Predicting the Chance of Developing Obesity or Type II Diabetes

Northwestern University - EECS 349

Alex Grimes[1] and Mikayla Litt[2]

[1]alexandragrimes2019@u.northwestern.edu
[2]mikaylalitt2019@u.northwestern.edu

Abstract

The United States is currently the most obese country in the world, with over a third of the population weighing in well above healthy rates [2]. Obesity also causes a slew of other health problems, including diabetes. Americans should to be able to know their risk for these potentially life-threatening conditions, and what features lead to a higher risk for these prevalent health problems. Using different machine learners, such as neural nets and decision trees trained on the food availability data from the United States Department of Agriculture [4], we can help overweight Americans asses their risk for further health issues.

Our research led us to a few interesting conclusions. Neural nets performed the best for this prediction task with very high correlation for obesity and diabetes on the test set, 0.7665 and 0.9001 respectively. Decision trees also performed quite well on the test set with correlations of 0.7667 and 0.872. The most important factors to predicting obesity and diabetes rates were poverty rates, average household income, and percentage of students eligible for free lunch. As these factors are all strongly related to financial status of a community as a whole, it brings into question bigger problems about the correlation between poverty, effective government assistance, and making healthy options more accessible.

Introduction

Obesity and diabetes rates across the United States are increasing at an alarming rate [1]. Americans are falling susceptible to fast food branding and convenience, and the US government has started a plethora of health initiatives to help people live healthier lives [3]. As optimistic as these programs are, health issues in the US are still on the rise.

Machine learning can enable us to predict a persons risk for developing obesity or diabetes later in life and well as analyze the factors that have the highest influence. Additionally, the United States Department of Agriculture publishes food environment data for counties around the country, providing a means for developing our models and comparing relevant features.

Through our work, we aimed to help identify the features that most highly correspond to higher obesity and diabetes rates, in the hope to provide Americans with insight about their health risks and food decisions. Here, we propose that there is a strong relationship between financial status and obesity and diabetes risks and that there may be a bigger underlying problem to the increasing health problems in the US.

Sample Collection

The data set we used was acquired from the United States Department of Agriculture, which records data on the food environment of 3,144 counties for various years between 2009-2014, including the obesity rates and diabetes rates of the counties. Each county was used as an example for our machine learning algorithms to be trained and tested on, each with characteristics of the food environment. Food environment included many features from access and proximity to different food sources, (i.e., grocery stores, fast-food restaurants, full-service restaurants, etc.) to median household income and socioeconomic data. The initial dataset had over 200 attributes, many of which were missing values for a good portion of the examples. We decided to narrow down the number of attributes we would be looking at to 30 by eliminating features with many missing values or that did not prove to be important predictors from Wekas preprocessing tool.

After processing the data, we set aside 100 random samples for testing our models. Since we considered obesity and diabetes predictions separately, we had two pairs of training and test sets, one for each output. For validation, we trained all of our models with 10-fold cross-validation.

Algorithms

In our initial approach, we built decision tree models for predicting obesity rates and diabetes rates using Wekas M5P algorithm on our training sets. These models were then evaluated on the corresponding test sets. The correlation coefficients of the training and test data for obesity and diabetes rates can be seen in Figure 1. Next, we built nearest neighbor models for our data using Wekas IBk nearest neighbor algorithm, the training and testing results of which can again be seen in Figure 1. Finally, the last models we built were

neural network models using Wekas MultilayerPerceptron algorithm.

Obesity Rates (30 attributes)					
M5P		**IBk**		**Multilayer Perceptron**	
10-fold CV	0.7988	**10-fold CV**	0.7006	**10-fold CV**	0.7618
Test	0.7667	**Test**	0.7293	**Test**	0.7665
Diabetes Rates					
M5P		**IBk**		**Multilayer Perceptron**	
10-fold CV	0.8325	**10-fold CV**	0.7976	**10-fold CV**	0.8449
Test	0.872	**Test**	0.773	**Test**	0.9001

Figure 1: Correlation coefficients for models using 30 attributes.

Obesity Rates (4 attributes)					
M5P		**IBk**		**Multilayer Perceptron**	
10-fold CV	0.5552	**10-fold CV**	0.4313	**10-fold CV**	0.4194
Test	0.5201	**Test**	0.1637	**Test**	0.4528
Diabetes Rates					
M5P		**IBk**		**Multilayer Perceptron**	
10-fold CV	0.6786	**10-fold CV**	0.5448	**10-fold CV**	0.6225
Test	0.7043	**Test**	0.5063	**Test**	0.6923

Figure 2: Correlation coefficients for models using 4 attributes.

