Heart Disease Prediction on Medical Data Using Ensemble and Deep Learning

Sagnik Majumder (sagnikm@stanford.edu) Eric Wang (ejwang@stanford.edu) Peter Dun (bodun@stanford.edu)

Motivation

Around 1 in every 4 deaths in the United States is related to cardiovascular disease. This unfortunately common occurrence can be mitigated with intelligent forecasts of heart disease based on related medical data.

Data Preprocessing

We collected 899 healthy and unhealthy patient datapoints containing medical and demographic features from Cleveland Clinic, University Hospital in Switzerland, and Hungarian Institute of Technology.

- Imputed missing data with the average in continuous features of the dataset.
- Augmented data with two features corresponding to the aggregate frequency of heart disease given age and biological sex.
- 4:1 train/test split with 5-fold cross validation for hyperparameter tuning
- Breakdown of categorical features into dummy variables (except for random forests)

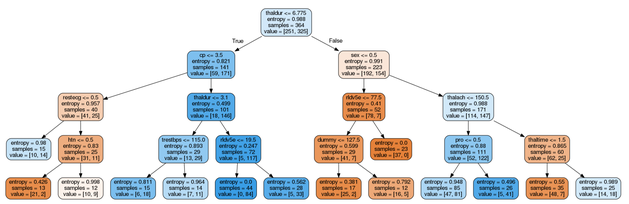
Feature Exclusion

We manually excluded features from our dataset corresponding to:

- Identifying information such as name and social security number
- Logistical information such as dates and locations of medical examinations
- Direct symptoms rather than causal indicators of heart disease

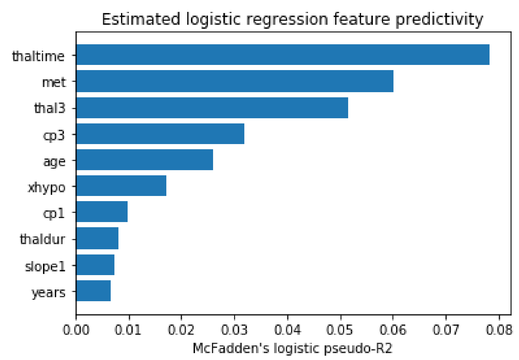
Random Forest

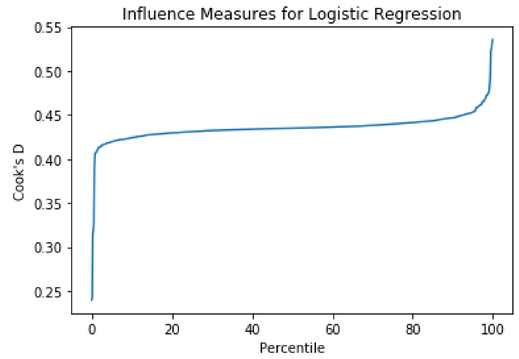
- Iterative hyperparameter tuning and backward feature selection to maximize validation set accuracy
- Predictive features: reported chest pain, max heart rate, cholesterol level, age, sex



Logistic Regression/SVMs

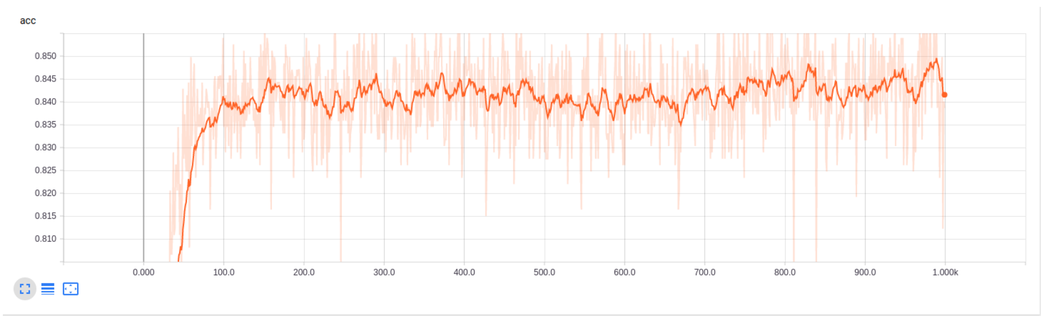
- Backward feature selection, dummy variables, L_2 regularization, removal of highly influential observations
- Predictive features: endurance, history of heart problems, reported chest pain, age
- Predictive power measured by McFadden's pseudo-R^2: $\tilde{R}^2 = 1 - \frac{\ell_{full}}{\ell_{sub}}$
- Attempted heavily L_2-regularized SVM with linear kernels
- Polynomial kernels have more parameters than there are datapoints so they don't perform as well due to overfitting





Neural Network

- Fully-connected 4-layer neural network with Xavier/He initialization
- BatchNorm after first dense layer to minimize covariate shift
- Adam optimizer at learning rate 1e-3 with binary cross-entropy loss
- Architecture: Input $\rightarrow$ Dense(64) $\rightarrow$ BatchNorm $\rightarrow$ Dense(32) $\rightarrow$ Dense(16) $\rightarrow$ Sigmoid
- Deeper neural nets severely overfit the data. Both L1 and L2 regularization hyperparameters needed to be extensively tuned to have comparable training and validation accuracy



Discussion

Similarity between the distributions of misclassified examples in models suggest a degree of irreducible error

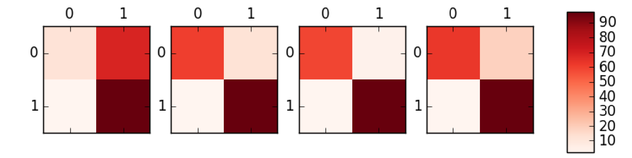


Figure: Confusion matrices of linear SVM, random forest, logistic regression, and neural network. 0 = healthy, 1 = diseased.

Linear SVMs, logistic regression, random forests, and neural networks all perform at a similar level. Radial basis function SVMs overfit greatly to the training set and hence have lower dev and test accuracies.

Model	Train	Dev	Test
RBF SVM	99.8	57.0	57.8
Log. Regression	80.0	78.9	75.0
Linear SVM	80.3	78.9	75.0
Random Forest	84.1	83.0	77.2
Neural Network	84.7	83.3	78.3
Majority Vote			77.0

Table: Different model performances

Future Work

- Acquire data with balanced demographics from more diverse geographical regions
- Expand feature set to mitigate irreducible error

Reference

Janosi, A et al. UCI Machine Learning Repo. [archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Summary Health Statistics: National Health Interview Survey 2015. [ftp.cdc.gov/pub/Health_Statistics]. Center for Disease Control.