STATS 503 Research Report on Obesity

Yuxiang Gao, Chenfei Wang, Tongyao Jiang

1. Introduction

Obesity is a chronic disease of adults and children, which continues to increase in the United States. It can lead to severe diseases such as diabetes, metabolic disorders, and some cancers. According to the CDC, 1 in 5 children and more than 1 in 3 adults struggle with obesity. In addition, a recent study has highlighted that young adults are more vulnerable to obesity and its complications. These findings underscore the importance of the prevention and control of obesity.

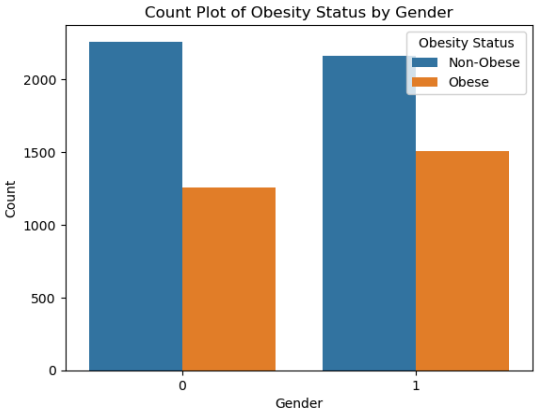
In this report, we utilized machine learning methods to examine two scientific questions related to obesity. In the first part, we built a classification model to predict obesity based on personal living habits such as diet, smoking, and sleep cycle. We used the Body Mass Index (BMI)[1] score to determine each person's **obesity status**. In the second part, we analyzed the relationship between the change in weight in the past year and different strategies. We explored which strategy may be statistically significant in decreasing body weight and which might be helpful for someone to **lose weight**.

2. Data

This report combines several datasets in 2017-2020: *Weight History*, *Demographic Variables*, *Sleep Disorders*, *Diet Behavior & Nutrition*, *Physical Activity*, *Smoking - Cigarette Use*, *Alcohol Use,* and *Body Measures*. The first dataset is *Weight History*. It records some loss strategies people took or did not take and their weight change in the past year. As the names suggest, the rest of the datasets represent an aspect of people's living habits, including potential variables influencing weight.

Part1: Predicting obesity

In the first part, we chose the samples that recorded interviewees' demographic information and their dietary, eating, sleeping, and exercise habits. The effective sample size is 7182 and contains 47 features. We've done visualizations based on different variables. For example, from the plot below, it seems that more males are considered obese than females in our study cohort.



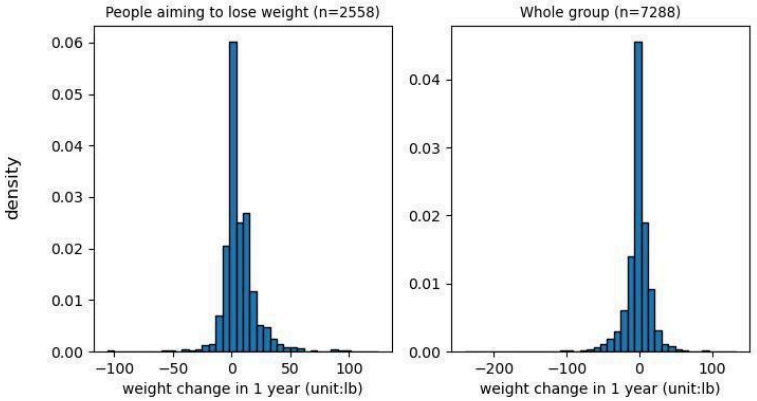
Picture1. The distribution of weight changes

Part2: Weight-loss strategies

In the second part, we chose the interviewees who claimed that they tried to lose weight in the past year, and the effective sample size was **2558**. The response variable was the weight

