Prediction of Diabetes Through Supervised Machine Learning

Abstract

Diabetes is rapidly reaching epidemic proportions in America with roughly 1.4 million new diagnoses every year [1]. Since complications associated with diabetes are the leading cause of mortality, promptly identifying patients with warning signs not only can help reduce exacerbations of diabetes but also avoid potential high cost in the long run. In this study, different classifiers were used to predict the incidence of diabetes. Model performance was evaluated in terms of classification accuracy, AUC, speed and memory usage. My finding indicates that random forests and generalized boosted trees outperformed SVM and decision trees with regards to overall discriminative ability (AUC). All models show improvement in predictive power when applying resampling method (SMOTE) to the original training data.

Introduction

Diabetes was reported to affect 29 million Americans and remains the leading cause of death among U.S population [1]. Globally, the number of people with diabetes has increased to 422 million in 2014 from 108 million since 1980 [9]. Diabetes is a chronic condition in which the body is unable to regulate blood sugar levels. As a result, abnormally high level of glucose remains in the bloodstream. If left uncontrolled, diabetes can lead to long-term complications, such as blindness, heart disease, stroke, nerve damage, and kidney failure [2]. These complications are often linked to increased morbidity and the mortality rate among diabetic individuals is almost double compared to those without diabetes. In addition, the cost associated with this growing health epidemic is also staggering with global spending estimated to reach 825 billion dollars every year [9]. Due to the gravity of diabetes and its complications, early detection of individuals who are at an increased risk for the development of diabetes is an important issue that should not be underestimated.

Predictive models provide vast benefit in guiding clinical decision-making for physicians in which an individual's health outcome can be predicted based on a combination of patient risk factors. Those predictions could enable timely and effective modifications to lifestyle or medical care for high-risk patients [10]. Several machine learning techniques have been widely used for the prediction of disease risks. For example, Yu et al. established SVM (support vector machine) models to classify persons with diabetes and without diabetes [5]. Khalilia et al. employed different classification methods, such as SVM, bagging, boosting and random forest to predict individual risk of eight chronic diseases [11]. Lee et al. applied tree-based models to identify individuals with undiagnosed diabetes [4]. Both of the latter studies also addressed the class imbalance problem.

Class imbalance arises when one class is underrepresented compared to the other class. This problem often results in unsatisfactory performance of the learning algorithms where the classification is skewed toward the large class. Khalilia and colleagues found that random forest ensemble learner yielded higher accuracy compared to the rest of the learners, when used in conjunction with repeated random sub-sampling [11]. Similar results were found in Lee's study where the prediction power of the models is improved when using either over-sampled or under-sampled datasets [4]. Both over-sampling (the small class) and under-sampling (the large class) are common resampling approaches used to overcome class imbalance. Another approach is called SMOTE (Synthetic Minority Over-sampling Technique). This method combines oversampling of the small class from which synthetic samples are generated through nearest-neighbor searches as well as random under-sampling of the large class [12]. The combination tends to outperform under-sampling technique while reduces overfitting often associated with random over-sampling.

In the current study, I aim at examining the performance of different classifiers in terms of prediction accuracy and the cost of classification (e.g. speed and memory usage). Machine learning algorithms including SVM, decision tree, random forest and generalized boosting are applied to predict the incidence of diabetes based on a set of predetermined risk factors commonly associated with the diabetes risk. A 2013-2014 dataset from the National Health and Nutrition Examination Survey (NHANES) is used to generate these algorithms. The effect of resampling with SMOTE is also investigated. The dataset has an imbalance rate of 17% (the percentage of the individuals with diabetes). I hypothesized that the resampling strategy will increase accuracy in diabetes prediction for these models.

