PREDICTING HEART DISEASES

Shijie (Brandon) Bian, Huy Nguyen, Sophie Yun, Pinyi Zhen

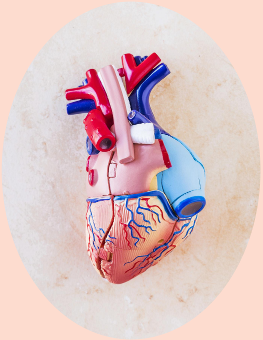


TABLE OF CONTENTS

01 Introduction

Background, Motivation, Dataset

1.1 Background & Goal

- **Heart disease** is a leading cause of death in the United States for decades.
    - **Every 36 seconds,** 1 person dies from cardiovascular disease (CVD)
    - **659,000 people** die from heart disease each year
    - Monetary cost: **$363 billion** a year
- Heart diseases are terrifying but preventable
    - Smoking, hypertension, and high cholesterol levels
    - Socioeconomic status -> nutrition, life styles, physical activities
    - Other diseases history
- Current study
    - **Goal: leverage large datasets to predict heart disease**
        - Healthcare providers can intervene early towards potential patients
        - Bring health to more people (Clinical Applications)

1.2 Dataset

4220 × 19 → Heart Disease

<u>Observations</u>

Each represents one person in the United States

<u>Predictors</u>

Detailed info of each person (eg. age, cholesterol, blood sugar, hypertension)

<u>Response Variables</u>

Categorical (Yes or No)

02

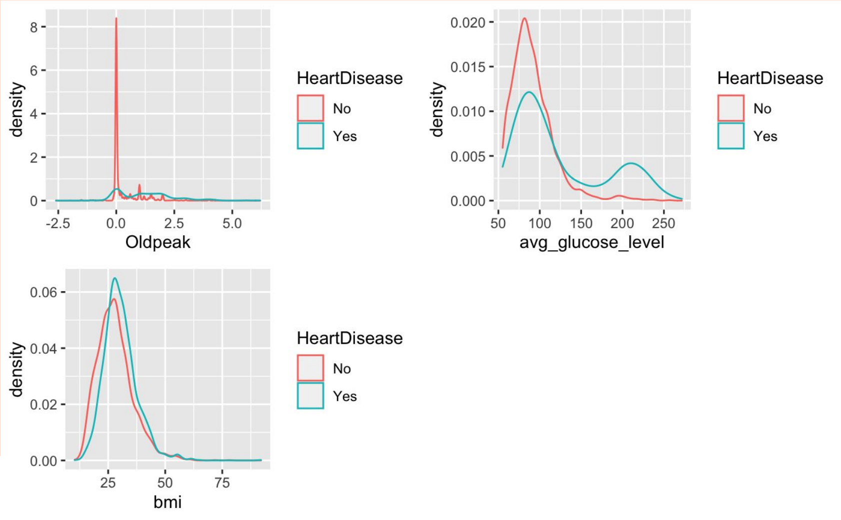
Data Analysis

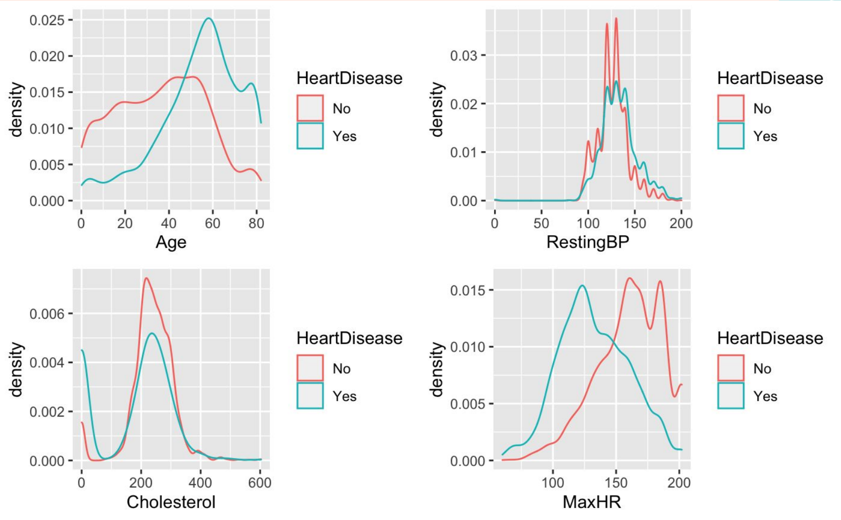
Exploratory Data Analysis, Data Cleaning and Modeling

2.1 Numerical Predictors

Density Plots:

- **Motivation**: Significant predictors should show minimum overlap between class.
- **Observation**: Predictors "RestingBP, Cholesterol, bmi" are the worst.





