Service Level Sensitivity

Overview:

This Service Level Sensitivity Analysis applies multiple machine learning models to determine the sensitivity and interconnectedness of various service level metrics, providing deep insights into call center operations. This project aims to highlight and understand how changes in your Service Level Agreement (SLA) can significantly affect your overall staffing requirements.

Problem:

Lowering SLAs often reduces the staffing requirements for meeting said SLAs, but in reducing headcount, there can be impacts that may or may not actually drive lower staffing requirements. For instance, lower service levels impact other call center metrics, such as increased average handle time (primarily wrap), increased occupancy, and increased shrinkage. All of those metrics are generally drivers of increased staffing.

This analysis attempts to answer the following questions:

1. What is the relationship between service level and other call center metrics?

2. How much will staffing requirements decrease at different SLA targets?

3. Does lowering SLAs generate actual cost savings from a headcount perspective?

Call Center Metric Relationships

Several different models were employed to explore the relationships between the different call center metrics and predict staffing requirements. Models developed were standard linear regression, multivariate regression, gradient boosting regression, and random forest regression, which is an ensemble method that combines multiple decision trees to make predictions. The Random Forest Regression had the greatest predictive capabilities with the highest level of accuracy and was used for the final models.

The models were initially trained on just the sales group to test the predictive capabilities of the model, but the correlations in both the smaller training set and overall model remained fairly consistent. Here is the correlation matrix for the combined dataset.
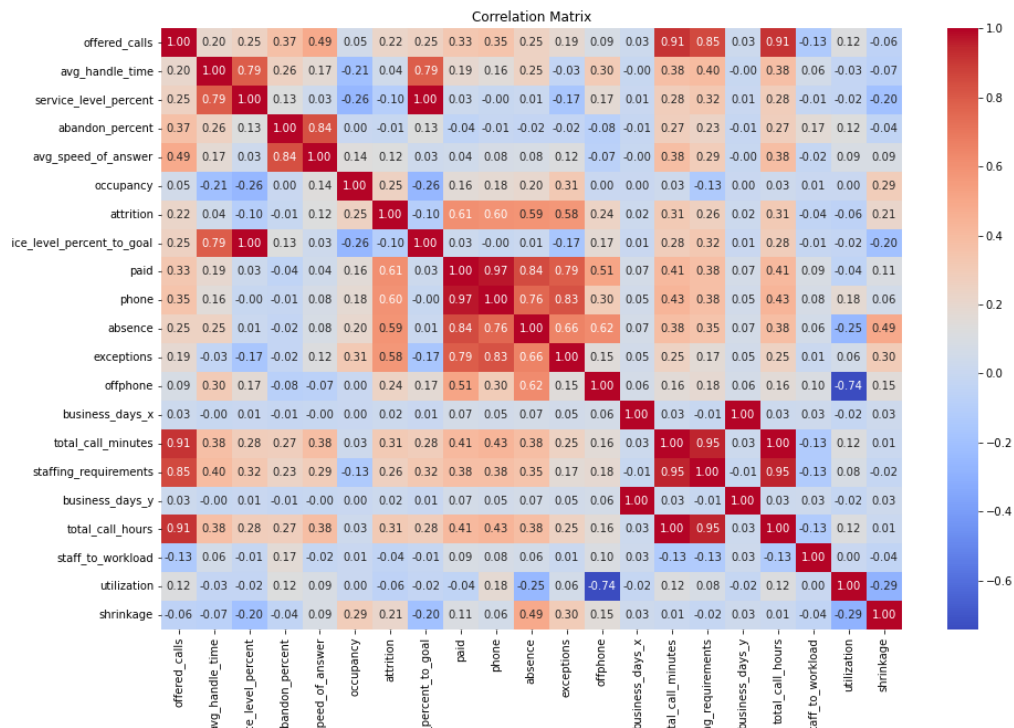
Correlation coefficients range from -1.00 to +1.00, where:

- -1.00 indicates a perfect negative correlation (as one variable increases, the other decreases by a proportional amount).

- 0. 00indicates no correlation (variables are independent of each other).

- +1.00 indicates a perfect positive correlation (as one variable increases, the other increases by a proportional amount).

Interesting findings in the correlations matrix were that both abandonment percentage and average speed of answer (ASA) had moderate negative correlation with staffing requirements, suggesting that as staffing goes up abandonment and ASA go down. This is definitely the case with service level, but it's interesting the model picked up a separate relationship with staffing requirements as well. Additionally, occupancy is typically more of an agent metric than a queue metric; therefore, the model did not pick up on strong correlation between occupancy and staffing requirements. For CCO, occupancy is calculated as the total handle time / total handle time + idle, indicating how busy agents are when logged into the queue. Differing call priorities and multiskilling likely further complicated this relationship within the models. As expected, offered calls, average handle time, and the derivative workload calculations (calls * handle time) had strong positive correlations with staffing requirements.

