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.

Heart disease rates by location

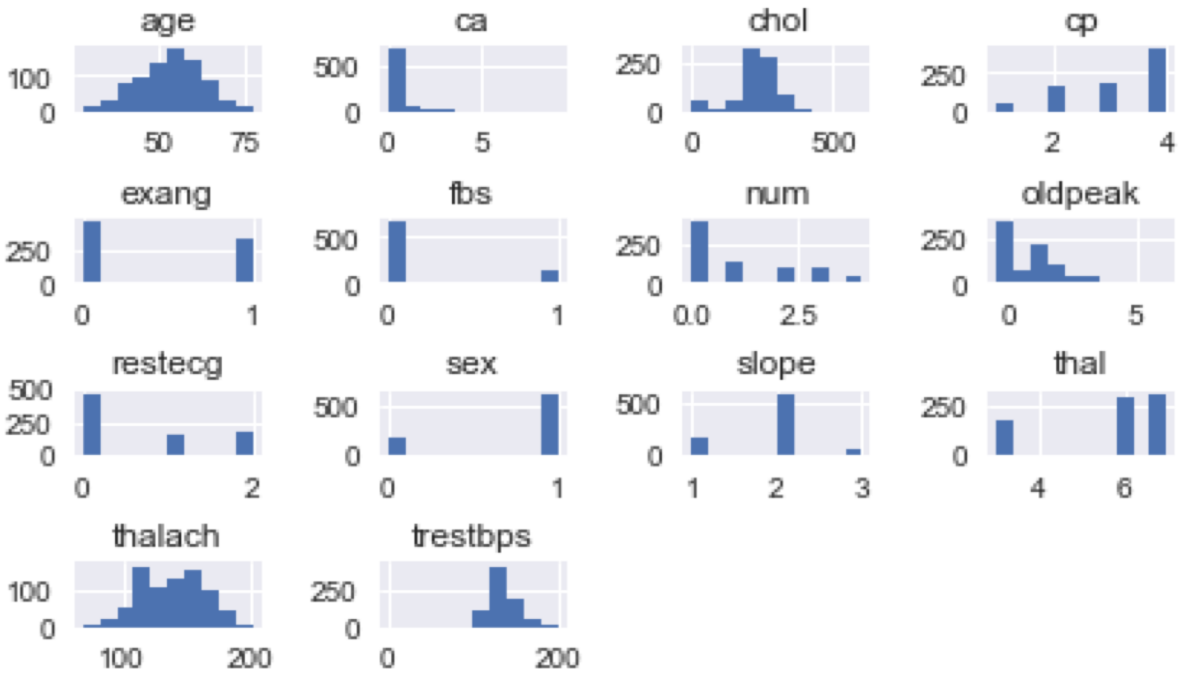
Location	Heart Disease Rate
Ohio	7.62
Cleveland, OH	8.76
California	5.30
Long Beach, CA	5.46

- Heart disease rates (above) 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 the table to the right with their definitions. The target variable is “num”, or “numerical heart disease score”.

