Using Social Determinants of Health to Assess Risk of Heart Disease

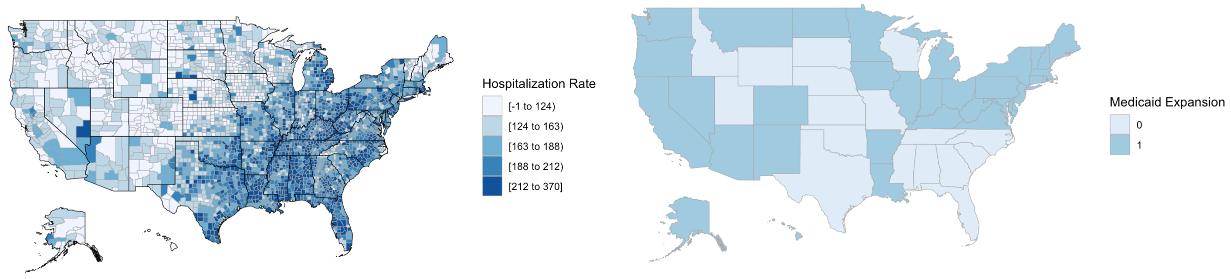
Jack Collison
jack10

Athreya Steiger
athreyas

Department of Computer Science
Stanford University
Autumn, 2019

1 Introduction

Social determinants of health are recognized as deeply predictive of an individual's risk of illness. Particularly, these determinants have been shown to have very high predictive power in assessing an individual's likelihood of developing heart disease over the course of their lifetime. Our project uses machine learning techniques to predict county-level incidence of heart disease, determine the causal impact of the expansion of Medicaid in the Affordable Care Act, and assess optimal policies for its subsequent expansion. We hope that the results of this project will allow us to better understand the principle factors that are relevant to heart disease and the policies that are effective in improving health outcomes. This will facilitate conversations between patients and physicians about the risk of developing heart disease, thereby allowing earlier adoption of preventative lifestyle shifts that can mitigate this risk and provide insight to influence legislative action.



(a) Hospitalizations by County 2014-2016

(b) Medicaid Expansion in 2013

Figure 1: Maps of Response and Treatment

2 Literature Review

There have been many previous studies — at the NIH, CDC, and various medical schools — on heart disease. However, the vast majority of these studies have been purely descriptive and have not implemented novel methods such as the ones utilized in this class. Additionally, many studies examining the impact of social determinants of health on heart disease focus on a particular determinant, without examining how multiple determinants intersect to influence heart disease risk.

For instance, one epidemiological study done by Schulz et al. found that socioeconomic status is largely predictive of one's risk of outcomes of cardiovascular disease [1]. Namely, individuals from lower socioeconomic status groups have poorer outcomes, including higher hospitalizations, higher rates of co-morbidity, and higher mortality rate. Additionally, the researchers in Schulz et al. found differential outcomes based on sex, employment status, psychological factors, and environmental factors controlling for SES [1]. These factors are further influenced by dietary habits, including the types of foods an individual eats, which largely vary based on one's culture. [5]. While providing valuable information, this study does not account for the summative effect of multiple risk factors and demographic characteristics to predict outcomes. This is an opportunity to apply artificial intelligence to extend the findings of this study and to thereby see how all of these factors intersect to predict the outcomes of heart disease.

Indeed, to generate a useful predictor of one's risk of heart disease, it is important to account for all of an individual's demographic and socioeconomic characteristics and risk factors. A CDC study found that race and ethnicity influence other the likelihood that an individual has other risk factors. One key finding from this study is that one's race and ethnicity influence the likelihood that an individual has more than one risk factor for cardiovascular disease. This reaffirms the importance of examining the intersecting influence of all predictive factors of one's risk of heart disease [3].

Nonetheless, it's also important to consider that some factors may outweigh the risk conferred by multiple other factors. For instance, one study by researchers at UC Davis Health found that lower socioeconomic status was strongly linked with heart disease regardless of any improvements in other factors [4]. This is largely indicative that a machine learning algorithm to learn weights of each of these risk factors would be an excellent approach to model risk of heart disease dependent on the presence or absence of such risk factors.

Additionally, while these demographic and socioeconomic characteristics and risk factors can be predictive of one's cardiac outcomes, policy changes that expand or contract access to healthcare (e.g. through implementation of a more robust government-funded insurance program) affect access to care, which largely influences outcomes. Namely, the Medicaid expansion in 2013 expanded the reach of government-funded insurance across the country in states that implemented these expansions. However, the downstream effects of Medicaid expansion on outcomes of individual conditions is yet to be seen [9].

Once the impact of Medicaid expansion is known, this would be informative about next steps for further expansion of Medicaid into states that have not yet adopted it. Researchers Athey et al. have developed algorithms that quantify the causal effect of a treatment and learn the optimal policies for implementing the treatment while mitigating the biases of an observational study [7]. Thereby, algorithms such as these can be used to determine the causal effect of and the best policy for implementing Medicaid, which would allow advocates for Medicaid expansion to specifically target the states and their individual counties that would benefit most from this expansion.

3 Data Exploration

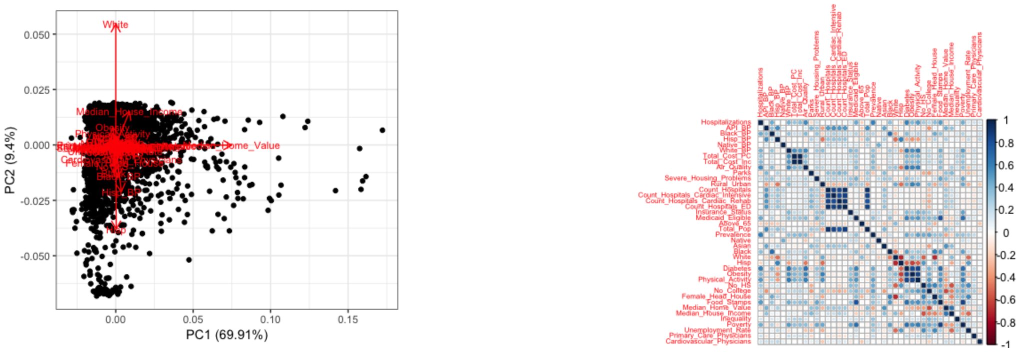
3.1 Data

Our study leverages county-level reports of hospitalizations related to heart disease derived from reports to the CDC. More specifically, for each county in the United States, our response variable operationalizes the risk of heart disease as the number of hospitalizations per one thousand individuals. Our independent variables include county-level demographic characteristics (race, socioeconomic status, education levels, etc.) and risk factors (diabetes and obesity status, median income, pollution index, etc.). We supplement this with state-level data on Medicaid expansion in order to assess average treatment effects and optimal policies.

A limitation of this dataset is that we have county-level data rather than individual-level data. This limits the model's ability to attribute county-wide hospitalizations to individual social determinants. However, given the demographic diversity and range of heart disease incidence seen across counties, the model is able to parse out the most predictive risk factors.

3.2 Exploration

As an initial step in our data exploration, we generated and visualized a covariance matrix to determine which variables are most strongly correlated with one another. This is important as multicollinearity could affect the results of any model built. While this does not affect the predictive power of the model, it could skew the values of the coefficients as well as confounding the statistical significance of any given model parameter. As our project seeks to determine which variables are important in predicting county-level incidence of heart disease hospitalizations, this confounding could lead us to believe that certain variables are statistically significant, although they do not hold predictive power (and vice versa). We find that a number of variables have strong correlations. Thus, we use regularization to mitigate the problem of multicollinearity.



