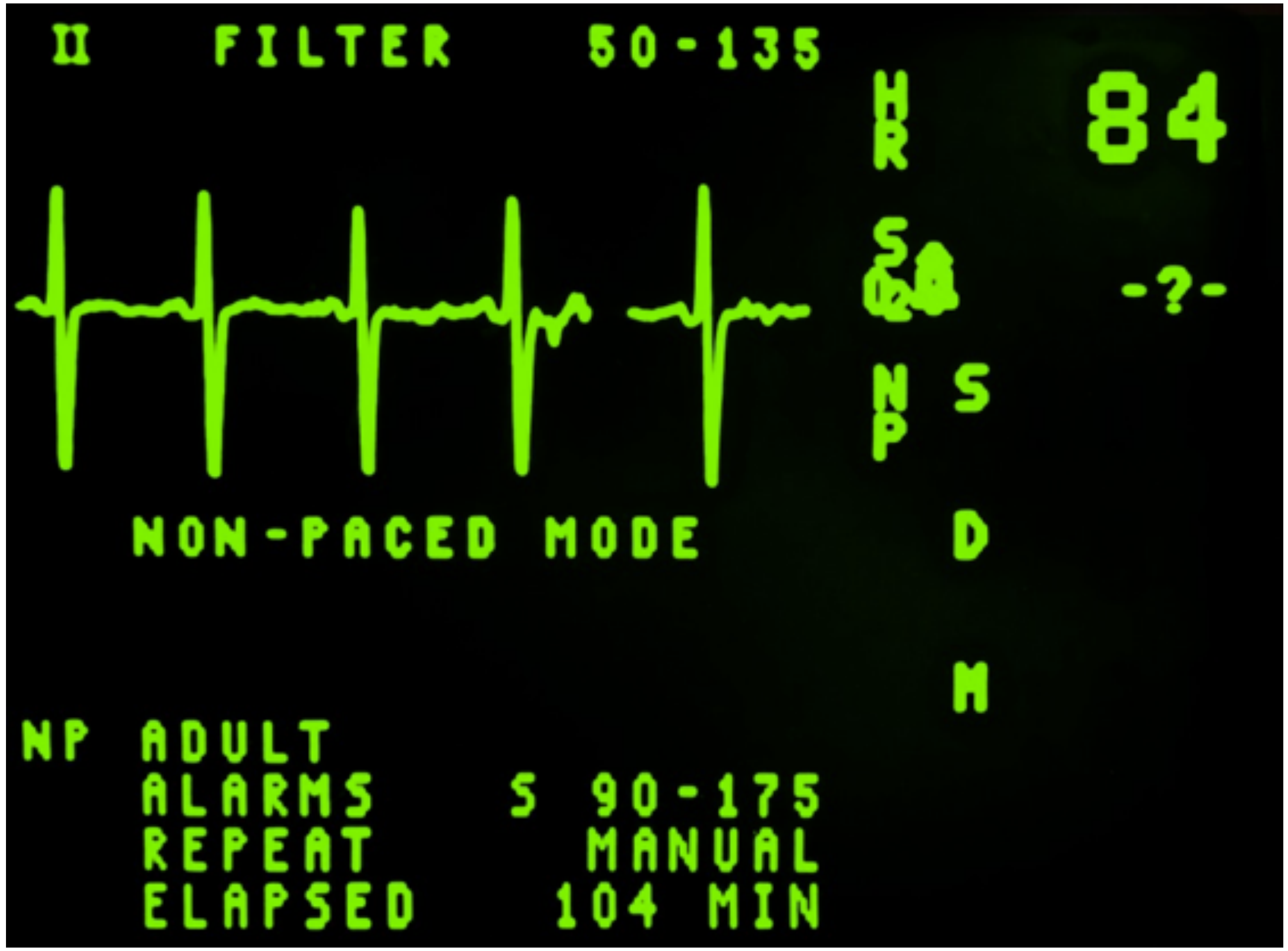
Can we use lifestyle data to accurately predict risk of heart attack?



Inspiration: EKG prediction

Can machines beat humans in medical diagnosis to improve patient outcomes?

New project

- Analyzing EKG data is hard!
- South African heart attack dataset is accessible
- Would this give us any useful results?

Problem

- Can we use lifestyle data (physical characteristics and living habits) to accurately predict risk of heart attack?

