Onset of Type-II Diabetes Mellitus (T2DM) Predictions

Qihang Wu [1]

[1]Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Abstract

Diabetes is a global health issue with high incidence and mortality rates, affecting millions of people worldwide. Effectively predicting diabetes using clinical and demographic data can enhance early detection and prevention strategies. This project evaluated eight machine learning techniques, including logistic regression (both standard and penalized), linear discriminant analysis, naïve Bayes, decision trees, boosted trees, random forests, and support vector machine with linear kernels. We use grid search and 10-fold cross-validation to obtain the optimized tuning parameters for each model, if possible. Our results show that penalized logistic regression and SVM with a linear kernel performed best, achieving AUCs of 0.96 and 0.957 on our test dataset, respectively. These models have proven to be robust and reliable in predicting diabetes, demonstrating their potential for practical deployment in clinical settings to help manage this prevalent disease.

Introduction and Problem Statement

Type-II Diabetes Mellitus (T2DM) is the most prevalent form of diabetes and arises from the body's ineffective use of glucose. Over time, this results in excessive glucose levels in the bloodstream, which can eventually lead to serious complications affecting the nervous and immune systems. Accurate prediction of T2DM using advanced statistical and machine learning models can significantly achieve the public health benefits. Early detection facilitates timely interventions and personalized treatment strategies, enhancing patient outcomes. Consequently, this can decrease the overall incidence of T2DM and mitigate its complications. In this project, we would like to answer the following questions:

- **[Prediction Accuracy]** Can we accurately predict the onset of T2DM in at-risk populations via several statistical learning methods?
- **[Model Comparison]** What are the strengths and limitations of different modeling techniques?
- **[Optimal Modeling and Predictors]** Which model is preferred to make predictions of T2DM? Based on which metric(s)? What are the most significant predictors in this population? Any other insights from the data?

Data Description

The dataset is collected from Mendeley, a free and secure cloud-based data repository. The data is currently available online and is originally from the Vanderbilt datasets. It contains demographic factors (e.g., age, gender, height, and weight, etc) and some lab results (e.g., total cholesterol, glucose level, and HDL, etc) from a total of 390 (228 females and 162 males) rural African Americans in Virginia, USA. Data preparation steps include:

- Import the original data and omit the rows with NA values although the original data set doesn't contain NA values for any observation;
- Convert the response variable and the gender variable into factor variables for further data analysis;
- Exclude several variables including the patient number, a unique patient identifier;
- Separate the data into two data sets: **65%** training data and **35%** test data. Thus, the training data contains **14** predictors and **254** observations. The response variable is *fc_diabetes*.

We perform exploratory data analysis (EDA) with the correlation plot (**Figure 1**), the bar plot for gender, the box plot for cholesterol stratified by gender, and the box plot for all numerical predictors (**Figure 2**).

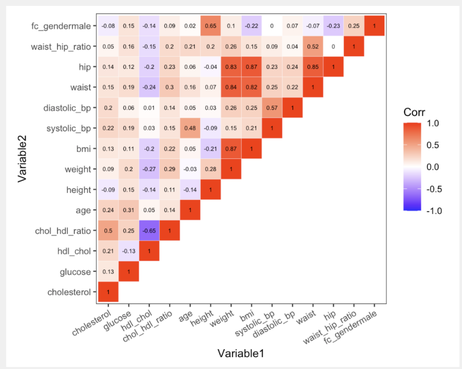

Figure 1. Correlation between Predictors

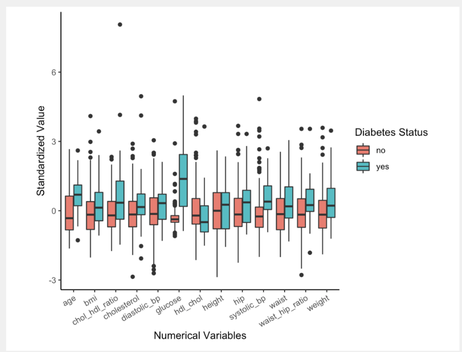

Figure 2. Box Plot for All Numerical Predictors (stratified by Diabetes Status)

Methods

We applied a suite of supervised learning methods to predict the occurrence of diabetes. We believe the chosen models represent a broad range of algorithmic complexity and interoperability. The following methods were implemented using the *caret* package in R, with their tuning parameters optimized via 10-fold cross-validation (CV).

