E. Chris Lynch

Predicting the hyper-local prevalence of chronic kidney disease

A Novel Approach Using Census Data and Stochastic Gradient Boosting

ABSTRACT

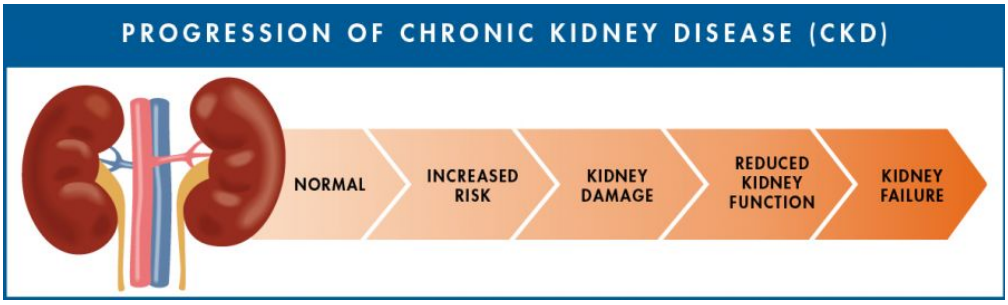
Chronic kidney disease (CKD) has been on the rise in recent years and is a major cause of mortality and health expenditure in the United States. This project uses 235 features extracted from the U.S. Census Bureau to test whether hyper-local rates of CKD can be determined using readily available demographic data. These features include data on age, sex, marital status, disability, employment, profession, household type, housing costs, and type of insurance. Regression and ensemble methods were used to predict rates of chronic kidney disease. Ultimately, gradient boosted decision trees proved to be the best prediction model with a predictive accuracy of 83.94%.

All code available at github.com/TheeChris/springboard/tree/master/predicting_chronic_disease.

INTRODUCTION

The Problem

According to the National Kidney Foundation, 30 million Americans suffer from chronic kidney disease, a life-threatening disease that impairs the kidneys' ability to remove waste products from the body, maintain a proper balance of bodily fluids and chemicals, regulate blood pressure, produce vitamin D, and control the production of red blood cells [1]. Although CKD can be prevented and easily managed if caught in the early stages, it is often the case that it goes unchecked until the disease has progressed to more advanced stages. Recent estimates put the annual cost of treating kidney failure at $31 billion with 89,000 deaths per year [2]. Using machine learning to develop a model that can assist in early detection of CKD can help save money, save lives, and improve the livelihood and productivity of millions of Americans.



Machine learning has been used to detect CKD[3,4] and its progression[5] using labwork data. While this is potentially useful for those patients that find themselves getting bloodwork done at routine checkups or in tandem with other health-related issues, it poses two problems: first, it requires individual patients to have regular lab work done, and second, it does nothing to predict the disease at the population level.

The Centers for Disease Control (CDC) recognize chronic kidney disease as a public health concern that requires population-level surveillance and prevention[6]. However, the hyperlocal surveillance of CKD is difficult and costly. This limits our ability to effectively target campaigns to prevent CKD and its progression. In order to predict hyperlocal disease prevalence, the model presented in this paper uses readily available Census Bureau and CDC data on known and hypothesized risk factors with stochastic gradient descent to better identify census tracts where aggressive public health campaigns and healthcare initiatives can positively affect the early detection and treatment of CKD. Further investigation into this model may also give insight into potential risk factors.

This project was conceived to help city, county, and state public health departments decide how best to use limited resources on public health campaigns regarding the prevention and management of chronic kidney disease. Being able to target high-risk areas and gain insight into potential risk factors could help public health campaigns more effectively prevent chronic kidney disease and its progression. Local governments could also use the data in state and federal grants proposals in order to better make their case for the need for financial assistance. By using the individual and population-level prediction models in tandem, public health efforts may be more effective in terms of health outcomes and spending.

About the Data

This project combines data from the Center for Disease Control's 500 Cities: Local Data for Better Health and the U.S. Census Bureau's American Community Survey 5-year Data (ACS, 2015). The 500 Cities Project contains 28,004 census tract-level observations on CKD prevalence as well as 12 additional health outcomes, 5 unhealthy behaviors, and 9 prevention measures. While models built with these additional features from the 500 Cities Project proved to be more accurate in their predictions, they were ultimately left out of the final model as this data

is not provided on a regular basis. Therefore, features were chosen that could be more easily acquired for future models.

