
AI-Based Embodied Carbon Estimation from Bill of Quantities Data in Indian Construction Projects

Ankith Prabhu

*Dept. of CSE (AI&ML), Dayananda Sagar University, Bangalore,
Karnataka, India, eng23am0286@dsu.edu.in*

Ritika Reddy Alla

*Dept. of CSE (AI&ML), Dayananda Sagar University, Bangalore,
Karnataka, India, eng23am0289@dsu.edu.in*

Shreya Rao

*Dept. of CSE (AI&ML), Dayananda Sagar University, Bangalore,
Karnataka, India, eng23am0278@dsu.edu.in*

Ankita Thakkar

*Dept. of CSE (AI&ML), Dayananda Sagar University, Bangalore,
Karnataka, India, ankitath1989@gmail.com*

Yisol Choi

*Dept. of CSE (AI&ML), Dayananda Sagar University, Bangalore,
Karnataka, India, eng23am0285@gmail.com*

Lakshmanan M

*Dept. of CSE (AI&ML), Dayananda Sagar University, Bangalore,
Karnataka, India, lakshmanan1909@gmail.com*

Abstract.

No practical method exists for estimating embodied carbon at the Bill of Quantities (BOQ) stage even though construction projects produce significant amounts of it. This paper proposes using BOQ data from five real Indian construction projects to directly predict Carbon Cost Intensity (CCI) which is the cost per unit of embodied carbon. Emission factors are drawn from ICE DB V4.1. Item descriptions are vectorized using TF-IDF, combined with one-hot encoded material and unit categories, concrete grade, and log-transformed quantity. Five regression models were trained and evaluated under 5-fold cross-validation: XGBoost, Support Vector Regression, Decision Tree, Random Forest, and Linear Regression. XGBoost performed best, achieving a mean absolute error of 37.56 INR/kgCO_{2e} and an R² of 0.594. A GGBS substitution analysis showed that replacing 30% and 50% of cement with GGBS cut total portfolio carbon by 24.1% and 40.2% respectively. The outcome is a working, reproducible framework for embodied carbon assessment built around the realities of Indian construction practice.

Keywords. embodied carbon, bill of quantities, machine learning, XGBoost, carbon cost intensity, GGBS, Indian construction, emission factors, ICE DB

1. INTRODUCTION

Buildings and construction account for roughly 37% of global energy and process-related CO₂ emissions [1], while embodied carbon is largely determined during design and procurement [2]. In India, standardized project-level assessment tools remain limited, and conventional LCA requires detailed material inventories often unavailable at the BOQ stage [3,4]. We therefore investigate whether BOQ data can directly support early-stage embodied carbon estimation using Carbon Cost Intensity (CCI), defined as cost per unit of embodied carbon. This study develops a BOQ-based ML pipeline using ICE DB V4.1 emission factors, evaluates regression models for CCI prediction, and examines portfolio-level carbon reductions under GGBS substitution scenarios.

2. RELATED WORK

Embodied carbon estimation falls under process-based LCA, input–output analysis, and hybrid methods, with machine learning gaining attention as a data-driven route [7], as in Pomponi et al. [8] and Kayaoglu et al. [9]; neither confronts the inconsistent unit conventions typical of Indian BOQs. Factors here are sourced from ICE DB V4.1 [10], and XGBoost [11] is generally the strongest choice for tabular data of this kind.

3. DATASET AND METHODOLOGY

The dataset comprised 1161 BOQ items from five Indian construction projects. Descriptions and units were standardized, while material categories and concrete grades were extracted using keyword-based rules. PA (321 rows) is the only priced BOQ; CCI = Unit Rate (INR) / Emission Factor (kgCO₂e/unit), using ICE DB V4.1 values (e.g., M20 concrete: 284.9, steel Fe500: 1.720 kgCO₂e/kg). Raw CCI is heavily right-skewed (skewness = 6.95); log transformation reduces this to 0.49, and models predict log(CCI), converted back before evaluation. Features combined TF-IDF representations of item descriptions with unit, material category, concrete grade, and log-transformed quantity.

Five models were compared — XGBoost, SVR, Decision Tree, Random Forest, and Linear Regression — alongside Ridge, Lasso, and an MLP. All went through 5-fold CV stratified by CCI quartile, with MAE/RMSE/R² reported on the original scale. A leave-one-project-out test also checked cross-project transfer, since all labeled data comes from the single priced source, PA.

4. RESULTS AND DISCUSSION

Table I shows cross-validated results for all seven models. XGBoost performed best (MAE 37.56 INR/kgCO₂e, R² 0.594); SVR came closest (MAE 39.04, R² 0.546). Linear models and the MLP generalized worse, with the MLP yielding a negative R². Performance is most sensitive to items lacking a clear material keyword and to pricing outliers in PA, but the MAE falls within 12–18% of median CCI, within the tolerance expected of early-stage Indian estimates.

TABLE I
5-FOLD CROSS-VALIDATION RESULTS

Model	MAE (INR/kgCO ₂ e)	RMSE (INR/kgCO ₂ e)	R ²
XGBoost	37.56	89.18	0.594

SVR	39.04	94.30	0.546
Lasso Regression	53.35	109.89	0.384
Decision Tree	45.26	111.57	0.365
Ridge Regression	60.14	117.68	0.293
Random Forest	54.03	118.37	0.285
Neural Network (MLP)	64.72	171.02	-0.493

Material category and unit type account for over 60% of XGBoost's prediction gain. CCI was predicted for 577 line items across the four unpriced projects; concrete and steel together account for over 80% of predicted embodied carbon, while items with no ICE DB match are excluded from totals.

