Group 2: Predicting Calorie Burn

Jessica Reyes, Seth Hillis, Niamh Corrigan, Andi Mellyn

2025-12-02

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.

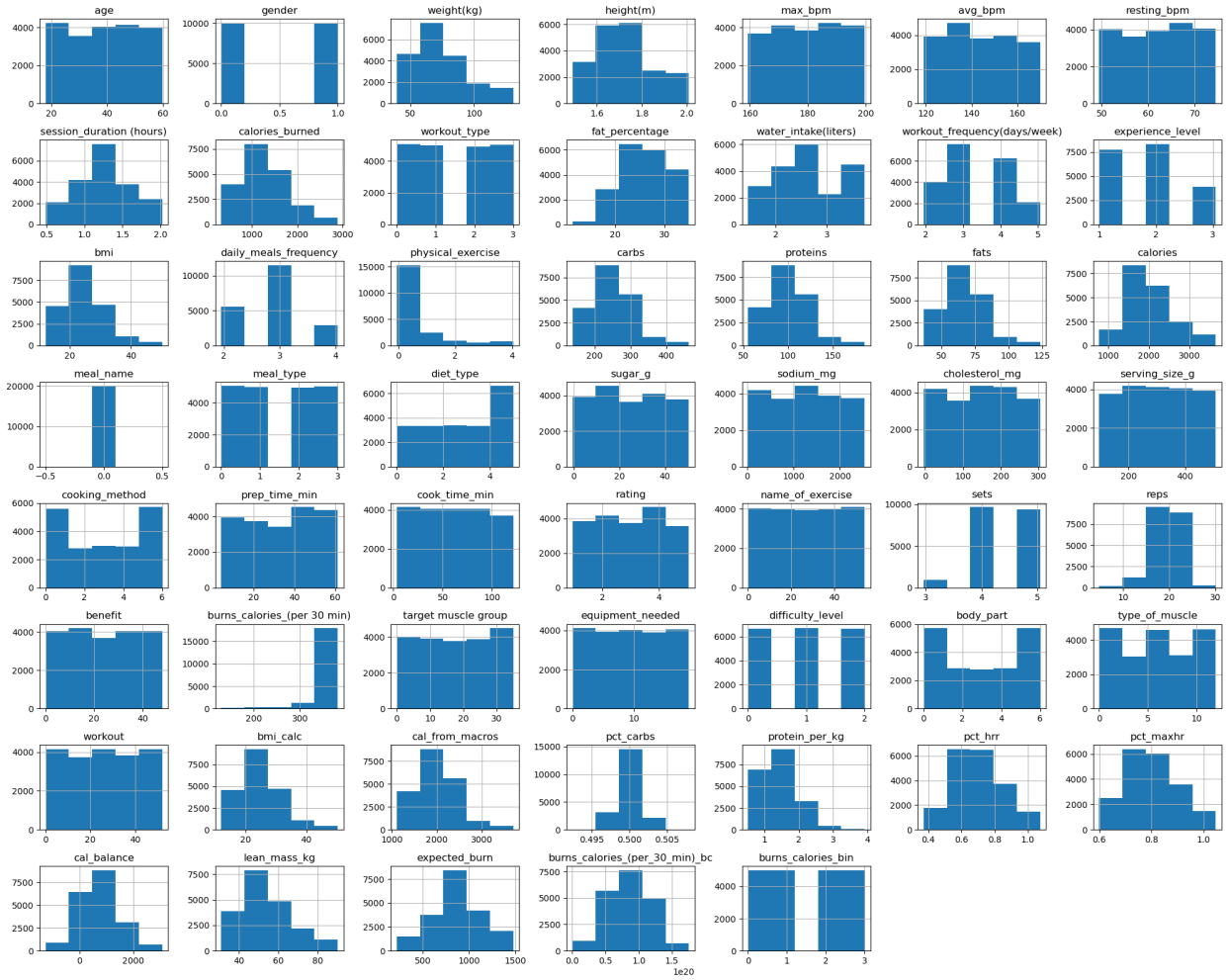
An additional motivation for selecting this dataset was its natural structure. The variables can be divided into two distinct feature groups, diet related variables and exercise related variables which would allow us to work with the data and run models on different parts for comparison. We were even 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.

