IJSES
Volume 9, Issue 6, pp. 18-22, 2025.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

Data-Driven Decision-Making Frameworks for Large-Scale Electric Infrastructure Programs

Wanqiu Chen
Exponent - Senior Associate - Construction Consulting
Oakland, California, United States

Abstract— *This article examines the specific characteristics of decision-making systems built on data-driven methodologies within the context of large-scale electric infrastructure development programs. The study underscores the relevance of leveraging big data, machine learning, and optimization techniques to overcome the limitations of traditional approaches, offering improved forecasting accuracy and greater transparency in results. The research outlines the full workflow—from data collection, preprocessing, and integration to predictive modeling using the XGBoost algorithm, optimized through Bayesian Optimization, and interpretability enabled by SHAP values. The proposed methodology is validated through a case study focused on evaluating the cost and operational performance of electrical substations, yielding strong performance metrics ($R^2 \approx 0.9567$, RMSE ≈ 0.8690, MAE ≈ 0.4875). These findings confirm that a data-driven framework can reduce costs, enhance resource allocation, and increase the reliability of decision-making processes in the energy sector.*

Keywords— *Data-Driven decision-making, electric infrastructure, XGBoost, Bayesian Optimization, SHAP, machine learning, big data, optimization, integrated methodological framework.*

I. Introduction

Electrical networks and infrastructure systems are inherently complex and dynamic, necessitating adaptive decision-making frameworks to optimize resource allocation and reduce operational costs. Traditional methods—primarily based on expert judgment and static models—often fall short in processing the vast, heterogeneous data now available, leading to inefficiencies in planning and execution of large-scale infrastructure projects [1].

The academic literature in this domain generally splits into two major streams. The first focuses on the development of methodological foundations and algorithmic solutions for optimizing infrastructure management processes. For instance, Chen Y. et al. [1] propose statistical analyses and demand-side management models that support energy transition efforts in urban energy hubs, significantly improving resource allocation efficiency. Zhang J. et al. [3] introduce a conceptual cost assessment model that combines the XGBoost algorithm with Bayesian optimization, showcasing the utility of machine learning in forecasting and risk mitigation for large infrastructure developments. Antwarg L. et al. [4] emphasize anomaly detection using autoencoders and the enhancement of model interpretability via Shapley Additive Explanations (SHAP), which strengthens model credibility and facilitates real-world deployment.

The second group of studies addresses the socio-economic dimensions of infrastructure, including the impact of transport and urban policies. Van Thang N. [2], for example, applies data-driven techniques to identify socio-economic inequalities in access to urban transportation, enabling targeted interventions in underserved areas. Fontoura W. B., Ribeiro G. M., and Chaves G. D. L. D. [5] present an analytical framework for evaluating the dynamic effects of urban mobility policies on socio-economic systems, highlighting the intricate interplay between infrastructure decisions and public welfare. Tucho G. T. [6] underscores the role of informal transport networks in advancing sustainability and equity, arguing that such systems can complement conventional models of urban mobility.

Together, these studies reflect both theoretical and applied perspectives on data-driven governance of large infrastructure programs. On one side, research focused on methodological and algorithmic development provides high-tech tools for real-time, data-informed decision-making. On the other, socio-economic analyses stress the importance of embedding technological solutions within socially equitable and policy-aligned frameworks. The literature also reveals a notable disconnect between algorithmic models and empirical evaluations of how infrastructure decisions affect socio-economic systems. Key issues such as the integration of heterogeneous data sources, uncertainty estimation in cost forecasting, and the contextual adaptation of technical models amid shifting political and economic conditions remain insufficiently explored—underscoring the need for interdisciplinary research to produce balanced and actionable solutions.

The objective of this study is to examine the use of data-driven decision-making systems within the scope of large-scale electric infrastructure development programs.

The contribution of this research lies in its integration of modern methods for optimization, forecasting, and interpretability into a unified methodological framework capable of processing diverse data streams while maintaining transparency in strategic decision-making.

The central hypothesis is that the integration of data-driven techniques with model explainability mechanisms can enhance the quality of strategic decisions, reduce costs, and improve resource distribution in the planning and execution of major electric infrastructure initiatives.

