Section A
Team 9: Yan Guan(yg238), Erik Henig(eh307), Jingjing Hu(jh892), Chuangfa Liang (cl670), Nupur Shah(ns417)

DECISION 520Q Team Project Final Report

I. Business Understanding

a) Identify, define, and motivate the business problem that you are addressing.

GoalZone is a fitness club chain in Canada. It offers a range of fitness classes in two capacities, 25 and 15. Some classes are always fully booked. However, fully booked classes often have a low attendance rate (DEE, 2023).

b) How (precisely) will a data mining solution address the business problem?

Data mining solution could help predict whether the member will attend the class so that GoalZone can make another space available if they predict a member will not attend the class. As a result, GoalZone could offer more class spaces to members who would attend. They also would gain insight into what classes are most popular and could use this to increase profits. They can also gain insight into the most popular times to optimize labor and reduce costs.

We hypothesize people are more inclined to attend a class when they see it is extremely popular (hard to register, always fully booked, & always fully attended). By carefully selecting class offerings and designing the class schedule to be filled, GoalZone could develop a competitive advantage and increase its brand recognition by creating trendy classes, meaning that they will be the only gym that clients go to for HIIT classes. Once they know when to schedule a class, the labor cost will be allocated correctly. The labor cost associated with all fitness classes will only be incurred when classes are full. There will be no empty classes and thus no unnecessary labor cost.

II. Data Understanding

a) Data description

i) This dataset includes data about GoalZone's members who registered a fitness class and data about the class that member has registered. Below are the detailed descriptions of the variables:

1) *booking_id*: a unique identifier for each booking record for the fitness class (numeric values from 1 to 1500)
2) *months_as_member*: the number of months of being GoalZone's member (numeric values in months, the minimum is 1 month)
3) *weight*: member's weight (numeric values in kilograms & "NA" represents missing values)
4) *days_before*: the number of days before the class when the member registered (a single numeric value in days or a categorical value in the format of "x days", x is a numeric value)
5) *day_of_week*: which day of a week is the class (categorical variable, from Monday to Sunday, representing in different formats)
   (a) Mon & Monday & Fri.
6) *time*: whether the class is AM or PM (categorical variable, representing as AM or PM)
7) *category*: the category of the fitness class (categorical variable, representing as Aqua, Cycling, HIIT, Strength, or Yoga, "-" represents missing values)
8) *attended*: the **outcome** of this dataset, whether the member attended the class or not. (binary variable, 0 = the member did not attend the class, 1 = the member attended the class).

ii) Since *booking_id* does not have an influence on a member's decision to attend the class, we decided to exclude this variable for the following analysis and modeling. All other variables are meaningful features to predict the outcome.

b) Data source: we obtained all of the data from Kaggle (DEE, 2023)

c) Potential bias

i) Weight is a continuous variable that will **fluctuate with time**. It could also be **self-reported**, leading to slight errors in the reported value.

ii) **Availability bias**: Some other variables (not listed in the current dataset), such as the number of kids and work hours each week, might also influence members' decision to attend the class. However, we will not be able to obtain that data.

iii) The majority of the values in the *attended* column are 0s (1046 out of 1500), implying that GoalZone might have **selection bias** while collecting the data by focusing on members who did not attend the class or on fully booked classes as they claimed that the issue they desire to solve is that fully booked classes often have a low attendance rate.

III. Data Preparation

To clean the data we evaluated both the unique values in each column and the data type. When evaluating the unique values in each column we were looking for issues pertaining to "NA" values or user indicated null values. When evaluating the data type for each column we were ensuring that columns reflected the data they contained. For instance, you would not want a column containing numbers to be stored as the categorical data type. Techniques such as data transformation and data imputations are implemented to address the two aspects of concerns mentioned above.

