Risk Factors for Heart Disease Classification

Peter Gatto

Motivation/Introduction

- Heart disease is the highest cause of death in the United States
- Can heart health be predicted by survey responses?
- What risk factors have the highest predictive value?

Data and Feature Engineering

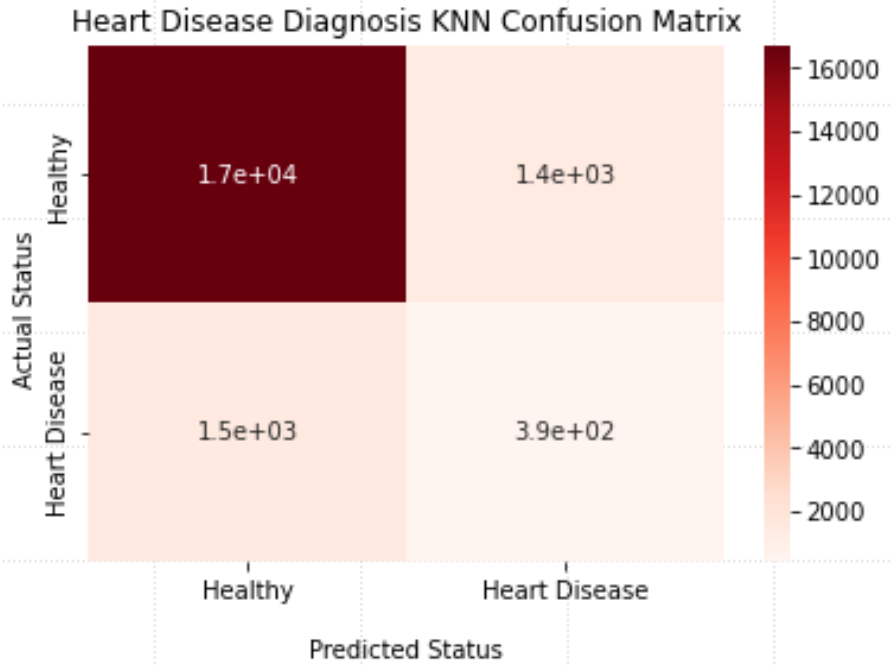
- 253,680 respondents to CDC phone surveys
- Target feature: history of heart disease or heart attack
- Risk factors include high blood pressure, high cholesterol, BMI, smoking history, stroke history, diabetes, heavy alcohol consumption, sex, age
- BMI and age transformations

Methodology

- Hard classification
- Metrics: log loss and recall scoring
- Three models: KNN, Logistic Regression, Random Forest

