What Makes You Obese?

Introduction to Data Science Capstone Project

Members: I-Chieh Kuo (ik2437@nyu.edu) Bangyan Chen (bc3517@nyu.edu)

Introduction

Overweight and obesity are common health conditions in the United States. It describes a person's weight as higher than what is considered normal for a given height. According to 2017-2018 data from the National Health and Nutrition Examination Survey (NHANES), nearly 1 in 3 adults (30.7%) are overweight. More than 2 in 5 adults (42.4%) have obesity. About 1 in 11 adults (9.2%) have severe obesity. Overweight and obesity are risk factors for non-communicable diseases such as cardiovascular diseases and diabetes, which greatly threaten public health.

Many factors contribute to overweight and obese conditions. The contributors can be divided into two categories -- diet habits and lifestyles. This report investigates what contributes to overweight and obesity in the following two parts. In the first part, we describe our data. In the

first part, we apply statistical methods such as hypothesis tests. In the second part, we utilize linear regression to determine trends and make predictions. We use machine learning models for projections and decisions in the third part.

Data and codebook

The data we use in this project is from research in this area. The data contains 17 attributes and 2111 records. The meaning of each column and its answers are in the code book in Table 1. Fortunately, there is no NAs in the data set. The data collector used some methods to deal with some missing data. Therefore, some cells supposed to be integers are float values in the dataset.

We then calculate the BMI by this equation.

$$BMI = \frac{Weight}{Height^2}$$

With the BMI index, all the records are labeled with the class variable NObesity (Obesity Level). The criteria are in Table 2. For the convenience of hypothesis testing. We then split the

NObesity into two tyes. Those who are obese (Obesity I, Obesity II, Obesity III) are labeled “1” in the new variable “is_obesity” while respondents of normal- or overweight are “0.”

Table 1. Questions of the survey used for initial recollection of information.

Question	Column in dataset	Answers
What is your gender?	Gender	Female; Male
what is your age?	Age	Numeric value
what is your height?	Height	Numeric value in meters
what is your weight?	Weight	Numeric value in kilograms
Has a family member suffered or suffers from overweight?	family_history_with_overweight	Yes; No
Do you eat high caloric food frequently?	FAVC	Never; Sometimes; Always
Do you usually eat vegetables in your meals?	FCVC	Never; Sometimes; Always
How many main meals do you have daily?	NCP	Between 1 & 2; Three; More than three
Do you eat any food between meals?	CAEC	Never; Sometimes; Frequently; Always
Do you smoke?	SMOKE	Yes; No
How much water do you drink daily?	CH2O	Less than a liter; Between 1 and 2 L; More than 2 L
Do you monitor the calories you eat daily?	SCC	Yes; No
How often do you have physical activity?	FAF	I do not have; 1 or 2 days; 2 or 4 days; 4 or 5 days

How much time do you use technological devices such as cell phone, videogames, television, computer and others?	TUE	0–2 hours;
		3–5 hours;
		More than 5 hours

how often do you drink alcohol?	CALC	I do not drink;
		Sometimes;
		Frequently;
		Always

Which transportation do you usually use?	MTRANS	Automobile;
		Motorbike;
		Bike;
		Public Transportation;
		Walking

Table 2. Body Weight Types by BMI Range

NObesity	BMI Range
Underweight	Less than 18.5
Normal	18.5 to 24.9
Overweight I	25.0 to 26.9
Overweight II	26.9 to 29.9
Obesity I	30.0 to 34.9
Obesity II	35.0 to 39.9
Obesity III	Higher than 40

Inference Statistics

Question

Do people who are obese differ from those not in diet habits and lifestyles?

Approach

To answer these questions, we make two kinds of hypothesis tests for categorical (is_obesity) - numeric (FCVC, NCP, CH2O, FAF, TUE) comparisons and -categorical (is_obesity) – categorical (family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC) comparisons. The significant level is set at 0.05, which is commonly used.

For the first kind of hypothesis test, we check the distribution of each column grouped by is_obesity. Figure 1 shows that they are not normally distributed. Therefore, we use the Mann-Whitney U test rather than the T-test. The null assumption is that the probability is 50% that a randomly drawn member of the non-obese population will exceed a member of the obese population.

For the second kind, we choose the Chi-Square Test, as we want to compare the observed and the expected frequencies of the set of individuals who are obese and who are not (the selected features are categorical).