Methods

Data preparation

Data source

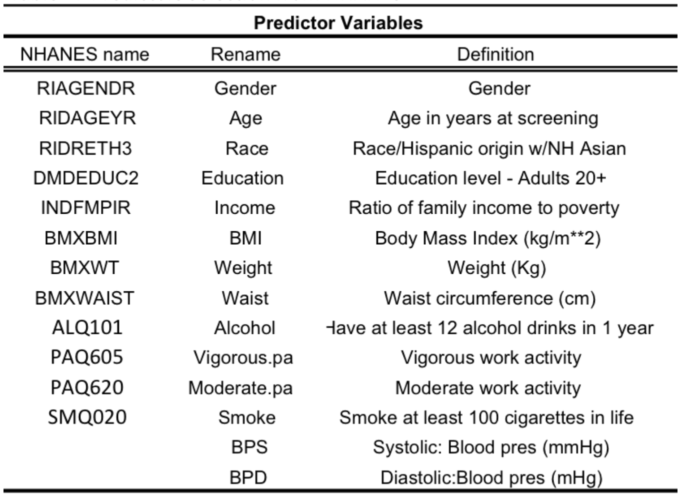
National Health and Nutrition Examination Survey (NHANES) 2013–2014 dataset was used in this study to generate predictive models [3]. Conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention, the biannual sample survey was designed to evaluate the health and nutrition status of the United States population. Out of 10175 participants who completed the survey, only non-pregnant participants with aged 20 and above, and with at least two blood pressure readings were included in the analysis. This results in a final sample of 5297 respondents.

Variable selection

A participant was classified to have diabetes if he/she met any of the following criteria 1) answering “yes” to the question “Doctor told you have diabetes”, 2) fasting glucose ≥ 126 mg/dL, 3) non-fasting glucose ≥ 200 mg/dL, 4) hemoglobin A1c ≥ 6.5% [4,5]. Participants with and without diabetes were coded as “Yes” and “No”, respectively in a new variable called “response”.

A total of 14 attributes representing commonly associated diabetes risk factors were used for classification in the study [4,5]. These include gender, age, race, education level, BMI and blood pressure. A complete set of predictors was shown in Table I. Variables BPS and BPD represent the average of two or three blood pressure readings obtained from the participants.

Table 1. Predictors selection from NHANES



Predictor Variables		
NHANES name	Rename	Definition
RIAGENDR	Gender	Gender
RIDAGEYR	Age	Age in years at screening
RIDRETH3	Race	Race/Hispanic origin w/NH Asian
DMDEDUC2	Education	Education level - Adults 20+
INDFMPIR	Income	Ratio of family income to poverty
BMXBMI	BMI	Body Mass Index (kg/m**2)
BMXWT	Weight	Weight (Kg)
BMXWAIST	Waist	Waist circumference (cm)
ALQ101	Alcohol	Have at least 12 alcohol drinks in 1 year
PAQ605	Vigorous.pa	Vigorous work activity
PAQ620	Moderate.pa	Moderate work activity
SMQ020	Smoke	Smoke at least 100 cigarettes in life
	BPS	Systolic: Blood pres (mmHg)
	BPD	Diastolic:Blood pres (mHg)

Model generation and evaluation

Predictive modelling was conducted using R (version 3.2.3). 70% of the data (N = 3709, of which 640 had diabetes) was randomly selected for training. The remaining 30% (N =1588, of which 273 had diabetes) was used as testing set. Prior to data split, missing values[1] were replaced with column medians for numeric attributes or with most frequent factor level for categorical attributes using rough imputation method ("na.roughfix" function in *randomForest* package). Learning algorithms, including support vector machines (SVM), decision trees, random forest (RF) and generalized boosting (GBM) were employed to generate predictive models using the *caret* R package. A 10-fold cross-validation with three repeats was used in the training data set to estimate predictive accuracy of each model.

SVM algorithm performs classification by constructing a hyperplane with maximized margin in order to optimally separate between two classes [7]. High discriminative power is achieved by mapping observations into a multidimensional space using either linear or non-linear kernel functions. In the current study, two SVM kernel functions, including linear and radial basis functions (RBF) were tested. Both training datasets were scaled and centered before SVM classification.

Classification and regression tree (CART) algorithm was used to create tree-based classification models. CART tree is generated by recursive partitioning of the data into more homogeneous subset based on variable importance [8]. *rpart* inside the *caret* package was used to build the decision trees. The complexity parameter and minimum number of observations in a node were set at 0.01 and 20, respectively. In addition, random forest and GBM in the *caret* package were used to generate the ensemble models. Both methods are extension of CART by creating multiple decision trees and aggregate their results to improve prediction accuracy [4].

In addition to the original training set, the models were fitted using a balanced training dataset generated by SMOTE algorithm (DMwR package). The resampling method was employed in order to test whether the predictive power would be improved with balanced class distribution.

Overall model performance was compared using evaluation methods, such as the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, sensitivity, specificity, positive predictive value (PPV), and negative predictive value.

