CITADEL

CITADEL INVITATIONAL DATATHON

SUMMER 2024
TEAM NINE

Fast Food, Socio-Economics, and Obesity: A Predictive Analysis

Bisheshank C Aryal | Heon Lee
Ziao (Ollie) Zhang | Zhengyu Zou

ABSTRACT

This study investigates the relationship between the accessibility of fast food, socioeconomic factors, and obesity rates across state and county levels. Our study seeks to answer three core questions: (1) How do the prevalence and concentration of fast food outlets, in conjunction with socioeconomic and demographic variables, correlate with obesity rates? (2) To what extent does fast food consumption drive obesity, and are there other, more influential factors at play? (3) Finally, can predictive modeling, informed by these variables, forecast future obesity trends? By constructing a time-series model that incorporates these elements, we aim to highlight the impact of the food environment and socioeconomic conditions on obesity trajectories. This predictive framework will equip policymakers with insights to devise targeted public health interventions. Ultimately, we aspire to determine whether regulating fast food density, enhancing access to healthier alternatives, or addressing socioeconomic disparities is the most effective strategy for mitigating obesity and fostering healthier communities.

Keywords: Obesity, Fast Food, Socioeconomic Factors, Public Health, Time-Series Analysis, Predictive Modeling
Github: JackeyLove36/Citadel-Datathon-Summer-2024

Contents

1 NON-TECHNICAL EXECUTIVE SUMMARY

How does the interplay between the food environment, specifically the availability of fast food restaurants and accessibility to grocery stores, and socioeconomic factors influence obesity rates in marginalized communities across the United States over time?

Obesity has emerged as a critical public health concern, linked to a myriad of chronic diseases. This study examines the relationship between the food environment, socioeconomic factors, and obesity prevalence. By analyzing time-series data on fast food restaurant density, grocery store accessibility, and sociodemographic characteristics, we aim to explicate the factors driving obesity rates, particularly in marginalized communities.

1.1 Background

Obesity has reached epidemic proportions in the United States, with far-reaching consequences for public health [Far23]. The prevalence of obesity has surged, contributing to a significant burden of chronic diseases such as heart disease, diabetes, and certain cancers [Pat+23]. This study delves into the complex factors driving obesity rates, with a particular focus on the interplay between the food environment and socioeconomic conditions.

1.1.1 The Food Environment and Obesity

The availability and accessibility of food play a critical role in shaping dietary habits and overall health. Research consistently demonstrates a link between increased access to fast food and higher obesity rates. Fast food, characterized by its reliance on highly processed, calorie-dense foods with limited nutritional value, has contributed significantly to the obesity epidemic. The convenience, affordability, and aggressive marketing strategies employed by fast food chains further exacerbate the problem. A study [Ric+11] found that children living near schools with a higher concentration of fast food outlets were more likely to consume unhealthy meals. A one standard deviation increase in fast food restaurants within 1600 m of individual residences (or 5.2 restaurants) increases BMI by 1.0% with respect to the sample mean, and by 0.5% within the 1600 m school-buffer. This highlights the impact of proximity and density on eating habits. Conversely, communities with abundant grocery stores offering fresh produce, whole grains, and lean proteins tend to have lower obesity prevalence. This access to healthy food options is crucial for promoting healthier dietary patterns.

The uneven distribution of food outlets further complicates the issue. Food swamps are areas characterized by an abundance of fast food restaurants and convenience stores. Studies [Hag+16] demonstrate that residing in a food swamp is associated with higher consumption of unhealthy snacks and desserts compared

to fruits and vegetables. This lack of access to healthy options creates a challenging environment for maintaining a healthy diet.

Figure 1: Childhood Obesity Linked to Proximity of Fast-Food Restaurants. Source: [Pos19]

1.1.2 Socioeconomic Disparity and Obesity

Socioeconomic disparities exacerbate the impact of the food environment on obesity. Disadvantaged communities often face a double burden: limited access to healthy food options and a higher concentration of fast food outlets. This phenomenon, often referred to as a "food desert," contributes to higher obesity rates among vulnerable populations.

Research consistently highlights the correlation between low-income neighborhoods, particularly those with a predominantly Black population, and a higher density of fast food restaurants [Jam+14]. Moreover, the detrimental effects of fast food consumption on Body Mass Index (BMI) are amplified among individuals from households with lower maternal education levels [Rei+14]. This finding aligns with broader research indicating that educational disparities within families magnify health inequalities from early childhood (Deaton, 2003; Marmot, 2010) [Lib+23].

Compounding these challenges, aggressive marketing tactics employed by fast food chains disproportionately target low-income and minority communities, promoting unhealthy, calorie-dense options [Gri+07] [GK08]. These factors collec-

tively underscore the critical role of income, education, and employment levels in determining access to nutritious food and subsequent health outcomes.

1.2 Model Building

By analyzing time-series data on fast food restaurants, grocery stores, obesity rates, income, education, and other socioeconomic indicators, we aim to identify key patterns and relationships. By analyzing the feature importance, we can better understand the complex factors driving obesity and inform strategies to address this public health challenge.

The following are the input features for our models:

- **Year**: To capture changes over time
- **Number of fast food restaurants adjusted to population size:** To measure exposure to unhealthy food options
- **Number of grocery stores adjusted to population size:** To assess access to healthy food
- **Obesity rate:** The primary outcome of interest
- **Median per capita income:** To reflect economic conditions
- **Percentage of population with health insurance:** To assess access to healthcare
- **Education level:** To measure educational attainment
- **Percentage of population below poverty line:** To indicate economic hardship
- **Unemployment rate:** To indicate economic hardship

By analyzing these variables, we aim to develop a predictive model that can help identify communities at greatest risk for obesity and inform targeted interventions to improve public health.

1.3 Key Findings

1.3.1 Impact of Fast Food Density on Obesity

Our analysis reveals an association between the increase in the number of fast food restaurants per capita and higher obesity rates. This correlation suggests that communities with a higher density of fast food outlets are more likely to experience rising obesity rates, regardless of other socio-economic factors. The availability and convenience of fast food, which is often calorie-dense and nutrient-poor, contribute to unhealthy dietary patterns that lead to obesity. This finding underscores the critical need to address the proliferation of fast food outlets as part of broader public health strategies to combat obesity.

1.3.2 Role of Grocery Stores

The study indicates that a greater availability of grocery stores per capita is correlated with lower obesity rates. Access to grocery stores, which typically offer a wider variety of healthier food options compared to fast food restaurants, appears to be a protective factor against obesity. This relationship highlights the importance of ensuring that communities have diverse food environments where residents can access nutritious foods. Policies aimed at increasing the number and accessibility of grocery stores in underserved areas could play a pivotal role in promoting healthier dietary habits and reducing obesity rates.

1.3.3 Marginalized Communities and Obesity

Our findings show that marginalized communities, characterized by lower median income, higher unemployment and underemployment rates, and lower educational attainment, are particularly vulnerable to increases in fast food density and decreases in grocery store availability. We found that educaiton levels were especially important. These communities exhibit higher obesity rates when the food environment is dominated by fast food outlets and lacks sufficient grocery stores. This vulnerability highlights the compounded impact of socio-economic disadvantages and limited access to healthy food options. Targeted interventions that improve food environments and address socio-economic disparities are crucial for mitigating obesity in these communities.

1.3.4 Feature Importance in Community Healthiness

The analysis identifies several socio-economic indicators, such as median income, education level, and insurance coverage, as significantly correlated with obesity rates. However, it is essential to note that these correlations do not imply causation. For instance, simply increasing income levels may not directly reduce obesity rates without concurrent improvements in other factors like education and healthcare access. This nuanced understanding is vital for developing effective public health policies. It suggests that comprehensive approaches that simultaneously address multiple determinants of health are more likely to succeed in reducing obesity rates.

1.3.5 Recommendations for Community Resilience

Based on our findings, effective strategies to combat obesity should include efforts to reduce the density of fast food restaurants and increase the number and accessibility of grocery stores. Additionally, enhancing community education seems to be an important aspect. Interventions should be multifaceted and tailored to address the specific needs and vulnerabilities of each community, which requires more data collection and validation. Policymakers must be cautious in assuming that addressing a single factor, such as income, will be sufficient to improve health outcomes without broader systemic changes. By adopting a holistic

approach that considers the interplay of various socio-economic and environmental factors, communities can become more resilient to obesity and related health issues. This report aims to provide a data-driven foundation for such comprehensive strategies.

2 EXPLORATORY DATA ANALYSIS (EDA)

2.1 Overview of EDA

Our team's long-term goal is to implement **High Performance computational tools** or **Deep Learning Theory models** to solve this **Regression Problem** based on **Supervised Machine Learning**. However, complexity of the datasets and limitations in scope make the predictions difficult. Our Team, therefore, investigates the obesity rates with consideration of many factors beyond the provided datasets and explores the future prediction in both **State Level** and **County Level**. However, raw datasets need to be processed through various techniques to ensure they are suitable for our models. This involves data cleaning, feature engineering, normalization, and possibly augmentation to enhance the quality and relevance of the data.

Currently, we have successfully parsed the state-level data, but county-level data processing is still an ongoing process. We categorized the data into three main types: geospatial, health, and socio-economic data.

We have datasets at both the state and county levels. The data cleaning process for these datasets will be discussed in the next section. For now, we have chosen to focus on the state-level data due to its immediate availability and ease of management. The analysis of county-level data is still in progress. Below is a table showing all the features we have collected:

Features	Feature Type	Data Type	Description
Year	category	int	year when the data is collected
State	category	str	state from which the data is collected
#_fast_food	continuous	int	number of opening fast food restaurants
#_groceries	continuous	int	number of opening grocery stores
PCT_UNE	continuous	float	percentage of individuals unemployed
MEDIAN_HOUSE_INC	continuous	int	median household income
MEAN_HOUSE_INC	continuous	int	mean household income
PER_CAPITA_INC	continuous	int	per capita income
PCT_NO_INSUR	continuous	float	percentage of people without health insurance coverage
PCT_FAM_POVERTY	continuous	float	percentage of families below poverty line
PCT_PPL_POVERTY	continuous	float	percentage of individuals below poverty line
PCT_OBESITY	continuous	float	percentage of individuals that are obese
Highschool	continuous	float	percentage of adults who received a high school education
Under_Highschool	continuous	float	percentage of adults who received an education below high school
Higher_Education	continuous	float	percentage of adults who received higher education

Table 1: Data Table of Pre-processed State Level Dataset

2.2 Data Collection

2.2.1 Geospatial Data

Features covered: #_fast_food, #_groceries

We used geospatial data to analyze the number of fast-food restaurants and grocery stores on a per-county basis. Specifically, we used the County Business Patterns (CBP) data with NAICS codes for limited-service restaurants and supermarkets/groceries [Bura]. The following figures illustrate the differences in the number of fast-food restaurants and grocery stores in 2021 versus 2003:

