Economic Sciences
https://economic-sciences.com
ES (2025) 21(1S), 102-111| ISSN:1505-4683

Modelling Construction Sector Output Using Steel and Cement Indicators: A Machine Learning Time Series Approach

Sunita Daniel[1], Abin Sam[2]
[1] Visiting Faculty, FORE School of Management, New Delhi – 110016.
[2]Founder and Principal Architect, Sam & Sons, New Delhi – 110066.

Abstract:

Cement and steel play an important role in the construction industry. In this paper, we study the relationship between key steel and cement indicators and construction sector output in India, using the Index of Industrial Production (IIP) for Infrastructure/Construction Goods as a proxy for sector performance. The monthly data such as the steel and cement production, their respective indices and growth rates, the monthly IIP For Infrastructure was collected data from April 2011 to March 2025. The data was cleaned and pre-processed and missing values were handled. Feature engineering was done and since the data was a time series, lag features were introduced to capture temporal dependencies. Bagging and boosting machine learning models were applied to the dataset to model and predict construction sector output. The top five and top ten most important features were identified and used for retraining and hyperparameter tuning. Among these, steel-related features—particularly Steel Index, Steel Growth, and Steel Production—along with lagged IIP values emerged as the strongest predictors of construction sector output. Cement-related variables had marginal influence by comparison. This machine learning approach demonstrates its potential in economic modelling and can assist policymakers and industry stakeholders in making data-driven decisions.

Keywords: *Construction Sector, IIP, Time series, Machine Learning, Hyperparameter Tuning,*

1 Introduction

One of the key economic indicators is the Index of Industrial Production (IIP) which measures the growth and performance of various industrial sectors in an economy. The IIP tracks changes in the volume of production across three main sectors: Mining, Manufacturing, and Electricity. This report is released monthly by the National Statistical Office (NSO). Within the IIP framework, detailed indices are also provided for specific use-based sectors, one of which is the Construction sector. The IIP for the Construction sector reflects the production and supply of infrastructure-related goods such as steel, cement, and other construction materials, making it a crucial measure of activity in infrastructure development and building projects. The Index of Industrial Production (IIP), particularly the sub-index related to infrastructure/construction goods, is widely recognized as a reliable proxy for assessing the output of the construction sector. Thus, the construction industry plays a pivotal role in India's economic development, contributing significantly to GDP and providing employment to millions. As of 2021, the construction sector has contributed around 8% to India's GDP and employed more than 40 million individuals (Business Wire, 2024). It is closely interlinked with core material industries—most notably, steel and cement—which are essential inputs for infrastructure development. Steel and Cement are not only essential in construction industry but they are also important indicators in the economic growth of the country. They are two of the eight core industries of India, the others being coal, crude oil, natural gas, refinery products, fertilizers and electricity.

Steel serves as a critical backbone for construction due to its durability, structural integrity, and versatility. It is extensively used in bridges, roads, high-rise buildings, and other infrastructure projects. India is currently the second-largest producer of steel globally, with an output of over 154 million tonnes in 2023 (Wikipedia, 2024a). However, challenges such as competition from cheaper Chinese imports and fluctuating global prices have pressured the domestic industry (Reuters, 2024). Cement is another material,

primarily consumed in the building and infrastructure segments. It accounts for nearly 75% of total usage in construction activities in India (Kansal, 2021). With an installed capacity of over 500 million tonnes per annum, India ranks as the second-largest cement producer globally (Wikipedia, 2024b). Yet, the sector also experiences price volatility and uneven supply due to changes in fuel and raw material costs (Times of India, 2024).

Given the importance of steel and cement in construction industry and their influence in the IIP of the construction sector, predicting the IIP of the construction sector is crucial for predicting the overall IIP of the country which in turn helps us predict the economic health of the country. In general, traditional statistical or econometric models have been used for forecasting or modelling industrial output. These models do not work well with high-dimensional, non-linear interactions between multiple economic indicators which are very common in real-world construction and economic data. In recent years with the sudden growth in big data and machine learning (ML) techniques, forecasting in economic and industry have changed dramatically. This is because these ML techniques are good in identifying complex non-linear patterns and can deal with high dimensional data.

