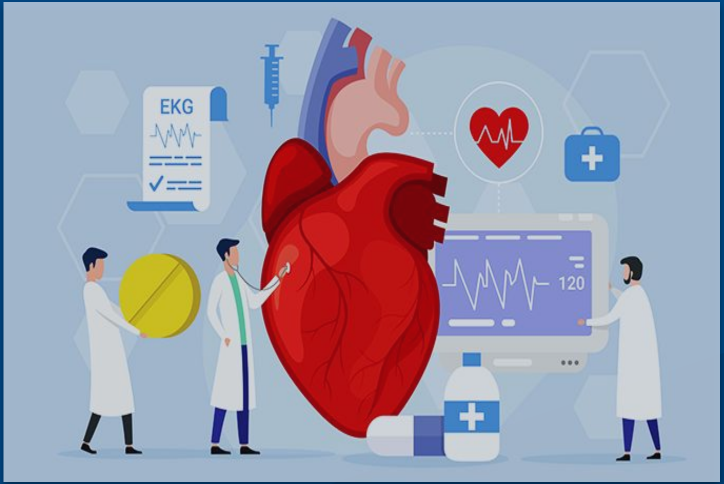
Life-Saving Data Models

Apr 29, 2022

Muhua Chen, Yanqi Liu, Tianxiao Zhang



Who We Are

Muhua Chen

SEAS DATS ‘22

Yanqi Liu

SEAS DATS ‘22

Tianxiao Zhang

SEAS DATS ‘22

Agenda

Project Background

❖ Goal of Study

❖ Data Intro

❖ Evaluation Metric

Exploratory Analysis

❖ Distributions

❖ Variable Relationships

Model Analysis

❖ Logistic Regression

❖ LASSO Regression

❖ Tree-Based Models

Wharton
UNIVERSITY *of* PENNSYLVANIA

Project Background

-- Tianxiao Zhang

Goal of Study

- Heart disease is the leading cause of death in the US
  - Accounts for more than 20% of deaths in most racial groups
- It requires timely diagnosis and treatment
- We want to identify potential heart disease in advance
  - Predict whether a person is likely to have heart disease (binary outcome) given other physical measures of the person that could be tracked earlier

Link: https://www.cdc.gov/heartdisease/facts.htm

Data Information

- Obtained from Kaggle.com
- 18 Variables and 320,000 observations
  - Physical measures & Other disease history
- Clean data & no missing values
- Imbalanced outcome distribution
  - Over 90% without heart disease, only 9% with heart disease
  - Could lead to biased predictions
  - Data resampling (downsample the majority group)
  - 55,000 observations after resampling

91% Accuracy

Data Source: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download

Evaluation Metric (for Model)

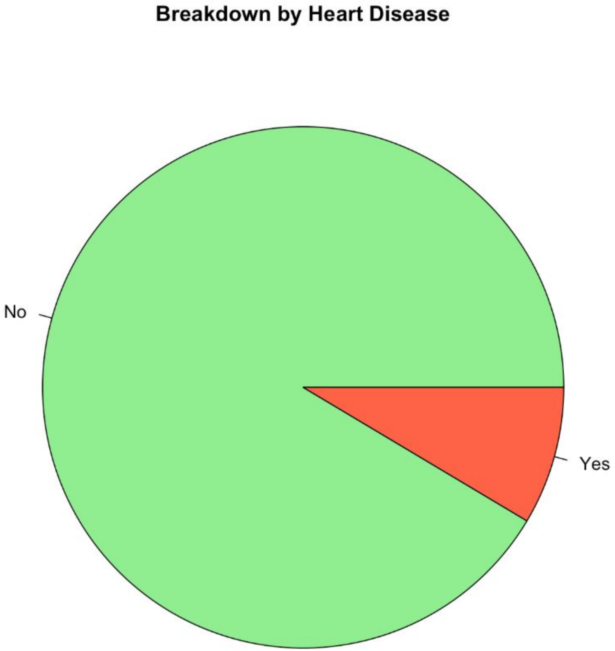
- Usually use Accuracy and F-1 score
- Recall is the most relevant metric for our data
  - Recall = TP / (TP + FN)
    - Recall is the largest when FN is minimized
  - The cost of FN is much higher than the cost of FP
    - FN means unable to identify the patient who will actually get heart disease
      - Miss the best treatment time
    - FP means false alarm when the patient will not actually get heart disease
      - Go to doctors for preventative measures