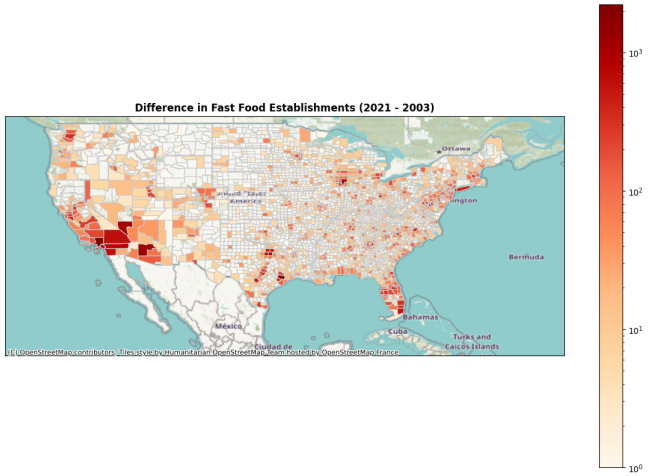


Figure 2: Number of Fast-Food Restaurants (2021 vs. 2003)

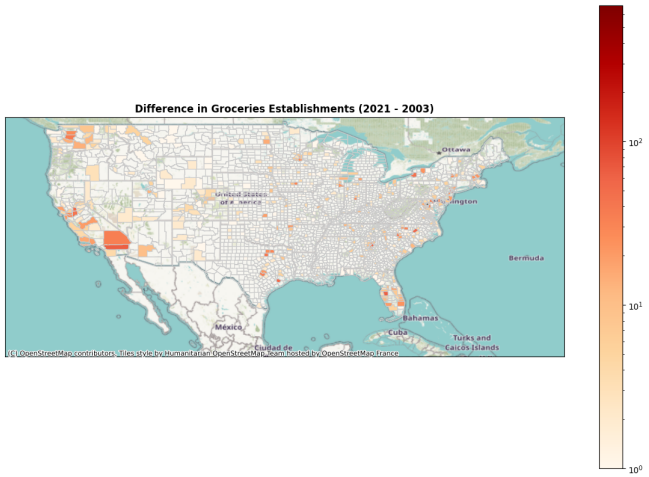


Figure 3: Number of Grocery Stores (2021 vs. 2003)

The trend of new fast food locations clearly higher than the grocery stores. Since we also needed state level data, we aggregated the number from each county for the state.

2.2.2 Health Data

Features covered: PCT_OBESITY

We obtained both state-wise and county-wise obesity rate. For the state level obesity rate, we used the provided nutrition dataset [Dis]. To explore more through the State Level, our team makes plots to figure out the potential factors for the obesity rate. Firstly, Figure 4 shows a histogram of **Target Variable** in the raw dataset:

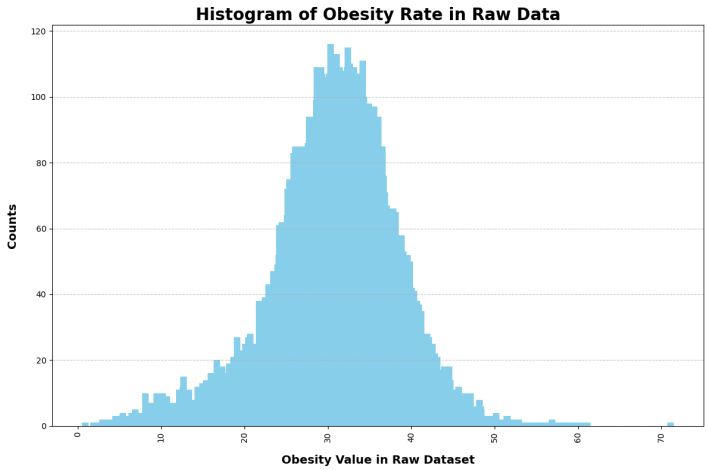
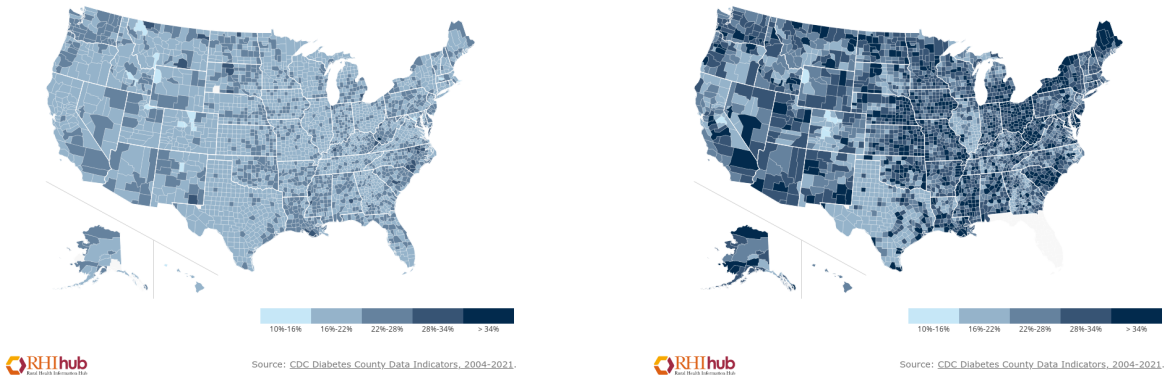


Figure 4: Histogram Distribution in Obesity Rate in Raw Dataset

From Figure 4, we can see that the number of obesity value reflects **symmetrical and is approximately distributed by a normal distribution.**

The health data was sourced from the Rural Health Information Hub, supported by the Health Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services (HHS) under Grant Number U56RH05539 [Burd]. We also used data from the CDC Diabetes County Data Indicators, 2004-2021, which provided the obesity prevalence score for each county:



(a) in 2004 (b) in 2021

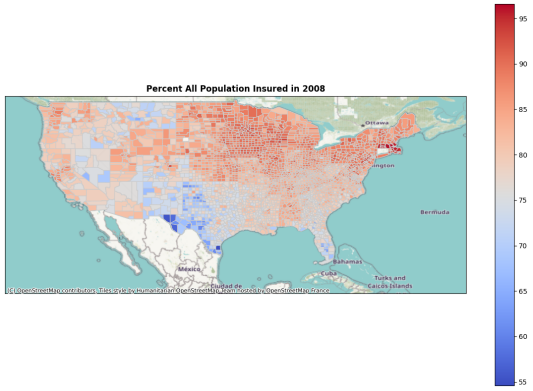
Figure 5: Obesity Prevalance Score over the Years

2.2.3 Socio-Economic Data

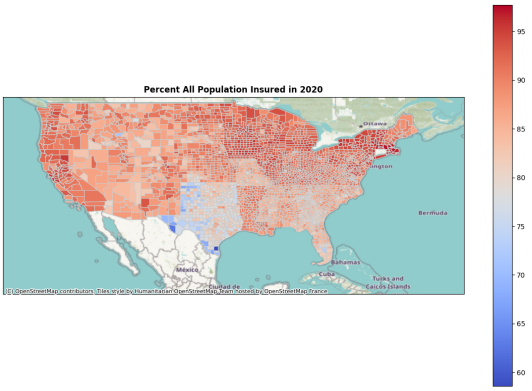
Features covered: PCT_UNE, MEDIAN_HOUSE_INC, PER_CAPITA_INC, PCT_NO_INSUR, PCT_FAM_POVERTY, PCT_PPL_POVERTY

We extracted state-wise unemployment, income, poverty, and health insurance coverage data from the American Community Survey - Selected Economic Characteristics [Burc] dataset. The data includes information for the entire population and smaller demographic groups. For convenience, we used data from the entire population of each state, converting percentage data into floats between 0 and 1, and converting income data into integers by removing commas for cleaner analysis.

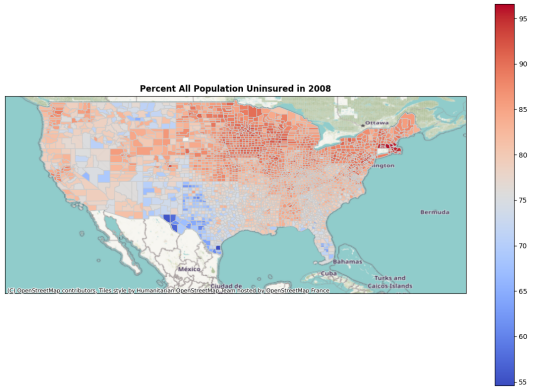
For education demographic data, we used the American Community Survey 1-Year estimates [Burb]. The dataset is divided into two age groups: 18-24 years and 25+ years. We calculated the weighted proportion of these categories to determine the overall education level for adults in a given state or county.



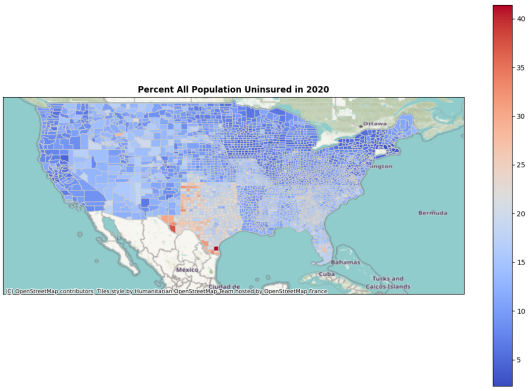
(a) % of all population insured in 2008



(b) % of all population insured in 2020



(c) % of all population uninsured in 2008



(d) % of all population uninsured in 2020

Figure 6: Insurance Coverage by Population Percentage

The insurance coverage data was quite inconclusive for our study as they were usually on the correct trends across the years. We then create plots of various features against the obesity rate in the raw dataset to gain a general understanding of the potential factors that correlate with the obesity rate as shown in Figures 7.

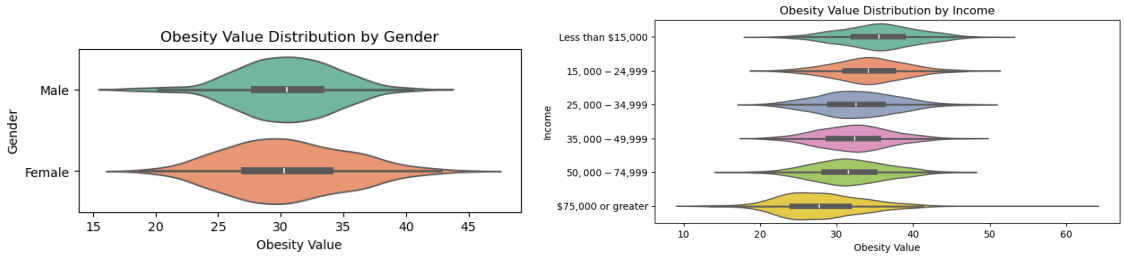


Figure 7: Visualization of Gendar and Income respected to Obesity Value

The income vs obesity shows a clear positive correlation in the mean with higher income being associated to a lower obesity level. We do note that the obesity for each income bracket is less than the lower income bracket in both the maximum

and median except for the highest bracket of $\$75,000$ or greater. This could potentially be explained by the fact that high income individuals have greater freedom to choose to indulge in food. However, the median and the quartiles for the highest income bracket are clearly lower than the other brackets.

