**Exploring Attendance Patterns and Predictive Modeling for Fitness Club Members**

|  |
| --- |
| ABSTRACT  The project is about analyzing attendance patterns and predictive modeling for fitness club members.  We particularly focus on Goal Zone, a chain of fitness clubs in Canada. We have done detailed examination of member data and applied machine learning techniques like Logistic Regression, Random Forest Classifier and K-Nearest Neighbors (KNN). This study identifies that the weight and membership duration are factors that significantly influence the attendance. By using the results fitness club can refines their strategies, enhance member engagement, and optimize operations. This will leads to more vibrant and engaging environment for their members. |

|  |
| --- |
|  |
| Table of contents  [Introduction: ------------------------------------------------------------------------------------------------------------ 1](#_Toc165779537)  [Methodology ----------------------------------------------------------------------------------------------------------- 1](#_Toc165779538)  [Data Exploration ------------------------------------------------------------------------------------------------------ 1](#_Toc165779539)  [Descriptive Data Analysis Summary: ------------------------------------------------------------------------- 1](#_Toc165779540)  [Questions for valuable insights --------------------------------------------------------------------------------- 2](#_Toc165779541)  [Preprocessing -------------------------------------------------------------------------------------------------------- 4](#_Toc165779544)  [Data Wrangling: Preparing the Dataset for In-Depth Analysis: --------------------------------------- 4](#_Toc165779545)  [Handling Outliers -------------------------------------------------------------------------------------------------- 4](#_Toc165779546)  [Handling Imbalance Dataset ----------------------------------------------------------------------------------- 4](#_Toc165779547)  [Scaling Dataset ---------------------------------------------------------------------------------------------------- 5](#_Toc165779548)  [Exploratory Data Analysis ------------------------------------------------------------------------------------------ 5](#_Toc165779549)  [Histograms and Box Plots --------------------------------------------------------------------------------------- 5](#_Toc165779550)  [Box Plot -------------------------------------------------------------------------------------------------------------- 6](#_Toc165779551)  [Correlation Heatmap --------------------------------------------------------------------------------------------- 7](#_Toc165779552)  [Machine Learning Techniques ------------------------------------------------------------------------------------ 8](#_Toc165779553)  [Logistic Regression ----------------------------------------------------------------------------------------------- 8](#_Toc165779554)  [Random Forest Classifier --------------------------------------------------------------------------------------- 8](#_Toc165779555)  [K-Nearest Neighbors (KNN) ------------------------------------------------------------------------------------ 9](#_Toc165779556)  [Grid Search CV ---------------------------------------------------------------------------------------------------- 9](#_Toc165779557)  [Comparative Analysis of Machine Learning Models: -------------------------------------------------------- 9](#_Toc165779558)  [Assessment of Logistic Regression: -------------------------------------------------------------------------- 9](#_Toc165779559)  [Random Forest Assessment: ----------------------------------------------------------------------------------- 10](#_Toc165779560)  [Assessment of KNN ----------------------------------------------------------------------------------------------- 11](#_Toc165779561)  [Grid Search CV Assessment: ----------------------------------------------------------------------------------- 11](#_Toc165779562)  [Identifying Features Importance ---------------------------------------------------------------------------------- 12](#_Toc165779563)  [Conclusion -------------------------------------------------------------------------------------------------------------- 13](#_Toc165779564) | |

|  |
| --- |
| Introduction: Understanding member behavior is key for effective and better management in a fitness club. This project focuses on this issue and analyze and predict attendance of fitness club members. For this task we choose Goal Zone, a chain of clubs in Canada. By examining different features like weight, attendance rate and months of member, we can easily predict the attendance behavior of members and improve planning for club. We have done data cleaning and exploratory analysis which uncovers trends and hidden patterns in data. After that we applied machine learning techniques like Logistic Regression, Random Forest Classifier and K-Nearest Neighbors (KNN) to predict attendance behavior. After assessment of models and tunning hyperparameters of modals we highlighted the importance of features for prediction. Finally, this provides some valuable information for fitness club to optimize and improve their strategies. Ultimately, club improve their anviornment and members experience. Methodology The technique that we used for this project includes several procedures for cleaning, analyzing and predictive modeling. The data set contains data from Goal Zone, a chain of clubs in Canada. In cleaning process we have done