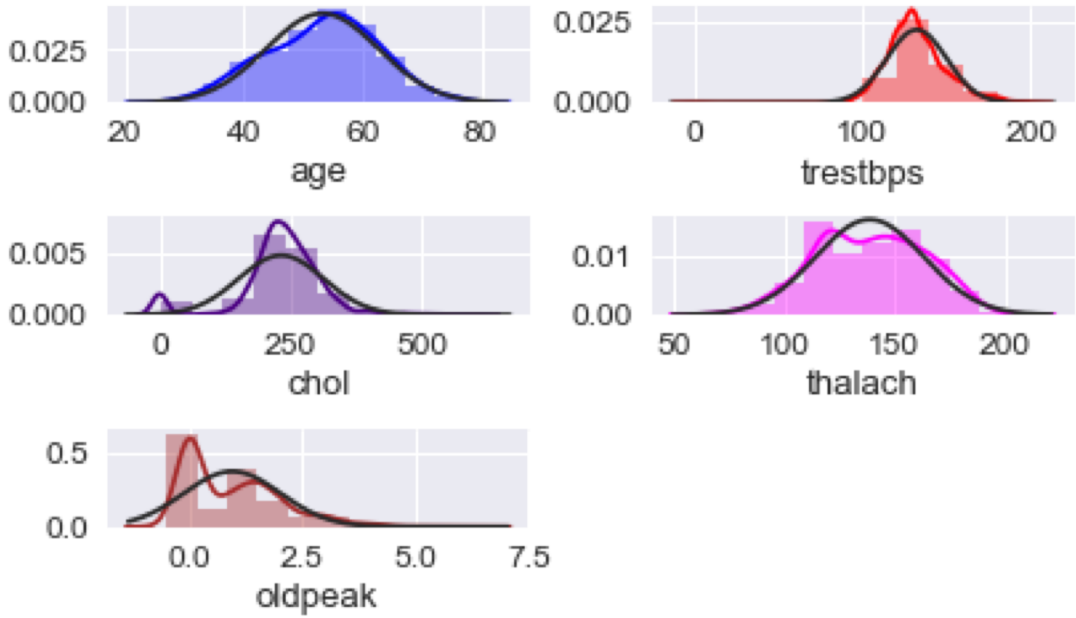
Variable names and definitions

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



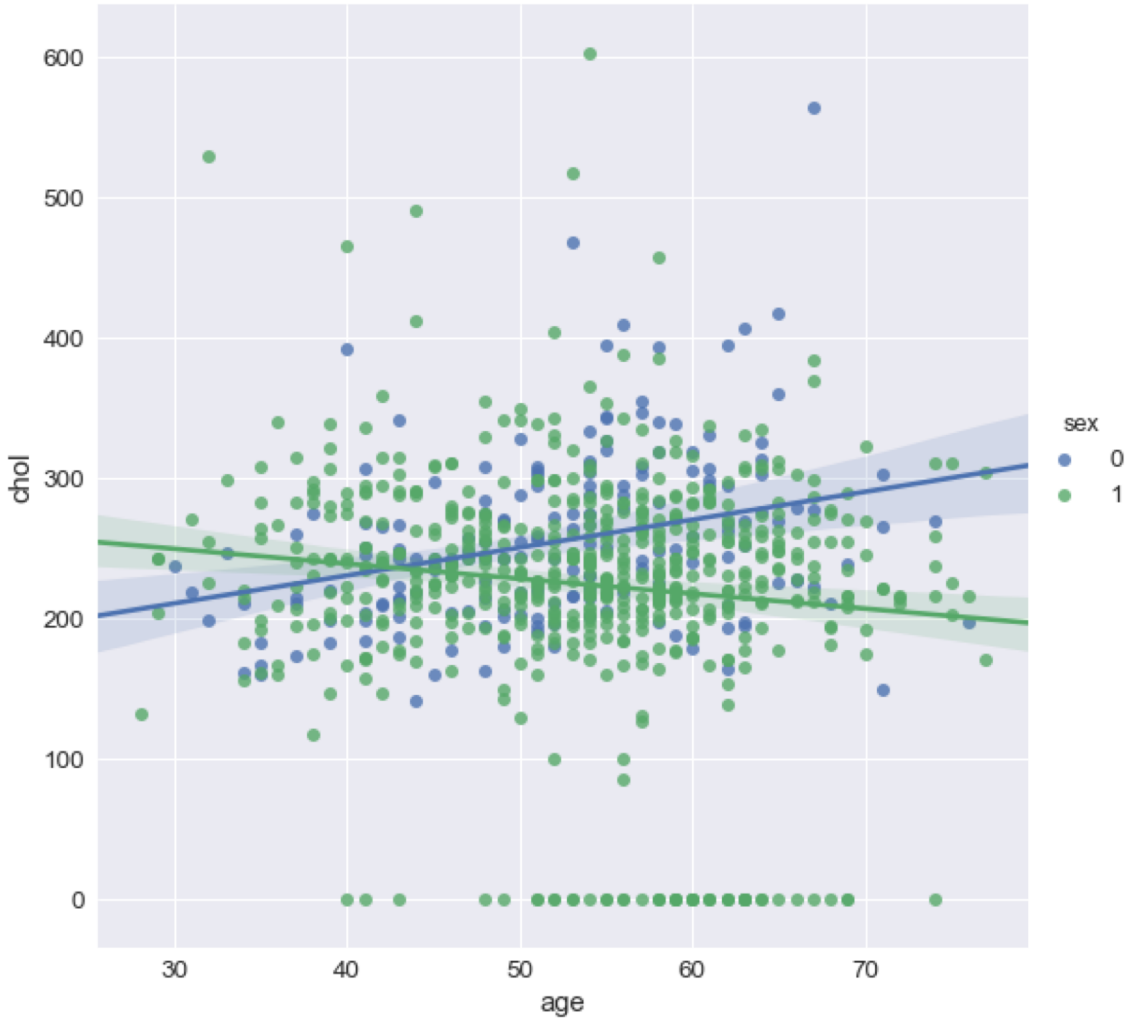
For quick glances at the form of the data for each feature.



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

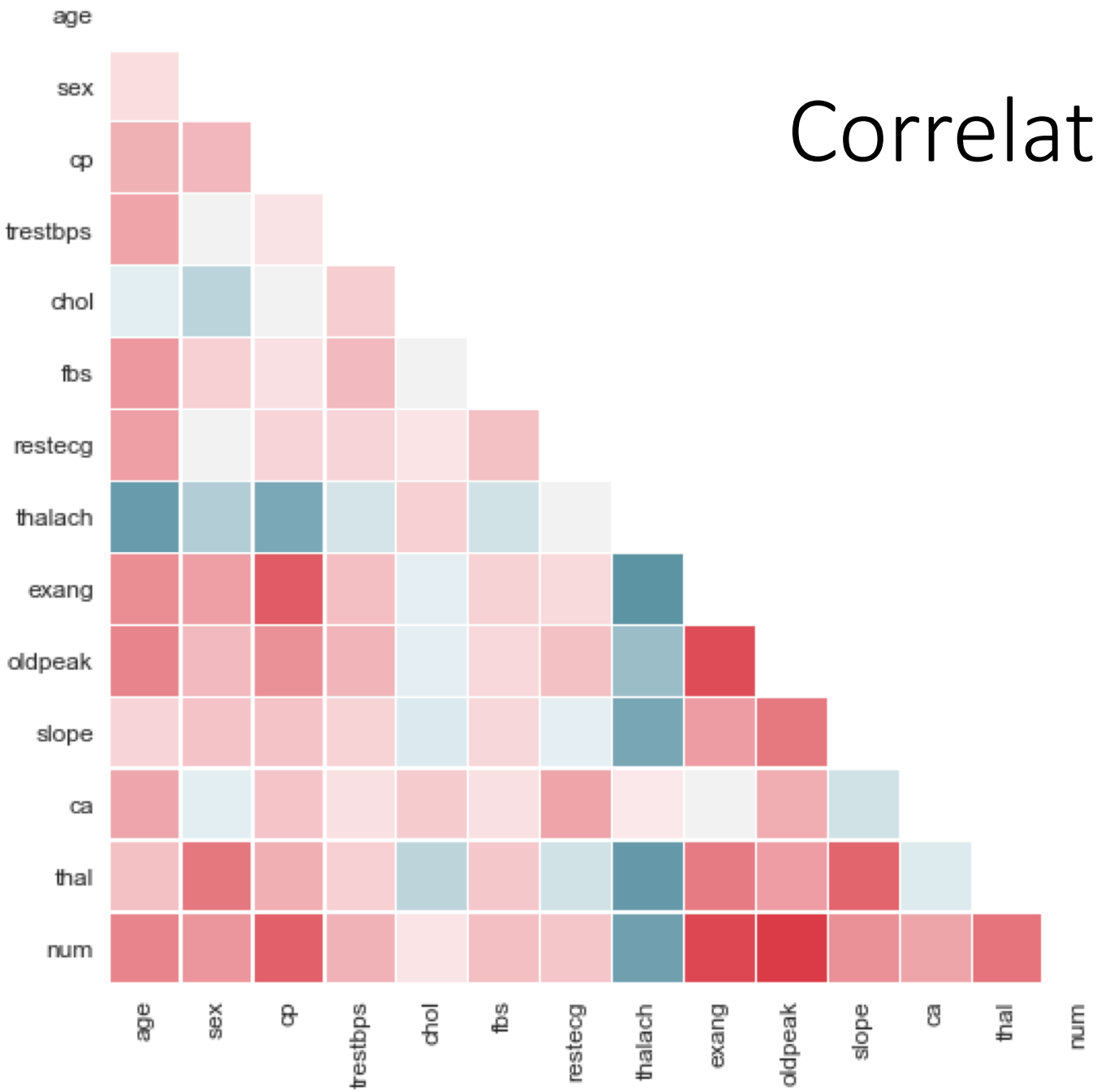
Exercise-induced ST depression versus resting blood pressure

