Social Determinants of Health Impact on Quality of Life in US Counties: A Predictive Analysis

Final Written Deliverables

Authors: Steven Uzupis, Udumaga Onyeukwu, and Miranda Gemme
Each author contributed equally to the design, coding & development, analysis, and writing of this project

Data Science Master's Program
Merrimack University
Course: DSE6311 Data Science Capstone
Supervisor: Dr. Katherine Geist
Date: August 25, 2024

Table of Contents

Abstract......3
Background & Question......3
Data...... 4
Data Acquisition...... 5
Data Cleaning......5
Data Exploration...... 8
Models...... 11
Pre-processing and dimensionality reduction and feature engineering...... 11
Dimensionality Reduction...... 13
Algorithm(s) Selection......14
Final Model Selection......18
Conclusions......23
Discussion...... 28
Appendix......33
Code and Documents......33
References......34

Abstract

This report analyzes how Social Determinants of Health (SDOH) shape the quality of life across U.S. counties, using predictive modeling to uncover the most significant factors. Data from the Agency for Healthcare Research and Quality (AHRQ) and the County Health Rankings (CHR) were analyzed to predict self-reported health status, a key measure of well-being. The study identifies significant socioeconomic and environmental determinants, such as income, education, and household size, which correlate with health outcomes. The Gradient Boosting Machine (GBM) model showed superior predictive accuracy, emphasizing the critical role of these factors in public health. The findings emphasize the need for targeted, data-driven interventions to reduce health disparities and improve quality of life, offering useful insights for policymakers and health practitioners.

Background & Question

The Centers for Disease Control and Prevention (CDC) has extensively studied social determinants of health (SDOH), providing valuable insights into how factors including income, education, and access to healthcare influence health outcomes. These findings highlight the potential of addressing SDOH to improve health and quality of life, serving as a crucial foundation for this project. Inspired by the CDC's work, this project aims to answer the research question: “How do social determinants of health affect quality of life in different localities?” The primary objective is to predict the self-reported health status of the adult population in US counties using available data on SDOH and County Health Rankings (CHR). Self-reported health status, which reflects the percentage of adults that rank their health as fair or poor, “is a simple, easy to administer measure of general health” and a good indicator of overall well-being, encompassing physical, mental, and social aspects (Bombak, 2013). The project will also evaluate the relative importance of various social behaviors and environments on these health outcomes.

Hypothesizing that social determinants such as economic stability, social connectedness, access to healthcare, and neighborhood environment significantly predict self-reported health status in US counties, the project predicts that counties with higher economic security, robust social support, better access to healthcare, and safer neighborhoods will report better overall health status compared to those lacking these factors.

By investigating this research question, the project aims to provide localized, data-informed insights that can guide the allocation of resources and interventions based on the most significant and impactful determinants. This approach can help decrease health disparities by identifying and addressing the causes of poor health outcomes. While the relationship between SDOH and health outcomes is well-documented, this project's use of predictive modeling and focus on county-level analysis offers valuable contributions to public health strategies and social research.

Data

The data for this project were sourced from two primary datasets:

1. **Social Determinants of Health (SDOH)** Database provided by the Agency for Healthcare Research and Quality (AHRQ) from the year 2020.
2. **County Health Rankings (CHR)** dataset provided by the University of Wisconsin Population Health Institute from the year 2020.

** See References for more information, including URLs (Uniform Resource Locators) for datasets.*

Data Acquisition

These datasets were chosen because they provide comprehensive and well-documented variables that are crucial for analyzing the impact of social determinants on health outcomes. The SDOH database covers various domains, including social, economic, healthcare, and environmental factors, which are all critical to understanding health disparities. The CHR dataset includes detailed health outcomes at the county level, such as the percentage of adults reporting fair or poor health, which serves as the response variable.

The SDOH dataset contains 3,229 observations with 682 features, while the CHR dataset consists of 3,194 observations with 720 features. The original data from the SDOH and CHR datasets were collected through surveys.

**Please refer to the Discussion section for more information about the concerns and caveats regarding the methods used for data collection.*

Data Cleaning

Step 1: Using Domain Knowledge to Remove Unwanted Features.

The first step involved consulting the 'stakeholders' to identify and remove irrelevant features. Following this 'discussion,' a total of 582 features were removed from the SDOH dataset, and 659 features were removed from the CHR dataset.

Step 2: Combining the Two Datasets into One.

The SDOH and CHR datasets were then merged using the county FIPS code as a common identifier, resulting in a unified dataset called "qol_data" (quality of life data). This merged dataset offers a comprehensive perspective on social determinants across U.S. counties, making it suitable for our analysis of health outcomes.

Step 3: Removing Features with Significant Missing Data.

After combining the two datasets into one, titled “qol_data,” features with a significant number of missing values were identified. The goal was to retain approximately 3000 unique observations by removing features with extensive missing data. This process resulted in the elimination of an additional 50 features.

Step 4: Identifying and Removing Highly Correlated Features.

After analyzing the dataset, highly correlated features (|r| > 0.7) were identified and evaluated for high collinearity. Features with redundant information were removed to avoid collinearity in the models, removing 27 features. Table 1 below shows a sample of highly correlated features that were considered for removal.

Table 1. *Collinearity of Predictors (Sample of Features with High Correlation).*

Variable 1	Variable 2	Correlation
pct_home_owner	weighted_population	0.9768797
pct_female	pct_male	-1
pct_naturalized_citizens	pct_not_citizens	0.7460535
pct_adult_citizens	pct_not_citizens	-0.7604839
pct_no_english_spoken	pct_not_citizens	0.8039729
pct_hispanic	pct_not_citizens	0.7112025
pct_asian	pct_naturalized_citizens	0.7205063
pct_0_17_age	pct_adult_citizens	-0.7853593
pct_hispanic	pct_no_english_spoken	0.7028031
pct_white	pct_black	-0.8307521

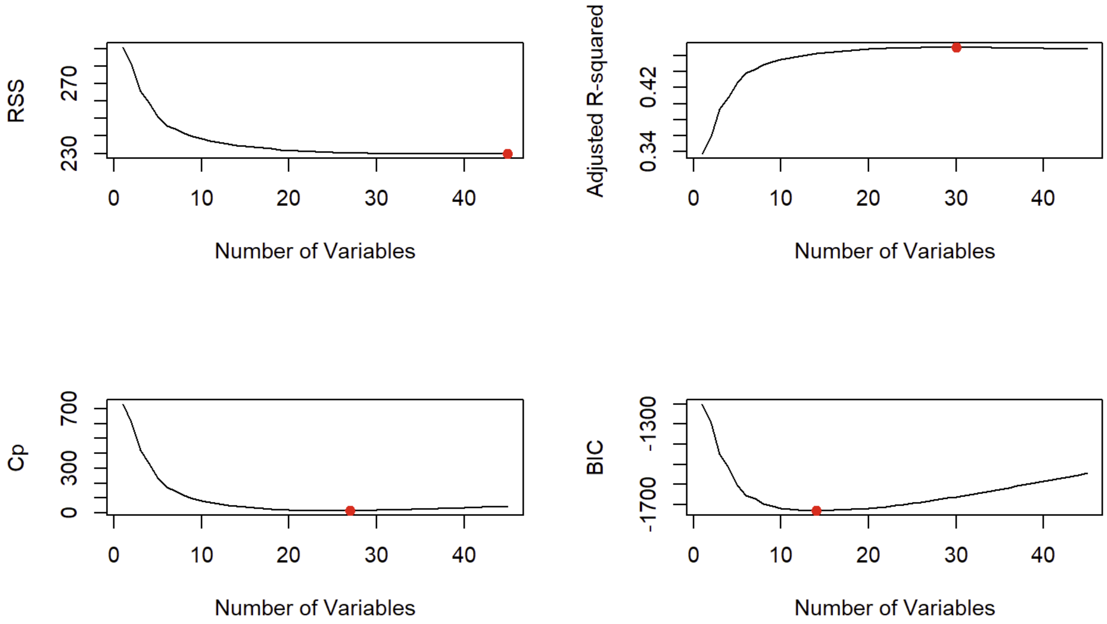
Step 5: Removing Features with Near Zero Variance.

