What are the Greatest Risk Factors for Heart Disease?

Cardiovascular diseases are the leading cause of death worldwide, accounting for about 18 million, or about 1 in 3 deaths yearly. Heart diseases are a subcategory of cardiovascular disease, and include things like coronary heart disease and heart failure. In our analysis of a heart disease data set we found that a portion of the electrocardiogram (ECG) called the ST segment was by far the most important feature for predicting heart disease. Specifically, an upward facing ST slope, combined with its position below the baseline, were the most important features for predicting heart disease. What follows is a report on our data, the changes we made to the data, our modeling process, and our insights.

The Data

This is a composite data set of heart disease data that comes from the United States, Hungary, and Switzerland. It consists of 918 patients with information across twelve features. Our **dependent variable** is **heart disease** (HD). This is the variable we'd like to predict. Some of the independent variables of this data set are a little less straightforward than others. Age and sex, for example, are self explanatory. Here are some of the more interesting or complicated independent variables:

- ChestPainType: Typical Angina, atypical angina, non-anginal pain, and asymptomatic. Angina is a kind of chest pain, and is often described as a squeezing or pressure around the heart area. Angina can be considered atypical if it occurs less predictably and in varying degrees of pain. We take non-anginal pain to be any other kind of bodily pain that isn't obviously not related to the heart. For example, the pain of a stubbed toe does not count as non-anginal pain, but perhaps an aching thigh might. We take asymptomatic to mean the complete absence of angina or any pain that might be related to it.

- Cholesterol: Serum cholesterol (mg/dl). Cholesterol is a waxy fatty substance present in our bloodstream. It is absolutely essential for survival, being involved in things like hormone regulation, cell structure, and vitamin synthesis. It is important to note here that serum, or blood, cholesterol is not the same thing as dietary cholesterol. If serum cholesterol is found to be a risk factor for HD, it doesn't necessarily say anything about dietary cholesterol. A sample of serum cholesterol is measured in milligrams of cholesterol per deciliter of blood.

- FastingBS: Fasting blood sugar. Blood sugar is the measure of glucose (in milligrams) concentrated in (1 deciliter) of blood. A fasting blood sugar level of less than 100 is normal. 100-120 is thought of as pre-diabetic levels. A person with a fasting blood sugar over 120 is diabetic.

- RestingECG: Resting electrocardiogram results. Divided into normal, ST (T wave inversions and/or ST elevation or depression of > 0.05 mV, see figure 1), and LVH

(showing probable or definite left ventricular hypertrophy by Estes' criteria). Cardiograms measure electrical impulse vs. time.

- Oldpeak: Oldpeak = ST. This feature takes non zero values when the resting ECG value (previous bullet point) is ST. The ST segment is a portion of an ECG that occurs after the QRS complex, and before the T wave. Normally, with a healthy heart, this segment would be mostly flat on the baseline. It is when it is measured below the baseline that HD may be a concern. Since the oldpeak metric here is tracked in positive millimeters below the baseline, negative measurements must mean millimeters above the baseline. See figure 1.

- ST_Slope: The slope of the peak exercise ST segment. Up means upsloping, flat is flat (laying flat on the baseline), and down has a negative slope. Again, this is the same portion of the ECG as referred to in oldpeak. This feature is measuring the slope of the ST segment, rather than its position on or below the baseline. A flat or up ST slope would be normal, as the ECG signal must increase positively to get back to the baseline from the end of the QRS segment, which is typically quite negative. See figure 1.

Changes to the Data

While cleaning our data, we found that 171 entries had a value of 0 for their cholesterol levels. This is unlikely, so we were faced with a serious decision: what to do with the missing cholesterol values, which, we found, all came from the portion of the data from Switzerland. We kept two versions of the data, one with the median cholesterol value imputed for those missing values, and one with all the data from Switzerland dropped. That is, 171 rows of information dropped.

Ultimately, if we were to base our judgments entirely around one of the data sets, we think the insights drawn from the data set with the imputed values is closer to any 'real' relationship that exists in the world between all of these features and heart disease. As we saw in our modeling section, cholesterol turned out to be a pretty unimportant feature no matter which data set we used. Including those values, then, is helpful for the model learning about the other features since we didn't lose any data.

In the preprocessing stage of our capstone we OneHot encoded all our non-numeric variables specifically to prepare for using a Logistic regression machine learning (ML) model, which is a derivative of the Linear Regression ML family adapted specifically for predicting categorical outcomes. We will also use a random forest of decision trees model, which is an ensemble method for classification. We feel decision trees are a good point of comparison to other ML models because they're computationally simpler.

