If Obesity Correlates With Community Characteristics

A case study in Los Angeles, U.S.

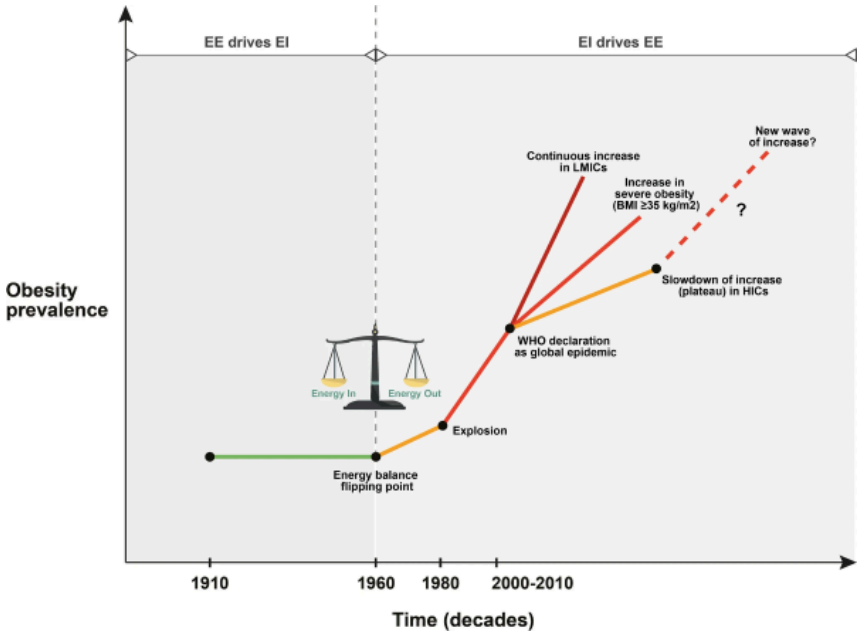
Daoyang Li & Yishan Wang
April 25, 2024

Abstract

This project aims to explore whether there is a significant correlation between obesity rates and community characteristics, including residents' income, lifestyle habits (such as smoking, drinking, and sleep duration), transportation availability, and accessibility to food sources. After an in-depth analysis of the data set using advanced statistical methods such as correlation heatmaps, random forests, XGBoost, and Generalized Linear Regression, the results indicate a significant association between obesity rates and various community characteristics, providing an essential basis for formulating effective public health policies and obesity prevention measures.

Introduction

Obesity is defined as a body condition where excessive fat accumulation has adverse health effects, with a Body Mass Index (BMI) over 30 considered obese. According to the World Health Organization, in 2019, around 5 million deaths from non-communicable diseases were due to obesity (WHO, 2024).



Obesity Prevalence (Koliaki, C., 2023)

Hence, obesity has become a serious global public health issue. However, obesity is largely preventable. It is crucial to examine residents' income, lifestyle habits, transportation availability, and food source accessibility in different communities to identify key community characteristics contributing to high obesity rates and propose targeted public health interventions. These measures can help reduce the health burden of obesity and potentially improve the overall well-being of community residents, providing data support and scientific evidence for formulating more effective health promotion policies.

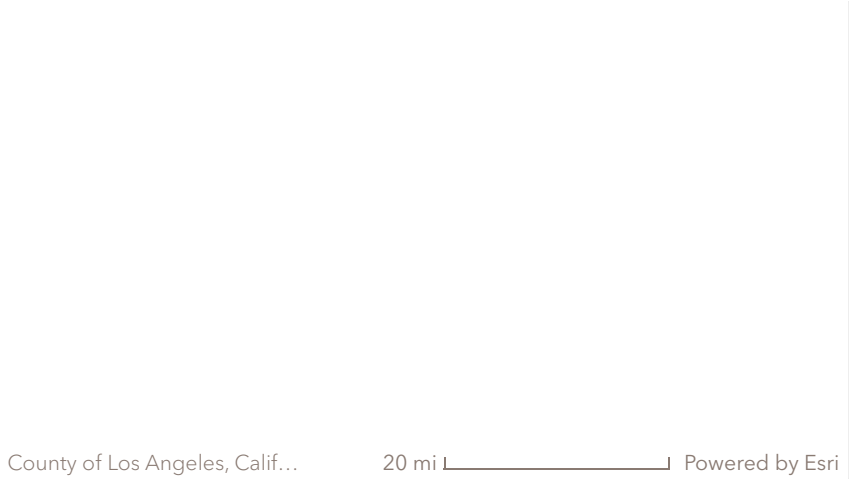
Previous Work

Kim et al. (2017) discovered that urban environments significantly influence obesity trends, Day et al. (2013) examined the impact of urban design on obesity and physical inactivity. Additionally, Plantinga and Bernell (2007) analyzed how residential location choices potentially affect health and weight. These studies collectively emphasize the potential of community characteristics in preventing obesity.

