ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

Incorporating Machine Learning into a Data Warehouse for Real-Time Construction Projects Benchmarking

Zhe Yin[1], Deborah DeGezelle[2]*, Kazuma Hirota[3], and Jiyong Choi[4]

[1] *Construction Industry Institute, The University of Texas at Austin, Austin, TX 78759, U.S.A.,* Email: zhe.yin@utexas.edu
[2] *Construction Industry Institute, The University of Texas at Austin, Austin, TX 78759, U.S.A.,* Email: ddegezelle@cii.utexas.edu
[3] *Walker Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX 78759, U.S.A.,* Email: kazhirota7@utexas.edu
[4] *Dept. of Manufacturing and Construction Management, Central Connecticut State University, CT 06050, U.S.A.,* Email: jaychoi@ccsu.edu

Abstract:
Machine Learning is a process of using computer algorithms to extract information from raw data to solve complex problems in a data-rich environment. It has been used in the construction industry by both academics and practitioners for multiple applications to improve the construction process. The Construction Industry Institute, a leading construction research organization has twenty-five years of experience in benchmarking capital projects in the industry. The organization is at an advantage to develop useful machine learning applications because it possesses enormous real construction data. Its benchmarking programs have been actively used by owner and contractor companies today to assess their capital projects' performance. A credible benchmarking program requires statistically valid data without subjective interference in the program administration. In developing the next-generation benchmarking program, the Data Warehouse, the organization aims to use machine learning algorithms to minimize human effort and to enable rapid data ingestion from diverse sources with data validity and reliability. This research effort uses a focus group comprised of practitioners from the construction industry and data scientists from a variety of disciplines. The group collaborated to identify the machine learning requirements and potential applications in the program. Technical and domain experts worked to select appropriate algorithms to support the business objectives. This paper presents initial steps in a chain of what is expected to be numerous learning algorithms to support high-performance computing, a fully automated performance benchmarking system.

Key words: Machine Learning, Construction, Data Warehouse, Real-Time Benchmarking

1. INTRODUCTION

Machine Learning (ML) is the study of computer algorithms that can improve automatically through experience [1]. Nowadays, ML is incredibly important because it can solve complicated real-world problems in a scalable way [2]. As a subset of artificial intelligence (AI), its applications have become more technical and highly practical [3]. ML is a field that has been used by many industries for better results and efficiency [4]. In construction, ML is can be used to monitor progress, assess risks, notify issues, improve activities, and predict more streamlined workflow [5], it plays a pivotal role in making the construction “smart” [6].

Any good ML strategy needs data to work [7]. Limited data is considered the very first challenge; it hampers employment and constrains the potential for applications [8]. Importantly, appropriate models must be developed on good data. The two main things that can go wrong with machine learning are “bad algorithm” and “bad data” [9]. The Construction Industry Institute (CII), a leading construction research organization based at The University of Texas at Austin has developed a sophisticated, external construction project execution benchmarking program for more than twenty years. It serves multiple industry sectors and supports benchmarking for a variety of goals from managing predictability to providing external estimate validation. The program has historically relied on human effort to validate the submitted data. In its effort to develop a next-generation benchmarking program, the Data Warehouse, CII partnered with Texas Advanced Computing Center (TACC) to fully utilize the world’s most powerful computational capabilities for benchmarking [10]. Because real-time ML relies largely on the infrastructure [11], CII’s Data Warehouse is a first-of-its-kind system that deploys machine learning algorithms to a rich and diverse construction dataset for real-time project benchmarking and analytics [12]. This study discussed the first steps in developing an autonomous and external benchmarking platform.

2. BACKGROUND

In construction research, ML, including both shallow and deep learning, has been explored by researchers [6]. The applications focus largely on prediction, detection, modeling, integration, and assessment in different project elements and aspects [5]. Recent research includes delay risk prediction in construction projects [13], activity recognition of construction workers and equipment [14], construction cost items modeling [15], integrating construction documents to the schedule [16], and project defect risk assessment [17]. It can be seen that all these applications need data to enable effective and accurate ML [18]. In other words, data is the requirement [19].

