HEALTH CAPPYBARA TEAM

Health Accessibility

Within Cook County

Group Members

(Alphabetically Ordered Last Names)

Ruihou Guo

Yijia He

Yue Jian

Qi Zhao

Contents

Overview

In the United States, a significant portion of the population faces challenges in accessing essential health care services. Defined by Healthy People 2030, Healthcare Accessibility refers to the ability of individuals to procure timely, high-quality, and economically viable health care services. This project is aligned with the objective of improving healthcare accessibility and fostering healthier communities, particularly focusing on the diverse communities within Cook County.

To achieve this, we have employed a multifaceted research approach, leveraging data from various sources, including public databases and innovative web scraping techniques. Our study encompasses an in-depth analysis of healthcare facilities, healthcare professionals, insurance costs, and population census data. Utilizing advanced methodologies such as machine learning, data visualization, and Geographic Information System (GIS) mapping, we aim to offer a comprehensive assessment of healthcare accessibility in Cook County.

This project not only seeks to identify the gaps in healthcare accessibility but also aims to provide actionable insights that can guide policy-making and community initiatives to address these challenges effectively. Through this systematic exploration, we endeavor to contribute to the broader goal of enhancing healthcare access and ensuring that all individuals, regardless of their socio-economic status or geographical location, can lead healthier lives.

Project Structure

Analysis: Contains py files used to clean the csv data sources, calculating the metrics, and build the prediction models.

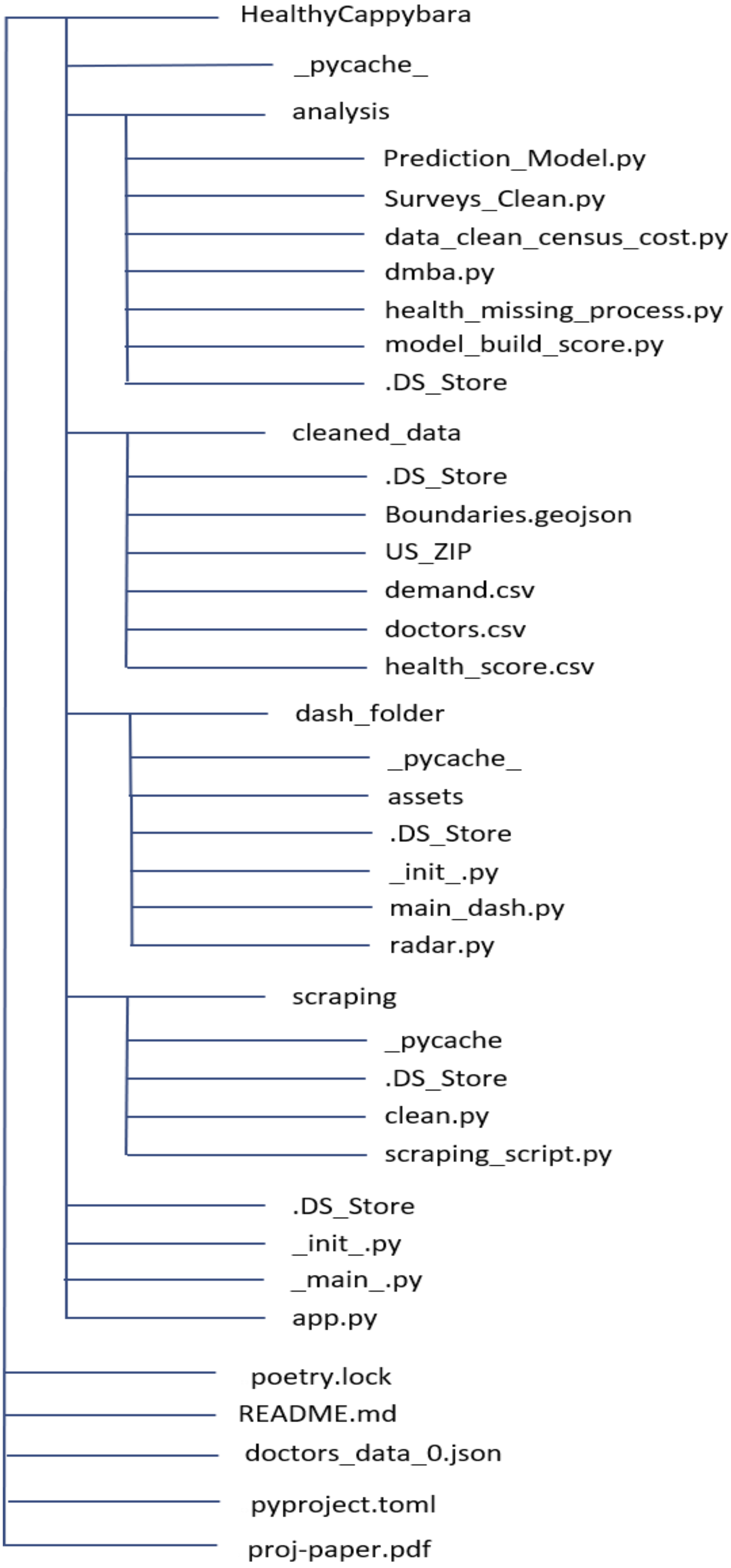
Cleaned_data: Cleaned data outputs of health facilities ratings, healthcare costs, socioeconomics data, and geographical data.

