Identifying patients with a risk of cardio disease

Iryna Mishiev
February 2021

Opportunity

Objective:

Alert the medical practitioners about patients with a high risk of having heart diseases.

Practical impact:

By highlighting patience with a high or medium-high risk of having the cardio disease, doctors will have an opportunity to prevent the development of the disease or put additional attention to this patient's condition.

Process and Tools

Data:

Data from Kaggle consists of 70000 reports with patient notes, including physical exam findings, analysis reports, and diagnoses.

Methodology:

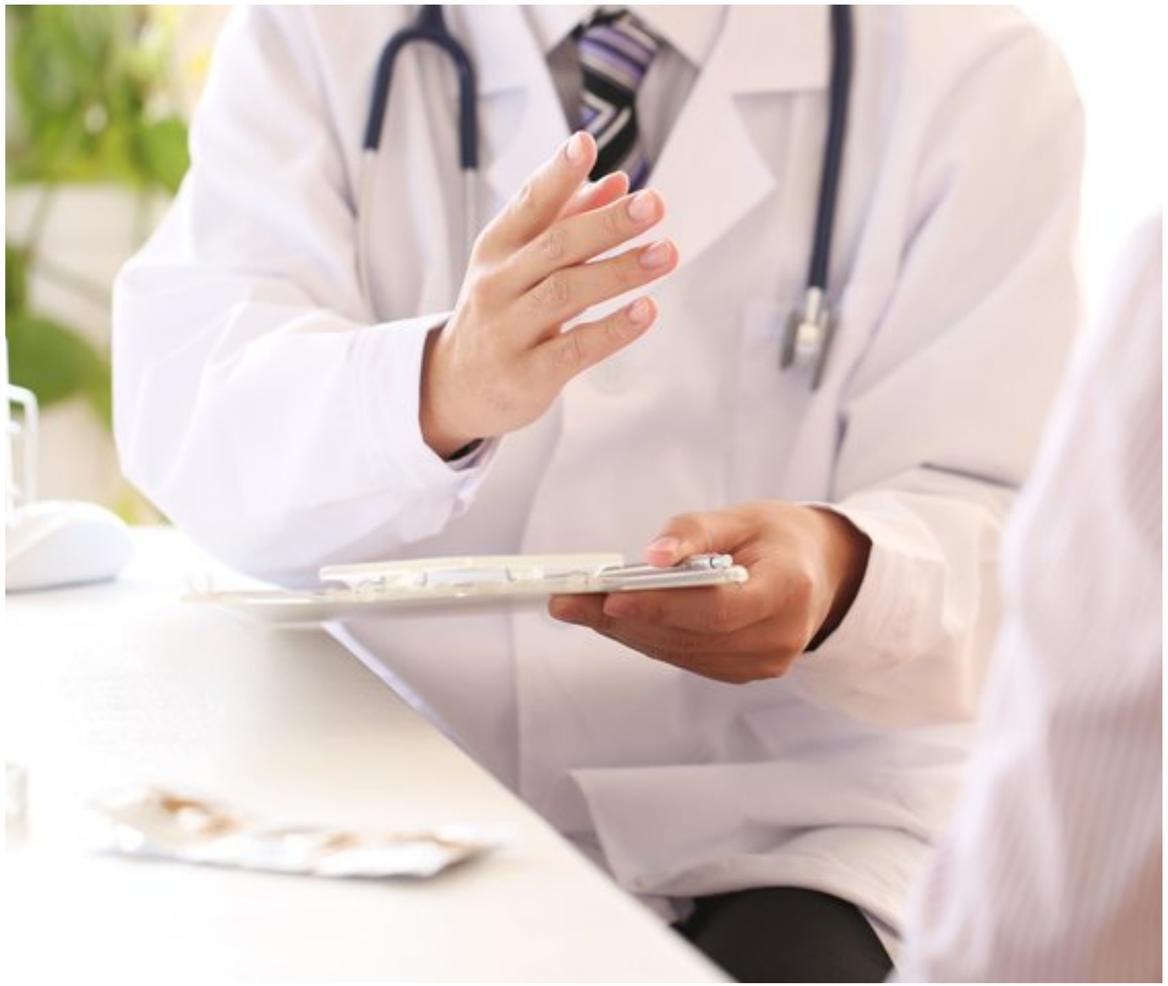
EDA and data cleaning,

Feature engineering,

Baselining - model Linear Regression,

Modeling - kNN, Random Forests, Extra Trees, GBMs, Naive Bayes.

