Predicting Heart Disease

ML Modeling Project

Miguel Ayala, Madeline Couse, Marinus de Beer, Mihir Garikiparithi, Amir Landage

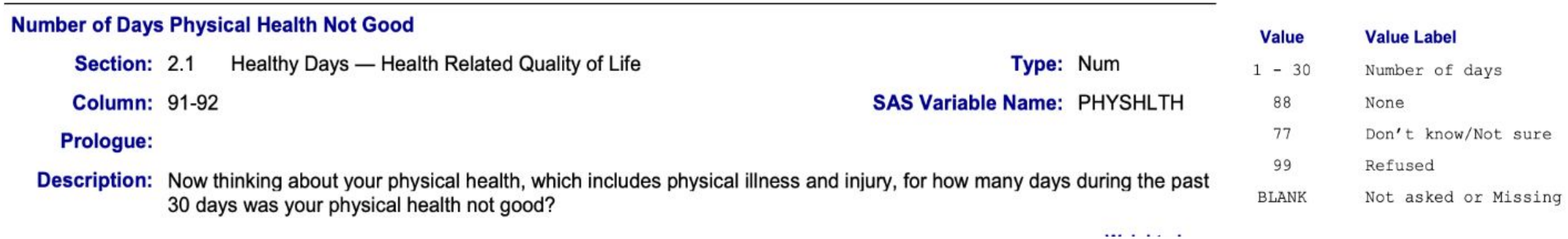
Heart disease is the leading cause of death in the United States

- Heart disease cost the US about $239.9 billion in 2018/2019
    - Lifestyle interventions for those at risk may reduce healthcare system burden
- **The Behavioral Risk Factor Surveillance System (BRFSS) Dataset**
    - Yearly telephone survey of ~400,000 US individuals
    - Health-related information, including if a respondent has heart disease

Objective: build a binary classifier to predict heart disease based on the BRFSS survey data

Data pre-processing

- **327 features** including high cholesterol/blood pressure, BMI, diet, exercise, income, race
- Excluded features with > 30% missing values or those clearly unrelated to heart disease -> **109 features -> 20 top features**
- Dataset balancing:
  - **38,633 with heart disease**
  - 398,881 without heart disease -> **38,633 without heart disease**



