Factors Analyzed for their Contributions to Heart Disease

Capstone 1 Report, V. Moore

Introduction

Predicting heart disease risk

The problem

Heart disease is one of the greatest causes of death, though it is not necessarily clear which patient health factors contribute most to risk of heart disease.

Potential clients

Understanding of the relative risks leading to heart disease would be useful to multiple stakeholders. The patient care community, as well as patients themselves, could benefit from awareness of specific risk factors. Personalized medicine is also potentially aided by an interpretable model capable of ranking risk factors for heart disease. This could potentially improve diagnostics and enable a patient and his/her doctor to understand the patient's particular risks, possibly providing actionable insights for mitigating risks. Insurers, whether private or publicly administered, would have an interest in knowing which factors make people vulnerable to heart disease and its outcomes, both to administer resources and to incentivize strategies to lead to optimal health.

Heart disease is not simply a problem of industrialized nations, either, and non-governmental organizations could make use of knowledge of risk factors for treatment and prevention, including education, in areas where they are able to make an impact on public health. Publicly funded medical researchers and pharmaceutical companies could use information about the etiology of heart disease to develop treatments or preventive medications, since it is quite possible that heart disease is mediated not by lifestyle choices alone.

Data sources

Data used for this machine learning project come from the University of California, Irvine, Machine Learning Repository. Datasets regarding heart disease risk factors are used here from three distinct patient populations, including Cleveland, Ohio (N=282), Long Beach, California (N=200), and Hungary (N=294). There are processed datasets available (examining 13 possible contributing factors, in addition to the target variable), but they do contain missing data to be addressed in the process of preparing the data for machine learning methods.

Project expectations

The target variable to be predicted in our machine learning model is presence and level of heart disease, which is organized within the data as a classification system with five levels (including a category for absence of heart disease). As the structure of this target variable is imbalanced among several levels of severity, this can affect classifier performance, so analyses were split between binary target

classifications and multiclass classifications. For binary target analyses, we chose a binomial, categorical classification system (heart disease diagnosis likely versus not) based on labeled data and applied supervised classification models to determine probability of presence or absence of heart disease. Additionally, we tested model performance at categorizing predictions into any of the five possible classes using the same classifiers as for the binary target analyses, but with the dataset treated using interpolation by applying SMOTE to the training data set for multiclass analyses to counter the imbalanced nature of the target data.

After exploring classifiers for accuracy, we then determine global features of importance for key models, and then interpret models using SHAP values to ascertain individual features of importance. Such insight when applied to health or diagnostic models has potential to greatly enhance an individual's understanding of his/her own risk factors and lead to actionable guidelines for optimal health.

While it is useful to have three separate populations from two countries included in this data analysis, a caveat is that all have come from Western, industrialized nations, possibly limiting the power of any model arising from this analysis to extrapolate to other regions of the world.

This report is accompanied by the following items: code for importation of data, data cleanup, exploratory data and inferential analyses, and model development and interpretation. A project report and a slide deck summarizing the project and its findings is also provided.

Table 1 shows rates of heart disease in locations from the United States that are relevant to this study, from data collected by the United States Centers for Disease Control. The rates come from a prevalence calculated as described here. Statewide, Ohio shows a higher rate of heart disease than does California, while within these states Cleveland, Ohio, seems to show a higher rate of heart disease than does Ohio, while Long Beach, California, seems to show a rate of heart disease comparable to the rate for California.

Table 1. Heart disease rates in key locations (from CDC.gov)

Location	Heart Disease Rate
Ohio	7.62
Cleveland, OH	8.76
California	5.30
Long Beach, CA	5.46

Data description and cleanup

Data sources

The data for this analysis comes from three datasets that each contain the same columns of information, and the datasets are available from the University of California, Irvine Machine Learning Repository. The variables (listed in Table 2 according to definitions that accompany the datasets) encompassed by these

data include years of age (age), sex (sex), chest pain type (cp), resting blood pressure (trestbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate (thalach), exercise-induced angina (exang), ST depression induced by exercise and relative to rest (oldpeak), slope of the peak exercise ST segment (slope), coronary angiography score (ca), thalassemia defect (thal), and numerical score (num) representing the degree of heart disease detected by angiography. This last value is the target variable.

This list consists of both discrete variables and continuous variables. Age, trestbps, chol, thalach, and oldpeak are continuous variables, while the remaining variables are categorical in nature. The target variable, num, is a categorical variable on a scale of 0 to 5, with 0 meaning no heart disease and 5 being the level of greatest severity.

Table 2. Heart disease feature names and their abbreviations

Variable Name	Definition of Variable (from UCI Machine Learning Repository)
age	age
sex	sex
cp	chest pain type
trestpbs	resting blood pressure
chol	cholesterol level
fbs	fasting blood sugar
restecg	resting electrocardiographic results
thalach	maximum heart rate
exang	exercise-induced angina
oldpeak	ST depression induced by exercise and relative to rest
slope	slope of the peak exercise ST segment
ca	coronary angiography score
thal	thalassemia defect
num	numerical heart disease score

Data cleanup

The missing values were originally denoted differently according to data file. The code for a missing value in data from the Cleveland, OH, and Long Beach, CA, datasets was a "?", and a missing value in the Hungarian dataset (reprocessed, used here) was denoted as "-9.0". In each case, a missing values was replaced with "NaN" prior to further processing. Initially datatypes were all treated as floats for flexibility in assigning NaN entries, but later the datatypes were altered as appropriate for the data.

After replacing missing values with NaN's, the next step was to examine data distributions for each variable by viewing histograms to assess whether values show normal distributions, and to identify if there were obvious and inappropriate outliers (for instance, a missing value that should have been

noted as "-9.0" but was entered as "9.0" for a categorical variable whose values were not to exceed 3). Additionally, getting an idea of the nature of a variable enabled the decision whether to assign missing values the median or mode of population values for the variable. In general, if a variable showed a relatively normal distribution, particularly for continuous data, then the median was used to assign to missing values. If a variable was not distributed in this way, particularly for categorical data, the mode was typically chosen. Exceptions to this method of value-assignment were allowed as seemed appropriate for a variable, and all are subject change as appropriate according to machine learning model development. This step of value-assignment for missing values was taken prior to consolidating data among populations in case there were location-based differences in median or mode, which seemed likely given that histograms identified slight variation in populations for some variables. For the "num" variable, one entry was removed from the Hungarian dataset prior to dataset consolidation due to a single missing value, and for all datasets the "num" datatype was adjusted as an integer datatype.

Next, the datasets for each population were consolidated into a single dataframe. A new column was added with the location information for each entry for future reference. It should be noted that of all the source files, the Cleveland dataset was initially the most complete with the least missing values, so this population has potential to influence overall analyses most highly.

After consolidation of all data into a single dataframe, datatypes were adjusted if needed for the nature of the data in each column. Continuous variables were kept as floats, and categorical data were converted to integers to allow for flexibility with calculations with the data (rather than set to categorical or Boolean values at this point).

Exploring the data

Preliminary examination of the data

The initial interrogation of data surrounded visualization of qualitative patterns in data using basic exploratory data analysis tools. This was partially involved in the data cleaning steps described above, as population distributions of variables were used for preliminary value-assignments for missing data from each of the three component datasets. Figures 1 and 2 below show histograms for all variables (Figure 1) and with more detail for continuous variables (Figure 2), with curves overlain with normal curves in black for visual comparison. These allow visualizations of the nature of each variable.

In general, throughout this study a null hypothesis is that markers of poorer cardiac health, as well as increased age, are not related to presence or level of heart disease. This can be examined most accurately for categorical variables through chi-square tests of independence, as the target variable is also a categorical variable. Another null hypothesis is that more broadly none of the variables in this study influence each other. Categorical variables can be examined versus each other using chi-square tests of independence, and continuous variables can be measured against each other using Pearson correlation coefficient analyses. These will be shown in the next section.

A primary aim of this project is to use the variables in the dataset for quantitative predictive analysis through machine learning techniques. If a null hypothesis cannot be rejected, then it likely will be of no value for prediction. The strength of a statistically significant relationship between variables may indicate predictive utility for the machine learning model. However, strong relationships between variables other than the target variable may complicate the model.

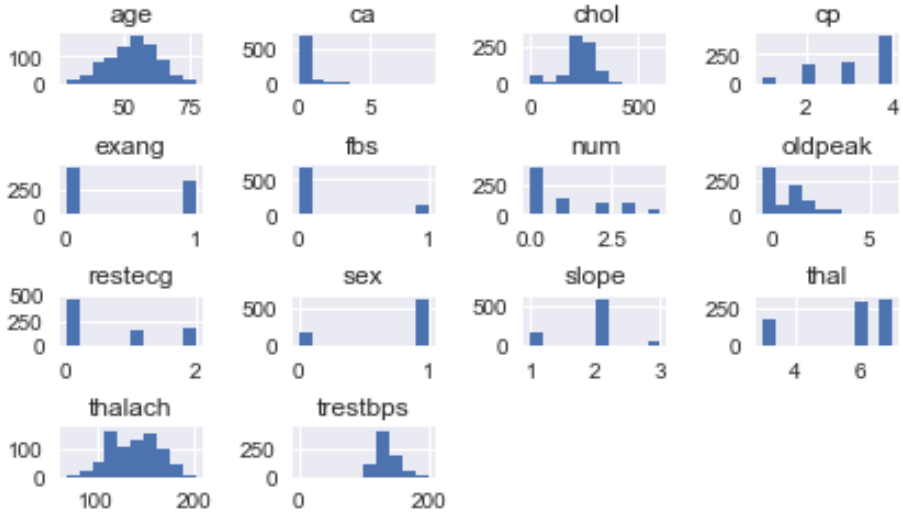
Apart from prediction of the target variable, one possible prediction could be that age itself is correlated with markers of poorer cardiac health. Another prediction could be that sex is an important factor in heart disease risk, with men showing greater proneness. It is important to keep in mind, though, that women may show different predictive features than men do.

Linear regression plots of some of the continuous variables (Figures 3-8 below) provide another preliminary-level analysis of some possible relationships of interest. Trendlines for these are differentiated by sex to see if there were differences between men and women for any patterns.

Not all possible comparisons appear here in the preliminary figures because most data are categorical, so they will be examined separately from continuous variables. However, continuous variables are compared, basic statistics for each variable are achievable, and some preliminary interpretations are possible. The later analyses in this notebook will explore correlation coefficients and chi-square contingencies between variables.

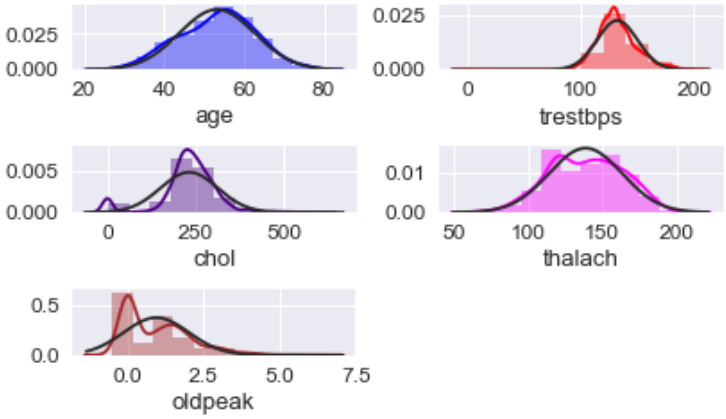
Figure 1 shows the general structure of data from each variable in the consolidated dataset, with most variables being categorical in nature. For many individuals the target variable, "num", shows the highest number of individuals to have a category of "0", which indicates lack of detected heart disease. However, there are many individuals in this dataset who appear to have detectable heart disease (categories 1-4).