Final Model selection and turning.



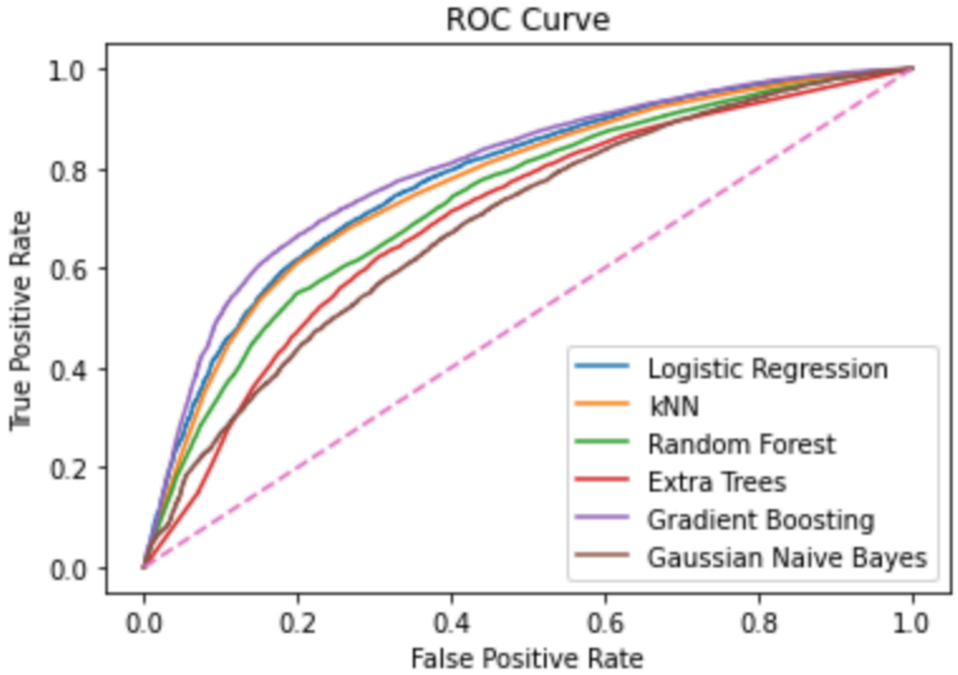
Final Model Selection

Recall from different models

Logistic Regression 62.45%	kNN: 69.50%	Random Forest: 69.69%	Extra Trees: 66.69%	Gradient Boosting: 68.36%	Naive Bayes: 29.56%

Random Forest start point

Training accuracy: 98.45%
Val accuracy: 65.65%
Precision: 0.6671
Recall: 0.6969
F1: 0.6619

