Heart Disease Prediction
UCI Study Data

Brad Doty
December 2021

Problem Statement

1. Find a classifier to predict the target, which is heart disease as measured by angiogram, based on the 13 diagnostic test results in this UCI dataset on Kaggle.
2. Identify which diagnostic tests in this study data, formulated as features in these classification models, carry the most predictive information for these models.

Datasets

The dataset actually deviates from the documentation in several columns. Some variables don't have the same number of values as described and one column contains floats with one decimal place, rather than integer categories. Several of the column names are cryptic. In the Kaggle comments, a user posts a "corrected" dataset, which he says is more accurate than the original post. I investigated on Kaggle and UCI.

Information Digging at UC - Irvine

I could not find any corroborating evidence for either dataset on Kaggle, so I did some digging into the UC Irvine ML Data Repository at http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

This study was done in 1988 and analyzed by many papers over the years. There were 76 attributes considered in the study, but only 14 attributes were reported. This dataset is from the Cleveland Clinic. There were 3 other datasets from Long Beach VA Hospital, Hungary, and Switzerland. The others are not readily apparent at UCI.

Found these improved feature descriptions at UCI. Applied improved explanations in the data wrangling notebook Column Description list and improved the name set.

This is the clincher on which dataset to use. The UCI archive states that column 58 of the full set is column 14, the predicted attribute, of the reported set and it has only 2 values.

> 58 num: diagnosis of heart disease (angiographic disease status) -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing

Data Science Approach

After verifying that the original dataset with the binary target variable was preferable to the alternative dataset, I performed Data Wrangling, EDA, Feature Engineering, and used each of the main supervised classification algorithms to find the best-fitting model to predict heart disease from the dataset. I tracked and compared the coefficients or feature importances available from each model for attribution of

feature importance. Each step is performed and documented in its own Jupyter notebook, available in my Github repository.

Data Wrangling

Several categorical features trend with or against the target. A few features had one value going against the trend, so I moved values to make them in line.

For instance, the 4 values of Chest Pain are 0: No Pain, 1: Typical Angina, 2: Atypical Angina, and 3: Other Pain. Patients reporting 0 had a much lower risk of the target, while patients with 1 through 3 had slightly decreasing risk. So the 0 fits a decreasing risk order as a 4. I replaced 0s in this column with 4s to make it a trending category.

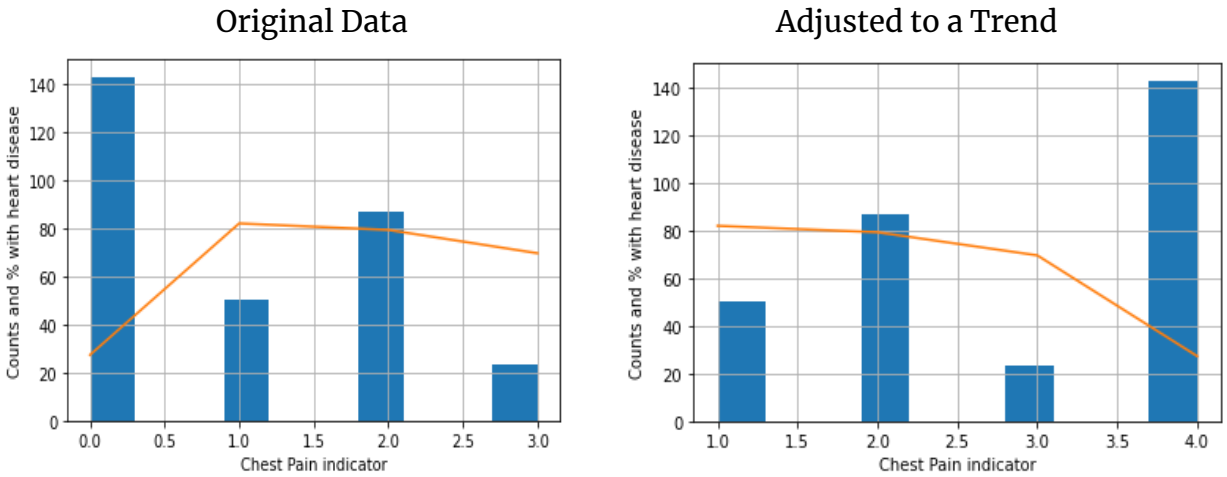

Figure 1. ChestPain: Raw and with 0 moved to 4.

STWave, a coding of characteristics of the ECG wave during a stress test, came in with many decimal values, rather than 4 integer categories as documented. I found that putting the histogram into bins greatly smoothed the percentage correlation with the heart disease target. So I rounded, with a couple of tweaks to balance out the categories, to make a nice categorical with an inverse relationship with the target.

