Cleveland Heart Disease Dataset : Analysing the Features Associated With Heart Disease

Karthik Sivaram

14th June, 2020

Executive Summary

In this report I summarise and distil the main insights from my analysis of associational relationships between different feature variables and the response variable – diagnosis of Coronary heart disease. In this pursuit, I approach the problem through 3 different techniques, each building on the conclusions from the former with increasing sophistication.

First, I perform exploratory analysis of the feature set, rationalise their coding, and identify naive associational relationship between each one and the response variable. I develop a metric framework to compare and rank features by their associational relationship.

Second, I examine the conditional effect of features on the response variable using generalized linear models. I discuss the meaning and interpretation of these effects and their limitations.

Third, I apply machine learning algorithms to build predictive models and from these I examine the relative 'Importance' of each feature. I discuss the advantages and limitations of these techniques for inference and use the results to build a full picture of the associational relationships in the data.

Based on these, below are my top 3 insights:

1. I can claim with reasonable certainty that the 3 most important features associated with an increased chance of heart disease diagnoses are: a.) Presence of Coloured Arteries in Fluoroscopy images, b.) detection of any defect through the Thalium Stress test, and c.) a flat or downward slope of ST-Segment during peak exercise. It is obvious that these are diagnostic tests to investigate the cardiac health of a patient and the positive association in all three cases is when they show a symptomatic result, that is the opposite result of a healthy person. Therefore, we can safely conclude that these tests work well in identifying coronary heart disease, at least better than any of the other features, including other medical tests, in the dataset.

2. There are only two features in the dataset that code for characteristics of the patient – Age and Sex. Among these sex seems to have the clearest association with heart disease. Men are much more likely than women to be diagnosed with heart disease holding all other features constant. Although there is a naive positive association between being older and an increased chance of heart disease diagnosis, this effect disappears when conditioned on other features. I wish there were more patient related features in the dataset, like family history of heart disease, obseity, or smoking habits. This would have allowed for a greater understanding of what characteristics are conditionally associated with heart disease diagnosis.

3. A counter-intuitive result from the dataset is that not reporting chest pain (asymptomatic) seems to have significant positive association with diagnosis of heart disease. Moreover, reporting Anginal pain, a typical symptom of heart disease, seems to have the lower positive association with heart disease than being asymptomatic._Prima facie_ this seems like a coding or data entry error. However, it is possible that it is not an error and that this variable has a non-linear effect or an interaction effect along with another feature on heart disease diagnosis. There is some suggestion of this from the application of machine learning, which is better at picking up such effects. Another explanation could be that those

with chest pain symptomatic of heart disease were selected out of this study or not in a condition to take part. It is also possible that since this feature seems to be self-reported that it might not actually reflect truth.

Scope of Analysis

There are three important caveats regarding the scope of this analysis that I wish to state at the very beginning:

1. Although I apply some techniques that are associated causal inference, it is **not possible** to make a causal claim regarding heart disease using this dataset for two reasons:

- First, the dataset deals with an observational study and not an experimental one. That is there is no randomly assigned treatment variable that we can exploit to estimate an average causal effect.
- Second, I have no domain expertise or specific knowledge of cardiology or for that matter medicine. So I am unable to build a structural casual model using the available features that would allow for estimation of causal relationships.

2. Therefore the scope of this study is strictly restricted to analysing the association relationship in the dataset using my knowledge of statistics and machine learning. Since the dataset comes from a hospital in Cleveland, it maybe safe to assume that it is a representative sample of patients at risk of heart disease. Therefore, the associational claims maybe generalisable to such a population.
3. The larger purpose of this exercise is to demonstrate how statistics and machine learning can be applied for inference and distilling insights from data.

In the following sections I present the top insight from each section.

