TCS-CMU Collaboration

Chronic Disease Forecasting

Based on Social Determinants of Health

05/14/2021 Final Presentation

Presenters

Ajay Valecha

Satvika Neti

Tianying Chu

Xiaojun Ma

Yilun Chen

CMU Primary Advisor

Jeremy C.Weiss

TCS Primary Advisor

Sundara R. Subramanian

Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

The Team

Heinz College - Master of Science in Public Policy and Management

Ajay Valecha
Data Analytics
Spring 2021

Satvika Neti
Data Analytics
Spring 2021

Tianying Chu
Data Analytics
Spring 2021

Xiaojun Ma
Data Analytics
Spring 2021

Yilun Chen
Regular Track
Spring 2021

Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

Collaboration

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

We embrace our differentiators: Quantitative rigor. Technological mastery. An emphasis on evidence. Cross-disciplinary inquiry and compassionate values. Heinz College is a public policy school unlike any in the world, because our people are dedicated to solving the important problems of our time, **problems that live at the nexus of people, policy, and technology.**

Tata Consultancy Services (TCS) is an IT services, consulting and business solutions organization that has been partnering with many of the world's largest businesses in their transformation journeys for over 50 years. TCS offers a **consulting-led, cognitive powered, integrated portfolio of business, technology and engineering services and solutions**.

Executive Summary

NO public tool to systematically **analyse associated risk factors** at county level

Offer TCS **evidence-based, personalized solutions** for **early preventive intervention**

Innovative, interactive dashboard integrating historical information and predicted outcomes using machine learning techniques

Agenda

01 Background

02 Methodology

03 Results

04 User Cases & Value Add

05 Discussion

Agenda

01 Background

02 Methodology

03 Results

04 User Cases & Value Add

05 Discussion

1.1 Problem Statement

Broader Context:

- Heart disease and other cardiovascular diseases are the number one killer in America with heart disease causing 655,000 deaths every year and stroke causing 137,000 deaths per year.
- There is a heart-related death **every 36 seconds** and stroke-related death **every 4 minute**

Definition of Problem:

- Tata Consultancy Services aims to build a machine learning solution that detects heart diseases early through a risk scoring prediction model to intervene with preventive and personalized care.
- Increase health outcomes and save costs worth billions of dollars to the healthcare industry.

1.2 Scoping

Scoping:

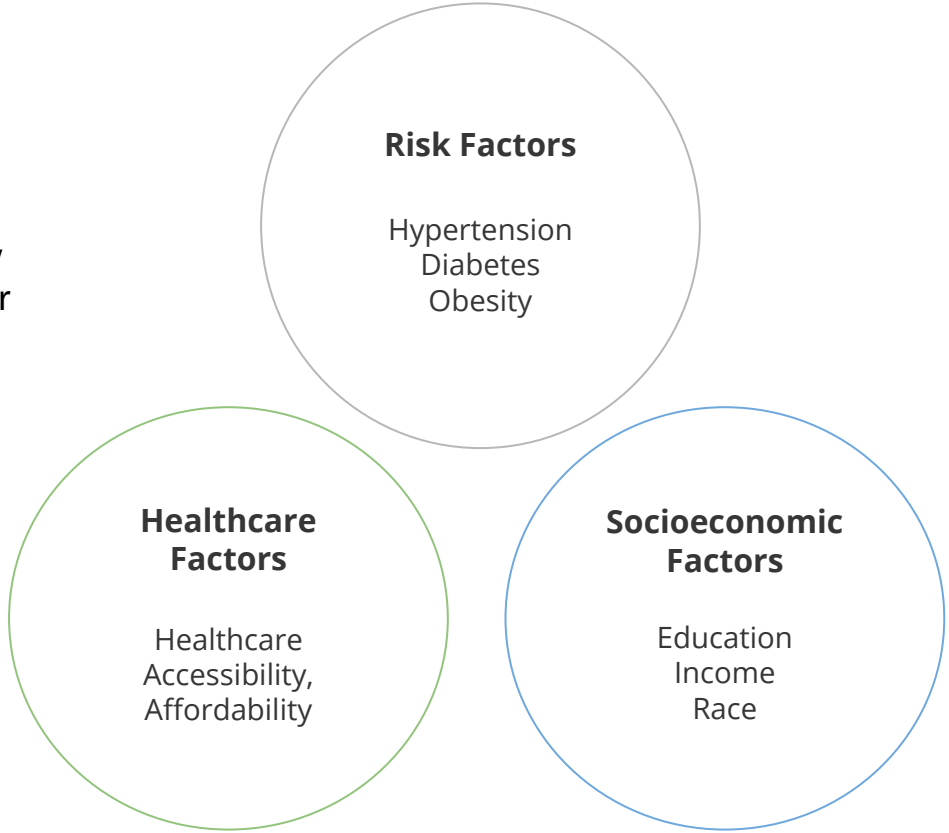
- ❑ Take risk factors such as the social determinants of health and generate a county level prediction for the hospitalization rate for both heart disease and stroke
- ❑ Provide the necessary intervention to prevent hospitalizations in the US

Goals:

- ❑ Machine Learning Models
- ❑ Dashboard to visualise the results and recommendations

Constraints:

- ❑ Data not available on individual basis



1.3 Data Sources

Outcome

Dataset	Information	Source	Release Interval Latest Version	Collect Method
Atlas - Heart Disease & Stroke	Hospitalization rate for 65+	CDC Centers for Disease Control and Prevention	Annually 2015-2017	Manual
American Community Survey	Demographics, employment, and more	Census Bureau	Annually 2019	API
Spending & Prevalence	Spending and prevalence for different chronic diseases	CMS Centers for Medicare & Medicaid Services	Annually 2018	csv files
Air Quality Index	Air quality data that measures the levels of pollutants in the air	EPA Environmental Protection Agency	Annually 2020	csv files

