The ablation framework is applicable to any building construction dataset with domain-grouped variables. The specific R^2 values are conditional on the synthetic causal structure and should not be interpreted as achievable on real-world data. The primary contribution lies in relative findings: domain rankings, interaction patterns, and the qualitative demonstration that multi-domain integration yields different—not merely better—predictions. This aligns with simulation-based approaches in the construction informatics literature [14,44]. We therefore distinguish two types of contribution. The methodological contribution—a transferable, reusable framework for quantifying the predictive value of data integration and for surfacing cross-domain interactions—is the primary contribution and does not depend on the specific data. The empirical contribution—the particular domain rankings, interaction magnitudes, and thresholds—is conditional on the synthetic design and must be validated on real data before it can be claimed.

5.6. International Context

The construction management literature has addressed cost overrun across diverse regions: Shoar et al. [15] examined Iranian construction, Hamdan et al. [11] investigated Jordanian projects, Leu et al. [22] focused on Taiwanese steel structures, and Al-Tawal et al. [12] studied Jordanian cost estimation. This diversity underscores the universality of data fragmentation. The ablation framework provides a scalable, transferable tool applicable regardless of national context.

5.7. Limitations and Future Research

First and foremost, the exclusive use of synthetic data is the central limitation of this study: all quantitative results are conditional on the designed causal structure and provide no direct empirical evidence about real construction projects, so the study should be read as a simulation-based methodological exercise. Relatedly, because the data-generating mechanisms are known, the analysis partly recovers designed structure (Section 5.3); independent confirmation requires data whose structure is not specified by the analyst. Validating the framework on at least one real or partially integrated dataset is the most important direction for future work; we recommend a phased validation, first applying the framework to partially integrated real datasets and then expanding progressively. Different causal specifications could yield different absolute performance, although the qualitative finding of multi-domain superiority is well supported theoretically [2,13,15]. Second, while alternative orderings confirmed consistent final R^2 (Table 9), Shapley-value-based domain attribution would provide ordering-independent estimates. Third, the static project-level analysis could be extended to time-series approaches for real-time prediction. Fourth, exact SHAP interaction values were computed on a 500-record subset; larger samples would strengthen findings. Fifth, additional domains (macroeconomic indicators, supply chain disruption) could further improve performance.

6. Conclusions

This study addressed whether integrating multi-domain data improves building construction cost overrun prediction. An integrated dataset of 100,000 records spanning five domains was analysed using domain ablation and validated through alternative orderings and hyperparameter sensitivity, with SHAP-based explainability, including exact interaction values.

The key findings are: (1) multi-domain integration yields a 220% relative R^2 improvement, robust across three alternative orderings; (2) Schedule ($\Delta R^2 = +0.312$) and Quality (+0.255) contribute the largest marginal improvements, providing—within the simulated causal structure—an illustrative quantification that schedule–data integration yields greater predictive return than resource–data integration; (3) cross-domain interactions (Schedule_Delay $\times$ Risk_Level, Project_Type $\times$ Change_Cost_Ratio) carry substantial predictive information confirmed by exact interaction values; and (4) emergent findings—including context-dependent marginal contributions and non-designed interaction patterns—demonstrate that the study’s contributions extend beyond recovery of designed causality.

The theoretical contribution lies in demonstrating that cost overrun is best understood as an emergent phenomenon arising from multi-domain interactions, not reducible to single-domain causes. This extends the systemic perspective of prior studies [15] with quantitative interaction evidence.

For building construction practitioners, the framework illustrates how an organisation could quantify which integration investments would yield the highest predictive return once it is applied to that organisation’s own data. The specific rankings and the cross-domain early-warning thresholds reported here are derived from synthetic data and are therefore illustrative; they would require recalibration and validation on real data before deployment as risk-management tools.

For the research community, this study establishes a replicable ablation-based methodology for quantifying data integration value. The theory-grounded synthetic data approach provides a template for future studies facing the data fragmentation barrier. Future research should validate this framework on real-world integrated datasets, incorporate dynamic time-series modelling, and explore additional data domains including macroeconomic and supply chain indicators.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/buildings16132517/s1>. Table S1: Descriptive statistics for all dataset variables; Table S2: Category frequencies for the categorical variables; Table S3: Complete Pearson correlation of every numeric predictor with Cost_Overrun; Table S4: Complete SHAP feature importance for all 21 predictors; Table S5: Domain-level cross-domain SHAP interaction matrix; Table S6: Single-domain predictive performance for all three models; Table S7: Computational and reproducibility settings; and Figure S1: Pearson correlation matrix among all numeric variables.