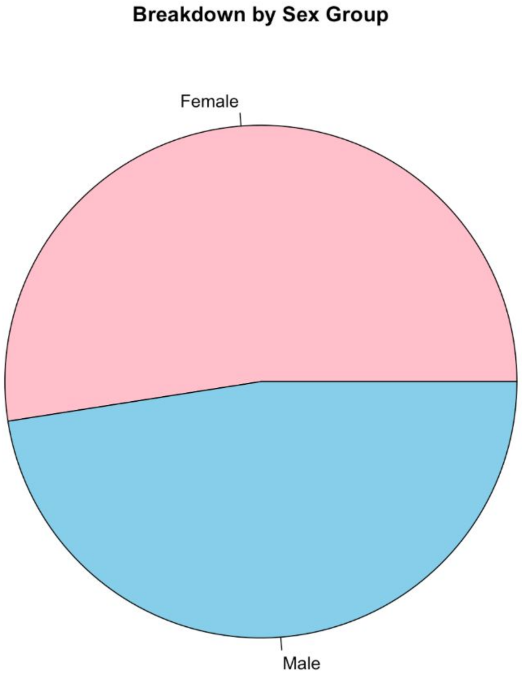
Exploratory Analysis

-- Yanqi Liu

A glimpse of the data

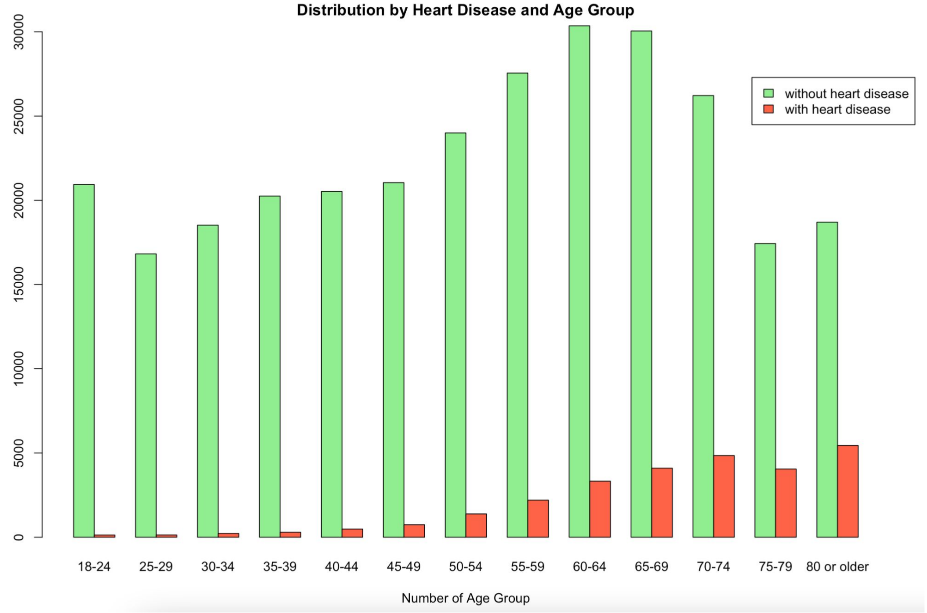
- The gender distribution is quite even with slightly more females than males.


