A Machine Learning Approach to Waiting Time Prediction in Queueing Scenarios

Athanasios I. Kyritsis and Michel Deriaz
Information Science Institute, GSEM/CUI
University of Geneva
Geneva, Switzerland
Email: {$athanasios.kyritsis, michel.deriaz$}@unige.ch

***Abstract*—Physically queueing is a reality on many industries that provide services or sell goods. Waiting in a queue can be stressful and exhausting for the clients because of the enforced idle time, and may lead to decreased customer satisfaction. Queueing theory has been widely used to assess client waiting times, to optimize staff schedules, and to increase the robustness of a queueing system against a variable demand for service. In this paper, we are exploring how multiple industries that require queues can benefit from machine learning to predict the clients' waiting times. We begin by predicting waiting times on bank queues, and then we propose how the procedure can be generalized to more industries and automatized. A publicly available dataset containing entries of people queueing in banks is initially utilized, and after training a fully connected neural network, a mean absolute error of 3.35 minutes in predicting client waiting times was achieved. We are then presenting a web application that is managing queues of different scenarios and industries. The queues may have unique parameters, and the system can adapt to each queue as it creates a per queue optimally trained neural network for waiting time prediction. The use and the capabilities of the system are validated with the use of a simulator. Machine learning, therefore, proves to be a viable alternative to queueing theory for predicting waiting time.**

***Keywords*—Machine learning, pattern recognition, queue system, queueing analysis, web application.**

I. Introduction and Related Work

Queueing problems exist when multiple people need access to a resource, and the service cannot match the level of demand. On the one hand, queueing is a necessary evil when accessing valuable resources like health-related ones, as the idle time of those resources is expensive [1]. On the other hand, queues may get unnecessary big even for this purpose and may lead to long idle times for the customers. Time is a valuable resource, and consumers have to make decisions regarding the use of it when purchasing services or goods [2]. The idea that "time is money" has a long history, and the earliest recorded version has been attributed to Antiphon of ancient Greece (ca. 430 BC) [3]. There is also a link between long waiting times and customer dissatisfaction [4], and thus, industries should strive for better resource allocation that will optimize waiting queues.

This work was co-financed by Innosuisse.

Queueing optimization techniques are used in several industries to improve customer service. In the healthcare sector, queueing models are used to improve resource utilization in hospitals [5], and to handle the tradeoffs that will improve the efficiency and the quality of the provided services of healthcare systems [4]. In pharmacies with high workloads or multiple points of service, queueing models can assess service and waiting times [6]. In airports, queue planning is necessary for security controls and the check-in process [7]. Queueing lines can also be observed in everyday life activities, such as paying at groceries, waiting for a table at a restaurant, or waiting to order at a fast-food restaurant. While research has shown that longer waiting times may sometimes lead to higher consumption in some cases [8], most queueing scenarios would benefit from queue management to reduce the cost on the used resources, in terms of personnel cost and consumer waiting times.

Queueing theory attempts to study the waiting lines through mathematical analysis. The earliest studied problems in this domain concerned the congestion of telephone traffic and were investigated by the mathematician A. K. Erlang [9]. A queue can be modeled as a First-In First-Out (FIFO) node in which clients arrive, possibly wait for some time, and take some time to be processed by the server before they depart from the queue [10]. The Kendall's notation [11] is typically used to describe a queueing node in the form of A/S/c, where A is the probability distribution of durations between each arrival to the queue, S the probability distribution of service times and c the number of servers at the node. Using queueing theory, queue managers aim to balance a queueing system's service to customers by keeping the queues short and thus the waiting times low (that might mean an increased number of servers), and the economic upkeep of the system (that would mean to keep the number of servers low). By constructing a queueing model, the waiting time for any client, and the queue length at any time can be predicted.

Machine learning techniques and simulation models can also be of use to deal with queueing problems. An improvement to the prediction of the overall waiting time for daily radiation treatment appointments, when compared to the rough estimates that are typically given to patients, was achieved by a regression model used in an oncology department [12]. Machine learning was also used in another

978-1-7281-4087-2/19/$31.00 ©2019 IEEE
DOI 10.1109/AI4I46381.2019.00013

study to predict patient waiting and facility delay times for radiology examinations [13]. Simulations can also be used to model queueing problems [14]. By simulating different distributions for client arrival and service times, as well as a variable number of available servers, it is possible to predict the expected performance of a queue system and to optimize it for specific scenarios.

