Simulation-free approach for wait time prediction in a tele-queue

Anonymous Author(s)
Affiliation
Address
email

Abstract

One of the most important performance measure of a call center is the timeliness of its service. The waiting times callers experience directly impact their tendency to abandon the queue, their patience in subsequent phone conversations and their propensity to call back. From a call center manager's point of view, understanding the relationship between time, resource and service output is important for setting performance targets and managing caller expectation. In this paper, such relationship is explored using machine learning classification techniques, specifically regularized logistic regression combined with online learning and a k-nearest neighbor classifier. The factors identified as influential coincides with the findings from previous sensitivity analysis performed on a Discrete Event Simulation (DES) model which in essence models the dynamics between time, resource and service directly by mimicking call processing events. The resulting predictions have absolute percentage errors between 0.15 and 0.2 which are up to two folds higher than the corresponding performance from DES, suggesting that more explicit modeling of the autocorrelation in the data set may be needed.

1 Background

Call center is a widely-used customer-facing channel for many businesses in the service industries such as banking, insurance, and tele-communication. One of the most important Quality of Service (QoS) measure for a call center is the waiting times customers spend in tele-queues. A major challenge in controlling this QoS measure while balancing expenditure is the stochastic demand for call service, both in terms of volume and complexity of service requested.

1.1 Related-work

Many recent studies on controlling call center tele-queues focus on mathematical models based on Queuing theories[1] and simulation-based approaches[2] [3]. On the other hand, little association-based studies exist in the literature, perhaps due to the dynamics between many business-specific factors such as workforce size, industry, nature of calls, caller demographics and so on which lack generality. However, for network traffic which is a similar setting, machine learning techniques have been applied for response time prediction[4]. Network queues and tele-queues share similar dynamics due to the limits in resources (infrastructure for both networks and call centers, and agents for call centers) and stochasticity in demand and workload. The presence of staffing limits in the call center scenario adds complexity to wait time analysis by introducing measurement errors and biases due to associated organizational complexity.

1.2 Overview

In this paper, the problem of predicting caller wait times in a tele-queue is explored using simulation-free approaches. Models are trained and tested on a data set containing over 310000 data points, each representing a call received by a call center operated by a Canadian insurance company primarily on homogeneous customer service issues. Instead of predicting the average waiting time, the percentage of customers waiting for no longer than a certain time threshold is estimated because a user rarely experience the average wait time which is a single-point estimate of the underlying wait time distribution.To fully take advantage of the granularity available from the large data set, online learning is used to handle memory limits. The initial set of feature identified are limited to factors which are quantitatively measurable. These include time of the call arrival, day of the week, month of the year, and holiday status. Staffing level information available to the call center is also considered both for causal simulation modeling in previous studies and the association modeling discussed in this paper; however, due to known problems in the accuracy of staffing level measurement complicated by break schedule adherence and discrepancies between operational and financial division within the company's multiple business units, the staffing level information is less reliable compared to the rest of the data set. Both relationship-based and memory-based approaches are explored. The findings of this study are as follows: information on the time of call arrival itself does not seem to suffice in predicting the service level of incoming calls to the call center studied; staffing information, despite having inherent noises in the record, does provide additional prediction power, but only in the existence of break schedule information. This is consistent with the high traffic intensity[1] the call center experience in most of its operational hours.

The rest of the paper is organized as follows. Section 2 introduces the data set used in this study as well as some major findings from exploratory data analysis which are used in model building. Section 3 describes the predictive models built on regularized binary logistic regression and on a simple Instance Based Learning (IBL) approach. Section 4 presents the validation results from the models followed by analysis. Conclusion and future directions are discussed in the final section.

2 Data

Records of over 310000 calls received at a call center operated by a Canadian-based insurance company spanning over a recent two-year period were collected by the automatic call distribution system as part of the call center's infrastructure. These call-by-call records provided the time stamp related variables used in this study. The staffing information of the corresponding time span was collected by an independent financial management system deployed by the same company. Due to the staffing-sharing practice between various business units and fuzzy adherence to schedule, the records are considered an indicative, but not completely accurate proxy to the actual staffing levels available to the call center. Based on the results from some exploratory analysis, day of the week and month of the year information are incorporated as dummy variables during the training and testing of the logistic regression models, and time of the day is discretized into 15-minute intervals for both approaches. The prediction of service level is calculated from prediction of probability that a call with certain characteristics will wait within the corresponding time threshold. For example, the prediction for service level with 60 seconds as threshold will be made by performing binary logistic regression with 'WaitWithin60Seconds' (meaning whether a call will wait for less than or equal to 60 seconds) as the response variable. The name, description and type of each pre-selected explanatory variable are presented in Table 1.