The dataset was further analyzed to identify features with near-zero variance that would add little to no value to predictive models. One feature met this criterion and was removed.

Step 6: Regression Model and Unsupervised Methods for Feature Selection

Following the initial feature reduction, a regression model was developed to assess the significance of the remaining predictors in forecasting the response variable. However, the code was run, and a computational limit wall was run into where local computers could not process the data. The large number of predictors added enough complexity to any models that further resources, such as AWS servers, would need to be used to process the code. To address this, a decision was made to streamline the model further by removing 32 additional features considered low relevance based on stakeholder feedback and group assessment. The regsubset() function was then applied to perform unsupervised modeling, aiming to identify the optimal number of features for the final model.

Figure 1. *Ideal Subset Selection.*

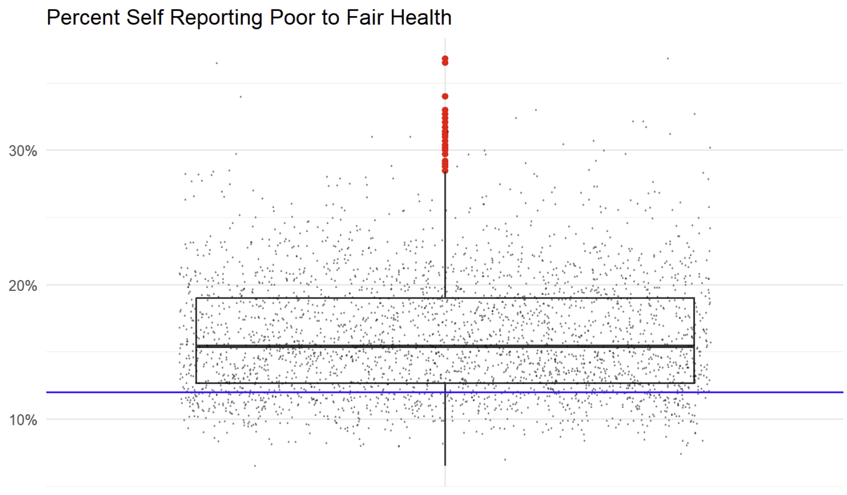


As seen in Figure 1 above, relying solely on metrics like the minimum residual sum of squares (RSS) and Adjusted R-squared could lead to overfitting, selecting as many as 45 and 38 features, respectively. Similarly, Mallows' Cp suggested 26 features, which might still be excessive. Instead, the Bayesian Information Criterion (BIC) was considered more appropriate at this point, as it recommended retaining just 14 features. This will be later examined further and refuted. Further analysis suggested an alternative number of features.

Data Exploration

Data exploration is a step in data analysis where visualization and statistical techniques are utilized to understand the characteristics of a dataset, including its size, structure, and accuracy (Data Exploration, n.d.). This process helps identify relationships between variables, detect outliers, and understand the distribution of data values. By revealing patterns and points of interest, data exploration provides a better view of the raw data, which helps guide further analysis and decision-making (Data Exploration, n.d.). The visualizations below were used during the data exploration process and was essential for further analysis steps.

Figure 2. *Distribution of the Target Response Vector.*



The boxplot shown on the left, *Figure 2*, is a critical part of the data exploration phase. Visualizing the distribution of the response variable, "Percent Self Reporting Poor to Fair Health," allows for observing skewness, identifying potential irregularities, and gaining insight into the trend of the data. The mean value is higher than the national average, represented by the blue line at 12%. There are also outliers above the mean. Based on these observations, a new response variable was considered based on whether the value was better or worse than the new median of 15.4%.

Figure 3. *Percentage of Smokers in a County vs. Percentage of Adults Reporting Poor to Fair Health.*

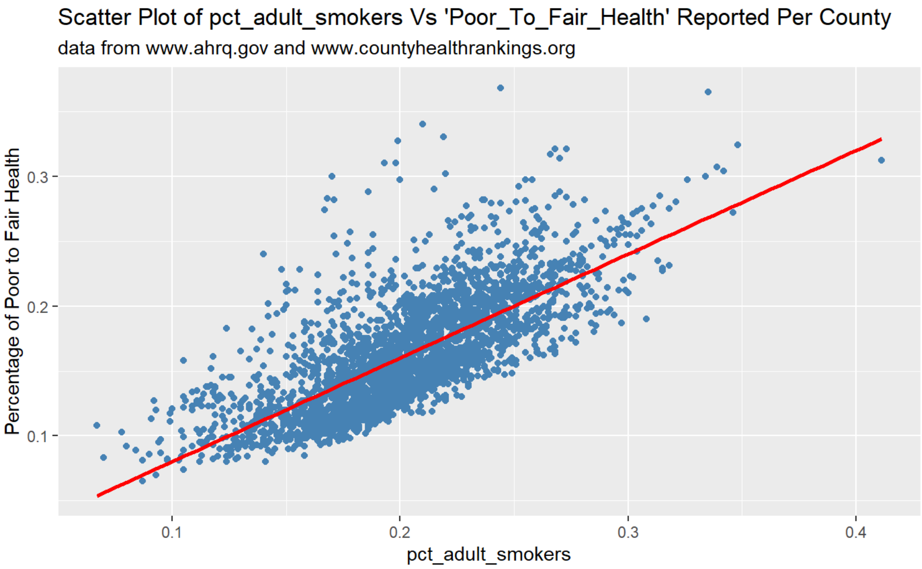
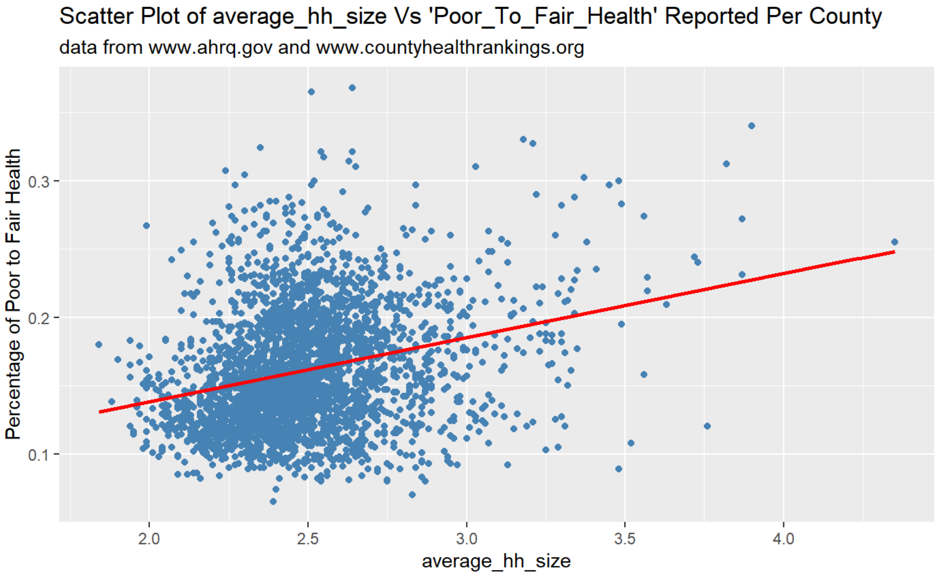