(a) Principal Components Analysis

(b) Correlation Coefficients

Figure 2: Data Exploration: PCA and Correlations

Next, we performed Principle Components Analysis on our data to identify which variables explain the most variance in our dataset. To compute the first principle component, we must maximize the variance with a given weight vector:

$$\mathbf{w}_{(1)} = \underset{||\mathbf{w}||=1}{\operatorname{argmax}} \, \mathbf{w}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\mathbf{w}$$

We find that the median value of homes, median income, education, access to cardiovascular physicians, and blood pressure are among the most important factors in determining the variance in the data; thus, they should be included in any model fit. We conclude by creating histograms of the variables that cause the highest variance to see if there are any highly skewed values. Detailed graphs and tables can be found in the Appendix.

3.3 Baseline & Oracle

In this context, it is very difficult to devise an oracle; there are so many covariates that even human assessment of the risk for heart disease based on social determinants will not be perfect. Further, there is so much heterogeneity across counties that the task becomes even more difficult.

As a baseline model, however, we implement a multiple linear regression that includes every variable in the dataset. This methodology provides a very basic understanding of which variables might be important and provide insight for future iterations of models. It was found that the test R^2 value from the initial regression was 0.6494.

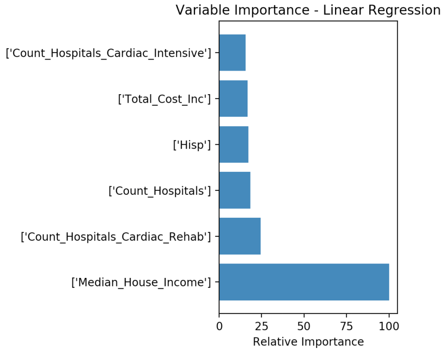


Figure 3: Baseline Model - 6 Most Predictive Features

As can be seen from Figure 4 above, the 6 most important features in the baseline linear model were — in decreasing order of importance — median home income, count of cardiac rehabilitation hospitals in the county, count of total hospitals in the county, percentage of Hispanic individuals in the county, total cost of insurance, and count of cardiac intensive hospitals in the county.

3.4 Problem Framework

The problem is defined as a two-pronged approach. First, we will develop predictive methodology to determine the risk of heart disease in a given county and which variables are instrumental in quantifying this risk. Analysis such as penalized regressions and gradient boosted trees will allow us to determine which counties may have higher future risk of hospitalizations due to heart disease. Next, we supplement our model with a treatment variable in the expansion of Medicaid in the Affordable Care Act in order to determine the causal effect that Medicaid expansion has had on hospitalizations due to heart disease and evaluate optimal policies. This is done with propensity score matching and random forests. Each of these prongs will be described more in depth in the following sections.

3.4.1 Example Input/Output Behavior

We include the input and output behavior of our model below. As we have many input variables, some have been omitted for readability. A comprehensive list of all input variables can be found in the Appendix.

Predictive Model Input: ‘Santa Clara’: {‘API_BP’: 22.2, . . . , ‘Total_Cost_PC’: 17494, ‘Total_Cost_Inc’: 12339, ‘Air_Quality’: 8.9, . . . , ‘Asian’: 35.2, . . . , ‘Count_Hospitals_Cardiac_Intensive’: 6, ‘Above_65’: 12.6, ‘Total_Pop’: 1911226, ‘Medicaid_Expansion’: 1}

Predictive Model Output: {‘Hospitalizations’: 121.3}

Causal Inference Input: The input for the causal inference portion of our model consists of a list of all counties in the United States and their respective features. Each individual county input follows the same format as that of the Predictive Model Input shown above and is supplemented by a boolean treatment.

Treatment: {‘State_Expanded_Medicaid’: 1}

Causal Inference Output: {‘Treatment_Effect’: -8.45 ‘Lower_Bound’: -10.78, ‘Upper_Bound’: -6.12}

Policy Optimization Input: The input for the policy optimization problem is the same as the causal inference part of the model. The output, however, is slightly different as it recommends thresholds of our input on which we would decide to implement a policy in the form of a decision tree.

Policy Optimization Output: {‘Expansion_Recommendation’: 1}

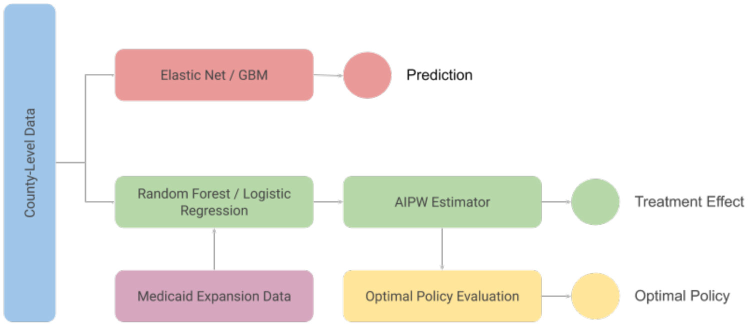


Figure 4: Model Framework

4 Models & Algorithms

4.1 Prediction

4.1.1 Regularized Linear Regression

As an initial model, we extend the vanilla linear regression by adding an elastic net regularization. This combination of L_1 and L_2 penalties will help facilitate the selection of the most predictive determinants (through generation of sparse solutions) while maintaining much of the robustness from a pure LASSO- or Ridge-penalized model. The formalized algorithm for this elastic net linear regression is as follows:

$$f(\mathbf{X}) = \beta^{\intercal}\mathbf{X} + \lambda\left(\alpha\sum_{i=1}^{p}|\beta_i| + (1-\alpha)\sum_{i=1}^{p}\beta_i^2\right), \qquad \alpha \in [0,1]$$

4.1.2 Gradient Boosted Trees

As a next step, we use gradient boosted trees to learn parameters rather than updating weights via a penalized linear regression. The intuition behind the choice of gradient boosted trees is that it is likely that the parameter space is not linearly separated. Decision trees are very flexible in that the data doesn't have have to have a linear relationship for the model to perform well. Further, using boosted gradient trees will help the model "learn" from the easy data points first and then move onto more difficult observations. The decision trees will output some value $y \in \mathbb{R}$ as a prediction of the number of hospitalizations due to heart disease. The gradient boosted trees are computed as follows:

```
#Initialize null model
f_hat = NULL
#Initialize residuals to actual values
r = [y[i] for i in range(len(y))]
#Fit B shrunken trees
for b in range(B):
    #Fit decision tree with d splits on training data and residuals
    f_hat_b = fit_tree(d, x, r)
    #Update model with shrunken decision tree
    f_hat += lambda * f_hat_b
    #Update residuals
    r -= lambda * dot(f_hat, x)
#Return boosted model
return f_hat
```

4.2 Causal Inference

The final section of the project is causal inference. In particular, we determine the effect that the expansion of Medicaid has had on hospitalizations due to heart disease. However, this is not a randomized control trial; it is likely that states that chose to implement Medicaid expansion policies are inherently different than those that did not. Therefore, we must control for this bias.

