Assessing the Impact of Social Determinants of Health (SDOH) on Life Expectancy by Census Tract in Chicago

Michael Goodman, Alejandra Manrique, Zihan Chen

www.github.com/mgoodman96/FinalProjectLNM

Agenda

1. Introduction & Background
2. Data Collection
3. Model Preparation
4. Methodology & Interpretations
5. Improvements & Future Use

Introduction & Background

Introduction: Life Expectancy in Chicago

- Life expectancy at birth, as defined by the World Health Organization, reflects the overall mortality level of a population. It summarizes the mortality pattern that prevails across all age groups - children and adolescents, adults and the elderly.[2]
- Life Expectancy is a key indicator of health and quality of life and can be used as a measure to gauge the overall health of a community.
- Our aim was to identify predictors of life expectancy across Chicago.

Background

- The Robert Wood Johnson Foundation (RWJF) and the University of Wisconsin Population Health Institute (UWPHI) collaborate in a program called the *County Health Rankings & Roadmaps* to help communities identify and implement solutions to make it easier for people to be healthy in their schools, workplaces, and neighborhoods.
- The Rankings are compiled using county-level measures from a variety national data sources
- Figure 2 shows the model and features they include to produce the rankings for each county.
- This model served as a starting point for us to determine what features we would include in our model.

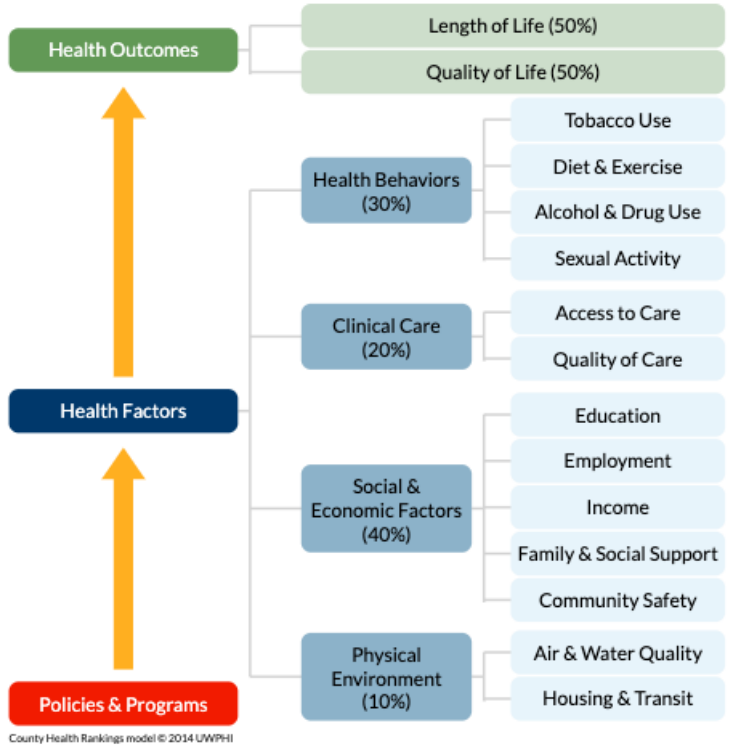


Figure 1. County Health Rankings Model by the University of Wisconsin Population Health Institute

Background

- In our research, we came across the **Social Determinants of Health (SDOH)** set by the Healthy People 2030 (HP2030). The HP2030 is an initiative set every decade by the US Department of Health and Human Services. The SDOH set by the HP2030 are the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks.[1]
- **Our team decided to explore how different features of each SDOH impact life expectancy in Chicago in each census tract.**

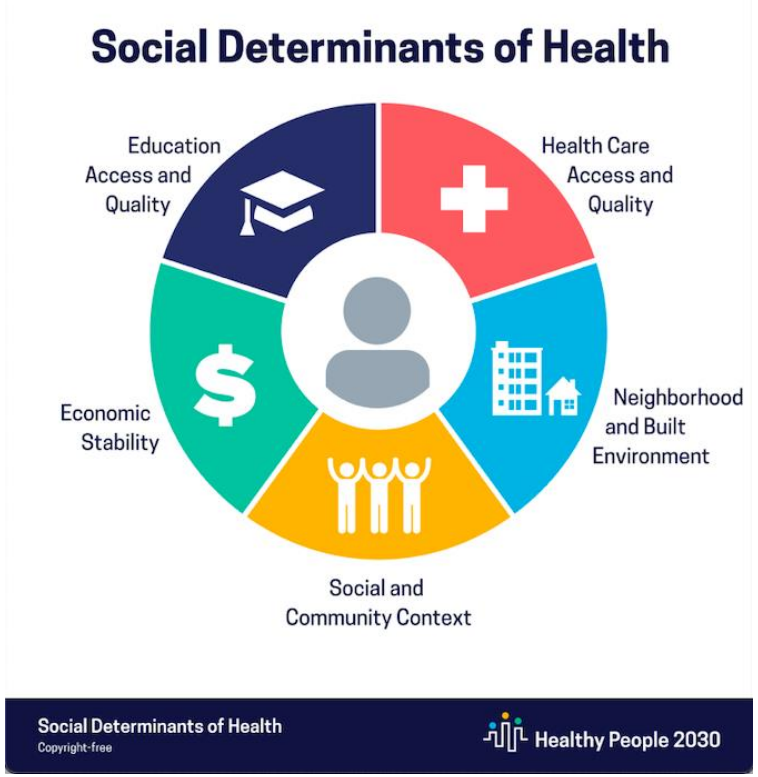
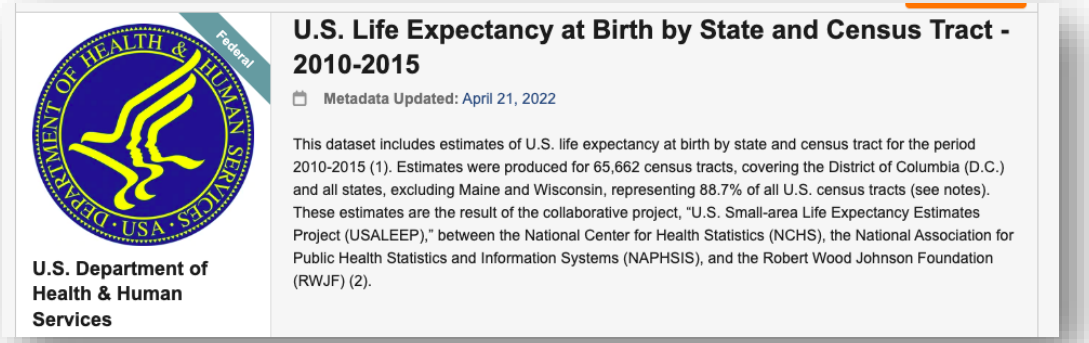


Figure 2. Social Determinants of Health as Defined by Healthy People 2030[1]

Data Collection

Life Expectancy

- Our team identified a dataset for the U.S. Life expectancy at Birth by State and Census Tract 2010-2015. The dataset includes estimates of U.S. life expectancy at birth by state and census tract for the period 2010-2015.