Dash_folder: Contains files used for dashboards and data visualization of healthcare accessibility.

Scraping: Contains the functions for web scarping on healthgrades.com and the clean function for web scraping data.

App.py: Contains functions to run the dash interface of our application, scarping the data from healthgrades, and clean the data available from web scraping.

README.md: Overview of our project and the instructions to install the portry virtual environment and run the application.



```
HealthyCappybara
    _pycache_
    analysis
        Prediction_Model.py
        Surveys_Clean.py
        data_clean_census_cost.py
        dmba.py
        health_missing_process.py
        model_build_score.py
        .DS_Store
    cleaned_data
        .DS_Store
        Boundaries.geojson
        US_ZIP
        demand.csv
        doctors.csv
        health_score.csv
    dash_folder
        _pycache_
        assets
        .DS_Store
        _init_.py
        main_dash.py
        radar.py
    scraping
        _pycache
        .DS_Store
        clean.py
        scraping_script.py
    .DS_Store
    _init_.py
    _main_.py
    app.py
poetry.lock
README.md
doctors_data_0.json
pyproject.toml
proj-paper.pdf
```

Code Roadmap

Code Responsibility

Data Collection & Data Cleaning

- Scraping doctors' information: Yijia & Yue
- Demographic Data & Health Expense Data: Qi
- Health Rating: Hourui
- Geographic Information: Qi

Data Processing

- Building Index for Population Demand: Qi
- Building Index for Health Demand: Qi

Data Visualization

- Geographic maps: Yijia
- Clustering & Correlation matrix graph: Yue & Hourui
- Radar Chart for health index: Qi

Data Analysis

- Supervised Machine Learning Models: Hourui
- Clustering Model for demand and health services: Yue

Interactive UI

- Dashboard: Yijia & Yue

Other

- Paper & Readme: Qi & Hourui

Code Roadmap

Let’s interact with our app

1, Clone the repostitory

```
git clone git@github.com:gagahe-cx/Healthy-Cappybara.git
```

2, Navigate to the repository

```
cd ./Healthy-Cappybara
```

3, set up and activate the virtual environment

```
poetry install
```

```
poetry Shell
```

4, Launch the App

```
Python3 -m Healthycappybara
```

5, Engage with the App (Using Alphabetical Inputs)

(a) The Dashboard

(b) Scraping Data

(c) Clean Data

(d) Quit App

6. Option 2 has three sub-options. Users have the capability to input their specific criteria for conducting web scraping.

