Cardiovascular Disease

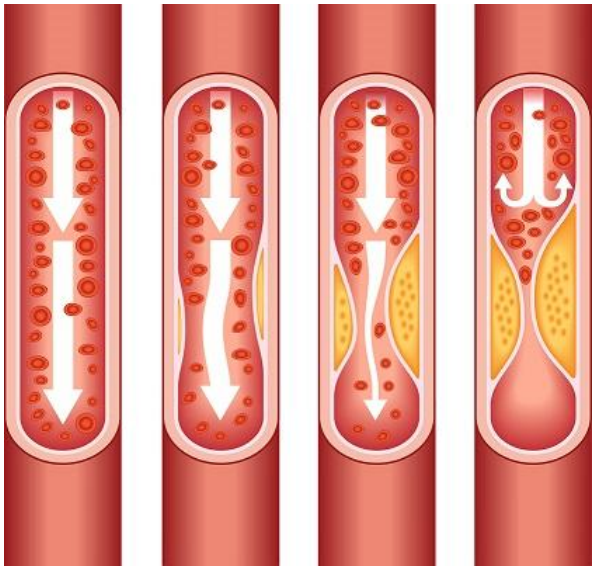
Catherine Clary, Jen Zieger,
Thomaz Moon, & Bassa Belhu

Scenario

- **Situation:** We are public health researchers who have been hired to investigate predictors of heart disease
- **Audience**: Medical professionals who are concerned with the high CVD mortality rates of patients
- **Goal:** Provide guidance to medical professionals on factors most closely correlated with CVD to help doctors identify CVD

What is Cardiovascular Disease

- Group of disorders of the heart and blood vessels
- Examples include:
  - Coronary heart disease
  - Cerebrovascular disease
  - Peripheral arterial disease
  - Rheumatic heart disease
  - Congenital heart disease
- According to the WHO, more than 80% of CVD deaths are due to heart attacks and strokes



Cardiovascular Disease Mortality: United States

- Heart disease is the leading cause of death in the United States
  - Leading cause of death for African American, American Indian, Alaska Native, Hispanic, and white men.
  - Second leading cause of death (behind cancer) for women
- According to the latest releases statistics from the CDC **659,000** people die from heart disease each year
- **Financial Burden:** Heart disease costs the US about $363 billion each year (CDC estimate 2016-2017)

Cardiovascular Disease Mortality: Global

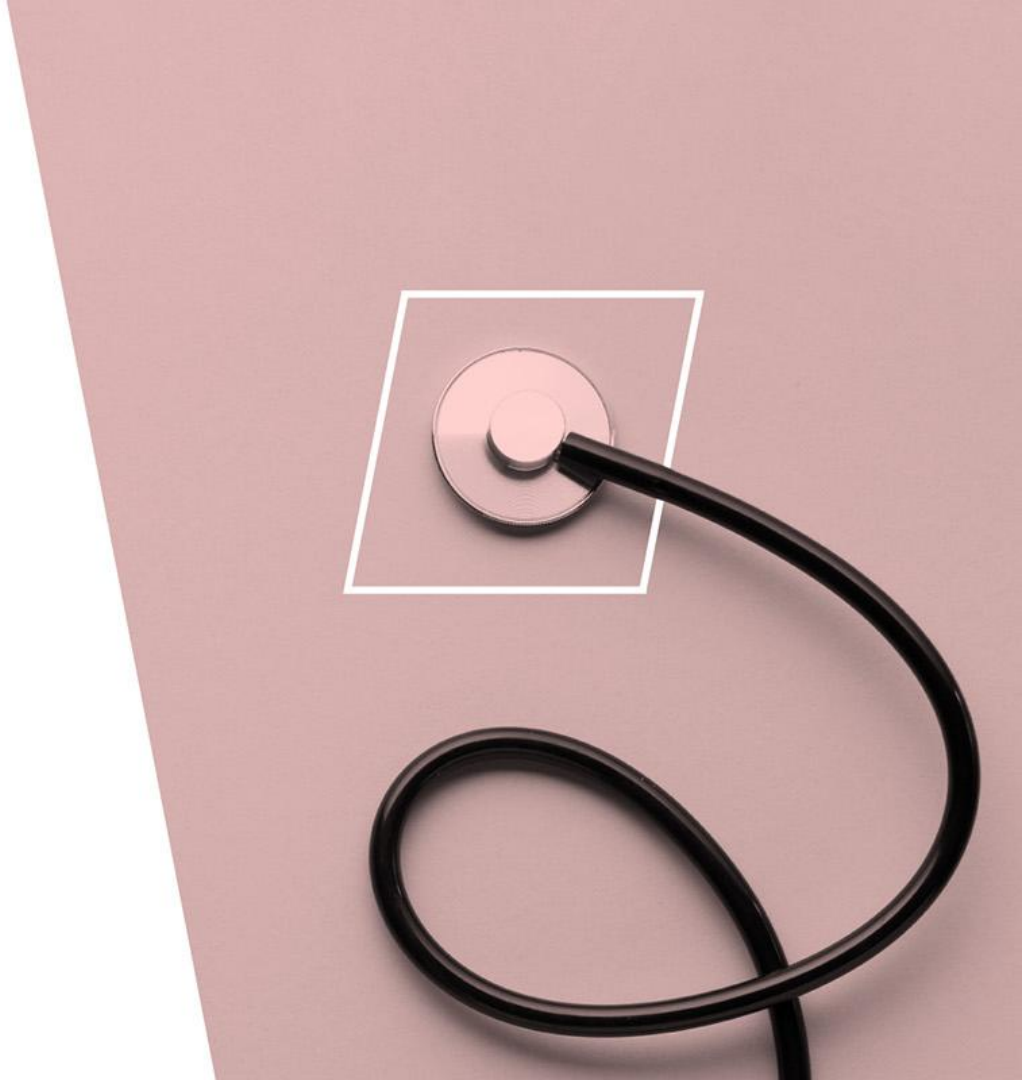
- CVD are also the leading cause of death globally, accounting for **17.9 million** deaths each year (likely underreported)
- Represents **32%** of all deaths worldwide
    - According to the WHO, **38%** of premature deaths were caused by CVDs
- Over **75%** of CVD deaths occur in low and middle income countries

