Predicting Cardiovascular Disease

Jonathan Murthy

Research Question

What model both explains the most variance and can most accurately predict whether or not a patient will have Cardiovascular disease?

The Data

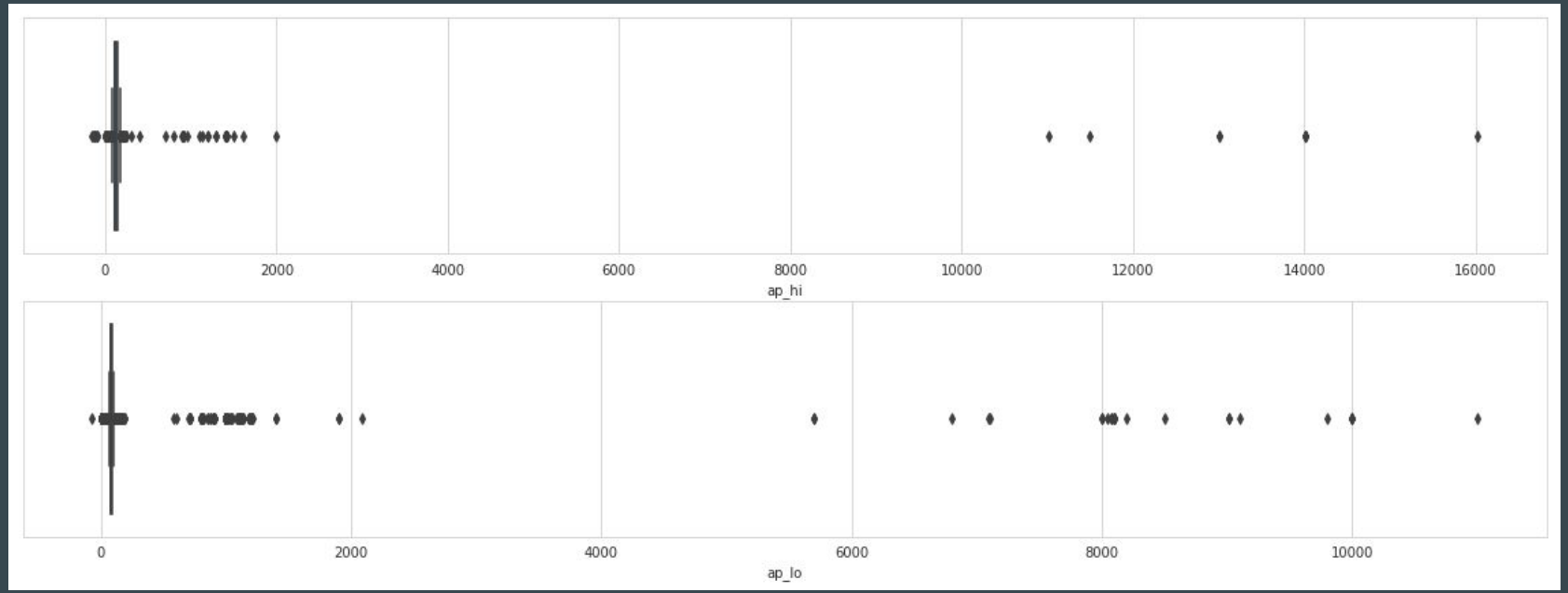
Source: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Features: Age (cont), Gender (cat), Height cm (cont), Weight kg (cont), Blood Pressure (cont), Cholesterol(cat), Glucose (cat), Smoker (cat), Alcohol use (cat), Active (cat)

