Kaggle Cardiac Data

Sam Loyd

The goal of my analysis of this dataset was to determine if being overweight, as measured by BMI, tended to increase the likelihood of having a cardiovascular disease. I also wanted to see if gender had any practical significance on the hypothesis above. I created a calculated variable from weight and height using the standard formula for determining body mass index or BMI. I started with a distribution analysis of the data. Most variables were found to have a non-normal distribution. Examples of this were provided in my power point and more fully explored in the notebook attached to the project. Extreme cases were removed using domain knowledge.

Spearman's method was selected for correlation analysis. I discovered that systolic blood pressure had the highest correlation to the binary variable for cardiac disease in this data set. BMI, while statistically significant, had a weak correlation that was confirmed by several hypothesis tests including Chai-squared. Two different models using multiple variable regression achieved a 72 percent accuracy at predicting cardiac disease. Correlation analysis and hypothesis testing of gender did not indicate a correlation with cardiovascular disease. A Chai-squared test also failed to prove a correlation for gender and cardio.

Unfortunately, the data set provided by Kaggle, did not provide much in the way of describing the population measured. The data was collected in a medical setting where measurements were taken by the examiner. In addition, survey answers were provided by participants. Age ranges would lead me to believe that the population is not strongly representative of the general population. That information is more detailed in the notebook attached to the project and beyond the scope of my summary.

In performing this analysis, I confirmed the importance of removing bias from my work. I was often surprised that the data was telling me something that showed less practical significance than what I expected going into the process. There were also several surprises. One

example from my regression analysis showed a relatively low pseudo r-squared value leading to a higher than expected accuracy value for the model. The categorical nature of the cardiovascular feature introduced challenges as well. I had to rely on new methods for visuals and testing not always provided for in the course.

I would like to feel more comfortable spending less time verifying findings and looking up ranges. For example, I ran tests that I knew would show strong correlation just to prove that I was performing them correctly. I am new to this and that comfort should come with time.

In summary, my analysis proved a significant, but weak correlation for BMI and heart disease. Two different regression models using systolic pressure alone in the form of simple regression had a 71 percent accuracy rating for predicting cardio disease with less risk of overfit than using multiple regression. And finally, the educational value of providing this analysis was invaluable to me.