Risk Factors for Cardiovascular Disease

- Past scientific literature has identified the following behavioral risk factors:
  - Unhealthy diet, physical inactivity, tobacco use, & harmful use of alcohol
- We seek to expand upon current knowledge, analyzing which medical predictors are associated with CVD

PROBLEM STATEMENT

What features best predict cardiovascular heart disease?



STRATEGY

1. Data Collection & Cleaning
2. Cleaning & EDA
3. Feature Engineering & Pre-Processing
4. Modeling
5. Evaluation
6. Conclusions & Recommendations

Data Source

- We used the Kaggle “Heart Disease Dataset” comprised of ~1200 observations as our primary dataset
    - Locations: Cleveland, Hungarian, Switzerland, Long Beach VA, Statlog Data Set combined to encompass one overall dataset
    - 11 columns

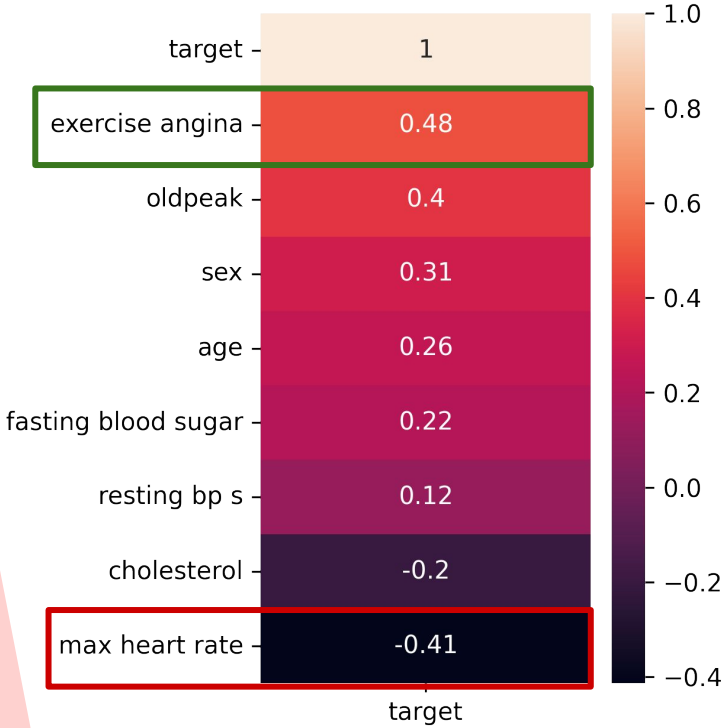
- The CHD dataset had around 72k rows, and 32 columns
    - Location were the U.S. States.

