Curiosity Team

Heart Disease Predictive Model

Predicting heart disease through personal key indicators of health

The problem

According to the CDC, heart disease is the leading cause of death in the United States, regardless of gender, race, or ethnicity. One of the most difficult aspects of treating heart disease is obtaining an accurate diagnosis from risk factors and comorbidities such as diabetes, asthma, alcohol consumption, and many other contributing factors. Machine learning algorithms have been employed to produce predictive models that would habilitate the proactive diagnosis of heart disease, such as the Heroku app. These predictive models have high accuracy but low true positive rates, which defeats the purpose of using data-driven methods in the first place.

The solution

We use the Heroku app model as our baseline logistic regression model. Our baseline model has a true positive value of only 12%. With the objective of developing a better performing model, we developed two new models: a challenger logistic regression model and a random forest model. With the scaling of the numerical variables and the adjusting of the class weight, we managed to create a logistic regression model with 78% true positive rate. By tuning model properties n_estimators and max_depth and applying balanced class weight, we managed to build a random forest with a true positive rate of 75%.