A Machine Learning Study to Enhance Project Cost Forecasting

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-10 (2022) 3286–3291

A Machine Learning Study to Enhance Project Cost Forecasting

Tolga İnan*, Timur Narbaev, Öncü Hazir *****

* *Electrical-Electronics Engineering Department, Çankaya University, Ankara, Turkey (e-mail: tolga.inan@cankaya.edu.tr)*

** *Business School, Kazakh-British Technical University, Almaty, Kazakhstan (e-mail: t.narbaev@kbtu.kz)*

*** *Supply Chain Management and Information Systems Department, Rennes School of Business, Rennes, France (e-mail: oncu.hazir@rennes-sb.com)*

Abstract: In project management it is critical to obtain accurate cost forecasts using effective methods. This study presents a Machine Learning model based on Long-Short Term Memory to forecast the project cost. The model uses the seven-dimensional feature vector, including schedule and cost performance factors and their moving averages as a predictor. Based on the cost variation patterns from the training phase, we validate the model using three hundred experiments in the testing phase. Overall, the proposed model produces more accurate cost estimates when compared to the traditional Earned Value Management index-based model.

Keywords: Cost forecasting, Earned Value Management, Estimate at Completion, Machine Learning, Project Management.

1. INTRODUCTION

Almost all the projects experience cost overruns irrespective of their size and type, as they face many uncertainties during their life cycles. Various project monitoring and control methodologies such as Earned Value Management (EVM) are commonly used to limit these cost overruns. These methods mainly support the organizations to monitor the progress of the projects and budget use effectively. During the project execution, at any time, project teams need to know about what has happened since the project start and, more importantly, be able to foresee what might happen in the remaining life of the projects. This makes accurate estimates critical to completing projects under budget and maintaining reliable communication with project stakeholders.

However, project monitoring and forecasting decisions are prone to the increasing uncertainties of today's data-rich business environments. In this respect, we observe the considerable potential for using Artificial Intelligence techniques in production (Cadavid et al., 2019; Rai et al., 2021) and project control (Munir, 2019; Chen et al., 2020; Ong and Uddin, 2020; Natarajan, 2022).

More specifically, Machine Learning (ML) algorithms can aid organizations in enhancing project cost forecasting, which we focus on in this study. Even though the potential benefits are remarkable, the literature is still scant (Willems and Vanhoucke, 2015; Pellerin and Perrier, 2019; Hashemi et al., 2020). The following section will briefly introduce some pertinent ML applications for cost forecasting in projects.

We first discuss the fundamentals of EVM, a Project Management (PM) methodology used to measure and forecast duration and cost in projects (Humphreys, 2018; Mahmoudi et al., 2021). We note that the conventional index-based EVM forecasting approaches assume linearity in cost growth (Anbari, 2003; PMI, 2019). However, cost growth is usually nonlinear in projects and often resembles an S-shaped curve pattern (Barraza et al., 2004; Narbaev and De Marco, 2014b). Moreover, the index-based methods may result in inaccurate forecasts in the early stages of a project (Kim and Reinschmidt, 2011; Warburton and Cioffi, 2016) due to a few data points available that make the extrapolation to the project's remaining part not reliable (Lipke et al., 2009; De Marco et al., 2016). These two limitations of the index-based cost forecasting methods motivate our study.

ML models have not been extensively applied for cost forecasting in ongoing projects. However, ML has a great potential to enhance decision making in PM (IPMA, 2020). Organizations have been carrying out more and more projects and gathering a tremendous amount of data from the undertaken projects. In fact, along with the data, know-how is also accumulated. Traditionally, this know-how is carried from project to project by senior managers. Project managers mainly depend on their experiences and implement traditional methods to estimate the total cost and completion time and take corrective actions using the predictions and relying on their experiences.

Our study aims to demonstrate how managers can benefit more from the data of the completed projects by using ML. In particular, the projects' data with similar characteristics are used to train ML-based estimators and support project

10.1016/j.ifacol.2022.10.127

managers in making estimations. These estimation methods can improve forecasting practices by considering the nonlinearity in cost growth and making estimates by studying results of our ML approach with the ones by the traditional index-based model.

