# Estimating Highway Rehabilitation Projects Using Machine Learning

Samantha Ocaña[1]; Silvia Flores-Osuna[2];
Gasser G. Ali, Ph.D.[3]; and Constantine Tarawneh[4]

[1]Graduate Student, Computer Science Department, The University of Texas Rio Grande Valley, Edinburg, TX. Email: samantha.ocana01@utrgv.edu
[2]Undergraduate Student, Computer Science Department, The University of Texas Rio Grande Valley, Edinburg, TX. Email: silvia.floresosuna01@utrgv.edu
[3]Assistant Professor, Dept. of Civil Engineering, The University of Texas Rio Grande Valley, Edinburg, TX (corresponding author). ORCID: https://orcid.org/0000-0002-9853-1082. Email: gasser.ali@utrgv.edu
[4]Louis A. Beecherl, Jr. Endowed Professor, Sr. Associate Dean for Research and Graduate Programs, CECS, Director, UTCRS, Director, NSF CREST MECIS, Distinguished Teaching Professor, Mechanical Engineering Department, University of Texas Rio Grande Valley, Edinburg, TX, USA 78539, Email: constantine.tarawneh@utrgv.edu

## ABSTRACT

Highways are a crucial part of the national infrastructure. As such, they are costly to construct and maintain; challenges such as deterioration, limited budgets, and cost uncertainty create a need for improved cost prediction of rehabilitation projects. This is especially true in the preliminary stage when all factors are unknown. Accordingly, this paper proposes a model to predict the cost of highway rehabilitation projects. A machine learning model is trained on large and publicly available datasets released by the Texas Department of Transportation (TxDOT), the National Oceanic and Atmospheric Administration (NOAA), and the Federal Reserve of Economic Data (FRED). These datasets include past projects, along with their maintenance dates, costs, annual average daily traffic, temperature data, and cost indices. Several models were tested, and XGBoost demonstrated the best performance, achieving an $R^2$ of 0.74. Further inspection of the inputs showed that the type of project and project length were the strongest predictors of project cost, with AADT and cost indices also contributing. Temperature variables had little to no influence. The findings show that an XGBoost model can provide reliable cost predictions, supporting informed decision-making and more efficient resource allocation in future highway rehabilitation projects.

## INTRODUCTION

Cost prediction is important for construction projects, especially at the preliminary stage. Inaccurate estimates can lead to project losses, delays, or even cancellation. Traditional approaches often only take direct project details into consideration, such as materials, project duration, and labor. In fact, a systematic review of 105 studies identified 41 cost drivers, grouped into project, organizational, and estimator-related categories (Awuku et al. 2024). The analysis showed that project-specific factors dominated, with project details such as size and location, but other project-related measures, such as inflation, were not among the top drivers considered at the preliminary stage (Awuku et al. 2024).

Many departments rely on expert judgment, software, or handbooks (Western Federal Lands Highway Division 2023) . While these approaches provide useful guidance, they may still produce

inconsistent results, especially in the early stages of project planning. A broader industry issue is the shortage of skilled labor. Shortages of skilled labor have led to an increase in costs by reducing productivity, driving wage growth and overtime use, and increasing the likelihood of schedule delays and quality problems (Karimi et al. 2018). Change orders are also frequent in construction projects and have been shown to significantly contribute to cost growth and schedule delays (Shrestha et al. 2025). These findings show the limits of traditional estimation and highlight the value of data-driven methods that can incorporate additional features beyond project-specific details.

Recent years have shown how abruptly costs can shift. Myrvang and Liu (2025) noted how post-COVID-19 recovery led to an inflation surge and created major challenges for estimators. They observed that past research relied primarily on labor and material cost indices. However, this narrow focus falters under such shocks and recommends integrating broader economic indicators into forecasting models (Myrvang and Liu 2025). In this study, that role is filled by the Consumer Price Index (CPI), Producer Price Index (PPI), and Construction PPI to capture economy-wide and producer-level price pressures that affect highway project costs.