2.2.4 Education Data

Features covered: Under_Highschool, Highschool, Higher_Education

For the education demographic data of state and county population, we used the *American Community Survey 1-Year estimates* [Burb]. The dataset is divided into two age groups: 18-24 years and 25+ years. We calculated the weighted proportion of these categories to determine the overall education level for adults in a given state or county. The data is presented as a percentage of adults with less than a high school diploma, a high school diploma, and a higher education degree.

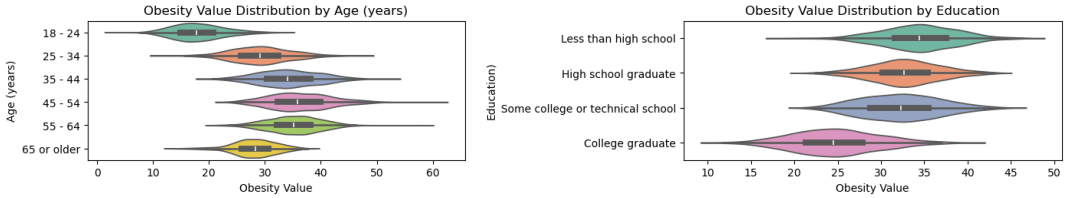


Figure 8: Visualization of Age and Education respected to Obesity Value

We observe that in each plot, there is a clear correlation between the potential factor with the obesity rate. The age vs obesity rate approximately looks quadratic with respect to the age brackets. The education vs obesity rate looks approximately logistical, with college graduates having a mean significantly below the other education levels and those without a high school degree having the highest obesity level.

2.3 Data Cleaning

Several considerations were addressed during data cleaning:

- **State vs. County-Level Data:** The socio-economic dataset and the nutrition dataset are measured at the state level, while our analysis requires county-level data. To address this, we decided to perform our analysis at two levels: state and county. We started with state-level data, which is more readily available and easier to manage, and we used this to establish our initial models and analyses. For county-level analysis, which is still a work in progress, we are applying a data disaggregation technique to estimate county-level values from state-level data. This involves using population-weighted averages, where county-level population data is used to proportionally allocate state-level metrics to each county.

- **Label Column Formatting:** The entries in the label column of the economic characteristics dataset contained extra spaces before and after the actual label string. To efficiently access and process these labels, we used string manipulation techniques to strip leading and trailing spaces from each entry.
- **Data Grouping:** Many economic characteristics are measured by grouping people into subcategories such as labor force, gender, and age. To simplify our analysis and focus on the broader trends, we used aggregate estimates derived from the entire population of a state instead of data on specific subgroups. This involved calculating weighted averages where necessary, and in cases where only subgroup data was available, we summed or averaged these groups to obtain a state-wide figure.
- **Inconsistent County Naming:** The datasets used different formats for county names, with some using FIPS codes and others using actual names. To standardize these and ensure consistency across all datasets, we created a mapping table that linked FIPS codes to their corresponding county names. This allowed us to convert all county identifiers to a common format.
- **Missing Data:** Dealing with missing data is crucial for maintaining the integrity of our analysis. We first assessed the extent and patterns of missing data. For features with less than 10% missing values, we opted to drop these entries to avoid introducing bias. For features with more significant missing data, we used imputation techniques. For numerical data, we applied mean or median imputation depending on the distribution of the data. For categorical data, we used the mode of the feature to fill in missing values.

2.4 Data Preprocessing

After the data cleaning process, we have the following stata:

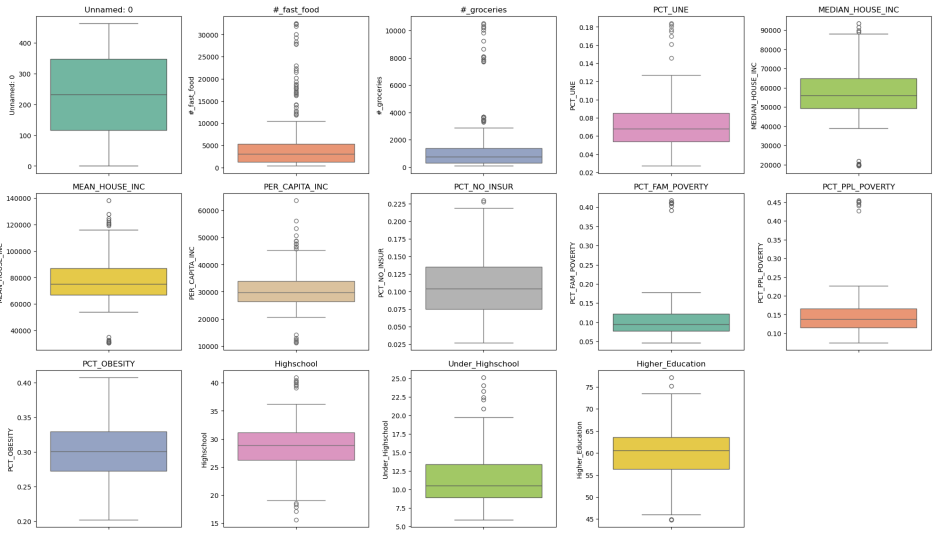


Figure 9: Boxplot of all Features

We then consider the data preprocessing: using OneHotEncoder( ) to all the categorical variables: State because this feature is unranked or unordered. And using StandardScaler( ) to all the continuous variables: since these features have a tailed distribution. After preprocessing, we used the GridSearchCV( ) to complete the cross validation.

The following diagram explains the process of cross-validation with 20%:

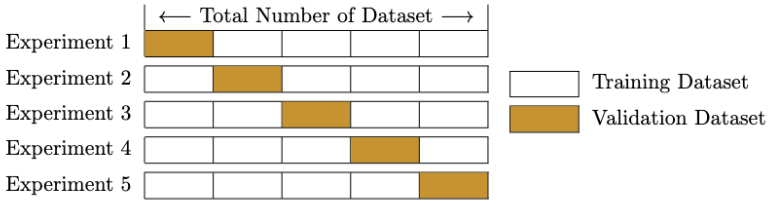


Figure 10: Flowchart in Cross Validation

3 METHODS

3.1 Problem Overview

Our group aims to explore this **Regression Problem** through multiple **supervised ML** techniques. The following explains how the predication of obesity rate is explained through mathematical concepts.

3.2 Algorithms

In this subsection, we let D denote our dataset where

$$D = \{(x_i, y_i) : 1 \leq i \leq n, x_i \in \mathbb{R}^k, y_i \in \mathbb{R}\}.$$

Here, x_i is the input features and y_i is the label corresponding to the obesity rate. Moreover,

$$\hat{y} = \mathcal{M}(x)$$

denotes the prediction by the model $\mathcal{M}$ given the input x.

3.2.1 Linear Regression

Linear regression models the relationship between a dependent variable and one or more independent variables as a linear function. The objective is to find the best-fitting hyperplane that minimizes the mean square error of the predicted and labeled data points. Given our dataset D and data points $(x_i, y_i) \in D$, we treat x_i as the independent variable and y_i as the dependent variable. We then aim to find a weight vector $w \in \mathbb{R}^k$ and a bias constant $\beta_0 \in \mathbb{R}$ that minimizes the mean square error

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where

$$\hat{y}_i = w^T x_i + \beta_0$$

is the prediction.

Linear regression is a suitable choice for our implementation because it serves as a strong baseline model for comparison with other methods and offers inherent interpretability, providing clear insights into the impact of each feature on the outcome.

3.2.2 Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees and combines their predictions. By training each tree on a random subset of data and features, Random Forest reduces overfitting compared to a single decision tree and improves model performance. This method can handle

Algorithm 1: LinearRegression

Input: data pairs $D = \{(x_i, y_i)\}_{i=1}^n$, learning rate η, loss threshold ε
Output: $w \in \mathbb{R}^k$, $\beta_0 \in \mathbb{R}$
Initialize w and β_0 ;
$L(w, \beta_0) \leftarrow \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - y_i)^2$; `means square loss`
while $L(w, \beta_0) > \varepsilon$ **do**
  $w \leftarrow w - \eta \cdot \frac{\partial L(w,\beta_0)}{\partial w}$; `gradient descent`
  $\beta_0 \leftarrow \beta_0 - \eta \cdot \frac{\partial L(w,\beta_0)}{\partial \beta_0}$;
  $L(w, \beta_0) \leftarrow \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - y_i)^2$;
end
return w, β_0

large datasets with higher dimensionality, which makes it suitable for predicting obesity rates.

Random Forest works by taking in the dataset D and creates $B \in \mathbb{N}$ bootstrap samples $D_1, D_2, \ldots, D_B$. By splitting the dataset and ensuring that each tree is trained on a different subset of the data, Random Forest can provide diversity among the trees and reduce overfitting. Each bootstrap sample D_b grows an unpruned decision tree T_b as shown in Figure 11. At each node, a random subset of m features from the total p features is selected, denoted as $M_b \subset \{1, 2, \ldots p\}$ with $|M_b| = m$. The node is then split using the best feature from M_b according to variance reduction for regression. This random feature selection at each split ensures that the trees are correlated. Once we train and get all B decision trees, the final prediction $\hat{y}$ for an input x is the average of the predictions from the trees:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

This is shown in the last layer of Figure 11.

Our time series model will leverage Random Forest's ability to analyze intricate relationships between numerous factors, such as fast-food and grocery store density, and socio-economic indicators. Random Forest's ensemble approach helps capture the non-linear relationships between these variables and obesity rates, which provides a more accurate and reliable prediction. Moreover, Random Forest's feature importance analysis will help identify key drivers of obesity. Its resistance to overfitting ensures reliable forecasts of future obesity trends.

3.2.3 XGBoost

XGBoost (Extreme Gradient Boosting) [CG16] builds an ensemble of decision trees sequentially, where each new tree aims to correct the errors made by the previous ones. The key strength of XGBoost lies in its ability to handle large datasets and complex interactions among features, which makes it particularly effective for tasks with intricate patterns.