Condition 1: How many medical category do you want to crawl?
Condition 2: How many cities do you want to crawl?
Condition 3: Do you want to crawl now?

Upon completion, you will get the message:

" Congratulations! The data has been successfully crawled and saved to {file location}!"

Our Goals

In the United States, many people face challenges in obtaining the health care services they need. Healthy People 2030 characterizes Healthcare Accessibility as the ability to acquire health care services that are prompt, of high quality, and within financial reach. In line with the goal to improve access to healthcare and foster healthier communities, our initiative is focused on creating a systematic method to enhance healthcare accessibility in Cook County. Our approach adopts a Visualization-Analysis-Prediction framework to find potential solutions.

Initially, we plan to consolidate various indicators and develop two models to assess the quality of health services and the demand of the population. The next step involves visualizing these assessments to identify and cluster the discrepancies between the needs of the population and the services provided.

Finally, we will delve into the specifics to pinpoint the most crucial indicators for improving health services, thereby gaining valuable insights for policymakers and individuals alike.

Accomplishments

Step 1: Building Scores to Evaluate Health Accessibility

We've developed two models to assess health accessibility in Cook County: Health Service Score and Population Demand Score, both grounded in the Analytical Hierarchy Process (AHP) and Entropy Weighting Method (EWM).

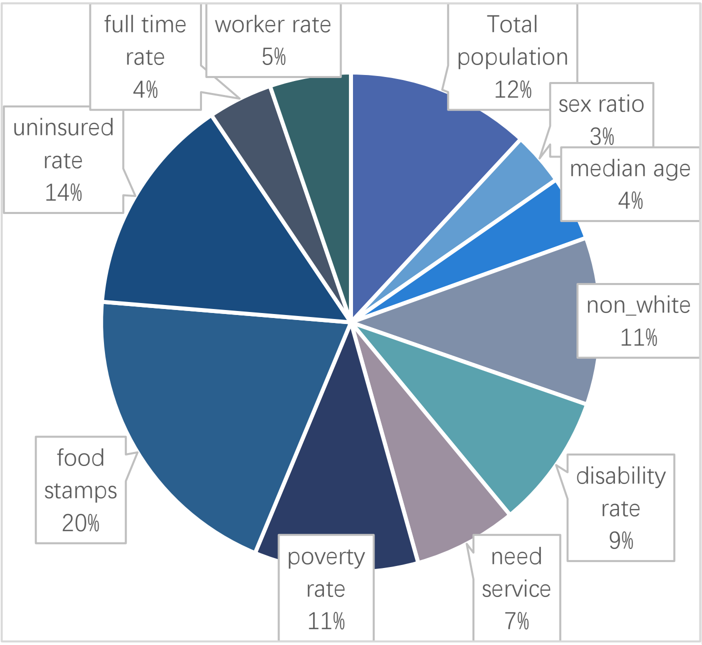
The Population Demand Score incorporates:
demographic, vulnerability, poverty, and development factors,
while the Health Service Score considers:
health service quantity, quality, and expenses.

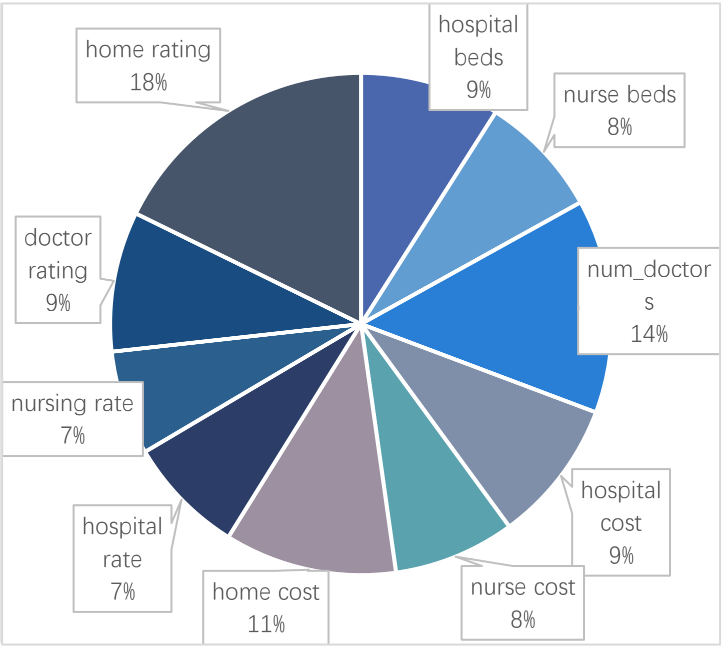
Post-data weighting, these models utilize 22 indicators to derive their respective weightings.