Hypothesis

- Initial thoughts: age and tobacco use seem quite correlated with incidence of heart attack

The data

- sbp
  - systolic blood pressure
- tobacco
  - cumulative tobacco (kg)
- ldl
  - low density lipoprotein cholesterol
- adiposity
- famhist
  - family history of heart disease
- typea
  - type-A behavior
- obesity
- alcohol
  - current alcohol consumption
- age
  - age at onset
- chd
  - coronary heart disease

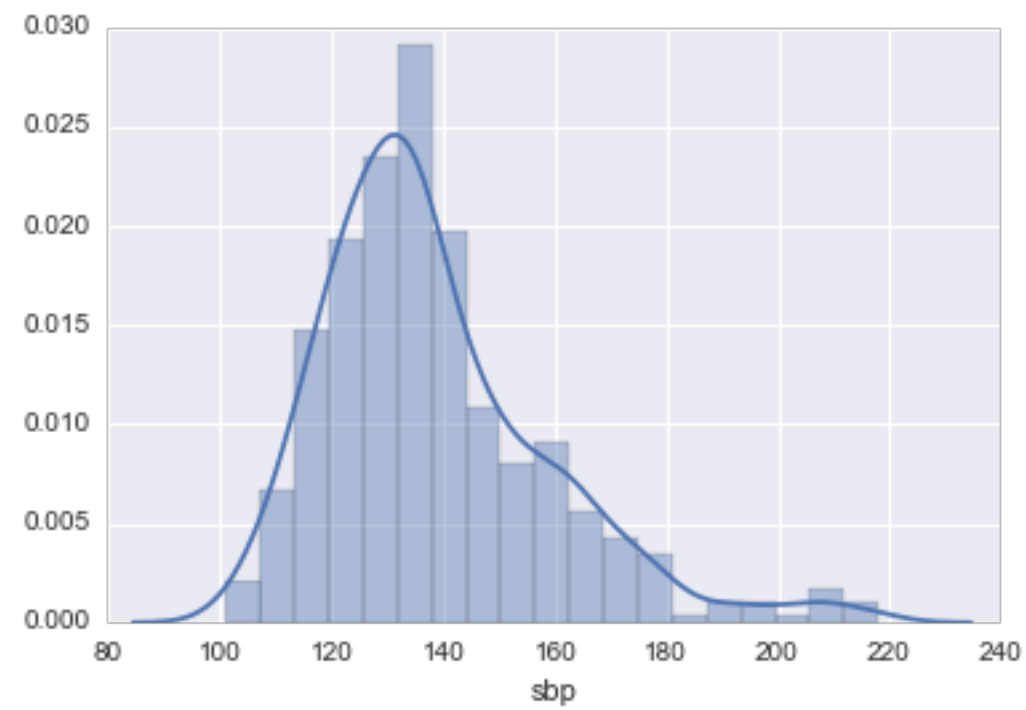
Data caveats

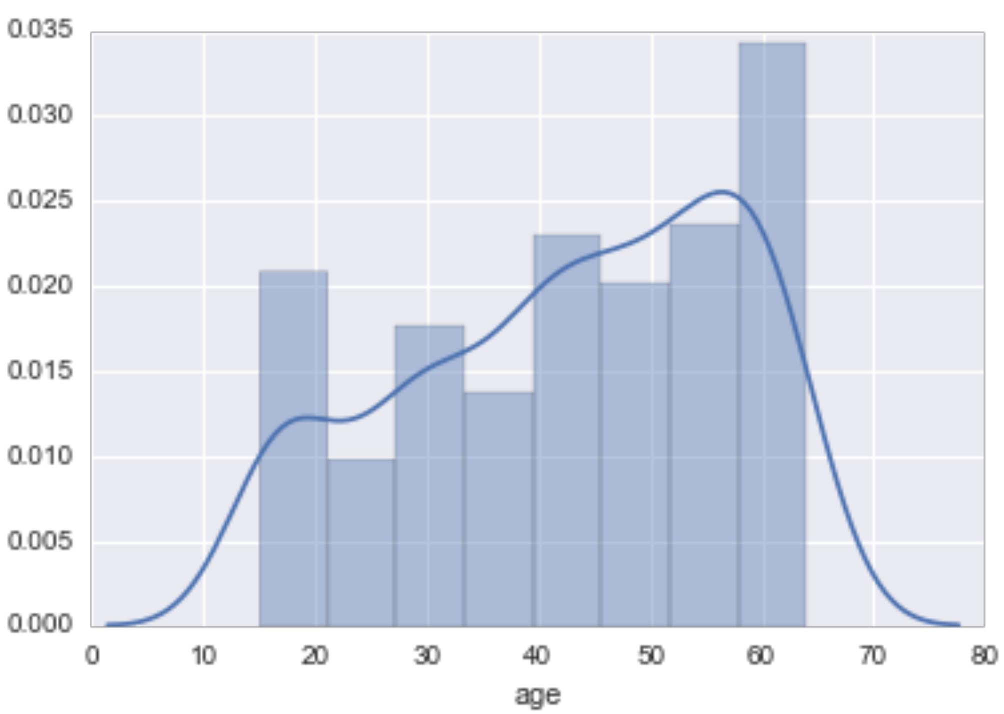
- Not generalizable or representative
  - Location
  - Measurements taken after heart attack
  - Some patients had even begun treatment!
- Limited inputs
  - No diet or other potentially important information