One patient had a record with 0 entered for their resting blood pressure. This record was dropped.

Exploratory Data Analysis

We saw some interesting distributions and correlations when exploring our data.

- Among the people who did suffer a heart disease, the vast majority of them experienced asymptomatic chest pain (non exercise angina)! We take this to mean they didn't feel chest pain at all, or any kind of pain typically symptomatic of a heart disease. We would have expected high correlation with heart disease and chest pain. See figure 2.
- ST slope flat dominated in the population that did suffer a heart disease. This is counter intuitive given that our model found ST slope up to be the most important feature for predicting heart disease. See figure 3.
- In a simple correlation heatmap, no variable had a pearson correlation coefficient higher than 0.55, which was, curiously, flat ST slope.

Modeling Insights

Our problem is a supervised learning problem. We already know what we're trying to predict (heart disease yes or no) and have the appropriate labels for this target variable. Therefore, we choose from the ML models in the supervised learning category. Secondly, our problem is a problem of classification, since the variable we're trying to predict, the presence of heart disease, is a categorical variable (yes or no). We are not trying to predict something continuous like height or weight, which can be anything from a continuous range of numbers. Therefore, we select classification models from the supervised learning family of models. We tried using a logistic regression model and a random forest of decision trees model.

In the end, the random forest model with a 75%/25% training/test split tied score wise with the logistic regression model. On both models we optimized hyperparameters with Gridsearch CV. For the logistic regression model we used ‘leave one out’ cross validation. For the random forest model, we used k = 5 folds cross validation. For a graphical illustration of the feature importance, see figure 4.

- ST slope up is by far the best predictor of heart disease.
- Oldpeak comes second.
- The other features seem to have decreased importance in chunks: The next important group of features are ST_slope flat, max heart rate, asymptomatic chest pain, age, resting blood pressure, cholesterol, and both kinds of exercise angina. A sharp drop in feature importance happens after exercise angina, and the remain features (notably sex) are relatively unimportant.

Having ST slope up and oldpeak together as the top two most important features makes some sense. Oldpeak is the magnitude of the distance of the ST segment below the baseline. As the ECG needs to travel back up to the baseline for the T wave, the slope would have to be positive.

Modeling Metrics

Our problem is a problem of classification. Therefore, the diagnostic metrics we used were accuracy, precision, and recall. The most important metric for our purposes is the recall score for the positive class (HD). In certain classification applications, like in medicine, it is most important to have a high recall score. The smaller your recall, the more false negatives are occuring in your model relative to true positives. It is much more costly to send someone on their way with no treatment because your model predicted they don't have heart disease when they actually do than it is to run further tests or begin treating someone your model predicted does have heart disease when they actually don't.

Both our tuned logistic regression model and random forest classifier had the same scores: 87% accuracy, 91% precision, and 88% recall. See figure 5.

Final Thoughts

Based on our model's feature importances, we'd like to recommend people do what they can to bring their ST segment more inline with the baseline of their ECG. Of course, this is a rather inactionable recommendation for most people. There may be an understood mechanism between 'within reach' changes like diet and exercise and the characteristics of one's ST segment, but we are neither biologists nor cardiologists and cannot make those kinds of recommendations. What we can recommend, however, is that careful evaluation of the ST segment of a patient's ECG be a part of screening for heart disease.

Certainly more work is needed to investigate the explanation for why ST slope flat could occur in the highest numbers with heart disease havers, yet ST slope up be the best predictor of heart disease as determined by a ML algorithm. We'd be interested in understanding any causal relationship between characteristics of the heart that give a depressed ST segment and the manifestation of HD. We would also like to see if anything would change if dietary cholesterol were an included feature of the data set, rather than just serum cholesterol. More data from other parts of the world would be interesting. Or perhaps a by region analysis, since diets and lifestyle habits vary so much in different parts of the world.

Figures

Figure 1: ST segment

The dashed line represents the baseline. This figure depicts ST depression, also known as oldpeak, with varying slopes.