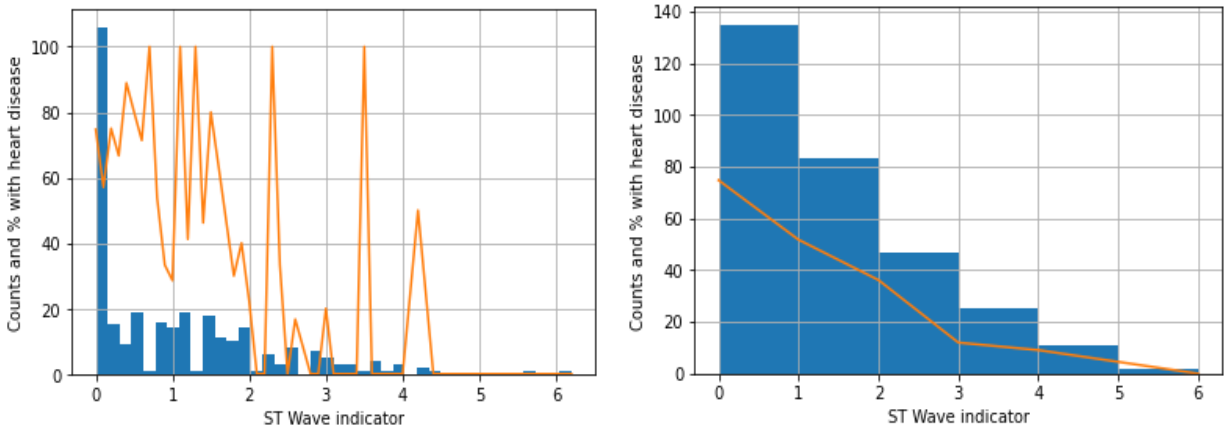

Figure 2. Raw STWave and STWave binned to integers

Exploratory Data Analysis

Exploratory Data Analysis with the adjusted columns shows that none of the candidate features are highly correlated to each other. The largest correlation is -0.55 between STWave and STSlope. The correlations with the target were not that high and surprisingly low or negative in a few cases, as shown in Fig. 3.