Study Area

The research area focuses on Los Angeles County, which is segmented into census tracts covering a diverse geographic span from urban to rural areas. The data set chosen reflects the most

recent obesity conditions and community characteristics collected over the past five years.



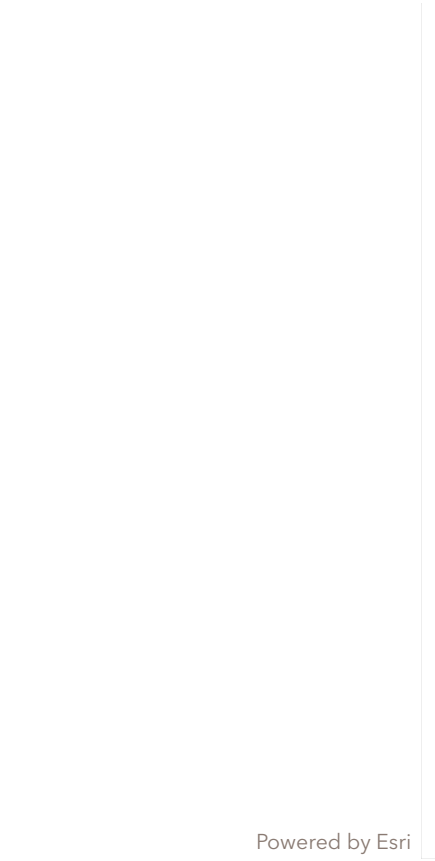
Los Angeles County

Data

Obesity Rate Data Set

This data set (Centers for Disease Control and Prevention, 2023). records the crude prevalence of obesity among residents of Los Angeles County (OBESITY_CrudePrev), determined by the

percentage of adults with a BMI over 30. Obesity rate data is a critical indicator in public health, directly linked to individual and population health risks. Additionally, the data set includes other behavioral risk factors related to obesity, such as binge drinking (BINGE_CrudePrev), current smoking status (CSMOKING_CrudePrev), sleep duration of less than 7 hours per night (SLEEP_CrudePrev), and crude prevalence of depression (DEPRESSION_CrudePrev). These variables are essential for assessing and understanding obesity trends and their influencing factors, and they are vital for public health interventions and policy-making. The left map shows the obesity rate and the right map show the current smoking rate.



Community Characteristic Data Set

The data set (Los Angeles County, 2024) from Los Angeles County encompasses community characteristics, including economic conditions, infrastructure availability, and demographic information. Key indicators such as median family income (MedianFamilyIncome), distribution of low-income families (LowIncomeTracts), and vehicle accessibility (LILATracts_Vehicle) provide insights into the economic environment of a community and the residents' ability to access health resources. Age and race data included help analyze how communities impact residents' health habits. These data are critical for revealing community factors related to obesity and guiding the formulation of health

policies. The left map shows the number of kids and the right map shows the median family income.

Data Preprocessing

We performed data preprocessing on two original datasets, including data cleaning, filtering, and merging. These datasets covered obesity rate data and community characteristics data for Los Angeles County, such as income levels, low income, and transportation accessibility.