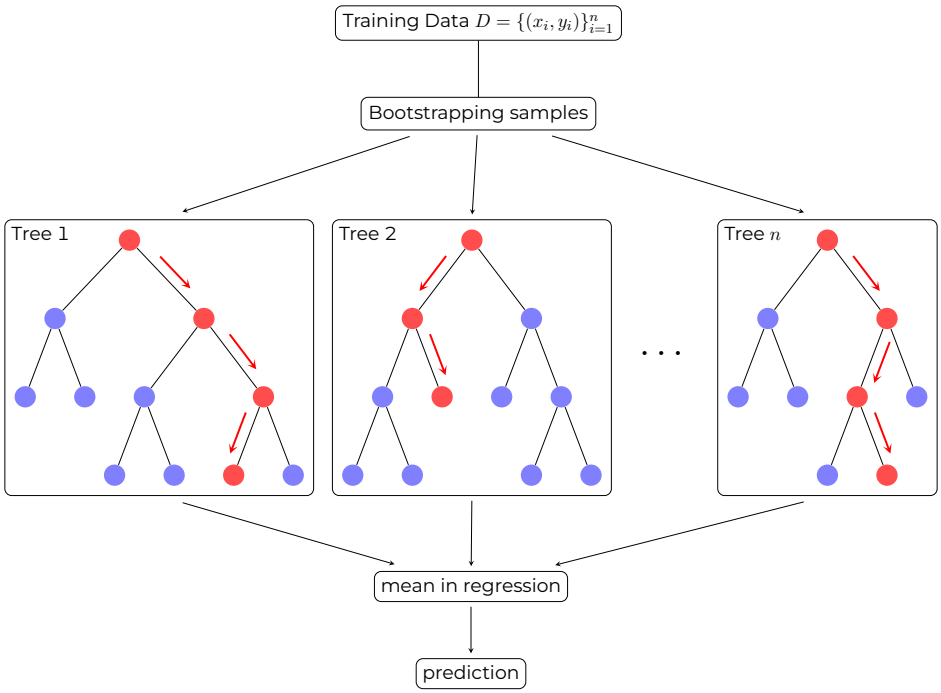


Figure 11: An example of a bagging ensemble method with B bootstrap samples and trees. Source: [RB20]

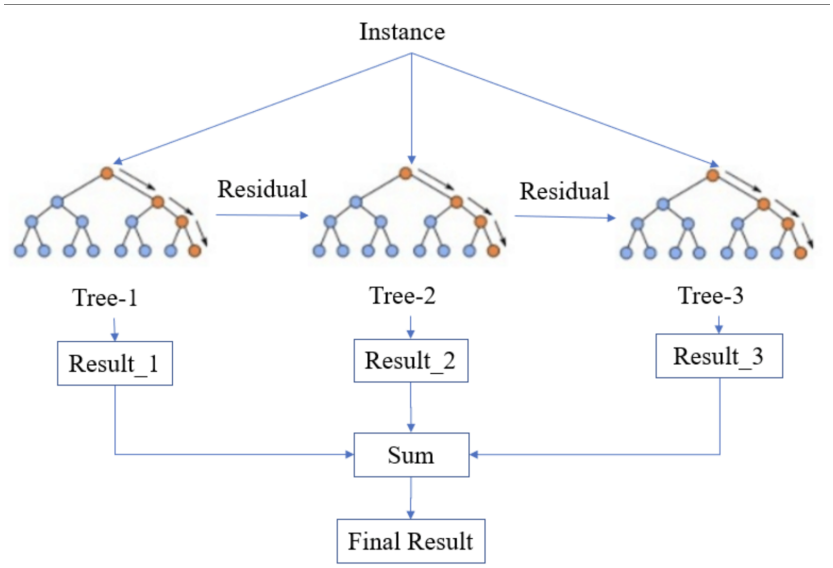


Figure 12: XGBoost Structure. Source: [WCC20]

Our implementation of XGBoost uses the mean square error as the loss function. Let f denote a tree, which is a function in some function space Φ. Consider $L_f(x_i, y_i) = (y_i - f(x_i))^2$. Then MSE is then given as $\text{MSE}_f = \frac{1}{n}\sum_{i=1}^n L_f(x_i, y_i)$. The

model starts with a constant tree f_0 and some predefined total number of trees N. It then iteratively learns a better tree f_i that minimizes the error generated by the previous tree f_{i-1} using the gradient of f_{i-1}. Finally, it takes the weighted average of the output of all N trees and returns the prediction. See a visualization of the algorithm in figure 12.

XGBoost is a highly suitable option for predicting obesity rates in our time series model due to its ability to handle complex interactions between features, such as the availability of fast food, grocery stores, and community health indicators. It provides robust feature importance metrics, which are crucial for understanding the relative impact of each factor on obesity rates. Furthermore, XGBoost's interpretability features allow us to analyze how different community health factors contribute to predictions, aiding in the identification of resilience factors and making targeted recommendations for improving community health.

3.2.4 Autoregressive (AR) Model

An AR model is a type of time series model used for predicting future values based on past values. It is effective for time series data that exhibit autocorrelation, where current values are correlated with past values. The underlying principle is that future values can be predicted by capturing the dependencies and patterns over time.

The AR model is defined by its order p and is denoted as

$$\mathrm{AR}(p).$$

The order p represents the number of lagged observations included in the model. We assume that the current value y_t can be predicted by the lagged values $y_{t-1}, \ldots, y_{t-p}$. Each value in the series is expressed as a linear combination of its previous values plus a random error term. In other words,

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t$$

where $\phi_1, \ldots, \phi_p$ are parameters of the model and ϵ_t is the noise error term at time t with the standard normal distribution. The parameters ϕ_i determine the influence of each lagged value on the current value. They are estimated from the data and indicate the strength and direction of the relationship between past and present values. The error term is added to accommodate for the randomness or noise in the data that cannot be explained by the lagged values. The assumption of white noise ensures that the errors are uncorrelated and have constant variance.

The AR model is suitable for our task of predicting obesity rates over time due to its ability to model temporal dependencies. Obesity rates are influenced by various socio-economic and environmental factors that exhibit patterns and trends over time. By using an AR model, we can capture these temporal patterns, allowing for more accurate predictions of future obesity rates. Given the historical data on obesity rates, fast food density, grocery store availability, and community health features, the AR model can utilize past values of obesity rates to predict future values, taking into account the time-dependent nature of these influences.

3.2.5 LSTM

The Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) designed to handle time series data and sequential tasks by addressing the vanishing gradient problem inherent in traditional RNNs. Compared to Vanilla RNNs, they are designed to learn long-term dependencies by incorporating memory cells that can maintain information over extended periods.

The LSTM model is highly suitable for our task of predicting obesity rates over time due to its ability to capture long-term dependencies and patterns in sequential data. Obesity rates are influenced by a complex interplay of historical data and temporal factors such as fast food density, grocery store availability, and socioeconomic indicators, which exhibit trends over time. The LSTM's architecture, with its memory cells and gating mechanisms, allows it to retain relevant information from past observations while discarding irrelevant data, making it adept at understanding and predicting these temporal dependencies.

Moreover, the LSTM model can handle varying time lags and patterns in the data, which is essential for accurately modeling the influence of past factors on current obesity rates. This capability ensures that the model can capture both short-term fluctuations and long-term trends, providing a comprehensive understanding of how different factors contribute to obesity over time.

3.2.6 Jump-Diffusion Model

Jump diffusion is a stochastic process used to model time series data that exhibit both continuous changes and sudden, discrete jumps. Unlike traditional diffusion models, which assume smooth, continuous paths influenced only by random fluctuations, jump diffusion incorporates occasional, abrupt shifts that can capture more complex behaviors observed in real-world data. This model combines a standard diffusion process, typically modeled by a geometric Brownian motion, with a jump component that accounts for sudden changes in the system. The jump diffusion model is particularly useful in scenarios where data display volatility or irregularities beyond what a simple continuous process can describe.

In reality, obesity rates could experience random jumps due to unforeseen events such as COVID. Jump diffusion can capture both smooth trends and sudden, significant changes in the data. By incorporating both continuous fluctuations and discrete jumps, the jump diffusion model can better reflect the complexities of real-world data, including the impact of increasing fast food availability or decreasing grocery stores. This approach allows for a more nuanced understanding of how these factors influence obesity rates, providing valuable insights for making recommendations to improve community health resilience.

3.3 Considerations

3.3.1 Metrics

We are evaluating our obesity prediction with MSE loss. We will train the models described in section 3.2 and compare the MSE loss of the testing data.

3.3.2 Hyperparameters

Table 2 lists the hyperparameters involved in our model and the descriptions.

Model	Hyperparameters	Description
Linear Regression	N/A	N/A
Random Forest	n_estimator random_state max_depth	the number of trees in the forest a random number generator seeded by the input integer maximum depth of a tree
XGBoost	max_depth	maximum depth of a tree
AR Model		
LSTM	Refer Table 3	model architecture
Jump-Diffusion	N/A	N/A

Table 2: Model Hyperparameters

Layer (type)	Output Shape	Param #	Hyperparameters
LSTM (LSTM)	(None, 1, 50)	14,200	units=50, activation='relu'
Dropout (Dropout)	(None, 1, 50)	0	rate=0.2
LSTM (LSTM)	(None, 50)	20,200	units=50, activation='relu'
Dropout (Dropout)	(None, 50)	0	rate=0.2
Dense (Dense)	(None, 1)	51	units=1

Table 3: LSTM Model Architecture

For the random forest implmenetation, we set n_estimator to be 100 and random_state to 42.

3.3.3 Pipeline

For model training, we decided to mainly focus on state-wise datasets. The detailed data description is given in table 1. For model training, we designed two ways to feed in the data to the model:

Scheme 1 We grouped the rows with the same "State" value together and trained separate models on each group. Within each group we separate our dataset for 80% training and 20% testing. We then feed the time series data we have

for each state into the model of our interest. We evaluate the training method by computing the MSE of the model prediction on the testing dataset.

Scheme 1 emphasizes specificity of each state at the expense of limiting the size of the training data. We used scheme 1 to train linear regression, random forest, XGBoost, AR, LSTM, and jump-diffusion.

Scheme 2 We extracted pairs of rows with the same "State" value and consecutive "Year" value. Call them year 1 data and year 2 data. We then completely discarded the "State" and "Year" features to obtain pairs of rows that contains obesity rate, socio-economic data, and restaurants/stores proximity of year 1 and year 2. We fed in the pairs of data into the model to predict the obesity rate of year 2.

Scheme 2 allowes us to obtain a larger data set, assuming that "State" is independent of the other feature variables. We used scheme 2 to train linear regression, random forest, and LSTM.

The choice of models for each scheme and the training results are discussed in Section 4.