Agenda

01 Background

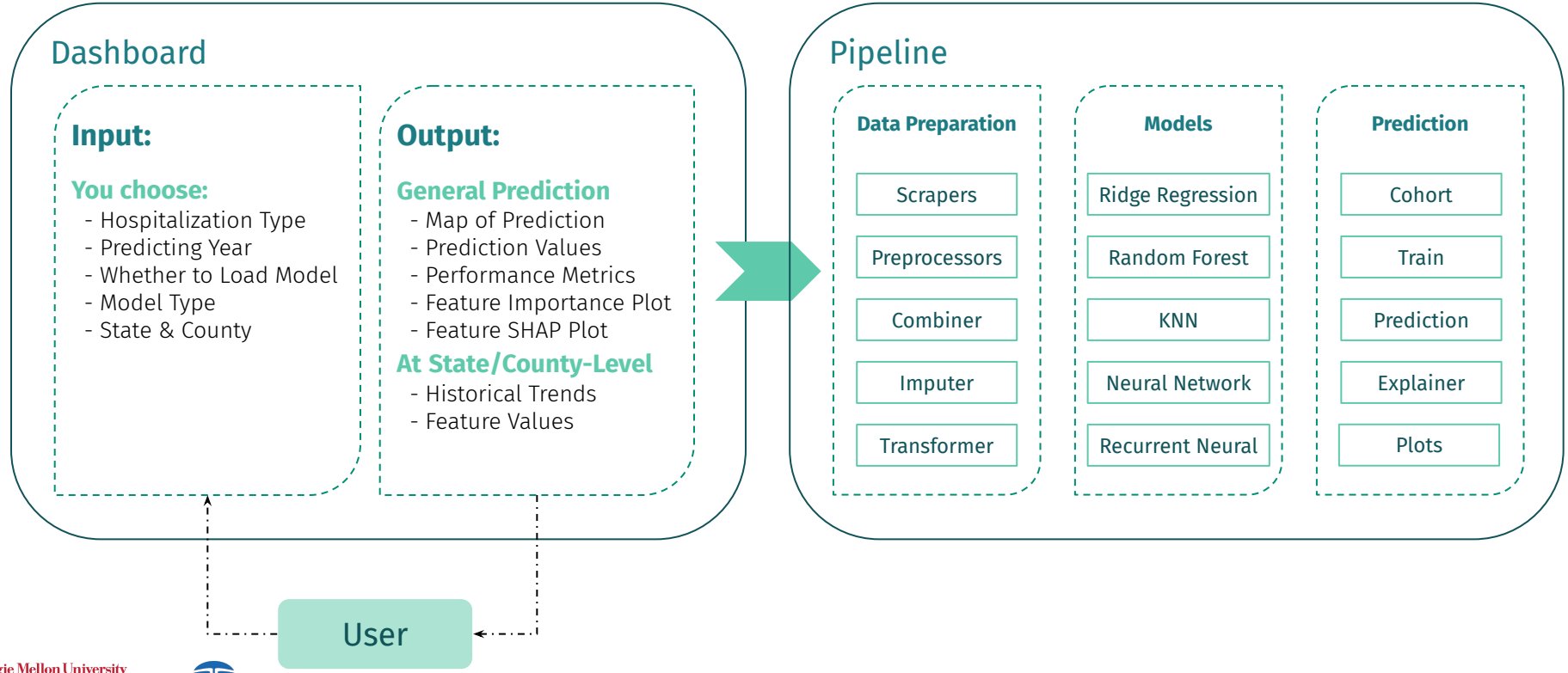
02 Methodology

03 Results

04 User Cases & Value Add

05 Discussion

2.1 Our Solution

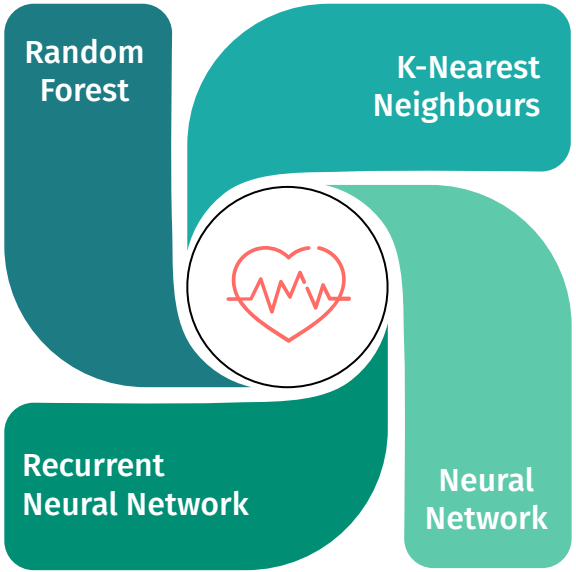


2.2 Models

Baseline

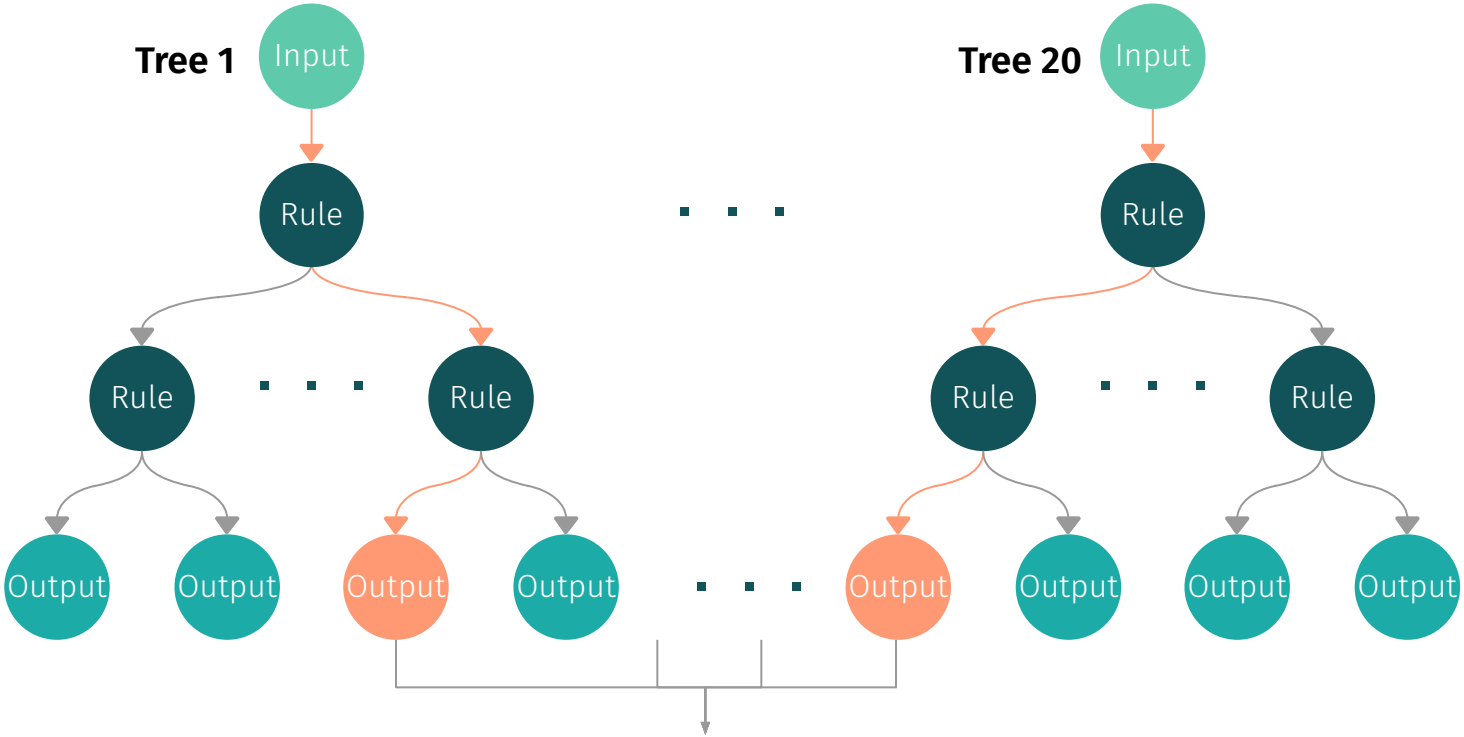
Ridge Regression

A linear model with regularized coefficients to combat overfitting



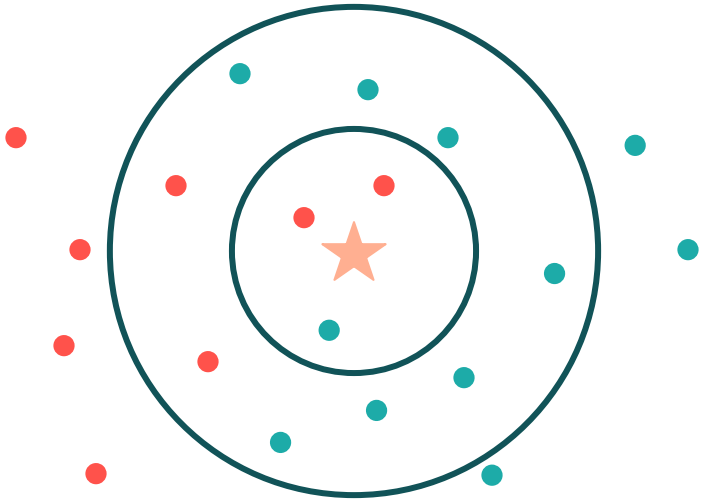
Training	2011-2013 2012-2014
Validation	2015-2017
Prediction	2019-2021
Evaluation Metrics	R-square

2.2.1 Random Forest



2.2.2 K-Nearest Neighbours

- Non-parametric, supervised learning algorithm
- Compute similarity between *x* and all examples in training data



Class A
Class B
Unknown Class

2.2.3 (Recurrent) Neural Networks

Neural Network

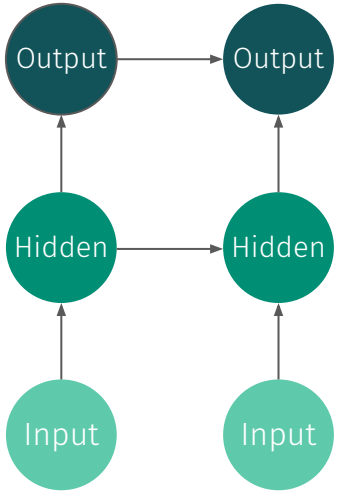
Second-order regularization

1 hidden layer, 20 hidden units

Year 2010 and 2011

sigmoid()

tanh()