Result and Discussion

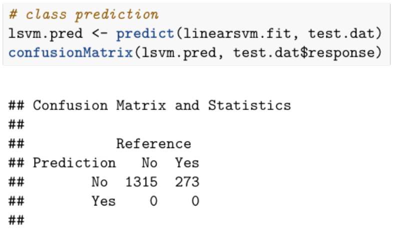
SVM classifier

In this study, a dataset of 5297 individuals was obtained from NHANES 2013-2014 dataset, with 913 diabetic individuals and 4384 non-diabetic individuals based on the selection criteria (in the method section).
The sample pool consisted of 2574 males and 2723 females between 20 and 80 years old. The original data set was randomly split (70:30) into a training and a test set.

For SVM classifier, two kernel functions were used to generate the predictive models and their performance was evaluated. With the imbalanced dataset (i.e. without resampling), the RBF function performed slightly better while the linear kernel function was not able to separate the two classes (with vs without diabetes) (Table 2). When fitting the models on the balanced training set, I found that both kernels performed roughly equally better than those trained on the original dataset. A total of 178 out of 273 diabetes cases were correctly identified with linear kernel and 166 cases were correctly identified with RBF kernel. The balanced accuracy for linear and RBF models increases from 50% and 54% to 71% and 70% (Table 7). It's worth noting here that balanced accuracy was used for evaluation instead of accuracy since the former takes class imbalance into account [6]. The overall discriminative ability assessed by the area under ROC curve for each SVM model was shown in Table 3 and Figure 1.

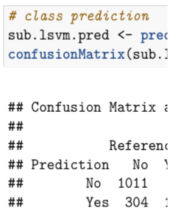
[1] Non-responses of "Don't know" and "Refuse to answer" were recoded as NA in this study.

Table 2. Confusion matrix for SVM with linear (upper) and radial (lower) kernel. Models were trained on the original dataset (left) and the balanced dataset (right).



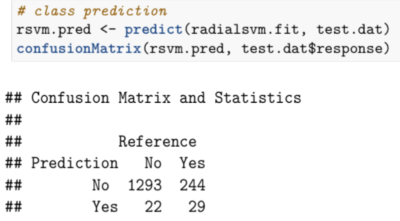
```
# class prediction
lsvm.pred <- predict(linearsvm.fit, test.dat)
confusionMatrix(lsvm.pred, test.dat$response)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction    No  Yes
##         No  1315  273
##         Yes    0    0
##
```



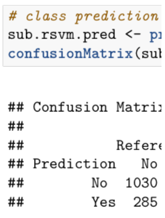
```
# class prediction
sub.lsvm.pred <- pre
confusionMatrix(sub.
```

```
## Confusion Matrix a
##
##             Referen
## Prediction    No
##         No  1011
##         Yes  304
```



```
# class prediction
rsvm.pred <- predict(radialsvm.fit, test.dat)
confusionMatrix(rsvm.pred, test.dat$response)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction    No  Yes
##         No  1293  244
##         Yes   22   29
```



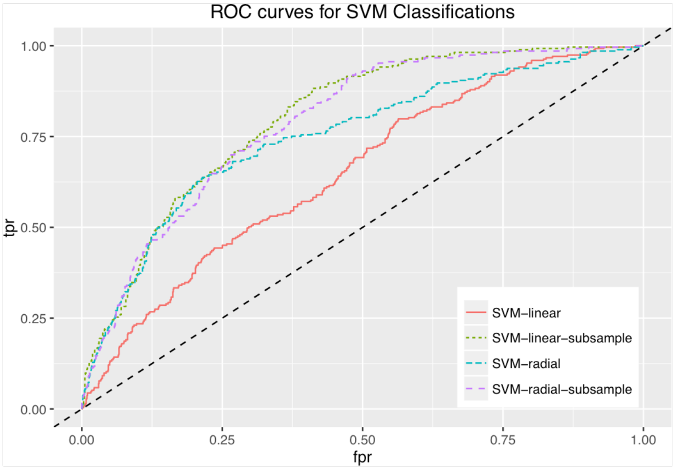
```
# class prediction
sub.rsvm.pred <- p
confusionMatrix(su
```

```
## Confusion Matri
##
##             Refere
## Prediction    No
##         No  1030
##         Yes  285
```

Table 3. Performance of SVM models

Model	Area under ROC curve	
	Linear	Radial basis
Original trainin set	0.65	0.75
Balanced trainin set	0.8	0.79

