Optimizing fitness Class Availability through Predictive Attendance Analysis at GoalZone Fitness Club, Canada.

AUTHOR
Jayesh Bhadane and Shubham Yedekar

17:59

Abstract:

GoalZone, a prominent fitness club chain in Canada, has experienced a surge in demand for its fitness classes. With classes offered in two capacities, 25 and 15 participants, certain popular classes are consistently fully booked. However, a significant discrepancy between booking rates and actual attendance has been observed. This under utilization of booked spaces leads to inefficiency and dissatisfaction among members who could otherwise participate. GoalZone seeks to address this challenge by leveraging predictive analytics to optimize class availability.

Introduction:

In the contemporary landscape of fitness and wellness, GoalZone stands out as a prominent chain of fitness clubs in Canada, renowned for its diverse array of fitness classes and commitment to enhancing the health and fitness of its members. At the heart of GoalZone's service offering are fitness classes, structured in two distinct capacities of 25 and 15 participants. These classes cater to a wide range of fitness enthusiasts, from beginners to seasoned athletes, providing a comprehensive fitness experience.

However, GoalZone faces a unique challenge that is increasingly common in the fitness industry. Certain classes, due to their popularity, are consistently fully booked. However, these fully booked classes often experience a lower attendance rate, leading to an inefficient utilization of resources and a missed opportunity for other interested members. This phenomenon not only impacts the operational efficiency of GoalZone but also undermines member satisfaction, as potential participants are unable to access these high-demand classes.

In response to this challenge, GoalZone has embarked on an innovative initiative aimed at optimizing class attendance and resource allocation. The core of this initiative is the development of a predictive model to forecast class attendance. By accurately predicting whether a member will attend a booked class, GoalZone aims to dynamically manage class capacities, thereby maximizing the availability of class spaces. This predictive approach represents a paradigm shift in how fitness clubs manage class bookings and attendance, offering a more agile and member-centric model.

The objective of this work is to explore the development and implementation of this predictive model within GoalZone. It will delve into the methodologies used for prediction, the challenges encountered, and the potential impact of this model on the operational efficiency of GoalZone and the overall satisfaction of its members. By addressing the gap between class bookings and actual attendance, GoalZone aims to not only enhance its operational efficiency but also to significantly improve the member experience, aligning with its overarching mission of promoting health and fitness.

Data Description:

Below is the description of the data:

Booking ID: A unique identifier for each booking.

Months as Member: The number of months the member has been associated with GoalZone.

Weight: The weight of the member in kilograms.

17:59

Days Before: The number of days the booking was made before the class.

Day of Week: The day of the week when the class is scheduled (e.g., Wednesday, Monday).

Time: Indicates whether the class is in the morning (AM) or evening (PM).

Category: The type of fitness class (e.g., Strength, HIIT, Cycling).

Attended: A binary variable indicating whether the member attended the class (1) or not (0).

Goal:

The primary goal/objective of this project is to develop a predictive model that accurately predicts if a club member is going to attend the fitness class or not. By predicting the likelihood of members attending booked classes, we aim to dynamically adjust class availability, thereby increasing overall access and efficiency.

Univariate exploratory data analysis:

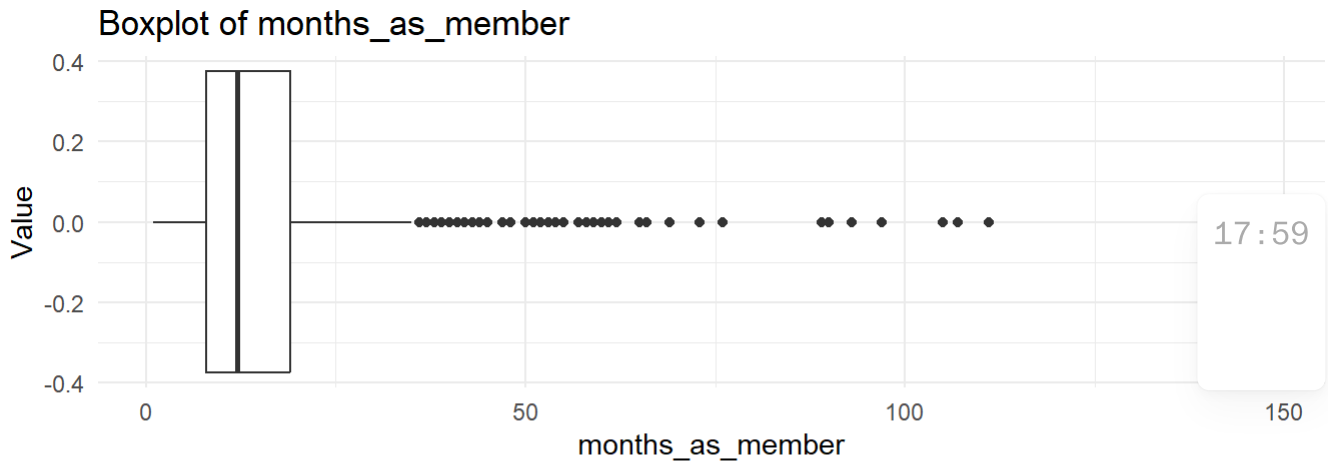
We will now perform univariate EDA to see the distribution of the data. We will check the distribution of continuous variables using boxplot and histogram. For categorical variables, we will use the table function and bar plot to check the frequency/counts..