Benchmarking is defined as measuring performance by using a specific indicator resulting in a metric of performance that is then compared to others [20]. In the construction industry, benchmarking is the systematic process of measuring an organization’s performance against that of industry peers to determine best practices that, when adopted and utilized, lead to superior project performance [21]. It is considered one of the key management techniques that allow companies and their projects to be compared for improvement [22] and it needs to be adopted to meet challenging new construction efficiency and productivity targets [23].

Table 4. Summary of the CII Performance Assessment System

#	Program	Start Year	Reference
1	Benchmarking & Metrics General Program	1997	[24]
2	Construction Productivity	2001	[25]
3	Small and Maintenance Projects	2004	[26]
4	Pharmaceutical and Biotech Facilities Projects	2005	[27]
5	Mega/Major Projects	2006	[28]
6	Engineering Productivity	2006	[29]
7	Health Care Facility Projects	2012	[30]
8	10-10 Program	2013	[31]
9	Federal Facilities Projects	2021	[32]

CII started its benchmarking efforts in the mid-1990s. **Error! Reference source not found.** shows the summary of CII's benchmarking programs. Several tailored metrics programs have been researched, developed, and deployed. The benchmarking framework is collectively referred to as the Performance Assessment System (PAS) [33] and resides on the Data Warehouse platform. The PAS includes a total of 2,416 projects. The 10-10 Program [34] collects data by project phase and it currently has 2,468 phase-based surveys. The PAS includes four sectors in the construction industry including Heavy Industrial, Light Industrial, Buildings, and Infrastructure.

3. RESEARCH METHODOLOGY

To accomplish the research effort, three major steps in Table 5 were taken to integrate the machine learning algorithms in the benchmarking program.

Table 5. Research Steps and Tasks

Steps	Research Tasks
1	Assess New Benchmarking Requirements: • Business Case • Data Acquisition and Processing Flow
2	ML Integration in The Benchmarking Program
3	Select ML Model for Application

A focus group consisting of industry members and data scientists was used in this research. The industry members are diverse and representatives of benchmarking users from owner, contractor, and consultant companies in the construction industry. CII and TACC data scientists brought expertise in data modeling for project assessment, big data architecture, web services, machine learning, and analytics. Combined, the focus group possesses the relevant expertise to move the research forward. Table 6 summarizes the background information of the group.

Table 6. Summary of the Focus Group

Company	Designation	Years of Experience
Owner 1	Competitive Intelligence Advisor	25
Owner 2	Section Manager	33
Owner 3	Project Measurement Improvement Manager	15
Contractor 1	Sr. Project Controls Manager	23
Contractor 2	Construction Engineer	15
Consultant 1	Chief Technology Officer	18
Consultant 2	Senior Manager, Business Development	35
CII and TACC	Manager, Data Management & Collections	22
	Data Systems Engineer	30
	Software Developer	17
	Programmer	5

4. ASSESSING MACHINE LEARNING REQUIREMENTS IN BENCHMARKING

4.1 Business Case

Human input can be incredibly valuable, and the performance assessment system has historically included qualitative assessments in conjunction with quantitative data as both explanatory and predictive inputs. However, the time and effort required to gather opinions constrict the volume of data that can be collected.

The business case of the new benchmarking system is aimed to reduce human efforts and unleash research potential. The Data Warehouse is configured to collect performance assessments in big data, high volume, high-velocity framework. Consequently, developing means to systematically extract project data via automated methods and to accurately assess the data with little human interaction is essential.

4.2 Data Acquisition and Processing Flow

The Data Warehouse is designed with a flexible Application Programming Interface (API). Participating organizations can map project data that resides in their construction management systems and export it via an automated or semi-automated workflow into the Data Warehouse. This enables organizations to provide data in quantities that were previously impractical, due to the manual data collection methods that were used in the past. The projects in the Data Warehouse contain enormous research potential and are consequently held highly confidential. As the system moves into a mode of increasingly automated data collection, it is essential to develop an intelligent system that can ingest data and assess its fitness with reduced human interaction.

