Heart Disease Prediction
UCI Study Data

Brad Doty
December 2021

Heart Disease Prediction Problem Statement

1. Find a classifier to predict the target.
   - 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.

UCI / Cleveland Clinic Study Data

The data at Kaggle deviates from the data description in several columns.

Investigated at http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

- Study was conducted in 1988, analyzed by many papers over the years
- 76 attributes considered from
  - Cleveland Clinic, Long Beach VA Hospital, Hungary, and Switzerland
- Only 14 attributes reported at UCI, only from Cleveland Clinic
  - Target variable is binary indicator of cardiac artery blockage

Data Wrangling

- There was no null data, 303 full rows
- Some cardinality and values deviated from the data description
- Investigated an alternative dataset from Kaggle, but kept original
- Corrected values, improved naming based on UCI data dictionary

Original

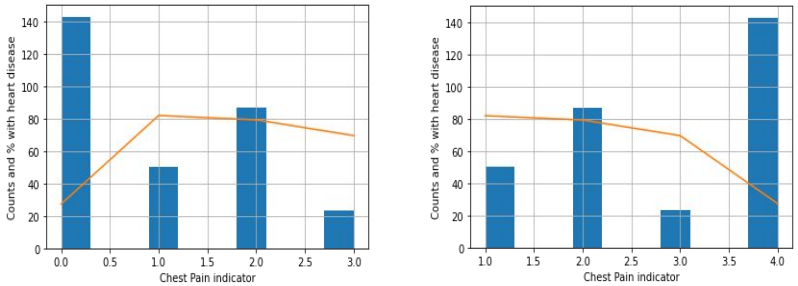
```
#   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
```

Renamed

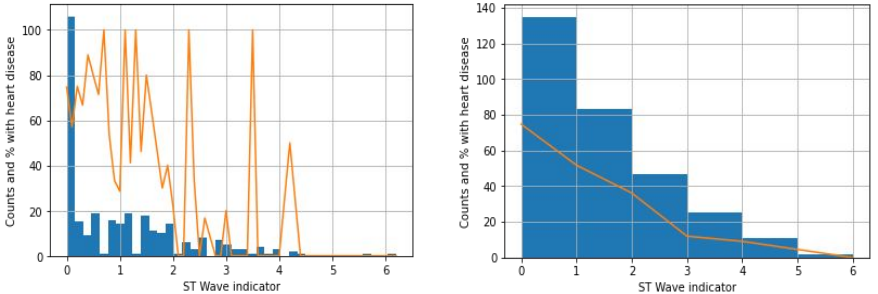
```
#   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Age         303 non-null    int64
 1   Sex         303 non-null    int64
 2   ChestPain   303 non-null    int64
 3   SystolicBP  303 non-null    int64
 4   Chol        303 non-null    int64
 5   Glucose     303 non-null    int64
 6   RestECG     303 non-null    int64
 7   STMaxRate   303 non-null    int64
 8   STPain      303 non-null    int64
 9   STWave      303 non-null    int64
 10  STSlope     303 non-null    int64
 11  NumColor    303 non-null    int64
 12  Defects     303 non-null    int64
 13  AngioTgt    303 non-null    int64
dtypes: int64(14)
```

Data Wrangling 2

- Some categoricals showed trend vs. target, some values broke trend.
- Adjusted some values to make monotonic trends.



ChestPain: Raw and with 0 moved to 4.