4 RESULTS

Below are the results and analysis of our implementations on the models mentioned in section 3.2:

4.1 Outcomes in Linear Regression – Scheme 1

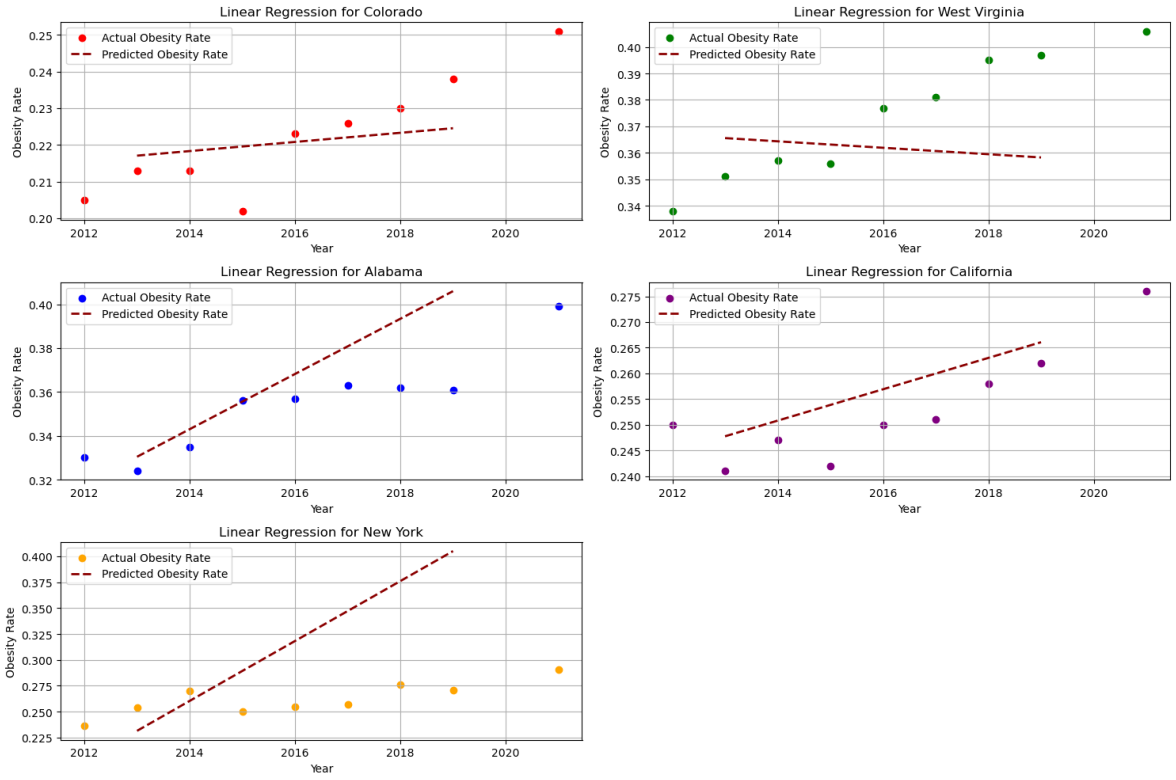


Figure 13: Results of Linear Regression for Five Chosen States: CO, WV, AL, CA, NY.

The linear regression analysis for the five states reveals varying trends. In Colorado, the actual obesity rate shows a slight upward trend, while the predicted rate is relatively flat. West Virginia exhibits a slight downward trend in both actual and predicted rates. Alabama's notable upward trend is well captured by the model, as is New York's significant increase. California shows a slight upward trend, with the model's predictions aligning reasonably well.

The model fits well in Alabama and New York but is less accurate in Colorado and West Virginia. It shows sensitivity to minor trends in Colorado and California and captures the downward trend in West Virginia.

4.2 Outcomes in Random Forest Regression – Scheme 1

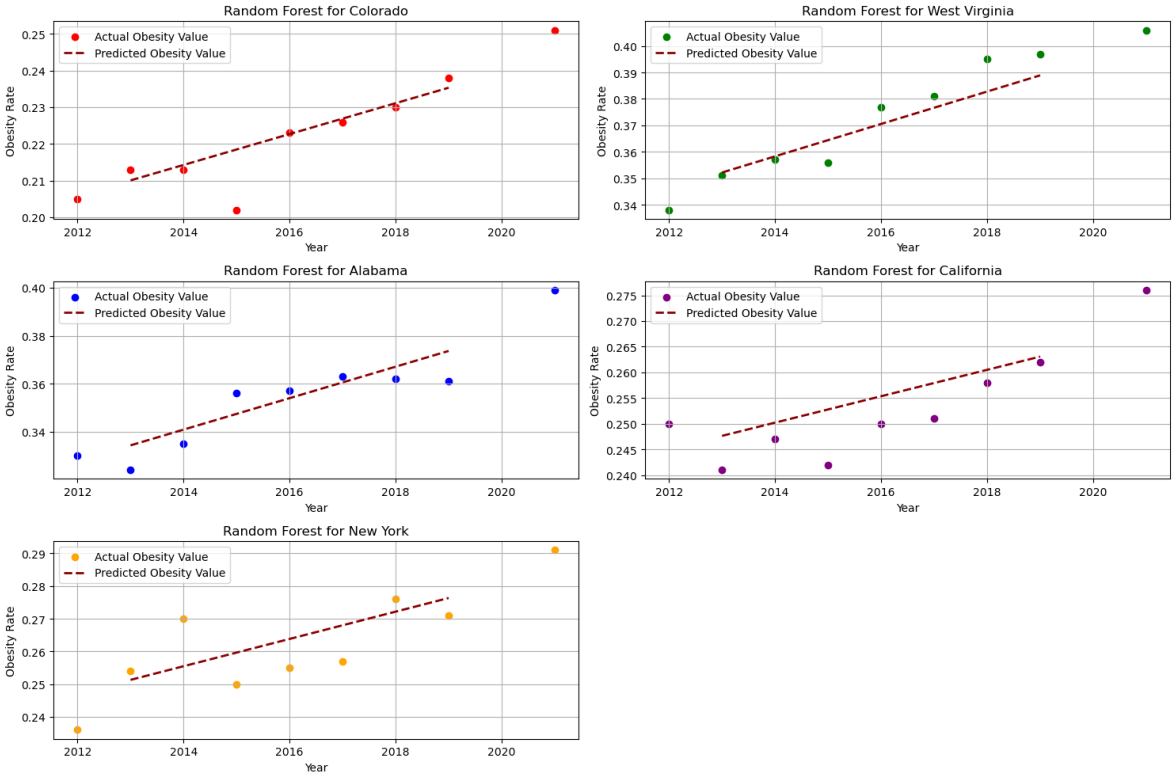


Figure 14: Results of Random Forest for Five Chosen States: CO, WV, AL, CA, NY.

The Random Forest Regression analysis for the five states provides valuable insights into the relationship between fast food accessibility, socioeconomic factors, and obesity rates. In Colorado, both actual and predicted obesity rates show an upward trend, with the model capturing the trend well. West Virginia exhibits an upward trend in both actual and predicted obesity rates, indicating the model's effectiveness in this context. In Alabama, the model accurately predicts the significant upward trend in obesity rates, while in California, the model aligns reasonably well with the slight upward trend observed in the actual data. New York's results show a moderate upward trend in both actual and predicted obesity rates, demonstrating a good model fit.

4.3 Outcomes in XGBoost – Scheme 1

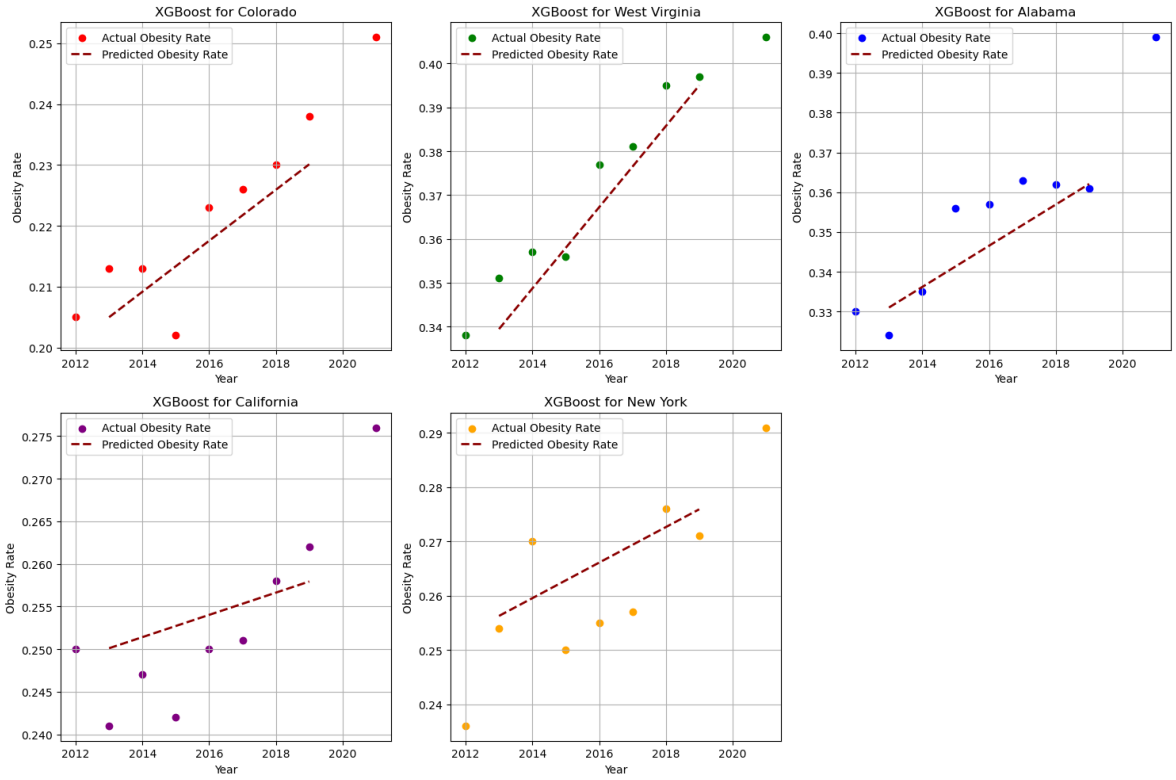


Figure 15: Results of XGBoost for Five Chosen States: CO, WV, AL, CA, NY.

The XGBoost Regression analysis for the five states provides further insights into the correlation between fast food accessibility, socioeconomic factors, and obesity rates. In Colorado, both actual and predicted obesity rates exhibit a clear upward trend, indicating the model's strong predictive power. West Virginia also shows an upward trend, with the model accurately reflecting this increase. Alabama's significant upward trend in obesity rates is well captured by the model, demonstrating its robustness. In California, the actual and predicted rates show a moderate upward trend, with the model providing reasonable predictions. New York's results also indicate an upward trend, with the model aligning closely with the actual data.

4.4 Outcomes in Autoregressive (AR) Model – Scheme 1

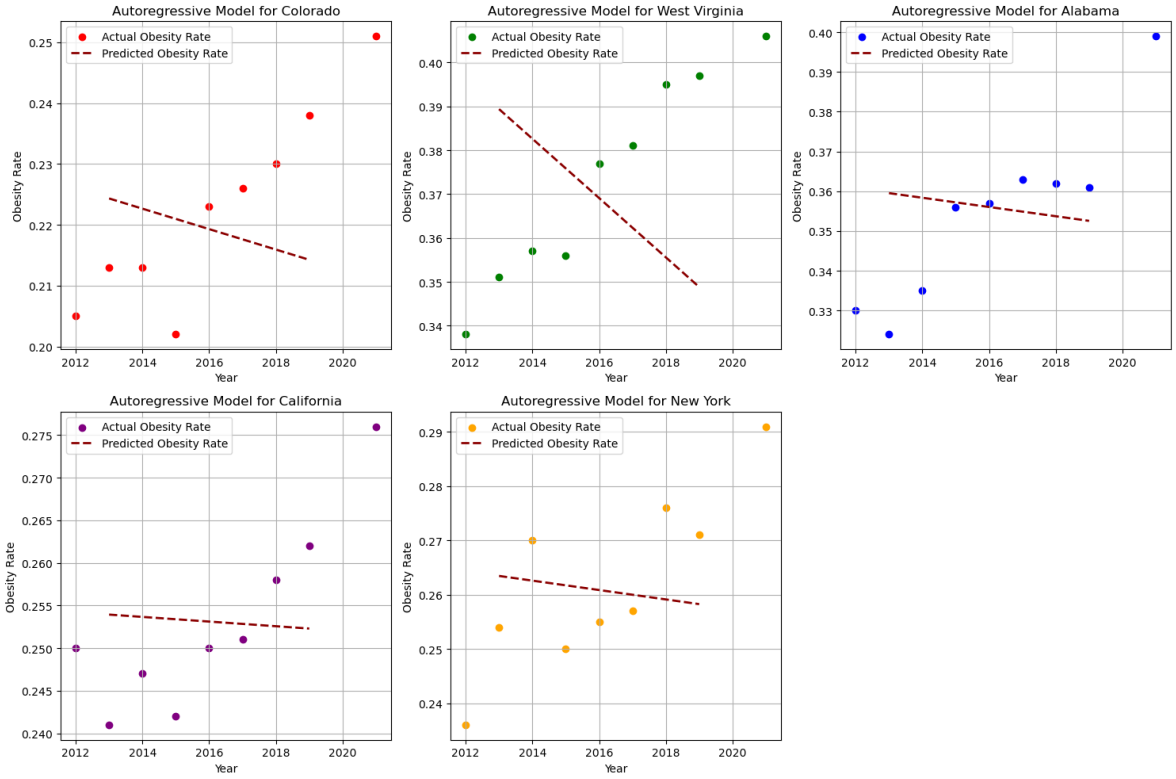


Figure 16: Results of Autoregressive Model for Five Chosen States: CO, WV, AL, CA, NY.

