Predicting the hyper-local prevalence of chronic kidney disease

A Novel Approach Using Census Data and Stochastic Gradient Descent

Why Chronic Kidney Disease?

- 30 million Americans
- $31 billion in annual treatment costs
- 89,000 deaths per year
- Easily managed if caught in early stages
- Essential to find affordable and effective modes of prevention

About the Data

Data Collection

500 Cities: Local Data for Better Living (Centers for Disease Control)

- CSV with census-tract level data on chronic disease, poor health indicator, and preventative behavior rates

5-year American Community Survey (U.S. Census Bureau)

- Pulled data on 235 demographic features from Census Bureau API

28,004 census tracts (observations) from 500 largest cities in the U.S.

Data Cleaning: 500 Cities

- Limited to 1 target variable to allow for future reproducibility
  - Explored unhealthy behavior and preventative services data for actionable insights
- Paired down to census tract-level data
- Pivoted from long to wide data
- Extracted Tract ID from UniqueID column
- Converted to numeric and categorical (state and city) data types
- Imputed missing data with the variable mean
  - Using median vs mean did not seem to change outcome

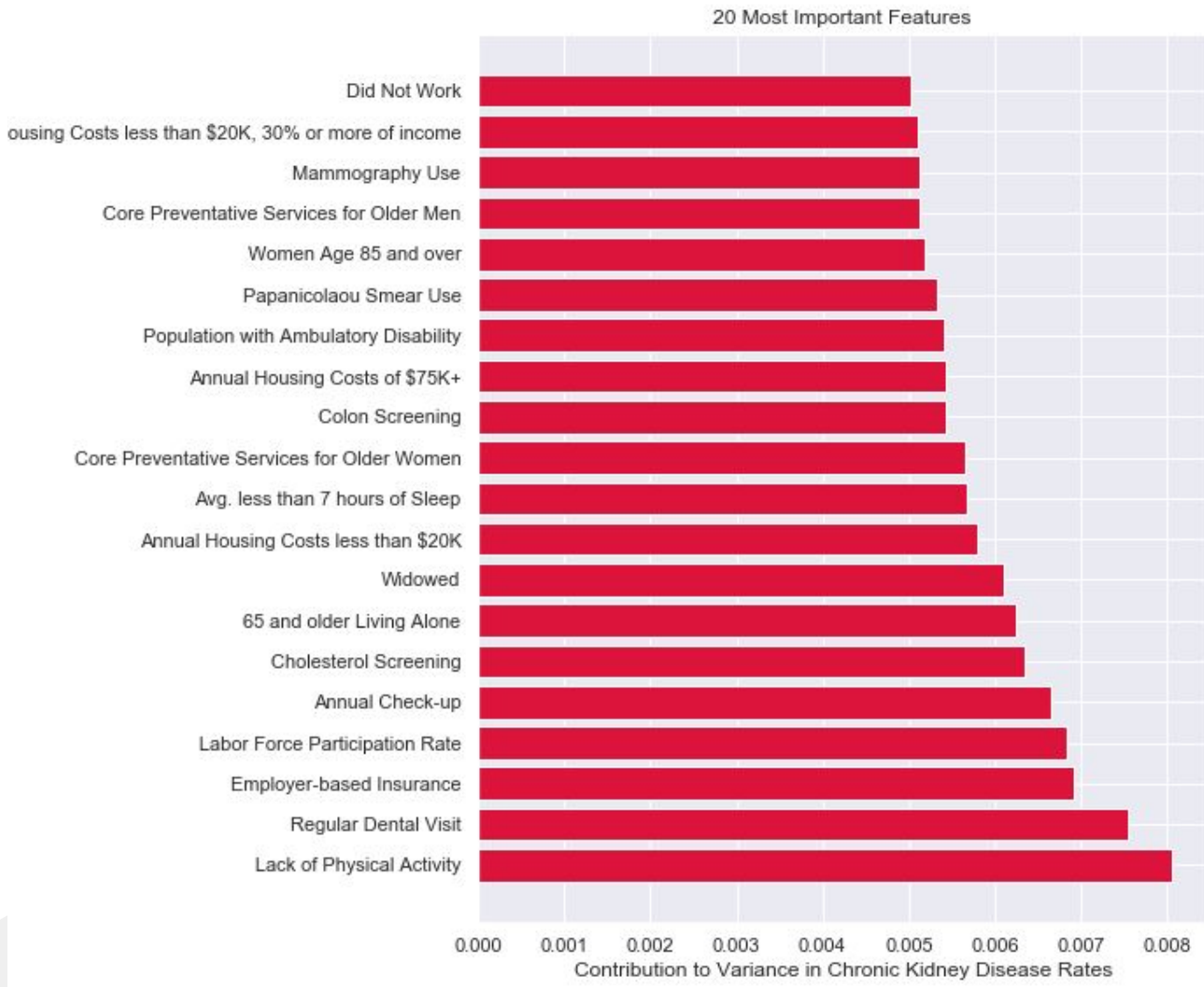
Data Cleaning: American Community Survey

- Dropped two empty and two redundant columns
- Converted negative numbers to NaN
- Dropped all columns missing more than 80% of data
- Imputed missing values with variable mean