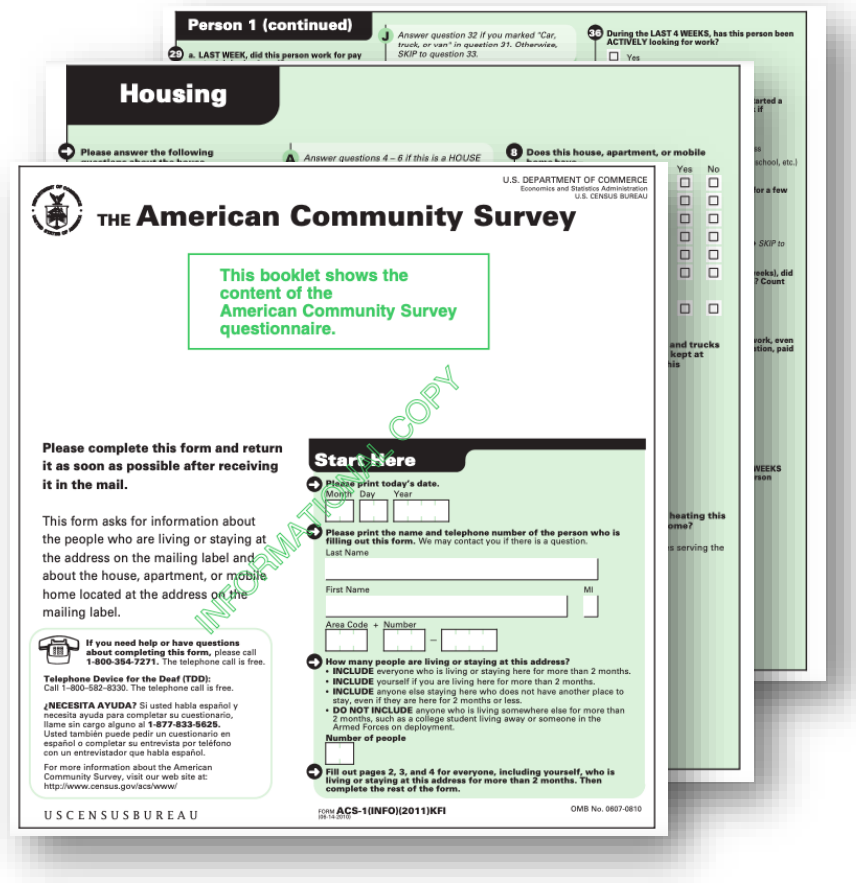


	State	County	Census Tract Number	Life Expectancy	Life Expectancy Range	Life Expectancy Standard Error
21114	Illinois	Cook County, IL	101.00	68.8	56.9-75.1	1.7306
21115	Illinois	Cook County, IL	102.01	77.3	75.2-77.5	1.9253
21116	Illinois	Cook County, IL	102.02	78.6	77.6-79.5	1.3567
21117	Illinois	Cook County, IL	103.00	70.0	56.9-75.1	1.0274
21118	Illinois	Cook County, IL	104.00	79.7	79.6-81.6	1.9647

Figure 3. Features Included in the Life Expectancy Dataset

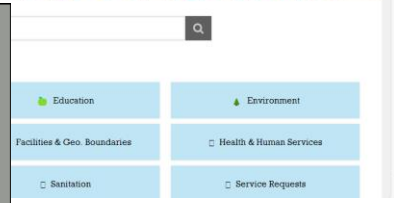
American Community Survey

- The American Community Survey (ACS) is an ongoing survey conducted by the U.S. Census Bureau. The ACS is conducted every month and provides detailed demographic, social, economic, and housing information about communities across the United States.
- The ACS collects data from a sample of households and individuals to produce estimates for various geographic areas, ranging from nationwide to specific neighborhoods.
- The ACS data is released annually in one-year, three-year, and five-year estimates
- We leveraged the 2011-2015 ACS 5-year estimates to estimate life expectancy from the US Department of Health & Human Services dataset.



Chicago Data Portal

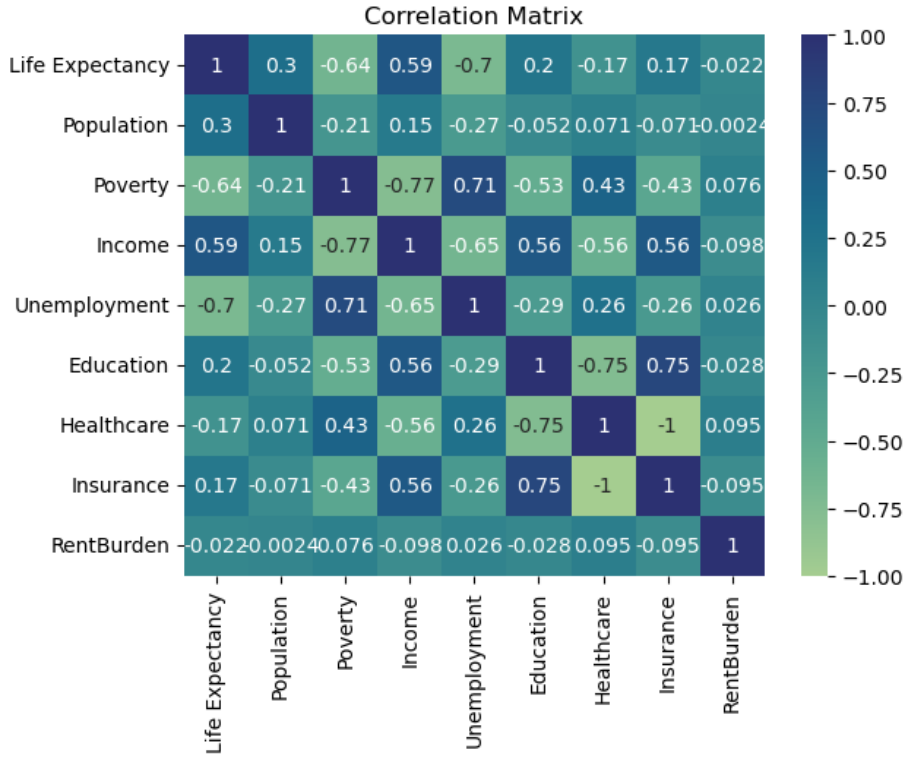
- Utilized the Chicago Data Portal to retrieve census tract information and their corresponding neighborhood/community area
- There are 77 community areas in Chicago, represented by 50 wards/alderman.
- Historically, these areas have been segregated in race, income, mobility **and as we predict**, health
- To reduce features but still preserve the influence of micro-location on the dataset, we built a mapping table to aggregate census data on the 9 geographic areas listed.

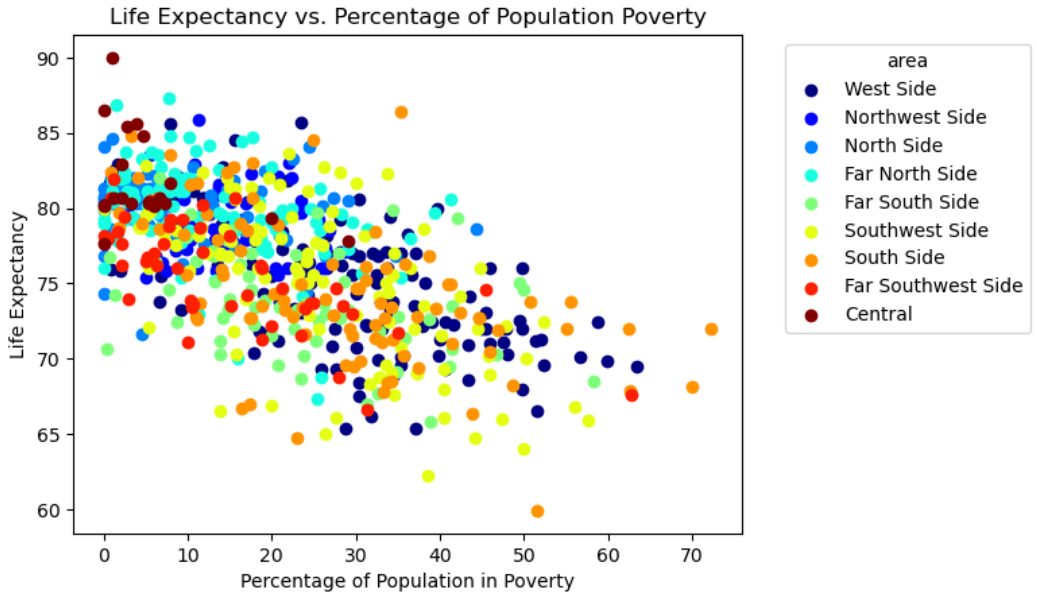


Model Preparation

Feature Selection – Initial Analysis

- We looked at the following 6 broad categories from the census tracts to determine significant areas in relation to life expectancy:
  - Poverty
  - Median Income
  - Unemployment
  - High School or higher %
  - % with healthcare
  - Rental Burden (% of income towards Rent)
- City location was also studied
- Final features were subcategories of the significant features above and aligned to the social and economic factors of health outlined by UW and the SDOH





