Assessing the Impact of Social Determinants on Obesity Rates Across U.S. Counties: Machine Learning Analysis

Yuhan Ma

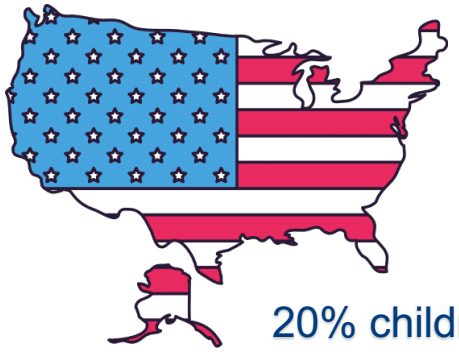
Data Science Final Project

GEORGETOWN UNIVERSITY

Agenda

- Introduction
- Research Goal
- Data Selection and Modeling
- Results
- Reflection and Policy Recommendation

Introduction



20% children
33% adults
Struggle with
Obesity!

147 billion spent on
Obesity!

40% households
have limited access
to get healthy food

Majorities don't meet
physical exercise standard

Research Goal

This project employs **machine learning to predict obesity rates across U.S. counties**, aiming to provide data-driven insights that help policymakers craft targeted, sustainable public health interventions. Through this approach, we seek to enhance the understanding of how **societal factors influence obesity** prevalence and to support the development of effective strategies to combat this pressing public health issue.

Data Selection and Methodology

Social Determinants of Health

County-level estimates of diabetes, physical inactivity, and obesity within the context of important social factors.

Data Resources

Social Determinants:

- Economics
- Food Environment
- Healthcare
- Physical Environment
- Housing-Neighborhood
- Urban-Rural
- Social Vulnerability Index (SVI)

Prediction Target:
Obesity (Percentage) 2021

Machine Learning Model

- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regression
- XGBoost Regression
- Support Vector Regression

Descriptive Stats

Variable	Count	Mean	Std	Min	50%	Max
Obesity (Percentage)	3065	28.17	6.52	12.20	28.60	46.50
Children in Poverty (Percent)	3065	20.00	8.43	2.80	19.00	58.50
Overall_SVI	3065	0.17	18.06	-999.00	0.49	1.00
Fast_Food_Restaurant	3065	0.62	0.28	0.04	0.60	5.81
soda_tax	3065	4.71	2.00	0.00	5.50	7.00
Primary_Care_Physicians_Rate_per_100	3065	55.24	34.82	3.52	50.20	576.43
No_Health_Insurance_Percent	3065	11.41	5.23	2.39	10.20	38.71
Severe_Housing_Cost_Burden_Percent	3065	10.63	3.50	0.69	10.15	31.25
Access_to_Exercise_Opportunities_Percent	3065	60.82	23.30	0.03	62.70	100.00
Isolation_Index_Hispanic	3065	0.13	0.15	0.00	0.07	0.99
Isolation_Index_Non_Hispanic_Asian	3065	0.03	0.05	0.00	0.02	0.52
Isolation_Index_Non-Hispanic_Black	3065	0.14	0.18	0.00	0.06	0.88
Isolation_Index_Non-Hispanic_White	3065	0.79	0.17	0.03	0.85	1.00
rural_urban_encoded	3065	0.43	0.49	0	0	1

Endogeneity

Endogeneity in my model stems from potential **bidirectional causation** between obesity rates and predictors such as fast food density and healthcare access, complicating the identification of direct effects.

- **Bias**: A simpler model may inaccurately approximate complex realities
- **Variance**: Highly flexible models are sensitive to training data nuances, which can cause overfitting and high variance if not properly regulated.
- **Interpretability**: Complex models act as "black boxes"

Results

Stage one: Without ”Sates”

Model	Data Set	R-Squared	MSE	MAE	RMSE	
Linear	Training	0.1672	34.8958	4.8568	5.9073	Too simple
	Test	-870.7886	38920.8649	12.9694	197.2837	
Lasso (α=0.5)	Training	0.1017	37.6394	5.1134	6.1351	Too simple
	Test	0.1179	39.3827	5.2067	6.2756	
Ridge (α=0.5)	Training	0.1672	34.8969	4.8577	5.9074	Too simple
	Test	-856.7761	38295.2802	12.9051	195.6918	
Random Forest	Training	0.9185	3.4150	1.4548	1.8480	Overfitting
	Test	0.4507	24.5226	3.9204	4.9520	
XGBoost	Training	0.9843	0.6578	0.5856	0.8111	Overfitting
	Test	0.4472	24.6790	3.9258	4.9678	
SVR	Training	0.4913	21.3171	3.4808	4.6170	Best Model
	Test	0.5384	20.6065	3.3194	4.5394	

