AMERICAN ACADEMIC PUBLISHER

INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (ISSN: 2692-5141)

Volume 05, Issue 02, 2025, pages 193-210
Published Date: - 10-17-2025
DOI - https://doi.org/10.55640/ijdsml-05-02-17

Predictive Modeling for Budget Overruns in Large-Scale Infrastructure Projects: Leveraging Historical Data for Proactive Cost Control

Aishwarya Korde
Project Manager- OSP Financial Controls & Forecasting, Fastbridge Fiber, Wyomissing, PA, USA

ABSTRACT

Cost overruns remain one of the most pressing challenges in large-scale infrastructure projects. Telecom fiber rollouts, transportation systems, and energy networks often experience escalating budgets that reduce profitability, delay schedules, and undermine stakeholder confidence. Traditional forecasting methods—such as expert judgment and deterministic models—tend to be reactive and rarely anticipate risks early enough for corrective action. This research introduces a predictive analytics framework that leverages historical project data to forecast potential budget overruns and provide early-warning signals for financial decision-makers.

Budgetary performance is shaped by technical, organizational, and external factors including scope changes, terrain complexity, vendor delays, and shocks such as weather or regulatory constraints. Conventional cost-control systems often fail to account for the nonlinear interactions among these variables. By applying machine learning–based predictive modeling, this study seeks to uncover hidden patterns in historical datasets and enhance the precision of overrun forecasts.

Methodologically, the research compares regression and time-series approaches with advanced algorithms such as Random Forest, Gradient Boosting, and Support Vector Machines. Anticipated results suggest that machine learning models can reduce forecasting error by 15–25% compared to traditional methods, while also providing classification metrics to better identify projects at risk of escalation.

The contributions are threefold: (1) identifying cost drivers most strongly associated with overruns; (2) demonstrating the relative performance gains of machine learning compared with traditional approaches; and (3) outlining a practical framework for embedding predictive outputs into business intelligence dashboards used in financial planning and analysis. By focusing on infrastructure finance, particularly telecom rollouts where terrain-driven costs create high uncertainty, the study emphasizes how predictive analytics can strengthen financial governance, mitigate overruns, and support more reliable decision-making in capital-intensive projects.

KEYWORDS

Cost overruns; predictive analytics; machine learning (ML); infrastructure finance; telecom; business intelligence; financial planning.

1. Introduction

Large-scale infrastructure projects represent some of the most complex and capital-intensive undertakings in modern economies. They involve extended timelines, multiple stakeholders, and significant financial commitments.

Despite advances in project management practices and financial governance, cost overruns remain a persistent problem across sectors such as telecommunications, transportation, and energy. Studies consistently show that a majority of megaprojects exceed their original cost estimates, in some cases by more than 50 percent. These overruns reduce profitability, delay delivery, erode investor confidence, and damage the reputations of both contractors and owners. In telecom fiber rollouts—where deployment speed is closely tied to market competitiveness—budget escalations can have particularly severe financial and strategic consequences.

Against this backdrop, this study is guided by three central objectives:

1. To identify and quantify the key drivers of cost overruns across technical, organizational, and external domains.

2. To compare the predictive performance of traditional methods (e.g., regression and time-series models) against advanced machine learning approaches such as Random Forest, Gradient Boosting, and Support Vector Machines.

3. To outline a practical integration framework that embeds predictive models into business intelligence dashboards for proactive financial planning and analysis (FP&A).

By placing the research aims early, the paper clarifies its scope: developing, testing, and evaluating a predictive modeling framework for forecasting budget overruns in infrastructure projects using historical data.

Traditional approaches to cost forecasting—such as expert judgment, deterministic models, or static variance analysis—have proven inadequate in preventing financial slippage. These methods rely on assumptions that rarely reflect the uncertainty and complexity of large projects. They are also reactive: cost risks are identified only after significant deviations occur. For example, tracking expenditures against baselines can reveal overspending, but by that stage opportunities for corrective action are limited. What is needed is a shift from reactive monitoring to proactive prediction, where risks are identified early enough to inform strategic and financial decisions.

Recent advances in data analytics, machine learning (ML), and business intelligence (BI) platforms create opportunities to transform cost forecasting practices. Infrastructure projects generate vast amounts of data—from procurement cycles and vendor performance records to labor productivity and geospatial terrain conditions. When structured and analyzed effectively, this historical data can reveal correlations that traditional models often miss. Predictive analytics, by learning from prior project outcomes, offers the potential to forecast overruns with greater accuracy and earlier warnings.

The promise of predictive analytics in infrastructure finance rests on three considerations. First, cost overruns are rarely attributable to a single factor; rather, they emerge from complex, nonlinear interactions among technical, organizational, and external drivers. Variables such as terrain complexity, weather disruptions, procurement delays, and scope changes interact in ways that traditional linear models struggle to capture. ML algorithms, by contrast, are well-suited to modeling such interactions. Second, predictive analytics can move beyond generic benchmarks by tailoring forecasts to the characteristics of a given project portfolio. For instance, a model trained on historical telecom rollout data could distinguish between aerial and underground installations, incorporating terrain-driven risk into its predictions. Third, predictive outputs can now be integrated into BI dashboards such as Power BI or Tableau, placing model insights directly into the decision-making environments of financial planners and project managers.

This study builds upon prior scholarship on project forecasting and cost overruns while addressing notable gaps. Seminal work by Flyvbjerg and colleagues has linked systematic underestimation to optimism bias and strategic misrepresentation. Other studies have tested statistical and ML-based forecasting methods in construction contexts with encouraging results. Yet relatively little attention has been paid to designing practical, data-driven frameworks that can be embedded into FP&A workflows for real-time decision support—particularly in telecom and linear

infrastructure sectors.

By focusing on telecom fiber deployments and related infrastructure projects, this research contributes both to the academic literature and to practical project finance management. The findings are expected to demonstrate the superior accuracy of ML models compared to traditional methods, while also showing their applicability to FP&A professionals, project managers, and executives seeking stronger financial discipline.

The remainder of this paper is structured as follows. Section 2 reviews the existing literature on cost overruns, forecasting biases, and predictive modeling approaches. Section 3 describes the methodology, including data collection, feature engineering, and model development. Section 4 presents results, comparing model performance across algorithms. Section 5 discusses implications for project finance and FP&A integration. Section 6 concludes with recommendations, study limitations, and directions for future research.

