**The Role of AI in Capital Asset Management at DCAMM**

Anuraag Macha, Anca Pitigoi, Matthew Branum

College of Professional Studies, Northeastern University

ALY 6080: Integrated Experiential Learning, Winter Term

Dr. Matthew Goodwin

March 29, 2025

**Abstract**

This study investigates the application of machine learning techniques to forecast total project costs for capital construction projects managed by the Division of Capital Asset Management and Maintenance (DCAMM). Historical spending data from two large datasets were cleaned, merged, and analyzed to examine trends in expenditure behavior and project characteristics. Results indicate a peak in project completions during 2018–2019, coinciding with increased overall expenditures. A positive correlation was observed between project duration and total cost, with variation in cost escalation patterns across construction types. Most projects exhibited spending concentrated around the middle phase, with a median 95% spending completion rate of approximately 72%. New construction projects reached this threshold significantly earlier than renovations, as confirmed through ANOVA and supporting statistical tests.

Multiple machine learning models were developed and evaluated, including Linear Regression, XGBoost, Artificial Neural Networks (ANN), and Random Forest. The Random Forest model trained on detailed expenditure data achieved the highest predictive performance, with an R² score of 0.93. Feature importance analysis identified key cost drivers such as Building Construction, Engineering Services, and Advertising Expenses. While model accuracy was high, reliance on detailed cost inputs limits applicability in early-stage forecasting. The results demonstrate the potential of machine learning for modeling cost behavior in capital project data.

*Keywords:* capital projects, construction projects, cost forecasting, costs overruns, capital projects cost prediction, XGBoost, artificial neural networks, random forest, historical data analysis, operational efficiency, capital project management, machine learning models, data mining, design costs, infrastructure planning, Python

**The Role of AI in Capital Asset Management at DCAMM**

The Division of Capital Asset Management and Maintenance (DCAMM) seeks to improve spending forecasts for state facility projects using historical data and AI/ML models (DCAMM, 2025). This project begins by outlining the business problem, providing a clear context for the analysis. It then proceeds with a comprehensive literature review to establish a theoretical foundation and identify relevant research. Following this, an in-depth exploratory data analysis (EDA) is conducted, detailing each step taken to understand and prepare the data. The focus then shifts to the application of analytical tools and modeling techniques to derive insights. Finally, the project concludes with a presentation of the results, actionable recommendations, and a discussion of the limitations encountered.

**Executive summary**

DCAMM is seeking to leverage artificial intelligence and machine learning models to improve cost forecasting for its capital projects. This project addressed this objective by cleaning and analyzing two large datasets of historical spending to identify key cost trends and features, engineer inflation-adjusted expenses, and evaluate predictive models such as linear regression, Artificial Neural Networks (ANN), and Random Forests. The data revealed that most DCAMM projects are renovations, and that new construction projects tend to incur costs earlier in the timeline. Spending behavior varied across construction types and categories, with HVAC and detention facility projects often spending most of their budgets early, while office and roof projects showed steadier pacing. A significant finding from an ANOVA analysis showed that new projects reach 95% of spending significantly earlier than renovations, offering potential strategic insights for future budgeting and monitoring.

Model evaluation showed that while ANN and Random Forest models achieved high R² scores, they largely learned mathematical relationships from post-construction data rather than predictive patterns. Because the model relies on finalized cost data, its use for early-stage forecasting is limited. To unlock greater forecasting potential, it is recommended that DCAMM expands its dataset to include pre-construction variables such as building specifications (number of levels, height of the building, square footage of built/renovated projects, special equipment needed, etc). Incorporating these attributes will enable more proactive and data-informed decision-making across the capital project lifecycle.

**Business Problem**

Managing large-scale public construction projects requires careful budgeting, cost forecasting, and resource allocation to ensure timely completion while minimizing financial risks. Over time, inflation, labor costs, and material price fluctuations have significantly impacted the total cost of these projects. Without accurate cost adjustments, organizations risk underestimating expenses, leading to budget overruns, project delays, or reduced scope. This issue is particularly relevant to DCAMM, which oversees numerous infrastructure projects and must balance new developments with ongoing renovations (Karadimos & Anthopoulos, 2024; Narbaev *et al.*, 2024).

The goal of the project is to develop an AI forecasting model to improve cost predictions for capital projects managed by DCAMM. Machine learning techniques will be used, including linear regression, XGBoost, Random Forest, and Artificial Neural Networks (ANN). The chosen model would continuously integrate new data to improve predictions over time.

The objectives of the project are to find the most influential features for the project costs, and implement and compare multiple forecasting models to determine the most accurate and reliable approach. The forecasting models should improve the baseline methods used by DCAMM with at least 15-20% improved accuracy. The models are evaluated for performance using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and (Coefficient of Determination) R² to ensure accuracy and generalizability. The objective is to achieve a MAE within 10-20% of the actual project cost. RMSE should be as low as possible, but since it penalizes large errors, comparing it to MAE helps determine if outliers are a problem. For capital project cost forecasting, a MAPE of 10-20% would be a strong target, considering construction costs can be unpredictable due to external factors like inflation, material costs, and labor fluctuations. If the model significantly outperforms traditional cost estimation methods, even MAPE under 25% could still be useful (Montaño *et al.*, 2013). The last benchmark, R² for the model should be between 0.7-0.9 to be considered as having good predictive power (Kabacoff, 2022). Lastly, this project will provide clear documentation for DCAMM to use and maintain the model effectively.

**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**

Background research and literature review in government capital construction focus on identifying best practices, assessing historical project outcomes, and understanding policy, regulatory, and procurement frameworks. Studies often highlight the importance of cost estimation accuracy, risk management, stakeholder engagement, and sustainability in public infrastructure projects (See Appendix A). Reviewing past government-led construction initiatives provides valuable insights into common challenges such as delays, budget overruns, and compliance issues, which inform the planning and execution of future capital projects.

