Factors Analyzed for their Contributions to Heart Disease

Capstone 1 Report

V. Moore

6/11/18

Predicting heart disease risk

- The problem: heart disease is one of the greatest causes of death, though most influential factors are not necessarily clear.
- Understanding these factors could be useful to the patient care community, patients themselves, insurers, policy-makers, and drug researchers.
- Data used for analyses presented here come from the University of California, Irvine, Machine Learning Repository and included partially processed data from Cleveland, OH, Long Beach, CA, and Hungary.
- The goal of this project is to develop an accurate model for heart disease risk that can be interpreted globally and individually.

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

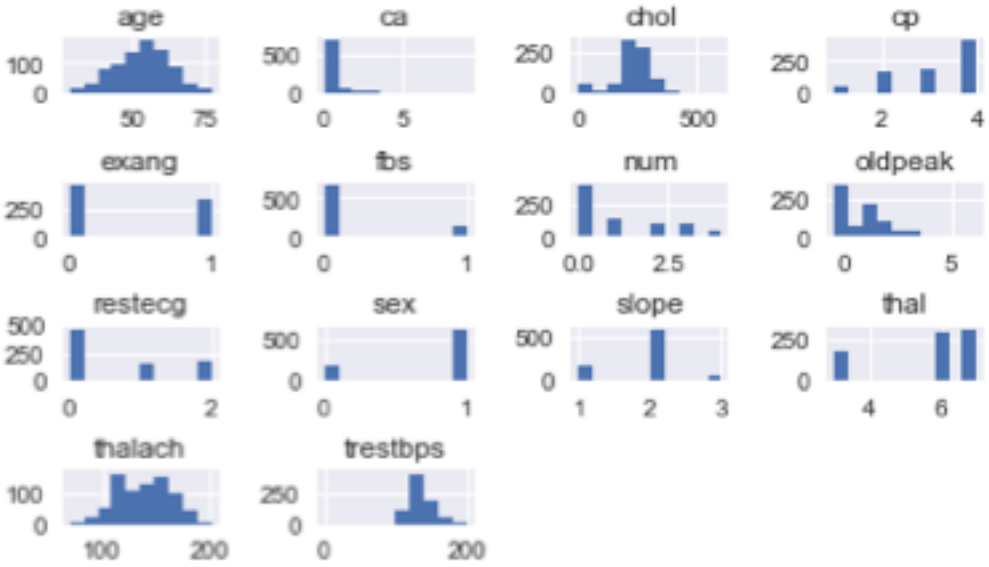
Location	Heart Disease Rate
Ohio	7.62
Cleveland, OH	8.76
California	5.30
Long Beach, CA	5.46

- Heart disease rates (Table 1) in some locations relevant to this study indicate prevalence of heart disease is somewhat higher in Cleveland, OH, than in its surrounding state and also higher than in Long Beach, CA, which shows heart disease at a rate similar to its surrounding state.
- Variables examined in this study are shown in Table 2 with their definitions. The target variable is “num”, or “numerical heart disease score”.

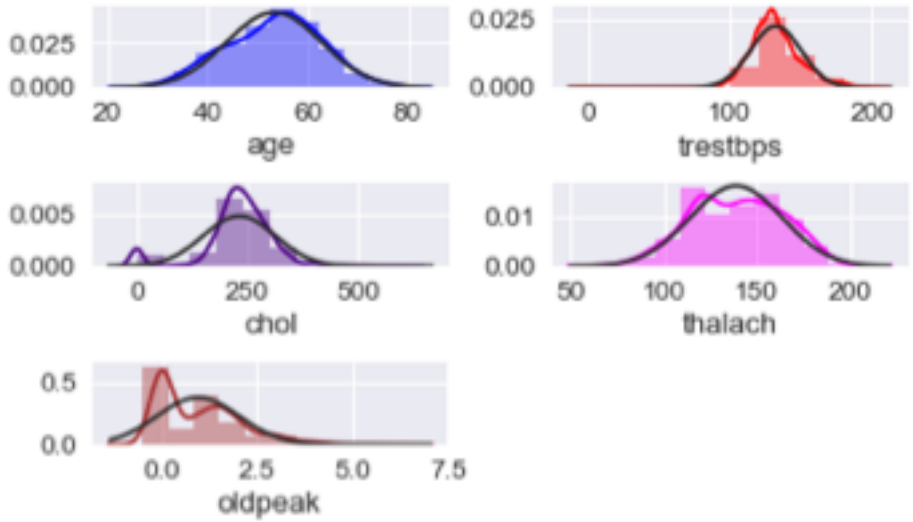
Table 2. Heart disease feature names and their abbreviations