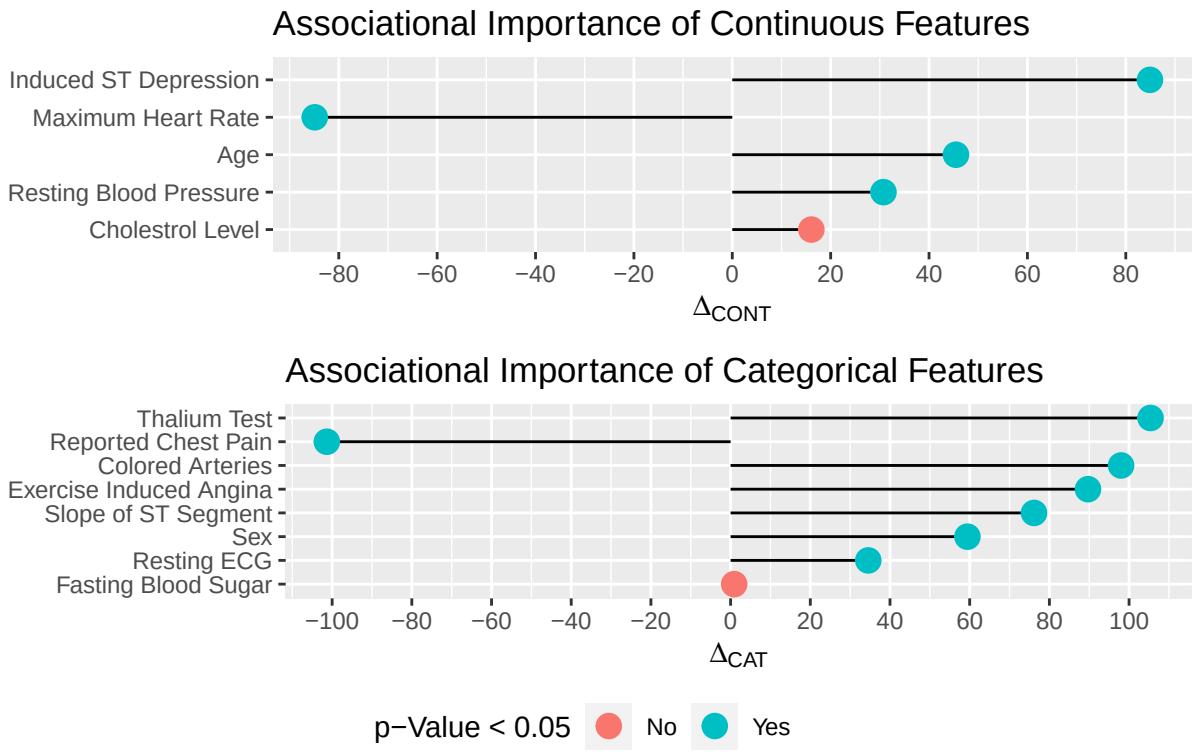
Approach 1: Naive Associational Relationships

I call this approach naive not in the literal sense of an endeavour being unsophisticated, but in the mathematical sense that I look at the assocaition between each feature and the response variable without conditioning on other features. I propose a framework with separate but related metrics for continuous and categorical variables that would allow for comparison of their naive assocaition with the response variable. I call these Δ_{CONT} and Δ_{Cat}. I define them mathematically below.

$$\Delta_{CONT}(Var) = \frac{(\hat{\mu_{Var|H=1}} - \hat{\mu_{Var|H=0}})}{\sqrt{\sigma^2_{Var}}} \times 100$$

$$\Delta_{CAT}(Var) = \frac{(\hat{\mu_{H|Var=1}} - \hat{\mu_{H|Var=0}})}{\sqrt{\sigma^2_{H}}} \times 100$$

H refers to the dichotomous response variable taking value 1 for diagnosis of heart disease and 0 for not. Below I show lollipop plots of these metrics for all features. From these it is possible to gather a basic understanding of how these variables are related. This feeds into my second approach.



Approach 2: Conditional Effect of Features Using Generalized Linear Models

In this section I shall print the results from two different logistic regressions. In the first one I condition on all feature variable and in the second I take a step-wise approach to determine the model of best fit by dropping and including various permutations of features and select the one with lowest AIC. I also take precautions to ensure that I am reasonably satisfying the assumptions of logistic regression.