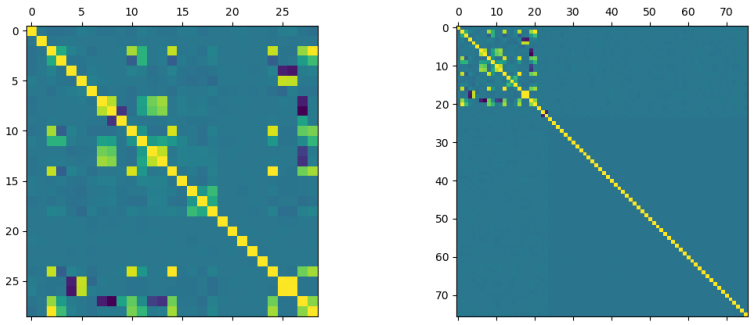
Data Preparation

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)

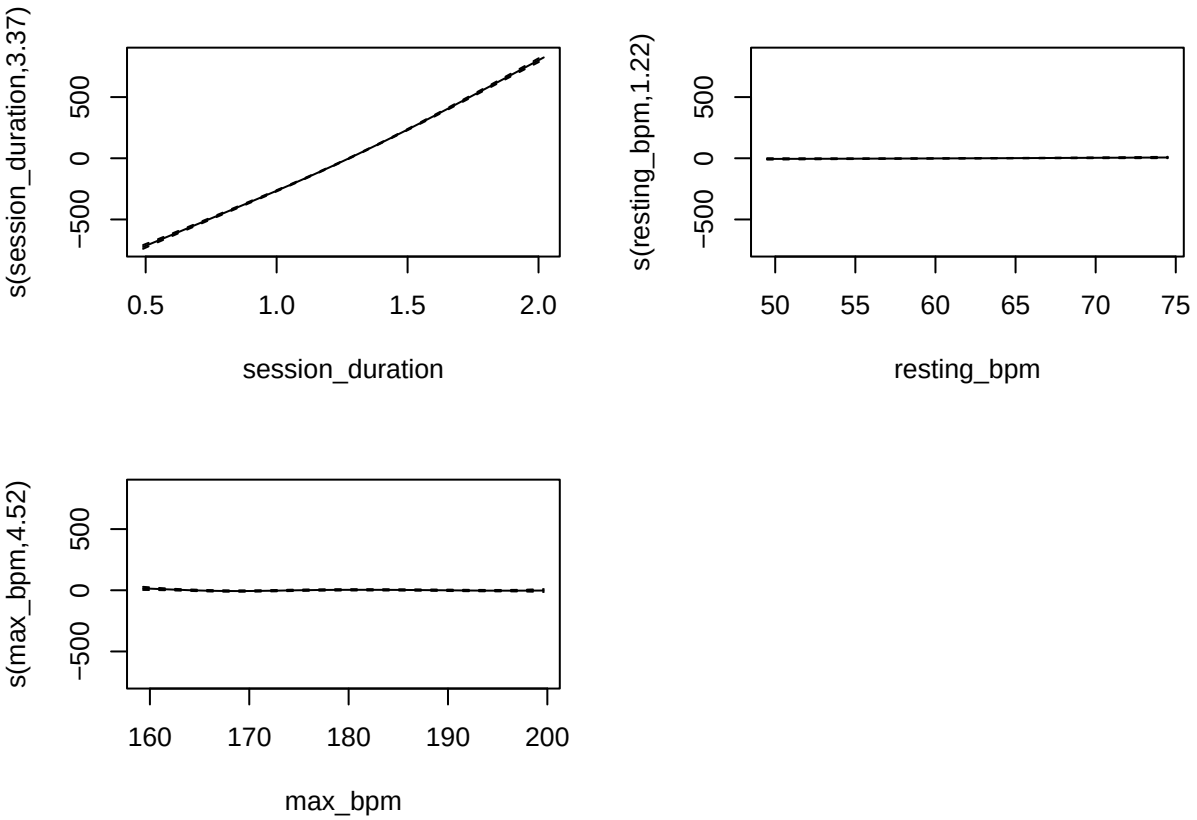
This model explains a substantial proportion of the variability in calories burned (84%), indicating that session duration, resting beats per minute, experience level, and workout type are important predictors.

```
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

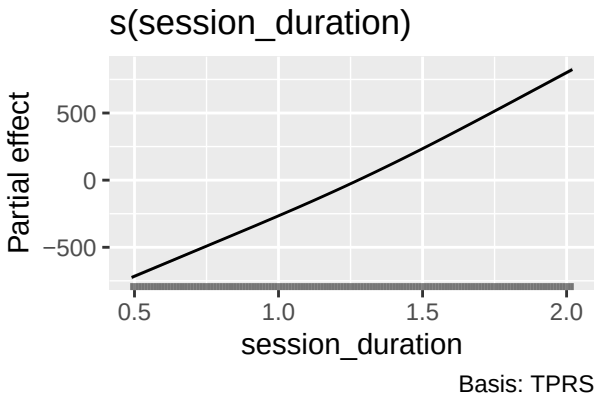