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	0.992	0.760
RF(tuned)	full	0.952	0.760
MNB	full	0.416	0.377
GNB	full	0.490	0.423
LogReg	full	0.494	0.476
RF	binary	0.995	0.790
RF(tuned)	binary	0.836	0.785
MNB	binary	0.760	0.735
GNB	binary	0.809	0.800
LogReg	binary	0.819	0.810
SVM	full	0.997	0.635
LSVM	full	0.439	0.403
SVM	binary	0.995	0.545
LSVM	binary	0.794	0.770
DT	full	1.000	0.677
DT	binary	1.000	0.745
GB	full	0.768	0.601
GB	binary	0.878	0.775
SGD	full	0.201	0.240
SGD	binary	0.637	0.635
ET	full	1.000	0.833
ET	binary	1.000	0.790
XGB	full	0.830	0.680
XGB	binary	0.916	0.784
NN	full	0.368	0.306
NN	binary	0.742	0.675

Initial model comparison

- Training and test accuracy for multiclass (full) and binary target variable predictions
- Tests with the multiclass target used SMOTE-treated data due to imbalanced nature of data in the target classes.
- 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

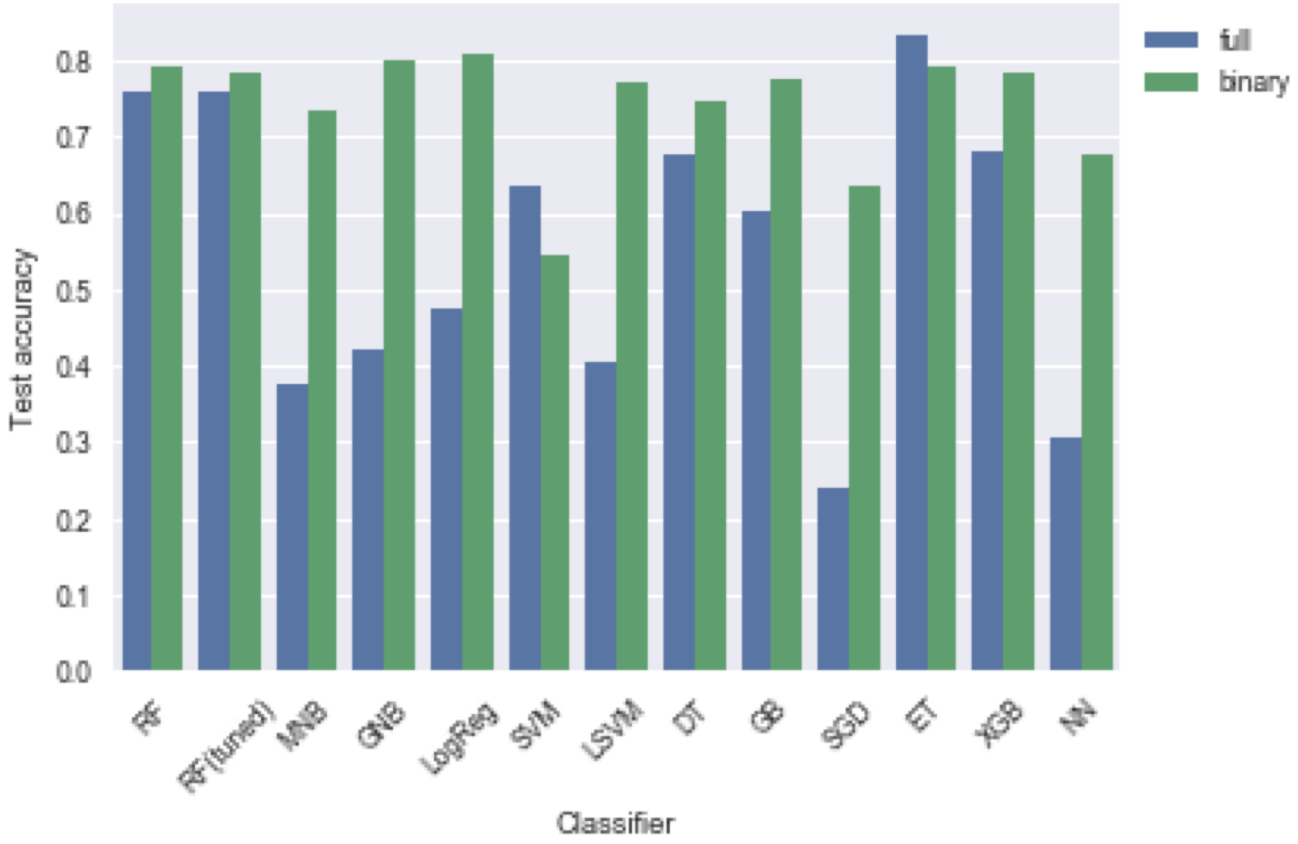
Tested on the full dataset

Classifier	Classification	Train accuracy	Test accuracy
RF	full	0.990	0.780
RF (tuned)	full	0.997	0.780
MNB	full	0.442	0.498
GNB	full	0.595	0.571
LogReg	full	0.571	0.537
RF	binary	0.982	0.816
RF (tuned)	binary	0.855	0.842
MNB	binary	0.758	0.803
GNB	binary	0.846	0.868
LogReg	binary	0.846	0.868
SVM	full	0.998	0.624
LSVM	full	0.511	0.454
SVM	binary	1.000	0.434
LSVM	binary	0.744	0.671
DT	full	1.000	0.707
DT	binary	1.000	0.803
GB	full	0.941	0.795
GB	binary	0.881	0.842
SGD	full	0.293	0.312
SGD	binary	0.577	0.447
ET	full	1.000	0.902
ET	binary	1.000	0.829
XGB	full	0.985	0.797
XGB	binary	0.975	0.860
NN	full	0.228	0.220
NN	binary	0.515	0.579

