The three stages of workforce optimisation: Moving beyond the industry standard

Received (in revised form): 21th March, 2022

Dakota Crisp
Data Scientist, RXA, USA

Dakota Crisp has a PhD from the University of Michigan. His hypothesis-driven approach to integrating concepts from neural engineering into the data science space provides a distinctive take on consumer behaviours.

RXA, 330 E. Liberty St. Ann Arbor, MI 48104, USA
Tel: +1 888-294-1512; Email: dakota.crisp@rxa.io

Jess Brown
Senior Data Scientist, RXA, USA

Jess Brown has a BS from the University of Michigan. Her background in computer science introduces a focus on bringing models to life through unique implementation strategies.

RXA, 330 E. Liberty St. Ann Arbor, MI 48104, USA
Tel: +1 888-294-1512; Email: jess.brown@rxa.io

Jack Claucherty
Data Scientist, RXA, USA

Jack Claucherty has a BS and MSE from the University of Michigan. He brings an industrial engineer's perspective to data science with a focus on complex systems and states.

RXA, 330 E. Liberty St. Ann Arbor, MI 48104, USA
Tel: +1 888-294-1512; Email: jack.claucherty@rxa.io

Davis Busteed
Senior Data Scientist, RXA, USA

Davis Busteed's background in information systems provides a creative approach to problem solving with cutting edge data techniques.

RXA, 330 E. Liberty St. Ann Arbor, MI 48104, USA
Tel: +1 888-294-1512; Email:davis.busteed@rxa.io

Anna Schultz
Marketing Coordinator,RXA, USA

Anna Schultz is a BBA graduate from the Ross School of Business. Anna has spent her career in marketing and analytics at RXA. She takes pride in building marketing strategies and materials that blend unique design with useful and direct communication. In analytics, she has worked with companies from a wide variety of industries to collect, transform and visualise their data to deliver actionable insights throughout the company.

RXA, 330 E. Liberty St. Ann Arbor, MI 48104, USA
Tel: +1 888-294-1512; Email: anna.schultz@rxa.io

Jonathan Prantner
Chief Analytics Officer and co-founder, RXA, USA

Jonathan Prantner's approach to applied mathematics has pushed analytics to the limits for over two decades. Jonathan's career has spanned educational research, automotive, CPG, travel and healthcare. At RXA he leads efforts surrounding applied artificial intelligence and machine learning as well as integrating advanced analytics with data visualisation platforms. Jonathan is a celebrated thought-leader and recipient of multiple data science patents.

RXA, 330 E. Liberty St. Ann Arbor, MI 48104, USA
Tel: +1 888-294-1512; Email: jp@rxa.io

Abstract Erlang-C has long been the industry standard for call centre staffing. As call centres evolve and staffing concerns move to other industries, the standard methods don't always work as expected. While traditional scheduling methods focused on translating historic demand into staffing needs, the estimation of demand and the logistics of scheduling employees have not always been equally considered. Through the analysis of multiple use cases from distinct industries, this approach evaluates all three stages of workforce optimisation and explores the gap between theory and reality.

KEYWORDS: optimisation, staffing, call centres, ensemble learning, Erlang-C, automotive, forecasting.

THE SITUATION

The influence of COVID-19 has highlighted the importance of workforce optimisation. Health precautions, staffing shortages and supply-chain disruptions have changed the way both businesses and consumers imagine the work week as well as what defines good and timely service.[1] Internally, workforce optimisation is designed to improve worker efficiency, productivity and performance, allowing companies to save resources and do their best with the limited staff they may have. Furthermore, optimisation can help minimise understaffing, which limits the stress workers experience and helps improve employee retention. Externally, in consumer-facing roles regardless of industry, workforce optimisation concerns itself with two main considerations: how many customers can be served and how quickly can and should they be served in a fiscally responsible manner. In times of stress and uncertainty, it is imperative that any company implement a strategy to address both these internal and external issues. However, these complex times are incompatible with a one-size-fits-all solution.

