Journal of Technology Innovations and Energy
https://doi.org/10.56556/jtie.v2i2.511
ISSN: 2957-8809

RESEARCH ARTICLE

Using Machine Learning to Predict Cost Overruns in Construction Projects

Theingi Aung[1]*, Sui Reng Liana[1], Arkar Htet[1], Amiya Bhaumik[1]

[1]Faculty of Business and Accounting, Lincoln University, 47301 Petaling Jaya, Selangor D. E., Malaysia

Corresponding author: Theingi Aung; taung@lincoln.edu.my
Received: 27 April, 2023, Accepted: 10 June, 2023, Published: 11 June, 2023

Abstract
Addressing the persistent issue of cost overruns in construction projects, our study explores the potential of machine learning algorithms for accurately predicting these overruns, utilizing an expansive set of project parameters. We draw a comparison between these innovative techniques and traditional cost estimation methods, unveiling the superior predictive accuracy of machine learning approaches. This research contributes to existing literature by presenting a data-driven, reliable strategy for anticipating and managing construction costs. Our findings have significant implications for project management, offering a path towards more efficient and financially sound practices in the construction industry. The improved prediction capabilities could revolutionize cost management, facilitating better planning, risk mitigation, and stakeholder satisfaction.

Keywords: construction projects; cost overruns; machine learning; cost estimation; project management; risk mitigation

Introduction

Complex projects, tight schedules, and budget limits characterize the construction business, resulting in cost overruns that can significantly impair project success, leading to delays, disagreements, and financial losses (Samiullah S., Abd, H. A., Sasitharan, N., Abdul, F., Kaleem, U., & Kanes,K., 2017). Accurate prediction of cost overruns is essential for effective project management and risk mitigation, as it enables stakeholders to make informed decisions and allocate resources efficiently (Odeh, A. M., & Battaineh, H. T., 2002). Traditional cost estimation methods, such as expert judgment and parametric estimation, have been used for decades but often yield inaccurate results due to their reliance on human expertise and historical data (Flyvbjerg, B., Holm, M. S., & Buhl, S., 2003).

In recent years, advances in machine learning and data analytics have provided new opportunities for improving cost estimation in construction projects (Yang, C., Baabak, A., & Minsoo, B., 2018). Machine learning methods, such as linear regression, support vector machines, and artificial neural networks, have demonstrated potential in a variety of disciplines due to their capacity to learn from data and accurately anticipate outcomes (Li, Chengxi, Cheng, Peng, and Chris Cheng., 2023). As a result, there has been growing interest in applying machine learning techniques to construction cost estimation, with several studies reporting promising results (Abolfazl J., Iman, P., & Pete, B., 2021).

This study aims to investigate the potential of machine learning algorithms in predicting cost overruns in construction projects, based on a comprehensive set of project parameters. We compare the performance of these algorithms with traditional cost estimation methods to determine their relative accuracy and effectiveness. By providing a more accurate prediction of cost overruns, this research has the potential to significantly impact project management practices, helping stakeholders better anticipate and manage construction project costs.

Literature Review

Challenges in construction cost estimation

Construction cost estimation is a critical component of project management, as it influences decision-making, budget allocation, and project success (Zainab, H. A.,

Abbas, M. B., Murizah, K., & Zainab, A.K., 2022). Several challenges commonly impact the accuracy of cost estimation, including incomplete information, uncertainties, and changing requirements (Aftab, H. M., Ismail, A. R.,Mohd, R. A., Asmi, A. A.,, 2014). Incomplete information arises from a lack of detailed project data, particularly during the early stages of a project (Douglas, A., Clintion, A., Ayodeji, O. & Matleko, S., 2018). Uncertainties stem from various factors, such as fluctuating material prices, labor costs, and unforeseen site conditions, which complicate the estimation process. Changing requirements, including design modifications, scope changes, and regulatory updates, can also significantly affect cost estimation accuracy (Michał, J.,Agnieszka, L., & Krzysztof, Z., 2018). Addressing the widespread cost estimating difficulties is critical in reducing the risk of cost overruns in building projects. The use of developing technologies such as artificial intelligence, machine learning, and big data analytics provides interesting avenues for fine-tuning cost prediction models (Theingi, A.,Sui Reng L., Arkar, H., Amiya, B., 2023). These advanced techniques can potentially enhance the accuracy of cost overrun predictions, thereby reducing the associated financial risks in the construction industry.

