Group 2: Predicting Calorie Burn

Jessica Reyes, Seth Hillis, Niamh Corrigan, Andi Mellyn

2025-12-13

Dataset

We used the *Life Style Data* dataset published on Kaggle by Omar Essa. This dataset contains approximately 20,000 observations across 54 variables, providing sufficient variablity and complexity for a machine learning research project. We deliberately selected a dataset that would require some cleaning on our parts in order to engage with the entire data science and modeling pipeline, from exploratory visualization and data cleaning through feature selection and predictive modeling.

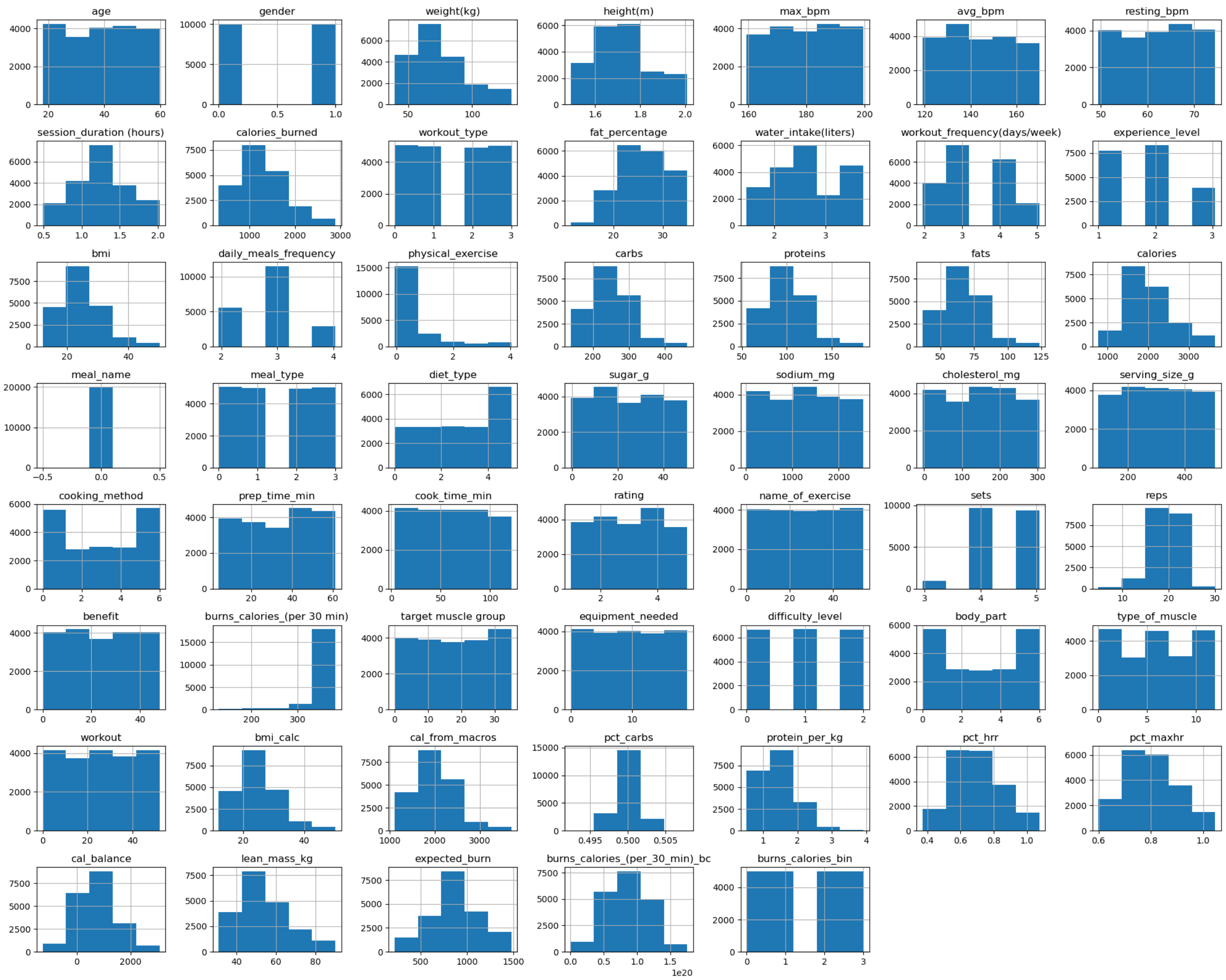
An additional motivation for selecting this dataset was its natural structure. The variables can be divided into two distinct feature groups, diet related variables and exercise related variables which would allow us to work with the data and run models on different parts for comparison. We were even able to choose our response variable as there were a couple different versions of calorie burn so the one we chose as the response for this project was "calorie burn in 30 minutes". Being able to divide the predictors into two groups allowed us to compare the predictive contributions of diet features, exercise features, and the combination of the two.

