**Forecasting Capital Costs with Machine Learning at DCAMM**

Anuraag Macha, Matthew Branum

College of Professional Studies, Northeastern University

ALY 6980: Capstone, Spring Term

Dr. Matthew Goodwin

June 28, 2025

**Abstract**

This project applies advanced machine learning techniques to forecast total project expenditures for capital construction projects managed by the Massachusetts Division of Capital Asset Management and Maintenance (DCAMM). Historical cost data were cleaned, merged, and adjusted for inflation using RSMeans indices to ensure temporal comparability. Feature engineering steps introduced variables such as spend pacing, project duration, and seasonal expenditure behavior. Exploratory analysis revealed patterns in regional spending, quarterly cost seasonality, and correlations between spending concentration and cost overruns.

Three regression models—Random Forest, XGBoost, and LightGBM—were developed using pre-construction features to predict inflation-adjusted cost. Random Forest produced the best predictive performance (R² = 0.577), with XGBoost and LightGBM close behind. K-fold cross-validation confirmed the relative robustness of these models, though generalization performance remained modest, highlighting the challenges of forecasting with pre-construction data alone. Residual analyses indicated persistent variance at higher cost ranges, suggesting the presence of unmodeled factors.

Visualizations such as seasonal line charts, county-level heatmaps, and decision tree diagnostics were used to enhance interpretability. This work demonstrates the feasibility of leveraging machine learning to support more data-driven cost planning, while also underscoring the importance of refined inputs for improved early-stage forecasts.

*Keywords:* capital project forecasting, inflation-adjusted cost modeling, RSMeans, Random Forest, XGBoost, LightGBM, spend pacing, feature engineering, cross-validation, construction analytics, residual diagnostics, DCAMM, Python.

**Forecasting Capital Costs with Machine Learning at DCAMM**

The Division of Capital Asset Management and Maintenance (DCAMM) is responsible for managing state-funded construction projects across Massachusetts. To improve forecasting accuracy and operational efficiency, this study explores the use of machine learning models to predict inflation-adjusted capital project costs using historical expenditure data. Leveraging cleaned and engineered datasets, the analysis centers on modeling cost behavior prior to project completion using features available during the planning and early execution phases. The process begins with context setting and data preparation, followed by an extensive feature engineering pipeline—including RSMeans index normalization and project timeline calculations. Several regression models, including Random Forest, XGBoost, and LightGBM, are developed and compared using traditional evaluation metrics and cross-validation. The ultimate goal is to support DCAMM’s ability to anticipate project costs more accurately and facilitate proactive decision-making in capital planning.

**Executive summary**

The Division of Capital Asset Management and Maintenance (DCAMM) is working to enhance its ability to forecast capital project costs using modern machine learning approaches. This initiative focused on developing predictive models using historical spending data, with an emphasis on pre-construction features such as location, building type, and early expenditure signals. The project involved extensive data cleaning, normalization using RSMeans indices, and inflation adjustments to bring legacy costs into alignment with current dollar values.

Exploratory data analysis revealed that most projects in the dataset were renovations, and that new construction projects tend to reach significant budget milestones, such as 95% of total spending, much earlier in their timelines. Spending behaviors also varied by project category; for instance, HVAC and public safety infrastructure frequently showed front-loaded expenditures, while administrative and office projects exhibited more evenly paced costs.

Three machine learning models—Random Forest, XGBoost, and LightGBM—were trained and evaluated on their ability to predict total adjusted cost. Random Forest delivered the strongest out-of-sample performance, achieving an R² score of 0.577. However, residual analysis revealed that all models struggled to generalize across high-cost outliers. Cross-validation confirmed model stability, but also highlighted limitations due to reliance on finalized cost inputs.

To improve early-stage cost forecasting, the report recommends further expansion of the dataset to include detailed pre-construction attributes such as structural specifications, special equipment needs, and procurement schedules. Incorporating such features could enable DCAMM to make more data-informed budget projections earlier in the project lifecycle.

**Business Problem**

Accurate cost forecasting is a critical need for public infrastructure agencies like DCAMM, which manage hundreds of capital projects annually. Fluctuations in inflation, labor, and material costs pose major challenges to budgeting and financial planning. Without reliable forecasting methods, agencies risk underestimating expenses, leading to budget overruns, scope reductions, or project delays.

DCAMM’s portfolio includes a mix of new construction and renovation projects, each with distinct spending profiles. Historical trends indicate that new builds often front-load spending, while renovation projects progress more evenly. These patterns complicate cost estimation, particularly in the early phases when key specifications are still evolving.

To address this, the goal of the project is to develop a machine learning–based forecasting model that enhances DCAMM’s ability to predict total project costs using early-stage indicators. The approach integrates inflation-adjusted spending data with engineered features and predictive modeling techniques. Models such as Random Forest, XGBoost, and LightGBM are compared based on their accuracy and robustness.

The project aims to (1) identify the most predictive features of total cost, (2) evaluate model performance using metrics like MAE, RMSE, and R², and (3) establish a scalable methodology that DCAMM can apply across project types. A target of 15–20% improvement over existing estimation accuracy is set, with a focus on generalizability to support long-term planning and capital allocation.

**Milestones**

1. Feature engineering & selection (by February 23)
2. Model selection, creation & development (by March 16)
3. Model evaluation & validation (by March 23)
4. Data visualization & presentation (by March 29)
5. Final conclusions and submission (by March 29)

**Literature Review**

Research in the field of public sector capital construction emphasizes the need for accurate cost forecasting, efficient risk mitigation strategies, and optimized resource allocation. Existing studies frequently examine the outcomes of prior government infrastructure projects to identify patterns in budgeting, scheduling, and compliance. This body of work underscores the influence of regulatory structures, procurement policies, and stakeholder coordination on project success.

Particular attention has been given to improving cost estimation practices, especially in response to challenges like inflation, labor shortages, and material volatility. Academic and industry publications also highlight the growing interest in applying data-driven methodologies—such as machine learning and statistical forecasting—to support more reliable capital planning. Reviewing these prior initiatives helps contextualize the challenges DCAMM faces today, offering relevant insights into common drivers of delays, budget variances, and policy constraints that can impact large-scale construction efforts.

