Food Insecurity: The Social Problem Causing Obesity

Fiona Baenziger

Abstract

In the global situation, there are more people becoming obese than ever before. This can be the result of a multitude of factors: food quality, fast food accessibility, exercise levels, diet culture and more. There is also a hidden underlying social issue that has become more prevalent in the past decade: food insecurity. As the rich are becoming richer and poor are becoming poorer, middle-class families are now facing an inability to provide financial resources for food at the level of the household. By identifying the factors that put an individual at risk of obesity, the effect of income will become clearer and provide useful information as to how to stop the growing trends.

1. Introduction

1.1 The Problem

Over the past 40 years, the global population continues to grow and expand – in more ways than one. Scientists are starting to see an alarming trend in rising obesity rates that begs the question of How and Why. The fact is that the proportion of obese individuals has increased 8.5% in women and 7.6% in men [1]. Within these numbers, children are most often the most affected by this trend – refer to Figures 1 and 2. Obesity is a risk factor for serious health problems including hypertension, type 2 diabetes, mellitus, coronary artery disease, sleep apnea, and osteoarthritis [3]. These consequences can be fatal and often are the catalyst for long-term poor quality-of-life implications. Potential explanations for the rising obesity rates are declining food quality, more processed food than before, more accessibility to fast food, more stagnant daily life, or the demanding diet culture – but these are all surface level issues. There is a cause at a foundational level that is causing the system supporting it to crumble. This social factor is the current global economic situation – especially in the United States. The rich are getting richer and the poor are getting poorer. As economic inequalities have increased, so have inequalities in weight. Those that are most effected? Minorities and the poor. Studies show that school-aged children of families between 100% and 130% of the poverty line were the most overweight [3]. This is due to something called 'food insecurity' or the lack of available financial resources for food at the level of the household. This is different from 'hunger' which is a personal, physical sensation of discomfort. It is this reason that the lower-middle class are the most affected as they still have the ability to afford food, but the nutritional value of the food consumed is a large contributor rising obesity rates. It can thus be

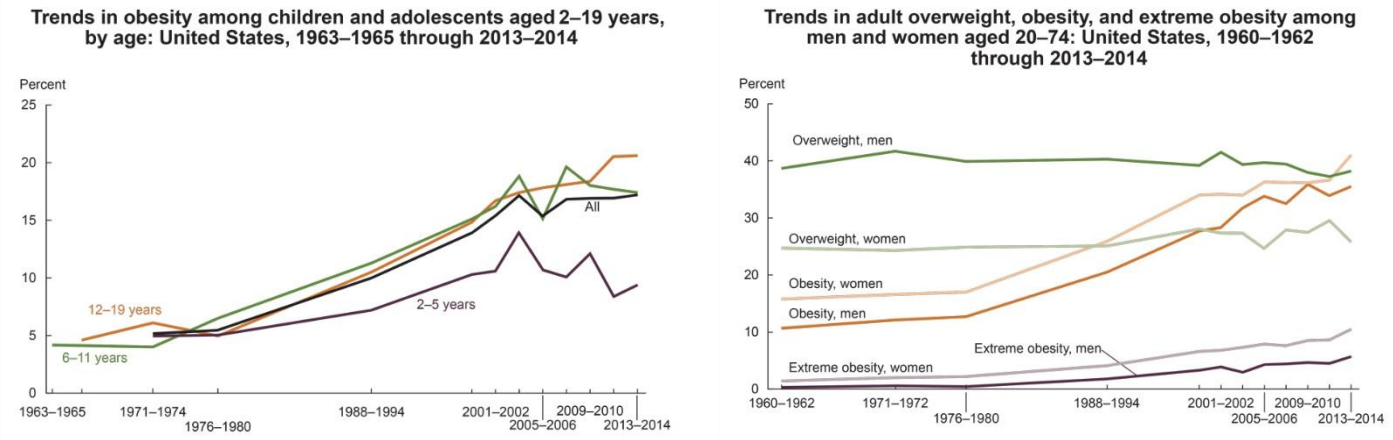

Fig. 1 and 2. A rise in obesity levels over the past 40 years in 2 – 19 years old and 20 – 74 years old [6]

concluded that poverty is an indicator that an individual is more likely to be obese because of food insecurity and lack of available financial resources for nutritional and sustainable food.