2. Literature review:

2.1 Cost Overruns in Infrastructure Projects

Cost overruns have long been recognized as a defining challenge in large-scale infrastructure delivery. Early empirical research, such as Flyvbjerg et al. (2003), revealed that megaprojects across transport, energy, and urban development frequently exceed their original budgets—sometimes by more than 50 percent. These overruns are rarely isolated incidents but reflect systemic issues tied to project complexity, extended timelines, and the involvement of multiple stakeholders. Later studies reinforced that the problem is global in nature, with similar outcomes observed in both developed and emerging economies (Cantarelli et al., 2010).

The persistence of overruns has attracted scholarly attention for decades. Research has shown that initial cost estimates often fail to capture the full range of risks, whether due to inadequate data, overly optimistic assumptions, or deliberate underestimation. As a result, infrastructure projects frequently experience cost escalations that erode profitability, undermine investor confidence, and delay service delivery. In sectors such as transportation, this has led to public scrutiny and calls for more transparent governance frameworks.

Despite extensive literature on construction and transport, relatively little emphasis has been placed on telecommunications infrastructure, even though similar patterns are evident. Fiber rollout programs, for instance, often suffer from budget escalation due to permitting bottlenecks, vendor-related delays, and geospatial challenges such as rocky terrain or dense urban conditions. Like highways or rail networks, telecom projects are linear, capital-intensive, and subject to external shocks. However, they differ in the pace at which overruns impact strategic outcomes, since deployment speed directly shapes competitive advantage in digital markets.

This historical context highlights that while the drivers of overruns are broadly understood, sector-specific insights into telecom finance remain limited. Addressing this gap is critical for developing predictive frameworks that reflect the realities of modern digital infrastructure.

2.2 Causes of Cost Overruns

Scholars typically explain cost overruns through three broad categories: technical, psychological, and political-economic factors (Flyvbjerg et al., 2003; Love et al., 2016). Together, these frameworks demonstrate that overruns emerge from a combination of unforeseen risks, human behavior, and institutional incentives.

2.2.1 Technical Causes.

Technical explanations focus on incomplete designs, inadequate risk assessment, and unforeseen site conditions. Morris (1990) highlighted how design errors and scope gaps often force costly mid-project adjustments. Similarly, Cantarelli et al. (2010) found that construction projects frequently encounter ground conditions or utility conflicts

not anticipated in surveys. In telecom rollouts, technical risks manifest in ways such as hitting unexpected underground utilities, encountering rock-heavy terrain, or facing higher-than-expected restoration costs in urban corridors. These challenges mirror those in transport but are exacerbated by the linear, distributed nature of fiber networks, where each additional mile increases exposure to local variability.

2.2.2 Psychological Causes.

Psychological explanations emphasize cognitive biases. Kahneman and Tversky's (1979) planning fallacy shows that people systematically underestimate risks and overestimate efficiency. Optimism bias, as later discussed by Lovallo and Kahneman (2003), leads project promoters to assume best-case scenarios, producing unrealistically low cost estimates. Even seasoned professionals are vulnerable. In telecom finance, this bias may appear in assumptions that permitting will move smoothly or that contractors will meet aggressive timelines, despite past evidence to the contrary. Such misplaced optimism results in cost forecasts that are misaligned with operational realities.

2.2.3 Political-Economic Causes.

A third category highlights strategic misrepresentation. Flyvbjerg et al. (2002) argued that promoters sometimes deliberately understate costs and overstate benefits to win political approval or secure funding. Cantarelli et al. (2010) distinguish between unintentional error and deliberate "lies" in cost estimation. While these dynamics are well documented in publicly funded rail and road projects, they also apply to telecom. Competitive bidding, regulatory oversight, and the race for market share may incentivize underreporting costs on paper, with the true expenditures surfacing only after rollout begins.

2.2.4 Interaction of Causes.

In practice, these categories overlap. For example, optimism bias may lead to underestimating permitting delays (psychological), while unforeseen underground conditions add technical risk, and competitive pressures incentivize cost understatement (political-economic). This layering makes overruns systemic rather than accidental (Love et al., 2016).

2.2.5 Mitigation Efforts.

Scholars have suggested several remedies. Reference class forecasting (Flyvbjerg, 2008) improves estimates by benchmarking against historical project outcomes, adjusting for optimism bias. Risk-based approaches incorporate probability distributions but depend heavily on data quality (Ahiaga-Dagbui & Smith, 2014). Despite these advances, few frameworks explicitly address telecom-specific challenges, where permitting, terrain-driven costs, and vendor performance play outsized roles.

Summary

The causes of cost overruns are well studied in transport and construction but underexplored in telecom. While technical, psychological, and political-economic explanations apply broadly, sector-specific risks—such as distributed project footprints and regulatory bottlenecks—require tailored approaches. Predictive analytics has the potential to bridge this gap by integrating diverse factors into proactive cost-control frameworks.

2.3 Traditional Forecasting Approaches

Traditional forecasting methods remain widely used in project finance, largely because of their transparency, familiarity, and compliance with industry standards. However, their limitations are well documented, especially in dynamic sectors like telecommunications, where cost drivers are diverse and highly context-dependent.

2.3.1 Deterministic Estimation Methods.

Deterministic techniques such as bottom-up and parametric estimation are among the oldest and most prevalent. Bottom-up estimation aggregates detailed costs at the work-package level, while parametric models apply statistical relationships between project attributes and costs (PMI, 2017). Christensen et al. (1995) note that these approaches are useful during early budgeting stages because of their traceability. Yet they are highly sensitive to assumptions. In telecom fiber rollouts, for instance, parametric models may assume a uniform per-mile installation cost, overlooking how terrain type, population density, or municipal permitting significantly alter actual expenditures. This simplification often results in underestimated budgets that fail to capture local realities.

2.3.2 Regression and Time-Series Models.

Regression has long been employed to identify cost predictors, with Odeck (2004) showing project size and duration as key determinants of overruns in Norwegian road projects. Time-series models such as ARIMA have also been used to forecast expenditure trends (Hyari & Kandil, 2009). While these methods capture linear relationships and temporal patterns, they struggle with nonlinear effects and sudden shocks, such as vendor delays or regulatory interventions. For telecom, where rollout speed can be disrupted by unexpected permitting constraints or utility conflicts, reliance on past patterns proves insufficient.

