Predicting Coronary Heart Disease

Henri Antikainen, Rachel Appel, Melanie King, Sriram Raghunath, Shanise Walker

The Erdős Institute Fall 2023 Data Science Bootcamp

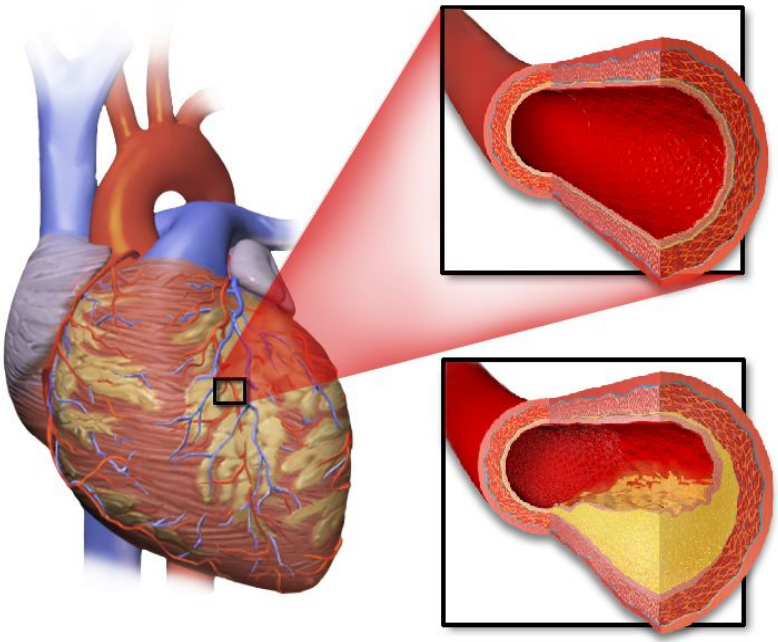
Overview: Coronary Heart Disease

Heart disease, a type of cardiovascular disease, is the leading cause of death in the United States. ***Coronary heart disease*** (coronary artery disease) is the most common type of heart disease and is responsible for over 365,000 deaths each year.

Stakeholders: People living in the United States, county officials, healthcare providers, policy makers

KPI: Mean Squared Error

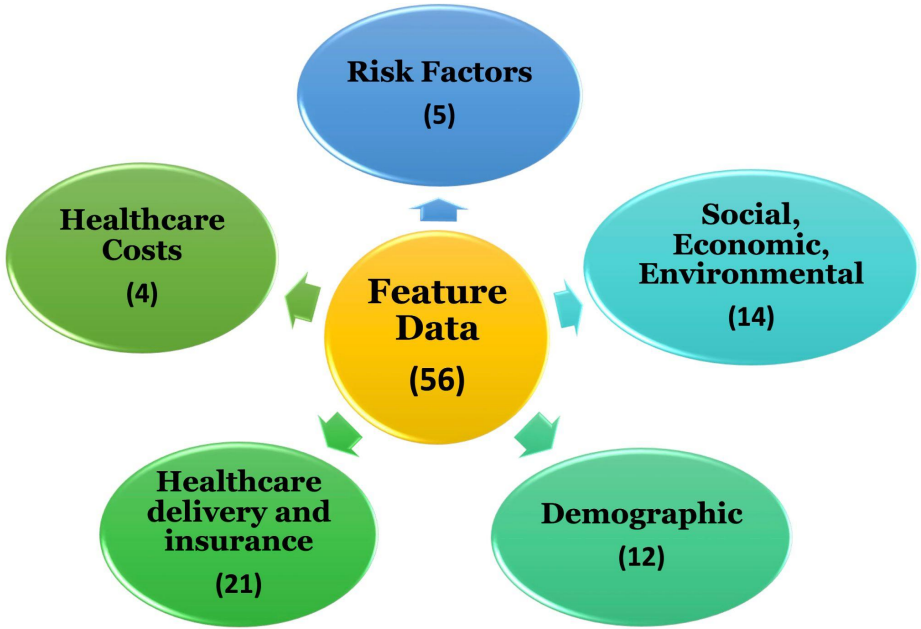
Goal: Develop a model that accurately predicts coronary heart disease and detects important features.