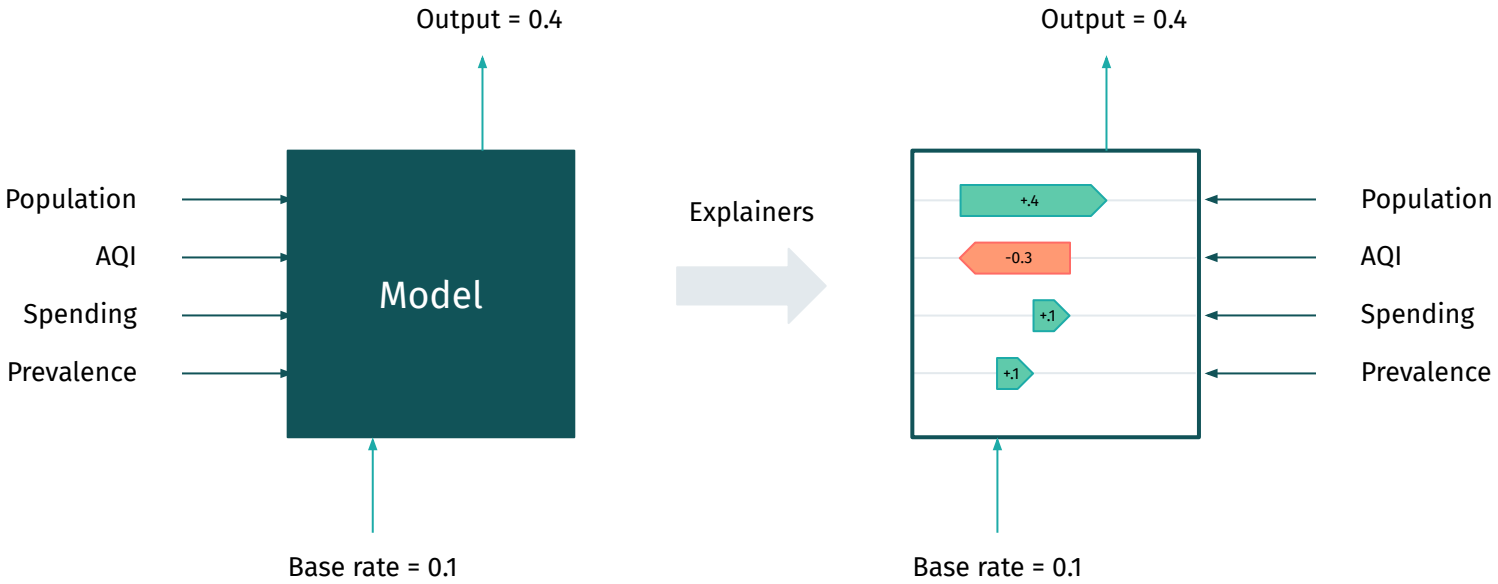


Year 2010 Year 2011

Recurrent Neural Network

Second-order regularization

2.3 SHAP Explainer



Agenda

01 Background

02 Methodology

03 Results

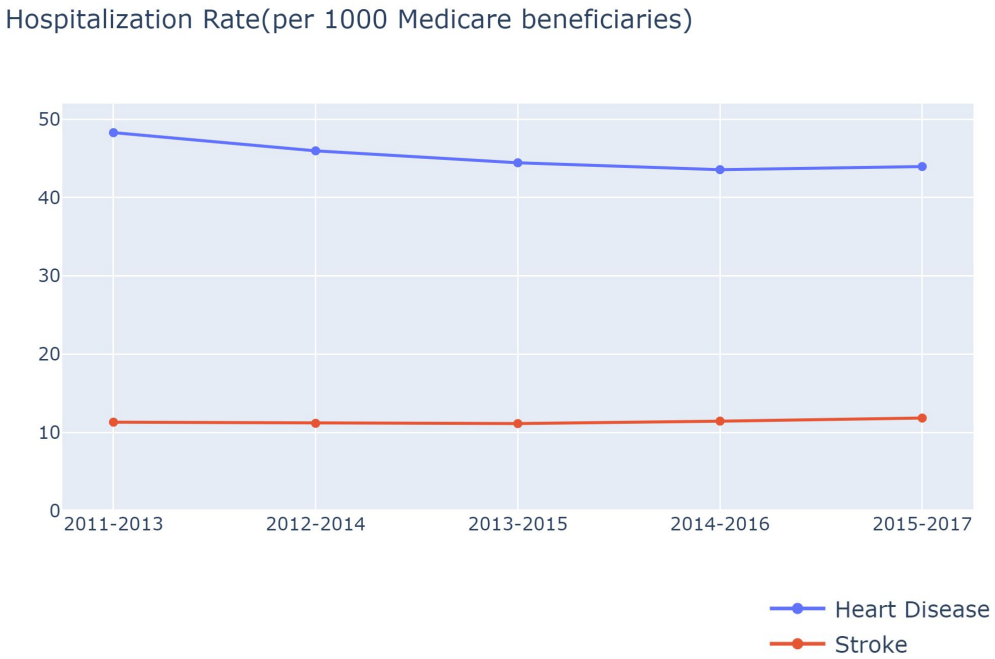
04 User Cases & Value Add

05 Discussion

3.1 Preliminary Analysis

On average, hospitalization rate of **heart disease** is over **297%** higher than that of **stroke**

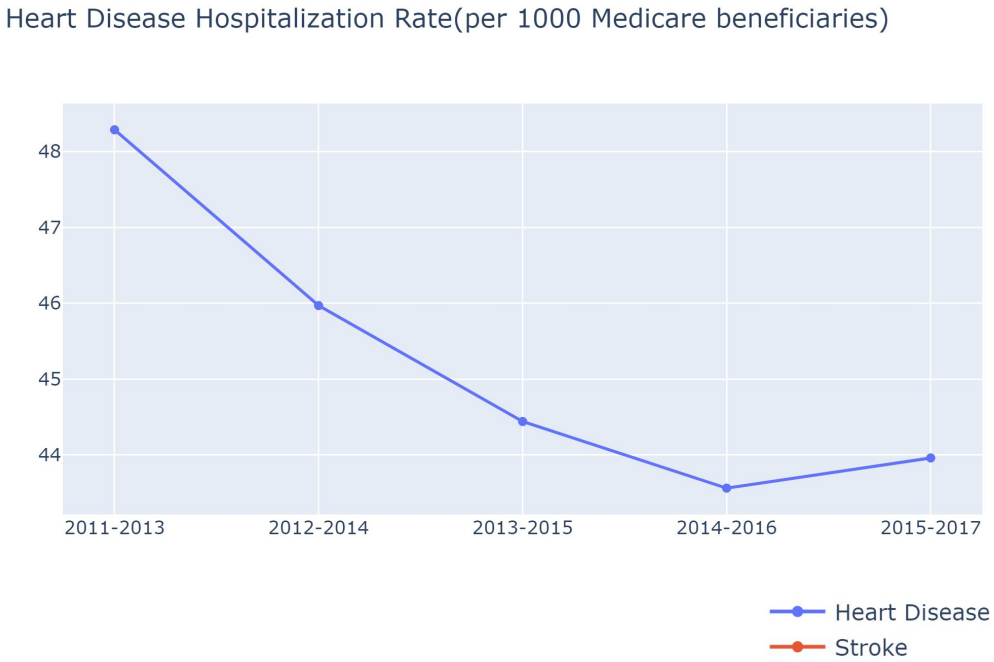
From 2011-2013 to 2015-2017, **heart disease** hospitalization rate **decreased** by **9%**, while that of **stroke** **increased** by **4%**



3.1 Preliminary Analysis

On average, hospitalization rate of **heart disease** is over **297%** higher than that of **stroke**