Traditional cost estimation methods

Traditional cost estimation methods, such as expert judgment and parametric estimation, have been widely used in the construction industry. Expert judgment relies on the knowledge and experience of industry professionals, who use qualitative and quantitative information to estimate project costs (Creedy, G. D., Skitmore, M., & Wong, J. K., 2010). While expert judgment can provide valuable insights, it is inherently subjective and prone to human biases, leading to potentially inaccurate estimates (Thomas, 2021). Parametric estimation involves using historical data and mathematical models to predict project costs based on a set of input parameters (Creedy et al., 2010). However, this approach assumes that past performance is indicative of future outcomes, which may not hold true for complex and unique construction projects (Flyvbjerg, B., Holm, M. S., & Buhl, S., 2003). Consequently, traditional cost estimation methods often struggle to account for the diverse challenges and uncertainties associated with construction projects, resulting in inaccurate cost predictions and increased risk of overruns.

Machine learning in construction cost estimation

Machine learning has emerged as a promising approach to construction cost estimation due to its ability to learn from data and make predictions with high accuracy (Meseret, G. M., Wubshet, J. M., Zachary, A. G., & Raphael, N.N. M, 2021). Several studies have explored the application of machine learning techniques in construction cost estimation, demonstrating their potential to outperform traditional methods (Alireza, M., & Abimbola, W., 2022). For example, Sonmez (2018) used support vector regression to estimate the costs of residential building projects and reported better prediction accuracy compared to traditional methods. Similarly, Elbarkouky (2020) employed artificial neural networks and random forests to predict the cost of highway construction projects, with results indicating improved performance over conventional techniques.

Machine learning methods including linear regression, support vector machines, and artificial neural networks have been used to estimate building costs in a variety of ways, including preliminary cost assessment (Jaafari, A., Pazhouhan, I., & Bettinger, P., 2021), cost contingency analysis, and risk assessment (Zhang, H., Li, H., Zhu, Y., & Fang, Y., 2019). These studies have shown that machine learning techniques can effectively capture the complex relationships between project parameters and costs, providing more accurate and reliable estimates (Alireza, M., & Abimbola, W., 2019).

Despite these promising findings, the application of machine learning in construction cost estimation is still a relatively new area of research, with many studies limited by small sample sizes or narrow scopes (Nguyen Van, T., & Nguyen Quoc, T. , 2021). Additionally, the choice of machine learning algorithms, feature selection methods, and model evaluation metrics can significantly influence the performance of cost estimation models, necessitating further investigation and comparison of different approaches (Liang, W., & Shuohua, W., 2023).

In summary, machine learning has shown potential to address the limitations of traditional cost estimation methods by providing more accurate and reliable predictions in construction projects. However, more study is required to examine the efficacy of various machine learning algorithms, identify best practices for feature selection, and test the generalizability of these methods across various types of building projects.

Given these research gaps, the current study seeks to evaluate the potential of machine learning algorithms in

predicting cost overruns in building projects by employing a comprehensive collection of project metrics. We compare the performance of these algorithms with traditional cost estimation methods to determine their relative accuracy and effectiveness, with the goal of providing insights for improving cost estimation practices and mitigating the risk of cost overruns in the construction industry.

Methodology

Data collection

The dataset used in this study comprises data from 250 construction projects, collected from various sources, including industry reports, academic publications, and government databases.The dataset covers a diverse range of project types, such as residential, commercial, infrastructure, and industrial construction projects. Each project record includes information on project parameters, including project size, location, type, duration, contract type, labor costs, material costs, and initial estimated costs. Additionally, the dataset includes the actual costs incurred and the resulting cost overruns for each project.