The remainder of the paper is structured as follows. Next, we introduce key EVM metrics and review relevant studies on ML applications in project cost forecasting. Then, we present our ML approach and the project dataset for calculating the cost estimates. Next, we report the results of our comparative analysis and discuss the main results. Finally, we conclude with the study summary, research limitations, and future research directions.

2. BACKGROUND

2.1 Key EVM metrics

According to the Project Management Institute (PMI, 2019), EVM is a methodology used by project managers to monitor and control the schedule and budget of a project. It is based on three key metrics: Planned Value (PV) – the budgeted value of the scheduled work; Actual Cost – the actual value of the performed work; and Earned Value (EV) – the budgeted value of the performed work (Anbari, 2003). Budget at Completion (BAC) is the project's total budget and Cost at Completion (CAC) is its total actual cost at completion. Planned Duration (PD) is the project's scheduled duration and Actual Duration (AD) is its actual duration at completion. To assess the project's cost performance (efficient use of BAC), Cost Performance Index (CPI=EV/AC) is used. To measure the project's schedule progress, Schedule Performance Index (SPI=EV/PV) is applied.

Finally, Cost Estimate at Completion (EAC($)) is the forecast that represents the final cost of a project. Our study uses ML to obtain a more accurate EAC($), and the index-based formula in (1) is used as a benchmark. We compare our EAC($) results with the ones calculated by (1).

$$EAC(\$)_t = AC_t + BAC - EV_t \quad (1)$$

The linear model (1) is selected as the benchmark following the study of Batselier and Vanhoucke (2015b), who conducted a comparative analysis of eight index-based EAC($) models. They used the EVM data of 51 real projects and associated simulations for comparison. Based on the accuracy results, measured by Mean Absolute Percentage Error (MAPE) (defined next in the paper), they found that the model by (1) showed dominance over the other seven models and produced the most accurate EAC($) estimates.

2.2 Brief review of ML applications for project cost forecasting

First, we note that ML models have not been extensively applied in project monitoring and control. Only a few studies developed ML models, specifically for project cost forecasting during the project execution phase. Table 1 provides a summary of these studies with a brief description of their models and contribution to the EVM body of knowledge.

the given data points. To show the effectiveness of the chosen approach, we will compare the accuracy of the cost forecasting

Among the first implementations, Pewdum et al. (2009) developed an Artificial Neural Network (ANN) model based on Backpropagation to improve the accuracy of duration and cost estimates. They integrated numerous variables. Among all the variables, the traffic volume, weather conditions, contract duration, construction budget, percent complete of the planned work, and percent complete of the actual work performed were the most influential. Their model produced accurate EAC($) when applied to highway construction projects in Thailand.

Table 1. A summary of the reviewed studies on ML applications for project forecasting

Study	Description	Contribution to EVM
Pewdum et al. (2009)	Several project performance factors are integrated into the ANN-based Backpropagation model	Duration and cost forecasting during project execution
Narbaev and De Marco (2014a)	Supervised regression approach that integrates the EVM cost data through the Gompertz Growth modeling	Cost forecasting during project execution
Elmousalami (2021)	Numerous ML methods, including the Ensemble-based, are compared using Fuzzy Logic	Cost estimation during project planning
Ottaviani and De Marco (2022)	Multiple linear regression model is proposed using the EVM cost data as independent (input) variables	Cost forecasting during project execution
Natarajan (2022)	The ANN and Reference class forecasting approaches are integrated to produce probabilistic estimates	Duration and cost forecasting during project planning and execution
Wauters and Vanhoucke (2016)	The four ML techniques are compared to produce more accurate duration estimates	Duration forecasting during project execution
The current study	The ANN-based Long-Short Term Memory model that uses the EVM-based CPI and SPI metrics and their derivatives	Cost forecasting during project execution

Narbaev and De Marco (2014a) adopted a Supervised nonlinear regression approach based on the Gompertz Growth model. Applying their model to nine construction projects, the authors compared the accuracy of their estimates with the ones produced by implementing the CPI-integrated index-based model. As their model fits better to the S-shaped curve observed in project cost growth, they obtained more accurate cost estimates.

