Westcliff International Journal of Applied Research
WIJAR

Westcliff International Journal of Applied Research
Vol. 7, No. 1. Fall 2023
https://doi.org/10.47670/wuwijar202371AMHKA

Improving Project Budgeting Systems by Developing Machine Learning Models

Dr. Ahmed Masry Hashala
Westcliff University

Dr. Kate S Andrews
Westcliff University

Abstract

The lack of an efficient budgeting system makes it more difficult for a business to satisfactorily execute projects or gain new business. To improve the accuracy of budgeting using the classical approach, a dynamic system is required. Building dynamic systems that apply machine learning techniques can support companies in improving their budgeting system. This quantitative study built five machine learning regression models: multiple linear regression, artificial neural network, support vector machine, k-nearest neighbors, and random forest. The five built models were used to predict the closing costs of 552 industrial automation projects that were carried out in Africa and the Middle East. Using root mean square error, the model forecast precision was compared to that of the classical system. The outcome shows that there is a significant difference between machine learning models and classical systems. Therefore, the use of machine learning techniques can improve the accuracy for businesses of their budgeting system.

Keywords: Artificial intelligence, machine learning, artificial neural network, support vector machine, random forest, k-nearest neighbors, multiple linear regression, projects, budgeting systems, benchmark

Introduction

The goal of this research project was to examine the potential of machine learning (ML) to improve the way industrial automation projects are budgeted. Companies develop a project budget during the bidding process with the understanding that the expenses will decide the selling price after adding a profit margin. Typically, the time allowed for the bidding process is not enough to accurately forecast future costs. As a result, the project costs may be calculated inaccurately using the existing budgeting system.

It is essential to determine the project budget accurately. Overestimating the selling price could result in missed chances. On the other hand, underestimating the project requirements might result in failing to satisfy clients. Therefore, maintaining business continuity implies the consistent need for an accurate project budget (Eyibio & Daniel, 2020).

The key to long-term corporate success is meeting customers' expectations. Companies may win bids but then without a sufficient budget are not able to fulfill consumers' needs. Contractors may make mistakes when estimating project budgets due to a lack of time during the bidding process. As a result, a quick and accurate budgeting system is required.

The budgeting system should take into consideration the lifecycle of projects and the nature of the industry (Kwon & Kang, 2018). Sophisticated technology initiatives make projects scope more variable. Accordingly, budgeting may not accurately respond to rapidly changing projects (Miandoab & Gharehchopogh, 2016).

Additionally, every project is unique in terms of timeframe, scope, risks, and other aspects. Although complex projects enhance the reputations of businesses, there is a significant risk that they may run over budget or behind schedule (Browning, 2019). Generally, projects are not completely defined during the initial phase. During the project execution, additional charges might be incurred. Businesses pursue minimizing the gap between the initial budget and the closing cost. Therefore, the budgeting process should consider the different aspects of a project, including a project timeframe, scope of work, market circumstances, and project complexity.

This research addresses a business need for many companies to enhance their ability to win projects with high customers' satisfaction. Unfortunately, there is not much research in using machine learning in the industrial automation budgeting system. The complexity and uniqueness of the industrial automation projects makes it difficult for companies to create a budgeting benchmark using the classical system. Therefore, this research utilizes machine learning techniques to cover this gap.

In the classical approach, the scope of work is divided into small tasks called a work breakdown structure (WBS) (Devi & Reddy, 2012). The project evaluates the adequate resources required to complete the WBS (Cerezo-narv et al., 2020). The basis of the estimate should include material, labor, freight, taxes, currency exchange, cost of finance and any other costs associated with the project execution (Greco, 2018). Fundamentally, material and engineering are a project two main cost components. For example, materials account for more than half of the project costs for the construction businesses (Mahagaonkar & Kelkar, 2017). All logistics activities, including shipping, transportation, and customs clearance, should be considered in the budget.

Initially, the project scope may not be precisely defined. Therefore, businesses may produce estimates of an opening budget (Srinivasan et al., 2021). During the project execution phase, a final solution is developed and submitted along with the final bill of material. Consequently, additional materials may be required to complete the technical solution and ensure that the system functions correctly.

According to a statistical analysis conducted in Hong Kong with a sample of projects with contracts value USD 14 billion, 47% of projects deviated from the planned budget (Love et al., 2019). Therefore, the budget should reserve an amount for risks and contingencies. Project management information systems can help with the risk quantification to be included in the budget (Besouw & Bond-Barnard, 2021).

