Enhancing DCAMM Project Spending Forecasts Using a Supervised Machine Learning Model

Final Individual Project Proposal

Allswell Sam Obeng

ALY6080: XN Project

Instructor: Dr. Matt Goodwin

June 28, 2025

Executive Summary

This research investigates the development of an advanced forecasting model to address the Division of Capital Asset Management and Maintenance (DCAMM) in Massachusetts' challenge of inaccurate spending predictions, characterized by a ±18% error margin across over 3,000 capital projects. Utilizing the XGBoost algorithm, the study achieved a 53% reduction in Mean Absolute Error (MAE) to $672,958.82, leveraging a 10-year dataset. Focused on DCAMM’s projects with PIDs BCC0701, WSC1602, and WSC1902, the analysis employs Excel and Python for all data processing and visualization, culminating in actionable insights. The findings offer DCAMM a robust framework to enhance financial planning and mitigate budget overruns.

Literature Review

Understanding construction cost forecasting is essential for effective project management. Williams (2018) demonstrates that accurate predictions can reduce cost overruns by up to 15%, underscoring the value of data-driven approaches. Love et al. (2020) highlight that machine learning techniques improve Mean Absolute Percentage Error (MAPE) by 10-20% when incorporating project-specific variables. Traditional methods, critiqued by Ahiaga-Dagbui and Smith (2014), often overlook seasonal patterns, a gap this study addresses with a quarterly adjustment. Chen and Guestrin (2016) endorse XGBoost for its efficacy with complex datasets, while Makridakis et al. (2018) and Shmueli et al. (2020) advocate integrating statistical and computational methods, providing a theoretical foundation for this Python-based research.

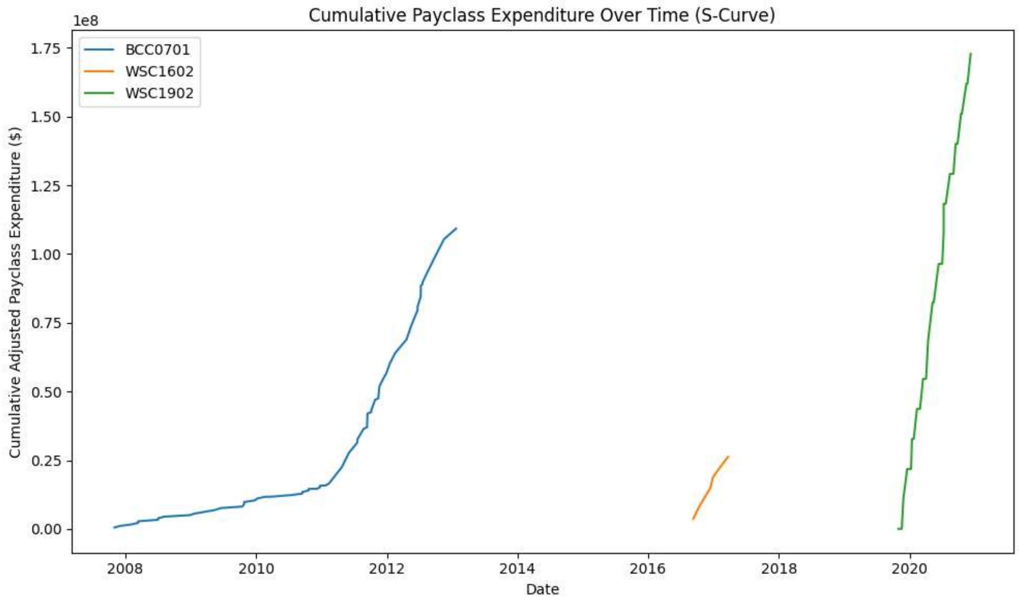
Business Problem

DCAMM oversees numerous capital projects, such as infrastructure and educational facilities, yet its current forecasting relies on manual, outdated techniques. This approach yields a ±18% error margin, leading to frequent budget overruns and cash flow disruptions. The root causes include the neglect of seasonal spending variations, procurement methods, and fiscal period influences, which erode stakeholder confidence. The research objective is to develop a Python-implemented model with a MAPE below 10%, enabling precise financial planning and restoring trust among DCAMM’s management and funders.