Key exercise-related predictors include session duration, heart rate (BPM) metrics, repetitions, sets, workout type, and experience level. Important diet-related predictors include carbohydrate, protein, fat, sugar, and sodium intake, as well as diet type and meal type. The dataset contains a mix of numerical and categorical variables, requiring appropriate preprocessing and encoding before modeling.

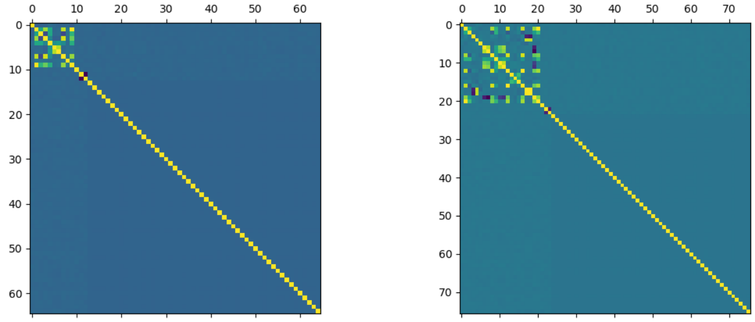
Data Preparation

We began the data preparation process with exploratory data analysis to understand the structure, scale, and behavior of the variables. We performed this using both Python and R, leveraging the complementary strengths of each environment as they bring different talents to the table. We examined the data through ranges, distributions and summary statistics with particular attention to variables we felt would play a meaningful role in predicting calorie burn.

To assess relationships within the data, we created several visualizations, including histograms and correlation plots. A primary histogram displayed multiple key variables side by side, allowing us to compare their distributions and identify skewness, outliers, and differences in scale.



In addition, a correlation plot provided insight into linear relationships among predictors as well as their associations with the response variable, calories burned in 30 minutes.



The variables that stood out after running the correlation plot:

Feature	Correlation with Burns_Calories_per_30_min
Burns_Calories_per_30_min	1.000

Feature	Correlation with Burns_Calories_per_30_min
Sets	0.465
Reps	0.343
lean_mass_kg	0.098
Weight_kg	0.091
Height_m	0.065
BMI	0.063
Fat_Percentage	0.054
Experience_Level	0.052
Workout_Frequency_days_per_week	0.044
Resting_BPM	0.038
Water_Intake_liters	0.026
Workout_Pull.ups	0.014
Difficulty_Level_Beginner	0.012
Workout_Crunches	0.011

Based on these exploratory analyses, we removed variables that appeared redundant or not directly relevant to the research question. We also converted data types as needed and encoded categorical variables to ensure compatibility with the modeling techniques under consideration. Multiple versions of the cleaned dataset were saved, as different models required different data representations and preprocessing steps.

We then prepared the data for modeling by splitting the dataset into training and testing sets using an 80/20 split. The models that we applied to this dataset included multiple linear regression, tree-based methods, and generalized additive models (GAMs), followed by cross-validation to assess out-of-sample performance.

Our motivation was to predict calorie burn based on a 30 minute workout session using progressively more flexible models that would allow us to compare predictive accuracy.

Research Question

What determines Calorie Burn? Do diet-related variables improve predictive performance beyond exercise variables alone?

Motivation

The goal of this project was to evaluate how well different classes of predictors and models can explain and predict calorie burn during a 30 minute workout session. In particular, we investigated whether diet related variables provide meaningful predictive information beyond exercise related features, and how model performance changes as we move from simple, interpretable models to more flexible machine learning approaches. To assess generalization and guard against overfitting, we compared models using 10-fold cross validation.

```
data <- read.csv("Data/new/data.csv")
```

Linear Regression

———- fill in ———

Generalized Additive Model (GAM)