Breakdown by Heart Disease


Breakdown by Sex Group

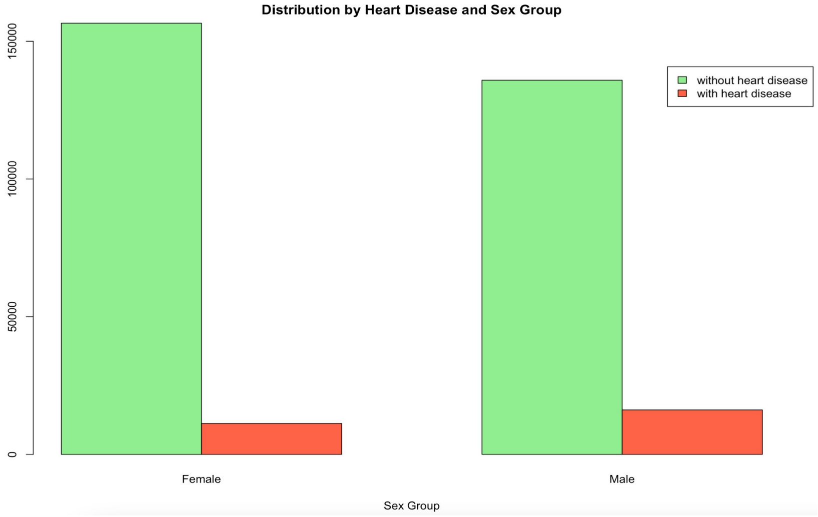
The relationship with age group

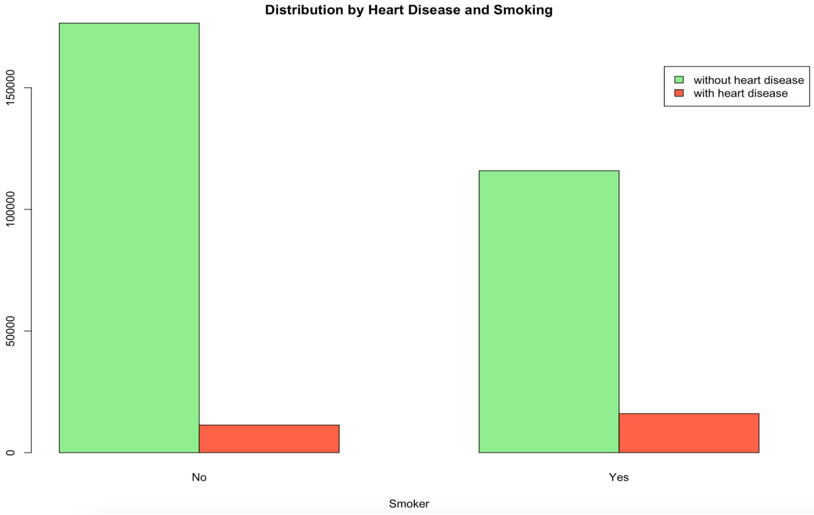
- Older people (> 60) have higher probability of getting heart disease



