Leveraging Artificial Intelligence to Mitigate Delays and Cost Overruns in Public Infrastructure Construction Projects

Data Science Capstone Project

Souleymane DOUMBIA

Abstract

Construction projects in the public infrastructure sector are frequently plagued by delays and cost overruns, leading to wasted resources, unmet community needs, and diminished trust in public institutions. These persistent challenges often stem from complex, interrelated factors such as regulatory delays, labor constraints, environmental disruptions, and insufficient early risk detection. In response, this study proposes a data-driven framework leveraging Artificial Intelligence (AI) and Machine Learning (ML) techniques to proactively classify and forecast project risks.

Grounded in systems thinking and complexity theory, the framework applies supervised learning models (Random Forest, Decision Tree, and XGBoost) to three prediction tasks: cost performance classification, schedule performance classification, and a combined project risk profile. The dataset integrates historical NYC public infrastructure records with contextual information, including borough-level weather and labor market statistics.

Results demonstrate that the models, particularly XGBoost and Random Forest, achieve high accuracy and balanced performance across multiple classes, successfully predicting cost deviations, schedule adherence, and risk profiles. Feature importance analysis reveals that variables such as project budget, spend-to-date percentage, weather risk category, and labor conditions play a key role in performance outcomes. These findings highlight the feasibility of using AI to develop early-warning tools that inform project planning and oversight.

By simulating separate and composite performance forecasts, the framework offers a scalable and interpretable approach for risk management and decision support in capital project delivery. This study contributes to bridging the gap between predictive modeling and public infrastructure

governance by providing data-informed insights to reduce inefficiencies and enhance accountability.

Introduction

Public infrastructure projects are critical to societal development, providing essential services such as transportation, water supply, education, and healthcare. Yet, despite their importance, these projects are frequently marred by delays and cost overruns. The consequences are far-reaching: extended inconvenience to the public, budget reallocations, diminished trust in public institutions, and underutilization of capital. A 2025 audit report by the New York State Comptroller's Office found that among 5,128 analyzed public infrastructure projects in New York City, 64 percent were delayed, which is defined as exceeding their planned completion date by at least three months, and nearly half were delayed by more than three years. Furthermore, over 50 percent of the projects were over budget, with some exceeding their original budget by 20 percent or more, highlighting systemic issues in capital project delivery and monitoring.

Numerous factors contribute to these inefficiencies, including poor project planning, inaccurate cost estimation, unpredictable regulatory processes, supply chain disruptions, and a lack of timely risk detection. Traditional project management approaches often rely on deterministic methods and expert judgment, which, while valuable, are limited in their ability to manage the inherent complexity and uncertainty of large-scale public works.

Emerging technologies, particularly Artificial Intelligence (AI) and Machine Learning (ML), offer a promising alternative. These tools can enhance decision-making by extracting patterns from historical data, forecasting risk trajectories, and recommending preventive actions. However, their adoption in public infrastructure construction remains limited due to concerns about transparency, interpretability, data availability, and integration into existing workflows.

This paper responds to this gap by proposing a data-driven, interpretable, and scalable framework that leverages AI and ML to identify, classify, and mitigate risks related to delays and cost overruns in public infrastructure projects. Rooted in complexity theory and systems thinking, the framework applies supervised classification models, specifically Random Forest, Decision Tree, and XGBoost to predict three dimensions of project performance: cost deviation, schedule adherence, and combined delivery risk. These predictions simulate early-warning tools that could support proactive interventions in the capital project lifecycle.

Despite growing interest, few studies have operationalized an AI-based framework tailored to the multifaceted needs of public infrastructure oversight. Most prior applications of AI in construction have focused narrowly on cost estimation, schedule monitoring, or labor forecasting, and rarely integrate environmental and contextual variables at scale.

By applying a structured machine learning approach to historical project data enriched with borough-level weather trends and labor market indicators, this study seeks to modernize infrastructure planning and delivery. The aim is to proactively identify risk-prone project profiles, forecast performance deviations, and provide interpretable insights that empower agencies to improve project delivery outcomes amid rising societal demands, fiscal constraints, and environmental disruptions.

I. Literature Review

1.1 Causes of Delays and Cost Overruns in Public Infrastructure

The challenge of delays and cost overruns in public infrastructure construction has been a longstanding focus in both academic research and government audits. Studies across disciplines have highlighted that construction projects, particularly those managed by public agencies, often suffer from inefficiencies due to fragmented oversight, complex stakeholder environments, and

poor upfront planning (Williams, 2016; Arantes et al., 2015). Moreover, projects of higher complexity tend to face greater risk of overruns, emphasizing the need for improved predictive and diagnostic tools (Sanni-Anibire et al., 2020).

1.2 Applications of AI and ML in Construction

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as transformative tools in construction management, enabling data-driven decision-making across planning, execution, and monitoring stages. Applications of AI in construction have included cost estimation using artificial neural networks and regression trees (Cheng et al., 2021), schedule risk prediction based on probabilistic modeling and machine learning classifiers (Viles et al., 2019), and labor productivity forecasting using backpropagation neural networks (Liu & Skibniewski, 2019). These models often rely on historical project data and input variables such as scope, location, regulatory timelines, labor availability, and environmental factors. AI-based decision support systems have shown promise in anticipating disruptions and optimizing resource allocation (Sener & Ozorhon, 2020).

1.3 Gaps in Existing Approaches

Yet, much of the existing literature focuses on isolated components of construction projects (only labor, or only cost), and often within the private sector context. Few studies operationalize integrated frameworks that holistically assess delays and cost overruns in public-sector infrastructure using both supervised and unsupervised learning. For example, Arantes et al. (2015) analyze project complexity and its relationship to performance outcomes in public construction projects, but do not incorporate temporal factors or project governance. Similarly,

Sanni-Anibire et al. (2020) explore risk assessment methods but stop short of applying predictive algorithms.

1.4 Environmental Factors and Weather-Related Delays

Shields and Nunemaker (2020) developed a process-based cost modeling tool (ORBIT) for offshore wind power plants that integrates time-series weather simulations to analyze installation delays. Their use of discrete-event simulation (DES) to account for hourly metocean constraints, such as wind speed and wave height which is demonstrating the importance of modeling environmental conditions in construction planning. While this work focused on offshore energy infrastructure, its methodology is applicable and instructive for urban infrastructure contexts, particularly in modeling the weather-induced uncertainties that contribute to project delays.

1.5 Toward a Comprehensive Predictive Framework

A 2025 audit by the New York State Comptroller's Office revealed that 64 percent of public infrastructure projects in New York City were delayed and over 50 percent exceeded their budgets, often without clearly documented causes. This underscores the need for proactive, data-driven methods to anticipate and address these risks early in the project lifecycle.

This study builds upon prior research by applying supervised machine learning techniques to forecast cost and schedule outcomes based on historical project data enriched with contextual variables such as weather and labor market conditions. Unlike previous studies that focused on isolated components or proposed clustering-based segmentation, this framework evaluates multiple dimensions of project performance (cost, time, and their intersection) using interpretable and validated classification models. By addressing the limitations of prior fragmented studies and

aligning with real-world audit insights, this research contributes a comprehensive, scalable methodology for public infrastructure risk management.

II. Purpose and Research Questions

The primary purpose of this research is to develop a data-driven framework that leverages artificial intelligence (AI) and machine learning (ML) techniques to identify, classify, and mitigate the risks associated with delays and cost overruns in public infrastructure projects. The framework aims to improve project planning and oversight by providing predictive insights into distinct project performance outcomes (cost deviation, schedule adherence, and combined risk profiles) enabling proactive interventions throughout the project lifecycle.