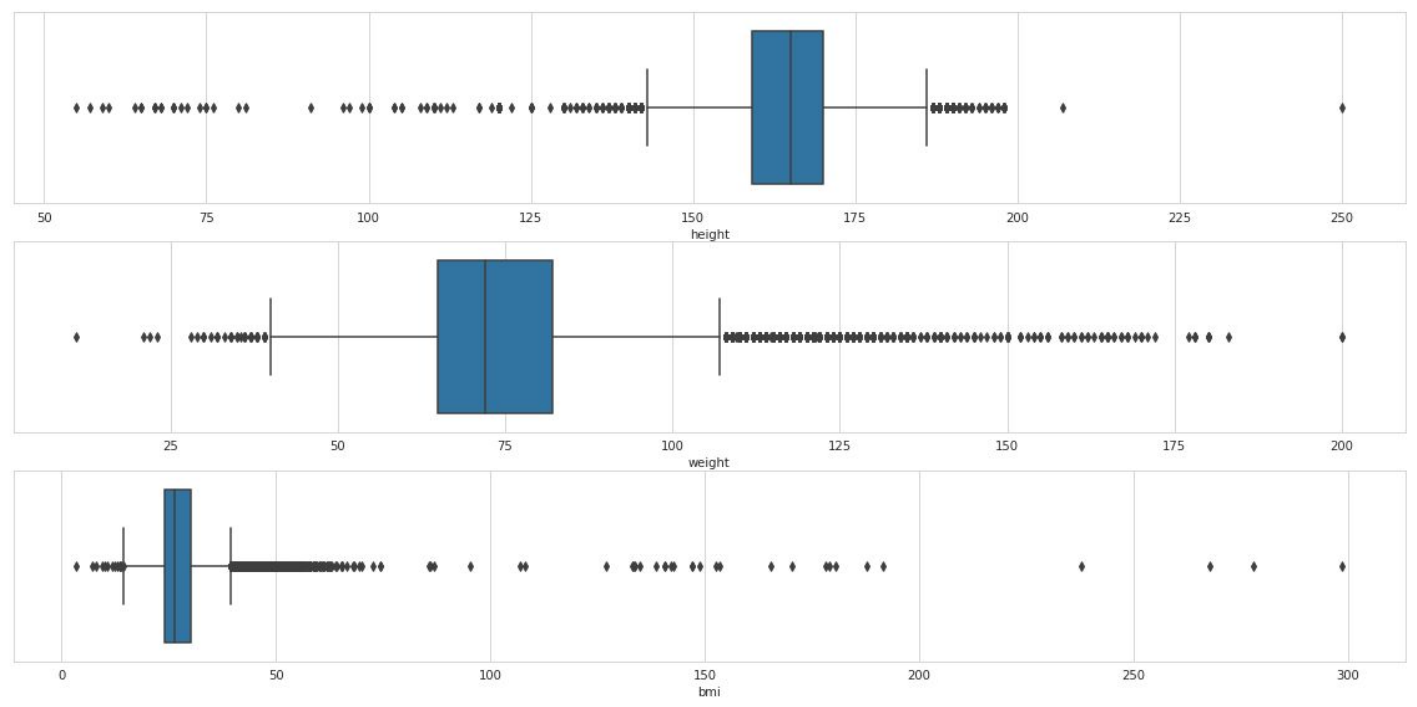
Target: Cardio (binary)