The relationship with gender and smoking history

- Males & Smokers are more likely to have heart disease

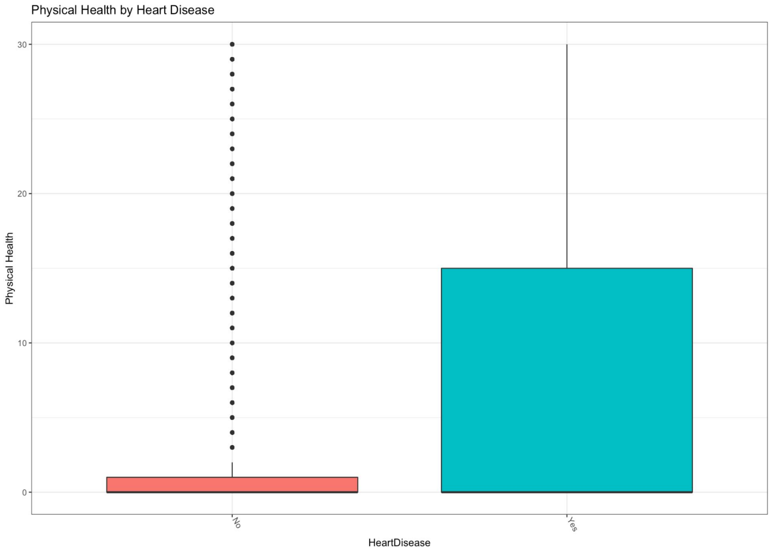


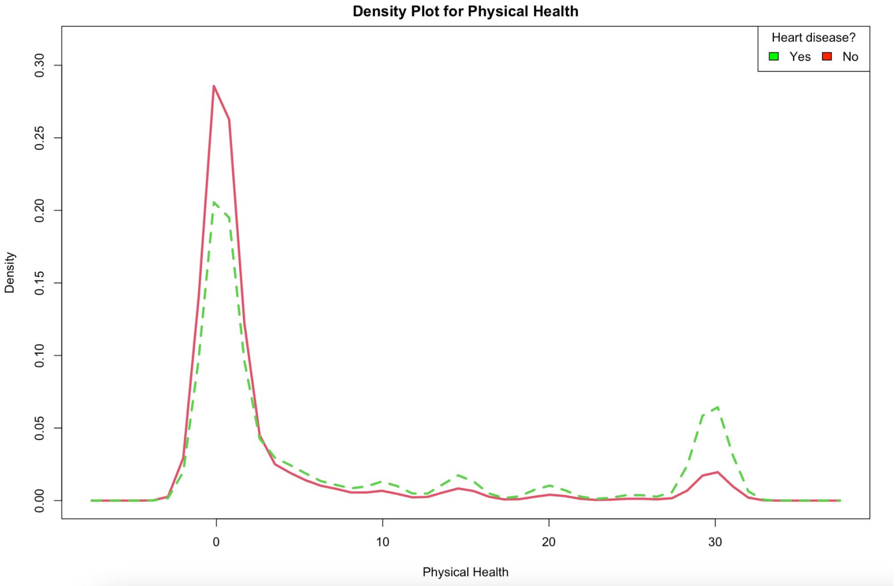


- No significant relationship between heart disease and other disease

The relationship with physical discomfort

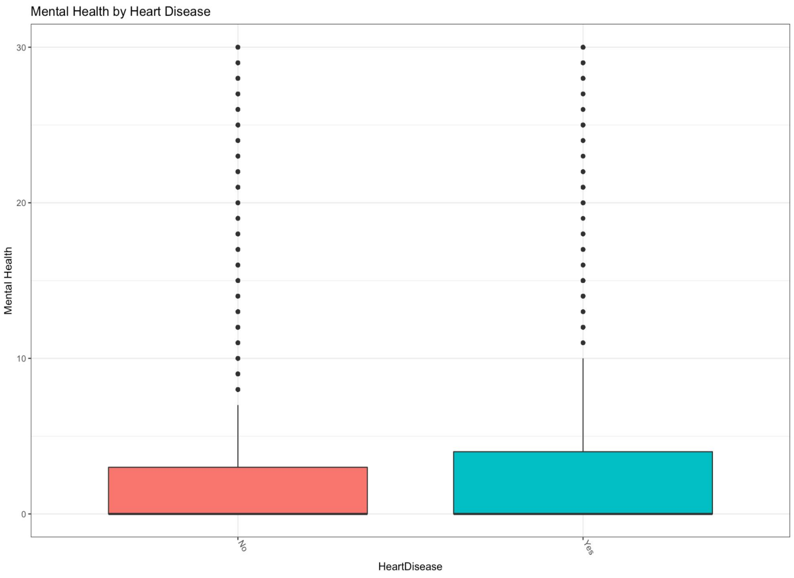
❖ Heart disease patients reported significantly more days of physical discomfort