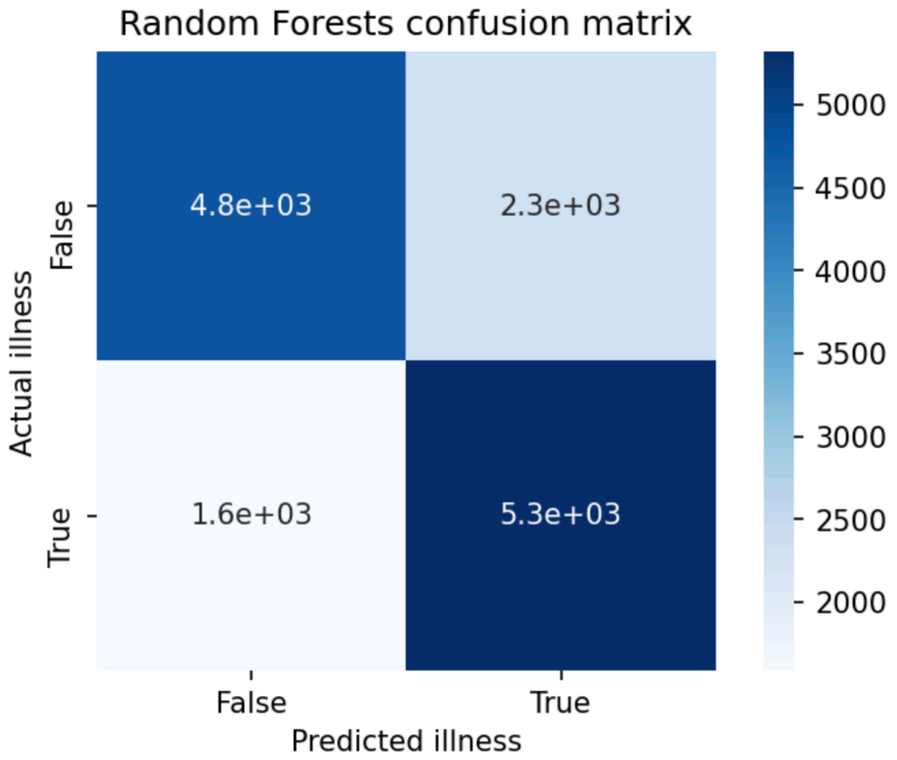


Random Forest. Summary Metrics

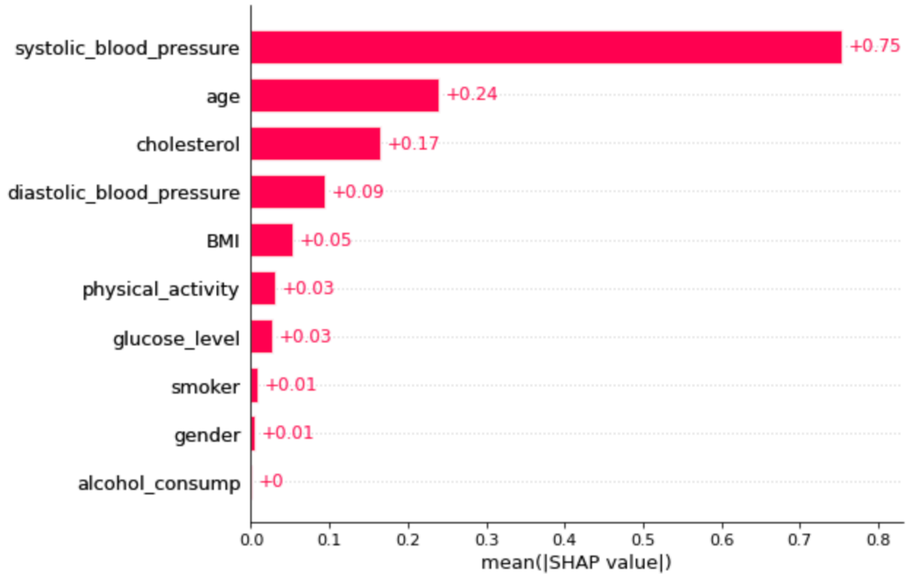
Random Forest accuracy after tune hyperparameters:
Training: 76.04%
Test set: 72.96%