Cleaning the Data

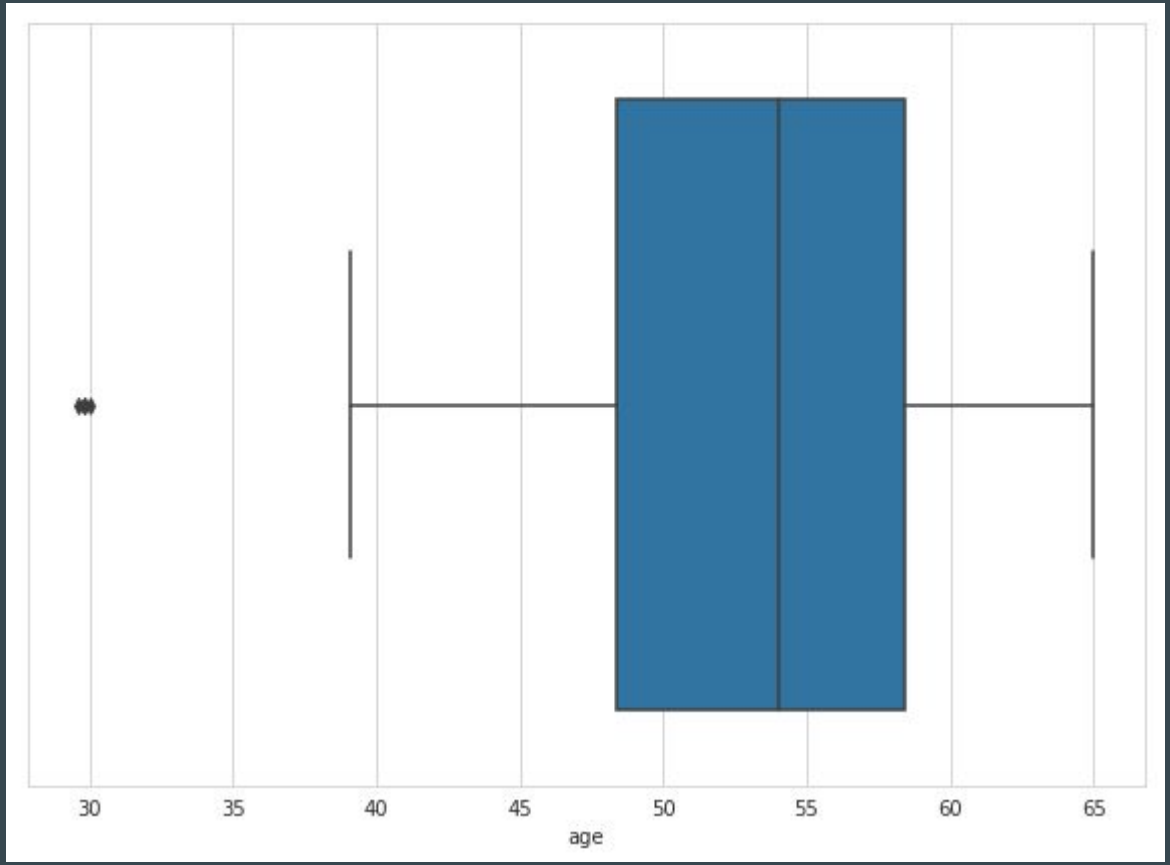
- Outliers
  - Blood Pressure



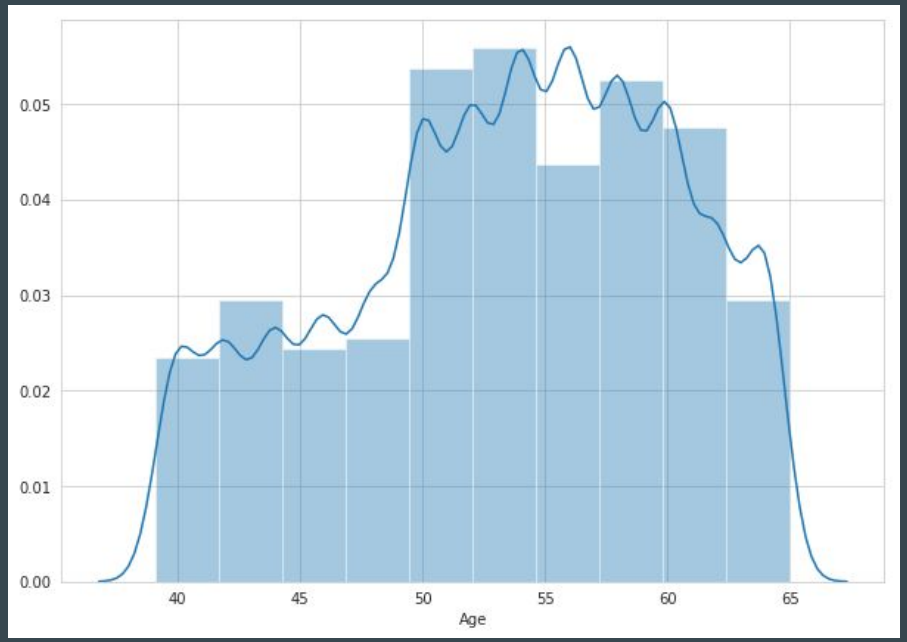
- Weight and Height

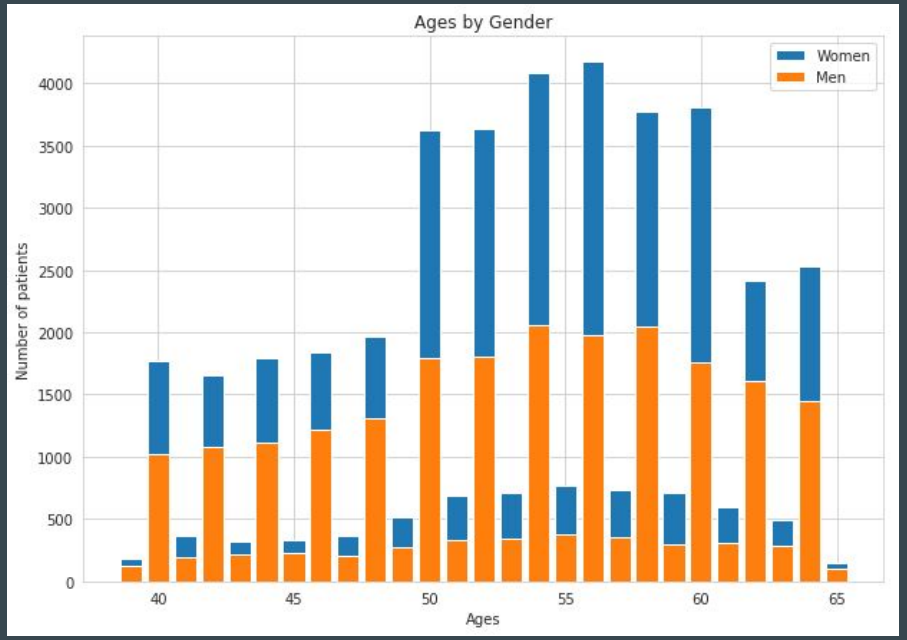


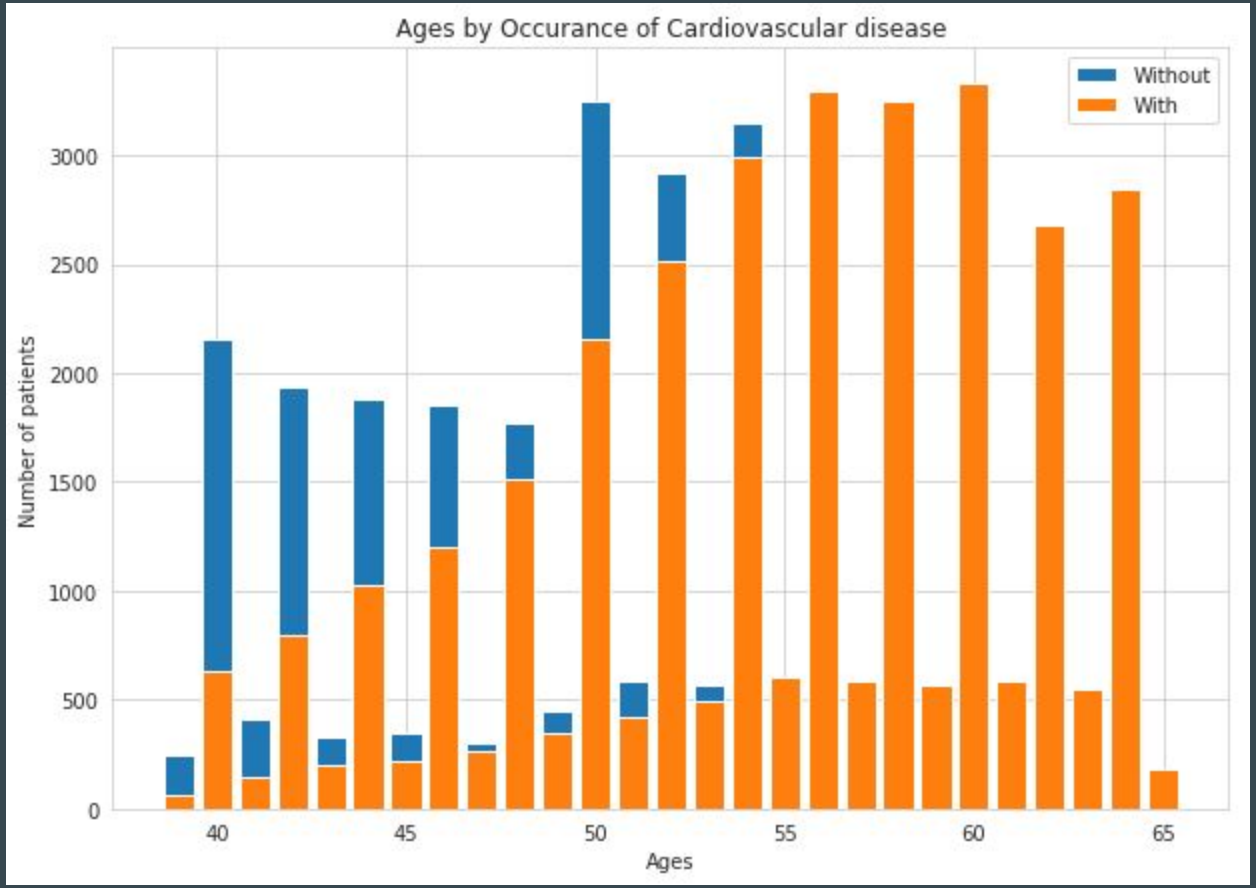
- Age

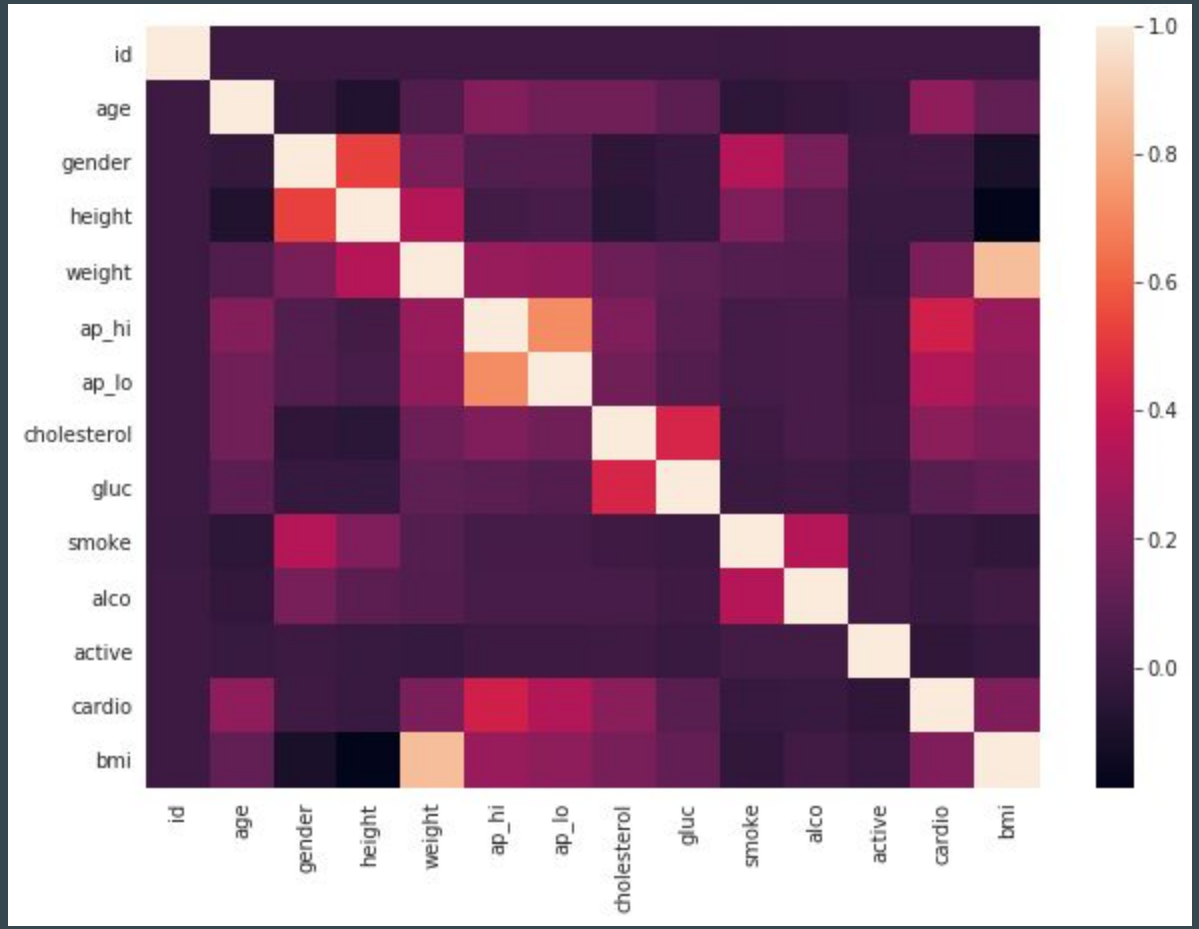


Exploratory Data Analysis






Ages by Occurance of Cardiovascular disease
Without
With
Number of patients
0
500
1000
1500
2000
2500
3000
40
45
50
55
60
65
Ages


id
age
gender
height
weight
ap_hi
ap_lo
cholesterol
gluc
smoke
alco
active
cardio
bmi
1.0
0.8
0.6
0.4
0.2
0.0

Feature Selection

- Full Feature Set
- SelectKBest
  - K = 10
- PCA
  - n_components = 10

Logistic Regression

- Default Parameters
- Clean Data:
  - All features