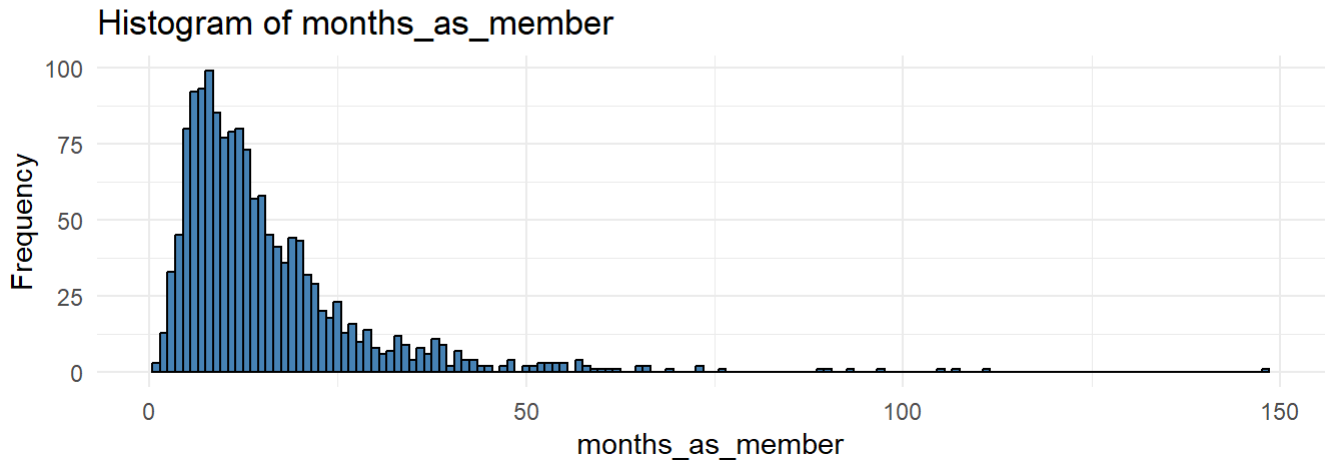
We will also have to check if there are any missing values present in the data set. After checking the missing values, we found that there are 20 missing values in the `weight`. The distribution for `weight` is slightly right skewed. Therefore, we can't impute the missing values using the mean/average value of `weight`. This is because, the mean value will be heavily affected because of the presence of the outliers. The median is more robust in skewed distributions. It is less affected by outliers and skewed data, making it a better measure of central tendency in these cases. Hence, we have imputed the missing values using the median of `weight`.

Let's move to the plots of the variable:

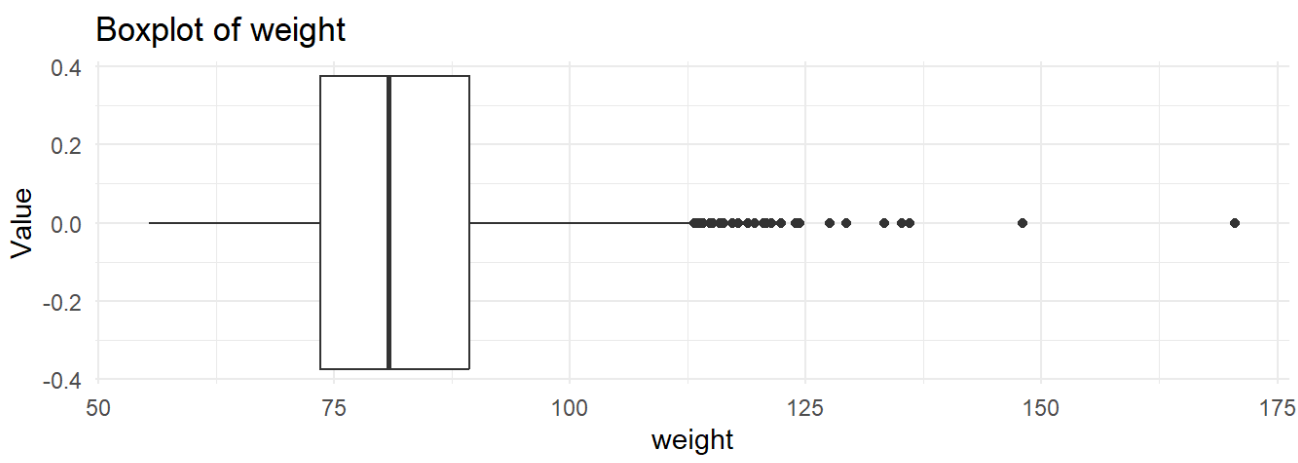
Months_as_member:

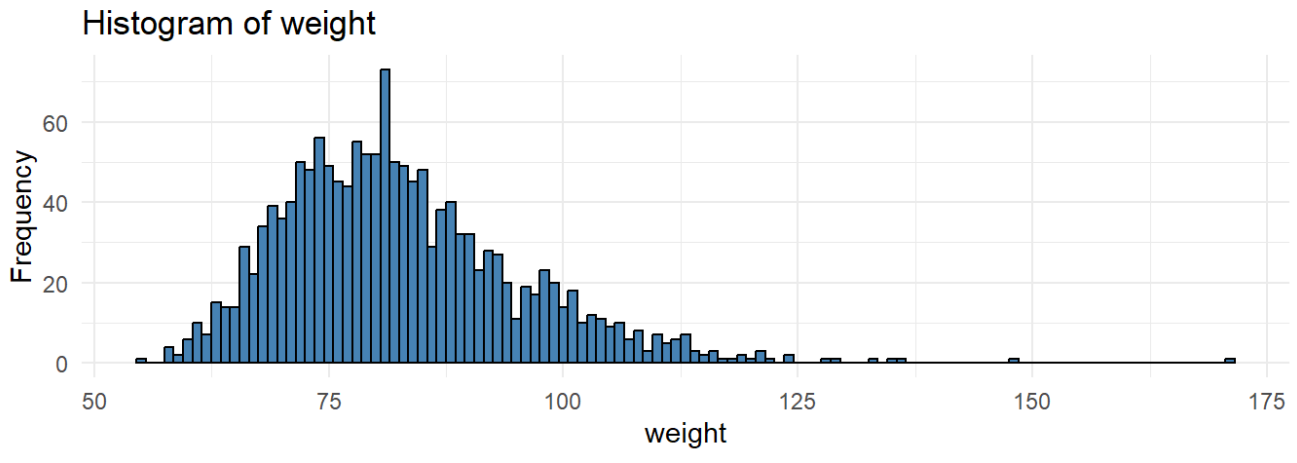
```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
ℹ Please use tidy evaluation idioms with `aes()`.
ℹ See also `vignette("ggplot2-in-packages")` for more information.
```





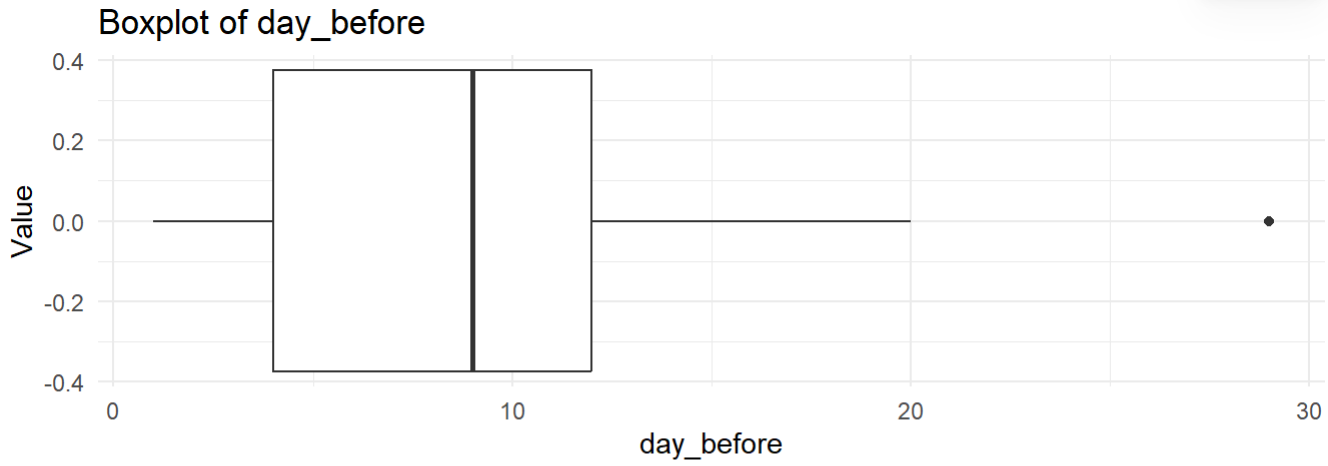
We can see that the variable `months_as_member` is somewhat right skewed. There are some outliers present in this variable.

