Fitness Class Forecast

Bo Baker

2023-03-01

Part 1):

Part a, b, c):

First, the dataset is read and the garbage values are checked in the dataframe. We found out that the days_before column has some values written as 8 days and some values written as just a number. Hence, str_replace_all function is used to clean this column. Same goes for day_of_week. There were some values written completely instead of 3 alphabets. They are re-coded. For missing observations in numerical columns, the missing observations are replaced with mean value and for categorical columns, missing observations are replaced with "unknown" value. After cleaning, all the data matched with the data description. The code being used throughout the process is attached below:

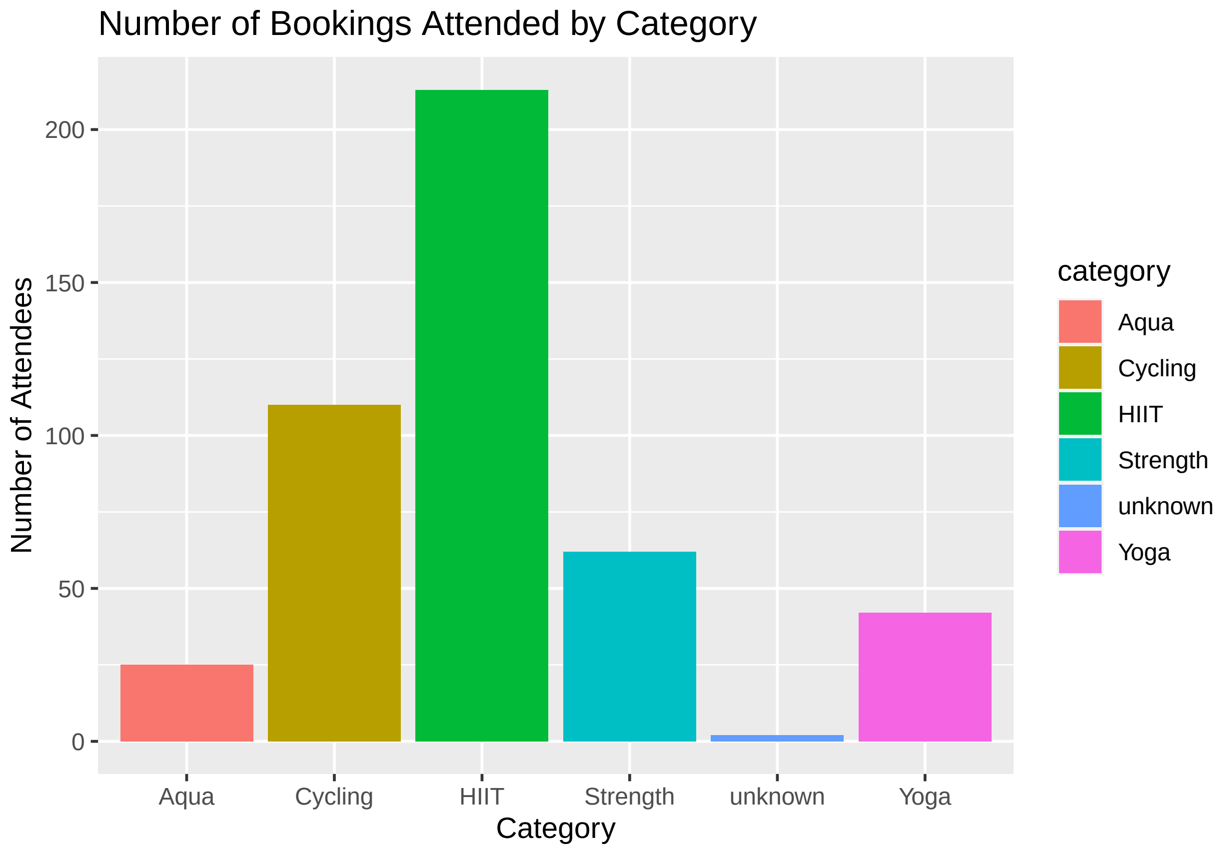
The final dataset after performing necessary preparation steps is seen below:

```
  booking_id months_as_member weight days_before day_of_week time category
1          1               17  79.56           8         Wed   PM Strength
2          2               10  79.01           2         Mon   AM     HIIT
3          3               16  74.53          14         Sun   AM Strength
4          4                5  86.12          10         Fri   AM  Cycling
5          5               15  69.29           8         Thu   AM     HIIT
6          6                7  93.33           2         Mon   AM  Cycling
  attended
1        0
2        0
3        0
4        0
5        0
6        0
```