5. ML INTEGRATION IN THE BENCHMARKING PROGRAM

The Data Warehouse has the potential to integrate numerous ML applications. In general, any ML application follows the process presented in Figure 7.

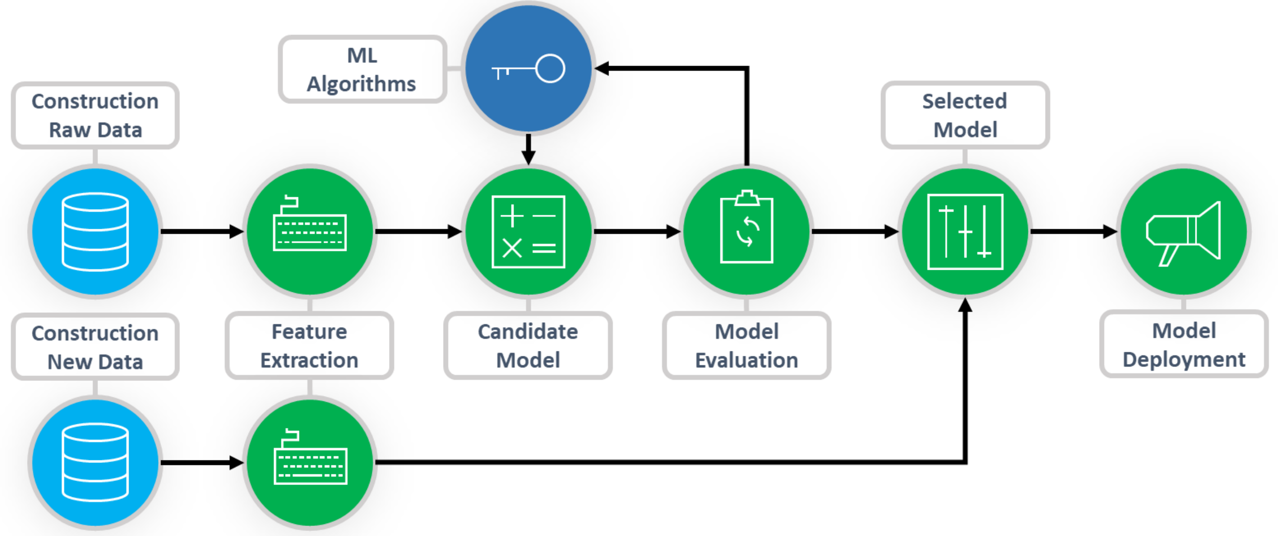

Figure 7. ML Application Process

The Data Warehouse requires numerous ML models, working together in a production environment. This study utilizes offline, supervised learning models. Further research will develop online, unsupervised learning models for the production environment. The focus group identified multiple use cases for ML that together, form an intelligent, autonomous benchmarking and performance assessment system.

Each use case will need to be researched. This can be a highly iterative process, but it is essential to find the best model. Otherwise, the resulting model can easily be overfitted, biased, or both. The process starts by developing a thorough understanding of the business need and considering how the available data might be utilized to serve that need. This can be a significantly time-consuming step if the data is highly complex and in need of dimensional reduction. However, it is essential to speed up training, as well as to apply the appropriate algorithm. It is generally expected that this step takes multiple rounds of testing and retesting.

Test results are evaluated and those that appear to fit the best are then selected. But deployment is not the end of the story. As new data come in, the system needs to be continually monitored. Not only does the world change, but models are often observed to decay over time. Monitoring and retraining are essential to a well-run system.