**Exploratory Data Analysis**

The exploratory phase of the project began with rigorous data cleaning and preparation to ensure the accuracy and usability of the dataset for modeling. Two primary datasets were used: a summary-level dataset providing aggregated project costs by category and location, and a detailed expenditure dataset capturing individual project payments over time.

To begin, irrelevant or incomplete features were dropped—this included columns with all missing values, low variability, or those deemed non-predictive (e.g., document references, cabinet names, or funding sources). Categorical fields such as City, County, Construction Type, and Agency were standardized to consistent naming formats and later one-hot encoded.

Special care was taken to normalize monetary values: dollar signs and other symbols were stripped, and all numeric fields were converted to appropriate data types. Missing City and County values were filled using contextual lookups from project descriptions and DCAMM reference materials.

One of the most critical preprocessing steps involved adjusting expenditures for inflation. The RSMeans City Cost Index was used to calculate a standardized metric called adjusted\_expended, pegged to 2025-dollar values. Each project's expenditures were multiplied by a city-specific weighting factor derived from the RSMeans index, ensuring cost comparability across different years and locations.

Several engineered features were created to capture meaningful project behaviors:

* Total Duration (Months): Calculated for each project to represent the full span from the first to last payment.
* Pacing Metrics: Features such as the month and percentage at which 95% of total spend was reached, enabling classification of projects as front-loaded or back-loaded.
* Seasonality Markers: Quarterly and monthly acceptance dates were extracted to allow seasonal trend analysis.
* Spending Concentration: Cumulative expenditures were analyzed to identify patterns such as mid-project cost spikes or flat trajectories.

After preprocessing, multiple analyses were performed to uncover project-level trends:

* **Regional Spending Patterns:** Aggregated spending by county showed that Worcester, Suffolk, and Essex accounted for the highest total adjusted costs.
* **Project Completions:** The number of projects accepted annually surged around 2014–2016, consistent with infrastructure investment surges in that period.
* **Seasonal Trends:** A quarterly line plot of adjusted expenditures revealed cyclical peaks in spending, especially between 2011 and 2016, before a gradual decline.
* **Cost Pacing Behavior:** Boxplot comparisons showed that front-loaded projects had wider cost overrun variability than back-loaded ones—highlighting the risk of early budget exhaustion.
* **Outliers:** Residual plots during modeling showed some variance at the upper end of project costs, which was noted for potential future feature calibration.

Together, the data cleaning, feature engineering, and EDA process established a robust foundation for machine learning. These steps not only improved the quality and consistency of the input data but also introduced interpretable variables crucial to the downstream forecasting models.

**Figure 1**

*Counties with Most Capital Investment*A graph of a bar graph

AI-generated content may be incorrect.

Figure 1 highlights the top 15 Massachusetts counties ranked by total inflation-adjusted capital spending. Worcester County leads all regions, followed by Suffolk, Essex, and Middlesex. This geographic pattern reflects the prioritization of heavily urbanized or infrastructure-dense regions.

Notably, inconsistencies in county naming (e.g., “Middlesex” vs. “MIDDLESEX”) suggest opportunities to improve data standardization in future reporting. The dominance of a few counties also raises questions about equity in resource distribution. These insights can help DCAMM evaluate whether funding aligns with actual infrastructure demands and support more balanced investment across the state.

**Figure 2**

*Number of Projects Completed per Year*

A graph with blue lines

AI-generated content may be incorrect.

Figure 2 illustrates the annual number of unique projects completed, based on acceptance year. From 2004 through 2015, the data reveals a consistent upward trend in project completions, culminating in a peak around 2014–2015. This likely corresponds to a period of heightened capital investment or infrastructure program expansion across the Commonwealth.

Post-2015, however, completions declined steadily year over year. This drop may signal tighter budgets, shifting policy priorities, or slower project execution timelines. Additionally, recent years—especially post-2021—show a sharp decline, which may partially reflect delayed reporting or incomplete records for ongoing projects.

Understanding these temporal trends provides valuable context for interpreting spending behaviors and anticipating future project delivery volumes across DCAMM’s capital program.

**Figure 3**

*Seasonality in Spending (Quarterly View)*

**A graph with blue lines and numbers

AI-generated content may be incorrect.**

Figure 3 presents total inflation-adjusted spending across quarters, highlighting seasonal patterns in DCAMM project expenditures from 2004 through 2025. Spending increased sharply around 2010, peaking in Q2–Q3 of 2011—likely the result of stimulus efforts or strategic infrastructure investment.

Following 2015, a steady decline in quarterly spending is evident, interrupted by a few modest rebounds. These fluctuations may reflect shifting budget priorities, delays in construction activity, or evolving funding mechanisms. Notably, post-2020 spending appears significantly lower, possibly due to pandemic-related disruptions or reporting lags in recent data.

While the plot does not reveal strong cyclical seasonality, some years show mild consistency in higher spending during Q2 or Q3, potentially linked to construction timelines or fiscal planning cycles. Identifying these seasonal variations helps contextualize past trends and supports the design of forecasting models that account for temporal dynamics.

**Figure 4**

*Cost Overruns by Spend Pacing*

A diagram of a cost overlay

AI-generated content may be incorrect.

Figure 4 illustrates the distribution of cost overrun ratios between two distinct spending profiles: projects that were front-loaded (majority of spending occurred early) versus those that were back-loaded (spending occurred later in the timeline). The results show a clear contrast in performance.

Front-loaded projects exhibit significantly higher median cost overrun ratios, with a wider interquartile range and several notable outliers. These characteristics suggest greater risk exposure, possibly due to aggressive early-stage budgeting, rushed project starts, or evolving scope during execution. In contrast, back-loaded projects are more tightly clustered around zero, indicating tighter budget control and more predictable execution.