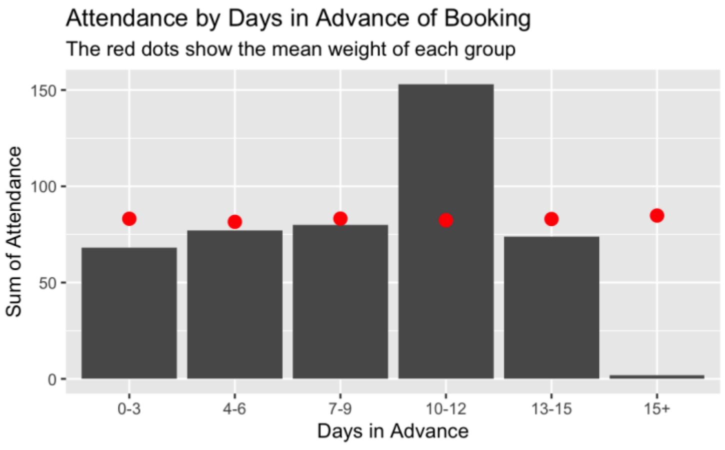
The first issue we resolved dealt with "NA" values in the *weight* column of our data. We alleviated this by finding the average of non-NA values of the *weight* columns and replacing the "NA" values with the calculated average. Mean imputation also allows us to preserve our data size, estimate a continuous variable, and is a very easily interpreted step. The next hurdle was to edit the *days_before* column which had numeric data stored as a string like "14 days". This resembled both a value error and data type error. We removed the string days from each entry and then converted this to a numeric value. The next error we dealt with was different data entries for days of the week (*day_of_week* column). We need to convert all "Wednesday" values to "Wed". Monday had a similar problem where we had to shorten "Monday" to "Mon". Finally, Friday had a period at the end of the string. We removed the period to preserve “Fri” as the entry here. The final error we had to remediate during data cleaning was removing all “-” from our *category* column. The *category* column is the type of workout class that is being held and options are strings like “Strength” or “Aqua”. To solve this problem, we found out which category was most popular and replaced the 4 dashes with “HIIT”. This method was chosen because of both how few dashes we had in the data and for its simplicity.

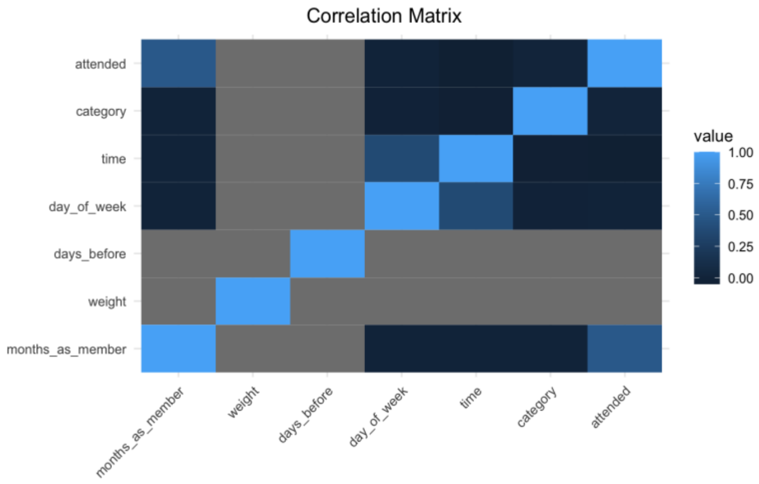
After making the above adjustments, we decided to further manipulate the data for an easier and smoother analysis process. We converted the categorical values in the *day_of_week*, *time*, *category* columns to numeric representations. For the *day_of_week* column, we randomly assigned a number between 0 to 6 to each day of the week to avoid any scaling effect. After this conversion, Monday is represented as 1, Tuesday is 5, Wednesday is 6, Thursday is 4, Friday is 0, Saturday is 2, and Sunday is 3. For the *time* column, we created dummy variables with 0 representing AM and 1 representing PM. For the *category* column, again, we randomly assigned a number between 0 to 4 to each category. After this conversion, Aqua is represented as 0, Cycling is 1, HIIT is 2, Strength is 3, and Yoga is 4.

Furthermore, to address the selection bias mentioned in Part I (the majority of the *attended* column is 0), we performed random sampling (oversampling in specific), which increases the sample size from 1500 to 2092, to combat the imbalance problem of the original dataset.

IV. Modeling

a) **Types of patterns mined:** Before modeling, we decided to conduct some preliminary analysis to better understand the data. The bar chart on the left shows that members who registered the class 10-12 days in advance of the class date are most likely to attend the class. We have also explored the correlation matrix for all the variables. We can see from the visualization on the right that the *attended* variable is quite highly correlated with the *months_as_member* variable.





b) Choices for data mining algorithms

i) NULL model: We did not expect the null model to be accurate, but it is always a good practice to include it to establish the baseline metric.

ii) Logistic Regression: Chosen for its simplicity and the fact that our target outcome is binary. It is easy to implement, interpret, and the algorithm can be regularized to avoid overfitting.

iii) CART (Classification Tree): Chosen for its interpretability and its ability to automatically detect and capture the non-linear relationships between variables. It doesn't assume anything about the underlying data distribution.

iv) Random Forest: it provides an improvement over simple Classification trees by reducing the variance. And we chose it because it runs all the probabilities of the combinations of variables.