Grounded in systems thinking and complexity theory, the study applies supervised classification models and simulation-based forecasting to evaluate how contextual, environmental, and operational factors influence project outcomes. The analysis uses enriched real-world datasets, including historical project records, budget and schedule information, labor conditions, and borough-level weather trends, to train and validate predictive models.

This research is guided by the following core questions:

1. **How accurately can machine learning models predict cost performance and schedule adherence outcomes for public infrastructure projects?**
2. **Which variables contribute most significantly to project performance classifications across cost, time, and combined risk dimensions?**
3. **How can AI-based simulations inform early warnings and support strategic decision-making throughout the project delivery lifecycle?**

By addressing these questions, the study contributes a modular and interpretable risk classification framework that supports data-informed capital planning and more resilient public infrastructure governance.

III. Methodology

This study employs a supervised machine learning approach to develop a predictive framework for identifying and mitigating risks associated with delays and cost overruns in public infrastructure projects. Grounded in systems thinking and complexity theory, the methodology enables a holistic analysis of interdependent variables such as cost, schedule, and environmental conditions.

3.1 Data Sources and Integration

The primary consists of historical public infrastructure project records from New York City, capturing planned and actual timelines, budget figures, project types, and location-specific metadata. To contextualize project performance, this data was enriched with borough-level labor market indicators (e.g., average construction employment and unemployment rates) and weather-related variables sourced from public climate and labor datasets. These were merged using borough and project start year as join keys. Extensive preprocessing included date normalization, missing value handling, feature engineering, and the imputation of unknown project themes based on borough-level mode frequencies.

To facilitate predictive modeling, the dataset was also restructured into a wider format, with one row per project (identified by fms_id). This was achieved by generating and merging several pivot tables:

- A main project pivot table using fms_id and maximum spend_to_date.

- Delay and cost category pivots aggregating delay_days and cost_diff respectively.
- A phase-based spend pivot reflecting cumulative latest_spend by project phase. Missing cells were filled with zeros to indicate absence of data (i.e., the event did not occur), and to ensure completeness. All outputs were merged into a unified wide-format dataset preserving the interpretability of measurement units.

3.2 Feature Engineering and Variable Categorization

Key variables were organized into:

- **Temporal and Financial Metrics**: durations, planned vs. actual time gaps, budget details
- **Contextual Factors**: borough, agency, fiscal year, weather and labor statistics
- **Derived Classifications**:
    - **Cost Performance Classification (cost_class)**: Indicates whether a project phase was Over Budget (final cost > 15% above planned), Under Budget (more than 15% below), or On Budget (within ±15% of the planned budget).
    - **Schedule Performance Classification (delay_class)**: Indicates whether a project phase was Delayed (completed >30 days late), Early (completed >30 days early), or On Time (within ±30 days of planned end date).
    - **Project Profile Classification (project_profile)**: A combined label reflecting both cost and schedule performance, with nine unique profiles such as High Risk (Delayed + Over Budget) and Exceptional Performance (Early + Under Budget).
    - **Project Theme (project_theme)**: A categorical variable reflecting the general type or scope of project work, such as Street Improvement, Facility Upgrade, or Public Space Renovation. Unknown themes were imputed using the most frequent value within each borough.

- An additional derived variable, weather_risk_category, was created by applying k-means clustering (k=3) to standardized weather variables (avg_cyclone_freq, avg_temp_anomaly, avg_max_wind_speed), resulting in categorical labels of Low, Moderate, or High risk weather.

Exploratory Data Analysis (EDA) was conducted to understand the distribution and variability of key project variables. Histograms and bar plots (see Figure 2 and 3 in appendix) were generated to explore cost differences, cost class distribution, and project profiles across phases. Boxplots (see Figures 1) highlighted the spread of delay durations across delay categories.

Feature transformations included one-hot encoding for categorical variables and standardization (using z-score normalization) of selected numerical inputs. All derived labels were created using reproducible rules documented in the data wrangling scripts (in Appendix Section D).

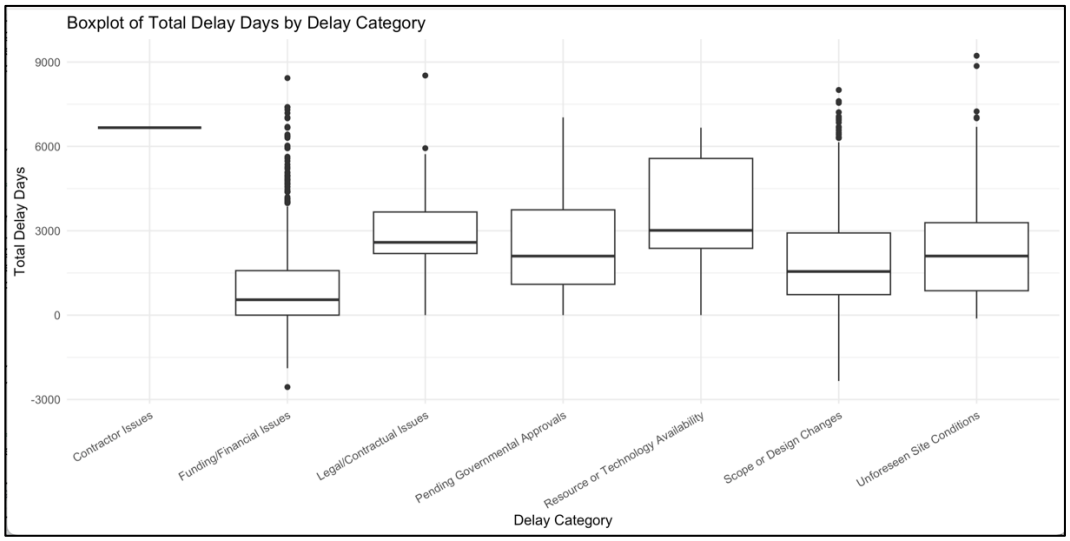


Figure 1: Variability in Schedule across Delay Category

3.3 Predictive Modeling Framework

To simulate project performance forecasting, three separate classification tasks were designed, aligned with the above labels. Each task was modeled independently using:

- **Random Forest** (for robustness and interpretability)
- **Decision Tree** (for transparency and simplicity)
- **XGBoost** (for high accuracy, especially in multi-class prediction)

Models were validated using 5-fold cross-validation and hyperparameter tuning. Performance metrics included accuracy, Kappa, and class-level balanced accuracy. Feature importance outputs from Random Forest and XGBoost were used to assess predictor influence. Decision trees were visualized using rpart.plot for model interpretability.

3.4 Simulation of Project Risk Dimensions

The classification tasks were conceptualized as predictive simulations: one for cost outcomes, one for schedule adherence, and one integrated profile combining both. This structure enabled targeted risk flagging while supporting unified project profiling.

3.5 Model Evaluation and Interpretation

Model performance was evaluated using standard classification metrics including overall accuracy, Cohen's Kappa, and class-level balanced accuracy. Each model was validated using 5-fold cross-validation and tested on a hold-out dataset (20% of the sample) to ensure generalizability. Feature importance scores from Random Forest and XGBoost were analyzed to identify the most influential predictors for each classification task. Decision tree structures were visualized allowing for easier interpretation of decision paths and helping validate how the models classified projects based on key predictive features.

This methodology supports a modular and scalable AI-based risk classification framework that may be adapted to public capital project oversight, resource optimization, and predictive risk flagging during project planning or monitoring stages.