Data Weighting Methods:

While the Analytic Hierarchy Process (AHP) aligns well with our health criteria due to its tailored weighting, it lacks objectivity. Conversely, the Entropy Weighting Method (EWM) is objective but doesn't account for our subjective health perspectives. To balance these strengths and weaknesses, we adopt a combined weighting approach, integrating AHP and EWM, to ensure a more holistic and accurate reflection of health priorities. The subjective and objective combination weights are as follows:

$$W_j = \frac{\sqrt{\alpha_j \beta_j}}{\sum_j \sqrt{\alpha_j \beta_j}}$$





Accomplishments

Dealing with missing data

To address missing data in our health service evaluation, which includes hospitals, nurses, and home services, it is essential to recognize that these services extend beyond their immediate communities. Following the National Institutes of Health (NIH) definition of Healthcare Service Areas, we consider these to be zones where hospital care is predominantly self-contained but may also extend across state boundaries. Consequently, we define a health service area to encompass both the community itself and its adjacent neighborhoods. This allows us to calculate a comprehensive data set for each indicator by summing the base data of the community with the corresponding data from neighboring areas (We use geographic polygon information for calculation here).

HSA_C represents the health service data for community c, I_C represents the base data for community c for a specific indicator, and N_c represents the set of neighboring communities to c, the equation would be:

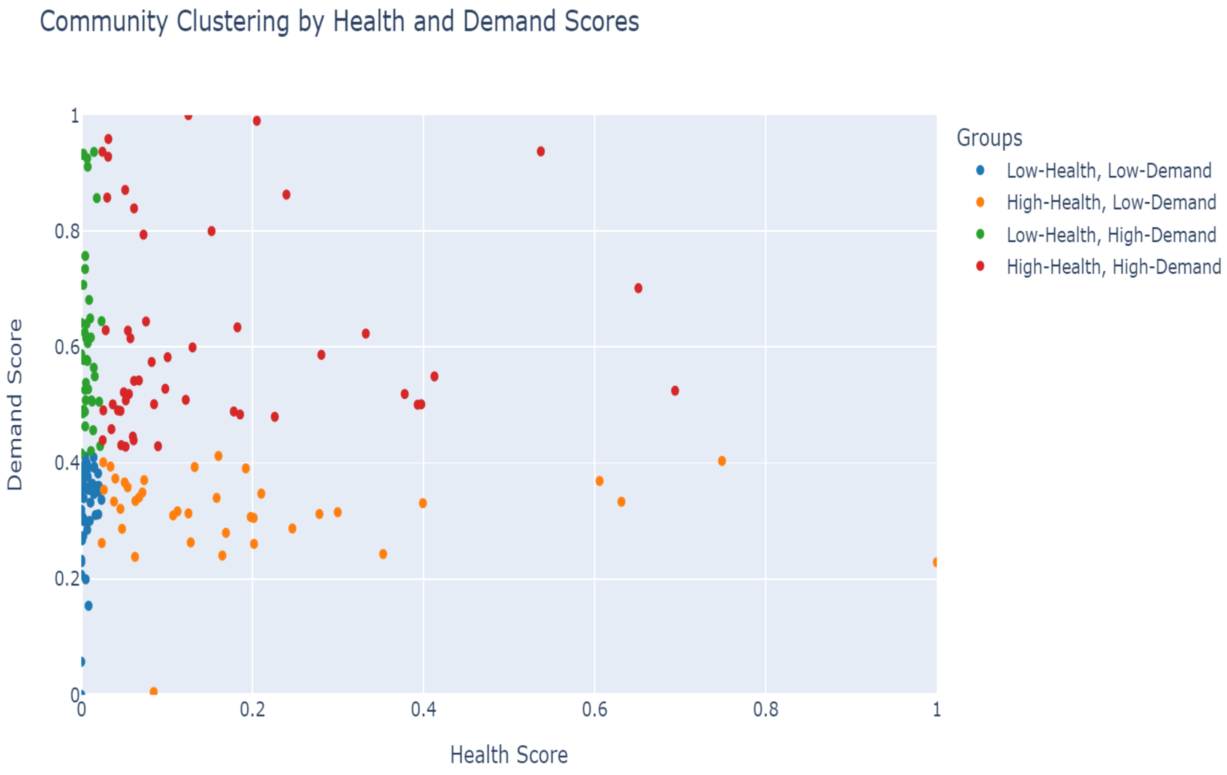
$$\boldsymbol{HSA_c = I_c + \sum\nolimits_{n \in N_c} I_c}$$

