Construction Check Civilian Area Cost Factors Utilizing Machine Learning (July 2025)

Michael Cheng, Casey Hatfield, Kartik Sundaram

***Abstract*— Accurately forecasting construction costs across regions is a persistent challenge due to variability in labor markets, material prices, geographic constraints, and economic conditions. This study introduces a data-driven Area Cost Factor (ACF) modeling framework tailored for U.S. civilian construction projects, building upon methodologies historically used by the U.S. Army Corps of Engineers. We assembled and integrated diverse datasets including project-level construction costs, metropolitan labor wages, localized material prices, natural hazard risks, weather events, and productivity indicators to construct a regionalized cost adjustment system. Inflation-adjusted price-per-square-foot estimates were used as the basis for modeling, normalized to 2025 dollars using Consumer Price Index (CPI) data.**

Following extensive preprocessing, including outlier filtering, log transformations, and feature engineering, we trained and evaluated a suite of regression models: Linear Regression, Random Forest, XGBoost, and LightGBM. Each model was benchmarked using cross-validation and a suite of error metrics (RMSE, MAE, MAPE, R-squared). Atlanta was selected as the baseline metro for cost normalization due to its consistent data coverage. Predicted ACFs for 102 U.S. metros were scaled relative to Atlanta, enabling interpretable regional comparisons.

The LightGBM model achieved the best predictive performance, with an R-squared of 0.665 and MAPE of 31.7%. Key predictors included project year, natural hazard frequency, and labor productivity indices. Comparison with government-published ACFs revealed general alignment, though discrepancies highlight the potential of machine learning to adaptively capture localized cost factors not reflected in static indices. The results demonstrate the feasibility and accuracy of dynamic ACF modeling using real-world inputs, offering enhanced decision-making tools for planners, developers, and public agencies involved in early-stage construction budgeting.

I. Introduction

Accurate and timely estimation of construction costs is a critical component of successful project planning and execution. The complexity of construction cost estimation arises from significant variations in local conditions, including labor markets, material availability, regional economic factors, and logistical considerations. To address this complexity, Area Cost Factors (ACFs) multiplicative adjustments that translate baseline or national-average construction cost estimates into accurate, region-specific forecasts are employed. ACF methodologies, notably utilized by entities such as the U.S. Army Corps of Engineers (USACE), allow for robust, standardized estimates across geographically diverse areas, including international contexts.

Recognizing the need for similarly rigorous estimation practices in the civilian sector, Construction Check, a professional construction cost estimation firm based in Atlanta, GA, seeks to adapt and refine an ACF model tailored specifically for civilian construction projects within the United States. The central goal of this project is the development of a data-driven, dynamic ACF model that delivers highly accurate early-stage cost estimates while accounting for regional pricing discrepancies. These models are especially relevant to local governments, state agencies, and commercial developers, and can substantially improve early-stage planning and budgeting for public and private infrastructure projects.

Moreover, implementing a civilian-oriented ACF model bridges the gap between military-based use cases, represented by methodologies developed by the USACE and the nuanced needs of civilian stakeholders. By utilizing real-world market inputs and systematically integrating factors such as regional labor and material costs, climatic conditions, natural hazards, logistics, and productivity adjustments, our model will empower Construction Check to enhance decision-making capabilities and reliability of project cost forecasts. Ultimately, the successful deployment of this tailored ACF methodology will enable stakeholders across the construction sector to achieve more accurate budgeting, mitigate financial risks, and drive efficient allocation of resources at critical early stages of project development.

II. Literature Review

Recent advancements in construction cost estimation underscore the necessity for dynamic, region-specific models to accurately predict building expenses across diverse metropolitan areas. Five pivotal resources, the ENR Q1 2025 Construction Cost Report, Estimating in Building Construction by Dagostino and Peterson, Kim et al. (2013), Elhag and Boussabaine (1999), and Horner and Zakieh (1996), offer critical insights into trends and methodologies relevant to the development of an Area Cost Factor (ACF) model.

The ENR Q1 2025 Construction Cost Report reveals a nuanced landscape of construction costs across the United States. The Building Cost Index (BCI) increased by 1.6% over the year, while the Construction Cost Index (CCI) rose by 0.9% during the same period. These indices reflect fluctuations in material prices and labor costs, which vary significantly across regions. For instance, certain metropolitan areas experienced higher increases in material costs due to supply chain constraints and localized demand surges. Additionally, labor shortages, particularly in regions with heightened construction activity, have driven up labor costs, further influencing overall project expenses. These findings underscore the importance of

incorporating region-specific data into cost estimation models to enhance accuracy.

Dagostino and Peterson's textbook, Estimating in Building Construction, provides a comprehensive overview of modern construction cost estimation practices. Their systematic approach emphasizes detailed quantity takeoffs, thorough labor and material cost analyses, and the integration of computer-assisted estimating tools. They advocate adjusting estimates based on regional economic indicators and local construction practices. This methodology aligns closely with the requirements of adaptable ACF models, emphasizing the need for responsiveness to the dynamic nature of construction markets across metropolitan areas.

Kim et al. (2013) conducted a comparative study on construction cost estimation methods, focusing specifically on regression analysis, neural networks, and support vector machines. Their research demonstrated the superior accuracy of neural networks in estimating school building construction costs. This outcome suggests that incorporating machine learning techniques into ACF models could significantly enhance predictive accuracy, particularly when addressing complex, nonlinear relationships inherent in construction data.

Elhag and Boussabaine (1999) explored various factors influencing construction costs, highlighting the importance of project size, complexity, and location. They employed artificial neural networks to effectively capture and model the intricate interactions between these cost drivers. Their findings reinforce the potential benefits of integrating advanced computational methods into ACF models, allowing for nuanced, precise, and dynamic cost predictions.