The Autoregressive (AR) Model analysis for the five states fails to accurately capture the trends in obesity rates. In Colorado, the actual obesity rate shows a slight upward trend, while the model predicts a downward trend. West Virginia's actual rates show an upward trend, but the model inaccurately predicts a decrease. In Alabama, the model fails to capture the variability in the actual data and predicts a slight downward trend, contrary to the actual data. California's actual rates show minor fluctuations, but the model predicts a consistent downward trend, failing to align with the observed data. Similarly, in New York, the model predicts a downward trend, while the actual rates fluctuate without a clear trend.

4.5 Outcomes in LSTM – Scheme 1

Due to the limiting number of data for each state, our model in LSTM for Dataset 1 performs badly. In this case, the idea of Dataset 2 is proposed and explored.

4.6 Outcomes in Jump-Diffusion Model – Scheme 1

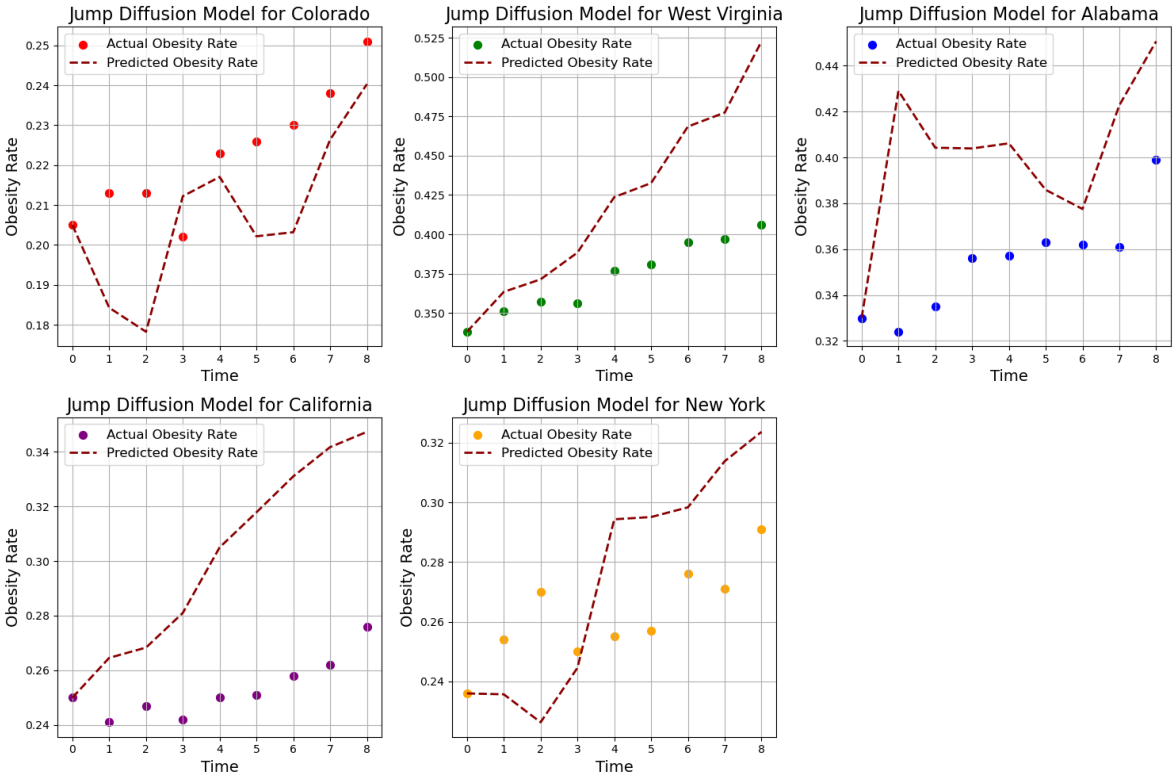


Figure 17: Results of Jump Diffusion for Five Chosen States: CO, WV, AL, CA, NY.

The Jump-Diffusion Model analysis for the five states provides an assessment of obesity rate trends, albeit with limited data points. In Colorado, the model captures some variability but tends to overestimate the later obesity rates. West Virginia's actual obesity rates show little change, while the model predicts significant jumps, indicating poor alignment. In Alabama, the model captures initial variability but fails to reflect the actual data's flatter trend later on. California's model predictions show a steep increase that is not reflected in the relatively stable actual data. Similarly, New York's model predicts significant jumps, which do not align well with the actual data showing minor fluctuations. The Jump-Diffusion Model struggles to accurately capture the trends in obesity rates due to the limited number of data points.

4.7 Outcomes in Linear Regression – Scheme 2

The linear regression analysis was constructed primarily as a baseline model. The MSE for this model is 22.005, which is considerably worse than the subsequent models. This high error rate indicates that the linear regression model is not suitable for this task. Given that we are measuring the obesity rate in percent and the obesity rate has a median value of approximately 30%, an MSE of 22 is particularly problematic. This large error margin suggests that the model's predictions

are significantly off from the actual values, reinforcing the notion that linear regression fails to capture the complexities and temporal dependencies inherent in the data.

4.8 Outcomes in Random Forest Regression – Scheme 2

The second model evaluated was a Random Forest, which achieved an MSE of 6.58. This performance was significantly better than the linear regression model, which had an MSE of 22.005. The Random Forest model, was more adept at capturing non-linear relationships in the data compared to linear regression.

4.9 Outcomes in LSTM – Scheme 2

The LSTM model was trained to predict the obesity rate of Year 2 using a set of 19 feature variables that combine information in both year 1 and year 2. The training process consisted of 200 epochs, with early stopping based on validation loss to ensure optimal performance. The early stopping mechanism restored the best weights once the validation loss ceased to improve, thus preventing overfitting.

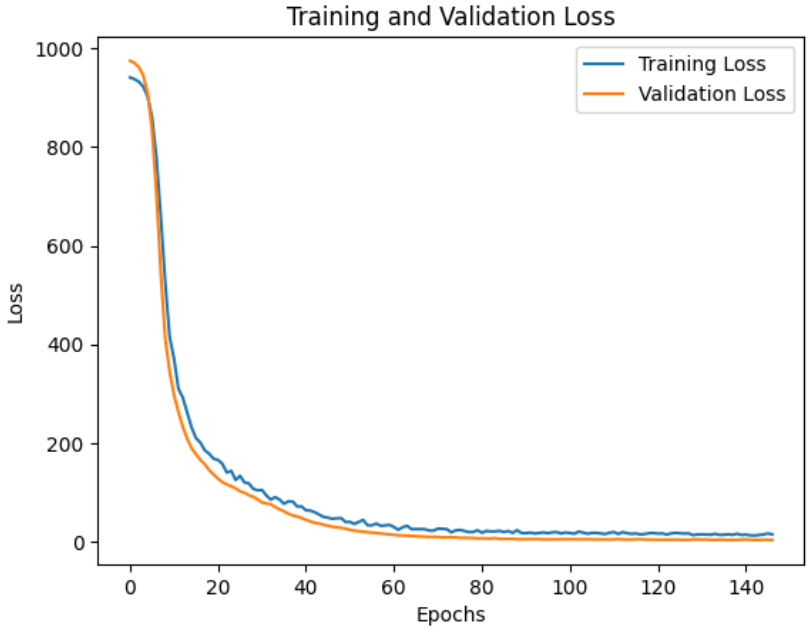


Figure 18: Training and Validation Loss for LSTM.

The training and validation loss curves, shown in Figure 18, exhibit a consistent decrease over the epochs. This trend indicates effective learning, with both curves converging smoothly, suggesting that the model avoids overfitting to the training data.

Upon evaluation on the test set, the LSTM model achieved a Mean Squared Error (MSE) of 3.276. This test loss reflects the model's ability to generalize well to unseen data. For comparison, the Random Forest model and a Linear Regression

model achieved test losses of 6.7 and 22, respectively. The LSTM model outperformed both baselines, demonstrating its ability to capture the temporal dependencies and complex relationships within the data.

This model can serve as a valuable tool for predicting obesity rates based on various socio-economic and health-related factors, especially with further feature analysis provided in subsequent sections.

In contrast, the LSTM model, designed to predict the obesity rate for Year 2 using 19 feature variables, demonstrated excellent performance. Trained over 200 epochs with early stopping based on validation loss, the LSTM model effectively learned from the data without overfitting, as indicated by the smooth convergence of the training and validation loss curves. The model achieved a low test loss with an MSE of 3.1677, reflecting its strong generalization capability.

When compared to baseline models, the LSTM model significantly outperformed both the Random Forest model, which had a test loss of 6.7, and the linear regression model. This superior performance underscores the LSTM model's ability to capture temporal dependencies and complex relationships within the data, making it a valuable tool for predicting obesity rates based on various socio-economic and health-related factors.

In summary, while the linear regression model serves as a baseline, its high MSE highlights its inadequacy for this task. The LSTM model's low test loss and effective learning process make it a far more reliable predictor for obesity rates, demonstrating the importance of choosing appropriate modeling techniques for accurate predictions in complex datasets.

5 EVALUATION

5.1 Global Feature Importance

We analyze the global feature importance for both the random forest and LSTM on dataset 2.

5.1.1 Random Forest Global Feature Importance

To further understand the Random Forest model's predictions, a feature importance analysis was conducted. The feature rankings by importance are shown in the accompanying plot in Figure 19. The features percent of Highschool graduates in year 1, percent of Highschool graduates in year 2, and number of groceries in year 1 were identified as the most influential in predicting the obesity rate for Year 2. These features have the highest importance scores, indicating that they play a significant role in the model's predictions. This insight helps identify the most influential factors contributing to obesity rates, providing valuable information for targeted interventions and policy-making.