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

Quick look at the data



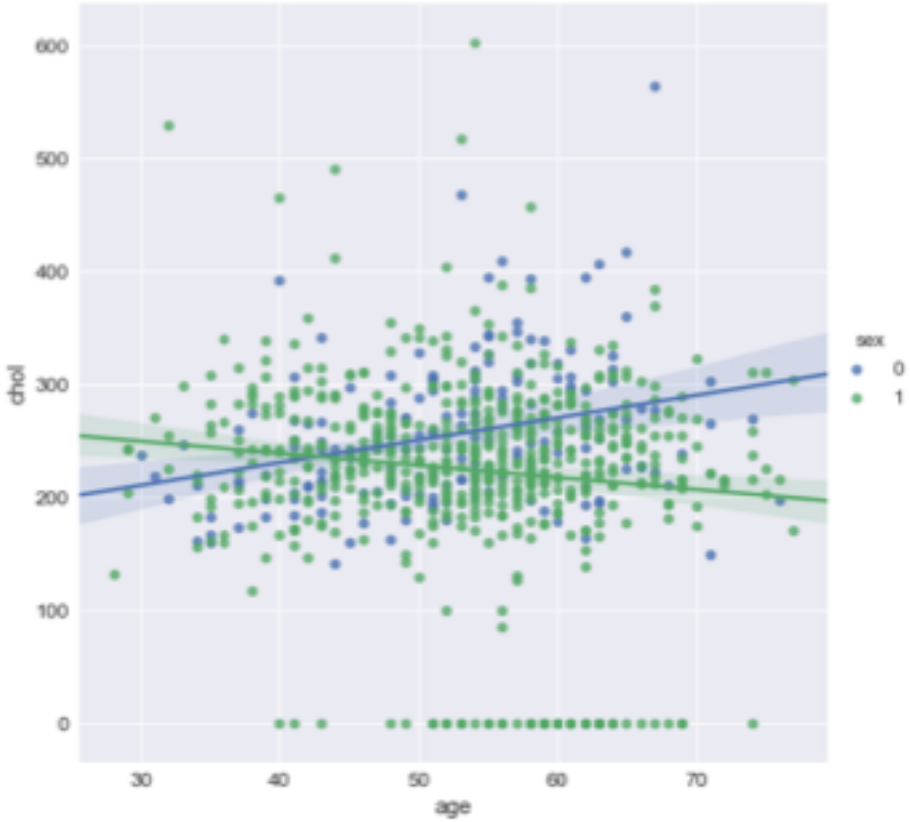
For quick glances at the form of the data for each feature.



To glimpse the distributions (versus normal distributions in black) of continuous variables.

Some relationships between variables

Age versus cholesterol level



Age versus maximum heart rate

Some relationships between variables

Age versus resting blood pressure

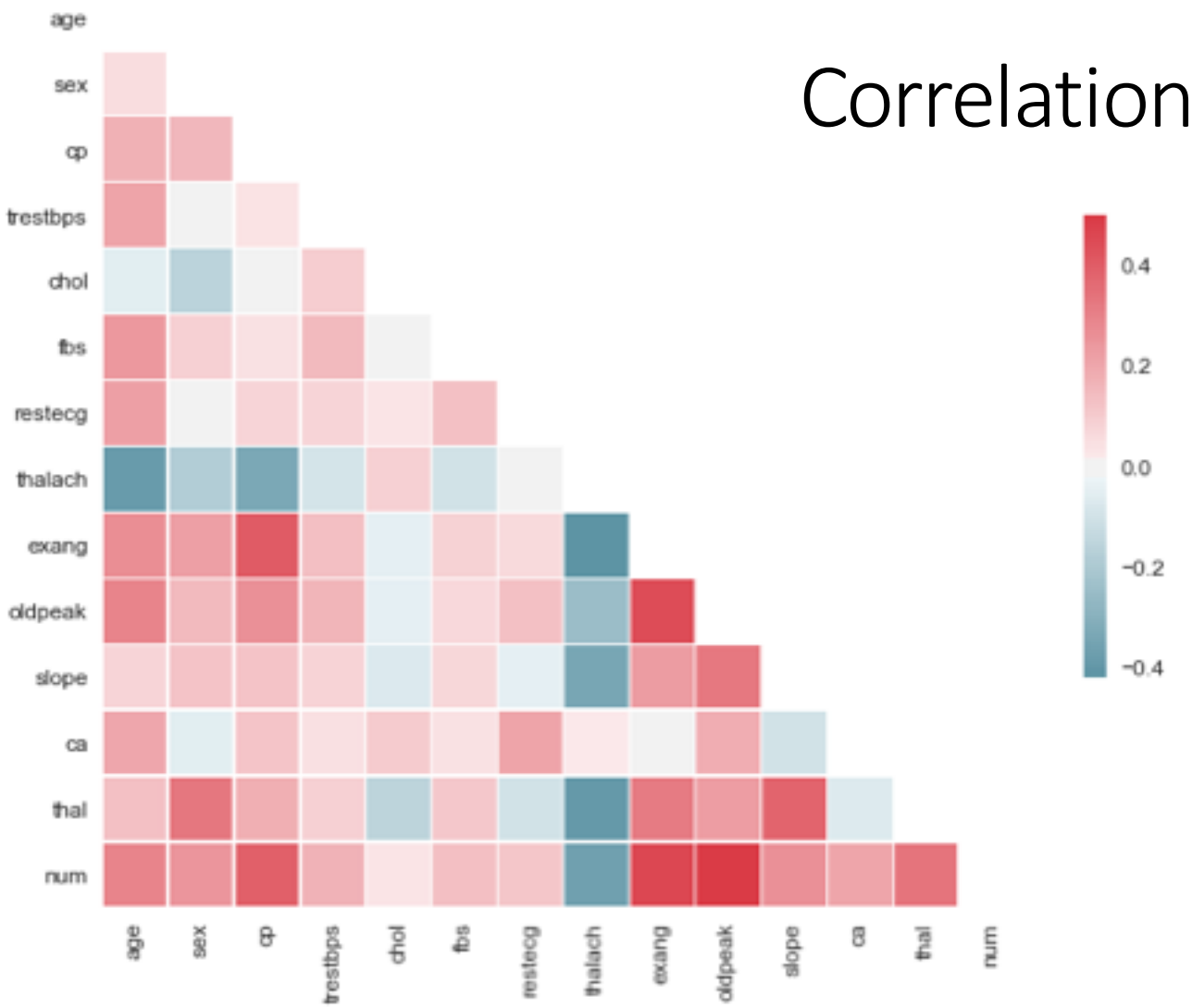
Exercise-induced ST depression versus resting blood pressure