Number of Days Physical Health Not Good

Section: 2.1 Healthy Days — Health Related Quality of Life

Type: Num

Column: 91-92

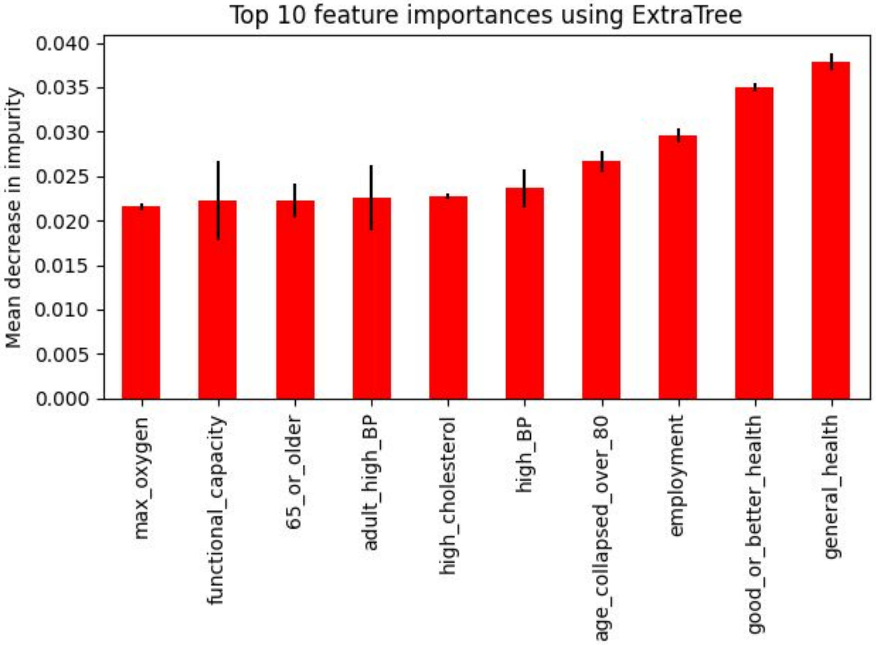
SAS Variable Name: PHYSHLTH

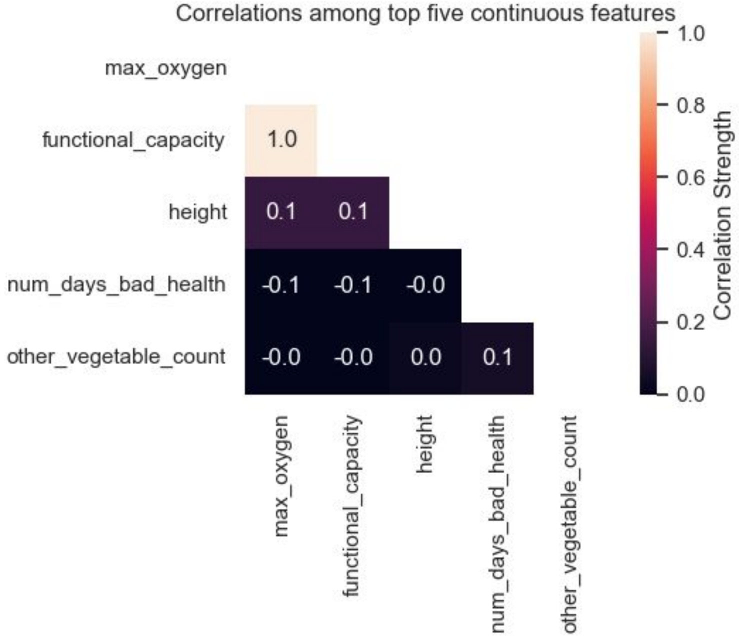
Prologue:

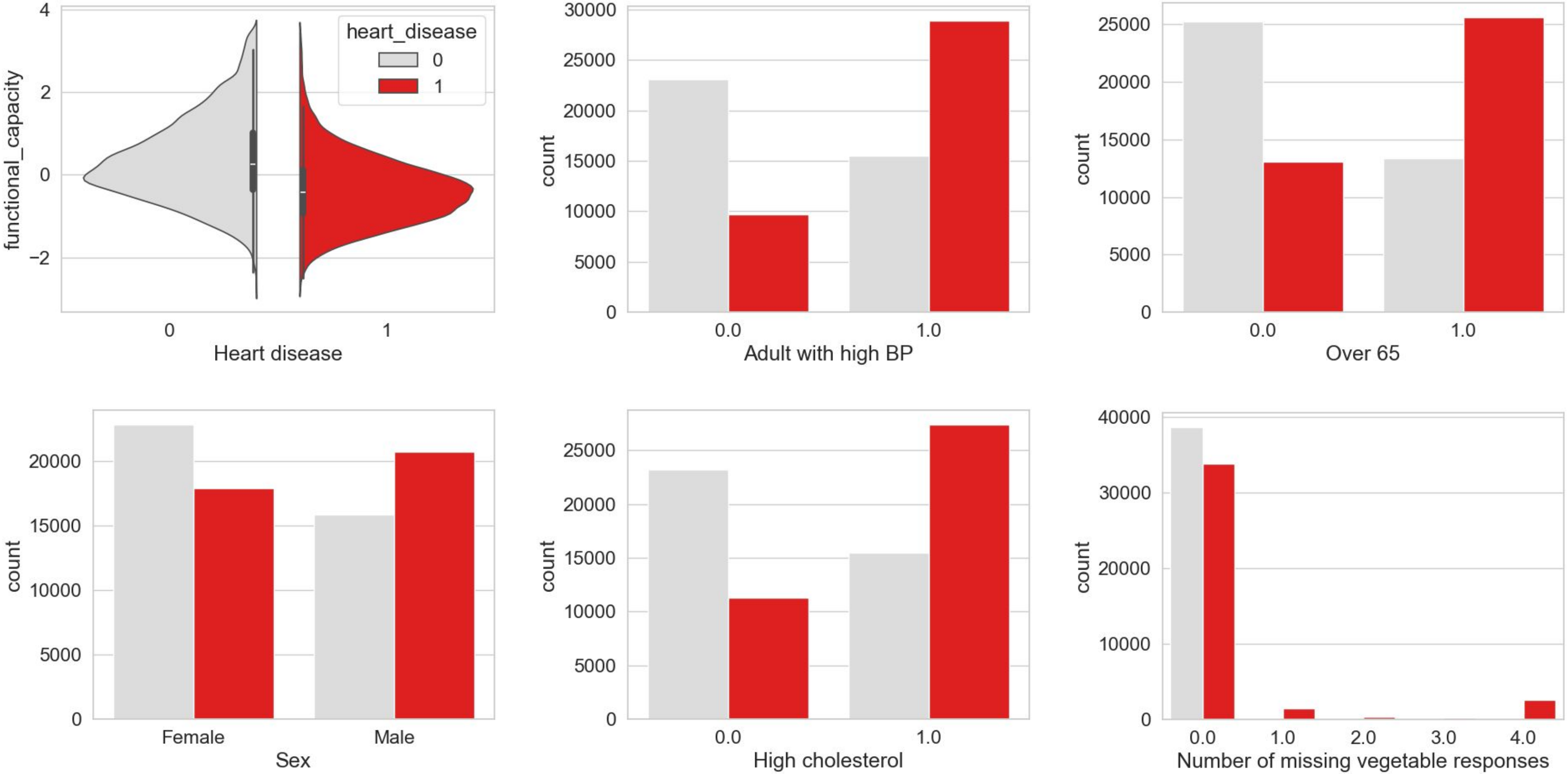
Description: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