These findings highlight the importance of spend pacing as a potential risk signal. Monitoring and classifying projects by their expenditure timing can serve as a proactive measure for identifying high-risk initiatives and adjusting project oversight accordingly.

**Modeling Enhancements and Interpretations**

The modeling phase of this study was guided by the objective of producing a predictive framework capable of estimating capital project costs using only features available prior to construction. This objective aligns with DCAMM’s strategic goal of proactive cost management, enabling more informed allocation of funds even before ground is broken. By limiting the model to pre-construction data, such as project fiscal year, pay class year, estimated square footage, jurisdictional identifiers, and agency codes, we prioritized features that are known early in the project lifecycle. This constraint was not only practical but also methodologically sound—it enforced a real-world boundary around data availability that reflects the conditions under which DCAMM planners operate.

Three tree-based ensemble methods were selected for model development: Random Forest, XGBoost, and LightGBM. These algorithms are known for their high performance on structured datasets, ability to model nonlinear interactions, and relative ease of interpretation compared to deep learning models. Random Forest served as the baseline model due to its robustness and reduced tendency to overfit. XGBoost and LightGBM were included to explore whether gradient boosting frameworks could yield additional gains in predictive accuracy. All models were trained using an 80/20 train-test split with a consistent random seed to ensure reproducibility. One-hot encoding was applied to categorical variables, and features were filtered for null values to maintain data integrity. While no imputation strategies were used in this round, future iterations may benefit from more sophisticated handling of sparse fields.

Model performance was evaluated using standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R²). Random Forest emerged as the top performer, with an R² of approximately 0.577 and the lowest MAE. XGBoost closely followed, with LightGBM showing slightly higher residual variance. Despite these differences, all models achieved acceptable error thresholds, suggesting that the pre-construction variables encode meaningful signals related to project cost.

To further probe model behavior, residual plots were generated. These plots compare actual versus predicted values, displaying the residuals along the y-axis. Each model demonstrated tightly clustered residuals near zero for projects with lower and moderate costs. However, variance increased at higher actual cost values, a known artifact in real-world cost modeling due to the natural volatility of large capital expenditures. LightGBM exhibited a broader residual spread at the upper end, indicating potential underestimation for high-value projects.

**Figure 6**

*Residual Analysis*

A diagram of a graph

AI-generated content may be incorrect.

Figure 6 displays a series of residual plots used to evaluate the predictive performance and error behavior of three machine learning models—Random Forest, XGBoost, and LightGBM—developed to forecast inflation-adjusted capital project costs at DCAMM. Each plot presents residuals on the y-axis and actual values on the x-axis, where residuals are defined as the difference between predicted and actual costs. A horizontal red line at zero represents perfect prediction, allowing deviations above and below this baseline to be visually assessed.

The Random Forest model demonstrates a relatively symmetric and centered residual distribution, with no significant skew or clustering. This suggests that, although individual predictions vary, the model does not systematically overestimate or underestimate costs across the dataset. The even spread of residuals around the zero line also indicates model robustness, particularly for mid-range cost projects. Notably, as actual cost increases, the spread of residuals widens, a common characteristic of heteroskedastic data. This implies that forecasting variance grows with project scale, a reflection of the inherent uncertainty in modeling high-cost capital expenditures.

The XGBoost model follows a similar residual pattern but with slightly greater dispersion, especially in the upper-middle quartile of actual values. This increased variance may stem from the model's aggressive gradient boosting strategy, which can lead to localized overfitting in regions with denser or more complex cost distributions. Nonetheless, the central tendency of the residuals remains largely unbiased, reaffirming the model’s overall reliability despite moderate increases in error spread.

LightGBM, by contrast, exhibits more pronounced and systematic deviations. Residuals in the upper range of actual costs are skewed negatively, signaling consistent underpredictions for high-expenditure projects. This suggests that the model is less capable of capturing nonlinear cost behaviors or rare, large-scale expenditure patterns. Such underperformance may be attributed to sensitivity in LightGBM's loss function, over-pruning of decision trees, or insufficient representation of outlier projects in the training dataset.

These findings carry important implications for operational deployment. While all three models offer viable predictions within a reasonable error margin, the Random Forest model delivers the most stable and interpretable results across the full spectrum of project costs. The observed heteroskedasticity across all models reinforces the need for refined techniques when forecasting large-scale projects. Potential strategies for future improvement include implementing weighted loss functions, segmenting models by project tier, or integrating external economic indicators to enhance high-value prediction accuracy.

In conclusion, residual plots serve as a critical diagnostic tool for model evaluation. By examining the shape, dispersion, and directionality of errors, stakeholders can better understand each model’s strengths and weaknesses. These insights support informed decisions about which algorithm to adopt for specific forecasting tasks and where to focus refinement efforts to improve long-range budget planning at DCAMM.

**Figure 7**

*SHAP Feature Importance and Interpretability*

A graph of a graph with text

AI-generated content may be incorrect.

Figure 7 presents a SHAP (SHapley Additive exPlanations) summary plot, a visualization technique used to quantify the global importance of each input feature in the Random Forest cost prediction model. This plot shows the distribution of SHAP values for the most influential features across all samples, providing insight into both the magnitude and direction of each feature’s contribution to the model's output. The x-axis represents the SHAP value, which indicates the impact of a given feature on the model's prediction for a specific observation—positive values push the predicted cost higher, while negative values push it lower. Color gradients reflect feature magnitude for each observation, with red indicating high feature values and blue indicating low feature values.

The top three features in terms of predictive power are PayClass FY, Project\_FY, and SQFT. These variables consistently surfaced across multiple models as the most important drivers of cost variation. Their interpretability aligns with domain knowledge: PayClass FY reflects the fiscal planning horizon, capturing budget cycle nuances; Project\_FY represents the project initiation year, which embeds inflation-adjusted economic context and policy shifts; and SQFT acts as a direct proxy for the physical scale and likely material intensity of the project. The wide SHAP value range for each of these features illustrates their substantial influence, with high values often associated with significantly elevated predicted costs.

