Weight Change Prediction – Final Report

Andy – Hanyu Chao, Michelle – Mengxi Lyu, Jacky – Pengyu Chen

1. Introduction

1.1 Research Background

Weight change is a crucial indicator of human health, and its variation is comprehensively influenced by multiple factors such as lifestyle habits, dietary structure, and physical activity levels. Notably, lifestyle habits are modifiable, making them potential and valuable targets for personalized health intervention. In the current context of increasing attention to public health, predicting weight change can provide strong support for formulating targeted personalized health intervention strategies, filling the practical gap of linking health data to actionable intervention measures.

1.2 Research Significance

This research is of both theoretical and practical significance. Theoretically, it enriches the research on weight change mechanisms by systematically exploring the correlations between demographic, dietary, activity, and lifestyle factors and weight change. Practically, it provides a data-driven foundation for subsequent predictive modeling, which can be further applied to clinical health guidance, public health promotion, and even the development of health management products, helping individuals and health institutions better understand and regulate weight changes.

1.3 Research Goals

The overall research goals are clearly divided into three core directions:

a. Explore the key factors contributing to weight change, focusing on analyzing the impact of variables such as age, gender, caloric intake, physical activity, sleep quality, and stress levels.
b. Conduct in-depth exploratory data analysis (EDA) to identify trends in weight change, distribution characteristics of influencing factors, and correlation relationships between variables.
c. Lay a solid foundation for subsequent predictive modeling through the above two aspects, ensuring that the subsequent modeling work is based on sufficient data understanding and factor screening.

2. Data Source

2.1 Basic Information of the Database

The research data is derived from the Kaggle dataset named "Diet Analysis, Predict The Weight", with the official access website: https://www.kaggle.com/datasets/abdullah0a/comprehensive-weight-change-prediction. The dataset is licensed under CC0 (Public Domain), ensuring legal and free access for research, and its usability is rated 10.00 (full score), indicating high reliability and applicability.

2.2 Dataset Characteristics

Sample Size: The dataset includes information from 100 participants, covering a certain range of demographic groups and lifestyle scenarios, providing a basic sample basis for analyzing weight change rules.

Included Variables: The key variables in the dataset are divided into four categories:

a. Demographic variables: Age, Gender;
b. Physiological and weight-related variables: Current Weight (lbs), Basal Metabolic Rate (BMR, in Calories), Final Weight (lbs), Weight Change (lbs), Duration (weeks);
c. Dietary variables: Daily Caloric Intake, Daily Caloric Surplus/Deficit;
d. Lifestyle variables: Physical Activity Level, Sleep Quality, Stress Level.

Data Purpose: The original purpose of the dataset is to analyze the effects of diet, exercise, and lifestyle on weight change, which is highly consistent with the research theme of this study, ensuring the relevance and effectiveness of the data used.

3. Data Analysis

3.1Explore the variable relationships using ggplot

To supplement the statistical modeling, this study employed the ggplot2 visualization method to explore the relationship between key predictor variables and weight changes. These graphic analyses visually reveal the correlation between linearity and nonlinearity, providing support for subsequent modeling:

3.11 Duration and weight changes

As the duration of the research extended, the trend of weight loss became more pronounced. This result is consistent with the conclusions of the Lasso and GAM models, that is, duration is a significant negative factor affecting weight change.

3.12 Physical activity levels and weight changes

Different activity levels exhibit differentiated patterns of weight change. Individuals with moderate activity levels experience the greatest fluctuations in weight, while very active individuals sometimes experience slight weight gain, which may be related to an increase in muscle mass. This is consistent with the small but explicable influence of active variables in the Ridge and Lasso models.

3.13 Sleep quality and weight changes

Participants with poor sleep quality tend to lose more weight, while those with good sleep quality experience more stable or smaller weight changes. This is consistent with the findings of the Lasso model, where sleep quality is one of the strongest predictors.

3.14 Stress levels and weight changes

Higher stress levels significantly predict greater weight loss, and the relationship shows a steep nonlinear trend. This echoes the results of the GAM model, where stress was identified as a variable that has a significant nonlinear impact on weight changes.

3.15Exercise level and calorie intake

Analysis shows a positive correlation between the two: individuals with a higher level of physical activity tend to consume more daily calories. This indicates the existence of compensatory behavior, that is, an increase in exercise may lead to an increase in food intake, thereby partially offsetting the weight loss effect.