**Exploratory Data Analysis**

Data cleaning is an important step in machine learning, ensuring that models are trained on accurate, consistent, and relevant data. Many analysts and data scientists often underestimate the importance of getting acquainted with their data, rushing into model building without thoroughly exploring and cleaning their dataset. For instance, a data entry error at Samsung Securities caused employees to mistakenly sell 5 million shares worth $187 million in just 30 minutes, leading to massive financial losses. Similarly, NASA's Mars Climate Orbiter disaster occurred due to a mix-up between metric and imperial units, resulting in the spacecraft burning up in Mars' atmosphere and a $125 million loss (Edge Delta Team, 2024).

There were two main datasets provided by the organization. The summary dataset contains aggregated costs by pay class (such as construction, land acquisition, f&e, etc.), durations by pay class and location of the projects. The detailed dataset contains cash expenses for each project down to the last cent.

Starting with the summary dataset, the following cleaning tasks were performed to prepare for machine learning methods:

* Feature names standardized to align with common practices.
* Removed features with no values and with only one unique value.
* Removed dollar signs from numerical columns.
* Converted the data to proper data type.
* Cities and counties were filled based on project descriptions, DCAMM reports, and Massachusetts Congressional Districts list (for instance, rows 18 & 97 were updated based on project ID and 2020 DCAMM report (DCAMM, 2021; Secretary of the Commonwealth of Massachusetts, 2021).
* Feature engineering was performed on bucket\_category to form broader, more generalized categories that better represent the overall project type.
* The total duration of the project in months was summed up across the duration of the pay class.

The detailed dataset had the following cleaning actions performed, mainly based on whether the feature contained useful information for cost predictions:

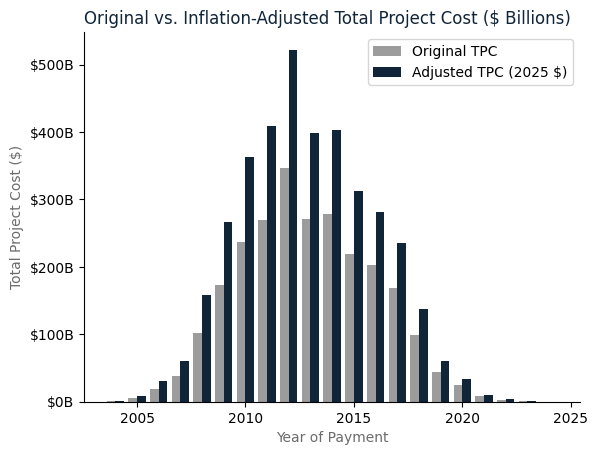
* Dropped features with empty columns, limited information, or low variability (cabinet\_name, department\_name, government\_branch\_name, etc.).
* Removed non-predictive payment document-related features (doc\_category, etc.), reference, and fund-related (funding source is generally not relevant for cost prediction).
* Introduced city feature from the summary dataset based on project ID to perform inflation adjustment.

Feature engineering included adjusting costs based on the RSMeans City Cost Index and instructions provided by DCAMM (Mendez-Carbajo, 2023). Because each city has a different inflation index, the city from the summary dataset was introduced in the detailed dataset. Since not all cities have an index in the table provided by DCAMM, the average for that year for the cities that do not have the corresponding index was used. Moreover, according to the instructions, if the payment was made between Jan-Jun previous year index is used, if July-Dec next year’s. The new expenses feature is called adjusted\_cash\_expense, and as a result the new total project cost was calculated.

The bar chart (Figure 1) compares the original total project cost (TPC) with the inflation-adjusted TPC (2025 dollars) over time. The data shows that project costs peaked between 2010 and 2015, with inflation-adjusted values consistently higher than the original costs. The gap between the bars highlights how much the purchasing power has changed.

**Figure 1**

*Original cost vs Inflation-Adjusted Total Project Costs*



***How many projects does DCAMM complete per year?***

While the inflation-adjusted total project cost highlights the increasing financial demands of construction projects over time, it is equally important to examine how project completions have fluctuated. Figure 2 presents the number of projects completed per year, revealing a steady rise in completions leading up to 2019, followed by a decline in recent years. The peak in project completions around 2018–2019 suggests a period of high construction activity, aligning with the increased total project expenditures observed in Figure 1. However, the subsequent downturn in completions may reflect broader economic constraints, shifts in funding allocations, or construction slowdowns. Understanding these trends provides valuable insight into the relationship between cost escalation and project delivery over time.

**Figure 2**

*Number of Projects Completed Per Year*

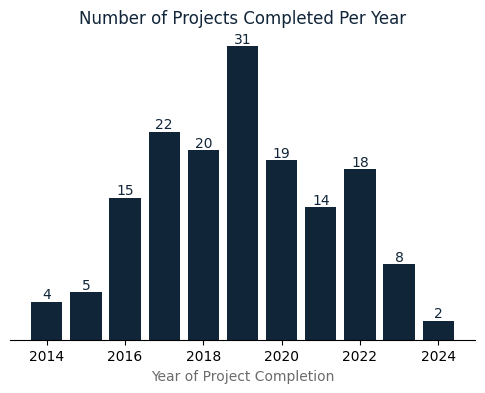
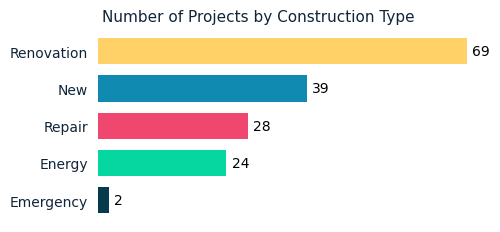


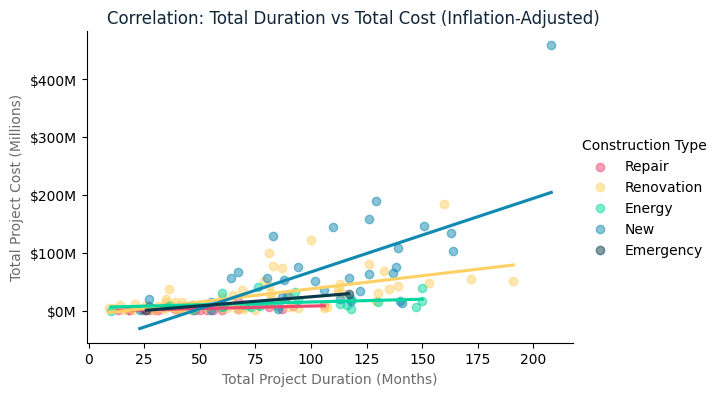
Figure 3 is based on the summary dataset and it shows that most of the projects performed by DCAMM are renovations, followed by new projects. Notably, only two emergency projects were recorded during 2004 - 2024, which indicates that the organization prioritizes proactive maintenance and long-term planning, minimizing the need for urgent interventions.