6. APPLICATION: CLASSIFYING TYPICAL AND NON-TYPICAL PROJECTS

This section outlines the initial efforts to select an appropriate ML model to assess the fitness of incoming data and to be able to establish whether the project is suitable for use in norms comparisons. Because the platform intakes data from many different participants, the quality of the data cannot be assumed to be consistent or suitable for norms analysis. Projects whose data are determined by the system to be robust, correct and representative of the current state of the industry can then be included in more sophisticated and value-generating predictive ML models. A statistically credible system therefore must utilize appropriately suitable projects for comparison and reject others that do not qualify. This paper covers the development of this evaluation model.

For simplicity, in this study, projects that are determined to be useful for norms comparisons, as well as for research and use in predictive models were labeled as “Typical” and those that are not, are labeled as “Not Typical”. Note that a project may perform very poorly and still be considered Typical. A project that should be labeled as Not Typical might refer to a project that was severely impacted by a catastrophic weather event, labor disruption, or if the project included the deployment of a novel technology or technique.

Many of the surveys in the Data Warehouse included manually provided responses by project managers rating whether their submitted projects are Typical or Not Typical. A total of 1,209 projects have been labeled as such in the Data Warehouse and this is used in the research for the training and test data sets. Three models, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Decision Tree (DT) were selected. All three of these supervised learning algorithms analyze data for classification and regression. The design of the study is to test each of the learning models using a variety of quantitative input factors to find the one that best classifies whether a project is “Typical” or “Not Typical” from the data alone, i.e., without asking the opinion of the project manager.

A simple method, k-NN is widely used as the first step in ML. It is often used as a benchmark for more complex classifiers. It utilizes a set of input values to predict output values and classifies a data point on how its neighbor is classified. SVMs are one of the most popular models used in ML and are well-suited for the classification of complex small and medium-sized datasets. SVMs set up decision boundaries and classify data depending on where it falls in relation to the decision boundary. Like SVMs, DTs are versatile ML algorithms that can perform both classification and regression tasks. DTs utilize a flowchart-like structure in which each internal node represents a "test" on an attribute. Because the data distribution representation of "Typical" and "Not Typical" is non-linear, this method is the most appropriate.

Several input variables were selected and tests were run in a variety of combinations. Input variables for rework, cost factor, scope change factor, cost growth, delta schedule growth, percent design complete at construction, percent design complete at authorization were all tested. Decision tree tests tended to expose From a series of tests, it was found that SVM was the best model for this instance because it accurately classified as "Typical" or "Not Typical" projects more often than other models. It was also found that the SVM tended to be a better predictor for contractor projects, so owner projects should likely consider other input factors in addition to the ones cited above.

7. IMPLICATION AND FUTURE RESEARCH

In this study, a relatively simplistic model was tested. In order for the system to accurately predict in production, the models need to be more sophisticated and take into account more factors, especially for owners. Further, the learning algorithm only addresses the first of several assessments that will be required within the Data Warehouse. Once a project is deemed Typical, for example, further models will be needed to understand other factors. Another limitation is that participating organizations are in the relatively early stages of integrating their connections with the CII Data Warehouse, so only limited data were available. In fact, it was found that many organizations themselves had immature and decentralized data management. Many organizations found that they needed to improve their internal systems before they would be positioned to fully integrate with the Data Warehouse.

Next, it must be acknowledged that moving from offline, supervised learning models to an online model is not an insignificant task. The group is already working on the next steps, intending to incrementally stitch together algorithms as they are proven to be valid. Continual learning remains difficult even for the best-funded enterprises, but a batch prediction can be usefully deployed, despite its limitations. As the system continues to be developed, the dream of a real-time, autonomous project performance assessment platform will become reality. Substantial research will be required along the way to ensure that the system is learning appropriately and isn't inappropriately biased. Further, the system will need monitoring and relatively continuous review and adjustments to ensure that the models stay relevant and useful for practitioners.