Horner and Zakieh (1996) introduced the concept of "characteristic items" as a novel method for pricing and controlling construction projects. They identified characteristic items as frequently occurring, high-value work packages that significantly influence overall project costs. Their research found a consistent linear correlation between the largest quantity item and total package cost, demonstrating that focusing on this dominant cost drivers could simplify estimation processes without sacrificing accuracy. This approach provides a pragmatic yet precise methodology for integrating simplified proxies into ACF models, potentially streamlining estimation and enhancing operational control.

Together, these studies collectively highlight essential considerations for developing a robust ACF model. The convergence of regional variability insights (ENR), detailed estimation frameworks (Dagostino and Peterson), advanced computational methodologies (Kim; Elhag and Boussabaine), and strategic simplifications (Horner and Zakieh) illustrate the multifaceted approach necessary to accurately predict construction costs across diverse metropolitan areas. Incorporating these combined methodologies will enhance the accuracy, adaptability, and overall robustness of future ACF models.

III. DATA DESCRIPTION

A. *Construction Check*

Project cost data was collected from Construction Check's database, which was limited to the Commercial construction category due to lack of geographic variation in the Civil, Infrastructure & Landscaping category. The total MLE of each project's line items were aggregated to get the total project cost, then divided by the project's square footage to control for project size. Project year was also extracted from the database and used as a feature.

B. *Labor Wage Data Collection and Processing*

To assess regional variability in labor costs critical to construction cost modeling, we implemented a data acquisition pipeline targeting occupational wage data from the U.S. Bureau of Labor Statistics (BLS) Occupational Employment and Wage Statistics (OEWS) program. Specifically, we utilized the annual metropolitan-level wage tables published by BLS, which provide detailed compensation data across hundreds of Standard Occupational Classification (SOC) codes at various geographic levels (https://www.bls.gov/oes/tables.htm). Our focus was on skilled trades relevant to general construction, aligning with labor roles emphasized in the Army Corps of Engineers' Area Cost Factors (ACF) methodology.

From the complete BLS dataset (`all_data_M_2024.csv`), we filtered to retain only metropolitan-level entries, excluding national aggregates, and isolated eight key construction occupations: Carpenters, Cement Masons and Concrete Finishers, Electricians, Structural Iron and Steel Workers, Construction Laborers, Painters, Plumbers, Pipefitters, Steamfitters, and Roofers. These roles represent a broad cross-section of labor-intensive construction activities and were chosen for their direct alignment with Tri-Service cost modeling conventions.

Each occupation's median hourly wage (`H_MEDIAN`) was extracted for all qualifying metropolitan areas. The dataset was then aggregated to create one row per city with individual columns for each trade's wage, enabling direct inter-city comparisons. Median wage values were coerced to numeric format, and rows with missing or non-numeric values were retained for consistency, with averages calculated across available trades to generate a composite labor rate (`H_MEDIAN_Avg`) per city.

To control for geographic representation, cities were grouped by state and sorted based on data completeness, favoring those with the lowest proportion of missing wage records. The top two cities per state, as measured by data availability (lowest missing count), were selected to serve as representative labor markets. The resulting output (`labor_by_city.csv`) serves as a geographically balanced panel of labor rates for downstream modeling of regional construction costs.

C. *Material Data*

To assess regional variability in material pricing relevant to construction cost modeling, we used consumer construction related data as a proxy for overall construction project material costs. A systematic data gathering strategy was implemented for various products from various Home Depot stores across 96 key U.S. metropolitan areas defined by the Department of

Defense Area Cost Factors (ACF) framework. The objective was to capture market pricing for a standardized set of building materials, forming a "market basket" representative of essential construction inputs. The market basket mirrors the Tri-Service Cost Engineering Committee's baseline labor-material-equipment (MLE) ratios and is intended to reflect the general needs of commercial and residential construction.

Using a predefined Army Corps of Engineers list of 96 base metro areas, we geocoded corresponding ZIP codes and fed them into the pipeline to anchor searches to specific store locations. To enable precise regional comparison of material pricing, each of the 96 metro areas defined by the Department of Defense Area Cost Factors (ACF) was assigned geographic coordinates using standardized geocoding methods. Latitude and longitude coordinates were used both for associating stores with metro areas and for enabling distance-based filtering during data collection.

A comprehensive list of US Home Depot Stores was extracted from SerpAPI (home-depot-stores-us.json) for each metro area centroid (latitude, longitude) and distance was computed to all available Home Depot stores. Stores were then sorted by distance, and the three closest stores were selected as the representative set for that metro. If a metro area had fewer than three nearby stores (within a 50-mile threshold), fallback logic was used to include the next nearest valid stores, prioritizing those within the same state.

To maintain temporal consistency, scraping was performed in a single batch window to avoid confounding price differences due to weekly promotions or regional sales events. Output from each run was serialized and consolidated into a flat CSV (hd_fullscale_final_output.csv), indexed by metro area, store ID, and item ID. Data normalization steps were applied to standardize unit pricing and reconcile discrepancies in packaging (e.g., cost per square foot vs. cost per roll). SKUs not found in a given location were logged for availability analysis, and in such cases, nearest-store substitution logic was optionally enabled to impute missing values.

D. Adjustment Factors

The USACE utilizes seven matrix factors in addition to normalized MLE that affect local construction costs: weather, seismic, climatic (exterior envelope zone, wind load), labor availability, contractor overhead and profit, logistics and mobilization, and local labor productivity (U.S. Army Corps of Engineers, 2025). We collected similar metrics at the local, state, and regional level to capture weather, natural hazard, labor productivity, and logistics and mobilization effects on construction cost.

At the city-level, adverse weather events were collected from the NOAA's Storm Events Database (National Centers for Environmental Information, National Oceanic and Atmospheric Administration, 2025). This dataset includes information for individual episodes of significant weather events starting in 1950, such as storms of enough intensity to cause loss of life, property damage, or disruption to commerce, rare phenomena, and other significant meteorological events. For each type of weather event, the mean number of annual episodes from 2020-2024 was aggregated at the city level, resulting in 16 features.