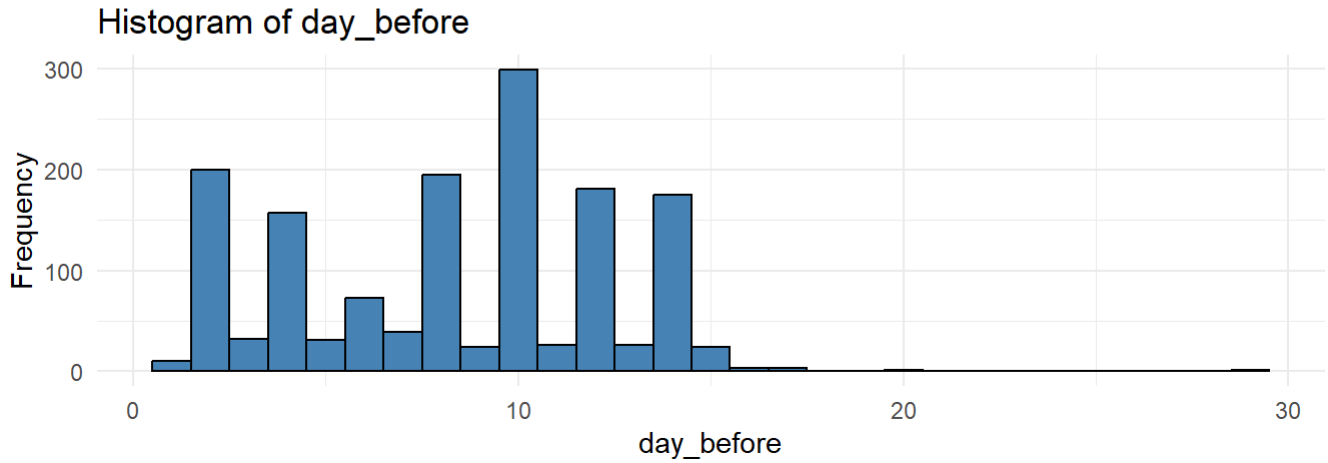




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  55.41   73.56   80.76   82.59   89.38  170.52
```

- The histogram illustrates the frequency distribution of weights. Most members have weights clustered around the 70-90 kilogram range, forming a bell-shaped distribution, which suggests normality.
- There is a long right tail, indicating that there are relatively few members with very high v This aligns with the outliers seen in the boxplot.





The histogram provided illustrates the distribution of the number of days before a class that members of GoalZone fitness club make their bookings.

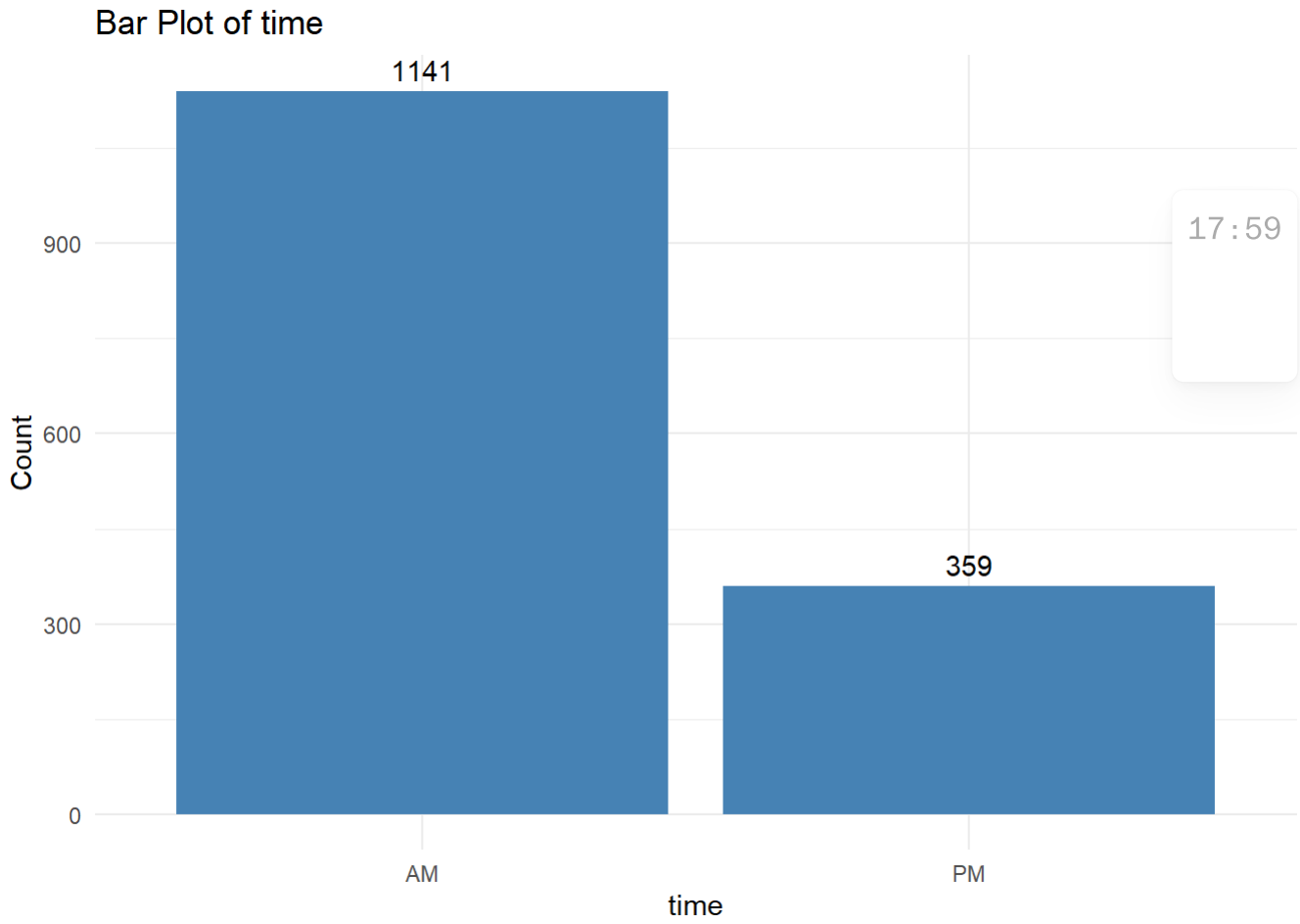
The most common time frame for making a booking appears to be around 7 to 12 days before the class, with the peak frequency occurring at around 10 days before. This could be a policy window where most members are actively booking classes.

There are fewer bookings made very close to the class date (0 to 2 days before) and also fewer made more than two weeks in advance.

The distribution is not symmetric and shows a preference for booking classes one to two weeks ahead.