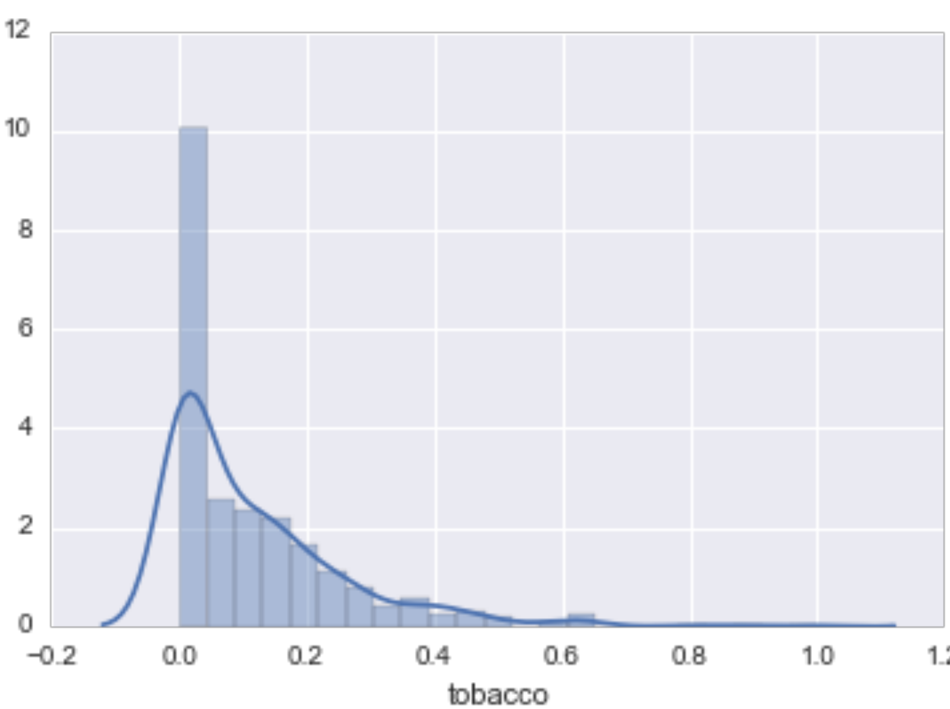
Exploring

- Most normally distributed (ex: sbp)