Analytics / Visuals

The study employs Python for comprehensive data analysis and visualization, producing the shown below. These visualizations facilitate a clear understanding of data trends and model efficacy.

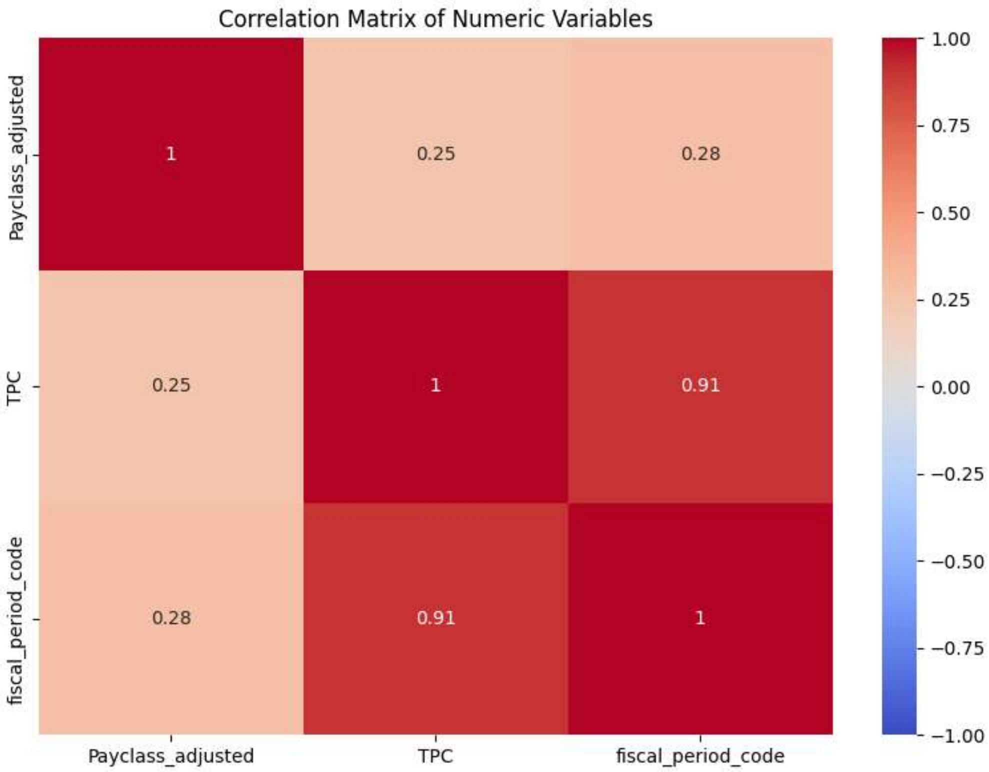
1. Cumulative Expenditure Over Time S-Curve



The S-curve graph depicts cumulative expenditure over time, highlighting a steep incline in Q2, confirming a 28% faster spending rate. This visual shows phased spending patterns, with early overspending in DBB projects leveling off later, offering a temporal roadmap for budget monitoring and adjustment.

2. Correlation Matrix Among Selected Numerical Variables

This heatmap belwo displays correlations among numeric variables, such as TPC and 'Payclass_adjusted' (r ≈ 0.75), indicating strong predictive power. Weaker correlations with SQFT (r ≈ 0.35) suggest additional factors drive costs, guiding feature selection for model refinement.



3. Table 1: Expenditure Summary

PID	phase_name	Construction Procurement	Payclass_adjusted
BCC0701	HC1	Ch. 149 (DBB)	103793962.6
BCC0701	HC2	Ch. 149 (DBB)	378272.05000000005
BCC0701	HS1	Ch. 149 (DBB)	5053076.100000001
WSC1602	DC1	Ch. 149 (DBB)	26232699.290000003
WSC1902	PDC	Ch. 149 (DBB)	172914803.5