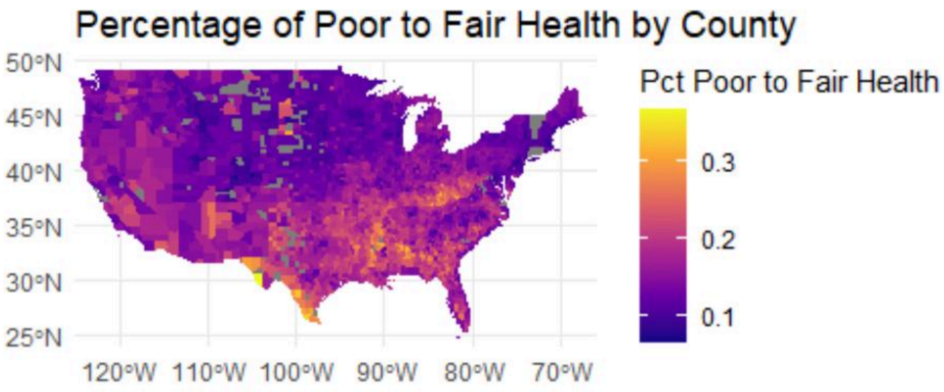
Figure 3 (on the left) illustrates the relationship between the percentage of smokers in a county and the percentage of adults reporting poor to fair health. The scatter plot reveals a positive correlation, indicated by the upward trend in the data points and the red regression line. This suggests that as the percentage of smokers in a county increases, the percentage of adults reporting poor to fair health also tends to rise. The strength and direction of this correlation show smoking as a significant factor contributing to negative health outcomes at the county level.

Figure 4. *Average Household Size vs. Percentage of Adults Reporting Poor to Fair Health*



The scatter plot on the right, *Figure 4*, reveals a positive correlation between average household size and the percentage of adults reporting poor to fair health. This not only suggests that larger household sizes are associated with higher percentages of poor health outcomes, but also underscores the broader implications. The trend indicates that households with more members face greater challenges related to resources, stress, or health management, contributing to the increased prevalence of poor health.

Figure 5. *Percentage of Poor to Fair Health by County.*



The heatmap on the left, *Figure 5,* shows the Percentage of Poor to Fair Health by County in the United States. The color gradient from purple to yellow represents varying percentages of the population reporting as poor to fair health. Areas in the Southeast and parts of the Midwest show higher concentrations of poor health outcomes, while the Northeast and West tend to exhibit better health outcomes. This shows potential disparities in health status across regions. The heatmap is helpful for identifying areas that may be requiring public health interventions to improve community well-being.

Table 2. *Summary of Health and Demographic Characteristics by Region.*

Characteristic	Midwest (N = 6,108)	Northeast (N = 1,194)	South (N = 8,124)	West (N = 2,304)
pct_poor_to_fair_health	0.14 (0.03, 4.90, 0.96)	0.12 (0.02, 2.66, 0.15)	0.19 (0.04, 3.68, 0.41)	0.14 (0.03, 3.62, 0.60)
life_expectancy_years	77.67 (2.60, 7.66, -0.55)	78.75 (1.75, 3.03, 0.19)	75.37 (2.67, 3.78, 0.30)	78.88 (3.58, 10.91, 1.09)
pct_voters	0.68 (0.08, 3.66, -0.23)	0.68 (0.08, 3.66, 0.25)	0.61 (0.09, 3.24, 0.09)	0.71 (0.10, 3.01, -0.25)

Above is Table 2, which summarizes the response variable broken up by region of each observation. The two response variables include: self-reporting poor to fair health and life expectancy in years. A predictor was included to determine any correlation, and the calculated values are listed in the attached table.

Models

Pre-processing and dimensionality reduction and feature engineering

Feature Engineering:

After conducting our Exploratory data analysis one new features was created. The feature created was “percent_grandparents_as_guardians”. The value of this new column is calculated by taking the percentage of children living with grandparents (ACS_PCT_CHILDREN_GRANDPARENT) and multiplying it by the sum of the percentages of grandparents who are responsible for their grandchildren (ACS_PCT_GRANDP_RESPS_P) and those who are not (ACS_PCT_GRANDP_RESPS_NO_P), divided by 100. The three constituent columns were then removed from the dataset. The new feature, representing, percent of grandparents who were raising children, was created to capture this unique family dynamic and support system within each county. Research[4] has shown that the presence of grandparents as primary caregivers can significantly impact the overall well-being and health outcomes of children. By incorporating this feature, we aimed to better understand the relationship between family structure and self-reported health status at the county level.

At this point a new dataset was created by merging the two datasets by the FIPS code which is a unique identifier for each count.

Preprocessing:

Split Data into Training and Test Datasets: The data split ratio was determined using the formula: $\frac{1}{\sqrt{P}} \times \text{nrow(df)}$

where P is the number of features in the dataset and nrow(df) represents the total number of observations. This formula ensures a data split that provides a sufficient sample size for training the model while reserving an adequate portion for evaluation. In our case, with 56 predictor variables and a total of 3,142 observations, the formula yielded a split ratio of approximately 86:14, resulting in a training set of 2,700 instances and a test set of 442 instances.

Encode Categorical Variables: One-hot encoding was used to encode the categorical features in the dataset, creating new binary features. For example, the "Region" feature was encoded into four binary features to represent the four regions of the country: North East, Midwest, South, and West.

Impute or Remove Missing Values: Missing values in the dataset were imputed using the Classification and Regression Trees (CART) method, chosen for its ability to handle both continuous and categorical variables. CART constructs decision trees to predict missing values based on the available data, aiming to preserve variable relationships and minimize the impact of missing data on the analysis. However, the feature “pa_pt” was excluded from the dataset. The decision to exclude the feature was made after a careful examination of its missing value pattern. It had a substantial proportion of missing values (>50%), and further analysis revealed that it did not significantly contribute to model's predictive performance. Therefore, excluding it helped reduce dimensionality and improve model efficiency without compromising accuracy. Additionally, features such as “fips_code,” “county,” and “state” were removed because they served as identifiers for instances or groups of instances and lacked predictive value. Removing these features further reduced dimensionality.

Normalize and/or scale the features: We performed centering and scaling of the predictors to have a mean of zero and a standard deviation of one.

Dimensionality Reduction

Unsupervised Methods - PCA (Principal Component Analysis): A Principal Component Analysis (PCA) was conducted to determine the optimal number of predictors for this dataset. PCA reduces the dimensionality of the data while retaining most of the variance. The PCA suggested that a model with 29 features could explain 90% of the variance in the response variable, while 35 features could explain 95%. However, a parallel analysis, which compares the eigenvalues of the data with those obtained from random data, indicated that using a subset of just 9 features would be sufficient to explain 95% of the variance. Ultimately, we decided to use algorithms that are robust to high dimensionality, such as boosting and ensemble methods, as they can effectively handle the complexity and interactions between a large number of features. By maintaining the large feature set while using methods robust to high dimensionality, we aim to capture the nuanced relationships in our data while still providing meaningful insights into the factors influencing county-level health status.

Algorithm(s) Selection

Models

This analysis developed and compared multiple regression models to identify the best fit for the data. Four initial models were evaluated: Ordinary Least Squares (OLS), Lasso Regression, Gradient Boosting Machine (GBM), and Random Forest. These models were trained and evaluated using a dataset with 56 predictor variables and a target variable. Performance was assessed using initial model results and cross-validation outcomes, except for the Random Forest model, which used out-of-bag (OOB) error estimation.

These models were chosen to cover a range of modeling approaches, from simple linear models to more complex ensemble methods, allowing us to compare their performance and suitability for our specific dataset.

Ordinary Least Squares (OLS): Chosen as a baseline linear model for its simplicity and interpretability.

