Cost Effective Multifactorial Prediction of Obesity Prevalence Using Machine Learning Models: A Societal Perspective

Leonardo Jia

Received June 12, 2024
Accepted September 04, 2024
Electronic access September 30, 2024

Background: obesity has grown significantly in the last few decades and become a "pandemic", leading to a broad range of health conditions and huge economic expenditures. Many factors play a role affecting obesity, making this problem complicated. Current literature typically focuses on factors associated with obesity prevalence across socioeconomic status groups, gender or race, or individual obesity classification/prediction. Objective: this paper targeted the obesity population to accurately predict prevalence on the local/regional level in order to assist cost-effective resource allocation for obesity management. Methods: the FCC 2023 data platform was analyzed, including 14 predictive parameters tested in 6 sets, combined with four machine learning (ML) regressors, to predict obesity prevalence of each US county. Results: XGBoost had the most robust and best performance (ranging from r^2=0.83 with only 3 lifestyle parameters to r^2=0.98 with all 14 parameters) across four ML regressors. The optimal model was XGBoost with six socioeconomic and three lifestyle parameters (r^2=0.96), with performance similar to that of the best model. Conclusions: this optimal ML model was accurate and well accommodated large variations in population size and demographics between different counties, making it flexible and scalable in applications. It could be used to help improve economic efficiency in deploying policies and resources in each area for obesity prevention, monitoring and control, as well as evaluate successes without the need to know race, age, gender, diet, health conditions, etc.

Introduction

Obesity is defined by The World Health Organisation (WHO) as "abnormal or excessive fat accumulation that may impair health"[1]. The Center for Disease Control (CDC) specifies overweight and obesity with body mass index (BMI): overweight is BMI of 25.0 to <30, and obesity is BMI of 30.0 or higher[2]. The prevalence of obesity has risen exponentially over the last 50 years and is now so widespread that many have announced a state of "obesity pandemic", owing to an obesogenic environment such as cheap calorie dense food, technologies that reduce or replace physical activity, handy inexpensive nonphysical entertainment, as well as excessive emphasis on low fat intake resulting in excessive intake of simple carbohydrates and sugar, which turns to be stored as fat in the body[3].

The numbers are shocking. Worldwide adult obesity has more than doubled since 1990, and adolescent obesity has quadrupled. In 2022, 2.5 billion adults (18 years and older) were overweight. Of these, 890 million were living with obesity. In 2022, 43% of adults aged 18 years and over were overweight and 16% were living with obesity[1]. Particularly in the US, the 2023 Trust for America's Health report mentioned that 41.9% of adults had obesity. Black and Latino adults had the highest obesity rates at 49.9% and 45.6% respectively. People living in rural communities had higher rates of obesity than people living in urban and suburban areas. Moreover, obesity rates were also increasing among children and adolescents nearly 20 percent in the US[4].

Not surprisingly, many health conditions result from obesity with significant mortality and morbidity including metabolic dysfunction (type 2 diabetes, hypertension, non-alcoholic fatty liver disease, polycystic ovary syndrome, and cardiovascular disease), mental disorders (depression and anxiety), dementia, joint problems (osteoarthritis), chronic kidney disease, obstructive sleep apnea, and at least thirteen types of cancer[5,6,7,8]. Specifically, type 2 diabetes relative risk was increased in those with obesity by a factor of 7[9] as compared to those with normal weight, 6% of cancers diagnosed were attributed to obesity[10] and a 2-fold increase in the risk of hypertension resulted from obesity[11].

Moreover, obesity puts a substantial burden on psychological and psychosocial functioning, and has profound consequences on global health economic expenditure. Obesity showed to have substantial economic impacts. According to the World Obesity Atlas 2023, a global economic impact was estimated to reach $4.32 trillion by 2035, comparable with the impact of Covid-19 in 2020[12]. In the US, diseases due to overweight and obesity were estimated to have $480.7 billion in direct health care costs, and an additional $1.24 trillion in indirect costs due to lost economic productivity, reaching a total of $1.72 trillion[13]. Compared to some other leading chronic diseases, the total estimated cost of diagnosed diabetes in the US in 2022

was 412.9*billion*, *including* 306.6 billion in direct medical costs and $106.3 billion in indirect costs attributable to diabetes [14]. Cardiovascular disease accounts for over $400 billion per year in direct medical spending and indirect costs [15]. Hruby and Hu estimated that for men and women with obesity, there was an additional $1,152 per year and $3,613 per year in medical spending respectively, with a total of $190 billion nationally going to the treatment of obesity and obesity-related diseases [7]. Possible reasons for the gender disparity in cost include a greater risk for women for developing obesity-related physical and psychological comorbidities, resulting in higher associated medical spending [16]. Wells et al. studied 68 countries and demonstrated that gender inequalities were associated with obesity in both developed and developing countries [17].