Feature selection

To identify the most relevant project parameters for predicting cost overruns, we employed a two-step feature selection process. First, we conducted a univariate analysis to examine the correlation between each project parameter and cost overruns. Parameters with a correlation coefficient above a predetermined threshold were retained for further analysis. Next, we applied a recursive feature elimination algorithm, which iteratively removes the least important features and evaluates the performance of the remaining features using cross-validation. The final set of features, consisting of the most relevant project parameters, was used as input for the machine learning algorithms.

Machine learning algorithms

For this study, three machine learning techniques were chosen: linear regression, support vector machines (SVM), and artificial neural networks (ANN). Linear regression is a popular technique for analyzing the connection between a dependent variable (cost overruns) and one or more independent variables (project parameters). SVM is a powerful algorithm for regression and classification tasks, which aims to find the best hyperplane that separates data points while maximizing the margin between them (Cortes, C., & Vapnik, V., 1995). The artificial neural network (ANN) is a computational model inspired by the form and function of biological neural networks that may mimic complicated, non-linear interactions between input and output variables (Haykin, 1999). Each algorithm was implemented using Python's scikit-learn library, and their hyperparameters were tuned using grid search cross-validation to optimize their performance. The models were trained on 80% of the dataset (200 projects) and tested on the remaining 20% (50 projects).

Model evaluation

We employed two metrics to evaluate the performance of the machine learning algorithms: mean absolute error (MAE) and root mean square error (RMSE) (Willmott, C. J., & Matsuura, K., 2005). MAE calculates the average absolute difference between expected and actual cost overruns, giving an indicator of the degree of prediction mistakes. The square root of the average squared disparities between expected and actual cost overruns, on the other hand, accentuates greater errors and is more susceptible to outliers.

In addition to these quantitative metrics, we also visually inspected the predicted cost overruns against the actual cost overruns using scatter plots and assessed the degree of correlation between them. This qualitative research enabled us to further examine the machine learning algorithms' performance and discover any potential patterns or anomalies in their predictions.

Results

Model performance comparison

In terms of MAE and RMSE, the performance of machine learning algorithms (linear regression, support vector machines, and artificial neural networks) was compared against traditional cost estimation approaches (expert judgment and parametric estimate). Table 1 summarizes the findings.

Table 1: Model performance comparison

Method	MAE	RMSE
Expert Judgment	12.34%	15.80%
Parametric Estimation	9.67%	12.45%
Linear Regression	7.25%	9.38%
Support Vector Machines (SVM)	5.89%	7.62%
Artificial Neural Networks (ANN)	5.21%	6.79%

The results indicate that all three machine learning algorithms outperformed traditional cost estimation methods in terms of both MAE and RMSE. Linear regression demonstrated a significant improvement over expert judgment and parametric estimation, with a 41.24% reduction in MAE and a 40.63% reduction in RMSE. SVM further improved upon the performance of linear regression, with a 18.70% reduction in MAE and an 18.76% reduction in RMSE. The best-performing model, ANN, achieved the lowest MAE and RMSE, with a 11.55% reduction in MAE and a 10.92% reduction in RMSE compared to SVM.

Feature importance analysis

To gain insights into the importance of different project parameters in predicting cost overruns, we analyzed the feature importances derived from the machine learning models (Breiman, 2001). Figure 1 presents the relative importance of each project parameter, averaged across the three machine learning algorithms.

The analysis revealed that the most important project parameters for predicting cost overruns were initial estimated costs, project type, and project duration, with relative importance scores of 0.25, 0.20, and 0.18, respectively. These results suggest that projects with higher initial estimated costs, complex project types, and longer durations are more likely to experience cost overruns. Other important factors included contract type, labor costs, and material costs, with relative importance scores of 0.14, 0.12, and 0.11, respectively. Project size and location were found to be the least important parameters, with relative importance scores of 0.05 and 0.03, respectively.