Several enterprise queue management solutions offer businesses the tools to manage queues, and thus reduce waiting times and improve service efficiency. To the best of our knowledge, there is currently no generic queueing system, where anybody can set up a queue that anyone can join. We are proposing QueueForMe, a system with which anyone can register as a creator and open a queue by defining an initial set of parameters unique to that queue, along with the response options for those parameters. Such a parameter could be the number of luggage for the scenario of a queue for checking in at the airport, for example. Each client that joins the queue responds to all the necessary parameters, and the expected waiting time for the client is inferred using machine learning. The machine learning model is taking into account the time the client joined the queue, the position of the client in the queue, the number of available servers for the queue, as well as the responses of the client to the aforementioned additional parameters. The model is continuously adapting to potential changes in queue patterns by using data from past clients.

The rest of the paper is organized as follows. In Section II, we are presenting a neural network-based waiting time predicting solution for the scenario of clients waiting to be served in a bank. We are then proposing our generic queueing system in Chapter III that can be used in various industries and can exploit queue specific parameters when predicting the estimated waiting time of the clients. We validate the learning capabilities of the queueing system with a simulator we have built. Finally, we conclude our work in Section IV.

II. Waiting Time Prediction in a Bank Scenario

1) Dataset: In order to lay the basis for how machine learning can be utilized to predict waiting times of new clients joining a queue, we are using a dataset published by Bishop et al. [15]. This dataset includes information regarding queues formed and served in 3 banks in Ogun State, Nigeria, over a period of 4 weeks. In total, 52444 clients are reported, and each entry includes the time the client joined the queue, the waiting time, the service time, and the total time in the system (waiting time + service time). Unfortunately, the number of servers is not reported in the dataset. It is evident from the throughput of the system that the number of servers is greater than zero for every case, but it is unknown if it was a constant number throughout the days.

There are no missing values in the dataset. The waiting time variable represents the amount of time in minutes each user had to wait in the queue before being served and is the output variable that our machine learning model is trained to predict. It has a mean value of 13.18, a median of 12, and a standard deviation of 5.95 minutes. The histogram of the waiting time values of the dataset is presented in Fig. 1.

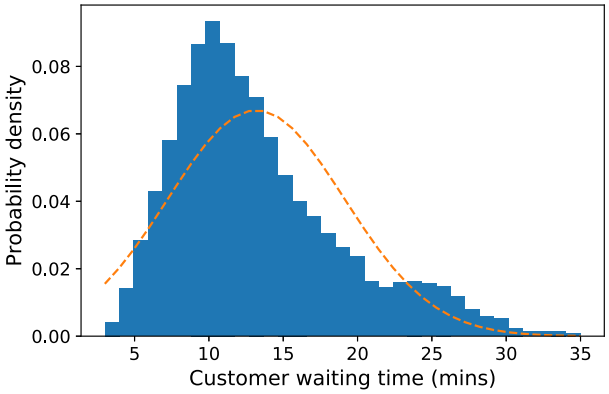

Fig. 1. Histogram of the customer waiting time in minutes.

2) Data Preprocessing and Feature Engineering: From the given dataset, for each client, we have calculated the number of people waiting in the queue at the time the client joined the queue. To do so, we have calculated the queue departure time for each client by adding the waiting time to the arrival time, and then we counted the number of people that were yet to depart from the queue at the time a new client joined the queue. Fig. 2 presents the histogram of the number of people waiting in the queue.

From the timestamp that the client joined the queue, we extracted 3 features; the day of the week (ranging from 0 to 4, for Monday to Friday), the hour (ranging from 8 to 14), and the minutes (ranging from 0 to 59). We used mean encoding, also known as target encoding, to encode new features from those 3 existing categorical features and the target variable. The idea of mean encoding for a regression task is simple. Let x be a categorical variable and y a target variable. For each distinct element in x we are computing the mean of the corresponding values in y. Then each entry of x_i in the feature vector is replaced with the corresponding mean.

In total, we are using 4 features as the input for modeling, the people waiting in the queue, the day of the week, the hour, and the minutes, and the waiting time variable as mentioned above forms the output.