Given the limits of the classical bidding process, ML regression helps to increase the accuracy of the budgeting system. ML models create a nonlinear link between dependent and independent variables (Antunes et al., 2021).ML provides regression models that may improve the precision of the budgeting system without going into the time-consuming classical method. ML uses data and algorithms to mimic the human brain to improve forecasting accuracy. ML technology is used in many fields: pattern recognition, medical applications, risk assessment, finance, and entertainment (El Naqa, 2015).

ML is classified as supervised, unsupervised, and reinforcement learning. Supervised techniques build algorithms from existing cases to automate decision-making (Burkart & Huber, 2021). The supervised learning technique develops prediction models by learning from many training instances, each containing a label identifying the output (Zhou, 2018).

Regression models use independent variables as inputs to forecast a project budget. ML techniques apply various types of regression: multiple linear regression (MLR), artificial neural network (ANN), support vector regression (SVR), k-nearest neighbor (KNN), and random forest (RF). These ML regression models may perform better than the classical budgeting method.

The different regression models draw a connection between cost factors and the projected budget. The ANN technique is highly effective in developing a mathematical equation used in predicting costs (Abd & Naseef, 2019; Balali et al., 2020; Tijanić et al., 2020). Among ML methods, the ANN and SVM techniques yield excellent estimation results (Hassim et al., 2018; Mohammed et al., 2021)

Additionally, ML can build a benchmark for future projects to increase computation speed as well as improving precision. A comparison of the classical and ML methodologies was done in order to evaluate the efficacy of the recently developed ML models.

Methods and Materials

The research entails developing ML models to forecast the project overall cost. The cost categories were regarded as the models input or independent variables. The actual project cost was regarded as the dependent variable.

The cost categories were considered as independent variables of the models. Under each cost category, there were expense items that stated the cost incurred under each item. The cost categories are hardware, engineering, logistics, cost of finance, risks, installation, and others. The dependent variable was calculated based on the aggregation of all actual cost items of each associated project. The initial budget of cost items was compared with the actual project cost upon closure of the sample projects to determine the actual budget error.

The initial budget and the actual cost were extracted from secondary data. Typically, businesses record the budget of projects on a project management information system (PMIS). Throughout the course of the project, the budget gets revised. Based on the project execution circumstances, some budget amounts can be moved from one cost center to another. In some cases, an additional budget is required to complete the project.

The secondary data were split into 70% for training and 30% for testing. The ML regression models used the project training data to update their algorithms. Consequently, the testing data set was used to verify the efficiency of the ML prediction models. Each ML regression model efficiency was compared with the actual budget error to determine which model provided the most accurate prediction.

The population of the secondary data consisted of industrial automation projects that had been completed during the previous 5 years. The sample was gathered from completed and closed projects that were received from 552 projects executed in five different countries. Every project differed in terms of its scale, schedule, stakeholder requirements, and environmental restrictions. However, the life cycle, engineering, logistics, and funding of projects were comparable.

The information extracted from the PMIS included both the initial budget for cost categories as well as the actual expenses that had already been incurred to complete the projects. The ultimate cost was used as the dependent variable to develop the ML models. The cost categories, including material, engineering, logistics, finance, risks, services, overheads, installation, and others, are considered the independent variables in the regression models.

Each cost category included sub sectors that are called cost centers. The cost centers were aggregated to comprise the cost categories. The data were extracted from the PMIS in transactional formats. Therefore, the data were validated and reshaped to be imported in the ML software. The regression models were developed using the R language to build the ML regression models: MLR, ANN, SVR, KNN, and RF.

Because the data contained information from different executed projects, the cost categories had different scales. For instance, the hardware budget could be given more consideration in one project than the logistical budget. Considering this, data preprocessing is crucial for enhancing ML performance. Scaling the various characteristics helped the models to run faster and perform better.