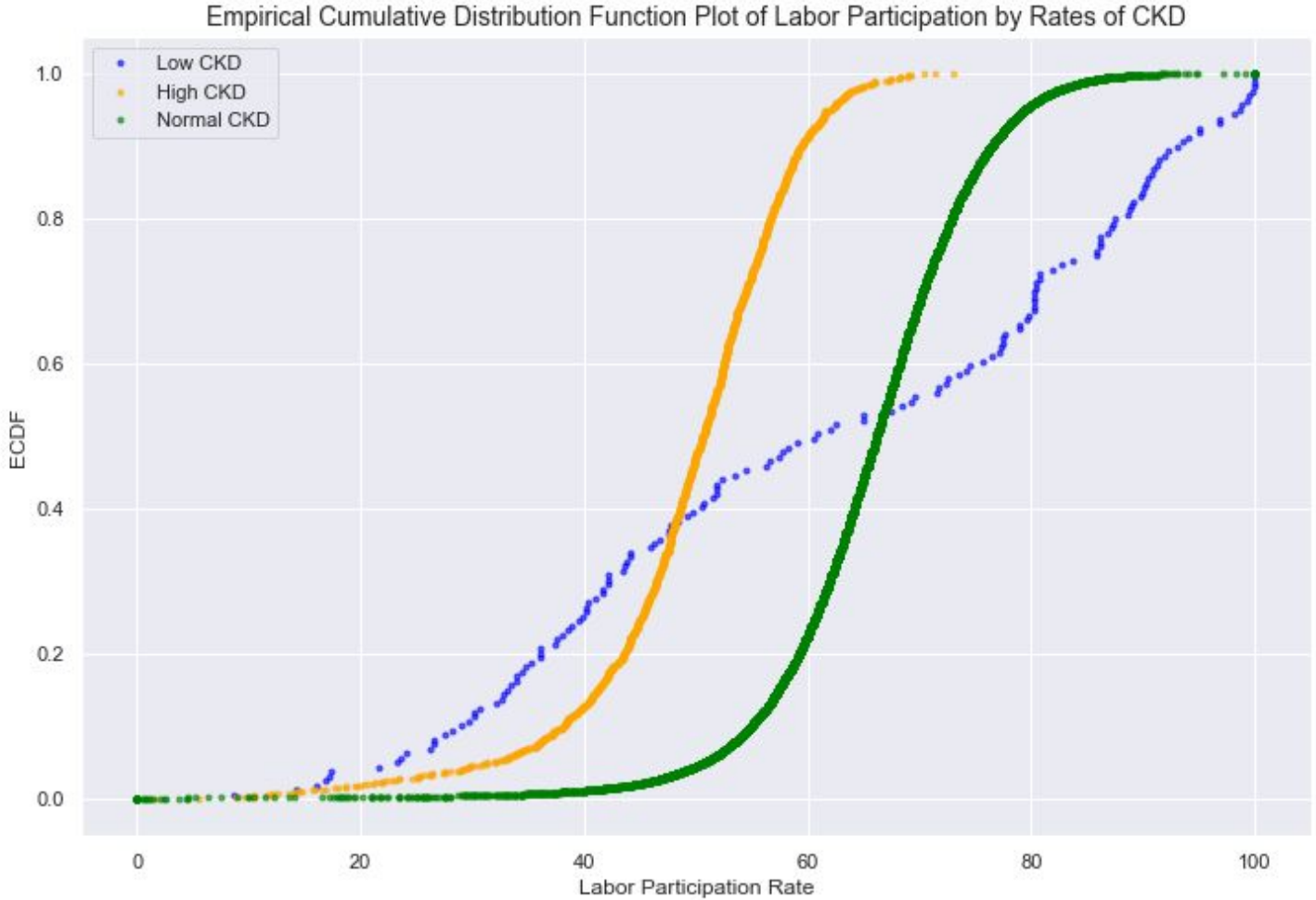
- Combined datasets on Tract IDs
  - Dropped all duplicate rows
- Final dataset:
  - 27,408 observations and 235 variables