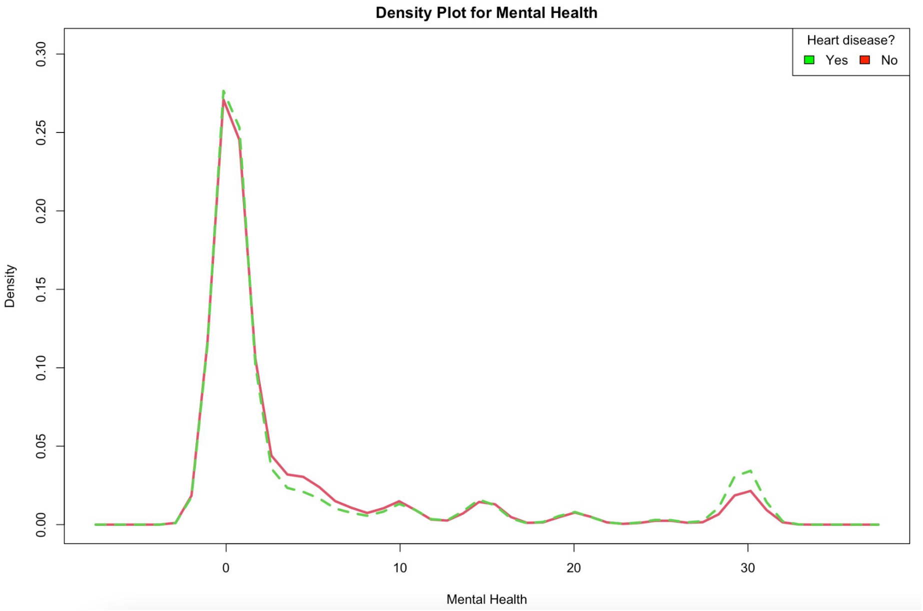

Physical Health by Heart Disease


Density Plot for Physical Health

The relationship with mental discomfort

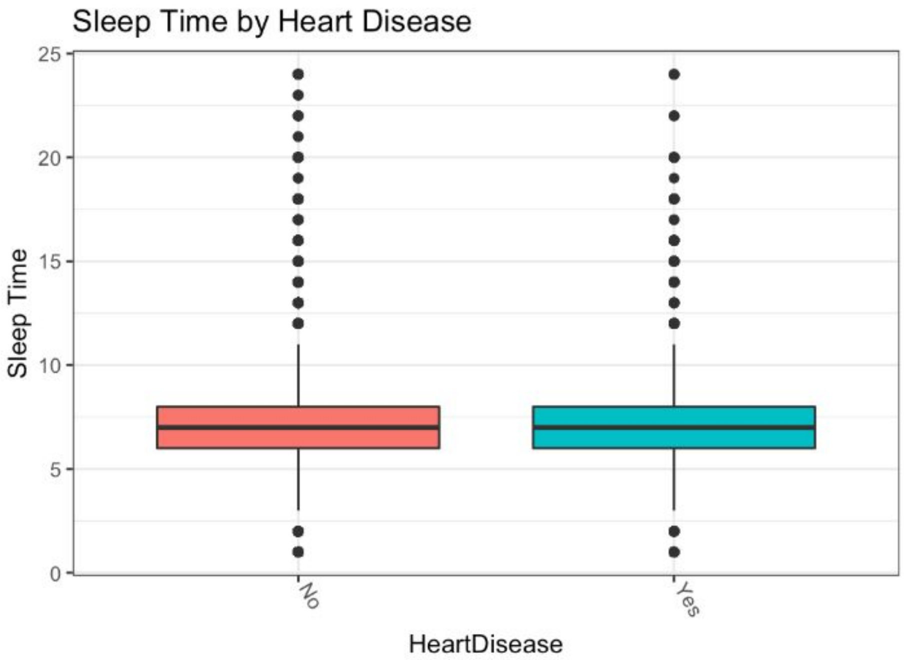
- Heart disease patients reported slightly more days of mental discomfort