Lasso Regression: Selected to address potential multicollinearity and perform feature selection.

Gradient Boosting Machine (GBM): Chosen for its ability to capture complex, non-linear relationships and handle interactions between features.

Random Forest: Selected for its robustness to outliers and ability to handle high-dimensional data.

The evaluation metrics used in model assessment are:

Mean Squared Error (MSE): MSE measures the average squared difference between the actual and predicted values, indicating the overall magnitude of prediction errors. Lower MSE values represent better model performance.

Root Mean Squared Error (RMSE): RMSE is the square root of MSE and provides a measure of the average magnitude of prediction errors in the same units as the target variable, making it easier to interpret. Lower RMSE values indicate better model performance.

R-squared (R^2): R-squared represents the proportion of variance in the target variable that is explained by the model. It ranges from 0 to 1, with higher values indicating better model performance. An R-squared value close to 1 means the model explains most of the variance in the target variable.

Initial Model Performance

OLS Model:

Fitted using R’s lm() function.

Showed evidence of heteroscedasticity as seen in its Residual Plot in Figure 2 below.

Achieved an R-squared of 0.9603, explaining 96.03% of the target variable's variance.

MSE: 3.97E-03, RMSE: 0.0630.

Heteroscedasticity indicated the need for more robust models.

Lasso Model:

Fitted using R's glmnet() function with cross-validation for optimal penalty parameter.

Achieved an R-squared of 0.9559.

Improved performance with lower MSE (9.04E-05) and RMSE (0.0095).

Selected 35 of the 56 predictor variables, enhancing interpretability.

GBM Model:

Fitted using R's xgboost package with cross-validated tree depth selection.

Achieved the highest R-squared (0.9699), explaining the most variance.

Lowest MSE (6.19E-05) and RMSE (0.0079) among the models.

Learning curve plot (Figure 8) shows RMSE for training and test datasets over boosting iterations, indicating good generalization without overfitting.

Random Forest Model:

Used to forecast the percentage of adults reporting fair to poor health in a county.

Evaluated with a dataset containing 56 predictor variables.

MSE: 8.10E-05, RMSE: 0.0090, R-squared: 0.9605.

High accuracy and strong suitability for predictive analysis.

Cross-Validation:

10-fold cross-validation was employed to assess the generalizability of the models. In this process, the dataset is randomly divided into 10 equal folds. Each fold is held out as a test set once, while the remaining 9 folds are used for training. This procedure is repeated 10 times, ensuring that each fold serves as the test set exactly once. Cross-validation provides a more robust estimate of model performance compared to a single train-test split and helps identify potential overfitting issues.

OLS:

Cross-validated OLS model showed improved MSE (7.91E-05) and RMSE (0.0089) compared to the initial model.

R-squared slightly decreased to 0.9599.

Improved MSE and RMSE suggest consistent performance across different data subsets.

Lasso:

Cross-validated Lasso model showed a slight increase in MSE (9.56E-05) and RMSE (0.0098) compared to the initial model.

R-squared decreased to 0.9525.

Indicates potential overfitting to training data, prompting exploration of other models.

GBM:

Cross-validated GBM model showed significant improvement in all metrics.

Achieved the lowest MSE (9.82E-06) and RMSE (0.0031), and the highest R-squared (0.9950) among all models.

Indicates high generalizability and strong performance on unseen data.

Final Model Selection

The GBM model was selected as the final model based on its superior performance in both initial evaluation and cross-validation. It consistently achieved the lowest MSE and RMSE values and the highest R-squared values, indicating its ability to make accurate predictions and generalize well to unseen data. Additionally, GBM's inherent ability to handle nonlinear relationships and its resistance to overfitting make it a suitable choice for this analysis. Table 1 compares the results of the evaluation of the models.

Table 3. Result of Evaluation Metrics

Model Type	***MSE***	***RMSE***	***R_squared***
OLS Initial	***3.97E-03***	***0.063014***	***0.9603253***
OLS Cross-Validated	***7.91E-05***	***0.008894***	***0.9599078***
Lasso Initial	***9.04E-05***	***0.00951***	***0.9559668***

Lasso Cross-Validated	***9.56E-05***	***0.009776***	***0.9524702***
GBM Initial	***6.19E-05***	***0.007865***	***0.9698829***
GBM Cross-Validated	***9.82E-06***	***0.003134***	***0.9950453***
RF Initial	***8.10E-05***	***0.009002***	***0.9605487***
RF Out-of-Bag	***8.39E-05***	***0.00916***	***0.9591531***

Overfitting Mitigation

We are confident that the GBM model will provide reliable predictions while minimizing the risk of overfitting, ensuring its generalizability and robustness. This is because we took several precautions to prevent overfitting. Firstly, we utilized early stopping rounds, which halt the training process if the evaluation metric fails to improve for a specified number of rounds namely ten(10), thereby preventing overfitting. Additionally, we conducted a learning curve analysis, where we monitored the training and testing RMSE over 100 boosting iterations. If the training RMSE continued to decrease while the testing RMSE started to increase, indicating overfitting, our early stopping mechanism would terminate the training to avoid overfitting. As stated earlier, we employed cross-validation on our model using repeated 10-fold cross-validation. This technique enabled us to assess our model's performance on different subsets of the data, providing a more robust evaluation and reducing the risk of overfitting.

A visual comparison of the performance models is shown in Figure 6 and Figure 7 below.