Figure 1. Histograms of heart disease analysis features



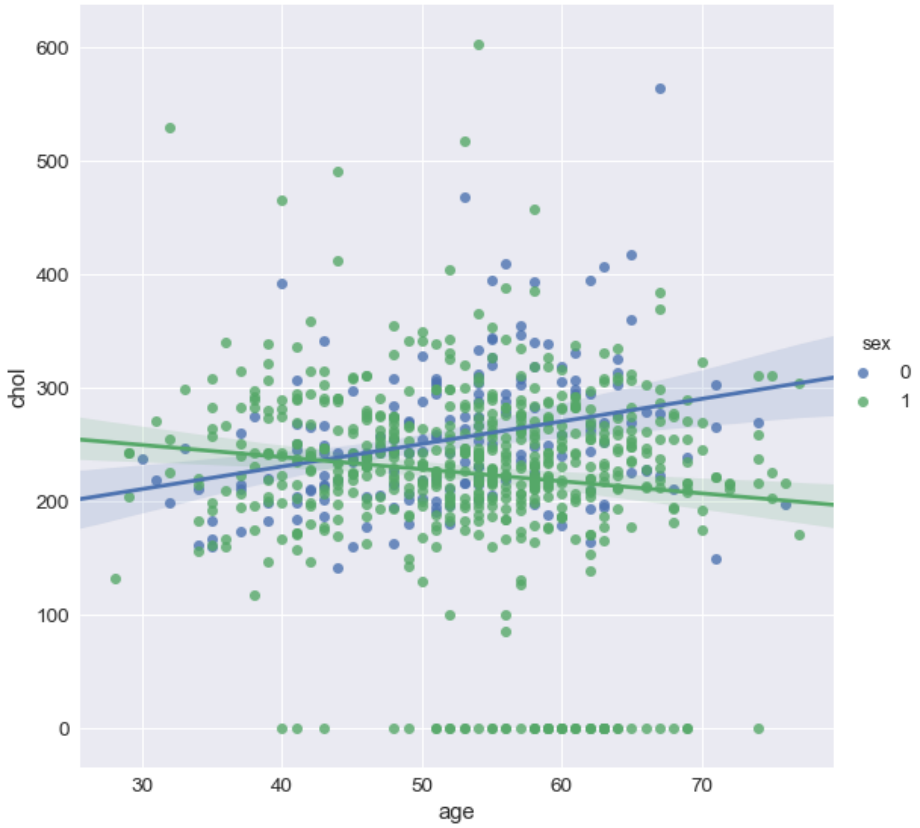
In Figure 2 continuous variables from this dataset are shown with normal curves overlain in black to enable visual comparison of data from each variable versus expectations if data are normally distributed. The variables age, "trestbps", and "thalach" show good agreement with normal curves, while "chol" and "oldpeak" diverge a bit more, though chol shows a mean close to that expected with a normal distribution.

Figure 2. Histograms of continuous variables with superimposed normal curves (black)



Figures 3-8 show relationships between continuous variables, examined as simple linear regressions but divided by sex.

Figure 3. Age vs. cholesterol level (sex: 0=female, 1=male)



In Figure 3, it is hard to distinguish an obvious relationship between age and cholesterol, though by sex it appears from the data available here that for women there is a trend between increasing age and cholesterol level. For men, the trend appears in the opposite direction using these data.

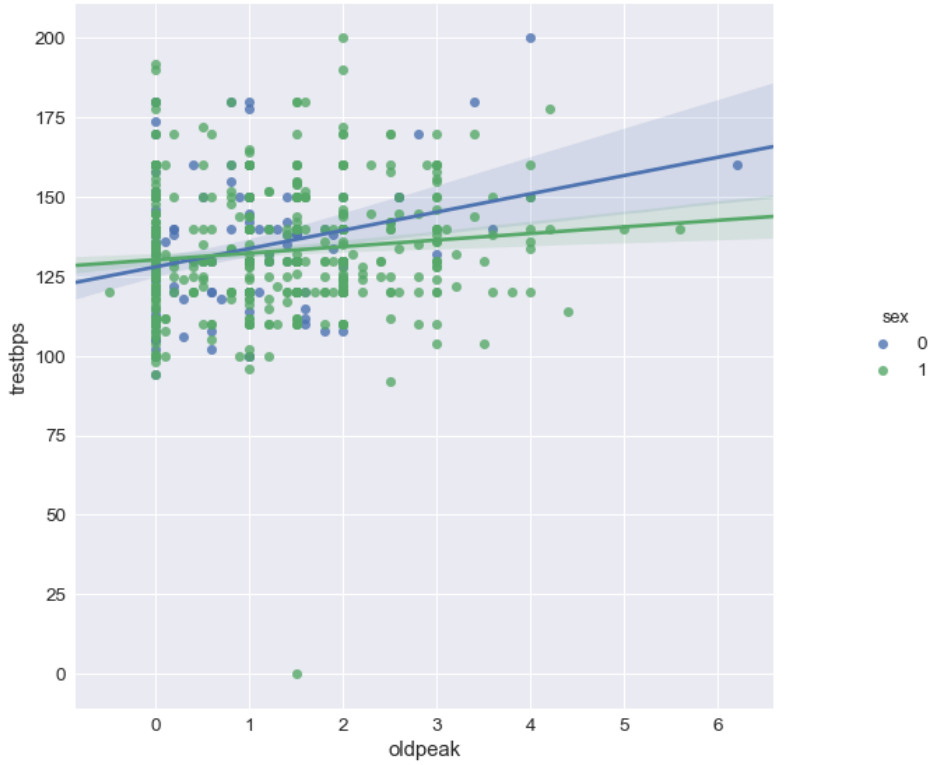
Figure 4. Age vs. max heart rate (sex: 0=female, 1=male)

In Figure 4, from these data there appears to be a slight downward trend in maximum heart rate ("thalach") with age. This is apparent for both sexes, possibly more for men.

Figure 5. Age vs. resting blood pressure (sex: 0=female, 1=male)

In Figure 5, from these data it appears that resting blood pressure ("trestbps") tends to increase with age for both sexes.

Figure 6. Exercise-induced ST depression vs. resting blood pressure (sex: 0=female, 1=male)



In Figure 6, it is difficult to discern a clear relationship between exercise-induced ST depression ("oldpeak") and resting blood pressure ("trestbps"). There is possibly a weak, positive trend, but this is difficult to assess from this plot with very large bands around trendlines.

Figure 7. Cholesterol vs. resting blood pressure (sex: 0=female, 1=male)

In Figure 7, with these data it appears that the relationship between cholesterol level ("chol") and resting heart rate ("trestbps") may be weak. This figure suggests a slightly stronger trend for women than for men for this relationship, though the confidence interval for women is quite broad.

Figure 8. Cholesterol vs. exercise-induced ST depression (sex: 0=female, 1=male)

In Figure 8, with these data it is difficult to discern a clear pattern for exercise-induced ST depression ("oldpeak") and cholesterol level ("chol").

Overall, most of the data available in this dataset do not show striking patterns between variables paired against each other one-on-one. Also, many variables are categorical, which are harder to plot and view patterns with, so they are excluded here. Among the continuous variables, though, there may be some patterns that show predictive value. More sophisticated analyses of variables such as cholesterol and sex may show stronger correlations with level of heart disease than what is viewable in the simpler exploratory analyses shown here. Some categorical variables in this study are derived from metrics that are continuous in nature but are here allocated into bins based on where the numbers fall in ranges.

It is hard to make much of these data at this point, but even though there is much noise in the data viewed here, the trendlines show potential for some relationships.

Exploratory data analysis approach

There is a total of 14 numerical variables, including the target variable, included in this dataset. Some are continuous variables, including age, "trestbps", "chol", "thalach", and "oldpeak". The rest are categorical variables, including sex (two categories as 0 and 1), "cp" (four categories as 1, 2, 3, and 4), "fbs" (two categories as 0 and 1), "restecg" (three categories as 0, 1, and 2), "exang" (two categories as 0

and 1), "slope" (three categories as 1, 2, and 3), "ca" (four categories as 0, 1, 2, and 3), "thal" (three categories as 3, 6, and 7), and the target variable "num" (five categories as 0, 1, 2, 3, and 4). Location is a separate variable to consider in the consolidated dataset, as the location associated with each subset is added in a new column in the consolidated set.

The continuous and categorical variables are explored here differently. Primarily, continuous variables are compared with each other using tests of Pearson's correlation coefficients, and categorical variables are primarily compared to each other using chi-square tests of independence.

A goal of inferential analyses in this study is to determine levels of correlation between variables and to judge the statistical significance of any of these relationships. The null hypothesis for each test is that there is no correlation between variables, with an alternate hypothesis for each being that there is a correlation. Figure 9 and Tables 3 and 4 will explore these relationships.

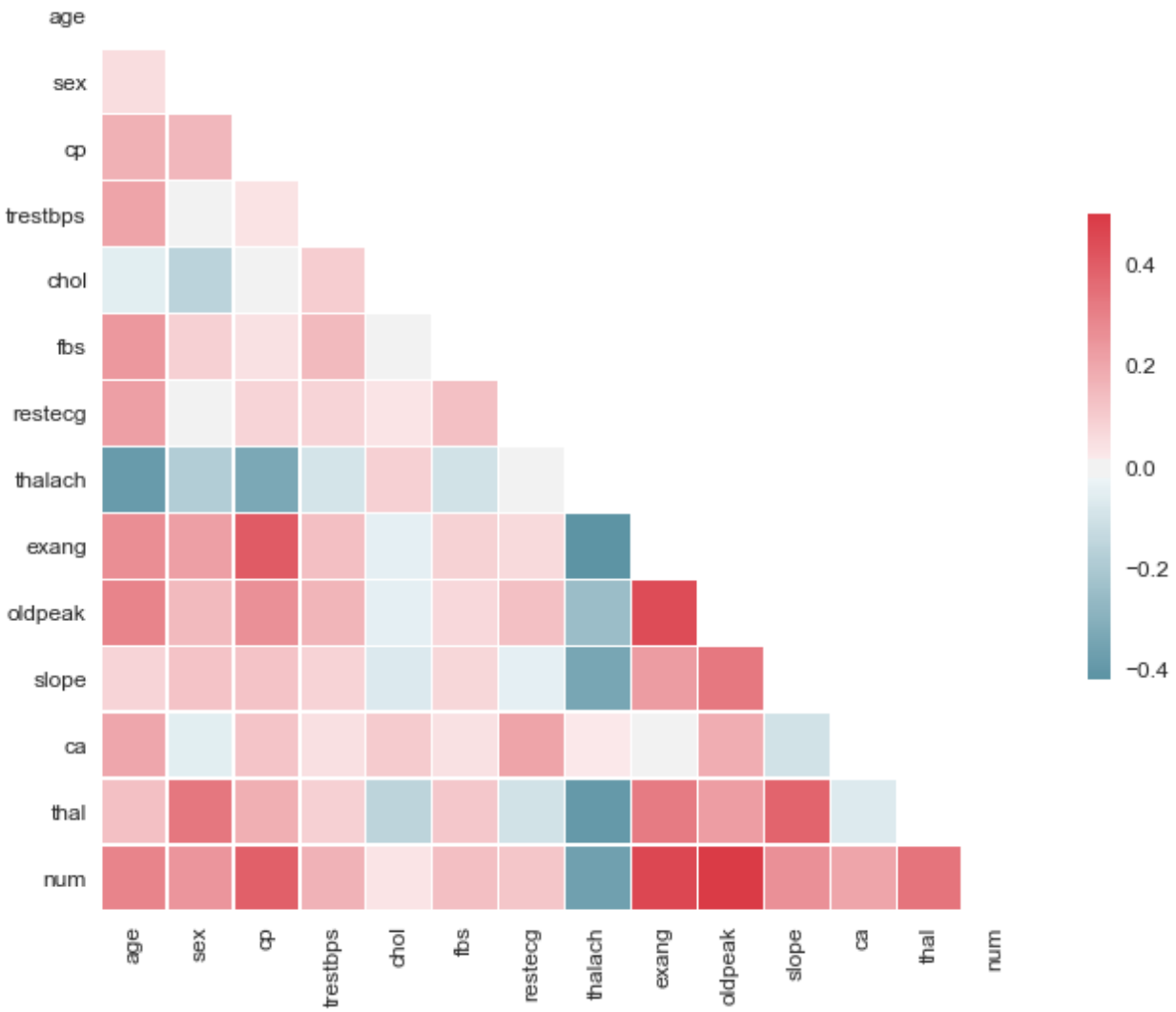
Figure 9 shows the strengths of relationships between variables paired against one another with darker red cells representing stronger, positive relationships. Darker blue cells represent stronger, negative relationships. The palest cells show the weakest relationships. The feature that shows the greatest amount of dark blue cells is "thalach" (maximum heart rate). This should be unsurprising, as higher values for many variables in this study are suggestive of less optimal health, and maximum heart rate is a measure of cardiac strength.