Erlang C Calculations

After establishing relationships, the dataset was run through Erlang C calculations at different SLA variations. The model then compared staffing results on the same data at the following SLAs:

- 65% in 30 seconds (current baseline)

- 60% in 30 seconds

- 55% in 30 seconds

- 50% in 30 seconds

- 60% in 45 seconds
- 60% in 60 seconds

The dataset for the model included the last 15 months of call data for every 30-minute interval. Staffing requirements were run for each interval, rolled up to a daily aggregation, and then averaged across the dataset as a whole to get a representative sample of impacts. The following heatmap shows the results of the Erlang differences for each team at different SLAs:
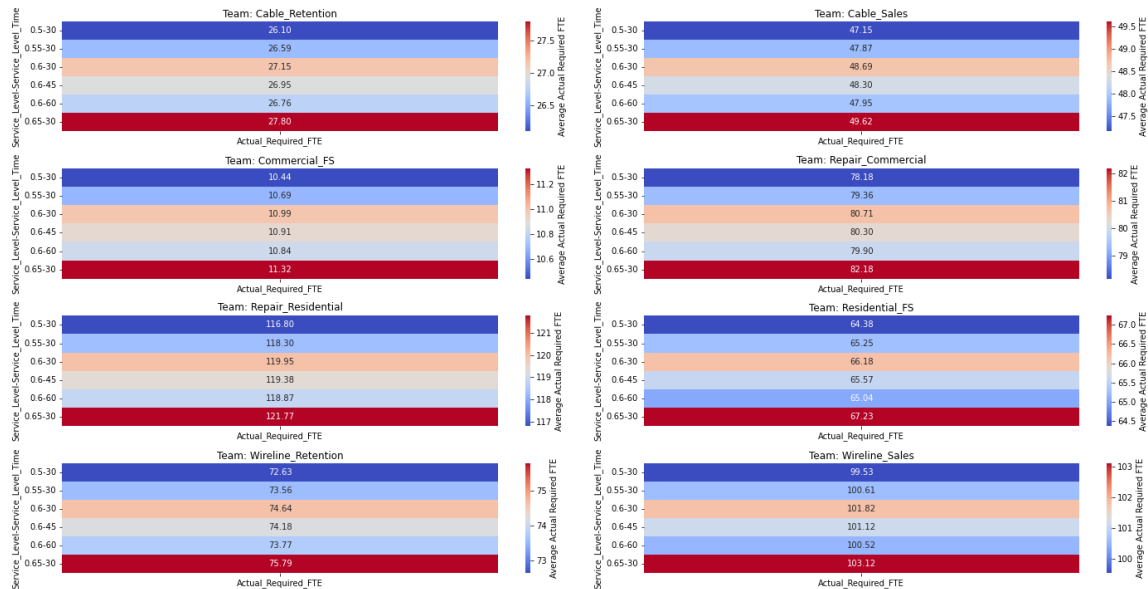

The same information is here represented in bar graph form:
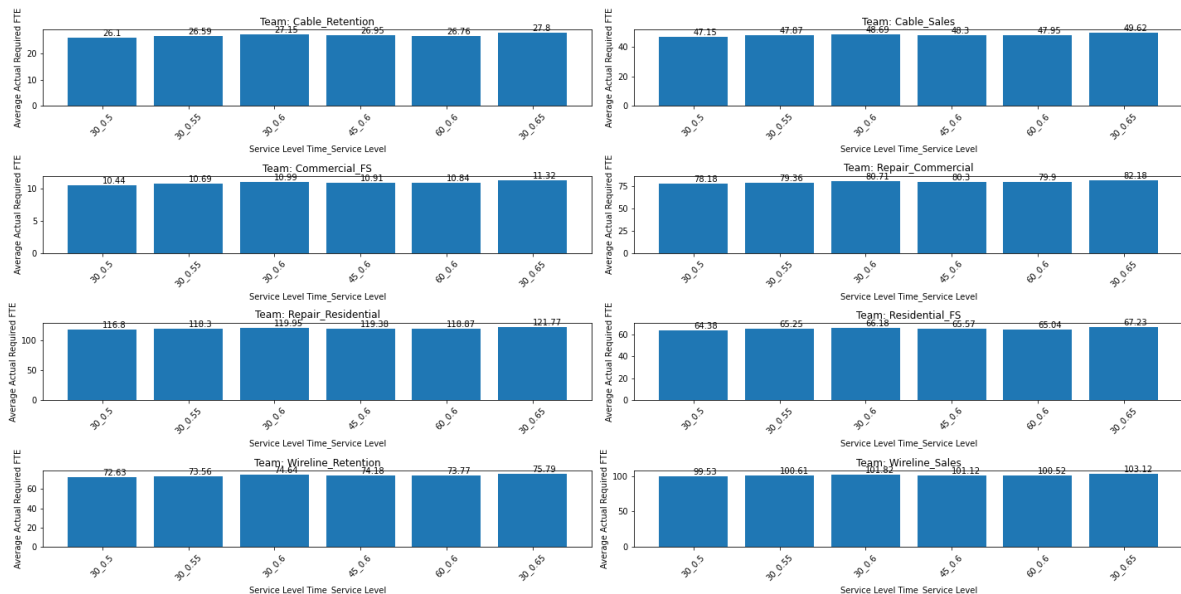
Regression Factors on Erlang Outputs