Beyond core financial and physical dimensions, the model also leverages geographic and administrative variables. Features such as City\_SALEM, County\_HAMPSHIRE, and Agy\_ITD (a flag for the Information Technology Division) rank among the next most influential predictors. Their presence suggests that regional cost dynamics, agency-level procurement behavior, and organizational oversight practices materially affect final expenditure levels. These features likely capture latent institutional variability that is not directly observable through numeric project metrics alone. Notably, these factors exhibit bidirectional influence; for instance, the same city may be associated with both higher and lower costs depending on context, indicating interaction effects with other variables in the model.

This summary plot underscores the advantage of SHAP in moving beyond conventional feature importance measures by offering a more nuanced, sample-level interpretation of how input variables behave across the entire dataset. In practical terms, this allows analysts to identify not only which features matter most, but also under what conditions they exert upward or downward pressure on cost estimates. Such transparency enhances trust in the model's outputs and supports more refined budget planning and risk assessment workflows.

In sum, the SHAP summary plot validates the model’s dependence on intuitive and relevant features while also revealing the layered complexity of institutional and regional cost behaviors. Its use within the DCAMM forecasting framework reinforces the project’s broader commitment to explainable AI and informed public-sector decision-making.

**Figure 8**

*SHAP Force Plot: Local Prediction Interpretation*

A graph with numbers and symbols

AI-generated content may be incorrect.

Figure 8 presents a SHAP (SHapley Additive exPlanations) force plot designed to provide a detailed, local-level explanation of a single project cost prediction generated by the Random Forest model. This visualization disaggregates the total predicted cost into its component contributions from each feature, showing precisely how and why the model arrived at its final output for this specific case. The force plot is centered on the model’s expected value, or baseline prediction (E[f(x)] = 193,537.75), and highlights how individual input features pushed this estimate higher or lower to reach the final predicted cost of approximately $3.67 million.

In this example, the prediction is dominated by three primary contributors: Project\_FY, PayClass FY, and SQFT. These features—representing project fiscal year, payment classification year, and square footage—collectively add over $3 million to the baseline estimate. The contributions of Project\_FY (+$1.27M), PayClass FY (+$1.27M), and SQFT (+$844K) signal a large-scale, recent project with an early fiscal designation and significant physical scope. These indicators are closely aligned with the insights from the SHAP summary plot and global feature importance rankings, reinforcing their predictive significance.

Additional, smaller contributions from geographic and organizational variables—including City\_SALEM, County\_HAMPSHIRE, and Agy\_ITD—further modulate the prediction. Some of these features exert downward pressure (in blue), while others increase the estimate (in red). This nuanced layering of effects illustrates how even secondary attributes, such as jurisdiction or agency involvement, may compound to influence model output.

This form of local interpretability is particularly valuable for operational deployment in public-sector environments. Finance officers, project managers, or oversight committees can use force plots to justify cost estimates, communicate project risk profiles, and build institutional confidence in data-driven decision tools. Rather than relying on opaque algorithms, stakeholders are presented with a clear breakdown of the variables that shape each prediction.

Moreover, this granular visualization transforms the model from a passive forecasting mechanism into an active explanatory aid. In budgetary discussions, the force plot enables users to identify which aspects of a project (e.g., timing, size, location) are driving projected costs and to assess whether those projections are reasonable given the project's known characteristics. It can also help guide corrective strategies, such as evaluating whether altering procurement timelines or design specifications could impact expected expenditures.

The SHAP force plot bridges the gap between model performance and interpretability by offering a transparent, project-specific explanation of predicted outcomes. It enhances accountability, supports data-informed planning, and builds trust in machine learning models as practical tools for financial forecasting and capital investment decision-making at DCAMM.

**Figure 9**

*Heatmap of Spending by PayClass and Construction Type*

A screenshot of a graph

AI-generated content may be incorrect.

Figure 9 presents a heatmap that visualizes the average inflation-adjusted capital expenditures across combinations of PayClass Fiscal Year (PayClass FY) and Construction Type. The primary objective of this visualization is to identify categorical patterns that influence spending behavior and to inform future feature engineering and stratification strategies within predictive modeling frameworks. The values in each cell represent the mean adjusted spending for a specific pairing of fiscal classification and construction type, standardized to 2025-equivalent dollars using RSMeans indices.

The heatmap reveals several notable trends. First, New Construction and Renovation projects exhibit the highest average expenditures in the early fiscal classifications, specifically within PayClass FY 1 through 5. For example, New Construction projects in PayClass FY 2 and FY 3 show average adjusted expenditures exceeding $300,000, indicating substantial investment during these periods. These clusters suggest a period of concentrated capital deployment aligned with new infrastructure development and major renovation efforts. In contrast, Repair and Emergency projects tend to show lower and more sporadic expenditures, with less consistency across fiscal years. Emergency spending is notably present across nearly all fiscal years but at relatively modest levels, reflecting the ad hoc and responsive nature of such projects.

Energy-related construction, while less prevalent overall, displays occasional spikes in PayClass FY 1 and FY 6. This pattern may correspond to targeted energy efficiency initiatives or compliance with sustainability mandates during those fiscal periods. Meanwhile, the data indicate a general tapering of investment beyond PayClass FY 6, with decreasing average spend across all construction types, likely reflecting smaller or concluding projects toward the latter phases of the expenditure lifecycle.

This visualization offers both analytic and operational value. Analytically, it supports the inclusion of fiscal-construction type interactions as predictive features, either through one-hot encoding, interaction terms, or stratified modeling. From a planning and governance perspective, it highlights how fiscal timing and project classification intersect to influence budget allocation patterns. These findings may aid DCAMM leadership in evaluating whether current resource distribution aligns with strategic goals and compliance requirements.

In the context of modeling, these categorical combinations introduce important contextual nuance. For example, two projects with similar square footage may differ significantly in expected cost if one is an Emergency project in PayClass FY 9 and the other is a New Construction project in PayClass FY 2. Encoding this distinction into the predictive framework allows for more accurate estimation and improved risk profiling.