Part 2):

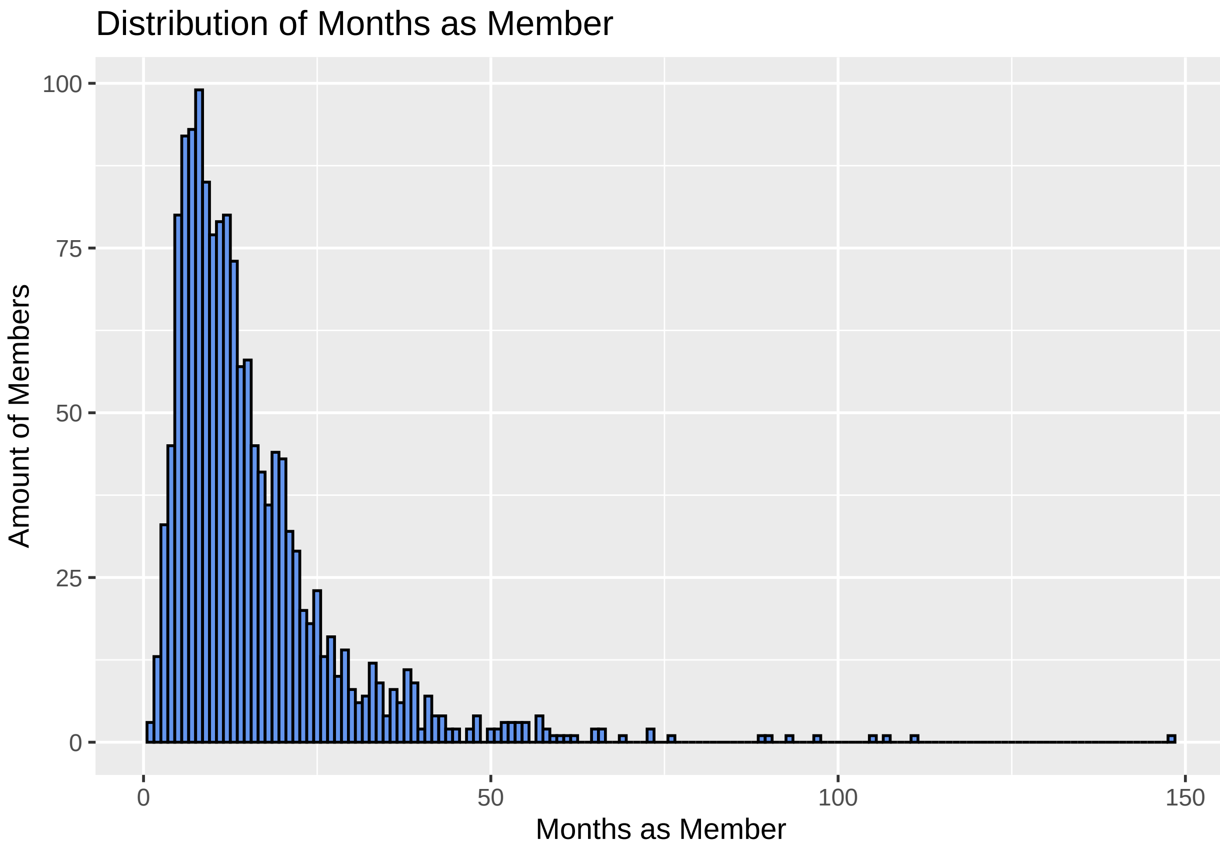
Part a & b):

HIIT has by far the most members attending its sessions. With the rest of the data we can see that it's not balanced either. HIIT is around a hundred observations higher than cycling which is also much higher than strength & yoga sessions with Aqua sessions having the least number of observations.