For relationships with the target variable, "num" or heart disease score, the strongest positive relationships appear to be with "oldpeak" (defined here as "ST depression induced by exercise and relative to rest") and "exang" (exercise-induced angina). Both of these are features that are either a sign or a symptom of possible cardiovascular distress, though it is not obvious why other features directly related to cardiac health in this dataset show less of a relationship to heart disease score. However, both are defined as exercise-related features, and perhaps that is a condition in which cardiovascular stress is especially prominent. "Oldpeak" and "exang" in this heatmap also show a relatively strong relationship to each other.

Cholesterol level is a variable that does not show strong relationships with other features in this study. Its strongest relationship, from the data available here, seems to be with sex, and it is a negative relationship. For sex this means that it is relatively higher with women than with men. Figues 3, 7, and 8 (above) indicate that cholesterol trends slightly higher for women than for men, especially with age. However, versus the target variable, here cholesterol seems to show the weakest relationship.

"Fbs" (fasting blood sugar) is another variable that shows relatively weak relationships with other variables in this study. Its strongest relationship is a possibly weak relationship with age. This is not surprising as insulin resistance may be more likely with age, though it does not seem to show a strong pattern with variables here that are directly tied to cardiac health.

Figure 9. Heatmap of correlations between variables



Some features in Figure 9 appear to show no relationships. It is possible, though, that interactions between variables may be important, such as the example in which there seems to be very little relationship between cholesterol and age, but when data are split by sex, as in Figure 3, opposite patterns appear. This heatmap is based on a correlation matrix, which is calculated most accurately for continuous variables, though many categorical variables in this study are bins for continuous data, so a heatmap may be used for a rough approximation of the strengths of relationships between variables. For calculated values, Table 3 shows Pearson correlation coefficients and significance values for relationships between continuous variables, and Table 4 shows chi-square statistics and significance values derived from chi-square tests of independence between categorical variables. For each of these tables, relationships with p-values that are set to <0.001 are shown in bold.

Table 3. Pearson correlation coefficients (r) and p-values (p) for relationships between continuous variables

Relationship	Pearson's correlation coefficient	p-value
age with trestbps	**r = 0.213**	**p = 0.0**
age with chol	r = -0.053	p = 0.13291
age with thalach	**r = -0.382**	**p = 0.0**
age with oldpeak	**r = 0.296**	**p = 0.0**
trestbps with chol	r = 0.1	p = 0.00459
trestbps with thalach	r = -0.091	p = 0.01022
trestbps with oldpeak	**r = 0.164**	**p = 0.0**
chol with thalach	r = 0.094	p = 0.00811
chol with oldpeak	r = -0.041	p = 0.2423
thalach with oldpeak	**r = -0.249**	**p = 0.0**

Pearson correlation coefficients in Table 3 show no especially strong relationships between continuous variables, but there appear to be some modest relationships, such as between age and "trestbps" (resting blood pressure), age with "thalach" (maximum heart rate), age with "oldpeak" (exercise-induced ST depression), "thalach" with "oldpeak", and a weaker relationship between "trestbps" and "oldpeak". Each of these correlations appears strongly statistically significant, with p-values below 0.001 for these.

Other statistically significant ($p < 0.05$) correlations appear for "trestbps" with "chol" (cholesterol), "trestbps" with "thalach", and "chol" with "thalach", but these are all with Pearson's r values of a magnitude at or < 0.1.

Two comparisons here have very weak relationships that lack statistical significance, being age with cholesterol and cholesterol with "oldpeak". The null hypothesis of no relationship for these pairs cannot be rejected.

Table 4. Chi-square statistics and p-values (p) for relationships between categorical variables

Relationship	Chi-square statistic	p-value
location with sex	**62.34**	**p = 0.0**
location with cp	**77.76**	**p = 0.0**
location with fbs	**64.51**	**p = 0.0**
location with restecg	**325.05**	**p = 0.0**
location with exang	**111.37**	**p = 0.0**
location with slope	**239.14**	**p = 0.0**
location with ca	**238.53**	**p = 0.0**
location with thal	**867.0**	**p = 0.0**
location with num	**77.33**	**p = 0.0**
sex with cp	**30.72**	**p = 0.0**
sex with fbs	6.38	p = 0.01153
sex with restecg	3.63	p = 0.16249
sex with exang	**37.7**	**p = 0.0**
sex with slope	12.9	p = 0.00158
sex with ca	6.25	p = 0.18145
sex with thal	**96.04**	**p = 0.0**
sex with num	**74.11**	**p = 0.0**

Relationship (cont'd.)	Chi-square statistic (cont'd.)	p-value (cont'd.)
cp with fbs	7.12	p = 0.0681
cp with restecg	21.66	p = 0.0014
cp with exang	**154.62**	**p = 0.0**
cp with slope	**25.86**	**p = 0.00024**
cp with ca	28.4	p = 0.00483
cp with thal	**99.84**	**p = 0.0**
cp with num	**218.6**	**p = 0.0**
fbs with restecg	**25.22**	**p = 0.0**
fbs with exang	5.72	p = 0.01679
fbs with slope	5.47	p = 0.06491
fbs with ca	4.22	p = 0.37673
fbs with thal	**30.59**	**p = 0.0**
fbs with num	**23.51**	**p = 0.0001**
restecg with exang	**24.0**	**p = 1e-05**
restecg with slope	**45.65**	**p = 0.0**
restecg with ca	**102.94**	**p = 0.0**
restecg with thal	**147.29**	**p = 0.0**
restecg with num	**28.41**	**p = 0.0004**
exang with slope	**43.45**	**p = 0.0**
exang with ca	2.37	p = 0.66865
exang with thal	**128.11**	**p = 0.0**
exang with num	**202.15**	**p = 0.0**
slope with ca	**31.22**	**p = 0.00013**
slope with thal	**228.24**	**p = 0.0**
slope with num	**61.85**	**p = 0.0**
ca with thal	**62.82**	**p = 0.0**
ca with num	**71.88**	**p = 0.0**
thal with num	**153.74**	**p = 0.0**

In Table 4 it appears that the variables location, "restecg" (resting electrocardiographic results) and perhaps "slope" (defined here as "slope of the peak exercise ST segment") and "thal" (relating to thalassemia) very frequently show high chi-square statistics with other variables - relationships that are highly statistically significant with $p < 0.001$.

It is possible that the nature of the location data somehow introduces artifacts into the dataset, though that is difficult to determine here.

The variable "fbs" (fasting blood sugar) shows weaker relationships with other variables. It does, however, bear a statistically significant relationship with the target variable "num" and thus may have predictive value. Importantly, here, all categorical variables tested with chi-square analysis show statistically significant relationships with "num", indicating these variables may all have use for prediction. However, many of them may be correlated with each other enough to complicate their individual predictive utility for heart disease.

Most of these factors are variables that directly relate to measurements of cardiovascular function, so this is unsurprising. "Fbs", which is fasting blood sugar, and sex are not as directly related to measurements of cardiovascular function, but they may be connected through other relationships to health. Age may prove to be important, though being a continuous variable, it is not paired with categorical variables for analysis at this point. Most variables included in correlation analyses with age,

however, do show strongly statistically significant relationships with it (“trestbps”, “thalach”, and “oldpeak”); the exception is cholesterol.

Predictive modeling

Upcoming analyses to be conducted with these data will incorporate machine learning techniques. It is not always clear at the outset of an analysis which machine learning approach will yield the best predictive model, but this will be ascertained next. Additionally, in order to make the model most useful, consideration of feature importance at the individual level will be used to apply meaning to the model, in terms of which factors for a given entry contribute most to the patient's risk.

The datasets used in this report are suited to development of a machine learning-based predictive model as each contains information that was collected in a similar fashion. Once a model is developed, and we have a better idea of the relative impact of each factor in influence heart disease risk, it may be appropriate to consider application to other datasets that may contain slightly different sets of information. Such data could come from information provided by the United States government, the World Health Organization, or state-level health data.

Machine learning model development

Before training the machine learning classifiers on the dataset, data may require more processing. Early on we already imputed missing data using medians or modes for features as appropriate from subpopulation sample estimates, but we have to address the fact that the "location" column contains non-numeric, categorical information that will not fit into machine learning algorithms well, so we create dummy variables out of location information with Boolean data type in one of three new columns, representing each location, and replacing the "location" column. We could go further and remove one of these columns (as a 0 in either of the two other columns implies assignment to the third location), but if we are examining feature importance later, we may as well keep each of these columns in, in case one of the locations is found to be important.

Since one of the populations (Cleveland) included in this dataset provides more complete information than the other two do, model testing will include both the entire dataset and parallel modeling with just the Cleveland component. Another level of testing within the dataset will involve the nature of the target variable, which is present as a multiclass variable with five classes. It may be difficult for even the best algorithm to predict within which level of heart disease severity an individual may possess, so multiclass and binary (absence or presence of heart disease) categorizations will face preliminary testing as well.

Next is the decision of which machine learning model(s) to use. There are many, and some have advantages over others in terms of computational speed and complexity, or in handling different data structures. The dataset used here is small enough (n=797) and with a manageable number of features (n=16 with dummy location variables) so as not to be unwieldy with most algorithms. Data are labeled and ultimately fit into a classification scheme for the target variable: primarily we want to know how likely it is for someone to develop heart disease, and, if possible, with which severity (which can be argued to be either quantitative or fitting into classes). This still leaves a lot of options, so several likely appropriate classifiers will be tested, with some hyperparameter tuning and cross-validation, to help

determine the best model(s) with which to move forward. Accuracy, precision, and recall will be examined at this point.

Finally feature importance is examined to determine which of the features of the model contribute the most to predictive power. Individual feature importance is enabled by a prediction tool in XGBoost that shows the contributions of each feature to any individual's target phenotype and visualized with SHAP values. This is where the power of a predictive model may be best realized, as having a model that is interpretable, particularly on the individual level, can provide specific and actionable information.