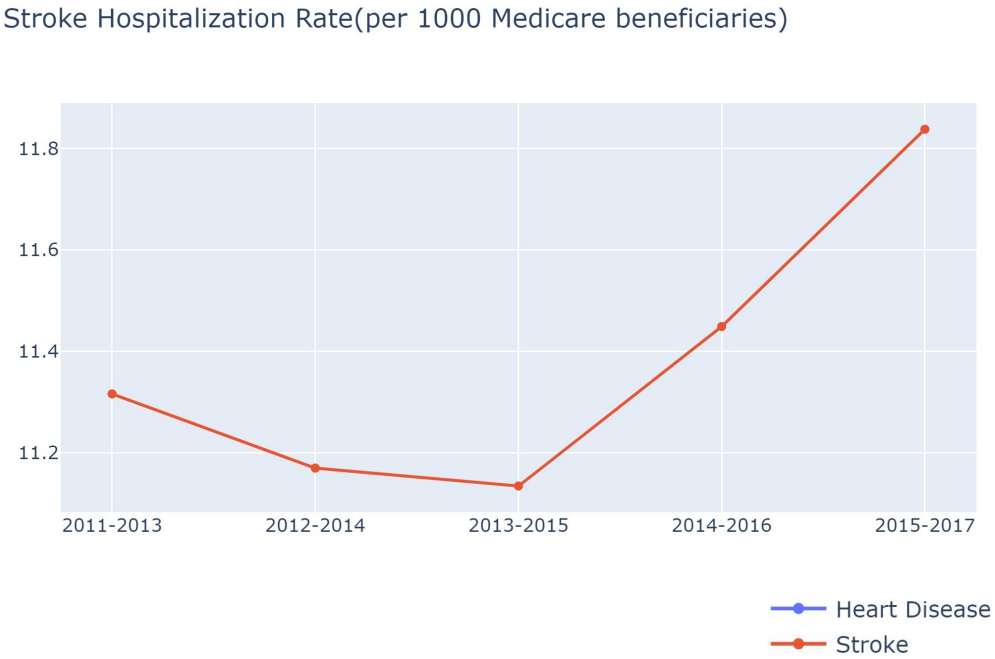
From 2011-2013 to 2015-2017, **heart disease** hospitalization rate **decreased** by **9%**, while that of **stroke** **increased** by **4%**



3.1 Preliminary Analysis

On average, hospitalization rate of **heart disease** is over **297%** higher than that of **stroke**

From 2011-2013 to 2015-2017, **heart disease** hospitalization rate **decreased** by **9%,** while that of **stroke increased** by **4%**



3.1 Preliminary Analysis

Education attainment and **household median income** are **negatively** correlated with hospitalization rates.

NOTABLE CORRELATION

	Education Attainment	Median Income
Heart Disease Hospitalization	-0.56	-0.41
Stroke Hospitalization	-0.46	-0.34

3.2 Model Performance

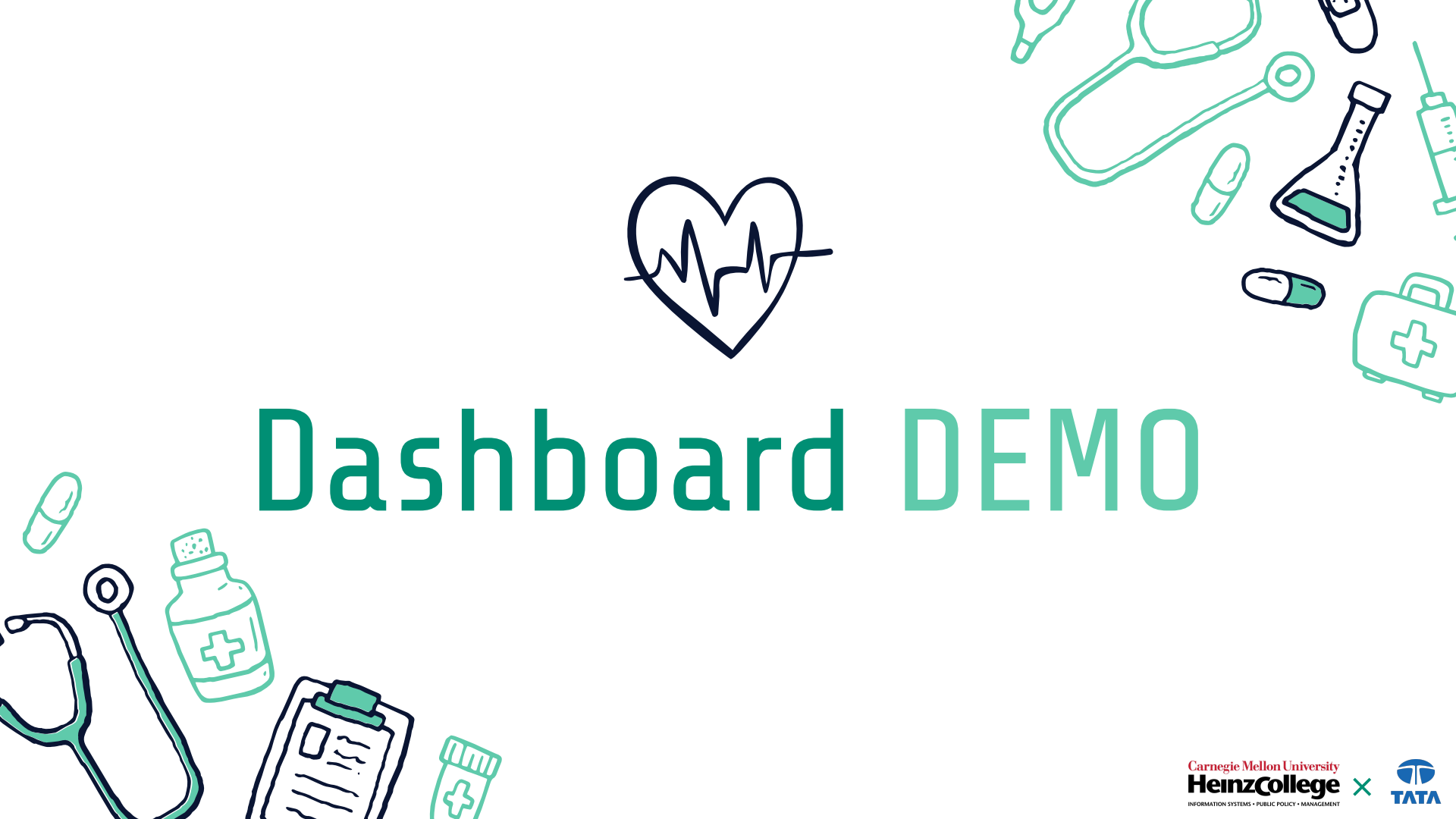
	Ridge Regression	Random Forest	K-Nearest Neighbours	Neural Network	Recurrent Neural Network
R^2 Heart Disease Hospitalization	0.44	0.68	0.67	0.69	0.71
R^2 Stroke Hospitalization	-0.08	0.39	0.45	0.22	0.52
Computation *How fast does it run?*	★★★★★	★★★★☆	★★★★☆	★★★☆☆	★★★☆☆
Interpretability *What do the results mean?*	★★★★★	★★★★★	★★☆☆☆	★★★☆☆	★★★☆☆
Temporal Variation *How does past data affect results?*	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	★★★★★
Generalizability *How will it work in the future?*	★★☆☆☆	★★★★☆	★★★☆☆	★★★☆☆	★★★★☆

Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

3.2 Model Performance

	Ridge Regression	Random Forest	K-Nearest Neighbours	Neural Network	Recurrent Neural Network
R^2 Heart Disease Hospitalization	0.44	0.68	0.67	0.69	0.71
R^2 Stroke Hospitalization	-0.08	0.39	0.45	0.22	0.52
Computation *How fast does it run?*	★★★★★	★★★★☆	★★★★☆	★★★☆☆	★★★☆☆
Interpretability *What do the results mean?*	★★★★★	★★★★★	★★☆☆☆	★★★☆☆	★★★☆☆
Temporal Variation *How does past data affect results?*	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	☆☆☆☆☆	★★★★★
Generalizability *How will it work in the future?*	★★☆☆☆	★★★★☆	★★★☆☆	★★★☆☆	★★★★☆

Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA



Dashboard DEMO

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT
×
TATA

Agenda

01 Background

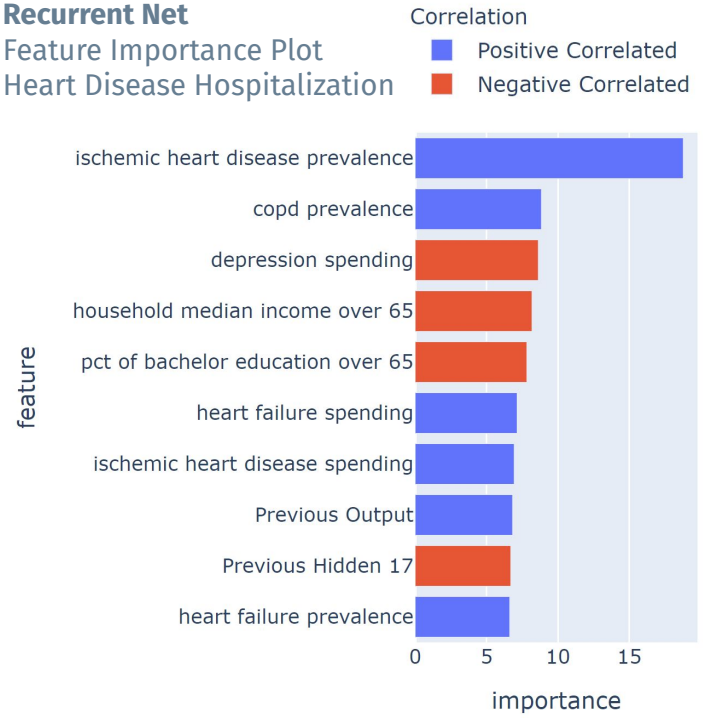
02 Methodology

03 Results

04 User Cases & Value Add

05 Discussion

4.1 Top Features



Common Set of TOP 20 Important Features for Heart Disease Hospitalization (RF, NN, RNN)

