Diagnosing Type II Diabetes based on Medical Records

Madeleine Gill
Stanford University
mmgill@stanford.edu

Katherine Holsteen
Stanford University
kholsteen@stanford.edu

Haju Kim
Stanford University
hajuk@stanford.edu

I. Introduction

Type II Diabetes plagues an estimated 9% of the U.S. population and gives rise to a variety of challenging medical complications. An estimated 28% of these individuals are living with the disease undiagnosed. The primary aim of this project was to use patient medical history to develop a machine learning algorithm to classify individuals with or without a diagnosis of type II diabetes. An accurate classification algorithm could serve as a clinically useful tool to advance the prevention and control of the disease by empowering doctors to prescribe preventative measures to at-risk patients and take earlier action to order diagnostic tests.

Our data source for this project was a 2012 Kaggle competition to solve the same classification problem. The available data included three years of medical records for 9,948 patients paired with a patient-level indicator for presence or absence of type II diabetes. The inputs to our algorithm were numeric features that we developed based on the medical record data. We used boosted trees, random forest, and support vector machines to predict the binary outcome of a positive or negative diagnosis of type II diabetes.

II. Related Work

A substantial body of literature has been published on predicting incidence of type II diabetes within five to ten years based on disease history and clinical measurements completed a systematic survey and external validation of 25 of these models [1] [2] [3] [4] [5]. Twelve relied on basic predictors (including BMI, blood pressure, waist and hip circumference, food frequency, and disease history), and thirteen extended models also included biomarkers (glucose and/or HbA1c). The highest AUC for 5-year risk of diabetes was 0.84 in the basic models and 0.92 in the extended models. The majority of these studies used logistic regression, which allows for easier interpretation of effect sizes but may fail to capture more complex nonlinear relationships and interactions among features. Our project adds a wider variety of machine learning methods including non-parametric tree-based methods and support vector machines with the goal of building a robust model with lower bias. Our project also contributes new utility in its cross-sectional nature and immediate applicability for current rather than future diagnosis. To the extent possible with our available data, we created features based on the characteristics identified as important in the literature, including older age [2] [6], higher BMI [3] [6], and high blood pressure (hypertension) [2]. We did not have access to some of the more specific metrics such as waist and hip circumference, lifestyle variables or diabetes-specific biomarkers.

In addition to scholarly studies, prior work on this same problem includes the 2012 Kaggle competition. Through interviews published on Kaggle, the top two teams reported investing substantial time into researching and developing features, including categorizing diagnoses based on publicly available groupings of ICD-9 codes and identifying medications with relevance to diabetes [7]. Each of the top three teams implemented gradient boosted models as part of their winning submission [7]. We tried to pattern our methodology in a similar way.

III. Dataset and Features

Our dataset is composed of medical records for the years 2009 through 2012 for 9,948 subjects, including physician visit transcripts with vital signs and diagnoses, lab test results, and medications. The median number of transcripts was 9 per subject. In preparation for the Kaggle competition, they were edited to remove diabetes-specific diagnoses, lab results, and prescriptions and assign each individual a binary indicator for presence or absence of diabetes. Of our sample, 19.1% of subjects have prevalent Type II diabetes; twice the estimated prevalence for the U.S. population. We divided this dataset into a 75% training sample (n = 7641) and 25% validation sample (n = 2307). For the initial analyses, we used a balanced subset of the training sample including all of the diabetic training patients (n =1409) and an equal number of randomly selected non-diabetic patients (n = 2818 total).

We spent significant time examining the available data and developing features with potential correlation to type II diabetes. The full set of 79 features is shown in table D1. These include medians of BMI, weight, height, and blood pressure across transcripts, and counts of lab tests and abnormal lab results over the three-year window. We created counts of diagnoses within clinically relevant categories, using the Clinical Classifications Software groupings of ICD-9 codes specifically designed for healthcare data analysis [8]. The counts of diagnoses within each of the 50 most common categories were included in our feature set. We also created counts of medications with potential relevance to diabetes; in particular, we counted specific hypertension prescriptions such as Lisinopril and Hydrochlorothiazide because of the established correlation between type II diabetes and hypertension. Missing data did not present a major obstacle. In the rare case of no data available to calculate a median, we imputed the median value for the training sample. The count variables did not require imputation because the absence of

specific diagnoses or medications corresponded to a count of zero.