Choosing a model

There are many possible suitable supervised learning algorithms that may help explain patterns within the heart disease dataset and be useful for prediction of heart disease risk based on features. Choice of algorithm relies on how each model handles the data present in the dataset upon which the model is trained and applied to test data. Each model may also fare better or worse in handling either a multiclass or binary target variable. Hyperparameter tuning is a process that enables an algorithm to perform at its best, but it is computationally costly, so for most tests it will be considered after a preliminary analysis of model performance.

Results of preliminary analyses of model performance are shown in Tables 5 and 6 (full dataset and Cleveland, respectively) with accuracies shown for both training and testing data. If accuracy approaches 100% for training data and is considerably lower for testing data, then it can be presumed that the model is overfitted to the training data.

Visual comparisons of accuracy on testing data are shown in Figures 10 and 11 (full dataset and Cleveland, respectively). In Figures 10 and 11 it can also be readily seen how well each model performs for the multiclass target variable (“full”) versus a binary target analysis.

In these figures and Tables 5 and 6 model classifier names are abbreviated, but abbreviations mean the following: "RF" represents random forest, "tuned RF" represents a random forest test with hyperparameter tuning using GridSearchCV, "MNB" represents multinomial naive bayes, "GNB" represents Gaussian naive bayes, "LogReg" represents logistic regression, "SVM" represents support vector machine, "LSVM" represents linear support vector machine, "DT" represents decision tree, "GB" represents gradient boosting classifier, "SGD" represents SGD classifier, "ET" represents extra trees classifier, "XGB" represents XGBoost, and "NN" represents a neural network classifier.

Table 5. Preliminary analysis of model performance on the full dataset

Classifier	Classification	Train accuracy	Test accuracy
RF	full	0.993	0.565
RF(tuned)	full	0.969	0.565
MNB	full	0.391	0.445
GNB	full	0.471	0.525
LogReg	full	0.509	0.520
RF	binary	0.985	0.760
RF(tuned)	binary	0.836	0.780
MNB	binary	0.506	0.505
GNB	binary	0.809	0.810
LogReg	binary	0.826	0.810
SVM	full	0.995	0.495
LSVM	full	0.300	0.400
SVM	binary	0.998	0.605
LSVM	binary	0.506	0.505
DT	full	1.000	0.530
DT	binary	1.000	0.750
GB	full	0.788	0.545
GB	binary	0.879	0.785
SGD	full	0.290	0.505
SGD	binary	0.603	0.585
ET	full	1.000	0.550
ET	binary	1.000	0.800
XGB	full	0.864	0.555
XGB	binary	0.913	0.785
NN	full	0.419	0.450
NN	binary	0.675	0.645

In both Tables 5 and 6, what is apparent is that for most models accuracy on training data is considerably higher than for test data. The pattern also emerges that when classifying as a multiclass target variable, test accuracy tends to be much lower than when the target variable is treated as a binary variable. The tuned random forest tests show much lower accuracy on training data than non-tuned tests do, so they are less likely overfitted to training data, though test accuracy does not go up with this ensemble classifier. However, for later modeling, hyperparameter tuning will be employed to some degree (with consideration of computational constraints).

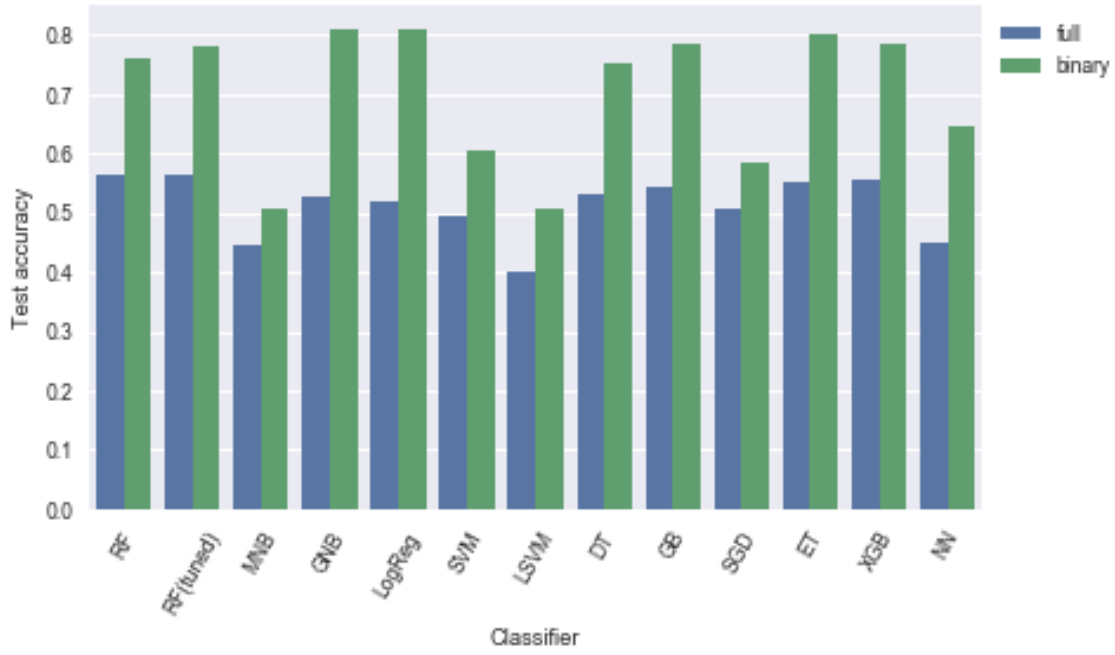
Table 6. Preliminary analysis of model performance on Cleveland data

Classifier	Classification	Train accuracy	Test accuracy
RF	full	0.997	0.566
Tuned RF	full	0.997	0.566
MNB	full	0.439	0.513
GNB	full	0.494	0.500
LogReg	full	0.551	0.605
RF	binary	0.991	0.803
Tuned RF	binary	0.859	0.868
MNB	binary	0.771	0.763
GNB	binary	0.841	0.868
LogReg	binary	0.846	0.882
SVM	full	1.000	0.539
LSVM	full	0.234	0.184
SVM	binary	1.000	0.553
LSVM	binary	0.819	0.842
DT	full	1.000	0.553
DT	binary	1.000	0.711
GB	full	0.959	0.579
GB	binary	0.965	0.829
SGD	full	0.216	0.553
SGD	binary	0.617	0.618
ET	full	1.000	0.592
ET	binary	1.000	0.803
XGB	full	0.984	0.579
XGB	binary	0.982	0.842
NN	full	0.200	0.039
NN	binary	0.718	0.658

In both Figures 10 and 11, it is easier to spot the profound differences in both accuracy with regard to the classification of the target variable as well as differences between models in their accuracy with these data. Accuracy here is based on scores for test samples rather than training data. Most classifiers perform fairly well with a binary target variable here, though linear SVM does poorly here even with the binary target. For linear SVM, however, there is a difference in performance between the full dataset and the Cleveland subset in these tests, with linear SVM performing comparably with other successful models for the binary target in this subset.

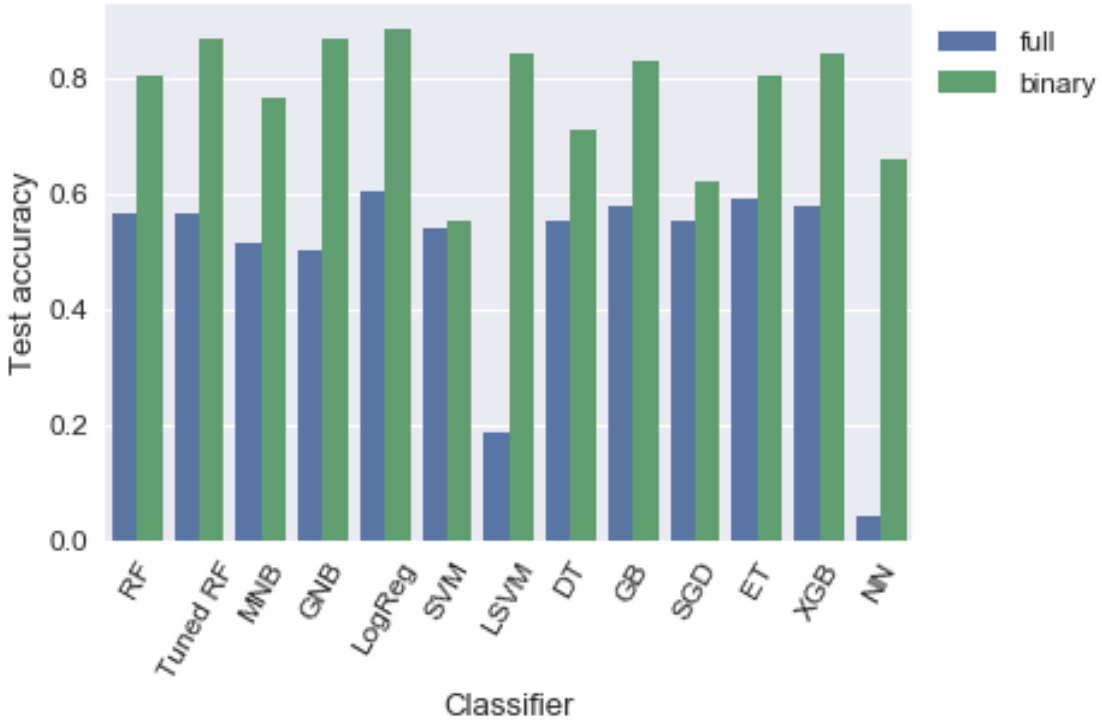
Many models performed similarly when used to predict the binary target variable, with test accuracies between 75-80% in this analysis. Most ensemble methods tended to perform better. Model performance among classifiers using SMOTE-treated data for multiclass target analyses was generally lower than with binary analyses. Since there is high similarity among classifiers, later in this report we will continue mostly with model interpretability tests using XGBoost with the binary target, as this is easy to work with and amenable to individual feature analysis.

Figure 10. Model performance on the full dataset



Most models tested as shown in Figures 10 and 11 performed at least slightly better with the Cleveland subset of data than with the full dataset that included relatively more imputed data. This is unsurprising for some models since the Cleveland subset could be said to contain relatively more accurate information. For other models it may have been sensible to not included imputed data at all because some models perform adequately with missing data. Overall, examining the binary target of presence or absence of heart disease and using extra trees, random forest, logistic regression, or a boosting algorithm seem to be supported as the most effective approaches with the heart disease data. Next is an examination of accuracy with these algorithm types with slight hyperparameter tuning and on the binary target.

Figure 11. Model performance on the Cleveland subset



Tables 7 and 8 feature more extensive comparisons of algorithms for modeling with the binary target, ranked by mean scores, and including some parameter tuning for each. Code for this follows upon code presented here and altered for Python 3 using code presented here, with further modifications introduced in this report in order to add XGBoost and logistic regression tests, to add more parameters to test, and to clarify that it is restricted to training data. Table 7 is a ranking for the entire dataset, and Table 8 is for Cleveland-only data.

Tables 7 and 8 contain a lot of detail, but the bottom line is that for the full dataset logistic regression, XGBoost, extra trees, and random forest models for the most part perform the best, at least on the training data, regardless of hyperparameters.