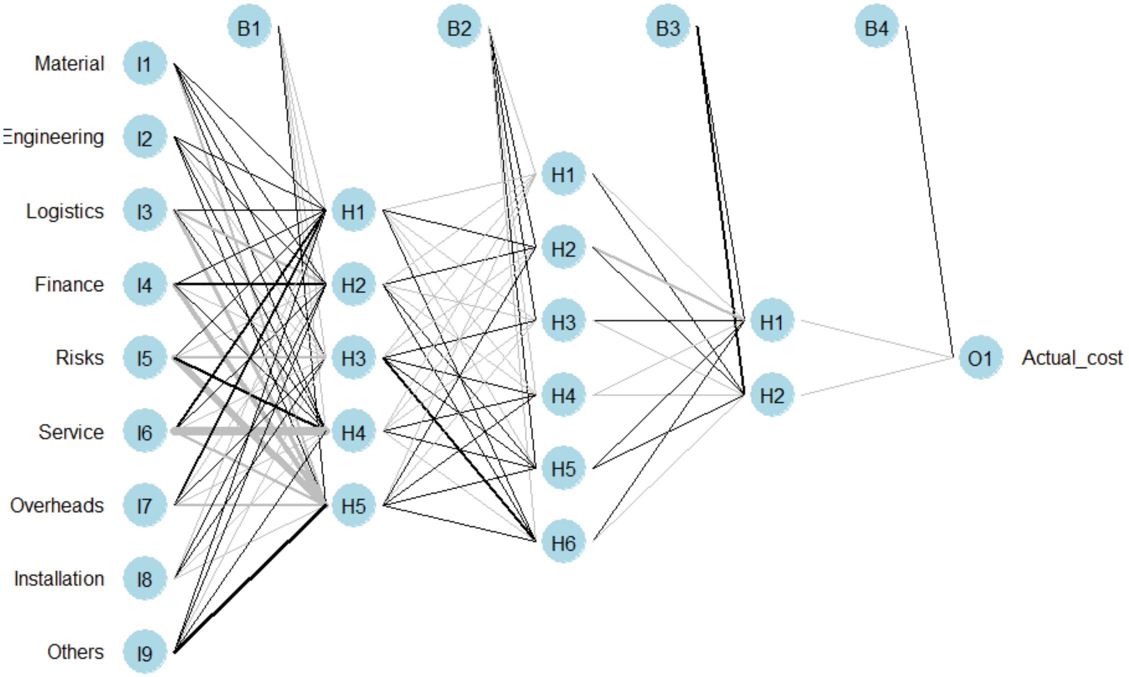
In a dataset, the information from several projects was kept as 30% of the dataset for testing, while 70% was used for training. The data of testing and training datasets were selected on a random basis from the secondary data. The algorithms developed independent variable weight using the training data. The performance of the models was evaluated using the root mean square error (RMSE) between the forecasted budgets and the actual costs listed in the testing datasets.

Data preprocessing is the next step in the modelling process after importing the data into R-Studio. The data were scaled. The code developed five groups of the regression models, MLR, ANN, SVR, KNN, and RF models. The code trained the models using different R-libraries, and then listed the regression outputs in a matrix for comparison. The optimal configurations of the regression models were determined by a tuning grid or nested loops. Accordingly, the forecast was created for each model using the testing dataset. The RMSE between the forecasted budget and the actual cost was calculated and listed in a matrix for comparison.

ANN models had been created using a variety of techniques. Model_ANN1 is run with just one hidden layer using the default settings. Two more ANN models were created and tuned using nested loops to determine the best fit to the training and the testing datasets. Model_ANN_Train showed the minimum RMSE between the training dataset and the actual cost. Nevertheless, the RMSE of the testing dataset of the model Model_ANN_Train showed a sign of overfitting. In other words, the Model_ANN_Train showed high ability to fit to the training dataset, however, it showed low generalizability performance.

In order to consider the generalizability, Model_ANN_Test was created and tuned to fit with the testing dataset. Figure 1 shows the configuration of Model_ANN_Test. The Caret library was used to develop ANN model called Model_Caret_ANN1. The H2O library for deep learning, a subset of the neural network, was used to create an ANN model called Model_H2O_DEEP1.

Figure 1
Artificial Neural Network Model



Note. The graph shows ANN with three hidden layers with neurons (5, 6, 2).

The support vector regression models were created. Model_SVM1 was built using E1071 library. Model_Caret_SVM1 and Model_Caret_SVM2 were created by the Caret library using svmRadial and svmLinear methods respectively.

KNN models were created using different libraries. Model_KNN1 has five nearest neighbors in the regression model configuration.

