Forecasting Childcare Staffing Needs at the Child Saving Institute: A Data-Driven Approach

ECON 8310 – Business Forecasting – Semester Project

Diane Mack, Taylor Hogan, Patrick Ndungutse

May 12, 2025

Introduction

The Child Saving Institute (CSI), a cornerstone of childcare services in Omaha, Nebraska since 1892, offers diverse programs and services aimed at strengthening families and communities in the Omaha area. CSI's mission includes leading with empathy, creating safe spaces, and meeting people where they are. For 133 years, CSI has provided early childhood development, which caters to children from six weeks to six years old across multiple centers. Ensuring optimal staff-to-child ratios is not only a regulatory mandate but also essential for delivering quality care. However, fluctuating attendance patterns pose challenges in staffing, which has the potential to result in understaffing or overstaffing of individual rooms. To address this challenge, this project harnesses historical attendance data to forecast staffing needs in 30-minute intervals, which provides a model for more efficient resource allocation and enhanced service delivery.

Project Goals

Accurately forecasting staffing needs will allow CSI to remain agile and responsive, and plan resources in the most appropriate and cost-efficient way possible while meeting regulatory requirements and excellence of care standards. The primary goals of this initiative are:

1. **Typical Week Forecast**: This identifies regular patterns in attendance that reflect common attendance trends, then generates a model that predicts average staffing requirements for a standard week, segmented into 30-minute intervals.
2. **Next Week Forecast**: This is a tactical tool that provides a precise forecast for the immediate week following the last available data point, detailing staffing needs in 30-minute increments. This will provide critical guidance for assigning staff when and where needed.

By achieving these goals, CSI can proactively manage staffing, ensuring compliance with mandated ratios, optimizing operational efficiency, and achieving their high standards for child care.

Data Overview

The dataset provided by CSI consists of three years of anonymized "check-in" and "check-out" records from two CSI childcare centers, ECEC and Spellman. Each record in the dataset includes specific information such as time stamp, child status (in/out), room name, and associated tags, all of which required decoding and transformation to be usable data. Note that the data did not include child names or other explicitly identifying information, but the data did consist of multi-year real-world attendance information with underlying patterns. In terms of creation of usable data, while time stamps indicate specific times of child attendance, they don't directly indicate which 30-minute intervals were occupied. Therefore, one of the first key data engineering tasks was to align presence with intervals of time. In addition, the required staffing ratios for each age group of children was a separate data set that was incorporated into the model. This data mapping was critical for producing valid staffing forecasts.

Methodology

Data Preprocessing

Due to the raw nature of the data, significant preprocessing steps were required to clean the data and create dataframes that could be utilized for analysis. Our team utilized pandas, datetime, regular expression, and other tools to convert timestamps, create binominals for data selection, time data, segment each day into 30-minute intervals, only count the weekdays (no weekends or holidays), count the number of children in each room per designated 30-minute interval, and apply the designated age-specific ratio of teachers to children for the purposes of determining staffing needs per interval.

We created flags for potential anomalies gaps in check-in/check-out pairing, empty entries (NaN), and duplicate entries. These flags informed decisions about whether to impute missing values or discard corrupted or empty rows. We created features such as day-of-week and hour-of-day, which assisted in the identification of cyclical patterns that are central to childcare operations. The transformation of the raw data to a usable dataset was an important technical milestone, and was critical to meaningful forecasting algorithms. The data was then restructured to include an attendance count during each 30 minute interval which took into account the student check In and check out times. Doing this allowed us to get a more accurate student count as it fluctuated throughout each day. This allowed us to create a more accurate measure of how many staff would be needed as this measure would automatically adjust every 30 minutes depending on how many students were currently checked into a classroom. For example, consider the Dinosaur Stomp Room that has a required student to staff ratio of 12:1. If on a given day 5 students checked in at 7:30 am, then the staff needs column that we added would automatically calculate that 1 staff member was needed. At 12:30 pm that same day, if 9 more students checked in to make a total student count of 13, the staffing needs for 12:30 pm would update to 2. Then if 2 students checked out at 2:30pm, decreasing the total students back down to 11, the staffing needs for that interval would adjust back down to 1.