1. **Logistic Regression (LR with and without penalization):** We began with a standard logistic regression model due to its simplicity, and particularly advantageous for binary classification problems. This model is also highly interpretable. To reduce the influence of less important predictors and improve the generalizability, we also adapt the penalty terms. The tuning parameters include 1) λ: the regulation term controlling the amount of shrinkage applied to coefficients, and 2) α: a parameter mixing between LASSO ($\alpha = 1$), Ridge ($\alpha = 0$), or intermediate values of α yielding Elastic Net. A sequence of 10 λ values equally spaced between 0.001 and 0.1.
2. **Linear Discriminant Analysis (LDA):** If the classes are well-separated, the parameter estimates from logistic regression will be unstable. In this case, linear discriminant analysis might be more appropriate when the distribution of predictors is normal in each of the classes, although it is unnecessary. In the analysis, we also provide a multiple-figure array which shows the classification of observations for every combination of two predictors. There is no tuning parameter for this model.
3. **Naïve Bayes (NB):** It is a simple and efficient algorithm. Meanwhile, it is not sensitive to irrelevant predictors. Here we compare the method using the Gaussian kernel and the kernel density estimator, without Laplace smoothing and a default bandwidth setting.
4. **Classification Tree and Boosted C-Tree:** In general, a single decision tree is simple, rule-based, and useful for interpretation. We consider a range of complexity parameter (cp) and the Gini index, a measure of node purity, i.e., a small value indicates that a node contains predominantly observations from a single class or a good split. However, a single tree generally does not have the same level of predictive accuracy as some of the other classification methods. Therefore, we use one of the ensemble methods: gradient boosting, where the trees are grown sequentially. Four parameters including the number of decision trees (n.trees), the maximum depth of each individual tree (interaction.depth), learning rate (shrinkage), and the minimum number of observations that must exist in a terminal node of the tree (n.minobsinnode), are selected using a grid-search strategy.
5. **Random Forest (RF):** This method is known for high accuracy and robust performance across diverse nature of predictors. It combines the predictors of multiple decision trees and reduce the risk of overfitting. However, it is difficult to interpret and can be computationally expensive. One key parameter is the number of randomly selected predictors (mtry) and a fresh selection is about $\sqrt{p}$ typically, where is the number of predictors.
6. **Support Vector Machine with Linear Kernel (SVMl):** It is a method that finds a plane that separates the class in feature space. For this reason, the decision function is fully specified by a small subset of training data, i.e., the support vectors.

We primarily evaluated the above models using accuracy and probably, the area under the ROC curve (ROC AUC). Accuracy provides an intuitive measure of overall correctness. For the validation of test error estimates, we rely on the 10-fold CV on our training set. By averaging performance across the 10 folds, we finalized the optimal parameters for each method and used them to predict the class on the test set.

- **Variable Selection:** In this project, we performed variable selection via the coefficients shrinkage method, especially setting some coefficients to zeros. For the completeness of the rest of analysis, we continued using all intended predictors from the data.
- **Variable Importance (VIP), Partial Dependence Plots (PDPs), and Individual Conditional Expectation (ICE) curves:** While the VIP identifies the predictors with the largest overall impact on the model's predictive performance, PDPs show how the predicted outcome changes as one single predictor varies when holding the other constant. Moreover, we provided ICE curves, which illustrated the dependence of predicted response on one selected predictor for each observation separately. These methods provide global interpretations for some black-box models like the random forests.

Results (1)

- The predictive performance across methods is shown in **Table 1** below.

