462 Final Project

Srinivasan Sathiamurthy and Rebecca Derham

April 27, 2023

1 Introduction

For this project we predicted whether or not an individual has heart disease based on 735 data points of 11 demographic and biomedical features, which are:

- Age (Numerical- in years)
- Sex (Categorical- Male or female)
- Chest Pain Type (Categorical- Typical Angina, Atypical Angina, Non-Anginal Pain, Asymptomatic)
- Resting Blood Pressure (Numerical)
- Cholesterol values (Numerical)
- Whether or not their fasting blood sugar was more than 120 mg/dl (Categorical)
- Resting electrocardiogram results (Categorical- normal, ST-T wave abnormality, probable or definite left ventricular hypertrophy)
- Maximum heart rate achieved (Numerical)
- Whether or not they had exercise-induced angina (Categorical)
- Oldpeak values (Numerical)
- The slope of the peak exercise ST segment (Categorical- up, down, or flat)

2 EDA and Feature Engineering

We first started out our data exploration by determining whether the data set is balanced with respect to the number of individuals with and without heart disease. We found this to be the case with 396 individuals with heart disease, and a similar number of 339 without.

We then looked at the distributions of each feature, and found two notable abnormalities. The first of which was that 138 data points, or 18.8% of the training data, had 0 cholesterol,

which biologically doesn't make too much sense. However, when we considered the relationship between cholesterol and heart disease, we found that in our training data, 87% of the individuals with zero cholesterol values had heart disease, while only 46% of the individuals with nonzero cholestrol values had heart disease. Moreover, when we applied a ks test to compare the distributions of individuals with nonzero cholestrol values with and without heart disease, we determined that there was not a statistically significant difference between the two densities, therefore we determined that the magnitude of an individual's cholesterol did not contribute any significant information about whether they had heart disease or not. This led us to transform the cholestrol feature into a simple indicator of whether or not their cholesterol value was nonzero.

The second notable observation pertained to oldpeak: 40% of the training data had oldpeak values of 0. We also observed that out of the ten negative oldpeak values, 7 of them had heart disease, only 31% of individuals with 0 oldpeak had heart disease, and a one-sided t-test on the individuals with positive oldpeak values showed that individuals with heart disease and positive oldpeak values had statistically higher mean oldpeak values than individuals without old peak. This led us to believe that heart disease was correlated with the deviation of oldpeak from 0. Since this feature was already unbalanced with a disproportionately large amount of zero values, we opted to turn the oldpeak feature into an indicator of whether or not the oldpeak value is nonzero instead of employing an absolute value measure. Besides these transformations, we decided to encode the categorical variables using dummy encoding in our logistic regression model, and one hot encoding in our random forest model (more on that in the next section). We included the remaining numerical features as is.

We also considered covariates. Our analysis for which covariates to include was split into three groups: categorical and categorical, categorical and numerical, and numerical and numerical.

For the first group, which consists of two predictors that are both categorical, we determined whether the covariate of the two categorical features combined was significant if the distribution of heart disease with respect to the first variable differed across the levels of the second variable using a simple histogram. The histogram on the left depicts an example of a covariate we deemed significant, and the one on the right is an example of a covariate that we did not deem significant:

In doing so, we decided to include the following covariates:

- Our indicator feature for Oldpeak and ST Slope

For the second group, which consists of a numerical and categorical predictor, we employed a similar approach, but instead considered the density plots of the first (numerical) variable by level of the second (categorical) variable. The plot on the left depicts an example of a covariate we deemed significant, and the one on the right is an example of a covariate which we did not deem significant:

In doing so, we decided to include the following covariates:

- Oldpeak and ST Slope

And finally, for the third group, which consists of two numerical predictors, we merely checked for multicollinearity by selecting the features with the highest correlations. We used a threshold of a magnitude of 0.25 to determine whether or not a correlation was large enough to include the covariate, and in doing so, decided to include the following covariates:

- Age and Max heart rate
- Age and Oldpeak
- Cholesterol and Max heart rate

3 Supervised Analysis

3.1 Introduction

Before we began the project, Cheenu wanted to use random forests for the dual purpose of feature selection and prediction, while Rebecca opted for logistic regression for the same purposes. Simply weighing the pros and cons of both models, we observe that random forests is a much more complex model than logistic regression, and is thus more prone to overfitting. As such, before we even trained our models on the data, we decided that we would only consider random forests as the better model if the following conditions were satisfied (in their order):

- Subject to the same train/validation data splits, Random forests has a better 10-fold cross validation accuracy than Logistic regression
- If our random forest model has a better 10-fold cross validation accuracy, we would bootstrap the standard deviation of the errors of the two models, and determine whether the difference in errors is significant. If so, only then would we choose Random forests over logistic regression, since logistic regression is the simpler model and is less prone to overfitting.