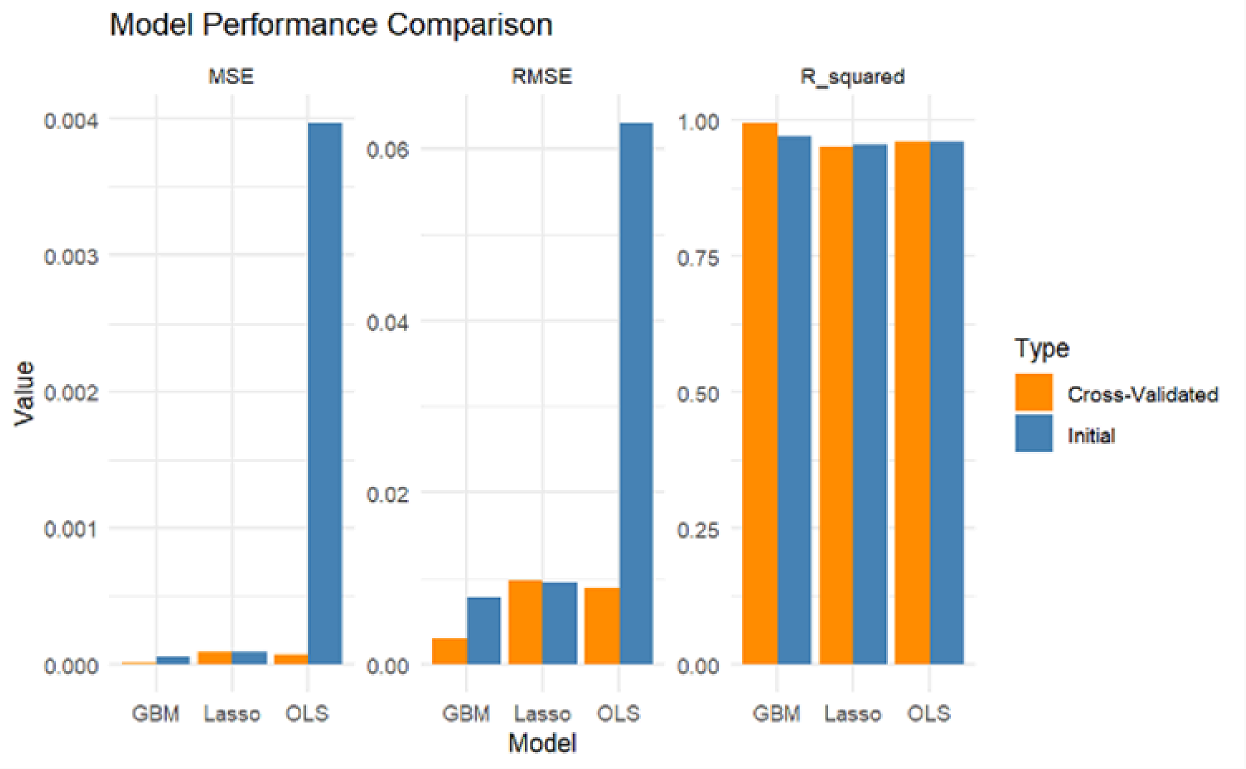


Figure 6. "Model Performance Comparison"*. This shows the Mean Squared Error (MSE), RootMean Squared Error (RMSE), and R-squared values for three models: GBM, Lasso, and OLS, with both initial and cross-validated results. The graph shows that the GBM model consistently outperforms the Lasso and OLS models across all metrics, with lower MSE and RMSE values and higher R-squared values, indicating better model performance and generalizability.*

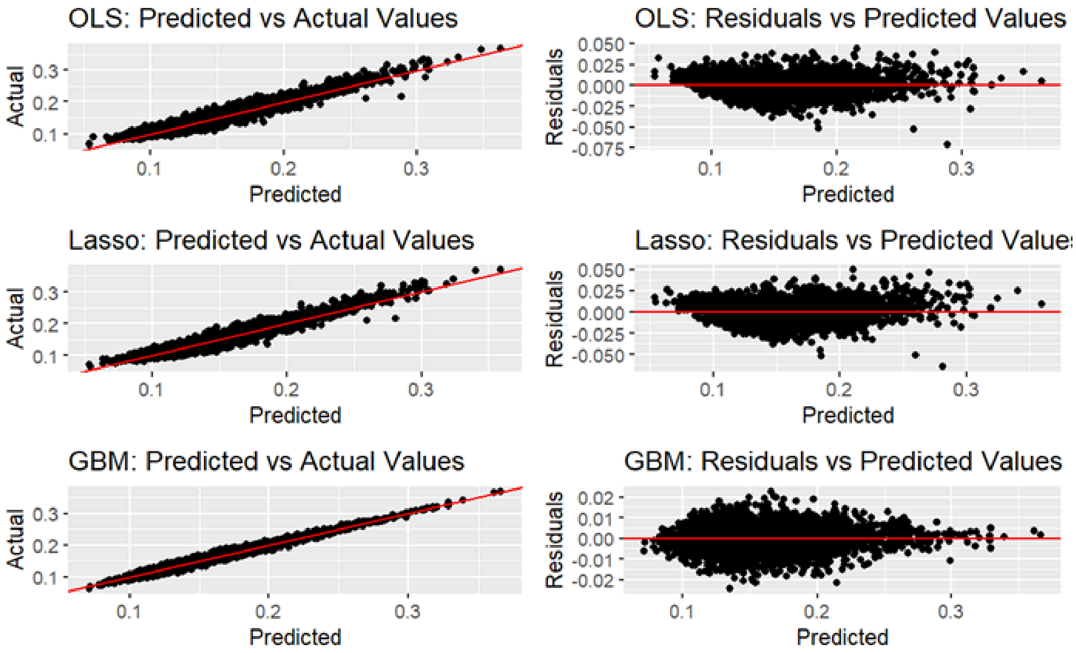


Figure 7. "Comparison of Predicted vs. Actual Values and Residuals for OLS, Lasso, and GBM Models". *The OLS model's predicted vs. actual values show some deviation from the reference line, and the residuals vs. predicted values plot reveals heteroscedasticity, with increasing variance of residuals. (The slight "u" shape of the residual vs Predicted values plot is evidence of heteroscedaticity)*

The Lasso model's predicted vs. actual values are closer to the reference line than OLS, but some deviation remains. The residuals vs. predicted values plot also indicates heteroscedasticity.

The GBM model's predicted vs. actual values align closely with the reference line, indicating good predictive performance. The residuals vs. predicted values plot shows more randomly dispersed residuals, suggesting better homoscedasticity.

Based on visual analysis, the GBM model is the best performer. It demonstrates better homoscedasticity, with evenly spread residuals, and higher predictive accuracy, with points closely aligned with the reference line. These observations confirm that the GBM model generalizes better and provides more consistent predictions across different levels of the target variable.

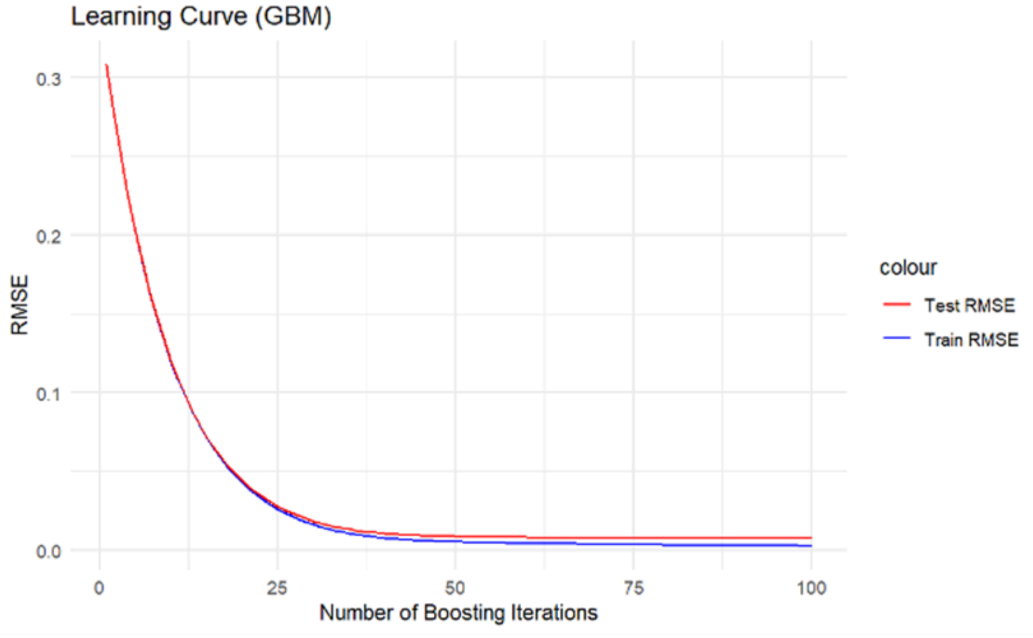


Figure 8. "Learning Curve (GBM)" *graph shows the Root Mean Square Error (RMSE) values for the Gradient Boosting Machine (GBM) model over 100 boosting iterations. The red line represents the training RMSE, which decreases sharply at the beginning and then flattens out,*

indicating that the model is learning and fitting the training data well. The blue line represents the test RMSE, which follows a similar pattern but remains consistently higher than the training RMSE, confirming that the model generalizes well

Conclusions

In conclusion, the GBM model predicted the county's response to the self-reported poor to fair health question. With this model, we were also able to determine what features were most important to this prediction, how the features influenced the prediction, and propose solutions to the client to improve the response, a decrease in the percent self-reporting poor to fair health. As discussed previously, the R-squared value for this model after cross-validation is 0.995, an extremely accurate model.

Category:

When the Domains of Social Determinants of Health are analyzed, the most important category is Education Access and Quality. This is followed by Economic Stability, Health Care Access, and Quality, Neighborhood and Build Environment, and lastly Social Community Context. An "Other" category was included to capture other features that described the observation region, race makeup, and sex makeup.