**Figure 3**

*Number of Projects by Construction Type*

Overall, there is a positive relationship between total project duration and total project cost (See Figure 4). The slopes of the regression lines differ noticeably by construction type, indicating that some project categories see a faster rise in total cost as duration increases. A steeper slope (New projects) means costs escalate more sharply with each additional month, while a flatter slope (Energy projects) suggests costs grow more gradually over time. It is good to mention that the category "Emergency" contains only two projects, therefore the regression line is unreliable and does not represent any underlying trend. From the same graph it also can be seen that a potential outlier is present (blue new project).

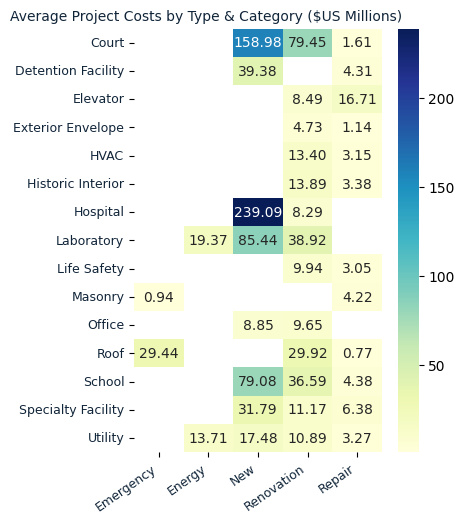
**Figure 4**

*Correlation: Total Duration vs Cost (Inflation Adjusted)*

After further investigation of the potential outliers, the results show that among the top three most expensive projects, the highest cost belongs to a hospital. This isn't surprising or an error; hospital projects are usually very costly because they require special designs, equipment, and safety standards. Meanwhile, the other two expensive projects, though still high in cost, do not reach the specialized expense level of hospital construction. This differentiation emphasizes the importance of considering project type when interpreting cost data, rather than relying solely on numerical thresholds for outlier detection.

**Figure 5**

*Average Project Costs by Type & Category ($US Millions)*

**

The heatmap (Figure 5) shows the average spending (in $US millions) for each combination of construction type and summarized category. Darker cells indicate higher costs. It can be seen that new projects cost on average more than repairs and renovations. With the majority of projects being renovations, a better glimpse into the averages is seen. Surprisingly, the emergency projects do not seem as costly compared to other categories of roof projects.

Furthermore, it seems like the elevator repair cost twice as much as the renovation, perhaps because this elevator was located in a historic building. It is recommended to be replaced with a more modern elevator that does not need specialized services, if the law allows.

***How many vendors on average per project exist? Is one vendor more prevalent?***

On average, each project involves approximately 17 different vendors, reflecting the collaborative and multidisciplinary nature of DCAMM’s capital projects. Among the vendors, a few stand out due to their repeated engagement across numerous projects. Boston Herald Inc is the most frequently involved, having contributed to 57 distinct projects. Creative Office Interiors Inc follows closely with participation in 43 projects, demonstrating consistent involvement in furnishing or interior solutions. W.B. Mason Co also ranks among the top vendors with 42 projects completed.

**Spending Analysis**

A deeper examination of project-level financial pacing was conducted to understand when DCAMM projects reach 95% of total expenditures, a key marker of budget usage. This threshold helps uncover trends in how quickly or gradually spending unfolds throughout a project timeline.

**Figure 6**

*Cumulative Spending Comparison*

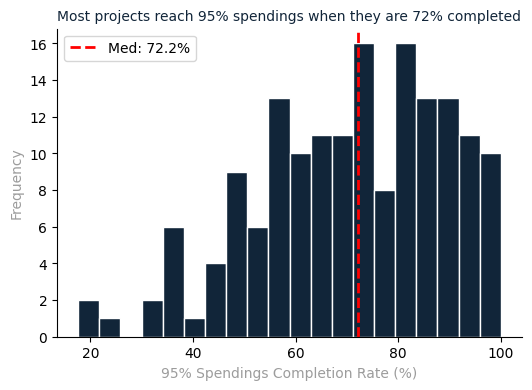


As shown in Figure 6, after looking at a couple of projects, it was noticed that some projects do not follow the traditional S-curve associated with construction spendings (Forbes & Riso, 2024). For instance, project S992 exhibited a sharp increase in spending early on, reaching the 95% spending threshold at just 18% of the total project duration. After reviewing the costs, the gym for the police officers acquired most of the equipment in the beginning of the project. In contrast, project POL1501 followed a more gradual spending trajectory, with 95% of expenditures occurring at 48% of the project timeline, aligning more closely with expected spending distributions, but still not the gradual increase as expected.

***When does DCAMM spend 95% of the funds on average?***

The histogram of 95% Spending Completion Rate (Figure 7) reveals several key insights about project spending patterns. The median completion rate is around 72% (marked by the red dashed line), indicating that many projects spend the funds well before 95% of the project is **Figure 7**

*Histogram of 95% Spendings Completion Rate*



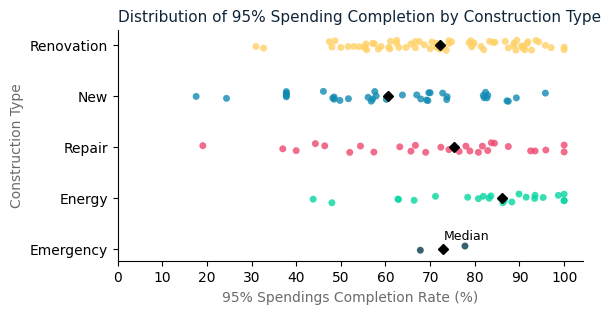
completed. The distribution is fairly wide, ranging from 20% to nearly 100%, with a notable concentration of projects between 60% and 90%, suggesting that while most projects progress well, a significant portion stall before reaching full completion. There is a clear peak around 80%, indicating a common stopping point. Additionally, a smaller but important subset of projects remains at very low completion rates (below 40%), which could point to funding issues, cancellations, project delays, or project nature - as seen in the case of the police gym.

