Methods for Predicting Type 2 Diabetes

Duyun Chen, Yaxuan Yang, Junrui Zhang

Stanford University, December 2015

Abstract

Diabetes Mellitus type 2 (T2DM) is the most common form of diabetes. More than 29 million people in the United States are affected by T2DM and another 86 million have prediabetes, a condition that exhibits high risk to progress into diabetes. Many T2DM cases can be prevented or avoided by improved awareness and lifestyle adjustments. Our project aims to improve T2DM diagnosis methodologies using supervised machine learning algorithms trained on electronic medical records (EMR).

Dataset

This project uses a publicly available EMR dataset released by Practice Fusion in 2012 for a Kaggle competition. It consists of electronic health records for 9,948 patients, among whom 1,904 have been diagnosed with DMT2. The data is very diverse in nature, ranging from physical attributes and demographics to personal habits such as smoking, medication history, *etc.*

Data Prep / Feature Selection

In the provided data, many features have missing or invalid values. All such cases and outliers were removed from analysis. The data was also skewed towards negative examples, and the *UnbalancedDataset* package (https://github.com/glemaitre/UnbalancedDataset) was used to perform resampling and balancing. Feature selection was done based on domain knowledge of team members (2 out 3 have background in biological/medical fields), literature review, and finally bottom-up and top-down sequential feature selection.

Methods

SVM with various kernels, Adaptive Boosting with Decision Trees, Random Forests, PCA, and Logistic Regression

PCA Results

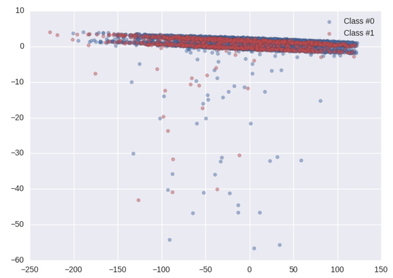

Figure 1: Raw Data Set

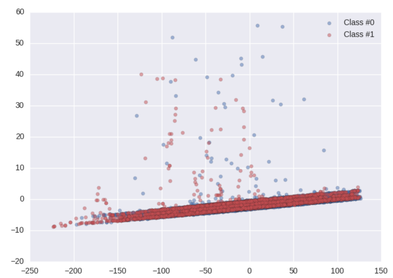

Figure 2: Undersampled Data Set

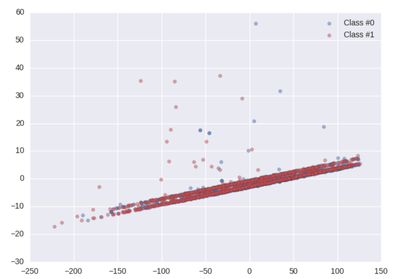

Figure 3: Oversampled Data Set

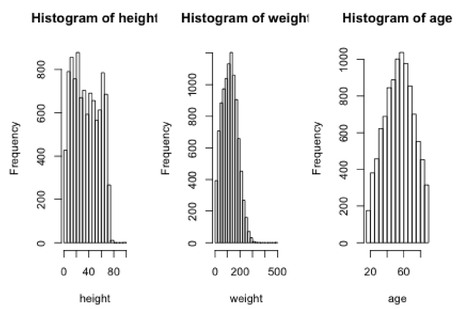

Figure 4: Oversampled Data Set

Model Evaluation Results

Table 1: SVM Results

Method	Kernel	C	Precision	Recall	F1 Score
SVM	RBF	0.1	0.1964	0.5431	0.2885
SVM	RBF	1.0	0.2212	0.5644	0.3183
SVM	RBF	10.0	0.1809	0.7110	0.2884
SVM	3rd Degree Poly	0.1	0.2127	0.6710	0.3230
SVM	3rd Degree Poly	1.0	0.3554	0.4510	0.3975
SVM	3rd Degree Poly	10.0	0.3554	0.4510	0.3975
SVM	Linear	0.1	0.2231	0.6543	0.3327
SVM	Linear	1.0	0.2231	0.6543	0.3327
SVM	Linear	10.0	0.2231	0.6543	0.3327
SVM	Sigmoid	0.1	0.1975	0.5414	0.2894
SVM	Sigmoid	1.0	0.2355	0.6968	0.3520
SVM	Sigmoid	10.0	0.2355	0.6968	0.3520

Table 2: Random Forest and AdaBoost Results

Method	Precision	Recall	F1 Score
RF Bottom-Up Undersampled	0.2894	0.7710	0.4208
RF Top-Down Undersampled	0.3333	0.4809	0.3938
RF Bottom-Up Oversampled	0.3345	0.7303	0.4588
RF Top-Down Oversampled	0.3307	0.7430	0.4577
Ada Bottom-Up Oversampled	0.3484	0.7277	0.4712
Ada Top-Down Oversampled	0.3515	0.6896	0.4656

Table 3: Logistic Regression with Lasso Regression (LRLR)* Results

Method	Cutoff Probability	Precision	Recall	F1 Score
LRLR	0.32	0.3086	0.8585	0.4673
LRLR	0.36	0.3256	0.7971	0.4624
LRLR	0.40	0.3425	0.7594	0.4721
LRLR	0.44	0.3636	0.7170	0.4825
LRLR	0.48	0.3782	0.6887	0.4883
LRLR	0.50	0.3857	0.6604	0.4870
LRLR	0.52	0.4006	0.6462	0.4946
LRLR	0.56	0.4040	0.5755	0.4747
LRLR	0.60	0.4104	0.5189	0.4583

* The value of the best lambda (the coefficient of the regularization term) is 0.001322.
This is determined by cross-validation.

Learning Curves

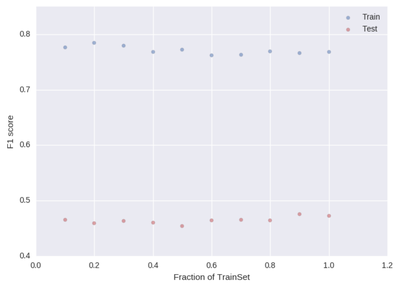

Figure 5: AdaBoost, Raw Data Set

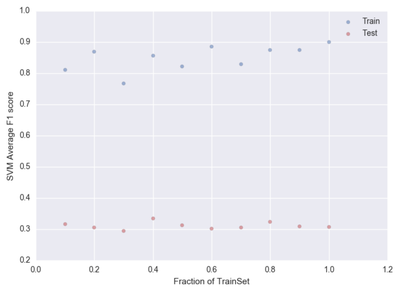

Figure 6: SVM, Raw Data Set

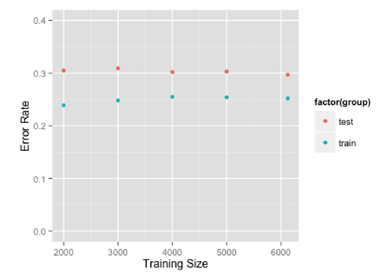

Figure 7: Logistic Regression, Raw Data Set

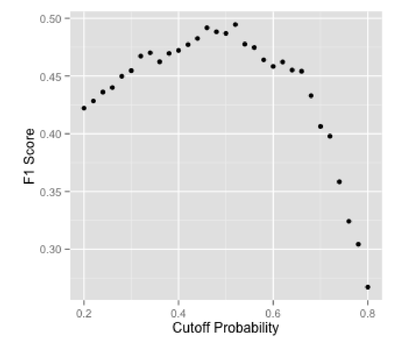

Figure 8: Logistic Regression, Raw Data Set