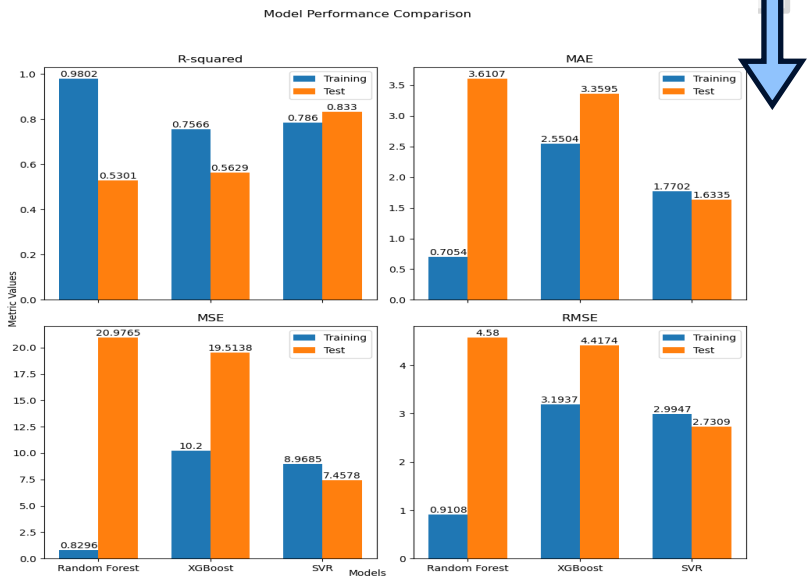
- The relatively stable performance on both the training set and the test set makes the SVR the preferable model in the initiative phase

Stage two: With ”Sates”

Model	Mean Squared Error	R-squared
Linear Regression	48359	-1082.19
Lasso Regression	45268.4	-1012.97
Ridge Regression	38449.6	-860.23
SVR	19.55	0.56
Random Forest	22.09	0.51
XGBoost	20.25	0.55

more precise hyperparameter tuning

The SVR model shows an R-squared of about 0.79 on the training set and about 0.83 on the test set. The results suggest that **SVR is the best** generalizing model among the three, with the highest and most stable R-squared and lowest RMSE, MSE, and MAE on average



Reflection and Policy Suggestion

Limitations:

- **Data Quantity**: Limited by the fixed number of U.S. counties, restricting sample size expansion and variability.
- **Data Granularity**: County-level data may mask critical nuances of individual obesity determinants.
- **Temporal Dynamics**: The cross-sectional study design may not fully capture the evolving nature of obesity trends.
- Model Interpretability: The SVR model's 'black box' nature complicates understanding the importance of features.
- Socioeconomic and Cultural Factors: Potential underrepresentation of cultural attitudes and localized determinants affecting obesity.

Future Research:

- **Dynamic Modeling with Time-Series Data**: To observe how policy and socioeconomic changes affect obesity trends.
- **Technological Integration**: Use of mobile health data and AI to enhance data analysis and feature integration.
- **Simulation and Effectiveness Studies**: To forecast and evaluate the impact of public health interventions and policies.

Several policy suggestions can be made:

- **Health Education and Promotion**: Develop community-specific health education programs that address local social determinants identified as risk factors for obesity.
- **Urban Planning**: Advocate for policies that encourage the development of walkable neighborhoods, and access to parks, and recreational facilities.
- **Food Accessibility**: Implement policies that increase access to affordable, healthy food options, especially in areas identified as 'food deserts'.
- **Socioeconomic Support Programs**: Financial stability can lead to better health outcomes.
- **Healthcare Access**: Work towards expanding access to healthcare services, including preventive care that can address obesity before it leads to more serious health issues.
- **Intersectoral Collaboration**: Encourage collaboration between different sectors such as health, education, and urban planning to address obesity more holistically.


Thank You !