2.1 Numerical Predictors

Student's T-Test:

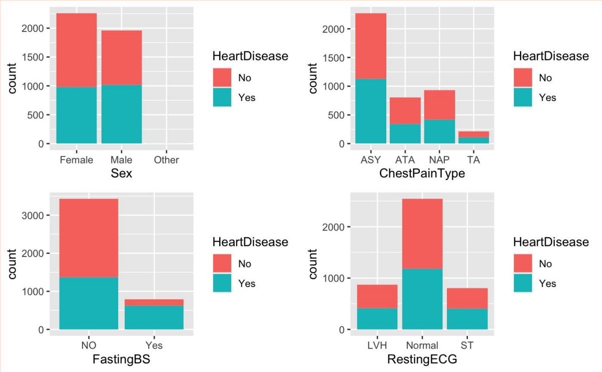
- **Motivation**: the larger the absolute t-statistic is, the more significant it is.
- **Observation**: Predictors "MaxHR, Oldpeak, Age" are the best.

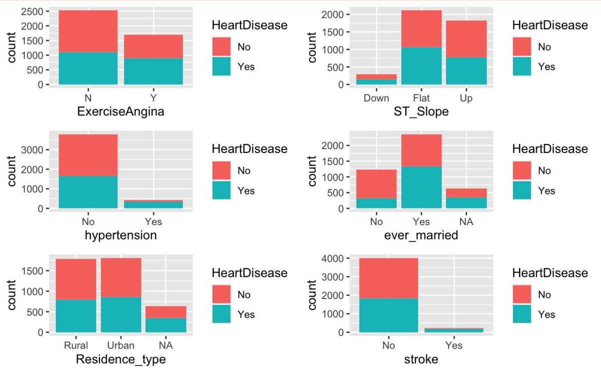
Predictor	t-statistic
MaxHR	33.616
Oldpeak	-33.425
Age	-28.913
avg_glucose_level	-25.071
Cholesterol	17.891
RestingBP	-12.315
bmi	-9.9527

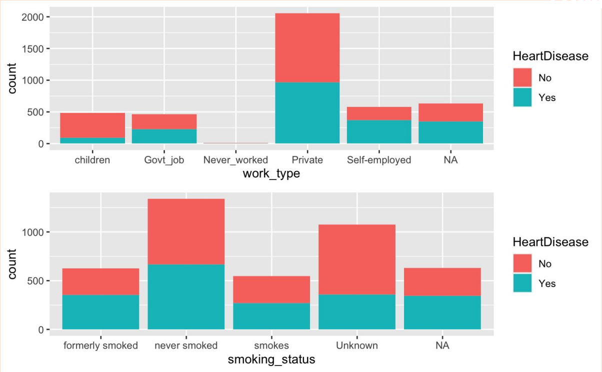
2.2 Categorical Predictors

Stacked Bar Plots:

- **Motivation**: Significant predictors should show difference in distribution between class.
- **Observation**: Predictors "RestingECG, ChestPainType, Residence_type" are the worst.







2.2 Categorical Predictors

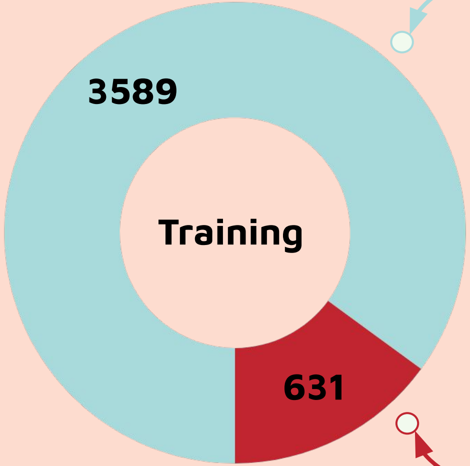
Chi-square Test:

- **Motivation**: the larger the absolute X-squared statistic is, the more significant it is.
- **Observation**: Predictors "FastingBS, ever_married, work_type" are the best.