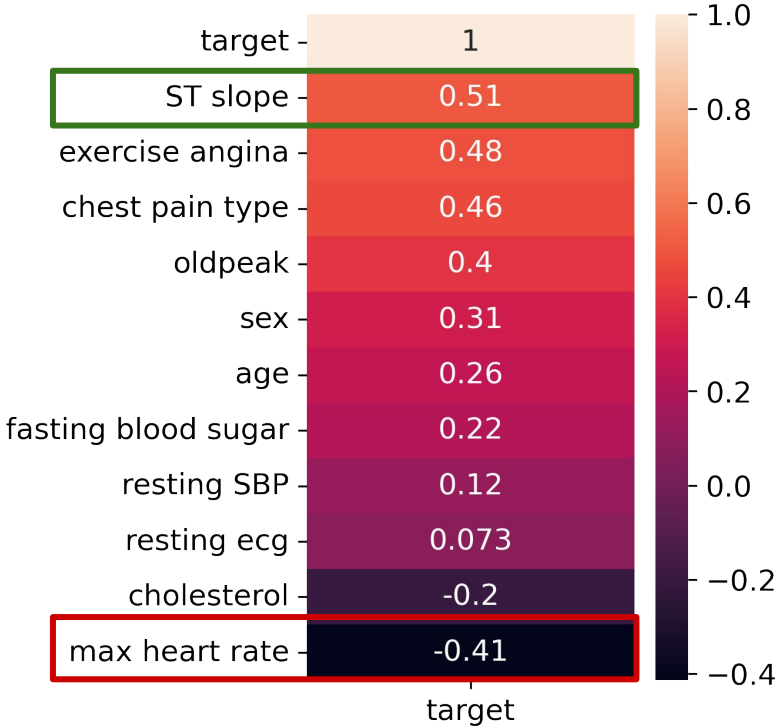
Preparing the Data Frame

Style	Description
Numeric Columns Only	*"Chest pain type", "resting ecg", "ST Slope"* column were made into categorical columns and not used.
Heart w/ Dummy	Same as the previous data frame, but included the categorical columns as well.
Poly Numeric	Did a poly transformation with only the numeric columns
Poly w/ Dummy	Used a Poly transformation on the entire *Heart* data frame
Select Feature Engineering	[ST slope^2], [Chest pain type^2], [ST slope * Chest pain type], [Cholesterol * ST Slope], [Max heart rate * Cholesterol]
Poly of Non - Categorized Data	Polyfit the data set without making the 3 columns into categories
Poly w/ Fewer Columns	Used Poly data from above, however only if their correlation was stronger than 0.20

Best One

Correlation Matrix



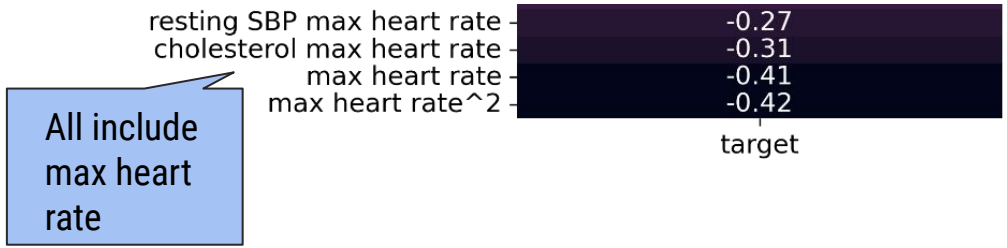


Correlation Matrix: Poly Features

- Other Than ST Slope, all other columns with a correlation of 0.5 or over were from feature engineering