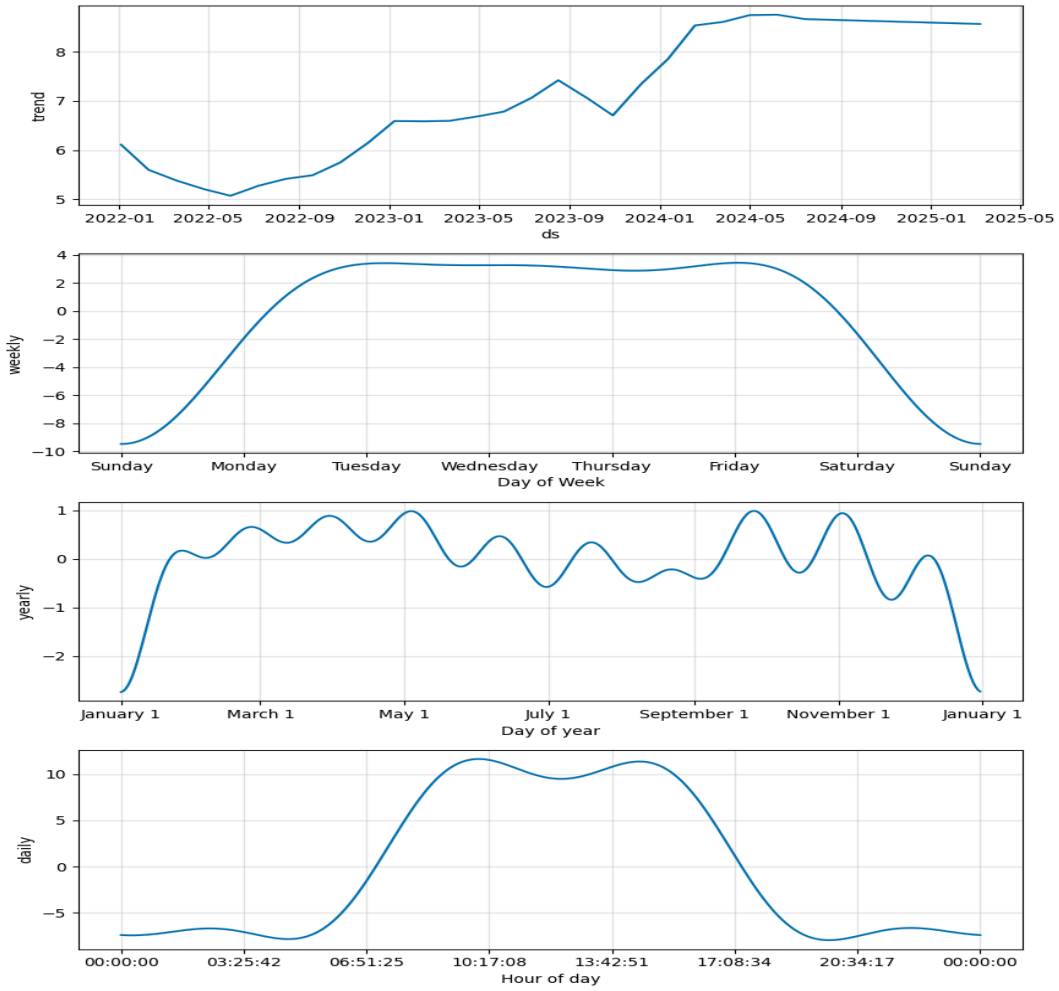
Due to the large amount of computing power it took to run our model we decided it was best to split it up. SpellmanModel was used specifically to process the Spellman data sets and ECECModel was used to specifically process the ECEC data sets. Functionally both are similar; they process separately.

Exploratory Data Analysis (EDA)

Immediately after preprocessing, we used matplotlib to visualize attendance trends in each room, Random Forest for capturing weekly trends, Prophet to confirm data continuity and assess seasonality, and statsmodels for smoothing of the data. These insights (both visual and calculated) shaped our understanding of childcare attendance behavior and our model selection strategy. See Figure 1 for Prophet modeling below. This analysis revealed certain patterns that align with anecdotal knowledge about childcare center attendance gained from the real world, including:

- **Daily Trends**: Peak attendance typically occurred mid-morning, with gradual declines post-lunch.
- **Weekly Variations**: Mondays and Fridays exhibited lower attendance, suggesting the potential for adjusted staffing on these days.
- **Seasonal Fluctuations**: A common trend emerged across all classrooms where attendance (and therefore staffing needs) tended to dip during major holidays and summer months, indicating the influence of external factors.

