International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 18s, 2025
https://theaspd.com/index.php

Predicting Coal Stock Levels In India's Thermal Plants: An Ensemble Machine Learning Approach And Policy Implications

Shashank Mishra[1]
[1]Academic Associate, Information Systems Area, Indian Institute of Management Indore , mishra.v.shashank@gmail.com

Abstract

Accurate forecasting of coal stock levels at thermal power plants is crucial for energy security and supply planning. This study uses daily coal stock data (2018–2023) from India's Ministry of Power to develop models for the prediction of next-year stock levels. We have used an ensemble of machine learning models—including Linear Regression, Support Vector Regression (SVR), XGBoost, and Neural Networks—and combined them in a stacking framework. The models are evaluated on regression metrics (R^2, MAE) and a derived classification task (flagging critical stock conditions). The stacked ensemble gives output as an R^2 of ~0.73, which significantly outperforms individual models, which achieve a value between 0.29–0.70 for the coefficient of determination. Figures summarize the pipeline and model performance. We discuss how such forecasts can inform policy: enabling planners to avoid shortages, optimise coal dispatch and maintain normal stock levels. Our research work provides a suggestive strategy that data-driven stock predictions can enhance energy resilience by guiding procurement strategies, coordinating rail and mining schedules, and mitigating supply-chain risks.

Keywords: *Ensemble Learning, Regression, Classification, Coal Stock, Forecasting, Machine Learning, Stacking, Policy*

INTRODUCTION

Electricity in India is mostly generated by coal, which powers about 55–74% of national output. Though renewables are expanding quickly, coal is still absolutely necessary; by 2030 it is expected to generate about 55% [1]. Such dependence creates vulnerability: for example, in late 2021 national coal stocks fell to only a few days' supply [2], which resulted into power shortages. Thus, maintaining sufficient coal inventories is essential for grid resilience. Normative stock requirements while making policy are practised (e.g. ~18–30 days of supply per plant), yet up-downs in demand, monsoonal disruptions, and rail bottlenecks create troubles in inventory planning and make it complex [3]. By maintaining accuracy in the forecast of coal stocks of future, it can help in building strategy while decision-making (e.g. timely procurement, logistic scheduling, contingency planning) and it will also help us to update energy policy with changing demand and supply scenarios [4].Machine learning (ML) offers powerful models for forecasting energy variables, which are capable of learning non-linear dynamics and seasonal patterns by time series data analysis [5]. The combination of multiple models together (generally known as Ensemble methods) often improves accuracy and robustness [6], [7]. For instance, stacking ensembles have been shown to achieve superior short-term power forecasts by integrating diverse base learners and then using their output as input for a meta-learner [7], [8]. In this study, we apply a stacking ensemble of Linear Regression, Support Vector Regression (SVR), XGBoost, and Neural Networks to predict next-year coal stock levels at India's thermal plants. We will observe how predicted stocks (and classification of critical shortages) can help in modifying policy on energy security and supply-chain resilience.The rest of the paper is organized as follows. Section II reviews relevant literature on coal demand, energy forecasting, and ensemble ML. Section III describes the dataset, pre-processing steps, feature engineering, and model design, including the stacking framework. Section IV presents regression and classification results, with evaluation metrics. Section V discusses policy implications and limitations. Section VI provides a conclusion to our research work.

LITERATURE REVIEW

Coal remains a central support for the India's energy mix. India is the world's second-largest coal consumer by consuming almost 14% of global demand [9]. Studies project that Indian coal consumption

will continue rising (~2.5% annual growth, 2018–2030). The paper by Yang and Li emphasises that forecast of future coal use will become helpful in forming environmental and energy policies [4]. On the policy side, recent government reports highlight stock levels in terms of days of supply; for example, as of mid-2024 stocks amounted to ~18.5 days of consumption (44.5 MT) [3]. The Coal Ministry also sets targets (55 MT by March 2025) and monitors deficits against normative requirements (~67 MT) [10]. These actual benchmarks highlight the need to foresee coal supply.

