Will Langas

Predicting Heart Disease

In 2020 alone, heart disease took the lives of nearly 700,000 Americans, making it the leading cause of death in the United States. Although the field of cardiology has advanced rapidly over the last few decades, heart disease remains a prominent issue in America. In this project, I will try to determine whether or not machine learning could be effective in predicting heart disease.

The dataset for my project comes from a study done by the University of California - Irvine, where 296 individuals were evaluated for the presence of heart disease, with 13 key measurements being recorded. Before I was able to start building predictive models, I first cleaned the data. This involved removing missing values, aligning the numeric column names with those from the original study, and filtering out outlier values amongst other tasks. This left me with 291 data points to work with which consisted mostly of men with heart disease (38.5% of the data).

I then proceeded to explore the relationship between heart disease and two factors that most Americans commonly associate with it, cholesterol and blood pressure. For cholesterol, the average level for an individual with heart disease was only slightly higher than those of individuals without heart disease (249 vs. 239). Similarly, the average resting blood pressure of someone with heart disease was only slightly higher than those of individuals without it (134 vs. 129). This relationship is further represented in **Figure 1**, where although we see that extremely high levels of cholesterol typically corresponding with heart disease, the rest of the graph is evenly distributed, and we don't see a significant correlation between higher blood pressure and heart disease. Additionally, a simple logistic regression model solely based on cholesterol and blood pressure yields low accuracy results of ≈50%, suggesting that a more holistic approach is necessary for predicting heart disease.

I then performed a principal component analysis over the 13 features included in the dataset. Standard scaling was necessary in this step, as values such as cholesterol typically have larger magnitude, ranging from 126 to 360, while other values such as the number of vessels discovered by fluoroscopy range from 0 to 3. Thus, the grey line seen in **Figure 2** can be discredited. With scaling, we can conclude that reducing the dimensionality of our data would have negative effects on the accuracy of any models constructed later on, as seen by the red line's constant increase when more components are included.

Lastly, I created a Sklearn pipeline that performed both standard scaling and logistic regression. This pipeline was then trained on 75% of the data, and tested on the remaining 25%. My model was able to typically achieve accuracy scores of ≈84%, with a cross validation that tended to range between 75% and 80%. I additionally performed a permutation test, which resulted in scores around 82% and a p value of 0.0099, implying that there was a real dependency between the features being tested and the target variable that was accurately captured by my model. I was able to further improve this model by setting the solver parameter of the logistic regression to lbfgs, which helped improve and stabilize the model around 85% accuracy, with cross validation scores hovering around 82%, and similar results in the permutation test. The coefficients for each feature as found by the pipeline are displayed in **Figure 3**, which suggests that gender, the number of blood vessels discovered, and the presence of Thalassemia tend to influence predictions the most, with age consistently being a relatively unimportant factor.

The relatively high accuracy of even a basic pipeline leads me to believe that at this scale, machine learning is effective at predicting heart disease, and with more data or more complex models, this project suggests that machine learning could play a significant role in heart disease research and prediction on a larger scale.

Figure 1: Blood Pressure vs. Cholesterol By Heart Disease

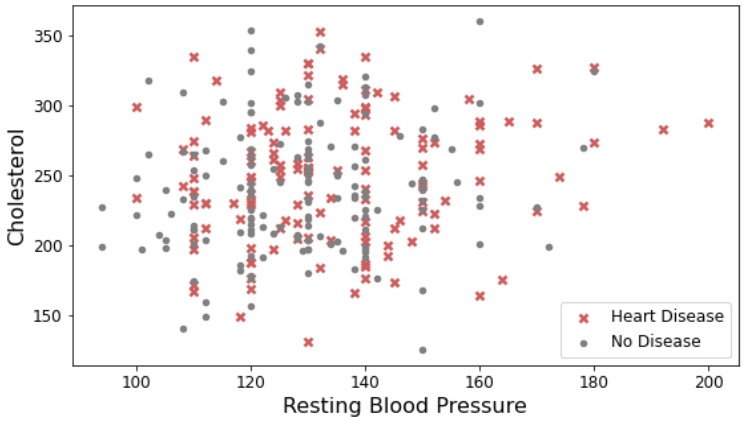

Figure 2: Number of Components vs. Explained Variance

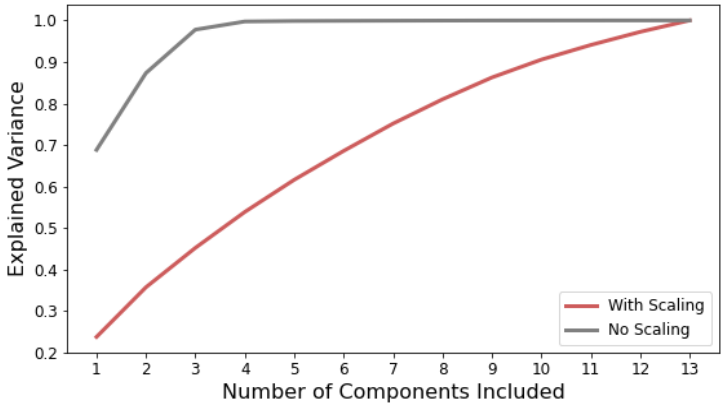

Figure 3: Logistic Regression Coefficients by Feature