- Some skewed

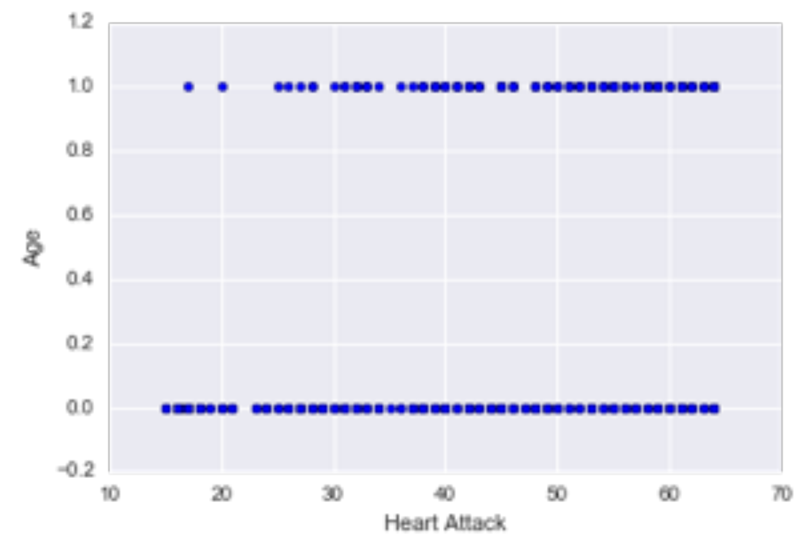


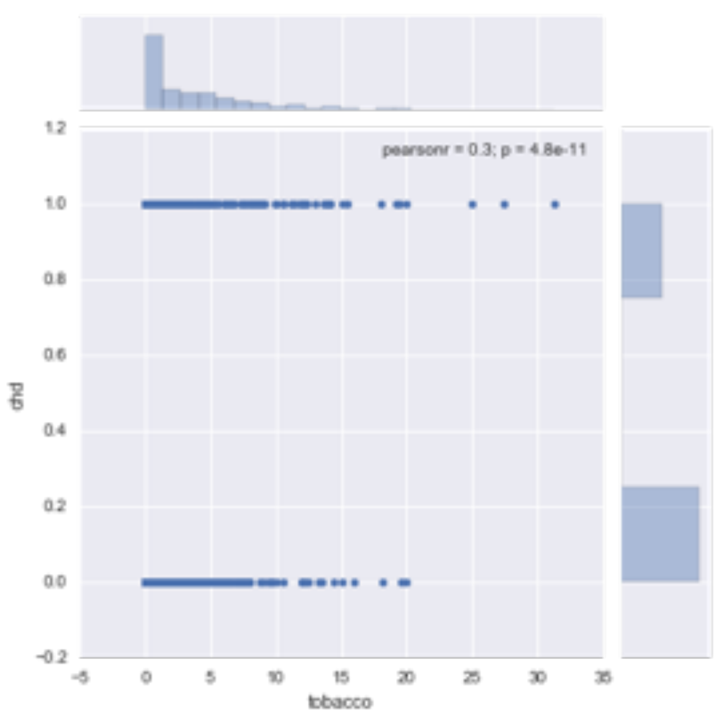


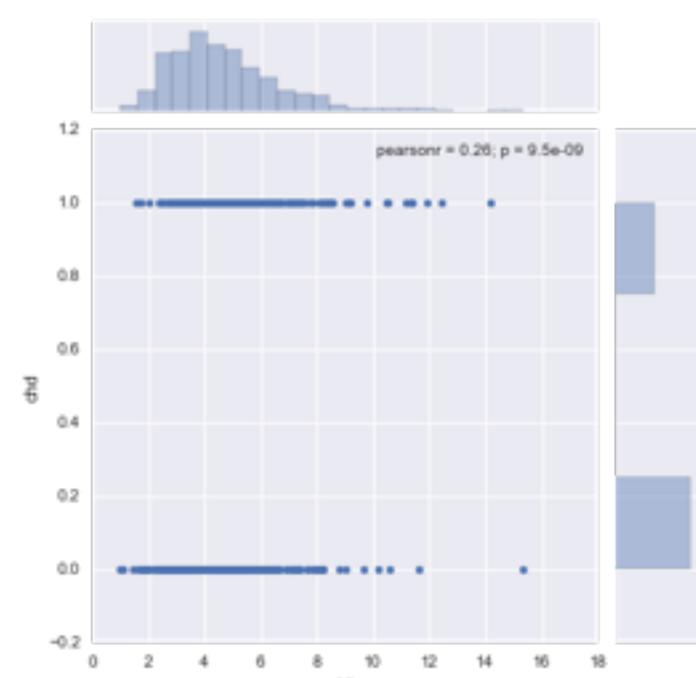


Exploring

- Mapping top 3 variables (p values < 0.001 and L1 penalty)
- Age is the only interesting result







Accuracy rate to beat

- Generally, 75-80% under ROC curve in published work
- 65% accuracy is the minimum for this dataset
  - Predicting nobody has CHD
- Korean study
  - Areas under ROC: 0.764 men, 0.815 women
- Second Korean study
  - Accuracy and receiver operating characteristic (ROC) curve: 69.51% and 0.594
- Classic study with Framingham data
  - c statistics (equivalent to area under ROC): 0.74 in men and 0.77 in women
- Study adding coronary artery calcium score as a factor
  - Area under ROC: 0.81