Accomplishments

Step 2: Data Analysis for Clustering

Utilizing the two scores, we can discern the correlation between population demand and health service provision. We categorize communities into four clusters: high demand-high service, high demand-low service, low demand-low service, and low demand-high service. This allows us to identify communities with resource misalignment. Through this clustering, we prioritize resource allocation to communities with high demand yet insufficient health services to enhance healthcare delivery.

Low Demand- High Service↵	High Demand- High Service↵
Low Demand- Low Service↵	High Demand- Low Service↵



Accomplishments

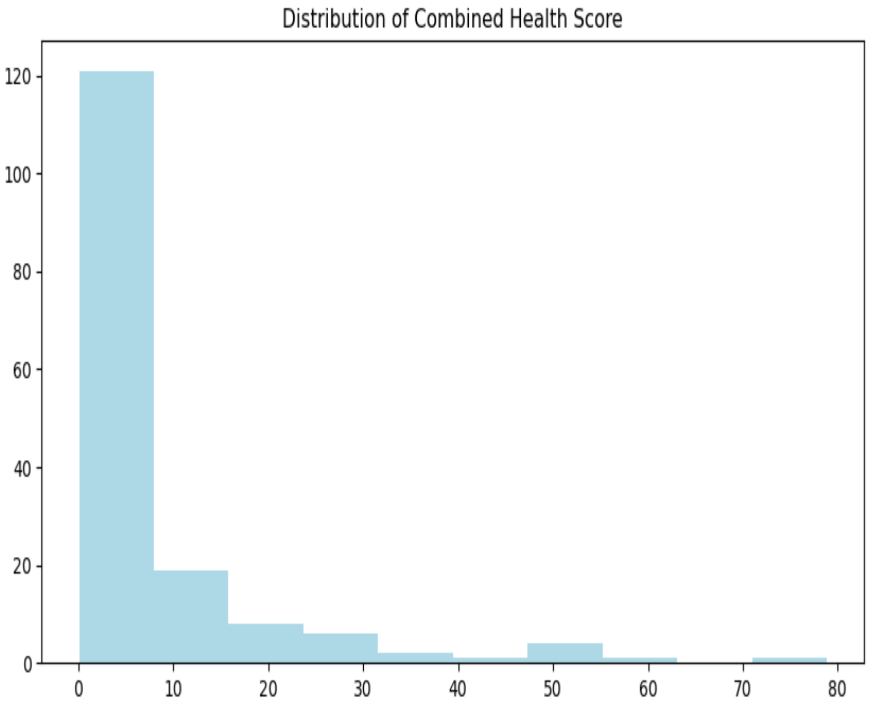
Step 3: Machine Leaning and Prediction

Part 1: To gain a deeper understanding of the interplay between combined health scores and census data across different communities, we embarked on an analytical journey employing data mining and machine learning techniques.