Previous literature found that obesity disproportionately affects women, non-Hispanic blacks, and Hispanics. Obesity prevalence is also positively associated with low socioeconomic status (SES), physical inactivity/sedentary behavior, lack of sleep, stress, smoking and alcohol consumption, an unhealthy diet and poor surrounding environments (such as high crime and rural areas) - which was also associated with SES, creating complex relationships between each of these characteristics and making this multifactorial disease complicated [4, 18, 19, 20].

SES factors

SES is an important aspect associated with obesity. SES can be determined using variables such as education, income, and occupation, with education considered to be the most stable variable over time. Zhang et al. used data from the National Health and Nutrition Examination Surveys from 1971 to 2000 in the US and found the prevalence of obesity increased among low SES groups, while increasing significantly among high SES groups, thus leading to a reduction in disparities in obesity rates across different SES groups [21]. This trend was consistent across ethnic/racial and gender categories. Odgen et al. used data from the National Health and Nutrition Examination Surveys 2005–2008 in the US and found that among men, obesity prevalence was generally similar at all income levels, but those with higher income were more likely to be obese than those with low income among non-Hispanic black and Mexican-American men. In contrast, higher income women were less likely to be obese than low income women. For obesity and education, there was no significant trend among men, but a trend among women: those with college degrees were less likely to be obese compared with less educated women [22]. In China, Wang et al. found in a retrospective study that education and per capita household income were positively associated with overweight and obesity risk in men but the association between education and obesity status was negative in women. Occupational status was only associated with general overweight in men [23]. Spinosa et al. found that there was a significant indirect effect of socioeconomic status (indicated by income and education level) on BMI via psychological distress and emotional eating. Specifically, lower socioeconomic status was associated with higher distress, which was associated with higher emotional eating, thus with higher BMI [24].

The factor of internet access

In recent decades, a quick advance on the internet has brought confounded effects to obesity. Some studies indicate a positive impact. Particularly, the internet provides access to tremendous amounts of online health information as well as telehealth services. For example, internet access enables the spread of intensive behavioral programs, which helps lower obesity by changing diet and exercise habits, lower the costs for patients and create community settings for the programs. Internet access also encourages technological innovations, such as self-monitoring data, online discussion forums, which allow for online obesity resources [25].

Moreover, broadband internet access is shown to impact the six social determinants of health (healthcare system, economic stability, education, food, social connection, and physical environment). For example, internet connection supports remote meetings necessary for telemedicine. Access to broadband is also extremely important for online job searching to secure employment for economic stability as well as to provide online learning resources to support education. There has also been an increase in online shopping for groceries, which supports food requirements via deliveries. Additionally, internet access allows people to stay connected even while physically distanced. Finally, access to information provided by the internet can be seen as a social determinant of health, as it is necessary to stay up to date with health recommendations during public health crises like the pandemic [26].

Although internet access offers significant benefits, it can also have a negative health impact. It can increase the risk of obesity by promoting sedentary "screen time". Compared to individuals with less than 2 hours of daily screen-time, the adjusted relative risks of obesity were 1.35, 1.70, 1.94 and 1.92 for 2-3, 4-5, 6-7 and 8 hours, respectively. Internet use can be easily addictive, such as watching movies and using social media, which can promote unhealthy lifestyles that reduce exercise, leading to higher chances of overweight and obesity. Snacking while browsing the internet is also another mechanism through which internet access can increase unhealthy behaviors and further the likelihood of obesity.

Predicting obesity

Due to the complex multifactorial nature of obesity, and single factors not having strong correlations with obesity or having different effects on obesity in different sub-populations,

traditional statistical analysis was hard to effectively isolate factors for obesity prediction, so most of previous studies were about factor associations with obesity without a prediction. Machine learning has advantages in processing this kind of complex system. Machine learning has been increasingly used in predicting obesity. However, previous literature mostly focused on using personal data for each individual's obesity risk prediction/classification. Ferdowsy et al. studied more than 1100 factors including gender, age, daily activities, height, weight, diet, stress, insomnia, time on social media, etc. from many varieties of people (both obesity and non-obesity) via applying nine machine learning algorithms for predicting obesity risks. The logistic regression algorithm achieved the highest accuracy of 97.09% as compared to the other classifiers [16]. Rodríguez et al. used data related to the physical condition and eating habits with eight machine learning algorithms to demonstrate the potential to identify people with obesity or overweight. Random forest had the best performance with a 78% accuracy . Thamrin et al. assessed the ability of machine learning methods for obesity classification with Logistic Regression, Classification and Regression Trees (CART), and Naïve Bayes using data of age, education, food and drinks, mental emotional disorders, diagnosed hypertension, physical activity, smoking, and fruit and vegetables consumptions etc. This study found that the Logistic Regression method showed the highest performance, but only moderate concordance between predicted and measured obesity [11].