IV. Results

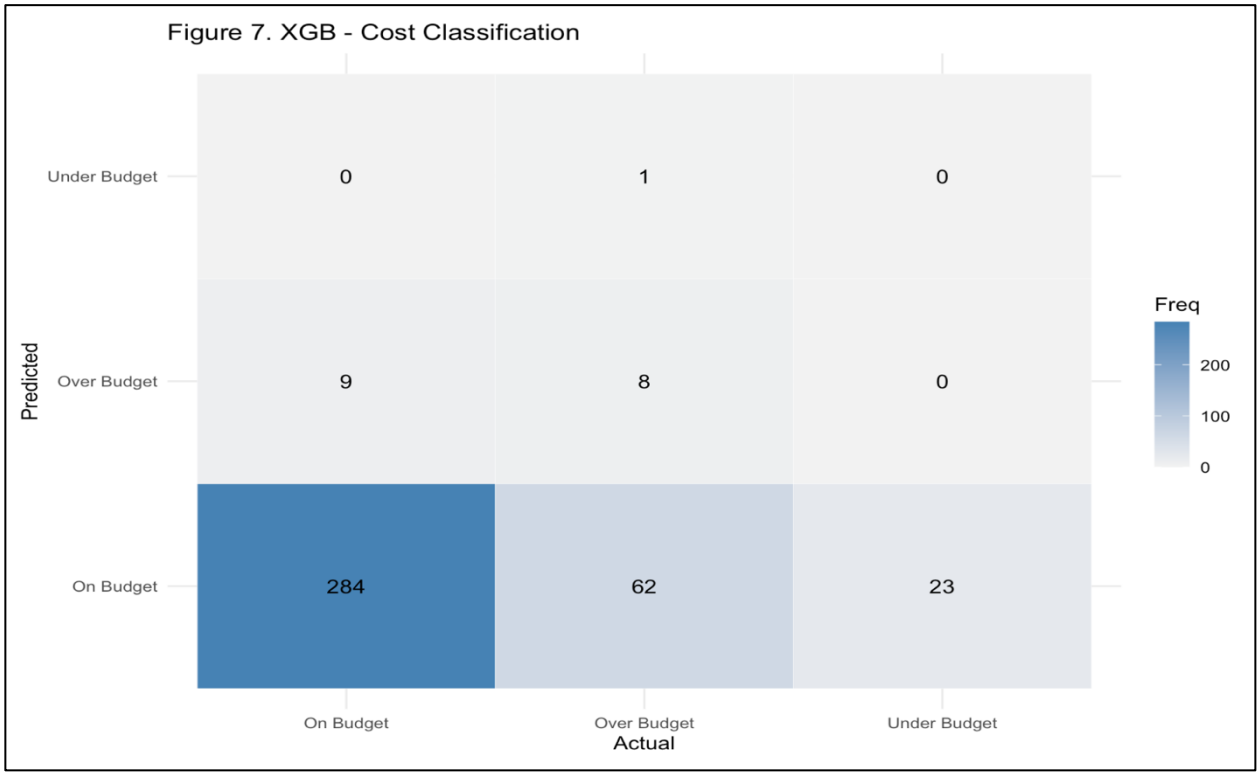
This study employed three supervised machine learning models, Random Forest (RF), Decision Tree (DT), and XGBoost (XGB), to predict project performance across three dimensions: cost_class, delay_class, and project_profile. These tasks correspond to cost performance, schedule adherence, and a combined risk classification, respectively. Each target variable was engineered from project schedule and financial metrics: cost_class was derived from budget deviation thresholds, delay_class from schedule adherence, and project_profile as a combined outcome reflecting both cost and time performance. These categories simulate core dimensions of project delivery risk, allowing the models to replicate how predictive analytics can inform intervention strategies. The three models were selected for their complementarity: Random Forest for balanced accuracy and robustness, Decision Tree for interpretability, and XGBoost for high predictive performance in multi-class settings. Below is a summary of performance results for each modeling task.

4.1. Cost Classification (cost_class: Over Budget, Under Budget, On Budget)

Among the three algorithms tested, Random Forest yielded an accuracy of **76.5%** and a Kappa score of **0.0606**. It performed exceptionally well in identifying "On Budget" projects, with a sensitivity of 0.9966. However, it struggled with minority classes, particularly "Over Budget" (Sensitivity = 0.0563) and "Under Budget" (Sensitivity = 0.0000), revealing class imbalance as a persistent challenge. (See Figure 4 for decision tree visualization in appendix.)

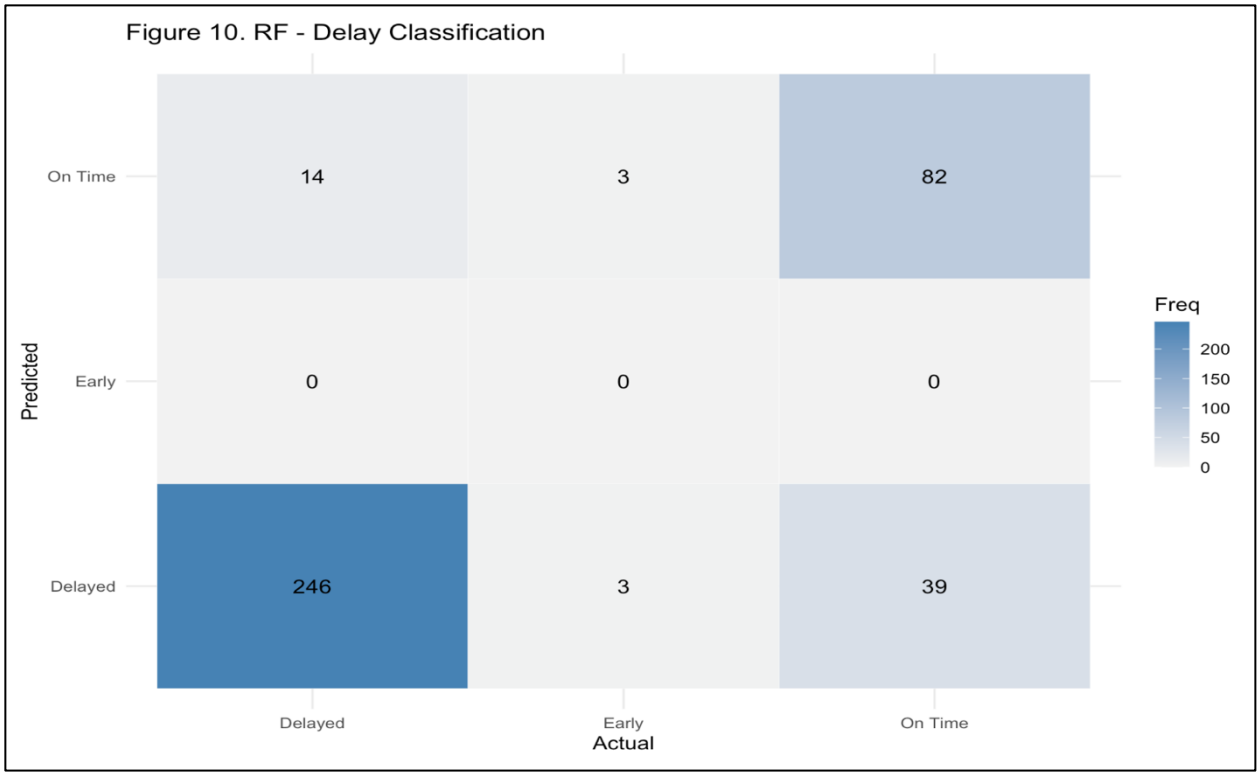
The Decision Tree model offered a similar overall accuracy of **77.0%** with a slightly higher Kappa score of **0.1277**, reflecting modest improvements in recognizing "Over Budget" cases. Still, no true positives were recorded for "Under Budget," limiting its effectiveness in capturing edge-case scenarios.