Initially, our exploratory data analysis (EDA) revealed foundational insights into our dataset, such as the maximum, minimum, and the distribution of combined health scores. Notably, the health score data exhibited a rightward skew, suggesting an imbalance that warranted further attention. To better understand the relationships between predictors and our target variable, we constructed a correlation matrix map, which is instrumental in highlighting the strength and direction of the associations within our data.

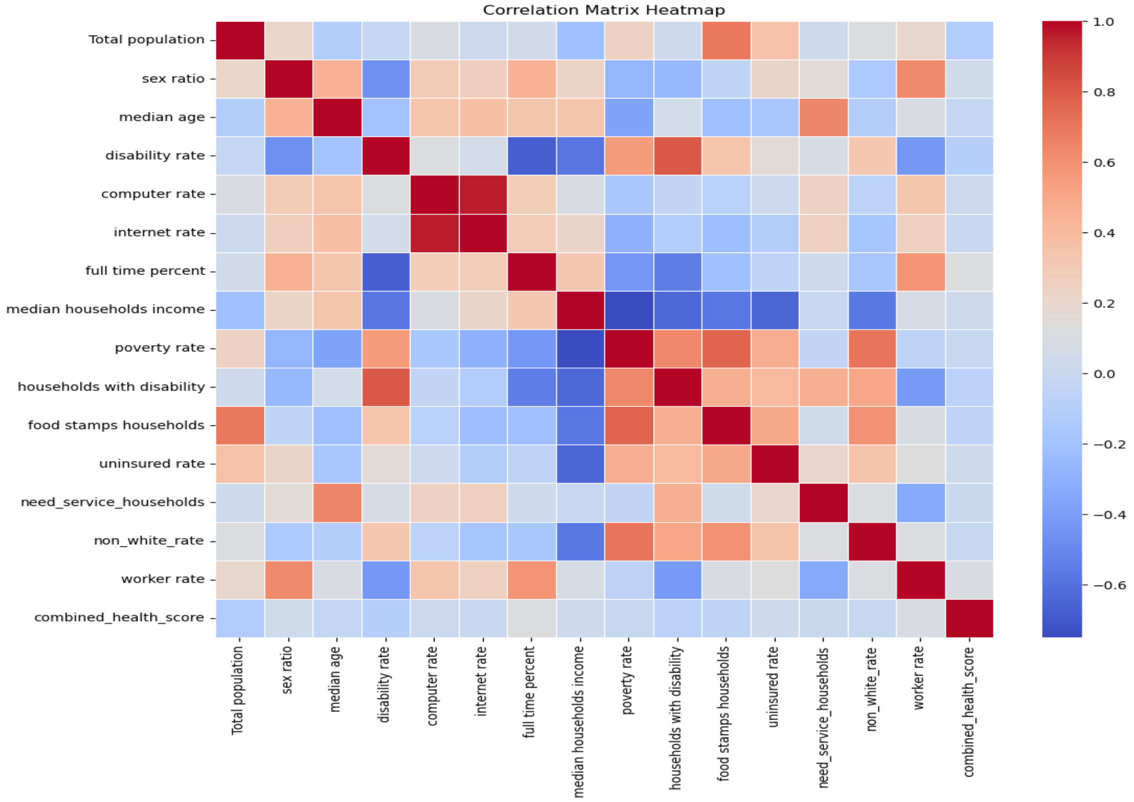
Given that the combined demand score and combined health score had been previously normalized, our preprocessing efforts primarily focused on addressing missing data, thus ensuring a robust dataset for model training.