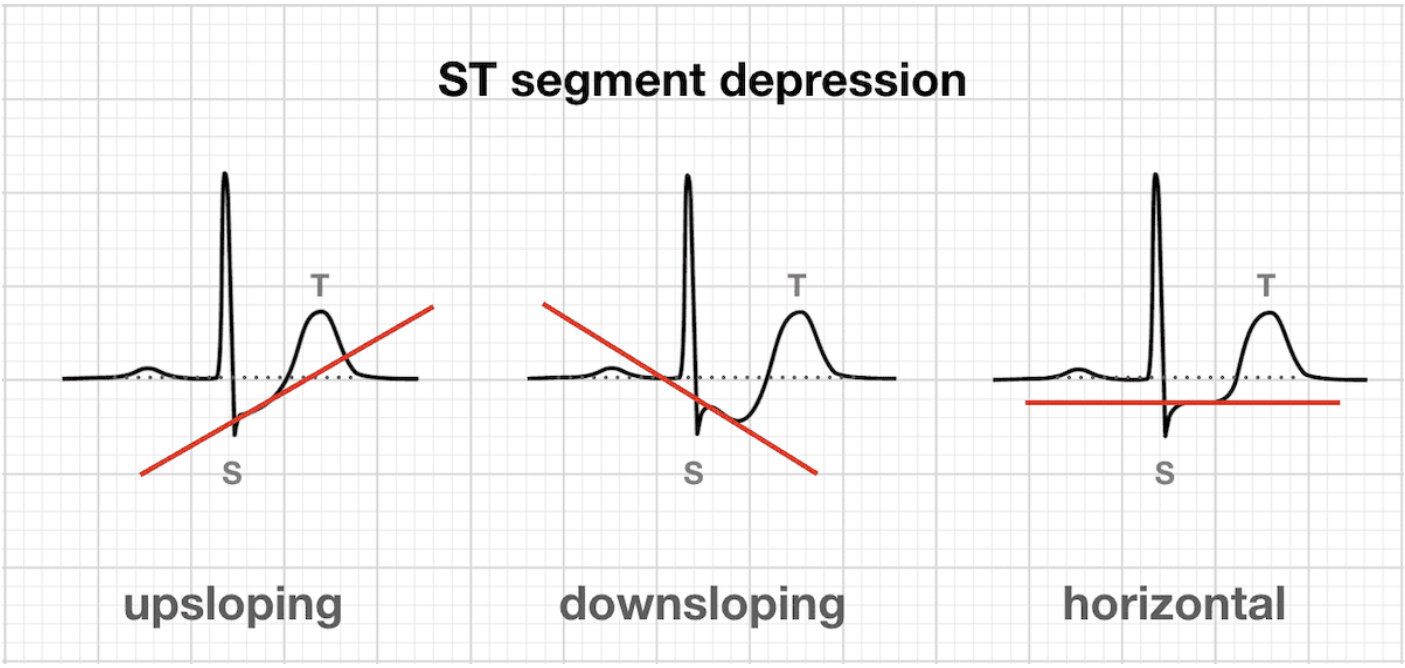


Figure 2: Distribution of heart disease by chest pain type

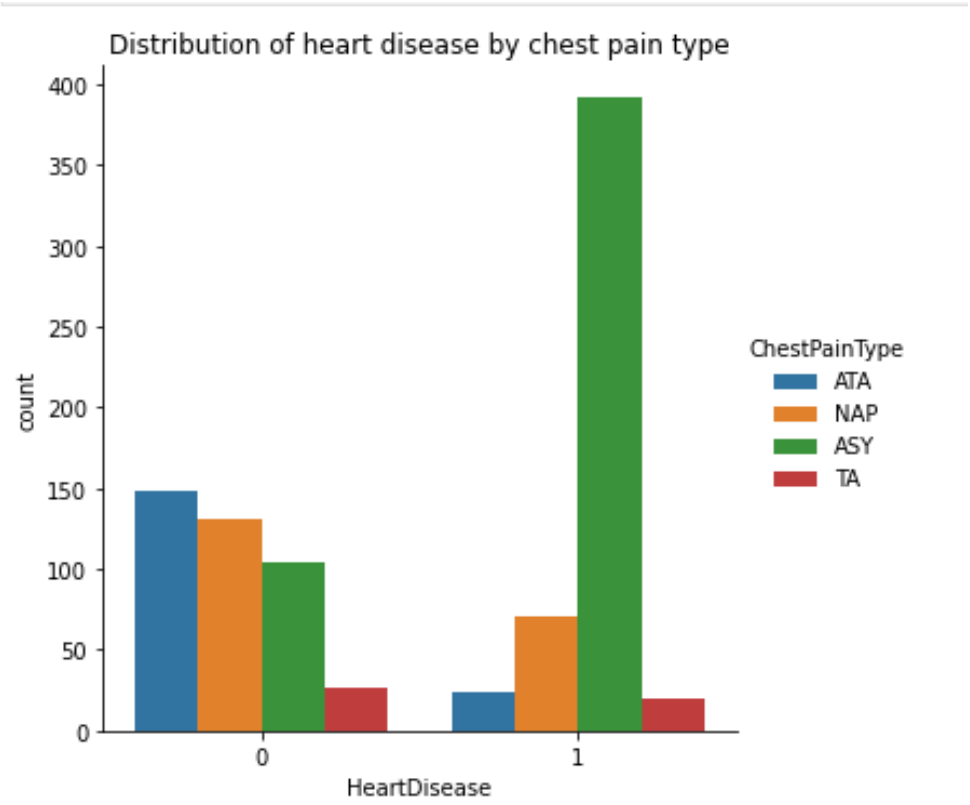