(-) Negative Correlated (+) Positive Correlated

Social Determinant

Household Median Income (-)
Education Attainment (-)
Percentage of Male (-)

Other Chronic Disease

Ischemic Heart Disease Prevalence (+)
COPD Prevalence (+)
Heart Failure Prevalence (+)
Hypertension Prevalence (+)
Diabetes Prevalence (+)
Hyperlipidemia Prevalence (+)

Temporal Dependency (RNN Only)

Previous Output (+)

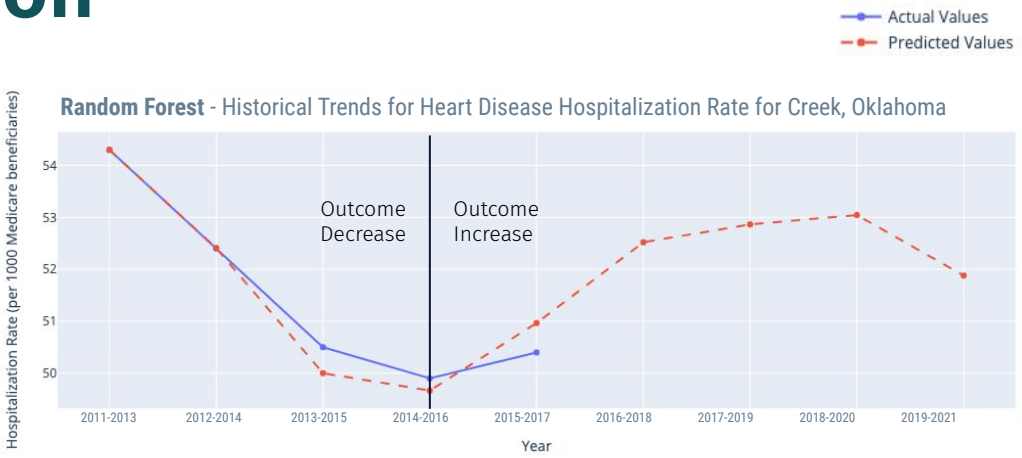
4.2 Early Intervention

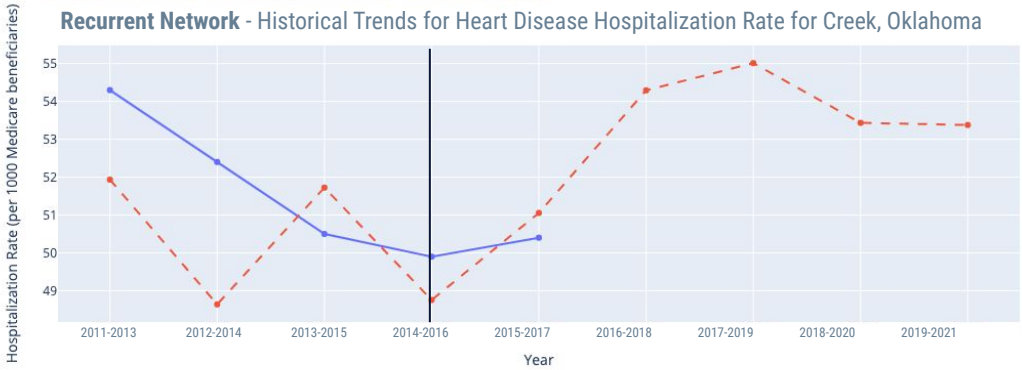
State: Oklahoma
County: Creek
Outcome: Heart Disease Hospitalization

- Correct model prediction.

- See a dramatic increase in predicted heart disease hospitalization rate after 2015-2017.

Which features are driving this increase? And how could policymakers intervene to prevent it?





Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

4.2 Early Intervention

State: Oklahoma
County: Creek

Observations

- Important social determinant factors (median income, education) don't explain this increase.

- Prevalence of the ischemic heart disease and heart failure followed a trend similar to hospitalization.

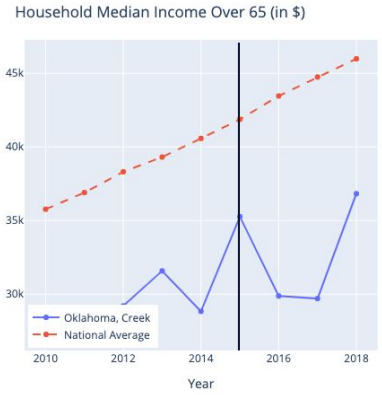
- Dramatic increase of the hypertension and diabetes prevalence starting from early years (2011, 2012).

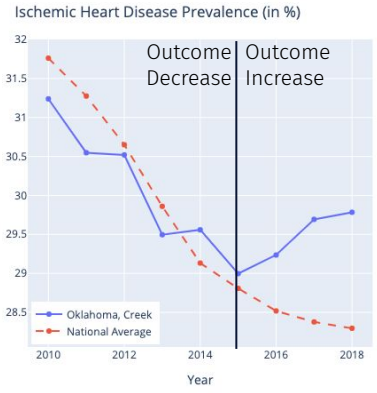
Conclusions

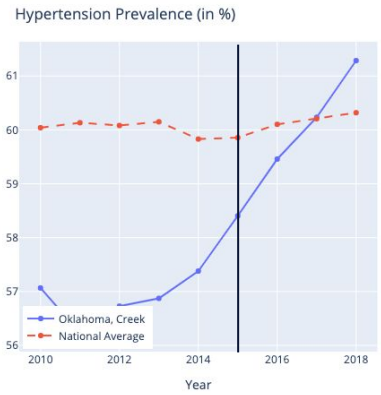
Early risk factors, such as hypertension and diabetes, seem to be the early signs of both prevalence and hospitalization for heart disease.

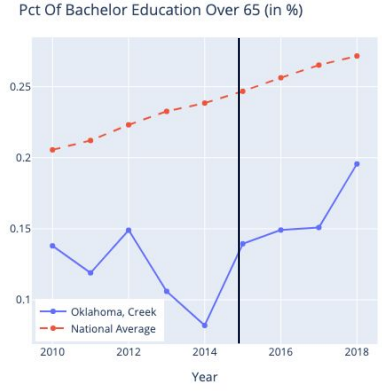
Intervention

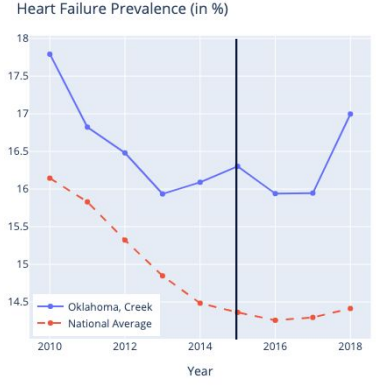
Provide early preventative intervention for hypertension or diabetes through education or laws.