XGBoost delivered an overall accuracy of **75.5%** and a Kappa of **0.0905**, with improvements in specificity over sensitivity for less frequent outcomes. While not surpassing Random Forest in balanced performance, it held steady in classifying dominant project categories. (See Figure 7)



4.2. Delay Classification (delay_class: Delayed, Early, On Time)

For delay classification, Random Forest achieved an accuracy of **84.8%** and Kappa of **0.6371**. It performed strongly on the majority class, accurately flagging delayed projects (Sensitivity = 0.9462), and reasonably identifying "On Time" completions (Sensitivity = 0.6777), though "Early" completions were entirely missed (Sensitivity = 0.0000). (See Figure 10)



The Decision Tree model maintained respectable accuracy at **78.8%** and Kappa of **0.4710**, again showing competent performance for "Delayed" cases but minimal predictive power for other categories. (See Figure 5 in appendix)

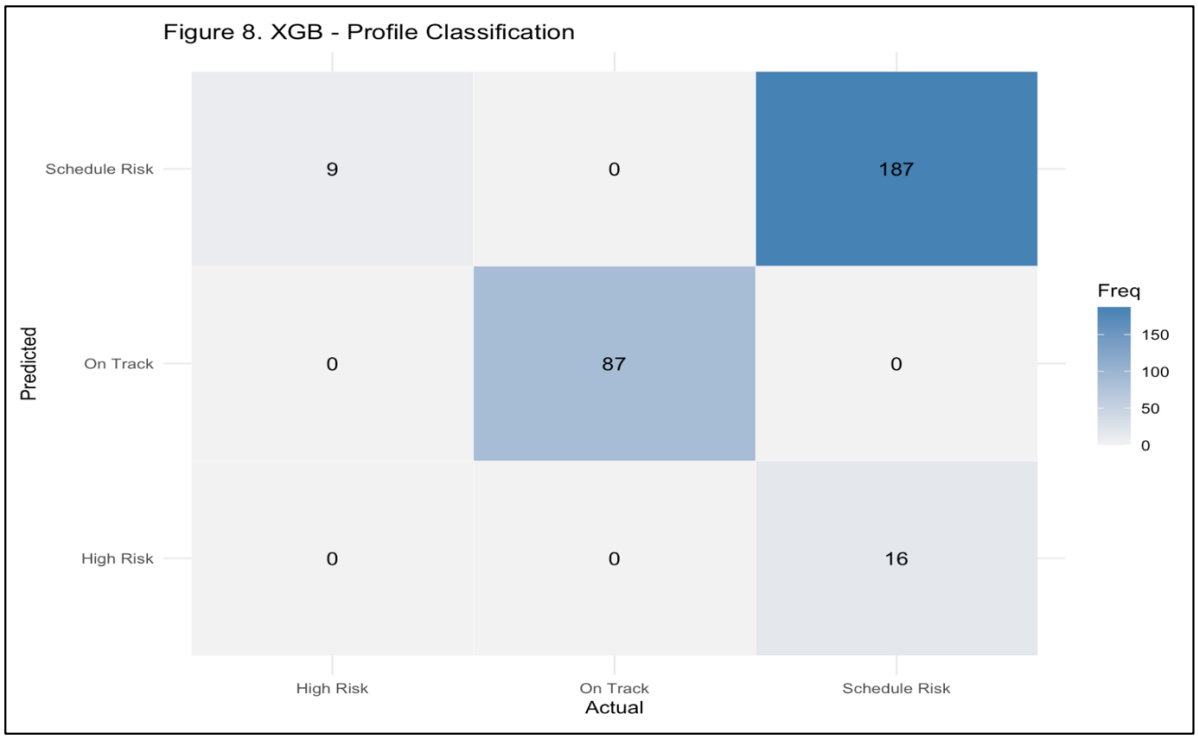
XGBoost led the delay classification task with an accuracy of **86.6%** and Kappa of **0.6908**. It balanced performance across major classes, attaining a balanced accuracy of 0.8453 for "Delayed" and 0.8445 for "On Time" predictions, confirming its robustness under class imbalance. (See Figure 9 in Appendix)

4.3. Project Profile Classification (project_profile: 9 multi-class labels)

In the most complex task, predicting nine possible project profile combinations, Random Forest scored an accuracy of **79.4%** and Kappa of **0.6569**. It handled dominant classes such as "Schedule Risk" and "On Track" with high reliability. (See Figure 6 in appendix for full tree structure.)

The Decision Tree produced **77.1%** accuracy and **Kappa = 0.6277**, slightly trailing Random Forest in overall balance but offering clarity through interpretable branching logic.

XGBoost outperformed both, reaching **85.9% accuracy** and a Kappa of **0.7801** (See Table 1 in Appendix). It demonstrated high sensitivity for multiple risk combinations, confirming its strength in managing high-dimensional classification tasks. (See Figure 8)



Across all three prediction tasks, XGBoost consistently outperformed other models in both overall accuracy and class-level balance, with Random Forest closely behind. Decision Trees provided helpful interpretability but had reduced performance, especially for underrepresented classes.

In terms of feature importance (See Table 2 in Appendix):

- **Cost-related predictors** such as initial_budget, spend_to_date_percent, and cost_cat_transportation (indicating investment in transportation-related projects) were dominant in predicting both cost_class and project_profile outcomes.
- **Schedule-related indicators** like delay_cat_funding_financial_issues (flagging delays due to funding problems), project_phase_procurement, and avg_labor_unemp_rate emerged as key drivers for delay_class predictions.
- **Combined predictors** such as weather_risk_category (clustered weather exposure), avg_construction_jobs, and project_theme (defining the general scope of the project) consistently influenced predictions across all three classification tasks.

These findings confirm that combining raw schedule and cost metrics with contextual environmental and categorical information enables robust and realistic risk classification. The ability to forecast project risk outcomes from these features demonstrates the value of structured data pipelines and model ensembles in capital project oversight.

V. Discussion

The predictive modeling results illustrate the viability of using supervised machine learning techniques to classify and forecast construction project performance across cost, schedule, and integrated dimensions. The consistent outperformance of XGBoost, particularly in multiclass settings like project profile prediction, reinforces its utility for high-dimensional, imbalanced construction datasets. Random Forest models delivered solid results, especially in delay prediction, while Decision Trees, despite lower overall accuracy, retained value for interpretability.

The variation in class-level metrics, especially the poor sensitivity for minority classes such as "Under Budget" and "Early", reaffirms the presence of class imbalance, a common issue in real-world public infrastructure data. XGBoost's high balanced accuracy on both delay and project profile tasks suggests its robustness under such imbalance, though further work with sampling strategies or cost-sensitive learning could strengthen performance on underrepresented categories.

A key contribution of this study is the structured decomposition of risk into three classification tasks: cost performance (cost_class), schedule adherence (delay_class), and an integrated profile (project_profile). This tripartite modeling approach enables stakeholders to pinpoint specific risk categories (e.g., projects likely to delay but remain on budget), making the system useful as a portfolio-level triage or early warning tool.