3) Experimental Setup: For the machine learning experiment presented in this study, we have used Python 3.7.3 and Tensorflow 2.0.0-beta1. To evaluate the performance of our system, we split the available dataset into a training set (80%) and a test set (20%), keeping the test set completely unseen during the training phase. We have decided to use a neural network over other machine learning models because of the continuous training capabilities of a neural network. When a new batch of training data is gathered, an existing neural network can be trained solely on those, and there is no need to train on all available training data regularly. This is ideal for a queue management system, as new data are consistently

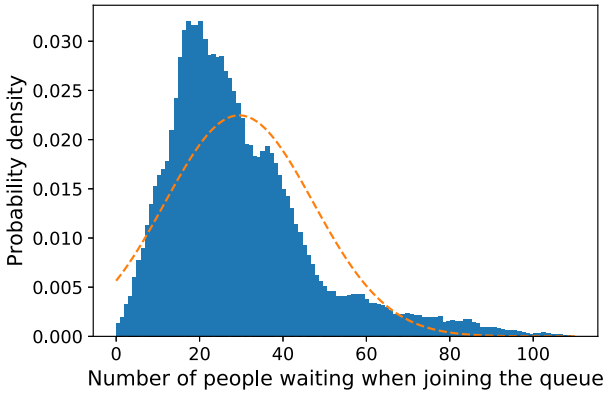

Fig. 2. Histogram of the people waiting in the queue.

being produced.

4) Results: We have trained a fully connected neural network with 2 hidden layers, the first with 12 neurons and the second with 8 ones. We have used the Rectified Linear Unit (ReLU) [16] as the activation function of all hidden layers, and the Adam optimization algorithm [17] for the iterative update of the network weights based on the training data. An exhaustive tuning of all related hyperparameters is out of the scope of this paper, and commonly used default values for the architecture of the neural network were used.

After 500 epochs of training, a mean absolute error of 3.35 minutes was achieved on the test set. To have an estimation of the predictive capability of the trained model, we are comparing against the naive mean and the naive median model; these are the models that always predict the mean and the median waiting time of the training set. The naive mean model had a mean absolute error of 4.71 minutes on the test set, and the naive median an error of 4.59 minutes. Our model has, therefore, achieved an improvement of 28.9% over the naive mean model and 27% over the naive median one. Unfortunately, since there is no information regarding the deployed servers, we can not compare the predictive performance of our model with one produced by queueing theory.

In order to make our research reproducible, we are sharing the code that we used for the tests of the current study here (https://doi.org/10.5281/zenodo.3378407). The dataset used in this analysis is available here (https://doi.org/10.1016/j.dib.2018.05.101).

III. Proposing a Generic Queueing System

1) QueueForMe: We are proposing QueueForMe, a web application that allows everyone to create a virtual queue allowing clients around to join. QueueForMe includes two types of users, the creators, and the clients. The creators are the ones that need to log in to our platform in order to create a queue by providing a queue name and a description. A client does not need any credentials for the platform, can search

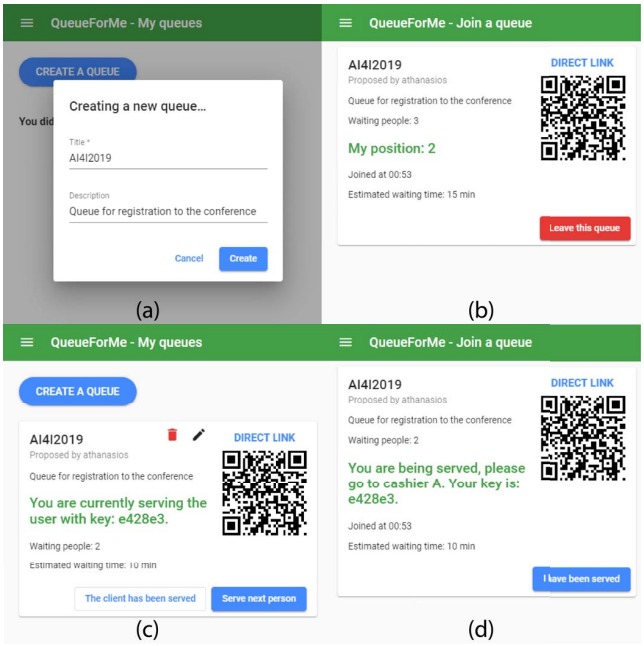

Fig. 3. Views of the QueueforMe web application when (a) the creator is creating a new queue, (b) the client has just joined a queue, (c) the creator is serving a new client, and (d) the client is being served by the creator.

for a queue by queue name or by the creator's name, and can finally join the queue. After joining the queue, the client is getting information about the queue including the position in the queue along with an estimated waiting time predicted by a neural network trained for the specific queue using past queue data. The creator can operate the queue by asking for the next available client in the queue, and the corresponding client is getting a notification to be served. Screenshots of the QueueForMe web application can be seen in Fig. 3.