As expected, session duration has a strong, nearly linear positive effect on calories burned. Longer sessions consistently result in more calories burned, showing a positive relationship.

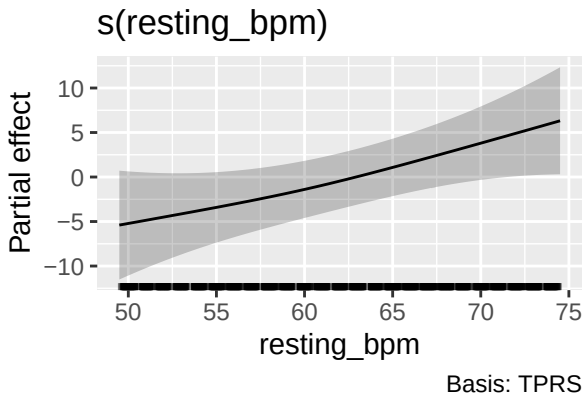
Resting BPM shows a moderate positive relationship with calories burned. As resting BPM increases, calories burned also tend to increase, though the effect is less pronounced than for session duration.

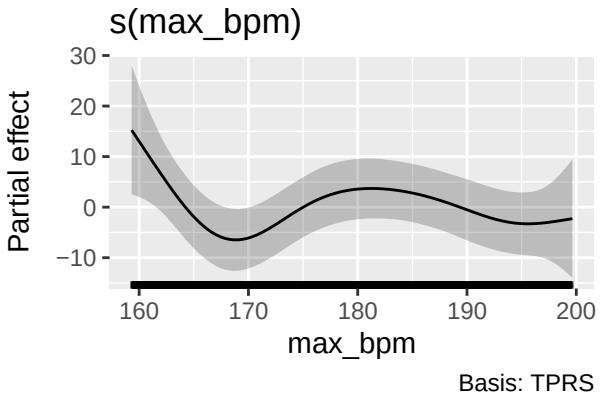
Maximum BPM demonstrates a complex non-linear effect with fluctuations across its range.

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