Author Contributions: Conceptualization, J.L. and W.P.; methodology, J.L.; software, J.L.; validation, J.L. and W.P.; formal analysis, J.L.; investigation, J.L.; resources, W.P.; data curation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L. and W.P.; visualization, J.L.; supervision, W.P.; project administration, W.P.; and funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MSIT) (2022R1C1C1009927).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The synthetic dataset and the complete Python code for data generation, model training, SHAP analysis, and figure reproduction are openly available in the Zenodo repository at <https://doi.org/10.5281/zenodo.20626385> (accessed on [10 June 2026]) under CC-BY-4.0 (the code is additionally provided under the MIT License) and are also available from the corresponding author. A single deterministic script regenerates every results table (Tables 4–12) and figure from the dataset using an 80/20 hold-out split with random seed 42. The full causal-mechanism specification—qualitatively in Table 2 and as explicit functional forms with the noise distribution in Appendix B—together with the data-generation code, enables exact reproduction of every reported result and allows other researchers to modify the causal assumptions for sensitivity analysis. The dataset is fully synthetic and contains no real project data.

Acknowledgments: The authors thank the editorial team and anonymous reviewers for their constructive feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Description of variables in the integrated dataset by data domain.

Domain	Variable	Type	Unit	Description
Project	Project_Type	Cat.	—	Building/Infrastructure/Remodelling
Project	Area	Cont.	m ²	Gross floor area
Project	Floor_Count	Disc.	floors	Number of floors
Project	Structure_Type	Cat.	—	RC/Steel/SRC/Composite/Masonry/Concrete
Project	Contract_Type	Cat.	—	Design-Build/Turnkey/DBB/CM-at-Risk
Project	Planned_Cost	Cont.	KRW	Initial budgeted cost
Project	Planned_Duration	Cont.	days	Planned duration
Schedule	Actual_Duration	Cont.	days	Realised duration
Schedule	Schedule_Delay	Cont.	ratio	(Actual – Planned)/Planned
Resource	Labor_Total	Cont.	persons	Total labour force
Resource	Equipment_Count	Disc.	units	Equipment deployed
Resource	Resource_Utilisation	Cont.	ratio	Utilisation rate (0–1)
Quality	Change_Order_Count	Disc.	count	Design change orders
Quality	Change_Cost_Ratio	Cont.	ratio	Change cost/Planned cost
Quality	Defect_Count	Disc.	count	Recorded defects
Quality	Defect_Severity	Cont.	index	Severity (0–1)
Quality	Repair_Cost	Cont.	KRW	Defect repair expenditure
Safety/Env.	Accident_Count	Disc.	count	Safety incidents
Safety/Env.	Risk_Level	Cont.	index	Composite risk (0–1)
Safety/Env.	Temperature	Cont.	°C	Average site temperature
Safety/Env.	Rainfall	Cont.	mm	Average rainfall
Target	Cost_Overrun	Cont.	ratio	(Actual–Planned)/Planned cost

Appendix B

This appendix documents the functional forms used to synthesise the dataset, complementing the qualitative causal mapping in Table 2 and meeting reproducibility expectations. The complete, executable implementation—including all numeric constants and the random seed—is provided with the data-generation code referenced in the Data Availability Statement.

Each project record was generated by first sampling the domain variables from their marginal distributions (summarised in Table 3) and then computing the cost-overrun outcome as an additive function of theory-grounded causal terms plus stochastic noise:

$$\text{Cost_Overrun} = \beta_0 + \beta_1 \cdot \text{Change_Cost_Ratio} + \beta_2 \cdot \text{Schedule_Delay} + \gamma \cdot 1[\text{Schedule_Delay} > 0.20] + \beta_3 \cdot \text{Accident_Count} + \delta \cdot 1[\text{Accident_Count} \geq 3] + \tau(\text{Project_Type}) +$$

$$\kappa(\text{Contract_Type}) + \lambda \cdot (\text{Remodelling} \cdot \text{Change_Cost_Ratio}) + \mu \cdot (\text{Schedule_Delay} \cdot \text{Accident_Count}) + \varepsilon.$$

Following Table 2, $\beta_1 = 1.8$ (design-change cost propagation [2]); $\beta_2 = 0.35$, with a threshold escalation γ activated when Schedule_Delay exceeds 0.20 [1]; $\beta_3 = 0.008$, with a multi-incident penalty δ for three or more incidents [27]; $\tau(\cdot)$ encodes the project-type differential (Remodelling > Infrastructure > Building [13]); $\kappa(\cdot)$ encodes contract-type moderation (DBB > Design-Build/Turnkey/CM-at-Risk [3]); and the product terms λ and μ implement the cross-domain amplification (Remodelling $\times$ Change and Delay $\times$ Accident [2,27]).