Highest positive correlation ^

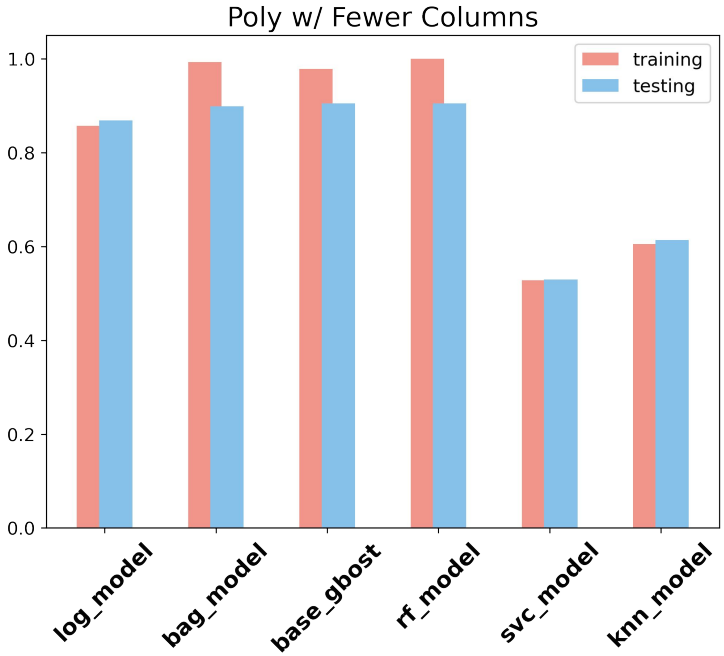
Highest negative correlation

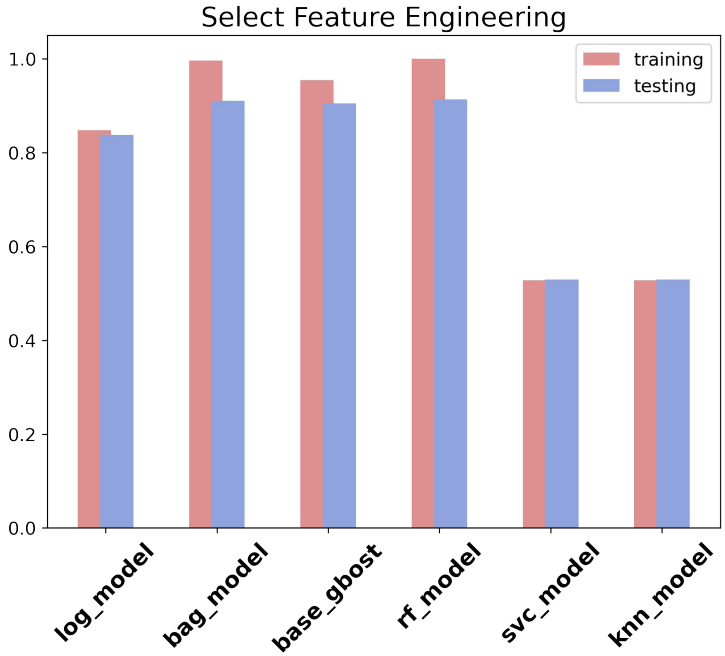


Developing Models

- In order to identify and build the best model, we tested out multiple models:
  - Logistic Regression
  - Bagging Classifier
  - Adaboost Classifier
  - Gradient Boosting
  - Random Forest
  - SVM
  - K-Nearest Neighbors
  - BernoulliNB
  - Decision Tree
  - Neural Network
- Performed Grid Searches on all models to identify best parameters

Top 2 Data Frames





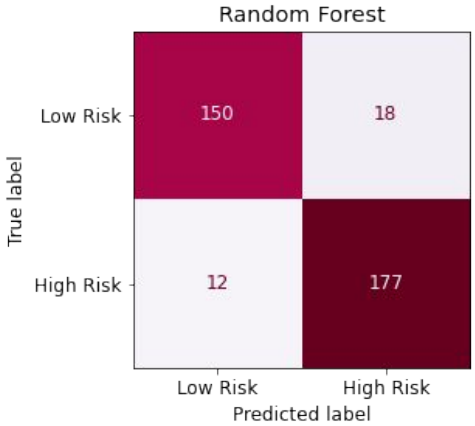
Models

Model	Log Reg	Bagging	Ada Boost	Gradient	RF	SVC	KNN	BNB	DT	NN	Voting
Base Models											
Train Score	0.857	0.992	0.983	0.977	1.0	0.876	0.891	--	--	0.894	--
Test Score	0.868	0.899	0.907	0.902	0.913	0.868	0.840	--	--	0.893	--
After Grid Search											
Train Score	0.858	1.0	0.941	0.991	1.0	0.724	1.0	0.745	1.0	--	0.946
Test Score	0.868	0.916	0.885	0.913	0.916	0.765	0.829	0.742	0.863	--	0.902

Best Models

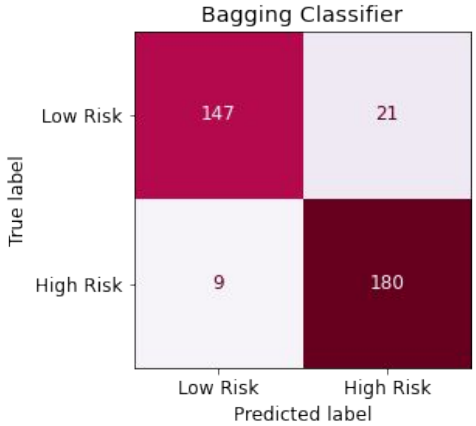
- Our three best performing models for recall were: Bagging Classifier, Gradient Boost, and Random Forest