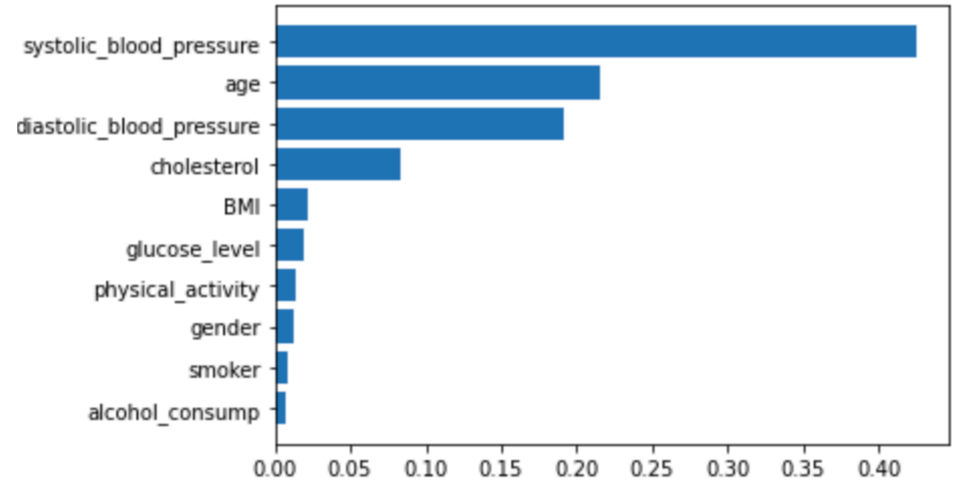
Recall score with threshold 0.4: 77.03%

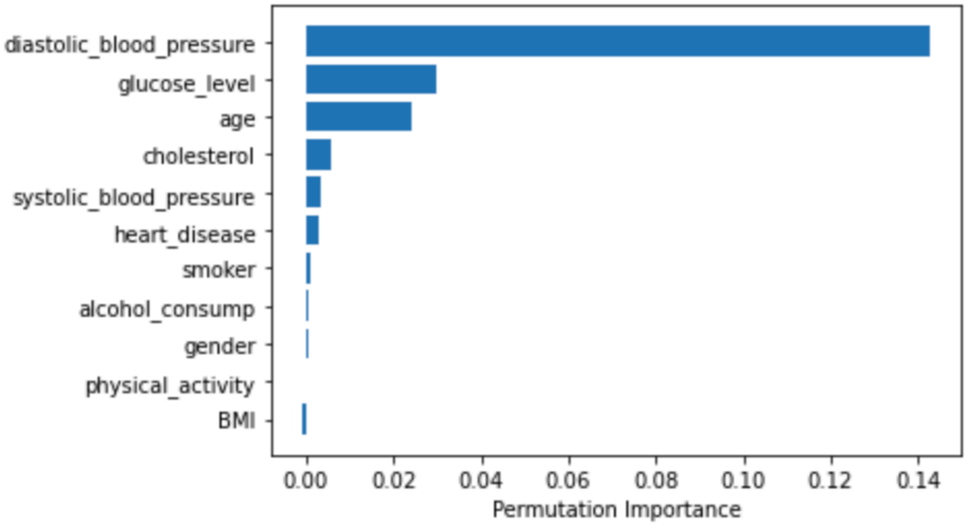
ROC AUC score with threshold 0.4 = 0.72



Feature importance and Permutation importance



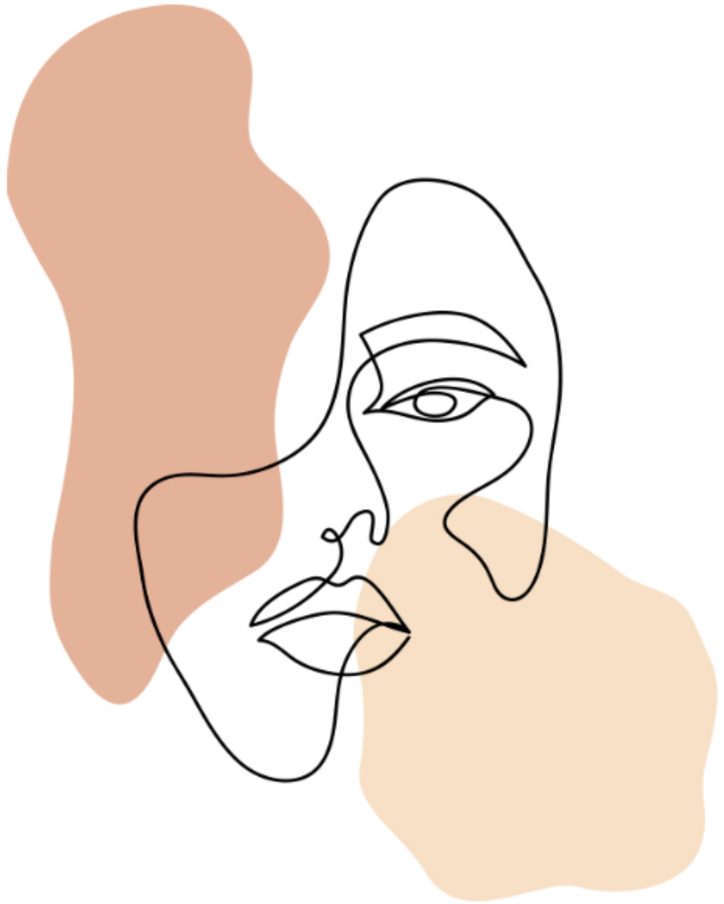




Systolic_blood_pressure
Age
Cholesterol
Diastolic_blood_pressure
Glucose level

High and Medium Risk Groups

Gender	Women
Age	60-65
Systolic_blood_pressure	140-150
Diastolic_blood_pressure	90
Glucose level	normal
Cholesterol level	above normal
Body index	overweight




Thank you!