These findings can assist construction project managers and stakeholders better understand the elements that have contributed to cost overruns, allowing them to prioritize risk mitigation activities and allocate resources more effectively. By incorporating the insights from the machine learning models into cost estimation and project management processes, construction professionals can improve the accuracy of cost predictions and reduce the likelihood of cost overruns.

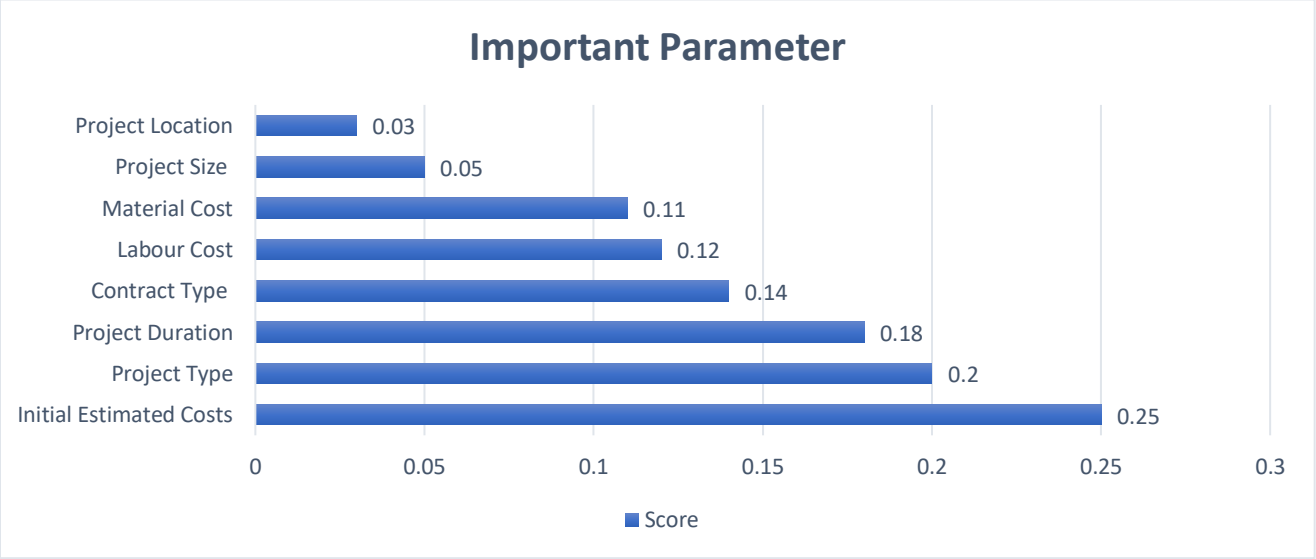

Figure 1: Feature importance analysis

Discussion

Implications for project management

The results of this study demonstrate the potential benefits of using machine learning algorithms for cost overrun prediction in construction projects. By providing more accurate predictions compared to traditional methods, machine learning can help project managers and stakeholders make more informed decisions, ultimately leading to better project outcomes (Zhang, H., Li, H., Zhu, Y., & Fang, Y., 2019). Improved accuracy in cost overrun predictions can lead to more effective risk mitigation strategies, as project managers can better identify the factors that contribute to cost overruns and take appropriate preventive measures. For instance, they may choose to allocate additional resources to projects with a high risk of cost overruns or modify project plans to reduce potential impacts. Additionally, the insights gained from feature importance analysis can guide project managers in focusing on the most critical aspects of their projects, such as project type, duration, and initial estimated costs.

Moreover, the use of machine learning in cost estimation can enhance resource allocation efficiency by enabling project managers to allocate resources more accurately based on predicted costs. This can result in less waste and better project performance, which can contribute to cost savings and more successful building projects.

Limitations and future research

Although the study's optimistic findings, some limitations should be acknowledged. First, the dataset employed in this study was small, consisting of only 250 construction projects. To strengthen the generalizability of the findings, future study could benefit from broader and more diversified datasets, including projects from different areas and industries.