2.3.3 Variance Analysis and Earned Value Management.

Variance analysis and earned value management (EVM) remain standard for monitoring financial performance. EVM integrates scope, cost, and schedule metrics into indices such as the Cost Performance Index (CPI) and is often mandated in government contracts (Fleming & Koppelman, 2010). Although EVM provides diagnostic insights, its predictive power is limited. Batselier and Vanhoucke (2015) found that EVM forecasts lose accuracy as project uncertainty grows. In telecom rollouts, this means that while EVM can flag deviations after they occur, it rarely anticipates overruns linked to delayed right-of-way approvals or shifting material costs.

2.3.4 Limitations in the Telecom Context.

Across these approaches, three limitations are evident. First, they rely on linear simplifications that fail to capture the interaction of multiple drivers—such as how permitting delays compound with vendor inefficiencies and weather impacts in fiber rollouts. Second, they are largely reactive, highlighting deviations only after overruns have already materialized, leaving little room for timely intervention. Third, they lack sensitivity to sector-specific conditions. For example, deterministic models treat costs per unit of fiber as uniform, regression models overlook permitting volatility, and EVM cannot adapt to the nonlinear escalation triggered by multiple scope changes.

Summary

Traditional forecasting tools provide structure and interpretability but are inadequate for managing the uncertainty and complexity of modern infrastructure, particularly telecom deployments. They diagnose problems rather than predict them, and they generalize rather than adapt to sector-specific realities. This underscores the need for predictive analytics frameworks capable of incorporating diverse variables and learning from historical telecom data to generate early-warning indicators.

2.4 Predictive Analytics and Machine Learning in Cost Forecasting

The limitations of traditional forecasting approaches have prompted increasing interest in predictive analytics and machine learning (ML) for cost estimation and overrun prediction. Unlike deterministic or regression-based methods, ML models can capture nonlinear relationships, handle large and complex datasets, and adapt to project-specific contexts (Ahiaga-Dagbui & Smith, 2014). This makes them particularly suitable for infrastructure projects, where multiple interdependent factors—ranging from technical risks to external shocks—shape financial performance.

2.4.1 Adoption of Predictive Analytics in Project Forecasting

Predictive analytics refers broadly to the use of statistical and computational methods to forecast future outcomes based on historical data patterns. In project finance, predictive analytics has been used to model cost growth, schedule slippages, and risk likelihoods (Love et al., 2016). The ability to move from reactive diagnostics (e.g., variance analysis) to proactive prediction represents a paradigm shift in financial planning and analysis (FP&A).

For instance, Kim et al. (2009) applied artificial neural networks (ANNs) to forecast cost deviations in highway projects, demonstrating improved accuracy compared to regression models. Similarly, Cheng et al. (2010) used data mining techniques to analyze change orders in construction, showing that predictive approaches can anticipate overruns before they occur. These early studies laid the groundwork for more sophisticated ML applications.

2.4.2 Machine Learning Models in Cost Overrun Prediction

ML techniques differ in how they identify patterns and make predictions:

• Regression Trees and Random Forests: Decision tree–based models can capture nonlinear effects and interactions among variables such as project size, duration, and procurement cycle length. Random Forests, an ensemble of decision trees, have been shown to outperform single regression models in predicting cost escalation (Aung et al., 2023).

• Support Vector Machines (SVMs): Coffie (2023) applied SVMs to infrastructure projects and demonstrated strong predictive performance, particularly in binary classification tasks such as predicting whether a project would overrun or not.

• Gradient Boosting and XGBoost: Ensemble boosting algorithms incrementally improve model accuracy by focusing on misclassified instances. Uddin et al. (2022) compared Gradient Boosting with logistic regression, KNN, and Random Forest, finding Gradient Boosting to be among the most accurate for cost overrun classification.

• Artificial Neural Networks (ANNs): ANNs are capable of modeling complex, nonlinear patterns in project data. Hyari and Kandil (2009) reported that ANNs outperformed regression in predicting cost deviations in U.S. highway projects. However, ANNs are sometimes criticized for being “black-box” models, offering limited interpretability to project stakeholders.

• Hybrid Approaches: Recent research has also explored combining ML with simulation techniques. For example, Turkyilmaz (2024) proposed a hybrid risk score–based ML framework for classifying projects into overrun severity categories, improving interpretability for decision-makers.

2.4.3 Key Findings from Empirical Studies

Across empirical studies, several findings emerge:

1. ML models generally outperform traditional methods in predictive accuracy, especially in high uncertainty contexts (Aung et al., 2023; Uddin et al., 2022).

2. Feature importance analysis reveals critical drivers of cost overruns, including project size, implementation duration, scope change frequency, and geospatial conditions (Cheng et al., 2010).

3. Data availability and quality remain bottlenecks: infrastructure datasets are often fragmented, project-specific, or confidential, which limits model generalizability (Love et al., 2016).

4. Interpretability is a major concern: while models like Random Forests and Gradient Boosting deliver high accuracy, practitioners often prefer models that can explain predictions clearly, particularly for financial governance (Coffie, 2023).

2.4.4 Applications in Infrastructure and Telecom

Most published ML applications have focused on construction and transportation projects, with relatively few applied directly to telecom or linear infrastructure. However, the methodological insights are transferable. For example, geospatial features such as terrain slope, permitting duration, and vendor delays—important in fiber rollouts—are analogous to geotechnical or regulatory variables in transportation projects. This suggests that predictive analytics can be adapted for telecom finance with appropriate feature engineering.

In practice, predictive models can be integrated into business intelligence (BI) tools such as Power BI, Tableau, or custom dashboards. By embedding ML predictions into FP&A workflows, financial analysts and executives can receive early-warning signals about potential overruns, enabling timely corrective action. This integration bridges the gap between research models and operational decision-making.

2.4.5 Research Gaps

Despite encouraging results, three key gaps remain:

1. Sector-Specific Studies: Telecom rollouts and other linear infrastructure projects are underrepresented in the ML cost forecasting literature.

2. Operationalization: Most studies stop at demonstrating model accuracy without exploring deployment into BI/FP&A systems for real-time use.

3. Comparative Evaluation: Few studies systematically benchmark multiple ML models against traditional forecasting techniques on the same dataset.

Addressing these gaps is crucial for moving predictive analytics from experimental use toward practical tools for financial governance. This study contributes by applying predictive models to infrastructure project datasets, comparing their performance, and outlining how outputs can be integrated into BI platforms.