Literature search revealed a lack of studies on local or regional obesity prevalence prediction, creating a gap to help local/regional obesity management such as resource allocations and distributions. Resource distributions without a good estimate can cause tremendous waste in one area but inadequate obesity management in another. Resources include medical supplies, staff, weight loss programs, etc. For example, needed medical supplies can sit in storage until expired without being used in an area with overestimations, while people with obesity cannot get enough medical supplies thus causing delays or insufficient management in another area with underestimations. This research aimed to tackle obesity from a societal perspective: to develop an accurate prediction model for local/regional obesity prevalence to optimize resource allocation and improve management outcomes. Although Logistic Regression has been shown as a top performer in individual obesity classification with physical conditions and demographic parameters, no literature was found with socioeconomic predictive parameters or broader applications. In addition, Logistic Regression is considered a generalized linear model assuming linearity between the dependent variable and the independent variables. Based on the general performance of ML models, we thus believed XGBoost would be the best one, because XGBoost optimizes the algorithm based upon gradient boosting, such as running a built-in cross validation at each iteration, using Lasso (L1) and Ridge (L2) regularization to counter overfitting and applying "depth-first" tree pruning called "max-depth" to get more optimal trees. Therefore, we wanted to study XGBoost and compare it to a few other gradient boosting ML models. In this paper, four different gradient boosting ML models with six sets of predictive parameters using data from a public database of US counties were trained and tested. This study was limited to using a handful of parameters easy to collect without clinical data and mainly focusing on socioeconomic factors. The optimal model should be flexible and scalable to apply to various demographics and population sizes. Cost and economic benefits of applying this model were also discussed.

Methods

Experimental data collection

Experimental data was collected from a public database: the Federal Communications Commission's Connect2Health dataset from 2023 from a total of 3142 US counties. Fifteen numerical variables of each county were used from the dataset, consisting of fourteen predictive factors and the dependent variable of obesity prevalence (i.e. % BMI>=30). These 14 predictive factors were grouped into three categories: socioeconomic parameters, lifestyle parameters, and demographic and health parameters. No race or diet information was available. Variables were abbreviated for ease of use with definition summarized in Table 1. Note that except for population, all other parameters were proportions. Data quality was checked for outliers, missing values and consistency and was found to be sufficient without further preprocessing needed.

To avoid redundancies, only one proportion parameter was chosen from tightly related parameters. For example, % female was chosen from the four parameters: of female, of male,% female and % male.

Basic statistics (mean, standard deviation, minimum and maximum) of the 15 studied parameters across all counties were summarized in Table 2. Some parameters had a relatively small range, such as unempl rate: 0.7-19.3% (4.0 $\pm$ 1.5%) which was probably because of the relatively stable economic status across the US. Some parameters showed large range but relatively narrow standard deviation, such as % housing problems: 0-69.1% (13.6 $\pm$ 4.5%), which might indicate that although economic inequality still existed among different counties, most counties didn't suffer a lot of housing issues as compared to less developed countries. However, some parameters had a large range and standard deviation, such as population: 86 - 10039107 (104468 $\pm$ 333457), % pop broadband: 0-100% (81.4 $\pm$ 22.2%) and % rural: 0-100% (59.9 $\pm$ 30.9%), which was probably due to large heterogeneity among different counties in size, population, geographic location etc. For example, Los Angeles County is very large and heavily urbanized with a population size of over

Table 1. Abbreviation and definition of each variable

Variable abbreviation	Definition
Predictive Factor Group 1	
% rural	percentage of the population residing in a rural census block
% poverty	percentage of population whose household income is below the poverty line
unempl rate	percentage of the civilian labor force aged 16 and older who are unemployment but seeking work
some college	percentage of adults with some post-secondary education
% housing problems	percentage of households with at least 1 of 4 housing problems of overcrowding, high housing costs, lack of kitchen facilities, or lack of plumbing facilities
% pop broadband	percentage of population living in census blocks with access to fixed broadband service at 25/3 Mbps or higher
Predictive Factor Group 2	
% smokers	percentage of adults who are current smokers
% drinkers	percentage of adults reporting binge/heavy drinking during the last 30 days
% phys inactive	percentage of adults who report no leisure time physical activity
Predictive Factor Group 3	
population	estimate of resident population
% female	percentage of population identified as female
% 65+	percentage of the population aged 65 and older
% diabetes	percentage of adults aged 20 and above with diagnosed diabetes
% poor health	percentage of adults considered to be in poor or fair health (age-adjusted)
Target Variable	
% BMI>=30	percentage of adults (aged 20 and above) that report a BMI greater than or equal to 30 kg/m2

Table 2. Summary of studied parameters. From left to right: parameter abbreviation, mean, standard deviation (SD), minimum (Min) and maximum (Max).