IV. METHODS

For the primary analysis, the full set of 79 features on the balanced training set (n = 2818) was used to train each of three supervised learning algorithms for classification: (1) gradient boosted trees, (2) random forest, and (3) support vector machines.

A. *Boosted Trees*

Boosted trees is an algorithm that sequentially combines weak learners and ensembles them through a weighted majority vote to predict the class. Boosted trees begins with a small tree and builds additive tree models. If an observation is misclassified, the observation receives more weight in the next iteration. The final classifier is a weighted sum of the decisions made by trees (the majority vote), for which the weights are assigned based on their prediction performance. Thus, boosting tries to reduce the bias of a large number of small trees that have low variance by finding the optimal linear combination of trees in relation to the training data. The gradient boosting algorithm is described below [9]:

Gradient Boosted Trees

1. Initialize $f_{k0}(x) = 0, k = 1, 2, \dots, K$.
2. For $m = 1$ to M:
   (a) Set
   $$p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{K} e^{f_l(x)}}, k = 1, 2, \dots, K.$$
   (b) For $k = 1$ to K:
      i. Compute $r_{ikm} = y_{ik} - p_k(x_i), i = 1,2, \dots, N$.
      ii. Fit a classification tree to the targets $r_{ikm}, i = 1, 2, \dots, N$, giving terminal regions R_{jkm}, $j = 1,2, \dots, J_m$.
      iii. Compute
      $$\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \in R_{jkm}} r_{ikm}}{\sum_{x_i \in R_{jkm}} |r_{ikm}| \, (1 - |r_{ikm}|)}, j = 1,2, \dots, J_m$$
      iv. Update $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x \in R_{jkm})$.
3. Output $\hat{f}_k(x) = f_{kM}(x), k = 1, 2, \dots, K$.

Boosted trees are known to be robust to outliers, missing data, heterogeneous predictors, and can automatically select variables [10]. For this project, boosted trees algorithms are implemented using the gbm and caret packages in R. When tuning the parameters, we evaluate different combinations of the interaction depth = {1, 5, 10, 15} and the number of trees to grow = {50, 100, .., 1450, 1500} with 10-fold cross validation with 5 replications.

B. *Random Forest*

Random forest is an algorithm based on bagging (bootstrap aggregation) that averages predictions from a series of decorrelated bootstrapped classification trees [11]. Thus, random forest tries to achieve low variance by reducing the correlation between the individual trees through bagging and random feature selection. When used for classification, random forest aggregates a class vote from individual trees and then predicts the outcome with the majority vote. One advantage of random forest is that random forest tends to work well with very little tuning required. The only parameter to optimize is the number of variables randomly sampled as candidates at each split, varied according to (m_{try}) = {4, 6, 8, 10, 12, 14}. Random forest is implemented through the randomForest and caret packages in R.

C. *Support Vector Machines (SVMs)*

Support vector machines constructs a hyperplane decision boundary in the feature space that maximizes the functional margin. Since we do not expect our data to be linearly separable, we implement the soft-margin SVMs and formulated SVM optimization problem with slack variables as below:

Soft-margin SVMs

Primal

$$\min_{\gamma,w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i$$
$$\text{s.t.} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, i = 1, \dots, m$$
$$\xi_i \geq 0, i = 1, \dots, m$$