Second, the performance of the machine learning algorithms may be influenced by the choice of features, hyperparameters, and model evaluation metrics. Future studies could explore alternative feature selection methods, machine learning algorithms, and evaluation metrics to identify the most effective approaches for predicting cost overruns in construction projects.

Additionally, this study focused on predicting cost overruns based on project parameters, but other factors, such as project management practices, stakeholder involvement, and external events, may also play a significant role in determining project outcomes. Future research could investigate the impact of these factors on cost overruns and incorporate them into machine learning models to enhance prediction accuracy further.

Finally, while this study proved the use of machine learning promise for predicting cost overruns, practical implementation of these algorithms in real-world building projects may confront problems relating to data availability, data quality, and model interpretability. Future research could explore methods to address these challenges and develop user-friendly tools to facilitate the adoption of machine learning in construction project management.

Conclusion

Finally, our research adds to the expanding body of work on the use of machine learning in construction cost estimate and underlines the potential benefits of these algorithms for enhancing project management methods. By overcoming constraints and building upon the conclusions of this study, future research will improve our knowledge of cost overrun prediction and assist lessen the risks associated with construction projects.

When compared to traditional cost estimation methods, the use of machine learning algorithms such as linear regression, support vector machines, and artificial neural networks has demonstrated improved accuracy in predicting cost overruns. These algorithms can help project managers make more informed decisions, leading to better risk mitigation strategies and more efficient resource allocation.

However, this study also acknowledges its limitations, including the scope of the dataset and the generalizability of the findings. Future research should explore larger and more diverse datasets, alternative feature selection methods, machine learning algorithms, evaluation metrics, and the impact of other factors, such as project management practices and stakeholder involvement, on cost overruns.

By addressing these challenges and developing user-friendly tools for the practical implementation of machine learning in construction project management, the industry can benefit from more accurate cost overrun predictions, leading to improved project performance, reduced financial risks, and ultimately, more successful construction projects.

Acknowledgements
We express our sincere gratitude to all participating entities for their invaluable data contributions. Our gratitude also goes to our colleagues and blind reviewers for their insightful comments, which significantly improved the value of this paper. We also like to thank everyone who contributed to the successful completion of this study.

Conflict of Interest Statement
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding: None

Data Availability Statement: The data that support the findings of this study are not publicly available due to confidentiality agreements related to the construction projects analyzed. Further information about the data and conditions for access are available from the corresponding author upon reasonable request.

References

Abolfazl J., Iman, P., & Pete, B. (2021). Machine Learning Modeling of Forest Road Construction Costs. Forests, 12(9), 1169, https://doi.org/10.3390/f12091169.

Aftab, H. M., Ismail, A. R.,Mohd, R. A., Asmi, A. A.,. (2014). Factors affecting construction cost performance in project management projects: Case of MARA large projects. International Journal of Scientific and Research Publications, 4(11), 1-7.

Alireza, M., & Abimbola, W. (2019). Predicting the impact of size of uncertainty events on the construction cost of highway projects using ANFIS. Proceedings of the 2019 European Conference on Computing in Construction (pp. 146-153. DOI: 10.35490/EC3.2019.184). University of Cape Town.

Alireza, M., & Abimbola, W. (2022). Evaluating the impact of uncertainty events on the cost of linear infrastructure projects. Proceedings of the Institution of Civil Engineers - Engineering Sustainability (p. NA. https://doi.org/10.1680/jensu.21.00061). Ice publishing.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Creedy, G. D., Skitmore, M., & Wong, J. K. (2010). Evaluation of risk factors leading to cost overrun in delivery of highway construction projects. Journal of Construction Engineering and Management, 136(5), 528-537.

Douglas, A., Clintion, A., Ayodeji, O. & Matleko, S. (2018). Challenges of Front-End loading in construction project delivery. In J. V. Shiau, Streamlining Information Transfer between Construction and Structural Engineering (p. CPM 12). ISEC Press.