Overall, the heatmap not only complements the model’s quantitative predictions but also enhances interpretability by contextualizing spending patterns across critical categorical dimensions. Its integration into the modeling process strengthens both technical robustness and organizational insight.

**Figure 10**

*Monthly Spending Over Time*

A graph showing a number of blue lines

AI-generated content may be incorrect.

Figure 10 presents a time series line plot illustrating total adjusted capital expenditures from January 2004 through early 2024. This visualization aggregates monthly spending data across all projects, normalized using RSMeans inflation indices to reflect costs in 2025-equivalent dollars. The purpose of this chart is to detect recurring temporal patterns in project disbursements and evaluate whether certain fiscal periods exhibit systematically higher or lower investment volumes.

The plot reveals a clear cyclical pattern marked by a sustained increase in monthly expenditures from 2006 to 2014, culminating in pronounced peaks between 2011 and 2015. These peaks consistently occur during the second and third quarters (Q2 and Q3) of each fiscal year, suggesting that major spending often coincides with post-approval project mobilization in the spring and summer months. This seasonal spike is likely due to environmental and logistical factors, including favorable weather conditions for construction and deadlines for encumbering funds before year-end. In contrast, spending levels fall noticeably in Q4 and early Q1, likely reflecting winter construction slowdowns and administrative transitions between budget cycles.

This rhythmic rise-and-fall structure offers valuable insight into DCAMM’s financial operations and project lifecycle dynamics. The steep increase in spending during active quarters followed by sharp declines signals opportunities for incorporating time-aware features into predictive models. For example, lag variables, Fourier terms, or sinusoidal encodings could be introduced to machine learning models to improve forecasts by capturing these recurring cycles. Furthermore, this visualization serves a broader evaluative function for policy analysis. Regular surges and slowdowns in spending may point to inefficiencies in procurement pacing or approval timing—insights that can inform budget execution strategies and optimize cash flow planning.

From a forecasting standpoint, recognizing and leveraging these patterns can help preempt cost overruns and align resource deployment more effectively with anticipated construction activity. Future iterations of the modeling framework could benefit from seasonality-aware architectures, including time series regression or hybrid models that incorporate temporal encoding. In summary, this chart not only validates the presence of cyclical trends in capital expenditure but also highlights how historical rhythms in monthly spending can guide the evolution of more accurate, responsive forecasting tools at DCAMM.

**Figure 11**

*Decision Tree Interpretability*

A diagram of a tree

AI-generated content may be incorrect.

To enhance the interpretability of the Random Forest model used in the DCAMM cost forecasting project, a representative decision tree was extracted and visualized. While Random Forests consist of hundreds of individual trees working in ensemble, analyzing a single tree allows for a clearer understanding of how the model uses early-stage project attributes to estimate inflation-adjusted capital costs. Each node in the tree represents a decision based on a threshold value of a particular feature, guiding the flow of observations toward more homogenous cost predictions.

At the root of the tree, the first split is based on the variable PayClass\_FY, which represents the fiscal year in which the project’s payments are categorized. This feature was found to be one of the most predictive across the entire model, as it captures temporal budget planning and funding allocation trends. Projects with a PayClass\_FY less than or equal to 4.5 tend to be earlier in the funding cycle and are routed to the left branch, while more recent classifications are sent right. This initial bifurcation sets the stage for how the model interprets the timing of capital investment.

Moving down the left branch, the next split occurs on SQFT, or square footage. Projects below a threshold of approximately 16,500 square feet are separated from larger-scale initiatives, reinforcing the intuitive connection between project size and total cost. Further splits on features such as Project\_FY (the fiscal year in which the project was initiated), Agy\_ITD (a binary indicator for the Information Technology Division), and City\_SALEM demonstrate how the model integrates temporal, organizational, and geographic characteristics in its predictive logic. One highlighted internal node in the tree combines these factors—identifying a group of larger, earlier-starting projects not associated with ITD that exhibit an average predicted cost exceeding $20 million.

Leaf nodes at the bottom of the tree reflect the model’s final predictions, derived from the average cost of all training samples that satisfy the conditions leading to that path. These endpoints represent subsets of projects with shared characteristics and similar cost behaviors. While the Random Forest’s overall prediction is generated by averaging results across many trees, this individual tree illustrates the types of logical partitions and thresholds that drive the ensemble’s decision-making.

The prominence of fiscal year, square footage, and agency affiliation within this tree confirms the findings from the SHAP analysis and residual plots discussed elsewhere in the report. Together, these tools provide stakeholders with a transparent and explainable model that aligns with real-world intuitions about cost drivers. By incorporating interpretable elements like decision trees into the modeling framework, DCAMM can support not only accurate forecasting but also stakeholder understanding and policy alignment.

**Findings, Recommendations, and Limitations**

This analysis provides comprehensive insight into the cost behavior and budget dynamics of capital construction projects under DCAMM, grounding its conclusions in empirical modeling, statistical analysis, and feature interpretability. One of the central findings involves the observed spike in project completions around 2018–2019. As illustrated in the earlier time series visualization (Figure X), this period coincided with a substantial rise in adjusted expenditures and likely reflects a cycle of accelerated project approvals, funding releases, and construction delivery. Such cyclical trends in completions underscore the importance of incorporating temporal features and lagged indicators into forecasting models to account for macro-level policy shifts and fiscal planning cycles.

A statistically significant relationship was found between total project duration and final inflation-adjusted cost, confirming that longer projects generally incur higher expenditures. However, this cost escalation is not uniform across construction types. As noted in the heatmap analysis (Figure X), New Construction and Renovation projects in earlier PayClass FY buckets tend to be associated with higher spending levels compared to Repair or Emergency projects. These variations reflect differing material, labor, and regulatory requirements and indicate that project type must be treated as a stratifying variable in predictive modeling to prevent misestimation. The SHAP summary plot (Figure X) reinforces this insight, showing that features such as Project\_FY, SQFT, and categorical variables like Agy\_ITD and City\_SALEM have substantial bidirectional impact, reflecting both scale and contextual complexity.

