Predicting Heart Disease

STEPHEN LANIER

PROBLEM STATEMENT

- 13 predictors
- Recall vs precision
- Combining models

BUSINESS VALUE

"Heart disease is the **leading cause of death** for men, women, and people of most racial and ethnic groups in the United States."

"Heart disease costs the United States about **$219 billion** each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death."

Citation: CDC

METHODOLOGY

01
Refine data

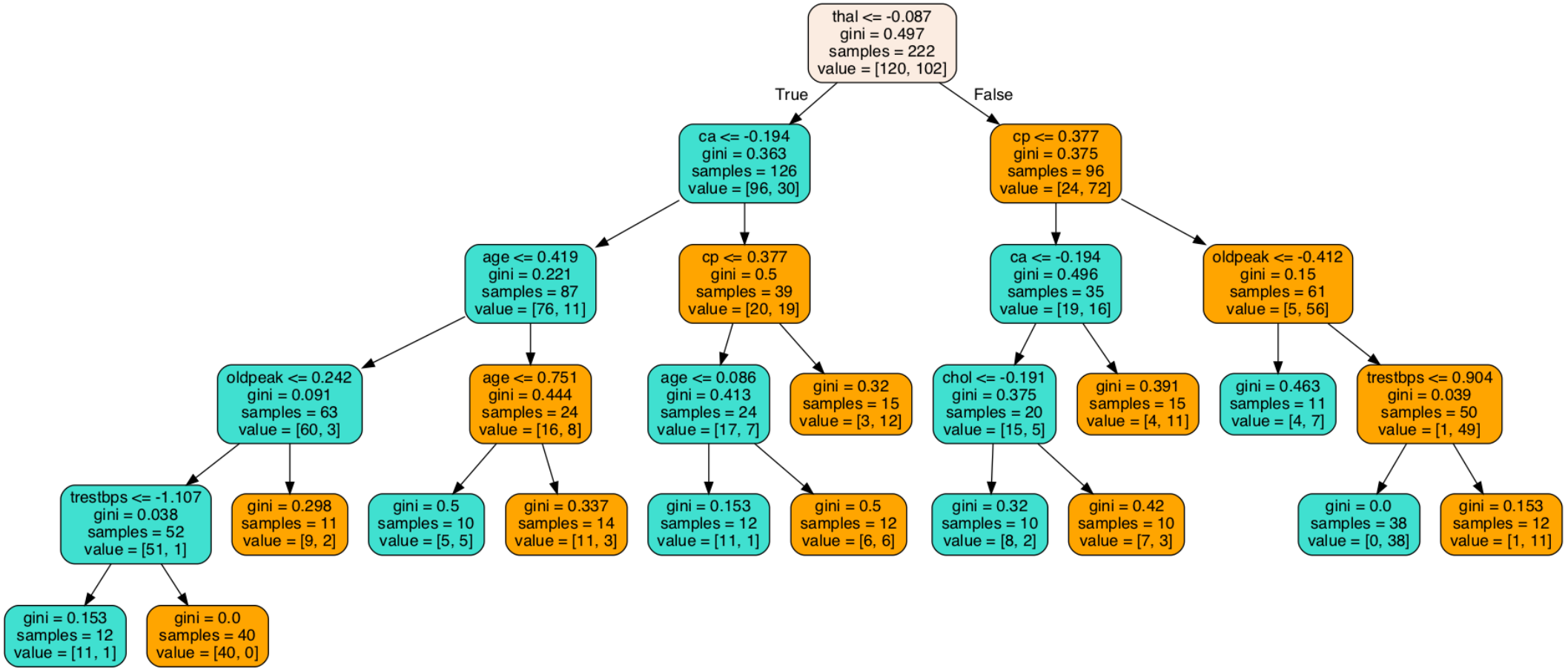
02
Tune models

03
Combine strongest models

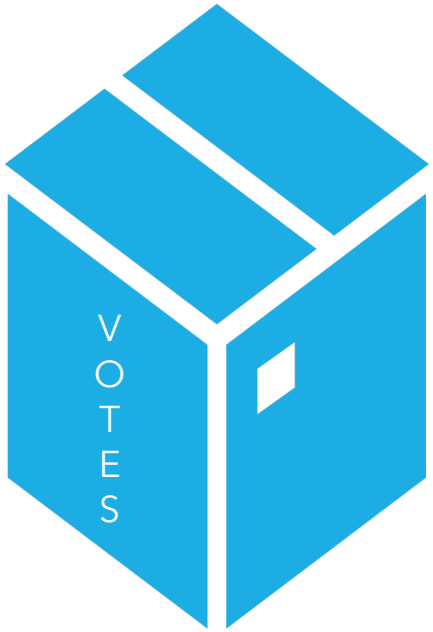
FINDINGS

1. **7** important predictors
2. Best single models: **88%** accuracy
3. Combined methods: **89%** accuracy