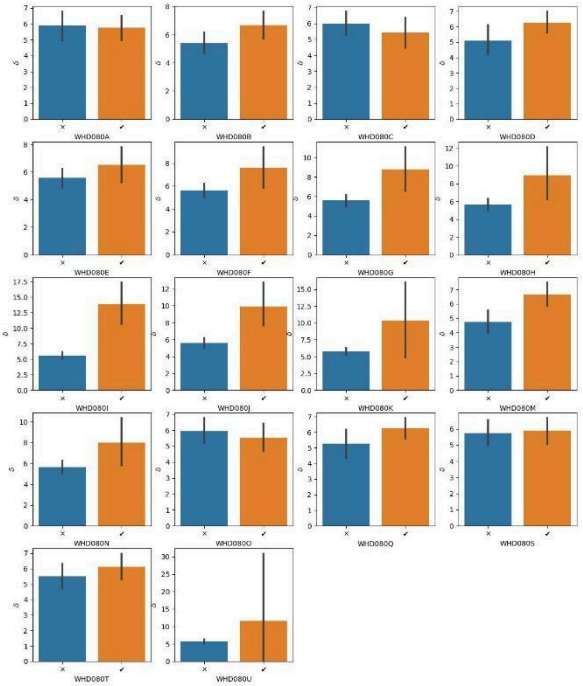
[1] We used weight and height to calculate each person's BMI, and turned it into a binary response variable for obesity status.

change in the past year[2]. Also, 40 independent variables[3] were chosen from the combined datasets.



Picture2. The distribution of weight changes

First, we plotted the distribution of the response variable (i.e., left). For comparison, we also plotted that of the whole group with the sample size being 7288 (i.e., left). As for the whole sample, the weight change was relatively symmetric, with some gaining and losing weight. On the other hand, most people aiming to lose weight actually gain weight. Probably, people aiming to lose weight belonged to a sample that tended to gain weight easily—this is called sample bias. But we could say our concentration was on this segment of people.



Picture3. The bar plot of average weight changes depending on whether taking the strategy or not

In Picture 3, there were 18 subplots, showing the relationship between the weight change (i.e., response variable) and one of the strategies[4]. We found that the **mean** was **similar** in many cases. The plot also provided the confidence interval of the estimated mean (i.e. the black line). Some black lines were long because few people took the strategy, so the number of sub-samples was small leading to high standard error.

[2] This equals WHD020 - WHD050, the current weight minus the weight one year ago [self-reported]
[3] Before one-hot coding, there are 40 variables including numerical and categorical ones.
[4] It is a binary variable. “Tick” indicates taking the strategy and “cross” indicates the opposite.

3. Methods

Part 1: Predicting obesity

In the first problem, we calculated each person's BMI score based on height and weight. According to the CDC criteria, anyone with a BMI score greater than 30 is considered obese. We followed this criterion and transformed the obesity score into a binary variable with 0 as non-obese and 1 as obese, which was used as our outcome variable. Because we have a binary outcome variable, the first problem is determined as a classification prediction problem. We chose logistic regression (with L1 penalties) as our baseline model and utilized 10-fold nested cross-validation to choose hyper-parameters. In addition to linear models, we also tried tree methods (random forest classifier), which have better prediction performance than linear methods. Finally, we will use various metrics, including accuracy and the AUC-ROC curve to examine model performance.

Part 2: Weight-loss strategies

Since the response variable was the weight change, a **continuous** variable, it was a **regression** problem. Therefore, LDA/QDA and SVM, mainly designed for classification, were not preferred in this scenario. In this part, we applied **three** types of tools to explore the effectiveness of strategies: 1) Linear models with lasso penalties, 2) splines and GAM, and 3) trees like random forest and boosting. Cross-validation, of course, is also used.

(1) Why use them?

First, there were more than 40 variables, so lasso penalties were useful for **dimension reduction** in the linear model and helped us filter more significant predictors. Of course, normalization was applied to ensure equal penalty weights on all variables.
Second, a non-linear relationship between the response and predictors can cause problems for linear models. Although polynomials could slightly deal with this, splines can better capture the non-linearity. That was the reason why we used GAM and included spline features.
Finally, we also considered boosting. Although it does not have the advantage of interpretability compared to linear models, it is good at learning residuals and improves the fit.

(2) How to use them

We did a train/test split: 70% of the data was used for training, and the rest for testing.
First, we started with the original linear model. We included all 40 predictors[5] in the linear model and trained it. Then we used the model to predict the test data. The original linear model was mainly used for comparison. Then we normalized all the predictors and added lasso penalties. Also, the lasso model was trained and used to predict the test data.
Second, we used GAM, which was convenient for generating spline features on numeric variables. Did the training and predicted the test data.
Finally, we tried tree methods. Random Forest was tried first. There were some tuning parameters, and we just used the default to start, except for *max_features*[6], where we used the thumb rule: the square root of the number of predictors was around 7. We trained the model and predicted the test data. During this process, we also adjusted the tuning parameter. The procedure for boosting was quite the same, starting with the default tuning parameters. When we trained the model, we adjusted the tuning parameters like the learning rate to balance good fit and overfitting.