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 the full dataset

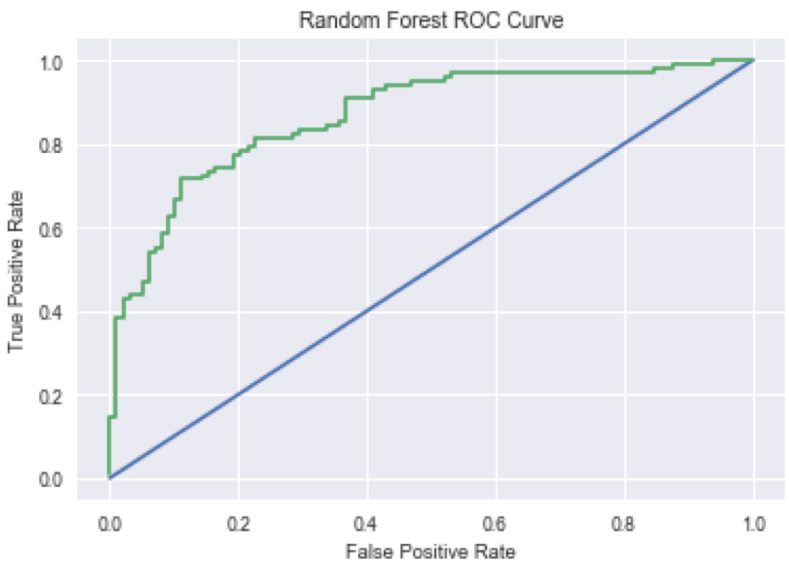
SMOTE-treated multiclass target analyses usually showed substantially greater accuracy than without SMOTE treatment, though for binary target analyses this was not the case (as the binary target arrangement of data was considerably more balanced). Most classifiers still performed better with a binary target, however, though results were similar for some ensemble methods, reaching above 90% for the multiclass target analysis using only the Cleveland subset.

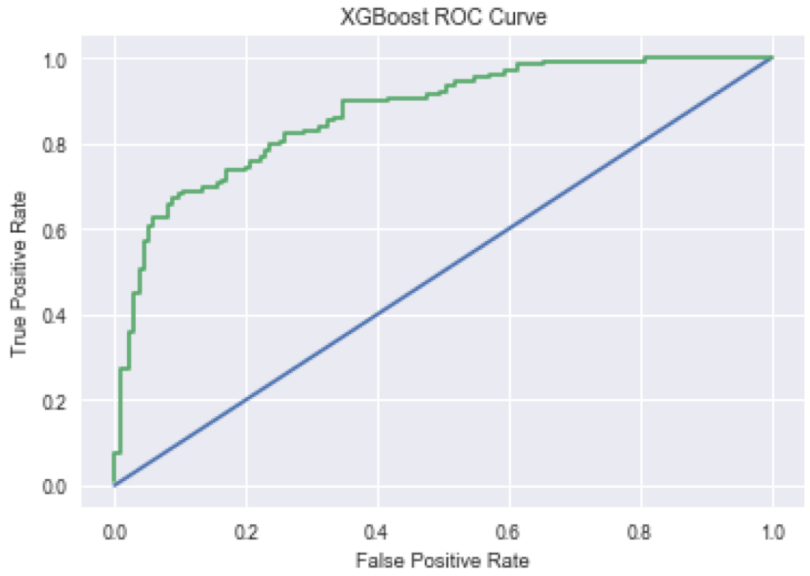
Promising models with tuning of some key hyperparameters, with a binary target

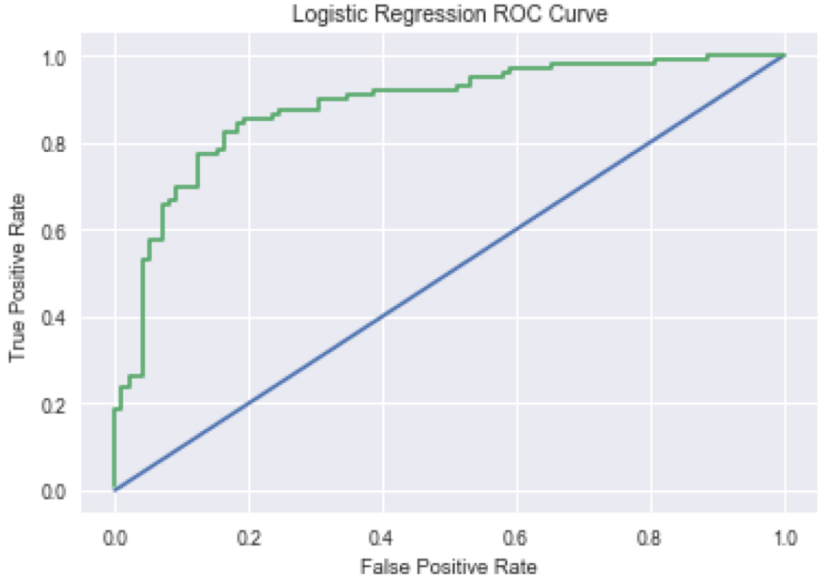
Logistic regression, XGBoost, linear SVC, and random forest classifiers usually outperform others in this analysis of the full dataset (training data), in many cases regardless of exact tuning.