Predictor	X-squared
FastingBS	375.829
ever_married	297.8327
work_type	227.5932
hypertension	186.6208
smoking_status	109.8935
stroke	107.8736
ExerciseAngina	33.31353

Predictor	X-squared
sex	31.07828
ST_slope	27.39078
ChestPainType	17.98923
Residence_type	3.098384
RestingECG	2.667704

2.3 Missing Values

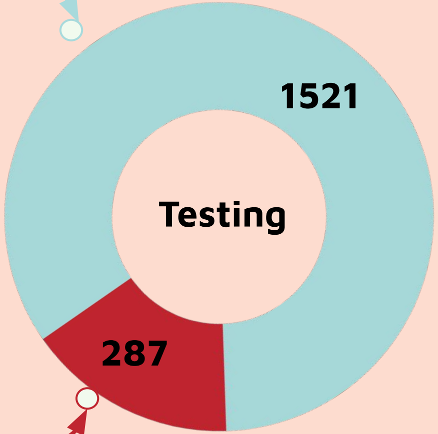


Valid Observations

85% **84%**

Missing Observations

15% **16%**



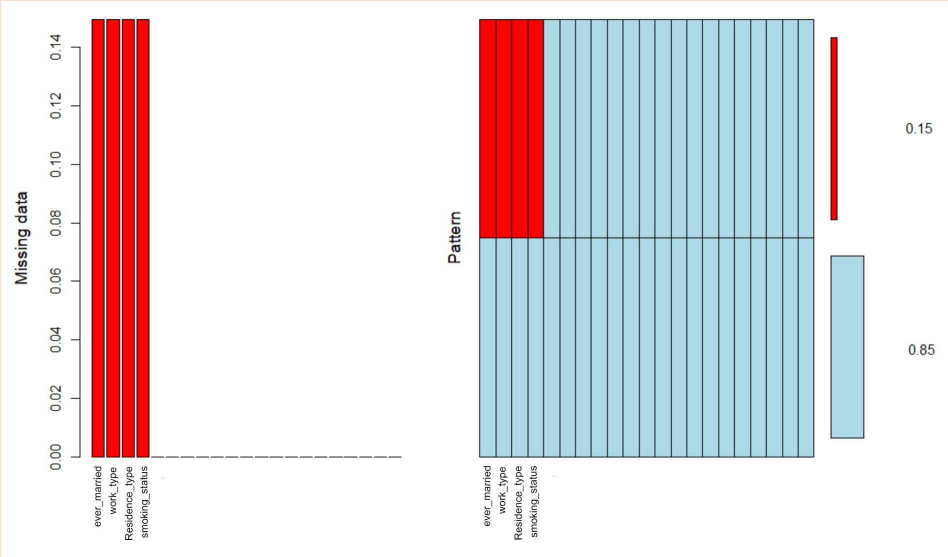
<u>Distribution of Missing Values</u>:

- **4 Predictors with NA values:** ever_married, work_type, Residence_type, smoking_status

2.3 Missing Values

Patterns of Missing Values:

- All 4 predictors with missing values are categorical.
- All 4 predictors with missing values have the same ratio of NA values.