With sufficient data and a mature, artificially intelligent learning platform, the system will be able to evaluate incoming projects for their fitness and applicability for benchmarking comparisons and furthermore employ predictive analytics to alert practitioners while the project is ongoing, to confirm that the project appears to be on track for success or to warn that it is at risk of failure. Such a platform will be transformational for the industry, as well as research open up new avenues for research.

ACKNOWLEDGMENTS

This research wouldn't make a progress without the support of the focus group consisting of volunteering industry professionals from the CII member companies. Researchers at CII are sincerely grateful for their dedicated efforts.

REFERENCES

[1] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[2] R. Iriondo, "What is Machine Learning?," *Towards AI*, 2019. https://towardsai.net/p/machine-learning/what-is-machine-learning-ml-b58162f97ec7

(accessed Nov. 09, 2021).

[3] G. Ellis, “How Machine Learning Is Making Construction More Human - Digital Builder,” *Autodesk*, 2021. https://constructionblog.autodesk.com/machine-learning-construction/ (accessed Nov. 09, 2021).

[4] M. K. Gourisaria, R. Agrawal, G. Harshvardhan, M. Pandey, and S. S. Rautaray, “Application of Machine Learning in Industry 4.0,” in *Studies in Big Data*, vol. 87, Springer, Singapore, 2021, pp. 57–87.

[5] A. Dashore, “Machine Learning and its Applications in Construction,” *The Constructor*, 2021. https://theconstructor.org/artificial-intelligence/machine-learning-applications-construction/553770/ (accessed Nov. 09, 2021).

[6] Y. Xu, Y. Zhou, P. Sekula, and L. Ding, “Machine learning in construction: From shallow to deep learning,” *Dev. Built Environ.*, vol. 6, p. 100045, May 2021, doi: 10.1016/j.dibe.2021.100045.

[7] K. Venkatasubramanian, “Machine Learning Is Transforming the Construction Industry,” *Construction Executive*, 2021. https://constructionexec.com/article/machine-learning-is-transforming-the-construction-industry (accessed Nov. 09, 2021).

[8] H. Reza Tizhoosh and L. Pantanowitz, “Artificial intelligence and digital pathology: Challenges and opportunities,” *J. Pathol. Inform.*, vol. 9, no. 1, Jan. 2018.

[9] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. 2017.

[10] TACC, “Stampede2,” 2020. https://www.tacc.utexas.edu/systems/stampede2 (accessed Nov. 12, 2021).

[11] C. Huyen, “Real-time machine learning : challenges and solutions,” 2022. https://huyenchip.com/2022/01/02/real-time-machine-learning-challenges-and-solutions.html (accessed Feb. 01, 2022).

[12] CII, “CII’s Data Warehouse,” 2020. https://www.construction-institute.org/resources/performance-assessment/cii-pdw (accessed Nov. 12, 2021).

[13] A. Gondia, A. Siam, W. El-Dakhakhni, and A. H. Nassar, “Machine Learning Algorithms for Construction Projects Delay Risk Prediction,” *J. Constr. Eng. Manag.*, vol. 146, no. 1, p. 04019085, Oct. 2020, doi: 10.1061/(asce)co.1943-7862.0001736.

[14] B. Sherafat *et al.*, “Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review,” *J. Constr. Eng. Manag.*, vol. 146, no. 6, p. 03120002, Apr. 2020, Accessed: Nov. 15, 2021. [Online]. Available: 10.1061/(asce)co.1943-7862.0001843.

[15] A. Mahdavian, A. Shojaei, M. Salem, J. S. Yuan, and A. A. Oloufa, “Data-Driven Predictive Modeling of Highway Construction Cost Items,” *J. Constr. Eng. Manag.*, vol. 147, no. 3, p. 04020180, Dec. 2021.

[16] M. Awada, F. J. Srour, and I. M. Srour, “Data-Driven Machine Learning Approach to Integrate Field Submittals in Project Scheduling,” *J. Manag. Eng.*, vol. 37, no. 1, p. 04020104, Nov. 2021.