The ACS data provides 16,557 census tract-level feature variables, of which 235 were ultimately selected. The ACS variables cover an array of demographic information in the following categories:

Variable Categories	Sub-categories
Age	By sex
Marriage Status	Heterosexual, homosexual, children, no children
Disability	By type
Employment	Labor force participation, hours worked per week, weeks worked per year
Profession	By sector
Household Type	Family, single, by sex, 65 and over living alone
Housing Cost	Annual and as a percentage of income
Insurance	Coverage by type, No coverage

Data Acquisition

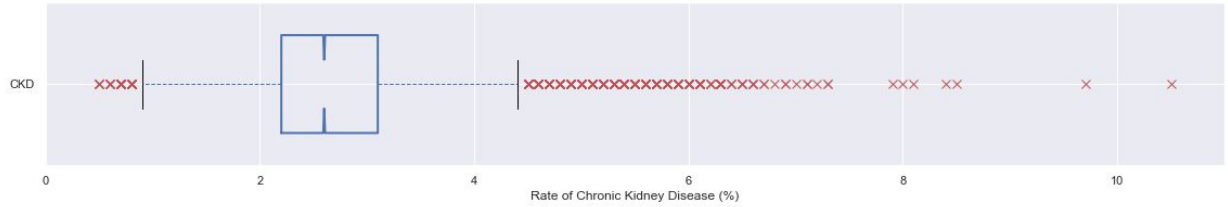
The CDC dataset was downloaded from Data.gov (https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-b32fd) on August 11, 2018.

The Census Bureau dataset was created by pulling 2015 American Community Survey 5-year Data from the U.S. Census Bureau API (https://www.census.gov/data/developers/data-sets/acs-5year.2015.html). Since this project involved tract-level data from all 50 states, it required 50 separate API calls. The API requires the desired variables to be explicitly called, so a dataframe was first created from the variable JSON file (https://api.census.gov/data/2015/acs/acs5/subject/variables.json), which includes all 16,557 features. The unwanted variables were systematically dropped from the dataframe. The variable ID column was then converted to a list and used in the API call. All 50 datasets were then merged into one dataframe, on which all further data wrangling was performed.

Data Pre-Processing

The CDC data was paired down to include only tract-level data. It was then pivoted so that each row represents a census tract and each column a variable. All chronic disease variables other than chronic kidney disease were dropped from the dataset. This was done to ensure that the final model could be easily reproducible for more than just 2015 data.

The data set was then merged with the population count, state, city, and geolocation data from the CDC data. Since the ACS dataset uses Tract FIPS numbers as unique identifiers, these numbers were extracted from the CDC's UniqueID column and saved as TractID (the same variable name given in the ACS dataset). Most of the data were converted to numeric data types, with state and city converted to categorical features. Outliers were kept since they seemed to correlate with actual data. There were only a few variable missing data and at most they were missing 1.97%, but any missing data were imputed using the mean value of the relative variable.

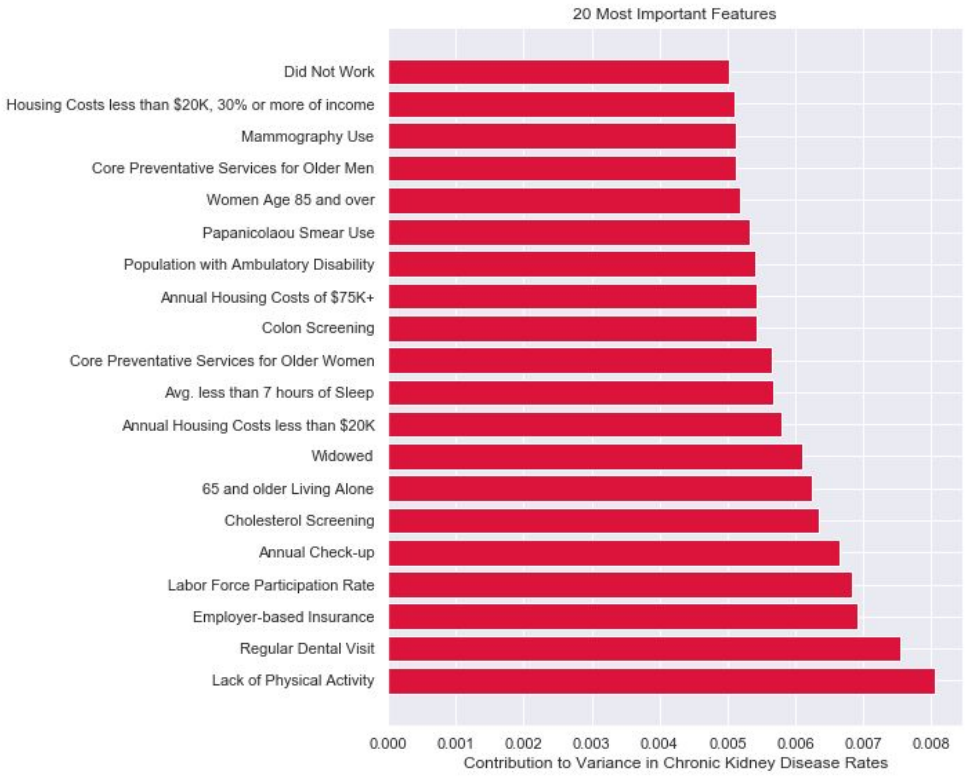


The ACS data was first converted to numerical data types. Two variables pertaining to the sexuality of unmarried couples were found to be completely empty and therefore dropped. Two additional variables (owner and renter-occupied housing units) were also dropped as they were found to be unscaled and repeated data. Missing values, which were marked with negative values, were converted to NaN. Columns missing more than 80% of their data were dropped. This brought the number of variable down to 237, and the remaining missing values were imputed with the variable mean. Despite known outliers, imputing with the variable median, rather than the mean, did not seem to affect the model output. The column names were then changed to be more descriptive (and then truncated to conserve space).