Minimum accuracy
> 65%

Fantastic accuracy
> 80%

Model 1: Logistic regression

- Second iteration: significant variables
- Cross-validation accuracy: 72.7%
- Tried setting C, but it didn't eliminate any variables. Boosted accuracy to 74%

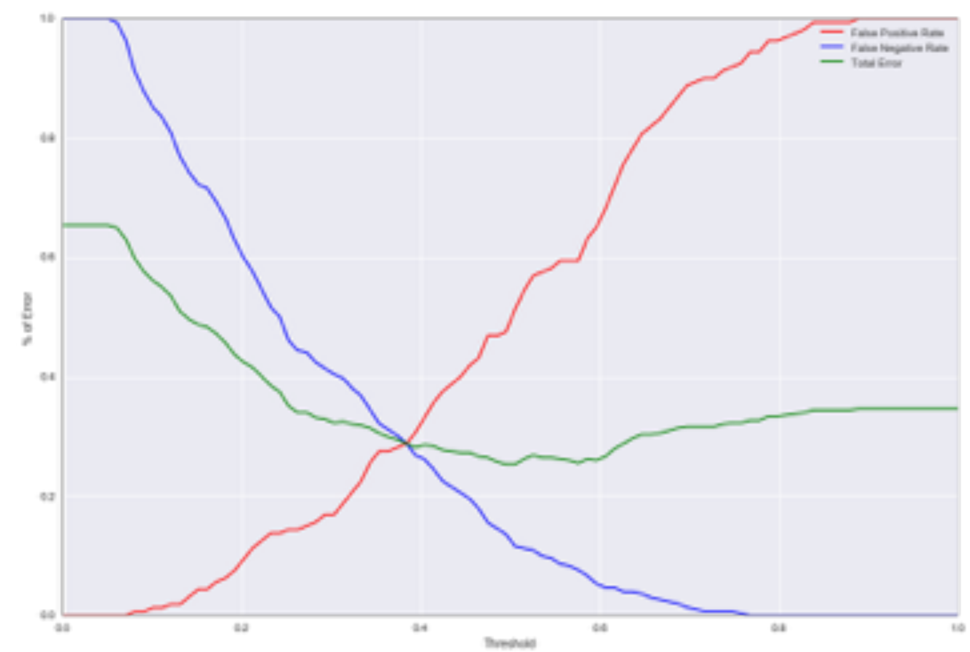
	coef	std err	z	P>|z|	[95.0% Conf. Int.]
tobacco	0.0804	0.026	3.106	0.002	0.030 0.131
ldl	0.1620	0.055	2.947	0.003	0.054 0.270
typea	0.0371	0.012	3.051	0.002	0.013 0.061
age	0.0505	0.010	4.944	0.000	0.030 0.070
famhist_present	0.9082	0.226	4.023	0.000	0.466 1.351