Methodologically speaking, time-series forecasting in the energy sector has progressively welcomed machine learning and ensemble techniques. Energy demand forecasting has utilized traditional statistical models (ARIMA) and also started incorporating advanced ML models (neural networks, SVM, tree ensembles) [5]. Recent work shows that combining models in a stacking ensemble framework may result in lower error than any single predictor. For example, The paper by Divina achieved highly accurate short-term load forecasts using a two-layer stacking ensemble of neural and regression models [6]. Similarly, the paper by Sakib developed an adaptive stacking ensemble (incorporating LSTM, BiLSTM, GRU networks) for forecasting energy demand, showing improved precision over individual models [5]. In his paper, Huang developed a data-driven stacking framework for forecasting renewable output that uses multiple base learners, finding it superior to benchmarked individual models [8]. These studies confirm that stacked ensembles can capture complex patterns and generalise well.

Even though a lot of work has been done on the energy forecasting, there are a few studies that focus specifically on coal inventory or stock levels. There are some reports too which inform us regarding the shortage of coal at various power stations in many states of India. Existing analyses (e.g., CEEW reports) have written about coal shortages and also given recommendations for the measures (e.g. conservation dispatch, rail prioritization) but relied on scenario analysis or proprietary models [2]. Starting in 2024, we anticipate a trend of falling global coal demand developing for our forecast period running until 2026. To our knowledge, the application of ensemble ML to predict coal stock level is quite new for India. This study fills that gap by leveraging a rich daily coal-stock dataset and state-of-the-art ML techniques.

Definitions

Linear Regression

Because of its ease of use and interpretability, linear regression is a fundamental statistical technique that has been widely applied in data analysis. Since Gauss developed the least squares method in the 19th century, it has been an essential tool for analysts and researchers to investigate and measure relationships between variables. For analyzing complex datasets, linear regression continues to provide substantial benefits as computing power and data volume have grown.

Fundamentally, the goal of linear regression is to use a linear equation to model the relationship between an independent variable (X) and a dependent variable (Y) [11]. Finding the best-fitting line, represented by a collection of coefficients, that minimizes prediction errors and provides an accurate estimate of the dependent variable is the goal.

Formally, the linear regression model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Where β_0 is the intercept, β_1 to β_n is the regression coefficient, X_1 to X_n is the independent variable, and ε is the error term.

XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable and efficient implementation of gradient boosted decision trees (GBDT), designed for speed and performance. It builds an ensemble of trees sequentially, where each tree corrects the errors of its predecessors by minimizing a differentiable loss function using gradient descent. The framework supports regularization (both L1 and L2) to reduce overfitting, handles sparse data efficiently, and incorporates advanced system optimization techniques such as cache-aware access patterns and out-of-core computation for large datasets [12].

Support Vector Machines (SVM)

The supervised machine learning technique known as Support Vector Machine (SVM) was first used in the early 1990s. Because of its efficacy and dependability, particularly when working with large datasets, it is frequently used to solve classification and regression problems. To improve prediction accuracy, SVM divides data into classes and seeks to maximize the margin between them by identifying the best

hyperplane between them. Additionally, SVMs can handle non-linear relationships by mapping the input data into higher-dimensional spaces using kernel functions [13].

Neural Networks (NN)

We used a Neural Network (NN) model in this study, a kind of statistical learning framework that draws inspiration from the architecture and operation of biological neural networks. The non-linear relationships found in energy-related data are especially well-captured and managed by this model [14]. Tasks requiring the estimation or approximation of unknown outputs that depend on a wide range of input variables are ideally suited for neural networks. An input layer, one or more hidden layers, and an output layer make up the structural core of a neural network.