Beyond economic conditions, climate can also influence construction and rehabilitation costs, even if indirectly. Weather patterns influence how often maintenance is required, and even how efficiently projects can be delivered. Prior studies have focused on pavement deterioration under climate conditions (Cui and Wang 2025), yet they suggest climate may also be a relevant factor in highway preliminary cost estimation. To examine the potential influence of environmental factors in cost estimation, NOAA climatic features: maximum and minimum temperatures, average temperature, precipitation, and degree days are added and evaluated. Because our dataset is based on TxDOT projects, this analysis also provides a region-specific perspective.

Machine learning (ML) has increasingly been applied to cost estimation and related infrastructure challenges. Using Florida Department of Transportation (FDOT) data, researchers modeled the prices of highway construction cost items with 69 variables spanning construction markets, energy, socioeconomic, and macroeconomic indicators. They found that linear models outperformed nonlinear methods, reflecting the relatively stable relationships at the item level (Mahdavian et al. 2021). However, unlike item-level studies, the present work focuses on predicting the entire cost of rehabilitation projects as a whole, offering insights at the project scale. Related work using Georgia Department of Transportation (GDOT) data on 539 pavement maintenance projects have used test tree-based predictive models, including Random Forest, Extra Trees, and XGBoost, alongside different feature selection methods (Paik et al. 2025). This study showed that tree-based approaches are well suited for project level cost estimation, paralleling our own finding that nonlinear models captured the complexity of project level features.

These studies demonstrate both the promise and the constraints of ML in this domain. Building on this literature, this study uses machine learning on a large TxDOT rehabilitation dataset. It expands cost estimation by adding economic indices and climate variables to project level features. The study shows which models perform best and which factors matter most.

## GOAL

The goal of this paper is to develop a machine learning model that predicts the estimated construction cost of a rehabilitation project using early-stage project information to possibly obtain more accurate estimates during the preliminary stage when uncertainty is highest. Cost estimates in construction are developed across multiple stages, with their accuracy increasing as the project progresses. This is often due to limited knowledge about the project, such as what materials will

be used and which contractor will be performing the construction project. The PMBOK "cone of uncertainty" illustrates this principle, showing that initial estimates can vary widely and only narrow as the project progresses (Project Management Institute 2020). The AACE International Cost Estimate Classification System (AACE International 2020) expands on this principle by incorporating classes 1 through 5, with Class 5 being the preliminary stage and Class 1 being the definitive stage where bidding takes place. Class 5 indicated the expected accuracy ranges from -50% to 100%. Given that predicting the cost of a project at an early stage can lead to cost savings, there is a need for a model that performs more accurately than traditional approaches.

## METHODOLOGY

The methodology consisted of six steps: three for data preprocessing (data collection, data cleaning, data labeling) and three for model development (model selection, model training, and fine-tuning) as seen in **Figure 1**. The following sections describe each step in detail.


**Figure 1. Sequential machine learning workflow of the six main stages from data preparation to model development.**

**Data Collection.** The primary dataset was collected from the TxDOT and contained a total of 69,783 projects. The main attributes collected from this dataset were construction cost, estimated let date, project status, and project length, as these are attributes known to directly influence cost estimation outcomes. The dataset included additional variables, but they were excluded from the model due to limited relevance. To incorporate other external factors that could potentially affect construction cost, the TxDOT projects were supplemented with temperature data from the NOAA, construction cost indices from FRED, and AADT data from TxDOT. Cost indices from FRED were merged with the project data using the estimating letting dates of each project to provide each project's economic analysis. Temperature data from the NOAA was joined with the TxDOT project records using FIPS county codes and dates to consider impact of local climatic conditions on project costs. AADT was integrated to capture the impact of traffic volume. An initial attempt to merge this data using highway numbers was not successful, as highways often span long distances and do not reliably pinpoint the location of a specific project, so the AADT shapefiles were joined to the project dataset using a nearest-neighbor spatial join. Each project was matched to the closest available AADT record within a 10 meter radius to provide a more localized estimate of traffic conditions for each project.