Figure 3: Distribution of heart disease by ST slope

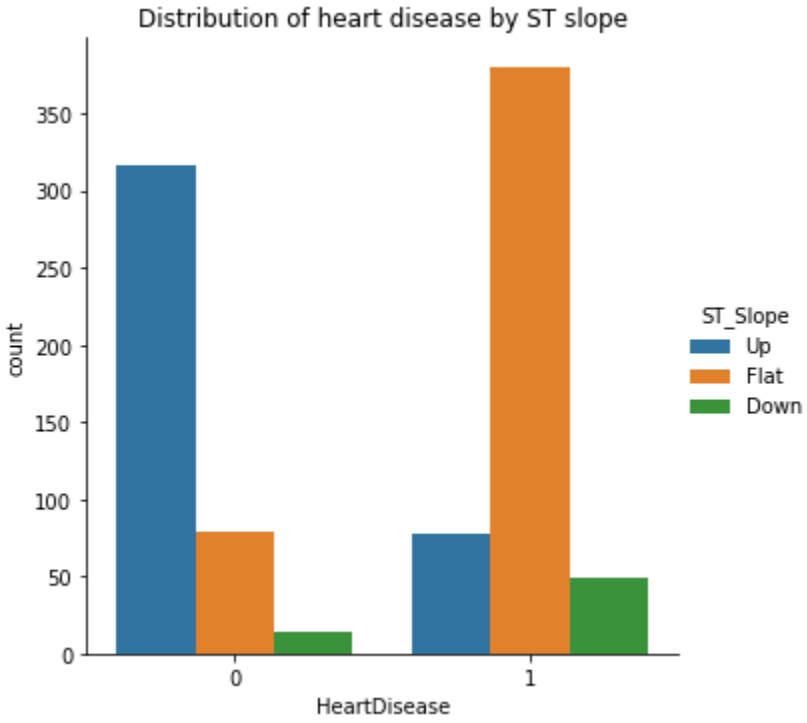


Figure 4: Feature importance (random forest model)

