Diagnosing Type II Diabetes based on Medical Records

CS 229: Machine Learning Project | Madeleine Gill, Katherine Holsteen, Haju Kim

MOTIVATION

- An estimated 9% of the US population has Type II diabetes, and 28% of them are living the with disease undiagnosed.
- The goal of this project is to develop a classification algorithm based on health history to help doctors identify presence of diabetes and promote early diagnosis and treatment.

DATA

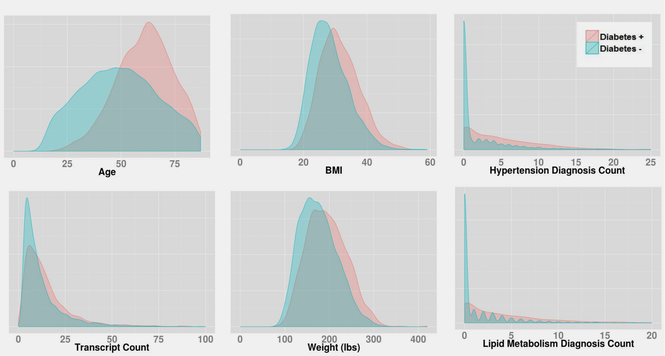
Raw Files

- De-identified medical records years 2009 through 2012 for 9,948 subjects, made public in a 2012 Kaggle competition
- 19.1% diagnosed with Type II diabetes
- Includes physician visit transcripts with vital signs and diagnoses, lab test results, medication history and smoking status
- The challenge: Diabetes-specific diagnoses, medications, and lab results (including glucose and insulin-related labs) were removed.

Feature Development

- Median values of BMI, weight, height, systolic and diastolic blood pressure across transcripts.
- Counts of diagnosis codes within clinically meaningful categories
- Counts of medications with potential relevance to diabetes
- Counts of physician visits by specialty
- Full feature set included 79 predictor variables.

Distribution of Key Features by Diabetes Status



APPROACH

METHODS

- Training set of 1409 (74%) of diabetic patients and a randomly selected subset of 1409 non-diabetic subjects undersampled to create balanced training data.
- Primary analysis involved estimating each model using the full feature set with 10-fold cross-validation for parameter selection.
- Secondary analysis evaluated each model using a reduced feature set based on a relative variable importance cutoff of 50.

Boosted Trees

- Learns slowly by sequentially adding shrunken values of the predictions from bootstrapped classification trees fit to the residuals of the previous predictions.
- A small value of shrinkage (0.001) allowed strong predictive performance at the cost of more computation time.

Random Forest

- Averages predictions from a series of decorrelated bootstrapped classification trees.
- A small number of predictors (4) considered at each split generated bias within each tree but low variance among the averaged predictions.

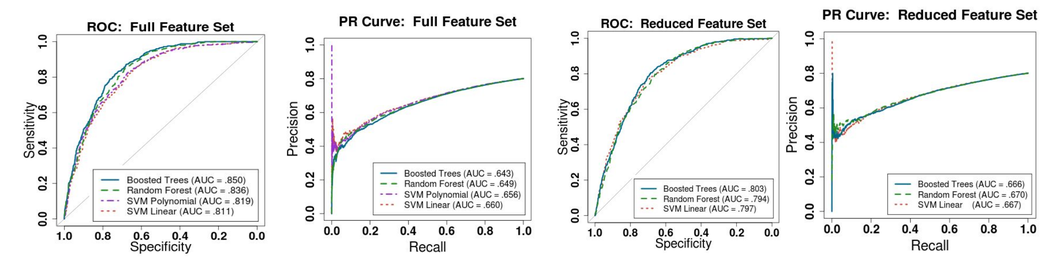
Support Vector Machines

- Constructs a hyperplane decision boundary in the feature space that maximizes the functional margin.
- Implemented linear and polynomial kernels of degree 2 and 3, with the best predictive performance from degree 2.

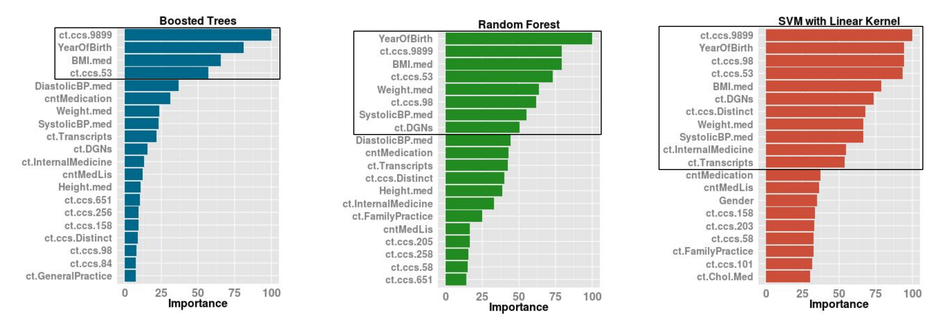
RESULTS

Predictive Accuracy by True States (Predictions on the held-out test set, n = 2487)

	Full Feature Set			Reduced Feature Set		
	GBM	RF	SVM	GBM	RF	SVM
Diabetes + (n = 495)	81.6%	85.1%	74.1%	79.6%	75.4%	76.5%
Diabetes - (n = 1992)	73.5%	68.3%	72.8%	69.1%	68.6%	70.8%



Variable Importance: Top 20 by Model



KEY TAKEAWAYS

- Top predictor variables show consistency across models and reflect the known correlates of diabetes.
  1. YearOfBirth (Patient birth year)
  2. Ct. ccs.9899 (Count of diagnoses of hypertension)
  3. BMI.med (Median BMI over all transcripts)
  4. Ct.ccs.53 (Count of diagnoses of disorders of lipid metabolism)
- Of the models considered, boosted trees performed best on the full feature set in terms of ROC AUC (0.849), and implementing the same model with only 4 predictors still provides an AUC of 0.803.
- These classification algorithms represent preliminary versions of potentially useful clinical tools.