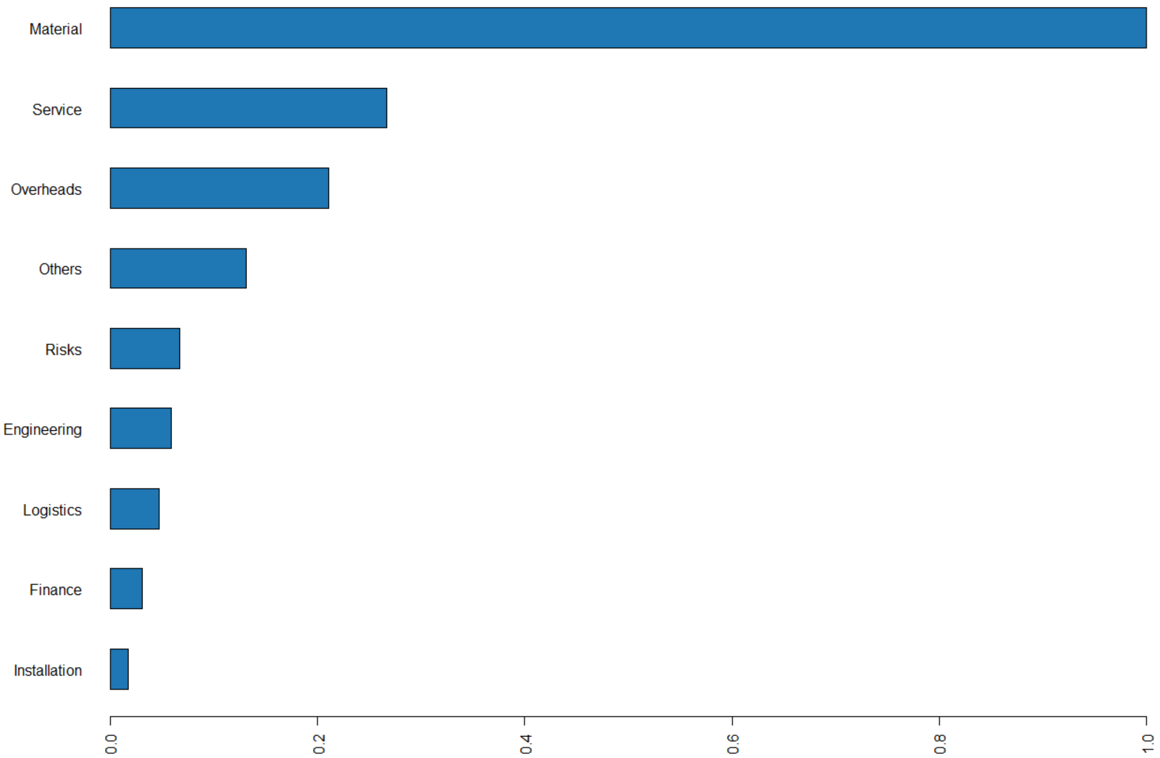
Caret library was used to develop Model_Caret_KNN1 and Model_Caret_KNN2. Model_Caret_KNN1 demonstrated higher precision in fitting the model to the training dataset. On the other hand, Model_KNN1 demonstrated a higher generalization competence.

Random forest regression models, Model_RF1, Model_Caret_RF1, and Model_H2O_RF1, were created using libraries Caret, Random Forest, and H2O respectively. Model_Caret_RF2 was created using a quantile random forest. The code showed similar accuracy between Model_RF1 and Model_Caret_RF1 as advantageous for extrapolation and fitting the training dataset.

The algorithm determined the importance of the variables to specify the most critical variables that contributed to the prediction. A budget system should concentrate on variables of the highest importance. In contrast, the less important variables may consume less time while preparing the budget. Given that bidding phases are time-limited, concentrating on the most important cost categories may be advantageous using the variable significance map. The map of variable importance of the model Model_H2O_RF1 is shown in Figure 2. According to the model, the most significant independent variables were material, service, and overheads.

Figure 2

Importance of Independent Variables



Note. Model_H2O_RF1 shows the importance of the independent variables.

Results

As discussed, five groups of regression models were created. Consequently, the RMSE between the forecasted budget and actual costs were calculated for both training and testing datasets. The RMSE were calculated for all models along with the real data to assess the performance of the models. The actual RMSE indicates the discrepancy between actual and budgeted expenditures while using the classical budgeting technique. Table 1 displays the RMSE of the models considering training and testing datasets.