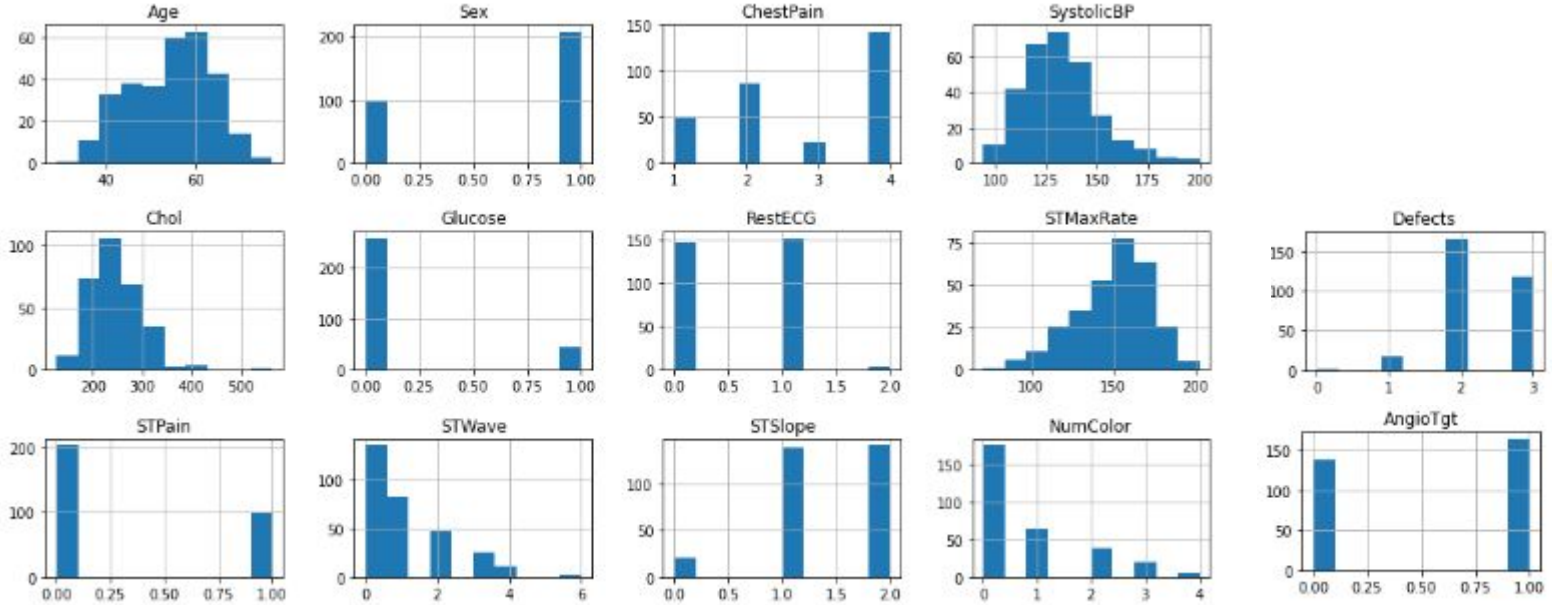


Raw STWave and STWave binned to integers

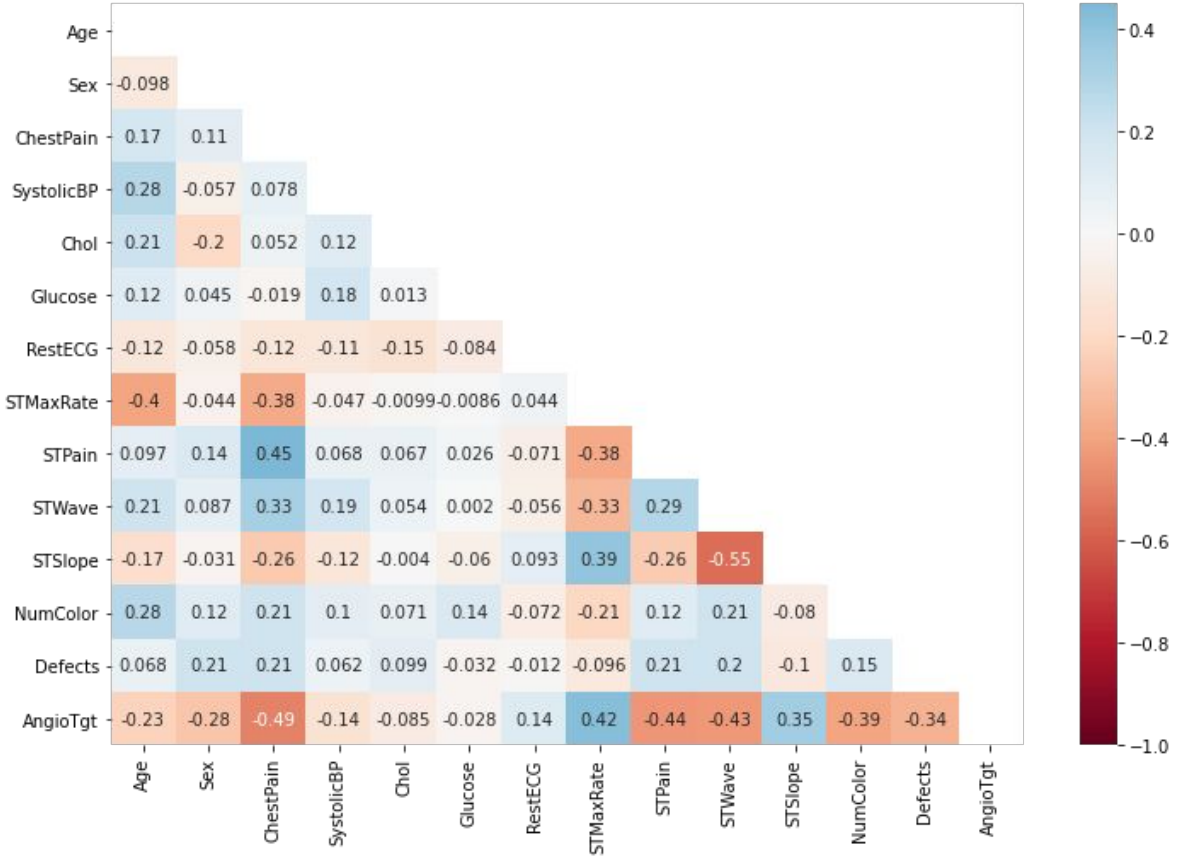
Exploratory Data Analysis - Distributions