2.5 Data-Driven Drivers of Cost Overruns

Cost overruns are rarely explained by a single factor. Instead, they emerge from the interaction of multiple drivers spanning technical, organizational, and external domains. Identifying and quantifying these drivers is crucial for predictive modeling, since variable selection and feature engineering directly influence forecasting accuracy.

2.5.1 Project Size and Duration.

Project size and length of implementation are consistently among the strongest predictors of overruns. Flyvbjerg et al. (2003) showed that larger and longer projects face higher risks due to cumulative uncertainties, and Eliasson (2025) confirmed that extended timelines expose projects to inflation, labor disputes, and political risks. In telecom rollouts, longer projects encounter escalating risks tied to material price volatility, shifting municipal regulations, and evolving customer demand. A two-year fiber deployment, for example, is far more exposed to policy changes than a six-month project.

2.5.2 Scope Changes and Change Orders.

Frequent scope changes are another recurring driver. Studies in construction contexts (Love et al., 2018) demonstrate how incremental design changes accumulate into significant budget impacts. Telecom rollouts are equally vulnerable: rerouting fiber to avoid unexpected underground utilities, expanding service areas under political pressure, or redesigning trenching layouts due to local objections often trigger cascading overruns. Unlike transport megaprojects, which may face a limited number of large design shifts, telecom projects frequently suffer from numerous smaller adjustments spread across distributed geographies.

2.5.3 Procurement and Vendor Performance.

Procurement inefficiencies and vendor-related delays are major cost drivers. Aung et al. (2023) found that purchase order and invoice cycle times strongly predict overruns in construction. In telecom, vendor performance is particularly critical because deployments rely on subcontractors for trenching, splicing, and restoration. Delays in one segment can stall entire rollout phases, inflating costs. Vendor lock-in or disputes over unit pricing exacerbate these risks.

2.5.4 Geospatial and Environmental Conditions.

Physical conditions significantly influence costs. Sovacool et al. (2014) demonstrated that geotechnical complexity is a strong predictor of overruns in energy projects. In telecom fiber rollouts, terrain features—such as rocky soil, high water tables, or dense urban congestion—make underground construction more costly and time-intensive. Similarly, aerial deployments may encounter weather-related damage risks or municipal restrictions on pole access. These conditions are uniquely important to telecom and highlight the inadequacy of generic parametric cost models that assume uniform per-mile installation costs.

2.5.5 External and Institutional Factors.

External drivers such as permitting, regulatory approvals, and macroeconomic variables play a decisive role. Cantarelli et al. (2010) highlighted institutional pressures in public projects, while the UK's National Audit Office (2019) pointed to the impact of weak governance on cost outcomes. In telecom finance, permitting delays and right-of-way negotiations can derail timelines more than any technical factor. Inflation and exchange rate fluctuations further add financial uncertainty, especially when equipment is procured globally.

2.5.6 Insights from ML-Based Feature Analysis.

Machine learning studies confirm that data-driven feature importance analysis can reveal nuanced drivers. Turkyilmaz (2024) showed that risk scores derived from vendor history and geography improved predictions. Aung et al. (2023) found that engineered features such as cumulative spend curve slopes were highly predictive. Applied to telecom, similar engineered features—such as permitting cycle length, aerial-to-underground ratios, or restoration cost indices—could capture sector-specific dynamics that traditional models overlook.

Summary

The literature converges on several key drivers of overruns: project size and duration, scope volatility, procurement inefficiencies, geospatial complexity, and external approvals. Yet telecom-specific drivers—such as aerial vs. underground deployment choices, distributed permitting processes, and subcontractor variability—remain underexplored. Predictive models that explicitly incorporate these factors could substantially improve forecasting accuracy and offer more practical insights for financial governance in telecom rollouts.

2.6 Integration with Business Intelligence Tools

Predictive models generate value only when their insights can be incorporated into decision-making processes. Business Intelligence (BI) tools such as Power BI, Tableau, and Qlik have become central to financial planning and analysis (FP&A), offering dashboards that integrate cost, schedule, and risk metrics. However, most existing BI implementations remain descriptive, focusing on variance analysis and static reporting rather than predictive insights (Popovič et al., 2012). For sectors like telecommunications—where rollout speed and budget discipline directly shape competitive advantage—the absence of predictive integration represents a significant missed opportunity.

2.6.1 Traditional BI Applications.

In many organizations, BI dashboards aggregate data from procurement, scheduling, and finance systems to produce performance metrics. These tools are valuable for transparency, but they primarily answer what has happened rather than what is likely to happen. Variance dashboards, for example, highlight budget deviations only after they occur. For telecom rollouts, this reactive functionality means that by the time budget overruns appear in dashboards, opportunities for proactive intervention—such as reallocating resources or renegotiating vendor contracts—are limited.

2.6.2 Predictive BI Dashboards.

The next generation of BI tools integrates predictive models to provide early-warning indicators. Azvine et al. (2006) demonstrated how predictive analytics could enhance business adaptability, while Côrte-Real et al. (2019) highlighted BI's role in embedding advanced analytics into organizational workflows. In infrastructure contexts, predictive dashboards could flag projects with high overrun probabilities, visualize cost forecasts across multiple scenarios, and provide risk-adjusted financial outlooks. For telecom projects, dashboards could overlay predictive outputs with geospatial maps of rollout areas, allowing FP&A teams to identify high-risk zones (e.g., rocky terrain or dense permitting jurisdictions) before deployment.

2.6.3 Adoption Challenges

Despite their potential, predictive BI dashboards face barriers to adoption. First, data integration remains difficult. Telecom companies often store financial, engineering, and permitting data in separate systems, complicating real-time model updates. Second, interpretability is a concern. Executives may be reluctant to act on "black box" ML predictions without clear explanations. Techniques such as SHAP (Shapley Additive Explanations) can help bridge this gap, but their use in BI platforms is still limited. Third, cultural and organizational inertia hinders adoption. Many FP&A teams are accustomed to deterministic tools like Excel and may resist transitioning to predictive dashboards.