Some relationships between variables

Cholesterol level versus resting blood pressure

Cholesterol level versus exercise-induced ST depression

Correlations between features



Derived from Pearson's correlation coefficients, these may be most accurate for continuous variables or categorical variables that are linearly arranged, but for all variables this heatmap can provided a quick guide to possible correlations.

Machine learning model development

- Which supervised classification model works best, and with regard to the following points?
- Predictive success with the multiclass target variable (severity of heart disease) versus binary target variable (presence vs. absence of heart disease)
- Full dataset with more imputed data points versus the Cleveland subset with relatively fewer imputed data points
- Both global and individual feature importance

Classifier	Classification	Train accuracy	Test accuracy
RF	full	0.977	0.540
Tuned RF	full	0.595	0.540
MNB	full	0.511	0.440
GNB	full	0.603	0.455
LogReg	full	0.610	0.530
RF	binary	0.985	0.770
Tuned RF	binary	0.972	0.800
MNB	binary	0.760	0.735
GNB	binary	0.809	0.800
LogReg	binary	0.819	0.810
SVM	full	0.990	0.510
LSVM	full	0.556	0.525
SVM	binary	0.995	0.545
LSVM	binary	0.791	0.760
DT	full	1.000	0.505
DT	binary	1.000	0.735
GB	full	0.782	0.540
GB	binary	0.878	0.775
SGD	full	0.501	0.455
SGD	binary	0.496	0.485
ET	full	1.000	0.570
ET	binary	1.000	0.790
XGB	full	0.889	0.511
XGB	binary	0.916	0.784
NN	full	0.129	0.130
NN	binary	0.682	0.640

Initial model comparison

- Training and test accuracy for multiclass (full) and binary target variable predictions
- Tests include:
  - Random forest (RF)
  - Multinomial naïve bayes (MNB)
  - Gaussian naïve bayes (GNB)
  - Logistic regression (LogReg)
  - Support vector machine (SVM)
  - Linear SVM (LSVM)
  - Decision tree (DT)
  - SGD Classifier (SGD)
  - Extra trees classifier (ET)
  - XGBoost (XGB)
  - Neural network (NN; MLP)
- Here, random forest is also compared after tuning with GridSearchCV