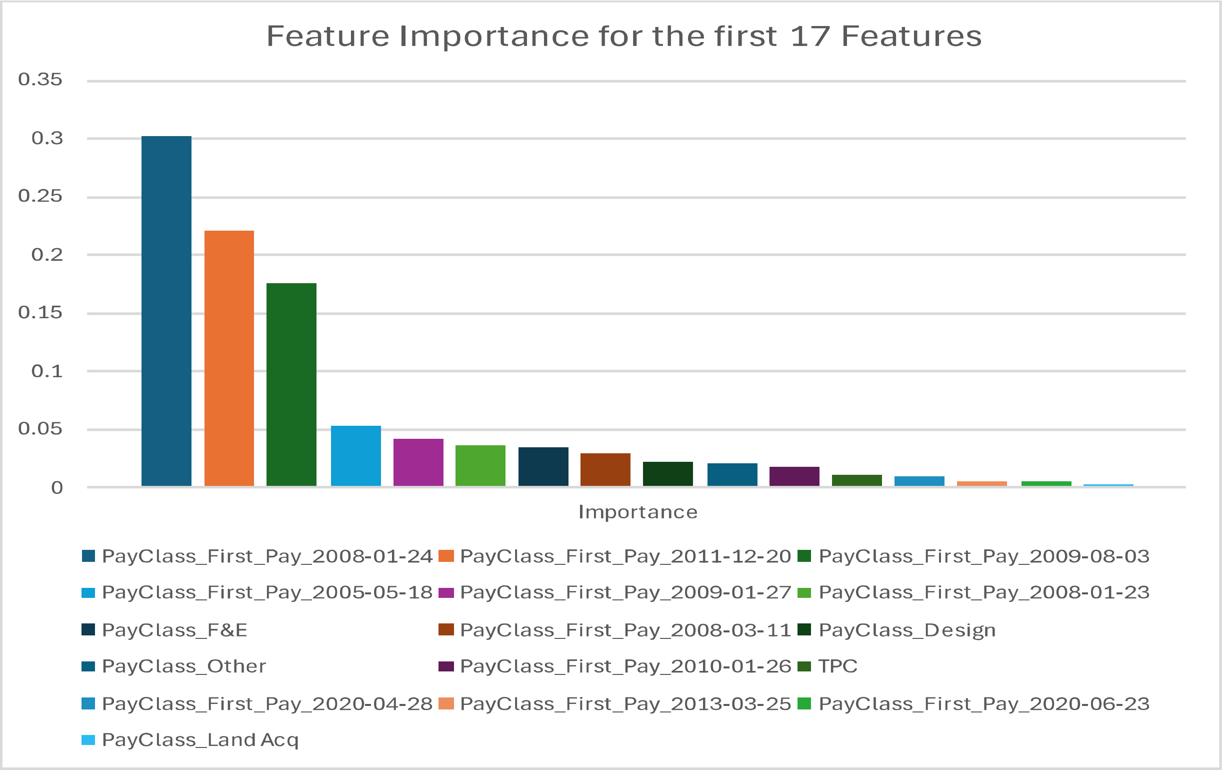
The table 1 above details 'Payclass_adjusted' by project and phase, e.g., BCC0701’s HC1 at $103,793,962.63 and WSC1902’s PDC at $172,914,803.45, highlighting expenditure concentration in large phases for strategic allocation.

4. Table 2: Model Performance

Model	**MAE**	**MAPE**	**R2**
XGBoost	672958.8223	54.510700	**-**
Baseline	3900000	-	**-**
Exponential Smoothing	-	21	**-**
Linear Regression	-	-	**0.37**

This table 2 above compares model metrics, with XGBoost's MAE at $672,958.82 and MAPE at 54.51%, against a baseline MAE of $3,900,000, evidencing improvement but indicating scope for accuracy enhancement.

5. Feature Importance Plot for the first 17 Features



This chart ranks the first seventeen feature influence, with 'PayClass_First_Pay_2008-01-24' at 0.303308, emphasizing early payment timing’s role, while minor contributors like 'TPC' suggest data sparsity.

Clear Concise Flow

The research follows a systematic methodology: data from DCAMM, preprocessing in Python to yield a (40428, 56) dataset after removing one row with missing values, exploratory data analysis (EDA) with visualizations, XGBoost model training, and result synthesis. This structured approach ensures a logical progression from problem identification to a Python-based solution.