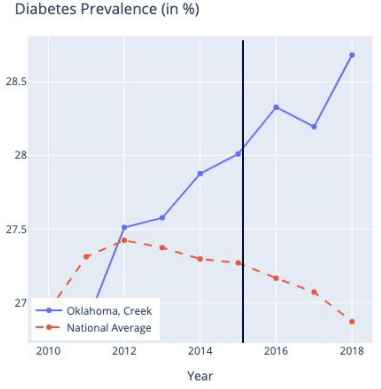








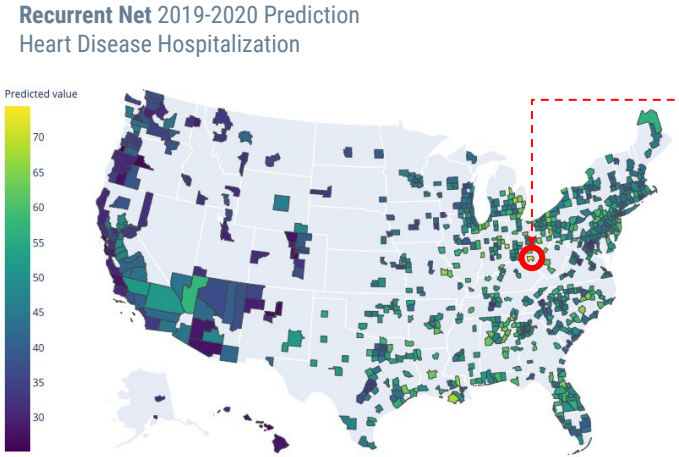




Carnegie Mellon University HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT
× TATA

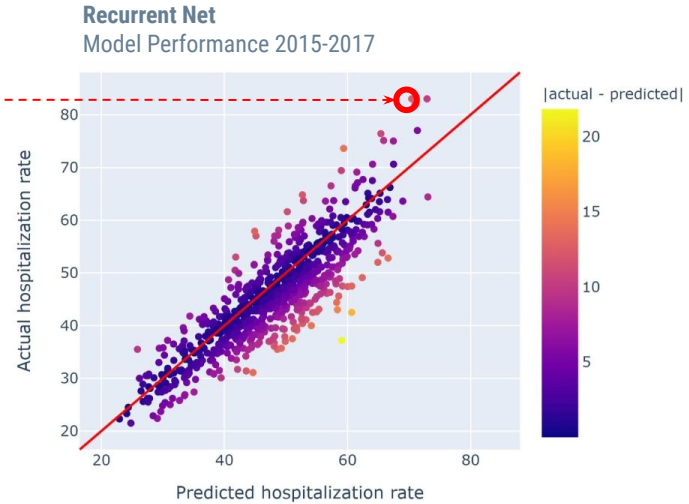
4.3 Outlier Exploration

Scioto, a double-fold outlier



High heart disease hospitalization rate

Scioto, Ohio



High |actual-predicted| value

4.3 Outlier Exploration

State: Ohio
County: Scioto
Outcome: Heart Disease Hospitalization

- Actual values first decreased from 2011-2013 to 2014-2016 and then increased quite dramatically.

- All of our models fail to predict such an abrupt increase, which results in Scioto being an outlier (in terms of model performance) in 2015-2017.

There may exist something driving this increase that is currently not captured in our feature set.

Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

4.3 Outlier Exploration

State: Ohio
County: Scioto

Observations

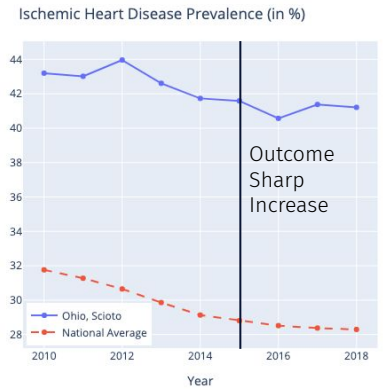
- The “important features” can explain Scioto as an outlier of extremely high outcome rate.

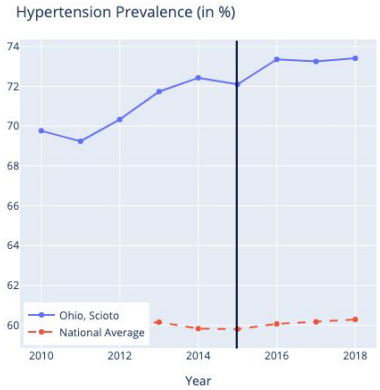
- The abrupt increase is not reflected in the “important features” from earlier. This can possibly explain the large prediction deviation of our model in 2015-2017.

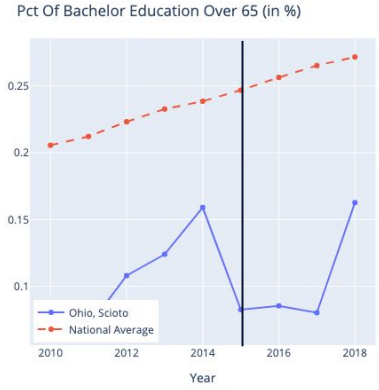
Conclusions

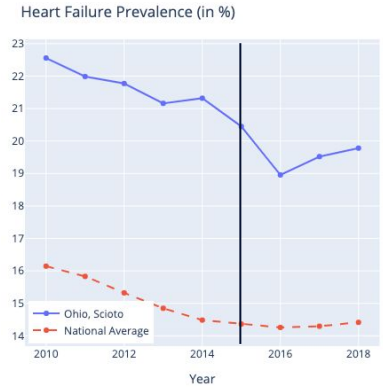
- In this case, the model can accurately predict a high baseline outcome of Scioto comparing to other counties. However, it may not be able to accurately predict the abrupt change of outcome that is being driven by features we don’t have.

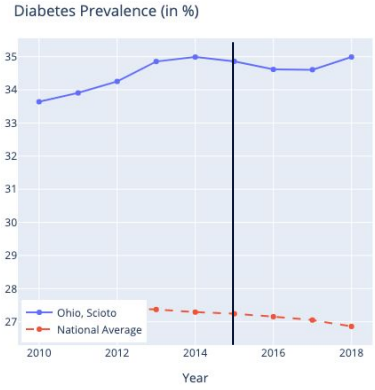
- The result here allows users to narrow down the possible cause of the increasing outcome and inspire further investigation into what happened to this county in 2015-2017.











Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

4.4 Value Added

Dashboard enpowers evidence-based policy decision making

Chronic Disease Prevention
Healthcare Education
...

Nationwide

State/County Level

Important Social Determinants of Health and Risk Factors

- Income, Education Attainment, Gender
- Diabetes, Hypertension, Hyperlipidemia, COPD
- Heart Failure, Ischemic Heart Disease, Stroke

→

Early Intervention

- Consistent, accurate model prediction
- Consistent, increasing hospitalization rate
- Investigate feature trend
- Formulate intervention Strategies

+

Outlier Exploration

- Outliers above diagonal
  - Account for unexpected nature shock
- Outliers below diagonal
  - Learn from good policy interventions

Agenda

01 Background

02 Methodology

03 Results

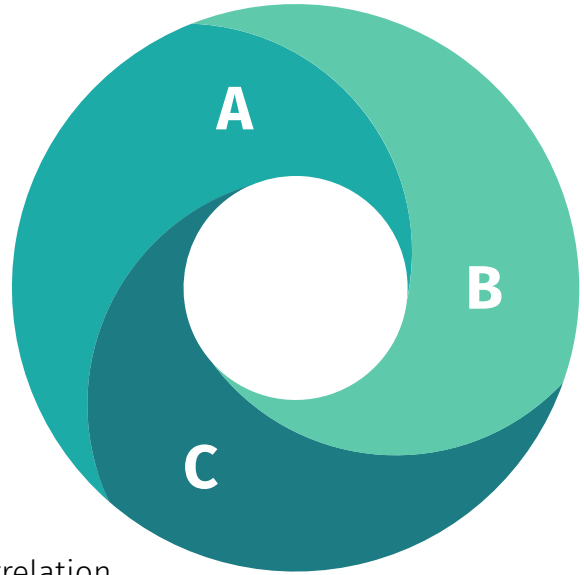
04 User Cases & Value Add

05 Discussion

5.1 Limitations

A. Data

1. Data not available on individual basis
2. Data not available for all counties in the US
3. Healthcare accessibility datasets are not available
4. Uncertainty over the release of data in the future from different data sources



B. Modeling

1. No proper validation dataset for hyperparameter tuning
2. Inability to train data from all available years due to overlap in the outcome variable
3. Stroke results consistently substandard for all machine learning algorithms

C. Implication

1. Results show correlation, not causation

5.2 Going Forward

TCS

- Start early intervention based on prediction
- Analyze feature causality
- Figure out unexpected shock

Future Team

- Find more risk factors
- Streamline the way to get data
- Publish the application to general healthcare community

Thank You !

Time for Questions

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

× TATA

Reference

[1] Kurian A.K., Cardarelli K.M. (2007). Racial and Ethnic Differences in Cardiovascular Disease Risk Factors: A Systematic Review. Ethnicity & Disease 17:144-152.