***Do the spending habits differ between construction types?***

The distribution of 95% spending completion rates by construction type (Figure 8) includes black diamond markers indicating the median for each group. The plot reveals clear differences in spending behaviour across project types.

**Figure 8**

*Distribution of 95% Spending Completion by Construction Type*

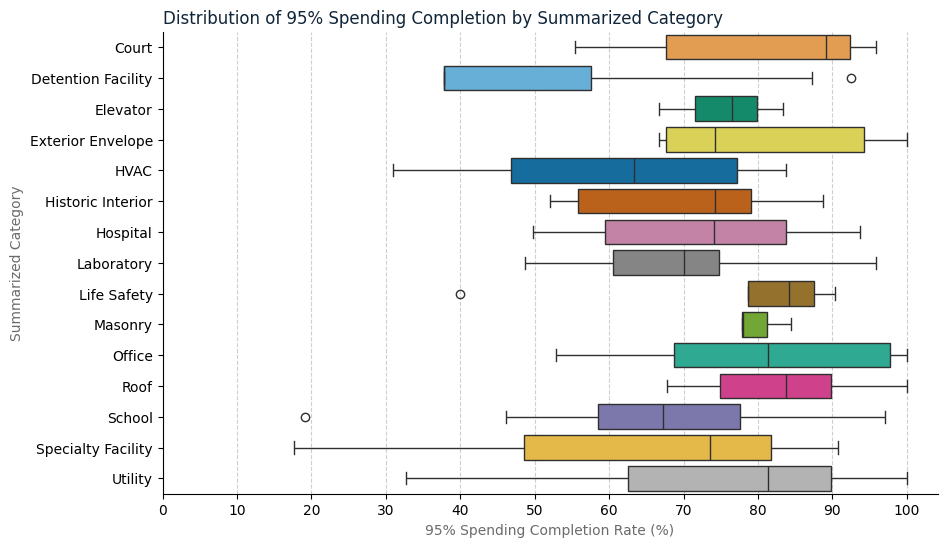


Renovation projects show a relatively tight clustering in the upper range, with a median close to 70%, suggesting that spending is spread steadily across the project lifecycle. New and repair projects display more variation. The new projects median falls earlier, at about 60%, indicating that these projects tend to spend most of their budget earlier. The repair projects follow a similar pattern to renovations, but with a median completion rate of around 75%. The energy projects show a slower burn rate, with the median around 85%. Lastly, there are only two emergency projects, so a statement should not be drawn.

Overall, the graph suggests that construction type significantly influences the timing of 95% spendings, with new construction projects dispensing the budget sooner than other types. The diamond marker helps highlight central tendencies, serving as a useful reference point for planning and forecasting.

Based on the summary statistics (Table 1, Appendix A) and Figure 9, an overview about how different project categories behave in terms of reaching 95% of their total spending can be seen.

**Figure 9**

*Distribution of 95% Spending Completion by Summarized Category*

Detention facility projects stand out with the lowest median (37.8%), lowest mean (53.74%), and high standard deviation, indicating these projects often reach 95% spendings, and there is large variation between the projects. HVAC projects also lean earlier, with relatively low 25th percentile (46.82%), and median (63.3%), suggesting that these projects spend heavily up front. This makes sense for HVAC projects, because the equipment could be ordered and partially paid in advance. Because of the high standard deviation for both of these categories (20+), it can be drawn that the spending patterns are inconsistent, likely due to differences in project scope, urgency, or timeline.

Roof, Office, and Court projects show high means and median (80%+), implying that spending is distributed more evenly across the project timeline.

Mansory shows extremely low variability (std = 3.75), with close percentiles, which indicates consistent and phased delivery approach.

Outliers are also evident in several categories. For example, Detention Facilities, Life Safety, and School projects each contain projects that reached 95% spending either unusually early or late compared to the norm. These outliers may reflect unique funding strategies, project delays, or unusual scopes of work.

This visualization not only helps highlight typical spending behavior within each category but also reveals which types of projects tend to be more predictable versus those that are more variable or exceptional in their spending patterns. Such insights can inform planning, budgeting, and risk assessment across future capital projects.

All in all, it begs the question, why do some of the projects take so long to pay the remaining 5% after the 95% spending threshold was reached so early in the project?

**Analytical tools and models**

The main tool used in this analysis is Python. Diverse modules are used to achieve different visualizations or ML models, such as pandas, matplotlib, seaborn, sklearn, torch, and keras. Moreover, ANOVA is used to identify if there is a statistical difference between construction types. Lastly, Tableau or PowerBI will be utilized to present the findings.

In terms of analytical models, a linear regression model was attempted before any other advanced models were performed. However, according to Karadimos and Anthopoulos (2024), when working with construction projects that have limited historical data, Artificial Neural Networks (ANNs) and hybrid models demonstrate high accuracy. Moreover, in another study performed by (Narbaev et al., 2024) XGBoost forecasting model combined with the traditional Earned Value Management (EVM) method also yielded good accuracy levels.

***Does the mean 95% spending completion rates differ significantly between Renovation, New, and Repair projects?***

In order to answer this question, hypothesis testing with ANOVA will be performed. The independence, normality, and homogeneity of variances assumptions need to be checked before applying ANOVA (Bluman, 2018; Kabacoff, 2022).

Since each project is independent and does not influence others, the assumption of independence is satisfied. The Shapiro-Wilk test was used to assess normality, and the results indicated that the data is normally distributed. Additionally, Levene’s test confirmed the homogeneity of variances across groups. With these assumptions met, the ANOVA test can be appropriately conducted.

