Project 3 – Predicting Heart Disease

Bianca Orozco

Project Design

Every year, 1 in every 4 deaths in the US is caused by heart disease making it the leading cause of death for both men and women. For this reason, I decided to make it the focus of my project.

I was able to get my data from the UC Irvine Machine Learning Repository which was provided to them by Robert Detrano, M.D., Ph.D. at the Cleveland Clinic. The dataset provided about 300 observations, 13 features, and a heart disease diagnosis. Originally, the target was a value between 0 and 4. Zero represented no presence of heart disease and 1-4 represented different levels of heart disease. For the sake of having a binary output, I changed all non-zero values to 1.

I trained with SVM, KNN, Random Forest, and Logistic Regression and I found Random Forest gave the best F1 Score. Out of all 13 features, the following 4 carried the most weight in the Random Forest model: chest pain (ranked 1-4), maximum heart rate, number of damaged major blood vessels (between 0 to 3), and a score indicating the extent of the damage (3-normal, 6-fixed defect, 7-reverse defect).

Tools

Jupyter Notebook
Data Analysis:

- Pandas
- Numpy
- Matplotlib
- Sklearn

Modeling:

- RandomForestClassifier
- LogisticREgression
- KneighborsClassifier
- SVC, Linear SVC

Data

After running my final train, fit and predict, I got an F1 score of 0.772 and the following Confusion Matrix:

		Actual	
		No Heart Disease Present	**Heart Disease Present**
Predicted	**No Heart Disease Present**	25	3
	Heart Disease Present	10	22

Interpretation

In my final test, ten patients have no heart disease but my model says they do. What is the cost? A model like this wouldn't determine whether a patient goes directly into surgery, but more than likely another test would be administered to confirm the presence of heart disease. The cost that would bring the most concern would be the False Negative values. These are the 3 patients who have heart disease but were predicted to have no heart disease present. On the humane side, you would be sending them home with a life-threatening disease. On the business side, this could result in legal action against the hospital.

What I Would Do Differently

After receiving feedback, I would have liked to change the threshold and therefore changing my model predictions and False Negatives. I would also like to improve my model's F1 score by adjusting/removing the features accordingly.