Normal and Partially Blocked Blood Vessels

Dataset Information

- Gathered from ***Center for Disease Control and Prevention's Interactive Atlas of Heart Disease and Stroke*** (IAHDS) online mapping tool
- ***County level data*** for 3226 counties in the United States, which includes data from all 50 states and all US territories
- ***59 total columns*** in the data set at the county level
  - County fips codes
  - County names and state
  - Coronary heart disease percentage
  - 56 possible modeling features



Data Cleaning and Challenges

- ***Missing data*** for some county features, especially US territories, identified in the data set as value -1
- States with ***small number of counties*** (i.e. less than five data points available)
- ***Modified data*** to remove all US territories, Alaska, Hawaii, Washington DC, and Delaware

	cnty_fips	display_name	heart_disease	high_cholesterol	diagnosed_diabetes	obesity	physical_inactivity	current_smoker	broadband_internet	computer	...
0	2013	"Aleutians East, (AK)"	5.9	31.2	9.9	27.2	21.5	18.5	42.1	11.5	...
1	2016	"Aleutians West, (AK)"	4.6	30.3	9.3	25.4	20.0	16.7	21.0	8.2	...
2	2020	"Anchorage, (AK)"	4.9	29.4	8.3	29.8	17.9	15.7	8.0	3.3	...
3	2050	"Bethel, (AK)"	8.1	28.7	8.8	23.8	22.0	34.0	26.6	10.2	...
4	2060	"Bristol Bay, (AK)"	7.5	32.3	9.2	24.6	20.9	17.8	19.0	7.0	...