Dual

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$
$$\text{s.t.} \quad o \leq \alpha_i \leq C, i = 1, \dots, m$$
$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Thus, our optimization problem minimizes the training error traded off against the margin. The parameter C is a regularization term that allows us to control the trade-off between the slack variable penalty and the size of the margin. The higher the value of C, the higher the variance of the function. We tried both the linear and polynomial kernels; the latter increases the flexibility of the decision boundary by introducing polynomials with an intercept. The polynomial kernel can be formulated as:

$$K(x, z) = (\gamma(x^T z) + 1)^d$$

When tuning the models, we limited the possible polynomial degrees to {2,3} because of a lack of prior theory suggesting higher-degree relationships among the features. We evaluate different combinations of the regularization term C = {0.001, 0.01, 0.1, 1, 10, 100}, γ= {0.001, 0.01, 0.1, 1, 10,

100}, and the degree of polynomials= {2, 3}. SVM is implemented with the kernlab and caret packages in R.

Based on strong cross-validation performance, secondary analyses used the boosted tree algorithm to estimate three additional models: (1) using only the top four most important features for the balanced training set; (2) adding in the full feature set for the full training sample (n=7641) with observations weighted inversely to outcome class size; and (3) the top four features for the full weighted training sample. As a follow-up analysis, we estimated and evaluated the boosted trees model separately for younger patients (<= 50 years old) and older patients (> 50 years old). These age-stratified models included the full feature set and the full weighted training sample.

We chose to optimize area under the ROC curve (AUC) as our metric of predictive utility. The AUC quantifies the trade-off between sensitivity and specificity at all cut-off thresholds. For each of the models estimated, 10-fold cross-validation was used to select parameters that maximize the mean AUC on the held-out fold. Finally, each model was evaluated on the held-out test set.

V. Experiments/Results/Discussion

A. *Initial Classification Models*

The following 4 models were run on the balanced subset of the training sample with the full feature set: (1) Boosted trees with 900 trees with an interaction depth of 10, a shrinkage rate of 0.01 and minimum 10 observations per node; (2) Random forest with 4 randomly selected predictors considered at each split; (3) SVM with a linear kernel with cost of 0.01, and (4) SVM with a degree-2 polynomial kernel, the cross-validation results suggested an optimal cost of 0.01. The results for these four models in terms of cross-validation AUC and test AUC are shown in Table 1. The test ROC curves are compared in Figure 1. ROC AUC on the test set is largest for the boosted trees (0.850), and smallest for the linear SVM (0.811), but the different algorithms showed relatively similar performance overall.

The area under the PR curves (not shown) ranges between 0.643 and 0.660 for the four models, indicating generally low precision in identifying prevalent diabetes cases in the test dataset. For the development of this classification tool, low precision is less of a concern than sensitivity and specificity in practice.

Table 1 Cross-validation AUC and test AUC for the four models using the balanced training set (n = 2818) and full feature set (p = 79)

	Boosted Trees	Random Forest	SVM Linear	SVM Polynomial
CV AUC (SE)	0.847 (0.019)	0.832 (0.024)	0.816 (0.034)	0.821 (0.032)
Test AUC	0.850	0.836	0.811	0.819

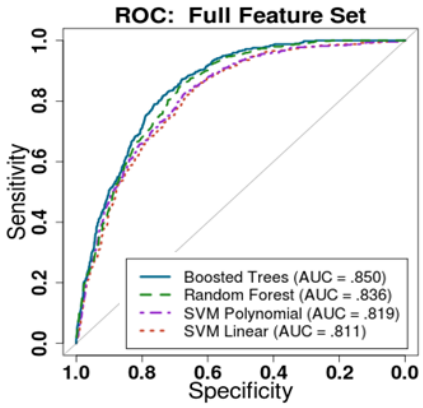

Figure 1 ROC curves for predictions on the held-out test set with each of the four initial models

Figure 2 shows the variable importance scores for the top 10 most important variables from each model. These charts highlight the consistency in the key predictor variables across algorithms, particularly in the standout four predictors ct.ccs.9899 (count of hypertension diagnoses), YearOfBirth, BMI.med (median BMI over all transcripts), and ct.ccs.53 (count of lipid metabolism disorder diagnoses). Indeed, these predictors reflect known patterns in diabetes epidemiology: older age and higher BMI independently increase one's risk for type II diabetes, while hypertension and hyperlipidemia are common comorbidities with overlapping risk factors [2] [3] [12].