Parameter	Mean	SD	Min	Max
% rural	59.8	30.9	0	100
% poverty	14.5	5.8	2.7	47.7
unempl rate	4	1.5	0.7	19.3
some college	58.1	11.9	0.8	100
% housing problems	13.6	4.5	0	69.1
% pop broadband	81.4	22.2	0	100
% smokers	21.3	4.2	7.1	44.6
% drinkers	19.1	3.4	6.5	31
% phys inactive	26.7	5.8	8.9	50.4
population	104468	333457	86	10039107
% Female	49.9	2.3	26.5	57
% 65+	19.8	4.8	4.9	58.2
% diabetes	12.3	3.7	2.4	29.5
% poor health	20.1	5.1	8.6	41.9
% BMI>=30	33.4	6	11	58.9

10 million, while Monroe County is small and fairly rural with a population of 6701.

Data Analysis

Pearson correlation Pearson correlation was used to determine a relationship between two variables (e.g. % poverty and % BMI>=30):

$$r = \frac{\sum(x_i - X)(y_i - Y)}{\sqrt{\sum(x_i - X)^2 \sum(y_i - Y)^2}}$$

where r is the correlation coefficient, x_i (y_i) is the ith value of the x-variable (y-variable), X (Y) is the mean of the values of the x-variable (y-variable). Based on Akoglu's "User's guide to correlation coefficients"[27], Pearson correlation coefficient's absolute value |r| was categorized in three levels: poor (0-0.2), moderate (>0.2 and <0.8) and strong (0.8-1). Parameters with a "strong" correlation with % BMI>=30 might play a more significant role in the prediction model than parameters with a "moderate" or "poor" correlation with % BMI>=30. The main purpose of using Peason correlation was to investigate which predictive parameters had poor/weak correlations with % BMI>=30 in order to reduce the number of predictive parameters included in a model for practical uses and those having poorest/weakest correlations with % BMI>=30 would be considered first. Before eliminating a parameter, the parameter was plotted against % BMI>=30 to confirm that the weak association was not due to a nonlinear relationship.

Prediction - Multivariable Regression

Each model was an ML algorithm combined with a predictive parameter set. A classic 5 fold cross-validation was used with 80% of the data for training and the remaining 20% for testing repeating 5 times. The symmetric mean absolute percentage error (SMAPE) was used to evaluate error (i.e. each model was trained to minimize the SMAPE):

$$\text{SMAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|P_i - A_i|}{\left(\frac{|A_i| + |P_i|}{2}\right)}$$

where A_i is the ith actual value, P_i is the ith predicted value and n is the total number of observations. There were many options of error metrics, SMAPE was chosen because it's not scale dependent.

Machine learning regressors

Four different open-source gradient boosting algorithms were used as the regressors: Gradient Boosting Trees (GBT), Categorical Boosting (CatBoost), Light Gradient Boosting Machine (LGBM), and Extreme Gradient Boosting (XGBoost). Training and testing were carried out by utilizing Symon.AI web-based analytics platform. No manual hyperparameter tuning was performed. Predicted obesity prevalence values were output in an excel spreadsheet for comparison with the actual obesity prevalence values from the database. Pearson correlation r^2 of the predicted and actual obesity prevalence (i.e. % BMI>=30) was studied to evaluate performance of the model, each of which was an ML algorithm with a set of predictive parameters, such as XGBoost with 6 socioeconomic parameters.

- **GBT:** GBT combines several weak learners into strong learners, in which each new model is trained to minimize the loss function or error function such as mean squared error of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize

this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met. GBT frequently makes use of decision trees. It is predicated on the hypothesis that when merged with earlier models, the best next model will minimize the overall prediction error[28].

- **CatBoost:** CatBoost is a decision tree gradient boosting technique that takes extremely little time to predict. Symmetric trees, also known as balanced trees, are used in CatBoost to describe trees where the splitting condition holds true for all nodes at the same level of the tree[28].
- **LGBM:** LGBM is a dispersed, strong gradient boosting framework based on the decision tree method and it requires less RAM to run while handling massive data sizes. When compared to other algorithms, LGBM grows trees vertically, or leaf-wise, as opposed to other algorithms, which grow trees horizontally[28].
- **XGBoost:** XGBoost is commonly known to offer smart solutions to structured data problems through the implementation of the gradient boosted trees technique. XGBoost reduces a formalized objective function by merging a convex loss function based on the difference between the observed and target outputs with a weighting parameter for model computational complexity[28].

Results

Pearson correlation between parameters Pearson correlation was calculated between parameters to study how each predictive parameter was associated with obesity prevalence (for both the strength and direction). Pearson correlation between socioeconomic parameters and % BMI>=30 (Figure 1a) showed that unempl rate, % poverty and % rural all had positive associations with % BMI>=30, whereas % pop broadband, % housing problems and some college had negative associations with % BMI>=30. In short, socioeconomic parameters had at best a moderate correlation (r$\leq$ 0.4) with % BMI>=30. This indicated no single socioeconomic parameter had a dominant effect on obesity prevalence, thus combinations were needed in predictions. When considering parameters to be eliminated in the prediction models, % housing prob, % rural and % pop broadband could be included due to weak correlations with % BMI>=30 (r=0.1 or -0.1). Pearson correlation was also calculated between lifestyle, demographic and health parameters, and % BMI>=30 (Figure 1b), which showed that % smokers, % phys inactive, % female, % diabetes, and % poor health all had positive associations with % BMI>=30, whereas % population, % 65+ and % drinkers had negative associations with % BMI>=30. Overall, compared to socioeconomic parameters, all three lifestyle parameters (% smokers, % phys inactive and