The Erlang models represent how staffing requirements change at different SLAs if all things remain constant. The complex relationship between variables in the regressions and machine learning models indicates those variables will not remain constant, however. The probability a call waits is an important variable to determine how some of the other relationships may change as we alter SLAs. Here is a heatmap of the waiting probability at the different SLA targets:
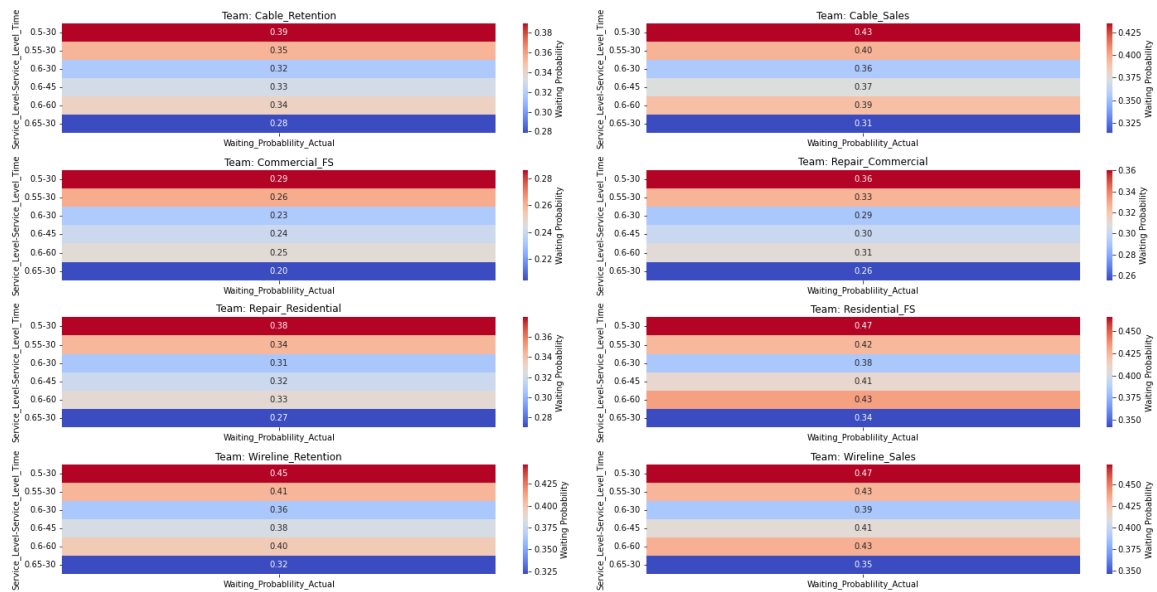
As the probability a call waits increases, so too does the abandonment percentage increase. Also, as we consider the probability a calls waits, we must also consider that as the same probability an agent has back-to-back calls, which typically also increases average handle time (primarily wrap) and occupancy. In other words, as the probability a call waits increases, it is also proportionately more likely that staffing requirements increase because the workload actually increases as a result of higher occupancy. Here are separate regressions for waiting probability vs service level and occupancy for reference:
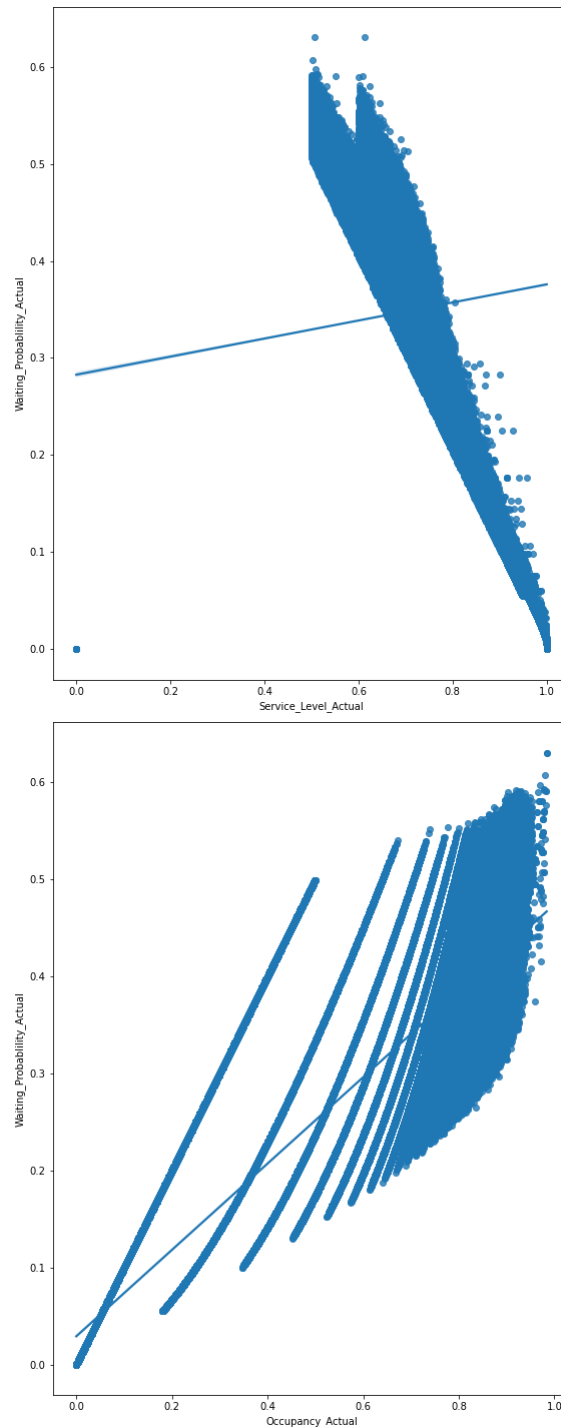
In terms of basic workload calculations, for every 100 calls and 30 seconds of handle time, the staffing requirements increase by 0.83. This does not factor in additional variables such as shrinkage or occupancy targets, which would further increase that number. Thus, for larger teams who take thousands of calls per day, increases to handle time quickly increase staffing requirements. Those increases likely exceed any FTE savings generated by lowering SLAs.