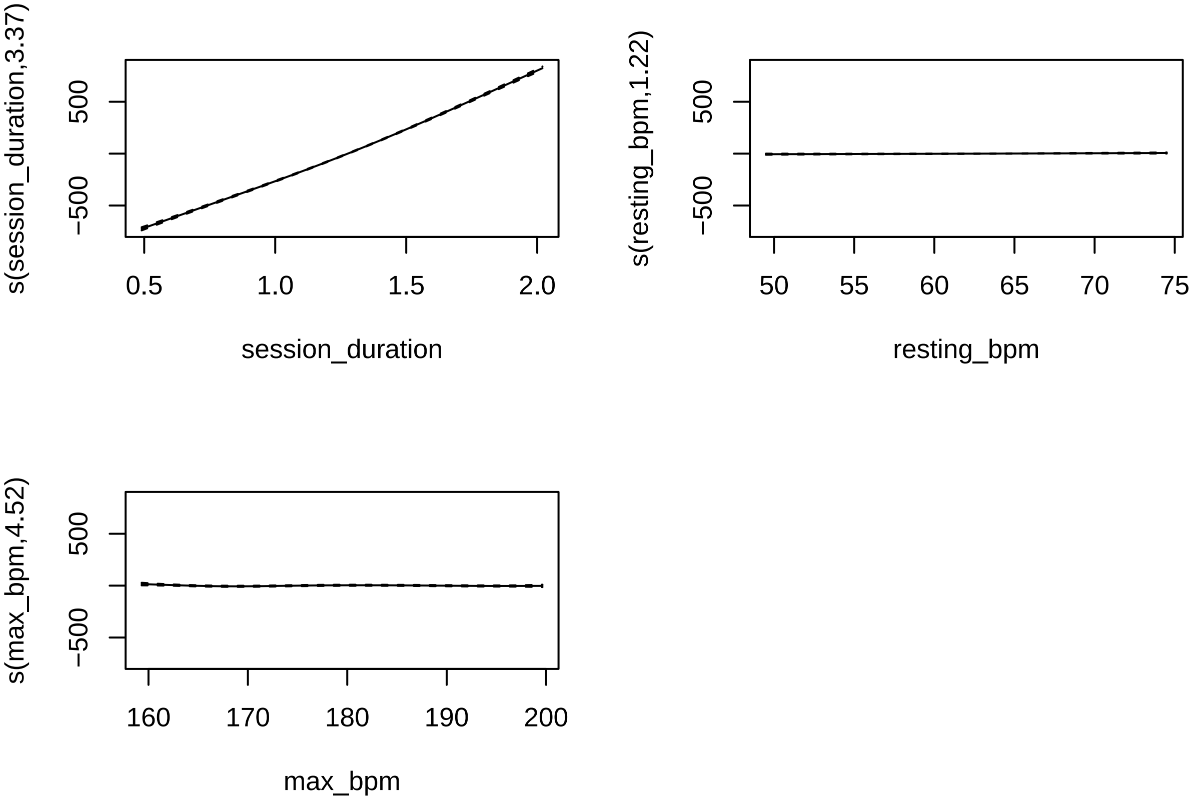
Exercise Model (GAM)

This model explains a substantial proportion of the variability in calories burned (84%), indicating that session duration, resting beats per minute, experience level, and workout type are important predictors.

```
data <- read.csv("~/project-2/Data/new/data.csv")
exercise_gam <- mgcv::gam(
  calories_burned ~
    s(session_duration) +
    s(resting_bpm) +
    s(max_bpm) +
    experience_level +
    workout_type,
  data = data,
  method = "GCV.Cp"
)

summary(exercise_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## calories_burned ~ s(session_duration) + s(resting_bpm) + s(max_bpm) +
##     experience_level + workout_type
##
## Parametric coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1345.000      6.215   216.4   <2e-16 ***
## experience_level  112.255      3.171    35.4   <2e-16 ***
## workout_type     -179.155      1.265  -141.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                        edf Ref.df        F p-value
## s(session_duration) 3.367  4.178 5703.610  <2e-16 ***
## s(resting_bpm)      1.215  1.398    4.833  0.0294 *
## s(max_bpm)          4.525  5.549    1.960  0.0665 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.84   Deviance explained =   84%
## GCV =  40328  Scale est. = 40303     n = 20000
```

```
plot(exercise_gam, pages = 1)
```