1.2 Goal

This paper proposes the usage of machine learning algorithms for exploring the factors of obesity and the relationship between poverty levels and risk of obesity. More specifically, the prediction of individuals who are more likely to be obese based on shopping habits, income, exercise levels, etc. The proposed machine learning algorithm models can be used to aid in uncovering the hidden social factors for increasing obesity levels.

1.3 Previous Work – Literature Review

Obesity is not a new issue in the world and scientists have been researching and brainstorming solutions for years to try to determine the reasons why the trend is growing. Psychologists, philosophers, doctors, nutritionist, dieticians and others are all looking at the problem in different ways to come up with a solution. The following published papers are just a small portion of those delving deeper into why this health risk is a national epidemic and why America should start caring now.

"Predicting Overweight and Obesity in Adulthood from Body Mass Index values in Childhood and Adolescence" is a study using the CDC BMI charts and definitions of overweight and obese children to predict the likelihood of becoming an overweight and obese adult. The paper proposes the use of logistic models to predict overweight and obese adults in the 75^{th}, 85^{th}, and 95^{th} percentiles on the CDC charts of adolescent BMI. [1]

"A fatter, healthier but more unequal world" is exploring the growing obesity trends in relation to other world and economic indicators such as income levels and global life expectancy. As the number of individuals classed as overweight are increased, the economic inequality is increasing as well – leading to a large inequality in weights. [2]

"Obesity and Poverty: Are Food Stamps to Blame?" is exploring the more specific nature of the American obesity epidemic and how individuals in the Food Stamp Program (FSP) are affected by obesity. The paper focuses on how single FSP participants in New York are consuming and how they are using their benefits. [3]

"Poverty, Obesity, and Malnutrition: An International Perspective Recognizing the Paradox" is exploring the global factors contributing to the rising obesity trends. The interrelationship of hunger and food insecurity is the hidden hunger of the world and the paradox of poverty and obesity. The exploration of how poverty is a factor connected to obesity and potential solutions to address food insecurity. [4]

"Predicting Obesity in Young Adulthood from Childhood and Parental Obesity" is a study of how parental obesity affects the chances of a child's likelihood of becoming an obese adult. The paper studies a cohort of young adults and their parents who were long-term members of a health maintenance organization and uses statistical analysis to compute weighted averages of BMI values throughout time. [5]

While there has been extensive research on indicators of obesity in economics, psychology, and philosophy using statistical analysis, the use of machine learning techniques in this topic is not widespread from the research. Not that this topic is groundbreaking, but the prevalence of published papers on machine learning and obesity indicators is low. This paper aims to use this research in aiding the factor exploration process in a more technical data approach than previous methods.

---

2. Methodology

2.1 Dataset

The data set used in this paper is the Eating & Health Module Dataset from the American Time Use Survey Eating & Health Module Files. The data files contain information related to eating, meal preparation and general health collected in the United States from questionnaires given in 2014. More specifically, we know the income, BMI, changes in income, time spent eating, dietary choices, exercise levels and a couple for demographic

information. There are approximately 37 features and 11.2 thousand entries.

2.2 Framework Overview

The methodology framework for this paper is simple. Try many things to determine what is going to give the best and most explainable results.

2.3 Preprocessing

Before creating and running the proposed models, there was an entire process of cleaning and manual feature selection to remove correlations, irrelevant and redundant information – as required when working with wonky questionnaire data. Correlated variables included Body Mass Index (BMI) with height and weight, which makes logical sense as BMI is a ratio of the two. In total, I removed 15 features after analyzing the correlation plot and conducting research on relevant information. After this process, I am working with 22 features.