Result

We run the test and list the result in Table 3. The test results are statistically significant for FCVC, CH2O, FAF, and TUE. However, the p-value for NCP is 0.096. Hence we fail to reject the null hypothesis. There is no significant statistical difference in the number of meals obese and non-obese respondents take every day. In addition, we check the effect size for each variable. We use rank-biserial correlation for dichotomous nominal data vs. rankings (ordinal). Three of the four significant results have a small-medium effect size, indicating there may be moderate differences in these aspects between the two weight groups.

Table 4 contains the result of the Chi-square test. Apart from smoking habits, every variable has a significant result with p-value close to 0. We use Cramer’s V as effect size. The effect sizes are all in the small-medium range of the test result, from 0.157 to 0.417.

In summary, some diet habits and lifestyles are statistically different between obese and non-obese people. Therefore, we want to learn more about to what extent these factors can have an influence on a person’s body weight.

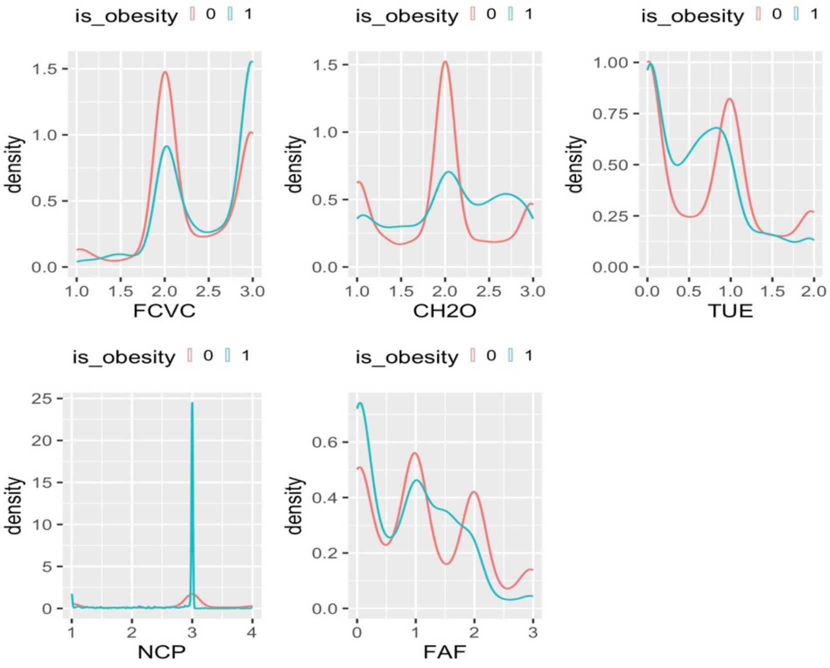
Table 3. Mann-Whitney U Test and Effect Size

group	group1	group2	n1	n2	p	Effect size
FCVC	0	1	1139	972	0.000	-0.218
NCP	0	1	1139	972	0.096	0.096
CH2O	0	1	1139	972	0.000	-0.139
FAF	0	1	1139	972	0.000	0.146
TUE	0	1	1139	972	0.004	0.071

Table 4 Chi-square Test Results and Effect Size

group	group1	group2	n1	n2	p	Effect size
family_history	0	1	1139	972	0.000	0.417
FAVC	0	1	1139	972	0.000	0.278
CAEC	0	1	1139	972	0.000	0.364
SMOKE	0	1	1139	972	0.705	0.012
SCC	0	1	1139	972	0.000	0.188
CALC	0	1	1139	972	0.000	0.157

Figure 1. Distribution of Numeric Variables



Prediction

Question

Can personal eating habits and physical conditions predict body fatness, controlling for gender and age?

Approach

To answer our research question, we computed the body mass index (BMI) of each person as the dependent variable (Y) and the 14 other eating habits and physical conditions as the independent variables (X) in our analysis. The BMI is calculated by dividing a person's body weight by the square of their height and is a reliable indicator of healthy weight.

Before we can make predictions and classify the data, we performed some feature engineering. Specifically, we converted the following columns from character to categorical variables:

- Family_history_with_overweight
- FAVC (frequent consumption of high-caloric food)
- CAEC (consumption of food between meals)
- SMOKE (smoking habit)
- SCC (calories consumption monitoring)
- CALC (consumption of alcohol)

If the variable is binary (i.e., it has only two categories), we converted "yes" to 1 and "no" to 0. If the variable is a frequency measure (such as "no," "sometimes," "frequently," and "always"), we converted it to a 4-point Likert scale. This will allow us to better analyze the relationships between these variables and the target variable. Also, we used one-hot encoding to convert the gender and transportation used variables into numerical form for the model. By this approach, the model would learn and make predictions based on the individual categories, rather than assuming a linear relationship between the categories.