We collected information at the state level for 18 natural hazards from FEMA's National Risk Index dataset (Federal Emergency Management Agency, 2023). For each hazard, we collected the Expected Annual Loss Score and Expected Annual Loss Rate – Building fields, as well as the Expected Annual Loss Rate composite score. These scores represent the average percentage losses to buildings, population, and agriculture each year due to the natural hazard.

11 measures of labor productivity were taken from the U.S. Bureau of Labor Statistics' Private Nonfarm Labor Productivity and Costs by State and Region dataset (U.S. Bureau of Labor Statistics, 2025).

20 features were provided by the local area factors in the US Army Corps of Engineers Engineering Pamphlet (EP) 1110-1-8 (U.S. Army Corps of Engineers, Walla Walla District, 2024). Local area factors provided transportation and logistics costs for 12 regions such as gasoline cost per gallon, diesel cost per gallon, and freight rates.

E. Combined Data

Metro areas of interest were selected by overlapping Construction Check historical project data with the three feature datasets mentioned above. These metro areas do not map one to one with the ACOE dataset as the construction database is not complete for all 96 metro areas. There were also several Construction Check projects that did not match a metro area in the three data sets directly. To account for this, we made the following manual changes to better merge the datasets between the feature data we collected externally and the Construction Check dataset.

Chicago (removed Elgin-Naperville)
Akron -> Pittsburgh, PA
Wheeling -> Morgantown, WV
Fayetteville, NC -> Durham, NC
Philadelphia, PA -> Allentown, PA

After merging, we had 102 metro areas for ACF prediction.

IV. EXPLORATORY DATA ANALYSIS OF CONSTRUCTION CHECK DATABASE

Project information in the Construction Check database was available across all states, and at the time of evaluation had cities for 1,300 cities populated for 2,603 total projects. As Georgia was the only state with projects in the Civil, Infrastructure & Landscaping construction category, we limited our analysis to the Commercial category.

Line items were coded at varying levels of CSI Code granularity, and market basket analysis was conducted at the division level. Using Pareto analysis, the top 10 divisions were identified based on mean project cost contribution, shown in *Table 1* These high-impact divisions were used to inform which MLE costs should be collected at the local level.

Division	Number of States		Project Count	Mean % Project Contribu
Electronic Safety and Security	50	1,444	73.4	
HVAC	42	416	54.7	
Material Processing and Handling Equipment	50	1,086	40.7	
Electrical Power Generation	8	19	35.6	
Exterior Improvement	5	126	26.6	
Fire Suppression	43	335	22	
Earthwork	22	169	19.6	
Integrated Automation	2	12	10.3	
Utilities	2	143	10.2	
Electrical	1	116	9.2	

Table 1: Top 10 Divisions by Mean Percentage of Project Costs Across all CSI Codes

We also found a wide variation in the number of project per project category within commercial construction, with educational, government, medical, offices & warehouses, religious, residential, retail, and sports categories having at least 100 projects available. High-impact line items between these project categories had some overlap for materials and equipment. However, the impact of items such as regulatory, scope of bids, and contracts differed, especially for government projects. These findings underscore the importance of adjustment factors beyond MLE for ACF prediction.

Although CSI codes and unit costs were provided by the Construction Check database, outside indexes such as RSMeans City Cost Index would be required for city-level analysis. To avoid an outside index biasing ACF prediction, we determined that outside data sources should be used for a more detailed cost comparison, prompting materials cost collection with Home Depot's API.

Below we have graphs of Construction Check data created during the exploratory data analysis phase.