3.16 Stress levels and daily calorie intake

Stress is also related to dietary behavior. Higher stress levels are usually accompanied by lower average calorie intake, which may explain why stress is associated with weight loss in multiple models.

3.17 Current weight and sleep quality

The visualization results show that individuals with higher body weight are more likely to report poorer sleep quality, suggesting that there may be an indirect path: body weight affects sleep, and sleep further influences body weight changes.

3.2 Ridge Regression:

The ridge regression model handles multicollinearity by adding a penalty (L2 regularization) to the standard linear regression objective function, thereby shrinking coefficients toward zero to reduce model variance and improve generalization to new data. It's used to address overfitting and improve the stability and interpretability of models when independent variables are highly correlated.

A Ridge regression model was performed to predict Weight.Change..lbs. using a randomly selected 20% of the weight dataset as training data and evaluating its performance on the remaining 80%. The analysis was conducted over 1,000 trials to ensure the stability and robustness of the results. The glmnet and ggplot2 R packages were used for model fitting and visualization.

Through performing the Ridge Regression, we get an Average Mean Squared Error (MSE): 404.82, Standard Deviation of MSE: 13.45766, Average R^2: 0.5584403. The results indicate that the Ridge regression model is both predictive and robust. The average MSE shows its predictive performance, while the low standard deviation demonstrates that this performance is highly reliable and consistent.

The average R^2 shows that 55.8% of the variance in weight change can be explained by the modelThe coefficients for each predictor in the linear model are distinct from those in other models (glm, lasso) because Ridge regression applies a penalty that shrinks the coefficient values toward zero, but does not force them to become exactly zero.

3.3 Lasso:

Lasso regression is a powerful statistical technique that not only predicts an outcome but also helps identify the most important predictors by shrinking the coefficients of less relevant variables to zero. Our analysis identified two primary models based on the regularization parameter, lambda (λ): the "best fitting" model and a "simpler" model.

The Best-Fitting Model (λ min= 0.5308461) uses the lambda value that provides the lowest cross-validated error, making it the most accurate in terms of predictive performance. It includes the most influential predictors, which are:

- Poor sleep quality: Individuals with poor sleep quality experienced an average of 8 pounds more weight loss.
- Stress level: For every one-point increase in stress, we saw an additional 1 pound of weight loss.
- Duration: Every extra week in the program was associated with 0.25 pounds of weight loss.
- Physical activity: Being very active correlated with a slight weight gain of about 1 pound, which could be due to muscle mass.

This model provides a comprehensive view of the factors most closely associated with weight change in the dataset.

The Simpler Model (λ 1se= 2.581293) uses a higher lambda value to achieve greater simplicity. It is considered more robust and less prone to capturing noise in the data because it only retains the most powerful predictors, while shrinking all others to zero. This model identified only two key predictors:

- Poor sleep quality
- High stress

This suggests that, of all the variables in the dataset, poor sleep quality and high stress are the two most consistent and reliable predictors of weight loss outcomes. The effect of other variables, while present in the more complex model, is not as robust.

3.4 Random Forest:
Random Forest is an ensemble machine learning model that leverages multiple decision trees to make predictions for both classification and regression tasks. It operates by building a "forest" of individual decision trees and then combining their results to produce a more robust and accurate final prediction.:
The random forest model, while flexible, performed very poorly in predicting weight change. The key metrics are:

- MAE: Approximately 5.0–5.5. This indicates that on average, the model's predictions were off by 5 to 5.5 pounds.
- MSE: Approximately 55–60. The high value here suggests that the model frequently made large errors.
- R^2: 0–0.1. This is the most critical metric. An R^2 close to zero means the model explains almost none of the variability in weight change. In essence, the random forest model is not a good fit for this data.

However, the model still provided valuable insight into feature importance. It identified Stress Level, Duration, and Current Weight as the most important predictors, while Gender was deemed the least important. This suggests that the relationships between these variables and weight change are complex and non-linear, which is what a random forest model is designed to detect, even if it couldn't accurately predict the outcome itself.

3.5 GAM (Generalized Additive Model):
A Generalized Additive Model (GAM) is a flexible statistical model that extends linear regression to capture non-linear relationships between predictor variables and a response variable by summing smooth functions of each predictor.
We used GAM to visualize each predictor variable's relationship with the response variable which is weight change, identifying both linear and non-linear relationships:
Variables with a Significant and Linear-like Effect:

- BMR (Calories): weak, negative relationship demonstrated.
- Duration (weeks): a clear negative relationship with the outcome.