c) Why & How this model solve the business problem

If the model can accurately predict if the person will attend or not, it brings a huge positive value to GoalZone. GoalZone can optimize its scheduling, avoid wasting resources on classes that members do not attend, and potentially offer more spots in popular classes, enhancing member satisfaction and maximizing profits.

i) Uncertainty: Attend (A) or Not Attend (!A)

ii) Expected profit maximization: E[Profit | X] = P(A | X) * q(X) * n - Cost

   1) P(A | X) is the probability that a member will attend the class

   2) q(X) is the revenue we could obtain from each customer

   3) n is the number of participants for each class

   4) Cost is the expense related to offering fitness classes

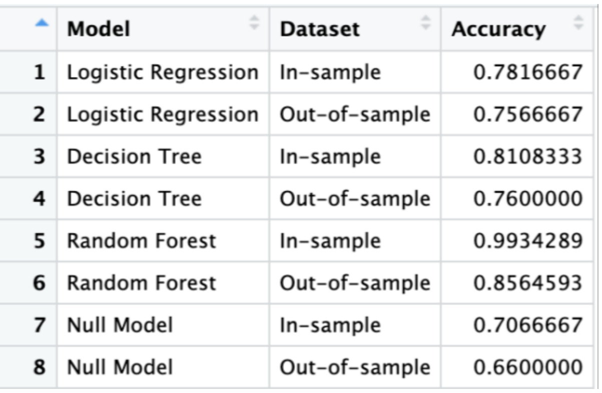
iii) Goal is to maximize E[Profit | X]: To maximize the profit, we should ensure the highest attendance rate for each of the class because the cost will stay consistent regardless of the number of participants.

V. Evaluation

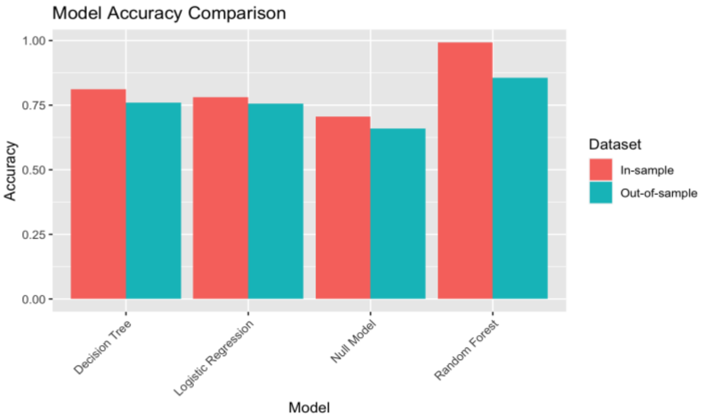
When evaluating prediction outputs from models it is important to consider the Out of Sample error of the model. We tested our outputs in two ways. First, we randomly splitted our dataset in 80% for training and 20% for testing, built the selected models with the training data and used it to forecast the testing dataset. Secondly, we use k-fold cross validation to further test the models separately. We used k=10 that divides the dataset into 10 pieces. And within each of the dataset pieces, we again divided it into a training part (80%) and a testing part (20%). Each dataset piece will generate an accuracy metric. And the average of these 10 accuracy metrics would be the final accuracy of the k-fold cross validation model.

The accuracy metric is calculated as below: (1) Generate the predictions for the outcome *attended* (will be either 0 or 1); (2) Compare the predictions with the actual outcomes & count the number of correct predictions (3) Divide the counts obtained in (2) by the total number of predictions to generate the accuracy. And the specific code to calculate the accuracy metric is: SUM(predicted_value == testing$value) / length.

The in-sample and out-of-sample accuracy for each model is displayed on the below table and graph. We can see that the baseline metric null model has an out-of-sample accuracy of 66.00 percent. And among all of the models that we have included in our analysis, random forest has the highest out-of-sample accuracy of 85.65 percent of accuracy. Therefore, we decided to ultimately use the random forest model for the following forecasting.