[2] Virani S.S., Alonso A., Benjamin E.J., Bittencourt M.S., Callaway C.W., Carson A.P., et al. (2020). Heart disease and stroke statistics—2020 update: a report from the American Heart Association external icon. Circulation 141(9):e139–e596.

[3] Tully P.J., Cosh S.M. (2013) Generalized anxiety disorder prevalence and comorbidity with depression in coronary heart disease: A meta-analysis. Journal of Health Psychology 18(12).

[4] Wu Y., Zhu B., Chen Z., Duan J., Luo A., Yang L., Yang C. (2021). New Insights Into the Comorbidity of Coronary Heart Disease and Depression. Current Problems in Cardiology 46(3):100413.

[5] Williams P.T. (2001). Physical fitness and activity as separate heart disease risk factors: a meta-analysis. Med Sci Sports Exerc. 33(5):754–761.

[6] Critchley J.A., Capewell S. (2003). Mortality Risk Reduction Associated With Smoking Cessation in Patients With Coronary Heart Disease. JAMA 290(1):86-97.

Reference (Continued)

[7] Havranek E.P. et al. (2015). Social Determinants of Risk and Outcomes for Cardiovascular Disease. Circulation 132:873–898.

[8] Dupre M.E., George L.K., Liu G., et al. (2012). The Cumulative Effect of Unemployment on Risks for Acute Myocardial Infarction. Arch Intern Med 172(22):1731-1737.

[9] Hyattsville: National Center for Health Statistics (US); 2012 May. Report No.: 2012-1232.
Health, United States.

Data Reference

AQI (Air Quality Index) Data:
https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual (link to download data)

ACS (American Community Survey) Data:
https://www.census.gov/programs-surveys/acs (link to homepage, can download specific tables from there)
https://www.census.gov/data/developers/data-sets/acs-1year.html (link to API)

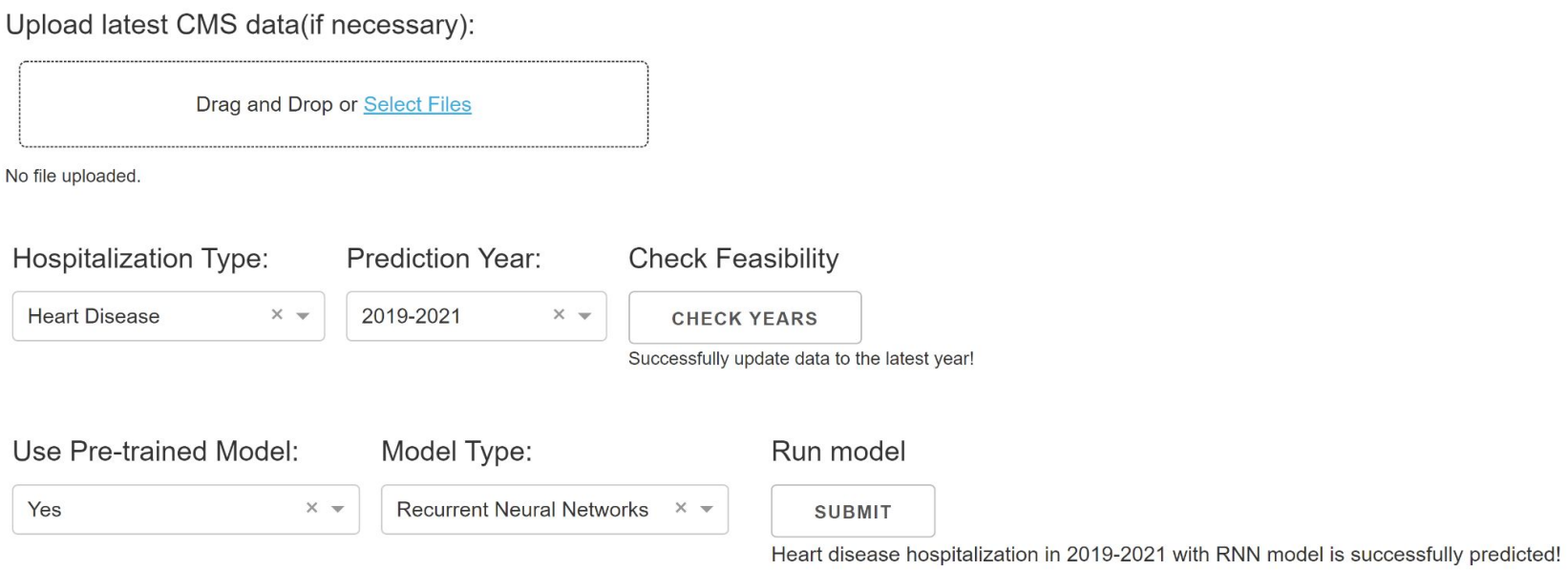
CDC (Centers for Disease Control) Atlas Data:
https://nccd.cdc.gov/DHDSPAtlas/Reports.aspx (link to download data)

CMS (Centers of Medicare and Medicaid Services) Spending and Prevalence Data:
https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CC_Main (link to download data)

Appendix - Aggregate Methods

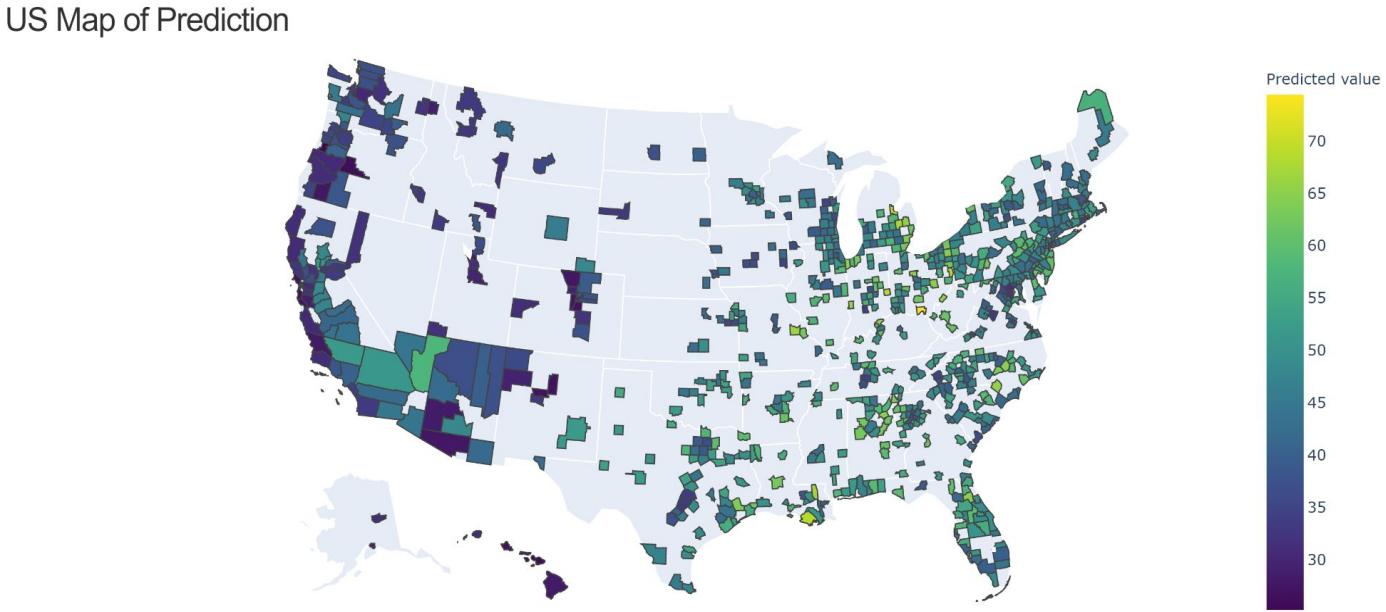
	National Average - County	National Average - State	County	State
Income-ACS	Mean by ALL county	Mean by county of each state then by ALL state	-	Mean by county of state X
Population-ACS	Mean by ALL county	Mean by county of each state then by ALL states	-	Mean by county of state X
Others - ACS	Weighted mean by ALL county	Weighted mean by ALL county	-	Weighted mean by county of state X
Spending	Weighted mean by ALL county	Weighted mean by ALL county	-	Weighted mean by county of state X
Prevalence	Weighted mean by ALL county	Weighted mean by ALL county	-	Weighted mean by county of state X
AQI	Mean by ALL county	Mean by county of each state then by ALL states	-	Mean by county of state X