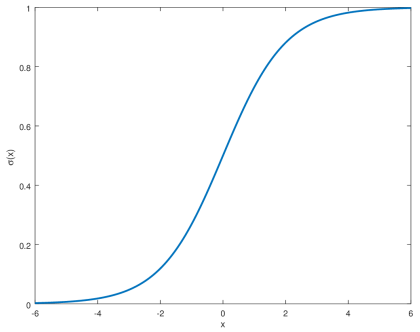
Decision Tree, 87% Accuracy



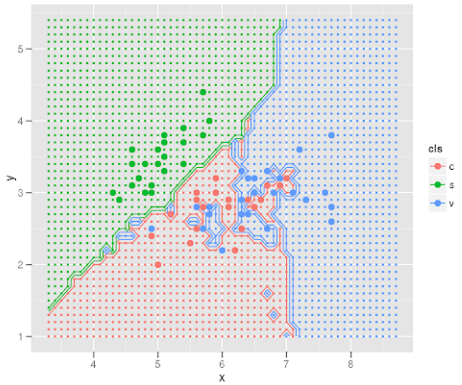
Voting Ensemble, 89% Accuracy



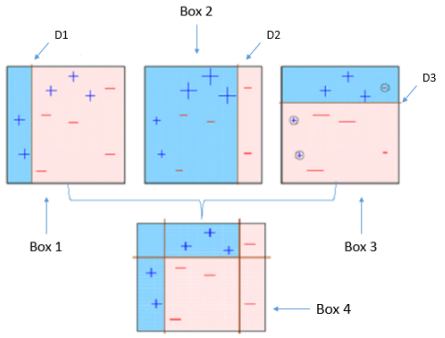
Logistic Regression



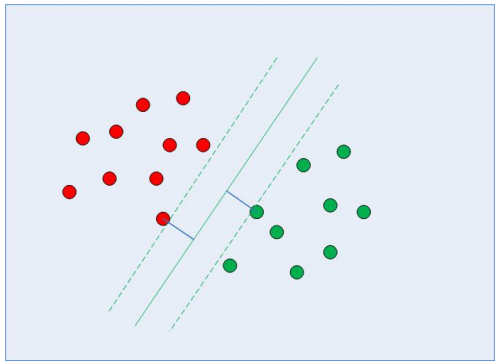
K-Nearest Neighbors



AdaBoost



Support Vector Machine



CONCLUSIONS

Able to predict heart disease from 7 measurements with ~90% accuracy.

Future Work

- Change metric from accuracy to recall for healthcare setting
- Develop more sophisticated models: clustering, perceptrons, neural networks

ACKNOWLEDGEMENTS

Kaggle for data associated with Heart Disease Ensemble Classifiers

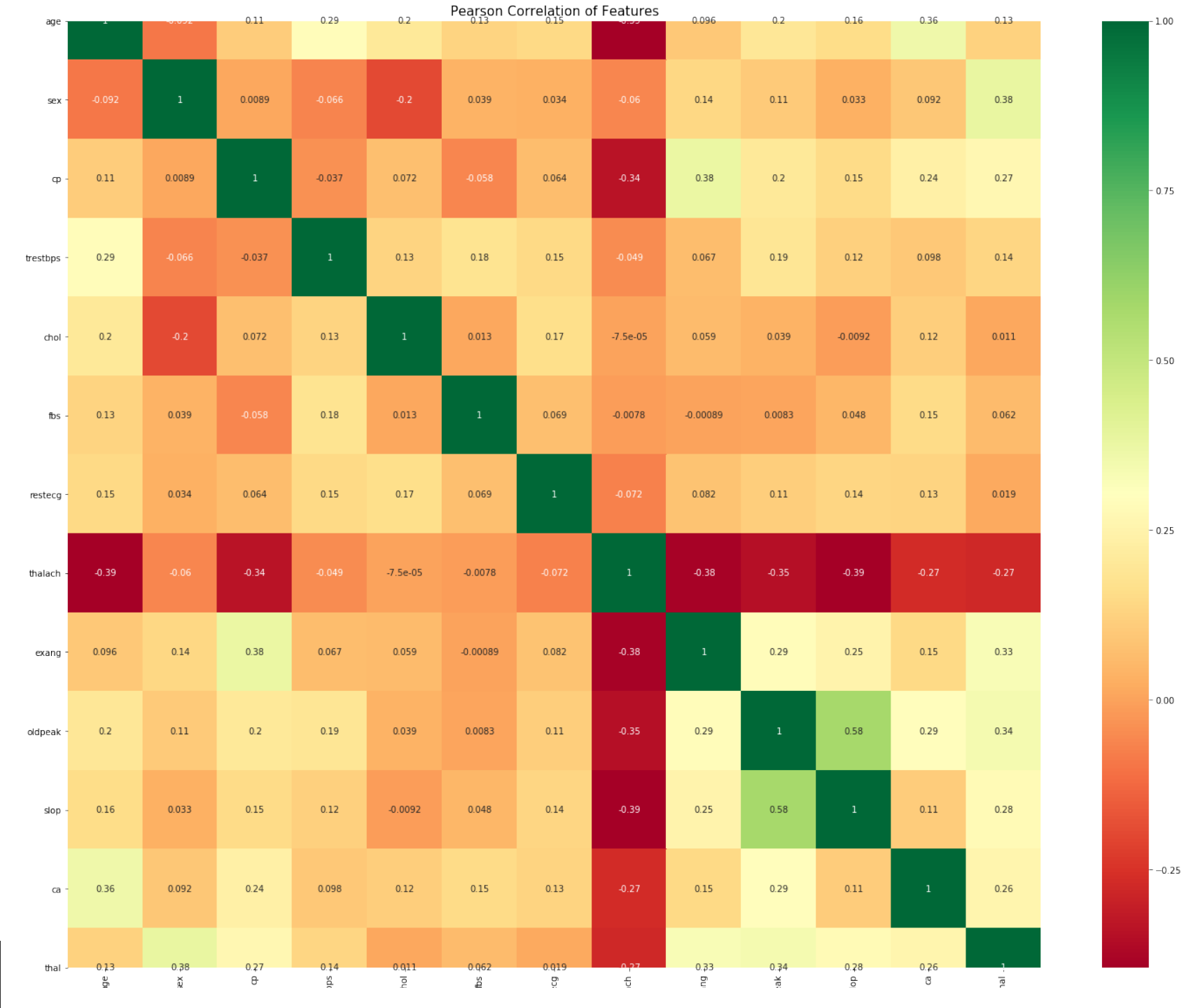
Thank you!