Unbiased average treatment effects are measured as the difference between the average among the treated population and the untreated population. Given that we have bias present in the data, this methodology will not provide accurate results. In order to address this issue, we propose propensity score matching and augmented inverse propensity weighting in order to control for the bias. We calculate propensity scores — the probability that a county (conditional on its characteristics) has been treated with Medicaid expansion — with an L_1 penalized logistic regression. The average treatment effect is then calculated by weighting each observation by the inverse of its given propensity score.

$$\hat{\tau} = \mathbb{E}\left[\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right]$$

$$\hat{e} = \frac{\exp[\beta_0 + \sum_{i=1}^{p} \beta_i X_i]}{1 + \exp[\beta_0 + \sum_{i=1}^{p} \beta_i X_i]} + \lambda \left[\sum_{i=1}^{p} |\beta_j|\right]$$

We then use the doubly robust augmented inverse-propensity weighted average treatment effects on the treated (ATT) to determine the causal effects. The propensities are estimated with L_1-penalized logistic regressions and causal forests and the ATT is calculated as follows:

$$\hat{\tau}_{ATT} = \frac{1}{n}\sum_{i=1}^{n} \hat{\Gamma}_i$$

$$\frac{\hat{\Gamma}_i}{n} = \frac{W_i(Y_i - \hat{\mu}_{(0)}(X_i))}{n_1} - \frac{(1-W_i)\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(Y_i - \hat{\mu}_{(0)})}{\sum_{W_i=0}\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}}$$

Double robustness is a nice theoretical quality in an estimator as it is consistent when either (1) the propensity score calculation is well calibrated or (2) the estimation of means (i.e. $\hat{\mu}_{(0)}$ and $\hat{\mu}_{(1)}$) is well calibrated; both are not required. In this way, we are able to control for the bias present in the data.

4.3 Optimal Policies

We use a similar doubly robust methodology to estimate the value of using a given policy in an observational study. The improvement yielded by the policy π over a random policy is estimated as:

$$\hat{A}(\pi) = \frac{1}{n}\sum_{i}\left(2\pi(X_i) - 1\right)\hat{\Gamma}_i$$

where $\hat{\Gamma}_i$ is the doubly robust score of the treatment effect

$$\hat{\Gamma}_i = \hat{\tau}^{(-i)}(X_i) + \frac{W_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left(1 - \hat{e}^{(-i)}(X_i)\right)}\left(Y_i - \hat{\mu}_{W_i}^{(-i)}\right)$$

where $\hat{\tau}(\cdot)$ are the point-wise estimates of the treatment effect, $\hat{e}(\cdot)$ are estimates of the propensity score $P(W_i = 1|X_i)$, and $\hat{\mu}_0, \hat{\mu}_1$ are estimates of $\mathbb{E}[Y|X_i, W_i = 0]$ and $\mathbb{E}[Y|X_i, W_i = 1]$ using out-of-bag estimates from causal and regression forests.

4.3.1 Plug-in Policies

A plug-in policy assigns a treatment to a given county as long as it is above the median out-of-bag treatment effect found in the causal forest. We find that there were no statistically significant results for any plug-in policies.

4.3.2 Learning Optimal Policies

We can learn an optimal policy from a constrained class of policies denoted Π. Using the doubly robust scores from before, we can find a $\pi \in \Pi$ that maximizes the following objective function:

$$\hat{A}(\pi) = \frac{1}{n}\sum_{i}(2\pi(X_i) - 1)\hat{\Gamma}_i$$

This is done by turning the maximization criterion denoted above into a classification problem. Each doubly robust score is decomposed into:

$$\hat{\Gamma}_i = |\hat{\Gamma}_i| \cdot \text{sign}(\hat{\Gamma}_i).$$

This is the same as maximizing the weighted correlation between the (transformed) assignment rule $(2\pi(X_i) - 1) \in \{+1, -1\}$ and the sign of its estimated effect $\text{sign}(\hat{\Gamma}_i) \in \{+1, -1\}$. Counties that respond very strongly to the treatment so that $|\hat{\Gamma}_i| \gg 0$ will receive larger weights learned through optimal regression and classification trees using evolutionary algorithms. The algorithm is roughly as follows:

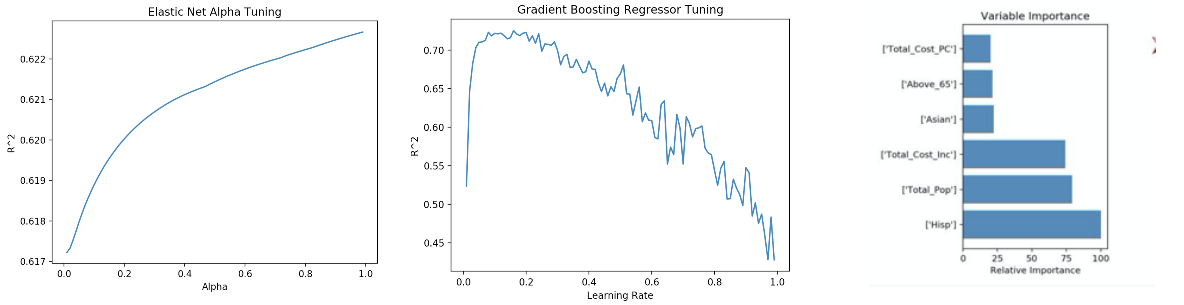
1. Set of trees initialized with random split rules in root nodes
2. Mutation and crossover operators applied to modify structure
3. Apply tests to internal nodes
4. Select best candidate for the next iteration until convergence

This algorithm allows the average quality of the population of trees to increase over time, only terminating with either (1) the quality of the best trees do not improve more than some threshold or (2) the maximum number of iterations is exceeded.

5 Implementation & Results

5.1 Prediction

We implemented a regularized linear regression with an elastic net using `scikit-learn`. We ran this elastic net model with multiple values of α (all 0.01 increments between 0 and 1, inclusive) to tune for an optimal R^2 and plotted (shown below) all resulting (α, R^2) pairs. From this graph, we can see that the optimal α (highest R^2) value is approximately $\alpha = 0.99$, yielding an R^2 of 0.607. This indicates that a LASSO-penalized model most accurately fits our data, although the magnitude of the change in R^2 is minimal.



(a) Elastic Net Regression

(b) Gradient Boosted Trees

(c) Variable Importance: GBR

Figure 5: Hyperparameter Tuning

Next, we implemented a gradient boosted tree model using `scikit-learn`. We increment the learning rate η from zero to one using a step size of 0.01 in order to maximize the training R^2 value. We obtain a maximum training R^2 of 0.8630 with $\eta = 0.16$, a relatively slow learning rate. In order of decreasing importance, the six most important features were the following: ['Hisp']: percentage of people in a given county who identify as Hispanic, ['Total_Pop']: total county population, ['Total_Cost_Inc']: average difference in insurance cost between those with heart disease and those without, ['Asian']: percentage of people in a given county who identify as Asian, ['Above_65']: percentage of people in a county over the age of 65, and ['Total_Cost_PC']: average cost of insurance per capita.

We create a table below comparing the tuned and un-tuned models. We found that the learning-rate tuned gradient boosted trees performed the best on both the training and the test sets. This is indicative that the problem space is highly non-linear.