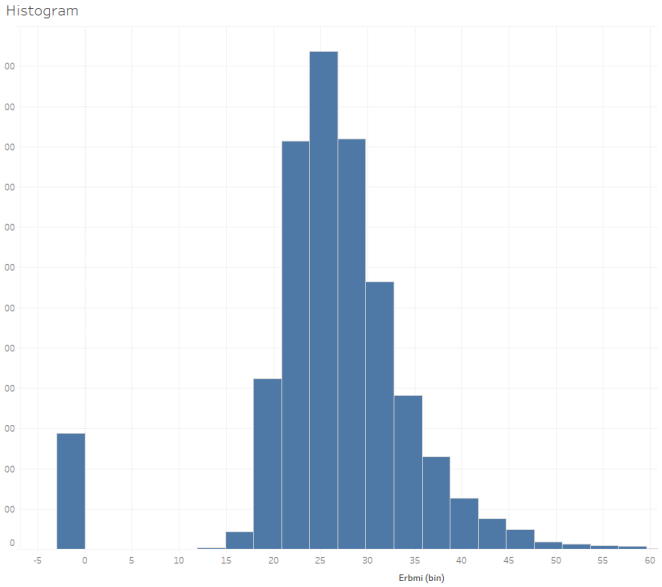

Fig. 3 A distribution of the BMI values

The feature that will be my target for feature selection and model creation is Body Mass Index, or BMI. As mentioned previously, this value is a height to weight ratio commonly used to estimate body composition in individuals. In total, there were 575 missing BMI values, so I removed the corresponding entries. The values ranged from 13.0 to 73.6 – refer to Figure 3 – so I attempted to bin the BMI into 4 different classes: Underweight, values below 18.5; Normal Weight, values between 18.5 and 24.9; Overweight, values between 25.0 – 29.9; Obese I, values between 30.0 - 34.9; Obese II, values between 35 – 39.0; and Obese III, values greater than or equal to 40.0 [6]. This made the bins very uneven and I lost a lot of information by binning as I turned meaningful continuous values into a multi-class problem. These issues made creating a model difficult, so I got rid of the bins. In addition, I removed all the values from class III obesity as those were outliers and were skewing the data – this is an issue that I will be discussing later in the paper. Since BMI is a continuous value, I decided to use that raw BMI value and turned my machine learning methodology to regression problems.

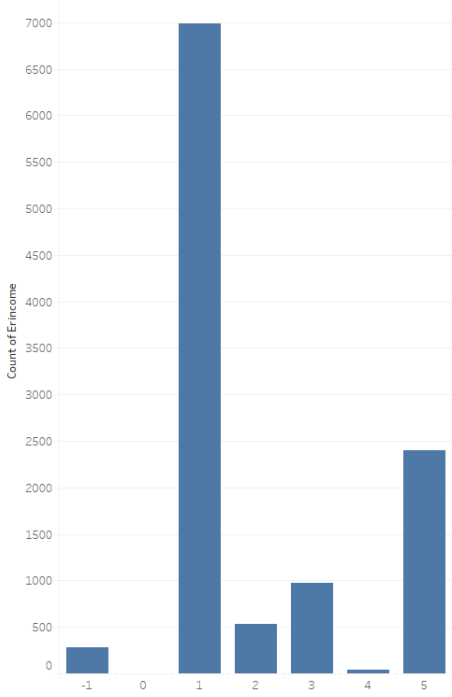

Fig. 4 A distribution of the Income values

A variable of interest to me is Income, which is based on the number of individuals in the household and it is classified as a percentage above or below the poverty line. There are 5 total values that this variable could appear as, 1-5 based on how far above or below the poverty line the family was – see Fig 4 above. It is overall very imbalanced, so I attempted to bin this data into two classes: Class One, with Income > 130% above the poverty threshold and Class Two, with Income <= 130% above the poverty threshold. The balance was 7000 to 4000, which was not turning out goo results. I decided to approach the problem differently and tried to under sample class

one. The overall number of samples decreased from 10,637 to 3,728 but now there is even sampling of those far above the poverty line and those who are not.

2.4 Feature Selection

As I am exploring the factors that are contributing to obesity, I tried many methods of feature selection to determine which one was going to return the best and most clear results. I attempted stepwise recursion – backwards, wrapper select, univariate feature selection with mutual information, and a full-blown wrapper select. The methods that were not returning great results were stepwise recursion – backwards and the full-blown wrapper select. I was not running it on a computer with the capacity for these methods to handle such a large dataset as they are computationally very heavy and were not able to be completed.

The two that I was choosing between was Univariate Feature Selection with Mutual Information and Wrapper Select via Support Vector Regression. They were producing similar results but after running both with the machine learning algorithms, Wrapper Select was outperforming the Univariate Method. After running the feature selection process with this method, there were a few variables that repeatedly showed importance in the models:

Feature	Question from the Survey
eugenth	In general, would you say your physical health was excellent, very good, good, fair or poor?
euexercise	In last 7 days, have you participated in any physical activities?
euffyday	Did you purchase any prepared food yesterday from a deli, carry-out, delivery food or fast food yesterday?
euwic	In last 30 days, did you or any member of your household receive benefits from the Women, Infants, and Children program?
eudrink	Not including plain water, were there any other times yesterday where you were drinking any beverages?
eueat	Were there any times you were eating meals or snacks yesterday, for example, while you were doing something else?

