Johnny Chapman, Conrad Frisch, Lai Pham

MGSC 310

Dr. Hersh

13 Dec 2019

Coronary Heart Disease

**Introduction**

The leading cause of death for men and women in America is heart disease. The most common form of heart disease, coronary heart disease (CHD), kills over 370,000 people annually. In addition to being common, having CHD is also expensive. Heart disease costs the US on average $200 billion annually, with an average out-of-pocket expenditure of $2,450. To cover the rest of the medical fees, insurance companies are constantly having to pay for very expensive procedures to help people with this disease.

This forms the basis of our project. By using the variables in the dataset, the model can predict those who are at risk of CHD for them to get preventive care, or a procedure needed to correct CHD. The variables that were used to predict those at risk for CHD, came from the framingham heart study dataset.

First many of the variables were transformed into factors, such as education, if the person is on blood pressure medication, and other medical based variables. Next, other classes of variables were created. For example, the category heavy smoker and overweight. An individual would be considered a heavy smoker if that person smokes more than 8 cigarettes a day, with 8 being the average number of cigarettes per day in the dataset, and overweight if that person has a BMI more than 25, with 25 as what the CDC considers as overweight).

The transformed data was then compared to real world statistics to see if the data collected reflects the population in America. The results were that while in the heart dataset, there were 49% smoked cigarettes, while in the real world only 17% smoke cigarettes (CDC) and 8% smoke some type of vape product (Gallup). On the contrary, the dataset had 54% being overweight, while in America, 70% are considered overweight (NHANES).

Using a correlation matrix to see the relationship of the variables in the model, there were no surprises. The highest positive correlated variables were blood pressure and hypertension and the highest negative correlated variables were age and cigs per day. As hypertension is a condition of having high blood pressure, the correlation matches our expectations. The female and male graph shows that on average men smoke more cigs per day and the individuals currently at risk of heart disease smoke more cigs per day as well. The pirate plot of cigs per day and education shows that education does not play a factor. This was unexpected, as we presumed those who are more educated to be aware of the harmful toxins in cigarettes, thus smoking less. The age and systolic blood pressure plot indicates that overall those at risk for heart disease have a higher blood pressure, steadily increasing with age.

**Logistic Regression**

We wanted to predict if a person is at risk of coronary heart disease. For this we decided to use a Logistic Regression model for this classification problem. In our model, we found that the three things that heavily increase chances of being at risk for CHD were being male, having experienced a stroke, and a higher log heart rate. We also discovered that glucose had a very little effect on increase your risk, despite the fact that in the summary of our model it was listed as statistically significant. Our ROC plot showing the false positive over the true positive, generated an area under the curve of 0.74. In this model, we chose a cutoff of 0.08, which was able to maximize our sensitivity rate and increase the number of false positives. We were able to get a sensitivity rate of 0.94 and a specificity rate of 0.35.

**Random Forest**

One of the models used in the project, was random forest model, which constructs multiple decision trees through random sampling and subsets. This method is well known in classification prediction model as the multiple diverse trees minimize the error of individual trees. First 1000 trees were randomly created with the errors plotted to verify how many trees minimum should be ran for the least error. This plot showed at least 50 trees should be used for the lowest error, which was used for the model. Using the mindepth plot, the random forest model concluded age as being the variable with the most predictive power, while cigs per day was not. The default cut off for random forest is 0.5, however, predicting with that model produced sensitivity of 0.02 and specificity of 0.99. As the purpose of the study is to identify the true positive, the cut off value was adjusted to increase the sensitivity. After changing the Random Forest model to 0.9 cut off value, the results were a sensitivity of 0.84 and a specificity of 0.42. This is a relatively high sensitivity with a slightly significant specificity value.

**Lasso & Elastic Net**

In order to help with variable selection, we used a lasso model. The model had low differences in error from the removal of multiple variables. At 1 standard error away from the lambda that offers the highest area under the curve - we have 4 variables that are selected. These variables did not come as a big surprise, they were if a person is a male, age squared, systolic blood pressure, and glucose levels of a person. Based on the ROC plot of the lasso model, we chose a cutoff of 0.1 in order to maximize sensitivity. From this we got a sensitivity of 0.96 and specificity of 0.20. Depending on what the model is used for (whether it be from an insurance standpoint, a doctor deciding to do a surgery), the model can be tuned by altering the cutoff point to a higher cutoff to get a higher specificity or try and optimize the values of both sensitivity and specificity.

However going from the Lasso model we decided to penalize multiple variables less by using an Elastic Net model with an alpha of 4. We maintained the same cutoff of our model, and got a slightly lower sensitivity of 0.92, but a much higher specificity of 0.35. As in the Lasso model the cutoff point for the Elastic Net Model can also be altered depending on what is needed from the model (high sensitivity, or high specificity).

**Conclusion**

All of the models that we have used to predict CHD were tuned to offer a high sensitivity. However, by altering the cutoff point for each model we can tune the models for a high specificity. If people are using the model to see if they are at risk they may want a model with high specificity, as that way they can make healthier decisions to prevent CHD. Surgeons on the other hand may be more interested in a model that offers high sensitivity as they do not want to perform a surgery unless the patient absolutely needs it.

Of all the models that we used, the Lasso model had was able to have the highest sensitivity, while also having the lowest specificity. The Elastic Net model wasn’t able to predict with as high sensitivity as the Lasso model, but was able to generate higher specificity. The Logistic Regression model had the next highest sensitivity with a good specificity. While the Random Forest model while not having a high sensitivity as the other models offered the best specificity while still having a high sensitivity. We purely measured these models based on how high their sensitivity was but with a different measure the ranking would be different. The models can also be tuned with different cutoff levels so they can achieve a high specificity, sensitivity, or optimize both sensitivity and specificity.

In the future, if we could gather data about the kind of heart disease a person is a risk for and the remedies for the different diseases, we could model this problem from a regression standpoint and see what type of treatment is optimal given a person’s underlying characteristics.