many steps including handling missing values, encoding categorical columns and handling missmatched data in some columns. After cleaning the data we have done some visualizations to study some patterns and insights. In Exploratory Data Analysis (EDA) we used methds to visualize like correlation heatmap and boxplot. We observerd some outliers in data and we use quartile method to detect and remove them from data. These EDA techniques provides many data insigths which then be used for further analysis. Then we have used machine learning techniques to predict attendance for members of club. Then model assessment is done and improvement is done by tunning hyperparameters and finding best parameters for models. For that task we have included Grid Search CV and find parameters for all models explicitly. Finally we find the importance of features for predicting attendance by using random forest feature importance feature. This study finally made it easier for club to study members behavior and making useful strategies for fitness club. Data ExplorationDescriptive Data Analysis Summary: To understand the features of the dataset we use descriptive data analysis which includes a variety of statistical techniques. Below are some results of this analysis:   * **Months as Member:** The average of this column is 15.63 months. With a standard deviation of 12.93, the range in membership duration. The maximum value of this column is 148 months. * **Weight Distribution:** Members weigh about 82.61 units on average. The weight data have a standard deviation of 12.68, which indicates a significant variation in this column. Most of the values in the column lie between 73.56 and 89.38 units. * **Days Before Events:** Reservations for events typically take place on an average 8.35 days in advance. The mean of column is 9.00 which is higher than median which indicates that the distribution is skewed toward the positive. * **Attendance Rate:** This column has binary values, showing whether a given event was attended (1) or not (0). The average shows that roughly 30% of events were attended.  Questions for valuable insights:How does the duration of membership correlate with event attendance? The goal of this question is to analyze the correlation between how long is the person part of club and how many time the person attend the events. The member with long-term involvement in the activities of event and club is considered loyal and committed. The aim of this is to evaluate if there is difference in the attendance patterns of long-term members and new members. The one’s with long term membership are loyal indicating high attendance and participation but decline in engagement. Relation between membership and event attendance can be better examined by this correlation.    Here we are examining relation between two variables, months\_as\_members and attended by creating box plot. But these box plots indicates some outliers by which the mean values and graph is difficult to read. Through analysis of graph, we may conclude that the members who attend more classes span for 15 months and other who do not attend the classes span 11 months indicating the loyality of long-term members. Is there a correlation between specific categories and higher attendance rates among fitness club members? The correlation between categories and high attendance rates among fitness club members has of great importance. The key points are broken out as follows:  The fitness center may improve member satisfaction and engagement by customizing event schedules to suit the preferences of frequent participants. Employing successful marketing strategies will showcase the club's appealing atmosphere and aid recruiting new members. By allocating resources based on high-attendance categories, the club can optimize member experiences and resource use. Extending offerings and concentrating on high-attendance regions can also improve community engagement and club loyalty. This will encourage new members and maintain the interest of current ones.  A graph of different colored bars  Description automatically generated with medium confidence  The above graph represents the relation between category and count, describing count of students who attend and miss classes in each category. Majority indicating that most classes are left unattended. HIIT category has more count as compared to other categories so it has likely more attendance than the other category.  **A chart of a number of people  Description automatically generated with medium confidence**  Above is the graph “percentage of attendance by category” representing percentage of each category that attends or miss classes. “Aqua” has high rate of attendance of “32.9%”. PreprocessingData Wrangling: Preparing the Dataset for In-Depth Analysis: We have taken many actions during data wrangling for this project to enhance the pre-processing and suitability of dataset for this research.   * Handling Missing values in the 'weight' column: There are twenty missing values in the 'weight' column of our dataset. The rows containing missing values can be eliminated, since these make up only 1.33% of the dataset. However, for numerical columns, it is suitable to impute missing values using the column mean. We used the mean value of weight column to impute missing values. * Numerical Conversion of 'days\_before': 'days\_before' column contains some values including text with values (such as '14 days' or '47 days'). Therefore, its datatype is object. We choose only the integer portion of the data from all rows and convert its datatype to integer. * Handling missing values from 'category' column: The 'category' column has a missing category (shown by '\_') there are only 13 instances of it. There are two solutions to this problem: first is to impute the missing values with the category that appears the most frequently or mode in our dataset, "HIIT," and the second is to mark it as a new category in the column. Since there are just 13 instances in our dataset or 0.86% of all rows, creating a new category is not suitable. Therefore, we imputed the most frequently occurring category in place of '\_'. * Making uniform representation for 'day\_of\_week' column: The days were written in a variety of ways, including "Monday," "Mon.," and "Mon.". We chose the first three characters of each day to handle these variances in values and produce a uniform representation for whole column. Now the format of the entire column is like ('Mon', 'Tue', 'Wed', etc.).  Handling Outliers We used the quartile approach to deal with outliers, which is figuring out which observations are above the third quartile (Q3) plus 1.5 times the IQR, or below the first quartile (Q1) minus 1.5 times the IQR. To see the association between the variables "months\_as\_member" and "attended," we first made a box plot. Nevertheless, we observed many outliers that warped the mean values and skewed the graph. We eliminated the outliers to remedy this. After removing outliers, the data showed that people who attend classes have been members for an average of more than 15 months, while people who don't attend have been members for an average of around 10 months. It results that the members that are from longer time are more likely to attend the classes. Handling Imbalance Dataset Handling Imbalance dataset is important for those datasets that have some dominating classes in it. These dominating classes can made models performance biased for dominating class and model cannot performe well on new data. To address this issue we have used Synthetic Minority Over-sampling Technique (SMOTE) which is a built-in method in scikit-learn to handle imbalanced datasets. This technique improve the representation of minority class instances by oversampling the minority class. By using this the model can learn more evenly distributed data that improve its performance. Scaling Dataset Scaling is important when there are some dominating features in term of scale because some models learn patterns on the bases of distances like KNN and such data can affect their performance. To handle this issue we used Standard Scaler from scikit-learn this normalizes the data and tranforme the mean to zero and standard deviation to one. Our models stability and convergence improved using these models. Exploratory Data Analysis Exploratory Data Analysis (EDA) is an important step for looks patterns and insigths of data. It includes tasks like understanding data structure and finding anomalies and visualizing features relationship. We do a range of statistical and visual analysis to extract relevant information. Some are listed below: Histograms and Box Plots The below histogram shows that the 'months\_as\_member' column has a distribution that is right-skewed. In other words, most members have only been involved in the club for a short period of time, but a small percentage have been there for much longer due to which the distribution have a tail on the right side.  Work with normal distribution is easier and preferred so we used a log transformation to remove this skewness and make distribution normal. Log transformation is a mathematical procedure that modifies the values so that higher values are compressed more than lower ones.  A graph showing a number of data  Description automatically generated with medium confidence  After the log transformation, the histogram is shown below. Which is more normal or has bell-shaped distribution. This normal distribution helps in analytical and modeling techniques that rely on the assumption of a normal distribution of data.  A graph with a curve  Description automatically generated Box Plot In addition to histograms, box plots provide a concise overview of important statistical features of dataset. The inter-quartile range (IQR) is represented by the box in a box plot which contains the middle 50% of the data. The main feature of Box plots is that they highlight the distribution of outliers in our data. Whenever results are unusual it is important to look at the spread of outliers and whiskers in our box plot.  Below is the box plot for the 'weight' column of our dataset. It describes the data variation and how it is related to attendance column. The plot tells that members of club who attend classes have less weight than those members who do not attend classes.   Correlation Heatmap Correlation Heatmap is an important way to analyze the correlations of columns of the dataset. Below is the correlation heatmap of our dataset which provides information of the relationships between various columns in our data. The graph shows some relationships, such as the inverse relationship between weight and months\_as\_member columns. This inverse relationship tells how long a person has been a member of this fitness club. Also, it shows a high correlation between the day of the week and the number of days columns. Moreover, it indicates that members who have been within the club for a long time are more likely to participate in events. So, this visual representation serves as a helpful tool that shows interesting relationships within our data and giving us hints for further analysis.   