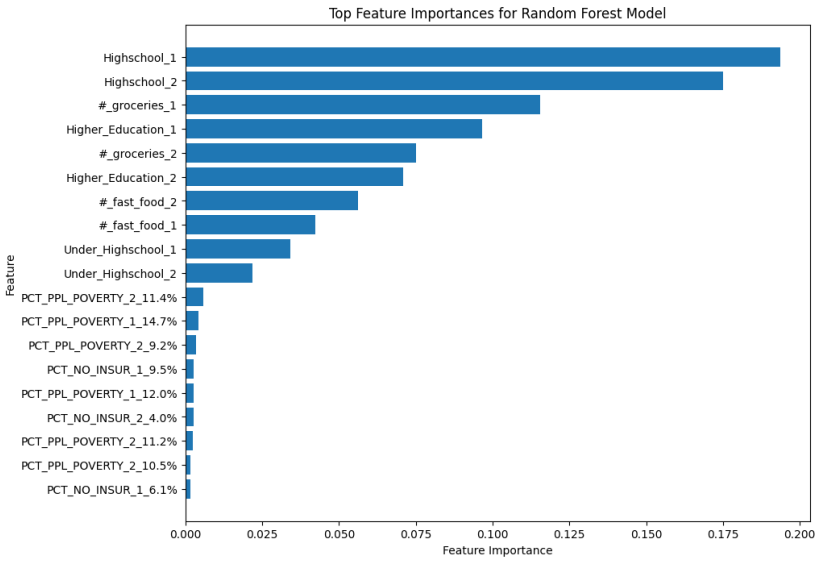


Figure 19: Global Feature Importance for Random Forest.

We identify that the number of fast foods in year 1 and 2 play a significant role in predicting the obesity rate in year 2. Furthermore, the number of fast food restaurants in year 2 has greater feature importance, which indicates that recent exposure to fast food establishments is more strongly correlated with obesity rates. This suggests that policies aimed at reducing the number of fast food outlets could have a more immediate impact on obesity rates, highlighting the importance of current food environment interventions in mitigating obesity.

5.2 Local Feature Importance

5.2.1 Random Forest Local Feature Importance

The figure presented is a SHAP (SHapley Additive exPlanations) summary plot, which displays the impact of various features on the output of a machine learning model. Each point on the plot represents a SHAP value for a particular feature and data instance. The features are listed on the y-axis, while the x-axis shows the SHAP value, indicating the impact on the model's prediction. The colors represent the feature values, with blue indicating lower values and red indicating higher values. This plot provides insight into how each feature contributes to the model's predictions, whether positively or negatively.

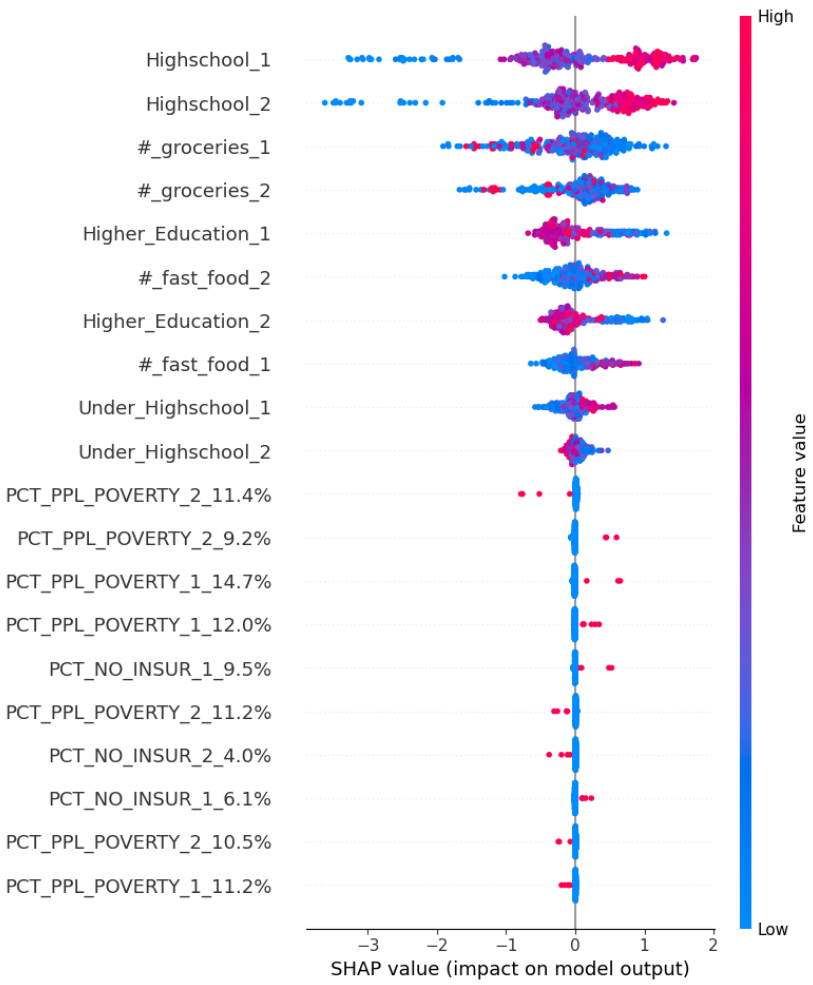


Figure 20: Local Feature Importance for Random Forest.

From above, "Highschool_1" and "Highschool_2" have a significant impact on the model output, with higher values (in red) having a positive effect and lower values (in blue) having a negative effect. Features such as "#_groceries_1" and "#_groceries_2" also show notable contributions but with a more concentrated range of SHAP values. Interestingly, "Higher_Education_1" and "Higher_Education_2" display a mixed impact, suggesting that their influence varies depending on their value. The presence of features like "PCT_PPL_POVERTY_2_11.4%" and "PCT_NO_INSUR_2_4.0%" towards the bottom of the plot indicates they have a less pronounced effect on the model's predictions. Overall, this SHAP summary plot provides a detailed overview of feature importance and interaction within the model, highlighting which features are most influential and how their values affect the model's output.

5.2.2 LSTM Local Feature Importance

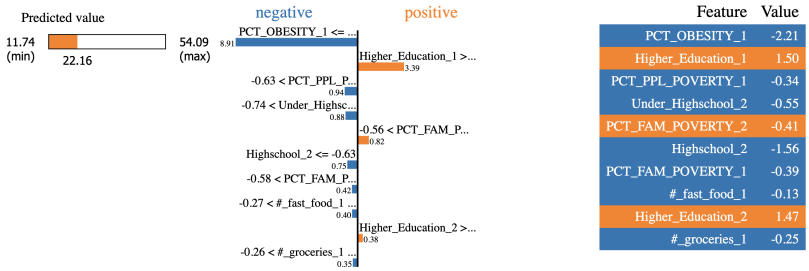


Figure 21: Local Feature Importance for LSTM Point 1.

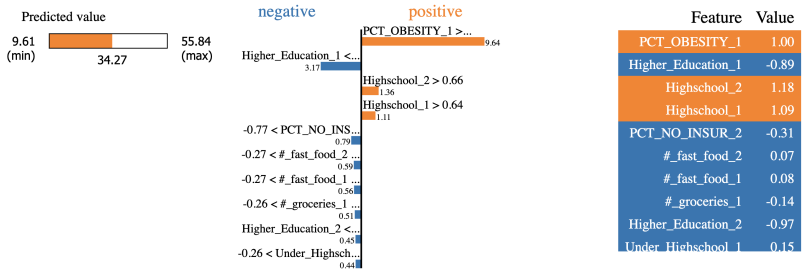


Figure 22: Local Feature Importance for LSTM Point 2.

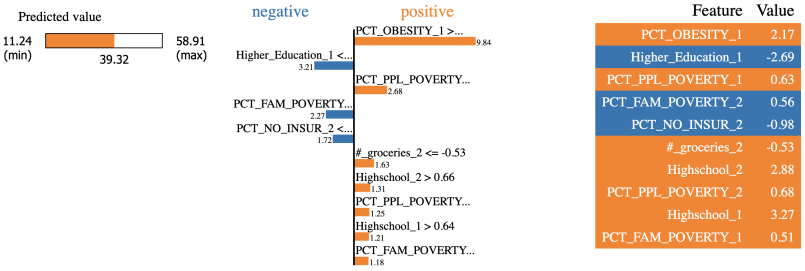


Figure 23: Local Feature Importance for LSTM Point 3.

LSTM models, known for their ability to handle sequential and time-series data, often function as black boxes, making it challenging to understand how input features influence their predictions. Local Interpretable Model-agnostic Explanations (LIME) is used for LSTM models to provide interpretable insights into the model's predictions. By employing LIME, we can generate local explanations for individual predictions, highlighting the contribution of each feature. This interpretability is crucial for validating the model, gaining insights into the factors driving predictions, and making informed decisions based on the model's outputs, especially in complex domains like public health and obesity research.

The LSTM model's performance was evaluated using the LIME explainer on three distinct data points representing low, medium, and high predicted obesity rates shown in Figures 21, 22, and 23, respectively. The analysis of these data points provides insights into the factors influencing obesity rates and highlights differences in resilience between marginalized and healthier communities.

For the instance with a low predicted obesity rate of 22.16%, the negative contributions include 'PCT_OBESITY_1' (previous year's obesity percentage), 'PCT_PPL_POVERTY_1' (percent of population in poverty in year 1), and 'Under_Highschool_2' (percent of population under high school education in year 2). These factors indicate that lower previous obesity rates and higher education levels contribute to better health outcomes. Positive contributors such as 'Higher_Education_1' and 'PCT_FAM_POVERTY_2' have a lesser impact, suggesting that healthier communities with higher education levels and lower poverty rates are more resilient to obesity, even when there are multiple fast food outlets present.

For the instance with a medium predicted obesity rate of 34.27%, factors like 'Higher_Education_1', '#_fast_food_2', and 'PCT_NO_INSUR_2' (uninsured individuals in year 2) contribute negatively, indicating that higher education and fewer uninsured individuals help mitigate obesity. The most significant positive factor is 'PCT_OBESITY_1', followed by 'Highschool_2' and 'Highschool_1', which reflect the previous year's obesity rate and high school education levels. Here, the previous obesity rate plays a significant role, but the presence of mitigating factors like higher education and better insurance coverage indicates some level of resilience.

For the instance with a high predicted obesity rate of 39.32%, the most influential negative factor is 'Higher_Education_1', showing that higher education strongly correlates with lower obesity rates. Significant positive contributors include 'PCT_OBESITY_1', 'PCT_PPL_POVERTY_1', and 'PCT_FAM_POVERTY_2', indicating that previous high obesity rates and higher poverty levels drive the obesity rate up. This suggests that marginalized communities, characterized by higher poverty rates and lower education levels, are less resilient to obesity. The presence of grocery stores and high school education does little to counteract the strong influence of poverty and previous obesity rates.