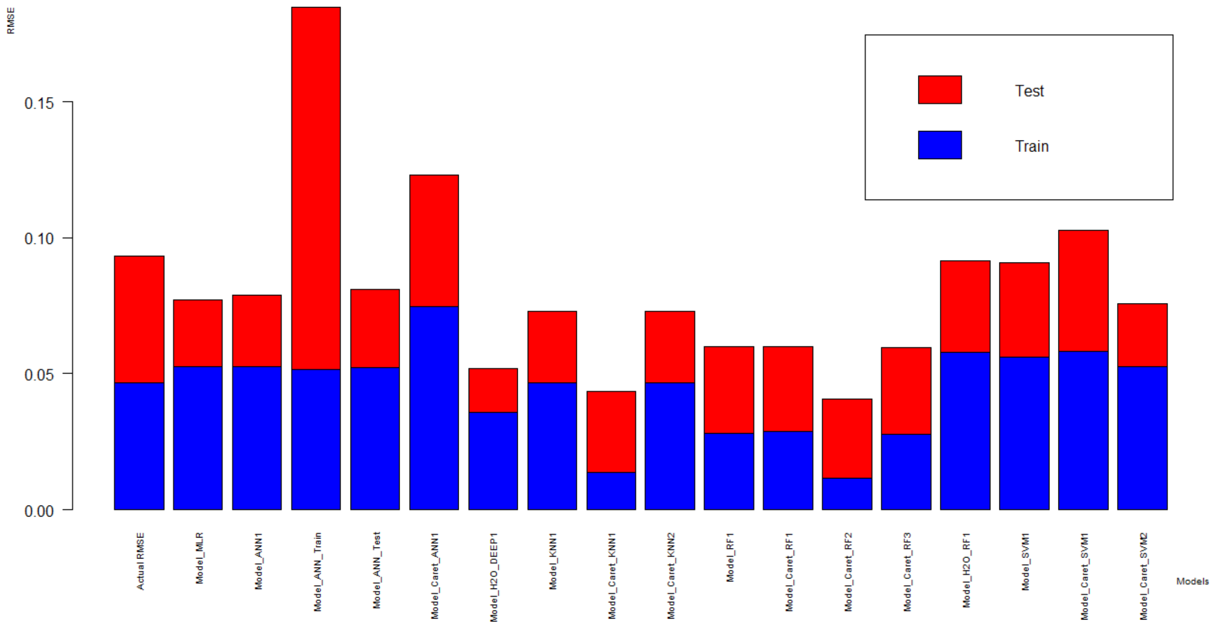
The table shows models with high prediction accuracy. Model_Caret_RF2 and Model_Caret_KNN1 showed high ability to fit with the training dataset. Model_ANN_Train showed overfitting to the training dataset. Model_H2O_DEEP1, Model_ANN_Test, and Model_Caret_RF2 showed high generalizability.

Table 1
Models RMSE

Model Name	Train RMSE	Test RMSE
Actual RMSE	0.046561753	0.046561753
Model_MLR	0.052411622	0.024788497
Model_ANN1	0.052560323	0.026308781
Model_ANN_Train	0.051655975	0.133026625
Model_ANN_Test	0.052399354	0.028609701
Model_Caret_ANN1	0.074509863	0.048495304
Model_H2O_DEEP1	0.035644294	0.016213948
Model_KNN1	0.046456911	0.026537751
Model_Caret_KNN1	0.013591195	0.030051811
Model_Caret_KNN2	0.046456911	0.026537751
Model_RF1	0.028033991	0.031779313
Model_Caret_RF1	0.028893984	0.031152761
Model_Caret_RF2	0.011702676	0.029131458
Model_Caret_RF3	0.027786526	0.031934766
Model_H2O_RF1	0.058015196	0.033510511
Model_SVM1	0.056112486	0.034586428
Model_Caret_SVM1	0.058096735	0.044510994
Model_Caret_SVM2	0.052615943	0.02294875

Figure 3 displays the comparison of the ML models RMSE. Each model RMSE is depicted on the Y axis. The training and testing RMSE are denoted by the colors blue and red, respectively.