METHODOLOGY

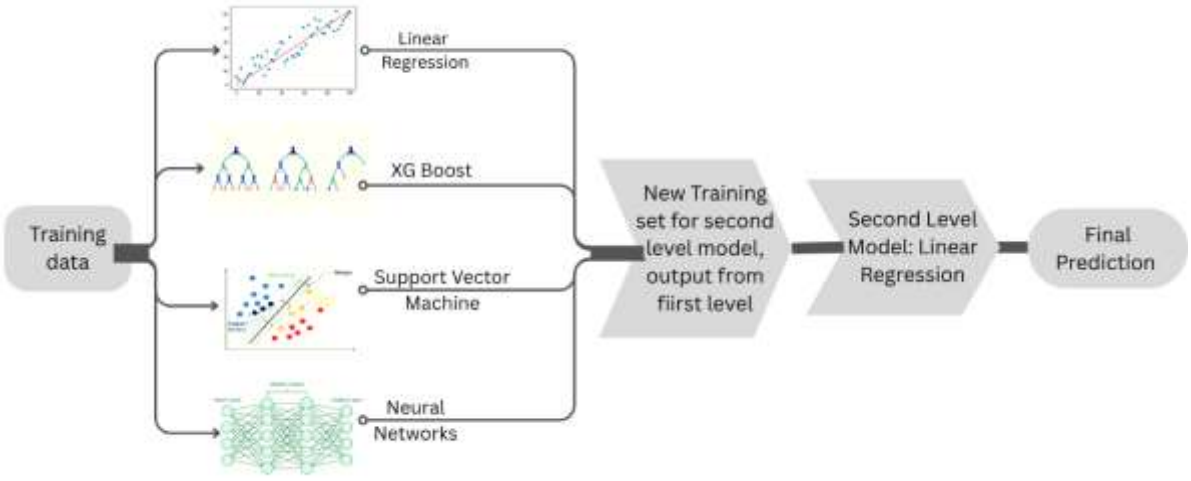
We use the Ministry of Power's daily coal-stock dataset (2018–2023), which records inventory and flow variables for each thermal plant per day. The raw data (≈253,600 entries) include plant attributes (state, capacity, sector), daily coal receipts, consumption, and current stock (indigenous, imported, total in thousand tonnes) and normative stock requirements (days and tonnes). We did data pre-processing by parsing dates, aggregating as needed (e.g. national or plant-level totals), and also handled missing values (e.g. filling gaps or omitting NaNs).

We obtain temporal and operational inputs for feature engineering. Seasonality is captured by date-related characteristics (day-of-year, month, and year). Encoded are plant characteristics (capacity, sector) and transport mode. The target variable for regression is the next-year coal stock level (e.g. total stock after one year) for each plant or national aggregation, as defined by shifting the time series forward by 365 days. For classification, we define a binary label "critical stock" if predicted stock falls below a critical threshold (e.g. normative stock days) in the following year.

We implement four base models: **Linear Regression (LR)**, **Support Vector Regression (SVR)** with polynomial kernel, **XGBoost** (gradient-boosted trees), and a **Neural Network (NN)** (multi-layer perceptron). Each model is trained on the same feature set. To leverage complementary strengths and to gain better results, we build a stacking ensemble: the first layer consists of the four base learners, which gave their predictions and we made those predictions (on validation folds) as become inputs for the second-layer meta-learner. We have used a simple linear model as the meta-learner for regression, learning to best combine fundamental forecasts [8]. (In practice, cross-validated predictions from base models form a new dataset to train the meta model.) A similar stacking approach is applied for the binary classification task, with a logistic meta-classifier.

Figure 1: Conceptual illustration of stacking ensemble learning.

This stacked approach is inspired from the ensemble theory, that is combining diverse learners often



results into generalization and robustness [6], [8]. Figure 1 gives conceptual illustration for this pipeline. We evaluate models using standard metrics: for Regression - the coefficient of determination (R^2) and Mean Absolute Error (MAE); for Classification - accuracy, precision, and recall. To simulate forecasting, model training and testing use time-aware splits (training on earlier years, testing on later years).