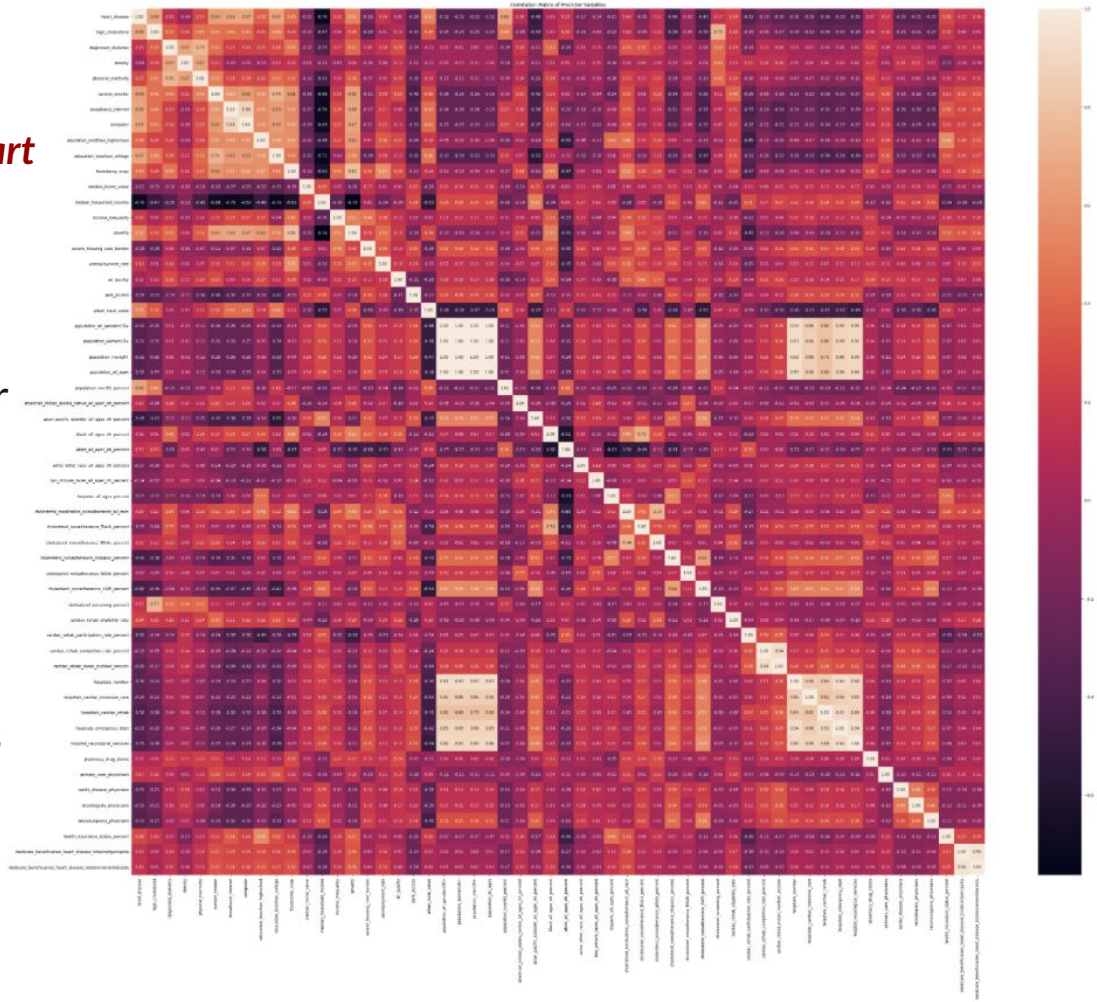
Exploring Data Set

Top features *strongly correlated with heart disease*:

- ❏ High cholesterol
- ❏ Households without a computer
- ❏ Individuals of age 25+ without 4 or more years of college