After feature engineering process, we examined the patterns of all predictors using a correlation matrix. The results are shown in figure 1. We found that there were a few correlated features, such as age and using public transportation (r = 0.60), and age and using an automobile (r = 0.55). However, these two features are simply two categories within the transportation used variable, and the correlations were not particularly strong. As a result, we decided not to use dimension reduction methods before training our predictive models.

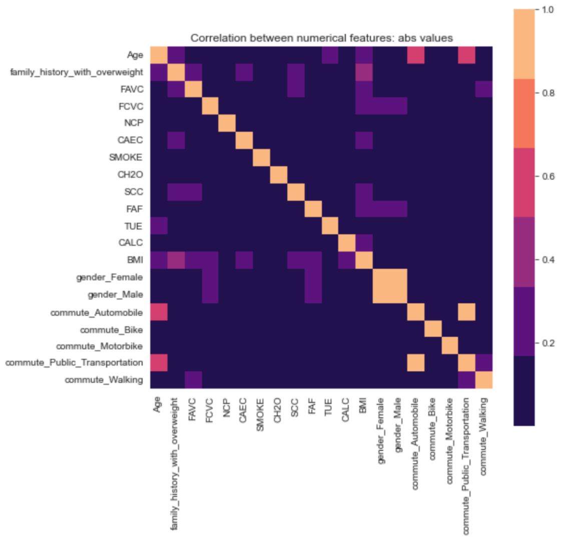


Figure 2 Correlation Heat Map

Result

We performed multiple linear regression on our data, using variables related to eating habits and physical conditions as predictors and body mass index (BMI) as the target. We first scaled the features matrix and the BMI vector, then divided the data into a training set (80%) and a test set (20%). We trained the model using the training set and used it to make predictions on the test set. To evaluate the model's accuracy, we calculated the root-mean-square error (RMSE = 5.6361) and the coefficient of determination (R^2 = 0.4689). We also compared the performance of the model with ridge regression, LASSO regression, and elastic net regression using hyperparameter tuning. The results of the models are summarized in table 1. The elastic net model had the lowest RMSE of 5.6259 and the highest R^2 of 0.4708, indicating it was the most effective.

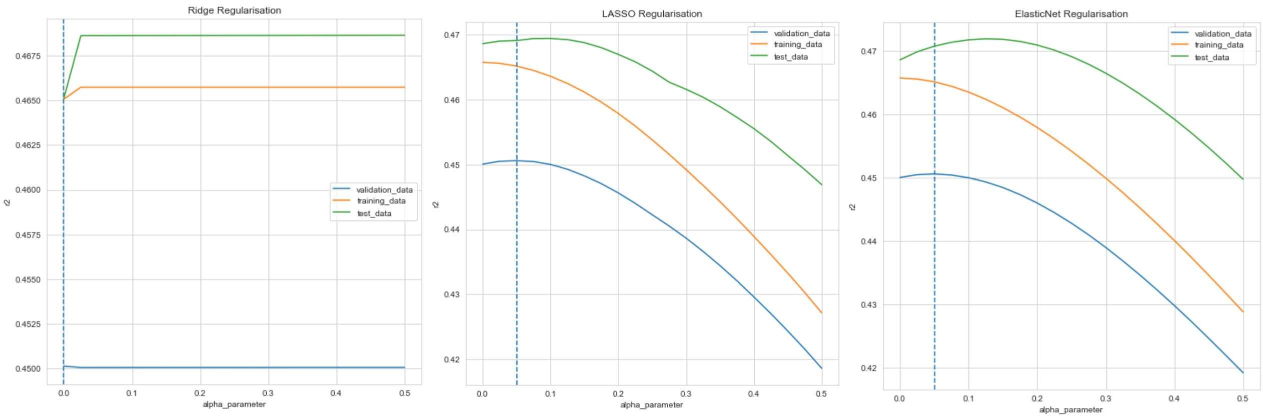
The best parameters using grid search methods are visualized as the following table and plots.

Table 5 Model of Prediction

Method	**alpha**	**R^2**	**RMSE**
Ridge	0.5	0.4651	5.6559

LASSO	0.05	0.4691	5.6349
ElasticNet	0.05	0.4708	5.6259

Figure 3 Models for Preiction



Classification

Question

Can personal eating habits and physical conditions predict obesity?