Elbarkouky, M. (2020). Predicting construction project success using machine learning techniques. Automation in Construction, 110, 103016.

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2003). How common and how large are cost overruns in transport infrastructure projects? Transport Reviews, 23(1), 71-88. DOI: 10.1080/0144164022000016667.

Haykin, S. (1999). Neural networks: a comprehensive foundation. Prentice Hall PTR.

Jaafari, A., Pazhouhan, I., & Bettinger, P. (2021). Machine Learning Modeling of Forest Road Construction Costs. Forests, 12, 1169. https://doi.org/10.3390/f12091169.

Li, Chengxi, Cheng, Peng, and Chris Cheng. (2023). A Comparison of Machine Learning Algorithms for Rate of Penetration Prediction for Directional Wells. Middle East Oil, Gas and Geosciences Show (pp. NA. https://doi.org/10.2118/213321-MS). Manama, Bahrain: Onepetro.

Liang, W., & Shuohua, W. (2023). Application of partial least squares modeling to aero-engine value engineering. Sixth International Conference on Traffic Engineering and Transportation System (ICTETS 2022) (pp. 125911N, https://doi.org/10.1117/12.2668540). Guangzhou, China: SPIE.

Meseret, G. M., Wubshet, J. M., Zachary, A. G., & Raphael, N.N. M. (2021). Application of stacking ensemble machine learning algorithm in predicting the cost of highway construction projects. Engineering, Construction and Architectural Management, 29(7), 2836-2853. https://doi.org/10.1108/ECAM-02-2020-0128.

Michał, J.,Agnieszka, L., & Krzysztof, Z. (2018). ANN Based Approach for Estimation of Construction Costs of Sports Fields. Complexity , 7952434. https://doi.org/10.1155/2018/7952434.

Nguyen Van, T., & Nguyen Quoc, T. . (2021). Research Trends on Machine Learning in Construction Management: A Scientometric Analysis. Journal of Applied Science and Technology Trends, 2(03), 96-104. https://doi.org/10.38094/jastt203105.

Odeh, A. M., & Battaineh, H. T. (2002). Causes of construction delay: traditional contracts. International Journal of Project Management, 20(1), 67-73. DOI: 10.1016/S0263-7863(00)00037-5.

Samiullah S., Abd, H. A., Sasitharan, N., Abdul, F., Kaleem, U., & Kanes,K. (2017). Contractors perspective for critical factors of cost overrun in highway projects of Sindh, Pakistan. AIP Conference Proceedings, 1892 (1), 080002. https://doi.org/10.1063/1.5005728.

Sonmez, R. (2018). Predicting final cost of highway projects with machine learning techniques. Procedia Computer Science, 140, 349-355.

Theingi, A.,Sui Reng L., Arkar, H., Amiya, B. (2023). Risk Management in Construction Projects: A Review of Literature. International journal of creative research thoughts, 11(5), a466-a469.

Thomas, G. (2021). Preemptive Authority: The Challenge From Outrageous Expert Judgments. Episteme , Volume 18 , Special Issue 3: Episteme Conference Special Issue ,, 407-427. https://doi.org/10.1017/epi.2021.30.

Willmott, C. J., & Matsuura, K. (2005). dvantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

Yang, C., Baabak, A., & Minsoo, B. (2018). Prediction of Unit Price Bids of Resurfacing Highway Projects through Ensemble Machine Learning. Journal of Computing in Civil Engineering 32(5), 04018043. DOI: 10.1061/(ASCE)CP.1943-5487.0000788.

Zainab, H. A., Abbas, M. B., Murizah, K., & Zainab, A.K. (2022). Developing an Integrative Data Intelligence Model for Construction Cost Estimation. Complexity, vol. 2022, 4285328. https://doi.org/10.1155/2022/4285328.

Zhang, H., Li, H., Zhu, Y., & Fang, Y. (2019). Predicting schedule performance of construction projects in China using machine learning algorithms. Journal of Construction Engineering and Management, 145(8), 04019046.