Parameter Name	Description
CT10	Census tract
LILATracts_1And10	Low income and low access tract measured at 1 mile for urban areas and 10 miles for rural areas
LILATracts_halfAnd10	Low income and low access tract measured at 1/2 mile for urban areas and 10 miles for rural areas
LILATracts_Vehicle	Low income and low access tract using vehicle access or low income and low access tract measured at 20 miles
HUNVFlag	Vehicle access, tract with low vehicle access
LowIncomeTracts	Low-income tract
MedianFamilyIncome	Tract median family income
LA1and10	Low access tract at 1 mile for urban areas and 10 miles for rural areas
LAhalfand10	Low access tract at 1/2 mile for urban areas and 10 miles for rural areas
LATracts_half	Low access tract at 1/2 mile
TLATracts1	Low access tract at 1 mile
TractLOWI	Tract low-income population, number
TractKids	Tract children aged 0-17, number
TractSeniors	Tract seniors aged 65+, number
TractWhite	Tract White population, number
TractBlack	Tract Black or African American population, number
TractAsian	Tract Asian population, number
TractNHOPI	Tract Native Hawaiian and Other Pacific Islander population, number
TractAIAN	Tract American Indian and Alaska Native population, number
TractOMultir	Tract Other/Multiple race population, number
TractHispanic	Tract Hispanic or Latino population, number
TotalPopulation	Total population estimates
ACCESS2_CrudePrev	Lack of health insurance crude prevalence
BINGE_CrudePrev	Binge drinking crude prevalence
CSMOKING_CrudePrev	Current smoking crude prevalence
OBESITY_CrudePrev	Obesity crude prevalence
SLEEP_CrudePrev	Sleep <7 hours crude prevalence
DEPRESSION_CrudePrev	Depression crude prevalence
Shape_Length	Shape_Length
Shape_Area	Shape_Area

Data Description

Methodology

We used preprocessed data to generate correlation heatmaps, which helped visualize the relationships between obesity rates and various community characteristics, providing an intuitive understanding for further analysis.

Create Heatmap

Here's a breakdown of each step and how it contributes to transforming raw data into heatmap:

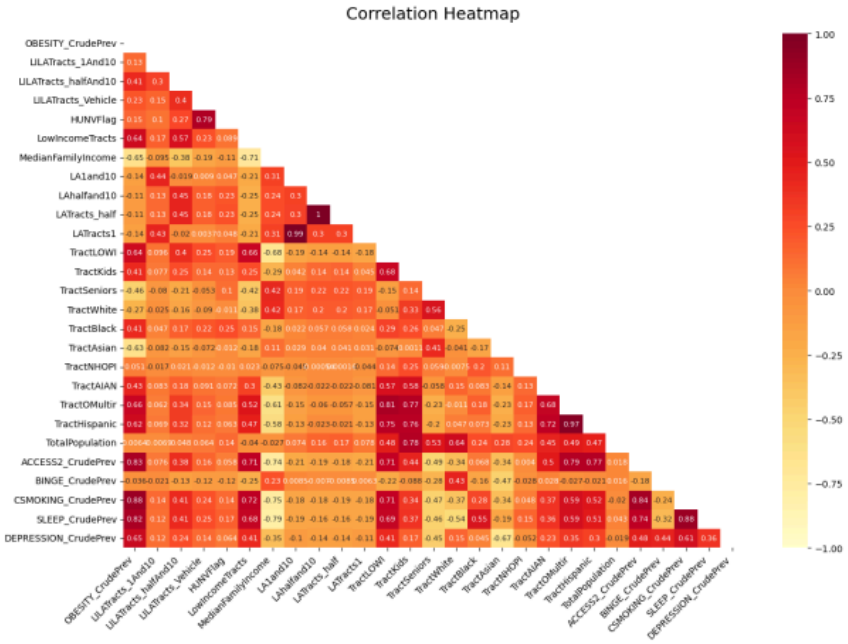
1. **Data Normalization with MinMaxScaler:** Numerical data in the DataFrame is scaled to a 0-1 range using MinMaxScaler to ensure uniformity and improve the effectiveness of statistical methods.

2. **Sorting Data:** The DataFrame is sorted in descending order by 'OBESITY_CrudePrev' to emphasize how other variables correlate with obesity prevalence.

3. **Correlation Matrix Computation:** A correlation matrix is calculated to explore the relationships between all pairs of numerical variables, highlighting potential influences on obesity.

4. **Heatmap Visualization:** A heatmap is generated to visually represent the correlation matrix, using color intensities to denote the strength of correlations and applying a mask to simplify the view.

In the heatmap, the intensity of the correlations was indicated by the depth of the colors in the heatmap, with deeper colors (closer to red) indicating stronger positive correlations and lighter colors (closer to yellow) indicating stronger negative correlations. From the heatmap, noticeable trends emerged, such as a strong positive correlation between obesity rates and low-income characteristics (e.g., LowIncomeTracts), suggesting that lower-income areas might have higher obesity rates. There was also a significant correlation between obesity rates and behavioral risk indicators like binge drinking (BINGE_CrudePrev), current smoking (CSMOKING_CrudePrev), insufficient sleep (SLEEP_CrudePrev), and depression (DEPRESSION_CrudePrev), implying potential links between these behaviors and obesity. This gave us a perspective on which community characteristics might significantly impact obesity, potentially guiding further causal research and the development of public health intervention strategies.