Model interpretability, while limited in XGBoost, was partially recovered through Decision Trees and feature importance scores from Random Forest and XGBoost. These revealed consistent drivers of project outcomes: cost-related features like initial_budget and spend_to_date_percent were pivotal for cost_class, while avg_labor_unemp_rate, delay_cat_funding_financial_issues, and weather_risk_category were more predictive of delay_class. For project_profile, performance reflected the interplay of cost and schedule variables, with context-rich attributes like project_theme and avg_construction_jobs further improving differentiation.

The simulation confirms that integrating categorical, environmental, and project-specific data enriches predictive modeling of infrastructure risk. Moreover, it points to a scalable architecture for AI-driven capital planning. Future research should explore integrating real-time project

updates via time series data, conducting causal inference to untangle risk drivers, and validating across external jurisdictions or cross-agency datasets.

Overall, this study demonstrates that machine learning, supported by thoughtful problem decomposition and feature design, can offer practical and interpretable insights for public infrastructure risk mitigation.

Conclusion

This study explored the application of artificial intelligence to mitigate cost overruns and schedule delays in public infrastructure construction projects in New York City. By structuring predictive modeling across three dimensions: cost performance, schedule adherence, and integrated risk classification, the study demonstrated how supervised machine learning can simulate proactive risk assessment scenarios. Random Forest and XGBoost models consistently delivered high accuracy and class-level balance, while Decision Trees offered transparent, interpretable decision paths.

The predictive tasks, classifying projects as over/under/on budget, delayed/on time/early, and assigning multi-class project profiles, serve as a proxy for real-world decision points in capital project management. The updated modeling results confirmed that XGBoost performed best across all tasks, particularly in capturing project profiles and delay classifications. Cost performance prediction remained more challenging due to imbalanced representation, especially for under-budget outcomes.

The modeling affirmed that key features such as planned budget, actual spend, labor market indicators, weather conditions, and categorical project characteristics are strong indicators of performance outcomes. By disaggregating risk dimensions and evaluating models across each, this approach offers a modular and scalable framework that could be adapted by public agencies

to support early detection of high-risk projects. This has implications for optimizing resource allocation, improving public trust, and reducing inefficiencies across infrastructure delivery.

The findings support the broader hypothesis that AI-driven predictive simulations can strengthen public construction oversight. Future implementations may benefit from real-time data feeds, expanded inter-agency datasets, cost-sensitive learning techniques, and incorporation of unsupervised learning to further refine risk detection.

Ultimately, the integration of machine learning into public project monitoring presents a meaningful opportunity to modernize infrastructure governance and elevate the effectiveness of capital investment planning in New York City.

References

Arantes, A., da Silva, P. F., & Ferreira, L. M. D. F. (2015). Delays in construction projects—Causes and impacts. In *6th IESM Conference*. Seville, Spain.

Dikmen, S. U., & Sonmez, M. (2011). An artificial neural networks model for the estimation of formwork labour. *Journal of Civil Engineering and Management*, 17(3), 340–347.

Durdyev, S., & Hosseini, M. R. (2018). Causes of delays on construction projects: A comprehensive list. *International Journal of Managing Projects in Business*, 11(2), 332–365.

Gunduz, M., & Almuajebh, M. (2020). Critical success factors for sustainable construction project management. *Sustainability*, 12(5), 1990.

Ramanathan, C., Narayanan, S. P., & Idrus, A. B. (2012). Construction delays causing risks on time and cost—A critical review. *Australasian Journal of Construction Economics and Building*, 12(1), 37–57.

Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2020). Causes of delay in the global construction industry: A meta-analytical review. *International Journal of Construction Management*. https://doi.org/10.1080/15623599.2020.1716132

Shields, M., & Nunemaker, J. (2020). Process-based balance of system cost modeling for offshore wind power plants in the United States. *Journal of Physics: Conference Series*, 1452(1), 012039. https://doi.org/10.1088/1742-6596/1452/1/012039

Waqar, A. (2024). Intelligent decision support systems in construction engineering: An artificial intelligence and machine learning approaches. *Expert Systems with Applications*, 249(Part A), 123503.

Williams, T. (2016). Identifying success factors in construction projects: A case study. *Project Management Journal*, 47(1), 97–112.

Viles, E., Rudeli, N. C., & Santilli, A. (2020). Causes of delay in construction projects: A quantitative analysis. *Engineering, Construction and Architectural Management*, 27(4), 917–935.

New York State Office of the State Comptroller. (2025). Capital Projects: Improving Timeliness and Cost Performance in New York City Public Infrastructure. Report 2025-001. https://www.osc.ny.gov/

Kor, M., Yitmen, I., & Alizadehsalehi, S. (2023). *An investigation for integration of deep learning and digital twins towards Construction 4.0*. Smart and Sustainable Built Environment, 12(3), 461–487. https://doi.org/10.1108/SASBE-08-2021-0148

Open Data NYC. (n.d.). NYC Capital Project Data. Retrieved from https://opendata.cityofnewyork.us/

U.S. Bureau of Labor Statistics. (n.d.). Local Area Unemployment Statistics (LAUS). Retrieved from https://www.bls.gov/lau/