Elmousalami (2021) integrated ML algorithms into the cost estimation efforts at the project development stage. Using the Fuzzy Logic, they embedded the uncertainty factors in their ML models and showed that the Ensemble methods were superior in predicting performance.

Recently, Ottaviani and De Marco (2022) developed a multiple linear regression model to assess the impact of input variables (CPI, original cost forecast, and percent of work performed) and improve the model fitting to the project's real CAC value. Using the data of 29 real-life projects, they showed that their model with the three variables provided higher accuracy and lower variance in EAC($) estimates.

Natarajan (2022) proposed a comprehensive model that integrated Reference class forecasting (the outside view of a project) and ML (the inside view from the project data) to improve schedule and budget planning and control. Using the cost data of 106 and the schedule data of 130 oil and gas projects, the author showed a higher predictive capability of the ML approach in predicting the most likely cost and schedule overruns in projects.

To forecast the project duration, Wauters and Vanhoucke (2016) applied Decision Tree, Bagging, Random Forest, and Boosting techniques. They compared their forecasting results with the ones by the conventional models (based on linear performance indexes). Using artificial project data, they showed that ML approaches had more accurate predicting capabilities than the traditional index-based methods.

Finally, we refer to some review studies. Willems and Vanhoucke (2015) examined the EVM methods and some ML applications in project control. Hashemi et al. (2020) discussed the ML applications for project cost forecasting. Ulusoy and Hazir (2021) listed many interesting application areas of ML in PM.

3. METHODOLOGY

3.1 Model

ML has been increasingly used in many fields, from computer vision to biometric recognition, from advertising to the defense industry. In the literature, ML approaches are classified into Supervised learning, Unsupervised learning, Semi-supervised learning, Reinforcement learning, and Dimensionality reduction (e.g., Panda et al., 2021).

Unsupervised learning methods use input data, mainly to find out the regularity in data. On the other hand, Supervised techniques use both input and output data. Depending on the problem, output data can be real numbers, integers, or categories. In our study, the cost figures (real numbers) constitute the output data.

Therefore, in this study, we focus on Supervised ML as there is output data. In this approach, the training phase is crucial as the patterns between the input-output data are found. On the other hand, the testing phase of the Supervised ML generates outputs following the input-output patterns determined during the training phase. In Supervised ML, approaches can be grouped into classification and regression subcategories. The classification algorithms generate discrete or categorical outputs. The regression algorithms generate continuous outcomes. Time-sequence regression is the type of Supervised ML that we implement in this study.

To explain how we used the time-sequence regression in our study, we describe our approach including the processes used in the ML training and testing phases. Fig. 1 presents how these two phases are used within our prediction algorithm. We learn the suitable ML model in the training phase, and the learned ML model is used as a predictor in the testing phase.

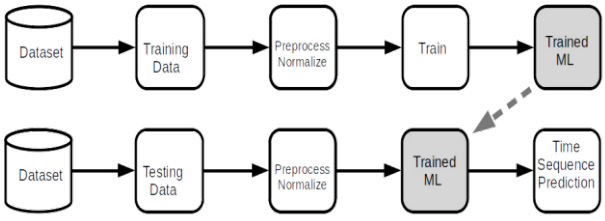

Figure 1. The training and testing phases of the proposed ML approach.

We employ the ML model of a recurrent ANN type, namely Long-Short Term Memory (LSTM). The LSTM networks are suitable for sequence-to-sequence regression problems. We refer to Hochreiter and Schmidhuber (1997) and Greff et al. (2016) for more information on the LSTM networks.