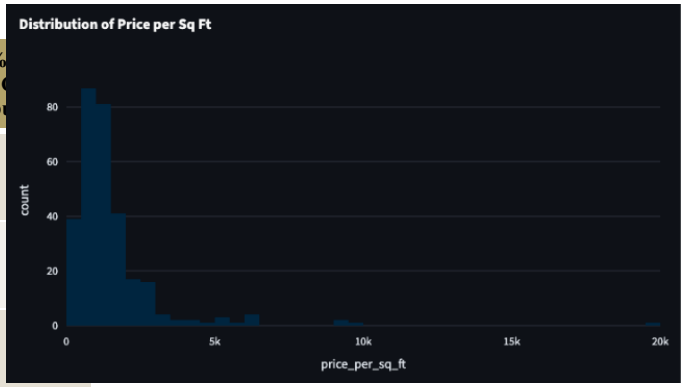

Figure 1: Distribution of price per square foot

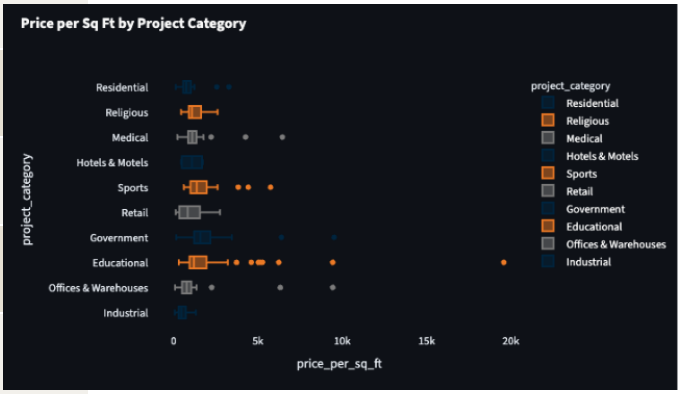

Figure 2: Price per square foot by project category

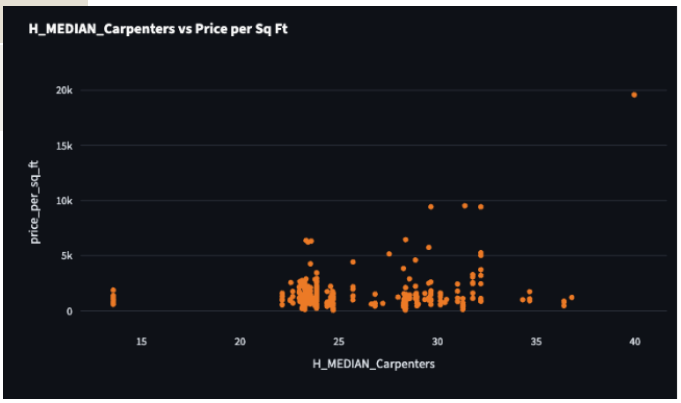

Figure 3: Carpenter Wage vs. Price per Square Foot

V. METHODOLOGY

A. Exploratory Modeling Attempts

The predictive modeling process began with a dataset comprising of 1,300 U.S. construction projects. These records included 117 features, ranging from basic project descriptors (e.g., size, year, state) to categorical identifiers such as project type and ownership. The response variable was the total cost of the project, representing the total reported cost of each project. Initial summary statistics revealed considerable skewness and dispersion in the target, with a mean of approximately $122 million and a maximum value exceeding $40 billion. The data was highly heterogeneous, with substantial variance in scale and category representation, highlighting the need for careful preprocessing and modeling design.

Our analysis aimed to estimate the Area Cost Factor (ACF) using normalized price per square foot data from U.S. construction projects. The response variable was initially defined as the total project price per square footage but was later inflation-adjusted to 2025 dollars using year-specific Consumer Price Index (CPI) values. This transformation yielded a new target variable, price_per_sq_ft_2025, allowing for cost comparisons across project years in constant terms.

The raw dataset was preprocessed to remove non-numeric columns and identifiers. Categorical variables, namely location key, construction category, and project category, were encoded using mean target encoding to preserve information about category-level cost variation. One-hot encoding was also selectively applied, particularly for metro area, to preserve geographic distinctions.

The initial baseline model employed was a simple linear regression applied to the raw dataset. Although this offered a starting point for understanding model behavior, the results were poor and uninformative. The model exhibited extreme error magnitudes and highly negative R-squared values, a clear indication that linearity assumptions were not appropriate given the underlying data structure. Performance metrics such as RMSE and MAPE confirmed that naive modeling would not yield practically useful results.

Recognizing the presence of significant outlier records, we turned to systematic outlier detection. A hybrid approach using the interquartile range (IQR), Z-score filtering, and Isolation Forest ensemble methods were employed. This process identified 147 observations, approximately 11.3% of the data, as outliers. These points were disproportionately influential and often associated with megaprojects or entries containing improbable cost-to-size ratios. Removing these records resulted in a cleaner dataset of 1,153 samples, which served as the foundation for subsequent modeling steps.

To address the skewness in the project cost distribution, a log transformation was applied. The original skewness of 1.443 was reduced to -0.755, yielding a more symmetric distribution of the target variable and improving model assumptions related to homoscedasticity. With the target stabilized, additional features were engineered to hopefully better capture domain-specific patterns. These included cost_per_sqft (cost normalized by project size), project_age (derived from the year of construction), and an interaction term combining square footage with project year. These features were informed by both literature research and preliminary correlation analysis and expanded the feature set to 108 columns.

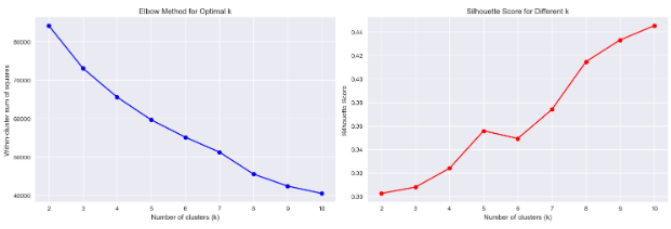

Figure 4: K means clustering analysis

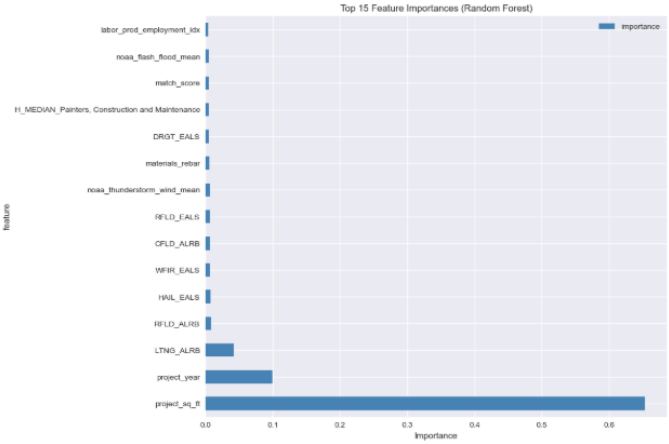

Figure 5: Feature Importance

To improve generalizability and capture latent structure in the data, we introduced unsupervised clustering. K-means clustering with k=10 was applied to segment projects based on multivariate similarity. The optimal number of clusters was selected using silhouette scoring. This new cluster feature was then used to stratify the training and test sets, ensuring that each subset preserved the diversity of project types present in the full dataset. The final stratified split allocated 80% of the data (922 samples) for training and 20% (231 samples) for evaluation.

Due to the high importance of the project square footage feature, we employed additional methods to uncover more meaningful predictors for our ACF model.

B. Final Model with Feature Normalization Relative to Atlanta Metro

To more completely capture the effects of location MLE and other adjustment factors, features were extracted from the raw dataset and normalized relative to the median values of Atlanta projects. There was not sufficient project data across all cities to use the national median as the benchmark. Due to the Atlanta-Sandy Springs-Roswell, GA metro area having the most robust data available across years and project types, it was used as our ACF benchmark. Projects for this metro were assigned a baseline value of 1.0 by definition. Outliers were again removed using IQR, Z-score filtering, and Isolation Forest ensemble methods. Projects were limited to the commercial construction category, and project year was the only project-specific feature included with the local MLE and adjustment factors. The resulting dataset had 906 samples and 103 features.