Images Used

- logistic regression: https://hvidberrrg.github.io/deep_learning/activation_functions/sigmoid_function_and_derivative.html
- KNN: http://dylanwiwad.com/project/predicting_car_prices/
- SVM: https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/
- AdaBoost: https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe

Appendix 1

thalach	0.128969
cp	0.122558
thal	0.112145
ca	0.106617
oldpeak	0.104328
age	0.091347
chol	0.083361
trestbps	0.080489
exang	0.064376
slop	0.046107
sex	0.027416
restecg	0.024662
fbs	0.007625


Pearson Correlation of Features
age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slop
ca
thal
1
0.11
0.29
0.2
0.13
0.15
0.096
0.2
0.16
0.36
0.13
-0.092
1
0.0089
-0.066
-0.2
0.039
0.034
-0.06
0.14
0.11
0.033
0.092
0.38
0.11
0.0089
1
-0.037
0.072
-0.058
0.064
-0.34
0.38
0.2
0.15
0.24
0.27
0.29
-0.066
-0.037
1
0.13
0.18
0.15
-0.049
0.067
0.19
0.12
0.098
0.14
0.2
-0.2
0.072
0.13
1
0.013
0.17
-7.5e-05
0.059
0.039
-0.0092
0.12
0.011
0.13
0.039
-0.058
0.18
0.013
1
0.069
-0.0078
-0.00089
0.0083
0.048
0.15
0.062
0.15
0.034
0.064
0.15
0.17
0.069
1
-0.072
0.082
0.11
0.14
0.13
0.019
-0.39
-0.06
-0.34
-0.049
-7.5e-05
-0.0078
-0.072
1
-0.38
-0.35
-0.39
-0.27
-0.27
0.096
0.14
0.38
0.067
0.059
-0.00089
0.082
-0.38
1
0.29
0.25
0.15
0.33
0.2
0.11
0.2
0.19
0.039
0.0083
0.11
-0.35
0.29
1
0.58
0.29
0.34
0.16
0.033
0.15
0.12
-0.0092
0.048
0.14
-0.39
0.25
0.58
1
0.11
0.28
0.36
0.092
0.24
0.098
0.12
0.15
0.13
-0.27
0.15
0.29
0.11
1
0.26
0.13
0.38
0.27
0.14
0.011
0.062
0.019
0.33
0.34
0.28
0.26
1
1.00
0.75
0.50
0.25
0.00
−0.25

Appendix 2

Model	Initial Test Accuracy	Final Test Accuracy
Ensemble Classifier (Soft, Unprocessed)	--	**89%**
Sigmoid SVC	55%	**88%**
AdaBoost	72%	**87%**
Decision Tree	77%	**87%**
Ensemble Classifier (Soft)	--	**87%**
Ensemble Classifier (Hard)	--	**87%**
Logistic Regression	87%	**85%**
Random Forest	83%	**85%**
XGBoost	81%	**84%**
KNN	59%	**82%**

Appendix 3