Null Hypothesis (H0): There is no significant difference in the mean 95% spending completion rates among Renovation, New, and Repair projects.

Alternative Hypothesis (H1): At least one group's mean 95% spending completion rate is significantly different from the others.

If p < 0.05, there is a statistically significant difference in the mean 95% completion rates across the three types. Because the p-value = 0.0094, which is less than 0.05, the null hypothesis is rejected. Therefore, there is a statistically significant difference in the mean 95% spending completion rates between at least one pair of project types.

***Which pairs of the project types differ in their mean?***

To better understand how spending patterns differ between types of projects, the average point in the project timeline when 95% of funds are spent was compared across Renovation, New, and Repair projects. Using Tukey HSD statistical test designed to compare groups, the following were found:

* New projects tend to spend their budgets much earlier than renovation projects (p<0.05). The difference is statistically significant, meaning it's unlikely to be due to random chance. New projects reach 95% of their spending about 11 percentage points earlier than renovation projects.
* While new projects also reach 95% spending earlier than repair projects, the difference is not strong enough to be considered statistically significant, so it cannot be said for sure that these two groups are meaningfully different.
* Renovation and repair projects have very similar patterns, with no meaningful difference in how quickly they reach 95% of spending.

***Artificial Neural Networks (ANN)***

The ANN models were performed only on the summary dataset. The predictors can be seen in Table 2, Appendix A.

Several iterations of the Artificial Neural Network (ANN) model were conducted to identify the most relevant features and optimize performance. Initially, the model underperformed, likely due to limited training data (70%) and insufficient epochs (1000 iterations). After increasing to 80% training and 10,000 iterations, performance improved significantly (RMSE = 0.491, R² = 0.70). However, due to the high number of one-hot encoded features, the adjusted R² was not applicable (Kabacoff, 2022).

Subsequent refinements showed that removing city, county, and phase improved the model significantly (RMSE = 0.302 and R² = 0.89), suggesting that these variables contributed little predictive power. When all character variables were removed, performance declined slightly (RMSE = 0.5816, R² = 0.58), suggesting that the one-hot encoded features contributed limited predictive value. These variables may have introduced more complexity than signal, making them less effective for this model. However, when all breakout costs (e.g., construction, land acquisition, fees, and other costs) were removed, model performance dropped drastically (RMSE = 0.7417 and R² = 0.31), confirming that these cost components were critical to the ANN's learning process.

Ultimately, the analysis revealed that the model was not truly predictive but rather learning an existing mathematical relationship, since total project cost is already the sum of its components. To develop a genuinely predictive model, it is recommended to use pre-project variables, such as square footage of built/renovated projects, to forecast costs rather than relying on post-construction financial breakdowns.

***Random Forest on Summary Dataset***

The Random Forest Regression model was performed on the summary dataset. The predictors can be seen in Table 2, Appendix A.

The Random Forest Regression model developed to predict total project costs demonstrates solid predictive power, with an R² score of 0.79, indicating that the model explains approximately 79% of the variance in project costs. The model's performance is quantified through key metrics: a Mean Absolute Error of 0.22 and a Root Mean Squared Error of 0.29. These figures suggest that while the model provides robust predictions, there remains some variability in cost estimations, which is not unexpected given the complex nature of project budgeting.

The feature importance analysis reveals a hierarchy of predictors, with construction\_costs\_final emerging as the most dominant factor, accounting for 32.6% of the model's predictive power. Construction-related features, including construction costs at various stages and final costs, collectively contribute significant explanatory value. Design-related metrics such as design duration, design costs, and design end dates also play meaningful roles in cost prediction. This suggests that project managers and financial planners should carefully track both construction and design-related expenses throughout a project's lifecycle.

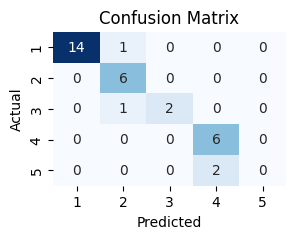
The learning curve analysis provides additional confidence in the model's reliability, showing that the training and cross-validation scores converge as the training dataset increases. This convergence indicates that the model generalizes well and is not prone to overfitting, a common concern in complex predictive modeling (Kabacoff, 2022). The model's ability to maintain consistent performance across different dataset sizes suggests it could be a valuable tool for cost estimation across various project scales and types. Practitioners in project management and financial planning could leverage this model to develop more accurate budget forecasts, with the understanding that while highly predictive, the model still requires careful interpretation and domain-specific expertise.

***Random Forest on Detailed Dataset***

Another Random Forest was performed, but on a different dataset (See Table 3, Appendix A). The dataset used for this Random Forest model is a project-level summary where each row represents a distinct project and each column corresponds to a specific type of project charge based on the object\_name (e.g., engineering services, equipment, salaries). These columns contain the actual amounts spent on each charge category, aggregated from detailed expense records. The target variable, tpc\_category, indicates the tier of project expenses ($0 - $10 Million).

**Figure 10**

*Confusion Matrix for Random Forest Model (Detailed Dataset)*

**

Based on the results in Table 4 (Appendix A) and the confusion matrix (Figure 10), the accuracy of this model is 87.5%, which is considered pretty good. Other few insights can be drawn:

* $0 to $10 Million (1): Very strong performance - precision = 1.00, recall = 0.93. Only one project was misclassified as category 2.
* $10 to $25 Million (2): Perfect recall (1.00), which means all actual class 2 were correctly identified, though one class 1 was wrongly predicted as class 2.
* $25 to $50 Million (3): Precision = 1.00, but recall = 0.67- the model missed one actual class 3 and predicted it as class 2.
* $50 to $100 Million (4): Great recall (1.00), precision = 0.75 - one project from class 3 was wrongly classified here.
* $100 to $150 Million (5): Model failed to predict this class (precision and recall = 0) - both projects in this category were classified as class 4. This is due to very few training samples in this range.

It is important to note that the $150+ million category was not predicted, as it contained only a single project in the randomly selected dataset, which resulted in it being included in the training set.