In this paper, we make use of ML methods such as Random Forest and XGBoost and model the construction sector's IIP using key indicators from the steel and cement industries. Data cleaning, handling of missing data, feature engineering are done before applying the models. The models are finetuned using the Random Search CV and Grid Search CV methods to reduce the error in prediction. By integrating time series techniques with machine learning, this study not only improves the accuracy of construction sector forecasts but also enhances interpretability for policymakers and industry leaders.

2 Literature Review

Over the years various research has been carried out and the dynamics of the construction industry, steel, and cement, both as individual sectors and as components of broader economic systems have been studied. The structure and challenges of the Indian construction industry have been detailed in foundational reports such as Construction Industry Structure in India (Planning Commission, 2007), which emphasized the role of material availability and sectoral integration. The relationship between steel consumption and GDP has been studied and it was found that steel does contribute to the infrastructure and economic development (Ghosh and Roy, 2012, Paul and Mitra, 2020). Similarly, cement has been identified as a critical construction input with its demand closely linked to housing and public infrastructure projects (Ghosh and Ghosh, 2011).

Concerning the methodologies used, traditional econometric models like ARIMA and VAR have been widely used to model industrial production and sectoral growth (Asteriou & Hall, 2015). However, these models are unable to handle complex patterns. Recently, ML techniques have widely used in economic and industrial applications. Predicting cement prices (Elshafey et al., 2022), forecasting concrete corrosion (Ali et al., 2023) and labour productivity (Nassar and Hegazy, 2023) were done using ML techniques. Ensemble methods like XGBoost and Random Forest outperformed traditional models in improving the accuracy of predicting manufacturing output (Choudhary and Chatterjee, 2020) and cost forecasting (Zhao and Qian, 2023). XGBoost was applied for sectoral forecasting and hyperparameter tuning and feature selection enhanced the model performance (Varma et al., 2021). In the context of construction forecasting, neural networks were applied to model construction cost estimations, showing the promise of data-driven models (Zhao and Magoulès, 2012).

However, there is no literature available linking steel and cement indicators to construction sector IIP using machine learning and hence this study integrates economic indicators from steel and cement sectors and applying machine learning models to predict the IIP for infrastructure/construction goods.

3 Methodology

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1S), 102-111| ISSN:1505-4683

In this section, we describe the methodology followed to model and forecast the output of the Indian construction sector—represented by the Index of Industrial Production (IIP) for Infrastructure/Construction Goods—using steel and cement indicators through machine learning techniques. Figure 1 shows a systematic workflow of the process.