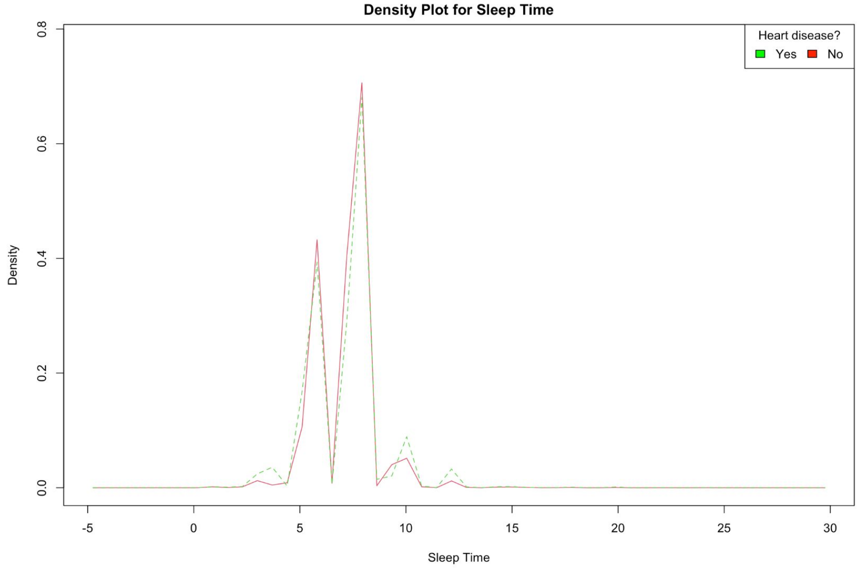




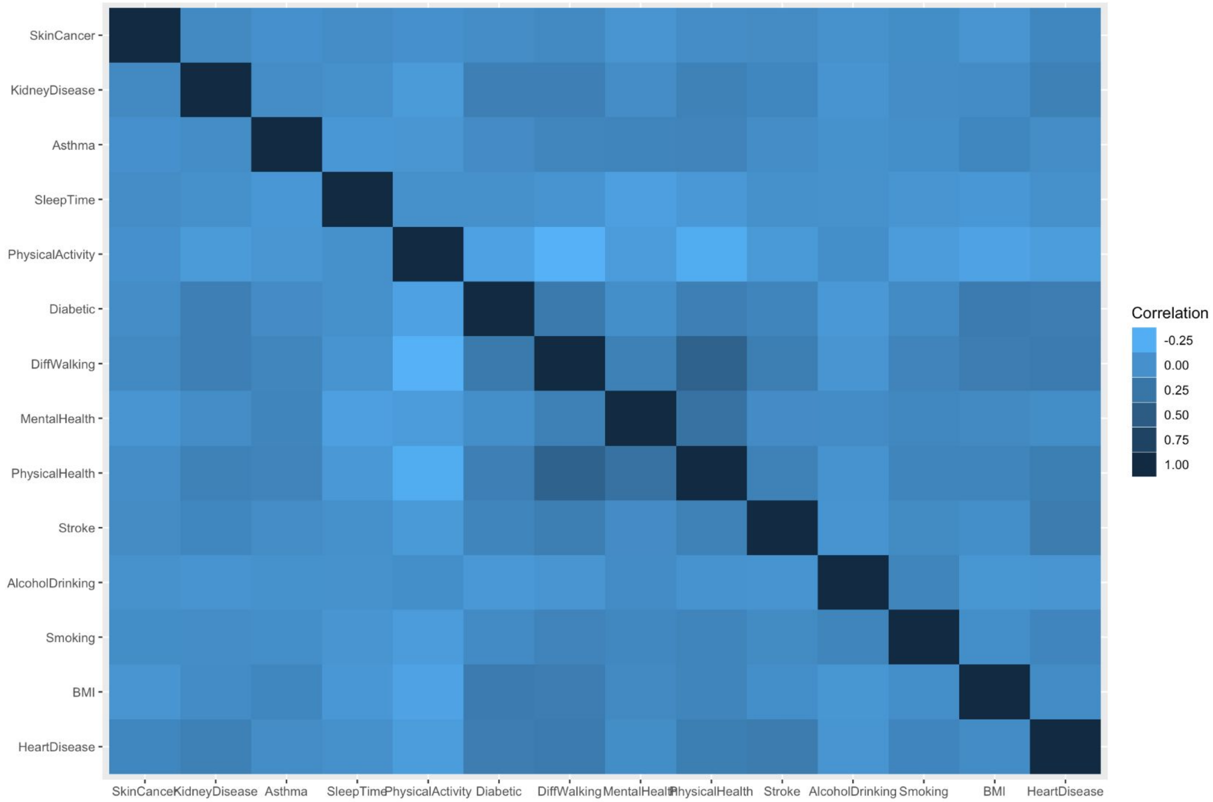
The relationship with sleep time

- Having good sleeping routine/habit does not keep someone away from heart disease.





Correlation between variables

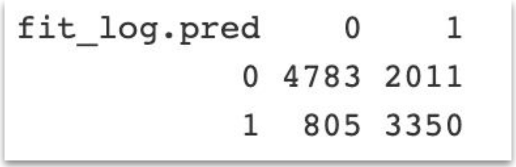


Model Analysis

-- Muhua Chen

Logistic Regression

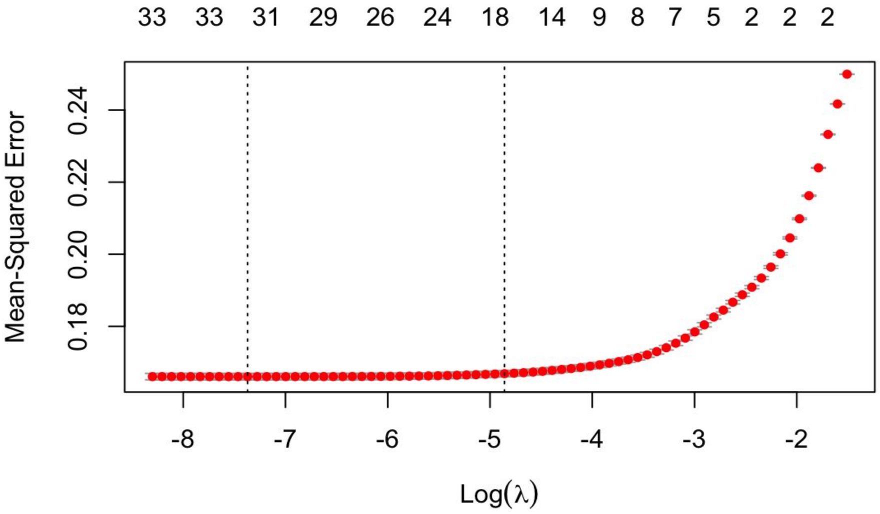
- Backward Elimination
  - Remove the most insignificant variable each time (largest P-value)
  - Final models contains variables significant at 0.05 level
- Important Features
  - **Physical features** such as age and sex
  - **Existing health conditions** such as stroke, generic health, asthma and kidney disease
  - **Lifestyle Habits** such as smoking
- Model Performance
  - **Accuracy**: 0.74; **Recall**: 0.86, **F1-score**: 0.77

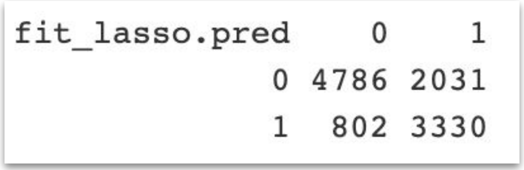


```
fit_log.pred    0    1
           0 4783 2011
           1  805 3350
```