1. Theoretical Foundations of Data-Driven Decision-Making Systems for Large-Scale Electric Infrastructure Development

Modern electric infrastructure development programs are characterized by their high complexity, dynamic behavior, and multidisciplinary nature, which necessitate innovative

approaches to support strategic decision-making. Data-driven decision-making systems (DDMS) represent integrated information-analytical platforms that merge heterogeneous data sources with advanced machine learning algorithms and optimization techniques to support management decisions in the energy sector.

At the core of data-driven methodologies is the use of large-scale data (Big Data) and sophisticated analytics to generate evidence-based insights. The theoretical basis of such systems is grounded in statistical learning, optimization theory, and artificial intelligence methods, allowing for the incorporation of both traditional data and real-time environmental dynamics. For instance, the XGBoost algorithm has proven highly effective in regression and classification tasks, as demonstrated across various infrastructure-related studies. Bayesian Optimization further enhances these models by fine-tuning hyperparameters, striking a balance between accuracy and computational efficiency [1].

Modern DDMS architectures consist of several interconnected components:

- Data Collection and Integration. The first stage involves aggregating information from technical specifications, operational metrics, IoT sensor data, and socio-economic indicators. This ensures a comprehensive and cohesive dataset for analysis.
- Data Preprocessing and Analysis. Techniques such as data cleaning, normalization, and correlation analysis are used to reduce noise and redundancy. Multi-step feature selection processes—combining correlation analysis with forward selection—are crucial to identify relevant inputs.
- Modeling and Forecasting. Machine learning algorithms including XGBoost, neural networks, and support vector machines (SVM) are used to develop predictive models for key infrastructure performance indicators. A notable innovation here is the use of Bayesian Optimization for hyperparameter tuning, which significantly enhances model precision.
- Result Interpretation and Decision Support. Explainable AI methods like SHAP (Shapley Additive Explanations) offer transparency in model predictions by quantifying each feature's contribution to the outcome, thereby strengthening strategic justification and fostering trust in automated decision support systems [2, 5].

These interrelated components are summarized in Table 1.

Theoretically, a DDMS can be viewed as a multi-layered architecture, where each level corresponds to a specific functional domain. The first level handles data aggregation and preprocessing; the second focuses on model construction and optimization; the third is dedicated to interpretability and visualization for end-users. This layered design ensures the system's flexibility, scalability, and adaptability to changing external conditions and user demands [5].

Furthermore, the adoption of data-driven methods supports a shift from static to dynamic control models—critical in the context of large-scale electric programs. Through ensemble methods such as XGBoost and hyperparameter optimization algorithms, the system can rapidly adjust to new data, minimizing the risk of overfitting and improving predictive accuracy [1].

TABLE 1. Main Components of a Data-Based Decision-Making System for Electrical Infrastructure [1, 2, 5]

System Component	Description	Method/Tool
Data Collection & Integration	Aggregation of technical, operational, economic, and social data	ETL processes, Big Data tools, API integration
Data Preprocessing & Analysis	Cleaning, normalization, noise reduction, and feature selection	Statistical analysis, correlation analysis, forward selection
Modeling & Forecasting	Development of predictive models for performance and cost evaluation	XGBoost, neural networks, SVM; hyperparameter tuning (Bayesian Optimization)
Interpretation & Decision Support	Model explainability and feature contribution analysis	SHAP, Explainable AI techniques

In conclusion, the integration of these components enables the creation of a robust DDMS capable of supporting evidence-based decisions in large-scale infrastructure initiatives. Such systems help reduce costs, improve resource allocation, and enhance the reliability of electrical system operations.

Ultimately, the theoretical foundations of data-driven decision-making systems for electric infrastructure development lie in the synthesis of advanced data collection, processing, predictive modeling, and explainability methods. This integrated framework not only increases forecast precision but also ensures transparency and accountability in management decisions—an essential requirement for the successful execution of complex infrastructure projects.

2. A Methodological Framework for Data-Driven Decision-Making

Modern decision-making systems for large-scale electric infrastructure development require the seamless integration of heterogeneous data sources, the application of advanced machine learning and optimization algorithms, and a strong emphasis on model transparency and explainability. A data-driven approach offers the ability to account for rapid changes in technical, economic, and social indicators, while enabling dynamic adaptation to evolving external conditions—critical for the effective governance of complex infrastructure projects.

The first stage of the proposed framework involves data collection and integration. In contemporary infrastructure initiatives, relevant data come from technical specifications, operational performance metrics, IoT sensors, and economic and social indicators. To ensure data completeness and currency, technologies such as ETL (Extract, Transform, Load), Big Data platforms, and API integration are used to consolidate disparate sources into a unified database for subsequent analysis.