2) Additional Queue Parameters: In a future version of QueueForMe, the creator of a queue will have the option to define a set of queue specific parameters along with the restricted set of responses for each parameter. Clients joining the queue will have to give a response to each of the defined parameters. These additional parameters set our solution apart since such customizable additional parameters do not exist in queueing theory, in a manner that can be automatically exploitable. The corresponding neural network of each specific queue will be able to extract more valuable information and detect patterns among those parameters.

3) Simulator: To verify the learning capabilities of each neural network that is created with every queue, we have developed a simulator with which we can simulate customers joining a queue over a specific period of time. In our tool, we can select the distribution at which new clients join the queue, the number of available lines and the distribution of the service time. A queue specific neural network is then retrained learning from the provided simulated customers that have passed from the queue. The trained model is also saving statistics regarding its predictive performance. Screenshots of

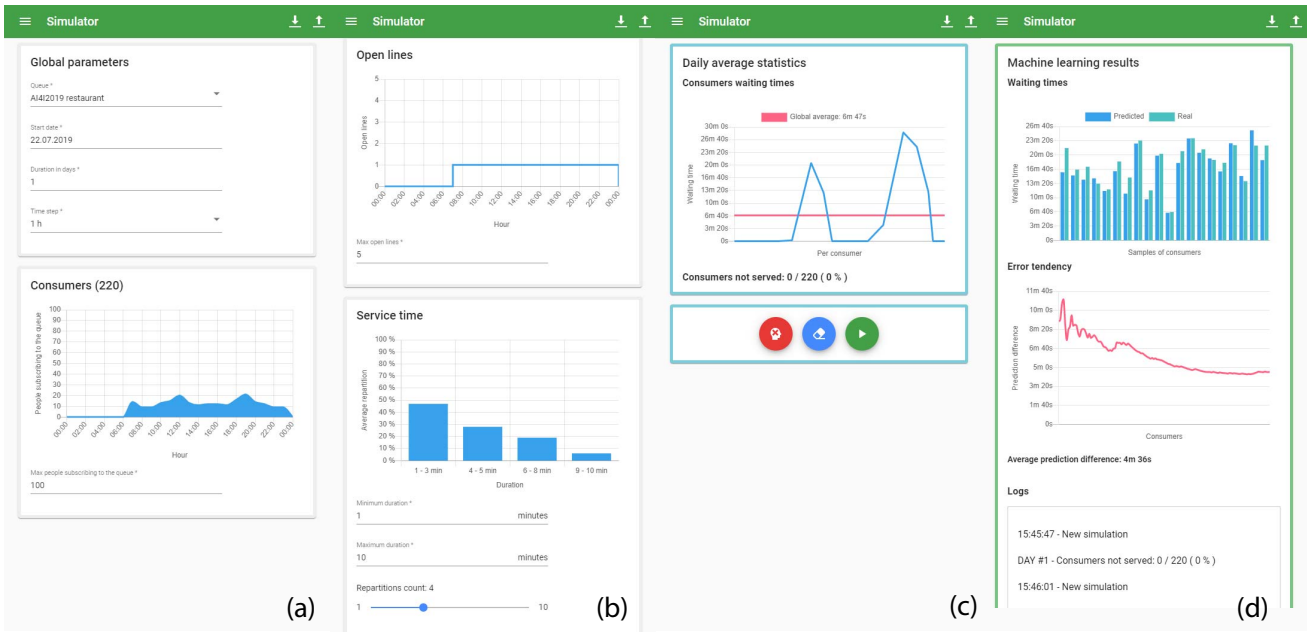

Fig. 4. Views of the simulator web application (a) setting global queue parameters with the customer arrival distribution, (b) setting the number of open lines and the service time distribution, (c) displaying the average statistics for the simulated dataset and starting a simulation, and (d) getting the error of the corresponding neural network.

the simulator can be seen in Fig. 4.