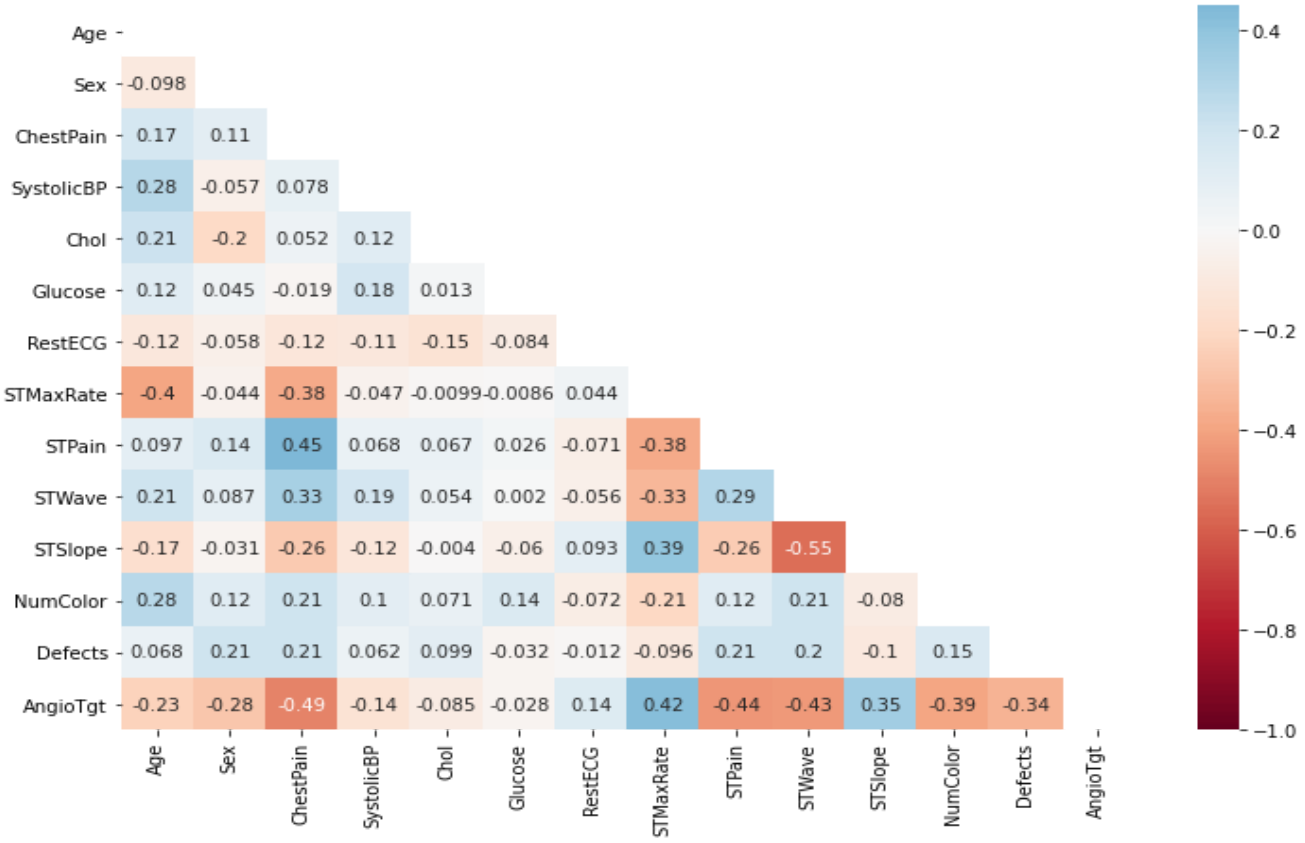

Figure 3. Correlation Half-Matrix

Feature Engineering

I used all 13 candidate features. Split the features into training and test sets. Scaled the training set and the test set separately to avoid any feedback from the test set to the training set.

Predictive Modeling

I used 5 algorithms to build 8 classification models.

Using this selection criteria to find the best prediction model:
1. Recall is most important for medical diagnoses.
2. F1 for a balanced fit is a tiebreaker
3. AUC of ROC is another tiebreaker.

Recall: SVM RBF, kNN k=7, and RF CV1 and CV2 tied at 83%. Weak for diagnostics.
F1: SVM RBF and kNN k=7 tied at 86%. Logistics was third at 85%.
AUC ROC: Logistic tied SVM Linear at 92%. SVM RBF, RF CV2, and RF D=9 tied at 91%.

Here are the fit metrics of each model.

Model	Recall	F1	AUC of ROC	Precision	Accuracy	R^2
Logistic Regression	79%	85%	92%	92%	87%	.474
Decision Tree	66%	73%	77%	83%	77%	.080
Random Forest (max d=9, gini, 100 trees)	79%	82%	91%	85%	84%	.343
#4. Random Forest CV1 (md=5, entropy, 75 trees, mxfeat=5, minsplit=5)	83%	83%	89%	83%	84%	.343
#3. Random Forest CV2 (md=4, gini, mxfeat=4, 130 trees, minsplit=3)	83%	83%	91%	83%	84%	.343
#2. K Nearest Neighbors k=7	83%	86%	90%	89%	87%	.474
#1. Support Vector Machine (RBF kernel)	83%	86%	91%	89%	87%	.474
Support Vector Machine (linear kernel)	79%	81%	92%	82%	82%	.277

Figure 4. Model Metrics

SUPPORT VECTOR MACHINES. I ran SVM with the default RBF kernel and also with the linear kernel to get feature coefficients. The RBF is best in all metrics except ROC AUC, where it is 1 percentage point behind SVM Linear and Logistic Regression. The Linear kernel fell off by 5-7 points for metrics except ROC AUC, where it tied for best. The SVM RBF is the best overall model, however the Recall of only 83% is disappointing for a diagnosis study.

K NEAREST NEIGHBORS. I ran K Nearest Neighbor against scaled data for K from 1 to 21, odd numbers only. Plotted accuracy versus K for training and test datasets to avoid overfitting with K too small. The best accuracy is at K=7, where it does slightly better on the test set than the training set. kNN k=7 exactly matches SVM RBF except on ROC AUC, where it loses out by 1 point to SVM RBF and by 2 points to SVM Linear and Logistic.

RANDOM FORESTS. I ran one random forest setting max depth to the depth of the decision tree, 9, and all other values at default, including Gini criterion and 100 trees. It beat the decision tree, especially in ROC AUC (9 pct. points), but still trailed logistics. Its Recall came in at 79% in a 3-way tie for 5th place.

The first cross-validation tied for best in the all-important Recall metric at 83%. Its ROC AUC also tied the champ, SVM RBF, at 91%, but trailed in other metrics. This cross-validation used this grid: n trees: 25, 50, ..., 175. Gini and entropy. Max depth: 5, 9, 10, 15. Min samples split: 2, 3, 5. Max features (to consider at splits, still using all 13 features): 5, 10, 13. This settled on (75 trees, entropy, max depth 5, min samples split 5, max features 5).