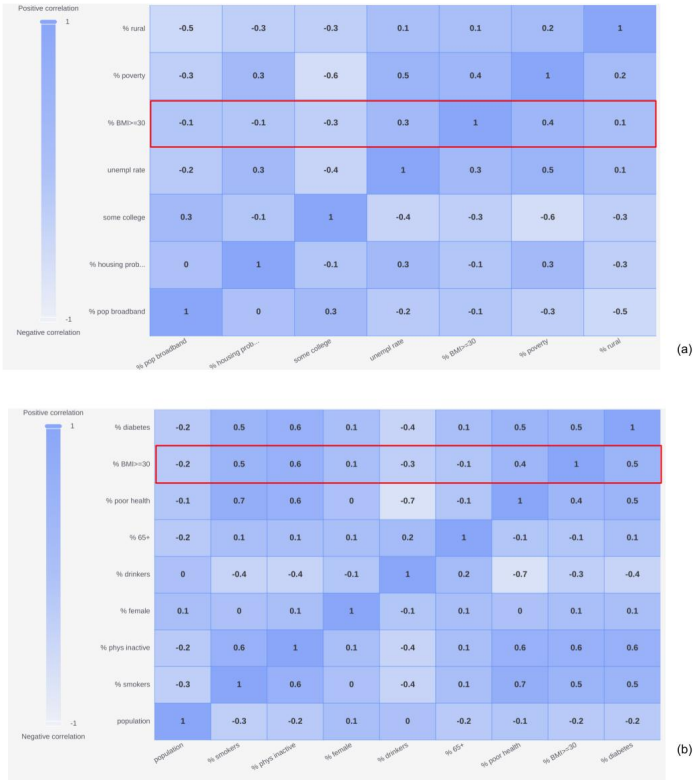

Figure 1. Pearson correlation heat map of (a) socioeconomic parameters and % BMI >=30, and (b) lifestyle parameters, demographic and health parameters, and % BMI >=30. Range of numbers: -1 to 1, 0: the weakest correlation, and +/-1: the strongest positive/negative correlation. Red box: correlations of parameters with % BMI>=30. All correlations with %BMI>=30 were statistically significant (p«0.05).

% drinkers) and two health parameters (% diabetes and % poor health) had stronger but still moderate correlations ($r \leq 0.6$) with % BMI>=30. Similar to socioeconomic parameters, no single lifestyle, demographic or health parameters had a dominant effect on obesity prevalence, leading to the needs of parameter combinations in predictions. In addition, % female and % 65+ had weak correlations with % BMI>=30 and could be eliminated in the prediction models as needed.

Performance of ML models

In order to determine the optimal parameter set for prediction based on socioeconomic factors, group 1 (6 socioeconomic parameters) and group 2 (3 lifestyle parameters) were studied separately, and then in combination together, as well as with group 3. Moreover, in order to reduce parameter numbers, parameters with weakest correlations with obesity prevalence were eliminated during combinations. In general, using more predictive parameters might make better predictions than

using less parameters. Parameters from the same group might have closer associations with each other (e.g. % poverty and unemply rate) thus affecting obesity prevalence less independently. Therefore, combining parameters from different groups might make predictions more robust. In addition, with the same number of predictive parameters, choosing those with stronger correlations with obesity prevalence might have better predictions. As a result, six parameter sets were formed (see details of which parameters are included in each set in Table 3). Each of the six parameter sets combined from the 14 predictive parameters with each of the four gradient boosting algorithms: GBT, CatBoost, LGB and XGBoost was a model under evaluation. The number of parameters as well as correlation strengths with obesity prevalence were analyzed.

The prediction performance (shown by r^2) was summarized in Table 3. Among the ML models, XGBoost produced the highest r^2 across various parameter sets, followed by LGBM, CatBoost and GBT. Across all parameter sets, overall, a set with more parameters/stronger parameters was better than a set with less parameters/weaker parameters. This was demonstrated by Set 6 (with 14 parameters across all three groups) with the best performance vs. Set 2 (only 3 parameters in one group) with the worst performance regardless of ML models. Moreover, comparing Set 1 and Set 3 (both with 6 parameters), Set 3 had stronger parameters and achieved overall better performance. However, comparing Set 4 and Set 5 (both with 9 parameters), the performance was very similar (note that Set 4 had stronger parameters), which indicated that a combination of 6 socioeconomic and 3 lifestyle parameters probably was good enough as an optimal set in real world applications. Although the highest performance was achieved by XGBoost with Set 6 ($r^2 = 0.98$), the optimal parameter set was determined to be Set 5, excluding health and demographic parameters due to greater difficulty in gathering the data and similar performance. In short, XGBoost with parameter Set 5 was the optimal model ($r^2 = 0.96$).