Feature Selection 2011-2015 ACS 5-year estimates

Table 1. Model Features by SDOH Category

SDOH Categories	Features
Education Access and Quality	• **Educational Attainment:** No high school, Completed Some High School, Bachelor's degree, Graduate/Professional degree
Economic Stability	• **Employment Status:** Unemployed, Employed
Health Care Access and Quality	• **Type of Health Insurance:** None, Public, Private
Social and Community Context	• **Method of transportation to work**: Drive alone, Carpool, Walk, Other Means, Work From Home (WFH), Public Transit
Neighborhood Built and Environment	• **Vehicle Density**: No vehicles, One vehicle, Two Vehicles, Three or More Vehicles

```
features = {
    'Labor Force': 'DP03_0002E',  # Population 16+ in labor force
    'Unemployed': 'DP03_0005PE',  # %Unemployed population 16+ in civilian labor force
    'Employed':'DP03_0004PE', # %Employed population labor force
    'Family Poverty': 'DP03_0119PE',  # % Families below poverty level
    'No HS':'DP02_0059PE',#No High school
    'Some HS':'DP02_0060PE',#Some High School
    'HS Graduates': 'DP02_0061PE',  # % Population 25+ with high school diploma
    'Bachelors Degree': 'DP02_0064PE',  # % Population 25+ with bachelor's degree
    'Grad/Prof Degree': 'DP02_0065PE',  # % Population 25+ with graduate or professional degree
    'Private Health Ins': 'DP03_0097PE',  # % Population with private health insurance
    'Public Health Ins': 'DP03_0098PE',  # % Population with public health insurance
    'No Health Ins': 'DP03_0099PE',  # % Population without health insurance
    'No Vehicles': 'DP04_0058PE',  # % % Households with no vehicles
    'One Vehicle': 'DP04_0059PE',  # % Households with 1 vehicle
    'Two Vehicles': 'DP04_0060PE',  # % Households with 2 vehicles
    'Three+ Vehicles': 'DP04_0061PE',  # % Households with 3+ vehicles
    'Drive Alone': 'DP03_0019PE',  # % Commute by driving alone
    'Carpooled': 'DP03_0020PE',  #  % Commute by carpooling
    'Walked': 'DP03_0022PE', #% commute by walking
    'Other means': 'DP03_0023PE', # % commute by other means
    'WFH': 'DP03_0024PE', # % WFH
    'Public Transit': 'DP03_0021PE',  # % Commute by public transportation
    'Total Population': 'DP02_0122E',  # Total population by ancestry
    'Male':'DP05_0002PE', #% of Population Male
    'Female':'DP05_0003PE' #% of Population Female
}
```

Figure 4. Features Included in the Final Model

Limitations: The ACS provides insight into many factors that fall within the SDOH. However, since it is survey data, some SDOH category features are limited. Even though there are many additional features our model could have included to give a holistic overview of each SDOH category, we chose to include these features since they represent factors of each category. Additionally, since they come from the ACS, they can be analyzed at a census tract level.

Exploratory Data Analysis (EDA)

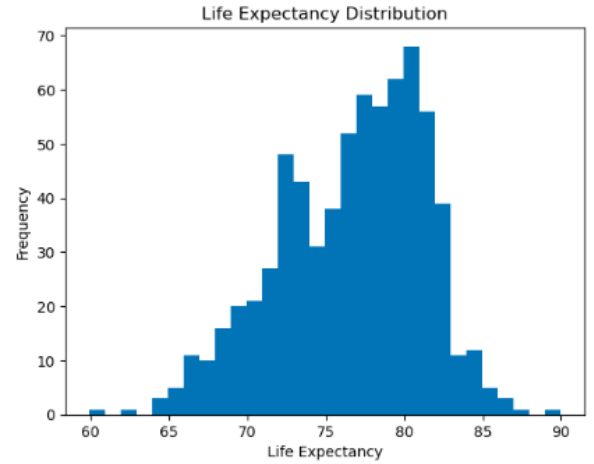
Dataset Size:
- 800 census tract records
- 46 columns

Data Quality:
- 99 missing values in life expectancy
- 3 census tracts with negative values across all features

Missing Data Handling:
- 3 census tracts with negative values were excluded from the analysis
- Life expectancy missing values were handled using KNN imputer

KNN Imputer
- Utilizes Euclidian Distance and k-nearest neighbor's algorithm
- This was useful imputation method as the missing y values were not distributed towards any particular x feature

A)



B)

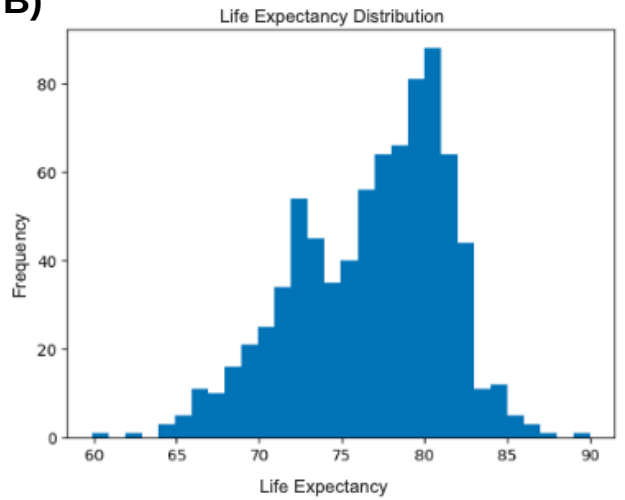


Figure 5. A) represents the distribution of life expectancy prior to KNN imputation. B) represents the distribution after KNN imputation