Correlation Heatmap of All Parameters

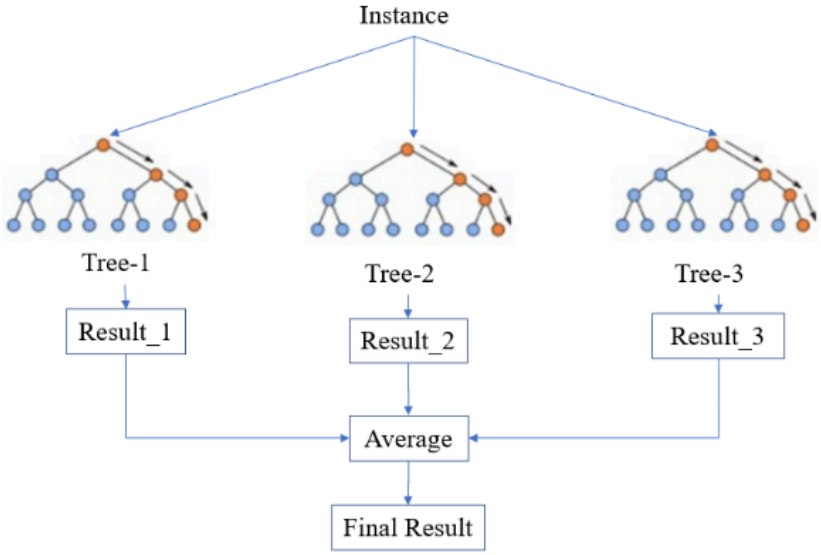
Subsequently, we constructed predictive models using three advanced statistical methods: ***Random Forest Regression***, ***XGBoost***, and ***Generalized Linear Regression*** (GLR), assessing the models' fit using their R-squared values. High R-squared values indicated that the models could explain the relationship between obesity rates and community characteristics well. This series of methods was aimed at ensuring the accuracy and reliability of our findings, providing profound insights into the relationship between obesity rates and community characteristics, thereby supporting robust data for the prevention of obesity in Los Angeles County.

Random Forest Regression

Random Forest Regression is a versatile machine learning technique that uses an ensemble of decision trees to make predictions about continuous target variables. It builds multiple decision trees during training and outputs the average of their predictions, providing a more accurate and stable result than a single decision tree could achieve. This approach effectively handles large datasets with numerous variables, manages overfitting by averaging multiple predictions, and is good at capturing complex nonlinear relationships in data.

Random Forest is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The image shows the process of a random forest algorithm where multiple decision trees are built

independently from random subsets of the data, and their individual predictions are then averaged to produce the final result.



Simplified Structure of Random Forest

We apply Random Forest Regression in a 6-step approach:

1. **Feature and Target Variable Selection:** A list of features relevant to the model is defined, and both features (X) and the target variable (y) are selected from the DataFrame.

2. **Data Splitting:** The dataset is divided into training and testing sets using the train_test_split method, with 20% of the data allocated to the test set to validate the model's performance.

3. **Model Initialization and Training:** A RandomForestRegressor is initialized and trained on the training data. This model is suitable for regression tasks and can handle complex interactions between features.

4. **Model Prediction and Evaluation:** The trained model makes predictions on the test set. Its performance is evaluated using the R^2 score, which measures the proportion of variance in the dependent variable that is predictable from the independent variables.

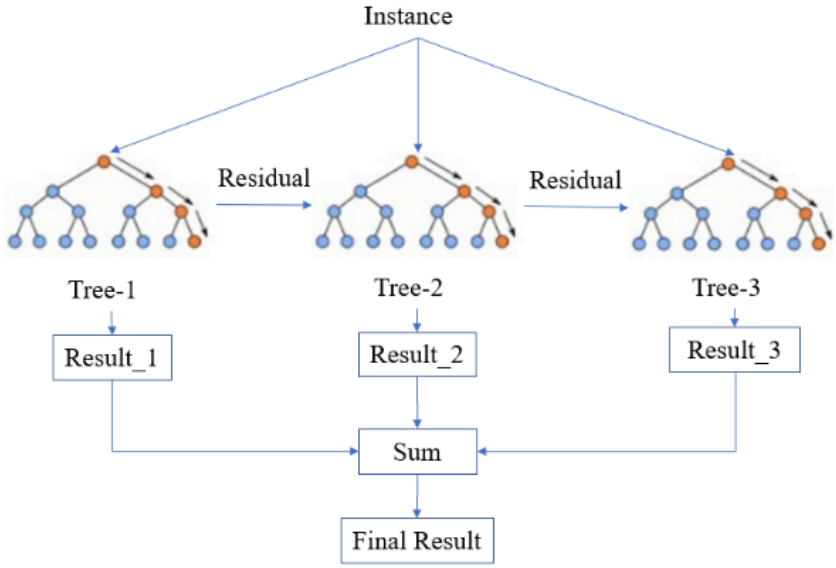
5. **Feature Importance Analysis:** Permutation importance is calculated to assess the impact of each feature on the model's predictive power. This method involves repeatedly shuffling individual features in the test set and observing the change in the model's accuracy.