**Data Cleaning and Labeling.** High-quality data is crucial in ensuring model output is reliable and accurate (Chen 2022). Because much of our dataset contained outliers, both heuristic approaches based on domain knowledge, and statistical methods, such as the interquartile range technique were employed to identify and address them. **Figure 2** shows how the data was filtered. A total of 16,487 projects remained after removing records outside the analysis years, incomplete entries, projects with unrealistic values, extreme outliers, and unrelated projects. The robustness

of outlier removal was assessed by conducting a sensitivity analysis. The interquartile range (IQR) multiplier was varied from 1.0× to 3.0× to examine how different cutoff thresholds affected dataset size. Across all tested thresholds, the total number of retained projects changed by only about 6%, indicating that the dataset remains stable and is not highly sensitive to the specific outlier cutoff chosen. Based on this stability and the conventional use of the 1.5× IQR rule in statistical practice, the 1.5× threshold was selected for the final dataset. In addition, a new feature engineered value, cost per mile, was added for each project by dividing the estimated construction cost by the project's length. This served as the target variable for the model and provided a consistent basis for comparing projects of different scales.

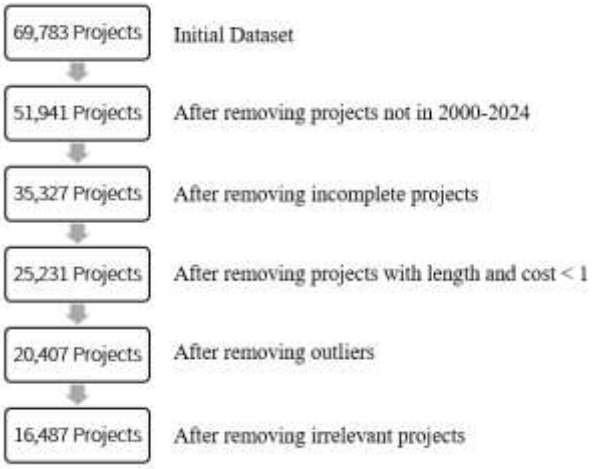


**Figure 2. Data filtering process for the TxDOT dataset, showing the reduction from 69,783 initial projects to 16,487 final projects.**

After collecting and cleaning the data, each project was assigned a label based on keywords found in the TYPE_OF_WORK and PROJ_CLASS feature columns. Both columns were manually reviewed by the authors due to overlapping and unstandardized descriptions, with different terminology often used to determine similar project types. Ultimately, a total of seven keywords were selected to act as the dataset labels: Overlay, Surfacing, Rehabilitation, Seal, Milling, Maintenance, and Repair. Because each project could belong to multiple categories, multi-hot encoding was used.

**Model Training and Selection.** Four machine learning models were used to predict project costs: XGBoost, Random Forest, Linear Regression, and AdaBoost. The choice of tree-based ensemble methods is supported by recent research that has demonstrated their effectiveness for cost estimation of public infrastructure projects. A study on pavement maintenance employed tree-based algorithms, including Random Forest, Extra Trees, and XGBoost, finding the approach to be suitable for the preliminary stage (Paik et al. 2025). Similarly, a study on bridge maintenance cost estimation compared several classifiers, including XGBoost and AdaBoost, to select an optimized algorithm for predicting unit costs, finding the XGBoost model to be superior to other models (Lee et al. 2024). Linear regression was included as a form of comparison. Another study noted how the linear regression model produced lower $R^2$ values and higher errors (Saeidlou and Ghadiminia 2024). Linear models can struggle to capture non-linear feature interactions that are common in construction cost data (Chen et al. 2025).

A pipeline was constructed using the cleaned dataset, following the outlier removal process described in the methodology. To handle remaining outliers, numerical features, including project length, letting year, economic indices, climate data, and AADT, were standardized using a 'RobustScaler'. Categorical features, including district and county names, were processed using one hot encoding to prevent models from assuming categorical order. The binary encoded project type features were passed through the pipeline without further transformation. **Table 1** provides a comprehensive explanation of the features used as model inputs. Following this, the dataset was split into two sets, 80% was reserved for model training and validation while the other 20% would serve as unseen testing data. For fair model comparison, each model underwent hyperparameter tuning before final evaluation. The tuning process was carried out using Optuna, a hyperparameter optimization framework, that evaluated different parameter combinations through repeated cross-validation. For each trial, the model was trained within the same preprocessing pipeline and assessed using Repeated K-Fold Cross-Validation with 10 splits and 3 repeats. The search space included parameters known to influence model behavior, such as tree depth, learning rate, and sampling ratios for XGBoost, and the number of trees and node-splitting rules for Random Forest and AdaBoost. Once the best parameters were selected, the corresponding model was refit on full training set and then evaluated with the test set. Model performance was then measured using $R^2$.