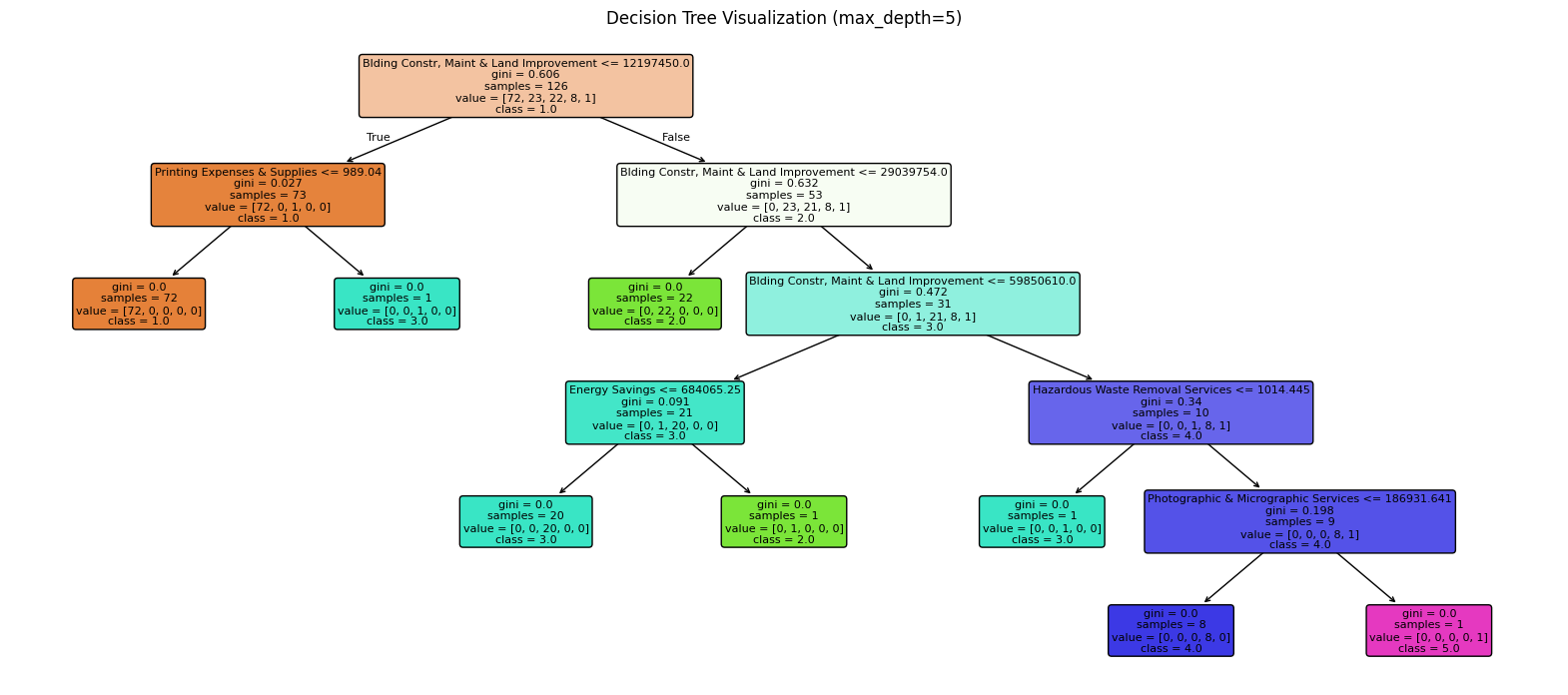
An interesting way to understand the Random Forest results is to look at a particular decision tree composition (See Figure 11). At its core, the model evaluates each project by sequentially assessing whether specific types of expenses exceed or fall below certain threshold values.

The decision process begins at the top node, where the model first considers whether the project’s spending on “Building Construction, Maintenance & Land Improvement” is less than approximately $12.2 million. Projects falling below this threshold are overwhelmingly classified into the lowest cost category, $0 to $10 million, indicating that this expense category is a strong initial indicator of project scale.

For projects exceeding that amount, the tree branches further, incorporating additional spending categories to refine its classification. For example, the model evaluates expenditures such as “Energy Savings,” “Hazardous Waste Removal Services,” and “Photographic & Micrographic Services” to distinguish between mid- and high-cost projects. These deeper nodes help the model differentiate more nuanced cases, such as distinguishing between projects in the $25 to $50 million and $50 to $100 million ranges, or identifying rare high-value projects in the $100 to $150 million category (Kabacoff, 2022).

Each terminal node (leaf) of the tree represents a final decision about a project’s TPC category. These decisions are based on the historical distribution of similar projects and their associated expenditures. The decision tree thus serves not only as a predictive model but also as an interpretable framework for understanding how different combinations of expenses contribute to overall project cost classification.

**Figure 11**

*One Decision Tree from Random Forest Model*

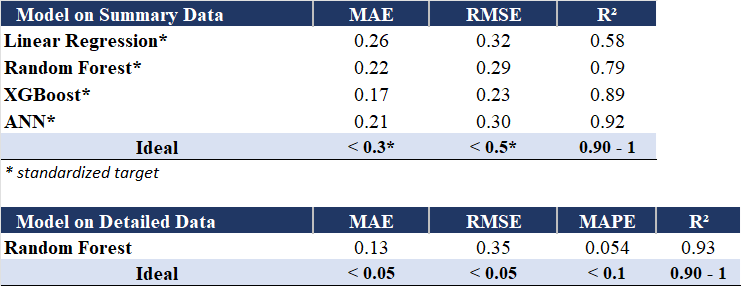
Lastly, the most influential factor in determining a project's cost category is Building Construction, Maintenance & Land Improvement, accounting for 27% of the model's decision-making. This is followed by Engineering, Research & Scientific Services (16%) and Advertising Expenses (8%), with smaller yet meaningful contributions from Hazardous Waste Removal Services (7%) and Photographic & Micrographic Services (6%). These features collectively reflect key spending areas that distinguish projects of varying scales.