Smooth Terms form GAM Model 1

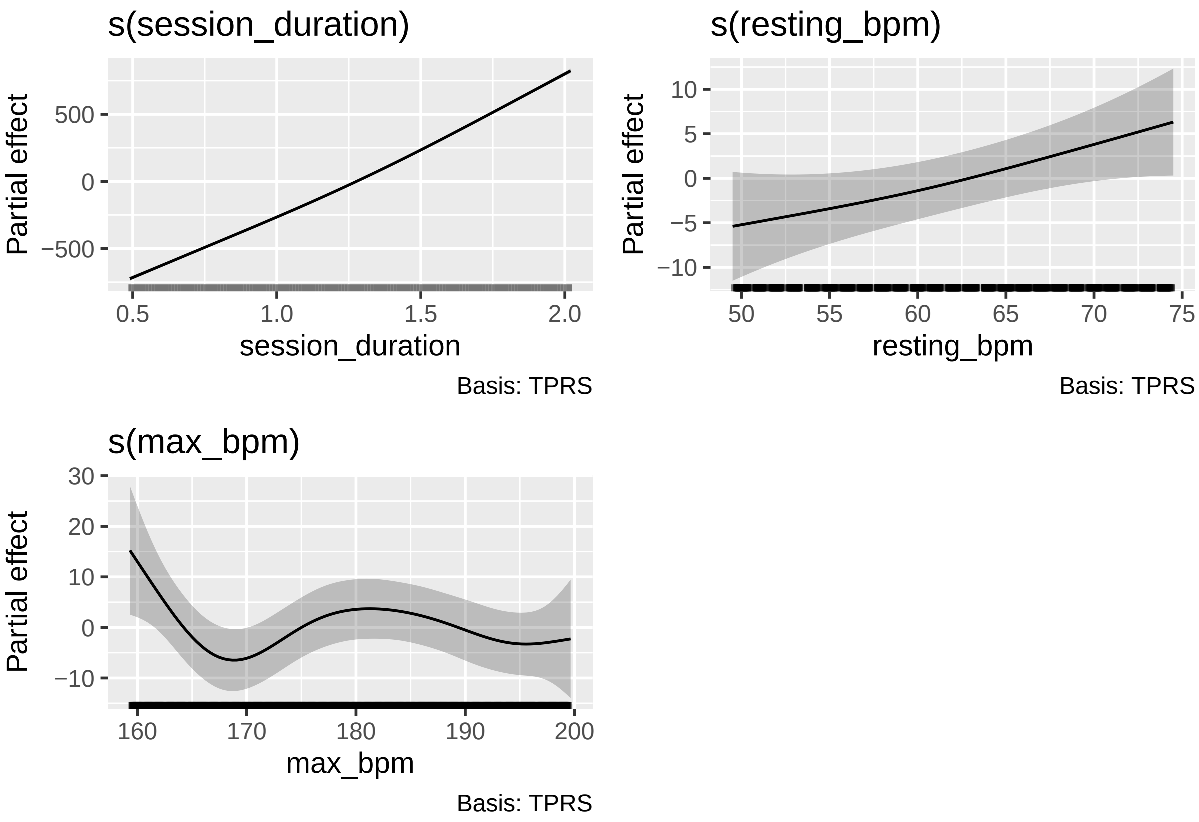
As expected, session duration has a strong, nearly linear positive effect on calories burned. Longer sessions consistently result in more calories burned, showing a positive relationship.

Resting BPM shows a moderate positive relationship with calories burned. As resting BPM increases, calories burned also tend to increase, though the effect is less pronounced than for session duration.

Maximum BPM demonstrates a complex non-linear effect with fluctuations across its range.

These patterns highlight the importance of considering non-linear relationships in understanding how physiological and workout factors influence calories burned.

```
draw(exercise_gam)
```



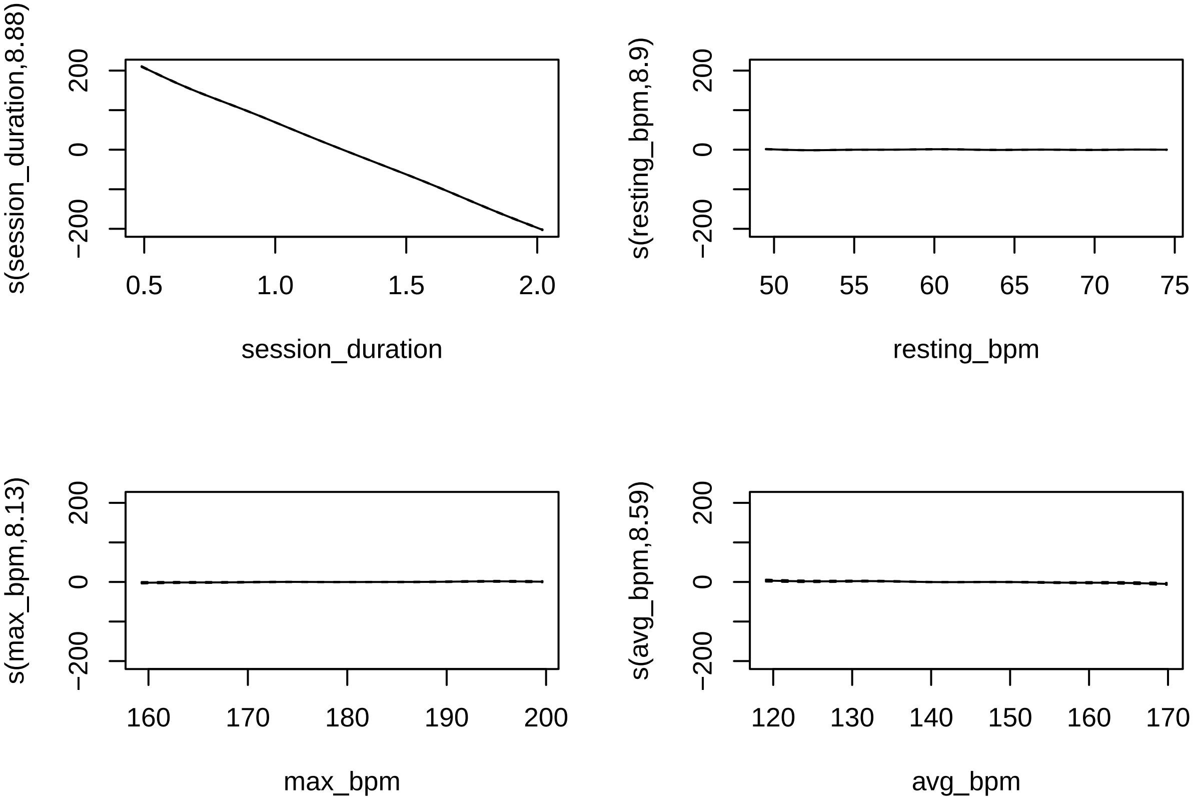
Generalized Additive Model (GAM)