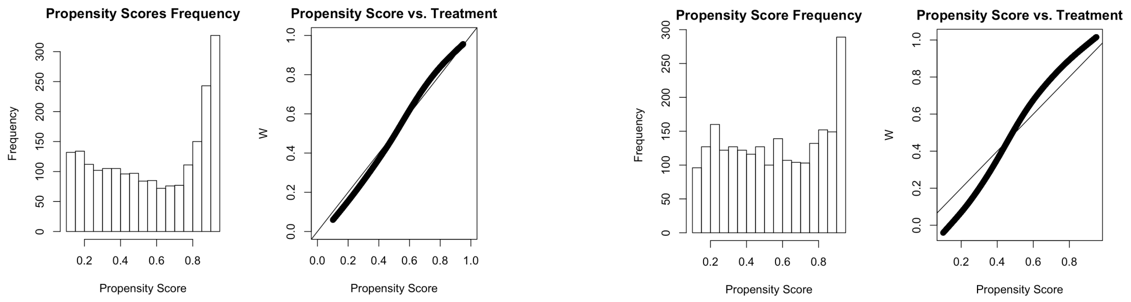
Model	**Test R^2**	**Training R^2**	**Tuned**
Linear Regression	0.6494	0.6638	N/A
Linear Regression (Optimal Elastic Net)	0.6507	0.6600	$\alpha = 0.99$
GBM	0.7313	0.8354	N/A
GBM (Optimal Learning Rate)	0.7467	0.8630	LR = 0.16

Table 1: Model Comparison

5.2 Causal Inference

The next section of this project is causal inference, which was implemented in R using various packages to simplify the optimization of the random forests, penalized logistic regression, and visualization of the analysis. We implement a variety of methods, each with a distinct way of controlling for the bias in the data. First, we use an unbiased estimate, which found a statistically significant decrease in the hospitalization rate in the counties with Medicaid expansion. However, this is not controlling for any bias. The next set of methods each control for the bias.

Next, we use inverse propensity score weighting to control for the likelihood of treatment. We estimate these propensity score first with a penalized logistic regression and then with a random forest. The results are each statistically insignificant and positive with very wide confidence intervals. We find that the calibration of the propensity scores calculated by the penalized logistic regression and random forests are both reasonable, as shown below.



(a) Logistic Regression Propensity Calibration

(b) Random Forest Propensity Calibration

Figure 6: Propensity Score Calibration

The best method for controlling for bias is augmented inverse propensity score weighting as it is doubly robust (the theoretical property described in the Algorithms section above). Although the results are not statistically significant (which we will delve into during our discussion) they show a negative average treatment effect on the treated with much tighter confidence intervals. Therefore, we see a decrease in hospitalizations that is borderline effective. This statistical insignificance could be caused by the lack of time between policy implementation and data collection.

Method	Average Treatment Effect	Lower-bound	Upper-bound
Unbiased	-6.83	-10.34	-3.32
IPW (Logistic)	2.12	-16.38	20.63
IPW (RF)	14.52	1.11	27.92
AIPW (RF, no weighting)	-2.34	-6.22	1.54
AIPW (RF, weighting)	-3.76	-11.19	3.67

Table 2: Average Treatment Effect Comparison

5.3 Optimal Policies

We also use R to implement the policy optimization segment of this project. Various packages allowed us to visualize and simplify some of the algorithms used. First, we fit a causal tree to determine which variables are most instrumental in driving hospitalization rates. In the image below, we can see that splits occur on various variables such as insurance status and costs, median income, housing, and blood pressure.

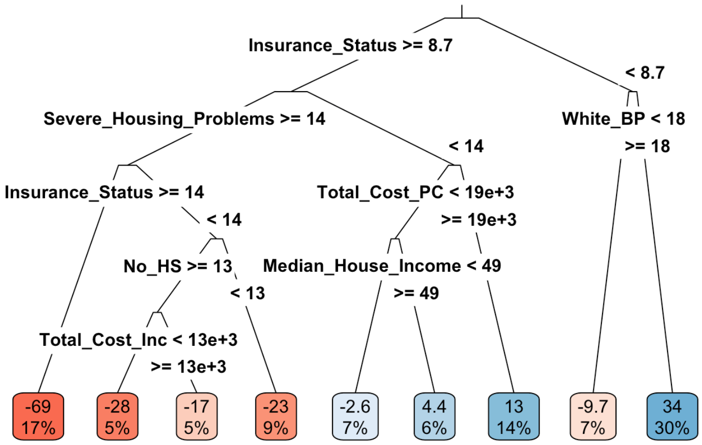


Figure 7: Policy Optimization: Causal Tree

We do not find any statistically significant results from the optimal policy. This could be due to the lack of time between the implementation of Medicaid expansion as the results have not had time to perpetuate.

6 Discussion

After using gradient boosted trees to learn parameters most predictive of a county's hospitalization rate, we find that the 6 most predictive features include the total cost of insurance, percentage of individuals over the age of 65, percentage of individuals who are Asian, percentage of individuals who are Hispanic, and total population. These demographic factors largely make sense for different reasons.

Simply, percent of individuals over the age of 65 makes sense as a factor predictive of a county's heart disease hospitalization rate as older individuals are more likely to have heart disease. Additionally, demographic factors such as race and ethnicity (Hispanic and Asian) influence one's diet and cultural practices and there exist genetic differences passed down through generations in these groups. For instance, there is a genetic marker in South Asians that has been strongly linked to an increased risk of cardiovascular disease [8]. Additionally, typical foods present in Hispanic and Asian cultures may differ from that of other cultures, influencing the risk of heart disease in these groups.

Moreover, beyond merely analyzing the implications of social determinants of health, we sought to understand how policy — namely Medicaid expansion — could have a causal effect on changing the rate of heart disease hospitalizations. While we found that expanding Medicaid had a negative effect on the rate of heart disease hospitalizations, it was not statistically significant, nor was the optimal policy. We anticipate that this is due to simply not having enough time for the downstream effects of the policy to take place. We expect to see a significant effect of Medicaid expansion on heart disease hospitalizations with more time.

7 Future Directions

As a future direction, we would like to extend our algorithm to predict an individual's risk of heart disease. To do so, we would need access to individual patient data, including their clinical history, demographic factors, and status of other risk factors. Doing so would allow us to make more granular predictions on an individual's risk of developing heart disease based on demographic characteristics and other risk factors. However, getting access to this data is certainly difficult, as much of it is protected patient information.

Further, we seek to tune our hyperparamters in the gradient boosted trees, including optimizing the number of trees, number of iterations, number of splits, and learning rate. We will use grid search or Bayesian optimization to do this. We would also like to explore different non-linear methods of modeling, such as random forests. This could allow for better accuracy in predicting the current risk level for heart disease.

In the future, when more data about changes in hospitalization rates in states that have adopted Medicaid becomes available, we would like to extend our analysis to see downstream effects of this policy implementation. It is our expectation that more time will allow any causal effects of Medicaid expansion on decreasing heart disease hospitalization rates to show. This is true because, following Medicaid expansion, it takes time for individuals to enroll in these insurance plans, to connect with primary care physicians and cardiologists, and to establish treatment plans to manage cardiovascular disease. Thus, we expect to see significant effects of Medicaid expansion on hospitalization rates for heart disease.

Link to code: https://drive.google.com/drive/folders/1lhjRi7dJZiW4R4OP7uICBXp_Fb3_5OH6?usp=sharing

References

[1] Schultz et. al. *Socioeconomic Status and Cardiovascular Outcomes.* AHA Circulation, June 2018.

[2] Psaltopoulou et. al. *Socioeconomic status and risk factors for cardiovascular disease: Impact of dietary mediators.* Hellenic Journal of Cardiology, February 2017.

[3] Hayes et. al. *Racial/Ethnic and Socioeconomic Disparities in Multiple Risk Factors for Heart Disease and Stroke.* CDC MMWR, February 2005.

[4] UC Davis Health. *Lower socioeconomic status linked with heart disease despite improvements in other risk factors.* UC Davis Health Newsroom, August 2011.