2.4 Data Cleaning

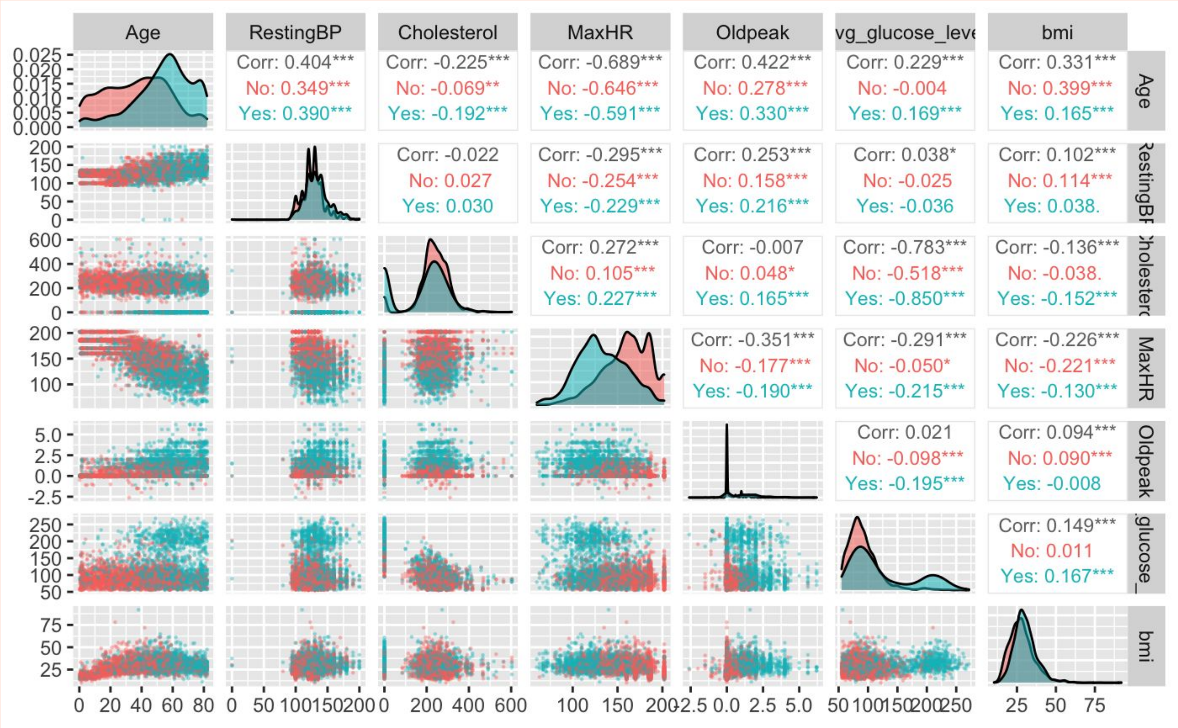
Solution 1: Deleting the columns with missing values

Solution 2: Imputation

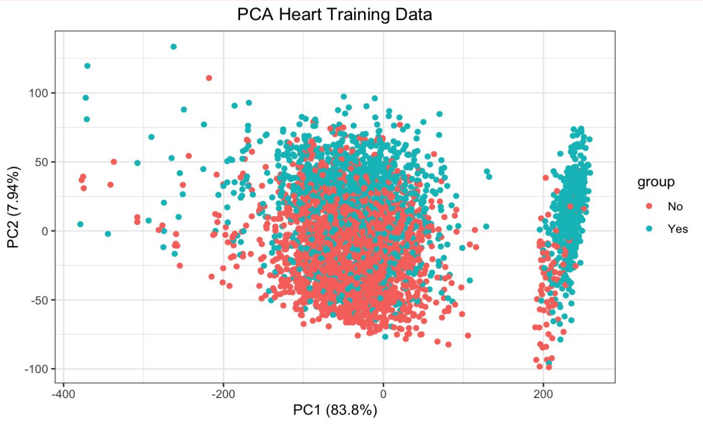
- Mice: Methods predictive mean matching (pmm) & random forest (rf)
- Imputing all missing values as "Unknown"

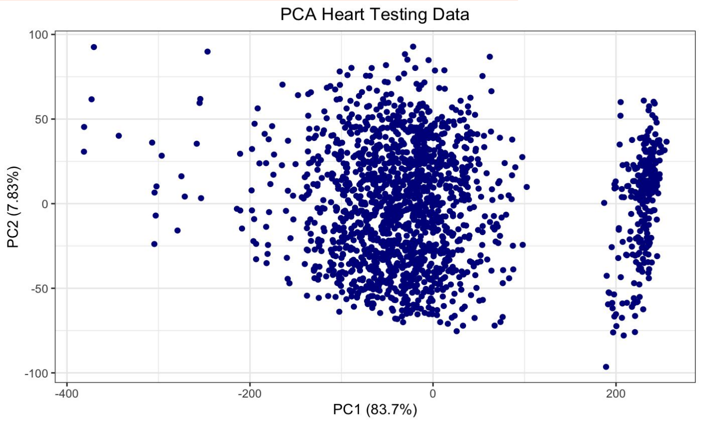
Observation: Method 1 produced the best result

2.5 Multicollinearity



2.5 PCA





Advantages: Eliminate multicollinearity of numerical variables, reduce dimension

Disadvantages: Difficult interpretability, loss of information

Observations:

- The PCA plots of training and testing data shows similarities.
- The PCA of training data shows that data with Yes for Heart Disease are mostly above and with No mostly below.
- There seems to be two clusters.

