Using Social Determinants of Health to Assess Risk of Heart Disease

Jack Collison, Athreya Steiger
CS 221, Stanford University

Stanford Computer Science

Background

Social determinants of health have a high predictive power in assessing an individual's likelihood of developing heart disease over the course of their lifetime. This project uses ML to predict county-level incidence of heart disease. The goal is to understand the principle determinants that are relevant to heart disease and the causal effects of the recent Medicaid expansion. This will facilitate conversations between patients and physicians about the risk of developing heart disease, thus allowing earlier adoption of preventative lifestyle shifts that can mitigate this risk.

Models

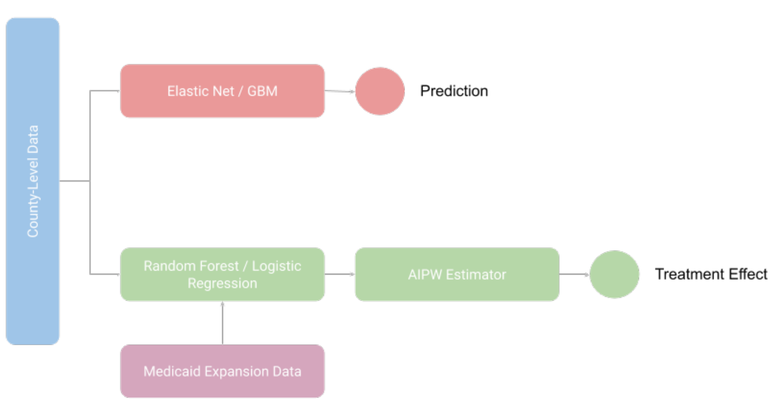
Data

- **Dataset Types**
  - County-level reports of hospitalizations related to heart disease derived from reports to the CDC.
  - Statewide Medicaid expansion reports
- **Response Variable**
  - Number of hospitalizations (per 1000 people)
- **Independent Variables**
  - County-level demographics (race, socioeconomic status, education levels, etc.)
  - Risk factors (diabetes and obesity status, median income, pollution index, etc.)
- **Treatment Variable**
  - Indicator of Medicaid expansion in 2013 per state

Limitations

- Only county-level data rather than individual-level; 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 can parse out the most predictive risk factors.

Model Outline



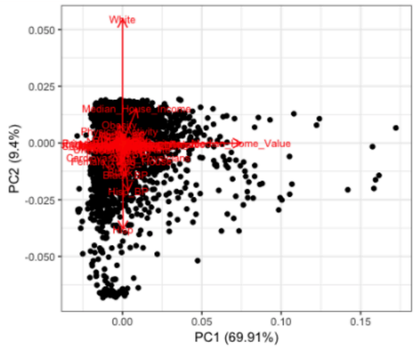
Implementation

- **Elastic Net**
  - Combo of L1 & L2 penalties used for dimensionality reduction (addresses multicollinearity)
  - Implemented using scikit learn
- **Gradient Boosted Regression Trees**
  - This is a highly nonlinear space; GB trees are especially useful for discovering nonlinear patterns
  - Implemented using scikit learn
- **Causal Inference**
  - Used to understand how policies (Medicaid expansion) have affected heart disease hospitalization.
  - Propensity scores (L1-logistic) used to estimate unbiased treatment effect as shown below

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

Results & Analysis

Predictive Results

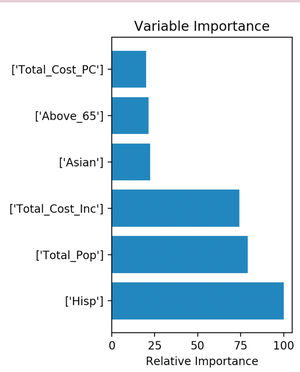


- **PCA**
  - Variables that account for the most variance in the data are as follows: **1)** Median Home Value; **2)** Median House Income; **3)** Presence of Parks; **4)** No College; **5)** Hispanic Blood Pressure; **6)** Number of Cardiovascular Physicians

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

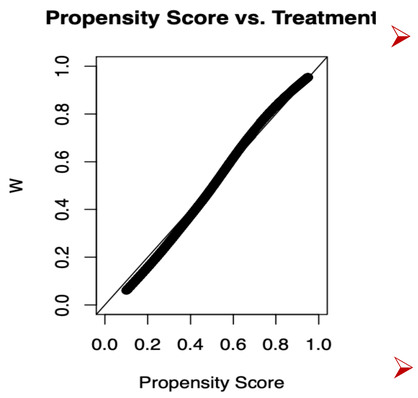
As shown above, the optimal α for Elastic Net was 0.01, indicating that a LASSO-penalized model most accurately fits the data. The optimal learning rate for gradient boosted trees was 0.85, although this overfit the training data.

Feature Importance



- **Gradient Boosted Trees**
  - The graphic to the left shows the 6 most important variables in predicting a county's hospitalization rate for heart disease

Causal Inference



	x
ATE.estimate	-2.323502
Lowerbound.estimate	-6.021404
Upperbound.estimate	1.374399

- **Propensity Scores**
  - Propensity scores are calculated with L1-penalized Logistic Regression & Random Forests
  - This plot shows that the scores are well calibrated as they follow the line y=x
- **Interpretation**
  - Average treatment effect is found to be negative (indicating that Medicaid expansion caused fewer heart-disease related hospitalizations
- Although the treatment effect is not statistically significant, this could be because downstream effects of Medicaid expansion have not taken full effect.
- Further, there is the question of unconfoundedness, as there may be outside variables affecting this relationship

Discussion

- Unsurprisingly, we find that features such as county-wide age, ethnicity & race proportions, and insurance cost are largely predictive of a county's heart disease hospitalization rate
- Though Medicaid expansion does not have a significant effect, it's likely these effects will be seen in several years

Future Directions

- Assess optimal policies for Medicaid expansion to see which counties would benefit most from this expansion
- Perform more granular individual predictions of heart disease risk (with access to patient datasets)
- Long-term effect of Medicaid expansion (after more time)

Acknowledgements & References

- We'd like to say thank you to our mentor Sharman Tan.

[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.