We used ACF as our new target by dividing project price per square foot by the median price of the benchmark Atlanta-Sandy Springs-Roswell, GA metro projects for the corresponding year. A crosswalk of benchmark projects was created by taking the median price per square foot of projects for all years 1971 onward. For years without a benchmark project, the price was calculated using the Consumer Price Index (CPI) and the median price per square foot for the most recent year with benchmark data (U.S. Inflation Calculator, 2025). The target was highly skewed, and applying a log transformation decreased the skewness from 4.101 to 1.438.

The data was split into 80% training data and 20% test data. We trained a suite of regression models, including Linear Regression, Random Forest, XGBoost, and LightGBM. These models were chosen to balance interpretability and predictive performance, and to test both linear and nonlinear relationships. 5-fold cross validation was applied for robust model R-squared comparison. Performance metrics include: mean squared error, root mean square error, mean absolute error, mean absolute percentage error, and R-squared.

To enable direct comparison of Area Cost Factors (ACFs) across metro areas, all predicted ACF values were normalized relative to Atlanta. Specifically, the "Atlanta-Sandy Springs-Roswell, GA" metro was selected as the baseline due to the consistency and completeness of its data across years and project types. For each model, Random Forest, XGBoost, and LightGBM, ACF values for every metro area were divided by the corresponding value predicted for Atlanta. This normalization procedure ensured that Atlanta had a cost factor of exactly 1.0 by definition, while other metros were expressed as a ratio of their predicted cost relative to Atlanta. As a result, normalized ACF values above 1.0 indicate higher relative costs compared to Atlanta, while values below 1.0 represent lower costs. The normalized ACFs were compiled into a final dataset containing one row per metro area and three model-specific normalized columns. This dataset was exported as a CSV and used for subsequent comparisons and visualization.

VI. KEY FINDINGS AND INTERPRETATION

The results of our project demonstrate the critical importance of iterative refinement in predictive modeling for construction cost estimation. The comparison between baseline models (trained on unprocessed data) and the final tuned models (trained on a log-transformed, outlier-filtered, and feature-enhanced dataset) shows substantial gains in both accuracy and stability.

Initially, model performance was unusable. The linear regression baseline completely failed to capture any meaningful variance in the target, producing highly negative R-squared scores and astronomically high error metrics due to the presence of extreme outliers and target skewness. Even ensemble methods such as Random Forest and XGBoost, though more robust, struggled in their initial implementation, achieving R-squared values around 0.56–0.58 and MAPE values exceeding 130%. These early results emphasized the data's inherent complexity and the insufficiency of applying models directly to raw values.

Through targeted iterative approach including log transformation of the cost variable, removal of statistical outliers, feature engineering of domain-relevant variables, and stratified clustering, our models produced much better results. Both Random Forest and XGBoost achieved R-squared scores greater than 0.85, with MAPE values improving to approximately 48–51%. This represents a 6-fold reduction in mean absolute percentage error and a dramatic reduction in absolute error (MAE), from approximately $18 million to just over $6 million.

Table 2 below summarizes this performance shift:

Model	R^2 (Ori.)	R^2 (Tuned)	MAE (Orig.)	MAE (Tuned)	MAPE (%) Orig.	MAPE (%) Tuned
Linear Regression	-7.42 x 1018	-0.529	$9.99 B	$30.0 M	2.38 x 1010 %	194.4 %
Random Forest	0.5799	0.8558	$18.6 M	$6.18 M	126.5 %	51.2 %
XGBoost	0.5649	0.8576	$18.8 M	$6.13 M	130.5 %	48.2 %

Table 2: Performance Shift after Initial Model Tuning

Overall, the tuned XGBoost model emerged as the best performer, offering the highest R-squared and lowest error metrics across all categories. These outcomes provide strong empirical support for using advanced ensemble models, coupled with domain-specific feature design and robust data cleaning, to improve forecasting in construction cost estimation contexts.

Beyond performance, analysis of feature importance offers insight into which variables most strongly influenced the model predictions. The Linear Regression model assigned astronomical coefficients to project_year and project_age (~$10 trillion), reflecting multicollinearity and poor model conditioning. These coefficients lacked meaningful interpretability, further emphasizing the limitations of linear methods in this context.

In contrast, both Random Forest and XGBoost revealed more stable and interpretable importance rankings. The interaction term sqft_year_interaction (calculated as project_sq_ft x project_year) emerged as the most important predictor across both models, contributing approximately 47.2% of decision splits in Random Forest and 34.7% of predictive gain in XGBoost. These results emphasized the high importance of project_sq_ft, and thus we sought out other models to identify additional predictors for our ACF model.

VII. RESULTS AND CONCLUSIONS

Table 3 summarizes the performance of our final four predictive models using Atlanta-scaled features. Consistent with exploratory analyses, Linear Regression performed the worst, yielding a negative cross-validated R-squared and the highest error rates across all evaluation metrics. In contrast, LightGBM demonstrated the strongest performance, achieving the highest test R-squared (0.665) and the lowest RMSE (0.265), MAE (0.206), and MAPE (31.7%). Random Forest and XGBoost followed in performance, with Random Forest slightly outperforming XGBoost across most metrics.

Model	Train CV R-squared +/- std	Test R-squared	RMSE	MAE	MAPE
Linear Regression	-0.234 +/- 0.537	0.143	0.423	0.307	54.0%

Random Forest	0.494 +/- 0.066	0.618	0.283	0.219	33.3%
XGBoost	0.389 +/- 0.098	0.522	0.316	0.236	36.0%
LightGBM	0.519 +/- 0.048	0.665	0.265	0.206	31.7%

Table 3: Evaluation Metrics Using Atlanta-Scaled Features

To interpret model decision-making, feature importances from the tree-based models were extracted, shown in Figures 6-8 below. Project year emerged as the dominant feature in both Random Forest and LightGBM, far surpassing all others in importance. However, project_year was not in the top ten feature importances for XGBoost. This model placed a heavier emphasis on natural hazard, weather, and labor productivity features and did not have any material cost features in the top ten. All three models ranked coastal flooding among their top two most influential features, highlighting its consistent relevance across modeling approaches.