Model and tuning parameters	10-fold CV Accuracy	10-fold CV Kappa
Standard LR	0.8943077	0.5567085
LR with penalty *(α=1, λ=0.034)*	0.9135385	0.5554446
LDA	0.9215385	0.6446788
NB *(Gaussian kernel, no Laplace smoothing)*	0.8861538	0.5695121
C-Tree *(cp=0.4786325)*	0.9255385	0.6815668
Boosted C-Tree *(shrinkage=0.1, interaction.depth=4, n.minobsinnode=10, n.trees=200)*	0.9173846	0.6596235
RF *(mtry=9, splitrule=gini, min.node.size=10)*	0.9098462	0.6309157
SVMl (C=0.006737947)	0.9216923	0.6561420

Table 1. Model performance Assessed by the 10-fold CV Accuracy

Results (2)

- **Variable Selection:** Although Ridge regression does not perform the variable selection, we selected LASSO regression with $\alpha = 1$ and performed the variable selection and coefficients estimation. The LASSO regression eventually selected only four predictors including *glucose*, *the ratio of total cholesterol to good cholesterol*, *age*, and *the upper number of blood pressure*.
- **VIP, PDPs, and ICE curves (Figure 3):** The random forest VIP demonstrates the *glucose* is by far the most influential predictor in determining diabetes, dwarfing the importance of all other predictors. Predictors like *systolic blood pressure* and *age* also contribute substantially. The PDP shows that when glucose levels are low, the model strongly favors the probability of being non-diabetic over diabetic. As glucose increases, there will be a higher chance of having diabetes overall. The ICE curves prove the consistent relationship across individuals.

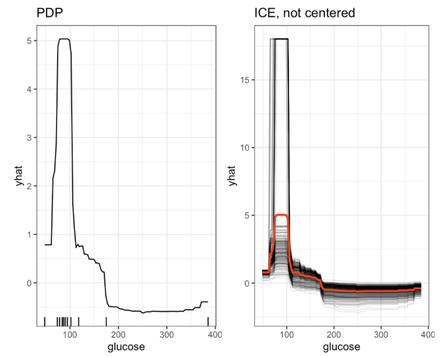

Figure 3. PDP and ICE Curves of Glucose Level

- **Test Performance (Figure 4):** After making the predictions on the test set, ROC Curves and AUC values indicate that most methods performed well. Among them, the penalized logistic regression (AUC 0.96) and SVM with the linear kernel (AUC 0.957) stand out as the top performers, while the classification tree is the weakest one when applied to the unseen test set.

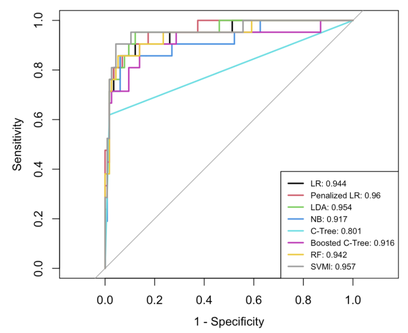

Figure 4. ROC Curves and AUC Values on Test Data Set

Discussion

- Among the methods investigated, the penalized logistic regression and SVM with linear kernel performed best. The former combats the overfitting issue by shrinking some coefficients towards zero, thereby reducing the model complexity and improving the generalizability. The linear kernel demonstrated that the classes can be easily separated by a straight line, i.e., a hyperplane in higher dimensions.
- Although Yan Y et al (2022)[1] concluded that the similar predictive performance between the standard regression and penalized regression, we have shown better performance regarding the penalized logistic regression model. Eilers P.H.C. et al (2001)[2] demonstrated the desirable performance of penalized logistic regression on the MIT ALL/AML data. While many considerations are required in the real-world data, these proposed models appear sufficiently reliable and could be practical choices in practice.
- The main challenge involved is the limited sample volume, which reduced the statistical power and increased the risk of overfitting. Moreover, we cannot deploy complex models like neural networks or extensive random forests as these models require large datasets. In this case, we may consider more external data in the future for further explorations.

References

[1] Yan, Y., Yang, Z., Semenkovich, T. R., Kozower, B. D., Meyers, B. F., Nava, R. G., Kreisel, D., Puri, V. (2022). Comparison of standard and penalized logistic regression in risk model development. JTCVS Open, 9, 303-316. https://doi.org/10.1016/j.xjon.2022.01.016.

[2] Eilers, P. H. C., Boer, J. M., van Ommen, G.-J., van Houwelingen, H. C. (2001). Classification of microarray data with penalized logistic regression. Proc. SPIE 4266, Microarrays: Optical Technologies and Informatics, June 4.

Acknowledgement

Heartfelt thanks to Dr. Adam Ciarleglio and Erika Hubbard for their outstanding teaching and commitment throughout this semester!