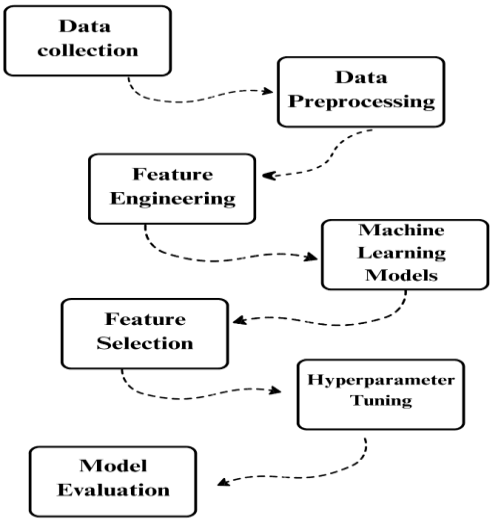

Figure 1: Flowchart showing the framework followed in analysis

3.1 Data Collection and Description

The time period for which the study was conducted is from April 2011 to March 2025, since currently the year 2011-2012 is used as the base year for calculating the IIP. The monthly reports of variables consisting of Steel Index, Cement Index, Steel Growth (%), Cement Growth (%), Steel Production (in thousand tonnes), Cement Production (in thousand tonnes), IIP (Construction Sector) were taken from various publicly available government and industry sources. The monthly steel/cement index is a number that reflects the average level of steel/cement production or price in a given month, compared to a base year. In fact, it shows how the steel/cement industry is performing each month, based on a fixed point of reference which is the base year (2011-2012=100). The indices show the trends in steel/cement production – whether output is increasing or slowing down, the industry health and are also used as a part of IIP calculations under the core industries. The monthly indices and the growth rates of steel and cement were taken from the Government of India website (Office of the Economic Adviser, 2025). Data on steel and cement production were taken from Trading Economics (n.d.-a, n.d.-b). The information regarding cement production from June 2021–March 2025 was not available. The monthly Indices of Industrial Production as per user-based classification for the infrastructure/ construction goods was also collected (Ministry of Statistics and Programme Implementation, 2025). However, there were missing values in the early time frame from April 2011 to March 2012.

3.2 Data Preprocessing

As discussed in the previous section, there were missing values in the dataset and the Month/Year column needed to be treated. Since the dataset was a time series, the missing values were imputed using linear interpolation, so that it does not affect the trends or seasonality. The ‘Months/Years’ column was converted into datetime format and set as the index to preserve chronological ordering which is critical for time series modelling.

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1S), 102-111| ISSN:1505-4683

3.3 Feature Engineering

This step is an important step in machine learning where the raw data is transformed in a format which can be used for applying the machine learning models. Since we are dealing with time series involving material indicators in the construction sector, these indicators may have delayed effect on the construction sector. In order to capture this, lag features of 1-month and 2-month for the indices (Steel Index and Cement Index), production (Steel production and Cement Production), Steel and Cement growth and also for the IIP was generated. Introducing the lag features created null values which were removed from the dataset.

3.4 Machine Learning Models

To analyze the influence of steel and cement indicators on the construction sector's performance, machine learning (ML) models with a time series regression approach were used. In contrast to econometric models which assume linearity, machine learning techniques, particularly tree-based ensemble models, are capable of modeling complex dependencies without rigid assumptions. Therefore, Random Forest and XGBoost regressors were chosen for this study, given their robustness, interpretability, and ability to identify key predictive features.

After data preprocessing and feature engineering the dataset had 152 rows and 21 columns. In order to apply the machine learning model, the dataset was first divided into two – the training set and the test set. The first 80% of the dataset formed the training test and the remaining 20% the test set. The model was built on the training set and then to check its robustness, it was used to predict for the test set.

3.4.1 Random Forest Regressor Model:

The Random Forest Regressor is an ensemble learning method that constructs multiple decision trees during training and outputs the average of their predictions for regression tasks. It belongs to the family of bagging algorithms, where the primary idea is to combine the results of many individual models to produce a more robust and accurate overall model (Breiman, 2001).

In Random Forest, each tree is trained on a random subset of the training data selected with replacement (a technique known as bootstrap aggregation or "bagging"). Additionally, at each split within a tree, a random subset of features is selected rather than considering all features, which introduces diversity among the trees. This randomness helps reduce variance and overfitting, which are common problems in single decision tree models. Each decision tree in the forest is trained to predict the target variable based on a set of features, and the final prediction is obtained by averaging the outputs from all trees in the forest. The averaging process helps in smoothing out predictions, reducing noise, and improving generalization to unseen data.

One of the key advantages of the Random Forest Regressor is its ability to handle non-linear relationships and interactions between variables without requiring explicit feature transformations. Moreover, it can provide estimates of feature importance, helping researchers identify which inputs have the most influence on the target variable.

3.4.2 XGBoost Regressor Model:

XGBoost (Extreme Gradient Boosting) is a powerful and efficient implementation of gradient boosting machines, widely recognized for its high performance in supervised learning tasks, including regression and classification. It builds an ensemble of decision trees in a sequential manner, where each new tree attempts to correct the errors made by the previous ones (Chen & Guestrin, 2016). Unlike Random Forest, which builds trees independently, XGBoost builds trees **additively**, optimizing a specified loss function at each step.