Now going back to the features we decided to include (for both models), to reiterate, we engineered our features from the data by taking everything except Cholesterol and Oldpeak as is, and turning those two into indicators of whether or not their values are equal to zero. For our covariates, we simply multiplied the features together for the numerical variables, and for categoricals, we encoded them via one hot encoding, and then multiplied that by the encoding of the other feature (simple multiplication if the second feature was numerical, and cross-product one hot encoding if the second feature was categorical- by cross product one hot encoding, we mean that we created an indicator feature for each combination of categories from each variable). Now for normal categorical encoding, we decided to opt for one hot for random forests due to interpretability, and dummy encoding for logistic regression to avoid multicollinearity issues.

Now that we basically used the same features with slight modifications for how we encoded

categorical variables, we also decided to split the training dataset into 635 points for hyper parameter tuning, feature selection, and model selection, and the remaining 100 points for estimating our model's test accuracy. With our 635 points, we also decided to use 10-fold cross-validation for consistency in how we measured our models' predictive power. For the remaining of this section, we will discuss our training procedures, results, and predictor relationships in more detail.

3.2 Random Forest

The main difference in the way we approached random forests versus logistic regression stems from the fact that fitting random forests require hyperparameter tuning in addition to feature selection. To account for this, we first grid searched through reasonable ranges of hyperparameters, picked a set of hyper parameters which had reasonable performance for sets of hyper parameters with similar values to make sure the hyperparameters weren't overfitting themselves, and then calculated the mean decrease gini score of each feature with the model fit on this set of hyperparameters as a measure of the predictors' importance within this intermediary model. We then selected a reasonable number of features by employing a threshold on this measure (selected all the features with an MDG score greater than a reasonable value, explained later), and again grid searched for a good neighborhood of hyperparameters. Our final model then consisted of this final set of features and hyperparameters.

Looking at this process in more detail, the hyperparameters we considered for our grid search process were the number of variables randomly sampled as candidates at each split of the decision tree (mtry), the minimum size of terminal nodes in each decision tree to control the depth of the decision trees (nodesize), and the number of trees to be grown in the random forest (ntree). Since fitting random forest models is relatively slow, we decided to first perform a grid search with a wide step size between the parameters, find a good performing neighborhood of hyperparameters, search through that neighborhood with a much smaller step size, and then choose the best-performing set of hyperparameters whose neighborhood of parameters also perform well (to avoid overfitting). A general rule of thumb for good mtry values is searching the range between a third of the number of our features, and the square root of the number of our features, which we widened slightly in the beginning to include all the integers between 3 and 8, inclusive. For nodesize, we decided to first iterate through from 1 through 10. For ntree, datasets with moderate size generally use between 100 and 500 trees, so we decided to loop through that range with a step size of 25. After doing this initial search and computing the tenfold cross validation for each combination of hyper parameters, we found that the best performing combinations tended to have nodesizes between 4 and 6, mtry between 3 and 6, and ntree between 215 and 290. We then looped through these ranges, with the ntree stepsize now 5, and finally selected (nodesize, mtry, ntree) = (4,6,245). After fitting this model, we then computed the MDG scores for each feature and obtained:

To understand what the mean decrease gini score is, note that gini impurity is a measure of how often a randomly chosen datapoint would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset, which is essentially the underlying

distribution of heart disease for each feature in our training data set. The Mean decrease gini score then measures the reduction in Gini impurity across all decision trees in the random forest model we fit that is attributable to each predictor variable, respectively. This essentially is measuring how much each predictor feature improves the prediction accuracy across all the trees from the base error rate.