ML models require features (inputs) to make the prediction. Therefore, we must define the features of our ML model. We design a seven-dimensional feature vector. Six dimensions of the feature vector consist of CPI and SPI metrics and their moving average filtered versions (having window sizes of two and three tracking points for each metric). The seventh and last dimension of the input vector is the normalized time. The normalized time is found by dividing AD by PD for a particular tracking point. The output (predicted value) of the ML model is the cost at completion.

The training-testing protocol we use for the ML is as follows. We use 12 projects in the training phase and three projects in the testing phase of our experiments. Projects in the training and testing phase are randomly selected. We repeat the experiment a hundred times for each of the three projects, covering the training and testing phases. Therefore, all projects are used in both training and testing phases. Accordingly, the results are reported independently of the training and testing sets.

The evaluation criteria to assess the accuracy of our model's cost estimate for a given project is the percentage error (the percent difference between the cost estimate and the actual cost of a project). We find the absolute average of these errors for all the projects in our dataset and measure this average with the Mean Absolute Percentage Error (MAPE), as in

$$MAPE = \frac{100\%}{n} * \sum_{t=1}^{n} \left| \frac{CAC - EAC(\$)_t}{CAC} \right| \quad (2)$$

Where t=1,…, n is the number of tracking periods for a project.

3.2 Dataset

We use the actual project data shared by the Operations Research & Scheduling Research Group of Ghent University (ORSRG, 2022; Batselier and Vanhoucke, 2015a). This database includes EVM data of 133 projects that have been executed and completed in different industries. The dataset mainly constitutes construction projects. Considering the database structure, we limit our scope only to construction projects, and our final dataset included the EVM data of 41 real-life completed projects.

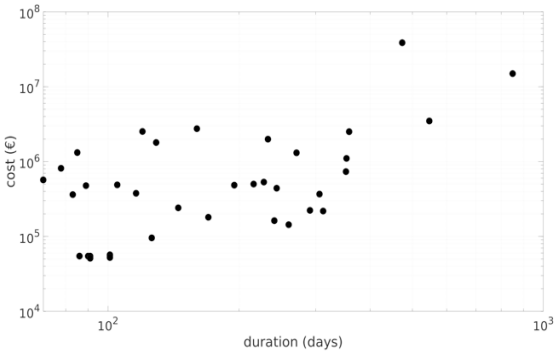

Figure 2. The cost and duration plot for 41 projects.

The total duration and cost data of the 41 construction projects extracted from the database are shown in Fig. 2. The projects have a large range of durations and budgets. Considering this variance, we chose the projects within a specific range. For the budget, we kept the upper limit to 3 million Euros. For the duration, we chose the projects with a maximum duration of 150 days and a minimum of four tracking points. The projects that fall in these ranges have some similarities, but the others are very small or big projects and quite different in project characteristics and resources. By setting budget and time limits, we generated a project pool of 15 projects. We randomly selected 12 projects for the training set, and the remaining three projects were reserved for testing.

4. SUMMARY of the RESULTS

Our results show that in 75.33% of the projects tested, the MAPE (2) obtained using our ML model was smaller than that obtained with the traditional index-based model (1). We found the difference between MAPEs and provide its results as a histogram in Fig. 3.

A positive difference in MAPE in this histogram shows a smaller MAPE of our ML method. The negative difference indicates the projects where our ML model produced a larger MAPE than the conventional index-based model. About 50.00% of 75.33% projects tested have a MAPE difference of about 1.00%. Even though this is a negligible difference in EAC($) estimate's accuracy between the two models, we note that the proposed model has a feature to learn from the given EVM data. This is because the EAC($) estimates calculated in the testing phase followed the input-output patterns of the EVM data of the projects analyzed in the training phase. Following this, during the training phase, the cost-related EVM data was utilized to build the proposed ML algorithm using LSTM network. Our ML model evaluated this input data repeatedly until learning its cost growth pattern (behavior).