Exercise Model with Calories Burned 30-minute as the Response Variable

```
data <- read.csv("~/project-2/Data/new/data.csv")
exercise_gam2 <- mgcv::gam(
  caloriesBurned_per30 ~
    s(session_duration) +
    s(resting_bpm) +
    s(max_bpm) +
    s(avg_bpm) +
    experience_level +
    expected_burn +
    pct_maxhr +
    workout_frequency +
    workout_type,
  data = data,
  method = "GCV.Cp"
)

summary(exercise_gam2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## caloriesBurned_per30 ~ s(session_duration) + s(resting_bpm) +
##     s(max_bpm) + s(avg_bpm) + experience_level + expected_burn +
##     pct_maxhr + workout_frequency + workout_type
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.1719198  6.9437571  -0.457   0.6478
## experience_level   0.3098357  0.1444273   2.145   0.0319 *
## expected_burn      0.3820983  0.0005409 706.394   <2e-16 ***
## pct_maxhr         19.5739384  8.6326410   2.267   0.0234 *
## workout_frequency -0.0333670  0.0888324  -0.376   0.7072
## workout_type       0.0218138  0.0387177   0.563   0.5732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df         F p-value
## s(session_duration) 8.883  8.996 47028.899  <2e-16 ***
## s(resting_bpm)      8.899  8.997    19.929  <2e-16 ***
## s(max_bpm)          8.130  8.794     6.579  <2e-16 ***
## s(avg_bpm)          8.587  8.951    14.149  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.963   Deviance explained = 96.4%
## GCV = 37.794  Scale est. = 37.717    n = 20000
```

```r
plot(exercise_gam2, pages = 1)
```



Smooth Terms form GAM Model 2

The second model extends the initial GAM by modeling calories burned per 30 minutes rather than total calories, and by incorporating additional variables. This allows for more standardized and interpretable outcome across workouts of different lengths.

Model 2 shows a substantial improvement in explanatory power compared to Model 1. The adjusted R^2 increases from 0.84 to 0.96, and deviance explained rises from 84% to 96.4%, indicating that the additional predictors are refined response variables capture much more of the variability in caloric expenditure. The lower GCV score also suggests a better model fit.

While Model 1 effectively captures total caloric expenditure driven largely by session duration, Model 2 provides a more detailed and physiologically meaningful explanation of calorie burn rate. The inclusion of average heart rate, expected burn, and intensity measures allows the model to uncover complex non-linear patterns and delivers a stronger fit. Overall, Model 2 offers a stronger and more interpretable framework for understanding how fitness, intenstiy, and cardiovascular response shape caloric efficiency during workouts.

Generalized Additive Model (GAM)

Exercise & Diet Model

Including diet type in the model yields a statistically significant but very small positive effect on calories burned per 30 minutes.