Random Forest



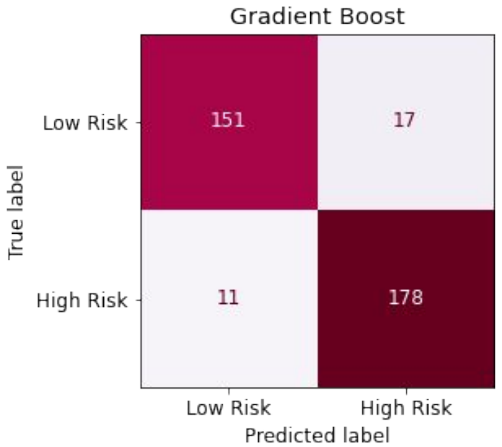
Recall: 0.937
Precision: 0.908

Bagging Classifier



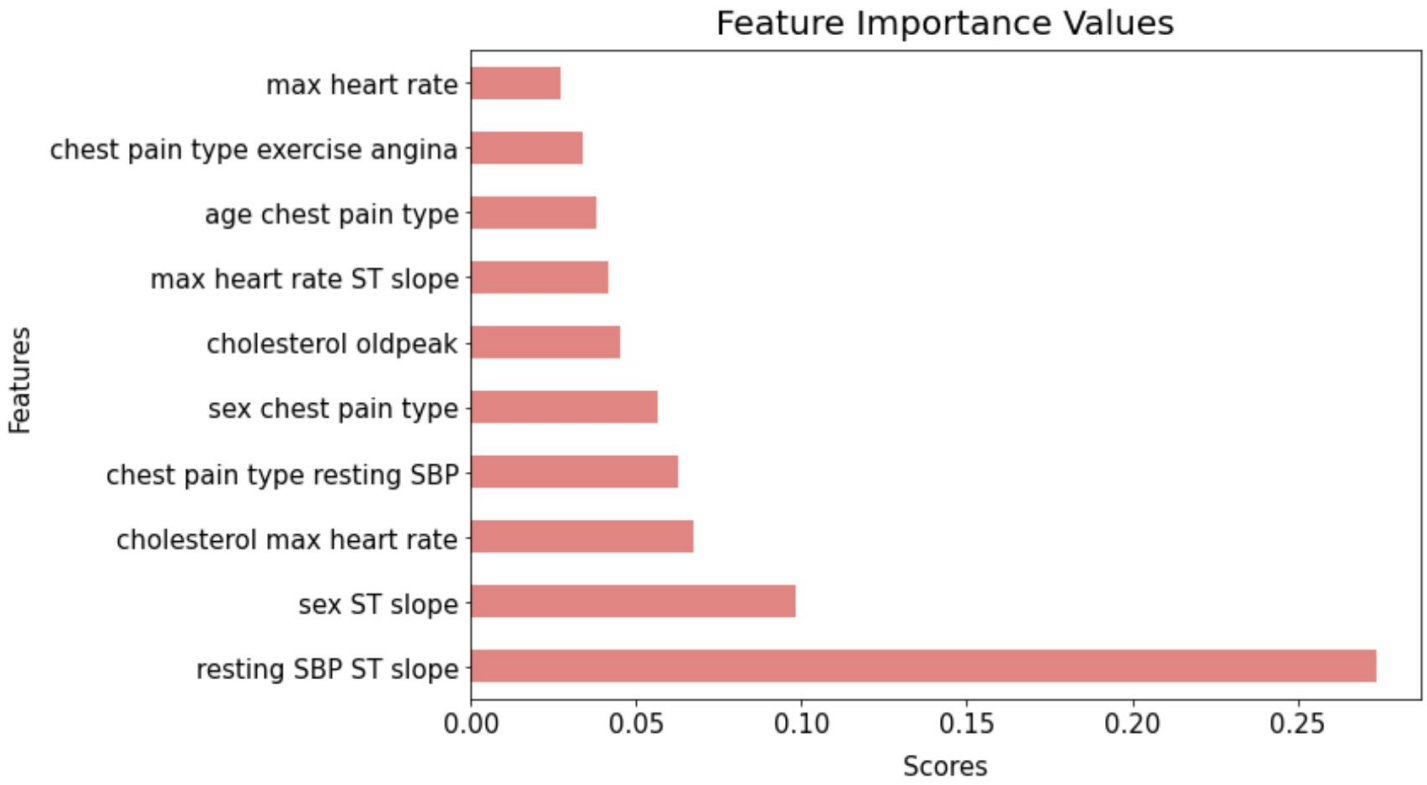
Recall: 0.952
Precision: 0.896

Gradient Boost



Recall: 0.942
Precision: 0.913

Evaluation


Feature Importance Values
max heart rate
chest pain type exercise angina
age chest pain type
max heart rate ST slope
cholesterol oldpeak
sex chest pain type
chest pain type resting SBP
cholesterol max heart rate
sex ST slope
resting SBP ST slope
Features
0.00
0.05
0.10
0.15
0.20
0.25
Scores