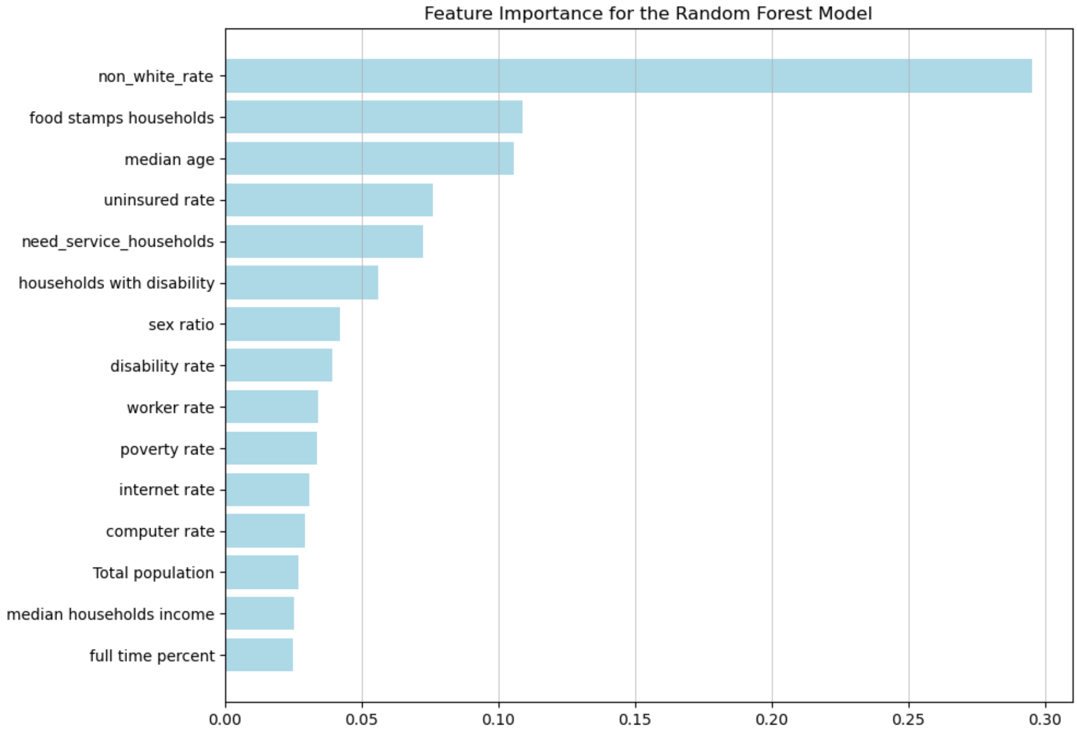
Part 2: Our modeling began with linear regression. However, the regression statistics and coefficients indicated that the model was not suited to our dataset, which was both small in size and exhibited signs of multicollinearity, as evidenced by the correlation matrix. We turned to regularized linear regression techniques, including Lasso and Ridge regressions. Despite their ability to mitigate overfitting and multicollinearity, these models underperformed, potentially due to the underlying non-linear relationships in our data.



Pivoting to more sophisticated models, we explored decision trees and the ensemble method of random forests. These models demonstrated a marked improvement over their linear counterparts after hypertuning. A particularly valuable outcome was the derivation of feature importance scores. These scores showed influential predictors such as the non-white rate, the prevalence of food stamp households, and median age, which interestingly align with findings from policy research.

In the subsequent sections of our report, we will delve into a comprehensive analysis of these findings, unpacking their implications and how they might inform future policy decisions.