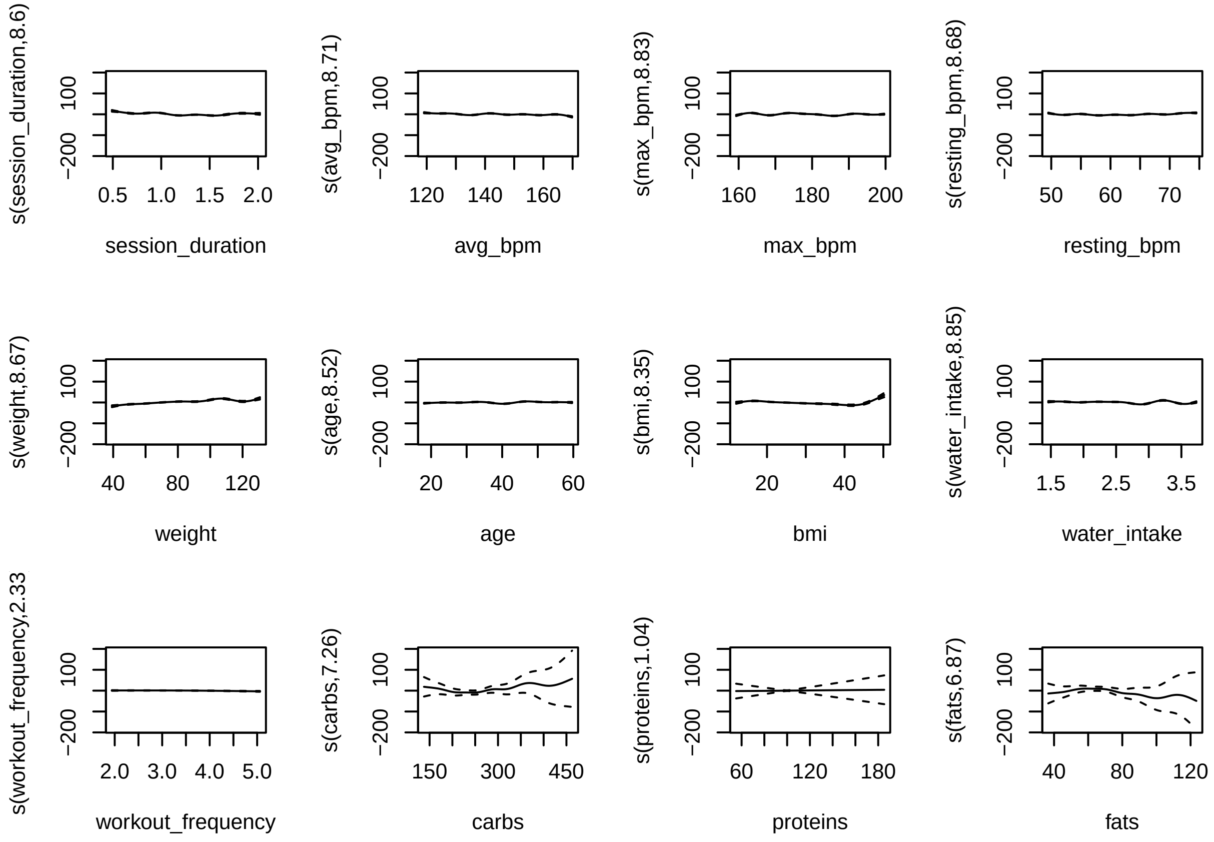
Despite this significance, overall model fit remains unchanged, indicating that diet type contributes minimal additional explanatory power beyond established predictors such as session duration, heart rate measures, experience level, and workout characteristics.

A separate comparison model using total calories burned was also evaluated; however, because the results were identical, it was not included.