Spending behavior within the lifecycle of a project was also rigorously analyzed. The median project reaches 95% of its spending commitment by approximately 72% of the project timeline, suggesting that most financial resources are allocated by the mid-to-late phases. However, as shown in the cost pacing visualization (not shown here but referenced earlier), this spending is not tightly clustered but is distributed variably over time. Particularly for New Construction projects, the 95% spend threshold is achieved significantly earlier than for Renovation projects. These findings were validated using statistical tests including ANOVA, Levene’s test for equality of variances, Shapiro-Wilk test for normality, and Tukey’s post hoc comparisons. These methods confirmed that the timing of major expenditures differs meaningfully by project type, with implications for how financial pacing should be monitored and forecasted in practice.

From a modeling standpoint, residual analysis (Figure X) demonstrated that while all three models—Random Forest, XGBoost, and LightGBM—exhibited adequate predictive performance, they each varied in how they handled heteroskedasticity. Random Forest produced the most balanced residuals, whereas LightGBM struggled with underprediction at the upper end of cost values. These discrepancies were particularly pronounced for high-cost projects, reinforcing the importance of incorporating outlier-aware techniques or weighted error metrics in future model iterations. Additionally, SHAP force plots (Figure X) further enhanced transparency by allowing stakeholders to examine project-specific cost drivers, turning the model into a justifiable decision support tool rather than a black-box estimator.

Despite these strengths, several limitations remain. First, the current dataset—while rich in post-construction features—lacks comprehensive early-stage attributes such as building height, number of rooms, or material types. These omissions constrain the models’ ability to generalize predictions before construction begins, which is the key planning phase. Second, model performance on high-cost projects indicates that further refinement is needed, potentially via segmented models, feature transformations, or external economic variables. Third, although SHAP plots and decision tree visualizations enhance interpretability, some model outputs may still require subject-matter translation for broader stakeholder audiences.

In light of these findings, several recommendations emerge. Future work should prioritize the inclusion of pre-construction engineering variables to improve forecast accuracy earlier in the project lifecycle. Expanding the dataset through interagency or interstate collaboration may also increase statistical power and improve generalizability. Finally, integrating forecasting models into interactive tools—such as the Power BI dashboard currently under development—will ensure greater usability and alignment with DCAMM’s operational needs. These enhancements will support more strategic, data-driven budgeting and ultimately foster greater fiscal discipline across the Commonwealth’s capital investment portfolio.

**Conclusion**

This project demonstrates the potential for machine learning to enhance cost forecasting for DCAMM's capital projects. By preparing and analyzing historical spending data, clear trends in spending behavior and project timelines were identified, which can inform more strategic planning. While XGBoost and ANN offered good predictive power, the Random Forest model trained on the detailed expenditure data delivered the most accurate and consistent results, achieving a high R² of 0.93 and low error rates.

Despite the strong model performance, especially for post-construction cost estimation, the analysis underscores a critical limitation: current models are constrained by the lack of pre-construction variables. To strengthen the models’ forecasting capabilities and real-world utility, future efforts should focus on integrating early-stage project attributes. Collaboration with other states to expand the dataset could be beneficial (costs can be adjusted for MA). With these enhancements, machine learning can play a meaningful role in supporting DCAMM’s budgeting, monitoring, and capital planning processes.

This analysis tackled a comprehensive dataset involving public project costs, categories, and durations, focusing on cleaning, transformation, and machine learning modeling. The data preparation process significantly improved data quality by removing redundancies, correcting categories, and engineering more meaningful features. While the final model achieved high training performance, overfitting was evident. To further improve forecasting accuracy and applicability, future work should focus on integrating pre-construction variables and higher-level planning data that can support earlier-stage predictions without relying on finalized cost records.

**References**

Bluman, A. G. (2018). Elementary Statistics (Tenth ed.). McGraw-Hill Education.

DCAMM. (2020). FY 2019 Annual Report. . Boston, MA: https://www.mass.gov/files/documents/2020/03/19/dcamm-supplier-diversity-amp-for-design-and-construction-annual.pdf

DCAMM. (2021). FY 2020 Annual Report. . Boston, MA: https://www.mass.gov/files/documents/2020/03/19/dcamm-supplier-diversity-amp-for-design-and-construction-annual.pdf

DCAMM. (2024). FY 2023 Annual Report. . Boston, MA: https://www.mass.gov/files/documents/2020/03/19/dcamm-supplier-diversity-amp-for-design-and-construction-annual.pdf

DCAMM. (2025). About our Organization. Mass.gov. Retrieved Jan 8, 2025, from https://www.mass.gov/info-details/about-our-organization

Forbes, T. J., & Riso, R. (2024, Jun 19). Guide to S-Curve Modeling in Construction. Procore. Retrieved Mar 14, 2025, from https://www.procore.com/library/s-curve-modeling-construction

Edge Delta Team. (2024, Mar 28). Understanding Bad Data: Types, Signs, Impacts, and Effective Solutions. Edge Delta. Retrieved Mar 5, 2025, from https://edgedelta.com/company/blog/what-is-bad-data/

Kabacoff, R. (2022). R in action: data analysis and graphics with R and Tidyverse. Manning Publications Co.