The error term ε is additive Gaussian, $\varepsilon \sim N(0, \sigma^2)$, with σ calibrated so that noise accounts for approximately 10% of the variance of Cost_Overrun, representing unobservable factors [1]. Actual_Cost is then obtained as Planned_Cost $\times$ (1 + Cost_Overrun) and is excluded from the predictor set to avoid mechanical target leakage. The exact values of β_0 , γ , δ , the categorical maps τ and κ , the amplification weights λ and μ , and σ , together with the random seed, are defined in the data-generation code, enabling exact regeneration of the dataset and modification of the causal assumptions for sensitivity analysis.

Appendix C

Although no real integrated dataset is available for direct validation, the synthetic variables were designed to be consistent, in direction and order of magnitude, with patterns reported in the construction-management literature. Table A2 summarises this face-validity comparison.

Table A2. Face-validity comparison between the synthetic variables and patterns reported in the literature.

Variable	Synthetic Data (This Study)	Direction/Magnitude Reported in the Literature	Source
Cost_Overrun	mean 0.246; range −0.005–0.550	Cost overruns of tens of percent are widely reported and are positively skewed.	[1,2]
Schedule_Delay	mean 0.170; range −0.056–0.450	Schedule slippage of comparable relative magnitude is reported and is positively associated with cost.	[1,27]
Change_Cost_Ratio	mean 0.032; range 0.000–0.185	Change-order and rework costs typically range from a few to tens of percent of contract value.	[2]
Accident_Count	mean 1.29; range 0–10	Per-project incident counts are low and are positively associated with cost outcomes.	[27]
Risk_Level	mean 0.206; range 0.000–0.503	Higher project risk is associated with larger overruns.	[13]
Project-type effect	Remodelling > Infrastructure > Building	Project-type differentials in cost overrun are reported.	[13]
Contract-type effect	DBB > Design-Build/Turnkey/CM-at-Risk	Procurement type moderates cost-overrun risk.	[3]

These comparisons establish face validity at the level of direction and order of magnitude only. The generator's coefficients were calibrated to the patterns in the cited sources (Table 2; Appendix B); the table is not a claim of distributional equivalence to any specific real dataset, which remains to be established through the real-data validation identified in Section 5.7.

References

1. Flyvbjerg, B.; Holm, M.S.; Buhl, S. Underestimating Costs in Public Works Projects: Error or Lie? *J. Am. Plan. Assoc.* **2002**, *68*, 279–295. [CrossRef]
2. Love, P.E.D.; Ahiaga-Dagbui, D.D.; Irani, Z. Cost Overruns in Transportation Infrastructure Projects: Sowing the Seeds for a Probabilistic Theory of Causation. *Transp. Res. Part A Policy Pract.* **2016**, *92*, 184–194. [CrossRef]