Exploratory Data Analysis

Extracting Important Features for Exploration



Labor Force Participation Rate and Chronic Kidney Disease

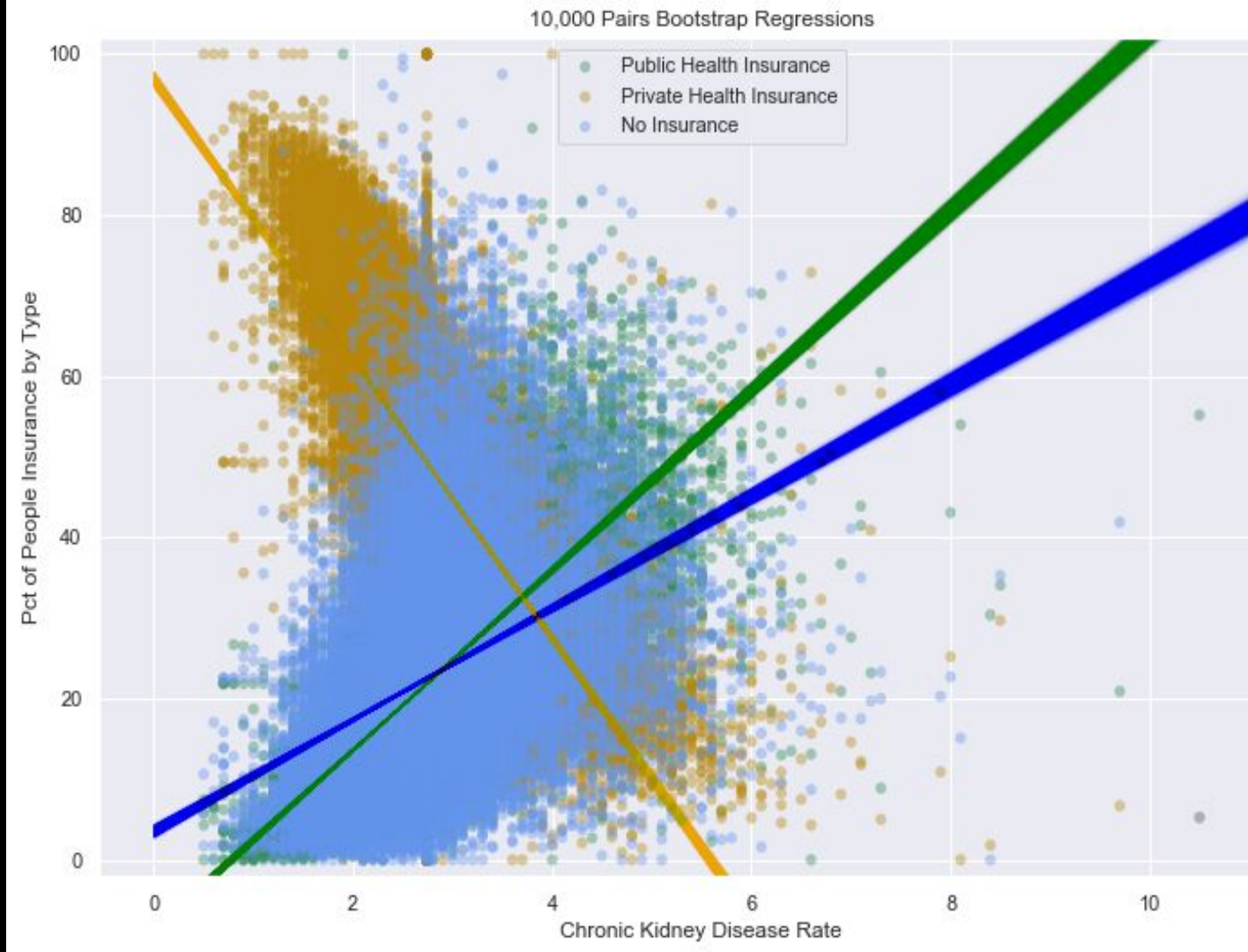


7.1 - 17.7% difference in labor force participation rate between high vs low rates of CKD (95% Confidence Interval)

CKD and Types of Insurance

Negative correlation with private health insurance

Positive correlation with public or no insurance



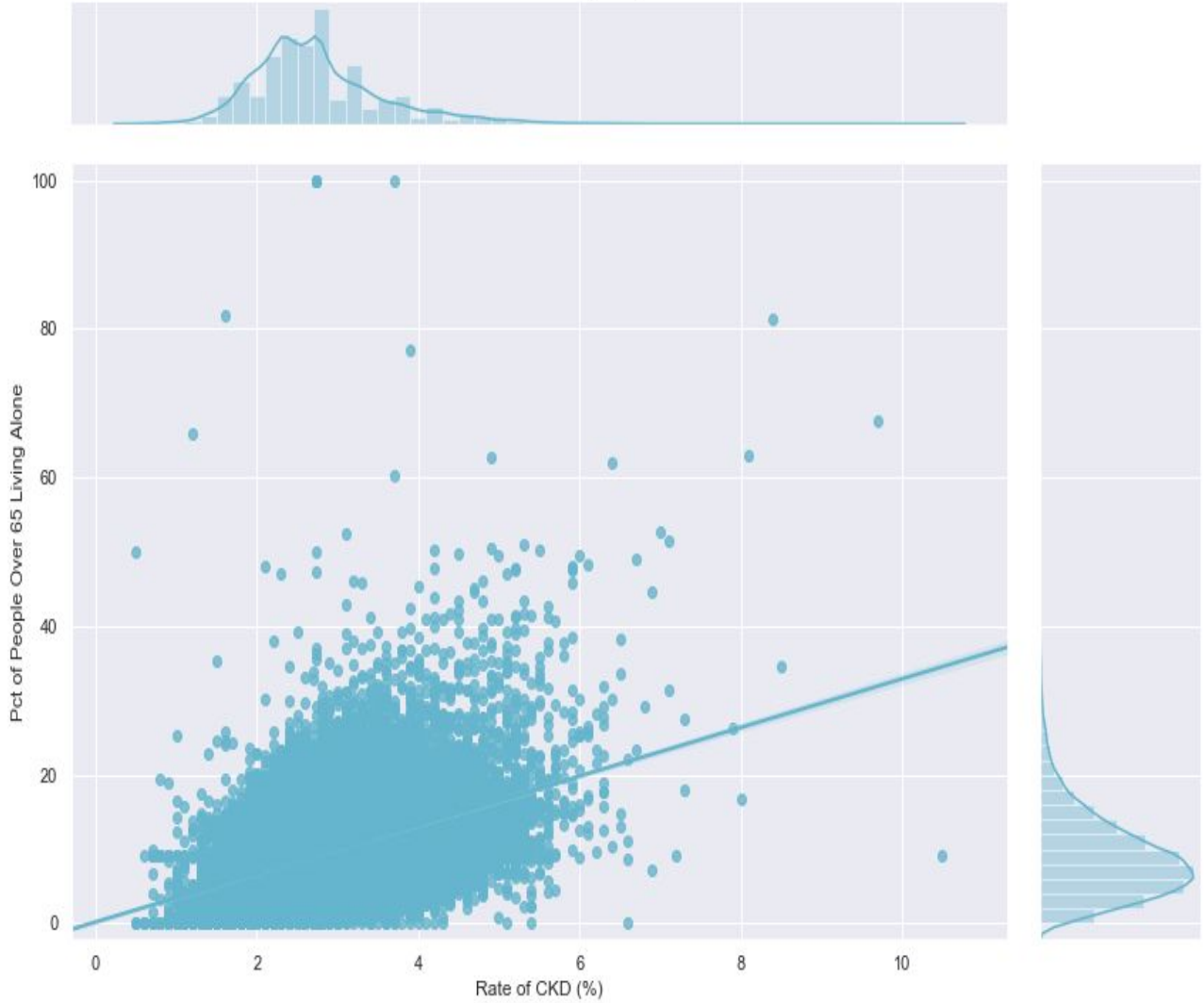
Comparing Rates of CKD Among Groups with Public Insurance vs No Insurance