**Findings, Recommendations, and Limitations**

This analysis offers valuable insights into the patterns and drivers of capital project spending and provides a strong foundation for predictive modeling at DCAMM. The observed peak in project completions around 2018–2019 aligns with a period of increased construction activity and total expenditures, highlighting a potential cycle of intensified project delivery. A clear positive relationship between project duration and total cost was identified, though the rate of cost increase varies noticeably across construction types, suggesting that certain project categories escalate in cost more rapidly over time.

Spending behavior across projects was also explored in depth. The median 95% spending completion rate is approximately 72%, closely aligned with the average, indicating that most project funds are committed by the mid-to-late stages. However, spending is not concentrated at a single point but rather distributed over the course of the project. Importantly, new construction projects tend to reach 95% of spending significantly earlier than renovations, a difference supported by statistical testing using ANOVA, Levene's test, Shapiro-Wilk, and Tukey's post hoc analysis. These findings emphasize the influence of construction type on financial pacing and offer actionable insights for budget planning and oversight.

**Table 5**

***Models Results Comparison***

This analysis compared predictive models using two types of datasets: one with summarized expense categories and another with highly detailed financial data. The summary dataset aggregated project expenses into broader categories such as construction, design, land acquisition, and furniture & equipment. Because these categories varied significantly in scale, the target variable was standardized to make patterns more comparable. Despite this transformation, models such as XGBoost and Artificial Neural Networks (ANN) performed well, demonstrating strong predictive accuracy (See Table 5).

The detailed dataset, in contrast, retained expense information down to the cent and categorized costs by specific charge types such as advertising and hazardous waste removal. The Random Forest model trained on this data showed the highest performance, with strong accuracy metrics even without standardizing the target. However, this result should be interpreted with caution. Many of the detailed charges directly contribute to the overall project cost, which raises the possibility of overfitting or data leakage - where the model learns to “predict” the target using inputs that already contain it. While the model performs well on this dataset, its ability to generalize to new, unseen data may be limited unless the same level of detailed cost information is available in advance.

The project faces risks related to data quality and availability, given the limited dataset of 160 completed projects over a decade. Moreover, the features included in the dataset are post-completion costs. It is recommended to add features such as, number of rooms, bathrooms, height of the building, square footage of built/renovated projects, special equipment needed, specialty rooms, cost of materials, type of materials, and so on. Housing prices might offer a glimpse into what predictors are used to foresee the cost of the house (See Table 6, Appendix A). After the new data collection is achieved, the models could be re-ran.

Lastly, integration and usability pose risks; some models are too complex or not easily interpretable, therefore may not be adopted. To address this, compensate for the model overfitting issue, and support more robust, practical decision-making, interactive dashboards are being developed using Power BI. The dashboard offers information on spending habits. In addition, a knowledge transfer plan will be implemented to ensure smooth adoption and effective use of these tools by the DCAMM team.

**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

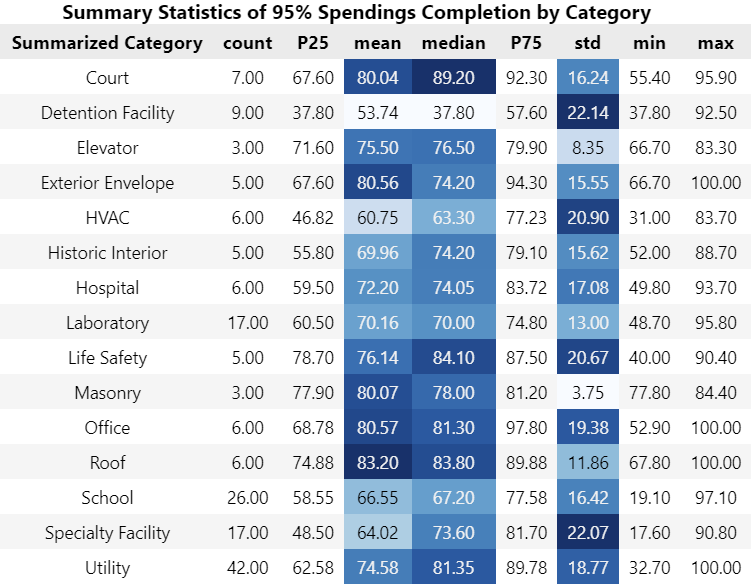
Narbaev, T., Hazir, Ö, Khamitova, B., & Talgat, S. (2024). A machine learning study to improve the reliability of project cost estimates. International Journal of Production Research, 62(12), 4372–4388. 10.1080/00207543.2023.2262051