5. GGBS SUBSTITUTION ANALYSIS

GGBS (Ground Granulated Blast-furnace Slag) can reduce embodied carbon in Portland cement concretes [12]. Three portfolio-wide cases were studied: Baseline (no GGBS), 30% GGBS (workable without affecting structural performance [13]), and 50% GGBS (practical for less demanding elements). The adjusted emission factor is:

$$EF_{blend} = EF_{baseline} \times (1 - f_{GGBS}) + EF_{GGBS} \times f_{GGBS}, \quad EF_{GGBS} = 0.083 \text{ kgCO}_2\text{e/kg}$$

TABLE II
PORTFOLIO-LEVEL GGBS SCENARIO RESULTS

Scenario	GGBS (%)	Total Carbon (tCO ₂ e)	Reduction (%)	Cost Premium (INR M)
A: Baseline	0	167,109	—	0.0
B: 30% GGBS	30	126,844	24.1	34.1
C: 50% GGBS	50	100,001	40.2	56.8

A 30% GGBS mix removes 40,265 tCO₂e (24.1%) for INR 34.1 million; 50% saves 67,109 tCO₂e (40.2%) at INR 56.8 million. Most savings are driven by M20 and M30 concrete. Carbon saved per INR million is nearly identical (~1181-1182 tCO₂e) at either level. Even a conservative 30% shift delivers substantial decarbonization at a cost well within standard contingencies, with no redesign or new materials required.

6. LIMITATIONS AND FUTURE WORK

All 194 labeled samples come from one priced project (PA), so it is unclear whether the learned CCI patterns generalize to other regional cost environments. So single sourcing is the most significant constraint. A leave one project out test offers only a partial check. ICE DB coverage is also incomplete, so the totals are lower bounds.

7. CONCLUSION

This pipeline estimates embodied carbon from BOQ data in Indian construction projects, mapping ICE DB V4.1 emission factors to line items and combining TF-IDF text features with structured attributes to estimate CCI without a full life cycle assessment. XGBoost performed best among seven models (MAE 37.56 INR/kgCO₂e, R² 0.594). The GGBS result is the most useful takeaway: replacing just 30% of cement with GGBS cuts portfolio

carbon by 24.1% with no redesign and a cost within typical Indian project contingencies. How well the model travels to other sites remains open — a question more priced BOQs should help resolve.

8. REFERENCES

- [1] United Nations Environment Programme, “2022 Global Status Report for Buildings and Construction,” UNEP, Nairobi, 2022.
- [2] F. Pomponi and A. Moncaster, “Embodied carbon mitigation and reduction in the built environment: What does the evidence say?,” *Journal of Environmental Management*, vol. 181, pp. 687–700, Oct. 2016.
- [3] K. G. Akshatha, S. Palaniappan, and M. K. Pradeep Kumar Rao, “Embodied carbon variability in Indian high-rise residential buildings: A case-based assessment,” in *Proc. 13th World Construction Symposium*, Sri Lanka, Aug. 2025, pp. 623–634.
- [4] C. De Wolf, F. Pomponi, and A. Moncaster, “Measuring embodied carbon dioxide equivalent of buildings: A review and critique of current industry practice,” *Energy and Buildings*, vol. 140, pp. 68–80, Apr. 2017.
- [5] Z. H. Ali and A. M. Burhan, “Hybrid machine learning approach for construction cost estimation: an evaluation of extreme gradient boosting model,” *Asian Journal of Civil Engineering*, vol. 24, pp. 2427–2442, 2023.
- [6] X. Zhang et al., “Predictive models of embodied carbon emissions in building design phases: Machine learning approaches based on residential buildings in China,” *Building and Environment*, vol. 258, 2024.
- [7] A. Chadha and A. Gupta, “Role of machine learning to analyze the impact of construction industry on the carbon footprint,” *SSRG International Journal of Civil Engineering*, vol. 12, no. 8, pp. 176–183, 2025.
- [8] F. Pomponi, C. De Wolf, and A. Moncaster, Eds., *Embodied Carbon in Buildings: Measurement, Management, and Mitigation*. Cham: Springer, 2018.
- [9] M. Kayaoglu, O. Tatari, and C. Skibniewski, “Machine learning for construction cost and embodied carbon estimation,” *Journal of Construction Engineering and Management*, vol. 148, no. 3, 2022.
- [10] C. Hammond and C. Jones, *Inventory of Carbon and Energy (ICE) Database V4.1*. Bath: University of Bath, 2019.
- [11] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 785–794.
- [12] K. L. Scrivener, V. M. John, and E. M. Gartner, “Eco-efficient cements: Potential economically viable solutions for a low-CO₂ cement-based materials industry,” *Cement and Concrete Research*, vol. 114, pp. 2–26, Dec. 2018.
- [13] Bureau of Indian Standards, *IS 456:2000 Plain and Reinforced Concrete – Code of Practice*, 4th rev. New Delhi: BIS, 2000.