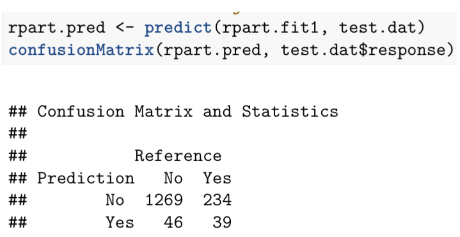
Fig 1. ROC curves for SVM classifications



Decision Tree classifier

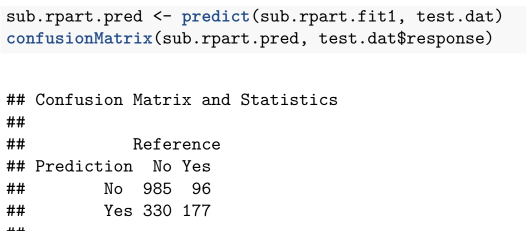
Figure 2 shows the decision tree models fitted on the original and balanced training datasets. Both models included age, waist circumference, race, BPD and income predictor variables for generating the trees. Furthermore, using a balanced training dataset seems to result in a less complex decision tree. Performance of the decision tree model with imbalanced data was found to be similar with that of RBF-SVM model. Both have an AUC value of 0.75 and a balanced accuracy of 0.55 (Table 7). Similar to SVM, the decision tree model also improves classification power when it was trained on the balanced dataset.

Table 4. Confusion matrix for Decision Tree. Models were trained on the original dataset (left) and the balanced dataset (right).



```
rpart.pred <- predict(rpart.fit1, test.dat)
confusionMatrix(rpart.pred, test.dat$response)
```

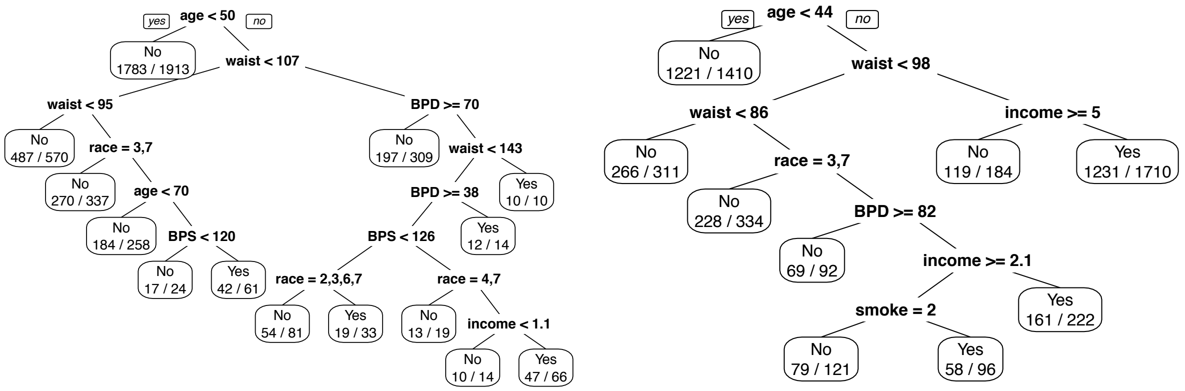
```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction    No  Yes
##         No  1269  234
##         Yes   46   39
```



```
sub.rpart.pred <- predict(sub.rpart.fit1, test.dat)
confusionMatrix(sub.rpart.pred, test.dat$response)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   No Yes
##         No  985  96
##         Yes 330 177
```

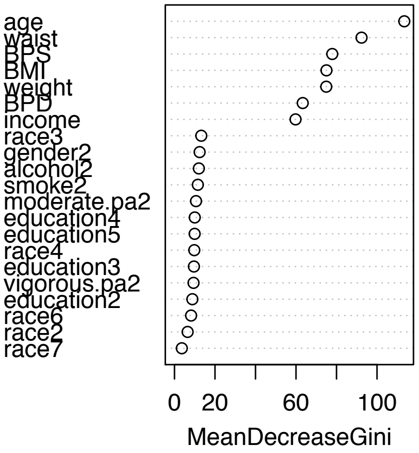
Fig 2. Decision tree model with the original training set (left) and balanced set (right). Values at the node: number of correct classifications/number of observations in the node.