Figure 1: Prophet modeling for staffing for all classrooms



Forecasting Models

Following the confirmation of solid data, the next step was creating forecasting models designed to harness existing tools in Python. SkiKit Learn is the tool we utilized to create training and testing data and evaluate R2 scores, PyGAM, and XGBoost, Prophet, and Random Forest were evaluated as modeling tools. Model selection was guided by performance metrics such as Mean Squared Error (MSE) and Root Squared. Both XGBoost and Prophet models/tools were considered and tested fully, but when attempting model creation with XGBoost we discovered subpar R-squared (ranging between .50 and .60) and accuracy scores (60% to 70%) compared to our findings with the Random Forest model. We were able to achieve relatively high accuracy scores across the models for each Room as well as relatively high R-squared values with the Random Forest modeling. For example, when using our model to create the weekly forecast for the “Monkeys” room, we were able to obtain an accuracy score of 94% and an R-squared of .82.

We found that the Prophet model revealed trends more than predicting staffing needs, which was helpful in obtaining a big picture of the model and predicting staffing long-term, but the rounding of predictions was not advantageous toward meeting our project objectives. Another factor in selecting a model was the ability to easily run multivariate regressions with multiple dependent variables, due to the forecasting need for both the number of students and the staffing. All of these factors ultimately drove our decision to base our overall forecasting model on Random Forest.

After the selection of Random Forest, we ran the forecasts based on each classroom individually. We felt this process served to provide more personalized forecast results to each classroom than an aggregate forecast created by running the model for all of the rooms together at once. The dependent variables consisted of staffing needs and student count, and the independent variables consisted of hour, year, month, and day. We trained the model on the separated hour/month/year/day variables as opposed to training it on the combined values.

Results

Typical Week Forecast

Figure 2 illustrates the trends for student attendance over the past three years. This provides an indication of the historical attendance, which also predicts the forecasted attendance.

The Random Forest model produced a "typical week" forecast, highlighting the highest staffing needs between 7:30 AM and 11:00 AM, tapering off in the afternoon with reduced staffing requirements after 3:00 PM. Across all rooms, there appears to be consistent attendance from Tuesday to Thursday, suggesting stable staffing needs.

Figure 2: ECEC and Spellman Attendance Trends, 2022-2025

Next Week Forecast

Among the rooms included as part of both the Spellman and ECEC data sets (Dinosaur Stomp, Rainbow Fish, Wild Things, Monkeys, Goodnight Moon, Pandas, Rabbits, Llamas Llamas, Hungry Caterpillars, House of Pooh, PreK-1, PreK-2, Pennie Preschool, Grampy Tom Multi-Age, Pennie Toddlers, Henry Toddlers, Grampy Tom Toddlers, Grampy Tom Preschool, Henry Infants, Henry Multi-Age, and Pennie Infants) we observed the following totals across all forecasts using our random first model over the predicted Next Week 3/1/2025- 3/8/2025:

- 3/1/2025: The model correctly predicted that this was a Saturday and therefore there are zero students and staff predicted throughout the day.
- 3/2/2025: The model correctly predicted that this was a Sunday and therefore there are zero students and staff predicted throughout the day.
- 3/3/2025 (Monday): The first group of 37 students (totaled across each room) are expected to check in between 7am and 7:30am, and the predicted number of staff needed (which is taking into account the different staff ratios required in each room) is 18 staff. The max number of checked in students is expected to peak to 190 students at 11 am with 33 staff members. A group of 40 students are predicted to be among the last to check out between 5:30pm-6:00pm with an expected staff count of 9.
- 3/4/2025 (Tuesday): The first group of 46 students are expected to check in between 7am and 7:30am, and the predicted number of staff needed is 18. The maximum number of checked in

students is expected to peak to 187 students at 10:30pm with 32 staff members. A group of 44 students are predicted to be among the last to check out between 5:30pm-6:00pm with an expected staff count of 17.

- 3/5/2025 (Wednesday): The first group of 22 students are expected to check in between 7am and 7:30am, and the predicted number of staff needed is 15. The maximum number of checked in students is expected to peak to 188 students at 11:30 am with 32 staff members. A group of 47 students are predicted to be among the last to check out between 5:30pm-6:00pm with an expected staff count of 17.
- 3/6/2025 (Thursday): The first group of 6 students are expected to check in between 6am and 6:30am, and the predicted number of staff needed is 2. The maximum number of checked in students is expected to peak to 191 students at 12pm with 34 staff members. A group of 38 students are predicted to be among the last to check out between 4:30p and 5:00p with an expected staff count of 19.
- 3/7/2025 (Friday): The first group of 38 students are expected to check in between 7am and 7:30am, and the predicted number of staff needed is 18. The maximum number of checked in students is expected to fluctuate between 183-181 students from 10am-12:30pm with staff fluctuating between 34-32 predicted members. A group of 32 students are predicted to be among the last to check out between 5:30pm-6:00pm with an expected staff count of 18.
- 3/8/2025 (Saturday): The model stops forecasting at 1am, and the day is a Saturday so it is a count of zero for both staff and students.