Following collection, the data must undergo preprocessing, including cleaning, normalization, outlier detection, and correlation analysis to reveal interdependencies among features. At this stage, correlation analysis, forward feature selection, and principal component analysis (PCA) are widely

applied to reduce dimensionality and eliminate redundancy [6]. This allows for the identification of the most relevant variables, optimizing predictive power while minimizing computational overhead.

The next phase involves modeling and optimization, which form the foundation for managerial decision-making. Within this framework, the XGBoost algorithm—demonstrated by Chen and Guestrin [2] to be highly effective for regression and classification on large datasets—is prioritized. To enhance predictive accuracy and mitigate overfitting, Bayesian Optimization is employed for hyperparameter tuning. Compared to conventional techniques like grid search, this method achieves higher efficiency by adapting model settings to the characteristics of the data while reducing computational demands [2].

A crucial component of data-driven decision-making is model interpretability. Transparency not only builds user trust in automated forecasts but also clarifies how individual features contribute to final outcomes. For this purpose, techniques such as SHAP and LIME are widely adopted. These tools quantify the influence of each variable, enabling stakeholders to justify strategic decisions and adjust project parameters based on objective insights [3].

The structure of this methodological framework is summarized in Table 2.

TABLE 2. The Main Stages of the Methodological Framework for Data-Driven Decision-Making [2, 3, 6]

Stage	Description	Methods/Tools
Data Collection & Integration	Aggregation of technical, operational, economic, and social data	ETL processes, Big Data platforms, API integration
Preprocessing & Feature Selection	Data cleaning, normalization, correlation analysis, dimensionality reduction	Correlation analysis, forward selection, PCA
Modeling & Optimization	Construction of predictive models; hyperparameter tuning to improve accuracy and adaptability	XGBoost, neural networks, SVM; Bayesian Optimization, Grid Search
Interpretation & Decision Support	Model transparency through quantitative assessment of feature importance to support strategic decisions	SHAP, LIME

The integration of these stages results in a robust decision support system tailored for large-scale electric infrastructure programs. By leveraging advanced methods of data collection and preprocessing, employing effective modeling and optimization algorithms, and ensuring model transparency through interpretability tools, the framework provides a strong foundation for a data-driven governance approach.

In sum, the proposed methodological framework constitutes an integrated system where data acquisition, processing, and analysis are closely aligned with optimization and explainability mechanisms. This synergy not only enhances the accuracy of performance forecasting but also strengthens the transparency and justification of managerial decisions—essential factors for the successful implementation of complex infrastructure development initiatives.

3. Practical Implementation and Performance Evaluation

As part of this study, an existing research project [3] was analyzed to examine the application of data-driven approaches for optimizing and supporting decision-making in large-scale electric infrastructure development programs. The practical implementation followed a framework that integrated data collection and preprocessing techniques, predictive modeling with hyperparameter optimization, and interpretability tools such as SHAP to justify strategic decisions. A case study was conducted to evaluate the operational and cost efficiency of electrical substations and energy hubs using real-world data.

To demonstrate the practical use of the framework, a case involving the estimation of conceptual costs and performance indicators of electrical substations was selected. The dataset—provided by industry partners—included technical specifications, economic performance metrics, IoT sensor data, and administrative records. These data were processed using ETL pipelines and Big Data platforms to construct a unified information environment for subsequent analysis.

Next, preprocessing steps such as data cleaning, normalization, and correlation analysis were applied to remove redundancy and minimize the risk of overfitting in predictive models. Key variables were selected using correlation-based and stepwise methods. Predictive tasks—such as estimating construction costs, operational expenses, and efficiency indicators—were addressed using the XGBoost algorithm, with hyperparameters fine-tuned via Bayesian Optimization. SHAP values were employed to interpret the results, providing insights into the contribution of each feature to the final predictions and ensuring transparency in managerial recommendations [2].

The case study placed particular emphasis on comparative analysis of alternative design decisions. For instance, modeling results enabled an evaluation of how different substation configurations or transformer types influenced the overall cost and performance of the project. These insights allow infrastructure development strategies to be rapidly adjusted, ensuring that technical and economic criteria are optimally balanced [3].