Consider the regression table in the next page. It shows the coefficients of each feature for two of the regressions specified. Empty cells indicate that the Step-wise selection method dropped the variable. p-Values < 0.05 indicate that we can reject the null hypothesis that the effect of the feature is 0 (or 1 in odds-ratio terms).

For continuous features, the odds-ratio maybe interpreted as the change in probability of being diagnosed with heart disease for a unit change in the concerned feature holding all other features at the same value. From the table we can see that none of the continuous features turn out to be significant. For the sake of understanding the logic of interpretation, let's assume that Cholesterol level feature was significant. This would mean that for a unit positive change in Cholesterol Levels holding all other features constant, the probability of being diagnosed with heart disease would increase by 109% or 1.09 times relative to the probability of no heart disease diagnosis. In this case since I standardized our continuous variables, a unit change would be a 2 standard deviation change. While this specification makes the coefficients of categorical and continuous features comparable,

For categorical features, the interpretation is similar but instead of a unit change, it would mean a change in category with all other features being held constant. From the table we can see that feature Colored Vessels has the largest statistically significant effect with an odds-ratio of 8.19 in the Step-wise model. This means that if a person were to go from no colored vessels(0) to at least one colored vessel (1) while none of the other features change, the probability of being diagnosed with heart disease would increase by 719% or 7.09 times relative to the probability of no heart disease diagnosis. Now consider the feature Reported Chest Pain which shows an odds ratio of 0.15, which is statistically significant. This would mean that a patient who

Table 1: Results from Logistic Regression Models

	Naive Logit			Step-wise Logit		
Characteristic	OR	95% CI	p-value	OR	95% CI	p-value
age	0.75	0.31, 1.78	0.5			
sex	3.65	1.37, 10.2	0.011	3.65	1.43, 9.82	0.008
trestbps	2.06	0.97, 4.53	0.065	1.82	0.91, 3.75	0.095
chol	2.05	0.90, 4.81	0.090	2.09	0.96, 4.68	0.066
fbs	0.48	0.15, 1.46	0.2			
thalach	0.50	0.19, 1.28	0.2	0.48	0.19, 1.14	0.10
exang	1.84	0.78, 4.32	0.2			
oldpeak	1.98	0.82, 4.97	0.14	2.15	0.91, 5.25	0.085
cp_clean	0.19	0.08, 0.41	<0.001	0.15	0.07, 0.32	<0.001
restecg_clean						
0						
1	1.50	0.71, 3.19	0.3			
slope_clean	2.66	1.10, 6.58	0.032	2.57	1.08, 6.18	0.033
ca_clean	9.92	4.31, 24.5	<0.001	8.19	3.82, 18.7	<0.001
thal_clean	3.52	1.54, 8.25	0.003	3.42	1.54, 7.75	0.003

[1] OR = Odds Ratio, CI = Confidence Interval, OR = Odds Ratio, CI = Confidence Interval

changes from not reporting chest pain to reporting chest pain with all other features remaining the same, the probability of being diagnosed with heart disease would decrease by 85% or 0.85 times relative to the probability of no heart disease diagnosis.

Using these regression estimates, we can conclude that when conditioned on all available features, an increase in probability of heart disease diagnosis relative to no heart disease diagnosis is associated with a change in category (from 0 to 1) of Colored Vessels, Thalium Test, Slope of ST-Segment, and Sex. The former three implies a symptomatic/abnormal test result whereas the latter implies a change in sex from female to male. Similarly, we can also conclude that a change from reporting no chest pain to reporting chest pain is associated with a decrease in probability of heart disease diagnosis relative to no heart disease diagnosis.

