**DSC530 Term Project**

Eric Dickey

Department of Data Science, Bellevue University

DSC 530: Exploratory Data Analysis

Dr. Shankar Parajulee

August 13, 2021

Question

The two questions I will be analyzing will be the following:

1. Is heart rate a good indicator of a heart attack?
2. Is high cholesterol a good predictor for a heart attack?

Dataset and EDA

The dataset that will be examined is from the University of California, Irvine, Machine Learning Repository. This dataset contains variables associated with heart disease and associated risks of having a heart attack. While there are a total of 14 variables associated with the dataset, a subset of 5 numerical variables were chosen to be analyzed. The subset of fields was picked based on familiarity with their meaning and support the questions proposed. The following observations were made regarding their distribution identified via histograms:

* Age is roughly bell shaped with a slight tail to the left
* Cholesterol is bell shaped with a slight tail to the right and outliers present outside the 99th percentile
* Heart rate is bell shaped with a tail to the left
* Resting blood pressure is roughly bell shaped with a tail to the right and outliers present outside the 99th percentile
* The sex has almost double the number of males (1)

The outliers present in the cholesterol and resting blood pressure were identified as being outside of the 99th percentile of data. To address the outliers, they could be dropped or set to the 99th percentile value without adversely affecting the analysis.

Examining the distribution of two variables, heart rate and cholesterol, produced valuable data that can assist with the question for the data. An increased heart rate is a high-fidelity indicator of a risk for a heart attack, while cholesterol by itself is not a good predictor to use. The CDF for heart rate is fairly linear with a higher risk of a heart attack represented and cholesterol is fairly linear but affected by the outliers. The standard normal quartiles for both datasets are both fairly concentrated within 1 quartile of the mean.

The scatterplots for heart rate and cholesterol paired with age were unremarkable and were widely dispersed. It does highlight the quality distribution of the data amongst age range. Separating the values of heart rate and cholesterol on a scatterplot by heart attack risk does show higher clustering for high heart rate values and a risk for heart attack. There were not significant correlations found between the variables and the low p-values indicating the null hypothesis test was statistically significant, indicating they were appropriate for a model. A logistic regression model had low accuracy (when considering the medical impact) when using only heart rate and only cholesterol as a predictor for a heart attack (72% and 57% respectively).

Methodology and Feedback

The use of just one model type, logistic regression model, may not be the best way to evaluate if a model will be useful in creating a prediction on outcomes. Testing multiple model types may have identified a better model type to use. Additionally, using a subset of variables was not the best approach to analyzing this dataset and reviewing all fields first would have produced better choices on variables to examine closer for relationships. Heart rate by itself does show promise of being a useful indicator of heart attack risk, especially if coupled with other high-fidelity fields. Cholesterol, however, was determined to be a weak field by itself. In this analysis, the use of several libraries allowed exposure to different Python libraries, which performed nicely. However, deeply researching each library may have allowed the reduction of redundancies between the libraries used.