Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

PREDICTING JOB WAITING TIMES IN A STOCHASTIC SCHEDULING ENVIRONMENT USING SIMULATION AND REGRESSION MACHINE LEARNING MODELS

Ivan Kristianto Singgih
Stefanus Soegiharto

Department of Industrial Engineering
University of Surabaya
Surabaya, INDONESIA

ABSTRACT

Scheduling real systems is complicated because of the consideration of various working conditions. Although various combinatorial optimization methods, ranging from mathematical models, heuristics, metaheuristics, etc., have been developed, these methods could require a long computational time due to the complexity of the problems. This study proposes a framework to understand the system's behavior using regression machine learning techniques. The considered system could be any type, e.g., the flow shop, job shop, and their variants, with a certain scheduling method. The framework consists of (1) the development of the simulation for generating the data and (2) how the data could be used for training the regression machine learning models. An example of the stochastic single-machine problem with the First-In-First-Out rule is considered. The framework could be used to simplify the process of understanding the system's behavior without solving the optimization problem, which could be time-consuming.

1 INTRODUCTION

The emergence of machine learning (ML) models is not negligible. It affects how people solve their daily problems through its combination with many technologies. The use of big data allows the ML models to understand systems much faster, which helps decision-makers make faster and better decisions.

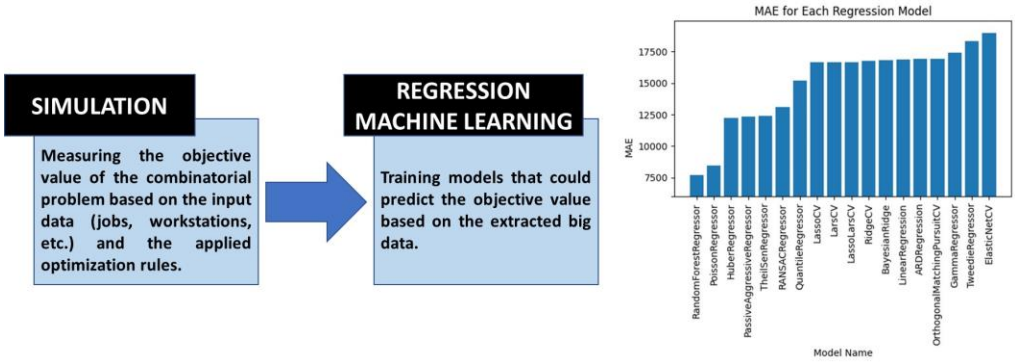
ML has also been used together with operations research (OR) techniques in several ways, e.g., (1) using ML to generate the input data for the OR method (Gumuskaya, et al., 2021), (2) using ML to improve the performance of OR methods or to solve OR problems, e.g., through heuristic selection in a hyper-heuristic and feature design in metaheuristics (Arnold and Sörensen, 2019), etc., and (3) using OR to improve the quality of ML techniques, e.g., metaheuristics for image processing cases. This study proposes a novel way to utilize simulation and regression machine learning techniques for understanding the behavior of a scheduling problem.

This study is closely related to Singgih (2021) because both studies analyzed big data generated from a production system using machine learning techniques to identify important features of the system. However, this study is different from Singgih (2021) because it (1) uses the regression models instead of the classification models in Singgih (2021) and (2) attempts to observe a simpler case (scheduling problem), which could allow researchers to identify more practical decisions for a specific problem, instead of a large black-box system, which was studied by Singgih (2021).

2 PROPOSED FRAMEWORK

The proposed simulation-machine learning framework is shown in Figure 1(a). In the first phase, a SimPy-based simulation is executed. For example, this study considered a single-machine scheduling problem as the testbed. The input data used for generating the production environment are: (1) interarrival time of the jobs that follows an exponential distribution, (2) mean job processing time, and (3) standard

deviation of the job processing time (assuming that the job processing time follows the uniform distribution). At the end of each 10,000 minutes-simulation run, the total waiting times of all jobs are measured. A job waiting time of is measured from the job's processing start time minus its arrival time. The number of simulation runs is 1,000.



(a) (b)

Figure 1: (a) The proposed simulation-machine learning framework; (b) Experiment results.

In the second phase, we use the three sets of input data for predicting the total waiting times. The portion of training and testing data are 75% and 25%. Before performing the prediction, the data are normalized using the MinMaxScaler Python library. Various regression machine learning models from scikit-learn (scikit-learn, 2023) are tested: (1) Random Forest Regression, (2) Linear Regression), (3) RidgeCV, (4) ElasticNetCV, (5) LarsCV, (6) LassoCV, (7) LassoLarsCV, (8) OrthogonalMatchingPursuitCV, (9) ARDRegression, (10) BayesianRidge, (11) HuberRegressor, (12) QuantileRegressor, (13) RANSACRegressor, (14) TheilSenRegressor, (15) PoissonRegressor, (16) TweedieRegressor, (17) GammaRegressor, and (18) PassiveAggressiveRegressor.

3 NUMERICAL EXPERIMENT RESULTS

The obtained mean absolute error (MAE) values for each regression machine learning model are shown in Figure 1(b). The required time to run all 18 regression machine learning models is very short, 36 seconds. The best models are the RandomForestRegressor and PoissonRegressor, with MAE values equal to 7669.86 and 8449.09, respectively.

4 CONCLUSIONS

This study introduces how simulation and machine learning models could be used to understand the behavior of production systems that consider uncertainty in job arrival times and processing times. It opens more opportunities to test how well various solution methods could be used in solving combinatorial optimization problems.

REFERENCES

Arnold, F., and K. Sörensen. 2019. "What Makes a VRP Solution Good? The Generation of Problem-Specific Knowledge for Heuristics." *Computers & Operations Research* 106:280-288. https://doi.org/10.1016/j.cor.2018.02.007

Gumuskaya, V., W. van Jaarsveld, R. Dijkman, P. Grefen, and A. Veenstra. 2021. "Integrating Stochastic Programs and Decision Trees in Capacitated Barge Planning with Uncertain Container Arrivals." *Transportation Research Part C* 132:103383.

scikit-learn: Machine Learning in Python. 2023. https://scikit-learn.org/, Accessed 6 October 2023.

Singgih, I.K. 2021. "Production Flow Analysis in a Semiconductor Fab using Machine Learning Techniques." *Processes* 9(3):407. https://doi.org/10.3390/pr9030407