The CDC and ACS dataframes were merged and duplicate rows were dropped. This brought the final shape of the data to 27,408 observations and 235 variables.

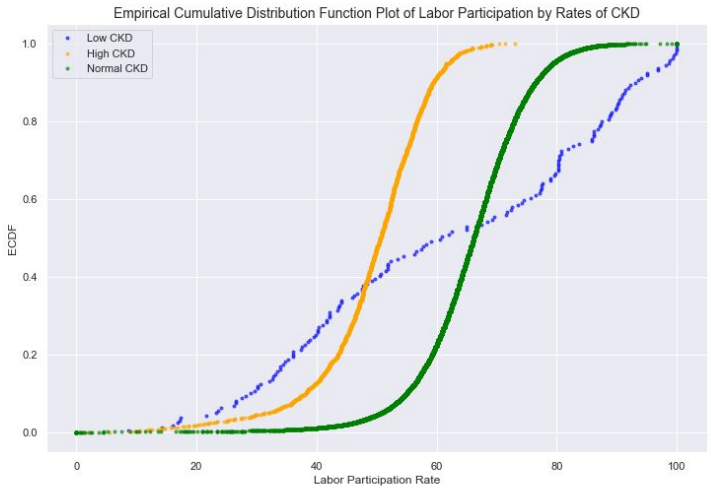
EXPLORATORY DATA ANALYSIS

First, a few different types of important features are identified using an extra trees classifier. This classifier fit 75 random decision trees (with a max depth of 20) on various sub-samples of the data in order to predict the features' contributions to the variance in rates of CKD. If you run this classifier with the other chronic diseases from the CDC dataset included, the top 10 features are always other chronic health diseases (with the top 3 always being stroke, diabetes, and coronary heart disease). Preventative services and unhealthy behavior data from the CDC dataset were left in to determine potentially useful campaign strategies.



Without the other chronic disease data, we see a more diverse set of features with unhealthy behaviors and preventative measures topping the list. While this data is also not regularly available, it may provide some useful information on how to lower future rates of CKD. We can see that campaigns focused on increasing physical activity and annual check-ups would likely be key to reducing rates of CKD. The fact that 'regular dental visits' appears so highly may be due to the fact that those who go in for regular dental visits are also taking other preventative measures. However, it may be worth doing further research into the connection between preventative dental health and lower rates of chronic kidney disease.

Two factors that we will examine further are the type of insurance and the labor force participation rate (both of which appear in the top 4 features). Another pattern of note from the feature importance ranking is the number of female-based variables that appear. This may be due to the fact that women suffer from chronic kidney disease at significantly higher rates than men[7].

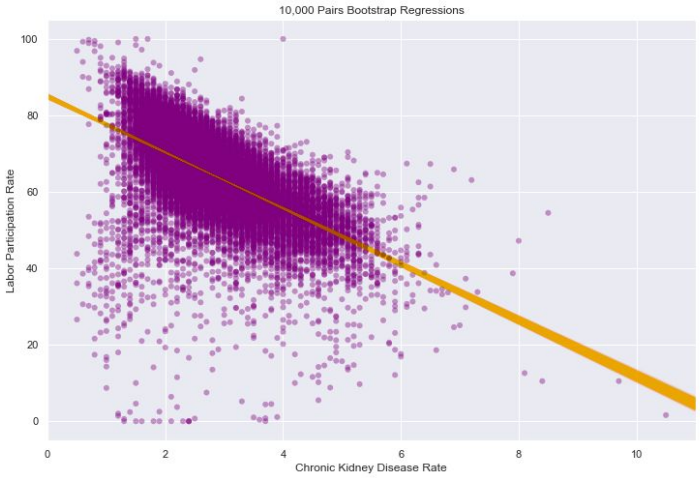


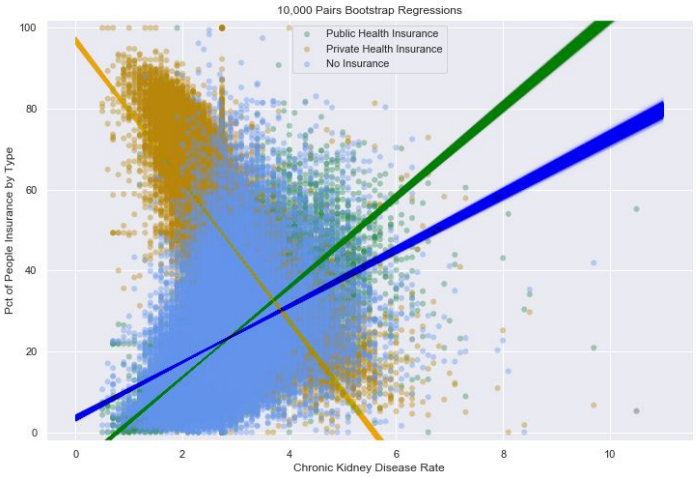
Whether the Labor Force Participation Rate had a significant impact on rates of CKD was examined by splitting the tracts into three groups of high, low, and normal

