Machine learning for early disease detection

Helen Lord • 11.02.2017
https://www.linkedin.com/in/helenlord27/
https://github.com/hslord

Simpatica
Medicine

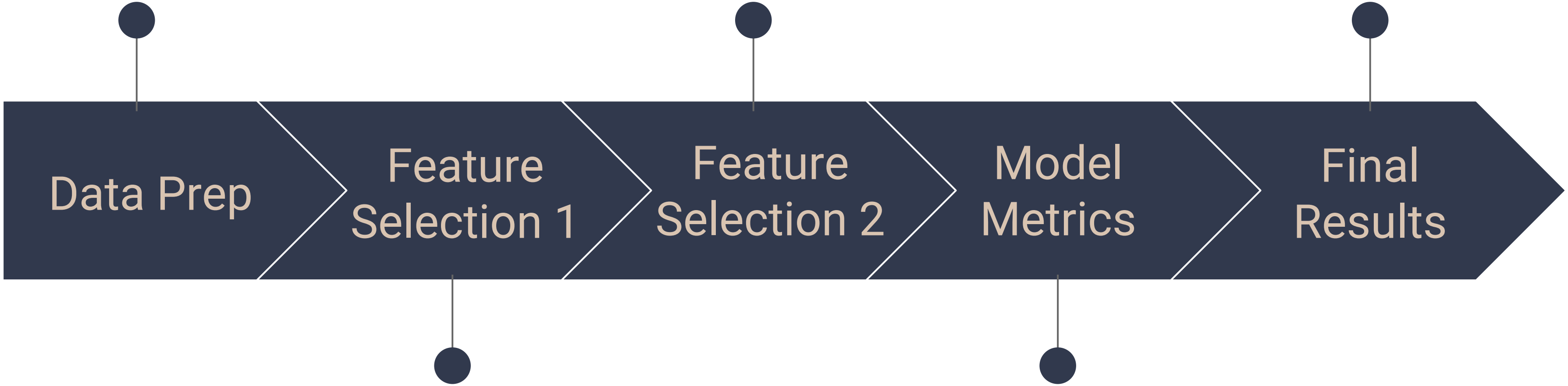
Overview

The Questions:

- What features are indicative of the disease?
- Can we accurately predict the presence of the disease?

The Data:

- 60k features per patient
- 116 labelled patients
  - 98 positive, 18 control
  - Baseline accuracy of predicting all positive: 84%



Reduced Features from 60k to 32k

Data Prep

Feature Selection 1

Compared Lasso and Random Forest models

Gradient Boosted model

Feature Selection 2

Model Metrics

Accuracy, Precision, Recall, and F1

Disease Detection and Feature Identification

Final Results

Initial Feature Selection

Random Forest

- Feature Importance identification
- Inconsistent feature selection
  - Independent of feature values
  - Model clusters related features and selects from clusters

Initial Feature Selection

Lasso - Regularized Linear Regression

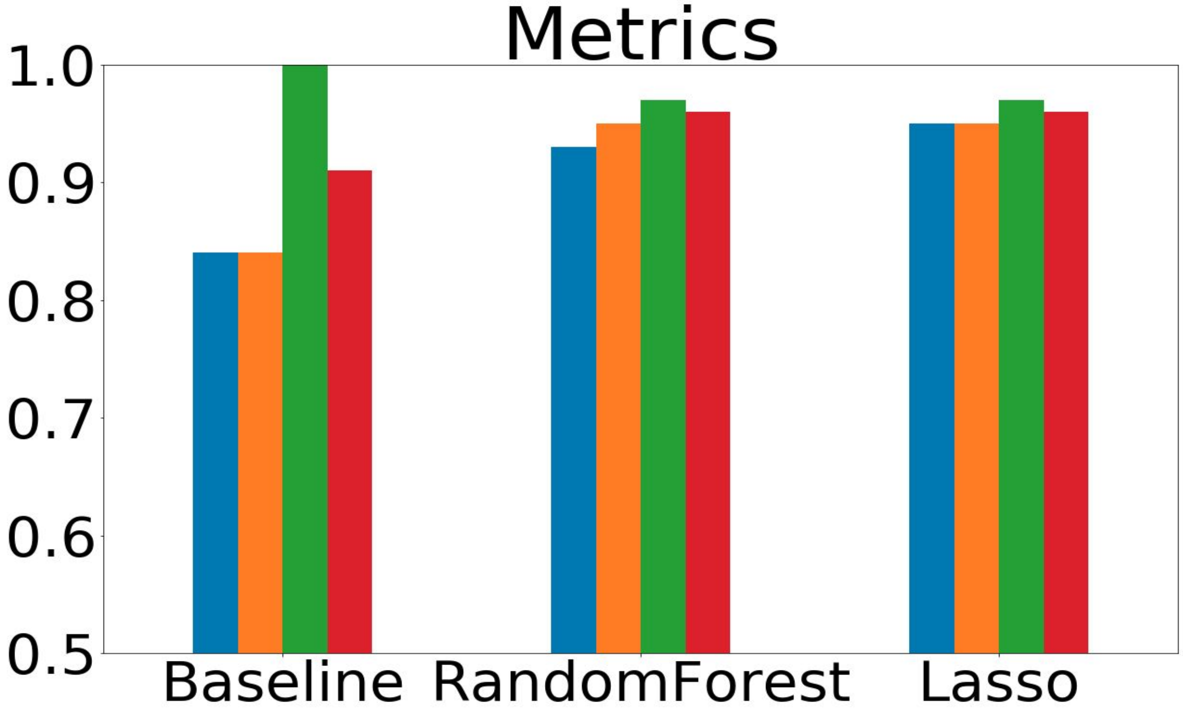
- Aggressive regularization to force low betas to zero
- Consistent feature selection
  - Biased towards high feature values
  - Model independently identifies strong feature relationships to label

Final Feature Selection

Gradient Boosting

1) Create two sets of features identified in previous models
2) Run feature sets independently through a Gradient Boosted model
3) Compare resulting top 25 feature importances identified
4) Iterate 50 times to find consistently important features

Gradient Boosted Model Prediction



Accuracy 🟦
Precision 🟧
Recall 🟩
F1 🟥

Next steps

Related Genes

- **Biological Approach:** Swap out biologically related features - see how model results change
- **Model Approach:** Find clustered features which can be swapped and maintain model results - research biological implications

More Data

- More stable, robust model
- More indicative of actual population