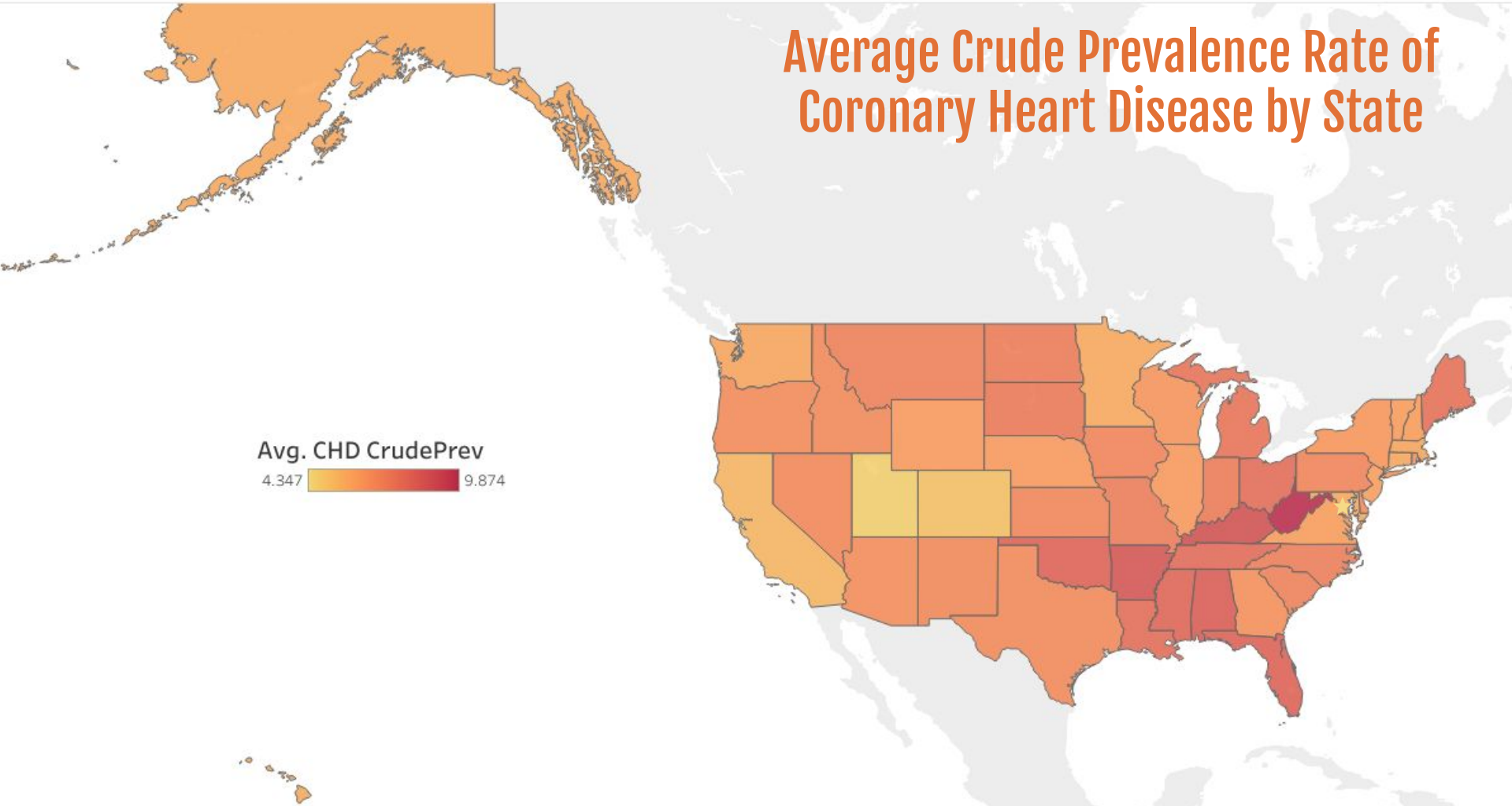
Confirmation of Results

In addition to our main dataset, we wanted to analyze a supplementary dataset to identify the features of a subset of CVD, Coronary Heart Disease (CHD)