```
  AM   PM
1141  359
```

```
Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
ℹ Please use `after_stat(count)` instead.
```



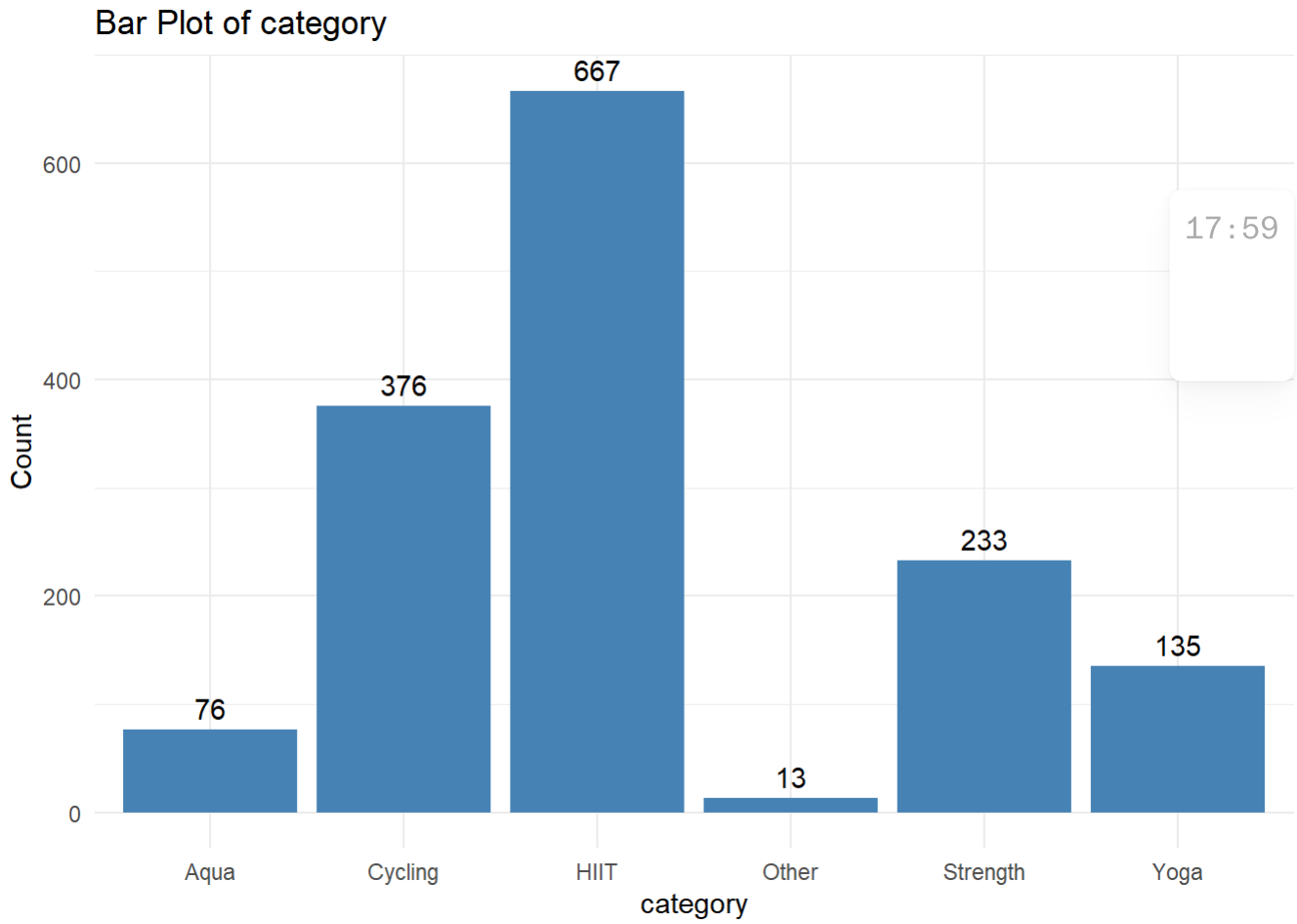
The bar plot depicts the count of fitness class bookings at GoalZone fitness club categorized by the time of day — AM or PM.

From the bar plot, we can observe the following:

There is a significantly higher number of bookings for classes in the AM (morning) with a count of 1141 compared to the PM (evening) with a count of 359.

This suggests a strong preference among the members for attending fitness classes in the morning over the evening.