Machine Learning Techniques We applied two different machine learning models, Random Forest Classifier and Logistic Regression to our dataset in this stage of analysis. These models or techniques were chosen because they are appropriate for the properties of our data and the goals of our study. Logistic Regression Logistic regression is baseline model that works well for binary classification issues like our problem where we are predicting attendance (attended or not). It helps in recognizing the variables that have a noticeable impact on the outcome by understanding the relationships between the dependent and independent variables of dataset. Logistic regression is a great option for our research or project because of its transparency and ease of use, especially when making sense of the data is a primary priority. Random Forest Classifier Random Forest Classifier is a flexible ensemble learning technique or model. It is suitable for complicated data sets like ours that have many features. It can capture non-linear correlations and relationships between variables in dataset. That’s why we chose it for our project. Random Forest Classifier also have a feature that determines the feature importance of columns that how much a column helps in prediction. This model is also robust for overfitting. By using this in our projecte we have find that which column has highest impact on attendance patterns. K-Nearest Neighbors (KNN) we also used K-Nearest Neighbors (KNN) algorithm also for prediction of attendance based on member attributes such as weight, number of months of membership, and days before. KNN works based on compareign it with its closest neighbors. We evaluated its performance with other models, such as Random Forest Classifier and Logistic Regression. Also adjusted its parameters to increase accuracy. Grid Search CV Grid Search CV was used in our project to tune the hyperparameters of our machine learning models. Grid Search CV searched through a grid of hyperparameter values to find the best combination. Using this it increased the predicted accuracy of the models. Using this method prevented overfitting or underfitting and ensured that our models were optimized for optimal performance on our dataset. It also made hyperparameter tuning more efficient by auto cross-validation and parameter selection. Finally it enhanced the overall performance of our prediction models. Comparative Analysis of Machine Learning Models:Assessment of Logistic Regression: • **Accuracy:** This model has an accuracy of 0.75.  • **Sensitivity (Recall):** This model has a recall of 0.74.  • **F-Score:** There is 0.73 F-Score.     Random Forest Assessment: • **Accuracy**: This model has an accuracy of 0.81.  • **Sensitivity (Recall**): This model has a recall of 0.84.  • **F-Score**: This F-Score is 0.80.  A screenshot of a graph  Description automatically generated   Assessment of KNN • **Accuracy**: This model has an accuracy of 0.77.  • **Sensitivity (Recall**): This model has a recall of 0.81.  • **F-Score**: This F-Score is 0.77.    A screenshot of a computer screen  Description automatically generated Grid Search CV Assessment:  * **Best Parameters LR**: C=0.1 * **Best Accuracy LR**: 0.73 * **Best Parameters RFC**: max\_depth=20, min\_samples\_leaf=1, min\_samples\_split=2, n\_estimators=200 * **Best Accuracy RFC**: 0.78 * **Best Parameters KNN**: algorithm=auto, leaf\_size=20, n\_neighbors=9, weights=distance * **Best Accuracy KNN**: 0.76   When comparing machine learning models for our project, Logistic Regression demonstrated a modest level of accuracy (0.73 F-Score), striking a balance between precision and recall. With greater accuracy and an F-Score of 0.80, Random Forest outperformed Logistic Regression in terms of resilience and predictive capacity. KNN performed competitively, with an F-Score of 0.77 and an accuracy of 0.77, indicating its efficacy in predicting event attendance. Even though Random Forest had the best accuracy, the best model to use will depend on the particulars of the project and how recall and precision are traded off. Identifying Features Importance Searching out for the most important characteristics in the dataset is essential before matching study with business goals. This means that the assessment of each variable's significance and impact considering the given goals. After analyzing the factors influencing attendance patterns among fitness club members using the Random Forest classifier, some interesting insights emerged. It turns out that weight had the greatest impact on attendance, meaning that individuals with lower weights tended to participate more frequently in events. Also, we found that being a member for a long time was the second most important thing.   Conclusion Our study offers useful information to fitness centers which helps them to enhance member experiences and strategies of their clubs. We found that weight distribution and membership duration were important factors I predicting attendance. A comparative study of the models' performance also showed that tells that Random Forest outperformed Logistic Regression and KNN in terms of accuracy and F-Score. The significance of weight and membership duration in forecasting attendance patterns was highlighted. These insights enable clubs to plan events effectively and allocate resources wisely. Implementing member engagement techniques based on these insights can create a better club perfromance. This can lead to increased overall members happiness. |