[5] Coffey et. al. *The role of social determinants of health in the risk and prevention of group A streptococcal infection, acute rheumatic fever and rheumatic heart disease: A systematic review.* PLOS Neglected Tropical Diseases, June 2018.

[6] Zeiher et. al. *Correlates and Determinants of Cardiorespiratory Fitness in Adults: a Systematic Review.* Sports Medicine Open, September 2019.

[7] Athey, Susan and Wager, Stefan, ”Efficient Policy Learning,” December, 2018.

[8] Palaniappan L, Garg A, Enas E, et al. South asian cardiovascular disease & cancer risk: Genetics & pathophysiology. *J Community Health.* 2018;43(6):1100-1114. doi: 10.1007/s10900-018-0527-8 [doi].

[9] Mazurenko O, Balio CP, Agarwal R, Carroll AE, Menachemi N. The effects of medicaid expansion under the ACA: A systematic review. Health Aff (Millwood). 2018;37(6):944-950. doi: 10.1377/hlthaff.2017.1491 [doi].

Appendix

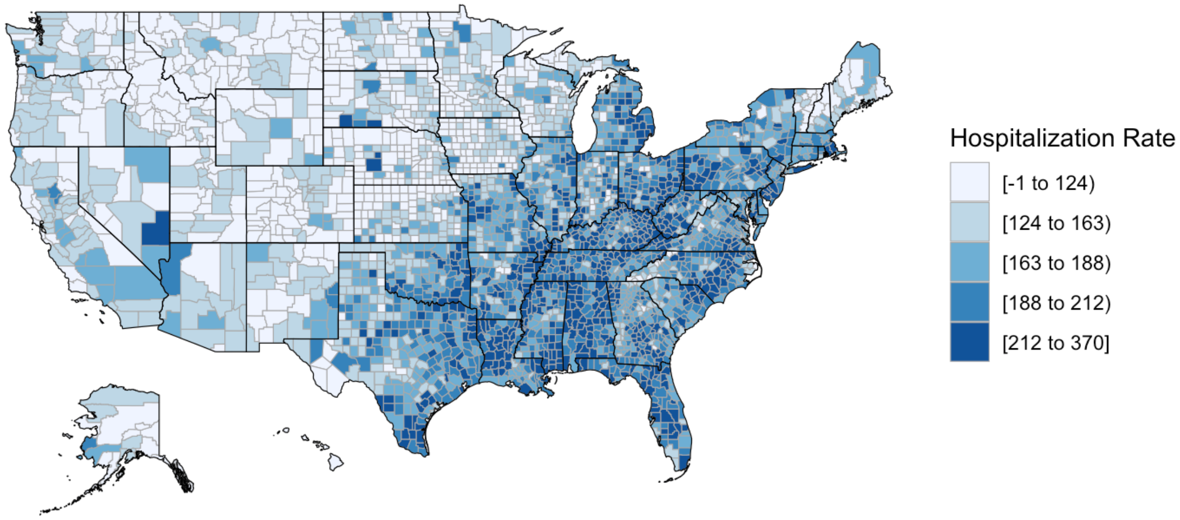


Figure 8: Hospitalizations by County 2014-2016

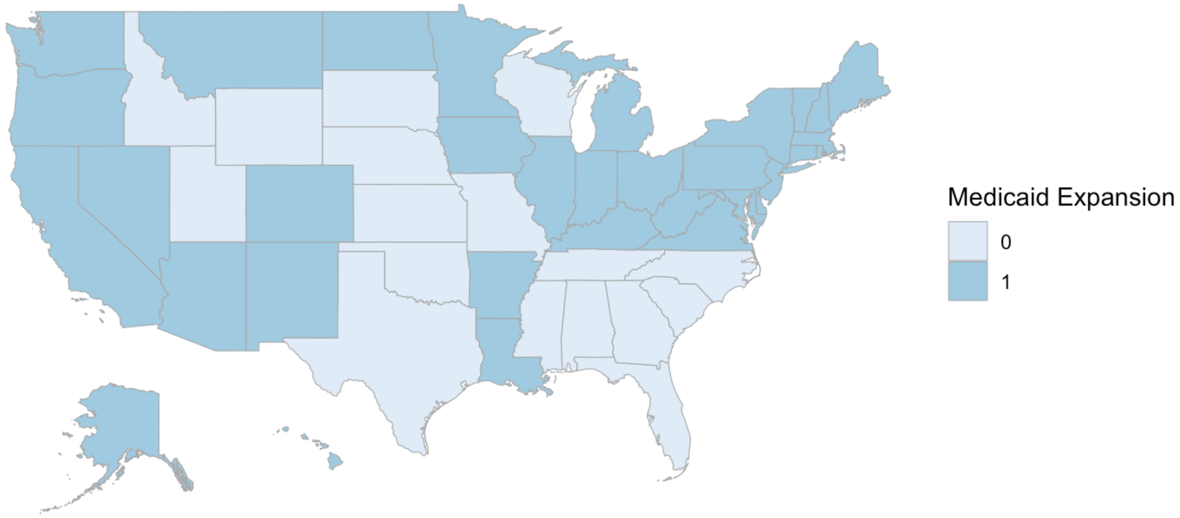


Figure 9: Medicaid Expansion 2013

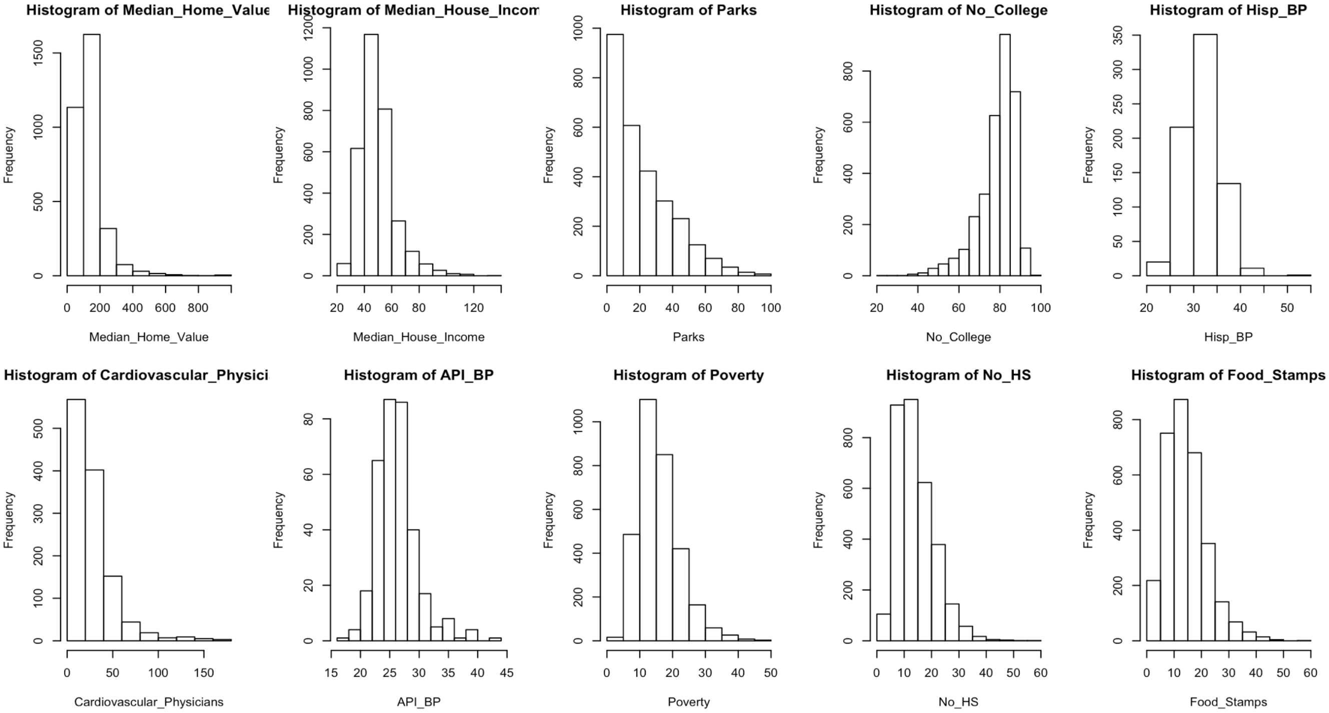
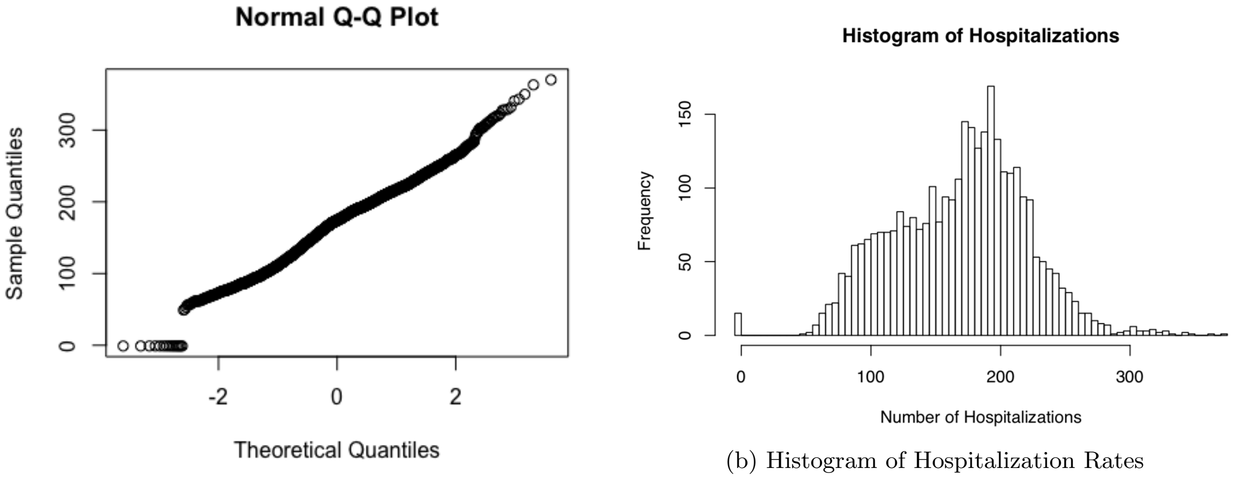


Figure 10: Histogram: Independent Variables

Variable	**Description**
FIPS	County FIPS code
County	Name of county
Hospitalizations	Heart disease hospitalizations (per thousand)
API_BP	Percentage of Asian Pacific Islanders in a given county who do not take prescribed heart disease medications
Black_BP	Percentage of African Americans in a given county who do not take prescribed heart disease medications
Hisp_BP	Percentage of Hispanics in a given county who do not take prescribed heart disease medications
Native_BP	Percentage of Native Americans in a given county who do not take prescribed heart disease medications
White_BP	Percentage of Whites in a given county who do not take prescribed heart disease medications
Total_Cost_PC	Average costs for Medicare beneficiaries with heart disease in a given county
Total_Cost_Inc	Difference in annual costs per capita with and without heart disease in a given county
Air_Quality	Air quality determined by amount of pollutants