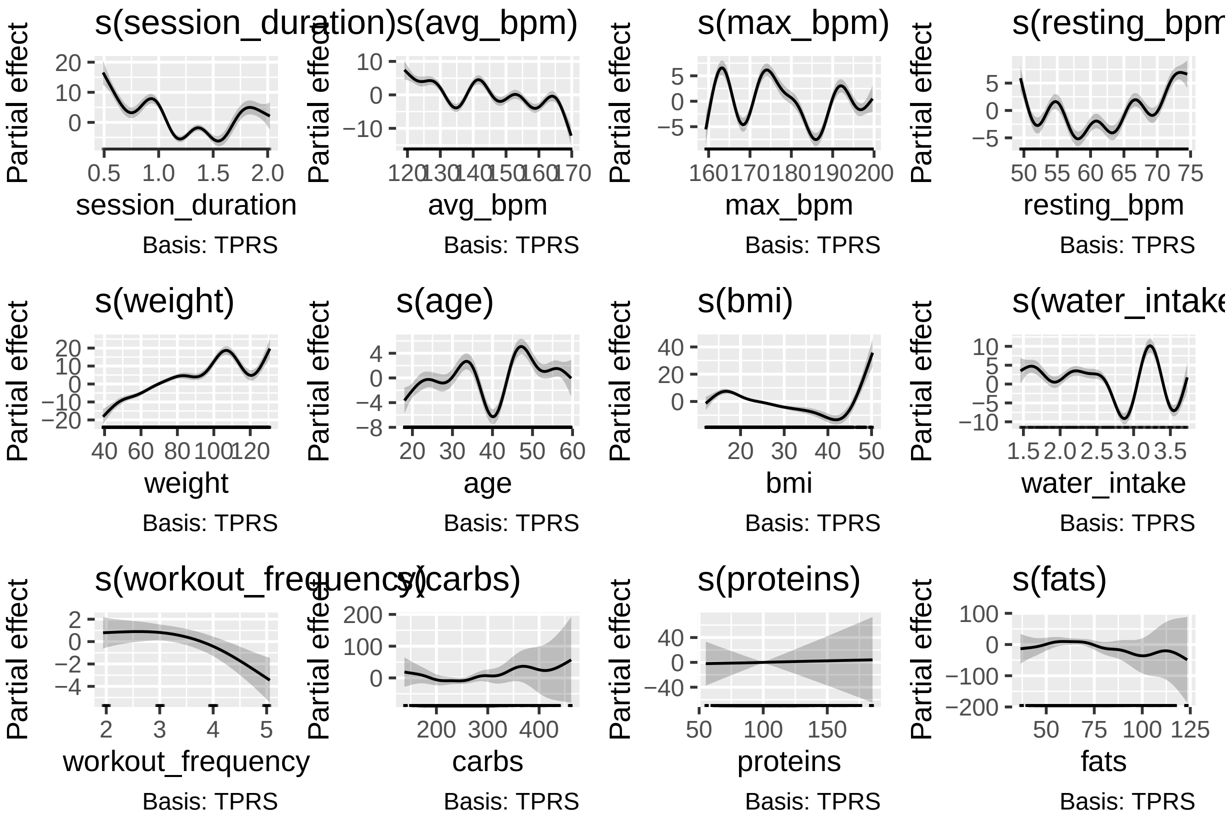
```r
combined_gam3 <- mgcv::gam(
  caloriesBurned_per30 ~
    s(session_duration) +
    s(avg_bpm) +
    s(max_bpm) +
    s(resting_bpm) +
    s(weight) +
    s(age) +
    s(bmi) +
    experience_level +
    workout_type +
    s(water_intake) +
    s(workout_frequency) +
    daily_meals_frequency +
    s(carbs) +
    s(proteins) +
    s(fats) +
    diet_type,
  data = data,
  method = "REML"
)


summary(combined_gam3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## caloriesBurned_per30 ~ s(session_duration) + s(avg_bpm) + s(max_bpm) +
##     s(resting_bpm) + s(weight) + s(age) + s(bmi) + experience_level +
##     workout_type + s(water_intake) + s(workout_frequency) + daily_meals_frequency +
##     s(carbs) + s(proteins) + s(fats) + diet_type
##
## Parametric coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           325.2611     1.7691 183.859  < 2e-16 ***
## experience_level        6.9894     0.7437   9.399  < 2e-16 ***
## workout_type           -0.2448     0.1924  -1.272   0.2034
## daily_meals_frequency   2.0501     0.3548   5.779 7.64e-09 ***
## diet_type               0.2544     0.1271   2.001   0.0455 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                         edf Ref.df      F p-value
## s(session_duration)  8.595  8.951 46.833 < 2e-16 ***
## s(avg_bpm)           8.712  8.975 25.988 < 2e-16 ***
## s(max_bpm)           8.830  8.991 31.698 < 2e-16 ***
## s(resting_bpm)       8.677  8.969 27.227 < 2e-16 ***
## s(weight)            8.670  8.967 34.791 < 2e-16 ***
## s(age)               8.518  8.933 18.355 < 2e-16 ***
## s(bmi)               8.348  8.872 23.438 < 2e-16 ***
## s(water_intake)      8.851  8.993 30.479 < 2e-16 ***
## s(workout_frequency) 2.333  2.736  4.382 0.00518 **
## s(carbs)             7.256  7.973  0.955 0.44883
## s(proteins)          1.045  1.083  0.014 0.96534
## s(fats)              6.866  7.748  0.714 0.67358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.101   Deviance explained = 10.5%
## -REML =  96885  Scale est. = 928.69    n = 20000
```

```
plot(combined_gam3, pages = 1, se = TRUE)
```



```
draw(combined_gam3)
```



