Titile: Death Prediction by Echocardiogram Data

Name: Xin Bu

According to the report of WHO(World Health Organization), heart disease has been the primary cause of human death for the past two decades. After the patient's heart attack, the doctor needs to determine the patient's condition based on the echocardiogram data and make sure to give enough follow-up attention to the high-risk patient. My goal for this project is to explore several factors' impact on predicting if patients will survive for at least one year after a heart attack, which would help doctors and patients to have a better understanding of the death risk by their echocardiogram data.

The UCI machine learning repository provides an echocardiogram dataset of 132 patients online at https://archive.ics.uci.edu/ml/datasets/Echocardiogram. The dataset contains 12 columns describing patients' personal and echocardiogram information, such as patients' names, left ventricular end-diastolic dimensions, scores of left ventricle's motion, whether patients are alive one year after a heart attack. For convenience, I call whether patients are alive one year after a heart attack as vital status in the following. There are a number of fields that are unrelated to vital status(for example, patients' names and research groups are unrelated to the result).

Six features of the data are selected for the analysis: (1) age when heart attack occurred, (2) Pericardial effusion, (3) fractional-shortening(a measure of contractility around the heart), (4) E-point septal separation(another measure of contractility), (5) left ventricular end-diastolic dimension(a measure of the size of the heart at end-diastole), (6) wall motion score(a measure of left ventricle's motion). Expect pericardial effusion, other features contain a few missing values. I replace those missing values with the median values for those features. To make sure we know patients' vital status for making an accurate prediction, I drop the rows that miss or contain abnormal values of the vital status column. Pericardial effusion is a binary variable to express either it is fluid or not, which is no need to duplicate divide it into two columns.

I start by exploring the impact of age and wall motion score on vital status. To make it explicitly, I divide age into three categories: below 60 as "60-", between 60 and 70 as "60-70", above 70 as "70+". The number of patients in each age category is close. For each age category, the patients are divide into two groups: one with a wall motion score below its median and another one with a wall motion score above the median. The survival percentage is the percentage that patients in each group alive one year after a heart attack.

From **Figure 1**, we can see there is a trend that the older the group of patients, the higher the survival rate they have. For instance, the survival rate of the groups of patients over 70 years old with wall motion scores above the median is 47% higher than the one below 60 years old. Moreover, **Figure 1** shows the higher the wall motion score, the higher the survival percentage, such as the patients between 60 and 70 years old with wall motion score above the median have 50% more survival percentage than those with wall motion score below the median.

Applying a principal component analysis on my 6 features is the way that go further into the dimensionality of my data. In **Figure 2**, the blue line shows that 3 columns capture 100% of the explained variance. However, it could be the reason that I have not transform the original data into a standard scale. When the feature column value in the original data is large, the weight of its proportion is greater. For instance, the age has a larger magnitude than the pericardial effusion column which only have 0-1 values. The red line(after using StandardScaler) shows that it is impossible to reduce the dimensionality of my data without lossing significant information.

To achieve the prediction for patients, I create a sklearn pipeline model consisting of (1) a StandardScaler to standard all the columns, (2) a LogisticRegression to estimate base on the data. To ensure the efficiency of this simple sklearn pipe, I made three comparisons about it. First, performing a PolynomialFeatures after scaling have the exact same performance as the simple one. Second, insteading of using the median to replace the missing value, I try a SimpleImputer with the "most frequence" strategy to fill in missing pipe dimensions, which have the exact same performance as the simple one. Third, using "balanced" strategy for the LogisticRegression to put more weight to patience that alive after one year, which also have the exact same performance as the simple one. Therefore, we can trust the efficiency of the sample pipeline model, which has a 0.74 score, 0.67 recall and 0.57 precision.

The coefficient weights for each of the features by the sample pipeline model are presented in **Figure 3**. We observed that the two features explored in **Figure 1** - age at heart attack and the wall motion score are playing the most significant role on estimate patients' vital status. Moreover, the epss(E-point septal separation) and pericardial effusion also have a high weight over 0.6. This means that doctors need to pay more attention to patients with outliers in those weighted features.

To conclude, my analysis shows that doctor should pay more attention on patience with fluid pericardial effusion, high wall motion score, high E-point septal separation and young age as they are more tend to die one year after heart attack. In the end, thanks for the help of Professor Tyler and the inspiration of the example provided by Wen Ye, Gautam Agarwal, and Bryan Jin.

Figures

Figure 1: Age & Wall Motion Score Effects on Survival Rate

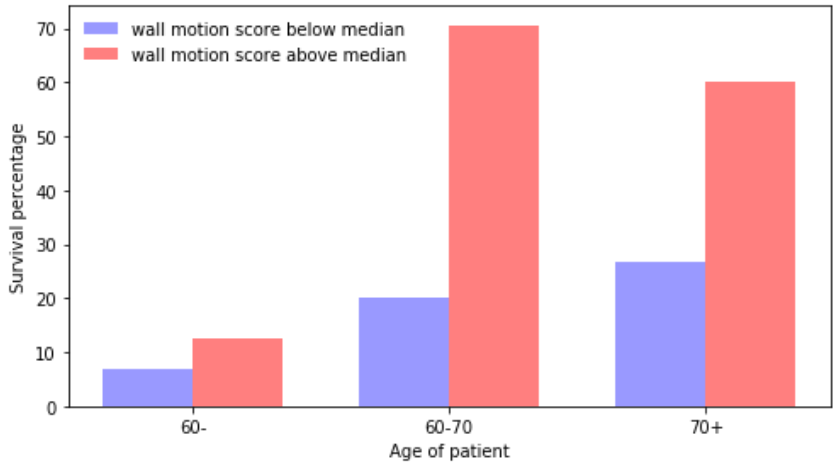

Figure 2: Principal Components of Breaks

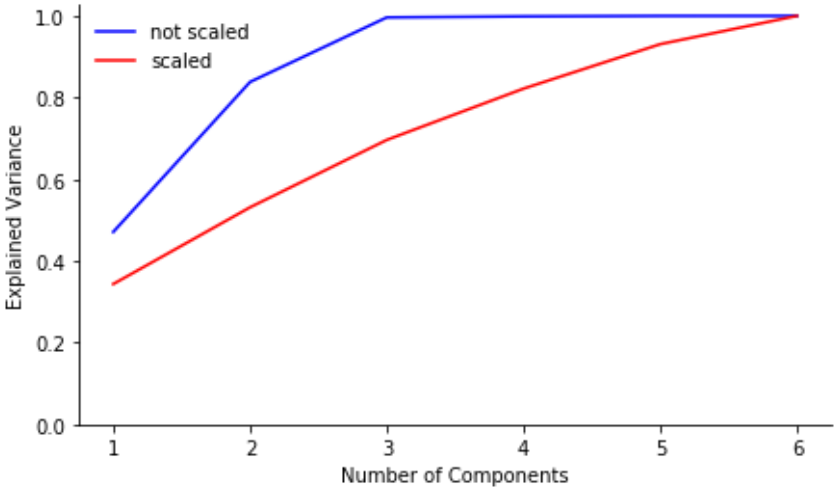

Figure 3: Logistic Regression Coefficients