Model Preparation Summary

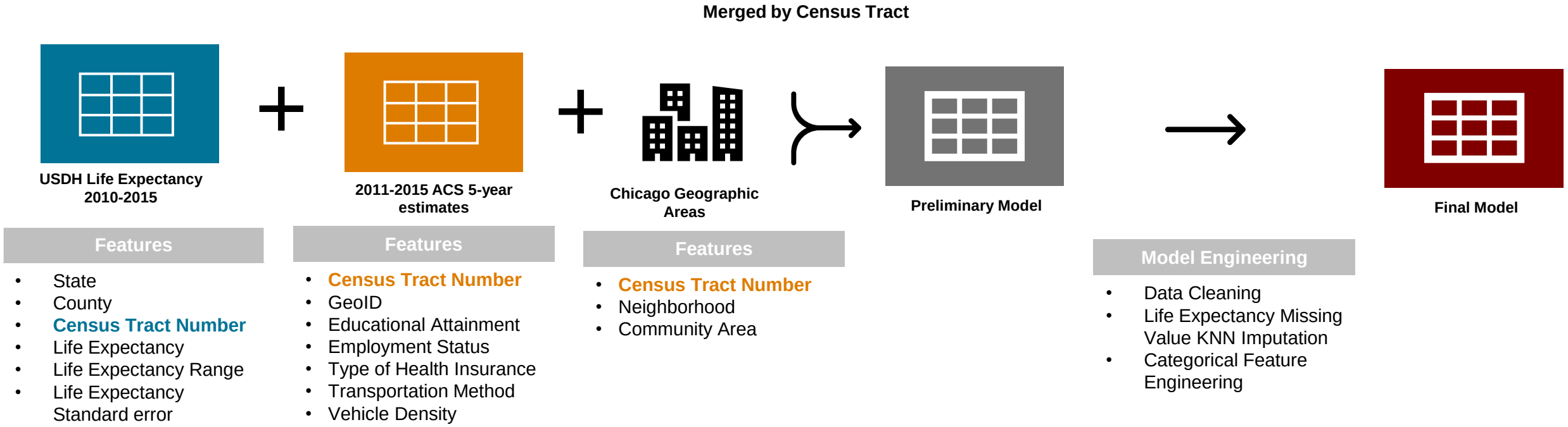


Figure 5. Visual Representation of Model Preparation

Methodology and Interpretation

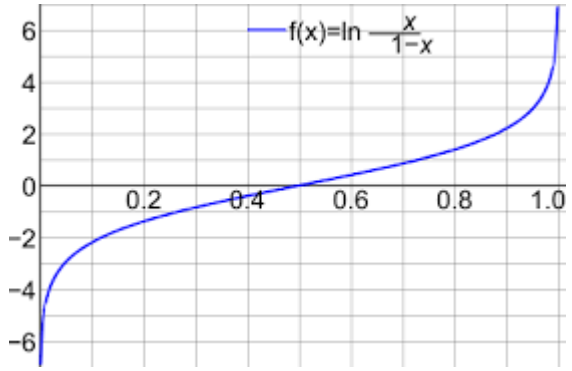
Methods Applied

- Logistic Regression
- Explainable Boosting Machine Classifier
- XGBOOST

Included analysis with and without Chicago Location information to understand how significant community area is on Life Expectancy

Logistic Regression

- To apply both a classification and logistic regression we **created a binary response** (above average or below average population life expectancy)
- Returns log odds for each coefficient
- Easily interpretable

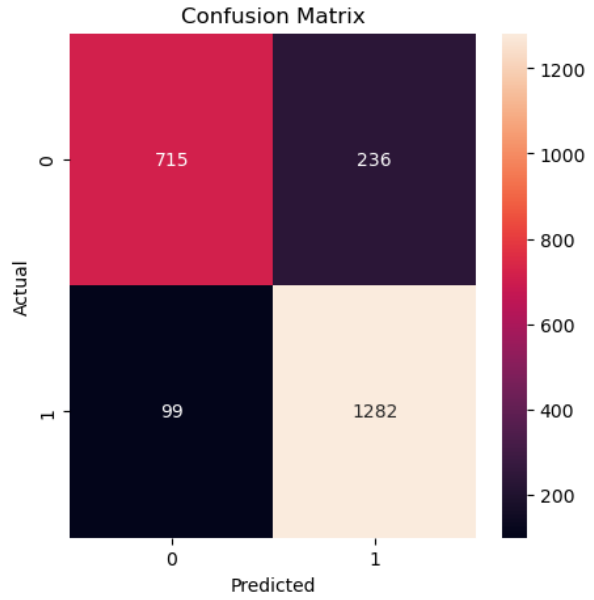



```
#Create a function that takes the logistic coefficient and returns the adjusted probability
def adj_prob(coef,data=df):
    init_prob=data['Life Expectancy_belavg'].sum()/data['Life Expectancy_belavg'].count()
    baseline=(init_prob/(1-init_prob))
    adj_odds=baseline*np.exp(coef)
    adj_prob=adj_odds/(1+adj_odds)
    return adj_prob
```

Logistic Interpretation

Model without Location Data

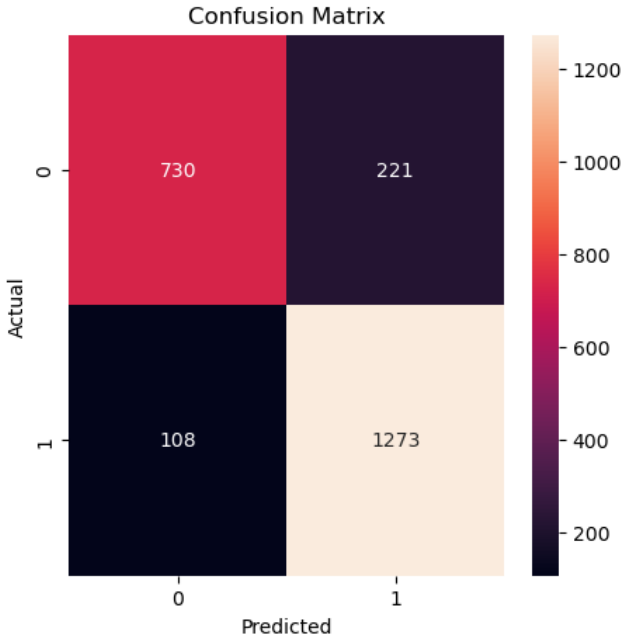
AUC: 0.904



	precision	recall	f1-score	support
0	0.88	0.75	0.81	951
1	0.84	0.93	0.88	1381
accuracy			0.86	2332
macro avg	0.86	0.84	0.85	2332
weighted avg	0.86	0.86	0.85	2332

Model With Location Data

AUC: 0.917



	precision	recall	f1-score	support
0	0.87	0.77	0.82	951
1	0.85	0.92	0.89	1381
accuracy			0.86	2332
macro avg	0.86	0.84	0.85	2332
weighted avg	0.86	0.86	0.86	2332

Continued

Logit with no location (.90 AUC)

	Coefficient	Adjusted Probability	prob_lowLifeExpectancy
Public Health Ins	-0.037199	0.392962	0.607038
Unemployed	-0.027277	0.395331	0.604669
Some HS	-0.022026	0.396587	0.603413
One Vehicle	-0.018639	0.397398	0.602602
Female	-0.017823	0.397594	0.602406

Logit with location (.92 AUC)

	Coefficient	Adjusted Probability	prob_lowLifeExpectancy
area_Far South Side	-1.679586	0.111326	0.888674
area_Far Southwest Side	-1.626125	0.116726	0.883274
vehicle_cat_Three+ Vehicles	-1.304747	0.154149	0.845851
area_South Side	-1.266939	0.159143	0.840857
area_West Side	-1.085144	0.185001	0.814999

TOP 5 Features for high LE (no location)