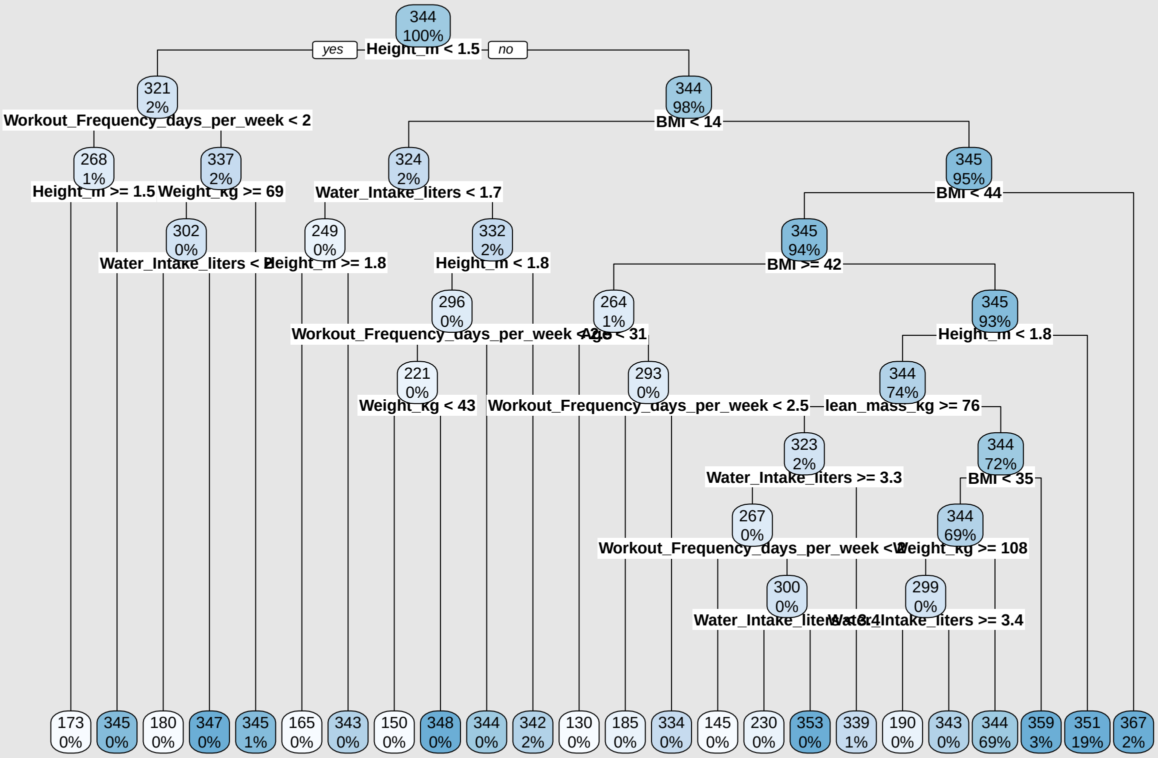
Regression Tree / Random Forest

I used tree-based regression to model **Calories Burned per 30 Minutes** against demographic and physiological predictors. These methods are well suited for fitness data, where relationships between variables are often nonlinear and involve interactions that are difficult to specify in advance.

Regression Tree

I fit a regression tree using the `rpart` package, treating **Burns_Calories_per_30_min** as the response variable. The algorithm partitioned the predictor space into disjoint regions by recursively selecting splits that minimized within-node variance. I controlled model complexity by setting a complexity parameter of **cp = 0.01** and a minimum split size of **minsplit = 20**. These constraints prevented the tree from overfitting and ensured that each split represented a meaningful number of observations.

```
library(rpart)
library(randomForest)
library(rpart.plot)
data_factor <- read.csv("./Data/clean/CleanLifeDataFactor.csv")

n <- nrow(data_factor)
train_idx <- sample(seq_len(n), size = 0.8 * n)

train <- data_factor[train_idx, ]
test  <- data_factor[-train_idx, ]
target <- "Burns_Calories_per_30_min"

tree_model <- rpart(
  Burns_Calories_per_30_min ~ .,
  data   = train,
  method = "anova",    # regression tree
  control = rpart.control(
    cp = 0.01,         # complexity parameter
    minsplit = 20
  )
)

rpart.plot(tree_model,cex = .5)
```



The regression tree also allowed me to calculate feature importance scores, I primarily used these importance scores as a diagnostic tool rather than as a basis for causal interpretation. These rankings identify which physiological variables contribute most to predictive accuracy.

```r
sort(tree_model$variable.importance, decreasing = TRUE)

##                             BMI                             Age
##                       2772206.8                       1926967.4
##                       Weight_kg             Water_Intake_liters
##                       1900895.4                       1767904.8
##                        Height_m                         Workout
##                       1666422.7                       1420435.3
## Workout_Frequency_days_per_week                    lean_mass_kg
##                       1156273.6                        845081.0
##                Experience_Level                               X
##                        565764.7                        197033.3
##                  Fat_Percentage                          Gender
##                        115707.9                         99865.9
```

The resulting tree structure showed that workout intensity and body composition dominate the primary nodes. These upper-level splits indicate that these factors exert the strongest influence on total calorie expenditure. Furthermore, the hierarchical branches reveal clear interaction effects; for instance, the impact of age on calorie burn changes significantly depending on the individual's heart rate. While this rule-based structure provides excellent interpretability, its greedy splitting mechanism makes it sensitive to minor fluctuations in the training data.

Random Forest

To stabilize these predictions, I implemented a random forest via the `randomForest` package. This ensemble approach constructs multiple decision trees using bootstrap samples and a random subset of predictors at each split. By averaging the results across the entire forest, the model reduces the high variance typical of a single tree while maintaining the ability to capture complex data patterns.

```r
p <- ncol(train) - 1

set.seed(123)
rf_model <- readRDS("Data/models/random_forest_model.rds")

#rf_model <- randomForest(
#  Burns_Calories_per_30_min ~ .,
#  data  = train,
#  ntree = 500,
#  mtry  = floor(sqrt(p)),
#  importance = TRUE
#)
#saveRDS(rf_model, file = "Data/models/random_forest_model.rds")

rf_model

##
## Call:
##  randomForest(formula = Burns_Calories_per_30_min ~ ., data = train,      ntree = 500, mtry = floor(
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 103.0987
##                     % Var explained: 90.2
```

Ultimately, the random forest provides a much more robust framework for predicting calorie burn across diverse exercise conditions, even if it sacrifices the immediate visual simplicity of a single tree.

Cross Validation

——— fill in ———

Conclusion

——— fill in ———

References

Essa, Omar. (n.d.) *Life Style Data* [Data set]. Available from Kaggle, Website: https://www.kaggle.com/datasets/jockeroika/life-style-data