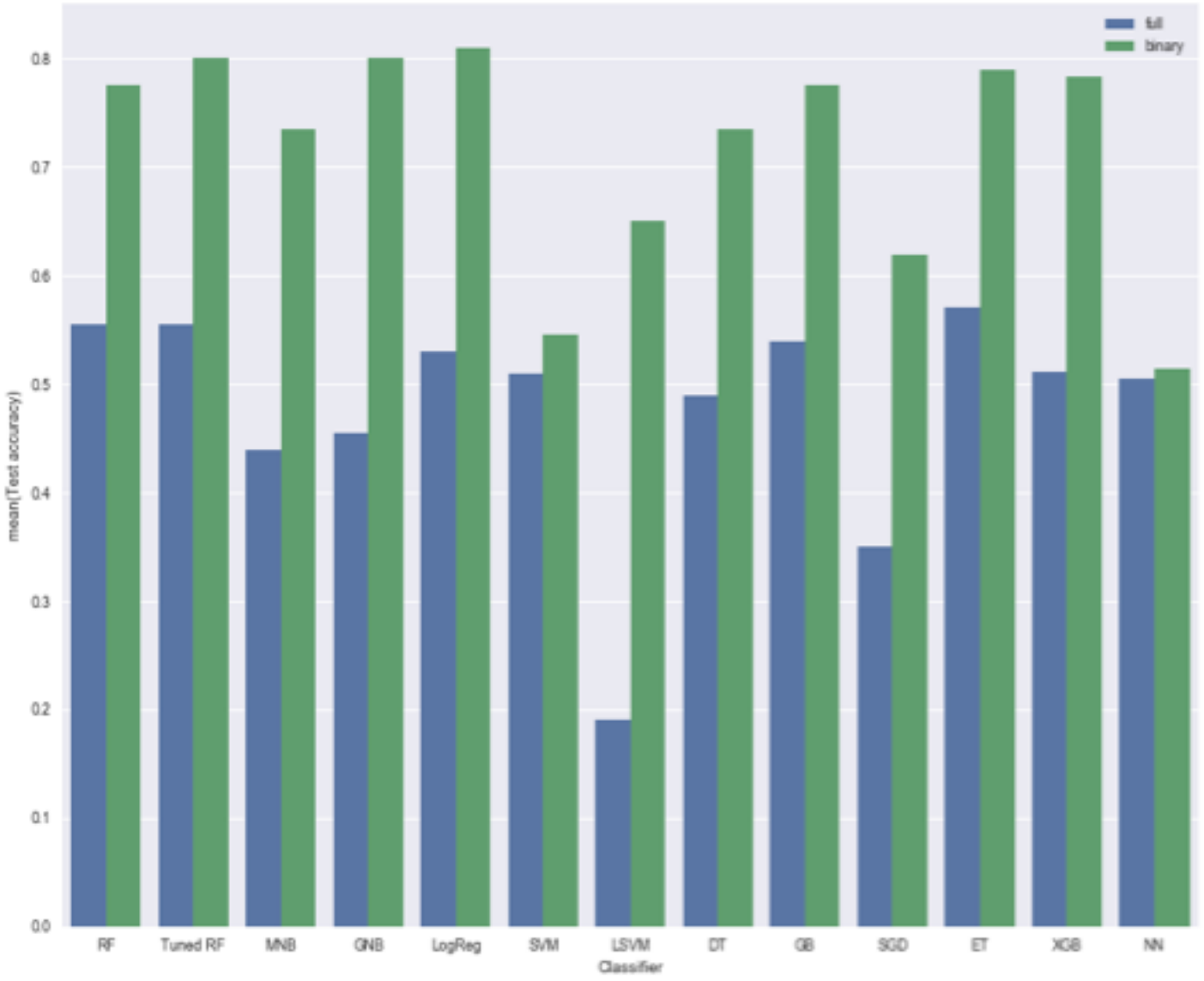
Tested on the full dataset

Classifier	Classification	Train accuracy	Test accuracy
RF	full	0.987	0.526
Tuned RF	full	0.648	0.526
MNB	full	0.586	0.526
GNB	full	0.678	0.513
LogReg	full	0.670	0.539
RF	binary	0.996	0.789
Tuned RF	binary	0.859	0.868
MNB	binary	0.758	0.803
GNB	binary	0.846	0.868
LogReg	binary	0.846	0.868
SVM	full	1.000	0.434
LSVM	full	0.678	0.539
SVM	binary	1.000	0.434
LSVM	binary	0.423	0.566
DT	full	1.000	0.500
DT	binary	1.000	0.789
GB	full	0.982	0.500
GB	binary	0.881	0.842
SGD	full	0.581	0.461
SGD	binary	0.445	0.579
ET	full	1.000	0.513
ET	binary	1.000	0.829
XGB	full	0.995	0.510
XGB	binary	0.975	0.860
NN	full	0.577	0.434

Initial model comparison

- Training and test accuracy for multiclass (full) and binary target variable predictions
- A binary target variable is much easier for all models to predict with these datasets.
- The Cleveland subset imparts a slightly higher test accuracy than does the full dataset, which contains more imputed data, but in the real world this may often occur.
- Many models show fairly similar results, while some clearly do not perform well with these datasets.