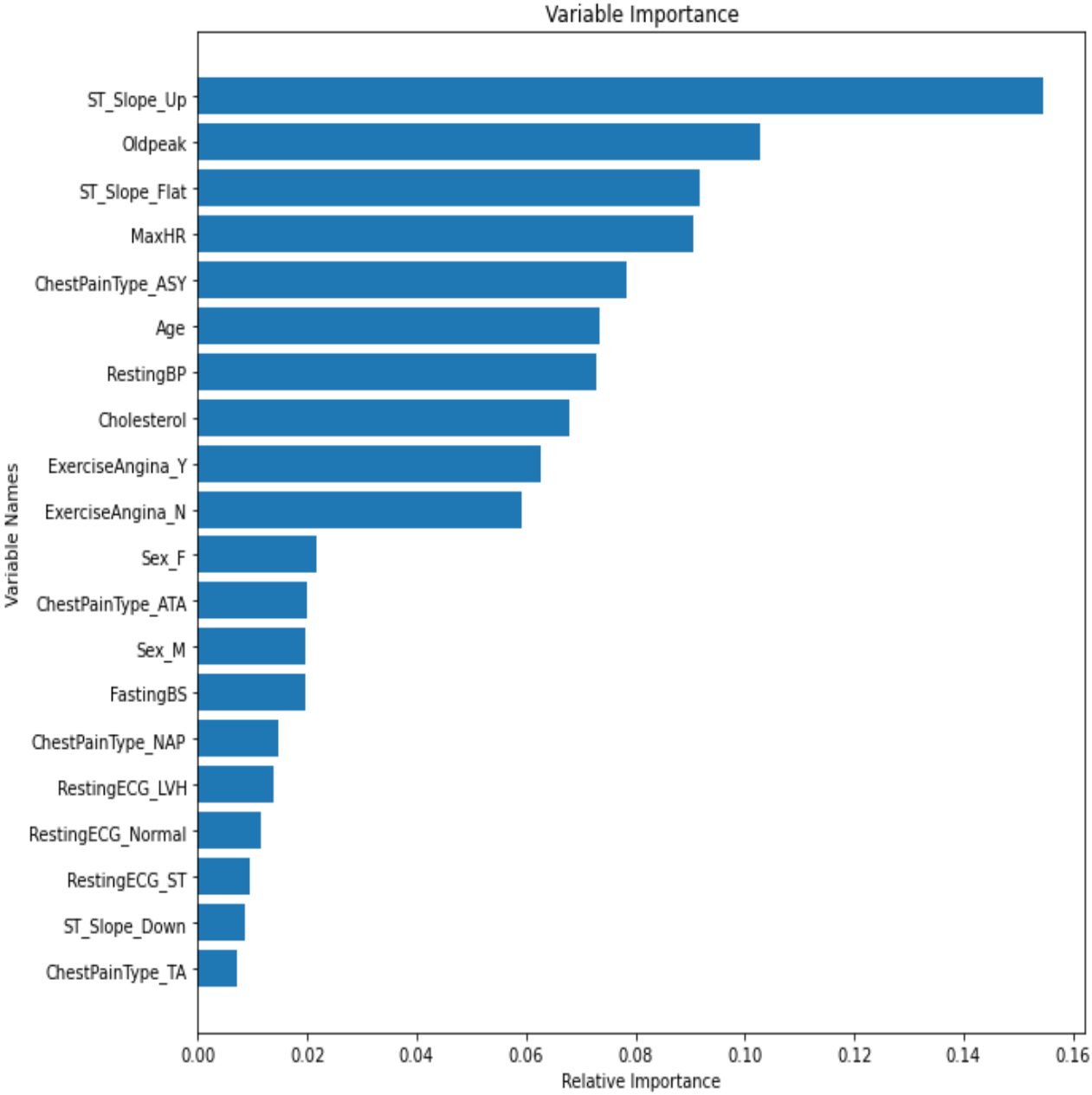
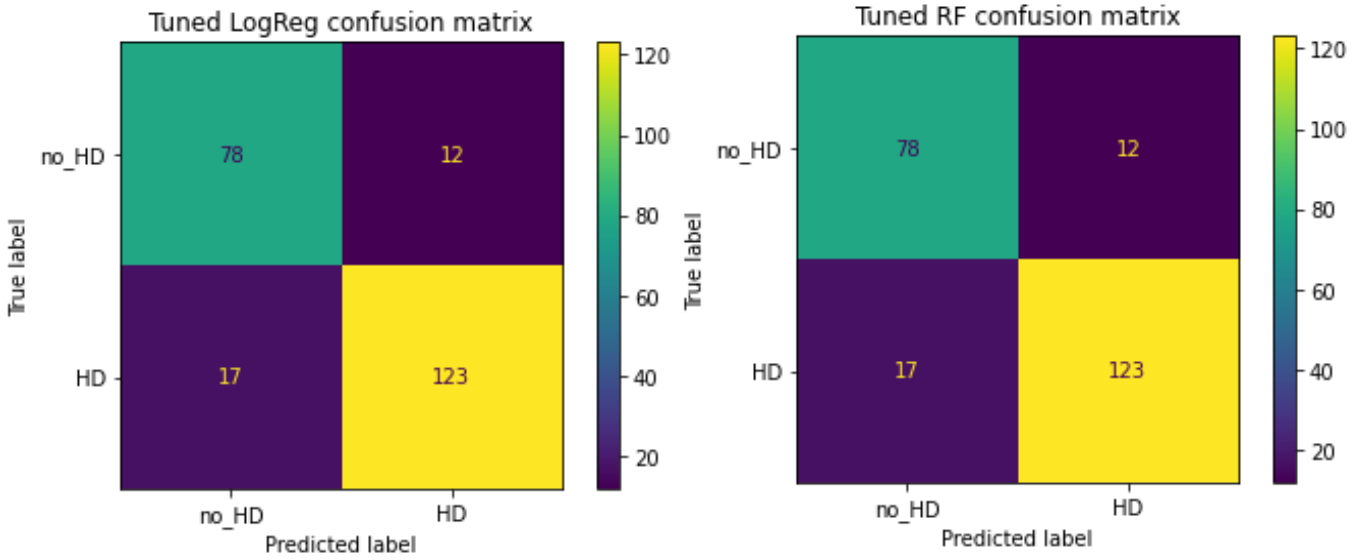


Figure 5: Confusion matrices and table of classification scores



	Precision	Recall	Accuracy
Tuned_LogReg	0.91	0.88	0.87
Tuned_RF	0.91	0.88	0.87