Predicted vs. actual obesity prevalence for all four ML models with the optimal parameter Set 5 were shown in Figure 2. Figure 3 showed XGBoost performance across all 6 different parameter sets.

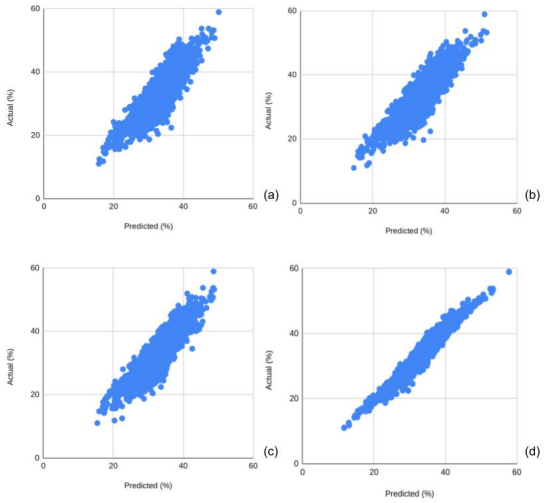

Figure 2. Predicted vs. actual obesity prevalence (%) for various ML models with parameter set 5: (a) GBT ($r^2 = 0.71$), (b) CatBoost ($r^2 = 0.77$), (c) LBGM ($r^2 = 0.80$) and (d) XGBoost ($r^2 = 0.96$).

Discussion

Different from individual's obesity classification/prediction, where each individual's physical condition and eating habits information is critical in order to pinpoint individual's risks and targeted obesity management, regional obesity prevalence prediction is mainly to help resource distribution such as appropriate amounts of medical supplies, staff, weight loss programs, etc. where the resource agencies don't need to know who has obesity but only what percentage of people need the resources. Therefore, more societal levels of information such as unemployment rate, poverty rate etc. plays an important role in prediction. In this study, high prediction accuracy (r^2=0.96) is demonstrated for regional obesity prevalence by using XGBoost with a small set of 6 socioeconomic and 3 lifestyle parameters, although each of these parameters has at most moderate correlations with obesity prevalence individually. The parameter set can be easily obtained by a simple 9 yes-no question survey, without asking for any clinical or demographic information. Although there are large variations in population size and demographics in different counties, this prediction model has accommodated all these factors pretty well: good performance (predicted vs. actual obesity prevalence) held across counties where large variations existed in population size: e.g. Los Angeles County with max population size of 10039107 (21.8% vs. 22.1%) and Kalawao County with min population size of 86 (13.0% vs. 11.8%), senior proportion: e.g. Sumter County with max senior proportion of 58.2% (29.8% vs. 29.3%) and Chattahoochee County with min senior proportion of 4.9% (40.5% vs. 43.0%), female proportion: e.g. Pulaski County with max female proportion of 57% (43.8% vs. 45.1%) and Crowley County with min female proportion of 26.5% (30.4% vs. 29.4%) as well as race dominance: e.g. Leslie County with White dominance (40.6% vs. 40.7%), Prince George's County with Black dominance (37.2% vs. 37.9%), Honolulu County with Asian dominance (24.4% vs. 24.4%), and Los Angeles County with Hispanic dominance (21.8% vs. 22.1%). Note that most predicted vs. actual obesity prevalence differences were no more than 1.3%, except for one: 2.5%.

The rough cost to implement this prediction model is about

	Set 1: 6 factors (group 1 only)	Set 2: 3 factors (group 2 only)	Set 3: 6 factors (three from group 1 + group 2)	Set 4: 9 factors (#3 + 3 from group 3)	Set 5: 9 factors (groups 1+2)	Set 6: 14 factors (groups 1+2+3)
GBT	0.61	0.59	0.65	0.72	0.71	0.77
CatBoost	0.65	0.6	0.7	0.78	0.77	0.83
LGBM	0.69	0.62	0.73	0.8	0.8	0.85
XGBoost	0.84	0.83	0.92	0.96	0.96	0.98

Table 3. Pearson correlation r^2 of predicted and actual % BMI>=30 for each model. Set 3 was the combination of three strong parameters in group 1 (some college, % poverty, unempl rate) and group 2, and Set 4 was Set 3 plus %diabetes, % poor health and population from group 3. All correlations were statistically significant (p«0.05).