Appendix

A. Illustration Visuals

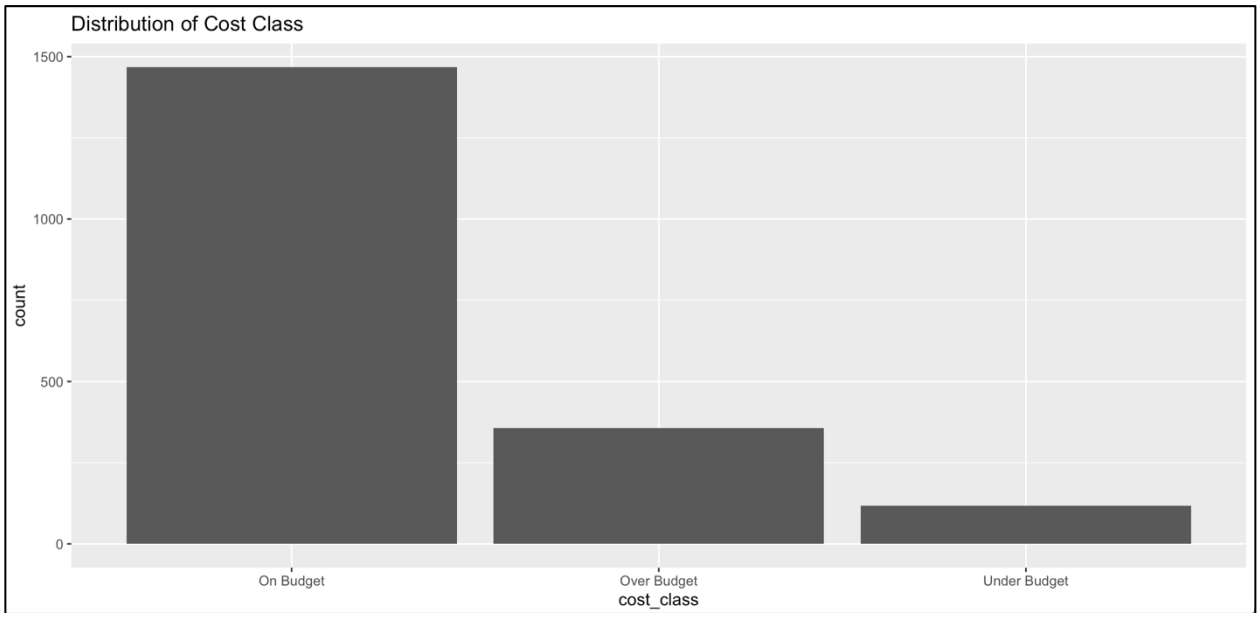


Figure 2: *Skewness and Concentration of Cost Overrun*

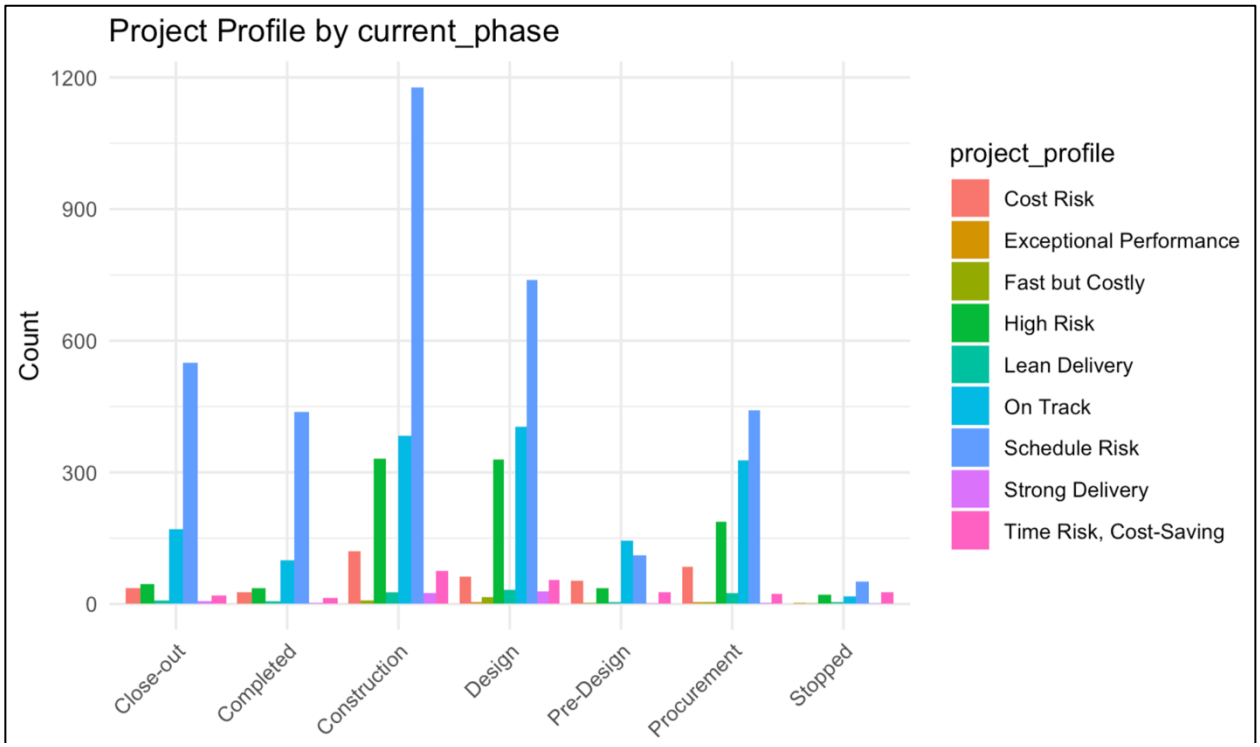


Figure 3: *Combined Risk Classifications across Phase*

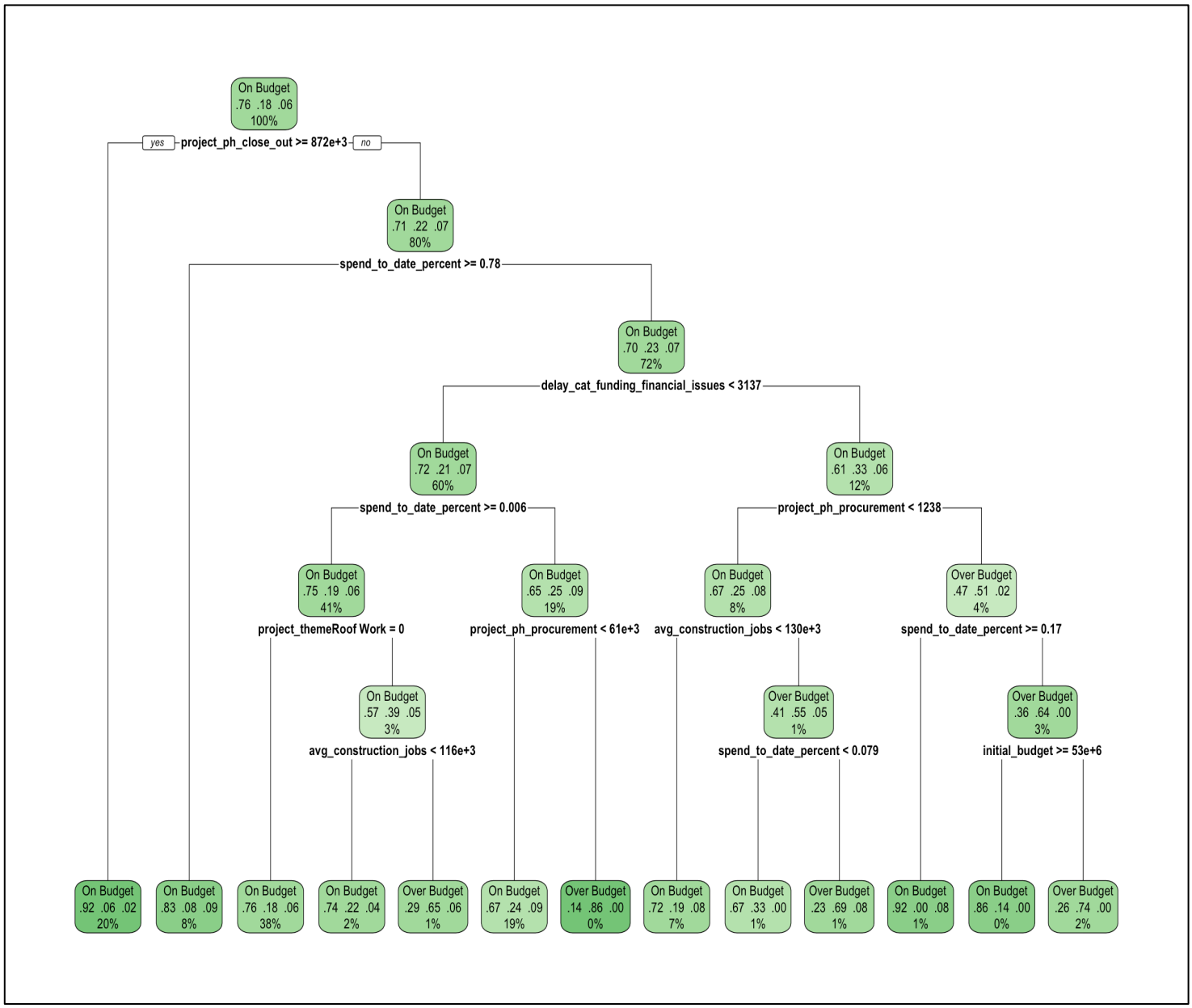