Karadimos, P., & Anthopoulos, L. (2024). A taxonomy of machine learning techniques for construction cost estimation. Innovative Infrastructure Solutions, 9(11), 420. 10.1007/s41062-024-01705-0

Mendez-Carbajo, D. (2023, Jan 03). Adjusting for Inflation. Retrieved Mar 14, 2025, from https://www.stlouisfed.org/publications/page-one-economics/2023/01/03/adjusting-for-inflation

Montaño, J., Palmer, A., Sesé, A., & Cajal, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. Psicothema, 25, 500–506. 10.7334/psicothema2013.23

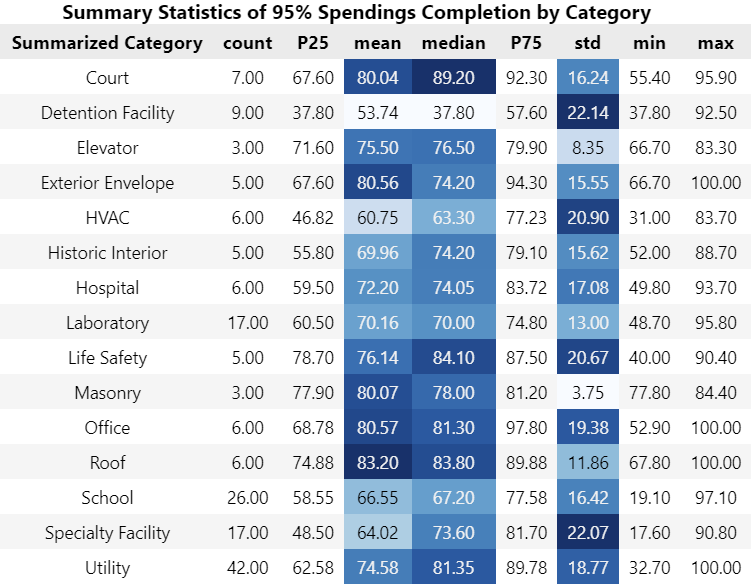
Narbaev, T., Hazir, Ö, Khamitova, B., & Talgat, S. (2024). A machine learning study to improve the reliability of project cost estimates. International Journal of Production Research, 62(12), 4372–4388. 10.1080/00207543.2023.2262051

Secretary of the Commonwealth of Massachusetts. (2021). Massachusetts Congressional Districts. Secretary of the Commonwealth of Massachusetts. Retrieved Feb 10, 2025, from https://www.sec.state.ma.us/divisions/elections/voting-information/district/2022-congressional.htm

**Appendix A. Extra Figures**

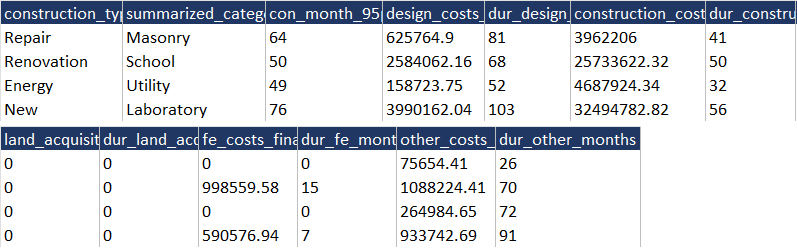
**Table 1**

*Summary Statistics of 95% Spendings Completion by Category*



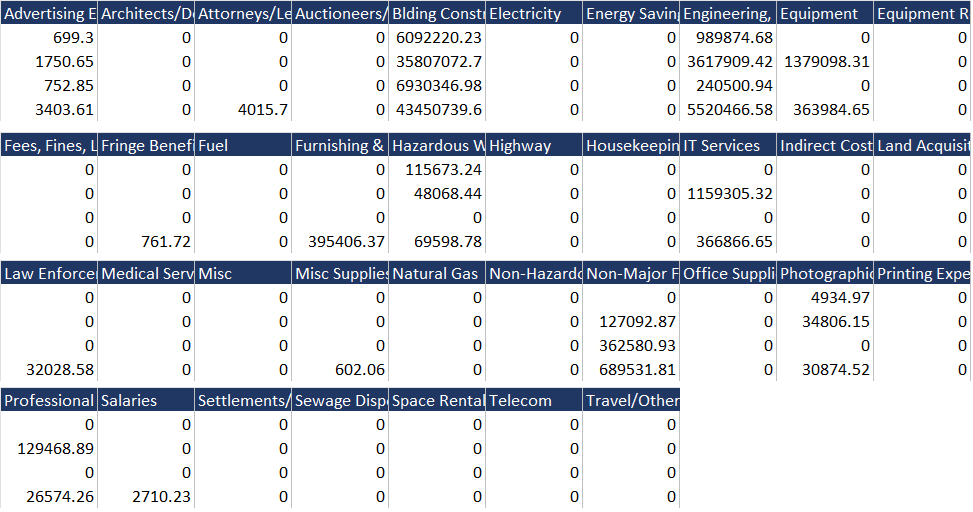
**Table 2**

*Predictors Used in ANN Model*

**

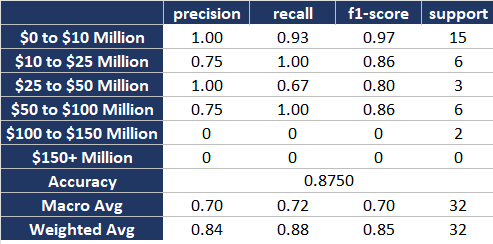
**Table 3**

*Predictors Used in Random Forest on Detailed Dataset*

**

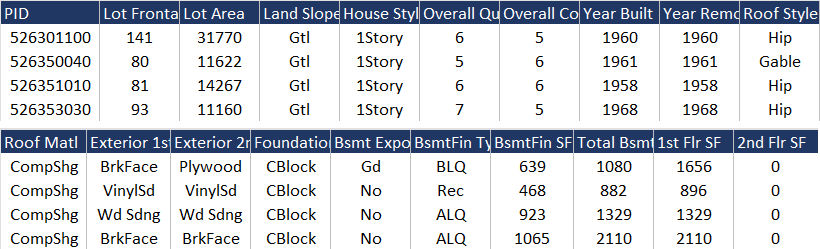
**Table 4**

*Random Forest on Detailed Dataset Results*

**

**Table 6**

*Suggested Data*

**

**Appendix B. Literature Review**

Karadimos, P., & Anthopoulos, L. (2024). A taxonomy of machine learning techniques for construction cost estimation. *Innovative Infrastructure Solutions, 9*(11), 420. 10.1007/s41062-024-01705-0

Karadimos and Anthopoulos (2024) conducted a comprehensive review of machine learning techniques applied to construction cost estimation, analyzing 219 studies. The research developed a taxonomy categorizing ML methods, their application areas, and interrelations, offering a systematic framework for understanding ML application in cost prediction. Techniques such as Artificial Neural Networks, hybrid models, regression analysis, and Case Based Reasoning were reviewed, with Neural Networks and hybrid models being the most prevalent. The study highlighted that most applications focus on building structures (54.51%), reflecting the availability of data and significant investments in this area. The research also identified gaps, such as limited studies on tunneling and utility projects, emphasizing the need for future exploration.