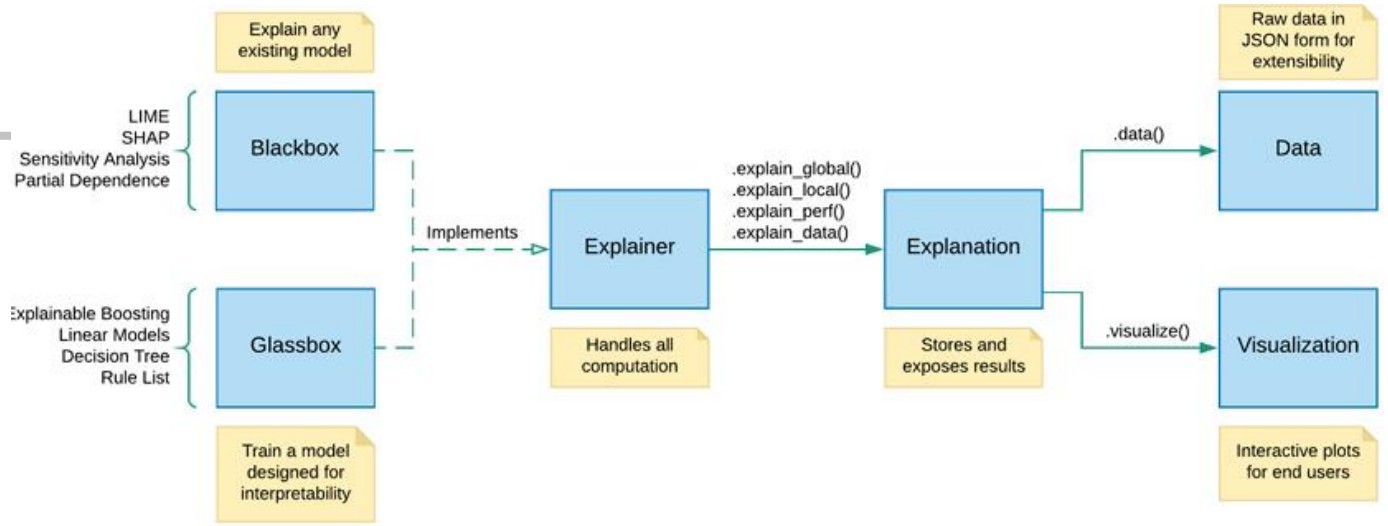
Private Health Ins	0.009863	0.404243	0.595757
Two Vehicles	0.017672	0.406125	0.593875
No Health Ins	0.020196	0.406734	0.593266
Bachelors Degree	0.028450	0.408727	0.591273
No HS	0.033055	0.409841	0.590159
Employed	0.042458	0.412117	0.587883

TOP 5 Features for high LE

health_ins_cat_Private Health Ins	0.970905	0.639507	0.360493
vehicle_cat_Two Vehicles	1.071727	0.662409	0.337591
Unemployed_cat_less than 5%	1.149809	0.679642	0.320358
Unemployed_cat_5-10%	1.294337	0.710262	0.289738
area_North Side	1.537302	0.757610	0.242390
area_Central	3.175962	0.941491	0.058509

InterpretML- EBM

- Utilizes General Additive Models
- Has the classification accuracy of XGBOOST and other Classifiers while maintaining the interpretability of other methods
- Helps demystify black box predictors and introduces more ‘glass box’ methods



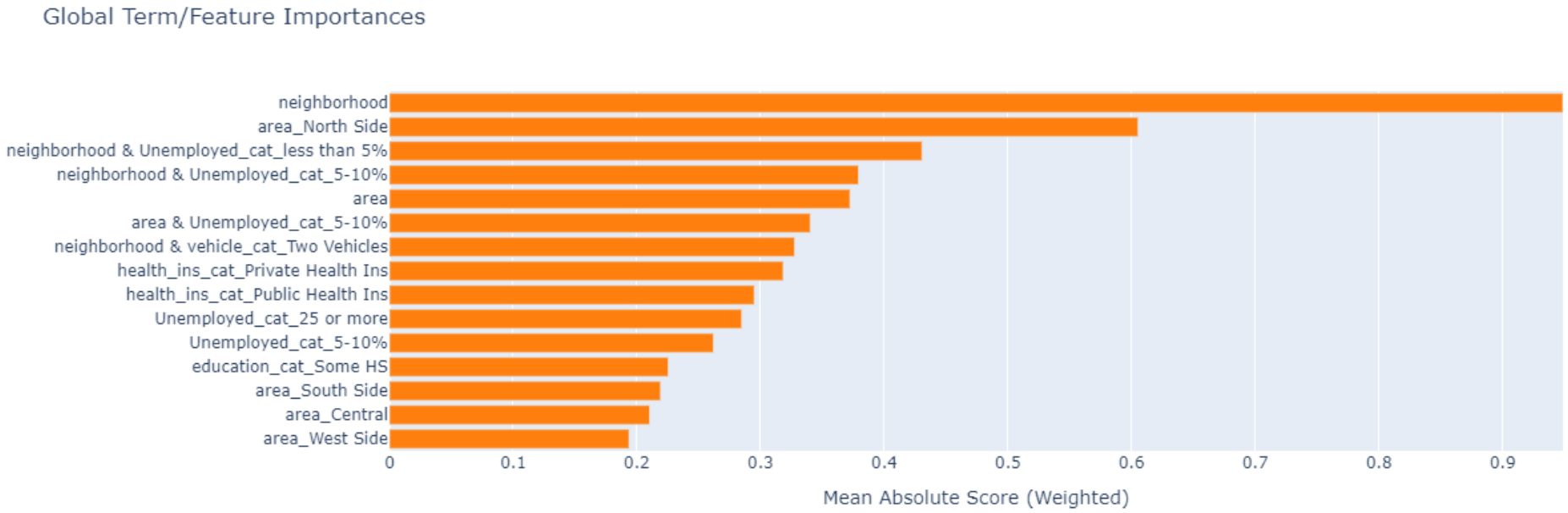
Classification Performance (AUROC)					
Model	heart-disease (303, 13)	breast-cancer (569, 30)	telecom-churn (7043, 19)	adult-income (32561, 14)	credit-fraud (284807, 30)
EBM	**0.916**	**0.995**	**0.851**	**0.928**	0.975
LightGBM	0.864	0.992	0.835	**0.928**	0.685
Logistic Regression	0.895	**0.995**	0.804	0.907	0.979
Random Forest	0.89	0.992	0.824	0.903	0.95
XGBoost	0.87	**0.995**	0.85	0.922	**0.981**

InterpretML: A Uni ed Framework for Machine Learning Interpretability
Harsha Nori Samuel Jenkins Paul Koch Rich Caruana

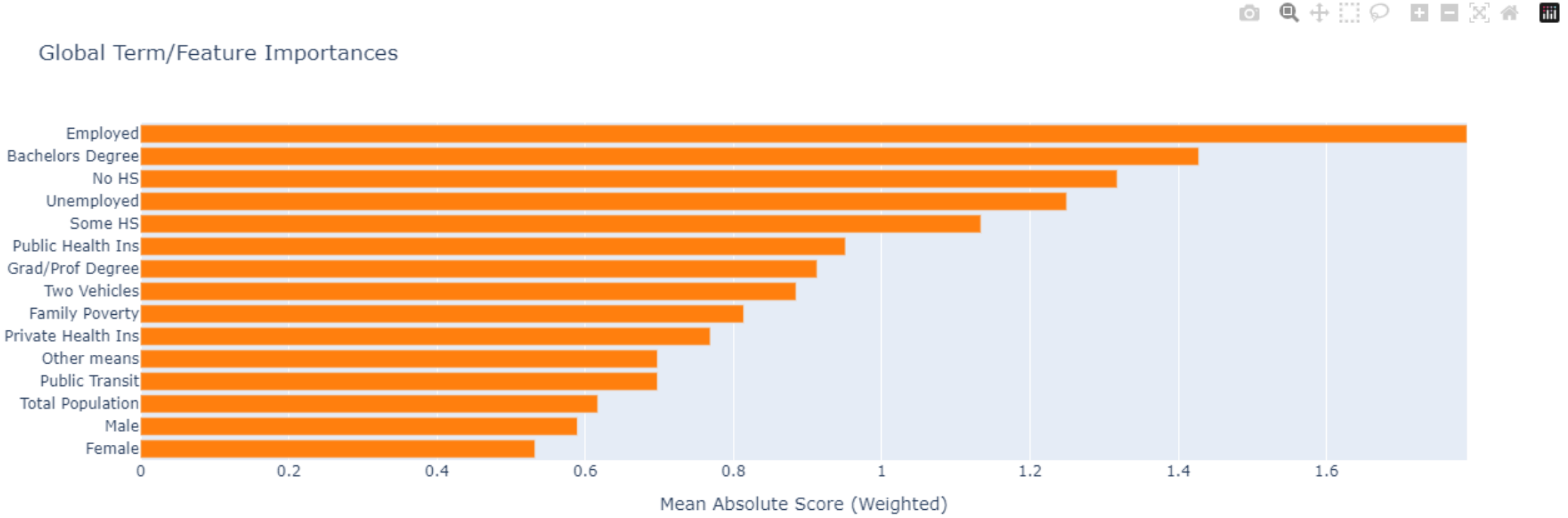
1909.09223.pdf (arxiv.org)