Conclusions & Recommendations

Limitations and Assumptions

The relationships between call center metrics is highly complex and difficult to model. Thus, it is extremely difficult to quantitively state that a 5% change to SLA equals an n% change to handle time or other correlative metrics. The only directly quantifiable metrics from changing service level are staffing requirements, occupancy, and waiting probability. This analysis makes inferences from general call center assumptions, specifically about the relationship between occupancy and agent-

related metrics, such as handle time, shrinkage, and attrition. Shrinkage in this context is meant as both unplanned absence time and out of adherence time. More data would be needed to specifically model those assumptions for the call data used in the model.

15 months of interval call data is still a relatively small data sample and could be expanded to yield more comprehensive results. The length of the dataset was chosen as a balance between computational time for the model and having sufficient time-series data to establish trends. Additionally, no normalization was applied to smooth out noise in the data from outages, system issues, etc. The call data is in its raw form and could be further transformed to possibly yield better results.

Conclusion

Staffing requirements do decrease as SLAs are lowered; however, the probability a call waits increases. Waiting probability positively correlates with occupancy, so increasing the probability a call waits also increases occupancy. In the scope of this model, it was not possible to make correlations between occupancy and other metrics even though relationships likely exist. We can infer, however, occupancy increases generally correlate with increases in other agent-based metrics, such as handle time and shrinkage. Further, high occupancy is generally assumed to increase agent burnout and attrition. Given those assumptions, increasing the probability a call waits will likely also increase handle time, shrinkage, and attrition. Of those metrics, handle time and shrinkage directly increase staffing requirements.

In conclusion, lowering SLAs would reduce staffing requirements by a few headcount (on average) per team, but the commensurate increases to call wait times, handle times, and shrinkage would very likely offset any headcount reductions from lowering SLAs. Accordingly, there is not a real cost savings associated with lowering SLAs. Further, given the high impact of handle time on staffing requirements, it would indeed be possible to lower SLAs and increase staffing requirements because of worsened agent conditions. Thus, lowering SLA targets is not recommended as a strictly cost-saving measure.

Recommendations for Further Analysis

- Based on these findings, further analysis is recommended to determine what the "right" SLAs would be for each team and/or service. Given more time and data, customer behaviors could be correlated with agent and call metrics to determine the "sweet spot" for SLAs, and right-size the metric per team and service.

- If LLMs could be applied to call recordings, customer sentiment could also be correlated with wait times with strong predictive capabilities for right-sizing the SLA metric.

- Neural network machine learning models could be applied to the dataset to help better model the highly complex relationship between staffing, agent, and call metrics. A more advanced model would better identify the interconnected correlations present in the data.