This is not the final features selected for the model, just the most popular features in all the created models. However, the variable of interest, income, is not present in the list of features selected which surprised me after the research I had done. I decided to see what features would be selected with different machine learning models and if income truly had a role in predicting BMI value.

2.5 Machine Learning Methods

The machine learning methodology I used is to try many different models to determine which one would fit my use case the best. After deciding on a model, then diving deeper into how to make that more effective, accurate and predictive.

There were 5 machine learning types that I tested: Random Forests, Gradient Boosting, Ada Boost, Neural Networks, and SVMs. For each of these, I tested two different training methods: Cross Validation, with both 5 and 7 folds, and a Test/Train split, with both 30/70 and 40/60 split. In total, I created 20 models and was able to determine the model and validation methods that worked best for the data and the use case. From there, I narrowed my testing to a specific model to determine the maximizing parameters.

3. Results

3.1 General Results

Of the models that I tested and tweaked, Gradient Boosting performed the best of them – See Table 1 for a more comprehensive breakout of my performance measures. The Neural Networks model performed similarly to the Gradient Boosting model, but the runtime was much slower, so I factored that into my decision.

After determining the superiority of the Gradient Boosting, I tweaked the model to maximize predictiveness and accuracy. The results of my Gradient Boosting models imply interesting conclusions.

Before analyzing the RMSE values, it is important to understand the number. I did not normalize the target BMI variable as it was performing poorly. Therefore, the final RMSE value is representing the range of values in the dataset – so a value of 4.0 is not

great but also not terrible in the scheme of things considering the range of values is nearly 25.0. The final Gradient Boosting model had an RMSE value of 4.45, which is approximately 20% of the range. So not fantastic but it performed the best of all the models.

To output these results, I tweaked the parameters of the model to see how I could do that. In the end, the model parameters are:

- n_estimators = 80, or 80 boosting stages
- loss = 'ls', optimizing with a least squares regression
- min_samples_split=3, a minimum of 3 samples required to split
- max_depth = 2, a maximum depth of 2
- random_state = rand_st, which is just a random number as the seed

These are the parameters I tested to create the best performing model. I came to these numbers because of manually attempting a wide variety of values. The RMSE value of this model is 4.45 (+/- 0.20) and an Explained Variance of 0.10 (+/- 0.04). The standard deviation of the BMI values is 7.86, which is only slightly less than the Explained Variance value which could mean that the model is merely guessing what the BMI value is and thus, the error rate is just the Standard Deviation. I attempted to account for this by removing the Obesity Class III values that represented the outliers in the situation. This model is not fantastic in the grand scheme of model creations, but it performs the best for the use case and the data it is trained on of my testing.

3.1 Feature Selection Results

In the end, I ended up using a Wrapper Select via Support Vector Regression (SVR). From the final model, the selected features are:

Feature	Question from the Survey
eugenth	In general, would you say your physical health was excellent, very good, good, fair or poor?
euexercise	In last 7 days, have you participated in any physical activities?
erincome	Relationship between income and poverty threshold
eusnap	In the past 30 days, did you or any member of this household receive [fill State SNAPNAME], SNAP, or food stamp benefits?
eueat	Were there any times you were eating meals or snacks yesterday, for example, while you were doing something else?

Table 1: A table of performance metrics from tests

Type	*Measure*	CV = 5	CV = 7	Test = 30%	Test = 40%
Random Forests	*RMSE*	4.55 (+/- 0.29)	4.48 (+/- 0.27)	4.4922	4.6226
	Expl. Var.	0.06 (+/- 0.03)	0.10 (+/- 0.07)	0.1185	0.0874
Gradient Boosting	***RMSE***	**4.44 (+/- 0.12)**	4.42 (+/- 0.40)	4.6179	4.6102
	Expl. Var.	**0.10 (+/- 0.07)**	0.10 (+/- 0.06)	0.0517	0.0724
Ada Boost	*RMSE*	4.53 (+/- 0.07)	4.50 (+/- 0.29)	4.5033	4.5759
	Expl. Var.	0.10 (+/- 0.01)	0.09 (+/- 0.03)	0.0746	0.0974
Neural Networks	*RMSE*	4.41 (+/- 0.18)	4.46 (+/- 0.27)	4.4372	4.4969
	Expl. Var.	0.10 (+/- 0.03)	0.10 (+/- 0.04)	0.1296	0.0989
SVMs	*RMSE*	4.47 (+/- 0.33)	4.46 (+/- 0.18)	4.5530	4.4394
	Expl. Var.	0.09 (+/- 0.02)	0.11 (+/- 0.03)	0.0993	0.1283

4. Discussion

4.1 Summarization of Findings

The conclusion of my results is that it is what I expected.