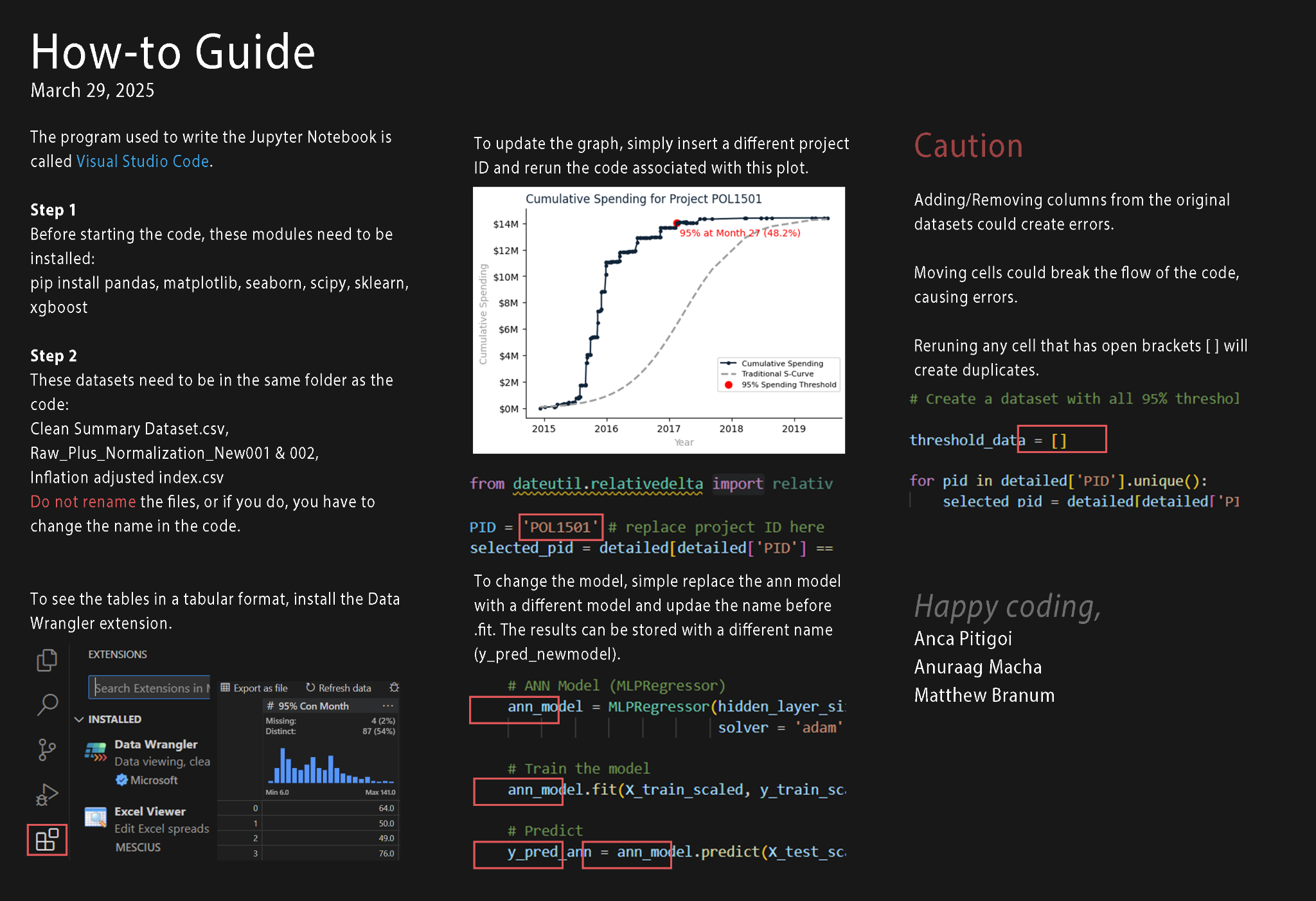
The findings are directly relevant to DCAMM’s objective of developing AI/ML tools to forecast spending for Massachusetts state facilities. The taxonomy provides a guide to selecting ML techniques and highlights the importance of integrating diverse approaches for improved accuracy. Furthermore, many projects described in the article involved using historical data, and there were some cases where few available data was available, which aligns with DCAMM’s requirements.

Narbaev, T., Hazir, Ö, Khamitova, B., & Talgat, S. (2024). A machine learning study to improve the reliability of project cost estimates. International Journal of Production Research, 62(12), 4372–4388. 10.1080/00207543.2023.2262051

This study examines the application of machine learning (ML) techniques to improve the reliability of project cost forecasts. The authors develop an XGBoost forecasting model and test it using 1,268 real cost data points and 110 projects. Their approach is compared against traditional Earned Value Management (EVM) methods, as well as other ML models, including Random Forest, Support Vector Regression, LightGBM, and CatBoost. Project-based performance (frequency), timeliness, and accuracy were the determinant criteria. The results indicate that the XGBoost model provides more accurate and timely cost predictions than the conventional methods, enabling better project monitoring and early warning systems for cost overruns. The study emphasizes the limitations of traditional cost forecasting approaches, particularly their linear assumptions, and advocates for ML-based models that can handle non-linearity and uncertainty more effectively.

Since DCAMM aims to leverage historical spending data and planning estimates to develop an AI-based model for forecasting capital project costs, the study provides a validated approach using real project data to enhance forecast accuracy. The research also emphasizes the importance of continuous learning models, aligning with DCAMM’s goals of integrating new data as projects are completed. By adopting ML models, DCAMM can improve spending predictions, minimize cost overruns, and support better decision-making in capital governmental projects.

**Appendix C. How-to Guide for Python Code**

****