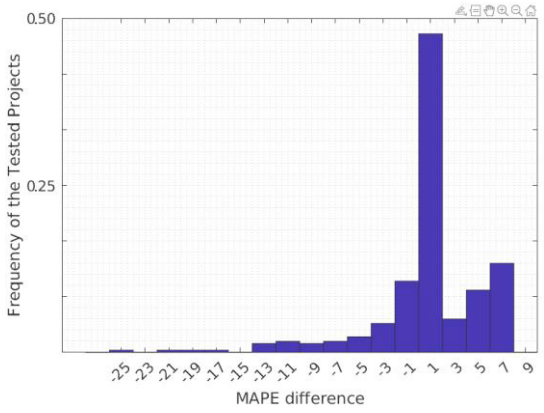

Figure 3. The MAPE difference between the proposed ML model and the EVM index-based model.

5. CONCLUSION

Cost overrun is a common problem in projects undertaken in various industries. To deal with this common problem, project managers opt for continuously monitoring the use of the project budgets. Many of them try to produce accurate cost estimates using the traditional EVM methods. These methods are mainly based on cost and schedule performance indexes which are linear. However, projects' budget acquisition and cost growth patterns are nonlinear and resemble an S-shape. Therefore, such methods have the inherent limitations in providing more reliable and accurate cost estimates that reflect the real cost growth behavior. Also, whatever the approach followed or the method implemented, having accurate cost estimates is critical to completing the projects successfully and maintaining loyal relationships with project stakeholders.

Considering this limitation of the existing index-based models and the importance of having accurate cost estimates, in this study, we developed an ML algorithm for estimating the total project cost more accurately. We employed a Supervised ML model based on the LSTM protocol to forecast EAC($). The EVM data of 41 real completed projects validated the proposed approach. The training phase of our approach with 12 projects allowed us to learn from the given dataset the patterns that characterized the changes in the project cost. We used the seven-dimensional feature vector that considered EVM metrics like CPI and SPI and their moving averages and the

normalized time as a predictor. Based on this, we used the learned patterns to calculate EAC($). In the testing phase, we validated our approach on three projects with an associated hundred experiments for each project. We compared our approach's EAC($) accuracy results with the ones computed using the widely used index-based model in practice (1). Overall, our model produced more accurate EAC($) results in 75.33% of project cases.

We acknowledge the following limitations that can potentially be addressed in future research. First, we conducted the experiments using a small dataset. We intend to extend the current research using a larger pool of projects and evaluate the model using additional forecasting criteria such as stability and timeliness of EAC($), in addition to the accuracy. Second, we will also work with projects from different industries, not only construction. However, the initial results obtained with the proposed ML approach are promising. The proposed method can be combined with other forecasting techniques to improve the solutions further.

ACKNOWLEDGMENTS

This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09259049).

REFERENCES

Anbari, F.T. (2003). Earned value project management method and extensions. *Project Management Journal*, 34(4), 12–23.

Barraza, G.A., Back, W.E., and Mata, F. (2004). Probabilistic forecasting of project performance using stochastic S curves. *Journal of Construction Engineering and Management*, 130(1).

Batselier, J. and Vanhoucke, M. (2015a). Construction and evaluation framework for a real-life project database, *International Journal of Project Management*, 33(3), 697–710.

Batselier, J. and Vanhoucke, M. (2015b). Empirical evaluation of earned value management forecasting accuracy for time and cost. *Journal of Construction Engineering and Management*, 141(11), 05015010.

Cacavid, J.P.U., Lamouri, S., Grabot, B., and Fortin, A. (2019). Machine Learning in Production Planning and Control: A Review of Empirical Literature. IFAC *PapersOnLin*e, 52(13), 385–390.

Chen, Z., Demeulemeester, E., Bai, S., and Guo, S. (2020). A Bayesian approach to set tolerance limits for a statistical project management. *International Journal of Production Research*, 58(10), 3150-3163.

De Marco, A., Rosso, M., and Narbaev, T. (2016). Nonlinear cost estimates at completion adjusted with risk contingency. *Journal of Modern Project Management*, 4(2), 24–33.

Elmousalami, H.H. (2020). Comparison of Artificial Intelligence techniques for project conceptual cost prediction: A case study and comparative analysis. *IEEE Transactions on Engineering Management*, 68(1), 183-196.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., and Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2222-2232.