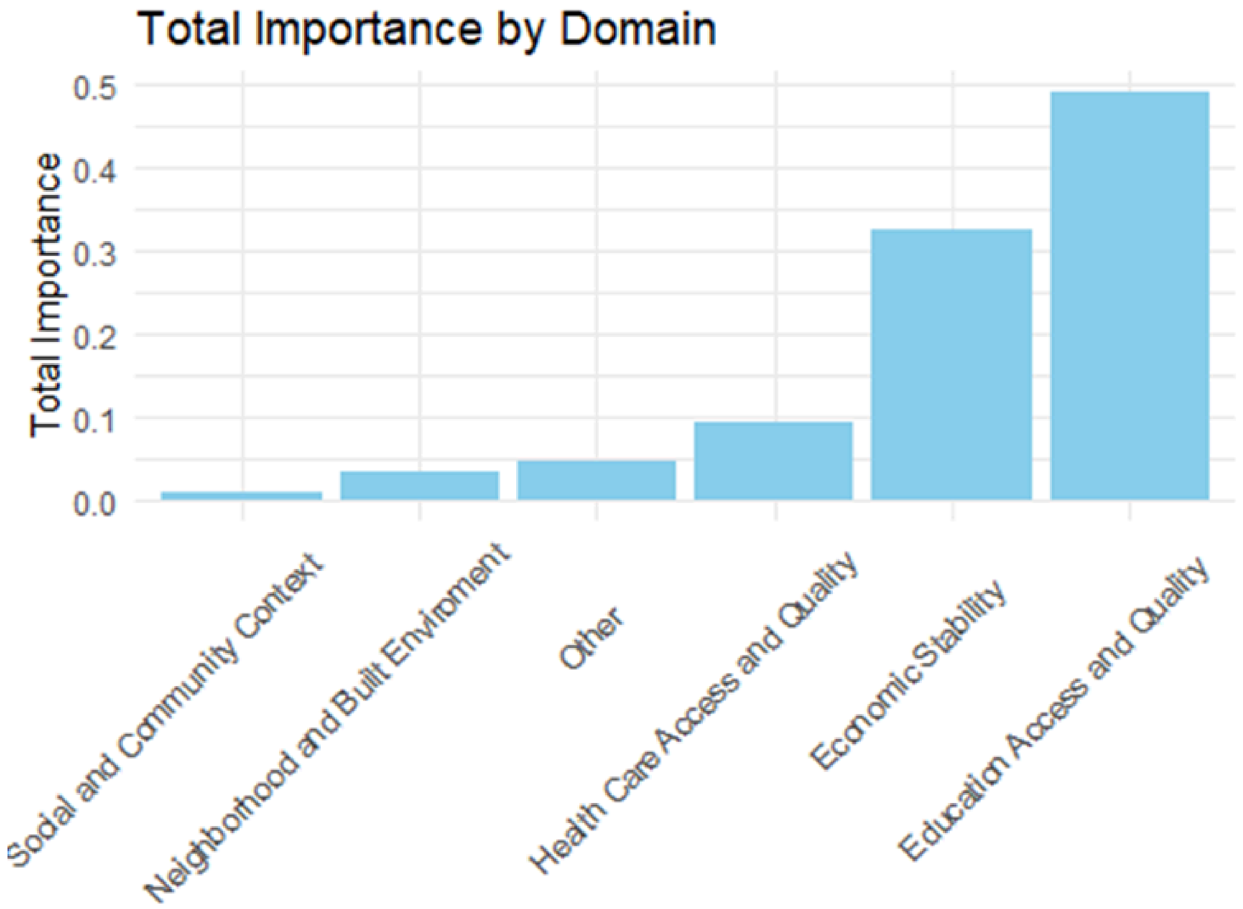


Figure 9. "Relative importance of SDOH Domain". *The figure shows the relative importance of each domain as defined by the CDC plus the "Other" damain, a catchall for remaining features remaining in the model.*

Feature Importance:

The most important feature is the percent of respondents reporting obtaining a high school degree or equivalent by the age of 25. This feature accounts for 49% of the models' predictive power. The next most important feature is percent reporting food insecurities explaining 30% of the models' predictive power.

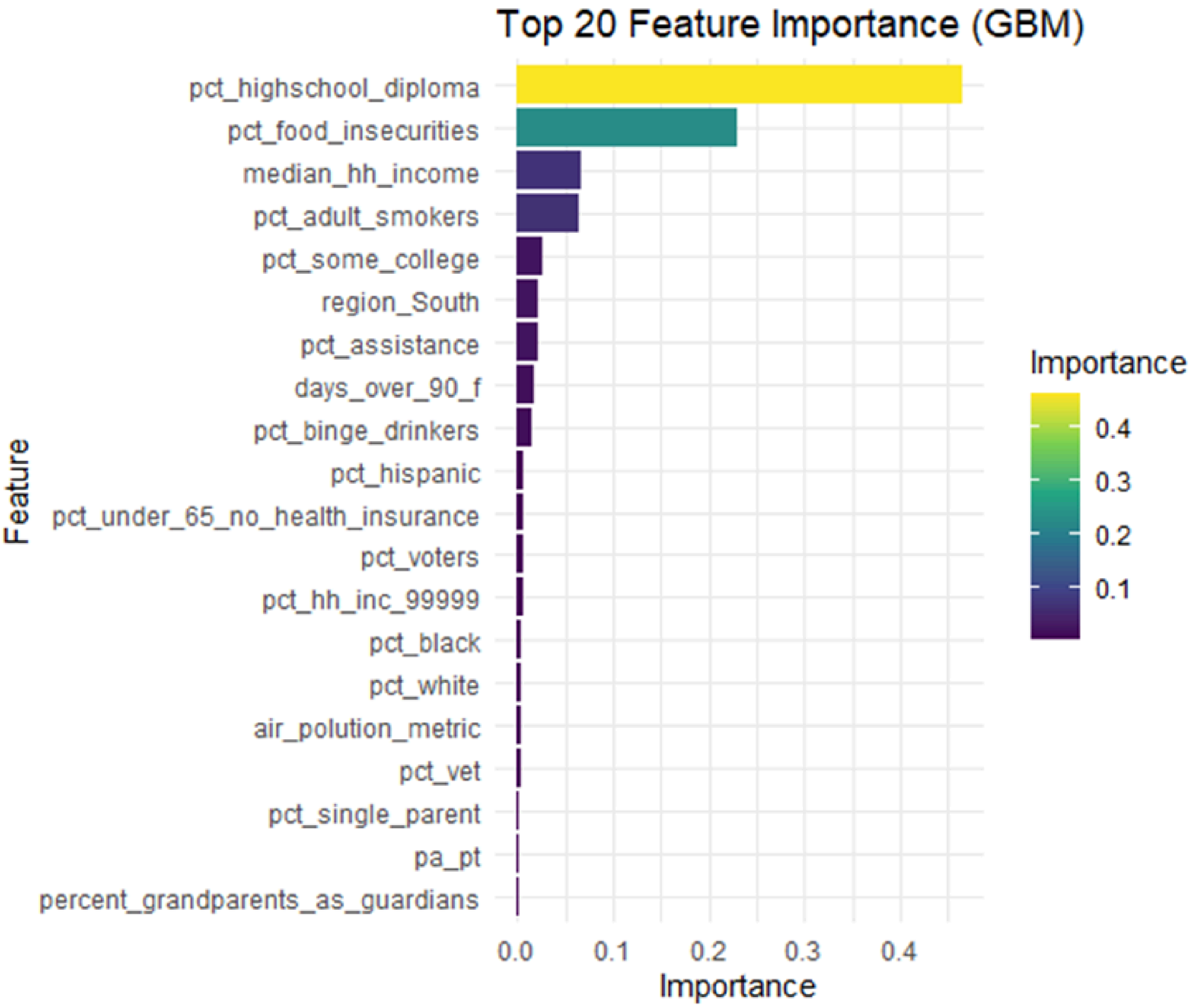


Figure 10. "Top 20 Features by Importance in GBM Model"*. The figure shows a chart of most important features in the GGM model from most important to least important.*

Feature Influence:

To understand how each feature effects the model, a partial dependence analysis was performed on dome of the feature excluding those in the “Other” category. The features mostly back up what was expected for the predictive analysis. The larger the rate of graduation in a county, the better the quality of life, the greater the food insecurities, the worse the response is to quality of life. The more people smoke in a county, the worse the quality of life. These are displayed bellow. If the line goes from bottom left to top right (roughly) the more likely that predictor is to affect people negatively, and if the line goes from top left to bottom right, the predictor is likely to improve the quality of life.