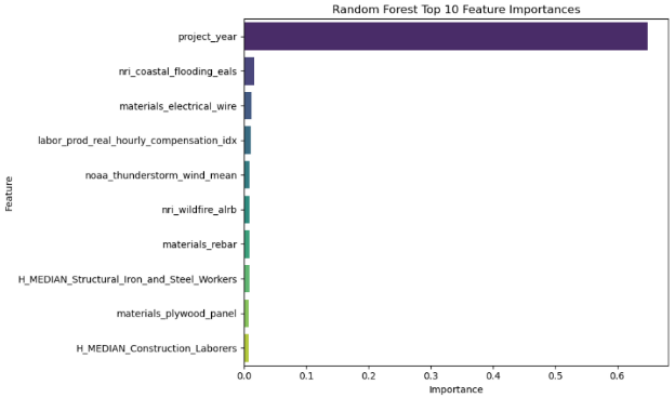

Figure 6: Random Forest Top 10 Feature Importances

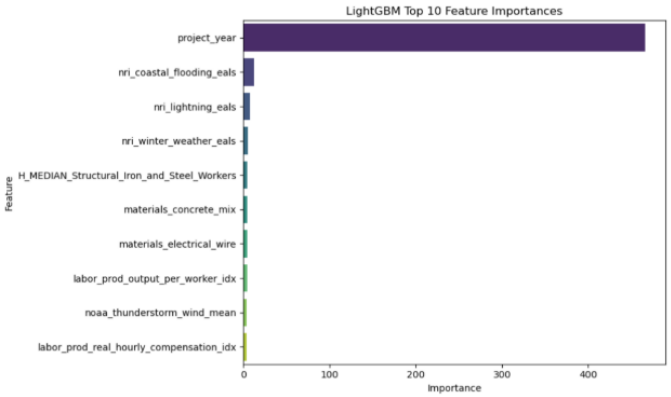

Figure 7: LightGBM Top 10 Feature Importances (Gain)

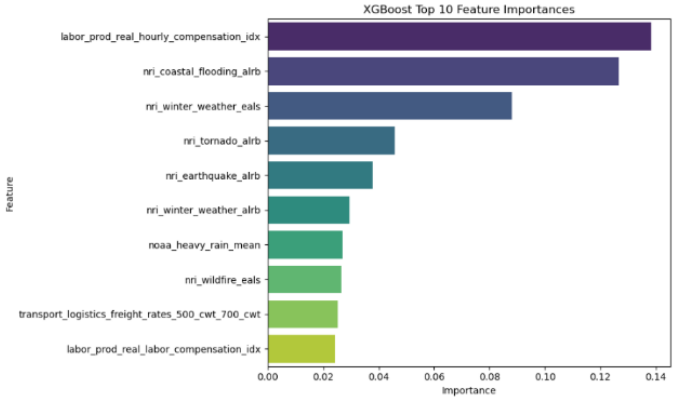

Figure 8: XGBoost Top 10 Feature Importances

Model-generated ACFs were generated for 102 metro areas using the chosen LightGBM model and are shown in Table 4 and Table 5 below. Predicted ACFs for civilian construction were compared with published military ACFs from the Army Corps of Engineers May 2025 PAX newsletter (2025) normalized for Atlanta for 46 overlapping metro areas. Although there are expected differences between civilian and military construction costs, this allowed us to look at similarities and differences between regional trends identified by the model and those identified by USACE's analysis. There were 12 cases where the model predicted different regional trends than what was published by USACE, with 3 predictions for lower costs than Atlanta when USACE predicted higher and 9 predictions for higher costs than Atlanta when USACE predicted lower. Importantly, only 8 out of the 102 metros evaluated had predicted ACFs below 1, indicating a general upward skew in predicted costs compared to Atlanta.

Metro Area	Predicted Civilian ACF	USACE Published ACF
Albany-Schenectady-Troy, NY	1.65	1.17
Albuquerque, NM	1.62	1.01
Anchorage, AK	1.85	2.58
Atlanta-Sandy Springs-Roswell, GA	1.00	1.00
Baltimore-Columbia-Towson, MD	1.49	1.02
Bangor, ME	1.65	1.21
Billings, MT	1.22	1.16
Boise City, ID	1.44	1.17
Boston-Cambridge-Newton, MA-NH	1.57	1.40
Bridgeport-Stamford-Danbury, CT	2.35	1.24
Buffalo-Cheektowaga, NY	1.74	1.18
Burlington-South Burlington, VT	1.37	1.20
Chattanooga, TN-GA	1.48	0.92
Cheyenne, WY	1.56	1.12
Davenport-Moline-Rock Island, IA-IL	1.29	1.12
Detroit-Warren-Dearborn, MI	0.91	1.18

Dover, DE	1.06	1.11
Duluth, MN-WI	1.12	1.25
Fairbanks-College, AK	1.60	2.64
Gulfport-Biloxi, MS	1.23	0.99
Jacksonville, FL	1.92	0.97
Kansas City, MO	1.16	1.08
Las Vegas-Henderson-North Las Vegas, NV	1.48	1.51
Lexington-Fayette, KY	1.15	0.98
Little Rock, AR	1.60	0.91
Louisville/Jefferson County, KY-IN	1.14	0.96
Madison, WI	1.07	1.24
Miami-Fort Lauderdale-West Palm Beach, FL	1.94	1.03
Minneapolis, MN	0.98	1.25
Mobile, AL	1.52	0.98
New Orleans-Metairie, LA	1.80	0.96
Ogden, UT	1.51	1.13
Oklahoma City, OK	1.63	1.02
Omaha, NE-IA	1.27	1.10
Phoenix-Mesa-Chandler, AZ	0.94	1.03
Portland-South Portland, ME	1.52	1.22
Portland-Vancouver-Hillsboro, OR-WA	1.57	1.25
Providence-Warwick, RI-MA	1.47	1.26
Rapid City, SD	1.30	1.12
Sioux Falls, SD-MN	1.21	1.17
Springfield, MA	1.48	1.22
Trenton-Princeton, NJ	1.39	1.33
Urban Honolulu, HI	1.68	2.36
Washington-Arlington-Alexandria, DC-VA-MD-WV	1.68	1.19
Wichita, KS	1.42	0.99