Historically, the translation of demand into staffing needs has taken the forefront, with the gold standard being Erlang-C. Developed in the call centre industry, Erlang-C addressed this issue by algorithmically determining the optimal staffing number for a given wait time depending on the expected incoming call volume.[2] While this provides a straightforward and quantitative approach to translating demand to required staff, this strategy cannot keep up with new industries and increasingly complex business environments.[3]

For example, in the automotive service space, both appointments and walk-in service can fall into one of two categories: the vehicle is dropped off or the consumer is waiting at the centre. Businesses must recognise that consumers in these categories have drastically different expectations, and any measure of demand must encompass

both the immediate demand of waiting customers and the stock demand of vehicle drop offs.[4] Therefore, workforce optimisation in this industry relies heavily on first understanding consumer demand then decrypting the appropriate staffing levels.

Keeping up with shifting landscapes is paramount for the call centre industry, as the success of call centre businesses hinges entirely on how well they can address the issue of workforce optimisation.[3] Their primary business model consists of correctly staffing enough workers to control customer wait times. Too few staff create large wait times, translating to unhappy customers and damaging the brand. Too many staff result in happy customers, but at inflated costs that reduce already tight margins. For years Erlang-C provided a straightforward and quantitative approach to translating demand to required staff, but the key assumptions limit efficacy in changing, real-world environments. Furthermore, as the industry has grown and began servicing a wider variety of clients, staffing issues arise outside the scope of the simple algorithm. Therefore, workforce optimisation in this industry goes beyond the standard translation of demand to staffing.

This work shows that a more holistic approach to workforce optimisation requires addressing three specific stages: estimating demand, translating demand into staff, and staff selection. This study sets out to show how machine learning can augment traditional staffing approaches concerning these three stages in two vastly different industries: automotive service centres and call centres.

AUTOMOTIVE USE CASE

The first case study focuses on the automotive service retailer, *Belle Tire*. In the past, *Belle Tire* exclusively used managerial discretion to estimate future demand and staffing requirements. These decisions centred around upcoming holidays and weather expectations and were difficult to standardise between scheduling managers. Senior management set out to optimise scheduling with three guiding principles: (1) increasing employee satisfaction, (2) encouraging more accurate tracking of future jobs, and (3) increasing profitability. Their ideal solution was to implement a transparent, monitorable system that would help assist with scheduling to increase efficiency.[4]

Belle Tire's needs were addressed in three separate ways: (1) leveraging their own extensive internal data to forecast demand, (2) supplementing their data with external control data, and (3) implementing an ensemble modelling approach for improved robusticity.

LEVERAGING CLIENT DATA

The approach leveraged highly detailed tire technician services data spanning several years. This data provided information such as the historic number of customers, tires and wait times, which could be considered as the input features for typical demand forecasting. This dataset differed from the norm because it also included technician 'on the clock' hours (specifying when technicians arrived and left work) as well as technician 'working' hours (the labour-specific punch times recorded per task by the technician). This differentiation reflects the inherent nature of how work is completed in the automotive service industry. In an automotive service centre, any time spent on a specific repair order (RO) is tracked similarly to the way billable hours are tracked within a law firm. This represents time tied specifically to revenue but does not account for the infrastructure-based work that needs to occur to enable the billable hours.[5] This would capture tasks like loading stock, cleaning workstations and other daily maintenance. Therefore, the blend of working time to total time should never reach one and this

blend will vary based on the intricacies of each individual store.

By relating these time-based features to the typical demand-based features, the data provided insight into technician efficiency and general work capacity for a given context. Therefore, instead of trying to predict whether a customer would come in for service (typical demand), the strategy was to forecast how busy the average tire technician would be for an allotted timeframe (data resolution allowed investigation at 30-minute intervals of each day). This data included multiple technicians and spanned a long enough history and turnover to reflect the 'typical technician'.

EXTERNAL CONTROL DATA

While client-specific data provides exceptional insight into the low-level business mechanisms, demand in the automotive services industry also benefits from considering higher-level natural and socioeconomic contexts. External factors constantly affect companies' sales, operations, marketing, staffing, and planning decisions. No company operates in a vacuum; all are subject to environmental changes that are outside of their control. Companies have turned to data analytics, statistical modelling, prescriptive analytics, and artificial intelligence to inform business decisions, but many models are built only from internal data (analysing how these internal components trend over time or by season). When models do not account for the complexities of external events, there is potential for false positives or negatives, resulting in a lack of comprehension concerning the causes of these business outcomes.[6] Therefore, control data is an incredibly important—though often overlooked—component of applied data science.