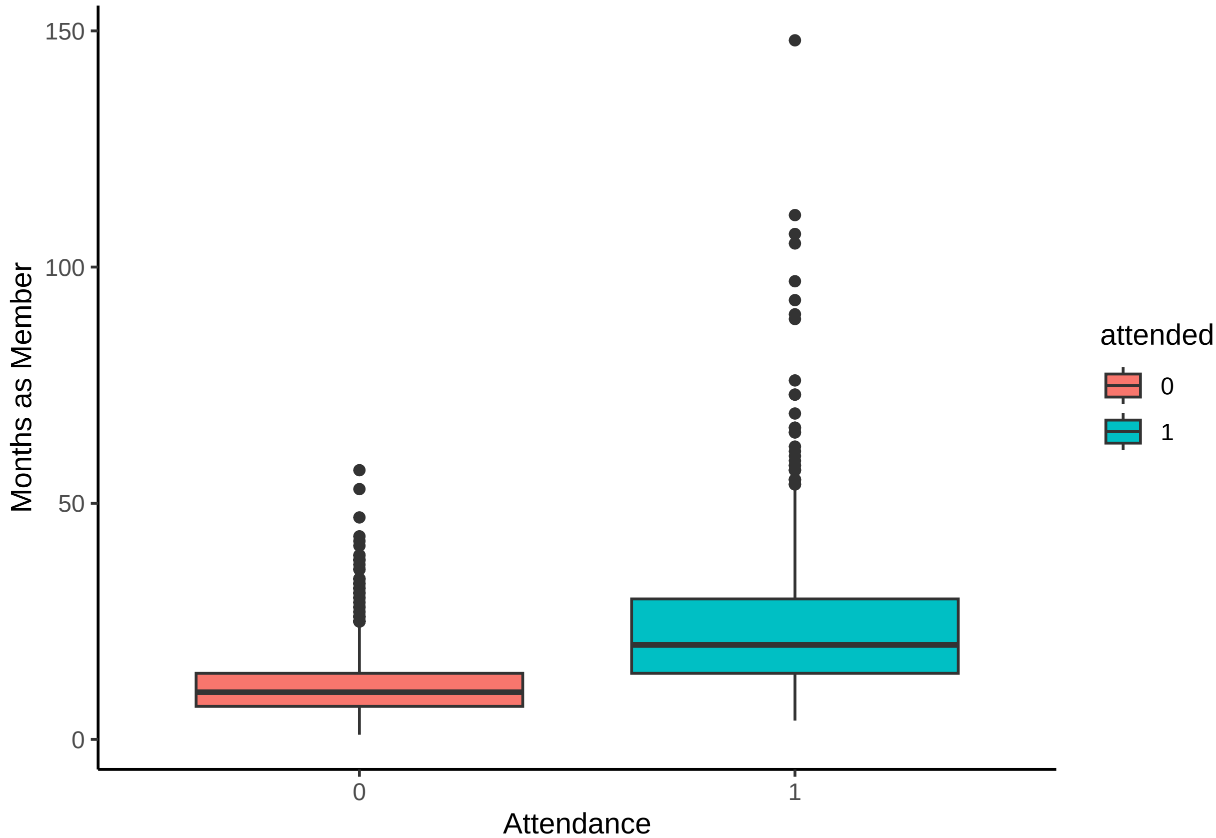
Part 3):

Below is a histogram which shows the distribution of 'Months as member'. It shows that a majority of members have around a dozen or so months as a member but very few have more than around 40 and very few have less than 10 or so.



Part 4):

From the boxplot seen below, we can see that the spread of those who attended is much higher than the spread of those who didn't attend. Similarly, we can see that the average months as a member are higher when the person is present and lower otherwise. Hence, we can say that people with higher months as members tends to be more present as compared to those having lower months as members.