With this in mind, we selected a threshold of 6, which gave us the final set of features:

1. Age
2. Gender
3. Whether or not an individual had Atypical Angina Chest Pain
4. Whether or not an individual had Asymptomatic Chest Pain
5. Resting Blood Pressure
6. Whether or not their cholesterol levels were nonzero
7. Max heart rate
8. Whether or not they had exercise-induced angina
9. Whether their ST-Slope was flat
10. Whether their ST-Slope was up
11. The covariate between their age and max heart rate
12. The covariate between their age and resting blood pressure
13. The covariate between their age and oldpeak value
14. The covariate between their cholesterol and max heart rate

With these fourteen features, we again grid searched locally optimal hyperparameters through the same procedure, and ended up with (nodesize, mtry, ntree) = (7,4,235). This final model had a 10-fold cross validation accuracy of 85.82341%.

3.3 Logistic Regression

For logistic regression, we didn't have to tune hyperparameters since the logistic regression model does not have any hyperparameters. For feature selection, we used glm's bidirectional stepwise method, which iteratively added and removed variables to maximize the AIC criterion, a measure that measures the goodness of fit of a model that penalizes the number of parameters to prevent overfitting. Doing so obtained a AIC score of 404.49, the following 12 features, and a 10-fold CV accuracy of 88.82192%

- Gender

- The first three chest pain types (dummy encoding)
- Cholestrol
- Whether or not their fasting blood sugar was more than 120 mg/dl
- Whether or not they have exercise induced angina
- Whether or not their oldpeak values are nonzero
- Their ST Slope (dummy encoding)
- The covariate between their age and oldpeak value
- The covariate between the oldpeak indicator and whether or not the st slope was up
- The covariate between their age and resting blood pressure

Another thing to note is that the p-value of all of these coefficients (including the intercept) were significant, except for the covariate between age and resting blood pressure. Additionally, the logistic model and the random forest model share many common features, but also have some differences, such as the fasting blood sugar feature in the logistic model, and max heart rate in the random forest model.

3.4 Model Selection

Since the 10-fold accuracy for the more complicated random forest model was less than that of the logistic model, it was straightforward to select the logistic model for our final predictive model.

3.5 Predictors versus Predictions

Analyzing the relationship between the predictors and predictions in our model is quite straightforward: variables with positive slopes are correlated with heart disease, and variables with negative slopes are correlated with the lack thereof. Looking at only the slopes with significant p-values, we see that these features are positively correlated with heart disease:

- High age and high oldpeak values
- Having nonzero oldpeak values and and Up ST-Slope
- High age and high resting blood pressure
- High fasting blood sugar levels
- Having exercise-induced angina

And these are correlated with the lack of heart disease:

- Female

- Having Atypical Angina Chest Pain
- Having Typical Angina Chest Pain
- Having Non Anginal Chest Pain
- Nonzero cholestrol levels
- Up ST Slope
- Down ST-Slope
- Nonzero oldpeak values
- Having nonzero cholesterol levels

Since we used dummy encoding, we can infer that having zero cholestrol values, having a flat ST-Slope (not Down or Up), having asymptomatic chest pain (not ATA, TA, NAP), and being male, all correlate with having heart disease as well, which logically makes sense.

4 Analysis of Results

4.1 Error analysis

Now when we performed an error analysis by looking at whether our final model (the logistic one) was

4.2 Next Steps

It would be nice to have more data to work with, since that would help with the overfitting issues of the random forest model, and perhaps even enable it to perform better than the logistic regression model. We would also like more data points with Females, high fasting blood sugar values, and more down ST-Slope values, since those features are quite heavily unbalanced. We would also like to perform more rigorous tests to determine whether the initial distribution of heart disease differed from the distribution of heart disease among misclassified points for each feature to see whether our model has biases in any feature, which would help us determine whether we need to add more features to remove such biases.

Unrelated to improving our predictive performance, it would also be nice to know what 0 meant in cholesterol, since it seems that a 0 cholesterol value is significant in our model's ability in predicting whether that person has heart disease or not, but biologically it doesn't make sense for a person to have absolutely no cholesterol.

Steeeeeps. Groovy. Steeeeeeeeeeieeeeeeeeeps.