**Table 1. Features used in the model**

| Category | Feature Name | Description | Type | Source |
|---|---|---|---|---|
| Project Details | DISTRICT_NAME | Geographical district where project is located. | Categorical | TXDOT |
| | COUNTY_NAME | County where project is located. | Categorical | TXDOT |
| | Shape_Length | GIS geometry | Numerical | TXDOT |
| | PROJ_LENGTH | Engineering estimate | Numerical | TXDOT |
| | PROJ_ESTMTD_LET_YEAR | Date project was opened for bids | Numerical | TXDOT |
| Economic Indices | CPI | Consumer Price Index | Numerical | FRED |
| | PPI | Producer Price Index | Numerical | FRED |
| | PPI_construction | Producer Price Index | Numerical | FRED |
| Temperature Info | temperature min | Lowest temperature during month | Numerical | NOAA |
| | temperature mean | | Numerical | NOAA |
| | temperature max | | Numerical | |

| | | | | |
|---|---|---|---|---|
| | precipitation | | Numerical | NOAA |
| | cooling degree days | | Numerical | NOAA |
| | hot degree days | | Numerical | NOAA |
| | County Name AADT | The AADT for a county | Numerical | Feature-Engineered |
| Keywords | OVERLAY | Type of work | Binary | TXDOT |
| | SURFACING | Type of work | Binary | TXDOT |
| | REHABILITATION | Type of work | Binary | TXDOT |
| | SEAL | Type of work | Binary | TXDOT |
| | MILLING | Type of work | Binary | TXDOT |
| | MAINTENANCE | Type of work | Binary | TXDOT |
| | REPAIR | Type of work | Binary | TXDOT |

## RESULTS

Our results are separated into three sections: Data Analysis, Model Performance, and Feature Importance.

**Data Analysis.** An initial analysis revealed the project distribution, as seen in **Figure 3**. The analysis showed that project categories were not mutually exclusive, instead projects could contain multiple keyword matches. The diagonal represents projects with solely one type of work while the values outside the diagonal represent projects with both the x and y axis' type of work. As seen in the heatmap, seal treatment projects dominated the data with more than 10,000 records, accounting for approximately 75% of the dataset.