Parks	Number of parks in a given county
Severe_Housing_Problems	Quantification of housing shortages or extreme costs
Rural_Urban	County density to determine how urban or rural a county is
Count_Hospitals	Number of hospitals in a given county
Count_Hospitals_Cardiac_Intensive	Number of hospitals with a cardiac intensive unit in a given county
Count_Hospitals_Cardiac_Rehab	Number of hospitals with a cardiac rehab center in a given county
Count_Hospitals_ED	Number of hospitals with an emergency department in a given county
Insurance_Status	percentage of individuals in a given county without health insurance
Medicaid_Eligible	percentage of individuals in a given county who are qualify for Medicaid
Above_65	Percentage of population above 65 years old
Total_Pop	Total population in a given county
Prevalence	percentage of individuals in a given county who have diagnosed heart disease
Native	Percentage of Native Americans in a given county
Asian	Percentage of Asians in a given county
Black	Percentage of African Americans in a given county
White	Percentage of Whites in a given county
Hisp	Percentage of Hispanics in a given county
Diabetes	Percentage of individuals with diabetes in a given county
Obesity	Percentage of individuals with obesity in a given county
Physical_Activity	percentage of individuals in a given county who are physically inactive
No_HS	Percentage of individuals who have not completed high school
No_College	Percentage of individuals who have not completed college
Female_Head_House	Percentage of households with a female head of house in a given county
Food_Stamps	Percentage of a county that receives food stamps
Median_Home_Value	Median home value in a given county (dollars)
Median_House_Income	Median household income in a given county (dollars)
Inequality	Gini index of inequality in a given county
Poverty	Percentage of individuals below the poverty line
Unemployment_Rate	Unemployment rate in a given county
Primary_Care_Physicians	number of individuals in a given county per primary care physician
Cardiovascular_Physicians	number of individuals in a given county per cardiovascular physician