Policy Implication

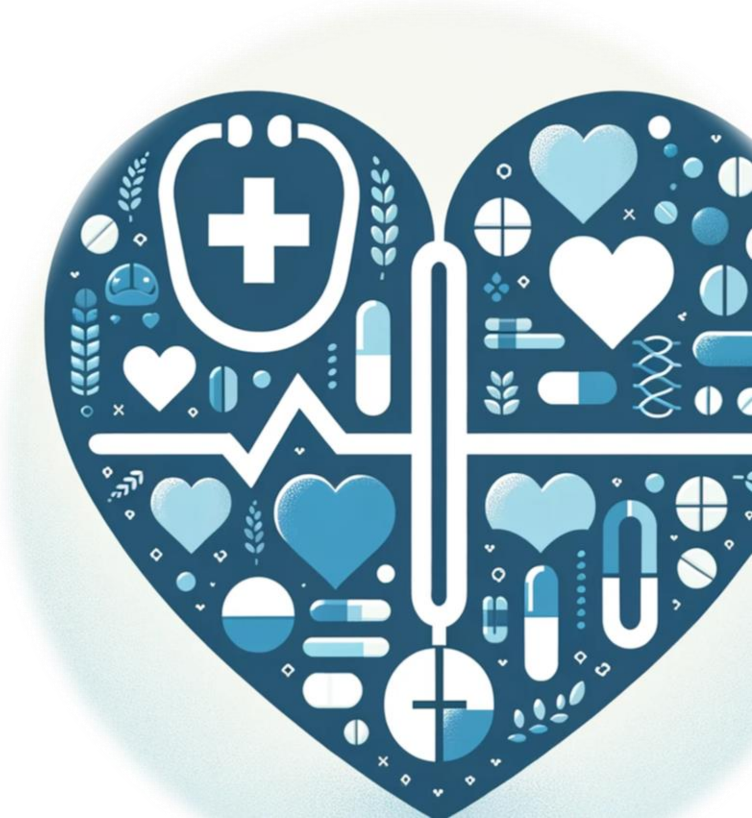
The insights garnered from our analysis—highlighting the strong relationship between non-white rates, median age, and food stamp households with healthcare accessibility—raise important considerations for policymakers.

The correlation with non-white rates suggests that healthcare accessibility is entwined with racial demographics, indicating a potential disparity that may require targeted interventions to ensure equitable healthcare distribution.

The significance of median age points to the necessity of age-sensitive healthcare provisions, ensuring that both younger and older populations have appropriate access to healthcare services. The association with food stamp households signals that economic factors and healthcare access are closely linked, suggesting that addressing socioeconomic barriers could substantially improve healthcare reach.

These findings advocate for policies that not only bridge the gap in healthcare accessibility across racial and age demographics but also address the underlying socioeconomic disparities.

Tailored community-based healthcare programs, increased funding for areas with higher food stamp usage, and policies that explicitly combat racial healthcare disparities could be effective measures to enhance healthcare equity and accessibility.



Future Improvement

In the process of data mining with our small dataset, we faced several challenges that have constrained the effectiveness of our models. The limited number of data points restricts the depth of learning our models can achieve, potentially missing out on capturing the full complexity of the underlying patterns.

Despite hyperparameter tuning efforts, we encountered the dual challenges of overfitting, where the model learns the noise instead of the signal, and underfitting, where the model is too simplistic to grasp the data's intricacies. The skewness towards certain population characteristics suggests an imbalanced dataset, which might have skewed the model's predictions. Multicollinearity among predictors further complicates the interpretation of the model's outputs.

To enhance the robustness of our findings, expanding the dataset either by collecting more data or integrating additional relevant features could be crucial. Sophisticated feature engineering, perhaps informed by domain expertise, could unveil latent variables and intricate non-linear relationships that our current models overlook. Exploring alternative algorithms, particularly those tailored for small or imbalanced datasets, alongside employing model interpretability tools that extend beyond SHAP and PDP, might provide greater clarity and yield insights that are more reliable. These enhancements are not merely technical adjustments but steps towards a more nuanced analysis that could inform more precise and effective policy interventions.

HEALTH CAPPYBARA TEAM

Thank You!

We hope you Live Healthier, Live Happier