```
draw(exercise_gam)
```







Exercise & Diet Model (GAM)

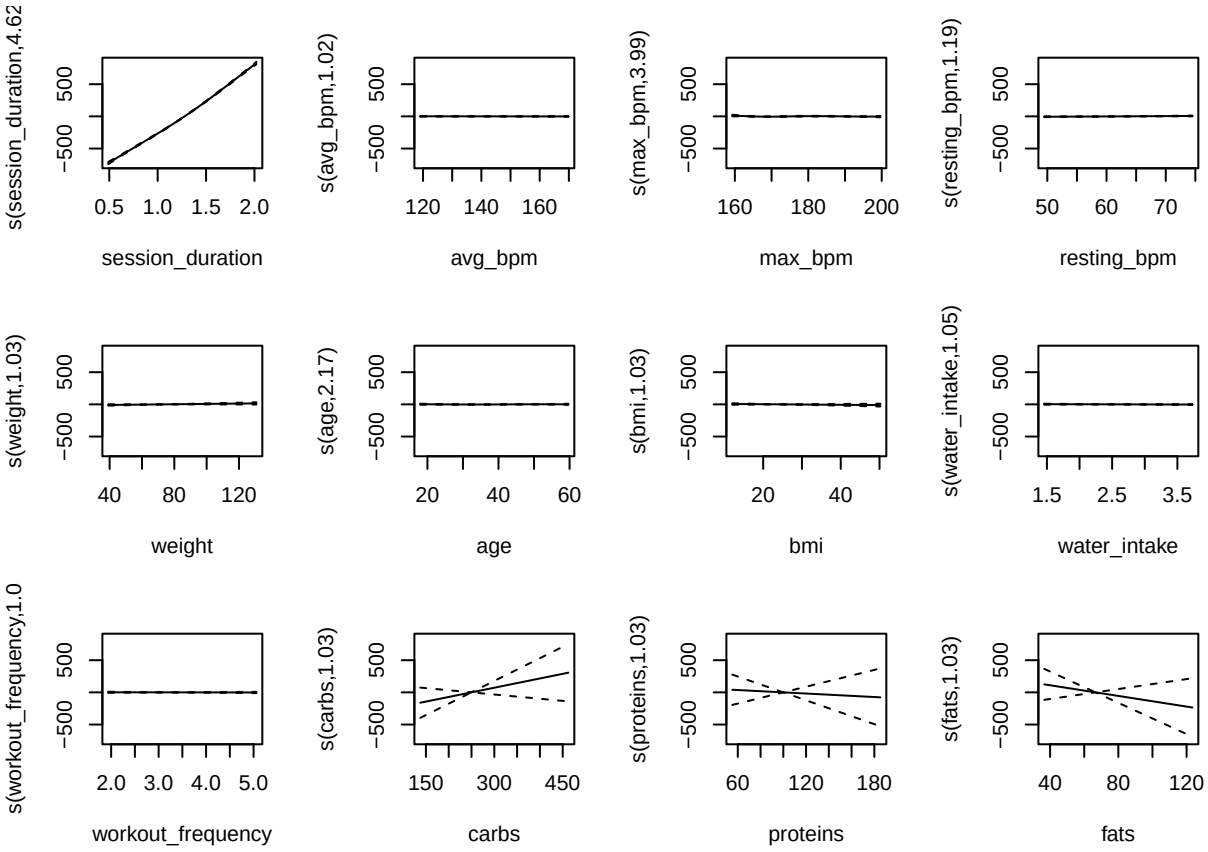
Including diet type in the model reveals a statistically significant but very small positive effect on calories burned. However, the overall model fit remains unchanged, suggesting that diet type does not substantially improve the model's ability to explain variability in calories burned beyond session duration, heart rate measures, experience level, and workout type.