Table 4: Civilian ACFs from LightGBM Model Compared with USACE ACFs from May 2025 PAX Newsletter Normalized for Atlanta

Metro Area	Predicted Civilian ACF
Allentown-Bethlehem-Easton, PA-NJ	0.95
Amarillo, TX	2.12
Atlantic City-Hammonton, NJ	1.43
Augusta-Richmond County, GA-SC	0.93
Austin-Round Rock-San Marcos, TX	1.76
Bakersfield-Delano, CA	1.62
Baton Rouge, LA	1.54
Bellingham, WA	1.73
Birmingham, AL	1.37
Bismarck, ND	1.21
Bozeman, MT	1.27
Canton-Massillon, OH	1.34
Cedar Rapids, IA	1.24
Champaign-Urbana, IL	1.20
Charlotte-Concord-Gastonia, NC-SC	1.20
Chicago, IL	1.19
Chico, CA	1.62
Coeur d'Alene, ID	1.27
Durham-Chapel Hill, NC	1.34
Elkhart-Goshen, IN	1.31
Fargo, ND-MN	1.16
Farmington, NM	1.70
Fayetteville, AR	1.39
Fayetteville, NC	1.34
Flint, MI	0.93
Green Bay, WI	1.19
Greenville-Anderson-Greer, SC	1.16
Hagerstown-Martinsburg, MD-WV	1.58
Harrisonburg, VA	2.07
Hartford-West Hartford-East Hartford, CT	2.56
Huntington-Ashland, WV-KY-OH	1.55
Indianapolis, IN	1.49
Jackson, MS	1.35
Kahului-Wailuku, HI	1.69
Kennewick-Richland, WA	1.73
Knoxville, TN	1.50
Lancaster, PA	1.27
Lincoln, NE	1.22
Lynchburg, VA	2.44
Manchester-Nashua, NH	1.52
Memphis, TN	1.60
Morgantown, WV	1.65
Nashville, TN	1.48
Norfolk, VA	2.44
Philadelphia, PA	0.95
Pittsburgh, PA	1.54
Provo-Orem-Lehi, UT	1.50
Reno, NV	1.48
Saint Louis, MO	1.35

Saint. Paul, MN	0.98
Salem, OR	1.58
San Juan-Bayamon-Caguas, PR	2.09
Sierra Vista-Douglas, AZ	1.16
Spartanburg, SC	1.30
Topeka, KS	1.55
Tulsa, OK	1.54
Wheeling, WV-OH	1.65

Table 5: Civilian ACFs from LightGBM Model Compared with no Corresponding USACE ACF

The final LightGBM model effectively captured regional cost differences, reflecting higher ACFs in high-cost metros and lower values in more affordable areas. However, the skew toward ACFs above 1 suggests potential overestimation in certain regions. The strong influence of project year in two out of three top-performing models also underscores the need for better temporal controls in future datasets. Additionally, the prominence of non-cost-related factors such as weather and natural hazards suggests that machine learning models may offer a more dynamic and context-sensitive alternative to traditional, government-published ACF indices. Overall, these findings support the feasibility of machine learning for civilian ACF estimation, enabling researchers to generate more nuanced and adaptive cost predictions when incorporating broader economic and environmental variables.

VIII. KEY CHALLENGES AND FUTURE IMPROVEMENTS

The development of an accurate, data-driven Area Cost Factor (ACF) model encountered several substantial challenges, particularly around data acquisition, data integration, and model performance. One primary issue was sourcing suitable datasets that provided comprehensive, metro-level detail while remaining accessible for research purposes. Many established resources for detailed construction pricing data, such as RS Means require costly subscriptions to be able to extract and use data. This limitation necessitated the development of alternative approaches to get data via SerpAPI and special allocation of tokens.

Another notable challenge was encountered during the integration of Construction Check's internal dataset with the externally collected material data from Home Depot. Variations in regional classifications and naming conventions complicated the merging process, requiring careful manual reconciliation and adjustments to standardize city-metro area relationships across datasets. Specific manual adjustments, such as redefining metro boundaries (e.g., replacing "Elgin-Naperville" with "Chicago," adjusting "Akron" to "Pittsburgh," or substituting "Wheeling" with "Morgantown"), illustrate the inherent complexities and limitations of combining diverse data sources.

From a modeling perspective, a persistent challenge was avoiding overfitting, especially due to the disproportionate predictive importance of project square footage (project_sq_ft) and related interaction terms (sqft_year_interaction). Despite various preprocessing techniques, including log transformations, outlier removal, and feature engineering, the dominance of project size consistently overshadowed other potentially meaningful variables. This effect limited the interpretability of certain model outputs and suggested the presence of underlying multicollinearity, reinforcing the need for careful feature selection and dimensionality reduction.

Significant efforts were required to address the skewness and heterogeneity in construction cost data. The application of multiple preprocessing methods, such as systematic outlier detection (Isolations Forest, Z-scores, IQR methods) and normalization, was necessary to achieve acceptable modeling performance. Yet, the resulting models still showed room for improvement, highlighting persistent complexities inherent in modeling construction cost data across heterogeneous geographic contexts.

Looking forward, several opportunities for future improvements are identified. First, enhancing data collection methods to systematically capture and integrate publicly available, standardized regional market data could significantly reduce the reliance on paid services and improve the scalability of the model. Additionally, establishing a robust, automated data normalization pipeline would streamline data integration, reduce manual intervention, and improve the consistency and repeatability of the modeling process.