Control data, often referred to as external data or third-party data, is data detailing the external factors that impact the outcome of variables in a model. Using control data in an applied model allows data scientists to understand which factors outside of the business have affected outcomes in the past, map out the correlation among those factors and inform how the business may perform in the future due to these same factors. Of course, correlation does not always prove causation, so it is important for humans (rather than solely AI) to analyse information for a deeper understanding of how and why these factors have a correlation in the first place.

Control data should be used during prediction when external factors can affect the outcome of a real-world scenario. This is true for all applied predictive analysis in business intelligence. All businesses experience environmental factors that can affect demand, inventory availability, costs, workforce availability and more. Even something as high-level as overall economic trends will affect a business and needs to be accounted for in a predictive model.

For example, in the case of a brick-and-mortar business, understanding the population growth in the surrounding area will be paramount to predicting demand; as the population of the target audience grows, sales may also grow. Weather is another important factor for a business operating out of a physical location, as patrons may be more likely to visit stores on warm days compared to days with blizzard conditions. Adding these factors to models quantifies their individual impact on the business and will enable planning for future changes.[7]

For an online business, overall economic growth or decline is more suitable for predicting sales than regional changes. If an input to a business is steel, understanding how government regulations have affected steel imports in the past will allow that business to plan for regulation changes in the future. These are a few simple examples of the way external factors affect businesses every day. Testing out various correlations

will allow companies to understand how each factor interacts to create an overall business outcome.

There are many online resources to source control data. Many government-related sets can be accessed for free or a small fee, such as economic data from the FRED database or weather data from the NOAA database. Other types of data, like consumer marketing data, can be found through paid services like Experian. Sourcing datasets in this way is useful for businesses when they already know what data they are looking for, are comfortable building connections to these data sources via their own resources and have the knowledge and training to transform these datasets to the correct dimensions and granularity required.[8]

Another option is through services like *Ready Signal*, a platform developed to source and deliver various control datasets. Users can start from scratch or receive dataset recommendations based on a type of analysis and industry. Then, datasets can be automatically transformed based on geographic area, timeframe or time granularity, through data science treatments like leads, lags and decays, and exported to a variety of destinations in real time. *Ready Signal* also provides the auto discovery tool, a patent pending technology that allows users to upload a target variable they wish to explain. The tool will automatically identify correlated control features to test and train models, saving the time and effort of having to know which external factors may be interacting with the business.

In this case, external control data was acquired via *Ready Signal* and leveraged to better characterise customer behaviours and influences on purchasing decisions, such as location-specific weather data and consumer sentiment.

Ensemble modelling

> 'Change is the only constant in life.'
> Heraclitus

An issue that repeatedly arises in demand forecasting is data non-stationarity. This was especially true for this client for two reasons.

First, their business consists of many store locations of varying age. The discrepancy between the oldest and youngest stores is 30 years, with the youngest being under five years old. From a business perspective, the amount of time a store location has spent establishing their brand and customer base can significantly influence performance. As time progresses, younger stores are most likely to experience greater change as they continue to adapt and establish themselves. In other words, this adaptation is store specific and nonlinear.

The second reason relates to the continued challenges resulting from the 2020 COVID-19 pandemic. This client, as many others, experienced a sharp decrease in demand for part of 2020 followed by a sharp increase in 2021. To address these concerns and buffer the client from future instability, analysts developed a three-stage ensemble model that was designed to detect and address changes in demand.