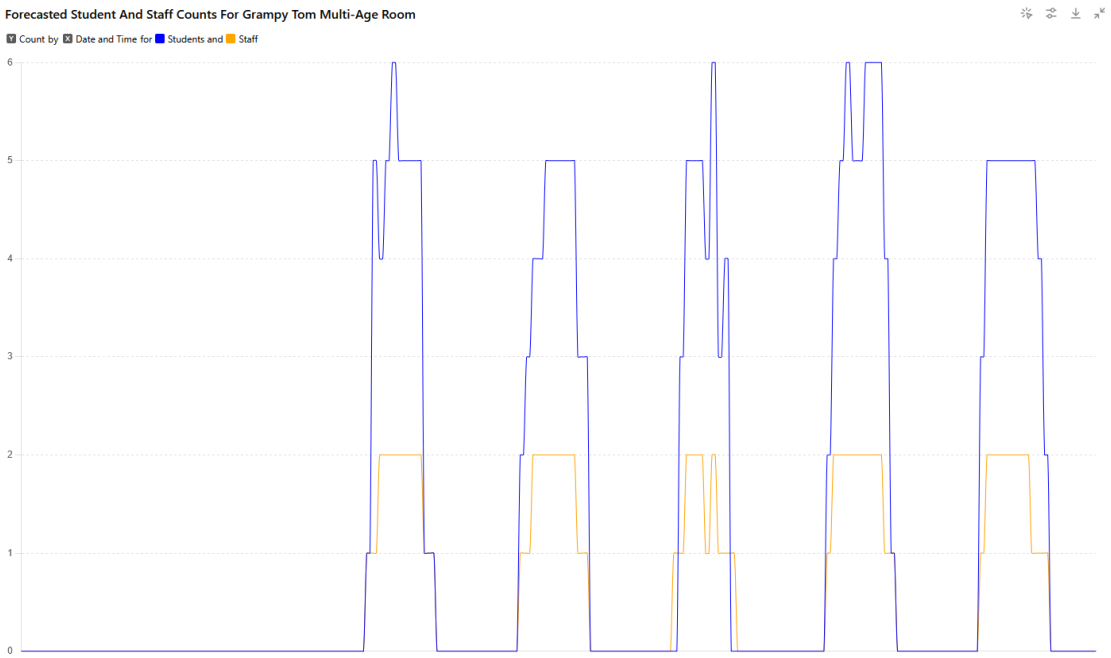
Figures 3 and 4 illustrate room-specific student counts and staffing for two separate rooms.

Figure 3: Predicted Student Count and Staffing for the Wild Things classroom on 3/3/2025:

Interval (in 30 mins)	Predicted Student Count	Predicted Staffing Needs
3/3/2025 7:00	1	1
3/3/2025 7:30	1	1
3/3/2025 8:00	13	2
3/3/2025 8:30	13	2
3/3/2025 9:00	19	2
3/3/2025 9:30	19	2
3/3/2025 10:00	20	2
3/3/2025 10:30	20	2
3/3/2025 11:00	21	2
3/3/2025 11:30	21	2
3/3/2025 12:00	21	2
3/3/2025 12:30	21	2
3/3/2025 13:00	21	2
3/3/2025 13:30	21	2
3/3/2025 14:00	21	2
3/3/2025 14:30	21	2
3/3/2025 15:00	19	2
3/3/2025 15:30	19	2
3/3/2025 16:00	10	2
3/3/2025 16:30	10	2
3/3/2025 17:00	8	1
3/3/2025 17:30	8	1

Figure 4: Grampy Tom Multi-Age Room Forecast for 3/1/2025 – 3/8/2025

*Note: 3/1/25 and 3/2/25 are zero due to the weekend. Data shown is 3/3/25-3/7/25, and the half-hour data is condensed. Blue is Students, Yellow is Staff.