rates of CKD (high and low were defined as being 2 standard deviations away from the mean). From the ECDF we can see that there are not a lot of census tracts with low rates of CKD (only 159 of 27,408) and the low rates do not appear to be normally distributed (in fact, they seem to be bimodal). However, the distribution does appear to be quite separate and descriptive statistics back this up, although there are large degrees of standard error.

	Mean CKD Rate	Mean Labor Participation Rate	Standard Error of Labor Participation
Low Rates of CKD	0.939%	61.62%	25.18%
High Rates of CKD	4.917%	49.32%	9.65%

With our sample difference of means of 12.30 ($p = 0.00$) and a 99% confidence interval indicating that the true difference of means is between 7.084 and 17.705. This means we expect to find a difference of 7.1% to 17.7% in the labor force participation rate between groups of high versus low rates of CKD. Therefore, we reject the null hypothesis and find that there is likely a difference in mean labor participation rates between groups of high and low CKD. A bootstrap analysis only further solidifies this assumption. This implies that low rates of labor participation can be correlated with higher rates of CKD. This relationship is most likely due to the fact that employment allows for access to healthcare, healthy food options, and other preventative care. Since labor force participation rate showed up as an important feature, but the total hours worked and income levels did not, we may reason that simply having a job, regardless of hours or pay, may reduce the risk of chronic kidney disease. Therefore, campaigns to boost employment rates may be effective in decreasing rates of CKD. Even part-time work may prove to be an important first step in disease prevention.



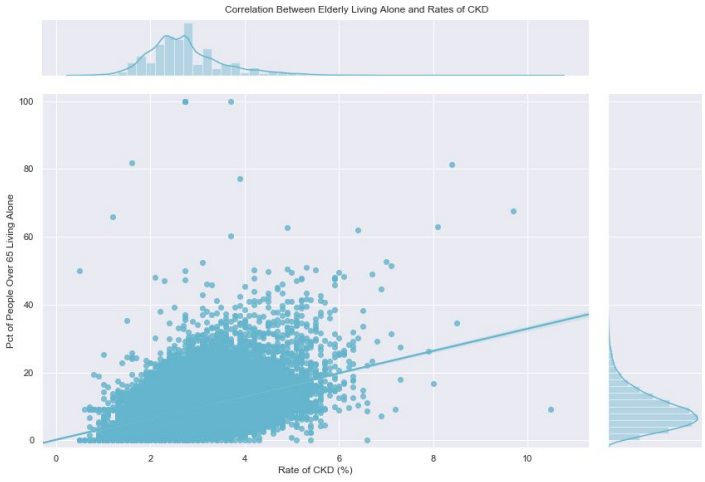


Next, we examine whether the type of insurance coverage (public, private, or no insurance) may have an effect on rates of CKD. Examining a scatter plot of the data with 10,000 bootstrapped slopes, we can see that private health insurance is very strongly negatively correlated with rates of CKD and nearly orthogonal to the slope of public health insurance.

Interestingly, public health insurance appears to be more positively correlated with CKD than no health insurance. This is corroborated by a bootstrap analysis of the difference of means and difference of mean slopes between these two groups.

	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

The difference between the slope of public and no insurance is likely due to the fact that older people are more likely to have public insurance (Medicare) than no insurance as compared to younger populations, and since age is strongly correlated with rates of CKD, there is a potential collinear effect. In addition, people with stage 5 kidney disease are eligible for Medicaid. A bootstrap analysis of private health insurance shows a 99% confidence interval for the slope between -17.78 and -17.13, indicating a strong negative correlation. We, therefore, find evidence to believe that the type of insurance may have an effect on rates of chronic kidney disease. It makes intuitive sense that the lack of health insurance would lead to higher rates of CKD. However, age alone does not seem to account for the fact that areas with high rates of public insurance appear worse in terms of CKD than those with no insurance. It may be worth investigating a cost-benefit analysis of offering public insurance to those that were at higher risk of chronic kidney disease instead of waiting until end-stage kidney failure in order to curb the negative trend with public health insurance.