	Sample Difference	99% Confidence Interval	z-score	p-value
Difference of Mean Slope	4.23	(4.228,4.236)	2751..27	0.0
Difference of Pearson r	0.2814	(0.281, 0.282)	4285.0	0.0

Age, poverty, and health status may explain the increase risk with public insurance

The Role of Seniors Living Alone

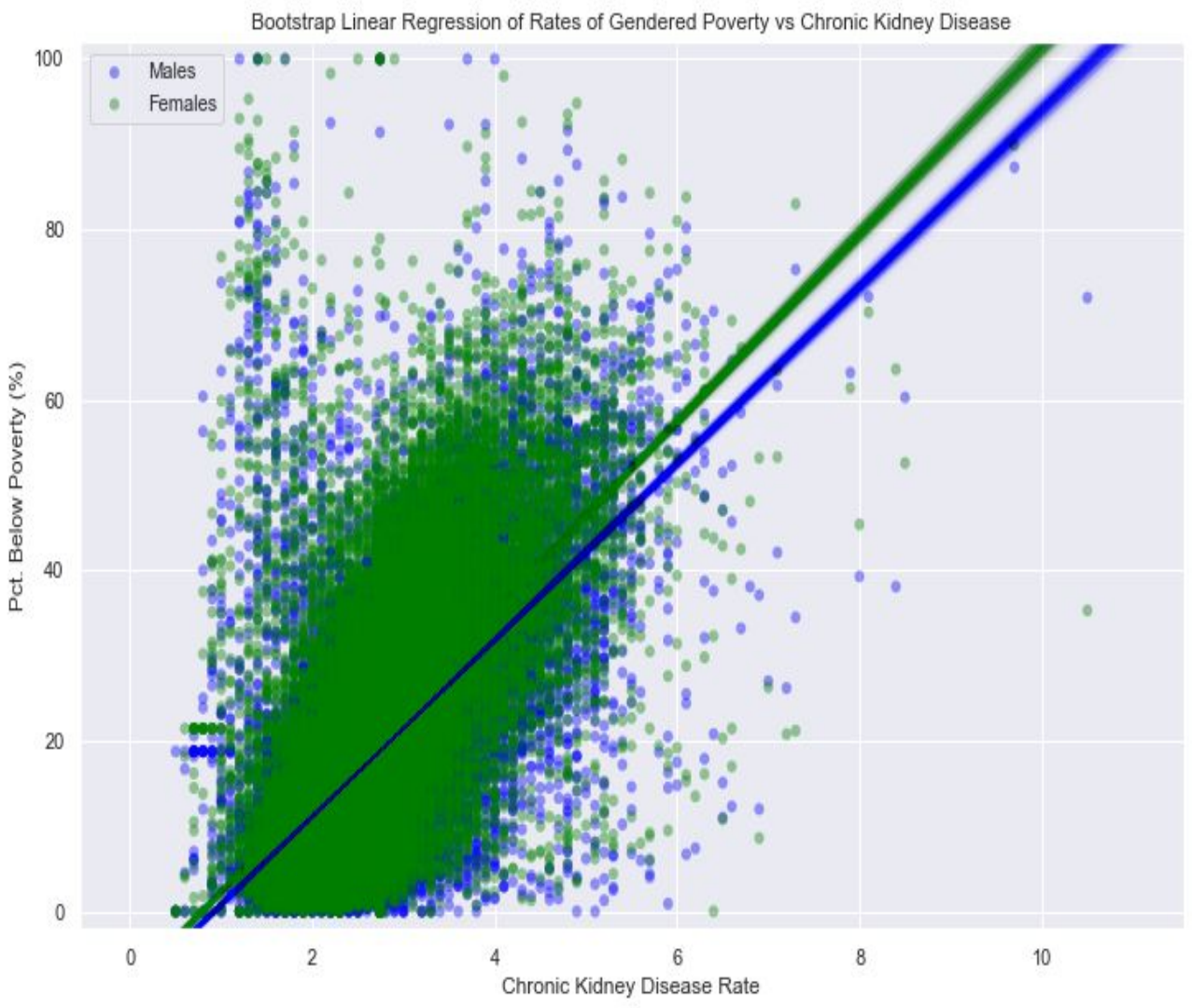
1% increase in CKD for approximately every 3.3% increase in percent of elderly living alone



The Combined Impact of Poverty and Sex

Poverty is highly correlated with CKD

On average, women in poverty are at higher risk of CKD than men in poverty.