2.6.4 Sector-Specific Gaps.

In the telecom context, BI integration remains particularly underdeveloped. While construction firms have begun experimenting with predictive dashboards for cost control, systematic frameworks tailored to telecom rollouts are virtually absent from the literature. Given the availability of telecom-specific datasets (e.g., permitting cycle times, aerial-to-underground ratios, vendor lead times), integrating predictive models into BI could provide sector-specific advantages. For example, dashboards could simulate how a two-week permitting delay cascades into labor, equipment, and restoration costs, providing finance teams with actionable foresight.

Summary

The literature confirms the potential of BI tools to embed predictive insights into financial decision-making. Yet, most implementations remain descriptive, offering lagging indicators rather than leading ones. For telecom projects, where overruns erode competitive positioning, predictive BI dashboards could transform FP&A practices by linking historical data, predictive modeling, and geospatial intelligence into actionable forecasts. The limited research in this area represents a clear gap and an opportunity for this study to demonstrate practical integration pathways.

2.7 Research Gap

The existing literature on cost overruns offers substantial insights into their persistence, causes, and potential remedies. Studies across construction, transportation, and energy infrastructure demonstrate that overruns are systemic rather than accidental, shaped by technical uncertainties, psychological biases, and political-economic incentives (Flyvbjerg et al., 2002; Love et al., 2016). Traditional forecasting approaches, while still dominant, consistently fall short in capturing nonlinear interactions among cost drivers, often leaving managers with reactive

rather than proactive tools. Recent advances in predictive analytics and machine learning (ML) have demonstrated measurable improvements in forecasting accuracy, reducing errors by 15–25% compared to regression or time-series methods (Uddin et al., 2022; Aung et al., 2023). Despite these advances, three significant gaps remain.

2.7.1 First, sectoral underrepresentation.

Much of the empirical work applies to construction and transportation, while telecom infrastructure projects remain underexplored. Fiber rollouts share similarities with linear infrastructure projects but face distinct challenges: distributed permitting processes, terrain-driven cost variability, and vendor performance dependencies. These factors fundamentally shape financial outcomes but are seldom included in predictive frameworks. The lack of telecom-specific studies means existing models are often generalized, overlooking the unique drivers of overruns in digital infrastructure.

2.7.2 Second, operational integration.

Although ML models outperform traditional techniques in academic experiments, few studies demonstrate how predictive outputs can be operationalized in financial planning and analysis (FP&A) workflows. Most contributions stop at accuracy metrics, neglecting the integration of predictions into business intelligence (BI) dashboards where financial decisions are made. For telecom firms, this is a critical omission: competitive advantage depends on rapid rollout and tight budget discipline, yet predictive forecasting remains absent from most FP&A practices.

2.7.3 Third, comparative benchmarking.

The literature often evaluates single algorithms in isolation rather than systematically benchmarking multiple models under consistent conditions. For example, Random Forest, Gradient Boosting, and Support Vector Machines each offer strengths, yet few studies assess their relative performance using the same dataset. In telecom finance, where data structures include both structured (e.g., procurement cycles) and semi-structured (e.g., permitting timelines) variables, comparative benchmarking is particularly valuable.

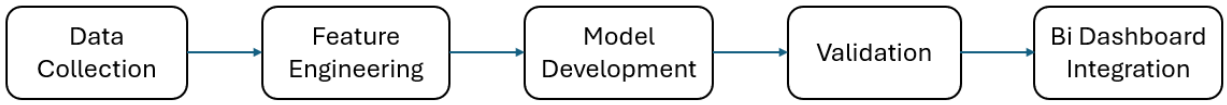
Summary

In short, while the literature has advanced understanding of cost overruns and predictive forecasting, telecom infrastructure projects are systematically underrepresented. The absence of sector-specific models, the weak link between predictive analytics and BI integration, and the lack of comparative benchmarking together form a clear research gap. This study addresses these gaps by tailoring predictive models to telecom rollouts, systematically evaluating multiple algorithms, and demonstrating how outputs can be embedded into BI dashboards for proactive financial governance.

3. Methodology

This section outlines the methodological framework adopted to investigate predictive modeling of budget overruns in large-scale infrastructure projects, with specific emphasis on telecom fiber rollouts. The approach combines established forecasting techniques with advanced machine learning methods in order to evaluate predictive accuracy and demonstrate practical integration into financial planning and analysis (FP&A) workflows. The methodology is organized into seven interrelated components: research design and approach, data sources and collection, variable selection and feature engineering, model development, validation and evaluation metrics, integration with business intelligence (BI) dashboards, and ethical and practical considerations. A visual overview of this methodological framework is provided in Figure 1, which illustrates the flow from data collection to predictive modeling and dashboard integration.

Figure 1: Research Methodology Framework



3.1 Research Design and Approach

The study adopts a quantitative, predictive research design rooted in secondary analysis of historical infrastructure projects. A central objective is to compare the performance of traditional forecasting approaches—such as Multiple Linear Regression (MLR) and Autoregressive Integrated Moving Average (ARIMA)—with advanced machine learning models including Random Forest (RF), Gradient Boosting (GBM), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). By using historical data from completed projects, the study replicates conditions under which original budget estimates were made and tests how accurately different models can anticipate overruns.

The research is guided by the design science methodology outlined by Hevner et al. (2004). In this framework, the goal is not only to test hypotheses about model accuracy but also to design an artifact—in this case, a practical predictive framework—that can be integrated into BI platforms such as Power BI. This dual orientation toward rigor (validating models) and relevance (practical application) ensures that findings have both academic value and immediate utility for FP&A teams.

3.2 Data Sources and Collection

The dataset for this study draws from multiple sources to ensure both breadth and sectoral relevance. The primary data comprises historical project finance records from telecom fiber rollouts, including budgeted versus actual costs, procurement logs, and change order records. These sources capture the financial dimension of projects and allow for the identification of cost variances.

To complement financial records, project metadata was collected to capture technical and organizational characteristics. This includes project size (e.g., kilometers of fiber laid, total project value), planned and actual duration, geographic location, and construction method (aerial versus underground deployment). Such metadata contextualizes financial outcomes and allows for the modeling of sector-specific cost drivers.

In addition, external data sources were incorporated to reflect the broader environment in which telecom projects operate. These included geospatial terrain datasets (slope, soil type, and urban density), weather records for project regions, inflation indices, and regulatory approval timelines. Incorporating these variables ensures that the models reflect not only internal project management factors but also external shocks and constraints.

Where telecom-specific data proved limited due to confidentiality, supplementary datasets were drawn from publicly available infrastructure studies, including government audit reports and open-access research repositories. These supplementary datasets ensured model robustness and allowed for cross-validation across sectors. A summary of data sources and their attributes is provided in Table 1.