6. **Visualization of Feature Importances:** A horizontal bar chart is plotted to visually represent the importance of each feature. This chart shows the mean decrease in accuracy (importance) along with its variability, providing insights into which features are most valuable for the model.

XGBoost Model

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms, designed to be highly efficient, flexible, and portable. It enhances the performance of gradient boosting framework through system optimization and algorithmic enhancements, specifically by utilizing a more regularized model formalization to control over-fitting, which gives it better performance.

The image illustrates the iterative process of XGBoost where each new tree is built to correct the residuals (errors) of the sum of the previous trees, with the final result being the cumulative prediction of all the trees.



Simplified Stucture of XGBoost

We apply XGBoost in a 6-step approach:

1. **Define Explanatory and Response Variables:** The factors list specifies the explanatory variables for the model, and X and y are defined to contain these variables and the target variable ('OBESITY_CrudePrev'), respectively.

2. **Data Splitting:** The data is divided into training and testing sets using train_test_split, assigning 20% of the data to the test set. This

separation helps in evaluating the model on unseen data to ensure its generalizability.

3. **Initialize and Train XGBoost Regressor:** An XGBoost regressor is initialized with specific hyperparameters such as learning rate, maximum tree depth, and regularization. It is then trained on the training data. XGBoost is known for its performance and speed in regression tasks.

4. **Prediction and Model Evaluation:** The model predicts outcomes for both the training and testing datasets. The R^2 score is calculated for these predictions to evaluate the model's performance. This metric indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

5. **Feature Importance Analysis:** It seems the code segment to calculate permutation importances (perm.importances_mean and perm.importances_std) is missing prior execution steps. Normally, you would calculate these using permutation_importance from sklearn.inspection, which assesses the importance by observing how random shuffling of each feature affects model performance.

6. **Visualization of Feature Importances:** The model predicts outcomes for both the training and testing datasets. The R^2 score is calculated for these predictions to evaluate the model's performance. This metric indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

Generalized Linear Regression

Generalized Linear Regression (GLR) in ArcGIS Pro is a powerful statistical tool used to model relationships between a dependent variable and one or more independent variables. When applied to studying obesity, GLR can help uncover patterns and predictors related to obesity rates across different geographical areas. By integrating various demographic, socioeconomic, and environmental factors, GLR allows us to assess the impact of these variables on obesity prevalence. This method supports different types of data distributions, making it versatile for addressing non-normal data patterns often seen in health-related datasets. Using ArcGIS Pro for GLR facilitates visualizing and interpreting the results geographically, which can be crucial for developing targeted public health interventions and policies aimed at combating obesity at local or regional levels.