The first stage utilised XGBoost, a highly accurate general model that had acted as the initial backbone of the implementation.[9] Similar to the standard gradient boosting approach, XGBoost trains a model sequentially, where each estimator builds off the success of the previous iteration. In the case of gradient boosting, the first 'round' of training begins with an initial prediction, which is used to calculate the residual for each observation. The subsequent rounds consist of fitting a new decision tree on the residuals of the previous iteration and calculating a new set of residuals. XGBoost provides several improvements to the standard gradient-boosting approach. Techniques such as the weighted quantile sketch, sparsity-aware split finding, and cache-aware access make XGBoost suitable for large, complicated datasets.[10]

The second stage utilised a Poisson regression, which further allowed for an

improved model fit and reduced prediction variability. The Poisson regression was selected to address the count-based structure of the data. The predictive target of these models met the requirements of the Poisson regression in that the observations were discrete, non-negative, and were measured within a predefined time interval. Although an exponential or polynomial model could have been used to approximate the fit of the observed distribution, these techniques assume a constant variation at each value of the target variable. By using Poisson regression, the underlying Poisson distribution is leveraged to provide a more realistic model of the data.[11]

The third and final stage was a simple linear regression model based on the two-week demand difference. Employee staffing is often subject to seasonal trends, and this use-case was no exception. Without adjusting for the seasonal trends in the data, the model may 'drift' and identify transient trends that would negatively impact future predictions. The differenced modelling approach reduced bias and grounded predictions in recent actual performance to capture steady states and transient trends.

These models engaged with the data from different perspectives, and their combined output created a more robust, holistic outcome.

Modelling outcome

The model was successfully implemented across all *Belle Tire* stores. Overall, there was found to be a 34% increase in accuracy (where an accurate prediction of demand was defined as +/- one tire technician from what was required for any given 30-minute window), bringing the new average accuracy up to 80%.

Of course, any algorithmic recommendation is only as useful as the willingness of scheduling managers to adopt it.[12] Therefore, dashboards were created to monitor both the managers' individual store performance over time (in weekly increments) compared to the algorithmic recommendation. This not only allows for greater transparency and adoption but helps companies to address issues more quickly.

Translating demand to staffing levels

Estimating future demand is just the first step in the comprehensive approach to staffing optimisation. The next step is translating this demand into the number of necessary workers to perform the job. Every staffing problem is a unique entity and there are many factors that will impact optimisation, from the distribution of demand to the way the company processes customers. Additionally, companies will have different sensitivities to key KPIs of employee utilisation levels and average wait times for customers. For example, some businesses may seek to increase employee utilisation to lower costs, while others may choose to increase staffing to maintain average wait times below a desired threshold. Therefore, understanding and aligning to the business' needs is paramount to the process.

Continuing the use case, a history of overstaffing necessitated a focus on employee utilisation levels. In this case, the business has a pool of manager employees that are available at any given time to step in and fill short-term understaffing issues. Recent trends have reduced the staffing pool and left managers with less than their desired number of employees. Therefore, it was ideal to aim closer to 100% utilisation, knowing that occasional over-utilisation could be handled without long-term repercussions.

The x-axis defines the difference or 'gap' between the number of staff members' activity working and the number of staff members scheduled to work for a given 30-minute increment. The histogram is partitioned by 'forecast' (the recommended staffing number of the model) and 'manager' (the decision of the scheduling manager without assistance).