3. Flyvbjerg, B. Over Budget, Over Time, Over and Over Again: Managing Major Projects. In *The Oxford Handbook of Project Management*; Morris, P.W.G., Pinto, J.K., Söderlund, J., Eds.; Oxford University Press: Oxford, UK, 2011; pp. 321–344. [\[CrossRef\]](#)
4. Choi, J.; Kim, H.; Kim, I. Multivariate Analysis of Construction Cost Overrun Using Bayesian Networks. *J. Constr. Eng. Manag.* **2019**, *145*, 04019048.
5. Akintoye, A.S.; MacLeod, M.J. Risk Analysis and Management in Construction. *Int. J. Proj. Manag.* **1997**, *15*, 31–38. [\[CrossRef\]](#)
6. Akinosho, T.D.; Oyedele, L.O.; Bilal, M.; Ajayi, A.O.; Delgado, M.D.; Akinade, O.O.; Ahmed, A.A. Deep Learning in the Construction Industry: A Review of Present Status and Future Innovations. *J. Build. Eng.* **2020**, *32*, 101827. [\[CrossRef\]](#)
7. Kim, G.H.; An, S.H.; Kang, K.I. Comparison of Construction Cost Estimating Models Based on Regression Analysis, Neural Networks, and Case-Based Reasoning. *Build. Environ.* **2004**, *39*, 1235–1242. [\[CrossRef\]](#)
8. Bilal, M.; Oyedele, L.O.; Qadir, J.; Munir, K.; Ajayi, S.O.; Akinade, O.O.; Owolabi, H.A.; Alaka, H.A.; Pasha, M. Big Data in the Construction Industry: A Review of Present Status, Opportunities, and Future Trends. *Adv. Eng. Inform.* **2016**, *30*, 500–521. [\[CrossRef\]](#)
9. Pan, Y.; Zhang, L. Roles of Artificial Intelligence in Construction Engineering and Management: A Critical Review and Future Trends. *Autom. Constr.* **2021**, *122*, 103517. [\[CrossRef\]](#)
10. Ahn, J.; Ji, S.H.; Park, M.; Lee, H.S.; Kim, S.; Suh, S.W. The Attribute Impact Concept: Applications in Case-Based Reasoning and Parametric Cost Estimation. *Autom. Constr.* **2020**, *118*, 103265. [\[CrossRef\]](#)
11. Hamdan, M.M.; Thneibat, M.; Hyari, K. Predicting Cost Overrun in Construction Projects Using Machine Learning Algorithms: The Case of Jordan. *Eng. Constr. Archit. Manag.* **2026**, *33*, 4732–4763. [\[CrossRef\]](#)
12. Al-Tawal, D.R.; Arafah, M.; Sweis, G.J. A Model Utilizing the Artificial Neural Network in Cost Estimation of Construction Projects in Jordan. *Eng. Constr. Archit. Manag.* **2021**, *28*, 2466–2488. [\[CrossRef\]](#)
13. Cantarelli, C.C.; Molin, E.J.E.; van Wee, B.; Flyvbjerg, B. Characteristics of Cost Overruns for Dutch Transport Infrastructure Projects and the Importance of Project Phase and Project Type. *Transp. Policy* **2012**, *22*, 49–56. [\[CrossRef\]](#)
14. Elghaish, F.; Mohandes, S.R.; Rahimian, F.; Abrishami, S.; Hosseini, M.R. Predictive Digital Monitoring of Construction Resources: An Integrated Digital Twin Solution. *Eng. Constr. Archit. Manag.* **2025**, *ahead-of-print*. [\[CrossRef\]](#)
15. Shoar, S.; Yiu, T.W.; Payan, S.; Parchamijalal, M. Modelling Cost Overrun in Building Construction Projects Using the Interpretive Structural Modelling Approach: A Developing Country Perspective. *Eng. Constr. Archit. Manag.* **2023**, *30*, 365–392. [\[CrossRef\]](#)
16. Ahiaga-Dagbui, D.D.; Smith, S.D. Dealing with Construction Cost Overruns Using Data Mining. *Constr. Manag. Econ.* **2014**, *32*, 682–694. [\[CrossRef\]](#)
17. Shane, J.S.; Molenaar, K.R.; Anderson, S.; Schexnayder, C. Construction Project Cost Escalation Factors. *J. Manag. Eng.* **2009**, *25*, 221–229. [\[CrossRef\]](#)
18. Cheng, M.Y.; Tsai, H.C.; Sudjono, E. Conceptual Cost Estimates Using Evolutionary Fuzzy Hybrid Neural Network for Projects in Construction Industry. *Expert Syst. Appl.* **2010**, *37*, 4224–4231. [\[CrossRef\]](#)
19. Kim, B.S.; Kim, T.H. Reducing Cost Overrun in Highway Projects Using Artificial Neural Networks. *J. Comput. Civ. Eng.* **2016**, *30*, 04015063.
20. Gondia, A.; Siam, A.; El-Dakhkhni, W.; Nassar, A.H. Machine Learning Algorithms for Construction Projects Delay Risk Prediction. *J. Constr. Eng. Manag.* **2020**, *146*, 04019085. [\[CrossRef\]](#)
21. Rafiei, M.H.; Adeli, H. Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes. *J. Constr. Eng. Manag.* **2018**, *144*, 04018106. [\[CrossRef\]](#)
22. Leu, S.; Huang, K.; Hung, C.C.; Wu, P. Revolutionizing Steel Building Project Cost Overrun Risk Assessment by Bayesian Network. *Eng. Constr. Archit. Manag.* **2024**, *31*, 4975–4987. [\[CrossRef\]](#)
23. Shoar, S.; Banaitis, A. Application of Fuzzy Fault Tree Analysis to Identify Factors Influencing Construction Labour Productivity. *J. Civ. Eng. Manag.* **2019**, *25*, 41–52. [\[CrossRef\]](#)
24. Choi, S.; Kim, D.Y.; Han, S.H.; Kwak, Y.H. Conceptual Cost-Prediction Model for Public Road Planning via Rough Set Theory and Case-Based Reasoning. *J. Constr. Eng. Manag.* **2014**, *140*, 04013026. [\[CrossRef\]](#)
25. Singh, V. Digitalisation, BIM Ecosystem, and the Future of Built Environment: How Widely Are We Exploring the Different Possibilities? *Eng. Constr. Archit. Manag.* **2019**, *26*, 993–1012.
26. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006. [\[CrossRef\]](#)
27. Hwang, B.G.; Zhao, X.; Toh, L.P. Risk Management in Small Construction Projects in Singapore: Status, Barriers and Impact. *Int. J. Proj. Manag.* **2014**, *32*, 116–124. [\[CrossRef\]](#)
28. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
29. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
30. Elfaki, A.O.; Alatawi, S.; Abushandi, E. Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey. *Adv. Civ. Eng.* **2014**, *2014*, 107926. [\[CrossRef\]](#)
31. Wang, Y.R.; Yu, C.Y.; Chan, H.H. Predicting Construction Cost and Schedule Success Using Artificial Neural Networks Ensemble and Support Vector Machines Classification Models. *Int. J. Proj. Manag.* **2012**, *30*, 470–478. [\[CrossRef\]](#)

32. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [[CrossRef](#)]
33. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates: Red Hook, NY, USA, 2017; pp. 4765–4774.
34. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [[CrossRef](#)] [[PubMed](#)]
35. Lee, J.; Lee, S.H. Explainable Pattern Analysis of Near-Miss Accidents in Construction Using Classifier Chain and SHAP. *Autom. Constr.* **2025**, *172*, 106010. [[CrossRef](#)]
36. Turkyilmaz, A.H.; Polat, G. Risk-Based Completion Cost Overrun Ratio Estimation in Construction Projects Using Machine Learning Classification Algorithms: A Case Study. *Buildings* **2024**, *14*, 3541. [[CrossRef](#)]
37. Lee, S.; Yoon, S.; Park, S. An Explainable AI-Based Framework for Predicting Construction Firm Profitability and Capital Structure Considering Supply-Chain Volatility. *Buildings* **2026**, *16*, 840. [[CrossRef](#)]
38. Park, M.; Kang, T.W.; Kwon, S.W.; Hwang, J.K. Post-Handover Housing Quality Management and Standards in Korea. *Buildings* **2023**, *13*, 1921. [[CrossRef](#)]
39. Liu, X.; Wang, Y.; Zhang, S. Interpretable Machine Learning Models for Prediction of UHPC Creep Behavior. *Buildings* **2024**, *14*, 2080. [[CrossRef](#)]
40. Mahmood, S.M.; Salman, S.; Almeida, P. Explainable Machine Learning to Predict the Construction Cost of Power Plant Based on Random Forest and Shapley Method. *CivilEng* **2025**, *6*, 21. [[CrossRef](#)]
41. Stefanini, L.; Badini, L.; Mochi, G. Rapid Visual Screening Feature Importance for Seismic Vulnerability Ranking via Machine Learning and SHAP Values. *Appl. Sci.* **2024**, *14*, 2609. [[CrossRef](#)]
42. Hossain, M.A.; Chowdhury, M.M.; Hossain, M.J. An Advanced Explainable Belief Rule-Based Framework to Predict the Energy Consumption of Buildings. *Energies* **2024**, *17*, 1797. [[CrossRef](#)]
43. Ma, Z.; Liu, Z.; Zhang, D. Data-Driven Approach for Construction Cost Estimation: State-of-the-Art Review. *J. Constr. Eng. Manag.* **2023**, *149*, 03123001.
44. Al-Tabtabai, H.; Kartam, N.; Flood, I.; Alex, A.P. Expert Judgment in Forecasting Construction Project Completion. *Eng. Constr. Archit. Manag.* **1997**, *4*, 271–293. [[CrossRef](#)]
45. Project Management Institute. *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*, 7th ed.; PMI: Newtown Square, PA, USA, 2021.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.