Table 7. Ranking of tuned models on full dataset (binary classification)

	estimator	min_score	mean_score	max_score	std_score	C	gamma	kernel	learning_rate	max_depth	min_child_weight	n_estimators
21	XGBClassifier	0.774869	0.810132	0.845771	0.0289469	NaN	NaN	NaN	NaN	NaN	10	NaN
31	LogisticRegression	0.789474	0.805243	0.822335	0.0134481	0.1	NaN	NaN	NaN	NaN	NaN	NaN
1	ExtraTreesClassifier	0.787565	0.803413	0.818653	0.0126989	NaN	NaN	NaN	NaN	NaN	NaN	32
32	LogisticRegression	0.772487	0.802329	0.838384	0.0272587	1	NaN	NaN	NaN	NaN	NaN	NaN
16	XGBClassifier	0.780488	0.802263	0.832487	0.0220533	NaN	NaN	NaN	NaN	NaN	NaN	32
4	RandomForestClassifier	0.753927	0.799072	0.845361	0.0373366	NaN	NaN	NaN	NaN	NaN	NaN	32
24	SVC	0.783069	0.797082	0.82	0.0163391	1	NaN	linear	NaN	NaN	NaN	NaN
34	LogisticRegression	0.78534	0.795922	0.812183	0.0116708	100	NaN	NaN	NaN	NaN	NaN	NaN
33	LogisticRegression	0.772487	0.795732	0.826531	0.0227007	10	NaN	NaN	NaN	NaN	NaN	NaN
5	RandomForestClassifier	0.761905	0.795567	0.835052	0.0301446	NaN	NaN	NaN	NaN	NaN	NaN	100
0	ExtraTreesClassifier	0.780749	0.794139	0.816327	0.0158009	NaN	NaN	NaN	NaN	NaN	NaN	16
15	XGBClassifier	0.764398	0.792576	0.829016	0.0270192	NaN	NaN	NaN	NaN	NaN	NaN	16
9	GradientBoostingClassifier	0.769231	0.790754	0.821782	0.022482	NaN	NaN	NaN	0.8	NaN	NaN	16
2	ExtraTreesClassifier	0.772487	0.790473	0.816754	0.019	NaN	NaN	NaN	NaN	NaN	NaN	100
25	SVC	0.776119	0.788956	0.81	0.0149999	10	NaN	linear	NaN	NaN	NaN	NaN
18	XGBClassifier	0.760976	0.786902	0.810256	0.0202008	NaN	NaN	NaN	NaN	4	NaN	NaN
19	XGBClassifier	0.76	0.785983	0.80203	0.0185414	NaN	NaN	NaN	NaN	20	NaN	NaN
12	GradientBoostingClassifier	0.756219	0.783747	0.815534	0.0244019	NaN	NaN	NaN	1	NaN	NaN	16
6	AdaBoostClassifier	0.768421	0.782579	0.795812	0.0112013	NaN	NaN	NaN	NaN	NaN	NaN	16
7	AdaBoostClassifier	0.767568	0.782459	0.80203	0.0144535	NaN	NaN	NaN	NaN	NaN	NaN	32
30	LogisticRegression	0.770833	0.782413	0.795918	0.010331	0.01	NaN	NaN	NaN	NaN	NaN	NaN
17	XGBClassifier	0.770833	0.780363	0.79798	0.0124706	NaN	NaN	NaN	NaN	NaN	NaN	100
20	XGBClassifier	0.770833	0.780363	0.79798	0.0124706	NaN	NaN	NaN	NaN	NaN	1	NaN
22	XGBClassifier	0.758621	0.777157	0.79798	0.0161495	NaN	0.01	NaN	NaN	NaN	NaN	NaN
10	GradientBoostingClassifier	0.744898	0.775079	0.829016	0.0382292	NaN	NaN	NaN	0.8	NaN	NaN	32
23	XGBClassifier	0.754902	0.773768	0.79798	0.0179882	NaN	0.1	NaN	NaN	NaN	NaN	NaN
8	AdaBoostClassifier	0.75	0.771126	0.80402	0.0235712	NaN	NaN	NaN	NaN	NaN	NaN	100
13	GradientBoostingClassifier	0.747368	0.768121	0.79602	0.0204945	NaN	NaN	NaN	1	NaN	NaN	32
3	RandomForestClassifier	0.743169	0.76496	0.793814	0.0212707	NaN	NaN	NaN	NaN	NaN	NaN	16
11	GradientBoostingClassifier	0.730964	0.764551	0.81	0.0333386	NaN	NaN	NaN	0.8	NaN	NaN	100
14	GradientBoostingClassifier	0.728205	0.764357	0.80198	0.0301365	NaN	NaN	NaN	1	NaN	NaN	100

Interestingly, with just the Cleveland subset (Table 8) considered in this analysis of hyperparameter optimization, the SVC classifier comes out on top for two different “C” parameter values, but below this the top classifiers continue to be logistic regression, random forest, and XGBoost classifiers.

Table 8. Ranking of tuned models on Cleveland subset (binary classification)

	estimator	min_score	mean_score	max_score	std_score	C	gamma	kernel	learning_rate	max_depth	min_child_weight	n_estimators
24	SVC	0.739726	0.806899	0.875	0.0552293	1	NaN	linear	NaN	NaN	NaN	NaN
25	SVC	0.722222	0.797907	0.888889	0.0688959	10	NaN	linear	NaN	NaN	NaN	NaN
31	LogisticRegression	0.735294	0.795471	0.875	0.0586532	0.1	NaN	NaN	NaN	NaN	NaN	NaN
34	LogisticRegression	0.704225	0.791114	0.875	0.0697508	100	NaN	NaN	NaN	NaN	NaN	NaN
5	RandomForestClassifier	0.742857	0.790467	0.870968	0.057239	NaN	NaN	NaN	NaN	NaN	NaN	100
3	RandomForestClassifier	0.724638	0.790249	0.852459	0.0522382	NaN	NaN	NaN	NaN	NaN	NaN	16
22	XGBClassifier	0.736842	0.788634	0.852941	0.0482163	NaN	0.01	NaN	NaN	NaN	NaN	NaN
23	XGBClassifier	0.736842	0.786338	0.852941	0.0489165	NaN	0.1	NaN	NaN	NaN	NaN	NaN
32	LogisticRegression	0.686567	0.785228	0.875	0.0771838	1	NaN	NaN	NaN	NaN	NaN	NaN
16	XGBClassifier	0.764706	0.784995	0.8125	0.0201681	NaN	NaN	NaN	NaN	NaN	NaN	32
20	XGBClassifier	0.736842	0.782453	0.852941	0.0505563	NaN	NaN	NaN	NaN	NaN	1	NaN
17	XGBClassifier	0.736842	0.782453	0.852941	0.0505563	NaN	NaN	NaN	NaN	NaN	NaN	100
4	RandomForestClassifier	0.753623	0.780842	0.819672	0.0281867	NaN	NaN	NaN	NaN	NaN	NaN	32
21	XGBClassifier	0.746269	0.778762	0.84375	0.0459531	NaN	NaN	NaN	NaN	NaN	10	NaN
18	XGBClassifier	0.72973	0.776401	0.811594	0.0343924	NaN	NaN	NaN	NaN	4	NaN	NaN
6	AdaBoostClassifier	0.707692	0.772633	0.825397	0.048818	NaN	NaN	NaN	NaN	NaN	NaN	16
2	ExtraTreesClassifier	0.695652	0.772307	0.875	0.0754979	NaN	NaN	NaN	NaN	NaN	NaN	100
33	LogisticRegression	0.647059	0.772059	0.875	0.0943548	10	NaN	NaN	NaN	NaN	NaN	NaN
19	XGBClassifier	0.72	0.77092	0.823529	0.0422826	NaN	NaN	NaN	NaN	20	NaN	NaN
15	XGBClassifier	0.72973	0.769583	0.825397	0.0406537	NaN	NaN	NaN	NaN	NaN	NaN	16
12	GradientBoostingClassifier	0.710526	0.768937	0.83871	0.0529437	NaN	NaN	NaN	1	NaN	NaN	16
11	GradientBoostingClassifier	0.717949	0.76015	0.78125	0.0298405	NaN	NaN	NaN	0.8	NaN	NaN	100
10	GradientBoostingClassifier	0.727273	0.755366	0.78125	0.0220914	NaN	NaN	NaN	0.8	NaN	NaN	32
13	GradientBoostingClassifier	0.72	0.751051	0.786885	0.0275144	NaN	NaN	NaN	1	NaN	NaN	32
1	ExtraTreesClassifier	0.685714	0.74359	0.830769	0.062739	NaN	NaN	NaN	NaN	NaN	NaN	32
30	LogisticRegression	0.676923	0.743201	0.78125	0.0470365	0.01	NaN	NaN	NaN	NaN	NaN	NaN
14	GradientBoostingClassifier	0.693333	0.73583	0.786885	0.0386688	NaN	NaN	NaN	1	NaN	NaN	100
8	AdaBoostClassifier	0.7	0.734234	0.8	0.0465165	NaN	NaN	NaN	NaN	NaN	NaN	100
7	AdaBoostClassifier	0.724138	0.732692	0.746667	0.0099638	NaN	NaN	NaN	NaN	NaN	NaN	32
0	ExtraTreesClassifier	0.636364	0.732146	0.806452	0.0710794	NaN	NaN	NaN	NaN	NaN	NaN	16
9	GradientBoostingClassifier	0.702703	0.72762	0.75	0.0193923	NaN	NaN	NaN	0.8	NaN	NaN	16
29	SVC	0.580645	0.654727	0.69697	0.0525559	10	0.0001	rbf	NaN	NaN	NaN	NaN

Figure 12 shows receiver operating characteristic (ROC) curves for five leading models (hyperparameter-tuned random forest (top left), non-tuned XGBoost (top middle), non-tuned logistic regression(top right), extra trees (bottom left), and Gaussian naïve Bayes (bottom right)), predicting a binary target with the full dataset. As expected from previous model tests (Figure 10, Tables 6 and 7), all of these models performed roughly similarly for this metric. A similar trend can be seen in Figure 13 with precision-recall curves of each of three of these models (random forest (left), XGBoost (middle), and logistic regression (right)).