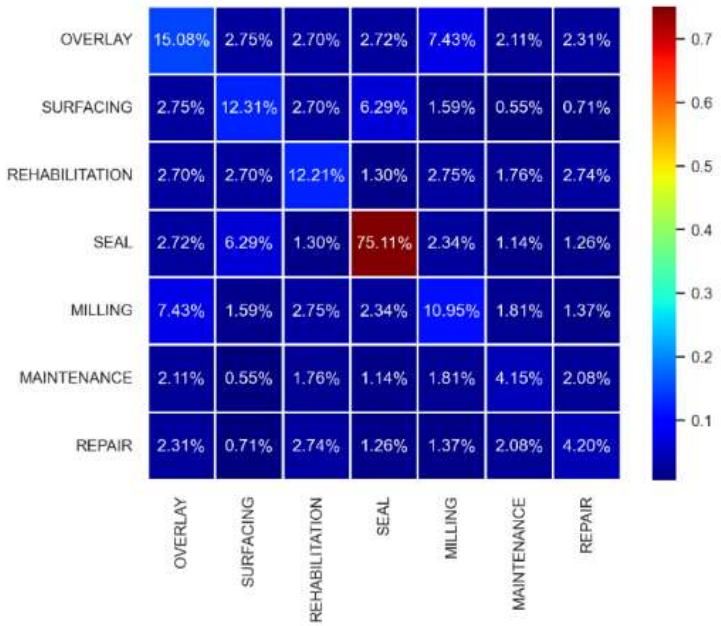


**Figure 3. Heatmap of project label distribution from the TxDOT dataset after categorization.**

**Model Performance.** The predictive performance of the four models varied. Cross-validation results show that the two ensemble methods, XGBoost and Random Forest, performed the strongest with an $R^2$ values of 0.7212 for XGBoost and 0.7217 for Random Forest. These results were consistent with the performance on the held-out test set that is summarized in **Error! Reference source not found.**. XGBoost achieved the highest test $R^2$ of 0.7393 with a test MAE of $31,513 and Random Forest produced a nearly identical test $R^2$ of 0.7377 and a test MAE of $31,618. The close performance between these two ensemble methods is a notable finding. The Median Absolute Error (MedAE) provides an important perspective, as it represents the typical prediction error for most projects and is not skewed by large outliers. XGBoost produced a MedAE of $9,989 and Random Forest produced a MedAE of $9,927, which corresponds to roughly one-sixth of $60,118, the median project cost per mile. The AdaBoost model performed better than Linear Regression achieving a test $R^2$ of 0.6122 with test MAE of $45,847 but remained below XGBoost and Random Forest performance. This pattern is consistent with its cross-validation results and suggests that although AdaBoost can capture some non-linear relationships, it is more sensitive to noise as it can place higher weights on weaker points and does not generalize as effectively as XGBoost or Random Forest for this dataset. The Linear Regression model's performance was weak, with a lower average test $R^2$ of 0.5635and an MAE of $49,586. This result is expected as project-level costs often involve complex non-linear dynamics that a simple linear model may not capture.

**Table 2. Test performance of all four models.**

| Model | $R^2$ | MAE | MedAE | Max Error |
|---|---|---|---|---|
| *Linear Regression* | 0.56 | 49,586 | 27,493 | 449,628 |
| *Random Forest* | 0.74 | 31,618 | 9,927 | 438,934 |
| *XGBoost* | 0.74 | 31,513 | 9,989 | 443,917 |
| *AdaBoost* | 0.61 | 45,847 | 24,613 | 547,073 |

The effect of class imbalance on model performance was evaluated by project type as seen in **Figure *4***. Seal projects dominate the dataset, yet their performance was not disproportionately better than the smaller categories. In fact, the distribution of $R^2$ values across categories was relatively uniform as even Milling, which accounted for a small sample of projects, achieved an $R^2$ close to that of Seal projects. This indicates that, despite the skewed distribution of project types, the model does not rely disproportionately on the majority class and generalizes reasonably well across categories.

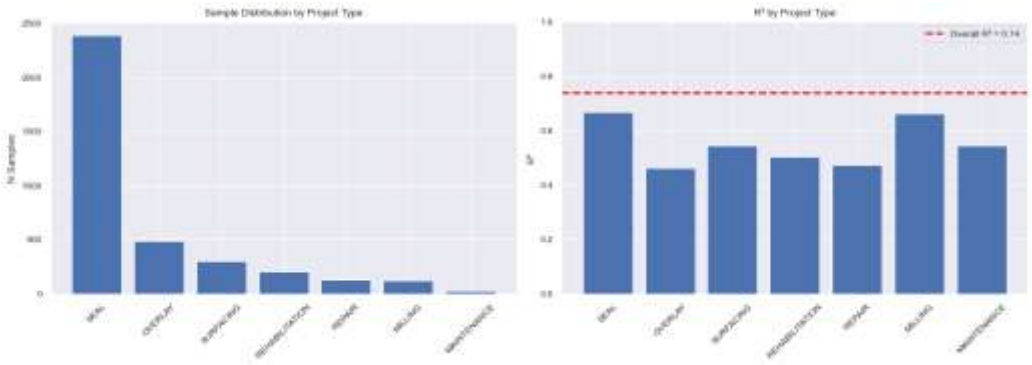

**Figure 4. Project category type performance.**

**Feature Importance.** Feature importance was evaluated using permutation importance on the best-performing XGBoost model to identify the key cost drivers. **Figure 5** shows the distribution of feature importance within the datasets. Project type had the largest effect on project cost, with the most frequent categories such as Overlay and Seal in highest ranking. This result is likely due to how project type reflects the resources, materials, and construction activities that a project will require which are strong indicators of overall cost. Some projects also include more than one keyword label showing that multiple types of work are being performed within the same project that can further vary costs. Additionally, project length ranked highly emphasizing the role of project details in shaping overall costs. AADT also made a meaningful contribution. Higher traffic volumes may indicate roads that experience more wear or require more effort to maintain. Cost indices were the next most influential features. This finding is consistent with the large inflation over the past years and reflects the importance of considering the present and forecasted economic conditions. Temperature data ranked low in importance. A plausible explanation is that, at the scale of Texas, temperature variation across counties is relatively small. There may not be enough variability to reveal meaningful relationships with project costs due to this.