3 Modeling

3.1 Benchmark

In the subsequent evaluation of performance by different models, the results from a DES model[5] built on the same data set will be used as benchmark; Mean absolute percentage error (MAPE)

[1] Traffic intensity can be interpreted as the ratio of arrival rate to service rate, or equivalently, the ratio of service time to inter-arrival time of a queuing system. Intuitively, higher traffic intensity means heavier workload on the servers, and the number of servers available becomes a more significant constraint on the system performance when the system is busy.

Table 1: Explanatory variables

Name	DESCRIPTION
TimeSlot	Index of the 15-minute time slot of a day a call arrived in (categorical)
Day	Day of the week (categorical)
Month	Month of the year (categorical)
StaffG	The gross staffing level of the day a call arrived on (numerical)
StaffB	Staffing level after adjusting for break schedule of the 15-minute time slot (numerical)
BeforeHoliday	Whether a call arrived within 2 days before a local statutory holiday (categorical)
AfterHoliday	Whether a call arrived within 2 days after a local statutory holiday (categorical)

will be used as the measure of prediction accuracy. The DES model has previously been validated against the historical data. For or the range of predictor variables for which historical data is sparse, the DES model was validated by subject-matter experts who are familiar with the operations of the call center the data set is captured from.

3.2 Model selection

Our interest in the predictor variables' influence naturally leads to a supervised learning setting. Besides simulation, two broad categories of techniques are available: memory-based approach where prediction is generated directly from examples without explicitly generalizing relationship between the predictors and the response; and rule-based (or model-based) approach where rules are established between the predictors and the response, ready to be applied to new values of predictor vectors. One special class of memory-based approach is IBL directly using training data [6], which is applied to the call center data set and compared to regularized logistic regression as an example of rule-based approach. One basic approach within IBL, K-nearest neighbors (KNN) is tested on this data set because of the abundance of training examples available, and the potential of speedup by reducing the size of training examples using many approaches such as the ones outlined by Wilson and Martinez [7]. Among the rule-based approaches, logistic regression is tested in this study primarily to provide insights directly related to measurable and controllable (staffing level and break schedule, for example) variables of the system that are easier to understand despite possible trade-off in classification performance.

3.2.1 Tuning of the hyperparameters

The regularization factor in the L1 logistic regression and the number of neighbors k for the KNN algorithm were selected by ten-fold cross validation. The original data set was divided into ten partitions and each trial value of hyperparameter was given a training set S consisted of nine of the ten partitions and returned the trained parameters p of the model based on S. Next, the trained parameters p are used to make predictions on the remaining partition (which is disjoint from S) and the MAPE in predicting service level is calculated. This process is repeated ten times so each partition is used once as the testing partition, and the average of the error from these ten tests are used to select among the trial values of hyperparameters.

3.3 Regularized logistic regression using Stochastic Gradient Descent

To take full advantage of the large number of data points available, model training was performed in an online fashion. Starting from a pre-selected set of features, feature selection is done automatically by L1 regularization. To circumvent the algorithm's sensitivity to feature scales, the numerical variables are scaled to [0,1] before training and testing. The dummy variables are already properly scaled by their definitions. The model is trained with the scikit-learn implementation [8] . To explore the extent to which the two human-captured inputs – staffing level and break schedule influence the prediction performance, a logistic regression model (Model 1) is first built without variable StaffG and StaffB, then StaffG is added (Model 2), finally a third model incorporates both StaffG and StaffB (Model 3).

3.3.1 Results of feature selection

Only dummy variables Monday and January were selected among the dummy variables of the same type for all three logistic regression models; BeforeHoliday and AfterHoliday were eliminated for all three models. StaffG was eliminated from Model 3, but was kept in Model 2 in absence of StaffB. StaffB turns out to contribute the most prediction power to the logistic regression models explored in this study. The signs of the trained coefficients are reasonable and consistent with the findings from simulation.