Results

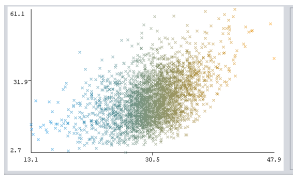
The decision tree algorithm produced the highest correlation coefficient predictive of obesity rates for the test data, although the correlation coefficient of the neural network algorithm for the test set was extremely close. For diabetes rates, decision trees and neural networks also produced the higher correlation coefficients than did the nearest neighbor algorithm, however, in contrast to predicting obesity rates, the neural network correlation coefficient was slightly higher in this case than the decision tree correlation coefficient.

Upon closer examination of the results, we found that four attributes in particular seemed to have high direct correlation with the output variables. Overall poverty rate, child poverty rate, average household income, and percentage of students eligible for free lunch were the four factors we observed. Creating models with only these four attributes yielded results not quite as accurate as with all 30, but diabetes predictions did particularly well, having correlation coefficients for neural nets and decision tree of 0.6923 and 0.7043 respectively. See Figure 2 for more comparisons between the 30-featured and 4-featured models.

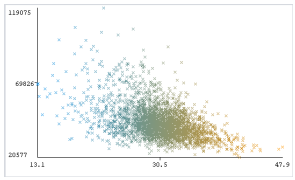
Analysis

Neural networks and decision trees proved to be very useful and perform well for our prediction task. Neural networks can learn complex functions and relationships, and were able to ignore the attributes in the feature set that did not have direct relationship to the output. Because of this, they consistently performed better on the test set than with 10-fold cross-validation. Additionally, they worked well because our task has no limits with training time or efficiency. We think that with more fine grain data, say town-wise instead of county-wise data, and more samples we could achieve even greater accuracy with neural nets.

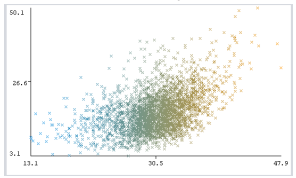
Decision trees did equally as well as neural nets with all attributes, and in the case where we used only the four most relevant attributes, decision trees were consistently the best predictor. Additionally, decision trees provided valuable insight into the features that most influenced a persons risk for obesity or diabetes, a main aim for our research. While nearest neighbor was included in our experimentation, we discovered for this task they did not perform nearly as well as decision trees or neural nets. We believe this is because of the large number of irrelevant features in the data set.



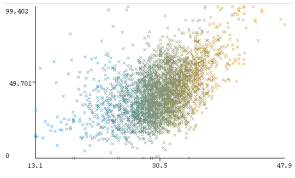
(a) Child Poverty



(b) Household Income

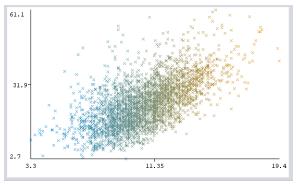


(c) Poverty

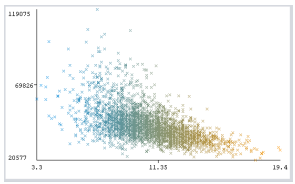


(d) Free Lunch

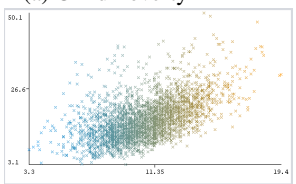
Table 1: Relevant attributes as they relate to obesity



(a) Child Poverty



(b) Household Income



(c) Poverty

(d) Free Lunch

Table 2: Relevant attributes as they relate to diabetes

Conclusion

When we originally set out to investigate obesity and diabetes rates in relation to food environments, we hypothesized that features of counties like the number of fast-food

restaurants they have, the number of participants in government food initiatives, etc., would be most predictive of obesity rates and diabetes rates. In the end though, we found that the main underlying feature that predicts rates of obesity and diabetes in a county appears to be financial status in general, which is closely associated with poverty rates and median household income. This suggests that the rising obesity and diabetes rates plaguing the United States is not going to be fixed simply by promoting healthy lifestyles and government-funded programs, but reflects more largely on an issue between poverty struggles and government intervention and help.

Future work on this projects subject area could dive further into the relationship between poverty rates and obesity and diabetes rates. If poverty rates are increasing in parallel to obesity and diabetes rates, it would provide even more reason to further investigate a potential correlation between the three. It would also be interesting to broaden the comparisons to a global scale to see how the rates of poverty, obesity, and diabetes in the United States are increasing relative to those same rates in other countries.

Contributions

For this research project, Alex Grimes conducted data collection and Mikayla Litt completed the data preprocessing. Work was done jointly to create the final models, conduct analysis, and complete the report.

References

CENTER FOR DISEASE CONTROL. Long term trends in diabetes, April 2017.

RENEW BARIATRICS. Report: Obesity rates by country - 2017, September 2017.

UNITED STATE DEPARTMENT OF AGRICULTURE. Obesity prevention healthy weight programs, 2012.

UNITED STATE DEPARTMENT OF AGRICULTURE: ECONOMIC RESEARCH SERVICE. Data access and documentation downloads, August 2015.