The XGBoost algorithm is based on the concept of **gradient boosting**, where weak learners (typically shallow decision trees) are combined to form a strong learner. In each iteration, the model learns from the residuals (errors) of the previous trees by minimizing a regularized objective function. This objective function includes a loss function (e.g., mean squared error) and a regularization term to penalize model

complexity, thereby improving generalization and reducing overfitting.

3.5 Feature Selection

Feature selection is a crucial step in machine learning, particularly when dealing with multiple input variables that may not all contribute equally to the predictive power of a model. It helps in improving model performance, reducing overfitting, and enhancing the interpretability of the results. In this study, feature selection was performed to identify the most relevant indicators influencing the IIP (Index of Industrial Production) of the construction sector, using steel and cement-related features as predictors.

After the initial models were trained using all available features, a tree-based method—Random Forest Regressor was applied to evaluate the importance of each feature. Tree-based algorithms are especially useful in this context because they provide a natural mechanism for ranking features based on how effectively they split the data and reduce error at each decision node. For feature selection, we included all the features - Steel Index, Cement Index, Steel Growth (%), Cement Growth (%), Steel Production (in thousand tonnes), Cement Production (in thousand tonnes), Lagged features for Steel Growth, Cement Growth, and IIP. The analysis helped us identify the top 5 or top 10 features of importance.

3.6 Hyperparameter Tuning

In machine learning, hyperparameter tuning plays a crucial role in optimizing model performance. Hyperparameters are parameters that govern the learning process and model architecture but are not learned from the data directly. Choosing the right combination of hyperparameters can significantly improve prediction accuracy and generalization ability.

To evaluate different combinations of hyperparameters effectively, cross-validation is commonly used. Cross-validation is a technique where the data is divided into multiple subsets or "folds," and the model is trained and validated multiple times, each time using a different fold for validation and the remaining folds for training. This helps ensure that the model performs well on different parts of the dataset and does not overfit to a specific portion. However, since this study deals with time series data, traditional k-fold cross-validation cannot be used directly, as it would violate the chronological order of the data. To address this, TimeSeriesSplit was used, which is a time-aware cross-validation strategy. In TimeSeriesSplit, the training set grows with each iteration, and the test set is always ahead of the training set in time, thereby maintaining temporal integrity and preventing data leakage (Arlot & Celisse, 2010).

Two popular approaches for hyperparameter tuning are RandomizedSearchCV and GridSearchCV. RandomizedSearchCV searches over a random combination of parameters from a defined distribution (Bergstra & Bengio, 2012). It is computationally efficient and especially useful when the hyperparameter space is large, as it evaluates a fixed number of random combinations. GridSearchCV, on the other hand, performs an exhaustive search over a manually specified grid of hyperparameters (Pedregosa et al., 2011). Though more computationally expensive, it ensures that all combinations are evaluated and can yield better results when the search space is small and well-defined.

3.7 Model Evaluation

Model evaluation is a critical component of the machine learning pipeline, especially in time series forecasting, where accurate predictions can directly inform economic policy and industrial planning. In this study, the performance of different regression models was assessed using two key evaluation metrics: Root Mean Square Error (RMSE) and R-squared (R^2).

- RMSE measures the average magnitude of the errors between predicted and actual values. It is particularly useful because it retains the original units of the target variable and penalizes larger errors more than smaller ones.
- R-squared (R^2) represents the proportion of variance in the dependent variable that can be explained by the independent variables in the

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1S), 102-111| ISSN:1505-4683

model. An R^2 value closer to 1 indicates a better fit, while a negative R^2 suggests that the model performs worse than simply predicting the mean of the target variable.