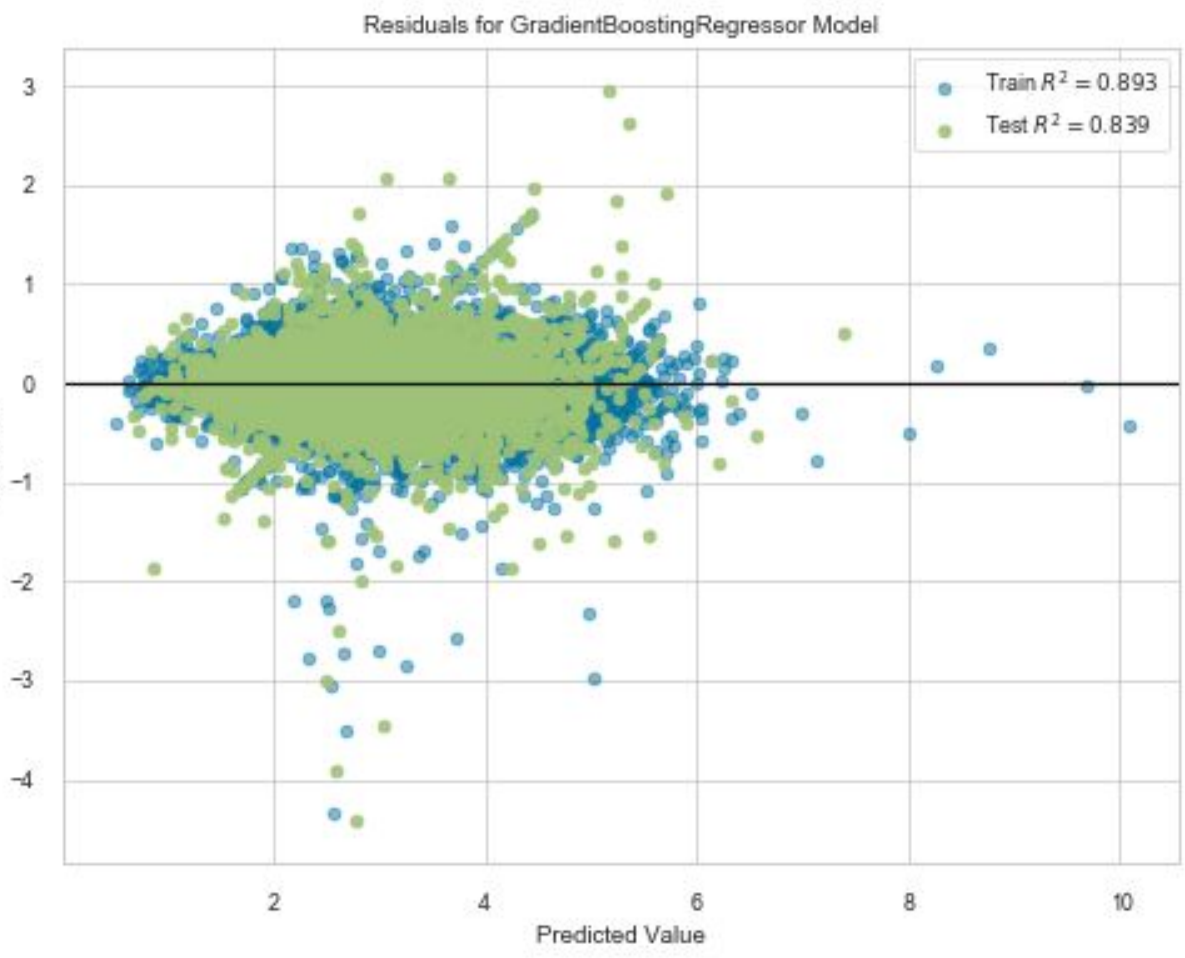
Predictive Modeling

Comparing Algorithms

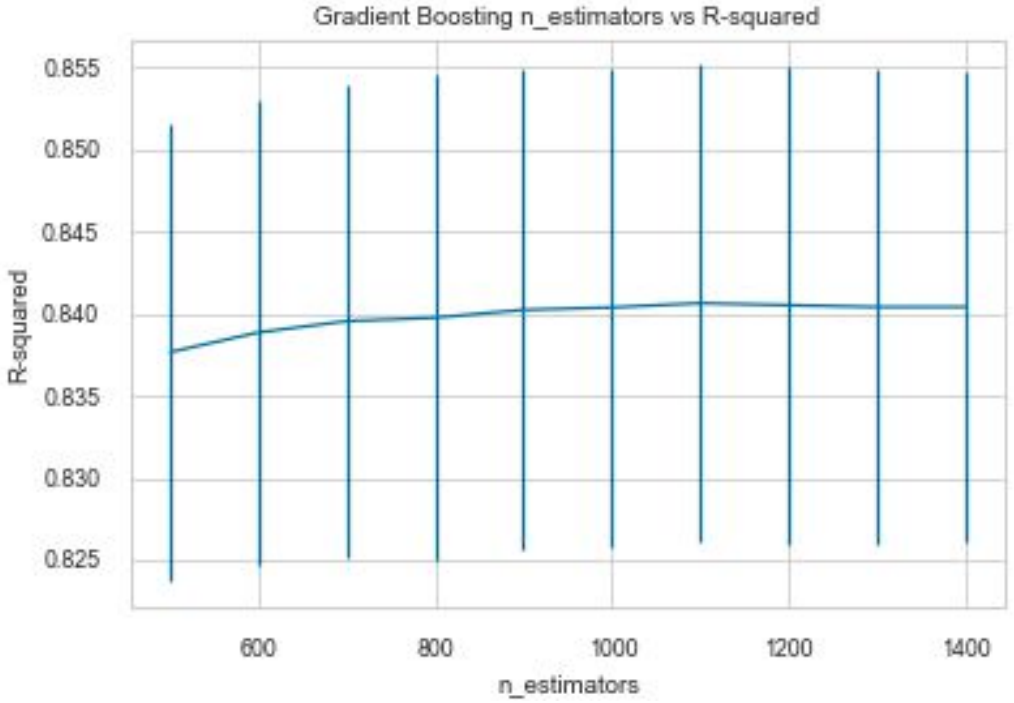
	Adjusted R^2	Mean Squared Error	Root Mean Squared Error
Stochastic Gradient Boosting	0.8394	0.1033	0.3214
Extreme Gradient Boosting	0.8377	0.1037	0.3220
Bayesian Ridge Regression	0.8153	0.1180	0.3435
Ridge Regression	0.8147	0.1184	0.3441
Ordinary Least Squares	0.8145	0.1185	0.3442
Random Forest	0.8093	0.1217	0.3489
AdaBoost	0.5994	0.2558	0.5058