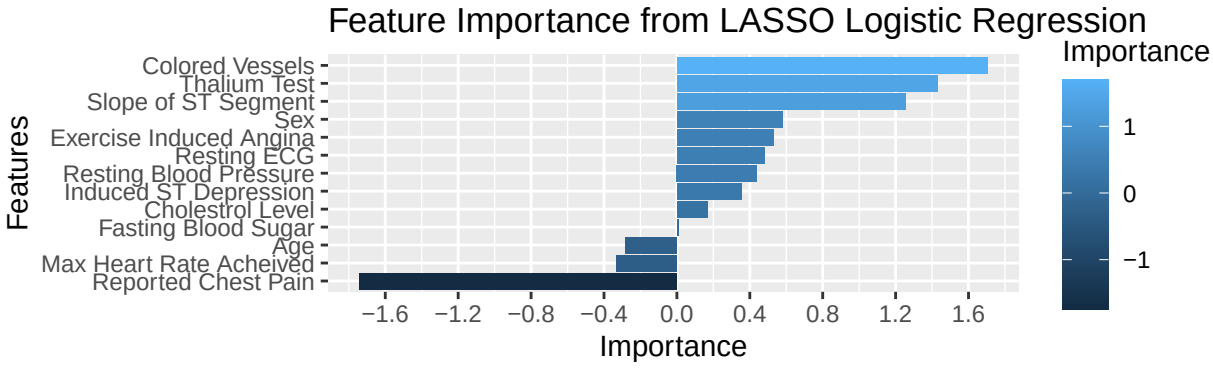
Approach 3: Feature Importance using Machine Learning Algorithms

In this approach, I use two machine learning algorithms optimized to predict diagnosis of heart disease. Although predictive power is not of importance for this analysis, these algorithms allow for ranking features based on their relative importance in predicting the response variable. These techniques are also better at identifying non-linear relationships and interactions between features compared to previous techniques. These models deal with the Bias-Variance trade-off much better than normal regression techniques which are constrained by seeking to compute unbiased estimates even at the cost of high variance. Therefore, this approach truly exploits machine intelligence with far less intervention from my end in contrast to previous sections.

LASSO Logistic Regression

LASSO Logistic Regression is a powerful yet simple machine learning algorithm that is similar to a logistic regression in its basic structure. However, it includes a penalty term hyper-parameter(λ) that works to shrink the coefficient of features to zero to ensure optimal trade-off between bias and variance to ensure maximum predictive power. Thus, it also works as a test of association between features and the response variable. However, the estimated coefficients of features are not very useful or interpretable since by design they are biased (due to the penalty). Therefore I shall not report these coefficient estimates. Nonetheless, there are

generalizable, model agnostic technique to determine the relative importance of the features. Below I show the the featured ranked by importance.