Figure 4: *Decision Tree Visualization for Cost Overrun Classification*

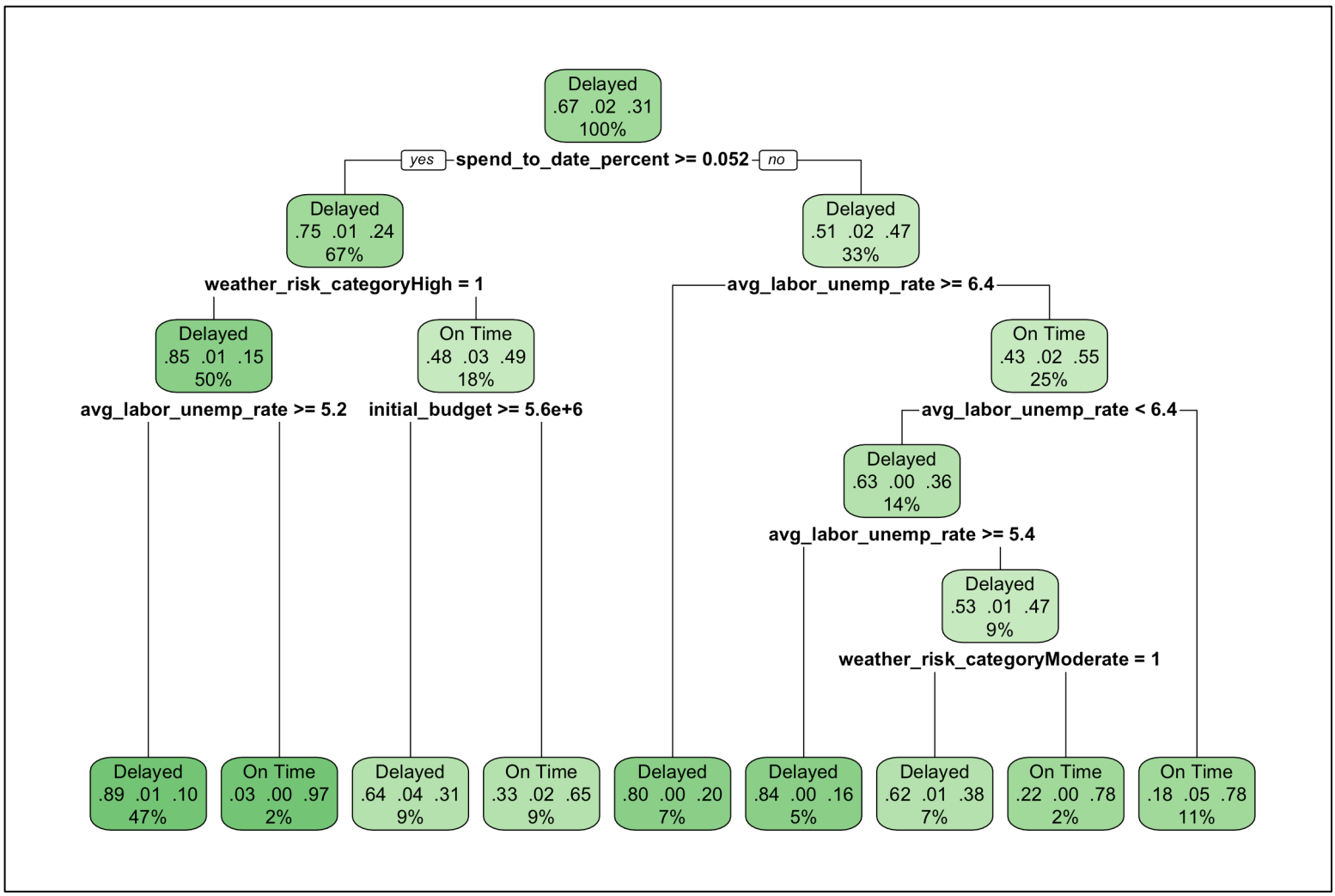


Figure 5: *Decision Tree Visualization for Delay Classification*

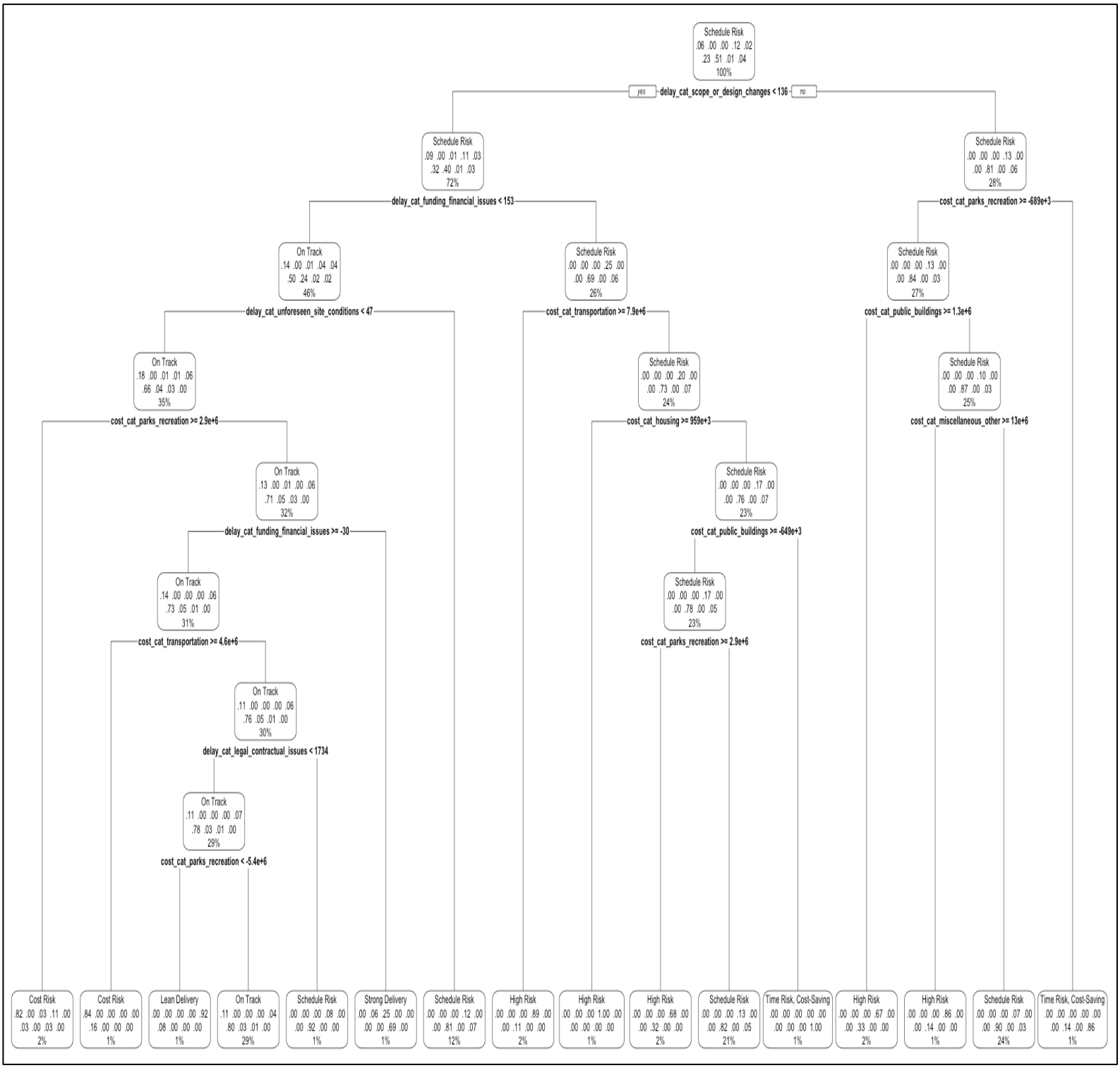
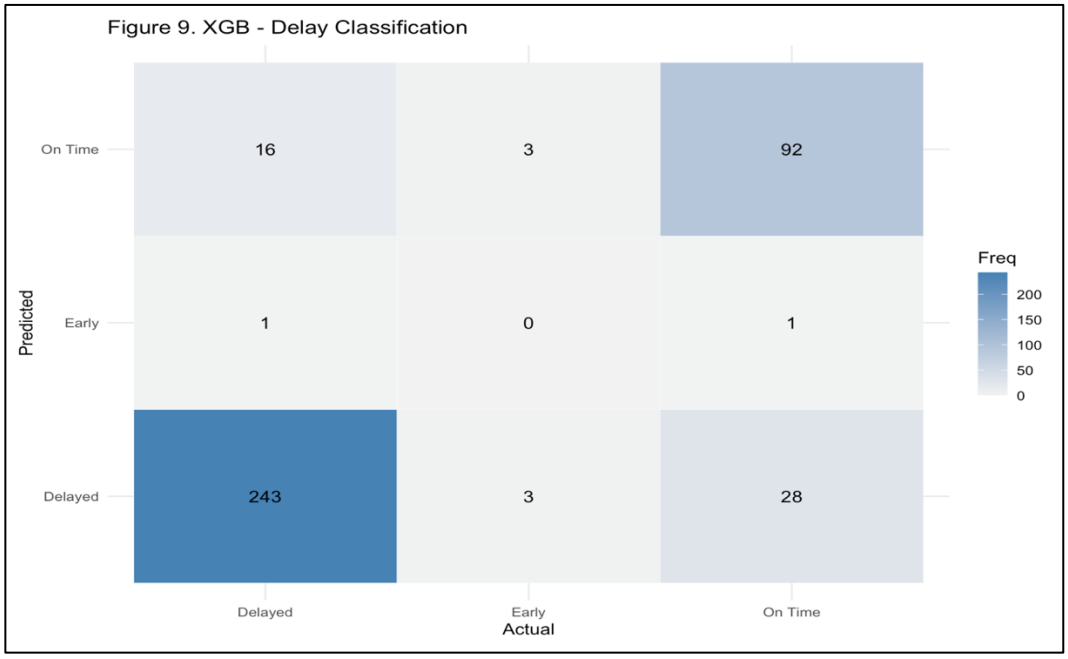