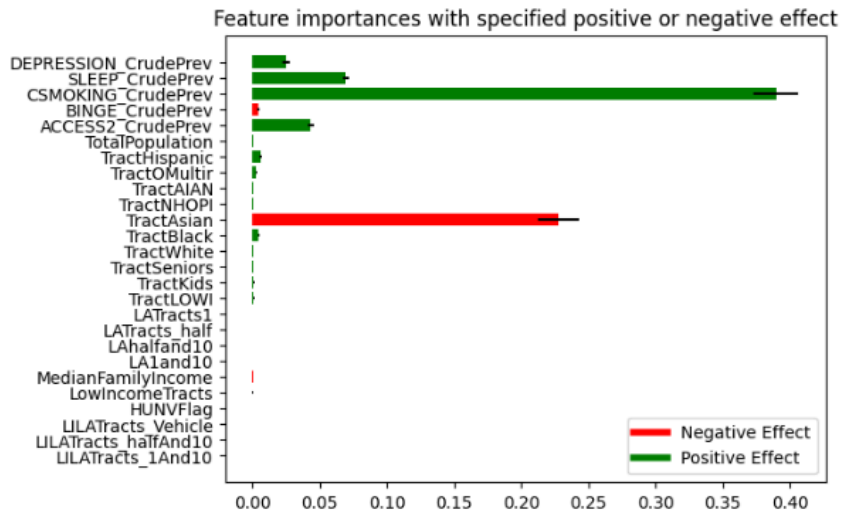
Results

Random Forest Regression

The random forest model yielded **R-squared values over 0.99 for the training set** and **about 0.98 for the test set**, indicating high accuracy and reliability in predicting obesity rates in Los Angeles County. Additionally, the permutation importance analysis identified lifestyle factors such as depression, insufficient sleep, current smoking, and binge drinking as critical predictors of obesity rates.

```
Train R2: 99.7075702993137
Test R2: 97.55049479482592
```

Random Forest Regression Model Performance



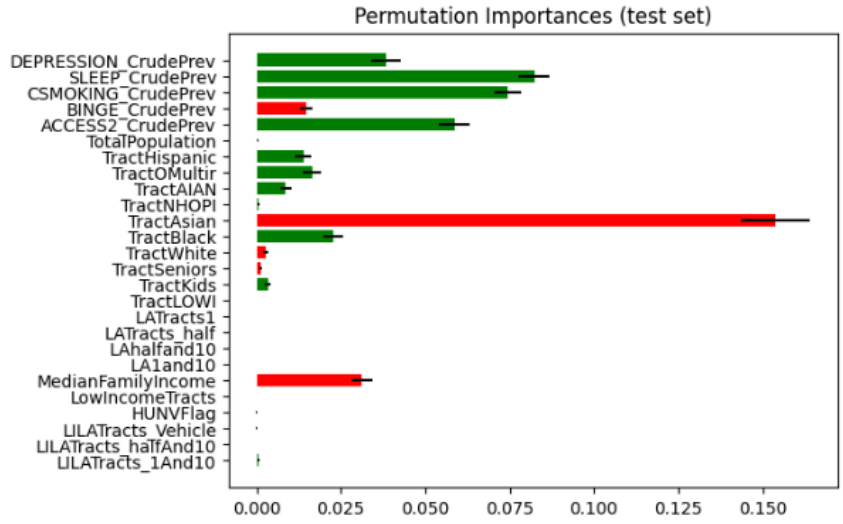
Random Forest Regression Feature Importances

XGBoost Model

The XGBoost model demonstrated a similar high accuracy level, delivering an **R-squared value of 0.92 for the training set and about 0.92 for the test set**. This result validates the XGBoost model's capacity to capture and interpret the complex relationship between obesity rates and community characteristics, especially when handling many nonlinear patterns.

```
Train R2: 92.01%
Test R2: 91.62%
```

XGBoost Model performance



XGBoost Model Feature Importances

Generalized Linear Regression

The Generalized Linear Regression (GLR) model produced an exceptionally high **R-squared value of 0.99**, indicating an excellent fit to the data. Moreover, diagnostic measures such as the Akaike Information Criterion (AIC) and various statistical tests (such as the joint F-statistic, joint Wald statistic, Koenker (BP) statistic, and Jarque-Bera statistic) all suggest a well-fitting model with no significant heteroscedasticity or non-normality.