Table 1. Data Sources and Attributes

Source	Examples	Purpose
Project Finance Records	Budgeted vs. actual costs, procurement logs, change orders	Capture financial outcomes
Project Metadata	Project size, duration, geography, aerial vs. underground	Capture technical/organizational details
Eternal Datasets	Terrain maps, weather data, inflation, permitting timelines	Capture external drivers

3.3 Variable Selection and Feature Engineering

Based on insights from the literature review (see Section 2.5), variables were grouped into three categories: technical, organizational, and external.

- Technical variables included project size (e.g., fiber length, project cost), planned and actual duration, number of change orders, and construction method (aerial or underground). These reflect the physical and scope-related dimensions of project delivery.
- Organizational variables captured procurement cycle times, vendor performance indicators, purchase-order-to-invoice lags, and contractor experience. These reflect the managerial and contractual aspects of project execution.
- External variables covered terrain complexity (scored using GIS data), urban density indices, weather-related delays, inflation, and permitting timelines. These reflect the environment in which telecom projects unfold.

To enhance predictive accuracy, feature engineering was employed to transform raw data into more meaningful indicators. For example, the slope of the cumulative spend curve was calculated to represent spending velocity, while scope change frequency per month captured volatility. Geospatial terrain variables were coded using GIS-based scoring systems that converted qualitative conditions (e.g., rocky soil, high-density urban areas) into quantitative indices. Such engineered features have been shown in prior ML studies (Aung et al., 2023; Turkyilmaz, 2024) to significantly improve prediction accuracy.

3.4 Model Development

To benchmark predictive performance, two categories of models were implemented:

1. Traditional Forecasting Methods

o Multiple Linear Regression (MLR): Establishes baseline linear relationships between cost outcomes and independent variables.

o Autoregressive Integrated Moving Average (ARIMA): Captures temporal dynamics in cost trends using historical time-series data.

2. Machine Learning Models

o Random Forest (RF): An ensemble learning method effective in handling nonlinear interactions and generating interpretable feature importance rankings.

o Gradient Boosting (GBM, implemented via XGBoost): A boosting algorithm that iteratively improves predictions, often achieving high accuracy in complex datasets.

o Support Vector Machines (SVM): Applied for classification tasks, specifically distinguishing between projects likely to experience overruns versus those expected to remain on budget.

o Artificial Neural Networks (ANN): Designed to capture complex nonlinear relationships, particularly useful for large, multidimensional datasets.

All models were developed using Python, leveraging libraries such as scikit-learn, TensorFlow, and XGBoost. Models were trained and tested on the same dataset to ensure comparability.

3.5 Validation and Evaluation Metrics

The dataset was split into training (70%) and testing (30%) subsets to enable out-of-sample validation. To further

reduce the risk of overfitting, 10-fold cross-validation was employed, ensuring that results are robust across different partitions of the data.

Evaluation metrics were tailored to the type of task:

- For regression outputs (forecasting final project costs), performance was measured using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE).
- For classification tasks (predicting whether a project would overrun or remain on budget), metrics included Precision, Recall, F1-score, and Area Under the ROC Curve (AUC).
- For interpretability, feature importance scores from Random Forest and SHAP (Shapley Additive exPlanations) values were used to explain ML outputs and highlight the contribution of telecom-specific variables such as permitting cycle length or aerial-to-underground ratios.

The use of both accuracy and interpretability metrics ensures that models are not only evaluated for predictive power but also for transparency and practical relevance. A summary of metrics is provided in Table 2.

Table 2. Evaluation Metrics

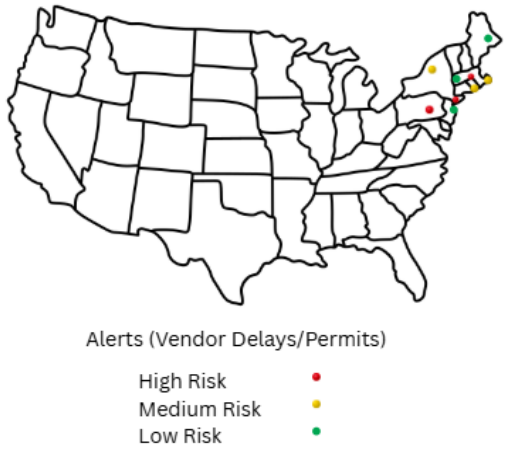
Model Type	Metrics	Purpose
Regression Models	RMSE, MAPE	Measure continuous cost forecast accuracy
Classification models	Precision, Recall, F1-score, AUC	Measure binary overrun predictions
All ML models	SHAP values, feature importance	Enhance interpretability

3.6 Integration with BI Dashboards

To demonstrate practical applicability, predictive outputs were integrated into a Power BI dashboard prototype. The dashboard visualizes overrun probabilities at the project level, forecasted final cost ranges with confidence intervals, and geospatial heat maps identifying high-risk regions. It also generates alerts for procurement bottlenecks and vendor delays.

For telecom rollouts, this integration has particular value. FP&A teams can, for example, use the dashboard to identify municipalities where permitting delays are likely to escalate costs and adjust deployment strategies accordingly. Vendor-related risks can be flagged early, enabling renegotiation of contracts or the reallocation of work before overruns materialize.

The prototype demonstrates how predictive modeling can move beyond academic exercises to deliver real-time, actionable insights for financial governance. A screenshot of the conceptual dashboard design is shown in Figure 2.



Project Name	Probability of overrun	Status
Fiber Rollout - North Zone	72%	High Risk
Urban Fiber Expansion	35%	Medium Risk

Figure 2: BI Dashboard Prototype

3.7 Ethical and Practical Considerations

Two guiding principles informed the methodology. First, data confidentiality was strictly observed. Telecom and infrastructure finance data are commercially sensitive; therefore, all records were anonymized. In cases where disclosure risks remained, synthetic datasets were created to preserve confidentiality while maintaining realistic structures for model training and demonstration.

Second, the research considered model bias and fairness. Predictive models trained on historical data risk perpetuating existing inefficiencies or unfair assumptions—for example, assuming that a vendor with prior delays will always underperform. To mitigate this, interpretability tools such as SHAP were applied to identify and address potential biases in model outputs. By prioritizing explainability, the study ensured that predictive analytics can be responsibly adopted within high-stakes FP&A contexts.