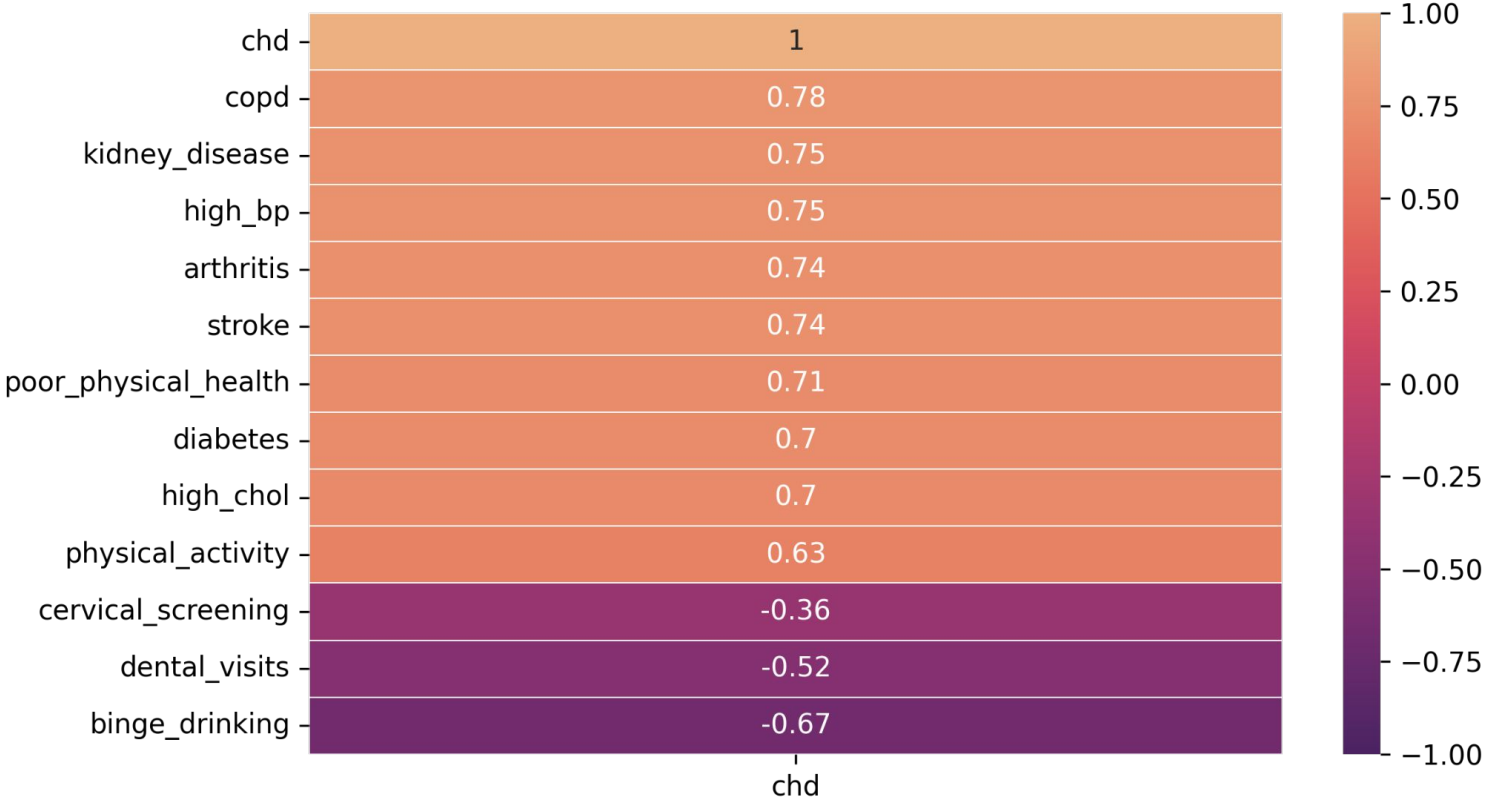
CDC Data Set from the PLACES project in 2020:

- Local health departments and jurisdictions better understand the burden and geographic distribution of health-related outcomes in their areas
- Provides city and census tract estimates for chronic disease risk factors, health outcomes, and clinical preventive services use for the 500 largest US cities