Tested on the Cleveland subset



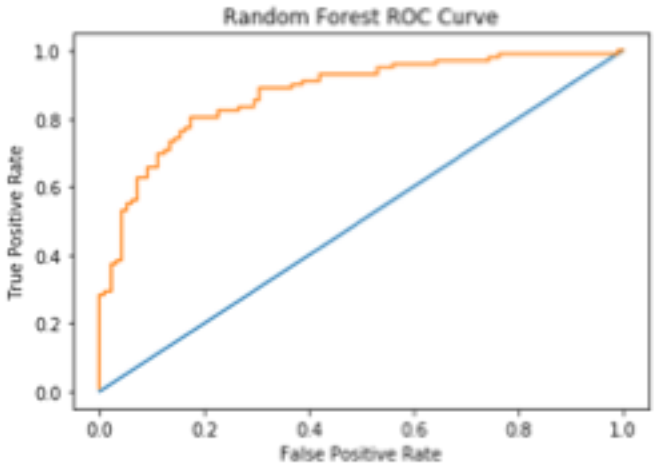
Model comparison for full dataset

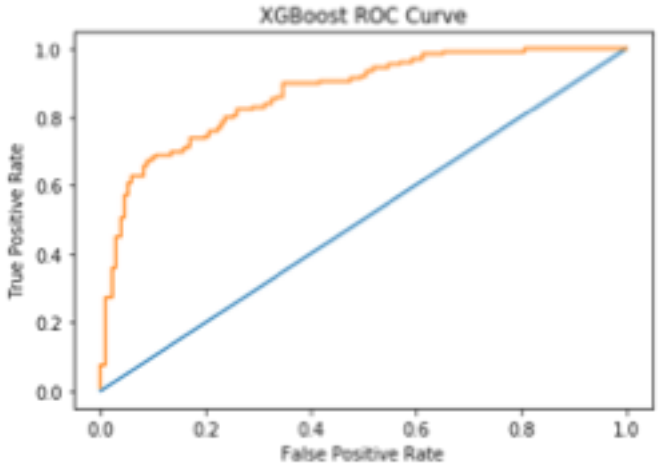
Promising models with tuning of some key hyperparameters

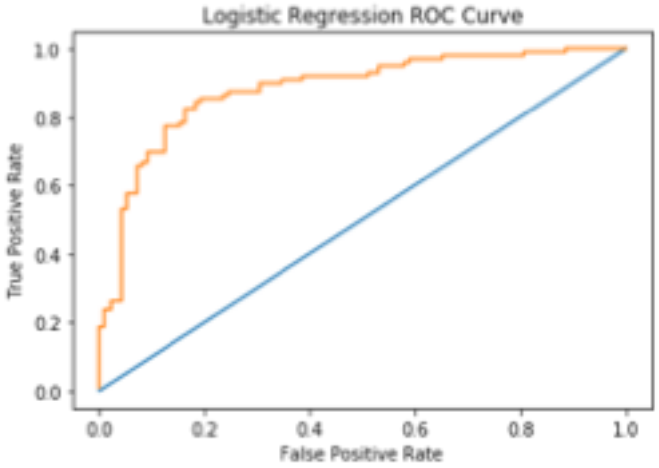
Logistic regression, XGBoost, and random forest classifiers usually outperform others in this analysis of the full dataset (training data), in many cases regardless of exact tuning.

	estimator	min_score	mean_score	max_score	std_score	C	gamma	kernel	learning_rate	max_depth	min_child_weight	n_estimators
30	LogisticRegression	0.660194	0.754902	0.815047	0.0678085	0.01	NaN	NaN	NaN	NaN	NaN	NaN
31	LogisticRegression	0.707547	0.751104	0.795031	0.0357161	0.1	NaN	NaN	NaN	NaN	NaN	NaN
16	XGBClassifier	0.696667	0.749507	0.825083	0.0544925	NaN	NaN	NaN	NaN	NaN	NaN	32
15	XGBClassifier	0.704301	0.748939	0.812903	0.0483948	NaN	NaN	NaN	NaN	NaN	NaN	16
3	RandomForestClassifier	0.696187	0.739537	0.785479	0.0354496	NaN	NaN	NaN	NaN	NaN	NaN	16
21	XGBClassifier	0.691943	0.739052	0.809365	0.0508675	NaN	NaN	NaN	NaN	NaN	10	NaN
5	RandomForestClassifier	0.698925	0.734797	0.801242	0.0472334	NaN	NaN	NaN	NaN	NaN	NaN	100
23	XGBClassifier	0.68932	0.732093	0.805112	0.0518847	NaN	0.1	NaN	NaN	NaN	NaN	NaN
32	LogisticRegression	0.696187	0.731592	0.790274	0.0415698	1	NaN	NaN	NaN	NaN	NaN	NaN
17	XGBClassifier	0.68932	0.730094	0.802548	0.0513671	NaN	NaN	NaN	NaN	NaN	NaN	100
22	XGBClassifier	0.68932	0.730094	0.802548	0.0513671	NaN	0.01	NaN	NaN	NaN	NaN	NaN
20	XGBClassifier	0.68932	0.730094	0.802548	0.0513671	NaN	NaN	NaN	NaN	NaN	1	NaN
24	SVC	0.686275	0.72612	0.791277	0.0464536	1	NaN	linear	NaN	NaN	NaN	NaN
25	SVC	0.686275	0.72611	0.796722	0.0514253	10	NaN	linear	NaN	NaN	NaN	NaN
33	LogisticRegression	0.686275	0.725133	0.797546	0.051269	10	NaN	NaN	NaN	NaN	NaN	NaN
34	LogisticRegression	0.686275	0.725091	0.797508	0.051251	100	NaN	NaN	NaN	NaN	NaN	NaN
19	XGBClassifier	0.689903	0.724308	0.806452	0.0590957	NaN	NaN	NaN	NaN	20	NaN	NaN
18	XGBClassifier	0.682692	0.721844	0.787879	0.0489616	NaN	NaN	NaN	NaN	4	NaN	NaN
4	RandomForestClassifier	0.679612	0.719634	0.779481	0.0425016	NaN	NaN	NaN	NaN	NaN	NaN	32
6	AdaBoostClassifier	0.646465	0.719022	0.819113	0.0731226	NaN	NaN	NaN	NaN	NaN	NaN	16
29	SVC	0.647059	0.717467	0.781006	0.0503492	10	0.0001	rbf	NaN	NaN	NaN	NaN
8	AdaBoostClassifier	0.646465	0.712921	0.807947	0.0699509	NaN	NaN	NaN	NaN	NaN	NaN	100
7	AdaBoostClassifier	0.646465	0.712486	0.80678	0.0694309	NaN	NaN	NaN	NaN	NaN	NaN	32
14	GradientBoostingClassifier	0.676056	0.710581	0.750789	0.0307832	NaN	NaN	NaN	1	NaN	NaN	100
1	ExtraTreesClassifier	0.679803	0.709073	0.757576	0.0345407	NaN	NaN	NaN	NaN	NaN	NaN	32
2	ExtraTreesClassifier	0.663366	0.704424	0.759678	0.0408948	NaN	NaN	NaN	NaN	NaN	NaN	100
0	ExtraTreesClassifier	0.669951	0.702119	0.742857	0.0303746	NaN	NaN	NaN	NaN	NaN	NaN	16
11	GradientBoostingClassifier	0.669856	0.69825	0.707965	0.0155267	NaN	NaN	NaN	0.8	NaN	NaN	100
12	GradientBoostingClassifier	0.643216	0.682486	0.709801	0.0277935	NaN	NaN	NaN	1	NaN	NaN	16
9	GradientBoostingClassifier	0.666667	0.676416	0.690217	0.0100323	NaN	NaN	NaN	0.8	NaN	NaN	16
10	GradientBoostingClassifier	0.644231	0.66824	0.688172	0.0181691	NaN	NaN	NaN	0.8	NaN	NaN	32
26	SVC	0.553991	0.668138	0.736508	0.0812229	10	0.001	rbf	NaN	NaN	NaN	NaN
27	SVC	0.594059	0.667969	0.723926	0.0545164	1	0.0001	rbf	NaN	NaN	NaN	NaN
13	GradientBoostingClassifier	0.623762	0.662171	0.696081	0.0303845	NaN	NaN	NaN	1	NaN	NaN	32
28	SVC	0.550265	0.648188	0.704819	0.069513	1	0.001	rbf	NaN	NaN	NaN	NaN