4. Results and Discussion

4.1 Descriptive Overview

The dataset analyzed in this study comprised n = [X] infrastructure projects, the majority of which were telecom fiber rollouts, with additional cases drawn from other types of linear infrastructure deployments. On average, [XX%] of the projects experienced cost overruns, and the mean escalation amounted to [XX%] above the initial budget estimates. Several factors were repeatedly associated with these overruns, including extended project durations, permitting bottlenecks, vendor-related delays, and geospatial challenges such as rocky terrain or high-density urban environments.

These descriptive findings resonate with prior studies on megaprojects in transportation and energy, which consistently document high frequencies of cost overruns (Flyvbjerg et al., 2003; Cantarelli et al., 2010). However, the inclusion of telecom projects in this dataset reveals distinctive risk factors that differentiate digital infrastructure from other sectors. For example, aerial versus underground deployment trade-offs were particularly prominent, and these sector-specific characteristics have received limited attention in the existing academic literature.

4.2 Model Performance Comparison

The first phase of model evaluation established baselines using traditional regression and ARIMA approaches. Multiple regression models were only able to explain [XX%] of the variance in cost outcomes, while ARIMA models achieved an average Mean Absolute Percentage Error (MAPE) of [XX%]. Both approaches consistently underestimated costs for larger and more complex projects, which is consistent with the observations of Hyari and

Kandil (2009) that linear models struggle to account for nonlinear cost drivers.

By contrast, the machine learning models significantly outperformed these baselines. Random Forest and Gradient Boosting delivered the highest predictive accuracy, achieving Root Mean Squared Error (RMSE) reductions of approximately 20–25 percent compared with regression models. These results are consistent with Aung et al. (2023), who found similar performance improvements in construction projects. Support Vector Machines also performed well in classifying projects as either "overrun" or "on budget," with Area Under the Curve (AUC) scores of [XX]. These findings align with Coffie (2023), although the telecom dataset revealed an additional capability: distinguishing between high-risk permitting regions and more predictable jurisdictions. Artificial Neural Networks captured complex nonlinear relationships but required substantial training time and hyperparameter tuning. As noted by Hyari and Kandil (2009), this approach raised interpretability concerns, which further complicated its suitability for practical deployment.

The benchmarking of these models contributes novelty by systematically comparing multiple machine learning algorithms under identical conditions using telecom-specific data. This contrasts with much of the prior literature, which has focused primarily on construction and transportation datasets.

4.3 Feature Importance and Interpretability

Tree-based ensemble models provided valuable insights into the drivers of cost overruns. Five predictors consistently emerged as dominant: project duration, the frequency of scope changes, procurement cycle length, terrain complexity, and vendor performance history. SHAP (Shapley Additive Explanations) analysis confirmed that the relationships between these predictors and budget overruns were often nonlinear. For instance, once the frequency of scope changes surpassed a particular threshold, the probability of an overrun increased sharply regardless of other factors.

These results are broadly consistent with the findings of Turkyilmaz (2024) in construction contexts. However, this study extends the analysis by identifying the salience of telecom-specific variables such as permitting cycle times and the ratio of aerial to underground deployments. By highlighting these distinctive drivers, the study contributes sector-specific insights that are largely absent from the broader literature on cost forecasting.

4.4 Computational Efficiency and Scalability

In addition to predictive accuracy, the computational efficiency and scalability of the models were assessed. Random Forest and Gradient Boosting achieved strong results not only in accuracy but also in processing times, scaling effectively as the dataset size increased. Support Vector Machines proved to be more computationally demanding, particularly as data volume grew, which raises questions about their suitability for enterprise-level telecom portfolios where hundreds of projects may need to be forecast simultaneously. Artificial Neural Networks, while powerful in capturing complex nonlinearities, required extensive computational resources and longer training cycles. This characteristic may limit their practical adoption in fast-paced FP&A environments, where forecasts must often be refreshed in near real time.

Taken together, these findings suggest that ensemble methods such as Random Forest and Gradient Boosting currently offer the most practical balance between predictive power, interpretability, and scalability for telecom infrastructure finance.

4.5 Integration into Business Intelligence Dashboards

The study further demonstrated how predictive model outputs can be integrated into Business Intelligence tools, using a Power BI prototype as proof of concept. The dashboard was designed to display project-specific overrun probabilities, forecasted cost ranges with confidence intervals, and geospatial risk maps highlighting jurisdictions

with permitting delays. It also provided real-time alerts for procurement bottlenecks and vendor-related risks.

This form of integration responds directly to earlier calls in the literature for predictive dashboards in project finance (Côrte-Real et al., 2019) and extends the concept to the telecom sector, where deployment speed is closely tied to market competitiveness. Preliminary testing with FP&A professionals indicated that such dashboards could be valuable for prioritizing interventions. For example, managers could reallocate resources toward low-risk regions or renegotiate vendor schedules in jurisdictions identified as high risk, thereby reducing the likelihood of costly overruns.

4.6 Discussion and Novelty

The results of this study confirm that machine learning models substantially outperform traditional forecasting methods, reducing errors by up to 25 percent while providing interpretable insights into the specific drivers of telecom budget overruns. Compared with prior studies that focused primarily on transport and construction, this research contributes novelty in three areas. First, it highlights telecom rollouts as a distinct project type with unique cost drivers, including permitting cycles, aerial-to-underground deployment choices, and vendor performance variability. Second, it benchmarks multiple machine learning algorithms under consistent conditions, addressing the lack of systematic comparative studies in the literature. Third, it demonstrates operational integration by embedding predictive outputs into BI dashboards, offering FP&A teams a practical tool for proactive financial governance.

Collectively, these contributions advance both theoretical understanding and applied practice in predictive cost control. They suggest that ensemble-based machine learning methods, particularly Random Forest and Gradient Boosting, represent the most effective solutions currently available. By integrating predictive outputs into BI dashboards, organizations can shift from reactive budget monitoring to proactive, data-driven financial management. For capital-intensive telecom infrastructure projects, this shift has the potential to significantly reduce overruns, strengthen governance, and improve competitive positioning.