Hashemi, S.T., Ebadati, O.M., and Kaur, H. (2020). Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. *SN Applied Sciences*, 2, 1703.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

Humphreys, G.C. (2018). *Project management using Earned value*. 4th edition. Humphreys & Associates, Inc.

IPMA. (2020). *Report on Artificial Intelligence impact in Project Management*. International Project Management Association. Amsterdam, The Netherlands.

Kim, B.C. and Reinschmidt, K.F. (2011). Combination of project cost forecasts in earned value management. *Journal of Construction Engineering and Management*, 137(11), 958–966.

Lipke, W., Zwikael, O., Henderson, K., and Anbari, F. (2009). Prediction of project outcome: the application of statistical methods to earned value management and earned schedule performance indexes. *International Journal of Project Management*, 27(4), 400–407.

Mahmoudi, A., Bagherpour M., and Javed, S.A. (2021). Grey earned value management: Theory and applications. *IEEE Transactions on Engineering Management*, 68(6), 1703-1721.

Munir, M. (2019). How Artificial Intelligence can help project managers. *Global Journal of Management And Business Research*, 19(4), 1-8.

Narbaev, T. and De Marco, A. (2014a). An Earned Schedule-based regression model to improve cost estimate at completion, *International Journal of Project Management*, 32(6), 1007-1018.

Narbaev, T. and De Marco, A. (2014b). Combination of growth model and earned schedule to forecast project cost at completion. *Journal of Construction Engineering and Management*, 140(1), 04013038.

Natarajan, A. (2022). Reference class forecasting and Machine Learning for improved offshore oil and gas megaproject planning: Methods and application. *Project Management Journal*, 53(OnlineFirst), 1-29.

Ong, S. and Uddin, S. (2020). Data science and Artificial Intelligence in project management: The past, present

and future. *Journal of Modern Project Management*, 7(4), 123–456.

ORSRG (2022). *Real data*. Operations Research & Scheduling Research Group. Ghent University. Available at https://www.projectmanagement.ugent.be/research/data/realdata

Ottaviani, F.M. and De Marco, A. (2022). Multiple linear regression model for improved project cost forecasting. *Procedia Computer Science*, 196(2022) 808–815.

Panda, S.K., Mishra, V., Balamurali, R., and Elngar, A.A. (2021). *Artificial Intelligence and Machine Learning in business management: Concepts, challenges, and case studies*. 1st edition. CRC Press Taylor&Francis Group. Florida, US.

Pellerin, R. and Perrier, N. (2019). A review of methods, techniques and tools for project planning and control. *International Journal of Production Research*, 57(7), 2160–2178.

Pewdum, W, Rujirayanyong, T., and Sooksatra, V. (2009). Forecasting final budget and duration of highway construction projects. *Engineering, Construction, and Architectural Management*, 16(6), 544–557.

PMI. (2019). *The standard for Earned Value Management*. 2nd edition. Project Management Institute (PMI). Newtown Square, PA.

Rai, R., Tiwari, M.K., Ivanov, D., and Dolgui, A. (2021). Machine Learning in manufacturing and industry 4.0 applications. *International Journal of Production Research*, 59(16), 4773-4778.

Ulusoy, G. and Hazir, Ö. (2021). Recent developments and some promising research areas. In Ulusoy, G. and Hazır, Ö. *An introduction to project modeling and planning*, 457-469. Springer Texts in Business and Economics. Springer, Cham.

Warburton, R.D.H. and Cioffi, D.F. (2016). Estimating a project's earned and final duration. *International Journal of Project Management*. 34 (8), 1493–1504.

Wauters, M. and Vanhoucke, M. (2016). A comparative study of Artificial Intelligence methods for project duration forecasting. *Expert Systems with Applications*, 46, 249-261.

Willems, L.L. and Vanhoucke, M. (2015). Classification of articles and journals on project control and Earned Value Management. *International Journal of Project Management*, 33(7), 1610–1634.