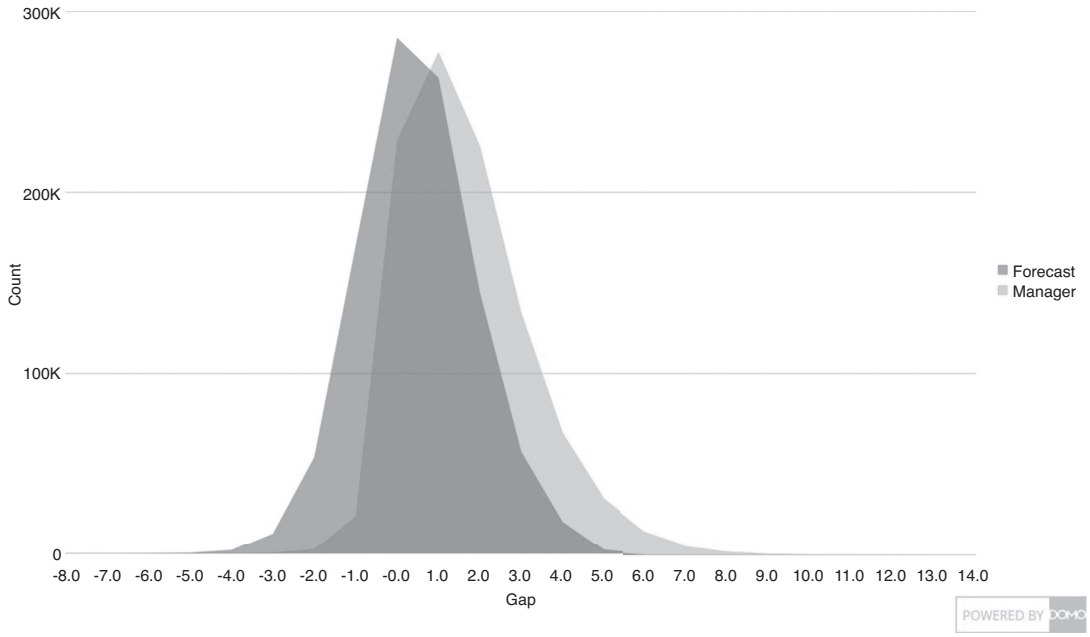

Figure 1: Staffing gap by scheduling strategy
Source: Author.

The frequency of overstaffing declined ($p < 0.001$, t-test, see Figure 1) with ~80% of instances falling between +/- 1 worker per 30-minute increment.

CALL CENTRE USE CASE

Translation via wait times

The second case concerns a credit union, a member of the call centre industry, whose business model depends on reasonable wait times. This issue is not new in the call centre industry, where the standard solution is to employ the Erlang-C formula.[13] The basic idea is to translate demand into workload per unit time. Using specified wait times and efficiency levels, the formula uses this workload to determine the number of staff members required to handle it.

Erlang-C is described in the formula below. The output P_W represents the probability a call waits which is a function of demand (A) and number of employees staffed (N).

$$\text{Formula 1}\quad P_w = \frac{\frac{A^N}{N!}\frac{N}{N\text{-}A}}{\left(\sum_{i=0}^{N-1}\frac{A^i}{i!}\right)+\frac{A^N}{N!}\frac{N}{N\text{-}A}}$$

Using this calculated probability that a call waits (Pw), average speed of answer or average wait time can be calculated with the formula below where traffic intensity is equivalent to demand (A) in formula 1.

Formula 2

$$\text{Average Speed of Answer (ASA)} = \frac{P_W \times \text{Average Handling Time}}{\left(\text{No. of Agents-Traffic Intensity}\right)}$$

While the power and simplicity of this formula cannot be understated, it is important to understand the limitations and assumptions. The Erlang-C method assumes an exponential relationship between utilisation and average wait time (ie, the fewer available operators, the higher the average wait time). While this approach is robust for simple