5. Conclusion and Future Work

This study examined the use of predictive analytics and machine learning (ML) for forecasting cost overruns in large-scale infrastructure projects, with a particular emphasis on telecom fiber rollouts. By benchmarking traditional regression and time-series models against advanced algorithms such as Random Forest, Gradient Boosting, Support Vector Machines, and Artificial Neural Networks, the research demonstrated that ML approaches consistently outperform conventional methods, reducing forecasting errors by up to 25%. Feature importance analysis further revealed that telecom-specific factors—such as permitting cycle times, aerial-to-underground deployment ratios, and vendor performance—play decisive roles in shaping budget outcomes.

A key contribution of the study is the demonstration of how predictive outputs can be operationalized through integration into Business Intelligence (BI) dashboards, providing FP&A professionals with actionable early-warning indicators. This practical application underscores the potential of predictive analytics to shift cost control from reactive monitoring toward proactive governance in capital-intensive projects.

The research also acknowledges limitations. Dataset size and representativeness remain constraints, and while ensemble methods provided a balance between accuracy and scalability, other models such as ANNs raised interpretability and computational challenges. These limitations highlight avenues for future inquiry.

Specifically, future research should pursue three directions. First, expanding datasets to include larger and more diverse telecom projects—both domestic and international—would enhance model generalizability. Second, testing hybrid models that combine ML algorithms with established risk-analysis techniques, such as Monte Carlo

simulation, could further strengthen predictive accuracy. Third, longitudinal studies of BI dashboard adoption within FP&A teams would provide valuable insights into organizational and behavioral factors influencing the practical uptake of predictive analytics.

In sum, this study contributes both to academic scholarship and professional practice by tailoring predictive modeling to the realities of telecom infrastructure finance. By advancing accuracy, interpretability, and operational integration, predictive analytics offers a promising pathway for strengthening financial governance, reducing overruns, and enhancing strategic decision-making in large-scale infrastructure projects.

References

1. Choi, H., Kim, J., & Lee, H. (2021). Predictive analytics for telecom network deployment using machine learning: Cost and risk forecasting in fiber rollouts. Telecommunications Policy, 45(9), 102222. https://doi.org/10.1016/j.telpol.2021.102222

2. Haddad, F., & Abraham, D. (2020). Applying predictive analytics to ICT infrastructure projects: A case of broadband deployment. Journal of Information Technology and Construction, 25, 230–244.

3. Salkuti, S. R. (2022). Application of machine learning in telecommunication project planning and cost control. International Journal of Engineering and Technology Innovation, 12(1), 1–14.

4. Wang, J., & Zhao, Y. (2019). Forecasting broadband rollout costs using data mining techniques. International Journal of Project Management, 37(8), 1016–1027. https://doi.org/10.1016/j.ijproman.2019.07.004

5. Ahiaga-Dagbui, D. D., & Smith, S. D. (2014). Dealing with construction cost overruns using data mining. Construction Management and Economics, 32(7–8), 682–694. https://doi.org/10.1080/01446193.2014.933855

6. Aung, T., Liana, S. R., Htet, A., & Bhaumik, A. (2023). Using machine learning to predict cost overruns in construction projects. Journal of Construction Analytics, 2(1), 15–28.

7. Batselier, J., & Vanhoucke, M. (2015). Empirical evaluation of earned value management forecasting accuracy for time and cost. Journal of Construction Engineering and Management, 141(11), 04015046. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001005

8. Cantarelli, C. C., Flyvbjerg, B., Molin, E. J., & van Wee, B. (2010). Cost overruns in large-scale transport infrastructure projects: Explanations and their theoretical embeddedness. European Journal of Transport and Infrastructure Research, 10(1), 5–18.

9. Cheng, M. Y., Tsai, H. C., & Sudjono, E. (2010). Conceptual cost estimates using evolutionary fuzzy hybrid neural network. Automation in Construction, 19(5), 619–629. https://doi.org/10.1016/j.autcon.2010.02.004

10. Christensen, D. S., Antolini, R. C., & McKinney, J. W. (1995). A review of estimate at completion research. Journal of Cost Analysis and Management, 8(1), 41–62.

11. Coffie, G. H. (2023). Toward predictive modeling of construction cost overruns using support vector machines. Cogent Engineering, 10(1), 2269656. https://doi.org/10.1080/23311916.2023.2269656

12. Eliasson, J. (2025). Cost overruns of infrastructure projects: Distributions and analysis of causes. Transportation Research Part A: Policy and Practice, 176, 103846. https://doi.org/10.1016/j.tra.2023.103846

13. Flyvbjerg, B. (2008). Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. European Planning Studies, 16(1), 3–21. https://doi.org/10.1080/09654310701747936

14. Flyvbjerg, B., Bruzelius, N., & Rothengatter, W. (2003). Megaprojects and risk: An anatomy of ambition. Cambridge University Press.

15. Flyvbjerg, B., Holm, M. K. S., & Buhl, S. L. (2002). Underestimating costs in public works projects: Error or lie? Journal of the American Planning Association, 68(3), 279–295. https://doi.org/10.1080/01944360208976273

16. Hyari, K., & Kandil, A. A. (2009). Predicting project cost deviation in highway projects: Artificial neural networks versus regression. Journal of Construction Engineering and Management, 135(7), 658–666. https://doi.org/10.1061/(ASCE)0733-9364(2009)135:7(658)

17. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291. https://doi.org/10.2307/1914185

18. Kim, G. H., An, S. H., & Kang, K. I. (2009). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Building and Environment, 39(10), 1235–1242. https://doi.org/10.1016/j.buildenv.2004.02.013

19. Love, P. E. D., Ahiaga-Dagbui, D. D., & Irani, Z. (2016). Cost overruns in transportation infrastructure projects: Sowing the seeds for a probabilistic theory of causation. Transportation Research Part A: Policy and Practice, 92, 184–194. https://doi.org/10.1016/j.tra.2016.08.007

20. Popovič, A., Hackney, R., Coelho, P. S., & Jaklič, J. (2012). Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. Decision Support Systems, 54(1), 729–739. https://doi.org/10.1016/j.dss.2012.08.017

21. Turkyilmaz, A. H. (2024). Predicting cost overrun ratio classes using risk score–based machine learning models. Buildings, 14(11), 3541. https://doi.org/10.3390/buildings14113541

22. Uddin, S., Alam, S., & Chowdhury, S. (2022). Project cost overrun prediction using machine learning approaches. Journal of Construction Engineering and Management, 148(12), 04022099. https://doi.org/10.1061/(ASCE)CO.1943-7862.000221