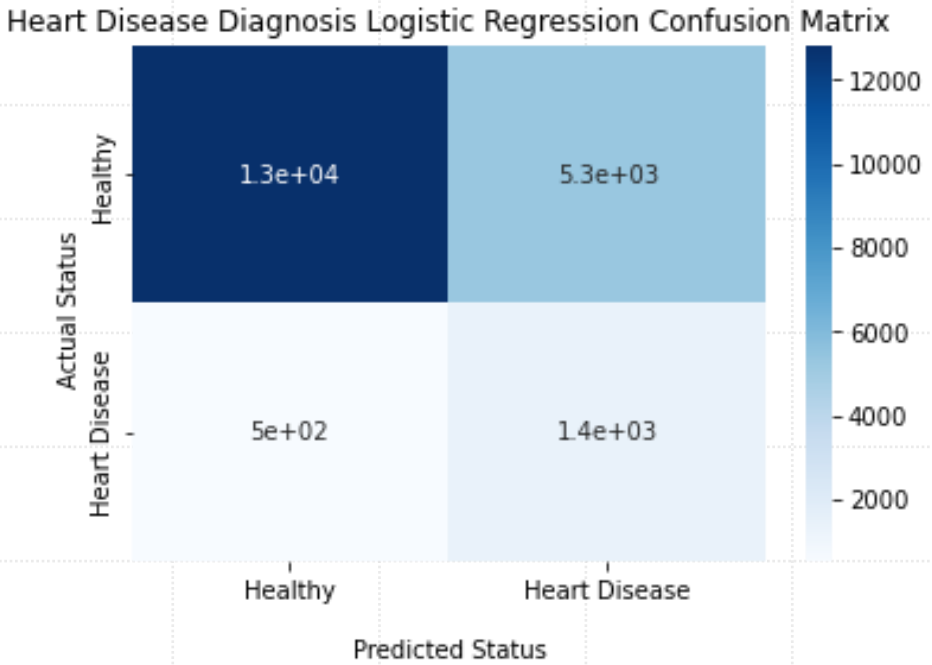
Model 1: K-Nearest Neighbors

- Recall score: 0.205
- K = 1
- Distance weighting



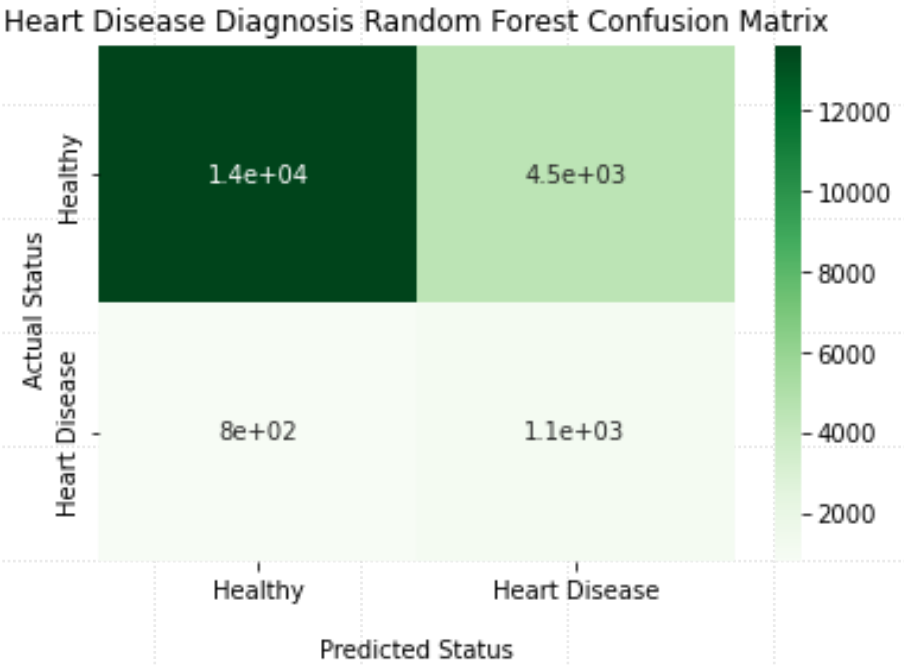
Model 2: Logistic Regression

- Recall score: 0.735
- Balanced class weighting



Model 3: Random Forest

- Recall score: 0.577
- Balanced class weighting



Feature Importance

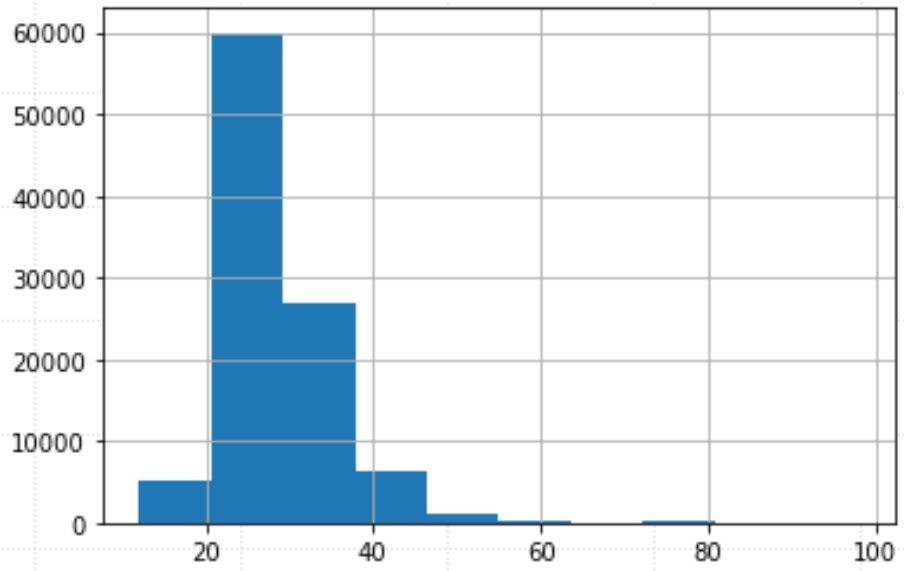
Feature	Log Loss	Importance
High blood pressure	0.292	0.147
Age	0.295	0.146
BMI	0.312	0.125
High cholesterol	0.298	0.098
Diabetes	0.302	0.086