4 skewed, unimodal numeric distributions: Age, SystolicBP, Chol, STMaxRate

6 categoricals, 3 binaries, plus the binary target: AngioTgt



Exploratory Data Analysis - Correlations



Feature Engineering

- AngioTgt is binary target. Do single classification.
- No highly collinear variables. Used all 13.
- One-Hot Encoding, drop first category on STSlope and Defects
- Left some categoricals alone:
  - AngioTgt monotonically trends over ChestPain, RestECG, STWave, NumColor
  - 3 binary features: Sex, Glucose, STPain
- Remapped values in Data Wrangling (pre-EDA) to make trends
  - NumColor 4 -> 0
  - STWave 6 -> 5
  - RestECG 2 -> 0
- Did 80/20 Training/Test Split
- Used Standard Scaler on continuous features for distance-based algos
  - Age, STMaxRate, SystolicBP, Chol

Modeling

- Built models using 5 supervised classification algorithms
  - Logistic Regression, Decision Tree, Random Forest, k Nearest Neighbors, SVM
- Used scaled data for distance-based algorithms
  - k Nearest Neighbors, SVM
- Selection criteria to find the best prediction model:
  - Recall is most important for medical diagnosis tool.
  - F1 for a balanced fit is a tiebreaker
  - AUC of ROC is second tiebreaker.
- Recorded Feature Attribution (feature importance or coefficents) where possible
  - Logistic Regression, Decision Tree, Random Forest, SVM Linear (not SVM RBF)

Modeling Metrics

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

4 models tied on Recall

SVM RBF, KNN, RF2, RF1 at 83%

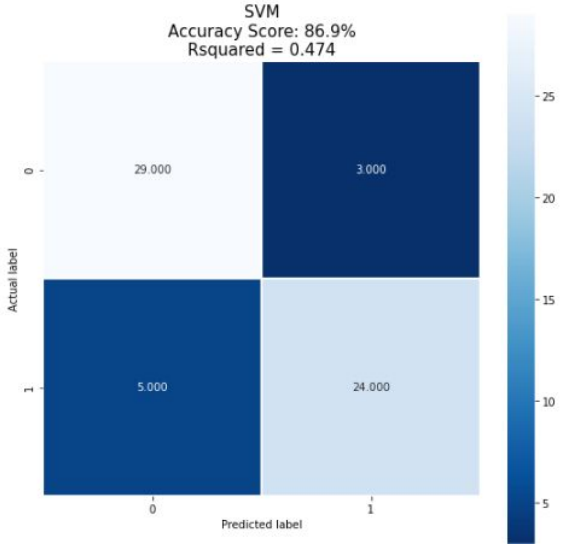
F1 tiebreak: SVM RBF, KNN

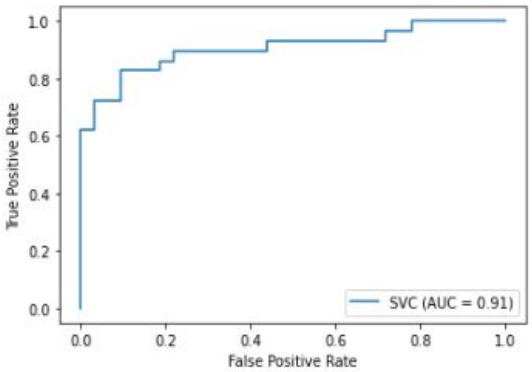
ROC AUC tiebreak:

SVM RBF

SVM RBF wins !

Support Vector Machine w/ RBF Kernel





Hyperparameters:
{'C': 1.0, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 'scale', 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': 21, 'shrinking': True, 'tol': 0.001, 'verbose': False}

	precision	recall	f1-score	support
0	0.85	0.91	0.88	32
1	0.89	0.83	0.86	29
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61

Feature Attribution

	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

#1 SVM RBF & #2 KNN do not yield attribution.

Is there a consensus among others?

No. NumColor is + for RF and - for Log & SVM Lin.

#3 RF2 & #4 RF3 agree:

NumColor, ChestPain, Defect 2 & 3, STMaxRate