	estimator	min_score	mean_score	max_score	std_score	C	gamma	kernel	learning_rate	max_depth	min_child_weight	n_estimators
31	LogisticRegression	0.77	0.819006	0.862944	0.0381131	0.1	NaN	NaN	NaN	NaN	NaN	NaN
32	LogisticRegression	0.78392	0.811953	0.833333	0.0207142	1	NaN	NaN	NaN	NaN	NaN	NaN
24	SVC	0.792079	0.811939	0.84	0.0204049	1	NaN	linear	NaN	NaN	NaN	NaN
34	LogisticRegression	0.772277	0.80938	0.837209	0.0273073	100	NaN	NaN	NaN	NaN	NaN	NaN
25	SVC	0.790244	0.805592	0.826531	0.0153325	10	NaN	linear	NaN	NaN	NaN	NaN
5	RandomForestClassifier	0.768421	0.80223	0.830189	0.0255536	NaN	NaN	NaN	NaN	NaN	NaN	100
33	LogisticRegression	0.772277	0.801915	0.818653	0.0210155	10	NaN	NaN	NaN	NaN	NaN	NaN
18	XGBClassifier	0.793814	0.801038	0.811321	0.00746698	NaN	NaN	NaN	NaN	4	NaN	NaN
21	XGBClassifier	0.782178	0.800261	0.818605	0.0148722	NaN	NaN	NaN	NaN	NaN	10	NaN
4	RandomForestClassifier	0.783505	0.800227	0.821256	0.0157099	NaN	NaN	NaN	NaN	NaN	NaN	32
16	XGBClassifier	0.780488	0.795499	0.804124	0.010654	NaN	NaN	NaN	NaN	NaN	NaN	32
6	AdaBoostClassifier	0.777202	0.791361	0.810811	0.0142217	NaN	NaN	NaN	NaN	NaN	NaN	16
2	ExtraTreesClassifier	0.746114	0.791001	0.822967	0.0326781	NaN	NaN	NaN	NaN	NaN	NaN	100
15	XGBClassifier	0.77451	0.789805	0.807339	0.0134959	NaN	NaN	NaN	NaN	NaN	NaN	16
23	XGBClassifier	0.778947	0.788506	0.794118	0.00679302	NaN	0.1	NaN	NaN	NaN	NaN	NaN
0	ExtraTreesClassifier	0.759358	0.787852	0.808081	0.0207316	NaN	NaN	NaN	NaN	NaN	NaN	16
1	ExtraTreesClassifier	0.748663	0.787467	0.81	0.0275574	NaN	NaN	NaN	NaN	NaN	NaN	32
22	XGBClassifier	0.770053	0.786766	0.8	0.0124705	NaN	0.01	NaN	NaN	NaN	NaN	NaN
17	XGBClassifier	0.770053	0.78614	0.798122	0.0118207	NaN	NaN	NaN	NaN	NaN	NaN	100
20	XGBClassifier	0.770053	0.78614	0.798122	0.0118207	NaN	NaN	NaN	NaN	NaN	1	NaN
30	LogisticRegression	0.753623	0.785383	0.810256	0.0236277	0.01	NaN	NaN	NaN	NaN	NaN	NaN
19	XGBClassifier	0.773869	0.783472	0.792627	0.00766421	NaN	NaN	NaN	NaN	20	NaN	NaN
7	AdaBoostClassifier	0.761421	0.777093	0.790909	0.0121096	NaN	NaN	NaN	NaN	NaN	NaN	32
8	AdaBoostClassifier	0.743455	0.776775	0.809091	0.0268049	NaN	NaN	NaN	NaN	NaN	NaN	100
3	RandomForestClassifier	0.731183	0.776404	0.8	0.0319864	NaN	NaN	NaN	NaN	NaN	NaN	16
11	GradientBoostingClassifier	0.75	0.773694	0.80829	0.0250145	NaN	NaN	NaN	0.8	NaN	NaN	100
12	GradientBoostingClassifier	0.76555	0.770937	0.779487	0.00611383	NaN	NaN	NaN	1	NaN	NaN	16
13	GradientBoostingClassifier	0.754717	0.767098	0.782178	0.0113724	NaN	NaN	NaN	1	NaN	NaN	32
14	GradientBoostingClassifier	0.748718	0.761462	0.782178	0.0147778	NaN	NaN	NaN	1	NaN	NaN	100
10	GradientBoostingClassifier	0.726316	0.756427	0.808081	0.036692	NaN	NaN	NaN	0.8	NaN	NaN	32
9	GradientBoostingClassifier	0.715026	0.751115	0.80402	0.038228	NaN	NaN	NaN	0.8	NaN	NaN	16
29	SVC	0.720379	0.742737	0.761421	0.0169556	10	0.0001	rbf	NaN	NaN	NaN	NaN
27	SVC	0.669903	0.69612	0.742268	0.0327321	1	0.0001	rbf	NaN	NaN	NaN	NaN
26	SVC	0.669903	0.677804	0.687179	0.00712984	1	0.001	rbf	NaN	NaN	NaN	NaN
28	SVC	0.639175	0.665919	0.688679	0.0204053	10	0.001	rbf	NaN	NaN	NaN	NaN