EBM Results

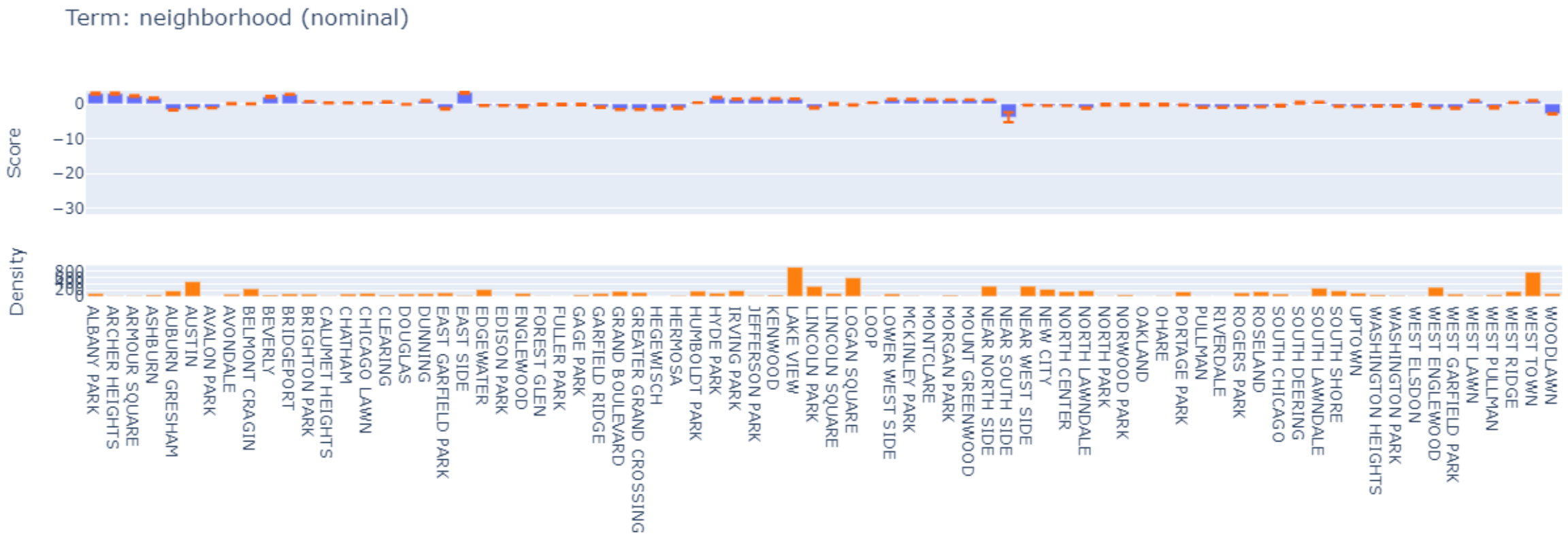
ExplainableBoostingClassifier_2



The term importances are the mean absolute contribution (score) each term (feature or interaction) makes to predictions averaged across the training dataset. Contributions are weighted by the number of samples in each bin, and by the sample weights (if any). The 15 most important terms are shown. Learn more



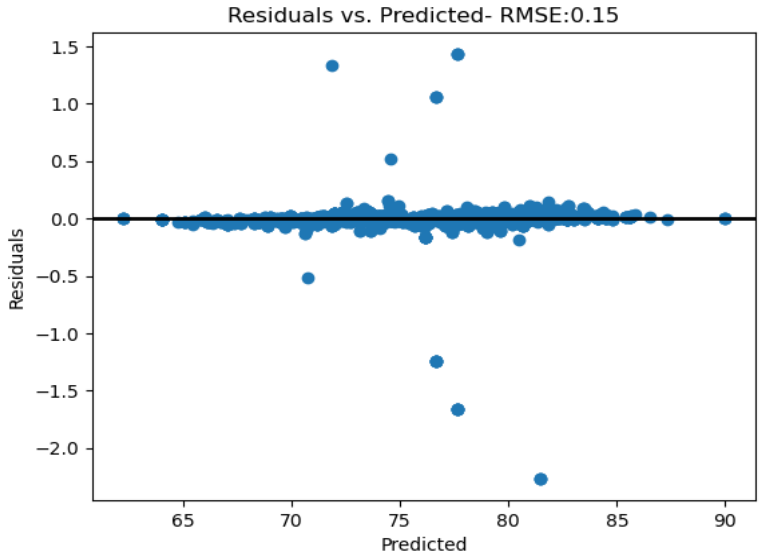
The term importances are the mean absolute contribution (score) each term (feature or interaction) makes to predictions averaged across the training dataset. Contributions are weighted by the number of samples in each bin, and by the sample weights (if any). The 15 most important terms are shown. Learn more



The contribution (score) of the term neighborhood to predictions made by the model. For classification, scores are on a log scale (logits). For regression, scores are on the same scale as the outcome being predicted (e.g., dollars when predicting cost). Each graph is centered vertically such that average prediction on the train set is 0. Learn more

XGBOOST

- **XGBoost Overview**: A high-performance gradient boosting library for regression, classification, and ranking problems, known for its speed and accuracy.
- **Key Features**:
  - **Efficiency and Scalability**: Handles large datasets with speed and efficiency.
  - **Built-in Regularization**: Reduces overfitting to improve model performance.
  - **Flexible and Portable**: Supports multiple languages and runs on various platforms.
- **Algorithmic Advancements**:
  - Utilizes advanced tree learning algorithms and parallel processing to enhance prediction accuracy and reduce computation time.
- **Practical Applications**:
  - Widely used in industries like finance, healthcare, and e-commerce for predictive modeling and data analysis.