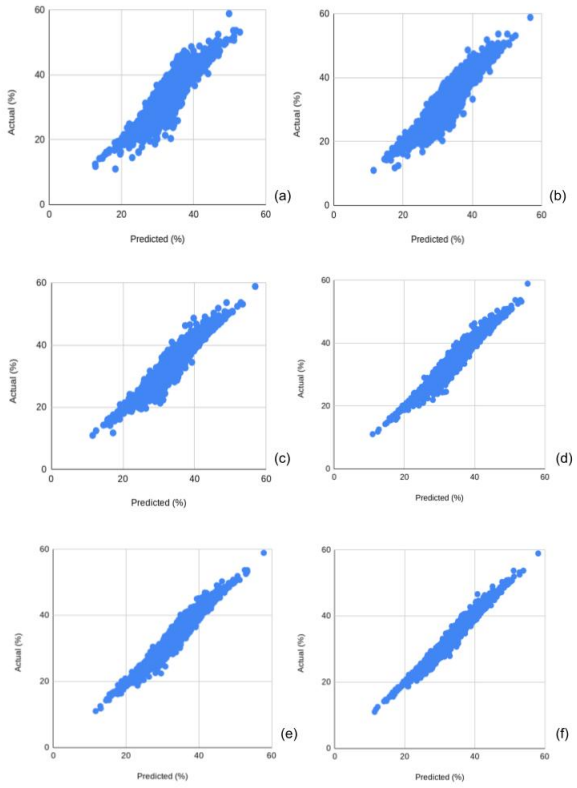

Figure 3. Predicted vs. actual obesity prevalence (%) for XGBoost across six parameter sets (a) Set 1 using 6 socioeconomic parameters ($r^2 = 0.84$), (b) Set 2 using 3 lifestyle parameters ($r^2 = 0.83$), (c) Set 3 using 3 socio-economic parameters and 3 lifestyle parameters ($r^2 = 0.92$), (d) Set 4 using 3 socioeconomic parameters, 3 lifestyle lifestyle parameters, and 3 health and demographic parameters ($r^2 = 0.96$), (e) Set 5 using 6 socioeconomic parameters and 3 lifestyle parameters ($r^2 = 0.96$) and (f) Set 6 using all 14 parameters ($r^2 = 0.98$).

$200. It is very helpful to provide a cost-effective estimate of needed resources in an area. For example, corresponding diet and exercise programs can arrange their capacities and distributions based on the size of needs in the area. Restaurants and grocery stores can estimate proportions of food for obesity needs. Local governments can also help lift the socioeconomic status, by providing more job opportunities and further education accordingly. Since obesity is a significant risk factor for many health conditions such as diabetes and cardiovascular disease, effective preventions and controls with low cost can have a tremendous economic impact as well. Using the data from obesity literature: $480.7 billion in direct healthcare costs in the US [13] even if 10% of US regions apply this prediction model to save 10% of costs, this model will save $4.8 billion of costs, not including indirect costs associated with obesity.

Besides the main findings, there were some interesting aspects of this research. Binge drinker prevalence (% drinkers) showed a negative association with obesity prevalence. Although consistent with previous literature, it was contradictory with the common knowledge that alcohol could cause weight gain. A possible explanation could be that % drinkers was negatively correlated with % smokers (r=-0.4), % phys inactive (r=-0.4), % poor health (r=-0.7) and % poverty (r=-0.5) which were all positively correlated with obesity prevalence, meaning that those binge drinkers had better lifestyles, better health and more money to take care of obesity, which outweighed the negative effects of binge drinking. Collins also found that with respect to household income, binge-drinking prevalence was highest among those with the highest income (> $75,000) [25]. Moreover, % pop broadband only had a poor correlation with decrease of % obesity (r = -0.1), likely because of no breakdowns of how individuals used the internet for. For example, whether the internet was mainly used for telehealth/exercise (i.e. related resources helping obesity) or games/movies (i.e. sedentary behaviors with high carb/sugar snacking worsening obesity), could have a totally opposite impact. As more e-health and obesity educational programs are introduced to people, this correlation can potentially be more significant.

There are a few limitations which can be considered to improve in future studies. First of all, there were no breakdowns in Internet use, such as the examples mentioned in the previous paragraph. Otherwise, correlation in each subcategory with obesity prevalence should be stronger and could be used for a more refined prediction model. In addition, there was no diet information, which could have probably increased accuracy in obesity prediction. However, although it's part of the lifestyle factors, diet information could be less practical to collect since it's changing constantly and hard to self-estimate accurately.

Moreover, there was no data on overweight prevalence, which is much bigger than obesity prevalence1 and is easier to reverse via appropriate diet and exercise. Predicting regional overweight prevalence could have a greater health and economic impact, which is very feasible under this prediction model's framework. Local programs and actions can be taken before overweight gets to obesity, which further helps alleviate the huge health and economic burdens associated with obesity. Furthermore, no manual hyperparameter tuning was performed to maximize ML model performance, thus the results could be suboptimal. However, XGBoost has achieved satisfactory results regardless. Finally, self-reported survey data could potentially have some inaccuracies which could make a county's prevalence calculation less accurate, although the survey questions were clear yes-or-no types like "are you a current smoker?"

Conclusion

Previous literature has studied many factors associated with obesity but focused mainly on individual's obesity classification instead of a regional obesity population prediction. This research filled in the gaps from a societal perspective, in order to assist local/regional obesity management with economic efficiency. Four different gradient boosting machine learning algorithms combined with six parameter sets were studied in each US county. The optimal prediction model was XGBoost with a set of 6 socioeconomic and 3 lifestyle parameters. This model could be potentially used in real world programs for cost-effective obesity prevention, monitoring and control.