2.6 Data Transformation

BMI: ≤30: Normal, >30: Obese

MaxHR: ≤150: Normal, >150: High

Age: ≤45: Not middle-aged and elder, >45: Middle-aged and elder

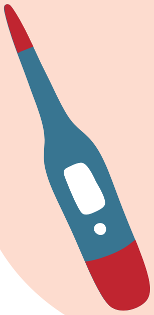
In the end, data transformation does not work for us

03

Models & Methods

Model Construction and Analysis

3.1
Basic Models



Basic Model: LDA & KNN

Linear Discriminant Analysis

- Find a linear combination of features that characterize groups in Y
- Continuous X, categorical Y
- **Pro**: works if the boundary is linear
- **Con**: not works if the boundary is non-linear
- **Assumptions**:
  - Independent subjects
  - Same within-group variance across groups in the response variable
- <u>Best model accuracy</u>: **0.80790**
  - After removing NA columns

K Nearest Neighbors

- Find the k closest neighbors to X and examining the corresponding Y
- Continuous X, categorical Y
- **Pro**:
  - Flexible, non-parametric, no assumption
  - Works better with non-linear boundary
- **Con**: Needs to tune k
- <u>Best model accuracy</u>: **0.76837**
  - With k = 3, after scaling all numerical variables

Basic Model: GLM & Splines

Logistic Regression

- Use logistic function to model the probability of classes
- Binary response variable Y
- **Pro**: can combine both numerical and categorical variables
- **Con (Assumptions)**:
    - No multicollinearity
    - Binary category
    - Linearity of variables
    - Independent observations
- Best model accuracy: **0.81106**
    - Use Principal Components and All Categorical Predictors

Splines

- Increase flexibility to reduce bias
- Use natural spline.
- **Pro**: Flexible
- **Con**: Large amount of parameters
- Best model accuracy:
    - Using degrees of freedom 4 for numerical predictors, the result is comparable to GLM

Basic Model: A Summary

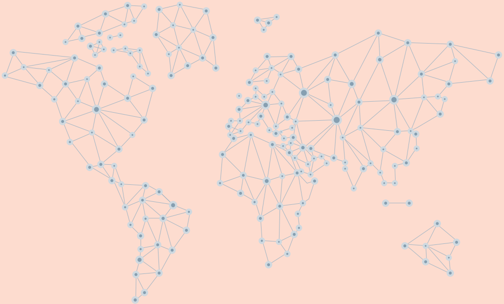
GLM model using Principal Components and all Categorical Variables yielded the best result

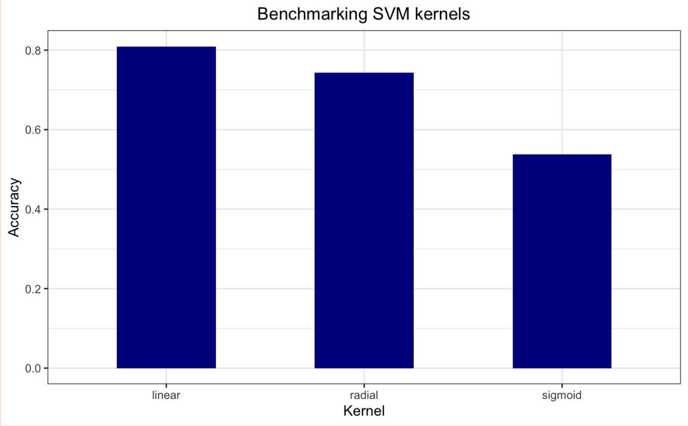
3.2
Advanced Models

Advanced Model: SVM

Support Vector Machine

- Finding hyperplane that best separates between the correctly classified points and that the points on the wrong side are not too off.
- **Pro**:
  - Many types of kernels
  - Work with both categorical and numerical variables
- **Con:**
  - Sensitive to overfitting
- <u>Best model accuracy</u>: **0.80948 (Linear)**