The LIME analysis of the LSTM model's predictions reveals insights into the determinants of obesity rates. Higher education consistently contributes to lower obesity rates across all instances, highlighting its protective effect. Communities with better educational attainment are more resilient to obesity, even in the presence of fast food restaurants. Poverty levels, both family and general, are strong positive contributors to obesity rates. Marginalized communities with higher poverty levels show less resilience, with poverty exacerbating the effects of other negative factors. The obesity rate from the previous year is a significant predictor of the current obesity rate, indicating that interventions need to be sustained over time to have a lasting impact. Interestingly, the number of fast food outlets does not have a uniform impact. While it is a negative contributor in most cases, suggesting that healthier communities can mitigate its impact, marginalized communities often are not as resilient to its impacts.

The LSTM model's feature importance analysis underscores the need for targeted public health interventions focusing on education and poverty alleviation to effectively combat obesity, particularly in vulnerable communities. While there is strong evidence showing the adverse effects of fast food on obesity, the varying magnitude of its impact based on overall community health suggests that simply decreasing the number of fast food chains may not be the only solution

to the obesity epidemic. Furthermore, a deeper analysis of feature importance can guide future policies, enabling better allocation of limited resources, such as investments in education or employment. Although this is a complex issue and these predictors serve as indicators rather than definitive solutions, they provide valuable insights to consider when developing comprehensive strategies to address obesity.

5.3 Evaluation in MSE

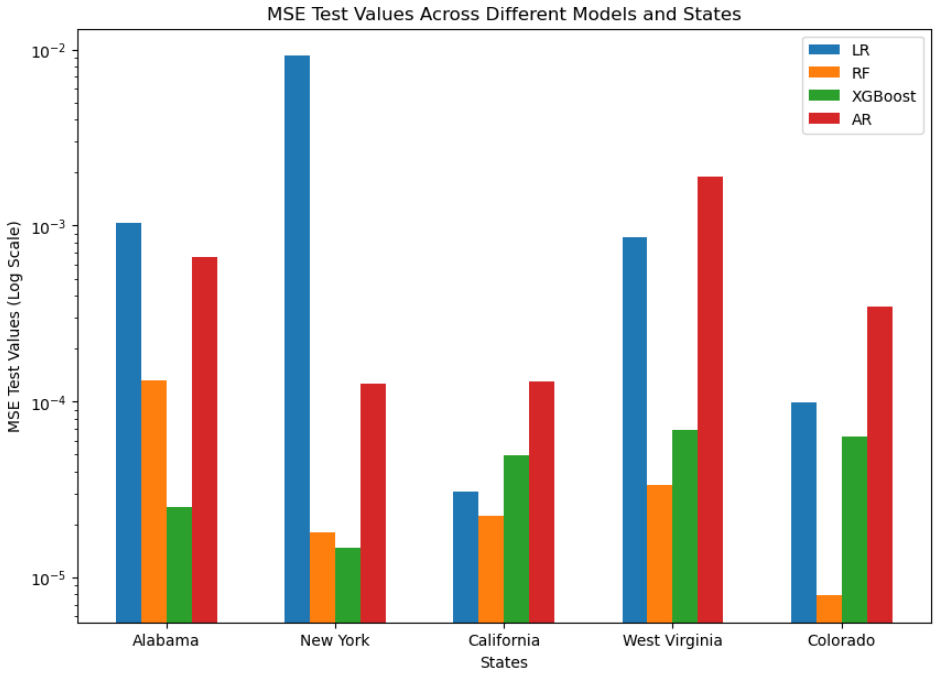


Figure 24: MSE of the Models

The Mean Squared Error (MSE) evaluation across different models and states highlights the varying performance of the Linear Regression (LR), Random Forest (RF), XGBoost, and Autoregressive (AR) models. The graph illustrates that the Random Forest and XGBoost models generally outperform the Linear Regression and AR models, evidenced by their lower MSE values across most states. Notably, Random Forest exhibits the lowest MSE in states such as Alabama, New York, and Colorado, indicating its superior ability to capture the underlying patterns in the data. XGBoost also demonstrates robust performance with relatively low MSE values, particularly in New York and California. In contrast, the AR model consistently shows the highest MSE, reaffirming its inadequacy in accurately predicting obesity rates. These results emphasize the effectiveness of ensemble methods like Random Forest and XGBoost in handling the complexities of the dataset, making them preferable choices for predictive modeling in this context.

5.4 Remarks

In our analysis, we utilized state-level data, which provided a higher level of aggregation than initially desired. Our original goal was to analyze data at the county level to gain more granular insights. However, due to time constraints and the complexity involved in handling county-level data, this was not feasible.

Despite this limitation, we evaluated several models to identify the best approach for predicting obesity rates. Among the models tested, the Random Forest model performed the best in terms of accuracy and capturing the trends in the data. However, it is important to note that even the best model had limitations and was not perfect.

The Random Forest model demonstrated robustness in handling the available state-level data, but the predictions were still constrained by the yearly scale of the data points. As we accumulate more data over time, particularly with more frequent data points, the accuracy and reliability of the model can be expected to improve. This highlights the ongoing need for data collection and model refinement to better inform public health strategies and interventions.

6 CONCLUSION

6.1 Considerations in Models

The analysis would have benefited from more granular data at the county level, which could provide a deeper insight into localized trends. However, the yearly scale of the data points presents challenges in making precise predictions. As more data becomes available over time, the models can be refined and improved for better accuracy.

6.2 Policy Suggestions at State Level

The LIME model suggests a high feature importance between fast food prevalence, education level, and obesity rates. This indicates a need for more research into these factors to develop targeted interventions. Policies at the state level should include:

- **Regulating Fast Food Density:** Implement zoning laws to limit the number of fast food outlets in certain areas, particularly near schools and residential neighborhoods.
- **Promoting Education and Awareness:** Develop statewide educational programs that emphasize the importance of nutrition and healthy eating habits. Schools should integrate nutrition education into their curriculums to reach children at a young age.
- **Improving Access to Healthy Foods:** Increase funding for programs that provide access to fresh fruits and vegetables in underserved areas. This can

include subsidies for farmers' markets and incentives for grocery stores to open in food deserts.

- **Economic Support in Low-Income Areas:** Implement policies that provide economic support to low-income families, such as food assistance programs and subsidies for healthy food purchases. Addressing economic disparities can help mitigate the effects of fast food availability in poorer regions.
- **Research and Monitoring:** Fund ongoing research to monitor the impact of these policies and identify new areas of concern. Continuous data collection and analysis will help refine strategies and ensure they are effective.

By implementing these policies, state governments can address the multifaceted issue of obesity and work towards healthier communities.

References

[Gri+07] Sonya A Grier et al. “Fast-food marketing and children’s fast-food consumption: Exploring parents’ influences in an ethnically diverse sample”. en. In: *J. Public Policy Mark.* 26.2 (Sept. 2007), pp. 221–235.

[GK08] Sonya A. Grier and Shiriki K. Kumanyika. “The Context for Choice: Health Implications of Targeted Food and Beverage Marketing to African Americans”. In: *American Journal of Public Health* 98.9 (2008). PMID: 18633097, pp. 1616–1629. DOI: 10.2105/AJPH.2007.115626. eprint: https://doi.org/10.2105/AJPH.2007.115626. URL: https://doi.org/10.2105/AJPH.2007.115626.

[Ric+11] Andrea S. Richardson et al. “Neighborhood fast food restaurants and fast food consumption: A national study”. In: *BMC Public Health* 11.1 (July 2011), p. 543. ISSN: 1471-2458. DOI: 10.1186/1471-2458-11-543. URL: https://doi.org/10.1186/1471-2458-11-543.

[Jam+14] Peter James et al. “Do minority and poor neighborhoods have higher access to fast-food restaurants in the United States?” en. In: *Health Place* 29 (June 2014), pp. 10–17.

[Rei+14] Lorraine R. Reitzel et al. “Density and Proximity of Fast Food Restaurants and Body Mass Index Among African Americans”. In: *American Journal of Public Health* 104.1 (2014). PMID: 23678913, pp. 110–116. DOI: 10.2105/AJPH.2012.301140. eprint: https://doi.org/10.2105/AJPH.2012.301140. URL: https://doi.org/10.2105/AJPH.2012.301140.

[CG16] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[Hag+16] Erin R Hager et al. “Food swamps and food deserts in Baltimore City, MD, USA: associations with dietary behaviours among urban adolescent girls”. en. In: *Public Health Nutr* 20.14 (Sept. 2016), pp. 2598–2607.

[Pos19] The Washington Post. Oct. 2019. URL: https://www.washingtonpost.com/business/2019/10/29/what-parents-should-know-about-how-living-near-fast-food-outlets-could-affect-their-kids/.

[RB20] Janosh Riebesell and Stefan Bringuier. *Collection of standalone TikZ images*. Version 0.1.0. 10.5281/zenodo.7486911 - https://github.com/janosh/tikz. Aug. 9, 2020. DOI: 10.5281/zenodo.7486911. URL: https://github.com/janosh/tikz (visited on 01/01/2023).

[WCC20] Weilun Wang, Goutam Chakraborty, and Basabi Chakraborty. "Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm". In: *Applied Sciences* 11 (Dec. 2020), p. 202. DOI: 10.3390/app11010202.

[Far23] Rhea Farberman. Nov. 2023. URL: https://www.tfah.org/report-details/state-of-obesity-2023/#:~:text=According%20to%20TFAH%27s%20analysis%20of,19%20states%20the%20prior%20year.

[Lib+23] Nicolás Libuy et al. "Fast food proximity and weight gain in childhood and adolescence: Evidence from Great Britain". en. In: *Health Econ* 33.3 (Nov. 2023), pp. 449–465.

[Pat+23] Sukanya Pati et al. "Obesity and Cancer: A Current Overview of Epidemiology, Pathogenesis, Outcomes, and Management". en. In: *Cancers (Basel)* 15.2 (Jan. 2023).

[Bura] U.S. Census Bureau. *County Business Patterns (CBP) APIs*. Accessed on 4 August 2024. United States Census Bureau. URL: https://www.census.gov/data/developers/data-sets/cbp-zbp/cbp-api.2022.html#list-tab-711980547.

[Burb] U.S. Census Bureau. *Educational Attainment*. U.S. Census Bureau. Accessed on 4 August 2024. URL: https://data.census.gov/table/ACSST1Y2022.S1501?q=state%20population%20by%20education&g=010XX00US$0500000.

[Burc] U.S. Census Bureau. *SELECTED ECONOMIC CHARACTERISTICS*. U.S. Census Bureau. Accessed on 4 August 2024. URL: https://data.census.gov/table/ACSDP5Y2020.DP03?g=010XX00US$0400000&y=2020&d=ACS%205-Year%20Estimates%20Data%20Profiles&tp=false.

[Burd] U.S. Census Bureau. *Small Area Health Insurance Estimates*. Accessed on 4 August 2024. United States Census Bureau. URL: https://www.census.gov/data/developers/data-sets/Health-Insurance-Statistics.html.

[Dis] Center of Disease Control. *Nutrition Physical Activity and Obesity Data*. Accessed on 4 August 2024. Behavioral Risk Factor Surveillance System. URL: https://www.cdc.gov/brfss/.