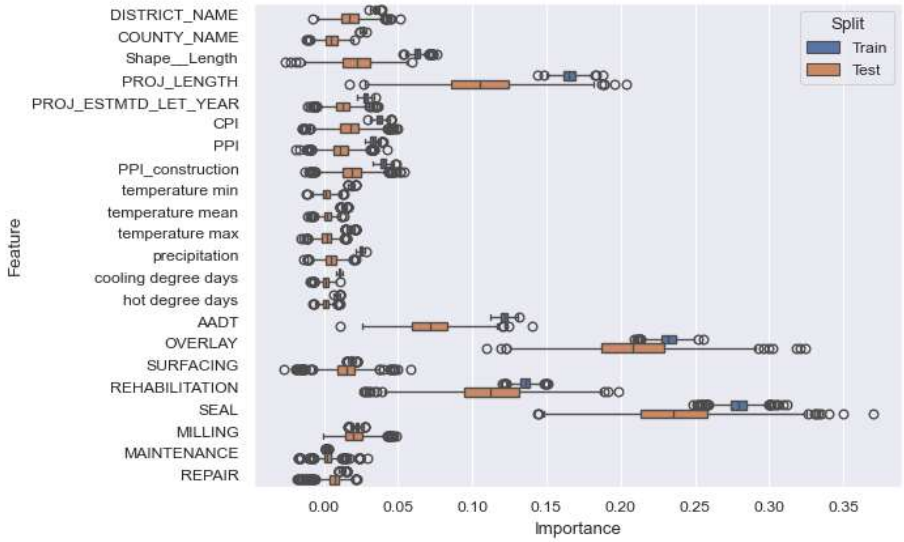


**Figure 5. Feature importance scores from the machine learning models.**

## CONCLUSION

This study developed and evaluated machine learning models for the preliminary construction cost estimate of highway rehabilitation projects by using project data from TxDOT combined with economic indices from FRED, climatic data from NOAA, and AADT data from TxDOT. By focusing on early-stage predictions, the work addresses a persistent challenge in transportation planning, improving the accuracy of estimates when the scope is known but project details remain unknown. Among the models evaluated, the XGBoost model demonstrated the strongest performance, followed closely by Random Forest. The results show these models are well-suited for capturing the complex, non-linear relationships that exist in project-level cost data. The inclusion of Linear Regression as a baseline reinforced this point, as its weaker performance highlights the limitations of linear methods for complex cost drivers. AdaBoost's higher errors suggest that not all ensemble methods are equally effective in this domain. A key finding of this research was the analysis of feature importance. The study confirmed that project-level attributes, AADT, and economic indices were influential features in predicting project costs. On the other hand, temperature data from NOAA had little to no influence on the model's predictions. Ultimately, the results demonstrate that XGBoost and Random Forest models can be utilized to generate cost predictions, supporting informed decision-making and the more efficient allocation of resources for future highway rehabilitation projects. The methodology developed here serves as a template for transportation agencies to leverage their own data to improve early-stage cost estimation. For future work, this model could be tested on data from other states to capture more diverse climates to further examine how temperature may influence project costs. Additional work could also incorporate methods such as SHAP values to examine feature-interaction effects and identify relationships that are not captured when features are assessed independently.

## ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support provided by the NSF Expand AI ARISE project under NSF Award No. 2434916, and the University Transportation Center for Railway Safety (UTCRS) at UTRGV under Grant No. 69A3552348340.

## REFERENCES

AACE International. 2020. 18R-97: Cost Estimate Classification System – As Applied in Engineering, Procurement, and Construction for the Process Industries. AACE International.