ROC curves

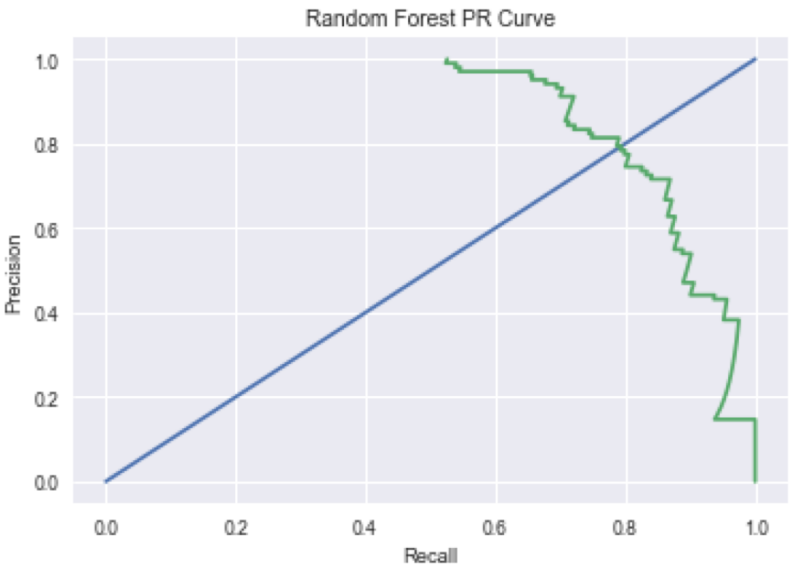


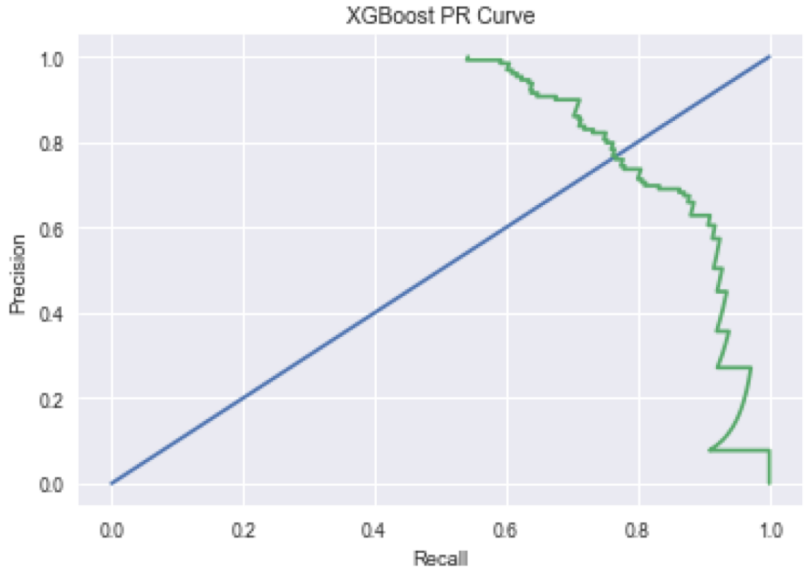




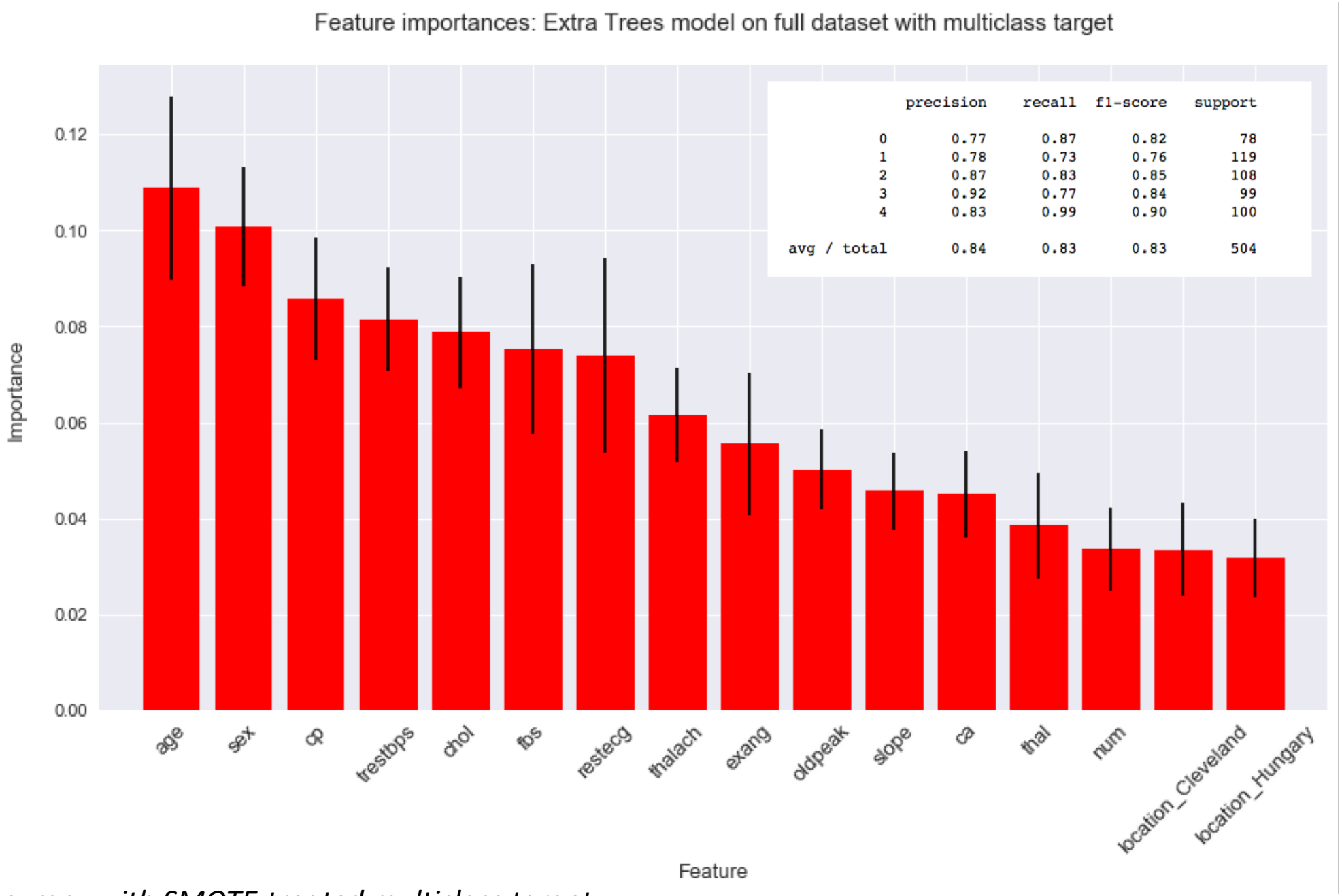
Each of the three most successful model types bears a fairly similar ROC curve, with logistic regression perhaps showing the best true versus false positive rate.