Figure 3
RMSE Statistical Test and the Results



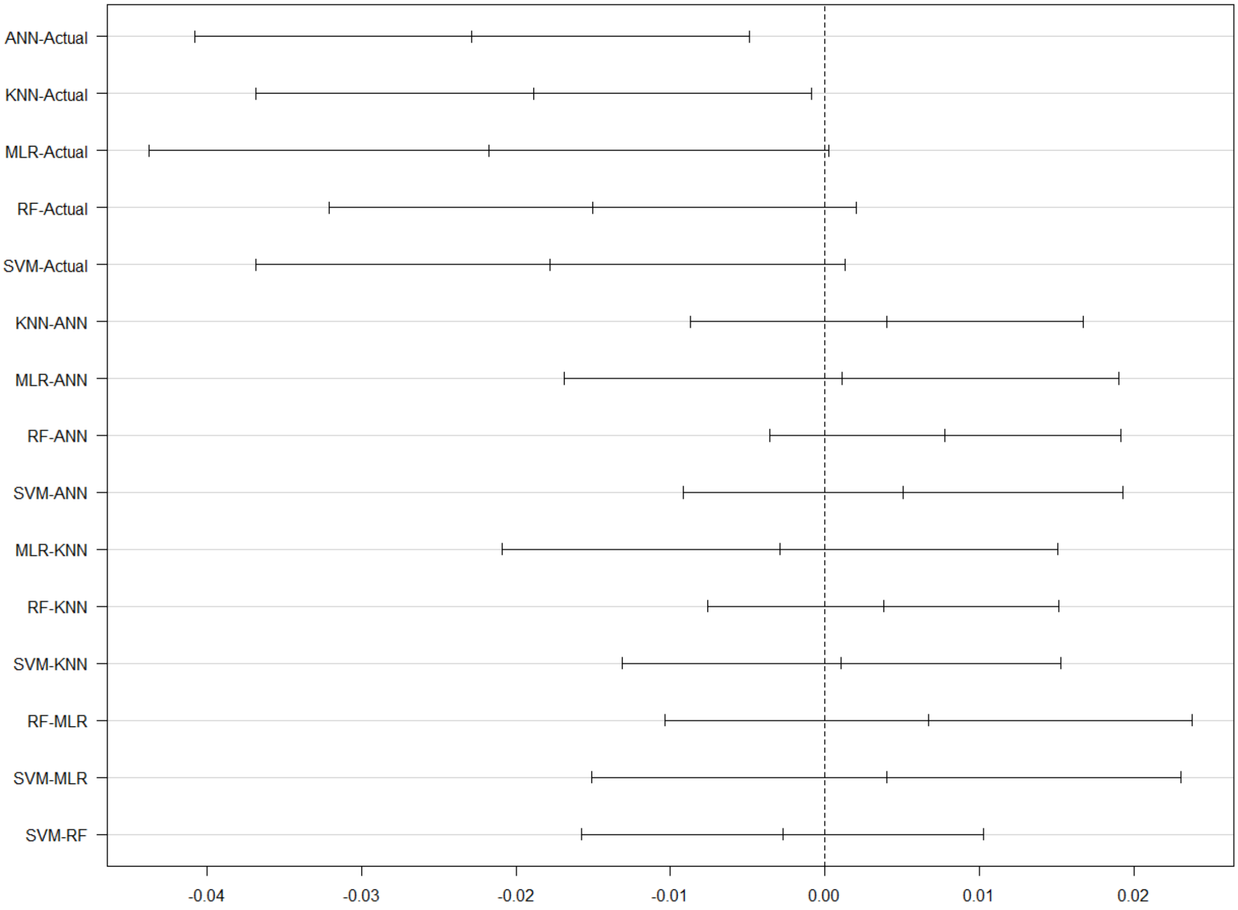
The models were aggregated into five groups: MLR, ANN, SVR, KNN, and RF to be compared with the actual RMSE. The actual RMSE is the error between the classical budgeting system and the actual cost. An ANOVA test was conducted to compare the variance across the mean of the five groups. An analysis of variance yielded significant variation among the ML models and the classical budgeting system. Table 2 presents the outcomes of the ANOVA test. The test revealed a significant difference between the ML RMSE and the actual RMSE.

Table 2
Analysis of Variance Test Result

	df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	5	0.0004443	8.887e-05	4.625	0.0229*
Residuals	9	0.0001729	1.922e-05		

A Tukey post hoc test was conducted to specify which model showed the significant difference with the actual RMSE. Figure 4 presents the findings. According to the test results, ANNs and KNN showed significant differences with the classical budgeting system.

Figure 4
Tukey Honestly Significant Difference Post Hoc Test



Note. 95% family-wise confidence interval.