Approach

The original data had 7 categories: insufficient weight, normal weight, overweight (levels I and II), and obesity (types I, II, and III). Our goal was to classify whether an individual had obesity, so we labeled our data into two categories: 1 if the individual had obesity and 0 if they did not. This column was used as the outcome variable (Y), and the rest of 16 variables related to eating habits and physical conditions was used as the predictors (X). The data was relatively balanced, with 972 individuals having obesity and 1139 individuals not having obesity.

Result

After scaling the eating habits and physical conditions features, we divided the data into a training set (80%) and a test set (20%). We then used logistic regression, K-Nearest-Neighbors (KNN), decision tree, and random forest to predict obesity. We calculated the area under the curve (AUC) values to assess the classification results, which are summarized in table 2. Overall, the four models performed very well, with the logistic regression model performing the best with an AUC value of 0.9999.

Table 6 Logistic Regression, KNN, Decision Tree and Random Forest

Method	AUC value
Logistic regression	0.9999
KNN	0.9682
Decision tree	0.9883
Random forest	0.9997

The ROC curve for the four models are shown below:

Conclusion

Obesity and overweight are major public health concerns that significantly impact both physical and mental health. They can lead to various health problems, including heart disease,

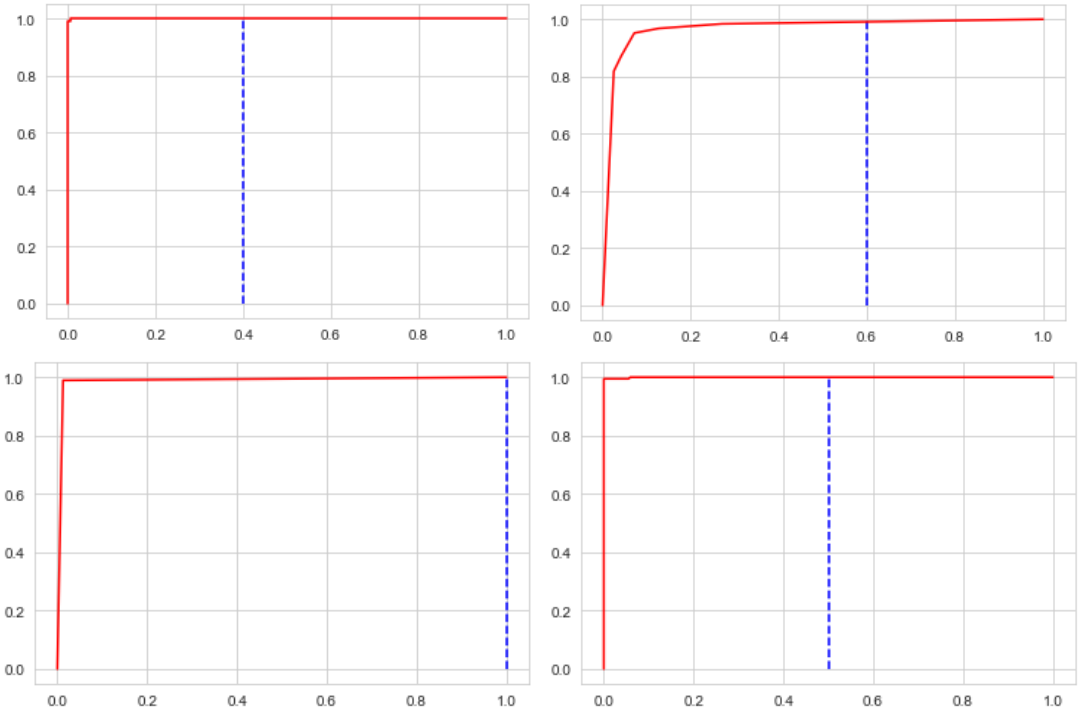


Figure 4 Logistic regression, KNN, Decision tree and Random forest

diabetes, and cancer. After learning the skills in Introduction to Data Science, we are eager to see if we could have some insight into the factors contributing to obesity. In the inference part, we test if obese people have different eating and lifestyles than others. The result tells that the answer is YES in most aspects we consider. Then we try to find the model that best predicts obesity and find the most suitable parameters.

However, our project still have some weakness. For example, our data consists of some imputed data, which harms its representativeness. We hope to dive deeper into this area and find out more about what contributes to obesity.

Reference:

Estimation of obesity levels based on eating habits and physical condition Data Set

Body mass index (BMI)