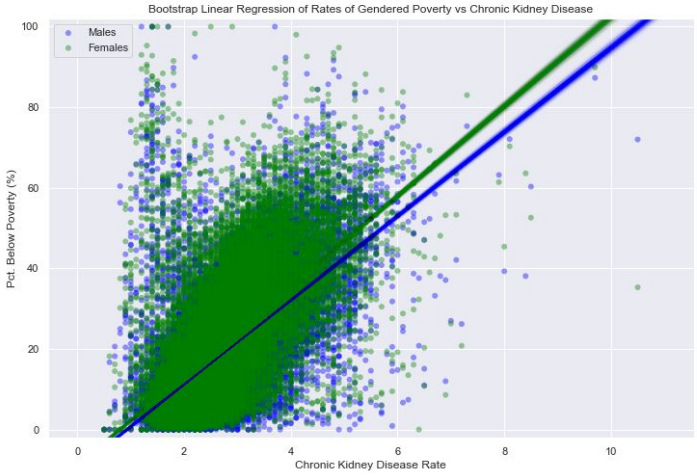


The third variable examined is the percentage of people over the age of 65 who are living alone. Examining the scatterplot with the regression line, we can see there appears to be a positive correlation. The hypothesis was first tested using the sample data and then 10,000 bootstrapped samples. The results are shown in the table below. Since the data appears to be skewed right for both variables, the Spearman ρ was also calculated (0.4033) and found to be slightly lower than the Pearson r, but with a complimentary p-value of 0.00. Since our mean slope and correlation coefficient both remain above 0 in the 99% confidence interval, we reject the null hypothesis. This implies that a campaign to target elderly people who are living alone may help to curb chronic kidney disease. It is difficult to know exactly what it is about living alone that appears to increase the risk of CKD,

but it could be several factors: assistance in getting to or remembering doctors appointments, social pressure to consume a healthier diet, or having someone with whom to be physically active. A campaign to examine the social aspect of disease prevention among the elderly may prove to be beneficial and cost-effective with volunteer involvement.

Sample Data	
Slope	3.29
Pearson r	0.4459
p-value	0.000
Standard error	0.0405

Bootstrap Samples		Confidence Interval
Mean Slope	3.29	(3.137, 3.454)
Pearson r	0.4459	(0.428, 0.463)
r^2	0.1989	(0.183, 0.215)



In order to examine potential interaction effects between economic factors and sex, we can examine the differences in the rate of CKD between males and females in relation to their respective rate of poverty. The results of the hypothesis testing lead us to assume that gender combined with poverty status may provide more insight into the rate of CKD than poverty alone. In other words, while focusing on reducing poverty (or the impacts of poverty) may have a positive effect on rates of chronic kidney disease, including a component that focuses on the gendered aspects of poverty (childcare, pay inequality, etc) may have an amplified effect.

	99% Confidence Interval	Mean Difference	z-score	p-value
Difference of Mean Slope	(0.6604, 0.6695)	0.6650	386.58	0.0000
Difference of Mean Pearson r	(0.0096 0.0101)	0.0098	110.38	0.0000

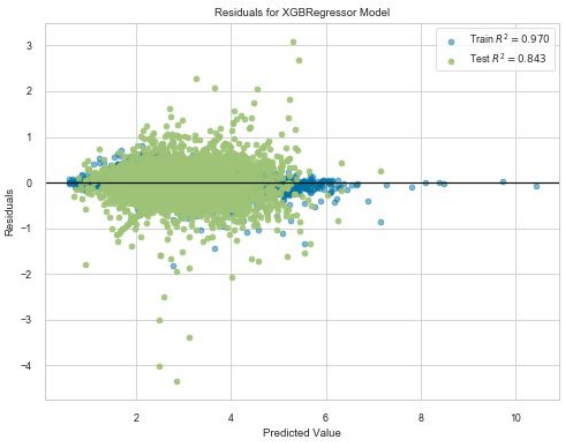
PREDICTIVE MODELING

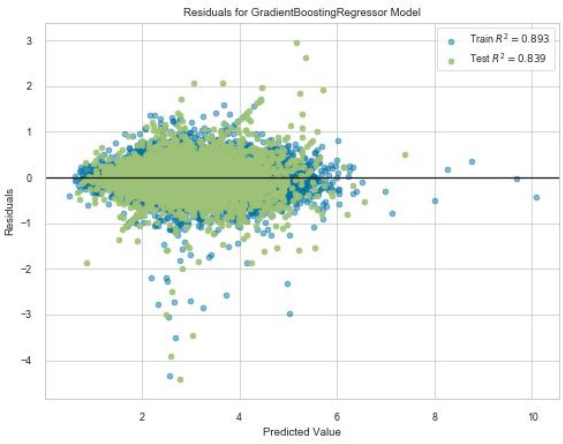
To predict the rate of chronic kidney disease (CKD) among adults, a series of linear regression and ensemble methods were evaluated (see table below). To assess generalizability, the dataset was split into a training and test set. The adjusted R^2 value was calculated to determine the

amount of variance in CKD explained by the model while taking the number of features into account. Mean squared error and the corresponding root mean squared error were calculated to measure the predictive accuracy and gain insight into the standard deviation of each model.