```
combined_gam2 <- mgcv::gam(
  calories_burned ~
    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) +
```

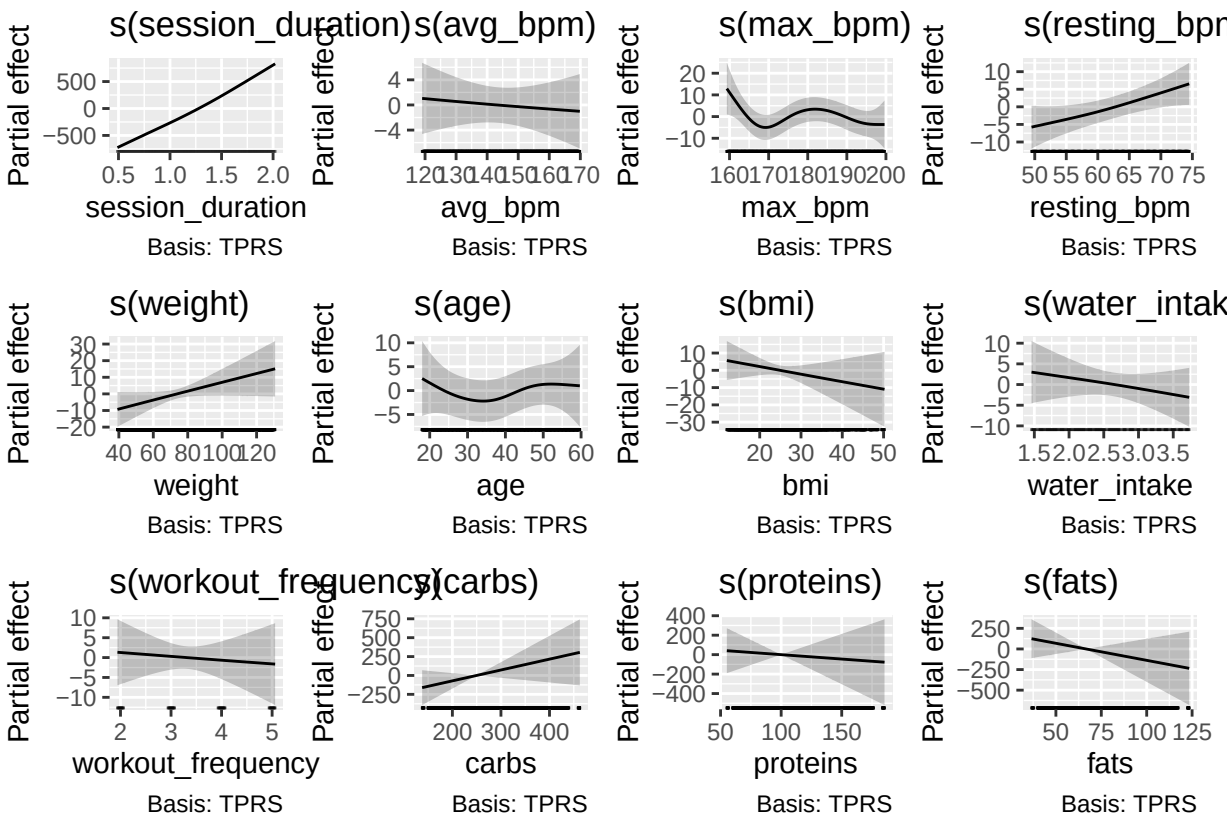
```
    diet_type,
  data = data,
  method = "REML"
)

summary(combined_gam2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## calories_burned ~ 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)           1333.4411    10.9645  121.615   <2e-16 ***
## experience_level       112.7252     4.4637   25.254   <2e-16 ***
## workout_type          -179.1810     1.2654 -141.601   <2e-16 ***
## daily_meals_frequency    2.2483     2.2661    0.992   0.3211
## diet_type                1.7295     0.8358    2.069   0.0385 *
## ---
## 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)  4.617  5.666 4183.865  <2e-16 ***
## s(avg_bpm)           1.023  1.046    0.142  0.7170
## s(max_bpm)           3.989  4.919    1.978  0.0940 .
## s(resting_bpm)       1.194  1.361    5.293  0.0218 *
## s(weight)            1.035  1.068    3.039  0.0755 .
## s(age)               2.171  2.706    0.811  0.6167
## s(bmi)               1.025  1.049    0.976  0.3198
## s(water_intake)      1.052  1.101    0.683  0.3930
## s(workout_frequency) 1.027  1.053    0.101  0.7811
## s(carbs)             1.031  1.061    1.761  0.1806
## s(proteins)          1.032  1.063    0.114  0.7795
## s(fats)              1.031  1.060    1.015  0.3148
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.84   Deviance explained =   84%
## -REML = 1.3439e+05  Scale est. = 40297     n = 20000
```

```
plot(combined_gam2, pages = 1, se = TRUE)
```




```
draw(combined_gam2)
```



Classification Tree / Random Forest

The single decision tree with CP of .01lead to RMSE of 23.670 and explains 41.298% of variance explained, which by itself is not too useful.

//Add model parameters from .fit

//Add graphs here

When the tree are made into a random forest of 500 trees the RMSE drops to 9.844 and explains 89.847% of the variance in the data.

//add dendragram here

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