Advanced Model: Trees

Tree Bagging

- Construct regression trees and average / find majority of resulting predictions
- **Pro**:
    - Low variance
    - More predictions
- **Cons**:
    - Seed-dependent
    - less interpretability
- Best model accuracy: **0.79683**
    - After removing NA columns

Random Forest

- Based on tree bagging, but considers random sample of predictors for tree splits
- **Pro**:
    - De-correlates trees
    - One of the most accurate classifiers
- **Cons**:
    - Seed-dependent
    - Possible overfitting
- Best model accuracy: **0.80869**
    - After removing NA columns

3.3
Exploring with Caret

Caret

- **CARET**: **C**lassification **A**nd **RE**gression **T**raining
- Functionality: Streamlined process of creating predictive models
- Motivation: Try out models based on **Extreme Gradient Boosting** & **Random Forest**

Train Control

10-folds
Cross Validation

Metric

Use "Accuracy" to
select optimal model

Method

Specify the
classification
model

Train

Train model on
entire training set

Caret - Extreme Gradient Boosting

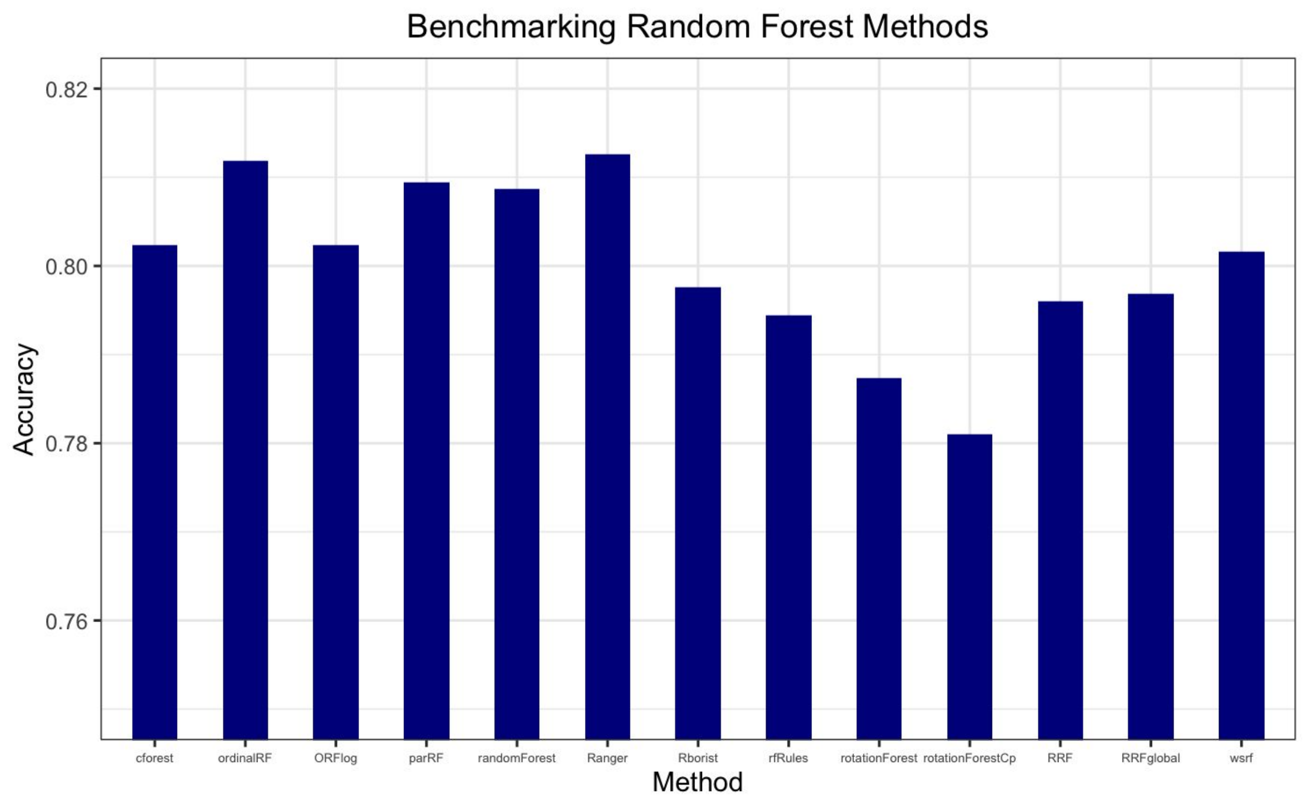
- **XGBoost**
- Functionality: Machine Learning algorithms under gradient boosting framework
- Motivation: Try out models focused on tree methods