Figure 6: *Decision Tree Visualization for Project Profile Classification*



B. Sensitivity and Performance Metrics For XGB – Profile Classification

Detailed evaluation metrics (including sensitivity, specificity, precision, and balanced accuracy):

```
Overall Statistics

               Accuracy : 0.8594
                 95% CI : (0.8205, 0.8926)
    No Information Rate : 0.5182
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7801

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Cost Risk Class: Exceptional Performance Class: Fast but Costly Class: High Risk
Sensitivity                   0.58333                            NA                0.000000           0.64444
Specificity                   0.99444                             1                0.997389           0.96755
Pos Pred Value                0.87500                            NA                0.000000           0.72500
Neg Pred Value                0.97283                            NA                0.997389           0.95349
Prevalence                    0.06250                             0                0.002604           0.11719
Detection Rate                0.03646                             0                0.000000           0.07552
Detection Prevalence          0.04167                             0                0.002604           0.10417
Balanced Accuracy             0.78889                            NA                0.498695           0.80600
                     Class: Lean Delivery Class: On Track Class: Schedule Risk Class: Strong Delivery
Sensitivity                      0.428571          0.9775               0.9397                0.750000
Specificity                      1.000000          0.9593               0.8703                0.997368
Pos Pred Value                   1.000000          0.8788               0.8863                0.750000
Neg Pred Value                   0.989501          0.9930               0.9306                0.997368
Prevalence                       0.018229          0.2318               0.5182                0.010417
Detection Rate                   0.007812          0.2266               0.4870                0.007812
Detection Prevalence             0.007812          0.2578               0.5495                0.010417
Balanced Accuracy                0.714286          0.9684               0.9050                0.873684
                     Class: Time Risk, Cost-Saving
Sensitivity                            0.46667
Specificity                            0.99187
Pos Pred Value                         0.70000
Neg Pred Value                         0.97861
```

Table 1: Accuracy, Sensitivity and Specificity by Class – Profile Classification

C. Top 20 Most Important Features Influencing Construction Project Delay and Cost Overrun from XGB-Classification

	Cost Risk <dbl>	Exceptional Performance <dbl>	Fast but Costly <dbl>	High Risk <dbl>	Lean Delivery <dbl>	On Track <dbl>	Schedule Risk <dbl>
delay_cat_funding_financial_issues	46.295253	4.454620	14.5228446	48.595912	24.947883	100.00000	78.936607
delay_cat_scope_or_design_changes	40.533866	6.789914	9.0212397	30.840295	23.566342	91.03962	69.035937
delay_cat_unforeseen_site_conditions	30.510862	5.800198	6.7899136	16.968546	16.711258	68.44188	53.004773
cost_cat_parks_recreation	57.571785	4.454620	6.1390084	47.082961	39.126357	42.85283	47.551021
cost_cat_transportation	47.465130	4.454620	0.9972339	36.631237	14.434415	27.89950	27.452946
cost_cat_public_buildings	14.629586	9.365578	5.8001979	45.871814	12.434204	28.67105	31.911466
cost_cat_housing	6.356313	4.454620	4.4546205	37.512751	6.276624	14.42715	21.626250
cost_cat_miscellaneous_other	2.625290	4.454620	7.7335450	33.837502	14.156870	23.32439	18.132683
avg_labor_unemp_rate	23.076261	7.475555	9.7195408	18.253599	16.228371	33.38678	32.541270
spend_to_date_percent	28.798037	6.789914	8.1677223	20.997890	3.500715	28.20747	32.197122
avg_construction_jobs	22.315125	7.153902	8.6586084	13.129796	13.331266	30.76867	22.707422
initial_budget	24.031441	4.454620	4.0219924	8.270703	9.233283	22.21232	29.892446
weather_risk_categoryHigh	9.493719	6.789914	4.8139070	16.063045	6.685077	28.23884	28.778887
project_ph_construction	12.132422	4.454620	7.3280478	6.672108	12.000244	27.74367	22.948634
project_ph_completed	11.644867	6.359464	6.6982904	14.238617	9.899321	18.65523	25.378622
cost_cat_infrastructure	10.066029	4.454620	11.4836419	23.867003	10.585039	15.79172	15.783463
cost_cat_environmental	23.577340	4.454620	4.4546205	9.234348	6.954514	17.40568	9.847063
project_ph_design	16.014798	2.549777	2.6730963	8.879762	10.573317	22.50903	15.903217
delay_cat_legal_contractual_issues	10.193180	4.454620	4.4546205	13.556905	7.126301	19.59923	22.372025
project_ph_close_out	12.700683	4.454620	6.7899136	16.795593	10.099634	14.92507	19.408517

20 rows | 1–8 of 9 columns

Table 2: Ranked Importance of Key Predictor in Construction Project Risk Modeling

D. Code Repository

The full codebase used for data preprocessing, feature engineering, modeling (Random Forest, Decision Tree, XGBoost), evaluation, and visualizations is publicly available at:

GitHub Repository: https://github.com/Doumgit/Capstone-Project-D-698/tree/main/Research%20Paper%20Code