Acknowledgments

I would like to acknowledge Ms. Cecilia Xi for guiding me through the process of conducting this research.

References

1 W. H. Organization, *Obesity and overweight*, 2024, https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

2 C. for Disease Control, *Overweight and obesity*, 2010, https://www.cdc.gov/ncbddd/disabilityandhealth/documents/obesityfactsheet2010.pdf.

3 D. R. Meldrum, M. A. Morris and J. C. Gambone, *Fertility and Sterility*, 2017, **107**, 833–839.

4 T. for America's Health, *State of Obesity 2023: Better Policies for a Healthier America*, 2023, https://www.tfah.org/report-details/state-of-obesity-2023/.

5 B. Lauby-Secretan, C. Scoccianti, D. Loomis, Y. Grosse, F. Bianchini and K. Straif, *New England Journal of Medicine*, 2016, **375**, 794–798.

6 G. M. Singh, G. Danaei, F. Farzadfar, G. A. Stevens, M. Woodward, D. Wormser, S. Kaptoge, G. Whitlock, Q. Qiao, S. Lewington, E. D. Angelantonio, S. V. Hoorn, C. M. M. Lawes, M. K. Ali, D. Mozaffarian and M. Ezzati, *PLOS One*, 2013, **8**, e65174.

7 A. Hruby and F. B. Hu, *Pharmacoeconomics*, 2015, **33**, 673–689.

8 M. Blüher, *Nature Reviews Endocrinology*, 2019, **15**, 288–298.

9 A. Abdullah, A. Peeters, M. Courten and J. Stoelwinder, *Diabetes Research and Clinical Practice*, 2010, **89**, 309–319.

10 A. P. Polednak, *Cancer Detection and Prevention*, 2008, **32**, 190–199.

11 D. Thompson, J. Edelsberg, G. A. Colditz, A. P. Bird and G. Oster, *Archives of Internal Medicine*, 1999, **159**, 2177–2183.

12 E. Mahase, *BMJ*, 2023, **380**, 523.

13 M. Institute, *Economic impact of excess weight now exceeds* 1.7*trillion* : *Costsinclude1.24 trillion in lost productivity, according to study documenting role of obesity and overweight in chronic diseases*, 2018, https://www.sciencedaily.com/releases/2018/10/181030163458.htm.

14 E. D. Parker, J. Lin and T. Mahoney, *Diabetes Care*, 2024, **47**, 26–43.

15 K. E. Joynt Maddox, *Circulation*, 2024, **150**, 419–421.

16 N. Kapoor, S. Arora and S. Kalra, *Journal of Midlife Health*, 2021, **12**, 103–107.

17 J. C. Wells, A. A. Marphatia, T. J. Cole and D. McCoy, *Social Science Medicine*, 2012, **75**, 482–490.

18 C. M. Hales, C. D. Fryar, M. D. Carroll, D. S. Freedman and C. L. Ogden, *JAMA*, 2018, **319**, 1723–1725.

19 P. H. Chyou, C. M. Burchfiel, K. Yano, D. S. Sharp, B. L. Rodriguez, J. D. Curb and A. M. Nomura, *Annals of Epidemiology*, 1997, **7**, 311–317.

20 A. Lee, M. Cardel and W. T. Donahoo, *Social and Environmental Factors Influencing Obesity*, MDText.com, Inc, 2000.

21 Q. Zhang and Y. Wang, *Obesity Research*, 2004, **12**, 1622–1632.

22 C. L. Ogden, M. M. Lamb, M. D. Carroll and K. M. Flegal, *Obesity and Socioeconomic Status in Adults: United States, 2005–2008*, 2010, https://www.cdc.gov/nchs/data/databriefs/db50.pdf.

23 K. Wang, C. Wu, Y. Yao, S. Zhang, Y. Xie, K. Shi and Z. Yuan, *Global Health Research and Policy*, 2022, **7**, year.

24 J. Spinosa, P. Christiansen, J. M. Dickson, V. Lorenzetti and C. A. Hardman, *Obesity*, 2019, **27**, 559–564.

25 M. J. Hutchesson, M. E. Rollo, R. Krukowski, L. Ells, J. Harvey, P. J. Morgan, R. Callister, R. Plotnikoff and C. E. Collins, *Obesity Reviews*, 2015, **16**, 376–392.

26 N. C. Benda, T. C. Veinot, C. J. Sieck and J. S. Ancker, *American Journal of Public Health*, 2020, **110**, 1123–1125.

27 H. Akoglu, *Turkish Journal of Emergency Medicine*, 2018, **18**, 91–93.

28 M. Fatima and M. Pasha, *Journal of Intelligent Learning Systems and Applications*, 2017, **9**, 1–16.