The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2023

Volume 24, Pages 96-100

IConTech 2023: International Conference on Technology

Enhancing Call Center Efficiency: Data Driven Workload Prediction and Workforce Optimization

Muhammet Ali Kadioglu
Dogus Teknoloji R&D Center
Istanbul Technical University

Bilal Alatas
Firat University

Abstract: Organizations can improve customer service quality, reduce wait times, and enhance overall operational efficiency by aligning staffing levels with predicted workload volume. Decision makers in the call centers gain valuable insights and practical guidance from the integration of workload forecasting and workforce optimization. Businesses can effectively utilize their personnel and resources by accurate workload forecasting and workforce optimization. Faster and more profitable services can be provided at customer contact points. It also increases employee satisfaction and enhances the organization's competitive advantage. A tailored solution is essential because every issue has its distinct dynamics. The two-layered pipeline known as "Predict and Optimize" is created by combining ML approaches for forecasting and mathematical programming techniques for optimization. The method offers a comprehensive solution for call center managers seeking to improve resource allocation and boost operational performance. In this study, we have tried to predict future workload levels by training a LSTM model and used integer programming techniques to optimize the allocation of available staff resources according to the forecasted workload. The workforce optimization model generates minimum staffing requirements by considering call center-specific various constraints.

Keywords: Predict and optimize, Time series, Integer programming

Introduction

Workload forecasting and workforce planning are critical for optimizing the operational efficiency of call centers. They are the two main parts of the proposed "Predict and Optimize" approach for call center operations. The effectiveness of the prediction layer can be quantitatively evaluated by comparing the forecasted and actual call volumes. The staffing levels suggested by the optimization model can be used to make predictions as valuable insights. A customized framework for managing data driven projects can greatly enhance innovations within organizations (Kadioglu & Takci, 2022). By operationalizing data science models, organizations can achieve exceptional AI outcomes. Researchers' interest has grown as a result of call centers' expanding role in operations management within the services industry. Most articles in this field focus on large call centers with several hundred connected agents. Chevalier and Schrieck, similar to us, deviate from this trend by focusing specifically on small-size call centers. These call centers may operate in business-to-business (B2B) companies or be part of small-scale business-to-consumer (B2C).

Planning in call centers leverage numerical analysis of queueing models, mathematical modeling, and decision-making based on economic and technical performance measures for improved outcomes (Stolletz, 2003). Workforce management typically begins with the Erlang formula or one of its generalizations (Koole, 2004). Nag and Helal present a comparative analysis of the Erlang A and Erlang C models for inbound calls. Prior to

this study, research was conducted in a different call center operating within a B2C. Forecasting for incoming and outgoing calls was made using machine learning (ML) models and planning was completed using Erlang C.

Hybrid solutions offer an adaptable approach to planning processes, allowing for better solutions. An approach can be created by a combination of methods and compatibility between the chosen methods is crucial to ensure seamless integration and cohesive functioning. By adapting and integrating these methods, it is possible to construct a high-performance pipeline that effectively utilizes the strengths of each approach. This enables the development of a robust and efficient solution for complex planning problems. The combination of Long Short-Term Memory (LSTM) based workload forecasting, and workforce optimization can provide call center managers with a data-driven approach to make informed decisions, optimize resource allocation, and improve operational performance. An integer program is typically used to solve the problem of choosing staff shifts that meet the requirements (Ertogral & Bamuqabel, 2008). Using real-world call center data, we evaluate the proposed approach and show how well it performs in terms of accurately forecasting workload and allocating the workforce. Operation Research (OR) and Machine Learning (ML) approaches are needed to overcome the limitations of existing methods. Linear programming offers a structured and mathematical foundation for optimization, while deep learning models can extract insights from data and make predictions. Organizations can achieve effective workforce management and better decision-making by combining these approaches.

Literature Review & Model Design

Workload Prediction

The time series prediction techniques can be classified as statistical approaches, ML approaches, and Deep Learning (DL) approaches. The statistical approaches employ the linear prediction structure, and they are widespread forms of predicting workload for stable datasets. In our first research for call center operations, we applied this approach to produce a quick win solution. On the other hand, we need to remember that many datasets are unstable. AutoML can be a substantial starting point in data science projects (Kadioglu, 2022). As well, ML approaches are widely used for workload prediction. So, we have improved our previous solution using ML algorithms. Finally, DL approaches are also applied for workload prediction in recent years. This study uses LSTM as a deep-learning method for workload prediction. According to Gao et al. (2020), they show LSTM can achieve higher accuracy than the Autoregressive method, artificial neural network (ANN), and autoregressive integrated moving average (ARIMA) in load prediction.

Additional information like holidays, promotional activities, and other special events may improve model predictions. These events and occasions often introduce variations and outliers in the data that can significantly influence patterns, trends, and behavior. (Albrecht et al., 2021) By incorporating holiday and event information into the analysis, models can provide more accurate predictions, identify outliers, and capture the nuances of customer behavior during these special periods. This additional context can improve decision-making for call centers.