    - R-squared: 0.7239
    - F1: 0.7029
  - PCA
    - R-squared: 0.7290
    - F1: 0.7076
  - KBest
    - R-squared: 0.7284
    - F1: 0.7080

- Default Parameters
- Full Data:
  - All features
    - R-squared: 0.7216
    - F1: 0.7097
  - PCA
    - R-squared: 0.7217
    - F1: 0.7065
  - KBest
    - R-squared: 0.7243
    - F1: 0.713

Random Forest

- Default Parameters
- Clean Data:
  - All features
    - R-squared: 0.6981
    - F1: 0.6786
  - PCA
    - R-squared: 0.6913
    - F1: 0.6736
  - KBest
    - R-squared: 0.6983
    - F1: 0.6793

- Default Parameters
- Full Data:
  - All features
    - R-squared: 0.6976
    - F1: 0.6842
  - PCA
    - R-squared: 0.6989
    - F1: 0.6834
  - KBest
    - R-squared: 0.6955
    - F1: 0.6835

KNN

- N_neighbors = 5, weights= ‘distance’
- Clean Data:
  - All features
    - R-squared: 0.6799
    - F1: 0.6674
  - PCA
    - R-squared: 0.6844
    - F1: 0.6761
  - KBest
    - R-squared: 0.6814
    - F1: 0.6707

- N_neighbors = 5, weights= ‘distance’
- Clean Data:
  - All features
    - R-squared: 0.6853
    - F1: 0.66683574
  - PCA
    - R-squared: 0.6853
    - F1: 0.6804
  - KBest
    - R-squared: 0.6817
    - F1: 0.6775

Gradient Boost Classifier

- N_estimators: 200
- Max_depth: 2
- Loss function: deviance
- Clean Data:
  - All features
    - R-squared: 0.7333
    - F1: 0.7202
  - PCA
    - R-squared: 0.7326
    - F1: 0.7202
  - KBest
    - R-squared: 0.7380
    - F1: 0.7245

- N_estimators: 200
- Max_depth: 2
- Loss function: deviance
- Clean Data:
  - All features
    - R-squared: 0.7369
    - F1: 0.7268
  - PCA
    - R-squared: 0.7336
    - F1: 0.7213
  - KBest
    - R-squared: 0.7343
    - F1: 0.7280

Conclusion

The Gradient Boosting Classifier explained the most variance in the dataset as well as having the highest f1 score.

Some of the other models were overfitting.

Practical Use

- Can be used to classify whether or not a patient is at risk of cardiovascular disease

Shortcomings

- There could be more interesting features for predicting cardiovascular disease (e.g. socioeconomic data, time series data, geographical/regional data, education, more lab tests)

- Computational limitations. More powerful computation could allow for more observations