Precision-recall curves



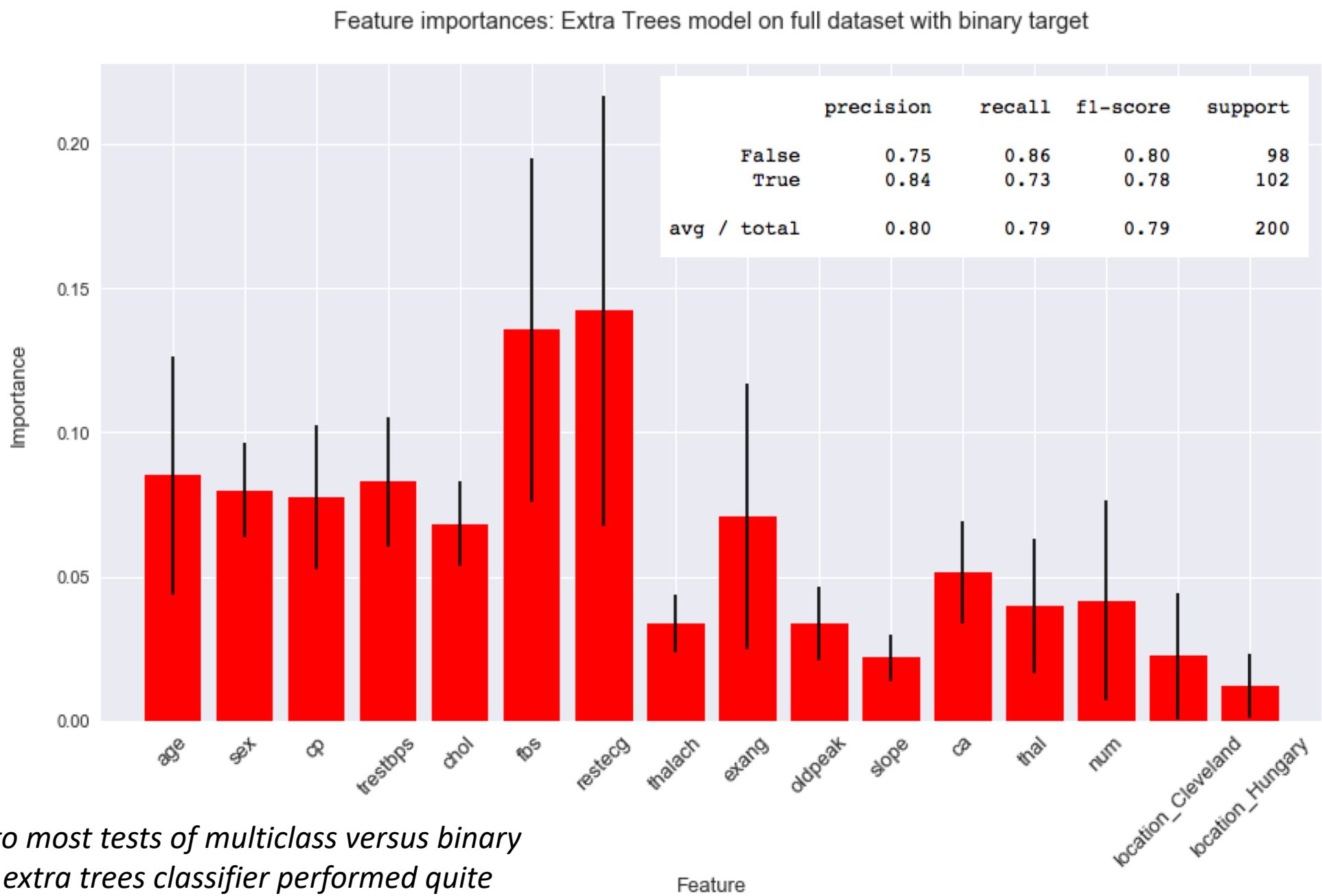


Likewise, each of the three most successful model types bears a fairly similar precision-recall curve, with logistic regression perhaps showing the best precision versus recall relationship.



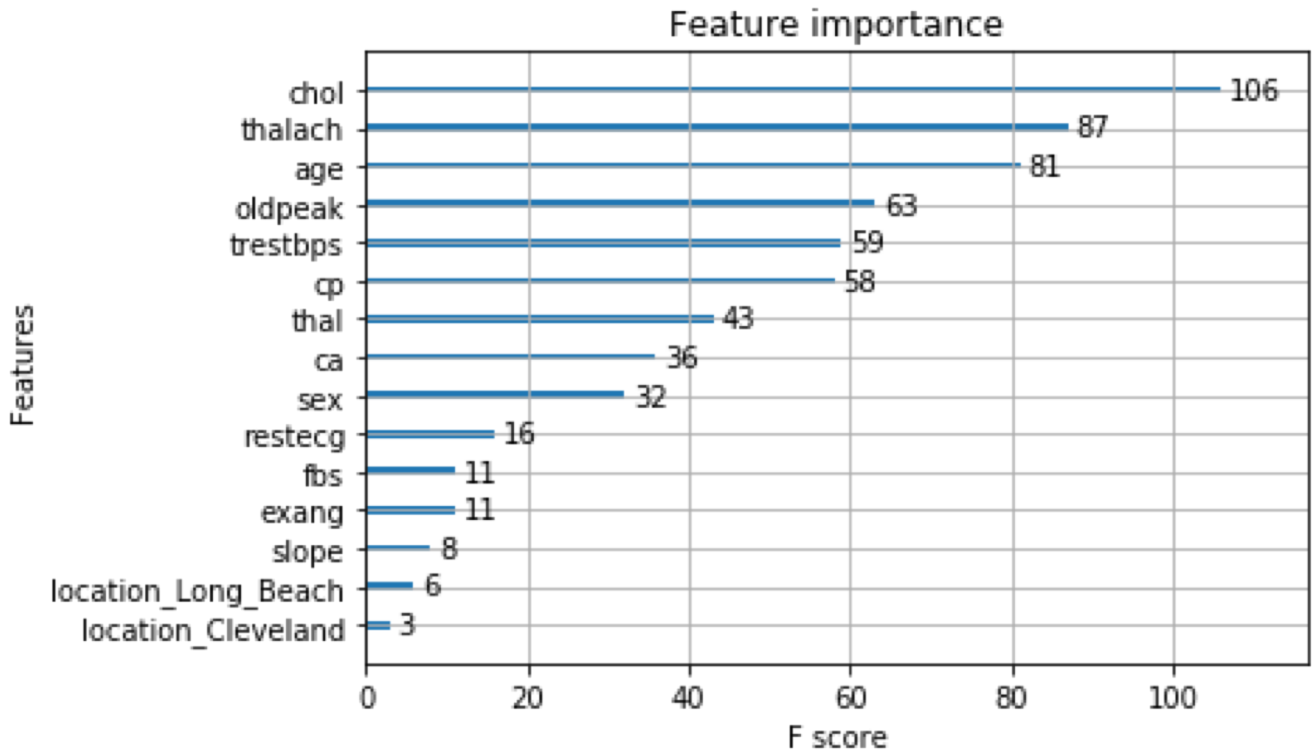
	precision	recall	f1-score	support
0	0.77	0.87	0.82	78
1	0.78	0.73	0.76	119
2	0.87	0.83	0.85	108
3	0.92	0.77	0.84	99
4	0.83	0.99	0.90	100
avg / total	0.84	0.83	0.83	504

Fairly high accuracy with SMOTE-treated multiclass target data using the extra trees classifier.



In contrast to most tests of multiclass versus binary targets, the extra trees classifier performed quite comparably between the SMOTE-treated multiclass target and the binary target.