Part 5):

As our target variable in this case is binary where the members are either present in the class or not. It can take only two values 0 and 1. Either the class is attended or not. Hence it is a binary classification problem.

Part 6):

First of all before building the baseline model, the categorical columns in the dataset are encoded, the booking ID column is useless and doesn't provide any useful information. Hence, that is removed from the dataset too.

```
  months_as_member weight days_before day_of_week time category attended
1               17  79.56           8           1    1        4        0
2               10  79.01           2           2    2        3        0
3               16  74.53          14           3    2        4        0
4                5  86.12          10           4    2        2        0
5               15  69.29           8           5    2        3        0
6                7  93.33           2           2    2        2        0
```

Next, the dataset is divided into training and testing sets. 75% of the data is used for training the model while the 25% of the data is set aside to test the performance of the model. The training set has 1125 observations while the testing set has 375 observations.

```
[1] 1125    7
```

```
[1] 375   7
```

The first baseline model that is built in this case is a logistic regression model with all available features as predictors and the summary of the model is seen below. It can be seen from the summary of the model that the predictors months as members and weight are statistically significant and has an impact on the target variable.

```
Call:
glm(formula = attended ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8253  -0.7124  -0.4951   0.6361   2.3192

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.80638    0.82987  -2.177   0.0295 *
months_as_member  0.12153    0.01056  11.512   <2e-16 ***
weight           -0.01585    0.00787  -2.014   0.0440 *
days_before       0.01504    0.01939   0.775   0.4381
day_of_week       0.01552    0.03554   0.437   0.6623
time              0.14626    0.18224   0.803   0.4222
category         -0.03275    0.06328  -0.518   0.6047
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1402.8  on 1124  degrees of freedom
Residual deviance: 1071.5  on 1118  degrees of freedom
AIC: 1085.5

Number of Fisher Scoring iterations: 5
```

Part 7):