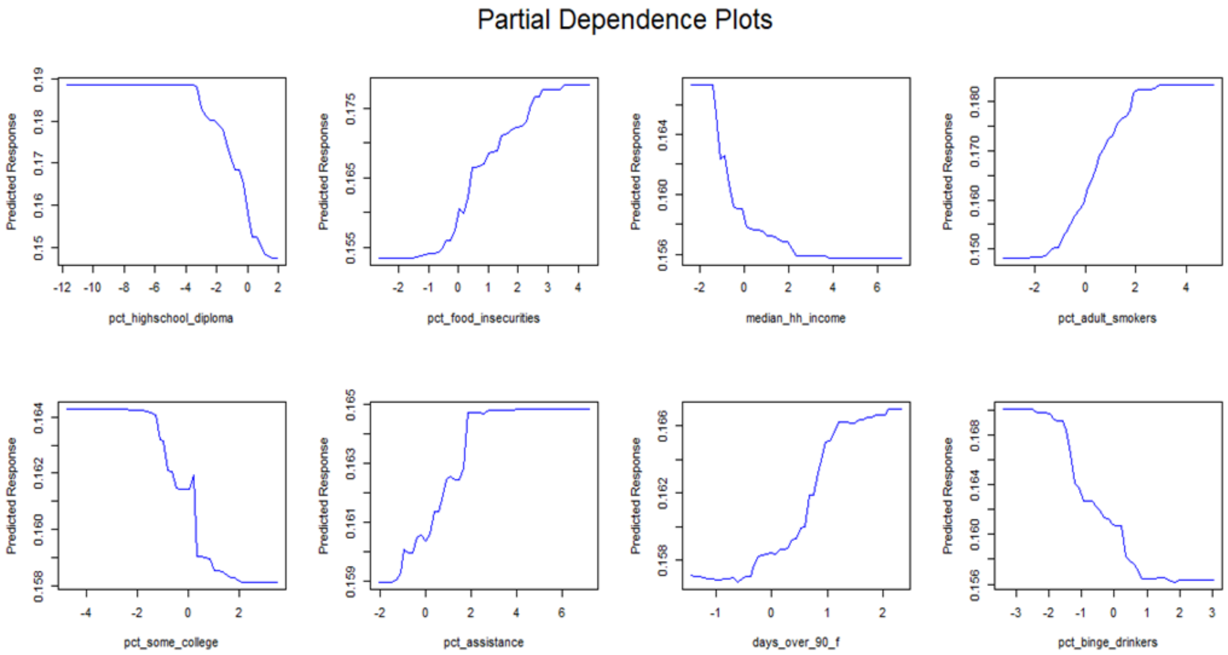


Figure 11. “Partial Dependence Plot of Top 8 Domain Related Featured in GBM Model”*. The figure shows a plot of multiple features importance vs response, from top left to bottom pct_highschool_diploma, pct_food_insecurities, median_hh_income, pct_adult_smokers, pct_some_college, pct_assistance, days_over_90, pct_binge_drinkers.*

A Unique response in binge drinking was observed that was not predicted, where the more people binge drink in a county, the better the quality of life. This may be due to a reporting bias where people who binge drink are not reporting this, a healthy user bias where binge drinkers who are healthy and binge drink are more likely to respond to the survey that binge drinkers who are unhealthy, confounding variables such as greater income where an individual with a higher income can afford more alcohol or a weighting error where the distribution of population density is skewing the response by oversampling small counties who are less likely to drink. Another possibility is that this bias is inherent in the relationship of the predictor and response so might not be able to be removed.

Answers to the Question:

With this information, we can propose ways each local, state or even federal government can improve the quality of life for the residents of each area. The largest predictor is the percentage of residents obtaining a high school diploma which would be an ideal place to start. Possible solutions could be to ensure adequate pay for those teaching at public schools to attract better potential teachers, increase public awareness of the county's education needs, and make efforts to support the school system with donation drives ensuring children have access to the tools needed for learning. Or promoting after-school activities, to give children in need further support and help with their education. Focusing on this alone would greatly improve residents' quality of life.

The next most important area for focus would be on ensuring that the residents have adequate access to food. There are multiple ways to do this, like promoting food drives, supporting the department that provides assistance to people in need of food by ensuring they have the resources needed as well as promoting this service to the citizens.

Another key area of focus is the economic factor, where providing access to higher jobs will increace individual's' quality of life. Local governments can do this by working to attract new businesses by promoting either the local workforce's ability to perform those jobs or other business-related incentives.

Discussion

There are multiple next steps to take for this project to better understand the relationship as well as identify and resolve internal bias and overfitting of the model. There were also limitations to the work done in this class on fully exploring the data.

Bias:

The distribution of densely populated counties to sparsely populated counties was extreme and may lead to skewness in the response data. The response itself has right skewness with a skew value of 0.73 and outliers to the right. An attempt was made to normalize this with the weighted population but this only increased the skewness of the response to 15.47