Figure 12. ROC curves for the full dataset with a binary target

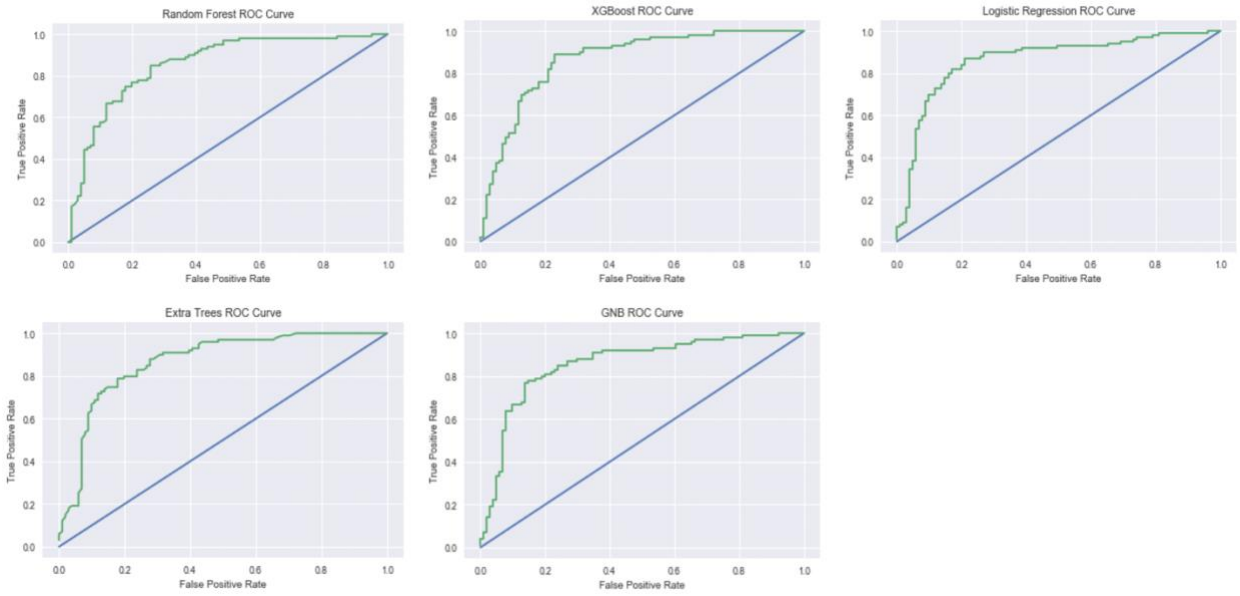
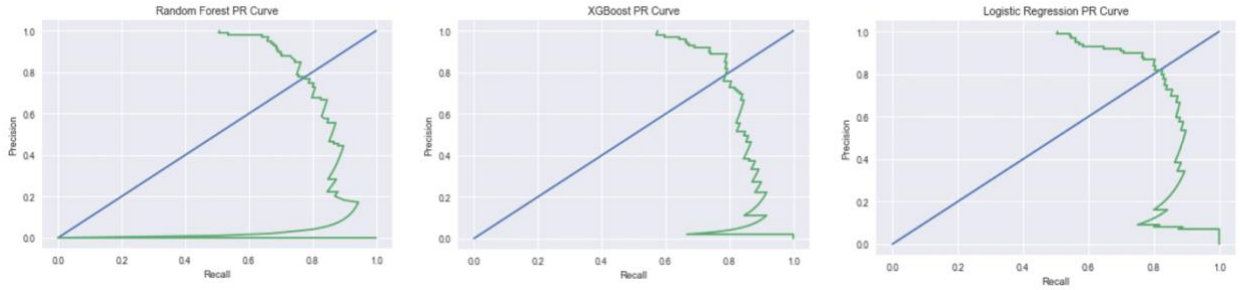


Figure 13. Precision-recall curves for the full dataset with a binary target



The previous results describe overall accuracy with the predictive models examined, but when using a predictive model it is often desired to know how each feature influences the target prediction and in what manner. Because it is a reasonably successful algorithm with this dataset and because certain methods work easily in scikit-learn with XGBoost, further explorations here will involve XGBoost results.

Figures 14-17 show feature importance metrics for various analyses with XGBoost. These models were trained on 2/3 of the dataset and default parameters, with 1/3 of the data set aside as the test set. With a multiclass target and using the full dataset (Figure 14), cholesterol, maximum heart rate ("thalach"), and age were the most prominent features influencing the model. These three features dominate, but with different order for the Cleveland subset, in each test (Figures 15-17). Following that, other features maintain roughly similar relative positions for each, though rankings vary slightly between tests. This does indicate that with XGBoost, the features are mostly telling the same stories for each analyses, even if accuracy differs between analyses. Not shown here, but a few tests involving removing some less influential features did not improve accuracy, indicating that even if some features seem unimportant, they may each offer slight value to the predictions.

Figure 14. Feature importance using XGBoost on the full dataset with a multiclass target

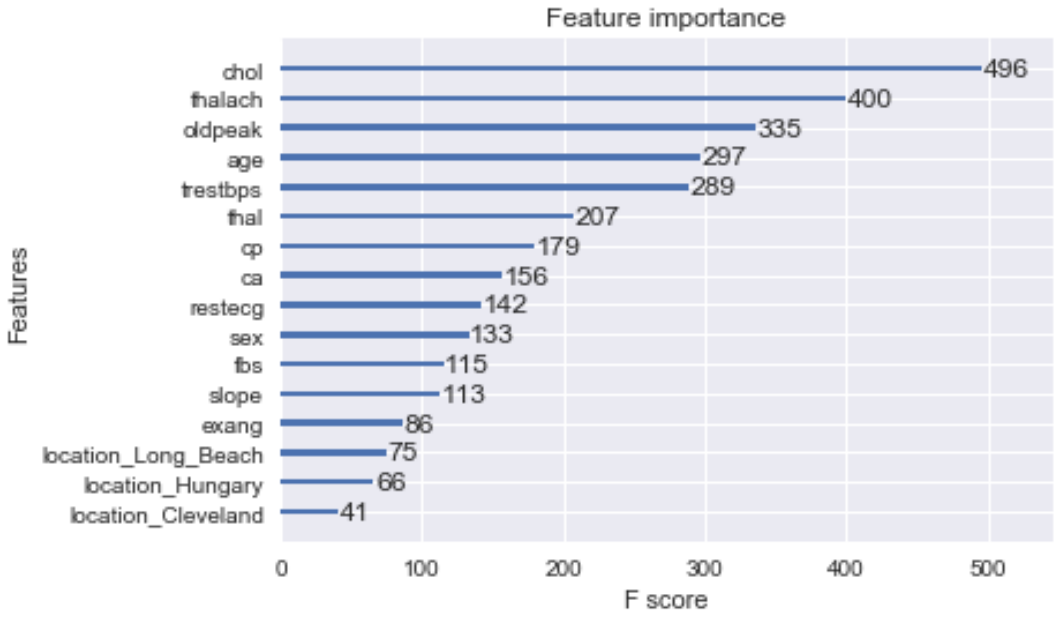


Figure 15. Feature importance using XGBoost on the full dataset with a binary target

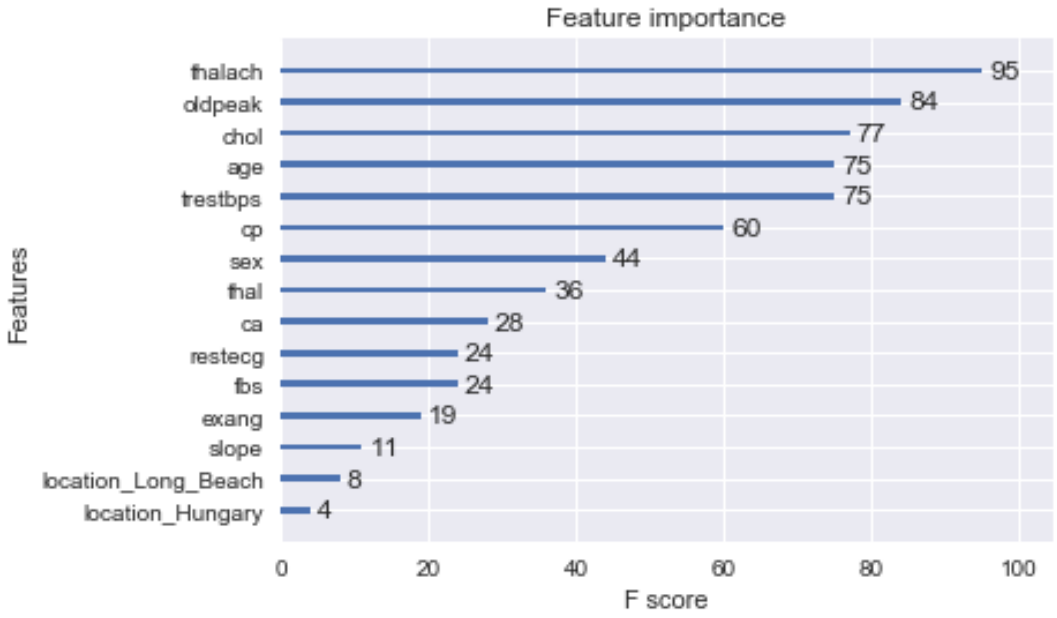


Figure 16. Feature importance using XGBoost on the Cleveland subset with a multiclass target

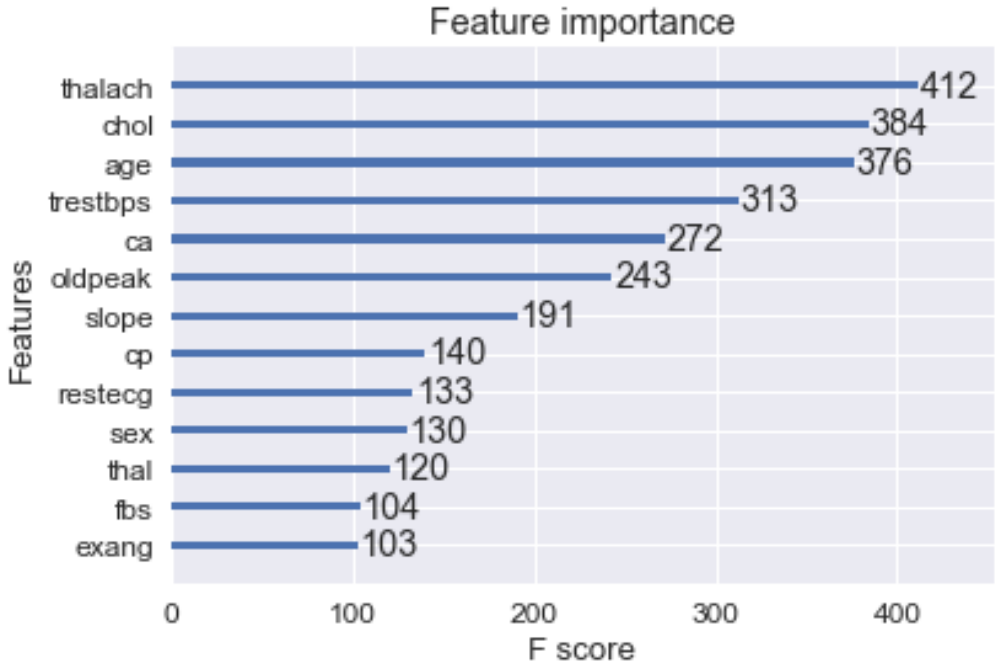
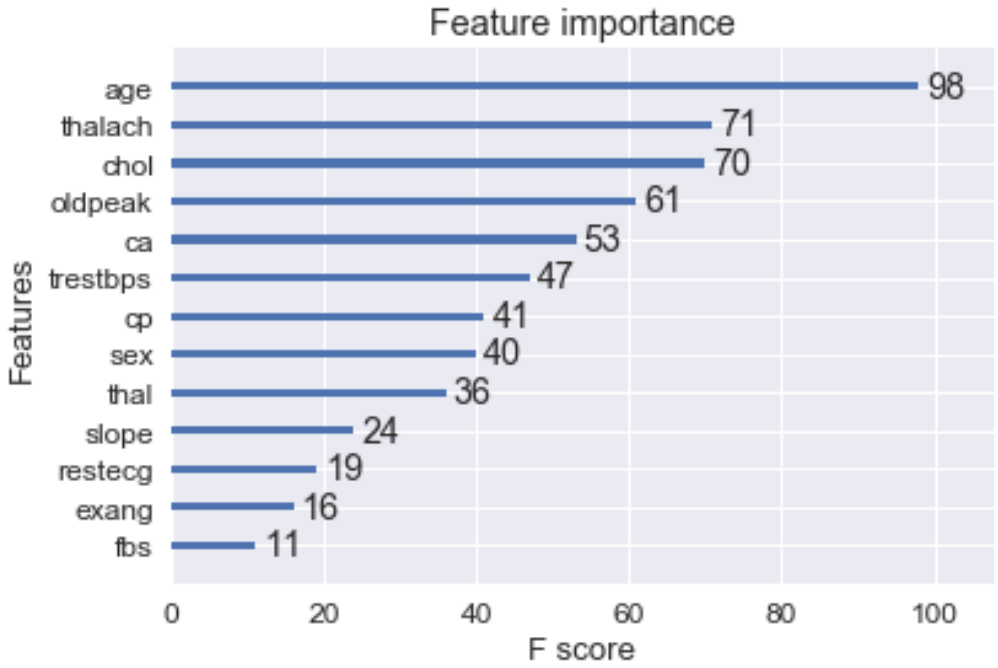


Figure 17. Feature importance using XGBoost on the Cleveland subset with a binary target



While feature importance is useful for indicating across a population what variables contribute the most to the model, this alone does not tell us how any individual's features contributed to his/her own predictive risk – information that may be quite useful for making any adjustments to one's risk.