Top features *negatively correlated with heart disease*:

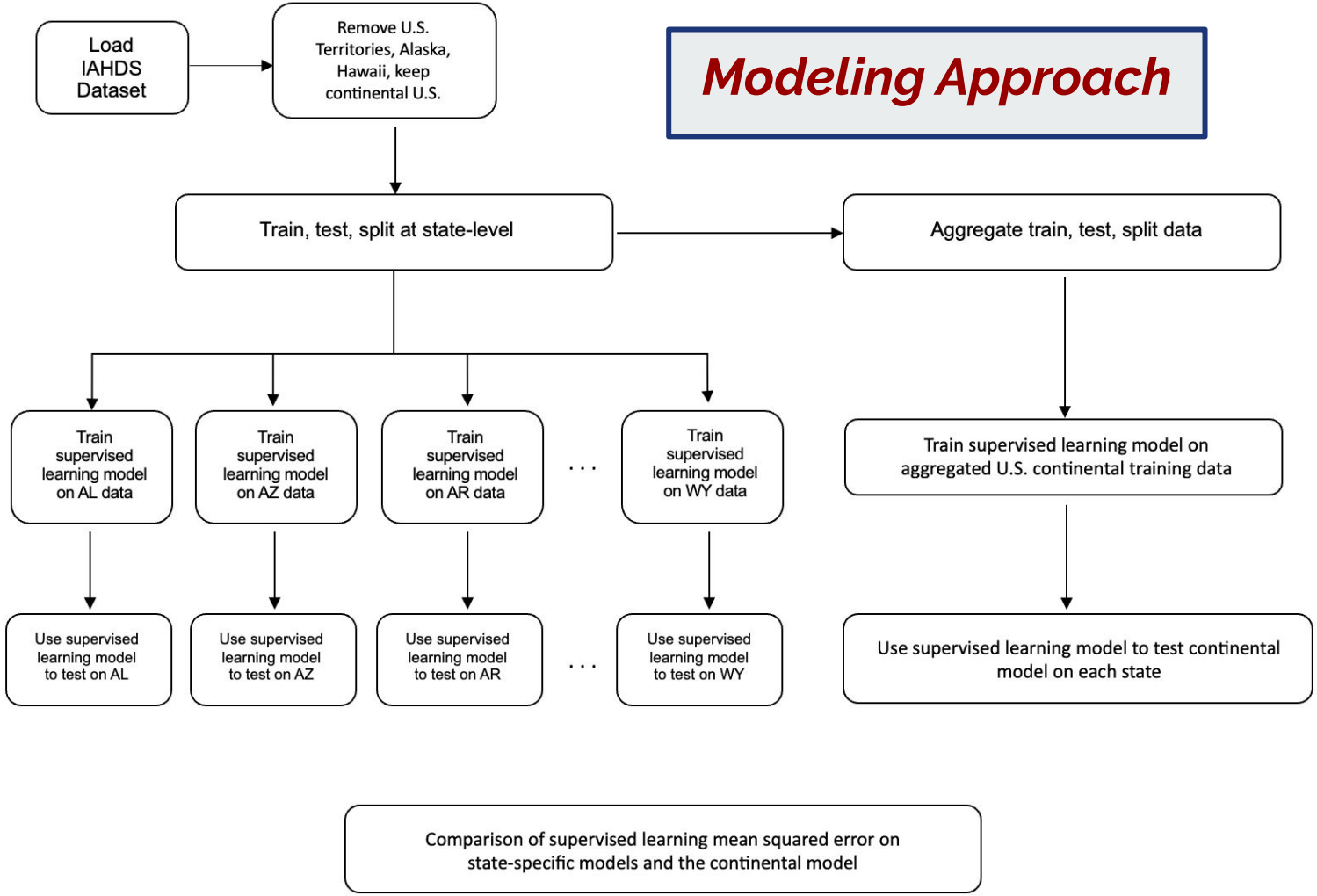
- ❏ Median household income
- ❏ Asian Pacific Islander race, all ages
- ❏ Asian and Pacific Islander cholesterol-lowering medication nonadherence, medicare beneficiaries Part D