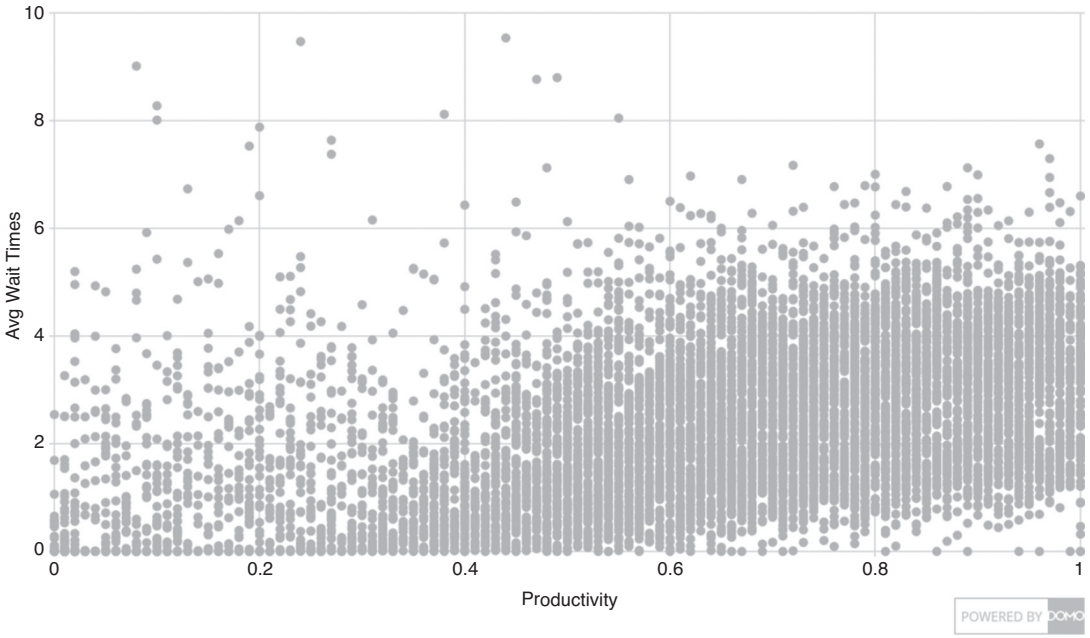

Figure 2: Scatter plot of productivity (working minutes/total minutes) and average wait times in half hour increments from July 2019 to August 2021
Source: Author.

queueing problems, this assumption has a narrow field of applicable business cases. As queues get more complex with the introduction of call-backs, multiple languages, etc., the assumption begins to break down, which was the case in this client's data: there was no clear relationship between utilisations and wait times (See Figure 2). When attempting to apply Erlang-C to predict wait times, the mean absolute error of the predictions was 3.09 minutes.

To address this dilemma, a deeper investigation of the historical data was needed. First when graphing average wait time over time, it was noticed that starting around March 2020 the average wait times become more sporadic and, on average, were 1.5 minutes longer than before (See Figure 3). This change was attributed to COVID-19 and the data was separated into pre-COVID-19 and post-COVID-19 segments.

Next, data was grouped by demand level to see if there was a more concrete relationship between average wait time and productivity at different demand levels (see Figure 4). In theory, 60% utilisation should result in the same wait time agnostic of the demand, however, it was theorised that due to the high fluctuation in demand the same utilisation at either end of the demand spectrum could be significantly different. After removing pre-COVID-19 data and grouping the remainder based on demand levels, the relationship between utilisation and average wait time started to become more understandable.

Each colour indicates how many hours of work were received in the half hour increments.

While this data is still sporadic, it is much more stable than in Figure 2 which enables more confident application of different models at each demand level. From here, Erlang-C or another model could be applied to explain these relationships at every level of demand. For this use case, linear regression at each demand level best explained the relationship, lowering the mean absolute error of the predictions to

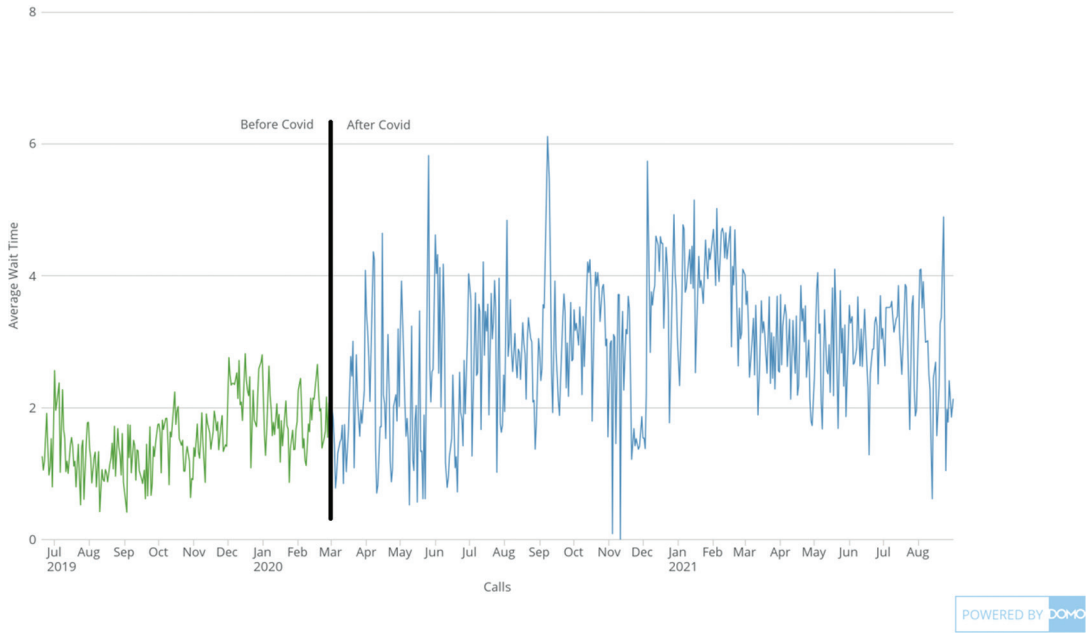

Figure 3: Average wait times by day from July 2019 to October 2020 showing the increased variance starting in March 2020
Source: Author.