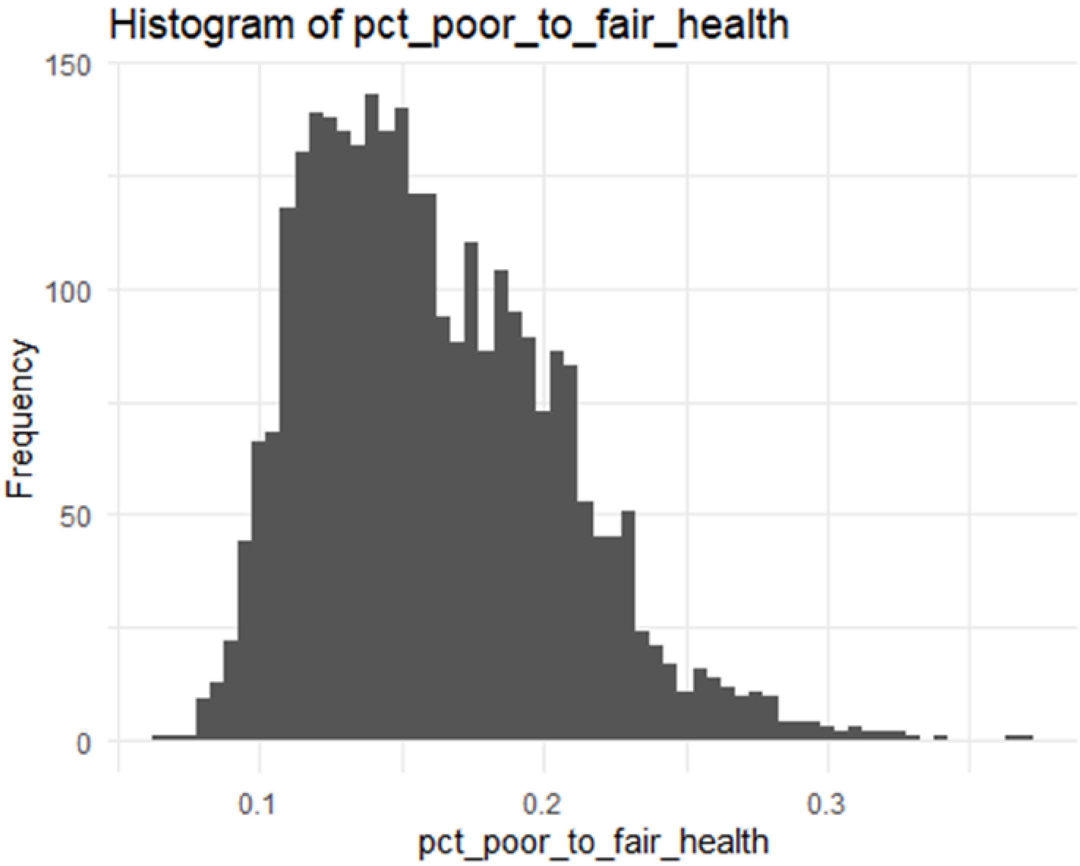


Figure 12. "Response Skewness"*. The figure shows a histogram of the response feature pct_poor_to_fair_health, which shows the data to be slightly right leaning. The feature has a skewness of 0.73.*

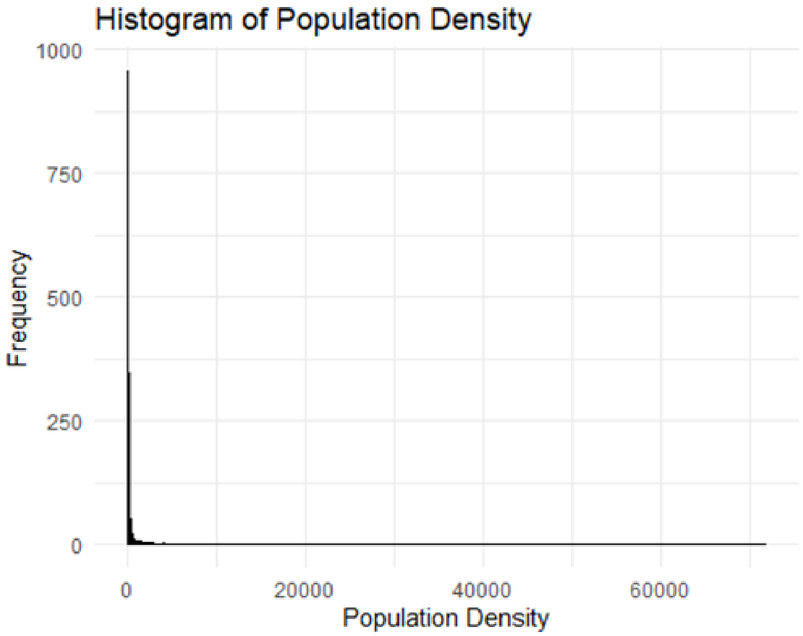


Figure 13. "Response Skewness". *The figure shows a histogram of the feature population_densiy, which shows the data is extremely right leaning. The feature has a skewness of 39.5.*

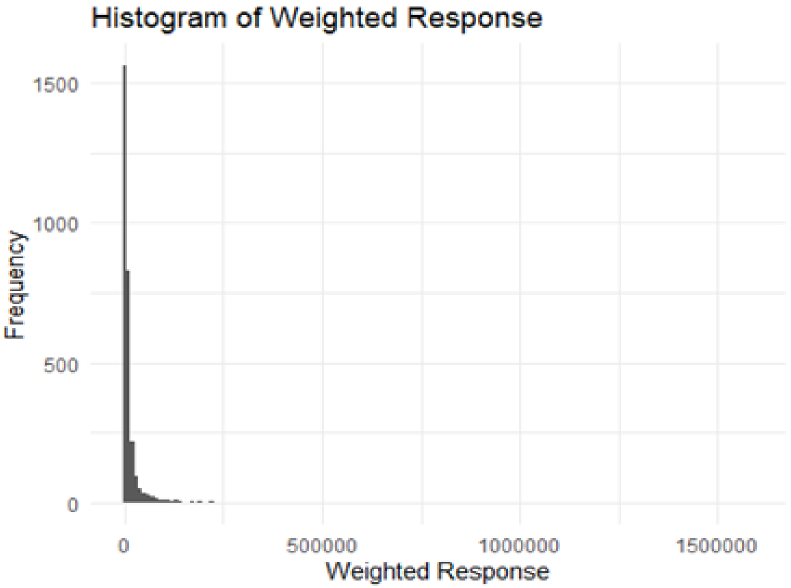


Figure 14. "Response Weighted with Population Density Skewness"*. The figure shows a histogram of the response feature weighted by multiplying pct_poor_to_fair_health by population_densiy, this also has extreme right skewness.*

Survey Bias:

A paper by Dan Lio and Robert Valliant proposes that survey data is typically biased in that the values within the model are highly correlated. They state that there are currently no statistical methods to remove this correlation in the data so the best solution is to remove features from the data. The PCA analysis was performed on the dataset and there are multiple outcomes to overcome this issue by feature removal for dimension reduction.

Reporting Bias:

As noted in the relationship of percent binge drinkers in a county and that county's reply to the poor to fair health question, there was an unexpected outcome. This type of bias is inherent in the data and not able to be removed.

Confounding Factors:

There is inherent collinearity in survey data, such as people with greater access to education are more likely to have higher incomes. Individuals with higher incomes are better able to afford food, and higher quality food. This is true for numerous features within the model.

Next Steps

The next step needed to improve the model would be to limit the number of features used to build the model. A PCA analysis was performed to identify what number of predictors would be best suited for this dataset. A subset was identified and a model with 29 features would be able to explain 90% of the response variable and 35 features would be able to explain 95% of the response variable. However a parallel analysis was performed and it identified that using a subset dimensionality of 9 features would be adequate for explaining 95% of the response with any additional features being equivalent to adding random values. Each model dimensionality should be explored.

Another area of improvement would be to identify ways to reduce the predictor's skewness. There are datasets available that with some manipulation could return this desired trait.

Appendix

beta_codebook_v2.pdf

Data Dictionary.xlsx

Code and Documents

GitHub:

https://github.com/suzupis007/capstone_project.git

Capstone Google Drive:

Capstone Team Beta

Capstone Final Written Deliverables:

Final Written Deliverables_Beta

References

Bombak A. E. (2013). Self-rated health and public health: a critical perspective. *Frontiers in public health*, *1*, 15. https://doi.org/10.3389/fpubh.2013.00015

Data Exploration - A complete introduction | HEAVY.AI. (n.d.). Retrieved from https://www.heavy.ai/learn/data-exploration

Liao, D, and Valliant, R. (2012), "Condition Indexes and Variance Decompositions for Diagnosing Collinearity in Linear Model Analysis of Survey Data," Survey Methodology, 38, 189-202.

National data & documentation: 2010-2022. (n.d.). Retrieved from https://www.countyhealthrankings.org/health-data/methodology-and-sources/data-documentation/national-data-documentation-2010-2022

Rapoport, E., Muthiah, N., Keim, S. A., & Adesman, A. (2020). Family Well-being in Grandparent- Versus Parent-Headed Households. *Pediatrics*, *146*(3), e20200115. https://doi.org/10.1542/peds.2020-0115

Social Determinants of Health Database (Beta version). (n.d.). Retrieved from https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html

Your Go-To Guide for English Grammar. (n.d.). Retrieved from https://www.grammarly.com/handbook/