Next, a comparison model called decision tree classifier is fitted to the dataset and the summary of the model is attached below. We can see that the top variables used in building the model are months as member, weight, days before and day of week. The variable with highest importance is months as member in this case. The lowest relative error is obtained for 7 splits.

```
Call:
rpart(formula = attended ~ ., data = train, method = "class")
  n= 1125

          CP nsplit rel error    xerror       xstd
1 0.25633803      0 1.0000000 1.0000000 0.04390914
2 0.01971831      1 0.7436620 0.7436620 0.04004045
3 0.01690141      4 0.6845070 0.7549296 0.04024882
4 0.01000000      7 0.6338028 0.7380282 0.03993479
```

```
Variable importance
months_as_member           weight      days_before      day_of_week
              74               21                3                2

Node number 1: 1125 observations,    complexity param=0.256338
  predicted class=0  expected loss=0.3155556  P(node) =1
    class counts:   770   355
   probabilities: 0.684 0.316
  left son=2 (716 obs) right son=3 (409 obs)
  Primary splits:
      months_as_member < 15.5   to the left,  improve=112.3751000, (0 missing)
      weight           < 75.12  to the right, improve= 41.2650900, (0 missing)
      days_before      < 13.5   to the left,  improve=  1.8401580, (0 missing)
      day_of_week      < 1.5    to the left,  improve=  1.8163970, (0 missing)
      time             < 1.5    to the left,  improve=  0.5735898, (0 missing)
  Surrogate splits:
      weight      < 75.065 to the right, agree=0.730, adj=0.257, (0 split)
      days_before < 1.5    to the right, agree=0.637, adj=0.002, (0 split)

Node number 2: 716 observations
  predicted class=0  expected loss=0.146648  P(node) =0.6364444
    class counts:   611   105
   probabilities: 0.853 0.147

Node number 3: 409 observations,    complexity param=0.01971831
  predicted class=1  expected loss=0.3887531  P(node) =0.3635556
    class counts:   159   250
   probabilities: 0.389 0.611
  left son=6 (267 obs) right son=7 (142 obs)
  Primary splits:
      months_as_member < 26.5   to the left,  improve=18.3995300, (0 missing)
      weight           < 74.545 to the right, improve= 7.1234310, (0 missing)
      days_before      < 10.5   to the left,  improve= 2.7733930, (0 missing)
      day_of_week      < 1.5    to the left,  improve= 2.4702300, (0 missing)
      category         < 2.5    to the left,  improve= 0.5369331, (0 missing)
  Surrogate splits:
      weight      < 61.125 to the right, agree=0.667, adj=0.042, (0 split)
      days_before < 1.5    to the right, agree=0.658, adj=0.014, (0 split)

Node number 6: 267 observations,    complexity param=0.01971831
  predicted class=1  expected loss=0.4981273  P(node) =0.2373333
    class counts:   133   134
   probabilities: 0.498 0.502
  left son=12 (20 obs) right son=13 (247 obs)
  Primary splits:
      day_of_week      < 1.5    to the left,  improve=2.7430670, (0 missing)
      weight           < 74.445 to the right, improve=2.6794070, (0 missing)
      days_before      < 12.5   to the left,  improve=2.2600320, (0 missing)
      months_as_member < 23.5   to the left,  improve=1.6283180, (0 missing)
      category         < 4.5    to the right, improve=0.5725075, (0 missing)

Node number 7: 142 observations
  predicted class=1  expected loss=0.1830986  P(node) =0.1262222
    class counts:    26   116
```

```
   probabilities: 0.183 0.817

Node number 12: 20 observations
  predicted class=0  expected loss=0.25  P(node) =0.01777778
    class counts:    15     5
   probabilities: 0.750 0.250

Node number 13: 247 observations,    complexity param=0.01971831
  predicted class=1  expected loss=0.4777328  P(node) =0.2195556
    class counts:   118   129
   probabilities: 0.478 0.522
  left son=26 (147 obs) right son=27 (100 obs)
  Primary splits:
      weight           < 74.445 to the right, improve=2.5866250, (0 missing)
      days_before      < 12.5   to the left,  improve=1.7574300, (0 missing)
      months_as_member < 23.5   to the left,  improve=1.4717850, (0 missing)
      category         < 1.5    to the right, improve=0.6659803, (0 missing)
      day_of_week      < 3.5    to the right, improve=0.6318723, (0 missing)

Node number 26: 147 observations,    complexity param=0.01690141
  predicted class=0  expected loss=0.462585  P(node) =0.1306667
    class counts:    79    68
   probabilities: 0.537 0.463
  left son=52 (26 obs) right son=53 (121 obs)
  Primary splits:
      weight           < 76.13  to the left,  improve=1.5156450, (0 missing)
      category         < 1.5    to the right, improve=1.3977880, (0 missing)
      days_before      < 9.5    to the right, improve=1.1301330, (0 missing)
      day_of_week      < 4.5    to the left,  improve=0.9412893, (0 missing)
      months_as_member < 22.5   to the left,  improve=0.8166405, (0 missing)

Node number 27: 100 observations
  predicted class=1  expected loss=0.39  P(node) =0.08888889
    class counts:    39    61
   probabilities: 0.390 0.610

Node number 52: 26 observations
  predicted class=0  expected loss=0.3076923  P(node) =0.02311111
    class counts:    18     8
   probabilities: 0.692 0.308

Node number 53: 121 observations,    complexity param=0.01690141
  predicted class=0  expected loss=0.4958678  P(node) =0.1075556
    class counts:    61    60
   probabilities: 0.504 0.496
  left son=106 (103 obs) right son=107 (18 obs)
  Primary splits:
      weight           < 78.43  to the right, improve=2.1668490, (0 missing)
      days_before      < 9.5    to the right, improve=1.8247790, (0 missing)
      day_of_week      < 4.5    to the left,  improve=0.9272510, (0 missing)
      months_as_member < 23.5   to the left,  improve=0.6841188, (0 missing)
      category         < 3.5    to the left,  improve=0.1089296, (0 missing)

Node number 106: 103 observations,    complexity param=0.01690141
```

```
  predicted class=0  expected loss=0.4563107  P(node) =0.09155556
    class counts:    56    47
   probabilities: 0.544 0.456
  left son=212 (55 obs) right son=213 (48 obs)
  Primary splits:
      days_before      < 9.5    to the right, improve=3.9302810, (0 missing)
      day_of_week      < 4.5    to the left,  improve=1.6767130, (0 missing)
      months_as_member < 19.5   to the left,  improve=0.6776744, (0 missing)
      weight           < 84.17  to the left,  improve=0.6776744, (0 missing)
      category         < 3.5    to the left,  improve=0.4694335, (0 missing)
  Surrogate splits:
      day_of_week      < 4.5    to the left,  agree=0.699, adj=0.354, (0 split)
      weight           < 85.625 to the right, agree=0.612, adj=0.167, (0 split)
      time             < 1.5    to the right, agree=0.573, adj=0.083, (0 split)
      months_as_member < 16.5   to the right, agree=0.553, adj=0.042, (0 split)
      category         < 5      to the right, agree=0.544, adj=0.021, (0 split)

Node number 107: 18 observations
  predicted class=1  expected loss=0.2777778  P(node) =0.016
    class counts:     5    13
   probabilities: 0.278 0.722

Node number 212: 55 observations
  predicted class=0  expected loss=0.3272727  P(node) =0.04888889
    class counts:    37    18
   probabilities: 0.673 0.327

Node number 213: 48 observations
  predicted class=1  expected loss=0.3958333  P(node) =0.04266667
    class counts:    19    29
   probabilities: 0.396 0.604
```