Value	Value Label
1 - 30	Number of days
88	None
77	Don’t know/Not sure
99	Refused
BLANK	Not asked or Missing

Exploratory data analysis





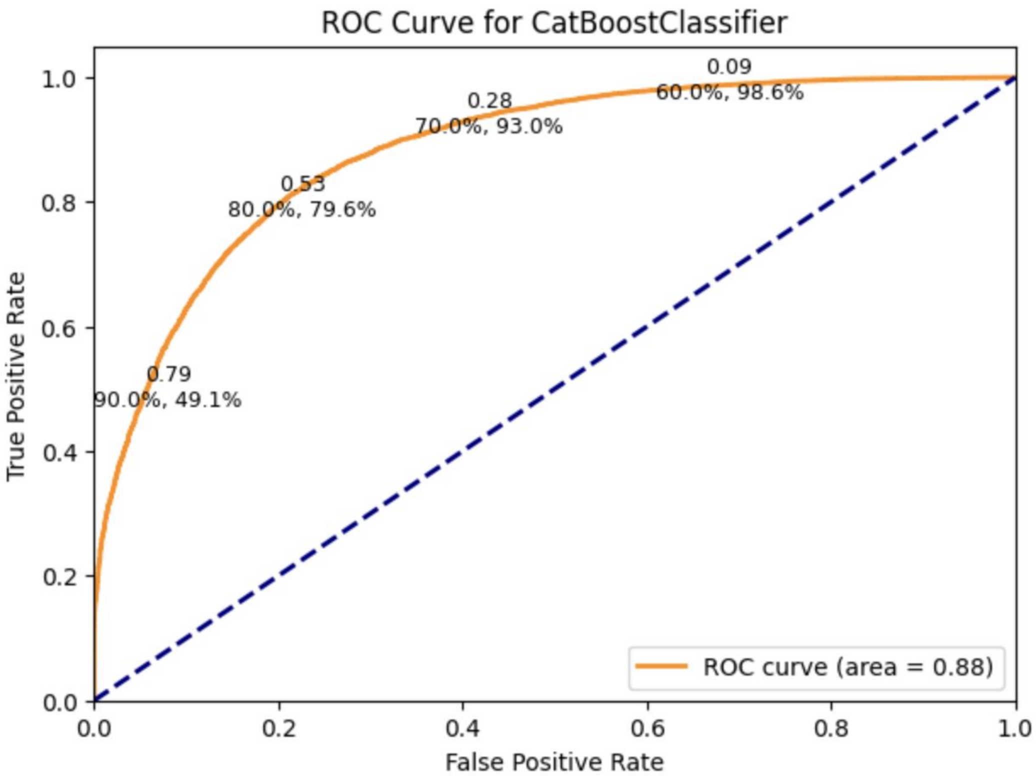

heart_disease
0
1
functional_capacity
4
2
0
−2
0
1
Heart disease
count
30000
25000
20000
15000
10000
5000
0
0.0
1.0
Adult with high BP
count
25000
20000
15000
10000
5000
0
0.0
1.0
Over 65
count
20000
15000
10000
5000
0
Female
Male
Sex
count
25000
20000
15000
10000
5000
0
0.0
1.0
High cholesterol
count
40000
30000
20000
10000
0
0.0
1.0
2.0
3.0
4.0
Number of missing vegetable responses

Training & Fine-Tuning

- CatBoost Classifier
- GradientBoosting Classifier
- Random Forest Classifier
- AdaBoost Classifier
- ExtraTrees Classifier

Results

Precision	70%
Recall	93%
F1_Score	80%



Conclusion

- The performance of our model is reasonable given the complex etiology of heart disease
- It could be utilized in a few ways
  - Early warning tool for a doctor
  - Incorporated into a system to inform insurance eligibility
  - Inform governmental programs
- Limitations:
  - We only used data from one survey year (2015)
  - Survey lacked some lifestyle details, e.g. drug use

Hope you eat your Veggies!