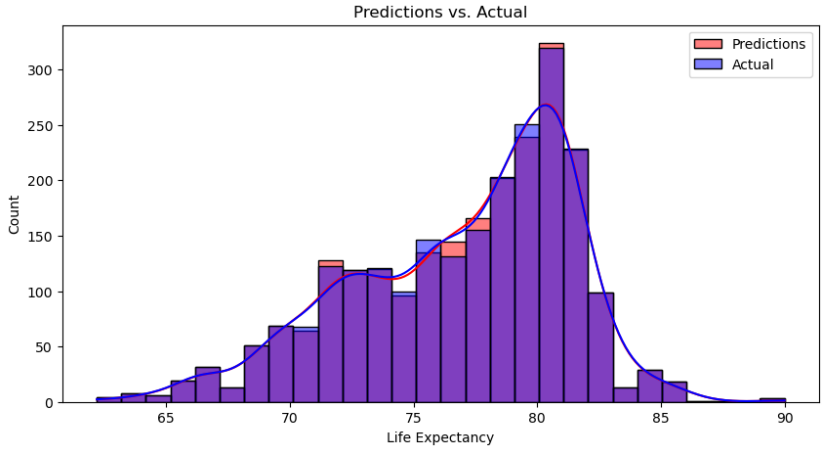
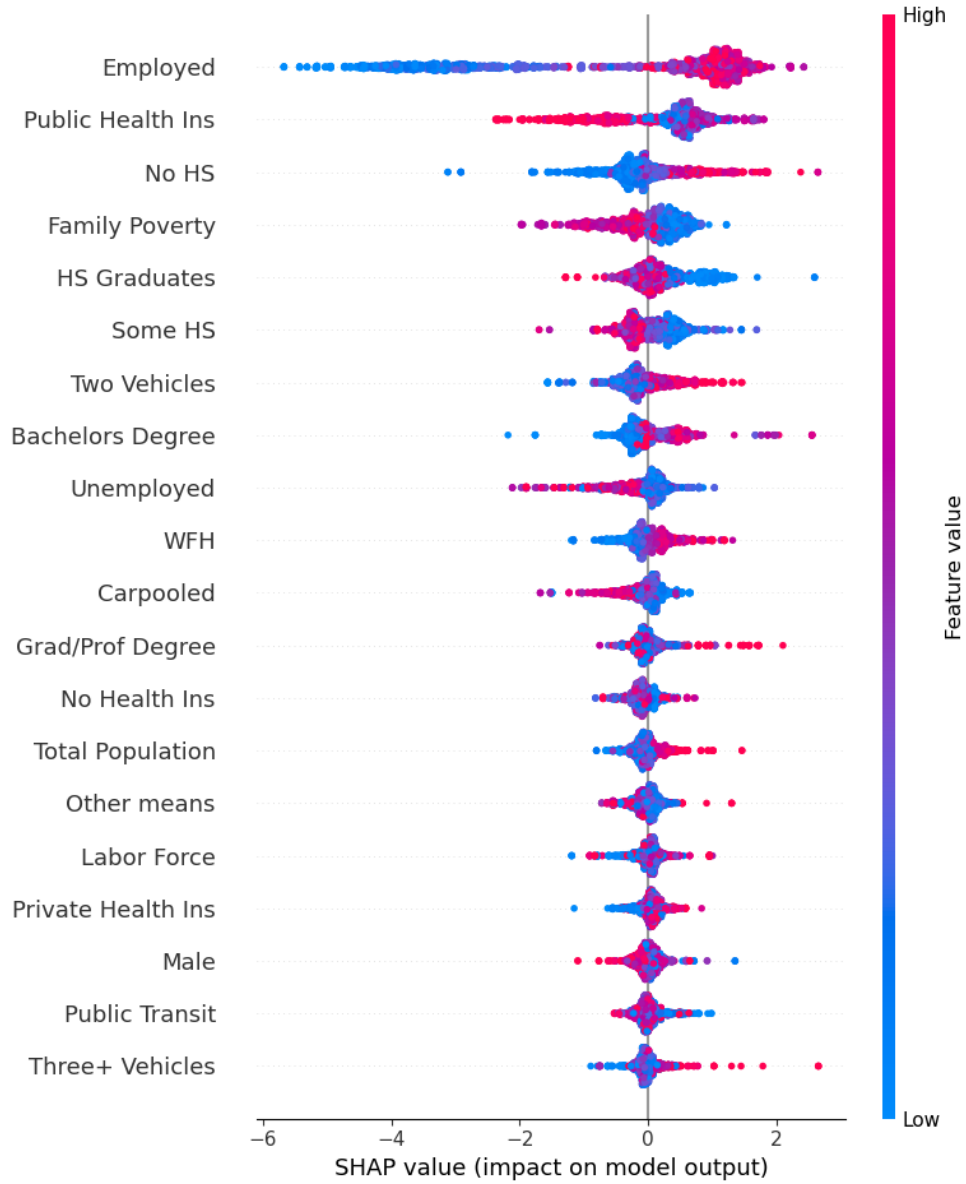
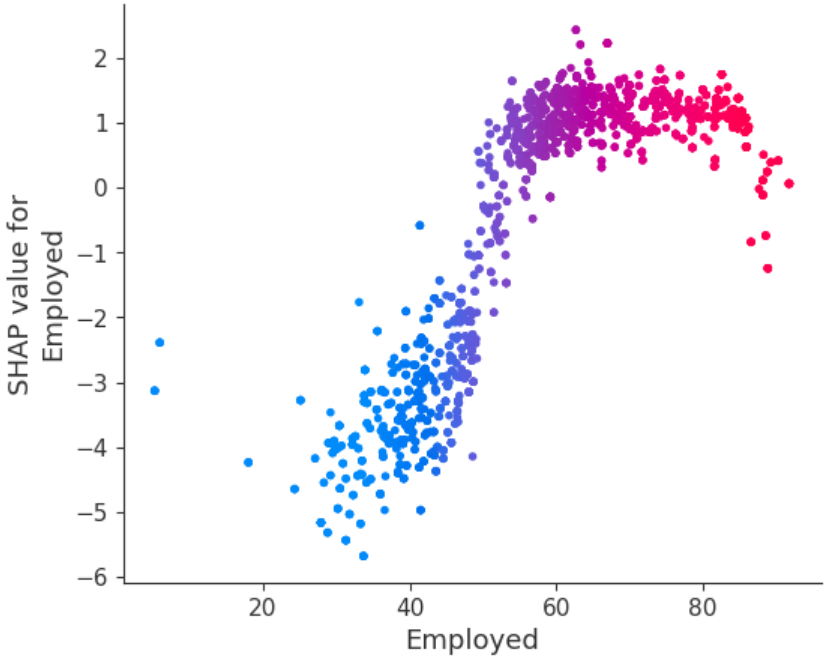


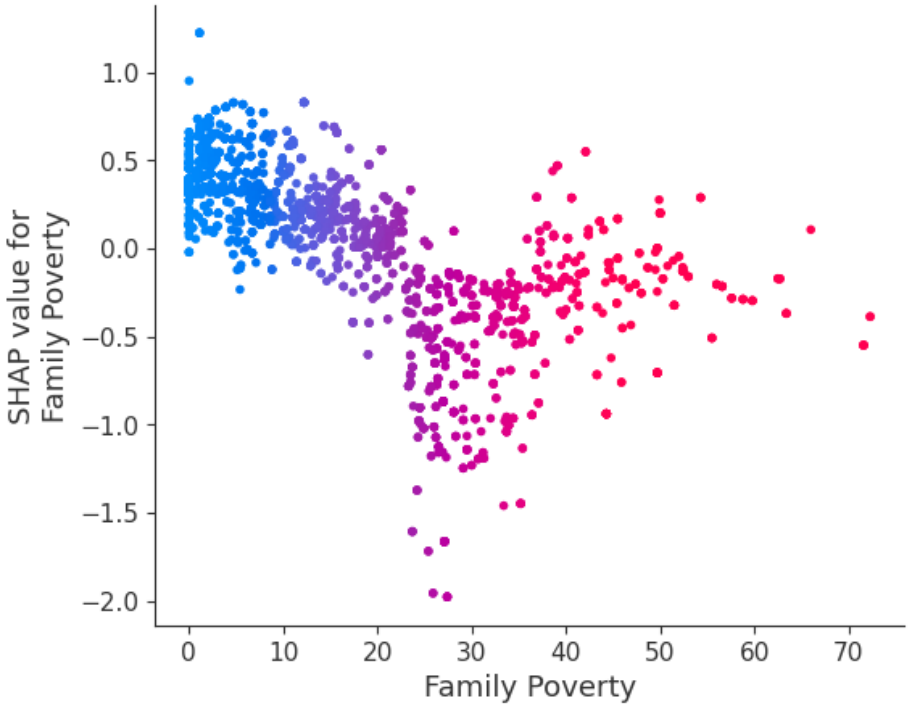
Figure 8. Represents our XGBOOST Model for predicting actual census life expectancy and its accuracy