Model performance was evaluated using standard forecasting accuracy metrics, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and adjusted R^2. Each of these has distinct interpretive benchmarks: R^2 values approaching 1 (typically >0.9) suggest a high degree of explained variance and model reliability, while values below 0.5 often signal the need for model refinement. Adjusted R^2 accounts for model complexity by factoring in the number of predictors, offering a more objective assessment of model fit. RMSE and MAE measure absolute forecasting errors, where lower values indicate better alignment between predicted and actual values. RMSE captures the standard deviation of prediction errors, while MAE reflects the average absolute deviation—values around or below 0.5 are generally acceptable for high-precision models, indicating minimal average error without heavy sensitivity to outliers.

In this case, the optimized BO-XGBoost model achieved the following results on the test set: $R^2 \approx 0.9567$, RMSE ≈ 0.8690, MAE ≈ 0.4875, and adjusted R^2 ≈ 0.9549. These figures

International Journal of Scientific Engineering and Science
Volume 9, Issue 6, pp. 18-22, 2025. ISSN (Online): 2456-7361

indicate a high-performing model capable of explaining nearly all data variance with minimal predictive error. For comparison, several other methods were tested, including Gradient Boosting Decision Trees (GBDT), Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Multiple Linear Regression (MLR), to assess their relative effectiveness within the same context [3].