Discussion / Implications

The results of the variance analysis show a significant difference between the ML models and the classical budgeting system. The findings disclose a significant variance between the actual RMSE and the ML RMSE with *F*(5, 9) = 4.625, $p < 0.05$. The results of a post-hoc Tukey test reveals that ANN and KNN models show a significant difference with the classical budgeting system with $p < 0.05$.

In conclusion, the precision and efficiency of budgeting systems can be improved by ML approaches. The Caret library RF performed best in terms of accurately fitting the model to the training dataset. The deep learning model developed by H2O library showed the best generalization accuracy.

The outcomes show that ML may enhance the budgeting process. The gap between the budget and actual costs may be reduced using ML models. As a result, businesses may improve their chances of winning projects without bearing a significant risk of going over budget. Using previous data, the ML models may provide a more precise projection of a project cost. Consequently, businesses may provide budgetary offers with minimum effort. Moreover, managers may validate the prices offered to the clients and highlight if there is an enormous discrepancy between their estimated cost and the anticipated.

Additionally, the algorithm specifies the weightages of the cost categories. Therefore, businesses may concentrate on the important elements affecting the budget. Hence the companies may prepare the budget more quickly without jeopardizing the budget accuracy. The least important factors can be estimated as a percentage of the most critical factors.

Companies can use ML models to create a benchmark. A company can specify a near estimate of a project using the ML models without going into detail by using prior project data. Companies may save time and effort when creating the cost estimate by offering a baseline. Conversely, the benchmark created by the ML model may indicate when the budget is overstated or undervalued.

In summary, businesses can benefit from their cumulative knowledge gained from executing projects in providing information to the ML model. This research supports companies in turning project lessons learned into quantified information to develop a machine learning model to adjust their pricing and costing system to win competitions. By analyzing business historical data, the model can support real-time decision-making in terms of pricing. Moreover, by using information collected from competitor pricing at open bids, the model can identify competitor pricing patterns. Accordingly, companies can predict the winning price.

This research provides an approach to develop project budgeting using ML techniques. ML can be utilized to provide a model to estimate the real cost. Each company may have different cost categories that regression models can accommodate. Additionally, the research offers a methodology for evaluating each model accuracy and comparing it with the actual error using RMSE to conclude which ML model best fits the company's nature of business.

Using previous data, companies can develop project benchmarks. Scientifically, accurate ML models are primarily dependent on reliable data and precise cost allocation. In this study, ANN and KNN demonstrated the most significance. In other businesses, the model may be different. Nevertheless, the concept still applies.

Many different sectors can benefit from the ML methodology outlined in the research. To do this, businesses should employ a project management information system to record precise historical data for the development of ML benchmarking models. In addition to large businesses, small and medium-sized businesses can develop their own ML models using the data developed on their systems. Although the benchmark of one company cannot be used for another, it provides a reliable estimate of the project market pricing.

This research used a limited sample of 552 projects to build the ML models. Accordingly, the models can be reproduced with more information from businesses in various regions. Additionally, this study utilized the field of industrial automation. The study can be expanded to include additional industries and different cost categories. Moreover, research can examine whether eliminating the least important expense categories compromises the accuracy of predictions.

The relationship between the expense categories is exceedingly important for businesses. For instance, demonstrating that logistics require a certain percentage of materials cost would enable companies to prepare budgets more quickly. As a result, creating a relationship between cost categories offers a research opportunity.

Consequently, the approach of using ML to forecast the actual cost can be applied to each cost category. Therefore, the independent variables can be transferred to be dependent variables. For instance, in the industrial automation field, it is possible to estimate the number of engineering hours by creating an ML model that includes engineering factors such as the number of input/output points, graphic pages, and hardware cabinets. Likewise, the cost of installation could be determined as a factor of site conditions, the hardware, and engineering hours.

Conclusion

Machine learning models reveal significant differences in forecasting the actual cost compared to the classical budgeting system. Machine learning techniques can be used to provide an estimate for project actual cost. Although different cost categories may be used, businesses can utilize the same methodology to conclude the machine learning model that best fits with their sector. RMSE can be used to assess the accuracy of machine learning prediction compared to the classical budgeting system. Therefore, companies may benefit from machine learning models to predict the final project cost to improve the budgeting system.