Training Three Models:

- *XGBoost*
- *Gaussian Naive Bayes*
- *Linear Regression (LassoCV)*

Modeling Approach



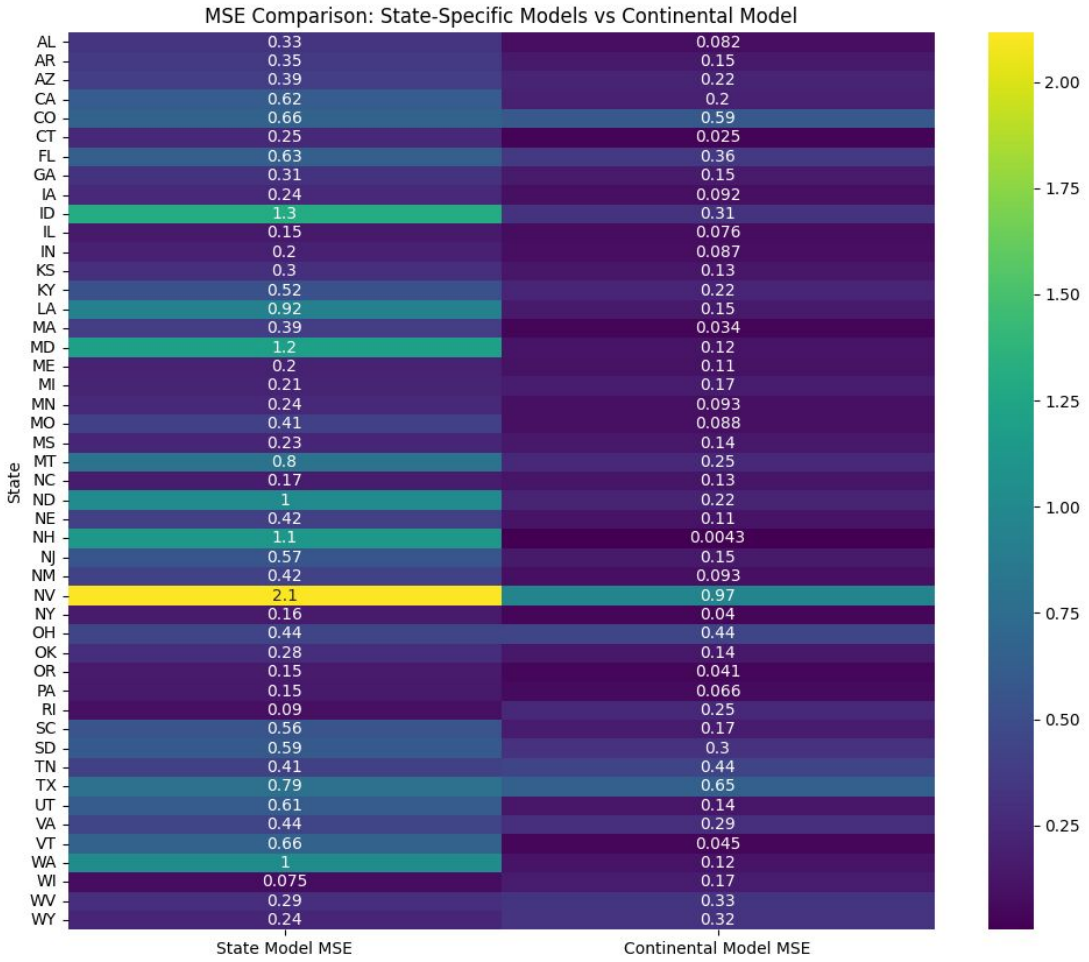
Comparison of Three Continental Models

Model	*MSE*	*MAE*	*R-squared score*
XGBoost	0.226	0.338	0.901
Gaussian Naive Bayes	5.440	1.667	0.338
Linear Regression (LassoCV)	0.278	0.323	0.879

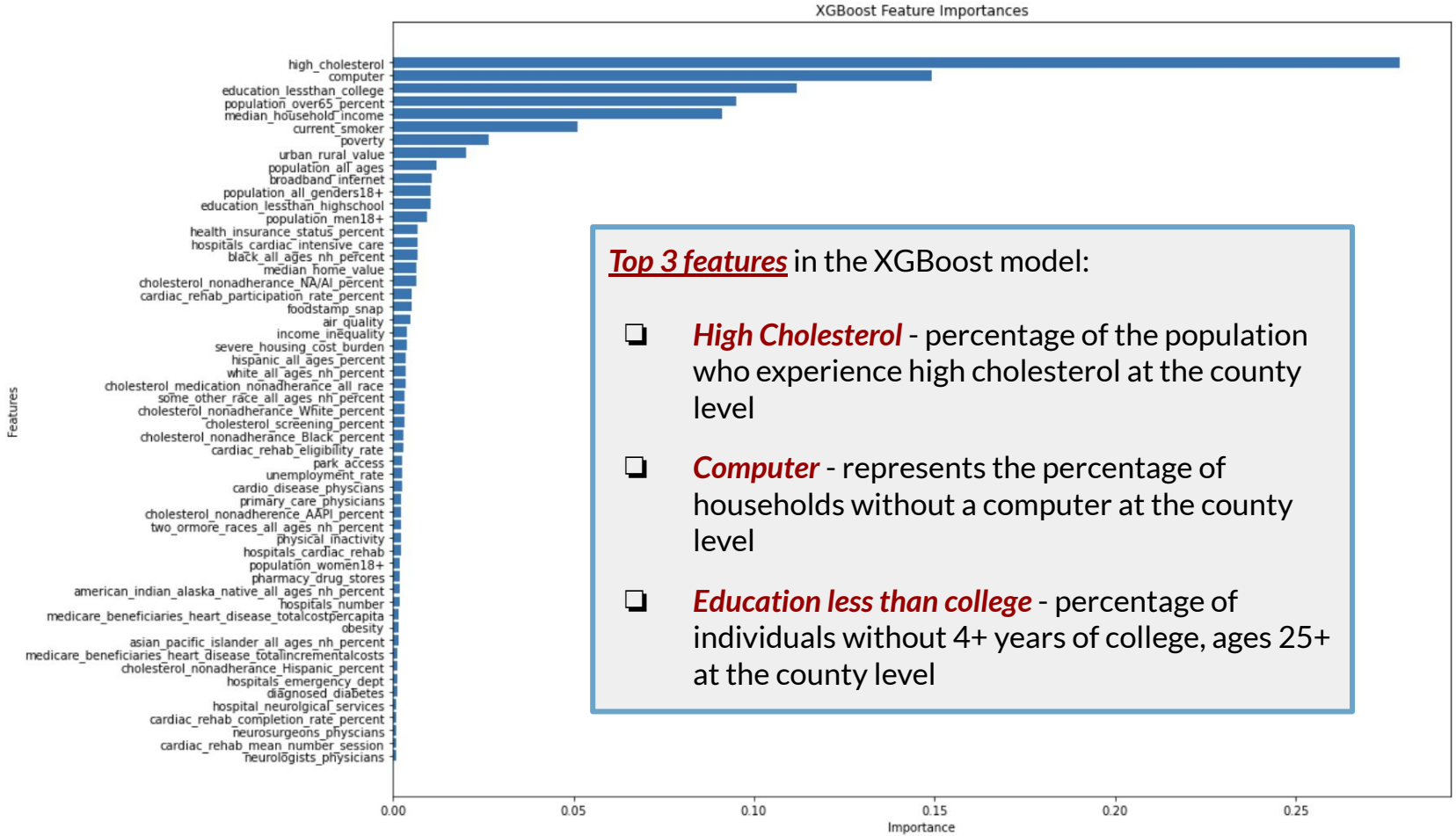
- State-level training data is constructed so that 80% of the available data from each state is represented in the set.
- The continental model is supervised learning on the aggregated state-level data.