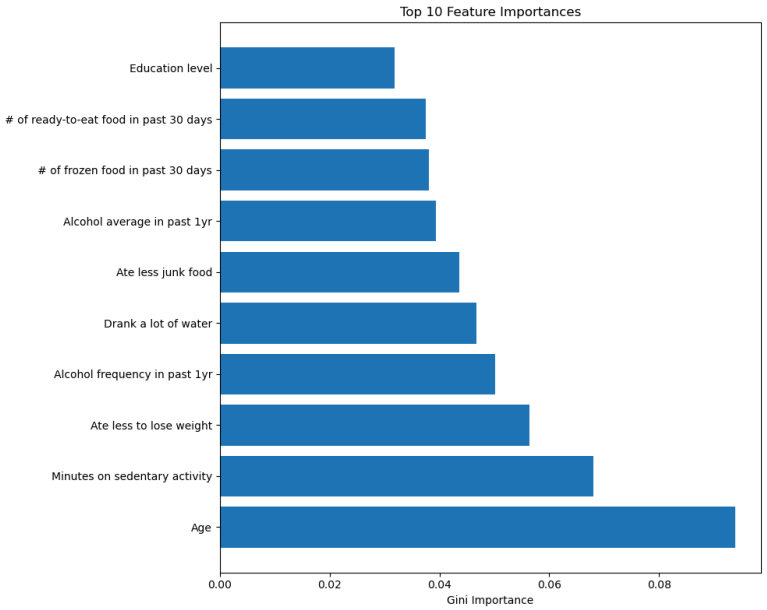
4. Results

Part1: Predicting obesity

In the first problem, we fitted a **logistics regression model** (without penalty terms) as the baseline, which achieved an accuracy of 68.2%. After selecting the best performance regularization strength C = 1000 by conducting nested 10-fold cross-validation, the **logistic**

[5] Before we start, we have done the one-hot encoding on categorical variables.
[6] It is the number of features the tree could choose each time when it splits.

regression with L1 penalty achieved an accuracy of 68.9%. Initially, the dataset contained 47 features. After adding the L1 penalty term, only 1 feature was 'shrunk' and 46 features were selected for prediction. These features could help us better interpret the model. For example, the coefficient of feature 'Age' is 0.009, indicating that the log odds ratio will be 0.009 higher when there is a 1 unit increase in age, provided other features are equal. The coefficient of the feature 'Alcohol frequency' is 0.09, indicating that the log odds of being obese will increase by 0.09 when there is a 1 unit increase in alcohol frequency, with other features being equal. The **random forest classifier** achieved an accuracy of 69.7% after hyper-tuning for the tree depth. We filtered and presented the 10 features with higher Gini importance in a bar plot. From Picture 4 below, features such as 'Age,' 'Minutes on sedentary activity,' 'Alcohol frequency,' 'Ate less to lose weight' and 'drank a lot of water' are the top 5 features with higher gini importance. We looked back at the coefficients of these features in the logistics model and recorded the results in Table 2 below.
In addition, we drew the ROC curve and calculated the AUC score, which is included in Pictures 5 and 6.



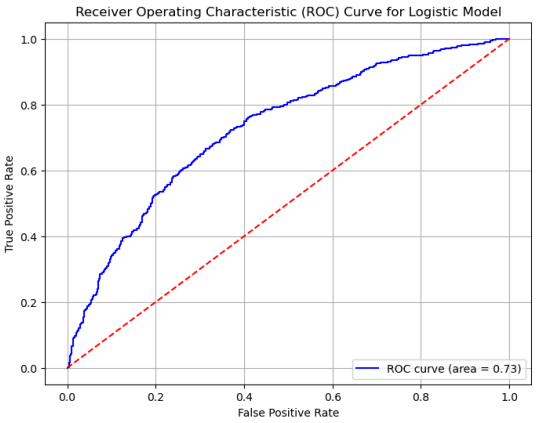
Picture4. Feature importance of random forest classifier