ROC curves

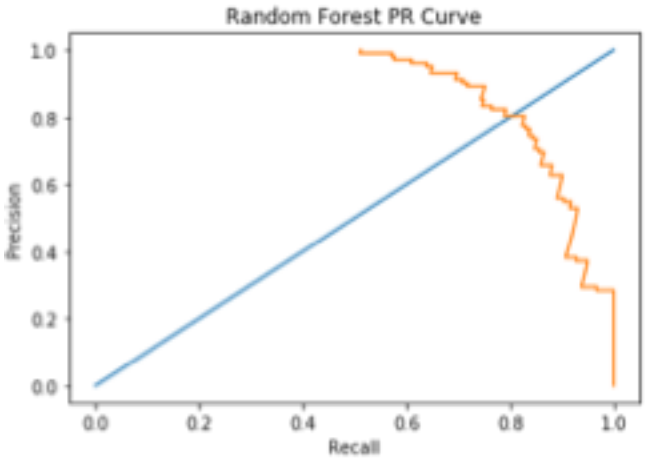


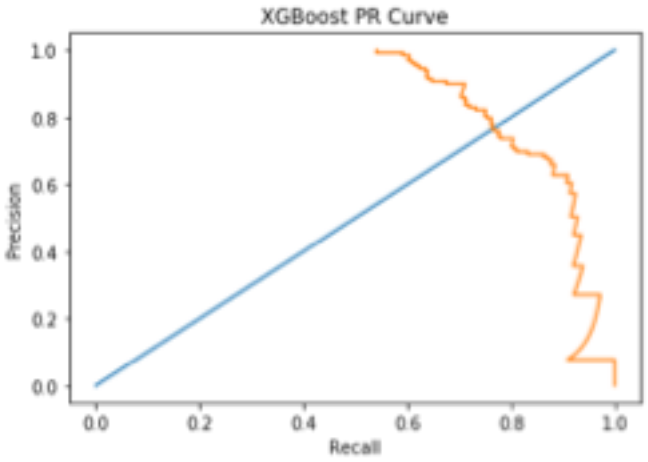


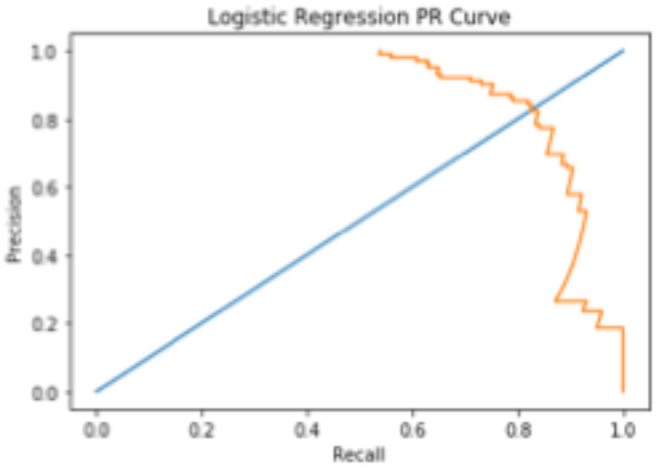


Each of the three most successful model types bears a fairly similar ROC curve, with logistic regression perhaps showing the best true versus false positive rate.