4 Results and Discussion

In this research, machine learning models were used to predict construction sector output (represented by the IIP for infrastructure/construction goods) based on material indicators such as steel and cement production, their growth rates, and price indices. Two popular ensemble machine learning models—Random Forest and XGBoost—were applied across different subsets of features to evaluate predictive performance. Initial model runs were performed using Random Forest Regressor and XGBoost Regressor on all features. The Random Forest model resulted in an RMSE of 92.07 and an R^2 of -6.03, whereas the XGBoost model gave a significantly lower RMSE of 16.20 and a slightly better R^2 of -0.23. These initial results indicated that XGBoost was more suitable for the data, although the negative R^2 values suggested further optimization was necessary.

To improve performance, feature selection was carried out using a Decision Tree Regressor, and models were retrained using the top 5 and top 10 most important features. the top 5 features were identified as IIP (lag 1 month), Steel Growth (lag 1 month), Steel Index, Steel Growth, Steel Production implying that steel related indicators are the important features in predicting the IIP. The top 10 features included cement-related indicators and their lagged values (Cement Growth, Cement Production, Cement Production (lag1 month), Cement Index, Cement Growth (lag1month). Table 1 shows the model performance of both the models for the different feature sets (All features, Top 5 features and Top 10 features).

Table 1: Model Performance before hyperparameter tuning

Model	Feature Set	RMSE	R^2
Random Forest Regressor	All Features	92.37	-6.03
	Top 5 Features	27.34	-1.08
	Top 10 Features	21.12	-0.61
XGBoost Regressor	All Features	16.20	-0.2
	Top 5 Features	20.23	-0.54
	Top 10 Features	21.09	-0.6

Following this, both RandomizedSearchCV and GridSearchCV were applied to tune hyperparameters for the Random Forest Regressor and XGBoost Regressor models for all features, the top 5 and top 10 features using TimeSeriesSplit as the cross-validation strategy. The results of the tuning method and their impact on model performance are shown in Table 2.

Table 2: Model Performance after hyperparameter tuning

Model	Feature Set	Tuning Method	RMSE	R2
Random Forest Regressor	All Features	Randomised Search CV /Grid Search CV	15.25	-0.22
	Top 5 Features		14.79	-0.19
	Top 10 Features		14.36	-0.15
XGBoost Regressor	All Features	Grid Search CV	12.06	0.03
	Top 5 Features		11.14	0.11
	Top 10 Features		14.73	-0.18

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1S), 102-111| ISSN:1505-4683

For the Random Forest Regressor, no significant difference was observed between the two tuning methods. Both RandomizedSearchCV and GridSearchCV yielded identical performance metrics for top 5 and top 10 features, suggesting the model was stable and not sensitive to small hyperparameter variations in this dataset. In contrast, for the XGBoost Regressor, GridSearchCV consistently outperformed RandomizedSearchCV, particularly with the top 5 features where it achieved the lowest RMSE of 11.14 and positive R^2 score of 0.11 across all models, indicating a relatively better model fit. However, for top 10 features, GridSearchCV still performed better than RandomizedSearchCV but did not surpass the top 5 feature results. Overall, these results confirm that careful feature selection and exhaustive hyperparameter tuning (especially with GridSearchCV) can improve model performance, particularly in non-linear models like XGBoost. Moreover, the top 5 features with XGBoost and GridSearchCV emerged as the best-performing combination in this study.

Table 3 gives the performance summary of the two machine learning models showing the final RMSE value and the R^2 score for different feature sets.

Table 3: Model Performance Summary

Model	Feature Set	RMSE	R^2
Random Forest Regressor	All Features	15.25	-0.22
	Top 5 Features	14.79	-0.19
	Top 10 Features	14.36	-0.15
XGBoost Regressor	All Features	12.06	0.03
	Top 5 Features	**11.14**	**0.11**
	Top 10 Features	14.73	-0.18

The Random Forest model achieved a Root Mean Squared Error (RMSE) of 15.25 and a R^2 score of –0.22, indicating poor explanatory power and generalization. In contrast, XGBoost performed relatively better with an RMSE of 12.06 and an R^2 score of 0.03, though still suggesting a poor model fit overall.