Table 1. Model Performance

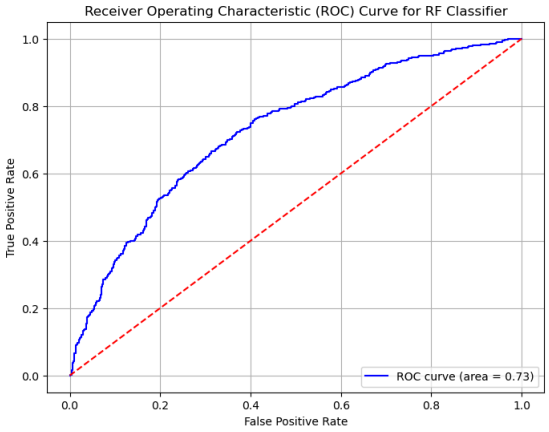
	Accuracy	AUC
(1) Logistic	68.2%	71.9%
(2) Logistic + lasso	68.9%	72.7%
(3) Random Forest	69.7%	72.8%

Table 2. Coefficients

	Coefficient
(1) Age	0.0093
(2) Mins on sedentary activity	0.0070
(3) Ate less to lose weight	0.4600
(4) Alcohol frequency	0.0900
(5) Drank a lot of water	-0.4400



Picture 5. ROC Curve for L1 Logistic Model



Picture 6. ROC Curve for Random Forest Classifier

Part2: Weight-loss strategies

(1) Model fit

The R^2 results of five models on the training and test data are given below. Overall, all of the model fits were **terrible**. The **lasso** model had the **highest R^2** on the held-out test dataset, followed by two tree models. Random forest and Boosting had roughly the same performance. GAM performed the worst and missed the non-linear relationship.

To get a sense of the quality of the estimate, we also used cross-validation on the 3 best methods mentioned above. MSPE and its standard deviation of the three methods were roughly the same.

Table 3. Result of model fit

	Train R2	Test R2
(1) Linear	0.089	0.015
(2) Linear + lasso	0.068	**0.039**
(3) GAM	0.115	0.009
(4) Random forest	0.318	**0.030**
(5) Boosting	0.134	**0.031**

Table 4. Cross-validation of the three best methods

	MSPE	standard deviation
(2) Linear + lasso	213.340	7.630
(4) Random forest	213.763	7.994
(5) Boosting	216.528	7.268

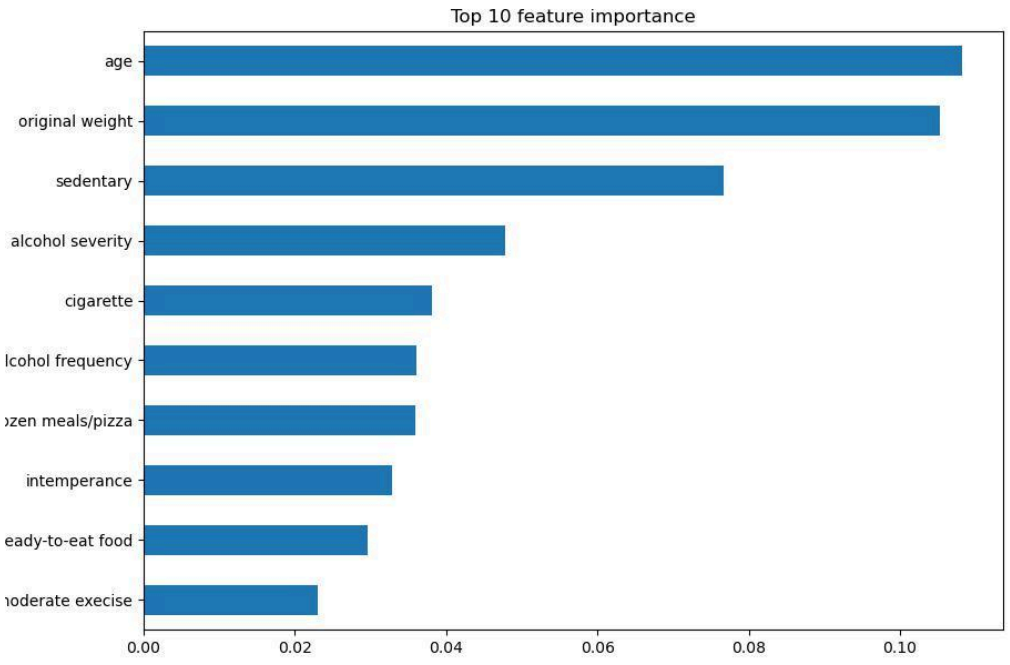
(2) Coefficient explanation[7]