XGBoost feature importance (binary target)



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.

Interpretations of feature importance

- The extra trees classifier models using the multiclass and binary targets differ in what is considered important.
  - Multiclass: on average, age and sex appear most important and contribute slightly >20% of a person's score for heart disease.
  - In the binary target analysis, fasting blood sugar and resting electrocardiogram results appear more important, though error bars are large.
  - Both sets of results contrast with XGBoost feature importances determined from the binary target analysis in which cholesterol stands out as important, followed by maximum heart rate and age.
  - Accuracy of these models is limited enough that it is difficult to determine if different feature importances for multiclass and binary target analyses are meaningful.
- Further analyses with improved model accuracy could explore the differences in feature importance between target variable types:
  - Do some features make heart disease more severe, beyond just contributing to the presence of heart disease?

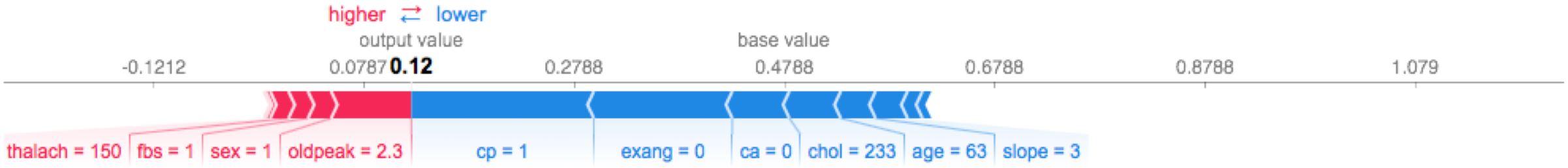
Individual predictive features

- 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.

An individual's feature contributions to the overall predictive score (SUM), here negative for heart disease, derived from this analysis with a binary target using XGBoost's "pred_contribs" function.

	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

Individual predictive features



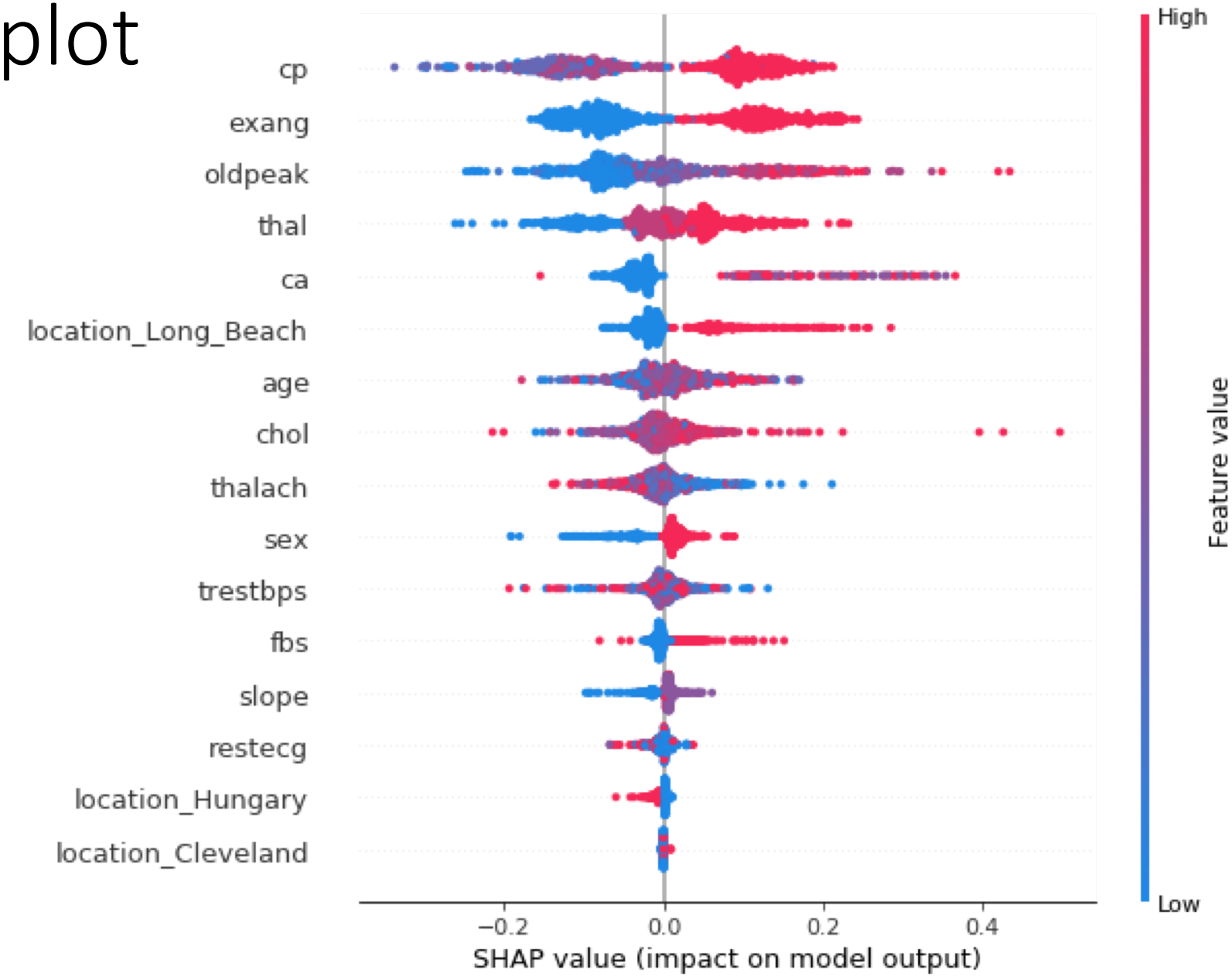
Using the SHAP python package, it is possible to examine predictions for each individual and with a quick glance at the features that contribute to the predictive score. In the case shown above, the individual has a low probability of having heart disease, with the features and values highlighted in 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.

According to the model used here, the individual whose data are shown here is at low risk of a positive heart disease diagnosis due, in part, to his low degree of chest pain, lack of exercise-induced angina, and low coronary angiography score. Potentially actionable features this individual may need to pay attention to are his cardiac metric ST depression with exercise relative to rest and high fasting blood sugar.

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. “Hyperparameter 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.