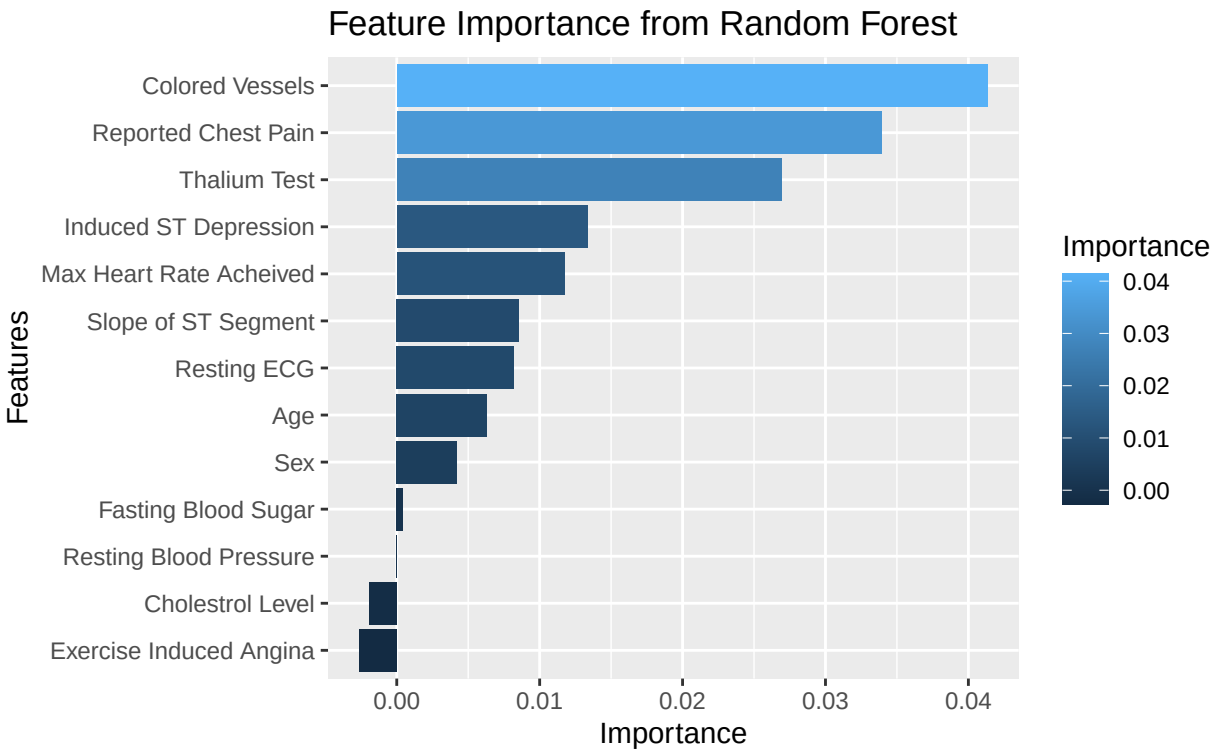


After hyper-parameter tuning through cross-validation, the LASSO model performs quite well in predicting diagnosis of heart disease. Even without tuning the threshold value (I round values > 0.5 to 1 and the rest to zero) the model has a accuracy of almost 80%. More importantly, the model provides an importance estimate for each feature as seen in the plot above. This result seems to be identical to the one from ordinary logistic regression. Colored Vessels, Thalium Test, Slope of ST Segment, and Sex seem to be the most important feature associated with increased chance of heart disease diagnosis; Reported Chest Pain seems to be the most important variable associated with a decreased chance of heart disease diagnosis. The fact that this model entirely agrees with the ordinary logistic regression should not come as a surprise since they both have similar structure, even if different optimization strategies. Nonetheless, this adds greater weight to the conclusions from the earlier section.

Random Forest Model

Random Forest is an ensemble learning method that has become one of the most, if not the most, popular machine learning technique. They are known to perform very well in classification tasks and provide lots of opportunities for optimization with at least three hyper-parameters that can be tuned. They are excellent at picking up on non-linear relationships and interactions between features that are beyond the scope of linear models. This is the main reason for them being included in this analysis. Through their application I hope to uncover insights that may have eluded earlier techniques.

Below I show the the featured ranked by importance.



After some tuning, I settle on a model that has a prediction accuracy of between 70 and 80%, which is not even better than LASSO.

However, the feature importance plot tells a far more interesting story. First of all, it is important to note that these estimates of importance do not seem stable. In running this code chunk a couple of time I have been shown quite different results even though the seed had remained the same. In fact I am not sure if the results will remain the same when I knit this document. Since the model relies on building decision path through random subsets of variables, this is not be expected. Therefore it'd be risky to generalize. This also suggests that the model needs to be tuned better.

Second, in comparison to the LASSO model, importance is more tightly distributed among the features; the model is not over-reliant on some features like the LASSO model. Although I am not sure if a direct comparison of values of the Importance estimate is warranted, it is still worth noting that the range of Importance is vastly smaller in scale compared to LASSO. This suggests that some features, although not significant by themselves, become important in conjunction with others. This suggests that there maybe significant non-linear effects and interactions between features that is associated with the response variable.

Third, although the most important features from the LASSO model continue to be important, there are some noteworthy changes in rank. The most conspicuous result seems to be for Reported Chest Pain. This feature has changed from having a strong negative association with the response to a strong positive association. This suggests that the feature maybe producing this effect through interactions with other features or in certain specific cases. Another feature that has similarly changed direction seems to be Maximum Heart Rate achieved which has flipped from being weakly negative associated to positive association.

In conclusion it difficult to take the estimates from the Random Forest model seriously given its instability, especially in comparison to LASSO and Logistic model which are far more stable.