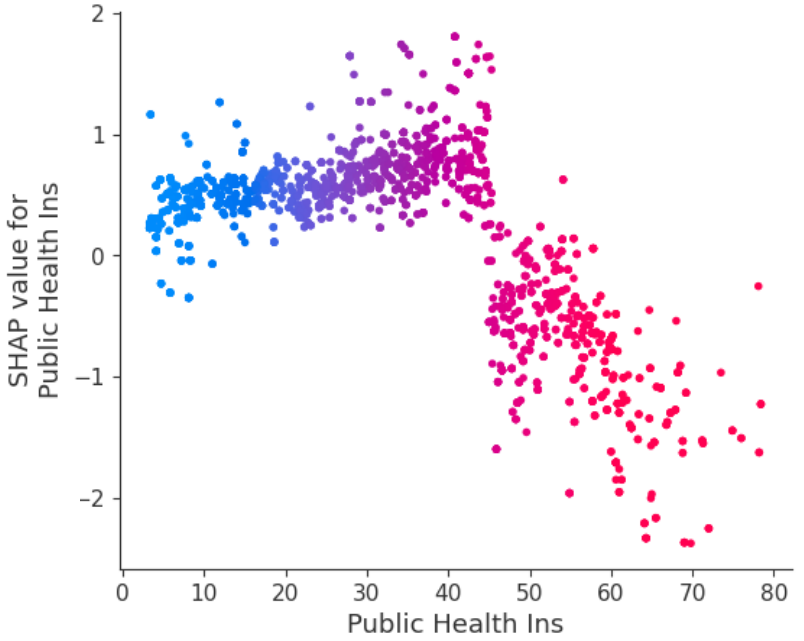
SHAP

- **Key Concepts**:
  - **Shapley Values**: Borrowed from cooperative game theory, they represent an equitable distribution of "payouts" (predictions) among "players" (features).
  - **Global and Local Interpretability**: Offers both overall importance of features across the model and specific impact of features on individual predictions.
- **Advantages of SHAP**:
  - **Consistency and Fairness**: Ensures consistent feature importance rankings, addressing inconsistencies present in other methods.
  - **Model Agnostic**: Works with any machine learning model, enhancing its versatility in application.




SHAP value for
Employed
Employed


SHAP value for
Family Poverty
Family Poverty


SHAP value for
Public Health Ins
Public Health Ins

Takeaway:

- **Improving life expectancy in a community area is primarily tied to the employment and income of the community.**
  - There are underlying features not captured in the census data, but associated with the neighborhood in Chicago.
  - Further studies should be done to identify what features of the neighborhoods result in the disparity, specifically in the south and west sides

Improvements & Future Use

Improvements and Future Use

- **Enhanced Data Collection**:
  - Expand the dataset to include **more recent years**, capturing the impact of the COVID-19 pandemic on life expectancy.
- **Incorporate More SDOH Factors**:
  - **Access to digital technology** which became increasingly important for education, work, and healthcare.
- **Interdisciplinary Collaboration:**
  - Foster stronger collaborations between public health professionals, urban planners, and technology experts to design integrated solutions that address the root causes of health disparities.
  - This cross-disciplinary approach can lead to innovative strategies that are more effective in improving life expectancy and resilience to health crises.
- **Policy Development:**
  - Inform policymakers on features that need to be addressed to improve public health outcomes, both in normal times and during emergencies
- **Health Equity Initiatives**
  - Leverage findings to design targeted health equity initiatives

Thank You

Appendix

References

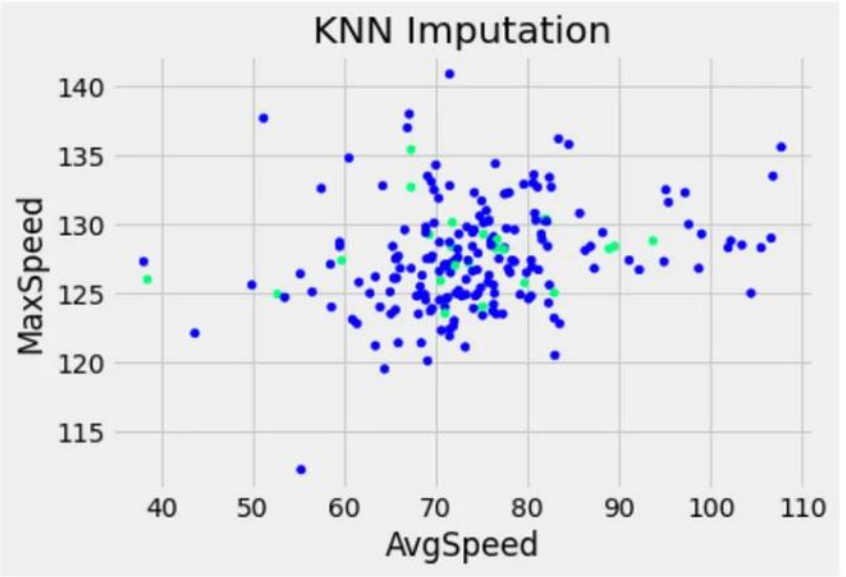
1. Healthy People 2030, U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion. Retrieved [26-Feb-2024], from https://health.gov/healthypeople/objectives-and-data/social-determinants-health
2. World Health Organization. (n.d.). World Health Organization. Retrieved from https://www.who.int/data/gho/indicator-metadata-registry/imr-details/65
3. University of Wisconsin Population Health Institute. County Health Rankings Key Findings 2014.
4. U.S. Census Bureau. (2024, February 26). American Community Survey (ACS) and the Census. Census.gov. Retrieved February 26, 2024, from https://www.census.gov/programs-surveys/acs/about/acs-and-census.html

Categorical Feature Engineering

Category	Categories
Unemployment	New column that populated the unemployment range that captures the unemployment percentage observed in that census tract • Less than 5% • 5-10% • 10-15% • 15-20% • 20-25% • 25 or more
Sex	New column that populated the sex that was more predominant in that census tract • More Men • More Women
Education	New column that populated the education level that was more predominant in that census tract • eg. If census tract had the highest percentage in bachelor's degree, then new column would be populated with "Bachelors Degree"
Health Insurance	New column that populated the type of health insurance that was more predominant in that census tract • eg. If census tract had the highest percentage in private health insurance, then new column would be populated with "Private Health Ins"
Mode of Transportation	New column that populated the mode of transportation that was more predominant for commuting to work in that census tract • eg. If census tract had the highest percentage in public transit, then new column would be populated with "Public transit"
Vehicle Density	New column that populated the vehicle that was more predominant in that census tract • eg. If census tract had the highest percentage in two vehicles, then new column would be populated with "two vehicles"
Life Expectancy	The average life expectancy was obtained from the mean of observations in our data set. A binary categorical variable was created to determine if the life expectancy was above or below the average. (1 = above average, 0 = below average)

KNN Imputer

- Utilizes k-Nearest Neighbors method
- Replaces missing value based on the mean value from the parameter n_neighbors found in the dataset using Euclidean Distance
- Each tract has relatively similar population and is on the same scale
- A majority of other features were populated and no specific features required more imputation than others



Imputing Missing Data with Simple and Advanced Techniques | by Idil Ismiguzel | Towards Data Scienc