Conclusions and Future Studies

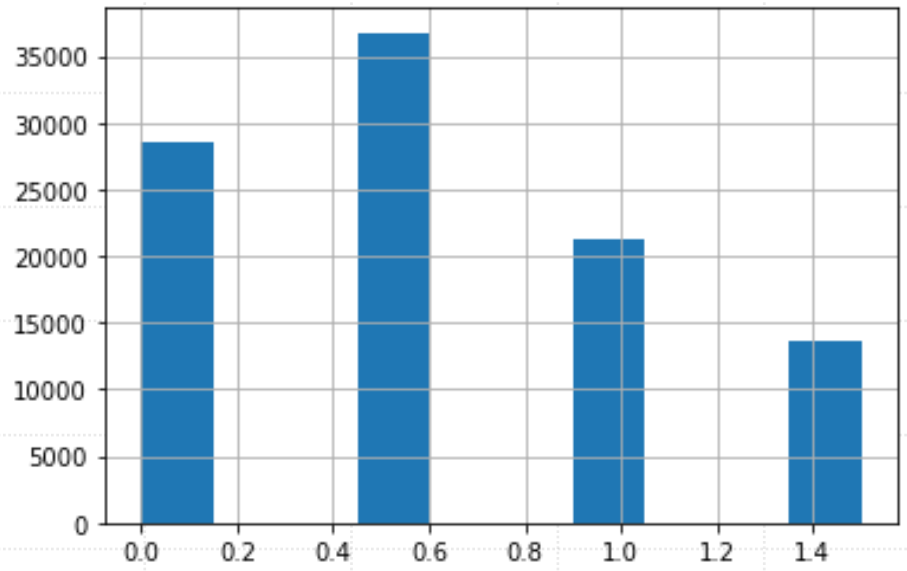
- Logistic regression model outperforms KNN and Random Forest approaches
- Custom weighting or resampling KNN to combat effects of class imbalance
- Consider ensemble classification methods

Appendix

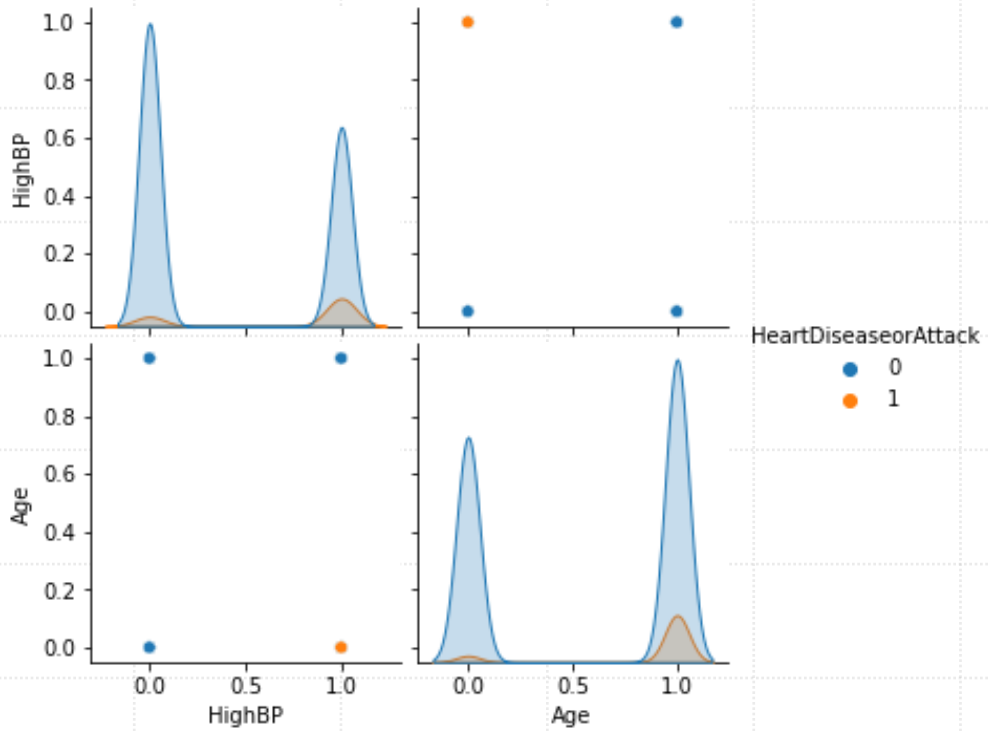
BMI distribution



Engineered feature



Appendix


HighBP
Age
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
HighBP
Age
HeartDiseaseorAttack
0
1