Precision-recall curves

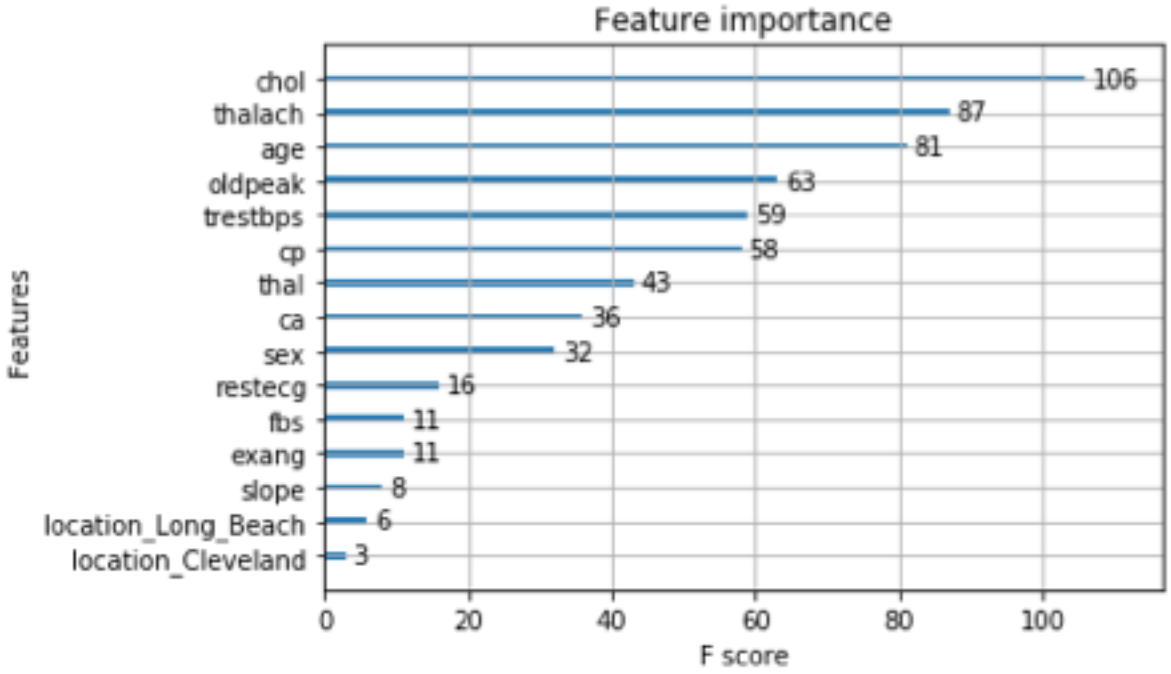






Likewise, each of the three most successful model types bears a fairly similar precision-recall curve, with logistic regression perhaps showing the best precision versus recall relationship.

XGBoost feature importance



Using XGBoost on the full dataset and with the binary target variable, cholesterol is shown to be the most important feature in this comparison. Maximum heart rate and age follow, with other features showing decreasing importance.

Removal of less important features, however, does not necessarily improve the model's accuracy and in fact can reduce accuracy, indicating that each of these features contributes some information that may be useful.

Individual predictive features

- It is helpful to know overall trends in impactful features, but it is potentially quite useful for an individual or care provider to know which features contribute what risk to a patient.
- XGBoost contains its own SHAP-based method of scoring the contributions of each feature for an individual prediction score.
- The SHAP python package also does this and provides visualization tools.

	prediction contributions
age	0.051798
sex	0.226797
cp	-0.845668
trestbps	0.099361
chol	-0.185660
fbs	0.088252
restecg	0.027044
thalach	-0.049714
exang	-0.509166
oldpeak	0.570319
slope	0.075837
ca	-0.266031
thal	-0.145798
Cleveland	0.000000
Hungary	0.002905
Long_Beach	-0.079323
bias	-0.000661
SUM	-0.939708

An individual's feature contributions to the overall predictive score (SUM), here negative for heart disease, derived using XGBoost's "pred_contribs" function.

