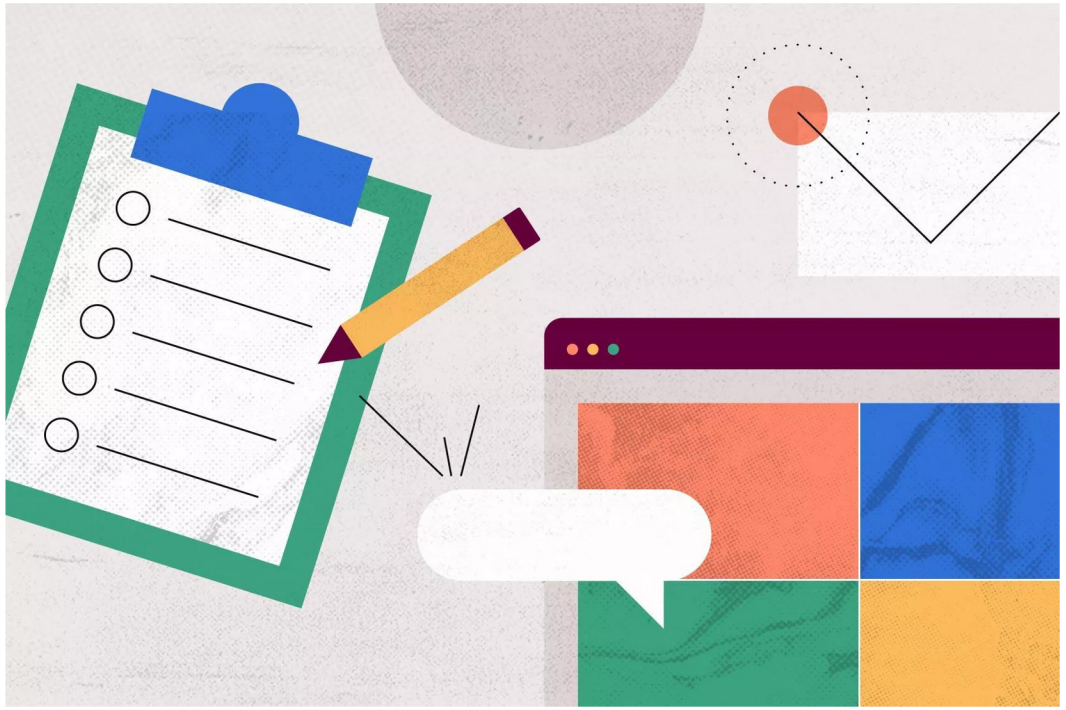
Evaluating Predictors of Heart Disease

Josiah Chung
jchung03@uchicago.edu

Agenda



1. Overview
2. Preparing the Data
3. Methodology and Modeling
4. Conclusions and Recommendations

Overview – Problem Statement

The goal of this project is to **develop a model** that can predict whether someone is at risk for heart disease.

According to the CDC:

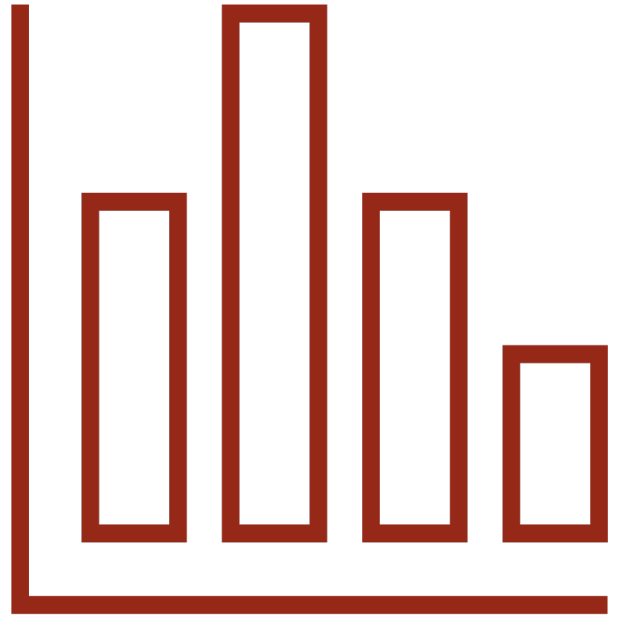
- Heart disease is the leading cause of death in the United States.
- About **695,000** people in the United States died from heart disease in 2021 (**1 in every 5 deaths**).
- Knowing what risk factors contribute to heart disease can help you take the steps to prevent it