```
Aqua  Cycling     HIIT    Other Strength     Yoga
  76      376      667       13      233      135
```



The bar plot shows the number of bookings for different categories of fitness classes at GoalZone.

From the above bar plot, we can observe that:

- The HIIT (High-Intensity Interval Training) category has the highest number of bookings, with a count of 667.
- Cycling comes second with 376 bookings, followed by Strength with 233 bookings.
- Yoga classes have a relatively moderate number of bookings at 135.
- Aqua classes appear to be the least popular with only 76 bookings.
- There is a category labeled "Other" which has very few bookings, only 13, suggesting these may be specialized or less frequently offered classes.

Outlier treatment

As we observed, there are some outliers present in the data. We have removed such potential outliers using the Interquartile Range (IQR) method, specifically from the `months_as_member` and `days_before` variables. The reason why we did not remove outliers from the `weight` is that there is a high chance those outliers represent obese and overweight customers of GoalZone fitness gym, which indicate potentially unhealthy customers who tend to attend the fitness sessions organized at the GoalZone gym.

Initially, the dataset contained 1500 customers. After addressing the outliers in `months_as_member`, the size of the data reduced to 1397. Further treating the outliers in 'days_before' brought the dataset size down to 1396.

Modeling phase:

To achieve the goal of accurately forecasting member attendance at fitness classes, we employed a suite of machine learning algorithms, each offering unique strengths in handling classification problems. The models were chosen based on their popularity, performance, and interpretability in similar domains. The models included:

- **Logistic Regression**: As a statistical model that predicts the probability of a binary outcome, logistic regression was used as a baseline model. It's particularly advantageous due to its simplicity, interpretability, and efficiency in terms of computation.
- **Decision Tree Classifier**: This model was chosen for its interpretability and ease of use. Decision trees split the data into subsets based on the value of input features, which makes it simple to understand and visualize.
- **Random Forest Classifier**: An ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes as the prediction. The random forest classifier was selected for its robustness to overfitting and its ability to handle non-linear relationships.
- **XGBoost (Extreme Gradient Boosting)**: An implementation of gradient boosted decision trees designed for speed and performance. XGBoost was included due to its reputation for outperforming other algorithms on a variety of machine learning benchmarks.

Model Training and Evaluation

Each model was trained on a dataset comprised of the features mentioned earlier, A hold-out validation set (20% of the cleaned data), separate from the training data, was used to evaluate the performance of each model. The metrics used for evaluation included accuracy, precision, recall, and the confusion matrix.

Logistic regression:

Null Model Benchmarking

As a preliminary step in our predictive analysis, we established a baseline by fitting a null model, which includes no predictors and only the intercept. The null model serves as a reference point to assess the improvement added by including predictor variables. The residual deviance of the null model was 1285, with an Akaike Information Criterion (AIC) of 1287. The intercept was found to be statistically significant ($p < 2e\text{-}16$), indicating that even without any predictors, the model already provides some information about the likelihood of members attending a class.

Full Model with All Predictors

We progressed to a full model incorporating all available predictors to capture as much information as possible. The inclusion of these predictors significantly improved the model's fit, as evidenced by a reduced residual deviance of 1060 and an AIC of 1092. This substantial decrease in deviance compared to the null model suggests that the predictors contribute meaningfully to explaining the variation in class attendance.

Stepwise Selection for Optimal Model

To refine our model further, we employed stepwise selection, a systematic method of adding and removing predictors based on their statistical significance and contribution to the model. The stepwise procedure converged on a model that includes 'months_as_member' and 'time' as significant predictors. The residual deviance for this stepwise-selected model is 1078.2, with an AIC of 1084.2.

While the AIC is higher than that of the full model, it is still lower than the null model, indicati er balance between model complexity and explanatory power. The slight increase in deviance cc o the full model suggests that while some predictive power may have been lost, the stepwise m more parsimonious, potentially improving its generalizability to unseen data.

Model Evaluation:

Full Model Performance

The full model's performance was evaluated using a confusion matrix, which is a table used to describe the performance of a classification model. The confusion matrix for the full model is as follows:

	Predicted No	**Predicted Yes**
Actual No	196	49
Actual Yes	10	25

From the confusion matrix, we calculated the accuracy of the full model, which is the proportion of the total number of predictions that were correct. The model achieved an accuracy of 78.93%. This means that approximately 78.93% of the class attendance was correctly predicted by the full model.

Stepwise Model Performance

The stepwise selection model was also evaluated using a confusion matrix:

	Predicted No	**Predicted Yes**
Actual No	199	46
Actual Yes	7	28

The accuracy of the stepwise model was found to be 81.07%. This is an improvement over the full model, indicating that the stepwise model was better at classifying the correct outcomes for class attendance.

When comparing the two models, the stepwise model not only has a higher accuracy but also shows an improvement in both types of correct predictions — true negatives (where non-attendance is correctly predicted) and true positives (where attendance is correctly predicted). Specifically, the stepwise model correctly predicted 3 more true negatives and 3 more true positives than the full model.

The lower number of false negatives (actual attendance that was predicted as non-attendance) in the stepwise model suggests that it is less likely to predict that a member will not attend when they actually