RESULTS

Table 1 & 2 summarizes the performance of the ensemble model for next-year coal stock prediction. The ensemble results into strong regression accuracy with an R^2 score of approximately 0.73 and a mean absolute error (MAE) of 16.46 thousand tonnes. For the binary classification task—predicting whether next year's stock exceeds a predefined threshold—the stacked classifier (trained on the base regressors' predictions) attains an accuracy of 80%, with both precision and recall at approximately 83%. These findings show that the ensemble provides consistent performance for both threshold-based classification and continuous forecasting by effectively capturing pertinent trends and patterns.
Code: https://github.com/shashank-v-mishra/Predicting-Coal-Stock-Levels.git

Metric	Value
R^2 Score	0.7319
Mean Absolute Error (MAE)	16.4640

Table 1 & 2. Regression & Classification results (predicting next-year coal stock).

Metric	Value
Accuracy	0.8000
Precision	0.8333
Recall	0.8333

DISCUSSION

The results show that the ensemble approach gives better and meaningful forecasts in comparison to the traditional approach using single models. In practical terms, these predictions can help in designing multiple policy and operational areas. First, knowing projected stock levels in advance helps planners pre-position coal supplies. For instance, if the model predicts a plant's coal days-of-supply will drop below normative levels, authorities can become proactive in getting extra rail shipments or releases from strategic reserves. This can be helpful in avoiding situations like the 2021 coal crisis, when rapid demand growth became more than the supply and stocks fell to ~4 days [2].Second, accurate stock forecasts will help in maintaining supply-chain coordination. The Ministry of Coal and Railways can use these forecasts to optimize train schedules and mine dispatch plans on time. For instance, coal flows can be rerouted from surplus areas if regional shortages are forecasted during monsoon months, using the "33% increase" in mine-level stocks already attained by recent government actions [3]. Similarly, power utilities and distribution companies can take care of demand-side measures (load shedding or imports) if critical shortfalls are observed.By highlighting days or plants that are likely to encounter critical stock levels, the classification results provide value. We can obtain easily actionable alerts (such as "super-critical" status) instead of just numerical estimates by approaching the shortage prediction as a classification task. These alerts could be communicated via dashboards to grid operators, triggering contingency protocols. As recent government reports note, even though stock levels have improved (55 MT by FY2025 vs 47 MT a year ago [10]), they remain below the normative requirement (~67 MT). Warnings based on predictions help in keeping stocks above the normative levels.Figure 2 illustrates the broader energy infrastructure (transmission grid), which contributes to sustaining the coal supplies. Machine learning forecasts thus contribute to overall energy resilience and availability: by ensuring coal-based generation can meet demand, they indirectly keep the grid stable. Practically speaking, including our forecasting model into current planning systems might improve resilience. For instance, the central dispatcher in their daily operational bulletins could suggest "coal priority" for particular units or ask for demand response based on model results.

Figure 2: Transmission network of a thermal power grid. Maintaining coal supply to such infrastructure is critical for energy resilience.

Limitations: Our study has limitations. The scope of features and the quality of the data limit the forecasting accuracy. It is challenging to forecast unplanned outages, extreme weather events (cyclones impacting coal transportation), and policy changes (such as import restrictions) based solely on historical stock data. During anomalous times (like lockdowns), model performance may deteriorate. Additionally, we make the assumption that past trends will recur; therefore, structural shifts in coal production or demand may weaken the model's dependability. Furthermore, the coarse national-level target (annual stock) may mask plant-specific nuances. Any type of research in future needs to incorporate additional data (coal auction prices, rainfall, plant outages) and must use more improved methods like sequence models (LSTM). Even though forecasting can be a wonderful technique but the expert judgement should never be ignored. Still, they offer a quantitative basis: if applied regularly, they would allow proactive policy, such as changing rail quotas or blending thresholds depending on expected shortfalls. Such models assist stakeholders move from reactive crisis management to anticipatory planning by linking policy and data analytics.

CONCLUSION

This paper presents an ensemble machine learning framework for forecasting next-year coal stock levels at India's thermal power plants. We predict future coal stocks and identify critical shortages with high accuracy using a stacking ensemble of Linear Regression, SVR, XGBoost, and Neural Networks and daily inventory data (2018–2023). The advantage of combining diverse learners is validated by the stacked model, which performs significantly better than individual base models.