Source: https://www.cdc.gov/heartdisease/about.htm

Data source: UCI Machine Learning Repository

Data Cleaning Steps:

- Importing and concatenating the data
- Resolving missing values
- Converting integer labels to categorical values
- Transforming target variable to binary
- Checking for data imbalance



Methodology and Modeling – Comparison of Models

Logistic Regression

- Accuracy: 0.792
- F1: 0.805
- AUC: 0.795

Random Forest

- Accuracy: 0.844
- F1: 0.864
- AUC: 0.839

XGBoost

- Accuracy: 0.818
- F1: 0.829
- AUC: 0.822

Methodology and Modeling – Limitations

Random Forest (Validate)
• Accuracy: 0.844
• F1: 0.864
• AUC: 0.839

Random Forest (Test)
• Accuracy: 0.766
• F1: 0.82
• AUC: 0.739

Overfitting?

- Random Search
- Model Choice
- Sample Size

Conclusions and Recommendations

Choosing a Model

- AUC
- Computational Efficiency
- Overfitting
- Explainability

Areas for Improvement

- Larger Dataset – total of about 400 rows of data, closer to 370 after resolving nulls
- Including more features – Exploring race, socioeconomic status, smoking history, etc.
- Feature selection – L1/L2 logistic regression, stepwise selection
- Exploring more model types – SVM's, KNN's, maybe neural networks

Thank You

6. Appendix

Appendix – Github

- https://github.com/josiah-chung/heart_disease/tree/main

Appendix – Data features

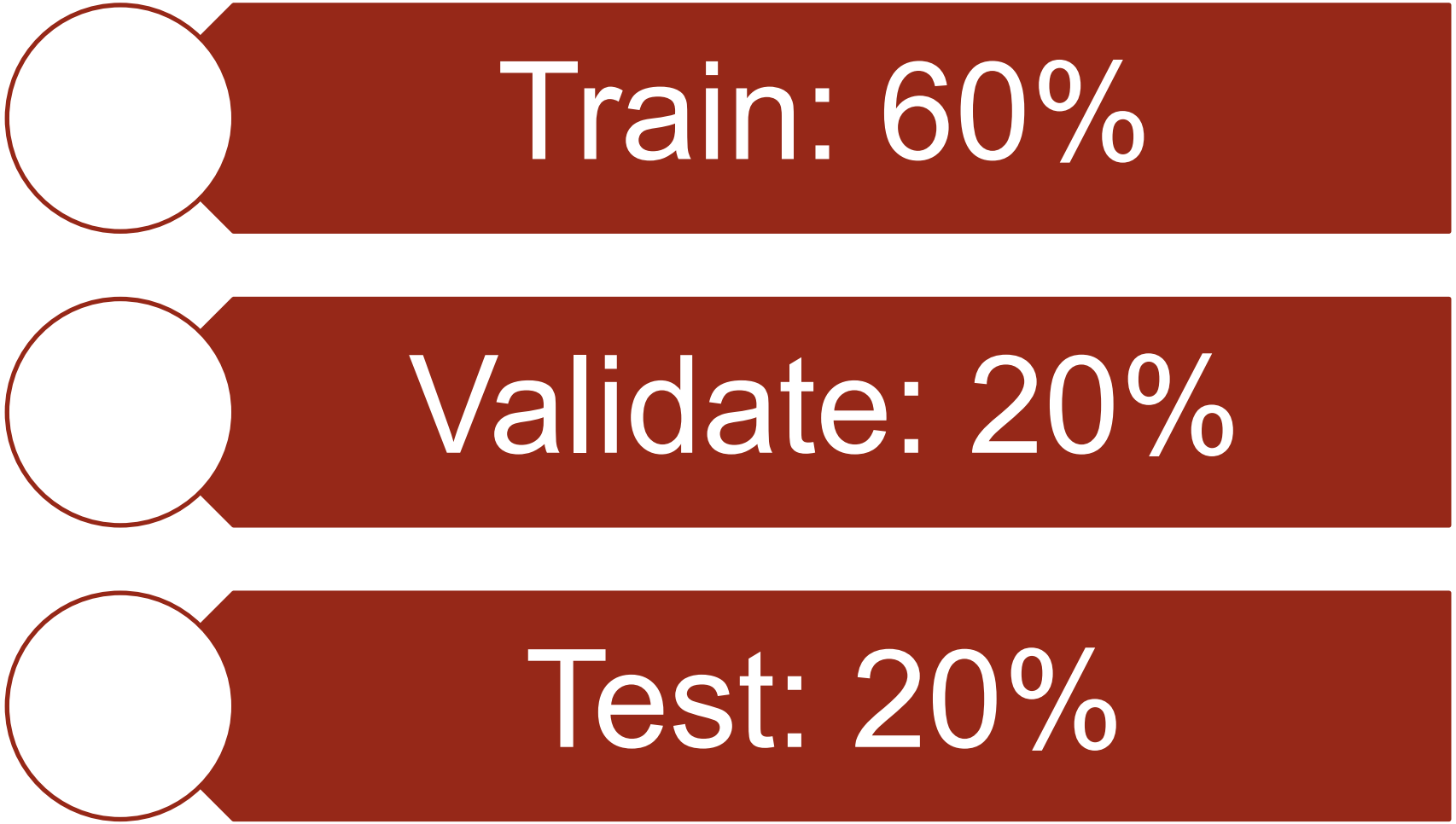
Columns in data:

- `age` : age in years
- `sex` : sex (1 = male; 0 = female)
- `cp` : chest pain type
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
- `trestbps` : resting blood pressure (in mm Hg on admission to the hospital)
- `chol` : serum cholesterol in mg/dl
- `fbs` : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- `restecg` : resting electrocardiographic results
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- `thalach` : maximum heart rate achieved
- `exang` : exercise induced angina (1 = yes; 0 = no)
- `oldpeak` : ST depression induced by exercise relative to rest
- `slope` : the slope of the peak exercise ST segment
  - Value 1: upsloping
  - Value 2: flat
  - Value 3: downsloping
- `ca` : number of major vessels (0-3) colored by flourosopy
- `thal` : 3 = normal; 6 = fixed defect; 7 = reversable defect
- `num` : diagnosis of heart disease (angiographic disease status)
  - Value 0: < 50% diameter narrowing
  - Value 1, 2, 3, or 4: > 50% diameter narrowing

Appendix – Final Clean Data

	age	sex	cp	trestbps	chol	restecg	thalach	exang	oldpeak	num
0	40	M	atypical angina	140.0	289.0	normal	172.0	No	0.0	0
1	49	F	non-anginal pain	160.0	180.0	normal	156.0	No	1.0	1
2	37	M	atypical angina	130.0	283.0	ST-T abnorm	98.0	No	0.0	0
3	48	F	asymptomatic	138.0	214.0	normal	108.0	Yes	1.5	1
5	39	M	non-anginal pain	120.0	339.0	normal	170.0	No	0.0	0
6	45	F	atypical angina	130.0	237.0	normal	170.0	No	0.0	0
7	54	M	atypical angina	110.0	208.0	normal	142.0	No	0.0	0
8	37	M	asymptomatic	140.0	207.0	normal	130.0	Yes	1.5	1
9	48	F	atypical angina	120.0	284.0	normal	120.0	No	0.0	0
10	37	F	non-anginal pain	130.0	211.0	normal	142.0	No	0.0	0

Appendix – Train Test Splitting



Random Forest

```
params = {
    'n_estimators': [10, 50, 100],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2],
    'bootstrap': [True, False],
    'max_features': ['auto', 'sqrt']
}
```

XGBoost

```
params = {
    'n_estimators': [50, 100, 200],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [3, 4, 5],
    'gamma': [0, 0.1],
    'colsample_bytree': [0.6, 0.7],
    'subsample': [0.6, 0.7]
}
```