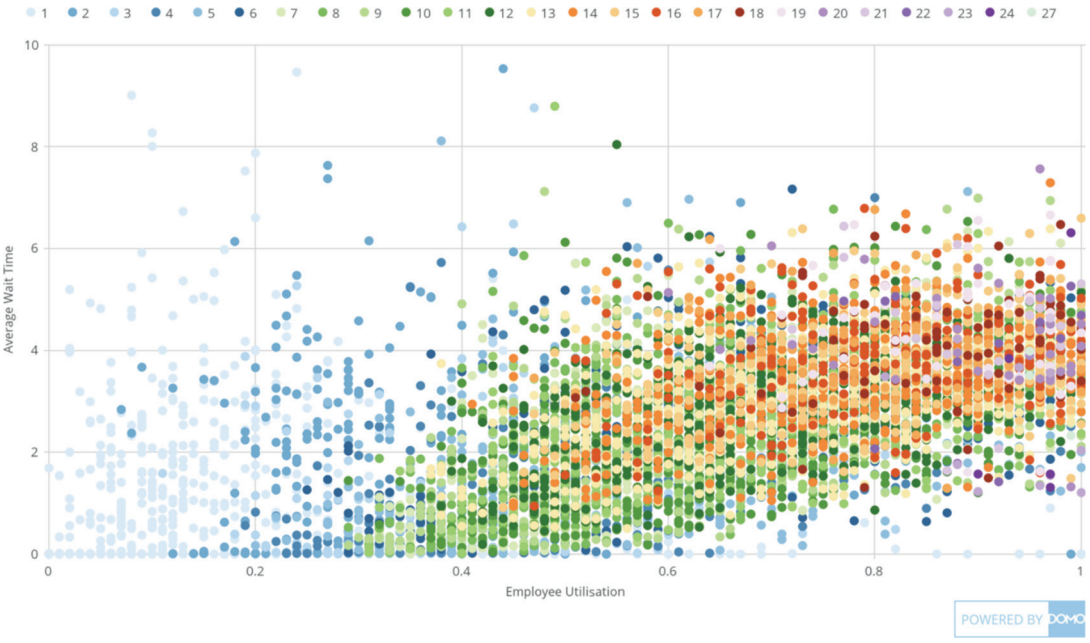

Figure 4: Scatter plot of productivity (working minutes/total minutes) and average wait times in half hour increments from March 2020 to August 2021 split by demand levels
Source: Author.

0.83 minutes. After finding this relationship, the demand estimated by the XGBoost model above can be used as an input to staff at a productivity level and control for desired wait times. The use of XGBoost or similar algorithms to estimate demand allows for a more dynamic way to view data and craft a staffing equation to staff at the optimal level.

Overall, when estimating the ratio of how many employees to staff, every use case is different. From determining business-specific KPIs to evaluating the system and historic data, it takes a hand-crafted approach to truly optimise staffing. Combining advanced algorithms to predict demand with this hand-crafted approach allows for more dynamic scheduling to improve workforce morale, increase company productivity, and improve customer experience.

CONCLUSION: WHO TO STAFF

The first cornerstone to staffing optimisation is understanding the demand and need for employees. Without understanding demand, employers are faced with times where their customers' needs are not met and when staff are not fully utilised. This leads to reduced customer and employee satisfaction. Thankfully, advances in machine learning allow organisations to better understand, anticipate, and plan for future demand, leading to a better understanding of staffing needs.

Schedules can be made weeks in advance and optimised with new data. In the automotive service use case, this took the form of ensemble models to estimate employee utilisation through the ratio of billable hours to on clock hours. In the call centre use case, this took the form of demand segmentation through both ensemble learning and linear modelling before employing the Erlang-C calculation. In both cases, machine learning was implemented to provide the companies with confidence concerning their staffing needs.