Appendix - Dashboard



Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

Appendix - Dashboard



Appendix - Dashboard

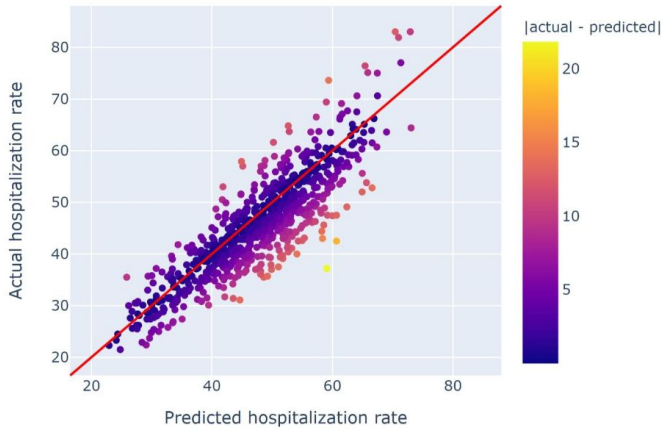
Table of Prediction

DOWNLOAD CSV

	fips	state	county
	filter data...		
	01003	Alabama	Baldwin
	01015	Alabama	Calhoun
	01043	Alabama	Cullman
	01049	Alabama	DeKalb
	01051	Alabama	Elmore
	01055	Alabama	Etowah
	01069	Alabama	Houston
	01073	Alabama	Jefferson
	01077	Alabama	Lauderdale
	01081	Alabama	Lee
	01083	Alabama	Limestone
	01089	Alabama	Madison
	01095	Alabama	Marshall
	01097	Alabama	Mobile
	01101	Alabama	Montgomery

Model Metrics

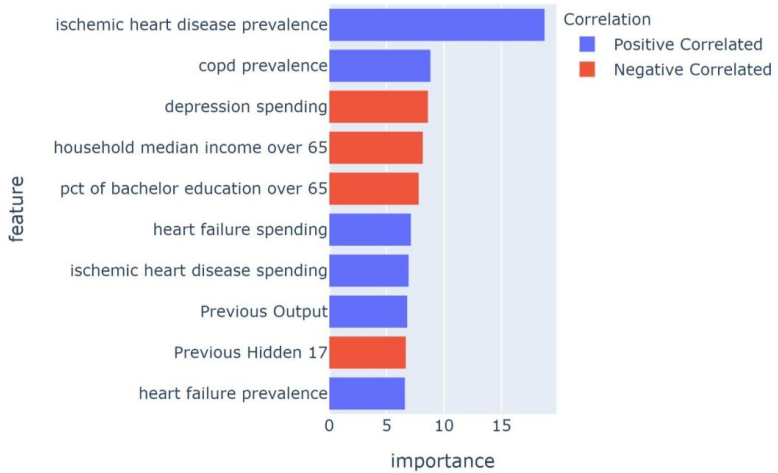
Unadjusted R square for the model is: 0.7



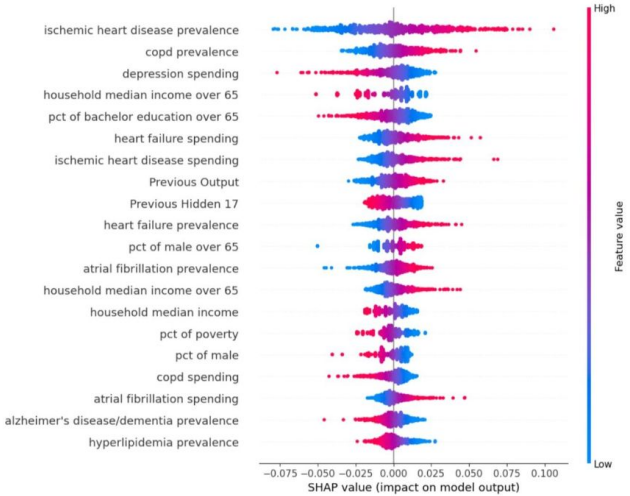
Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

Appendix - Dashboard

Feature Importance Plot



Feature Contribution Plot with SHAP



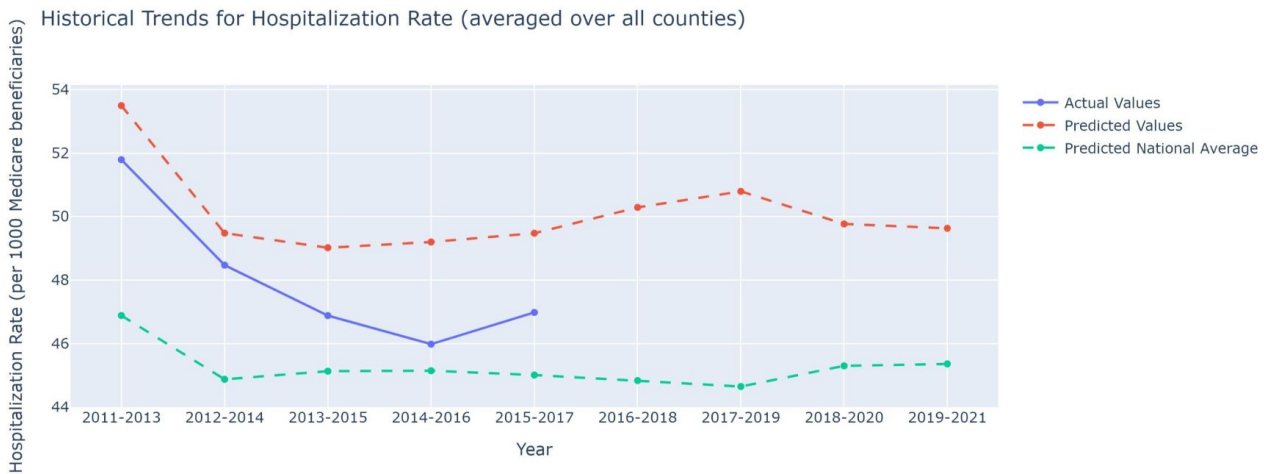
Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

Appendix - Dashboard

State

Oklahoma

Historical Prediction Trends



Appendix - Dashboard

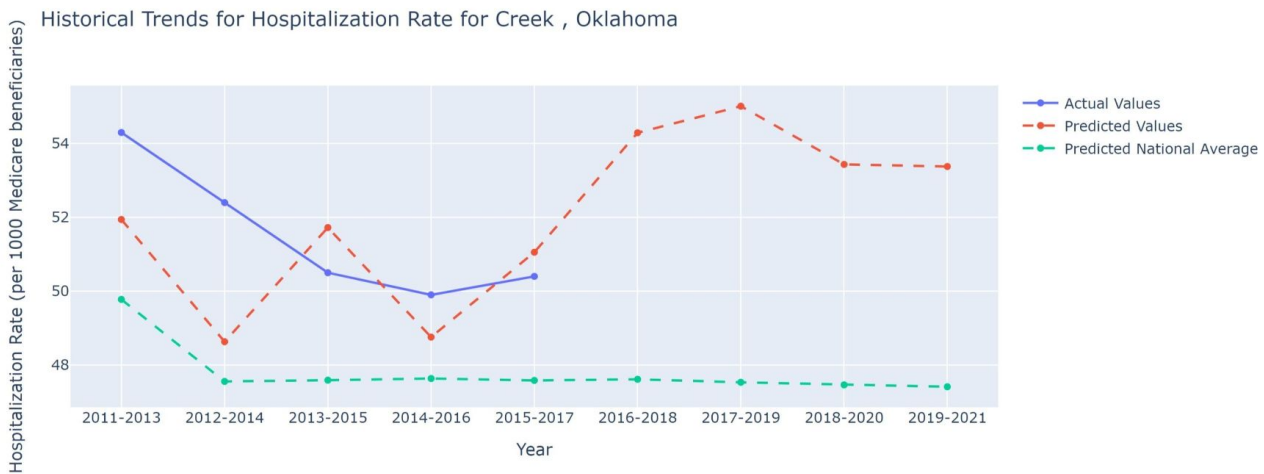
Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

Appendix - Dashboard

State: Oklahoma

County: Creek

Historical Prediction Trends



Carnegie Mellon University HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT × TATA

Appendix - Dashboard