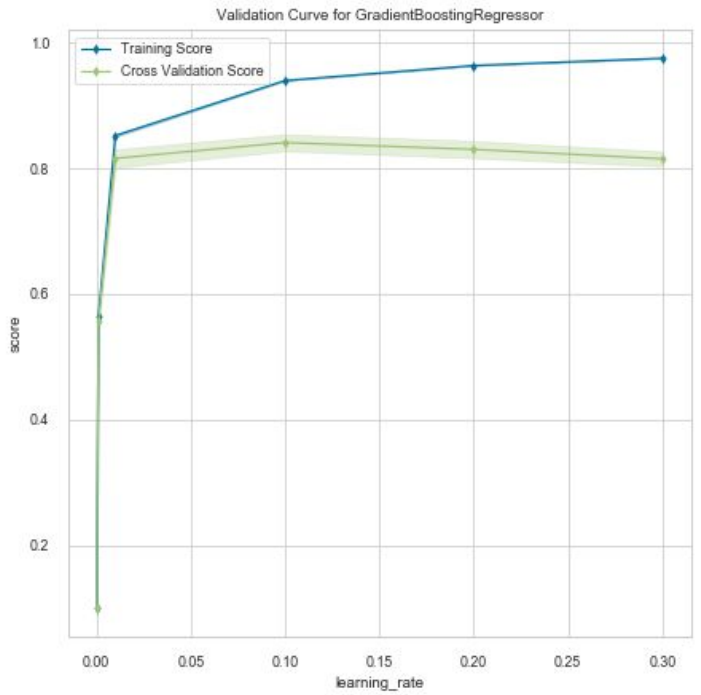
Visualizing Predictive Error

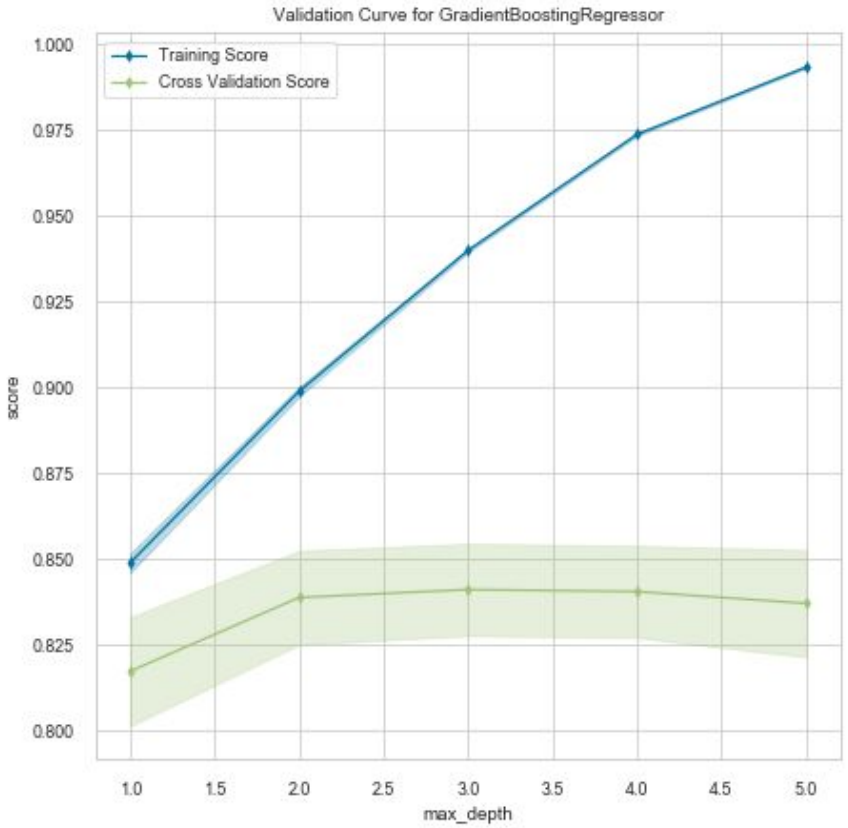
We expect 95% of the predicted rates to be within 0.64% of the actual rate of chronic kidney disease

