ds2_final_report

Weijia Xiong, Yimeng Shang, Xue Jin

5/15/2020

Introduction

Heart disease is one of the leading cause of death for both men and women all over the world. Heart disease detection based on typical clinical features is being increasingly critical to population health. It is a desirable way to help people enhance self protection ability and prevent faults. In this report, we built several predictive models to predict the heart disease.

There are 303 observations, 14 predictions in the dataset and 2 missing values.

- age: age in years **continious variable**
- sex: (1 = male; 0 = female) **categorical variable**
- cp: chest pain type **categorical variable**
- trestbps: resting blood pressure (in mm Hg on admission to the hospital) **continious variable**
- chol: serum cholestoral in mg/dl **continious variable**
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) **categorical variable**
- restecg: resting electrocardiographic results **categorical variable**
- thalach: maximum heart rate achieved **continious variable**
- exang: exercise induced angina (1 = yes; 0 = no) **categorical variable**
- oldpeak: ST depression induced by exercise relative to rest **continious variable**
- slope: the slope of the peak exercise ST segment **categorical variable**
- ca: 3 = normal; 6 = fixed defect; 7 = reversable defect **categorical variable**
- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect **categorical variable**
- target: 1 = disease; 0 = not disease **response**

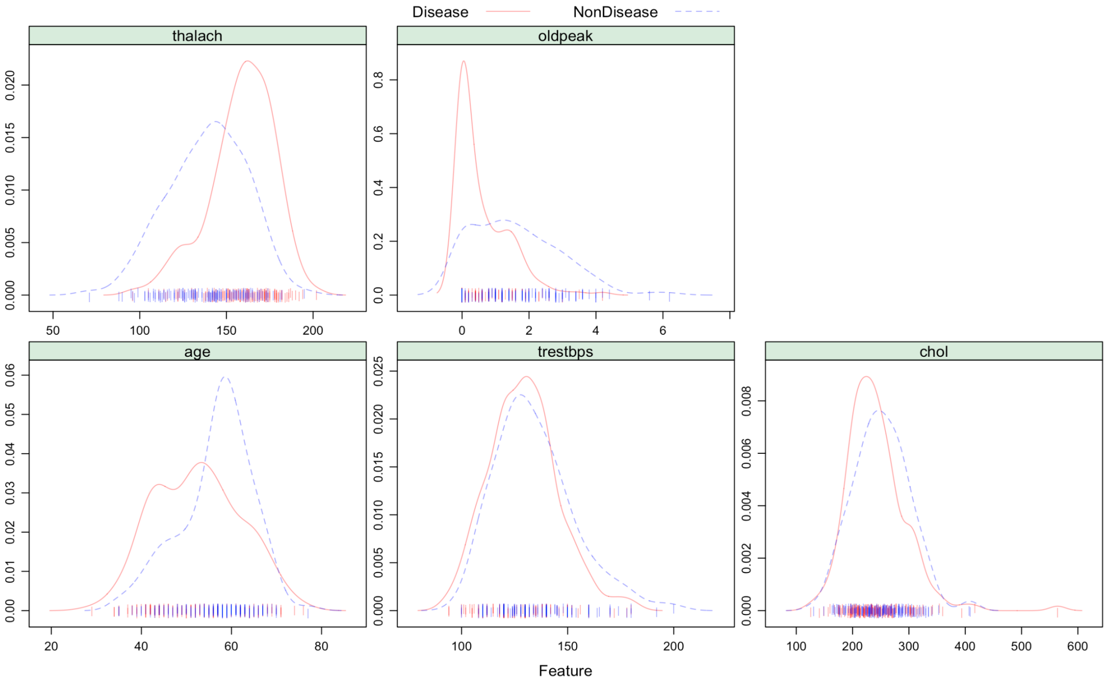
Among the predictors, there are 8 categorical variables and 5 continious variables. We first recoded all the categorical variables with its true meaning for further usage. Then, we removed missing values and changed all charactor variables into factor.

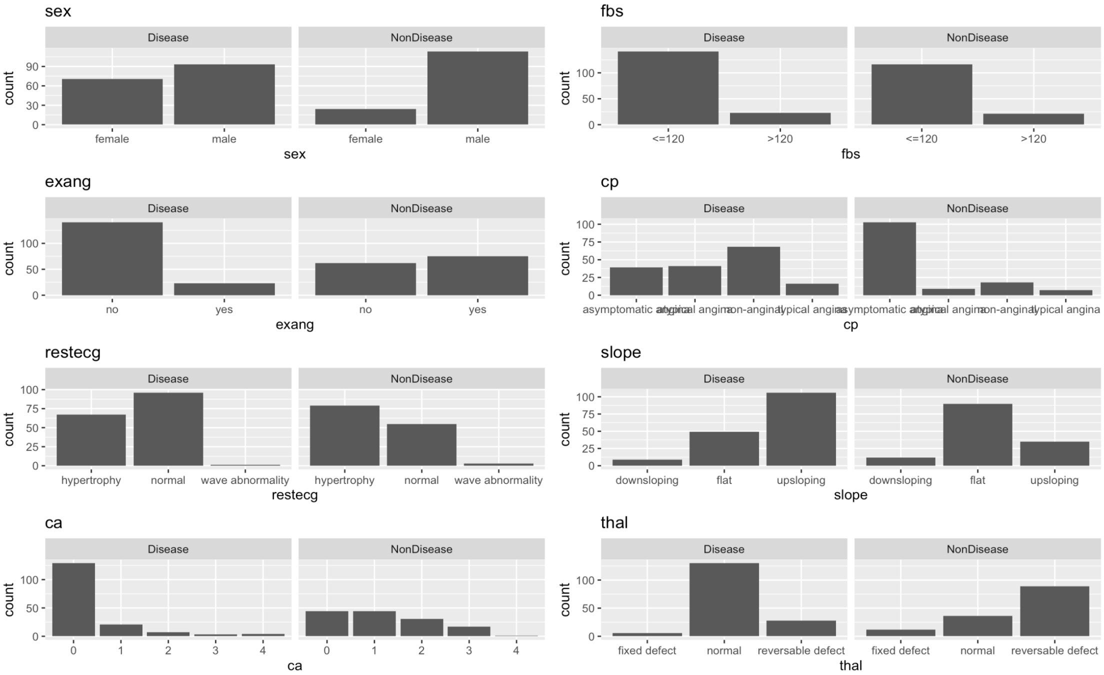
In this report, we aim to built a model to predict whether a person is likely to get heart disease or not. We use cross validation to compare the prediction performance of these models.

Exploratory analysis/visualization

Among the continious variables, thalach/oldpeak/age have very different distribution among those disease and non-disease population. Disease group tend to have higher *thalach* and smaller *oldpeak*.

Among the categorical variables, the distribution of fbs is similar among disease and non-disease groups while others are not similarly distributed.





Models

What predictor variables did you include? What technique did you use? What assumptions, if any, are being made by using this technique? If there were tuning parameters, how did you pick their values? Discuss the training/test performance if you have a test data set. Which variables play important roles in predicting the response? Explain/visualize the final model you select. What are the limitations of the models you used (if there are any)? Are the models flexible enough to capture the underlying truth?

Conclusions

What were your findings? Are they what you expect? What insights into the data can you make?