Variable with a Significant and Nonlinear Effect:

- Stress Level: higher stress levels are associated with a significant decrease in the outcome variable, but the effect is not uniform across all stress levels.

3.6 *Logistic Regression: weight gain (1) vs weight loss (0)
Logistic regression is a fundamental machine learning algorithm used for classification, not for predicting continuous values like the Weight Change (lbs) column. While its name includes "regression," it's used to model the probability of a categorical outcome.
To perform a logistic regression on the dataset we first transformed the continuous Weight Change (lbs) variable into a binary, or categorical, variable. We classified each observation as either having "Lost Weight" or "Gained Weight." Therefore, a positive weight change is classified as "Gain" and a zero or negative weight change is "Loss."

However, when using this model, we encountered an error called "Singular Matrix", indicating that there is complete multicollinearity among the independent variables, or that the classification level of some variables leads to "perfect separation" (one or more independent variables can perfectly predict the outcome), preventing logistic regression from converging. Through the IVF test we can confirm there's no significant multicollinearity among the variables. As a result, we believe the error may be due to perfect separation with one predictor having an overwhelming influence on weight change. A primary reason for this is the small sample size (100 observations) of the dataset. This can be fixed by collecting more observations to prevent coincidences of a trend or a variable being too prominent.

3.7 Glm+Vif

When we were working on this model, a negative number appeared in R-squared. R-squared =1 - TSS/RSS. When RSS > TSS, the numerator is larger than the parent, resulting in 1 - TSS/RSS<0. In other words, the predictive effect of the model is even worse than that of a simple "predict by mean" model.

4. Research Limitations

Based on these findings, we've identified several key limitations of this study that may have contributed to the models' performance:

- Sample Size: The dataset contains only 100 participants, which is a relatively small sample size for building a robust predictive model. A larger dataset would likely reveal more stable and generalizable relationships.
- Data Quality: Several important variables, such as Sleep Quality, Stress Level, and Physical Activity, were self-reported. This method can introduce human bias and inaccuracies, potentially affecting the model's ability to learn true relationships.
- Unobserved Factors: The dataset does not include crucial variables that are known to impact weight change, such as diet composition, genetic factors, and other lifestyle habits. The lack of these variables makes it impossible for the models to capture the full complexity of the problem.
- Simulation vs. Reality: While our simulated trials on the dataset showed low error for the random forest model, this doesn't guarantee the same performance on new, real-world data, which may contain more noise and unpredictability.

5. Conclusion:

This study compared the performance of GLM, Ridge/Lasso/Elastic Net, random Forest (RF), and generalized additive model (GAM) in predicting weight changes. The results show that the linear model with regularization is generally superior to the baseline GLM, but the R^2 on the test set is still relatively low. RF and GAM have revealed some nonlinear effects, but their predictive capabilities are limited. Overall, daily calorie surplus/deficit and its duration are the main factors determining weight changes, while the influence of other variables (current weight, age, stress, gender) is relatively weak and unstable.

The simplified regression equation is:

$$\text{Weight Change} = \beta_0 + \beta_1 \cdot (\text{Caloric Surplus/Deficit}) + \beta_2 \cdot \text{Duration}$$
$$+ \beta_3 \cdot \text{Current Weight} + \beta_4 \cdot \text{Age} + \beta_5 \cdot \text{Stress}$$
$$+ \beta_6 \cdot I(\text{Male}) + \varepsilon$$

From a classification perspective (weight gain = 1, weight loss = 0), it can be expressed as:

$$\text{Weight Change} = \beta_0 + \beta_1 \cdot (\text{Caloric Surplus/Deficit}) + \beta_2 \cdot \text{Duration}$$
$$+ \beta_3 \cdot \text{Current Weight} + \beta_4 \cdot \text{Age} + \beta_5 \cdot \text{Stress}$$
$$+ \beta_6 \cdot I(\text{Male}) + \varepsilon$$

The limitations of this study lie in the small sample size, the insufficiently precise quantification of behavioral variables (such as activity level and sleep quality), and the limited accuracy of the data, all of which weaken the extrapolation of the results. Future work should focus on expanding the sample size and improving the measurement of behavioral variables.