XGBoost contains within it an option to display a matrix of SHAP values for an individual sample's features, plus a bias value, which all add up to an overall prediction score for the individual. Table 9 shows an example using sample 0 from the full dataset with XGBoost trained using 80% of the data with 20% reserved for testing. Each feature is listed with the contributive value associated with that feature, followed by a bias term and finally the sum. In this example, a negative sum indicates an individual considered more likely to not have heart disease, while a positive value would indicate the individual is predicted to have heart disease. Features that show the greatest magnitude for this individual are "cp", with a fairly strong negative predictive value, while this individual's "oldpeak" value shows added risk of heart disease. Overall, the features add up to a prediction of not having heart disease.

Table 9. Individual prediction score and feature contributions using XGBoost on full dataset

	prediction contributions
age	-0.169331
sex	0.186544
cp	-0.709835
trestbps	-0.074752
chol	-0.136371
fbs	0.246492
restecg	0.002791
thalach	-0.258700
exang	-0.355108
oldpeak	-0.282242
slope	-0.300144
ca	-0.068493
thal	0.588108
Cleveland	0.000000
Hungary	-0.002698
Long_Beach	-0.074609
bias	-0.023274
SUM	-1.431624

The method used to generate Table 9 is convenient for allowing a glimpse at specific values for each feature contribution for an individual, but the SHAP python package itself can also provide more visual information about both global and individual patterns within a dataset using F1 scores. Figure 18 shows the SHAP value and contributions for the same sample (sample 0) as in Table 9, using an XGBoost model

trained on the full dataset (trained on 80% of the data and tested on 20%). It is clear that the two methods scale or normalize SHAP values differently, but the overall contributions seem to show relatively similar impacts.

Figure 18. Individual prediction score for heart disease presence using SHAP with XGBoost

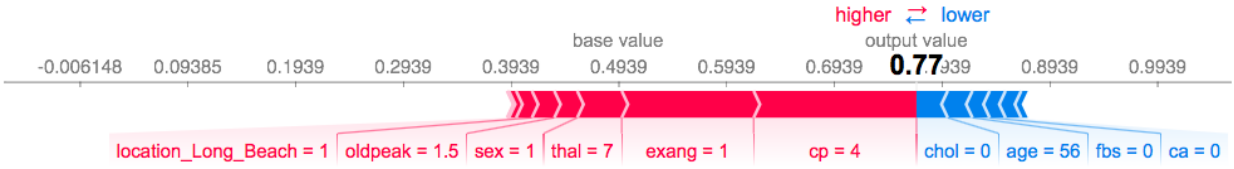


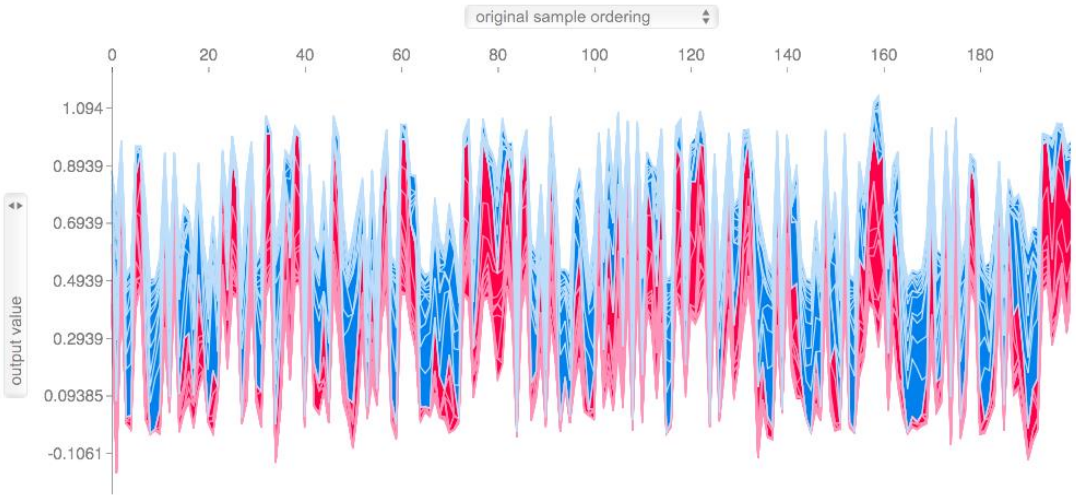
Figure 18 shows this individual's predicted risk of heart disease to be low based on features shown in this plot (and to a lesser degree other features that have a weak enough influence to not be named in the plot). The low "cp", "exang", and "ca" values, along with the cholesterol level, age, and "slope" value are the features most strongly contributing to a lower risk for this individual. The "thalach", "fbs", "oldpeak" and sex values most strongly contribute to some level of risk for this individual, but on balance this individual's features present a low risk, according to the model.

According to the model used here, the individual whose data are shown in Figure 18 is at low risk of a positive heart disease diagnosis due, in part, to his low degree of chest pain, lack of exercise-induced angina, and low coronary angiography score. Potentially actionable features this individual may need to pay attention to are his cardiac metric ST depression with exercise relative to rest and high fasting blood sugar.

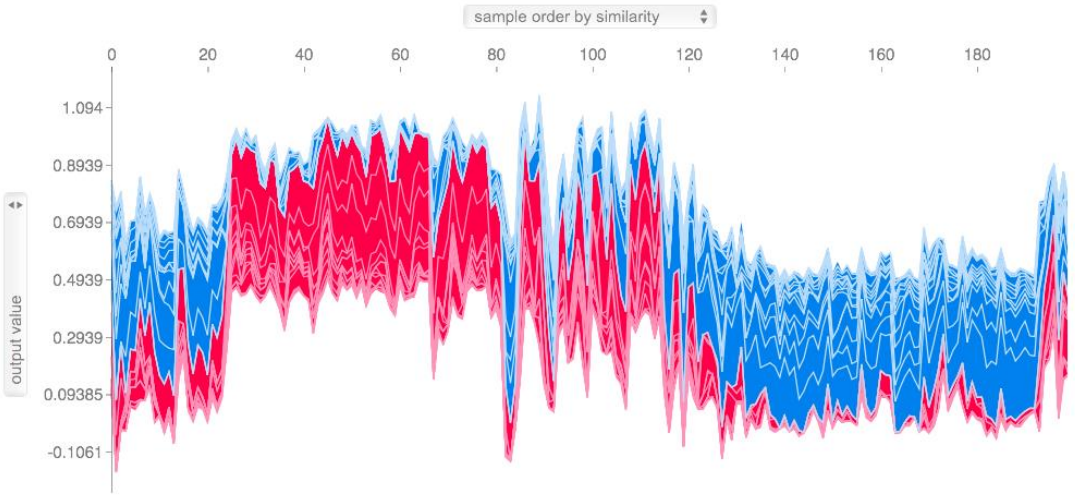
Figure 19 shows the entire dataset's SHAP values visualized and ordered by sample order (A) or in terms of overall similarity (B), with (C) showing a portion of the latter that shows interactivity in iPython with this figure. It is possible to mouse over the graph and get a quick look at each individual's main contributions to SHAP value and the direction of impact the most prominent features have for the individual, with (C) highlighting features for sample 111. For this individual, a chest pain ("cp") level of 3 lowers the risk of heart disease, while the features shown in magenta increase the individual's predictive risk of heart disease.

Figure 19. Global display of SHAP values with XGBoost (full dataset, binary target)

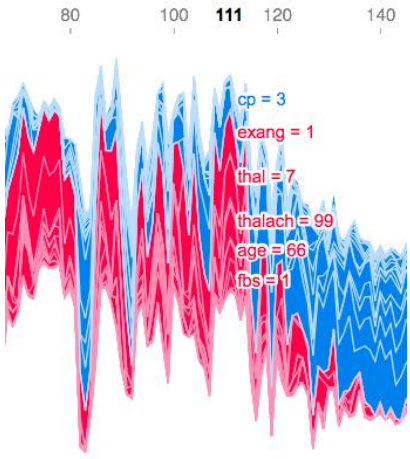
A.



B.



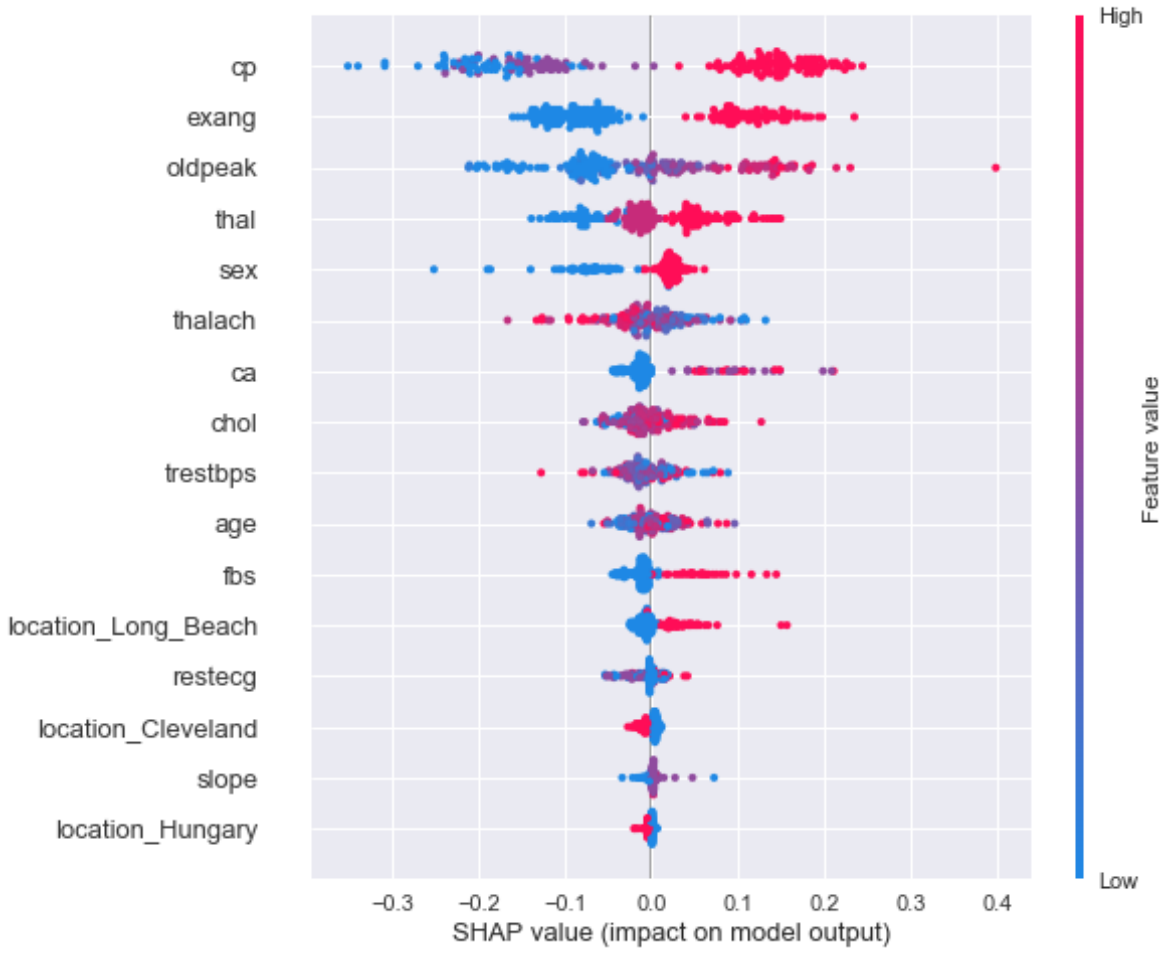
C.