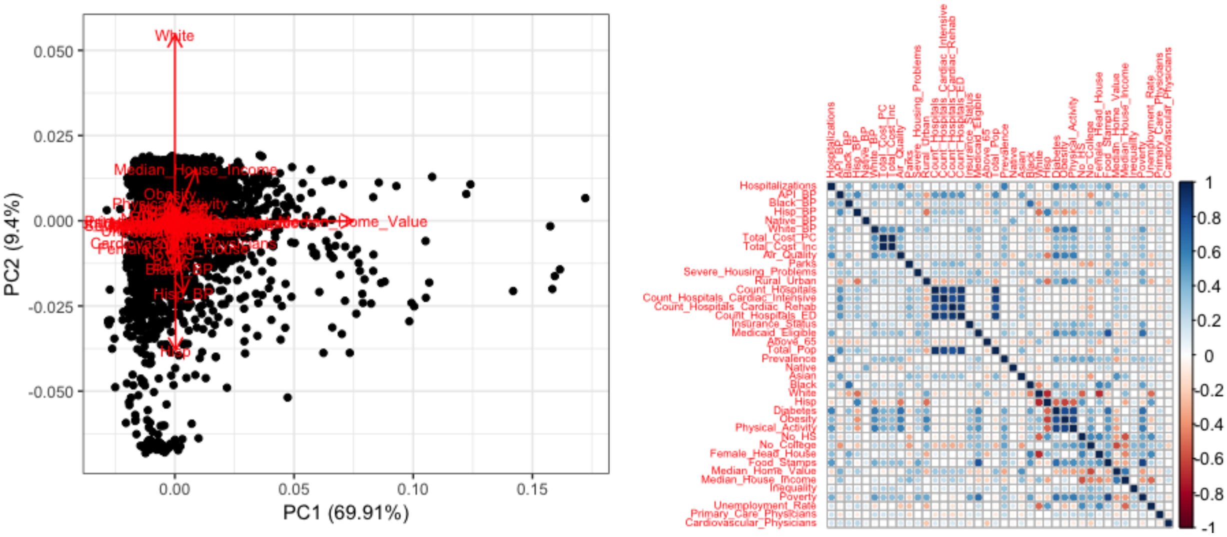
Figure 11: Variable Descriptions



(a) Evaluation of Normality

(b) Histogram of Hospitalization Rates

Figure 12: Normality and Histogram of Response



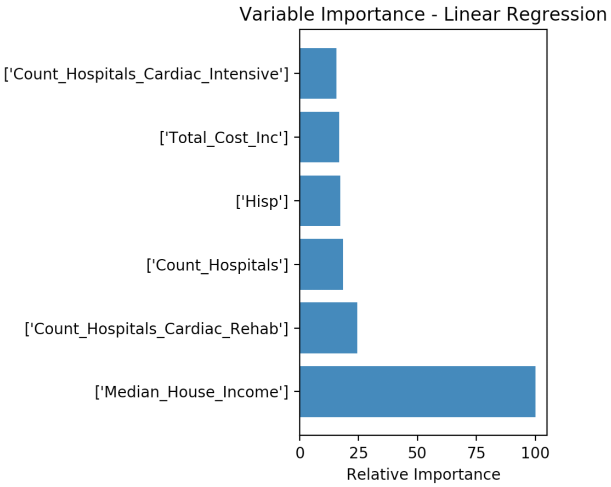
(a) Principal Components Analysis

(b) Correlation Coefficients

Figure 13: Data Exploration: PCA and Correlations

Model	**Test R^2**	**Training R^2**	**Tuned**
Linear Regression	0.6494	0.6638	N/A
Linear Regression (Optimal Elastic Net)	0.6507	0.6600	$\alpha = 0.9$
GBM	0.7313	0.8354	N/A
GBM (Optimal Learning Rate)	0.7467	0.8630	LR = 0.16

Table 3: Model Comparison

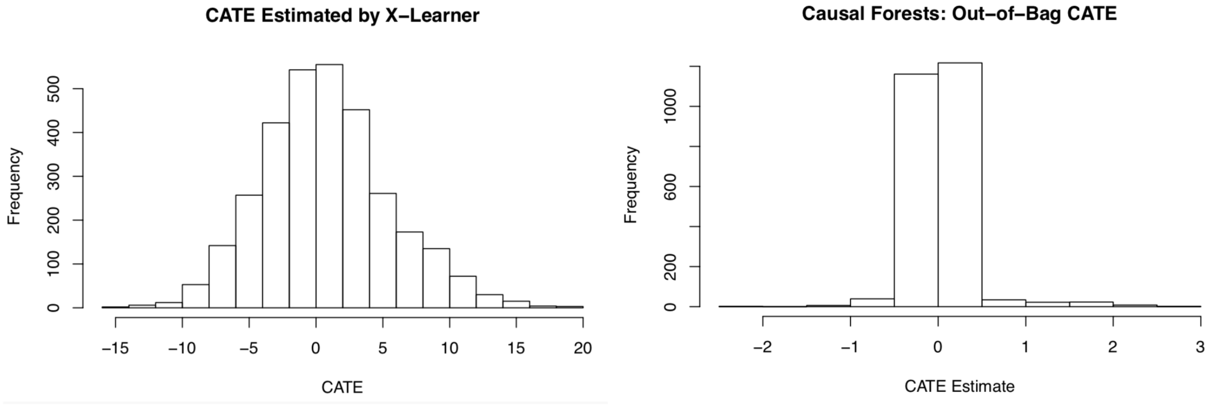


(a) Regression Importance

	x
Median_Home_Value	0.9797352
Median_House_Income	0.1170081
Parks	0.0825298
No_College	-0.0736841
Hisp_BP	0.0482002
Cardiovascular_Physicians	0.0481548
API_BP	0.0434000
Poverty	-0.0295831
No_HS	-0.0295279
Food_Stamps	-0.0283870
Physical_Activity	-0.0268881
Medicaid_Eligible	-0.0265065
Obesity	-0.0254321
Black_BP	0.0251568
Prevalence	-0.0233282
Severe_Housing_Problems	0.0213752
Asian	0.0192116
Black	-0.0174133
Insurance_Status	-0.0125385
Count_Hospitals	0.0095255
Female_Head_House	-0.0090532
Diabetes	-0.0087196
Above_65	-0.0086611
Native_BP	0.0073548
Count_Hospitals_ED	0.0069498
Native	-0.0050153
Unemployment_Rate	-0.0047346
Count_Hospitals_Cardiac_Rehab	0.0047306
Primary_Care_Physicians	-0.0045661
Hisp	0.0043208
Count_Hospitals_Cardiac_Intensive	0.0040747
Rural_Urban	-0.0034370
White_BP	-0.0027082
Air_Quality	-0.0012446
White	0.0003976
Inequality	0.0000749

(b) PCA Importance

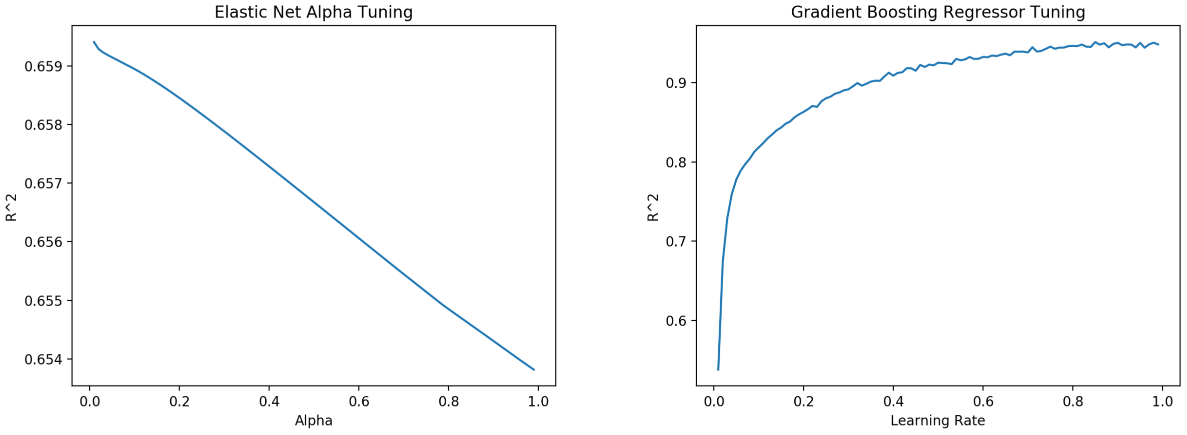
Figure 14: Importance Tables: PCA and Regression