	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

Gradient Boosted Decision Trees appear to produce the best fit models, with stochastic gradient boosting showing a slight improvement in prediction accuracy (and a decrease in overfitting) compared to XGBoost. It should be noted that the standard R^2 score of the XGBoost model is higher than the stochastic gradient boosting. However, when we use the adjusted R^2 to penalize features that do not add information to the model, we see that XGBoost's R^2 score was being artificially inflated due to the sheer number of features. The residual plots provide a visual sense of how close the model's predictions were to the actual values. The root mean squared error (RMSE) tells us that we expect 95% of the predicted rates to be within 0.64 of the actual rate of chronic kidney disease (in other words, just over half a percent off of the actual value).

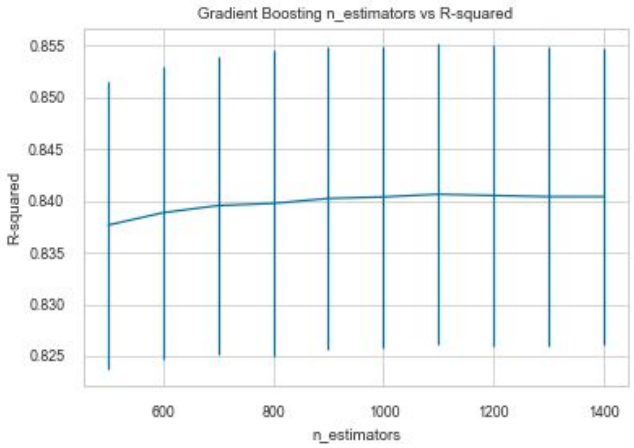


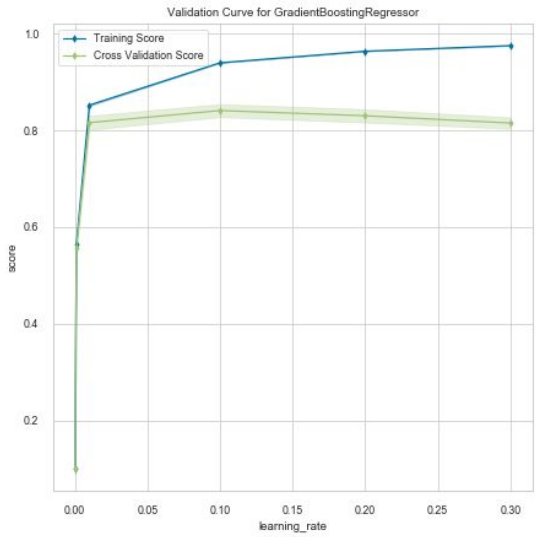


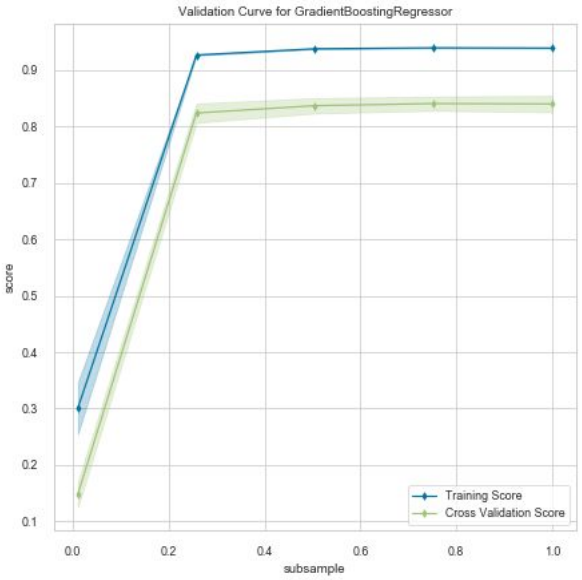
Improving the Model

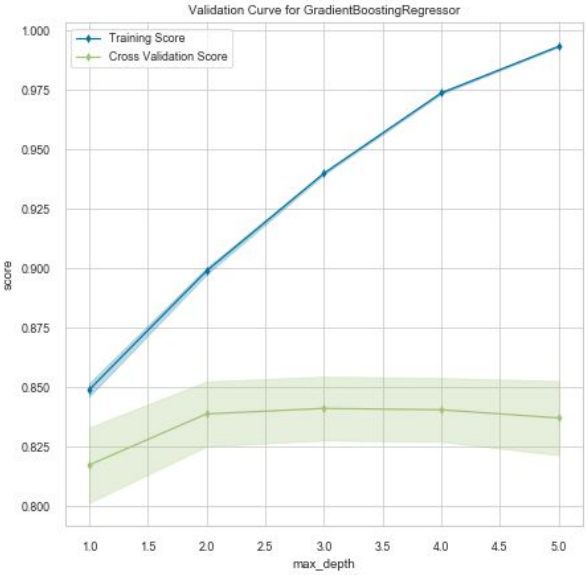
To improve the stochastic gradient boosting model, four hyperparameters were chosen for tuning due to their effect on reducing overfitting: the size of the tree (n_estimators), the maximum depth of the tree (max_depth), the fraction of samples to be used for fitting (subsample), and shrinkage

(learning_rate). Grid search using 5-fold cross-validation was used to determine the best performance at 900 estimators. Setting the number of estimators higher produced greater overfitting. Validation curves were visualized to determine the remaining hyperparameters.







