506 Mid Semester Report

Afzal Khan, Haider Khan, Mohammed Murshid, Rashfiqur Rahman, Maaz Shahzad

04 November 2024

Mid Semester Presentation: https://youtu.be/BasrIvTvmOg
Google Form (for data collection): https://forms.gle/FkYz5mwEaPcQJF4Z8

Project Title: Predicting Optimal Times and Days to Visit the Fitrec

At Boston University, FitRec is the primary gym facility on campus and is often crowded. The goal of this project is to analyze how different factors—such as weather, time of day, day of the week, and academic season (e.g., midterms, finals)—affect gym crowdedness. By identifying patterns in usage, we aim to predict the best times for students to visit FitRec with the least crowd, improving their gym experience.

Data Collection

For our data collection, we used a google form which had questions regarding Fitrec. We used the google form to assess various factors such as respondents' school year, weather conditions affecting the gym, time they most consistently go to the gym, etc. We made sure to ask a variety of questions to get a holistic view of the variables affecting gym attendance. We collected data by asking on various platforms such as Reddit, class Discord, Piazza, and asking people at Fitrec.

Preliminary Visualizations of Data

Weekly Gym Visit Frequency: A count plot was created to show the distribution of students' weekly gym visits. This highlights how frequently students visit the gym and identifies the most common frequency categories, helping us understand general gym usage patterns. We divided the frequencies as 0-2 times (low), 3-5 times (moderate) and 6-7 times (very frequent).

Timeslot Preferences: A bar plot showing attendance by time slot reveals which times are most popular. This helps us target specific times when students are more likely to visit, providing insights into peak hours. Our data reveals that 6-9pm is the most popular time slot for people to work out whereas 6-9 am is the least busiest.

Average Crowdedness Per Time Slot: This bar plot illustrates the average crowdedness rating (1 to 5) for each time slot. It helps identify which times are generally more crowded and which may be optimal for a less crowded experience. People who attended the time slot of 6 -9 am had an average response of 2.5 crowdedness rating, whereas people who attended the timeslot of 6-9 pm had an average of 4.1 crowdedness rating. This supports our weekly gym visit frequency data as well.

Impact of Weather: A pie chart was used to display the proportion of students affected by weather conditions, such as rain or snow, which influences whether they decide to skip the gym. This factor is significant for predicting attendance on particular days. Two thirds of people said that the weather did in fact affect their visit to the gym.

Impact of Exams on Gym Visits: Another pie chart displays whether exams and project deadlines impact gym attendance. This factor may be essential in understanding seasonal patterns during midterms or finals. 88.4% of people said that exams affected their gym visits which is a lot larger than weather impact.

Day of the Week Preferences: A bar plot shows the frequency of gym visits for each day of the week, arranged from Monday to Sunday. This plot helps identify which days are most and least crowded, guiding students on optimal days to visit. Monday was the most popular, while Sunday was the least.

Detailed Description of Data Processing

Data Preprocessing Steps

Timestamp Conversion: Converted the *Timestamp* column to datetime format for easy extraction of time-based features.

Crowdedness Conversion: Converted the crowdedness ratings to a numeric format to allow for statistical and predictive analysis.

Categorical Encoding:

-*Year* was encoded numerically, with each year represented as a category code.

- *Weekly_Frequency* was mapped to numerical values to represent different gym visit frequencies.

- *Time Slots* were converted to numeric codes to represent different time slots, making them suitable for model input.

Handling Missing Data: Used median imputation for missing values in the feature set, which is robust to outliers and ensures no loss of data.

Detailed Description of Data Modeling Methods

Logistic Regression

We utilized a logistic regression model to predict whether a given time slot at the gym would experience "High" or "Low" crowdedness, providing students with insights on optimal visiting times. To set up the binary classification, we defined "High" crowdedness as any rating of 4 or above, while ratings below 4 were labeled as "Low." This transformation created a new binary target variable, *Crowdedness_Binary* where 1 indicated "High" crowdedness and 0 indicated "Low."

The dataset was split into training and test sets, allowing the model to learn patterns in gym usage on one portion of the data and be evaluated on unseen data. We trained the logistic regression model to classify crowdedness, then evaluated it using metrics such as accuracy, precision, recall, and F1 score. These metrics helped gauge the model's effectiveness in distinguishing between high and low crowdedness levels.

The logistic regression model was also set up to make predictions for new inputs, such as specific time slots and day combinations. This allows gym-goers to receive straightforward predictions on whether a particular time slot is likely to be crowded, helping them make data-informed decisions about when to visit the gym for a quieter experience. We got an accuracy of 54% on this model.

Random Forest

Random Forest Regressor was also used to predict the crowdedness level based on features such as year, weekly frequency, hour, day of the week, and timeslot. The model provides continuous predictions for crowdedness, which were later analyzed to find optimal low-crowd times. The data was split into training and test sets (75% training, 25% testing). The model was trained on the feature set X to predict crowdedness levels.

The model's performance was evaluated using the Mean Squared Error (MSE), a common metric for regression tasks. The MSE achieved was 0.80, indicating the average squared error between predicted and actual crowdedness levels.

Grid Search and Hyperparameter Tuning: Future steps include fine-tuning the model's hyperparameters to improve performance. Grid Search or Randomized Search will help explore optimal settings for parameters like the number of trees and max depth.

Preliminary Results

Logistic Regression Results: We got an accuracy of 54% and precision, recall and f1-score of 50%. This model was able to predict whether the gym was crowded or not the majority of the time. It is a good baseline for us, and we plan to improve on it significantly for the final project.

Random Forest Results: Mean Squared Error (MSE): The Random Forest Regressor model yielded an MSE of 0.80 on the test set, indicating the average squared error between predicted and actual crowdedness ratings. This result gives us a baseline for further improvements. The Random Forest Classifier achieved an accuracy of ~50%. Although preliminary, this result indicates that improvements in feature engineering and model tuning could further enhance the model's ability to distinguish between high and low crowdedness.

Optimal Gym Times Prediction: Using the regression model, we generated predictions for crowdedness across all combinations of day and time slot. The model suggests that Thursday and Friday mornings (6 AM to 9 AM) are the least crowded times, providing an optimal window for students seeking a quieter gym experience. Throughout our data, we noticed a **clear pattern** such that the gym was the busiest during 6-9pm on most days and least busy during 6-9am on most days. We plan to be more specific in the final project as well.

Future Steps

Data: We would like to collect more data for the final project in order to generalize our findings more. This will allow us to increase both our training and testing data set size.

Improve Model Accuracy: We would like to experiment with additional features like weather influence, exam impact, and seasonality to improve prediction accuracy. We also want to perform hyperparameter tuning on the Random Forest Classifier to enhance classification metrics.

Explore Other Models: Try advanced ensemble methods such as Gradient Boosting or XGBoost, which may provide better accuracy on this tabular dataset.

Cross-Validation: Implement k-fold cross-validation to ensure that model results generalize across different data splits, minimizing overfitting.

Further Analysis of Seasonal Patterns: Explore patterns by academic season (midterms, finals) and weather conditions to provide additional insights for students planning their gym visits around these factors throughout the semester.

By implementing these improvements and further analysis, we hope to provide Boston University students with actionable insights on the best times to visit FitRec based on data-driven predictions.