Awuku, B., E. Asa, E. Baffoe-Twum, and A. Essegbey. 2024. "Conceptual cost estimation of highway bid items – A systematic literature review." Engineering, Construction and Architectural Management, 31 (3): 1187–1221. Bradford, United Kingdom: Emerald Group Publishing Limited. https://doi.org/10.1108/ECAM-03-2022-0266.

Chen, H. 2022. "Why Does Data Quality Matter? How to Evaluate and Improve Data Quality for Machine Learning Systems." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Chen, L., S. S. Tiang, K. S. Chong, A. Sharma, T. Berghout, and W. H. Lim. 2025. "Predicting Construction Costs with Machine Learning: A Comparative Study on Ensemble and Linear Models." JEEEMI.

Cui, B., and H. Wang. 2025. “Predicting Asphalt Pavement Deterioration Under Climate Change Uncertainty Using Bayesian Neural Network.” IEEE Trans. Intell. Transport. Syst., 26 (1): 785–797. https://doi.org/10.1109/TITS.2024.3505237.

“Federal Reserve Economic Data | FRED | St. Louis Fed.” n.d. Accessed September 11, 2025. https://fred.stlouisfed.org/.

“Index of /pub/data/cirs/climdiv.” n.d. Accessed September 12, 2025. https://www.ncei.noaa.gov/pub/data/cirs/climdiv/.

Karimi, H., T. R. B. Taylor, G. B. Dadi, P. M. Goodrum, and C. Srinivasan. 2018. “Impact of Skilled Labor Availability on Construction Project Cost Performance.” J. Constr. Eng. Manage., 144 (7): 04018057. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001512.

Lee, G., T. Chang, and S. Chi. 2024. “Data-Driven Bridge Maintenance Cost Estimation Framework for Annual Expenditure Planning.” J. Manage. Eng., 40 (2): 04023068. https://doi.org/10.1061/JMENEA.MEENG-5706.

Mahdavian, A., A. Shojaei, M. Salem, J. S. Yuan, and A. A. Oloufa. 2021. “Data-Driven Predictive Modeling of Highway Construction Cost Items.” J. Constr. Eng. Manage., 147 (3): 04020180. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001991.

Myrvang, R., and C.-Y. A. Liu. 2025. “Beyond Traditional Methods: Enhancing Cost Escalation Forecasting in Commercial Construction amid Economic Turbulence.” J. Constr. Eng. Manage., 151 (2): 04024201. https://doi.org/10.1061/JCEMD4.COENG-15598.

Paik, Y., F. Chung, and B. Ashuri. 2025. “Preliminary Cost Estimation of Pavement Maintenance Projects through Machine Learning: Emphasis on Trees Algorithms.” J. Manage. Eng., 41 (4): 04025027. https://doi.org/10.1061/JMENEA.MEENG-6623.

Piryonesi, S. M., and T. El-Diraby. 2021. “Climate change impact on infrastructure: A machine learning solution for predicting pavement condition index.” Construction and Building Materials, 306: 124905. https://doi.org/10.1016/j.conbuildmat.2021.124905.

Project Management Institute. 2020. Practice Standard for Project Estimating.

Saeidlou, S., and N. Ghadiminia. 2024. “A construction cost estimation framework using DNN and validation unit.” Building Research & Information, 52 (1–2): 38–48. Routledge. https://doi.org/10.1080/09613218.2023.2196388.

Shrestha, R., T. Ko, and J. Lee. 2025. “Quantifying Project Uncertainties: Leveraging Historical Bid and Change Order Data for Automated Detection of Cost and Schedule Impacts in New Projects.” J. Constr. Eng. Manage., 151 (4): 04025017. https://doi.org/10.1061/JCEMD4.COENG-15689.

“TxDOT Annual Average Daily Traffic Counts (Public).” n.d. Accessed September 12, 2025. https://gis-txdot.opendata.arcgis.com/datasets/TXDOT::txdot-annual-average-daily-traffic-counts-public/about.

“TxDOT DCIS All Projects.” n.d. Accessed September 12, 2025. https://gis-txdot.opendata.arcgis.com/datasets/4a20e2a5f2ef462f9358d177d860df15.

Western Federal Lands Highway Division. 2023. “Estimating Handbook.” Federal Highway Administration (FHWA), U.S. Department of Transportation.