References

Abd, A. M., & Naseef, F. S. (2019). Predicting the final cost of Iraqi construction project using artificial neural network (ANN). *Indian Journal of Science and Technology*, *12*(28), 1-7. https://doi.org/10.17485/ijst/2019/v12i28/145640

Antunes, E., Vuppaladadiyam, A.K., Sarmah, A.K., Varsha S. S. V. , KishorePant, K., Tiwari, B., & Pandey, A. (2021). Application of biochar for emerging contaminant mitigation, *Environmental Management and Protection* (pp. 65–91). https://doi.org/10.1016/bs.apmp.2021.08.003

Balali, A., Valipour, A., Antucheviciene, J., & Šaparauskas, J. (2020). Improving the results of the earned value management technique using artificial neural networks in construction projects. *Symmetry*, *12*(10) , 1-16. https://doi.org/10.3390/sym12101745

Besouw, J. Van, & Bond-Barnard, T. (2021). Smart project management information systems (SPMIS) for engineering projects. *Project Performance Monitoring & Reporting*, *9*(1), 78-97. https://doi.org/10.12821/ijispm090104

Browning, T. R. (2019). Planning, tracking, and reducing a complex project's value at risk. *Project Management Journal*, *50*(1), 71-85. https://doi.org/10.1177/8756972818810967

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245-317. https://doi.org/10.1613/JAIR.1.12228

Cerezo-Narv, A., Pastor-Fernández, A., Otero-Mateo, M., & Ballesteros-Pérez, P. (2020). Integration of cost and work breakdown structures in the management of construction projects. *Applied Sciences*, *10*(4), 1-33. https://doi.org/10.3390/app10041386

Devi, T. R., & Reddy, V. S. (2012). Work breakdown structure of the project. *International Journal of Engineering Research and Applications*, *2*(2), 683-686.

El Naqa, I. (2015). Detection and prediction of radiotherapy errors. *Machine Learning in Radiation Oncology: Theory and Applications*, 237-241. https://doi.org/10.1007/978-3-319-18305-3

Eyibio, O. N., & Daniel, C. O. (2020). Effective resource budgeting as a tool for project management. *Asian Journal of Business and Management*, *8*(2), 15-20. http://doi.org/10.24203/ajbm.v8i2.6190

Greco, A. (2018). *Direct costs vs indirect costs*. Project Cubicle. https://www.projectcubicle.com/direct-costs-and-indirect-costs-cost-classification/

Hassim, S., Muniandy, R., Alias, A. H., & Abdullah, P. (2018). Construction tender price estimation standardization (TPES) in Malaysia: Modeling using fuzzy neural network. *Engineering, Construction and Architectural Management*, *25*(3), 443-457. https://doi.org/10.1108/ECAM-09-2016-0215

Kwon, H., & Kang, C. W. (2018). Improving project budget estimation accuracy and precision by analyzing reserves for both identified and unidentified risks. *Project Management Journal*, *50*(1), 86-100. https://doi.org/10.1177/8756972818810963

Love, P. E., Sing, M. C., Ika, L. A., & Newton, S. (2019). The cost performance of transportation projects: The fallacy of the Planning Fallacy account. *Transportation Research Part A: Policy and Practice, 122*, 1-20. https://doi.org/10.1016/j.tra.2019.02.004

Mahagaonkar, S. S., & Kelkar, A. A. (2017). Application of ABC analysis for material management of a residential building. *International Research Journal of Engineering and Technology*, *4*(8), 614-620.

Miandoab, E. E., & Gharehchopogh, F. S. (2016). A novel hybrid algorithm for software cost estimation based on cuckoo optimization and k-nearest neighbors algorithms. *Engineering, Technology & Applied Science Research*, *6*(3), 1018-1022. https://doi.org/10.48084/etasr.701

Mohammed, S. J., Abdel-Khaled, H. A., & Hafez, S. M. (2021). Predicting performance measurement of residential buildings using machine intelligence techniques (MLR, ANN, and SVM). *Iranian Journal of Science and Technology - Transactions of Civil Engineering*. *46*, 3429–3451. https://doi.org/10.1007/s40996-021-00742-4

Srinivasan, N. P., Gowtham, T., Kumar, A. R. R., Baalaji, M., & Kumar, S. N. (2021). A review on feasible cost prediction model in construction projects. *Information Technology in Industry*, *9*(3), 537-543.

Tijanić, K., Car-Pušić, D., & Šperac, M. (2020). Cost estimation in road construction using artificial neural network. *Neural Computing and Applications*, *32*(13), 9343-9355. https://doi.org/10.1007/s00521-019-04443-y

Zhou, Z. (2018). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44-53. https://doi.org/10.1093/nsr/nwx106