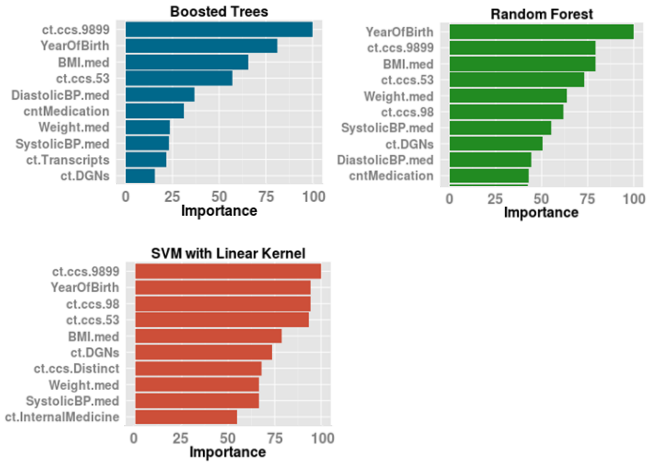

Figure 2 Variable importance for boosted trees, random forest, and SVM models applied to the balanced training dataset (Variable importance for linear and polynomial kernels were identical)

B. *Variations on the Boosted Trees Model*

Based on the strong cross-validation performance of boosted trees in the initial model, we estimated three variations on the boosted tree model to explore the predictive power with a reduced predictor space and with the full training set. First, we used a variable importance threshold of 50 to select the top 4 variables from the initial boosted trees model as the reduced predictor set. We estimated a reduced model with the balanced training set using only the top 4

features using 850 trees with an interaction depth of 1. Next, we added the 4643 additional non-diabetic training subjects back into the training dataset (total n = 7641) and estimated two boosted trees models including observation weights inversely proportional to class sizes. These models utilized (1) full feature set with 900 trees and interaction depth of 15, and (2) top 4 features with 300 trees and an interaction depth of 5. The AUC of these three models for cross-validation and for the test set are shown in Table 2.

With only four features, the boosted trees model estimated on the balanced training set provided an AUC of 0.803 on the test dataset. This result achieved from a very limited subset of meaningful predictor variables shows the potential for a useful classification model requiring only a few pieces of a patient's medical history. The addition of negative training examples did not significantly change the AUC for the full feature set or the reduced feature set, suggesting that our initial models were not overfitting to the smaller balanced training data (See Table 2). The breakdowns of misclassification rates by age and BMI are shown in Table 3. False positive rates are higher among those patients displaying the key risk factors of older age and higher BMI, while false negatives are more common among patients without those characteristics.

Table 2 Cross-validation AUC and Test AUC for boosted tree models on balanced and weighted training dataset with full and reduced feature sets.

	Full Feature Set		Top 4 Features	
	Balanced ($n = 2818$)	Full ($n = 7461$)	Balanced ($n = 2818$)	Full ($n = 7461$)
CV AUC (SE)	0.847 (0.019)	0.844 (0.017)	0.798 (0.025)	0.796 (0.021)
Test AUC	0.850	0.855	0.803	0.802

Table 3 Classification Rates by Age and BMI

	Diabetes	N Test Obs	Pred+	Pred -
Age <= 50	D +	1038	57.7%	**42.3%**
	D –	104	5.9%	94.1%
Age > 50	D +	954	80.3%	19.7%
	D –	391	**37.3%**	62.7%
Normal Weight (BMI <25)	D +	69	55.1%	**44.9%**
	D –	613	9.1%	90.9%
Overweight (25 <= BMI < 30)	D +	165	72.1%	27.9%
	D –	676	18.3%	81.7%
Obese (BMI >= 30)	D +	261	83.1%	16.9%
	D –	703	**33.7%**	66.3%