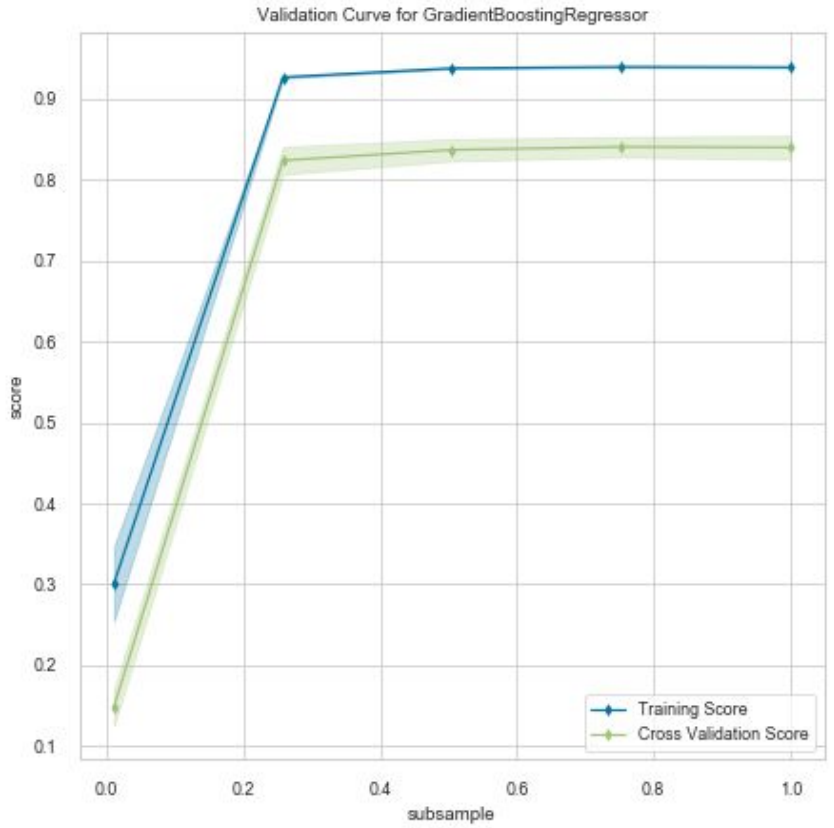


Hyper-parameter Tuning with Validation Curves







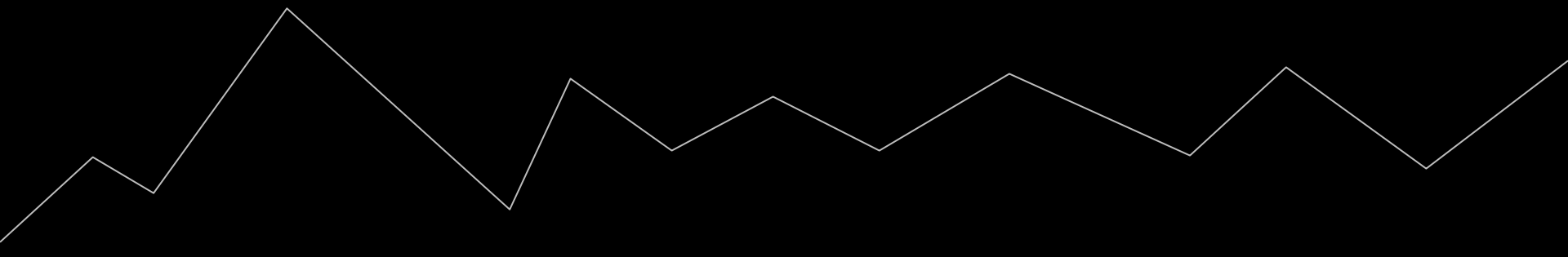


Validation Curves for Max Depth and Subsampling Rate

Improved Model Parameters

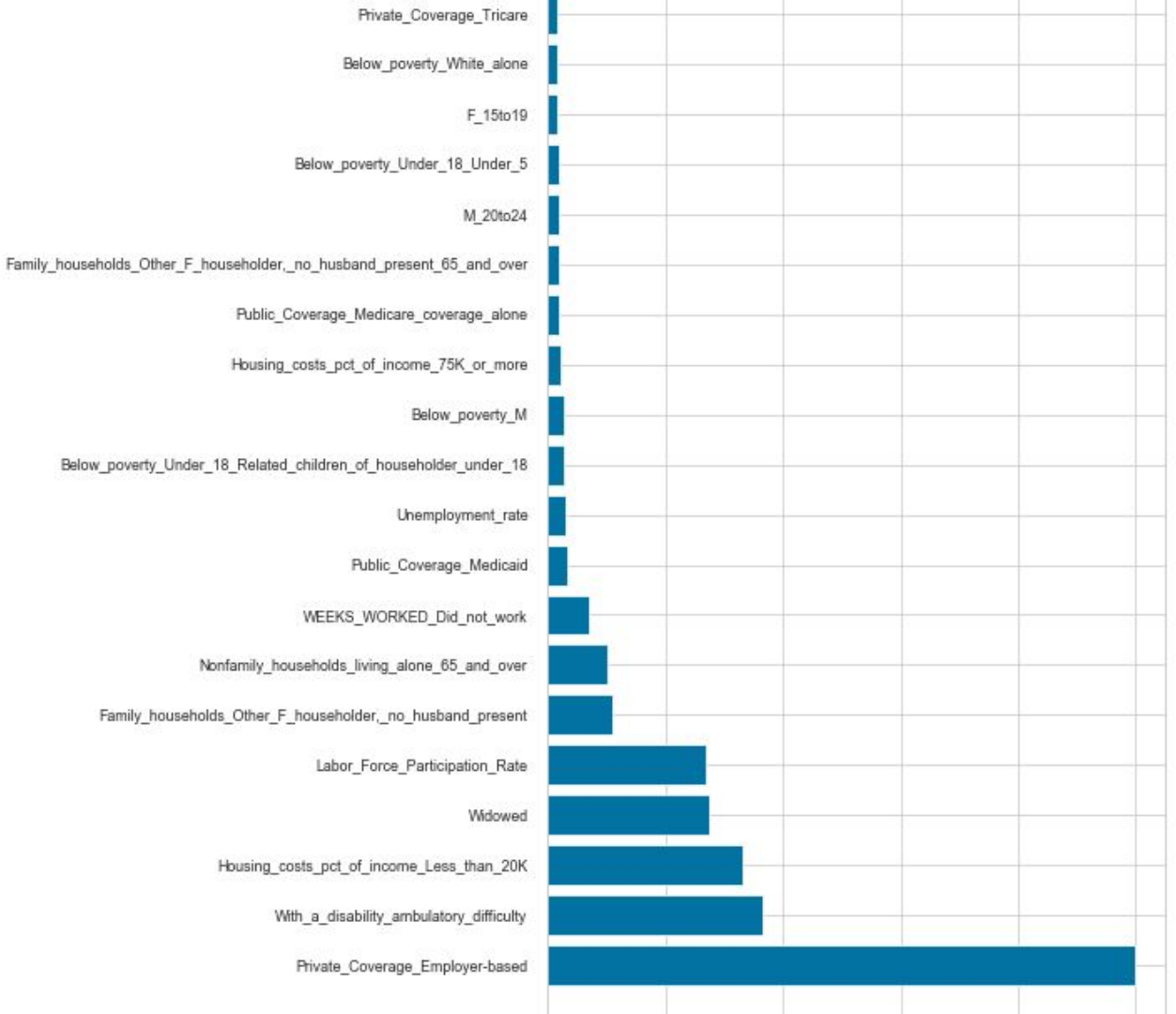
Parameter	Value
Size of Tree (n_estimators)	900
Max Depth of Tree (max_depth)	3
Learning Rate	0.1
Subsample	0.8