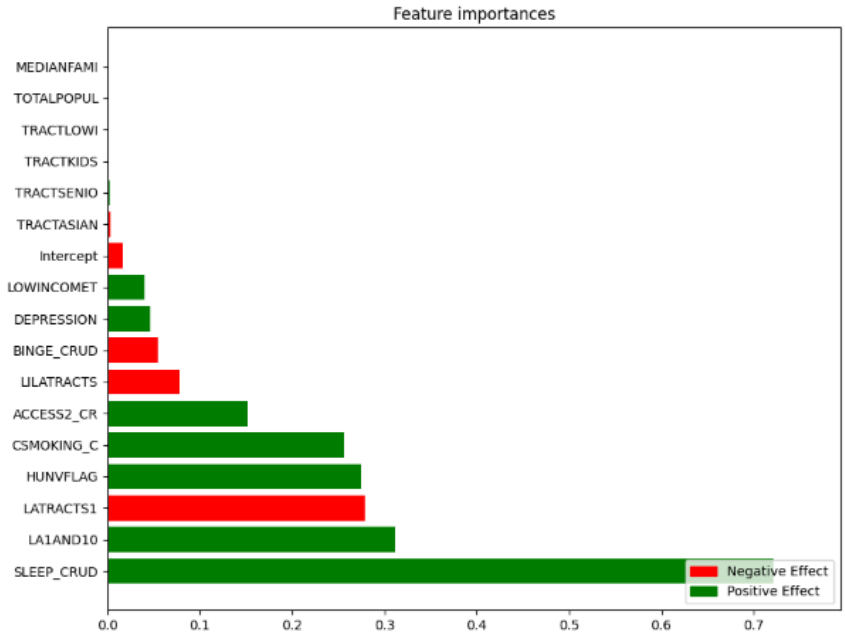
Variable	Coefficient[a]	StdError	t-Statistic	Probability[b]	Robust_SE	Robust_t	Robust_Pr[b]	VIF[c]
Intercept	0.010225	0.057092	0.42/988	0.897591	0.009569	4.0/1151	0.000075*	
LILATRACTS	-0.070256	0.156552	-0.500444	0.616818	0.169676	-0.461868	0.644798	1.568554
HUNVFLAG	0.274557	0.085166	3.223708	0.001207*	0.112047	2.450379	0.014334*	1.117576
LOWINCOMET	0.040192	0.070568	0.569560	0.569033	0.075816	0.530133	0.596076	3.768520
MEDIANFAMFU	0.000001	0.000001	0.926574	0.354221	0.000001	0.750096	0.455380	6.258578
LA1AND10	0.311699	0.643196	-0.484610	0.628008	1.037586	0.300488	0.763980	110.784892
LATRACTS1	-0.279046	0.642323	-0.434435	0.664027	1.035956	-0.269361	0.787680	109.571058
TRACTLOWI	-0.000272	0.000054	-5.030047	0.000001*	0.000084	-3.230436	0.001236*	10.254791
TRACTKIDS	0.001104	0.000113	9.782341	0.000000*	0.000241	4.604304	0.000006*	11.870472
TRACTSENIO	0.001735	0.000157	11.071268	0.000000*	0.000379	4.584017	0.000007*	6.327055
TRACTASIAN	-0.003548	0.000044	-80.272066	0.000000*	0.000096	-36.954863	0.000000*	2.734589
TOTALPOPUL	0.000068	0.000048	1.401302	0.161259	0.000120	0.565660	0.571692	20.722019
ACCESS2_CR	0.151288	0.005457	27.721702	0.000000*	0.007949	19.030485	0.000000*	9.123021
BINGE_CRUD	-0.054565	0.021695	-2.515372	0.011951*	0.054015	-1.010199	0.312484	83.419412
CSMOKING_C	0.250666	0.021604	11.600366	0.000000*	0.047639	5.387681	0.000000*	48.804923
SLEEP_CRUD	0.721787	0.009678	74.641021	0.000000*	0.017545	41.134487	0.000000*	50.607468
DEPRESSION	0.045980	0.022123	2.078397	0.037781*	0.058430	0.786928	0.431385	73.707427

GLR Model Performance

GLR Diagnostics

Input Features	new_obesity_food	Dependent Variable	OBESITY_CR
Number of Observations	3560	Akaike's Information Criterion (AICc)	8754.533583
Multiple R-Squared	0.947286	Adjusted R-Squared	0.945176
Joint F Statistic	12840.343733	Prob(>F), (16,3543) degrees of freedom	0.000000*
Joint Wald Statistic	3961155.555269	Prob(>chi-squared), (16) degrees of freedom	0.000000*
Koenker (BP) Statistic	566.515489	Prob(>chi-squared), (16) degrees of freedom	0.000000*
Jarque-Bera Statistic	11193.678598	Prob(>chi-squared), (2) degrees of freedom	0.000000*

GLR Model Diagnostics



GLR Feature Importances

Powered by Esri

Result of GLR

Discussion

In our experiment conducted within the relatively small confines of Los Angeles County, the Random Forest model notably outperformed XGBoost. This outcome is likely due to the characteristics of our dataset, which is noisy and includes outliers. Random Forest is inherently more robust against overfitting and

noise, thanks to its bagging technique that averages results across multiple decision trees to reduce variance and enhance generalization. In contrast, XGBoost, which uses a boosting technique, tends to overfit by excessively focusing on noisy data, potentially amplifying errors. Notably, in the feature importance analysis, Random Forest identified fewer key parameters as critical, suggesting a more focused model, whereas XGBoost highlighted a broader array of important features, indicating its sensitivity to a wider range of data nuances.