3.4 KNN

A simple KNN classifier is applied using uniform weights across features with Euclidian distance measure and implementation by Pedregosa et al.[8] . Features are included incrementally in the same way described in the 'Regularized logistic regression using Stochastic Gradient Descent' section. The model including feature StaffB yielded the highest prediction performance. For the simplicity and conciseness of presentation, only this model (Model 4) will be highlighted in the following section.

4 Results

Two levels of prediction performance are examined: the predicted service level for a given day of the week, month of the year, time of the day (the index of 15-minute times lot) and available staffing level (gross staffing level or net staffing level, depending on the model evaluated); and the predicted aggregated daily service level for a given day of the week. Service level is used instead of misclassification rate of individual calls because individual caller characteristics are not used to build the predictive models.

4.1 Predicting aggregate daily service level

The predicting performance of the logistic regression model improves as more detailed staffing information is incrementally provided (see Figure 1); the KNN approach yields performance similar to that of Model 2, possibly due to the limitations in the trial values of its hyper-parameters in a simple cross-validation setting. There is still some gap between the best performance among the four models tested and that of the benchmark.

4.2 Predicting service level of individual 15-minute intervals

Not only does Model 3 exceed Model 2 in predicting aggregate daily service level, incorporating the break-adjusted staffing level also leads to more accurate prediction of service level in individual 15-minute intervals. Both models' performance in predicting service levels of 15-minute intervals on Mondays in January are shown as an example in Figure 2. Performance in predicting other days and months exhibit similar patterns.

4.3 Extrapolation performance

Despite the less competitive performance of the KNN model, which has improvement potential given more careful setup and training[2], one obvious disadvantage of memory-based learning in this setting is the dependency on the support of historical data in the target input range. Specifically, the applicability of the IBL approach diminishes as the new input deviates from the range of the training input, thus limiting its usefulness in answering 'what-if' questions such as 'what will the service level be if 20 agents are added to the workforce?'. Logistic regression, on the other hand, admits new input even if it is out of the training range. However, a comparison of the predictions by Model 3 and the validated DES model for staffing levels outside the training range on both high and low ends shows that disagreement between the two sources of predictions widens as the new input deviates further from the training range.

[2] See section 'Alternative modeling'

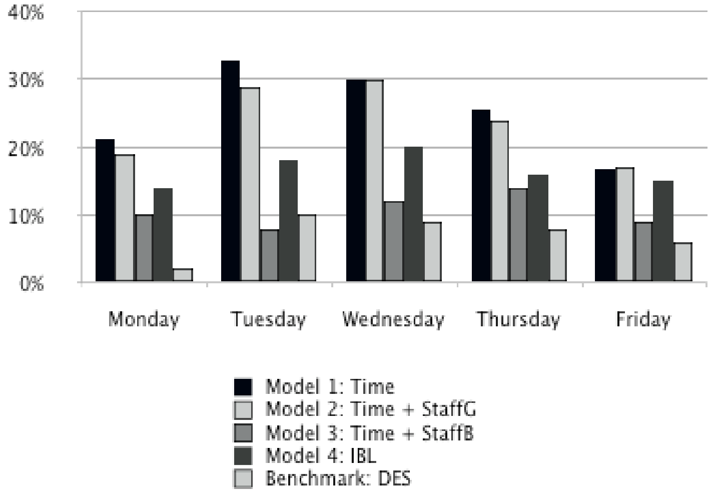

Figure 1: Mean absolute percentage error of daily aggregate service level predictions on Mondays in January

A clear drop in MAPE across all days of the week is visible as detailed staffing information (StaffB) is added to Model 2. On the other hand, adding the gross staffing level (StaffG) does not correspond to an effect nearly as drastic. Comparing the best performance among the three logistic regression models to that of the benchmark, the difference in MAPE is around 3-8 percent. The KNN (marked as Model 4 : IBL) performance is close to that of Model 2.