Random Forest classifier

The variable importance plot (Fig 3.) shows that the top five most important predictors for classification were age, waist circumference, BPS, BMI and weight for both training sets, as measured by the mean decrease Gini scores. Random forest classifier with original training set performed poorly and yielded a sensitivity of 3% but increased to 56% when trained on the balanced set. The balanced accuracy was improved from 51% to 69%. The running time was slow compared to the other models (Table 7).

Fig 3. Mean decrease in node impurity for Random Forest model trained on the original dataset (left) and the balanced dataset (right).



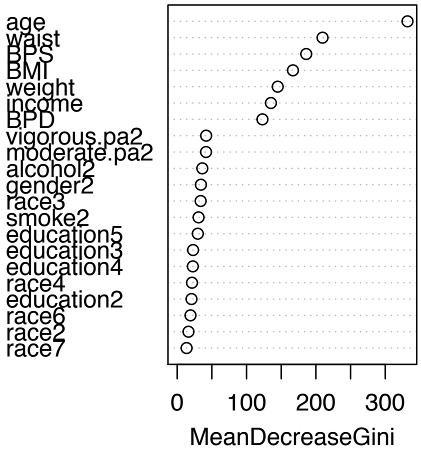
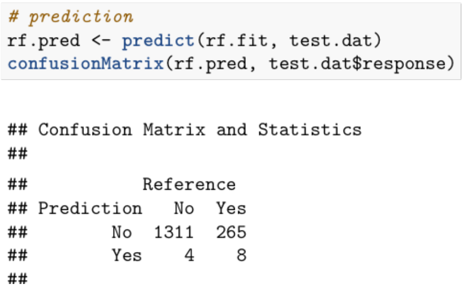
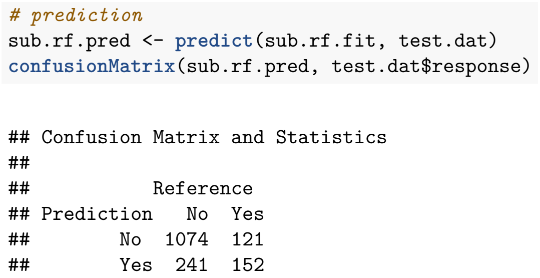


Table 5. Confusion matrix for Random Forest. Models were trained on the original dataset (left) and the balanced dataset (right).



```
# prediction
rf.pred <- predict(rf.fit, test.dat)
confusionMatrix(rf.pred, test.dat$response)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1311  265
##        Yes    4    8
##
```



```
# prediction
sub.rf.pred <- predict(sub.rf.fit, test.dat)
confusionMatrix(sub.rf.pred, test.dat$response)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1074  121
##        Yes  241  152
```

Generalized boosted trees

The variable importance for each GBM model was shown in Figure 4. Both include age, waist circumference and BPS among the top 5 most important predictors. Similar to the decision tree and random forest, age remains the most important variable associated with the outcome prediction. This agrees with the result reported in Khalilia's study in which they found age carried the most weight in diabetes prediction [11]. The best fit models on the original and resampled data have a number of trees equal to 100 and 50, respectively. Both have a tree depth of 3 and a shrinkage rate of 0.1. The performance of GBM model was found to be the best in terms of identifying participants with diabetes. The algorithm yielded the highest sensitivity and balanced accuracy among all the models tested on the original training set. The running time is much shorter compared to that of random forest ensemble learner. The resampling method also improved the balanced accuracy from 58% to 71%.

Fig 4. Variable importance for GBM model trained on the original dataset (left) and the balanced dataset (right).