do, which is an important consideration for minimizing the chance of underbooking a class.

Decision trees:

The root node error of the decision tree was found to be approximately 0.2654, indicating the proportion of misclassified instances at the initial stage before any splits. As the tree was cons the complexity parameter (CP), which governs the tree's growth by evaluating the improveme model fit with each split, was set to a minimum of 0.001 to prevent overfitting.

17:59

The tree's splits were based on the values of these variables that best segregated the attendance (yes or no). The 'rel error' column in the output indicates the relative error of each split, which decreased as the tree depth increased, showcasing the model's ability to learn from the data.

Model Evaluation

Confusion Matrix

To evaluate the effectiveness of the decision tree model in predicting class attendance at GoalZone fitness clubs, we analyzed the model's confusion matrix and calculated key performance metrics. The confusion matrix is as follows:

	Predicted No Attendance (0)	**Predicted Attendance (1)**
Actual No Attendance (0)	171	40
Actual Attendance (1)	35	34

Interpretation of Metrics

- **Accuracy**: The model achieved an accuracy of 73.21%, indicating that roughly 73 out of 100 predictions were correct. This metric reflects the overall ability of the decision tree to correctly classify both attendance and non-attendance cases.
- **Precision (Predicted Attendance)**: Precision for predicted attendance (class 1) was 81.04%. This means that when the model predicted a member would attend the class, it was correct about 81% of the time. High precision is indicative of a low false positive rate.
- **Recall (Actual Attendance)**: The recall for actual attendance was 83.01%. This suggests that the model successfully identified 83% of all actual attendance cases. High recall indicates the model's strength in minimizing false negatives.
- **F1 Score**: The F1 score, which is the harmonic mean of precision and recall, was 82.01%. This score is particularly useful because it balances the trade-off between precision and recall, providing a single measure of the model's accuracy in cases where the class distribution is imbalanced.

Pruned trees:

Pruning a decision tree is a technique used to reduce the size of the tree by removing sections of the tree that provide little power to classify instances. This process simplifies the tree, avoids overfitting to

the training data, and can improve the model's ability to generalize to new data. Pruning can be done by cutting off branches (nodes) that have low importance or by setting a minimum threshold on the gain of a node to continue splitting.

The metrics for your pruned tree are as follows:

	Predicted No Attendance (0)	Predicted Attendanc
Actual No Attendance (0)	200	52
Actual Attendance (1)	6	22

- **Accuracy**: After pruning, the accuracy of the tree is approximately 79.29%. This means that nearly 79 out of every 100 predictions made by the model are correct. The accuracy has improved compared to the unpruned tree, indicating that pruning has likely removed overfitting and enhanced the model's generalization.
- **Precision**: The precision is approximately 79.37%. This indicates that when the tree predicts that a member will attend a class, it is correct about 79% of the time. This is slightly lower than the unpruned tree, which might be due to the pruned tree making fewer positive predictions overall but being more conservative and accurate when it does.
- **Recall**: The recall has significantly increased to 97.09%, meaning that the pruned tree is very effective at identifying true attendees. It correctly identifies 97% of all actual attendance cases, which is an increase from the unpruned tree.
- **F1 Score**: The F1 score is 87.34%, which is higher than the unpruned tree, indicating a better balance between precision and recall. This suggests that the pruned tree, while making fewer positive predictions, is making them more accurately, and missing very few actual attendees.

In short, the pruned tree has shown improvements in accuracy and recall, with a slight trade-off in precision. The higher F1 score indicates a better overall balance between precision and recall, suggesting the pruning was beneficial for the model's performance.

RandomForest classifier:

We utilized the `ranger` package in R to create a Random Forest model for classifying the likelihood of members attending their booked classes at GoalZone. Random Forest is an ensemble learning method known for high accuracy and robustness, which works by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees.

Model Configuration

The model was configured with the following parameters:

- **Number of Trees**: 500 trees were generated, providing a diverse set of classifiers to vote on the final prediction.
- **Sample Size**: All 1116 observations in the training set were used to grow the trees.

- **Number of Variables Tried at Each Split**: 3 variables were randomly selected at each split point, providing a good balance between prediction accuracy and model simplicity.
- **Minimum Node Size**: Set to 1, which means that the model will attempt to split until each node has only one observation or until it cannot improve the prediction, ensuring maximum depth and detail in the trees.
- **Variable Importance**: Measured by impurity, specifically using the Gini index, which prov insight into which variables are most influential in predicting the outcome.

17:59

Model Evaluation:

The out-of-bag (OOB) prediction error was reported as 25.63%, indicating that approximately 74.37% of the time, the model correctly predicted whether a member would attend the class based on the OOB sample. This error rate is derived from predictions on the training set itself but using only the trees that did not have the particular sample in their bootstrap sample.

The Random Forest model's effectiveness was rigorously assessed using a confusion matrix, which quantifies the model's predictions against actual outcomes. The following results were obtained:

	Predicted No Attendance (0)	**Predicted Attendance (1)**
Actual No Attendance (0)	190	43
Actual Attendance (1)	16	31

Interpretation of Performance Metrics

- **Accuracy**: The model has an accuracy of approximately 78.93%, indicating a high level of overall predictive success. This means that the model correctly predicted attendance status for nearly 79 out of every 100 instances.