Insights



Extracting Important Features

Economic, social, and physical factors appear to play a role in CKD



Economic Factors

- employer-based health care
- labor force participation rate
- low income (relative to housing costs)

Propose: economic development, incentivization, or assistance programs

Disability: Ambulatory Difficulty

- Likely correlated with an economic factors and age component
- Possible interaction effect between economic, social, and physical factors

Propose: provide social assistance to prevention programming in the form of support groups and transportation

Social Factors

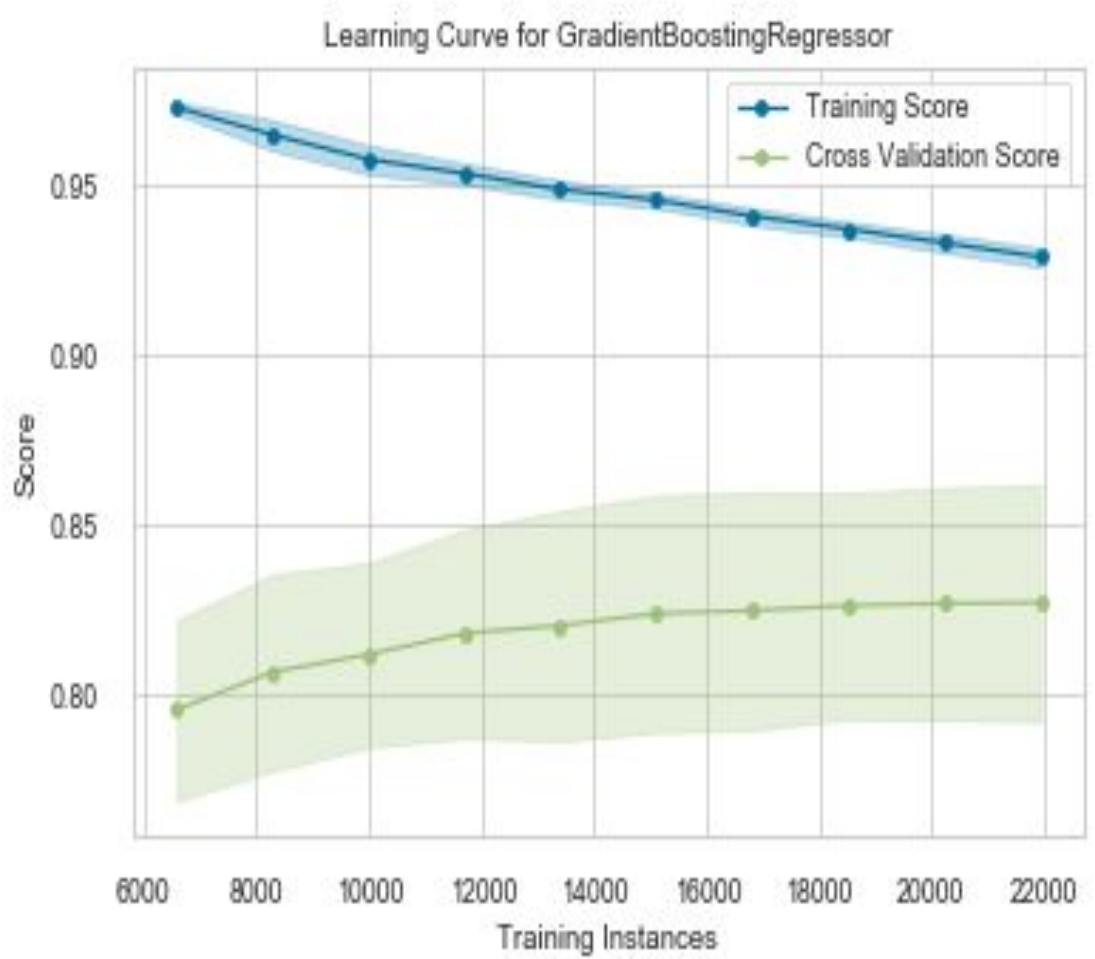
- Widowed
- Single mothers
- Elderly living alone

Propose: researching the effect of social interaction and intervention in the prevention of chronic kidney disease

Moving Forward

Improve the Model

- As usual, more data would likely help
- Reduce overfitting
  - Dimensionality reduction
  - Regularization
  - Hyper-parameter tuning
- Greedy algorithm - validate random seed



Any Questions?

E. Chris Lynch

echrislynch@gmail.com

github.com/TheeChris

linkedin.com/in/echrislynch