Each individual's values can also be located on a SHAP summary plot (Figure 20) that shows SHAP values spreading horizontally (similar to the manner of a feature importance graph) and density of samples for each position grouped vertically for each feature (similar to a violin plot) (Lundberg, Erion & Lee, 2018). Coloration shows feature value, with magenta indicating high value and blue indicating low value. Long tails in Figure 20 indicate values that lower risk of heart disease (far left) and values that raise the risk of heart disease (far right).

In Figure 20 we see that chest pain ("cp") is ranked most highly in feature importance, followed by exercise-induced angina ("exang") and ST depression with exercise versus rest ("oldpeak"). This differs from the feature importance graph, also generated using XGBoost, from Figure 15. One advantage to viewing a SHAP summary plot is that it is easier to judge each feature; maximum heart rate and cholesterol rank among the top three features in Figure 15, but they are less important in Figure 20. Most individuals are clustered in the middle of the pack for cholesterol, and high values can affect results for some individuals. These patterns could be weighted differently by software packages, leading to disparate rankings.

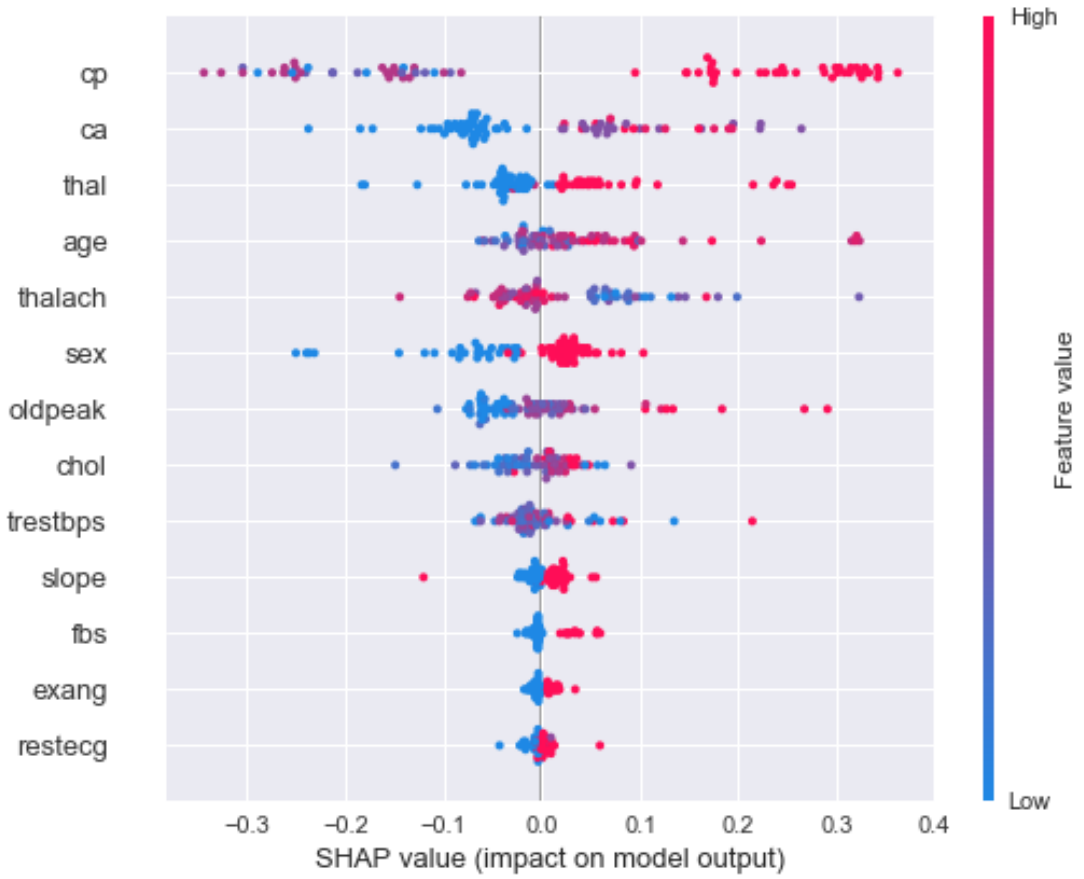
Figure 20. Summary plot of SHAP values with XGBoost and the full dataset (binary target)



Another observation from Figure 20 is that the top features listed resemble the most prominent features positively associated with the target variable in the correlation heatmap from Figure 9.

Figure 21 shows another SHAP summary plot with a binary target generated from an XGBoost model with the Cleveland subset alone. This subset contains relatively fewer imputed values overall than in the full dataset used for this report, and it is also specific to one clinic at one location and contains a much smaller number of samples. The overall feature importance ranking in this SHAP summary plot shows a number of features in slightly different order than in Figure 20, with variables such as “exang”, “ca”, and others populating very different positions.

Figure 21. Summary plot of SHAP values with XGBoost and the Cleveland subset (binary target)



Conclusions based on SHAP values

While the plots in Figure 19 provide a map of how each sample's feature values contribute to their overall predicted target values, the visualizations can also be adjusted to show how each feature's values trended toward a target value. For instance, by looking at effects of "cp" specifically, it appears that a "cp" value of 4 is associated with presence of heart disease, whereas "cp" values from 1-3 are associated with absence of a heart disease diagnosis. The most severe category of chest pain is what trends with heart disease, according to this model. This relationship seems to hold true whether examining the entire dataset or the subset from Cleveland with relatively fewer imputed values.

In this model on the full dataset, for exercise-induced angina, a value of 0 corresponds to absence of heart disease in this model, while 1 corresponds to presence of heart disease. A continuous variable, the values of "oldpeak" (ST depression induced by exercise and relative to rest) trend toward presence of heart disease the farther from 0 they are. Variable "thal" (thalassemia defect) correlates with predicted presence of heart disease if the value is 7, or absence of heart disease if the value is 3 (and usually for 6).

Also, in this model on the full dataset, coronary angiography score, or "ca", shows that any value above 0 more substantially raises heart disease risk.

Age is a continuous variable with a broad range. Using this model with the full dataset, overall the higher the age the more likely a prediction of heart disease, though for some individuals age was only weakly influential on this risk, while for others it was associated with substantial risk. This is unsurprising as age itself is not necessarily a causal mechanism of cardiac vulnerabilities, but age is associated with multiple factors that themselves can weaken cardiac tissue. For some individuals of advanced age, markers of good cardiac health can outweigh the influence of age, and vice versa.

Cholesterol is another continuous variable inhabiting a broad range, and throughout most of the population it appears to exert little effect on prediction of heart disease. However, it exerts a particularly substantial influence on this prediction for those individuals with very high cholesterol levels. This scenario could possibly account for the fact that with some feature importance estimations from different models (most not shown here) cholesterol shows greater prominence, while for other models it appears unimportant as most individuals do not inhabit the extremes for this feature.

Maximum heart rate ("thalach") is another continuous variable that has greater or lesser importance for prediction of heart disease depending on the individual and other features. However, while strength of this feature in the predictive model can vary for individuals, the relationship is not complex; higher maximum heart rate is associated with being less likely to have a heart disease diagnosis, and vice versa.

A different cardiovascular feature, resting blood pressure ("trestbps"), however, bears a more complex relationship to prediction of heart disease in this model on the full dataset as well as with the Cleveland subset. For most individuals this feature shows a modest effect on their risk, but for many it is a strong effect, and the relationship to value is unclear. Some high values for this feature appear to correspond to lower risk, while lower values correspond to higher risk, and vice versa. This variable appears to show a possible weakness in this model's predictive utility, or at least something to examine further for its relationship to other features. A similar SHAP analysis (not shown) using a logistic regression model on just the Cleveland subset of data, "trestbps" similarly showed odd characteristics with prediction.

Sex shows a more pronounced effect on lowering prediction of heart disease when the individual is female, using this model on the full dataset, though this effect varies in relative strength for each individual. The gender effect is relatively less influential overall for males.

Fasting blood sugar ("fbs") >120 mg/dl appears associated with prediction of heart disease. A fasting blood sugar below this threshold, contrarily, pushes one's risk down in most cases, though this influence is very mild compared with that for higher blood sugar.

A cardiac measurement, slope of the peak exercise ST segment ("slope"), shows variable influence on this model's prediction for different individuals in the full dataset and is hard to interpret in this model.

For resting electrocardiographic results ("restecg") most individuals with a value of 2 show higher risk of prediction of heart disease. This variable is categorical with only 3 values possible: 0, 1, or 2, with 2 reflecting left ventricular hypertrophy.

Location-wise, an individual sampled from the Long Beach population is considerably more likely to show presence of heart disease, though not being sampled from Long Beach only slightly lowers one's association with predicted heart disease. Being sampled in Hungary has the opposite effect. Apart from a few individuals, being sampled in Cleveland appears to have very little effect. Whether these results truly relates to geography or reflects bias in sampling by locality is unclear. However, more imputation occurred for samples from Long Beach and Hungarian subsets, and the values used for imputation could bias results.

Perspective

While the heart disease dataset cannot account for all characteristics of an individual's unique medical history and risk factors, it does lend itself to reasonably accurate predictive models. A reason for testing the full dataset in this series of analyses is that in addition to it presenting more observations than the Cleveland subset does, it also is a considerably less complete dataset overall than the Cleveland subset is. In the real world, most people will not be subjected to the more invasive technique of angiography, for instance, during an initial assessment of heart disease risk. Many of the other features considered in the dataset likewise are not likely be fully explored initially and for all patients. One risk assessment approach based on insights from this study could be to initially consider a patient's predicted heart disease risk in the context of the binary classification of presence or absence of heart disease, using the full dataset with XGBoost (or another ensemble method) and SHAP analysis. As an aid along with the patient's care provider's judgment, such an analysis, if it signals heart disease risk, could indicate the individual should be assessed further for severity of heart disease. In this instance there could be a practical use for a model trained on the Cleveland data subset in the context of binary classification.

References

Batista, D.S. “Hyperparameter optimization across multiple models in scikit-learn.” Web. Accessed: 6/11/18. http://www.davidsbatista.net/blog/2018/02/23/model_optimization/ .

Centers for Disease Control and Prevention. National Environmental Public HealthTracking Network. Web. Accessed: 3/1/18. www.cdc.gov/ephtracking .

Katsaroumpas, P. “Hyperparameter Grid Search across multiple models in scikit-learn.” Web. Accessed: 6/11/18. http://www.codiply.com/blog/hyperparameter-grid-search-across-multiple-models-in-scikit-learn/ .

Lundberg, S.M., & Lee, S.I. 2017. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.* Pages 1-10.

Lundberg, S.M., Erion, G.G., & Lee, S.I. 2018. Consistent individualized feature attribution for tree ensembles. eprint arXiv:1802.03888. Pages 1-9.

University of California, Irvine Machine Learning Repository. Heart Disease Data Set. Web. Accessed: 3/11/18. archive.ics.uci.edu/ml/datasets/Heart+Disease. Principal investigator information for component datasets used in this study: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.; V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.