Income, as researched, does appear to be a factor in predicting BMI and is thus correlated to each other. This means that higher BMI values are related to closer percentages to the poverty line; and inversely, lower BMI value are related closer to incomes higher above the poverty line. Another factor related to income is the Snap feature, which indicates if someone in the household is receiving SNAP or food stamps. This is an interesting factor that could correlate even closer to 'food scarcity' as there are limits as to what can be purchased with food stamps. Those on food stamps are more likely to be obese than those who are not [3].

Other features that are predictors of BMI values relate to a general awareness of health. If an individual feels healthy, it is more likely that they are. If an individual is exercising, it is more likely that they are healthier. If an individual is dedicating time to meals and less snacking, it is more likely that they are healthier. These are more of the obvious factors that I expected from the creation of this model.

The Gradient Boosting model has room for improvement to become more predictive and accurate with the predictions.

4.2 Future Work

There are a few action items that I would like to address for future work on this problem.

One of them is to clean up the model and improve the predictive accuracy. A suspicion that I have is that the other features are not balanced correctly and has the potential to be skewing the model. Income is a factor in the final model, but I spent time fixing and Undersampling that which could be a reason as to why it's considered predictive. In addition, I did not pull out a validation test set which is important when I am rebalancing the dataset. I would like to address this in the future.

Most of the focus of this paper was on determining the features that relate most closely to BMI, but this is only accounting for an individual's current situation. I would like to focus more time on predicting whether individuals are on a path of obesity or not – what are the indicators and habits that someone will become obese in the future.

Also, I would like to obtain a more specific income value as opposed to a class based on how above or below an individual is above the poverty line. Other factors of interest to explore are age, gender and the role in the household, whether a parent, child or other relative living in the house. I would like to focus more specifically on the differences in obesity factors between income levels. Do those below the poverty level experience different predictors of obesity compared to those above the poverty line? These are the questions that I would like to focus more on in future work on this problem.

4.3 Conclusion

The 'big picture' conclusion from the research and the paper is that Income, as researched, does appear to be a factor in predicting BMI. The Body Mass Index (BMI) value is a height to weight ratio that estimates percentage of body fat in an individual. There are tiers to BMI, with Obese being the largest values.

As income is a feature of importance in predicting BMI, we can classify they are correlated to each other. While low income is just a red flag, and not the single cause, of the larger social problem of 'food scarcity', this model begins to uncover the hidden realities of the rising obesity rates. The current economic situation is creating income disparities between the lower and upper classes which lead to larger issues such as financial access to nutritional foods. The importance of the SNAP program in the prediction of BMI is an indicator that changes need to happen at more of a foundational, basic needs level to provide low-income citizens with more nutritionally dense food.

Obesity is a risk factor for a multitude of serious health problems as stated previously. By determining the inconspicuous factors that are contributing to our growing population, the society can begin to address the problem at the root – creating a domino effect of change across industries.

References

[1] Shumei Sun Guo, Wei Wu, William Cameron Chumlea, Alex F Roche; *Predicting overweight and obesity in adulthood from body mass index values in childhood and adolescence*; The American Journal of Clinical Nutrition, Volume 76, Issue 3, 1 September 2002, Pages 653–658

[2] George Davey Smith; *A fatter, healthier but more unequal world*; Lancet, 387 (2016), pp. 1349-1350

[3] Kupillas, L. M., & Nies, M. A. (2007). *Obesity and Poverty: Are Food Stamps to Blame?* Home Health Care Management & Practice, 20(1), 41–49.

[4] Tanumihardjo, Sherry A. et al. *Poverty, Obesity, and Malnutrition: An International Perspective Recognizing the Paradox*. Journal of the Academy of Nutrition and Dietetics, Volume 107, Issue 11, 1966 - 1972

[5] Whitaker RC1, Wright JA, Pepe MS, Seidel KD, Dietz WH; *Predicting obesity in young adulthood from childhood and parental obesity;* N Engl J Med. 1997 Sep 25;337(13):869-73.

[6] "Overweight & Obesity Statistics." National Institute of Diabetes and Digestive and Kidney Diseases, U.S. Department of Health and Human Services, 1 Aug. 2017