4) Industry Specific Queue Simulation: In a future version of the simulator we will also include the queue specific additional parameters as described above. Each queue will have a predefined set of parameters, and each parameter will have a set of valid responses. Using the simulator, the operator will be able to simulate different distributions for each response to each parameter. New features fed to the neural network will be engineered by aggregating the parameter responses of all previous queuers in relation to a current queuer.

IV. Conclusion

In this paper, we have explored how machine learning can be used for predicting the waiting time of people queueing in lines. We have started by using a publicly available dataset of queues in banks, and by training a neural network, we achieved a mean absolute error of 3.35 minutes, improving over the performances of naive models. Unfortunately, we could not directly compare against queueing theory because the dataset lacked information about the deployed servers. After presenting a specific case on how machine learning can be used to predict waiting times in queueing scenarios, we are generalizing on more industries. We presented our work on QueueForMe, a web application that allows everyone to create a virtual queue, and everyone around to join. Using a simulator, we can verify the predictive capabilities of the queue specific neural networks. As future work, we will include the ability to add queue specific parameters at the queue creation phase with predefined responses. This additional information will be exploited by the underlying waiting time-predicting neural network of each queue, and different distributions of each parameter response can be simulated using the simulator.

References

[1] D. Worthington, "Queueing models for hospital waiting lists," *Journal of the Operational Research Society*, vol. 38, no. 5, pp. 413–422, 1987.

[2] C. Mogilner, H. E. Hershfield, and J. Aaker, "Rethinking time: Implications for well-being," *Consumer Psychology Review*, vol. 1, no. 1, pp. 41–53, 2018.

[3] B. Perrin, *Plutarch's Lives*, ser. Loeb classical library. W. Heinemann, 1920, no. v. 9. [Online]. Available: https://books.google.ch/books?id=bSJgAAAAMAAJ

[4] A. Komashie, A. Mousavi, P. J. Clarkson, and T. Young, "An integrated model of patient and staff satisfaction using queuing theory," *IEEE journal of translational engineering in health and medicine*, vol. 3, pp. 1–10, 2015.

[5] S. Belciug and F. Gorunescu, "Improving hospital bed occupancy and resource utilization through queuing modeling and evolutionary computation," *Journal of biomedical informatics*, vol. 53, pp. 261–269, 2015.

[6] R. A. Nosek Jr and J. P. Wilson, "Queuing theory and customer satisfaction: a review of terminology, trends, and applications to pharmacy practice," *Hospital pharmacy*, vol. 36, no. 3, pp. 275–279, 2001.

[7] R. De Neufville, L. Budd, and S. Ison, "Airport systems planning and design," *Air transport management: An international perspective*, vol. 61, 2016.

[8] S. Ülkü, C. Hydock, and S. Cui, "Making the wait worthwhile: Experiments on the effect of queueing on consumption," *Management Science*, 2019.

[9] A. K. Erlang, "The theory of probabilities and telephone conversations," *Nyt. Tidsskr. Mat. Ser. B*, vol. 20, pp. 33–39, 1909.

[10] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of queueing theory*. John Wiley & Sons, 2018, vol. 399.

[11] D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain," *The Annals of Mathematical Statistics*, pp. 338–354, 1953.

[12] A. Joseph, T. Hijal, J. Kildea, L. Hendren, and D. Herrera, "Predicting waiting times in radiation oncology using machine learning," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 1024–1029.
[13] C. Curtis, C. Liu, T. J. Bollerman, and O. S. Pianykh, "Machine learning for predicting patient wait times and appointment delays," *Journal of the American College of Radiology*, vol. 15, no. 9, pp. 1310–1316, 2018.
[14] M. Bahadori, S. M. Mohammadnejhad, R. Ravangard, and E. Teymourzadeh, "Using queuing theory and simulation model to optimize hospital pharmacy performance," *Iranian Red Crescent Medical Journal*, vol. 16, no. 3, 2014.
[15] S. A. Bishop, H. I. Okagbue, P. E. Oguntunde, A. A. Opanuga, and O. Odetunmibi, "Survey dataset on analysis of queues in some selected banks in ogun state, nigeria," *Data in brief*, vol. 19, pp. 835–841, 2018.
[16] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.
[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.