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**

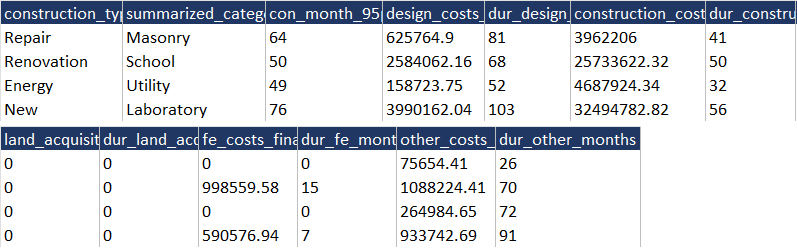
**Table 1**

*Summary Statistics of 95% Spendings Completion by Category*



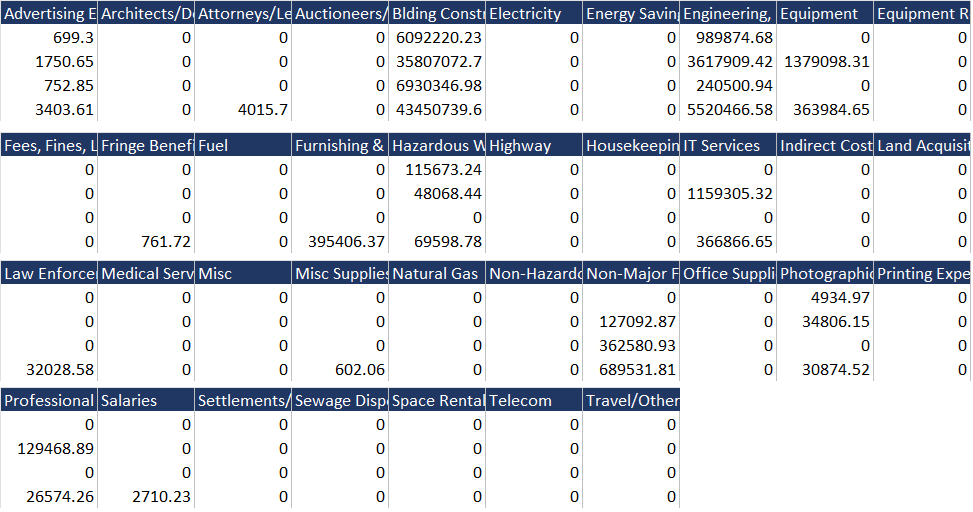
**Table 2**

*Predictors Used in ANN Model*

**

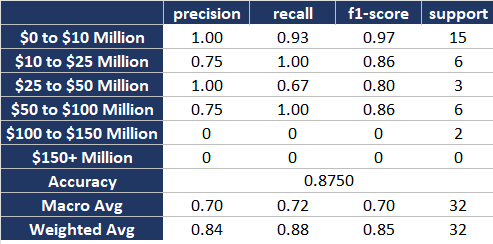
**Table 3**

*Predictors Used in Random Forest on Detailed Dataset*

**

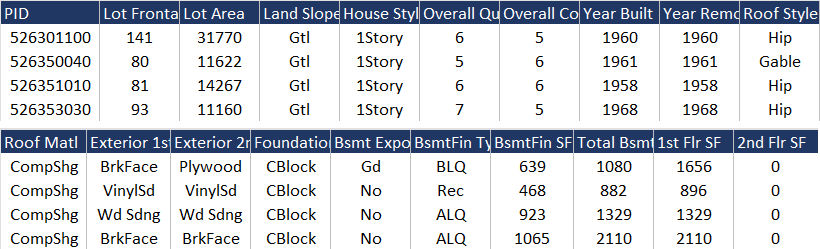
**Table 4**

*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.

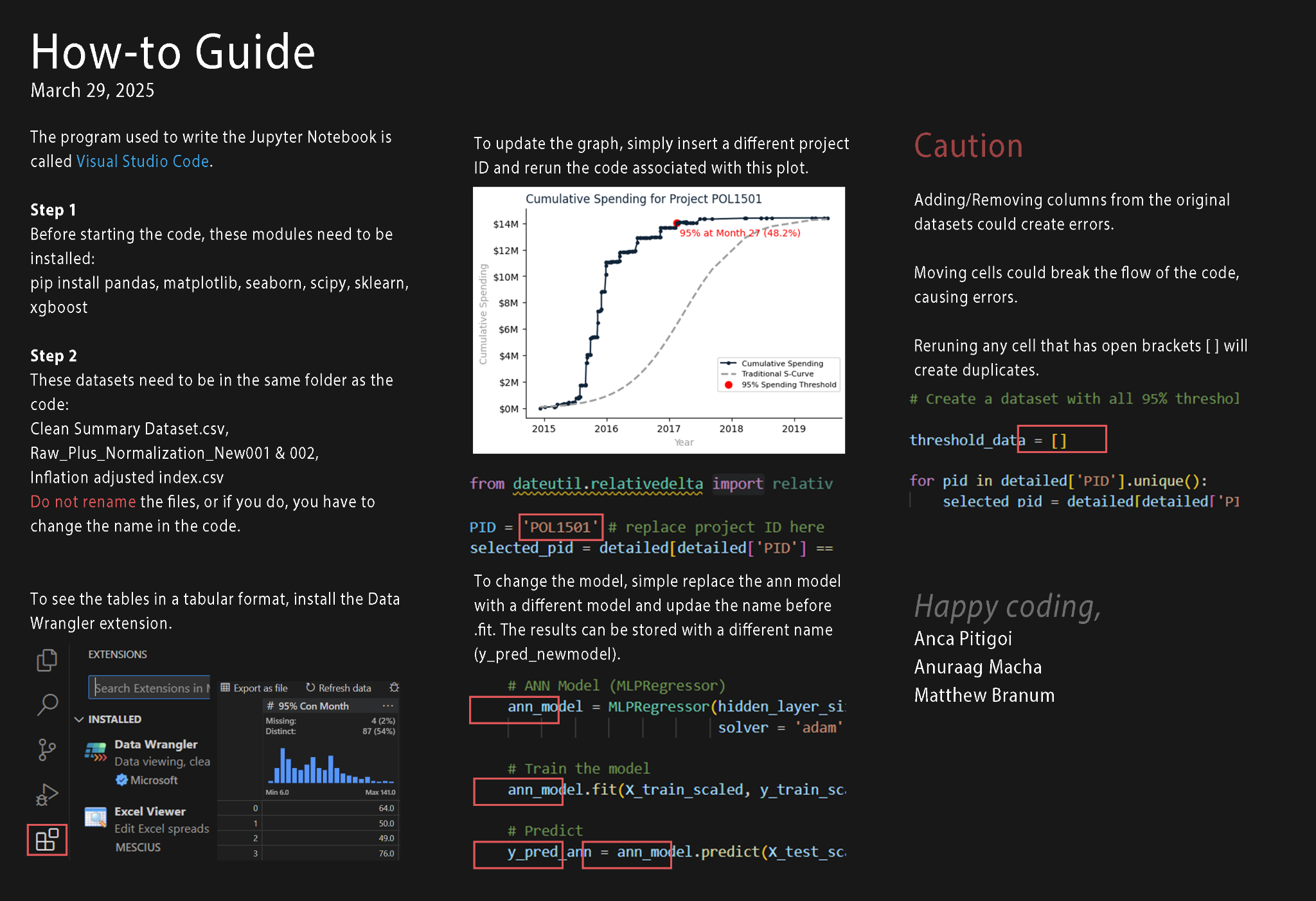
The findings are directly relevant to DCAMM’s objective of developing AI/ML tools to forecast spending for Massachusetts state facilities. The taxonomy provides a guide to selecting ML techniques and highlights the importance of integrating diverse approaches for improved accuracy. Furthermore, many 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**

****