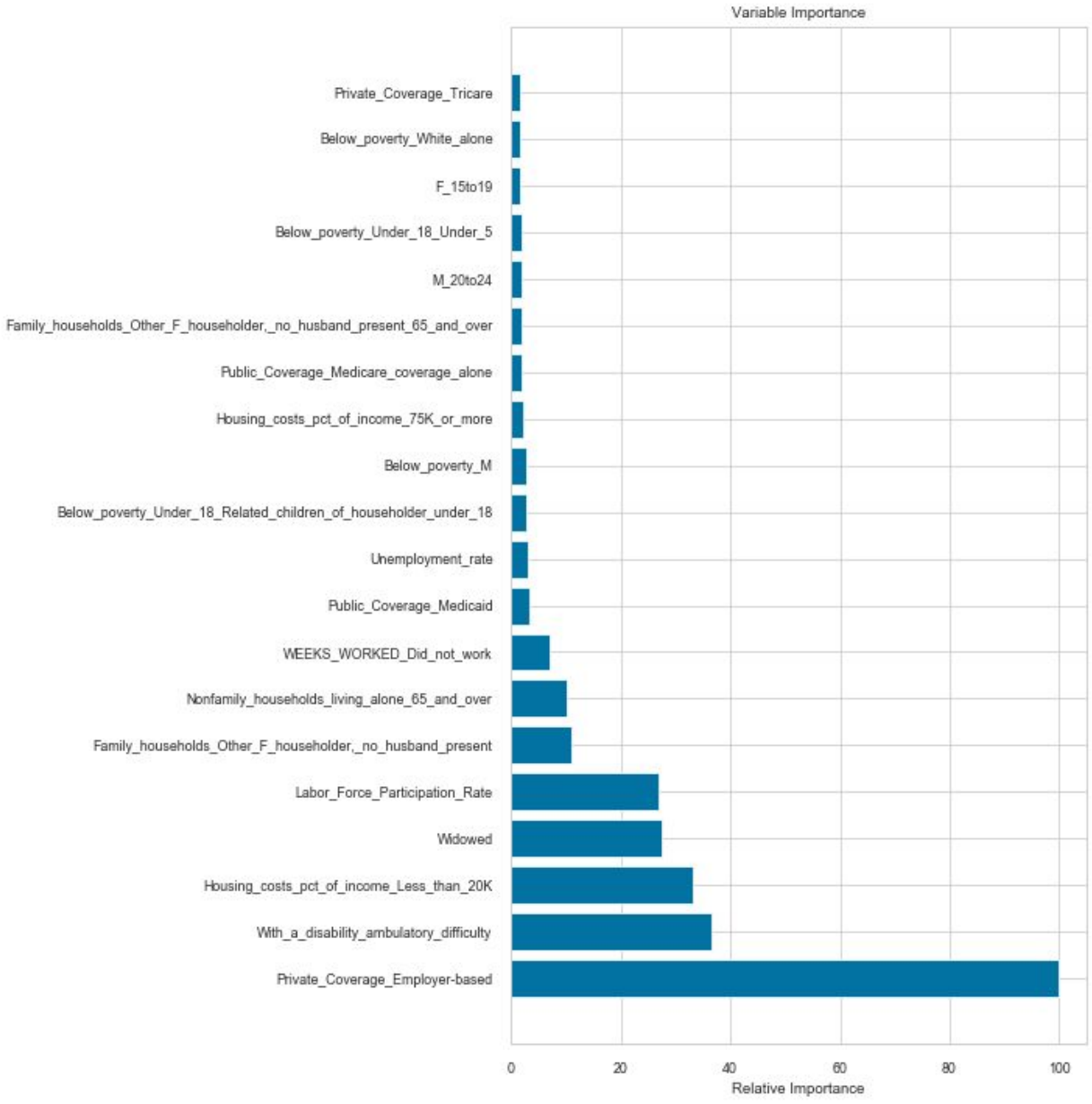


Max depth and shrinkage appeared to perform best when set to their respective defaults. The final model utilized the following hyperparameters:

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

ACTIONABLE INSIGHTS

While the predictive model can be useful, interpreting the model in order to gain actionable insights may prove to be more useful when designing public health campaigns. To do this, we extract the 20 most important features and rank them relative to all of the features.



First, we see that the model appears to mimic a Pareto distribution, where a small number of features account for most of the variation in rates of chronic kidney disease. This could prove to be useful as it will allow future efforts to focus on a few key areas to obtain the greatest effect. As we saw during our exploratory analysis, features such as employer-based health care, labor

force participation rate, and low income (relative to housing costs) near the top of this list. These can be grouped together as economic factors, indicating that areas with high predicted rates of CKD may benefit from economic development, incentivization, or assistance programs.