First, we examined the original linear model and the lasso model. The coefficients of the significant variable were quite similar, and we gave partial results of the lasso model here. It was terrible to find that eating fewer calories leads to increasing weight. This made no sense. Interestingly, the coefficient of **sleep** was **logical, as** sleeping sound can lead to weight loss. Other variables being equal, compared to people sleeping less than 7 hours every day, people who slept more than 7 hours lost an average weight of around 1.8 pounds in the past year.

Table 5. coefficients of the lasso model (partial)

	coef	p-value
less calorie	2.186	0.003
diet program	5.381	0.002
medicine	6.520	0.001
slee_best	-1.832	0.013

Take a look at the random forest's variable importance. The two primary variables were age and weight one year ago. Since our focus was strategy, it was rational to control these features. Apart from them, the most significant predictor is “long-time sitting.” In real life, the longer we sit and the fewer exercises we do, the more prone we are to gain weight.



Picture7. Variable importance of random forest

[7] The summary of linear model is too long to be put here.

5. Conclusions

Part1: Predicting obesity

(1) Model performance

In the first obesity prediction problem, the performance of the random forest classifier is slightly higher than that of the logistics regression model with an L1 penalty (68.2% vs. 69.7%). We also compared their AUC scores, where L1 logistic received 72.7% and random forest received 72.9% (Table 2 in the Results section). Both of them are fair classifiers. However, the L1 logistic regression model also has better interpretability than tree models, so we can understand how the features affect the odds of obesity by reading coefficients.

(2) Importance features

In examining the model performance, we also looked through the features that may influence the obesity status. By comparing feature coefficients in the logistic model with the feature importance graph generated from the random forest classifier, we may assume that age, alcohol drinking frequency, minutes of sitting, water drinking amount, and food intake amount are essential predictors for obesity. Specifically, the increase in age, alcohol drinking, and minutes of sitting will increase the odds of being obese while drinking more water will decrease it. However, for those who ate less to lose weight, the odds of being obese are higher than those who didn't (feature: 'ate less to lose weight'), which differs from our expectations. It may come from the fact that people who try to eat less are likelier to have a higher weight than others.

Part2: Weight-loss strategies

(1) Did the results shed light on this question?

Now, go back to the question at the beginning: which strategy may be statistically significant in decreasing body weight? We **cannot answer** this question based on this dataset without confidence. The R^2 on the test dataset was no more than 0.04 (see Table 1), which indicated a too-vulnerable correlation between the response and predictors[8]. Besides, the sign of many strategies' coefficients was positive (see Table 3). Eating food with fewer calories, people tended to gain weight—which was unreasonable. Given this situation, we should focus on why all models fail.

Limitations:

There is a really potential concern about data validity.

First, it is **doubtful whether people answered the questionnaire carefully.** Of the eighteen weight-loss strategies, none are effective.

Next, even if we assume the questionnaire is filled out in earnest, one of the most severe concerns is whether these people continue using weight loss strategies. Take smoking for comparison. Many people try to stop smoking, but they slide back to their bad old habits. Since the questionnaire only asked whether they had tried this method in the past year, we cannot know if this is true.

6. Contribution

Yuxiang Gao: Responsible for the 1st question: build models and write reports

Chenfei Wang: Data cleaning

Tongyao Jiang: Come up with the proposal, responsible for the 2nd question: build models and write reports

7. Reproducibility

This report explored two questions with two different response variables (Y). For simplicity, the codes are stored in two separate files. As for the 1st question–predicting obesity, one can run the code of the corresponding file. All the plots and models can be obtained without any effort. You only need to put some results into the tables. The logic for the 2nd question is the same. After that, you can reproduce this report.

[8] Even this 0.04 may be by chance. When you have large datasets, it is likely to have correlation.