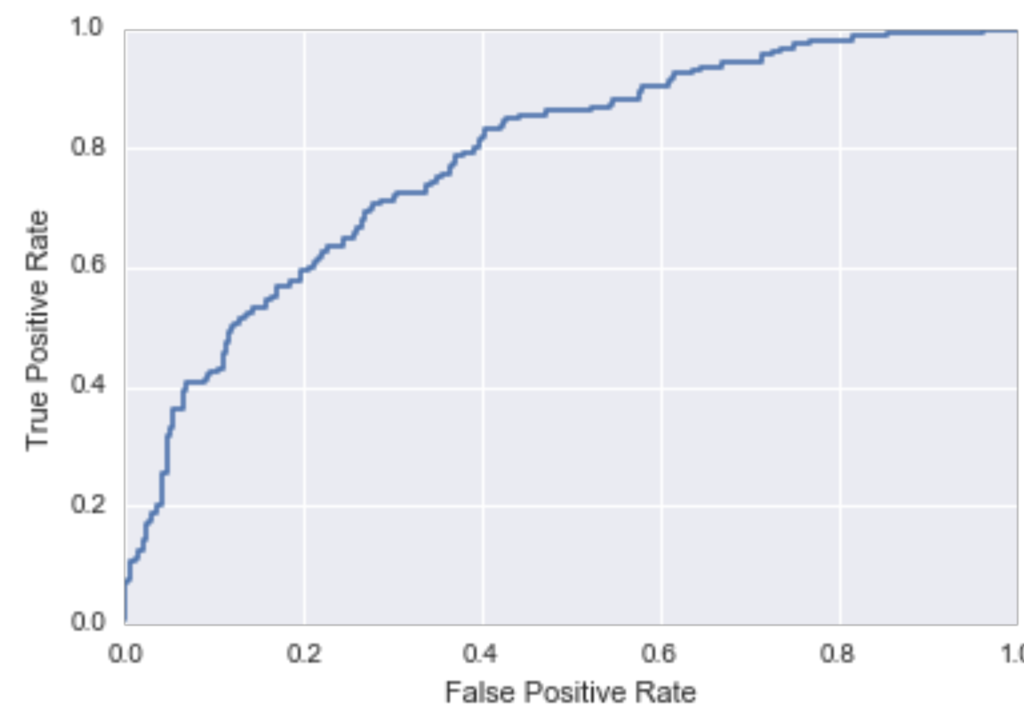
Model 1: confusion matrix

- Initial results are not great!
- We care most about reducing false negatives — these could be heart attack patients who weren't warned
- Err = 0.253247
- Acc = 0.746753
- FPR = 0.487500
- FNR = 0.129139

Model 1: adjusting threshold

- Default threshold of 0.5
  - False Positive Rate = 0.487500
  - False Negative Rate = 0.129139
  - Accuracy = 0.746753
- Updated threshold of 0.6
  - False Positive Rate = 0.662500
  - False Negative Rate = 0.05298
  - Accuracy = 0.735931
- Updated threshold of 0.7
  - False Positive Rate = 0.893750
  - False Negative Rate = 0.009934
  - Accuracy = 0.683983

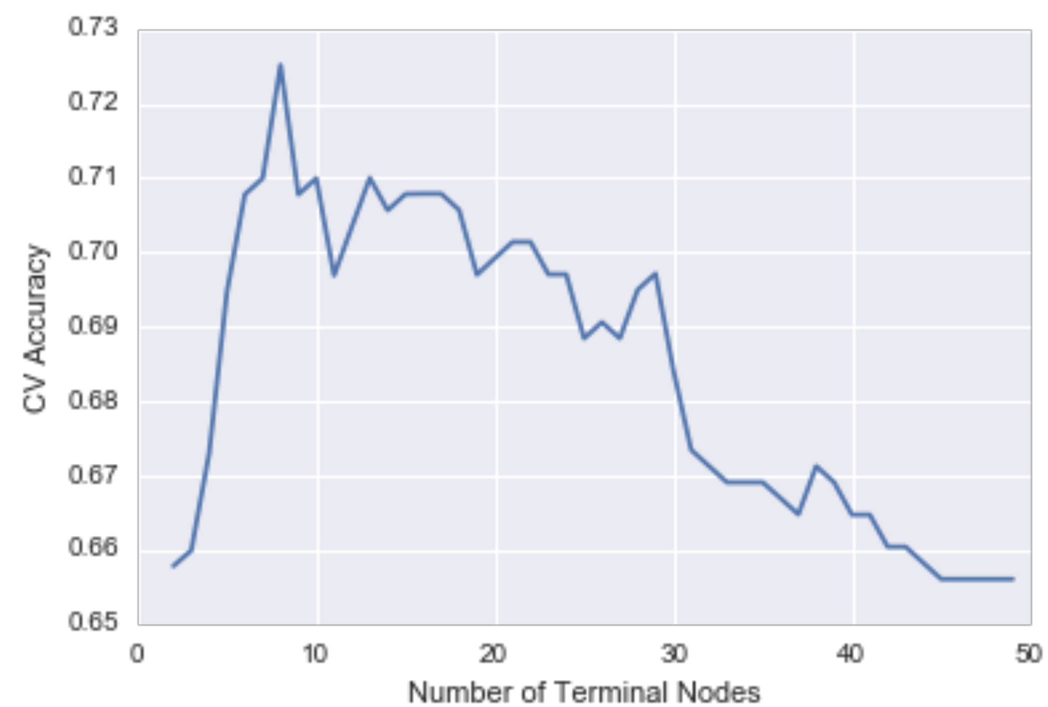




AOC: 78.35%

Model 2: decision tree

- Tuned based on max_depth and terminal nodes. Accuracy: 71.4% and 72.5%
- Similar variables were important