After selecting the top 5 features, the Random Forest model showed improved performance with an RMSE of 14.79 and R^2 of –0.19, while the XGBoost model further improved to RMSE 11.14 and R^2 0.11. This indicates that XGBoost was able to capture a modest portion of the variance in the target variable, outperforming Random Forest in both error minimization and explanatory power. Expanding the model to the top 10 features—which added Cement Growth, Cement Production, Cement Production (lag 1), Cement Index, and Cement Growth (lag 1), it can be observed from Table 3 that for Random Forest: RMSE = 14.36, R^2 = –0.15 and for XGBoost: RMSE = 14.73, R^2 = –0.18.

Interestingly, adding more features did not necessarily improve the model's performance significantly, and in the case of XGBoost, slightly degraded the R^2 value. These results emphasize that only a handful of steel-related features (especially Steel Growth and Steel Index) and the previous value of IIP appear to contribute meaningfully to forecasting the current IIP.

The negative R^2 values in most configurations—especially in Random Forest—indicate that the models are often worse than simply predicting the average IIP value. However, XGBoost with the top 5 features showed positive R^2 (0.11), suggesting that a leaner, carefully selected feature set can yield better results than using all available inputs indiscriminately.

The graphs in Figure 2 show the predicted values and the actual values of the Random Forest and XGBoost models for the top 5 features. It can be clearly seen that the XGBoost model closely tracked the trend of the

Economic Sciences
https://economic-sciences.com
ES (2025) 21(1S), 102-111| ISSN:1505-4683

actual values whereas the Random Forest predictions exhibited greater deviations from the observed data. This visual confirmation reinforces the conclusion that XGBoost is more suited for modelling this dataset.

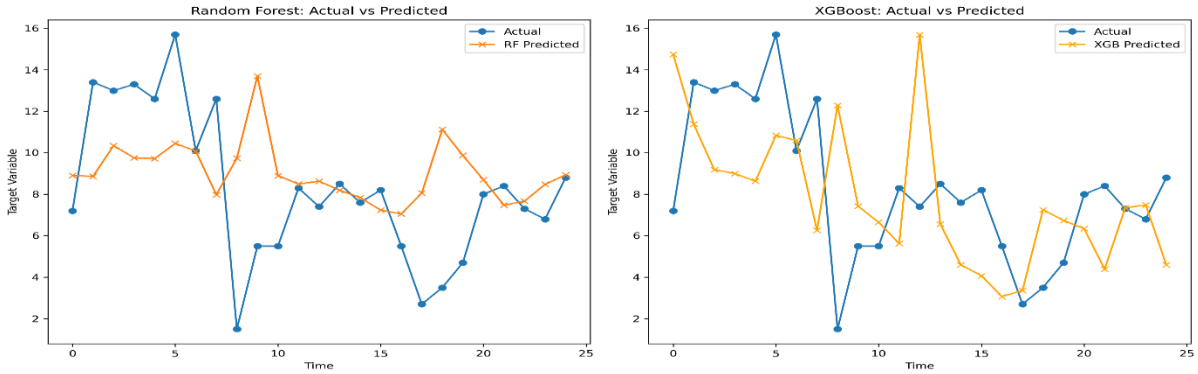

Figure 2: Graphs showing the actual and predicted values for both the Random Forest and XGBoost models for top 5 features

5 Conclusion

This study set out to explore the impact of steel and cement indicators on the construction sector output in India, using the Index of Industrial Production (IIP) for infrastructure and construction goods as the target variable. By leveraging machine learning techniques on monthly time-series data spanning April 2011 to March 2025, we aimed to model and predict construction activity using key variables such as steel and cement production, growth rates, and price indices. Two non-linear ensemble models—Random Forest and XGBoost—were employed for the analysis. Initial results using all available features revealed poor model performance, especially for Random Forest, which yielded a high RMSE and significantly negative R^2 values. XGBoost performed relatively better, indicating its suitability for this kind of structured, time-series data. To improve accuracy and interpretability, feature selection was conducted using a decision tree model. The top five and top ten most important features were identified and used for retraining and hyperparameter tuning. Among these, steel-related features—particularly Steel Index, Steel Growth, and Steel Production—along with lagged IIP values, emerged as the strongest predictors of construction sector output. Cement-related variables had marginal influence by comparison.