In our GLR model, we initially faced issues of data redundancy, which can lead to multicollinearity and affect the interpretability and stability of the model. To address this, we selectively pruned the list of predictors, retaining only those parameters that contributed unique and significant explanatory power to the model. This refinement process helped in enhancing the model's performance by reducing complexity and focusing on the most informative features. GLR differs from Random Forest and XGBoost in handling obesity data in LA because it assumes a linear relationship and is more affected by outliers. Random Forest and XGBoost can model complex interactions without needing explicit specification and are more robust against overfitting due to their ensemble nature. Consequently, they may yield different results and insights into feature importance and the overall relationship between variables. GLR may show a strong correlation between "low access" areas and obesity due to its sensitivity to linear relationships, which could reflect geographic patterns if such patterns exist. In contrast, Random Forest and XGBoost distribute effects across features, capturing complex interactions and potentially diluting the apparent impact of "low access" due to their non-linear modeling capabilities. The discrepancy suggests that GLR might be capturing a direct geographic correlation, while tree-based models consider a broader context of interacting factors.

Despite our models' predictive solid power and high statistical significance, there are still potential issues. For instance, high R-squared values indicate strong associations but do not establish causality. Future studies using causal inference methods, such as instrumental variables or randomized controlled trials, could help to identify and verify these relationships more accurately. Furthermore, considering the dataset's potential selection bias and measurement errors, improving data collection and processing methods could lead to more precise results.

To further deepen our understanding, the next step in research should explore individual-level data, considering more personal lifestyle habits, socio-economic status, and the complex interplay of community environments. In addition, studying the effects of community-level interventions, such as implementing diet and exercise plans to improve obesity rates, will be an essential direction to continue this research. Moreover, longitudinal studies of models tracking changes over time within the same community or individuals may reveal underlying dynamic relationships and trends. In summary, while our research has provided new insights into the relationship between community characteristics and obesity rates, much work must be done to uncover deeper mechanisms and develop effective intervention measures.

Conclusion

In this study conducted in Los Angeles County, we explored the correlation between obesity rates and community characteristics across various neighborhoods. Utilizing Random Forest, XGBoost, and Generalized Linear Regression (GLR) models, we derived compelling results that indicate significant positive correlations between obesity rates and lifestyle behavior factors such as depression, insufficient sleep, and current smoking. These findings affirm the substantial link between health behaviors and obesity, while also suggesting the influence of economic factors on this public health issue. Interestingly, our analysis revealed a negative trend between alcohol consumption and obesity rates. Additionally, areas with low access to food sources exhibited similar obesity rates to other areas, suggesting that food choice preferences remain consistent regardless of proximity to food outlets. Our dataset, comprising over 2,000 data points and 26 parameters, yielded exceptionally high R-squared values over 90%, indicating robust model fits across all three methodologies employed.

References

Kim, J., Shon, C., & Yi, S. (2017). The Relationship between Obesity and Urban Environment in Seoul. *International journal of environmental research and public health, 14*(8), 898. https://doi.org/10.3390/ijerph14080898

Day, K., Alfonzo, M., Chen, Y., Guo, Z., & Lee, K. K. (2013). Overweight, obesity, and inactivity and urban design in rapidly growing Chinese cities. *Health & place*, *21*, 29–38. https://doi.org/10.1016/j.healthplace.2012.12.009

Plantinga, A. J., & Stephanie Bernell. (2007). Can Urban Planning Reduce Obesity? The Role of Self-Selection in Explaining the Link between Weight and Urban Sprawl. *Review of Agricultural Economics*, *29*(3), 557–563. http://www.jstor.org/stable/4624865

World Health Organization. (n.d.). (2024). Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Centers for Disease Control and Prevention. (2023). Adult Obesity Prevalence Maps. https://www.cdc.gov/obesity/data/prevalence-maps.html

Los Angeles County (n.d.). (2024). Food Deserts. https://data.lacounty.gov/datasets/lacounty::food-deserts/about

Koliaki, C., Dalamaga, M. & Liatis, S. Update on the Obesity Epidemic: After the Sudden Rise, Is the Upward Trajectory Beginning to Flatten?. *Curr Obes Rep* **12**, 514–527 (2023). https://doi.org/10.1007/s13679-023-00527-y

Spatial Science Institute
University of Southern California

Daoyang Li & Yishan Wang

SSCI 575 Spatial Data Science