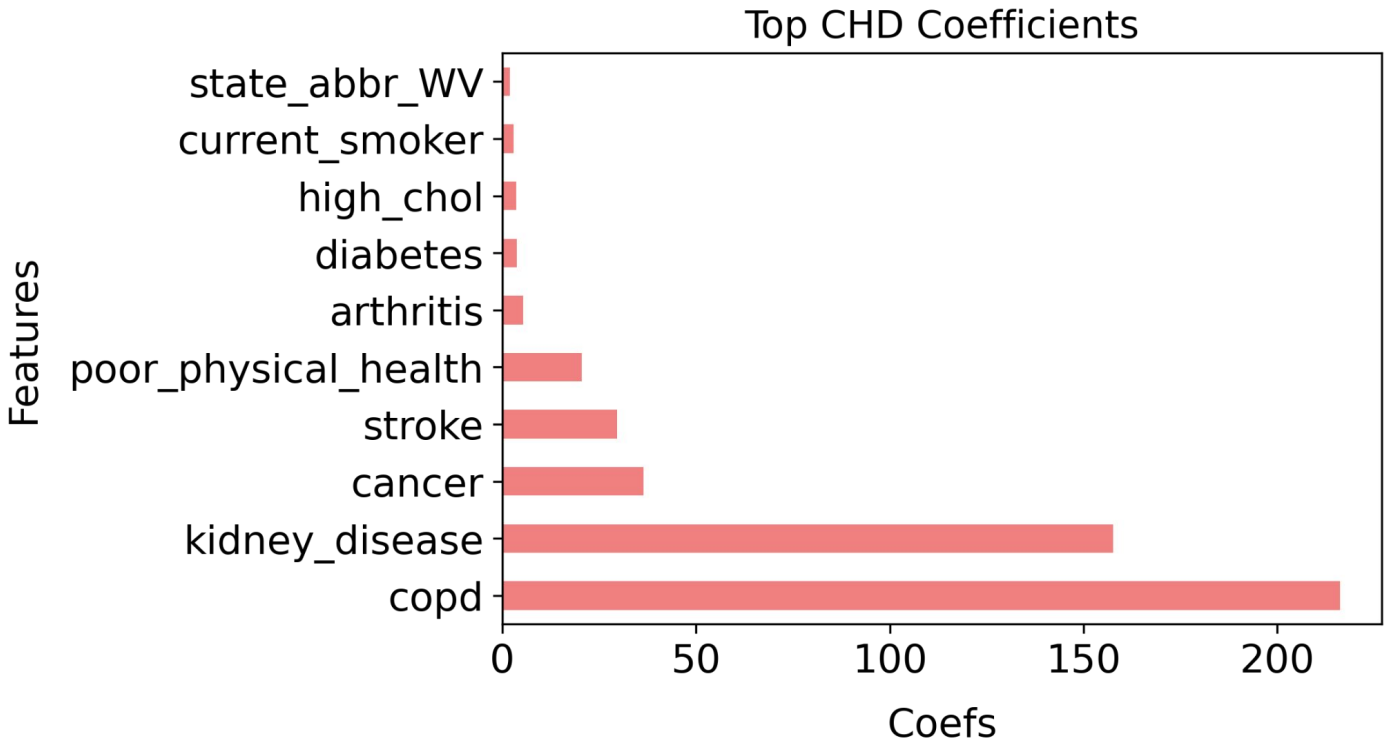
Average Crude Prevalence Rate of Coronary Heart Disease by State



CHD Correlation Matrix



CHD Top Coefficients



CHD Conclusions & Recommendations

- Crude prevalence of Chronic Obstructive Pulmonary Disease, Kidney Disease, and/or Cancer may lead to the increase of the crude prevalence of CHD risk
- West Virginia and Kentucky - higher crude prevalence of CHD risk
- CVD Comparison:
    - Having poor physical health, high cholesterol, and/or high blood pressure may lead to the increase of crude prevalence of CHD risk

Test patients more frequently if they have one of the high risk factors above and marking patient as high-risk for CHD if necessary

CVD Conclusions & Recommendations

- Resting blood pressure * ST Slope is the most important predictor of CVD
  - Take BP and if necessary, mark patient as high risk for CVD
- Sex * chest pain type
  - For all males, confirm chest pain type and mark as high risk for CVD
- Chest pain type * systolic blood pressure
  - Take BP and if necessary, mark patient as high risk for CVD
- Sex * ST slope
  - For all males, check ST slope first and if necessary mark patient as high risk for CVD
- Cholesterol * max heart rate
  - Perform cholesterol test and if necessary, mark patient as high risk for CVD

Next Steps

- Collect data on more diverse populations
  - Age, Sex, Socio-Economic Status, Race, and Location
- Collect data for a larger sample size
- Collect data on other subsets of CVD
  - Cerebrovascular Disease, Peripheral Arterial Disease, Rheumatic Heart Disease, Congenital Heart Disease

After collecting this data, we will need to train our model and fine tune the hyperparameters.

Deploy the model with the best performance to beta testing in hospitals or other health care facilities.

THANKS!

CREDITS: This presentation template was created by **Slidesgo**, incluiding icons by **Flaticon**, and infographics & images by **Freepik.**

Please, keep this slide for attribution.

REFERENCES

- Centers for Disease Control: Heart Disease Facts. https://www.cdc.gov/heartdisease/facts.htm
- World Health Organization: Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
- Amini, M, Zayeri, F. et al. “Trend Analysis of Cardiovascular Disease Mortality, Incidence, and Mortality-to-Incidence Ratio: Results from Global Burden of Disease Study 2017” *BMC Public Health*. 21:401(2021).
- PLACES: Local Data for Better Health. https://www.cdc.gov/places/index.html
- Heart Disease Dataset: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive
- CHD PLACES Dataset: CHD Places Dataset: https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Census-Tract-Data-GIS-Friendly-Format-2020-/yjkw-uj5s/data
- Framingham Risk Score: https://en.wikipedia.org/wiki/Framingham_Risk_Score