LASSO Regression

- Choosing Lambda
  - Lambda.min and Lambda.1se are similar in terms of error
    - Chose lambda.1se for a more parsimonious model
  - All variables are significant, no need for further backward elimination
- Difference from first model
  - Removes AlcoholDrinking, MentalHealth, and SleepTime
- Model Performance
  - **Accuracy**: 0.74; **Recall**: 0.86, **F1-score**: 0.77

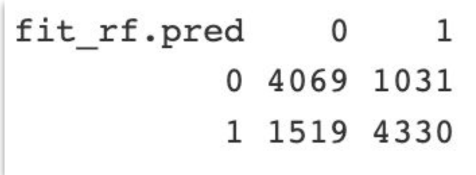




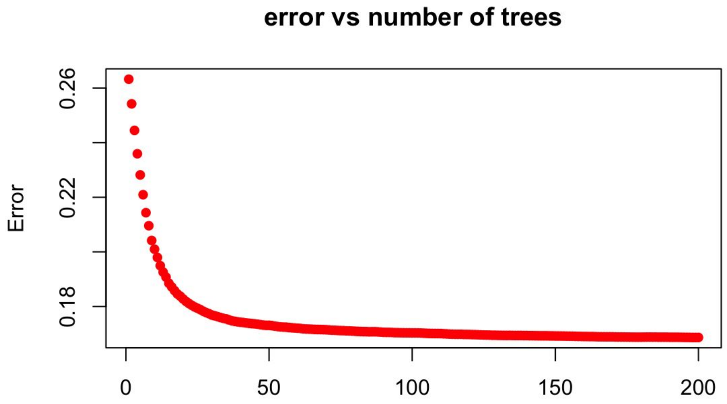
```
fit_lasso.pred    0    1
             0 4786 2031
             1  802 3330
```

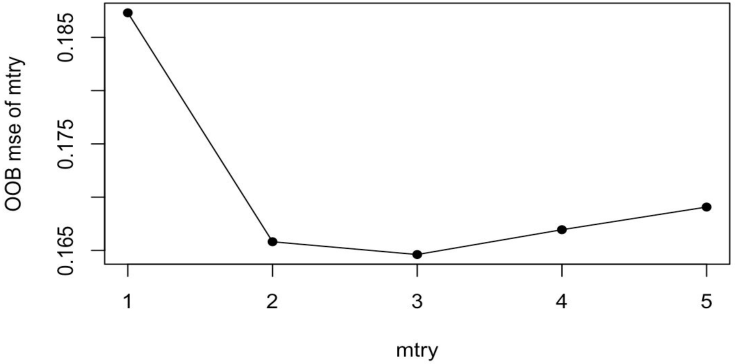
Tree-Based Models

- Decision Tree
  - Worse general health, stroke, males and older people are more likely to have heart disease
  - **Accuracy**: 0.71; **Recall**: 0.77, **F1-score**: 0.73
- Random Forest
  - Settled for 100 trees and 3 features sampled at each split.
  - **Accuracy**: 0.77; **Recall**: 0.73, **F1-score**: 0.76



```
fit_rf.pred    0    1
          0 4069 1031
          1 1519 4330
```





Final Recommendation

- Logistic Regression (LASSO)
  - Best performance for recall and F1 score
    - Accuracy is slightly lower than tree-based models, but recall is higher
  - Easier to interpret and more computationally efficient
- Potential Improvements
  - Try other models such as SVM and boosting
  - Collect more data variables
    - This dataset is focused on physical measures and other disease indicators of the patient
    - Conditions of patient's relatives could also be helpful

Wharton
UNIVERSITY of PENNSYLVANIA

Thank you for listening!

Questions?

Please enjoy the summer break!