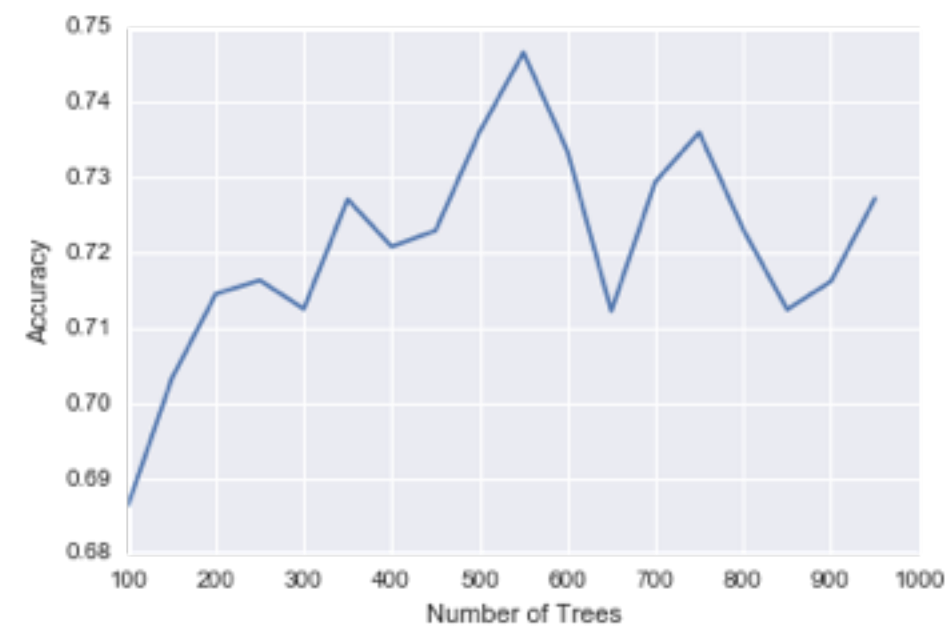
	feature	importance
3	age	0.456262
2	typea	0.202572
0	tobacco	0.156142
4	famhist_present	0.108380
1	ldl	0.076644

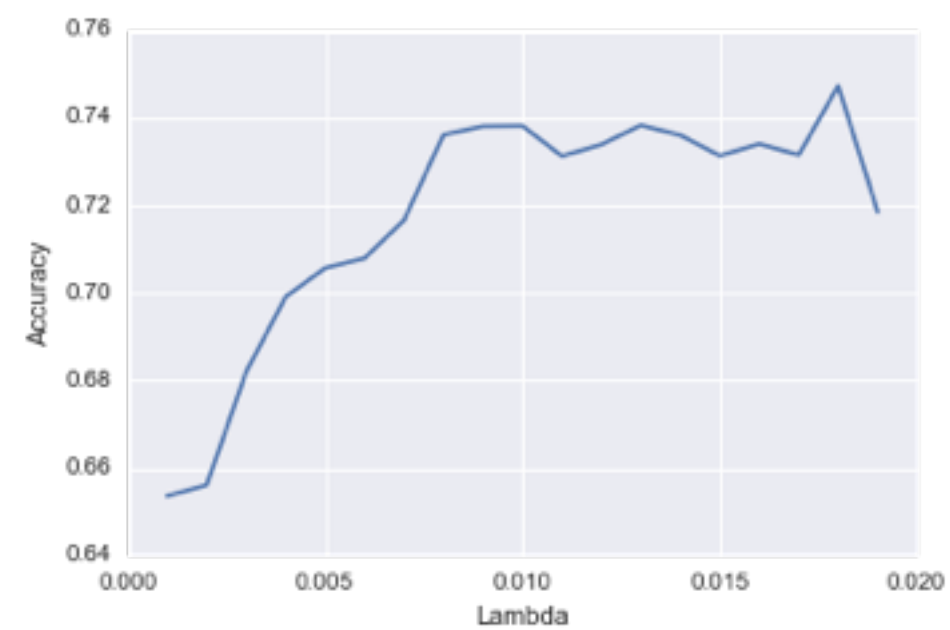
Model 2: confusion matrix

- Err = 0.220779
- Acc = 0.779221
- FPR = 0.475000
- FNR = 0.086093

Model 3: Random forest + boosting

- Regular random forest cross-validation accuracy: 71%
- Boosting: at first, 69.5% accuracy
    - lambda = 0.01, number of trees = 1000, depth = 2
- After tuning via semi-greedy approach: ~74% accuracy
    - lambda = 0.016, number of trees = 550, depth = 1