XGBoost Model Analysis

- Notable results:
  - The continental model ***outperforms*** the state specific model.
  - Both models perform the worst on ***Nevada*** data.
  - Both models perform the same on ***Ohio*** data.



XGBoost Model Feature Importances



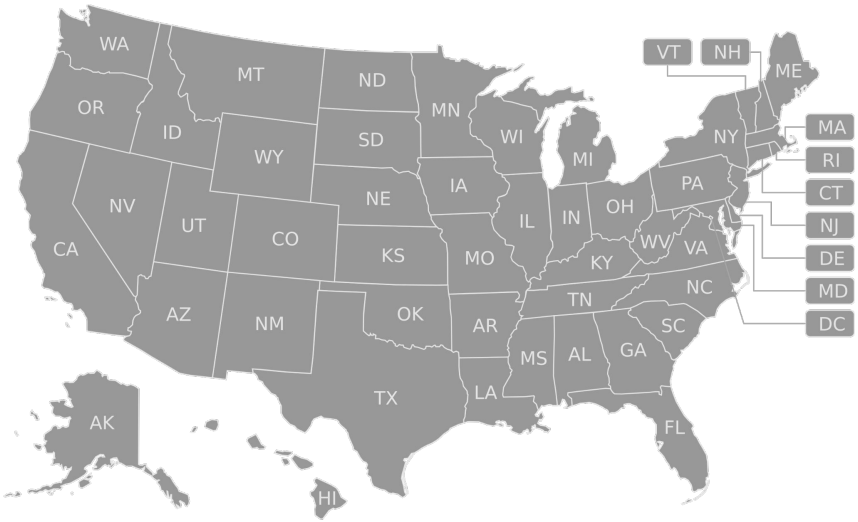
Top 3 features in the XGBoost model:

- ***High Cholesterol*** - percentage of the population who experience high cholesterol at the county level
- ***Computer*** - represents the percentage of households without a computer at the county level
- ***Education less than college*** - percentage of individuals without 4+ years of college, ages 25+ at the county level

Future Work/Next Steps

For the county officials and the general population, we will produce predictive information about coronary heart disease which is relevant to geographic location.

- ***Deeper Analysis of XGBoost Model***
  - Compare feature importance
- ***Improve Supervised Learning Models***
  - Include states with small number of counties
  - Add data for Hawaii, Alaska, and US territories
  - Better feature selection
- ***Disseminate Results***
  - Create an interactive map



https://en.wikipedia.org/wiki/File:Blank_US_Map_With_Labels.svg