From a policy standpoint, correct stock projections can greatly enhance resilience and energy security. They let planners maximize coal flows, keep normative reserves, and minimize supply interruptions in advance. This data-driven strategy enables rail operators, power companies, and government agencies to work more closely together. Integrating machine learning projections into decision-making will help India to maintain a consistent power supply during this vital time as it balances coal and renewables in its energy transition.

Acknowledgments: The author acknowledges the India Data Portal for providing the coal stock dataset, and the Ministry of Power for open data policy.

REFERENCES:

[1] "India's Coal Boom." Accessed: May 13, 2025. [Online]. Available: https://www.pib.gov.in/www.pib.gov.in/Pressreleaseshare.aspx?PRID=2118788

[2] "How Did India's Coal Stocks Fare in 2022 Post-Monsoon Season?," CEEW. Accessed: May 13, 2025. [Online]. Available: https://www.ceew.in/blogs/how-can-india-overcome-coal-shortage-crisis-and-build-stocks-for-thermal-plants

[3] "Adequate coal stock available, can meet 18.5 days' requirements of thermal power plants: Government," The Economic Times, Jul. 01, 2024. Accessed: May 13, 2025. [Online]. Available: https://economictimes.indiatimes.com/industry/energy/power/adequate-coal-stock-available-can-meet-18-5-days-requirements-of-thermal-power-plants-govt/articleshow/111411441.cms?from=mdr

[4] S. Li, X. Yang, and R. Li, "Forecasting Coal Consumption in India by 2030: Using Linear Modified Linear (MGM-ARIMA) and Linear Modified Nonlinear (BP-ARIMA) Combined Models," Sustainability, vol. 11, no. 3, Art. no. 3, Jan. 2019, doi: 10.3390/su11030695.
[5] M. Sakib, T. Siddiqui, S. Mustajab, R. M. Alotaibi, N. M. Alshareef, and M. Z. Khan, "An ensemble deep learning framework for energy demand forecasting using genetic algorithm-based feature selection", doi: 10.1371/journal.pone.0310465.
[6] F. Divina, A. Gilson, F. Goméz-Vela, M. García Torres, and J. F. Torres, "Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting," Energies, vol. 11, no. 4, Art. no. 4, Apr. 2018, doi: 10.3390/en11040949.
[7] "Stacking Ensemble Machine Learning With Python - MachineLearningMastery.com." Accessed: May 13, 2025. [Online]. Available: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/
[8] H. Huang, Q. Zhu, X. Zhu, and J. Zhang, "An Adaptive, Data-Driven Stacking Ensemble Learning Framework for the Short-Term Forecasting of Renewable Energy Generation," Energies, vol. 16, no. 4, Art. no. 4, Feb. 2023, doi: 10.3390/en16041963.
[9] "Demand – Coal 2023 – Analysis," IEA. Accessed: May 13, 2025. [Online]. Available: https://www.iea.org/reports/coal-2023/demand
[10]"Govt expects 55 MT coal stocks at power plants by March-end, higher than last year," Financialexpress. Accessed: May 13, 2025. [Online]. Available: https://www.financialexpress.com/business/industry-at-55-mt-month-end-coal-stocks-at-thermal-units-below-norm-3791909/
[11]K. Qu, "Research on linear regression algorithm," MATEC Web Conf., vol. 395, p. 01046, 2024, doi: 10.1051/matecconf/202439501046.
[12]T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[13]K. G, I. K P, J. Hasin A, L. F. J. M, S. Siluvai, and K. G, "Support Vector Machines: A Literature Review on Their Application in Analyzing Mass Data for Public Health," Cureus, Jan. 2025, doi: 10.7759/cureus.77169.
[14]S. Gupta, S. Kumar, and P. Kumar, "EVALUATING THE PREDICTIVE POWER OF AN ENSEMBLE MODEL FOR ECONOMIC SUCCESS OF INDIAN MOVIES," jpm, vol. 10, no. 1, pp. 30–52, Sep. 2016, doi: 10.5750/jpm.v10i1.1182.