Workforce Planning

To accurately estimate Customer Service Agent (CSA) requirements based on forecasted incoming call volumes, various mathematical and statistical methods can be employed. Among these, queuing theory, particularly the M/M/c model, also provides valuable insights by modeling the incoming call process. The Erlang C formula stands as a traditional choice, considering factors such as average call duration and time between calls to determine optimal staffing. Kanthanathan et al. (2020) used the Erlang C queuing theory and the forecasted incoming call rate per day to estimate the required number of CSAs per day.

More complex systems may opt for simulations, running numerous iterations to identify a range of outcomes and thus better predict staffing needs. Further granularity can be achieved through workload distribution, breaking down predictions by day and hour to accommodate fluctuating call volumes. Optimization algorithms can also be a highly effective approach for determining the optimal staffing levels in a call center based on projected call volumes. Optimization involves making the best (most profitable or least costly) use of resources under a set of constraints. While each method carries its assumptions and limitations, selecting or combining them based on specific call center attributes can facilitate efficient staffing, minimizing customer wait times and idle agent periods.

In the context of a call center, the objective might be to minimize the total staffing cost or to minimize the customer waiting time, subject to constraints such as the forecasted call volumes, the service rate of the CSAs, and the minimum and maximum number of agents available at any given time. An optimization model for a call center could have variables representing the number of agents working in each time slot, and constraints ensuring that the number of agents working at any time is enough to handle the projected call volume during that period. One of the advantages of using LP for this kind of problem is its flexibility. Optimization models can be easily adjusted to account for different types of shift patterns, different service level requirements, and other operational constraints. In addition, there are many efficient algorithms available for solving optimization problems, so solutions can often be obtained quickly even for large-scale problems.

Practical Implementation

Different forecasting methods can be used to predict the number of calls for short time intervals, such as mean values, exponential smoothing, ARIMA-models, and neural networks (Stolletz, 2003). In this study, we propose a comprehensive approach that combines time series forecasting with optimization techniques for call centers. The workload forecasting component utilizes LSTM, a powerful deep learning model, to capture the temporal patterns and dependencies in historical call volume data. LSTM model has become widely used in the field of DL. This neural network model is particularly well-suited for handling time-series data. By incorporating forget gates and input gates, the model effectively reduces the impact of pathological data, ensuring robust performance and high accuracy in prediction tasks. The LSTM model can efficiently identify long-term dependencies and patterns in sequential data due to its ability to selectively retain or discard information at various time steps.

When investigating prediction methods, two approaches have been identified. The first method, referred to as the "0-gap" prediction method, involves making predictions without any time gap between the input and output. This means that the prediction can be completed prior to the actual occurrence, but it leaves limited time for scheduling tasks before the actual workload appears. To address this limitation, an alternative technique known as the "m-gap" prediction method has been proposed. In the m-gap prediction method, a time window gap of m is introduced between the last time point of the input data and the output time point. This approach allows for workload forecasting to be performed a certain time before the predicted time point, thereby providing much time for task scheduling based on the predicted workload. Gao et al. (2020) presented the m-gap prediction method specifically for workload prediction, emphasizing the importance of leaving sufficient time for task scheduling based on the predicted workload.

Workload Prediction

The dataset used for experimentation was preprocessed. Any missing values were filled with zeros. Min-Max scaling technique was applied to normalize the data. The LSTM model was implemented using the Keras library, following a sequential architecture. It comprised three LSTM layers, each followed by a dropout layer to prevent overfitting. The number of LSTM units in each layer was set to 512, 512, and 256, respectively. The final layer of the model was a dense layer with a number of units corresponding to the prediction period, activated by the Rectified Linear Unit (ReLU) function.

Prior to training the model, the dataset was divided into training and testing sets. For the training set, input sequences (X_train) were created by selecting the previous n_steps values of all features, while the target variables (y_train) for the subsequent period time steps. Similarly, the testing set was constructed. The LSTM model was trained using the Adam optimizer with a custom learning rate of 0.001, optimizing the mean squared error (MSE) loss function. The training process was conducted for a total of 10 epochs, with a batch size of 64. Once the model was trained, it was evaluated using the testing data, and the loss value was computed and reported. Finally, the model was employed to generate predictions for the upcoming week using the testing data (X_test), with the predicted values stored in the prediction's variable.

Workforce Planning

Determining the number and types of agents who will be handling calls and their work schedules is one of the main optimization challenges that must be overcome to manage call centers. This task must adhere to restrictions on acceptable schedules and service quality (Buist et al., 2008). In our case, there is a single group of

agents that can serve all inbound calls. The call center operates dialogue management, roadside assistance, operational services. We present the staffing problem for a single-skill call center. The objective is to minimize operating costs. The objective function represents the total number of agents required. When the assignments' time slots are defined, the day is divided into hourly periods. Our model can easily handle additional constraints as needed.

In the second phase, the LSTM model's predictions for the number of inbound calls and AHT have been integrated into an optimization framework. The predictions for inbound calls and AHT have been assigned as inputs in the optimization model. This framework, implemented using the Pyomo library, aimed to determine the minimum number of agents required to handle the projected workload. Decision variables are defined to represent the number of agents assigned to each hour and day. The objective function is to minimize the total number of assigned agents.