Applying the model to predict the immediate next week following the end date of the data collection, we find that there is a slight increase in attendance on Wednesday, possibly due to midweek programs or other factors. There is a lower attendance forecasted for Friday, aligning with an upcoming public holiday. These anomalies provided evidence that the model is sensitive to recent data, which is desirable for short-term operational planning.

These insights, gleaned from the team's code included in the submitted GitHub Repository, enable proactive staffing adjustments, which assist CSI's effort to increase optimal resource utilization and provide a comprehensive and supportive learning environment.

Discussion

Implications

In terms of this specific data for CSI, the benefit of employing three years of check-in/check-out data is the enhancement of the available data pool and more accurate forecasting for future weeks and staffing needs. The data provided allowed us to include seasonality in the data analysis, verify the adequacy of the data, and provide the model with enough data for the training and testing process. Implementing data-driven staffing forecasts offers multiple benefits to CSI:

- **Operational Efficiency**: Aligning staff schedules with actual needs reduces idle time and prevents burnout.
- **Cost Savings**: Optimized staffing minimizes unnecessary labor costs.
- **Enhanced Care Quality**: Maintaining appropriate staff-to-child ratios ensures better supervision and care.
- **Regulatory Requirements:** Ensuring the appropriate staff-to-child ratios enables CSI to meet state requirements for child supervision.

The implications of this work also extend beyond CSI. Any childcare provider with dynamic attendance could benefit from a similar approach to forecasting. This method allows for the quantification of uncertainty, which may be critical in environments where staffing has life safety, regulatory, retention, and perception implications.

The forecasts provided in this project are not static, and they may be retrained and updated as new data is imported. With proper documentation and training, non-technical staff may be empowered to use and update the model going forward.

Challenges

Several challenges emerged during the project, the first of which was data cleaning due to the inconsistencies in the data entries, multiple pieces of information in fields, and missing or duplicate entries. A second challenge was the creation of 30-minute timeframes for data analysis, since the data fields did not specifically include this information. Adding to this challenge was accounting for the time that the child was in the classroom (after check-in and prior to check-out) to ensure the accurate number of children actually present in the room during a particular amount of time. Getting around this issue took a lot of cleaning and reprocessing the data to make it usable, especially considering that values like the check in and out time were already difficult to process due to the presence of text data. A third challenge was the variation of forecasting models with varying accuracy and error rates, as described above, with XGBoost, Prophet, and Random Forest models. Each has its own strengths and weaknesses, and contributed to a difficult decision process for the best forecasting model.

Future Directions

Going forward, there are several opportunities for improvement, which may have the ability to enhance forecasting accuracy and utility:

- **Ensure Standardized Inputs and Intentional Data Collection:** Ensuring that the check-in/check-out and overall data collection process is deliberate and completely standardized across every system, staff member, and parent will assist in the accuracy of the data, and enable improvements in forecasting accuracy.

- **Incorporate External Data**: Integrating information on school calendars, local events, and weather forecasts can refine predictions.
- **Real-Time Updates**: Developing systems for real-time data collection and model updates can improve responsiveness.
- **Advanced Modeling**: Though the Random Forest model was used here, there may be efficiencies gained by exploring machine learning techniques, such as Neural Networks or Bayesian Modeling, which may capture complex patterns more effectively.
- **Efficient Coding Methods:** In regard to our modeling methods with this data, much of the code built could be created more efficiently through the use of classes and functions to make the replication of equations a lot cleaner and faster. We copied the full code in order to replicate it across classrooms, which created needlessly long code. Cleaning up this process on our end could also have made replication of code and the overall project run quicker. In consideration of future uses of this model, it would be necessary to standardize our code to increase its replicability for the organization or other professionals.
- **Powerful Hardware:** We struggled with running a significant amount of code with large datasets because we lacked the computing power designed for this much technology. This resulted in many crashes and lost data. It also meant that we couldn't run the models for the two data sets (Spellman and ECEC) together, which may have sacrificed important insights on averages and totals in the whole datasets and more granular analysis such as for classroom types (i.e. Pre-K, Toddlers, etc.).

Conclusion

This project highlights the potential of leveraging historical attendance data to forecast staffing needs in childcare settings. By adopting a structured, data-driven approach, organizations like CSI can enhance operational efficiency, ensure compliance with care standards, and ultimately provide better services to the children and families they serve.

References

- Child Saving Institute. (2025). https://childsaving.org/
- GitHub Repository: https://github.com/Taytatat/Semester_ProjectV2
- Our Presentation Recording: https://youtu.be/Jfh4sogSXEo