- **Precision**: Precision, which assesses the model's ability to accurately predict positive attendance, is about 81.55%. This suggests that when the model predicts a member will attend, it is correct in that prediction more than 81% of the time.
- **Recall**: The model achieved a recall of 92.23%, indicating it is highly effective at identifying true attendance cases. In other words, it successfully captures 92% of the instances where members actually attend the class.
- **F1 Score**: The F1 score, which balances precision and recall, is 86.56%. This robust score underscores the model's balanced approach to both false positives and false negatives.

XGBoost (Extreme Gradient Boosting):

We then employed the XGBoost algorithm, a powerful and scalable machine learning technique for regression and classification problems. XGBoost is known for its performance and speed, often outperforming other algorithms on benchmark tasks.

During the training phase, we monitored the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for both the training and test datasets across ten iterations. The AUC-ROC is a performance measurement for classification problems at various threshold settings. It tells how much the model is capable of distinguishing between classes, with a higher AUC indicating a model that is better at predicting 0s as 0s and 1s as 1s.

17:59

Here are the key takeaways from the iterative training rounds:

- The AUC for the training data starts at 0.767787 and shows a general increasing trend, in[illegible] the model is effectively learning from the data. It reaches an AUC of 0.810508 by the 10th round.
- The AUC for the test data begins at 0.798183 and demonstrates fluctuations across the iterations, with an initial increase and then some variance in later rounds, peaking at 0.837346 in the 2nd round. By the 10th round, it settles at an AUC of 0.787621.

  As we increased the training round, the training AUC kept on increasing. However, the test accuracy was varying around 0.79 to 0.83. This suggests that if we keep on increasing the training round, the model will tend to overfit.

XGB Model Performance Assessment

Following the training phase, the XGBoost model's classification ability was quantified using a confusion matrix, and several key metrics were calculated to evaluate its performance:

	Predicted No Attendance (0)	**Predicted Attendance (1)**
Actual No Attendance (0)	191	45
Actual Attendance (1)	15	29

Key Performance Metrics

- **Accuracy**: The model achieved an accuracy of approximately 78.57%. This metric indicates the proportion of total predictions that the model got right, demonstrating a relatively high level of overall predictive success.
- **Precision**: Precision for the model stood at about 80.93%. This high precision rate suggests that when the model predicts a member will attend, it is correct in that prediction the majority of the time, which is crucial for planning and resource allocation.
- **Recall**: The recall for the model was around 92.72%. This indicates that the model is highly effective at identifying members who will actually attend, with more than 92% of all actual attendance cases being correctly recognized by the model.
- **F1 Score**: The F1 score, which considers both precision and recall, was 86.43%. An F1 score this high suggests that the model has a balanced approach to classifying both the positive class (attendance) and the negative class (no attendance), which is important for maintaining service quality and member satisfaction.

Results from the analyses:

We compared the metrics for all the models above:

Stepwise Logistic Regression Model

- Accuracy: 81.07%
- Precision: 80.00%
- Recall: 37.84%
- F1 Score: 51.38%

Full Logistic Regression Model

- Accuracy: 78.93%
- Precision: 71.43%
- Recall: 33.78%
- F1 Score: 45.87%

Decision Tree (Unpruned):

- Accuracy: 73.21%
- Precision: 81.04%
- Recall: 83.01%
- F1 Score: 82.01%

Decision Tree (Pruned):

- Accuracy: 79.29%
- Precision: 79.37%
- Recall: 97.09%
- F1 Score: 87.34%

Random Forest:

- Accuracy: 78.93%
- Precision: 81.55%
- Recall: 92.23%
- F1 Score: 86.56%

XGBoost:

- Accuracy: 78.57%

- Precision: 80.93%
- Recall: 92.72%
- F1 Score: 86.43%

For GoalZone's task of using predictive analytics to optimize class availability by predicting at in fitness classes, the choice of the best metric (accuracy, precision, recall, or F1-score) will de the specific aspects of the problem and the business goals:

17:59

1. **Accuracy**: Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. It is a good measure when the classes are balanced (i.e., the number of instances of the two classes is similar). However, in scenarios where there's an imbalance (like many more attendees than non-attendees or vice versa), accuracy might not be the best metric.
2. **Precision**: Indicates how many of the predicted positive (attended) cases actually turned out to be positive. High precision means that when the model predicts a member will attend, they are likely to do so. This is important if GoalZone aims to minimize instances where members are wrongly predicted to attend (thus potentially taking slots away from others).
3. **Recall (Sensitivity)**: Shows how many of the actual positive cases the model correctly predicted. High recall is crucial if the goal is to identify as many actual attendees as possible, even if it means some false positives (predicting attendance when it doesn't happen).

For this scenario, if the focus is on maximizing the utilization of each class (i.e., ensuring classes are as full as possible), **recall** might be the most important metric because it will focus on minimizing the number of missed attendees. However, if GoalZone is more concerned about ensuring that those who are predicted to attend are the most likely ones (to avoid overbooking or denying slots to others), then **precision** becomes more important.

We can also consider the f1-score to evaluate the model. F1-score is a good balanced metric if both aspects (minimizing missed attendees and avoiding overbooking) are equally important. It is a harmonic mean of both precision and recall and is a good choice when you seek a balance between precision and recall, especially if there is an uneven class distribution.

Conclusion:

Through the application of these predictive models, we have developed a robust framework that can forecast class attendance with a high degree of accuracy. The insights gained from these models will enable GoalZone to better manage class capacities, reduce inefficiencies, and enhance the overall member experience by ensuring that more members can attend the classes they desire.