	Model	Dataset	Accuracy
1	Logistic Regression	In-sample	0.7816667
2	Logistic Regression	Out-of-sample	0.7566667
3	Decision Tree	In-sample	0.8108333
4	Decision Tree	Out-of-sample	0.7600000
5	Random Forest	In-sample	0.9934289
6	Random Forest	Out-of-sample	0.8564593
7	Null Model	In-sample	0.7066667
8	Null Model	Out-of-sample	0.6600000



VI. Deployment

a) How The Result of The Data Mining Will Be Deployed

Our prediction will allow the company to accurately predict when a class will be attended. Our goal would be to have employees seamlessly interact with the model. As customers sign up for a workout class, the model should be fed their data and employees would have real-time predictions. Employees could also specify a certain time they would want to view the report before the class. For example, employees could view the output for an entire class 24 hours before the class was scheduled. With prediction information, GoalZone can overbook classes to increase profit. Since we can predict which sign-ups will actually attend, we can allow extra people to attend the session. In the opposite direction, we can replace low attendance classes. When a class is predicted to have low attendance, we can either replace it with a more successful class or cancel it altogether. The effect of canceling a class will lower our labor cost by avoiding spending money on instructors.

Another deployment of this data could be to enhance attendance by reminding the specific people that we predict will not attend that they have committed to the class. Similarly, we can ask members if they would like to cancel their signup because we are predicting that they will not be in attendance.

Depending on how members are charged for classes also has an impact on deployment. If classes are included with a membership fee, we should optimize our labor cost by canceling low attendance classes. If classes are a surcharge on top of the membership fee, then GoalZone should explore scheduling more classes as they can collect fees for people who choose to not attend classes. If classes are only charged when customers attend, then GoalZone is very interested in hosting as many fully attended classes as possible.

b) Issues The Firm Should Be Aware of Regarding Deployment

We can only predict whether or not a person will show up once they sign up. This delays the timeline of prediction and therefore all other analysis.

Another issue with the model is that if things change, the model might not hold all relevant information. For example, if GoalZone starts charging for missed classes the behavior of customers might change. Similarly, if GoalZone alters subscription prices the behavior of customers might change. This model operates on the current demographic that is our customer base. If this demographic was to change based on macro environmental factors, the value of our predictions would likely change.

Another issue regarding deployment stems from the data collection problem we alleviated through data cleaning. Since some of the format of the data had to be edited, it is likely in production we will see similar data entry. This data would have to be cleaned before it could be processed in a model. However, the fees are charged, we will be able to more accurately estimate either revenues or costs with these predictions.

c) Important Ethical Consideration

An ethical consideration for this business case would be data collection and data privacy issues. Members who are resistant to data collection would represent an unpredictable entity. In our dataset, we have assumed that the customers are participating voluntarily. We also do not have access to their names, just an ID associated with each individual. This allows us to observe anonymity for the modeling. We also recommend transparency during communication with the customers on how their data is being used.

Along the same lines, we would have to consider data privacy issues. As more and more data is processed, lawmakers could potentially become more focused on the aggregation and usage of this data.

We should keep in mind confidentiality concerns as well. The data that is collected and used in this model should not be pushed to any other service without informed consent from the customers.

Our group also wants to highlight the importance of data accuracy and integrity. The data should not be altered in any way that changes the original meaning of the entry. We want to remove any data that doesn’t reflect the real world and remove any misinterpretations this might cause.

d) Risks Associated With Our Proposed Plan & How We Would Mitigate Them

A risk associated with our proposed plan deals with the issue of possibly overbooking a class. To mitigate this, we recommend a testing period for our model. With our prediction in mind, we can leave our classes open to the amount of gym equipment we have available. We want to live test this model without negatively affecting the class. A negative experience in the class would look like having too many people per lane in an Aqua class or too few bikes in a cycling class. A protocol we would recommend is establishing a class size that is 3 people less than the available equipment, and then fill those three extra spots based on our predictions.

Another mitigation strategy our team recommends is paying more attention to data collection. In production, data cleaning will bottleneck the processing speed of our predictions. This could be avoided up front if we feed the production model the same type of data it is predicting with. A user-friendly protocol would be to take our data format and create a matching Excel Spreadsheet. The important piece of this implementation is using the Data Validation section of Excel. Data Validation allows our team to only accept specific data such as things like “Fri.” instead of “Fri” would no longer be an issue.

If the model accuracy begins to decline, it would be especially important to retrain the model based on all historical data. Our team suggests retraining the model at least once a quarter to integrate all new data into our predictions.

Appendix

I. Group Contributions

Business Understanding	Erik, Yan
Data Understanding	Jinjing, Nupur
Data Preparation	Erik, Nupur
Modeling	Chuangfa, Jingjing
Evaluation	Chuangfa, Yan
Deployment	Erik, Jingjing

II. Bibliography

DEE, D. (2023, August 31). *Fitness Club dataset for ML Classification*. Fitness Club Dataset for ML Classification. https://www.kaggle.com/datasets/ddosad/datacamps-data-science-associate-certification