The XGBoost model with the top five features achieved the best performance, with an RMSE of 11.14 and a positive R^2 of 0.11, indicating modest predictive power. This was a significant improvement over models trained with the full feature set. In contrast, adding more features did not substantially improve the results and, in some cases, led to overfitting and decreased model performance. Overall, the study demonstrates the potential of machine learning techniques in forecasting construction sector output, especially when guided by thoughtful feature selection and domain knowledge. While the predictive power remains moderate, this work is novel in applying time-series ML models to the intersection of material inputs and economic output in the Indian construction sector. Future work could explore additional economic indicators, alternative target variables, or deep learning architectures to further enhance forecasting accuracy.

6 Limitations and Future Work

While this study provides valuable insights into the relationship between construction sector output and key inputs like steel and cement, there are several limitations that should be acknowledged:

6.1 Limitations

6.1.1 Limited Predictive Power:

Although the XGBoost model with selected features performed relatively better, the overall R^2 values remained low, indicating limited explanatory power. This suggests that the chosen features, while

important, may not fully capture the dynamics influencing construction sector output.

6.1.2 Data Gaps and Interpolation:

Some variables, such as cement production and growth rates, had missing values for extended periods (e.g., June 2021 to March 2025). These were imputed using interpolation techniques, which may introduce bias or dampen variability, potentially affecting model accuracy.

6.1.3 Scope of Variables:

The study focused only on steel and cement-related indicators, excluding other influential factors such as government infrastructure spending, interest rates, foreign investment, labor costs, or global commodity prices. These could significantly impact the IIP and construction activity but were not available in the dataset.

6.1.4 Lag Feature Selection:

Lagged features were generated manually (e.g., 1-month and 2-month lags). More systematic techniques for selecting optimal lag orders (e.g., autocorrelation plots or information criteria) could enhance the model's temporal understanding.

6.1.5 Model Generalizability:

The models were trained and validated on a single national dataset. Regional construction patterns or sector-specific activities (e.g., residential vs. infrastructure) may not be accurately captured, limiting generalizability.

6.2 Future Work

6.2.1 Incorporate Additional Economic Indicators:

Future research could integrate broader macroeconomic variables such as GDP, inflation, policy spending, and interest rates to enrich the modelling framework and potentially improve predictive accuracy.

6.2.2 Regional and Sectoral Analysis:

Disaggregating the IIP data by region or type of construction (e.g., residential, commercial, infrastructure) may reveal more granular patterns and enable targeted forecasting.

6.2.3 Advanced Time-Series Models:

Deep learning models such as Long Short-Term Memory (LSTM) networks or Temporal Convolutional Networks (TCN) could be explored for capturing sequential dependencies more effectively, especially in longer time horizons.

6.2.4 Causal Inference and Economic Interpretation:

Incorporating causal modelling approaches (e.g., Granger causality, Structural VAR) may help establish directional relationships between inputs and outputs, going beyond predictive accuracy to provide economic interpretability.

6.2.5 Scenario-Based Forecasting:

Future studies could build scenario models (e.g., high investment vs. low investment in infrastructure) to simulate the impact of different economic policies or material shocks (e.g., steel price surge) on construction output.