C. Age-Stratified Model

Since age is such an important predictor, we stratified the training sample by age at study onset (year 2009) and estimated separate classification models on each strata to compare predictive performance and key variables between age groups. We estimated two separate boosted trees models using the full feature set (except YearOfBirth) on the patients from the full training set: (1) patient with age ≤ 50 (n = 3518), 600 trees with interaction depth of 5, and (2) patients with age > 50 (n = 3943), with 700 trees and an interaction depth of 15. Figure 3 shows the discrepancy in predictive performance across the two age groups. For patients ≤ 50 years old, the model predicts well with a test AUC of 0.883, but for older patients, the predictive performance deteriorates (test AUC = 0.787). The variable importance plots shown in Figure 4 reveal that BMI is by far the most important predictor among the older patients, followed by ct.ccs.9899 (hypertension). For younger patients, ct.ccs.53 (disorders of lipid metabolism) rose to the top followed by BMI and ct.ccs.9899. Four out of the five top predictors are identical across the age strata. The similarity of the predictors and difference in accuracy suggests common risk factors among the groups, but more noise in the older population due to an accumulation of different diseases and stresses over one's lifetime.

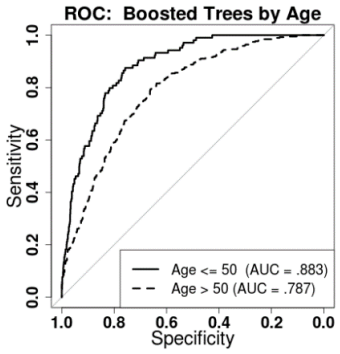

Figure 3 ROC for Boosted Tree models by age group

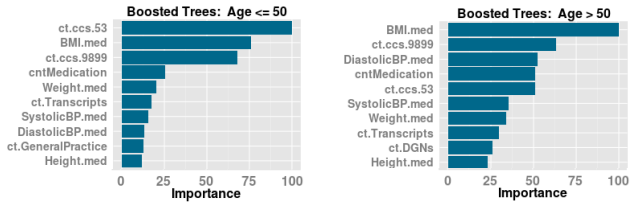

Figure 4 Variable importance for boosted trees by age group

Table 4 Cross-validation AUC and Test AUC for boosted tree models on full feature set except for year of birth by age group

Full Feature Set Except Year of Birth (p = 78), Weighted Training Set		
	Age ≤ 50	Age > 50
CV AUC (SE)	0.852 (0.034)	0.782 (0.024)
Test AUC	0.883	0.787

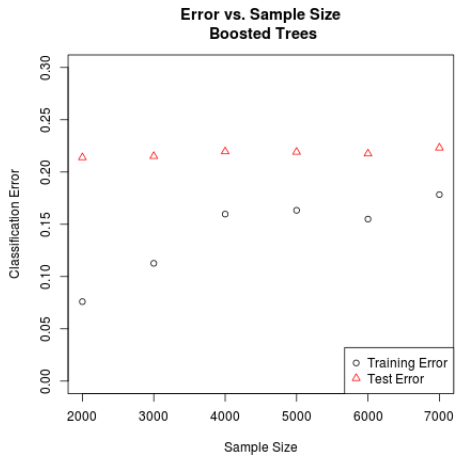

Figure 5 Classification Error vs. Sample size for the boosted trees model, run with the full feature set on progressively larger subsets of the full training dataset

As an investigation into the performance of our boosted trees models, we took a series of progressively larger samples of our full training dataset and conducted parameter selection and model estimation separately on each one. Finally we evaluated each resulting model on the held-out test set. The training and test misclassification error rates for each sample size are shown in Figure 5. The training error increases between 2000 and 4000 observations and then stabilizes around 16% as sample size continues to increase. The test error remains relatively constant around 22% across all sample sizes. The similar high rates of test and training error rates suggest that our model suffers from high bias rather than high variance. This is to be expected given that our feature space was missing some key diabetes diagnostics like glucose and insulin-related lab results, which in many cases are unexplained by the records we have available. Throughout the project we were aware of the limitations of our feature space and repeatedly created and tested new features including a wider range of diagnosis categories and ratios of diagnosis counts to the number of physician visits. In the end, these additions did not improve predictive performance over the core set of 79 features used in these analyses.