Secondly, refining the feature set through more sophisticated dimensionality reduction methods and regularization techniques could mitigate overfitting risks associated with dominant features like project size. Incorporating additional variables that more explicitly capture regional economic indicators, local regulatory environments, or labor market dynamics might also help balance predictive power across multiple explanatory variables.

Finally, employing ensemble or stacking techniques combining multiple modeling approaches, or integrating temporal modeling to explicitly account for changes in material and labor pricing over time, could further improve model robustness, predictive accuracy, and generalizability. These enhancements would strengthen the utility and adaptability of the ACF model, enabling more accurate and reliable forecasting of construction costs across diverse metropolitan areas.

REFERENCES

Dagostino, Frank R., and Steven J. Peterson. Estimating in Building Construction. 7th ed., Prentice Hall, 2010.

Elhag, T. M. S., and A. H. Boussabaine. "Factors Affecting Construction Cost Estimating: A Comparison of Neural Network and Regression Analysis." Journal of Financial Management of Property and Construction, vol. 4, no. 1, 1999, pp. 31–38.

Engineering News-Record. "1Q 2025 Cost Report: Growth for Some Materials Prices in 2024." ENR, 3 Mar. 2025,

www.enr.com/articles/60363-1q-2025-cost-report-growth-for-some-materials-prices-in-2024.

Horner, Malcolm, and Rashad Zakieh. "Characteristic Items: A New Approach to Pricing and Controlling Construction Projects." Construction Management and Economics, vol. 14, no. 3, 1996, pp. 241–252.

Kim, Gwang-Hee, et al. "Comparison of School Building Construction Costs Estimation Methods Using Regression Analysis, Neural Network, and Support Vector Machine." Journal of Building Construction and Planning Research, vol. 1, no. 1, 2013, pp. 1–7.

SerpApi. (n.d.). Google Search API. Retrieved June 7, 2025, from https://serpapi.com/.

U.S. Bureau of Labor Statistics. (2024). Occupational Employment and Wage Statistics (OEWS) Tables. Retrieved from https://www.bls.gov/oes/tables.htm.

U.S. Inflation Calculator. (2025.) Consumer Price Index and annual percent changes from 1913 to 2008. https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/.

U.S. Army Corps of Engineers. (2025). DOD Area Cost Factors (ACF) PAX Newsletter No 3.2.1, Dated 16 May 2025 Table 4-1, UFC 3-701-01. U.S. Army Corps of Engineers. Retrieved from https://usace.contentdm.oclc.org/utils/getfile/collection/p16021coll8/id/4831.

National Centers for Environmental Information, National Oceanic and Atmospheric Administration. (2025.) Storm Events Database [Dataset]. Retrieved from https://www.ncdc.noaa.gov/stormevents/.

Federal Emergency Management Agency (FEMA). (2023.) National Risk Index [Dataset]. Retrieved from https://hazards.fema.gov/nri/data-resources#csvDownload.

U.S. Bureau of Labor Statistics. (2025.) Private Nonfarm Labor Productivity and Costs by State and Region [Dataset]. Retrieved from https://www.bls.gov/productivity/tables/labor-productivity-by-state-and-region.xlsx.

U.S. Army Corps of Engineers, Walla Walla District. (2024.) EP 1110-1-8 Construction Equipment Ownership and Operating Expense Schedule (EP 1110-1-8). U.S. Army Corps of Engineers. Retrieved from https://www.nww.usace.army.mil/Missions/Cost-Engineering/EP1110-1-8/

Workload Distribution:

Task	Description	Team Member Contributions
Logistics, Sponsor Communication	Submission of project, updating sponsor on our progress and open line of communication with Construction Check	**Kartik**: Canvas submissions, sponsor communications
Construction Check EDA	Analyze Construction Check data warehouse for market basket and ACF validation	**Casey**: regional and project type distribution, line item frequency and importance across projects and states, missing values and other data limitations **Michael**: common work descriptions, unit costs by project and item
Materials and Equipment Market Basket	Evaluate the frequency of project line items, impact on project cost, and how they differ across project types	**Casey**: Pareto analysis, feature importance using LightGBM, and association rules using Apriori algorithm. These were conducted at division, subdivision, and item code levels.
External Materials Data Research	Researched various external data sources	**Michael**: Data sources included: RSMeans, Army corps of Engineers, Retail data APIs (Home Depot, Lowe's), PRISM Climate data, FAF5 Highway data
External Materials Data –Home Depot API	Build script to extract via SerpAPI, retail construction related materials data	**Michael:** Built basket of 10 common materials for building construction projects and extracted store-specific data for market basket products. Built model to determine closest 3 stores per metro area based on distance for extraction.
Labor Market Basket	Produce a weighted average for labor rate	**Kartik**: Research on Army Corps of Engineers labor crafts. Used as a baseline standard set of labor crafts for a construction project. Starting with even weighting, goal to map to project category-based labor mix.
External Data Collection and Preprocessing	Find external data sources for labor, materials, equipment, and additional matrix items like weather and labor availability	**Casey:** aggregated weather, natural hazard, labor productivity, and transportation & logistics data from various sources for risk adjustment features. Aggregated Construction Check project data.
Literature Review	Review of current research and market research surrounding construction cost estimation	**Michael**: Reviewed 10 construction cost-based research papers to understand past cost analysis modeling techniques, results and findings. Reviewed various articles on construction related data providers and industry trends.
Midterm Report	Presentation on project progress and next steps	**All**
Modeling	Several iterations of models to calculate area cost factors	**Michael:** Preliminary modeling for linear regression and logistic regression **Kartik:** Preliminary modeling for Random Forest **Casey:** used lessons learned from preliminary modeling and outside research to create the final model using ATL as a benchmark
Streamlit App	Streamlit dashboard for Construction Check to view EDA and results	**Kartik**
Github Repo	Compiling final code and ReadMe for sponsors	**Kartik**
Final Report	Final writeup of project background, methodology, and results	**All**