References

1. Ali, H., Singh, R., & Kumar, A. (2023). The application of machine learning techniques for forecasting corrosion in concrete structures. *Oriental Journal of Physical Sciences, 9*(2). https://www.orientaljphysicalsciences.org/vol9no2/the-application-of-machine-learning-techniques-for-forecasting-corrosion-in-concrete-structures/
2. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. https://doi.org/10.1214/09-SS054
3. Asteriou, D., & Hall, S. G. (2015). *Applied econometrics* (3rd ed.). Palgrave Macmillan.
4. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(1), 281–305.
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
6. Business Wire. (2024). *India Construction Industry Report 2024: Output to Grow by 11.2% this Year to Reach INR 25.31 Trillion - Forecasts*

to 2028. https://www.businesswire.com/news/home/20241119056246/en
7. 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 (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785
8. Choudhary, R., & Chatterjee, K. (2020). Forecasting India's manufacturing sector using Random Forest. *Economic Modelling*, 91, 88–97.
9. Elshafey, M., Kim, J., & Lee, S. (2022). Application of machine learning in cement price prediction through a web-based system. *ResearchGate*. https://www.researchgate.net/publication/364083427
10. Kansal, R. (2021). Cement consumption trends in India. *ACEEE Summer Study*. https://www.aceee.org/sites/default/files/pdfs/ssi21/panel-4/Kansal.pdf
11. Ghosh, M., & Ghosh, A. (2011). Cement demand in India and the economic impact. *International Journal of Economic Research*, 8(2), 45–53.
12. Ghosh, S., & Roy, S. (2012). Steel consumption and GDP: Evidence from India. *Journal of Economic Development*, 37(3), 1–18.
13. Ministry of Statistics and Programme Implementation. (2025). *Indices of Industrial Production (Base: 2011–12), Monthly and Annual – March 2025* [Data set]. Government of India. https://mospi.gov.in/sites/default/files/iip/IndicesIIP2011-12Monthly_annual_Mar25.xlsx
14. Nassar, K., & Hegazy, T. (2023). Machine learning-based forecasting of labor productivity in megaprojects. *SMU Data Science Review, 6*(1). https://scholar.smu.edu/datasciencereview/vol6/iss1/12
15. Office of the Economic Adviser. (2025). *Eight core industries data (Base year: 2011–12), up to May 2025* [Data set]. Ministry of Commerce and Industry, Government of India. https://eaindustry.nic.in/eight_core_infra/Core_Industries_2011_12_20250523.xlsx
16. Paul, P., & Mitra, P. (2020). Does the steel consumption affect the economic growth in India?. Journal of the Social Sciences, 48(3).
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
18. Planning Commission of India. (2007). *Construction Industry: Structure in India*. Government of India.
19. Reuters. (2024). Indian steel mills feel crunch of cheap Chinese imports. https://www.reuters.com/world/india/indian-steel-mills-feel-crunch-cheap-chinese-imports-2024-12-04/
20. Times of India. (2024). Surge in prices of building materials hits construction industry in Coimbatore. https://timesofindia.indiatimes.com/city/coimbatore/surge-in-prices-of-buliding-materials-hits-construction-industry-in-coimbatore/articleshow/121086805.cms
21. Trading Economics. (n.d.-a). *India cement production*. Retrieved May 14, 2025, from https://tradingeconomics.com/india/cement-production
22. Trading Economics. (n.d. -b). *India steel production*. Retrieved May 14, 2025, from https://tradingeconomics.com/india/steel-production
23. Varma, A., Srivastava, D., & Mishra, P. (2021). XGBoost-based sectoral forecasting in Indian economy. *Journal of Forecasting*, 40(5), 821–836.
24. Wikipedia. (2024a). *Iron and steel industry in India*. https://en.wikipedia.org/wiki/Iron_and_steel_industry_in_India
25. Wikipedia. (2024b). *Economy of India – Cement Industry*. https://en.wikipedia.org/wiki/Economy_of_India
26. Zhao, X., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6), 3586–3592.
27. Zhao, X., & Qian, Y. (2023). Predicting construction cost escalation using ensemble machine learning models. *Applied Sciences, 13*(5), 2261. https://www.mdpi.com/2076-3417/13/5/2261