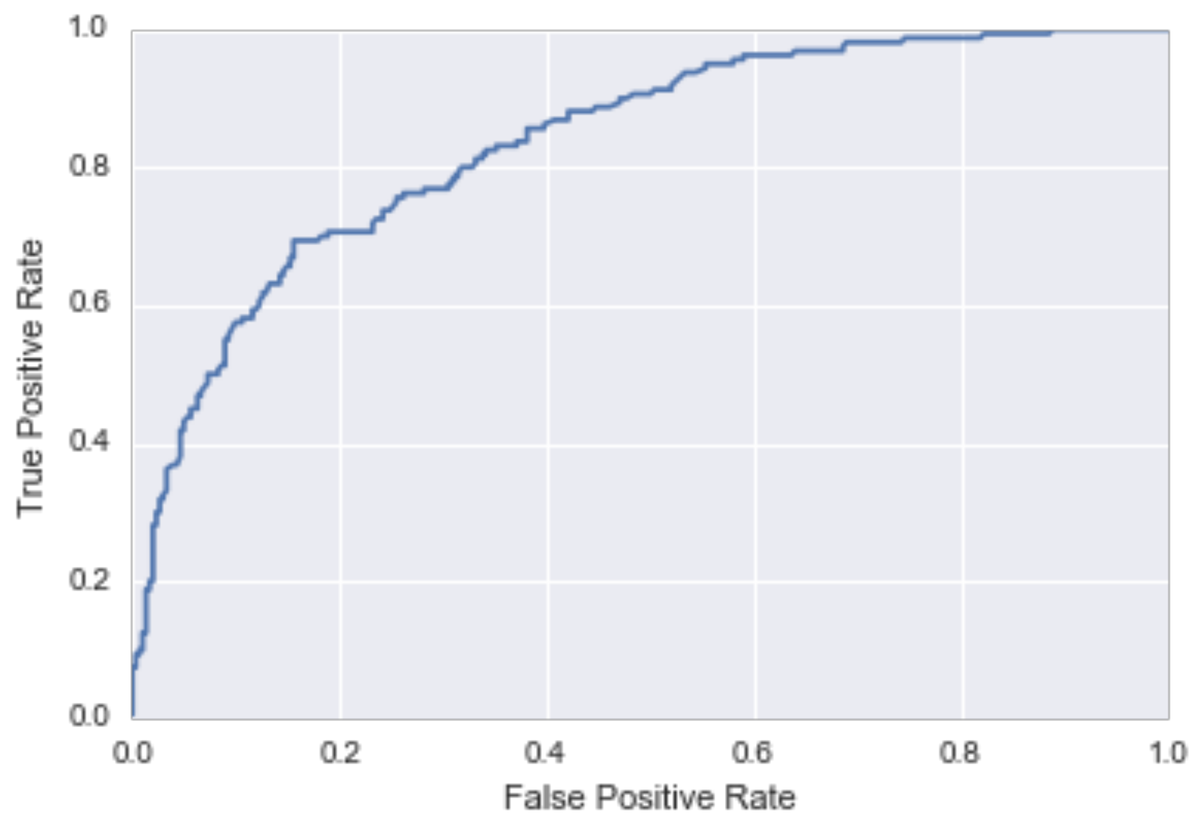




Model 3: confusion matrix

- Err = 0.216450
- Acc = 0.783550
- FPR = 0.418750
- FNR = 0.109272

Model 3: AUC



AUC: 83.6%

Meets our “fantastic accuracy” criteria: > 80%

Conclusions

- Boosting is great at predicting!
    - But is it correct?
- Significant and consistently important variables: age, family history, type A personality, LDL, and tobacco. (Not many!)
- It's hard to generalize these conclusions

Learnings

- Conclusions are only as good as your data
- Hypothesis: define this more clearly next time!
- Want to understand more of the math theory to validate findings before making solid recommendations

Next steps

- Predicting low, medium, high risk to reduce misclassification error
- Using a more comprehensive and diverse dataset
- Is lack of predictability really the biggest issue for preventing heart disease?
- If so, what is the most useful tool for sharing heart attack risk?

Thanks!

Keep in touch:
chloe.wood@gmail.com

References

- http://bmjopen.bmj.com/content/4/5/e005025.full
- http://www.ncbi.nlm.nih.gov/pubmed/26279953
- http://circ.ahajournals.org/content/97/18/1837.full
- http://jama.jamanetwork.com/article.aspx?articleid=185757
- http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.info.txt