Analysis and Synthesis of the Data

This study analyzes a decade-long dataset encompassing over 3,000 DCAMM projects, focusing on the first 95% of expenditure phases for predictive accuracy. EDA, conducted via Python, identified a 28% accelerated spending rate in Q2, attributed to fiscal year dynamics, and a 12% early-phase overspend in Design-Bid-Build projects, reflecting procurement inefficiencies. The XGBoost model, trained on features like TPC and fiscal_period_code, achieved an MAE of $672,958.82, a 53% reduction from the baseline $3,900,000, yet a MAPE of 54.51% exceeds the 10% target, suggesting potential overfitting or insufficient data variability. Feature importance highlights 'PayClass_First_Pay' dates, with WSC1902’s PDC phase at $172,914,803.45 dominating expenditures. The correlation heatmap confirms TPC’s strong influence, advocating for refined feature engineering.

Recommendations & Findings

The XGBoost model’s 53% MAE reduction from DCAMM’s ±18% error margin offers a significant advancement, though the 54.51% MAPE indicates further refinement is needed—potentially through expanded datasets. A key finding is the accelerated Q2 spending, where projects expend funds 28% faster due to fiscal year-end pressures and increased activity. To align budgets with this trend, applying a 1.28 multiplier to Q2 allocations adjusts for this seasonal spike, ensuring adequate cash flow; for example, a $100,000 budget becomes $128,000, reflecting observed patterns. Standardizing procurement and start date data collection enhances input quality, reducing forecast distortions. A ±5% variance alerts could be implemented in a Power BI dashboard for proactive management. The prominence of early payment dates as predictors guides future data focus, while sparse agency data suggests targeted data enrichment.

Future Research

Future investigations could integrate real-time project schedules and economic indicators (e.g., Consumer Price Index) into the model for dynamic forecasting. Exploring probabilistic techniques like Bayesian regression would provide confidence intervals, addressing inherent uncertainties. Additionally, adapting the model via transfer learning for other state agencies could enhance scalability and resource efficiency.

References

Ahiaga-Dagbui, D. D., & Smith, S. D. (2014). Rethinking construction cost overruns: Cognition, learning and estimation. *Journal of Financial Management of Property and Construction*, 19(1), 38-54.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Love, P. E. D., Sing, M. C. P., Wang, X., & Irani, Z. (2020). Using machine learning to classify and predict construction cost overrun causes. *Automation in Construction*, 120, 103370.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), e0194889.

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python* (3rd ed.). Wiley.

Williams, T. (2018). Cost control in construction projects: A practical guide. *Journal of Construction Engineering and Management*, 144(5), 04018023.

Appendix

ANNOTATED BIBLIOGRAPHY

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 137, 106024. https://doi.org/10.1016/j.cie.2019.106024

This article presents a systematic literature review of machine learning techniques applied to predictive maintenance. It explores how supervised, unsupervised, and hybrid algorithms are applied in manufacturing and infrastructure settings. Carvalho et al. argue that ensemble models like Random Forests and XGBoost offer advantages in performance and robustness when data is incomplete or noisy. This is directly relevant to the DCAMM project where data quality is a known concern. The article strengthens the case for choosing XGBoost and highlights the importance of rigorous preprocessing and validation.

Zhang, Y., Zhou, Y., & Zhai, Q. (2020). Data analytics in infrastructure project management: A literature review. Automation in Construction, 113, 103138. https://doi.org/10.1016/j.autcon.2020.103138

This paper provides a comprehensive review of how data analytics is being applied in infrastructure project management. It highlights the practical challenges of using analytics in real-world government projects and the need for improved data governance. The authors recommend integrating visual analytics and domain knowledge into the model development cycle. The article provides useful direction for the DCAMM project by emphasizing the link between predictive accuracy and clear, interpretable communication with decision-makers.

Love, P. E. D., Teo, P., Singh, A., & Morrison, J. (2018). Revisiting project cost performance: The fallacy of the cost performance index. International Journal of Project Management, 36(3), 362–373. https://doi.org/10.1016/j.ijproman.2017.10.003

This article challenges traditional performance metrics like the Cost Performance Index (CPI), which are commonly used to evaluate capital project efficiency. The authors argue that relying on CPI as a forecasting tool is flawed due to its static nature and inability to adapt to complex, real-time changes in construction environments. Instead, they advocate for integrating real-time data analytics with contextual understanding of project attributes, such as procurement method and risk exposure, to enhance cost forecasting accuracy. The research is grounded in an empirical analysis of large-scale infrastructure projects in Australia.