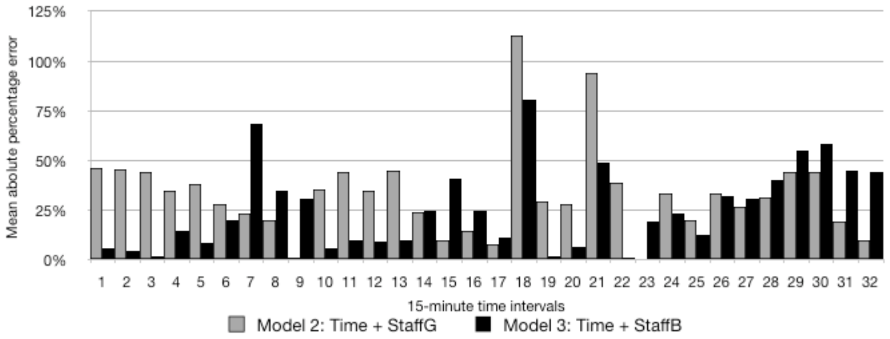

Figure 2: Mean absolute percentage error of service level predictions for each 15-minute interval on Mondays in January by Model 2 and Model 3

With the exception of the last six 15-minute time intervals, Model 3 yields a lower prediction error compared to Model 2 in most of the time intervals of the day. This means for this particular month and weekday the predictions by Model 3 tracks the trend of service level change throughout the day much closer compared to Model 2.

5 Discussion

5.1 Comparison with the simulation approach

The results presented in this paper shows that rule-based learning methods offer reasonable predictive power for the tele-queue studied. Although the performance is not as competitive as that offered by a validated simulation method, it does inspire the use of simulation-free learning methods in two ways: first, simulation method does not scale well to more complicated processes where heavy-weight general-purpose simulation package is needed and prediction not easily automated; Data-driven learning methods, on the other hand, is not challenged in the same way because it bypasses causal modeling of the call center process. Second, simulation methods have no readily available mechanism to incorporate new data points as they become available except off-line retraining of input parameters, while machine learning methods adapt easily to new data with online training algorithms.

5.2 Autocorrelation

One major challenge of modeling any first-in-first-out queues for services with limited resources is the natural dependency between consecutive data points. With the independence between the observations assumption violated, it is unnatural to directly apply many basic classes of classification methods relying on that assumption. The relatively high prediction error from the binary logistic regression model is likely due to this violation. In time series analysis, one usual practice is to take the first-order difference of the original data set and build prediction models on the difference instead. This approach is not applied in this study because of the difficulty in interpreting performance measure prediction after differencing.

5.3 Discretization of the response variable

In retrospect, arbitrarily setting the service level threshold as part of the data pre-processing might have limited the possibilities of applying a large class of machine learning techniques available; a model returning the full distribution of wait time as a numerical rather than categorical measure may be more flexible in terms of the variety of outputs it offers. However, the potential trade-off between added computational complexity and gain in prediction content and/or performance needs to be considered.

5.4 Alternative modeling

Many opportunities of improvement and fine-tuning exist for the logistic regression models and KNN method presented in this paper. For example, more intelligent selection of weights for variables' contribution to distance measure as well as number of neighbors k such as the ones presented in [4] could improve KNN's performance. In addition, prediction power of multiple approaches could be combined using ensemble methods such as building a boosting classifier. Furthermore, applying some alternative machine learning techniques with built-in features of dependency between observations (Markovian-based graphical model, for example) could further improve simulation-free prediction of wait time in tele-queues.

References

[1] W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. In *Production and Operations Management*, pages 88–102, 2006.

[2] J. Atlason, M.A. Epelman, and S.G. Henderson. Optimizing call center staffing using simulation and analytic center cutting plane methods. Technical report, Management Science, 2005.

[3] M.T.Cezik and P.L'Ecuyer. Staffing multiskill call centers via linear programming and simulation. *Management Science*, pages 310–323, 2008.

[4] H.Li, D.Groep, and L.Wolters. Efficient response time predictions by exploiting application and resource state similarities. In *Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing*, GRID '05, pages 234–241, Washington, DC, USA, 2005. IEEE Computer Society.

[5] A.M.Law and W.D.Kelton. *Simulation modeling and analysis*. McGraw-Hill series in industrial engineering and management science. McGraw-Hill, 1991.

[6] D.W. Aha, D.Kibler, and M.K. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991. 10.1007/BF00153759.

[7] D.R.Wilson and T.R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38:257–286, 2000. 10.1023/A:1007626913721.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.