Part 8):

As our problem is a binary classification problem which comes under the umbrella of supervised machine learning, hence the logistic regression and decision tree classifiers are the two best performing machine learning algorithms which are used.

Part 9):

The performance evaluation metrics obtained for the logistic regression model on the unseen test dataset are attached below. We can see that the accuracy of the model is 79.73% with 76 miss-classified observations. The confidence interval of the model states that we are 95% confident that the accuracy of model lies between 75.3% and 83.69% respectively. The kappa statistics is 40.96%. The higher the kappa statistics, the better is the model performance.

```
Confusion Matrix and Statistics


pred_logreg   0   1
          0 256  56
          1  20  43
```

```
               Accuracy : 0.7973
                 95% CI : (0.753, 0.8369)
    No Information Rate : 0.736
    P-Value [Acc > NIR] : 0.003479

                  Kappa : 0.4096

 Mcnemar's Test P-Value : 5.95e-05

            Sensitivity : 0.9275
            Specificity : 0.4343
         Pos Pred Value : 0.8205
         Neg Pred Value : 0.6825
             Prevalence : 0.7360
         Detection Rate : 0.6827
   Detection Prevalence : 0.8320
      Balanced Accuracy : 0.6809

       'Positive' Class : 0
```

The performance evaluation metrics obtained for the decision tree model on the unseen test dataset are attached below. We can see that the accuracy of the model is 77.07% with 86 miss-classified observations. The confidence interval of the model states that we are 95% confident that the accuracy of model lies between 72.47% and 81.23% respectively. The kappa statistics is 41.74%. The higher the kappa statistics, the better is the model performance.

```
Confusion Matrix and Statistics


pred_dt   0   1
      0 231  41
      1  45  58

               Accuracy : 0.7707
                 95% CI : (0.7247, 0.8123)
    No Information Rate : 0.736
    P-Value [Acc > NIR] : 0.06999

                  Kappa : 0.4174

 Mcnemar's Test P-Value : 0.74632

            Sensitivity : 0.8370
            Specificity : 0.5859
         Pos Pred Value : 0.8493
         Neg Pred Value : 0.5631
             Prevalence : 0.7360
         Detection Rate : 0.6160
   Detection Prevalence : 0.7253
      Balanced Accuracy : 0.7114

       'Positive' Class : 0
```

Part 10):

Based on the accuracy, confidence interval and number of miss-classified observations, we observed that the values obtained for logistic regression were better compared to decision tree model. Hence, the better performing model in this case is the logistic regression model.