Once staffing needs are understood, staff selection is one of the more deceptively challenging aspects of staffing optimisation. This complexity stems from the need to fill staffing gaps across various levels of requirements and limitations. For example, in addition to the overall staff-hours needed, it is important to consider how the company is staffed, whether it is by 'shift work' or hourly scheduling, hours of operation (client-facing as well as prep and closing hours), and how weekends and holidays are handled. With this basic structure in place, options are further limited to employee availability, how well the employee's skills and proficiencies meet this gap (be it training, language proficiency, etc.), and industry specified hourly minimums and maximums. Depending on the number of employees and size of the business, making an effective choice of who to staff can be a mammoth undertaking.

So how do businesses currently optimise staff selection? The short answer: most don't. Due to this complexity, most mid-level companies intentionally overstaff to minimise complications and potential liabilities. This short-cut strategy, while effective in the short term, wastes an incredible amount of resources over the long term. Therefore, the use of staffing optimisation is likely to pay for itself over time.

The practical approach used to translate staffing needs to employee schedules in these two uses cases relied upon two tenets. The first involves sitting down with the clients and establishing the staffing logistics of their specific businesses. This information was used to develop a set of modular conditional rules that can be applied and adapted as needed. The second tenet utilises these rules in an optimisation algorithm that can handle a wide variety of rules and systems, providing solutions that would be too difficult to compute by instinct alone.

By utilising augmented intelligence, a blend of machine learning, and specific

business constraints, a businesses' staffing needs can be better assessed and a more holistic approach to staffing optimisation can be provided. This approach segments these needs into specific, actionable stages: (1) estimating demand, (2) translating demand into staffing needs, and (3) staff selection. By making marginal improvements to all three stages, the end result is magnified, providing companies with long-term value.

References

1. Ivanov, D., (2021), 'Supply Chain Viability and the COVID-19 pandemic: a conceptual and formal generalisation of four major adaptation strategies', available at https://www.tandfonline.com/doi/full/10.1080/00207543.2021.1890852, last accessed 21st February, 2022
2. Chromy, E., Misuth, T., Kavacky, M., (2011), 'Erlang C Formula and its Use in the Call Centers', available at http://advances.utc.sk/index.php/aeee/article/view/34, last accessed 21st February, 2022
3. Robbins, T. R., Medeiros, D. J., Harrison, T. P., (2010), 'Does the Erlang C model fit in real call centers?', available at https://www.researchgate.net/publication/221526180_Does_the_Erlang_C_model_fit_in_real_call_centers, last accessed 21st February, 2022
4. Save Millions on Labor With Domo, Jason Harper, Founder and CEO, RXA Don Barnes, III, President and Chief Tire Guy, *Belle Tire*, Ben Schein, VP, Data Curiosity, Domo, available at https://www.domo.com/domopalooza/resources/save-millions-on-labor, last accessed 22 April, 2022
5. Maister, D. H., (1996), 'It's about time', available at https://www.proquest.com/openview/fa0f345479225f33982a46cb4c55755c, last accessed 21st February, 2022
6. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2256 last accessed 21 Feb 2022
7. Barnett, W., (1988), 'Four Steps to Forecast Total Market Demand', available at https://hbr.org/1988/07/four-steps-to-forecast-total-market-demand, last accessed 21st February, 2022
8. *Ibid*
9. Chen, T., Guestrin, C., (2016), 'XGBoost: A Scalable Tree Boosting System', available at https://doi.org/10.1145/2939672.2939785, last accessed 21st February, 2022
10. *Ibid*
11. Haipeng, S., Huang, J. Z., (2008), 'Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management', available at https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-2/Forecasting-time-series-of-inhomogeneous-Poisson-processes-with-application-to/10.1214/08-AOAS164.full, last accessed 21st February, 2022
12. Dietvorst, B. J., Simmons, J. P., Massey, C., (2014), 'Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err', available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2466040, last accessed 21st February, 2022
13. Reynolds, P., (2021), 'Essentials of Contact Center Staffing: Calculations and Staffing Models for Workforce Planning', available at https://swpp.org/on-target-summer-2021/essentials-of-contact-center-staffing-calculations-and-staffing-models-for-workforce-planning/, last accessed 21st February, 2022