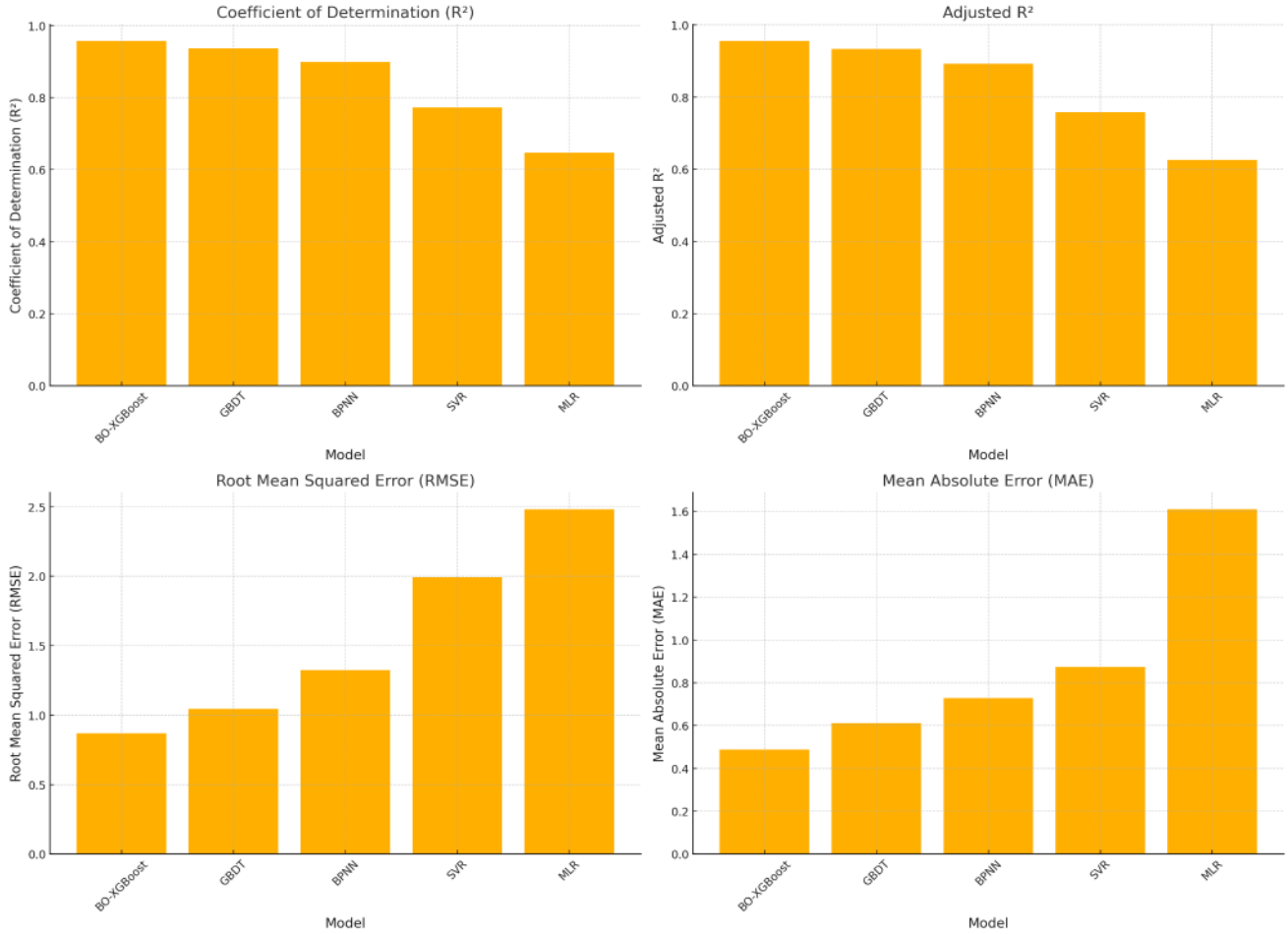

Fig. 1. Comparison of the Effectiveness of Predictive Models for Evaluating Electrical Infrastructure [3]

The proposed approach significantly enhanced the quality of managerial decision-making in the context of electric infrastructure development. The high predictive accuracy achieved by the models—evidenced by the performance metrics—provides a reliable foundation for selecting optimal technical and economic solutions. At the same time, the use of the SHAP method enabled the identification of key factors influencing project cost and operational performance, allowing for timely adjustments to design strategies. For instance, feature contribution analysis revealed that substation design type and transformer parameters had a substantial impact on total project cost, equipping planners with critical insights for investment planning.

Additionally, the ability to rapidly update models and fine-tune hyperparameters through Bayesian Optimization enables the framework to adapt to changes in the external environment and respond efficiently to new data. This adaptability is essential for managing large-scale programs where market dynamics and technological developments require continuous monitoring and strategic recalibration.

In summary, the practical implementation of the framework demonstrated its effectiveness both in terms of predictive precision and in its ability to deliver actionable insights for planning and operational management in electric infrastructure. This integrated data-driven approach supports cost reduction, resource allocation optimization, and improved reliability of infrastructure performance.

II. Conclusion

The approach presented integrates the collection and preprocessing of heterogeneous data, predictive modeling via the XGBoost algorithm optimized through Bayesian Optimization, and the use of interpretability tools such as SHAP, ensuring transparency and justification of forecasting results. Its practical application—focused on estimating the cost and operational characteristics of electrical substations—confirmed the high accuracy of the model and the effectiveness of the proposed methodology ($R^2 \approx 0.9567$, RMSE ≈ 0.8690, MAE ≈ 0.4875).

This methodological platform offers a dynamic and adaptable framework capable of integrating data from various sources and supporting informed decision-making based on big data analytics. As demonstrated, data-driven strategies improve resource distribution, lower operating costs, and enhance the reliability of electric infrastructure systems.

Nonetheless, further research should aim to expand the data foundation, incorporate additional information sources, and refine interpretability algorithms. These advancements will strengthen the framework's ability to support energy sector planning with even greater precision and robustness.

REFERENCES

1. Chen Y. et al. Data-driven approaches to achieve energy transition: Statistical analysis of demand-side management and smart decision making in urban energy hub networks //Sustainable Cities and Society. – 2024. – Vol. 101. – pp. 1-10.
2. Van Thang N. Data-Driven Insights into Socio-Economic Disparities in Urban Transportation Accessibility //Open Journal of Robotics, Autonomous Decision-Making, and Human-Machine Interaction. – 2025. – Vol. 10 (2). – pp. 1-8.
3. Zhang J. et al. A data-driven framework for conceptual cost estimation of infrastructure projects using XGBoost and Bayesian optimization //Journal of Asian Architecture and Building Engineering. – 2025. – Vol. 24 (2). – pp. 751-774.
4. Antwarg L. et al. Explaining anomalies detected by autoencoders using Shapley Additive Explanations //Expert systems with applications. – 2021. – Vol. 186. – pp. 1-10.
5. Fontoura W. B., Ribeiro G. M., Chaves G. D. L. D. A framework for evaluating the dynamic impacts of the Brazilian Urban Mobility Policy for transportation socioeconomic systems: A case study in Rio de Janeiro //Journal of Simulation. – 2020. – Vol. 14 (4). – pp. 316-331.
6. Tucho G. T. A review on the socio-economic impacts of informal transportation and its complementarity to address equity and achieve sustainable development goals //Journal of Engineering and Applied Science. – 2022. – Vol. 69 (1). – pp. 28.