(a) X-Learner Evaluation of Heterogeneity

(b) Out-of-Bag Causal Forest Heterogeneity

Figure 15: Assessment of Heterogeneity



(a) Elastic Net Regression

(b) Gradient Boosted Trees

Figure 16: Hyperparameter Tuning

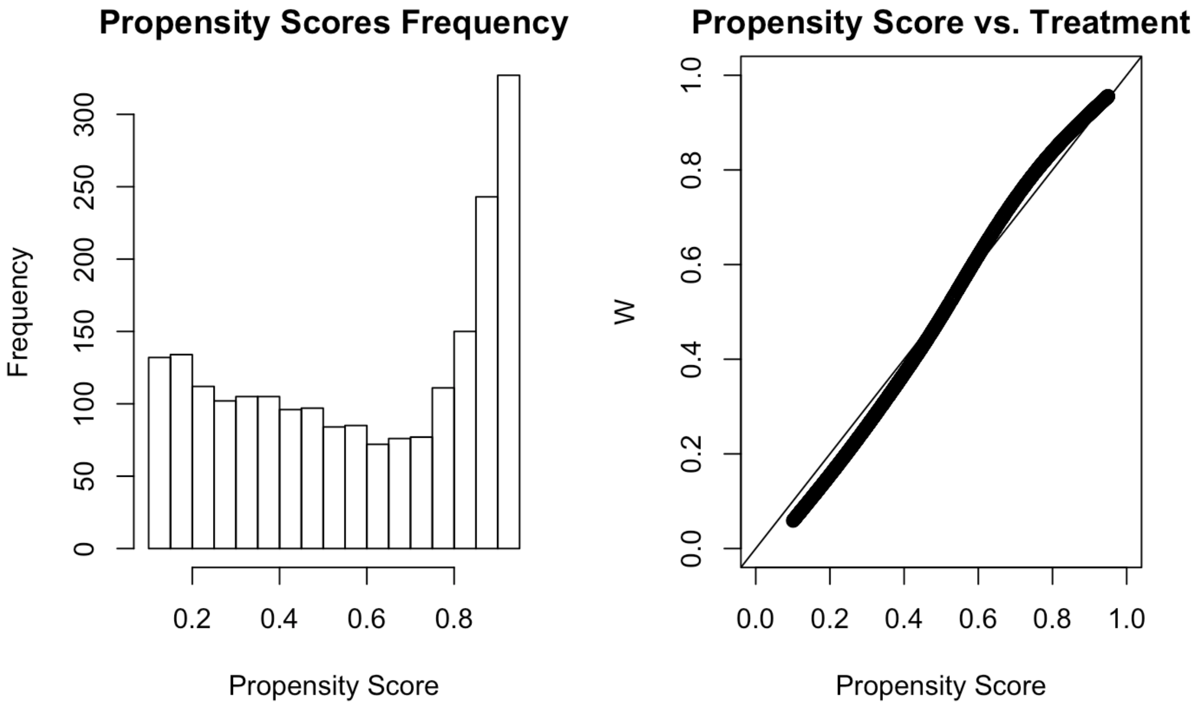


Figure 17: Logistic Regression Propensity Calibration

Method	Average Treatment Effect	Lower-bound	Upper-bound
Unbiased	-6.83	-10.34	-3.32
IPW (Logistic)	2.12	-16.38	20.63
IPW (RF)	14.52	1.11	27.92
AIPW (RF, no weighting)	-2.34	-6.22	1.54
AIPW (RF, weighting)	-3.76	-11.19	3.67

Table 4: Average Treatment Effect Comparison

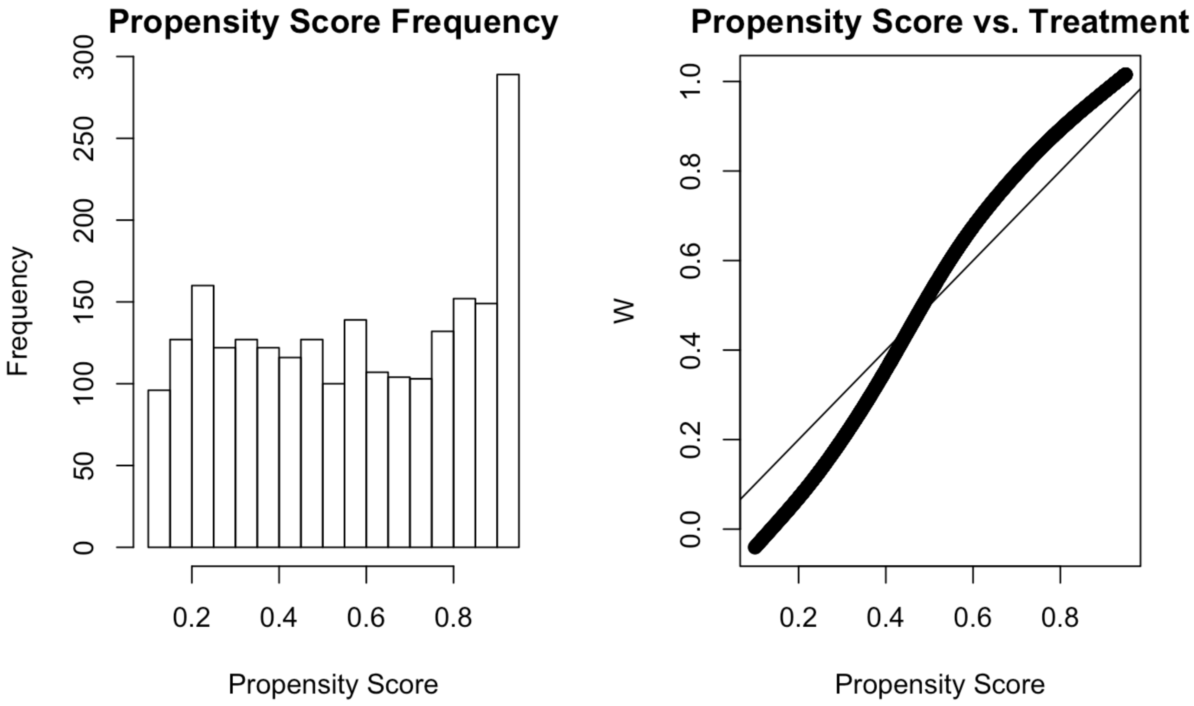


Figure 18: Random Forest Propensity Calibration

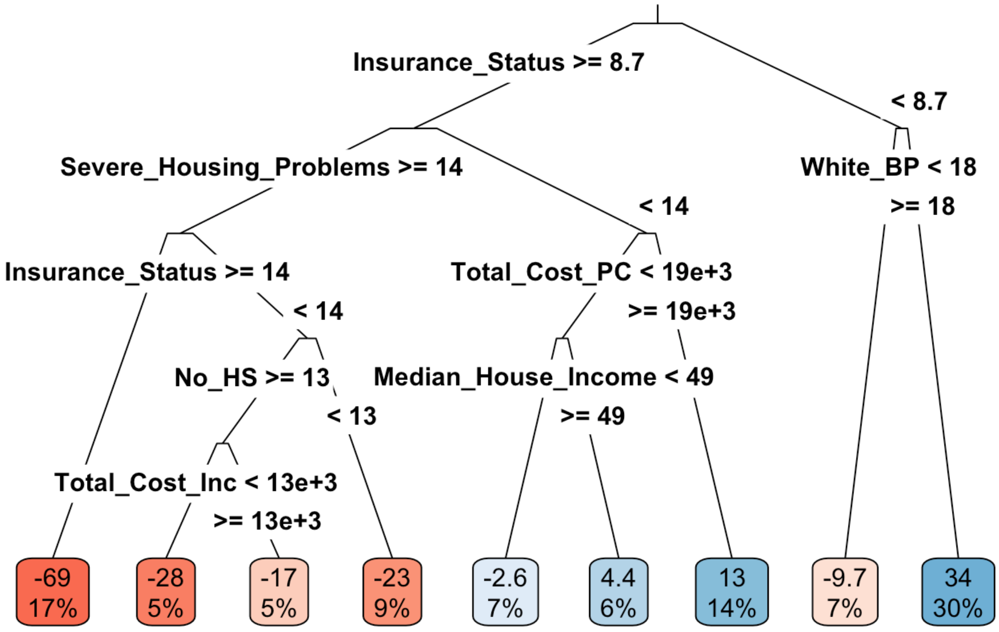


Figure 19: Policy Optimization: Causal Tree