A constraint was formulated by comparing the predicted call volume with the maximum call load an agent could handle within a given hour. Additionally, a constraint was imposed to ensure that at least one agent was assigned to each hour, guaranteeing the presence of an agent to handle the workload. The optimization problem was solved using the GLPK solver, which searched for the optimal solution that minimized the total number of assigned agents while satisfying the constraints. Upon successful optimization, the minimum agent requirements were determined for each hour and day. These results provided valuable insights into resource allocation, facilitating effective scheduling and management decisions within the call center.

Findings & Conclusion

While our approach demonstrates significant improvements in both forecasting accuracy and operational efficiency, the complexity of LSTM models and the need for sufficient historical data may limit its application in smaller or newer call centers. Further research could explore methods for streamlining the LSTM training process or for incorporating other types of data into the workload forecasting model. On the other hand, Pot et al. (2008) addressed the staffing issue in multiskilling call centers, focusing on the labor allocation process. Buist et al. (2008) discussed the problems of optimizing staffing and scheduling in large multiskilling call centers. Their key contribution lies in developing a staffing method that enables the creation of schedules in multiskilling call centers, ensuring a reasonable alignment between projected workload and available labor capacity. Additional constraints to the optimization model can be incorporated.

The obtained MSE and MAE values suggest that the LSTM model performed reasonably well in capturing the underlying patterns and trends in the time series data, yielding predictions that were relatively close to the true values. Upon evaluating the LSTM model's performance on the testing data, we computed the Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics to assess the accuracy of the predictions. The MSE value was found to be 96.4373, indicating the average squared difference between the predicted and actual values. MAE value, on the other hand, was 6.6996, representing the average absolute difference between the predicted and actual values.

The LSTM model was also executed to predict the AHT, and the evaluation metrics for this specific prediction task were computed. MSE was found to be 312.3277, indicating the average squared difference between the predicted and actual AHT values. Additionally, the Mean Absolute Error (MAE) was determined as 13.9535, representing the average absolute difference between the predicted and actual AHT values. These evaluation metrics shed light on the performance of the LSTM model in predicting the AHT. The integration of the LSTM model predictions into the optimization framework provides informed staffing decisions. By incorporating the forecasted call volume and AHT, the model optimized agent assignments to ensure that service levels were met while minimizing operating costs.

Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Technology (www.icontechno.net) held in Antalya/Turkey on November 16-19, 2023.

References

Albrecht, T., Rausch, T. M., & Derra, N. D. (2021). Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivals' forecasting. *Journal of Business Research, 123,* 267-278.

Buist, E., Chan, W., & L'Ecuyer, P. (2008). Speeding up call center simulation and optimization by Markov chain uniformization. *2008 Winter Simulation Conference*.

Chevalier, P., & Van den Schrieck, J. C. (2008). Optimizing the staffing and routing of small-size hierarchical call centers. *Production and Operations Management, 17*(3), 306–319.

Ertogral, K., & Bamuqabel, B. (2008). Developing staff schedules for a bilingual telecommunication call center with flexible workers. *Computers & Industrial Engineering, 54*(1), 118–127.

Gao, J., Wang, H., & Shen, H. (2020). Machine learning based workload prediction in cloud computing. *2020 29th International Conference on Computer Communications and Networks (ICCCN).*

Kadioglu, M. A. (2022). End-to-end AutoML implementation framework. *The Eurasia Proceedings of Science Technology Engineering and Mathematics, 19,* 35-40.

Kadioglu, M. A., & Takci, H. (2022). *A data science project management methodology: From development to production* (p.2). Proceedings Book.

Kanthanathan, C., Carty, G., Raja, M. A., & Ryan, C. (2020). Recurrent neural network based automated workload forecasting in a contact center. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 1423-1428). IEEE.

Koole, G. (2004). Performance analysis and optimization in customer contact centers. *First International Conference on the Quantitative Evaluation of Systems, 2004*. QEST 2004.

Nag, K., & Helal, M. (2017). Evaluating erlang C and erlang A models for staff optimization: A case study in an airline call center. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM).*

Pot, A., Bhulai, S., & Koole, G. (2008). A simple staffing method for multiskill call centers. *Manufacturing & Service Operations Management, 10*(3), 421-428.

Stolletz, R. (2003). *Performance analysis and optimization of inbound call centers*. Springer Science & Business Media.

Ta, T., l'Ecuyer, P., & Bastin, F. (2016). Staffing optimization with chance constraints for emergency call centers. In *MOSIM 2016-11th International Conference on Modeling, Optimization and Simulation.*

Author Information

Muhammet Ali Kadioglu
Dogus Teknoloji R&D Center
Istanbul Technical University
Istanbul, Turkey
Contact e-mail: *kadioglumuhammetali@gmail.com*

Bilal Alatas
Firat University
Elazig, Turkey

To cite this article:

Kadioglu, M.A. & Alatas, B. (2023). Enhancing call center efficiency: Data driven workload prediction and workforce optimization. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 24,* 96-100.