VI. Conclusion

We used three different classification algorithms to predict the presence or absence of a type II diabetes diagnosis based on features created from patient medical records. Among the three learning algorithms used, the boosted trees model performed the best, followed by random forest and last support vector machines. SVM may have fit the data less well because of the high level of overlap among the classes, even in a higher-dimensional feature space

We demonstrated that the boosted tree model represents a preliminary version of a potentially useful clinical tool. The ROC curve for the boosted tree model with the full feature set indicates that an optimal cutoff threshold could provide sensitivity of 85% and specificity of 70%. Given just a few pieces of information from one's medical record (age, BMI, number of hypertension diagnoses, and number of lipid metabolism disorder diagnoses), the boosted trees model yielded an AUC of 0.803. A doctor could use either the full or reduced model to obtain a relatively accurate prediction of a patient's diabetes status and help guide a choice to order diagnostic blood tests.

Analysis of the boosted trees model for different sample sizes revealed no improvement in test error with additional training examples. Though the boosted trees model shows promise, both test and training error remain higher than desired, indicating high bias in the model. In future steps to improve these models, we could incorporate clinical expertise to expand and refine the feature space to help reduce bias. This could be particularly helpful with a focus on risk factors specific to the older adult population, where the boosted trees model showed the greatest room for improvement.

References

[1] S. Mani *et al*, "Type 2 diabetes risk forecasting from EMR data using machine learning," *In AMIA Annual Symposium Proceedings,* vol. 2012, p. 606, 2012.

[2] J. T. Jaana Lindström, "The Diabetes Risk Score: A practical tool to predict type 2 diabetes risk," *Diabetes Care,* vol. 26, no. 3, pp. 725-731, 2003.

[3] S.G. Wannamethee *et al*, "The potential for a two-stage diabetes risk algorithm combining non-laboratory-based scores with subsequent routine non-fasting blood tests: results from prospective studies in older men and women," *Diabetic Medicine,* vol. 28, no. 1, pp. 23-30, 2011.

[4] J. Hippisley-Cox *et al*, "Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore," *Bmj,* p. 338:b880, 2009 Mar 18.

[5] A. Abbasi *et al*, "Prediction Models for Risk of Developing Type 2 Diabetes: Systematic Literature Search and Independent External Validation Study," *BMJ,* vol. 345, p. e5900–e5900, 2012.

[6] W. Rathmann *et al*, "Prediction Models for Incident Type 2 Diabetes Mellitus in the Older Population: KORA S4/F4 Cohort Study," *Diabetic medicine : a journal of the British Diabetic Association,* vol. 27, no. 10, p. 1116–23, 2010.

[7] M. Zwemer, "Practice Fusion Diabetes Classification - Interviews with Winners," 3 10 2012. [Online]. [Accessed 12 10 2015].

[8] Agency for Healthcare Research and Quality, Rockville, MD. , "Healthcare Cost and Utilization Project (HCUP)," June 2015. [Online]. [Accessed 9 December 2015].

[9] T. Hastie *et al*, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics, 2009.

[10] J. Friedman *et al*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics,* pp. 337-407, 2000.

[11] M. W. Andy Liaw, "Classification and regression by randomForest," *R news 2.3,* pp. 18-22, 2002.

[12] P. J. Lin, "Clinical Multiple Chronic Conditions in Type 2 Diabetes Mellitus: Prevalence and Consequences," vol. 21, no. 1, 2015.

[13] G. J Bosman and M. M. B. Kay, "Alterations of band 3 transport protein by cellular aging and disease: erythrocyte band 3 and glucose transporter share a functional relationship," *Biochemistry and Cell Biology,* vol. 68, no. 12, pp. 1419-1427, 1990.