I adjusted the grid to bracket the best model. n trees: 60, 70, 75, 80, 100, 120, 130, 140, 150. Gini and entropy. Max depth: 4, 5, 9. Min samples split: 2, 3, 5. Max features: 4, 5, 6. This came back with exactly the same fit metrics, except ROC AUC ticked up 2 to 91%. This yielded the third-best model (130 trees, gini, max depth 4, min samples split 3, max features 4).

LOGISTIC REGRESSION. This is not a distance-based algorithm, but it failed to converge with the unscaled data, even when I bumped up the iterations from 100 to 1000. I tried the scaled data and it converged. However the fit was exactly the same as it was with the unscaled data. This tied for the best accuracy, R^2 and ROC AUC at 92%, 1 point better than SVM RBF. It has clearly the best Precision at 92%, but the Recall is just unacceptably low at 79%.

DECISION TREE. The decision tree was clearly inferior to all other models on this data.

Feature Attribution / Explanation of Target

To get an idea of the relative importance of the diagnostic tests in this data study, I collected the feature coefficients/importance. Some models provide coefficients, others provide feature importance estimates. SVMs with kernels other than the Linear kernel do not provide this information. Neither do kNN algorithms. See Figure 5, extracted from section 7. Feature Attribution/ Importance Table in the Modeling notebook.

	LogReg	LR_rank	RF0	RF0_rank	RFcv1	RFcv1_rank	RFcv2	RFcv2_rank	SVM	SVM_rank
NumColor	-1.19	1	0.14	1	0.14	1	0.17	1	-0.81	1
Sex	-1.12	2	0.03	13	0.03	11	0.02	13	-0.43	7
STPain	-0.96	3	0.06	10	0.07	7	0.07	6	-0.77	2
Defects_2	0.78	4	0.09	4	0.10	4	0.12	3	0.75	3
Defects_3	-0.71	5	0.08	5	0.10	5	0.11	4	-0.45	6
STSlope_1	-0.68	6	0.03	11	0.02	13	0.04	11	-0.55	5
Glucose	0.56	7	0.01	15	0.01	15	0.00	15	0.38	8
Defects_1	0.53	8	0.00	16	0.00	16	0.00	16	0.69	4
ChestPain	-0.36	9	0.11	2	0.13	2	0.14	2	-0.17	12
STSlope_2	0.33	10	0.03	12	0.03	12	0.04	10	0.22	10
STWave	-0.19	11	0.06	9	0.05	8	0.04	8	-0.23	9
RestECG	0.16	12	0.02	14	0.01	14	0.01	14	0.14	14
STMaxRate	0.03	13	0.11	3	0.12	3	0.10	5	NaN	NaN
SystolicBP	-0.02	14	0.07	8	0.05	9	0.04	9	NaN	NaN
Age	0.02	15	0.08	6	0.09	6	0.06	7	NaN	NaN
Chol	-0.00	16	0.08	7	0.05	10	0.04	12	NaN	NaN
STMaxRate_SS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.21	11
SystolicBP_SS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-0.17	13
Chol_SS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-0.01	16
Age_SS	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.10	15

Figure 5. Available Feature Attribution

Unfortunately, these data do not support any definitive conclusions about the relative importance of the diagnostic test features in the data.

1. All the models except Decision Tree put NumColor as the most important feature. This would make sense because the target measure of heart disease is the narrowing of the cardiac artery via angiogram and the coloration is another way to measure the narrowing of 4 arteries in the area. **However, the models don't agree on the direction of its influence!!**
2. Logistic Regression, the only model of the top 3 fits which produces an attribution, puts Sex (M=1) at the 2nd most important, with male subjects less likely to have the target than the female subjects. The other models show Sex as one of the least important features
3. Logistic counts STPain (during the stress test) as the 3rd most important, the Linear Support Vector Machine puts it at 2nd. All the tree-based models rank it middling to low.
4. Defects 2 and 3 (Fixed or Reversible Defect) come in 4th and 5th or better in most of the models.
5. Chest Pain (in general, not during the stress test) comes in 2nd in the Random Forest models, but 9th in Logistic and 12th in SVM Linear.
6. None of the other features really stands out.

If one wants to rely on an average of our #3 and #4 models, which tied for Recall performance (RF2 and RF1), then we could say that the top 5 diagnostic features in this data are

1. NumColor 0.16 coefficient
2. ChestPain 0.135

3. Defect 2 (Fixed Defect) 0.11
4. STMaxRate 0.11
5. Defect 3 (Reversible Defect) 0.105

The other features have coefficients below 0.1.

Deliverables

A. Jupyter notebooks describing analysis and design decisions, Python code, and data visualizations for the Data Wrangling, EDA, Feature Engineering, and Predictive Modeling phases
B. This project final report
C. Model Metrics as a separate document
D. Presentation slide deck