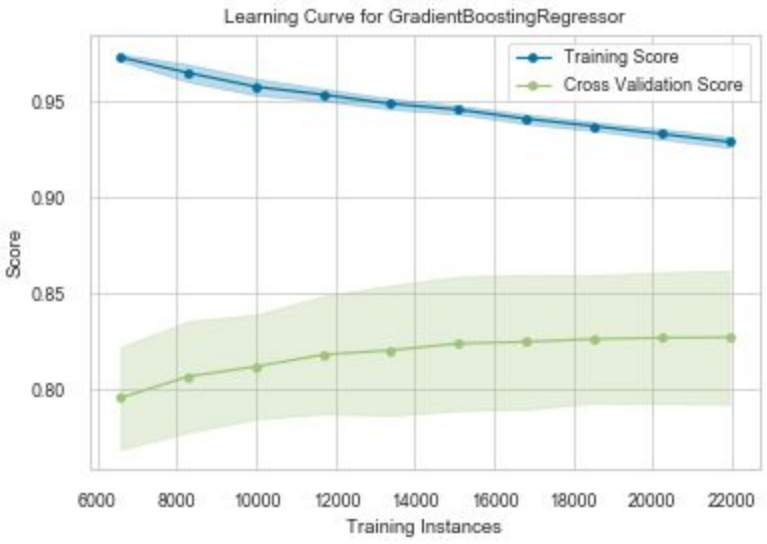
Those living with an ambulatory difficulty appear as the second most important feature. Although this is likely correlated with an economic and age component, there may also be an interaction effect between economic, social, and physical factors. This group may benefit from campaigns that combine these factors into programs that provide social assistance to prevention programming in the form of support groups and transportation.

Interestingly, there are three features (widowed, single mothers, and the elderly living alone) that indicate a social component to CKD prevention. Although single mothers are also likely correlated with an economic component, it may still be worth researching the effect of social interaction and intervention in the prevention of chronic kidney disease.

In summary, it appears that there are three barriers to overcome in working to curb and ultimately reverse the growing rates of chronic kidney disease. First, the economic barriers to healthcare and prevention services. Second, the physical barrier to health faced by those with disabilities, especially the elderly. And third, the social component of public health that encourages preventative behaviors. Using a predictive model to focus resources, we may see improved outcomes in overcoming these barriers to health and reducing the rates of chronic kidney disease.

Moving Forward

While the model appears to provide decent predictive power, there is still plenty of room to improve. We can see in the learning curve that although the model improvement seems to have slowed down, more training data could prove to be useful in improving the model accuracy.



There also remains a fair amount of overfitting that could potentially be reduced to help the model generalize to new data. This could be achieved through dimensionality reduction, regularization, or additional hyperparameter tuning. Additionally, because stochastic gradient boosting is a greedy algorithm, different seeds will lead to different results (and varying feature importance).

Before embarking on expensive campaigns, it may be best to compute models with various seeds in order to generate mean outputs. This could prevent the possibility of chasing a hypothesis that was an anomaly produced by one random seed.

Note: one additional stochastic gradient boosting model was built with a different seed. While this did produce a slightly lower adjusted R^2, the mean squared error stayed the same. Additionally, with some slight variation in order, the top 5 features (and 8 of the top 10) remained the same. However, it may be beneficial to continue testing with additional seeds and validation curves.

REFERENCES

1. National Kidney Foundation. About Chronic Kidney Disease. National Kidney Foundation A to Z Health Guide. https://www.kidney.org/atoz/content/about-chronic-kidney-disease. Last accessed October 30, 2018

2. National Kidney Foundation. End Stage Renal Disease in the United States. National Kidney Foundation website. https://www.kidney.org/news/newsroom/factsheets/End-Stage-Renal-Disease-in-the-US. Last accessed October 30, 2018.

3. Misir, R., Mitra, M., & Samanta, R. K. (2017). A Reduced Set of Features for Chronic Kidney Disease Prediction. Journal of pathology informatics, 8, 24. doi:10.4103/jpi.jpi_88_16

4. Soltanpour Gharibdousti, Maryam & Azimi, Kamran & Hathikal, Saraswathi & H Won, Dae. (2017). Prediction of Chronic Kidney Disease Using Data Mining Techniques.

5. Tangri N, Stevens LA, Griffith J, et al. A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure. JAMA. 2011;305(15):1553–1559. doi:10.1001/jama.2011.451

6. Centers for Disease Control and Prevention. About the CKD Initiative. Chronic Kidney Disease Initiative. https://www.cdc.gov/kidneydisease/about-the-ckd-initiative.html. Last accessed November 1, 2018.

7. Centers for Disease Control and Prevention. Age-adjusted prevalence of CKD Stages 1-4 by Gender 1999-2012. Chronic Kidney Disease (CKD) Surveillance Project website. https://nccd.cdc.gov . Last accessed November 6, 2018.