[17] C.-L. Fan, “Defect Risk Assessment Using a Hybrid Machine Learning Method,” *J. Constr. Eng. Manag.*, vol. 146, no. 9, p. 04020102, Jun. 2020.

[18] J. Brownlee, “How Much Training Data is Required for Machine Learning?,” *How Much Training Data is Required for Machine Learning*. p. 26, 2017.

[19] C. Mewald, “Data: A key requirement for your Machine Learning (ML) product,” *Medium*, 2018. .

[20] R. M. Fifer, “Cost benchmarking functions in the value chain,” *Plan. Rev.*, vol. 17, no. 3, pp. 18–19, Mar. 1989, doi: 10.1108/eb054255.

[21] M. S. El-Mashaleh, R. Edward Minchin, and W. J. O’Brien, “Management of Construction Firm Performance Using Benchmarking,” *J. Manag. Eng.*, vol. 23, no. 1, pp. 10–17, Jan. 2007, doi: 10.1061/(asce)0742-597x(2007)23:1(10).

[22] M. Bonilla and T. Castillo, “Benchmarking the construction industry: An adaptation of the world management survey methodology,” in *IGLC 28 - 28th Annual Conference of the International Group for Lean Construction 2020*, 2020, pp. 217–228, doi: 10.24928/2020/0057.

[23] S. McCabe, *Benchmarking in Construction*. Blackwell Science, 2001.

[24] CII, “Determinants of Jobsite Productivity,” 2001. Accessed: Feb. 01, 2022. [Online]. Available: https://www.construction-institute.org/resources/knowledgebase/knowledge-areas/construction-execution/topics/rt-143/pubs/rr143-11.

[25] H.-S. Park, “Development of a Construction Productivity Metrics System (CPMS),” 2002.

[26] L. Liang, “Small project benchmarking,” 2005.

[27] B. G. Hwang, S. R. Thomas, D. Degezelle, and C. H. Caldas, “Development of a benchmarking framework for pharmaceutical capital projects,” *Constr. Manag. Econ.*, 2008.

[28] COAA, “Alberta Report I - COAA Major Projects Benchmarking Summary,” Edmonton, Alberta, 2009. [Online]. Available: https://www.coaa.ab.ca/COAA-Library/COP-BEN-RES-01-2009-v1 The Alberta Report - COAA Major Projects Benchmarking Summary.pdf.

[29] P.-C. Liao, “Influence Factors of Engineering Productivity and Their Impact on Project Performance,” p. 251, 2008, [Online]. Available: http://books.google.com/books?id=xkseUGO9fKEC&pgis=1.

[30] V. Sharma, S. Yun, D. P. Oliveira, S. P. Mulva, and C. H. Caldas, “Development of a cost normalization procedure for national health care facility benchmarking,” *Proc. ICSC15 Can. Soc. Civ. Eng. 5th Int. Constr. Spec. Conf.*, 2015.

[31] S. Yun, J. Choi, D. P. Oliveira, S. P. Mulva, and Y. Kang, “Measuring project management inputs throughout capital project delivery,” *Int. J. Proj. Manag.*, vol. 34, no. 7, pp. 1167–1182, Oct. 2016, doi: 10.1016/j.ijproman.2016.06.004.

[32] CII, “Federal Facilities Data Analytics Research and Applications Program (FF-DARAP),” 2021. https://www.construction-institute.org/resources/performance-assessment/federal-facilities-benchmarking (accessed Feb. 01, 2022).

[33] J. Dai, S. Mulva, S.-J. Suk, and Y. Kang, “Cost Normalization for Global Capital Projects Benchmarking,” in *Construction Research Congress 2012*, 2012, pp. 2400–2409, doi: 10.1061/9780784412329.241.

[34] J. Choi, S. Yun, and D. P. de Oliveira, “Developing a cost normalization framework for phase-based performance assessment of construction projects,” *Can. J. Civ. Eng.*, 2016, [Online]. Available: 10.1139/cjce-2016-0223.