Model	Description	Acc.
xgbDART	DART booster (drop out trees to prevent over-fitting)	0.80079
xgbLinear	Linear boosting, catch linear link	0.79446
xgbTree	Tree boosting, catch non-linear link	0.80711

Caret - Random Forests

Model	Description	Acc.
Oblique Random Forest	Use oblique decision boundaries to simplify the boundary	**0.80237**
Ordinal Forest	Ordinal regression	**0.81185**
Rotation Forest	Fit tree on principle components of partitioned variables	**0.78735**
Parallel Forest	Running random forest using paralleled method	**0.80948**
Ranger Forest	Boosted random forest for classifying high-dimensional data	**0.81264**

Caret - Random Forests



Caret - Final Model

Model
Ranger Random Forest
(Seed: 7)

Observations
4220
(All training set)

Predictors
7 Numercial + 8 Categorical
(Removed predictors with NA)

Public Score
0.81264
(Ranking: 18 in class)

Private Score
0.8011
(Ranking: 8 in class)

Final Score
0.80918
(Ranking: 13 in class,
5 in Lecture 2)

04

Conclusion

Discussion and Reflections

4.1 Limitations

Limitation 1: Results are seed dependent

Limitation 2: Imputation

- Mice: Methods predictive mean matching (pmm) & random forest (rf)
- Imputing all missing values as "Unknown"
- Removed 4 categorical predictors

4.2 Conclusion

We created a model based on random forest methods to predict heart diseases with satisfactory accuracy. However, our methodology is limited since our result is seed-dependent, and our model is constructed by removing 4 categorical predictors (which may be correlated with Heart Disease). Future work should address these issues by exploring more robust modeling methods and considering more features related to Heart Disease. However, we believe our model will benefit patients and doctors by providing them with insights for reference.

We wish everyone good health!

4.3 References

[1] J Bolen, L Murphy, K Greenlund, CG Helmick, J Hootman, TJ Brady, G Langmaid, N Keenan, et al. Arthritis as a potential barrier to physical activity among adults with heart disease-united states, 2005 and 2007. Morbidity and Mortality Weekly Report, 58(7):165–169, 2009.

[2] Jeremiah R Brown and Gerald T O'Connor. Coronary heart disease and prevention in the united states. New England Journal of Medicine, 362(23):2150–2153, 2010.

[3] Centers for Disease Control and Prevention. Prevalence of coronary heart disease–united states, 2006- 2010. MMWR. Morbidity and mortality weekly report, 60(40):1377–1381, 2011.

[4] Centers for Disease Control and Prevention. Vital signs: avoidable deaths from heart disease, stroke, and hypertensive disease-united states, 2001-2010. MMWR. Morbidity and mortality weekly report, 62(35):721–727, 2013.

[5] Centers for Disease Control and Prevention. Underlying cause of death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention, 2018.

[6] National Center for Health Statistics. Health, united states, 2009: With special feature on medical technology. National Center for Health Statistics (US), 2010.

[7] Shaista Malik, Nathan D Wong, Stanley S Franklin, Tripthi V Kamath, Gilbert J L'Italien, Jose R Pio, and G Rhys Williams. Impact of the metabolic syndrome on mortality from coronary heart disease, cardiovascular disease, and all causes in united states adults. Circulation, 110(10):1245–1250, 2004.

[8] Michelle C Odden, Pamela G Coxson, Andrew Moran, James M Lightwood, Lee Goldman, and Kirsten Bibbins-Domingo. The impact of the aging population on coronary heart disease in the united states. The American journal of medicine, 124(9):827–833, 2011.

[9] SS Virani, A Alonso, HJ Aparicio, EJ Benjamin, MS Bittencourt, CW Callaway, et al. Heart disease and stroke statistics—2021 update: a report from the american heart association. Circulation, 143:254–743, 2021.

Thank You!

Made possible by slidesgo