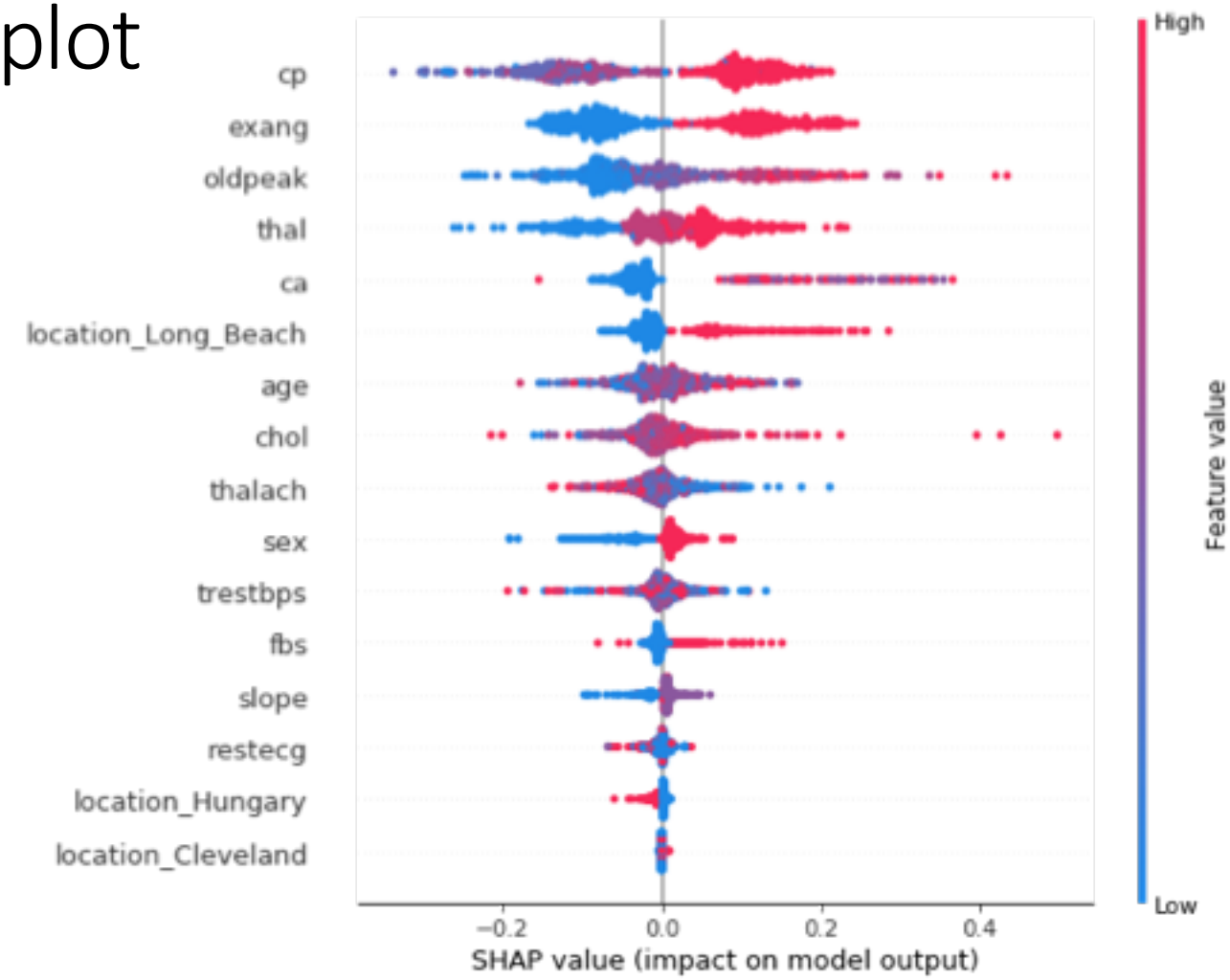
Individual predictive features

Using the SHAP python package, it is possible to examine predictions for each individual and with a quick glance at the features that contribute to the predictive score. In the case shown above, the individual has a low probability of having heart disease, with the features and values highlighted in blue indicating which of the individual's characteristics are contributing most to reducing risk and features and values highlighted in magenta indicating which are contributing most to the individual's very modest, possible risk.

SHAP summary plot

The summary plot to the right is similar to the feature importance plot from before, except that this plot also provides a density component akin to a violin plot, showing where individuals fall along SHAP values for each feature. Also, points in blue represent those lowering the predictive value, with those in magenta imparting greater probability of heart disease for the individual.

Some features appear ranked differently than in the previous feature importance plot. This may relate to different methods of handling extremes in data, but with a SHAP summary plot it is possible to ascertain this information.



Conclusions

- A binary target was much more likely to result in accurate predictions from the full dataset used in this report than a multiclass target was, and imputed data imparted slightly less accuracy than a relatively more complete subset of data (i.e., Cleveland data) could. This is unsurprising.
- Ensemble-based methods and logistic regression provided the greatest accuracy, and all performed fairly similarly.
- SHAP values provide a powerful way to examine both global and individual feature importance, providing a window into how each individual's prediction score is derived, by which features contribute in what ways, potentially providing actionable information.
- Analysis of SHAP values applied to the heart disease dataset allows us to realize certain details, such as that chest pain values below 4 are associated with absence of a heart disease diagnosis in this particular model, while a value of 4 is associated with heart disease.
- SHAP values also allow us to observe where a model may fall short; for instance, the model used in this report provides more complex results with some features in ways that may be unrealistic, such as resting blood pressure and resting electrocardiographic results. This provides insight into areas of possible model improvement.

References

- Batista, D.S. “Hyperparamter 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.