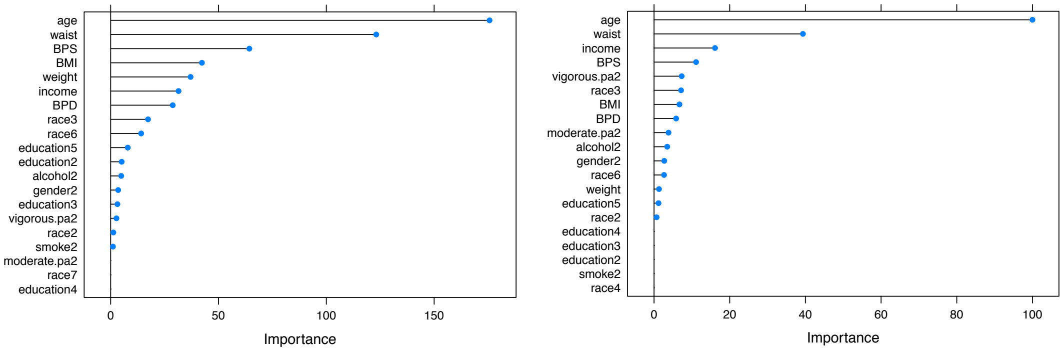
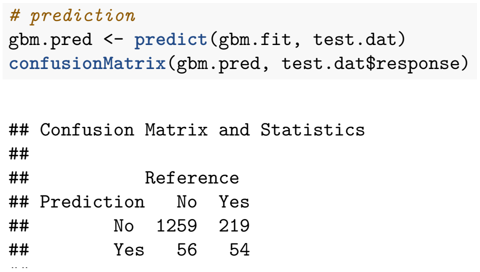
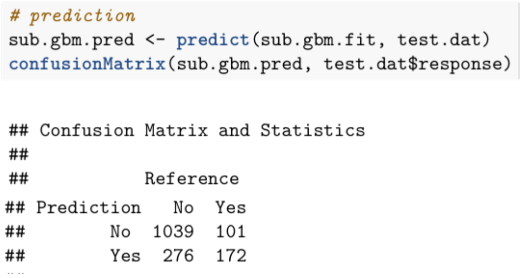


Table 6. Confusion matrix for GBM Models were trained on the original dataset (left) and the balanced dataset (right).



```
# prediction
gbm.pred <- predict(gbm.fit, test.dat)
confusionMatrix(gbm.pred, test.dat$response)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1259  219
##        Yes   56   54
```



```
# prediction
sub.gbm.pred <- predict(sub.gbm.fit, test.dat)
confusionMatrix(sub.gbm.pred, test.dat$response)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1039  101
##        Yes  276  172
```

Table 7. Model performance in the testing dataset

Model	Total run time (sec)	Memory usage (MB)	Sensitivity	Specificity	PPV	NPV	AUC	Balanced accuracy
Linear SVM								
original	1.92	247.9	0	1	NaN	0.8281	0.6522	0.5
balanced	5.10	515.6	0.652	0.7688	0.3693	0.9141	0.8008	0.71
Radial SVM								
original	2.28	457.1	0.1062	0.9833	0.5686	0.8412	0.7496	0.5447
balanced	4.36	997.6	0.6081	0.7833	0.3681	0.9059	0.7907	0.6957
Decision Tree								
original	0.8	99.6	0.1429	0.965	0.4588	0.8443	0.7567	0.5539
balanced	1.48	458.8	0.6486	0.749	0.3491	0.9112	0.74	0.6987
Random Forest								
original	9.32	372.7	0.0293	0.997	0.6667	0.8319	0.7976	0.5131
balanced	12.14	779.6	0.5568	0.8167	0.3868	0.8987	0.7916	0.6868
GBM								
original	1.16	274.3	0.1978	0.9574	0.4909	0.8518	0.8092	0.5776
balanced	1.68	665.4	0.63	0.7901	0.3839	0.9114	0.8011	0.71

Conclusion

This study illustrates the ability of different classifiers in predicting the occurrence of diabetes. Overall, I found that decision trees perform the best in terms of speed and memory usage. Both GBM and random forest models outperform SVM and decision trees when trained on the original training set with regards to discriminative ability (AUC value), although the sensitivity of the random forest model is low. All models show improvement in predictive power and have similar balanced accuracy (ranged from 69% to 71%) when applying resampling method on the training data. This is consistent with the previous findings in regard to the effectiveness of resampling approaches [4]. This study also further demonstrates the feasibility of using resampling with SMOTE to enhance the prediction accuracy of the rare class.

Reference

1. Statistics About Diabetes. American Diabetes Association. Available at: http://www.diabetes.org/diabetes-basics/statistics/. Accessed February 28, 2016.

2. Basics About Diabetes. Centers for Disease Control and Prevention 2015. Available at: http://www.cdc.gov/diabetes/basics/diabetes.html. Accessed February 28, 2016.

3. NHANES 2013-2014. NHANES -. Available at: http://wwwn.cdc.gov/nchs/nhanes/search/nhanes13_14.aspx. Accessed April 8, 2016.

4. Lee P. Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets. International Journal of Environmental Research and Public Health IJERPH 2014;11(9):9776–9789. doi:10.3390/ijerph110909776.

5. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak BMC Medical Informatics and Decision Making 2010;10(1):16. doi:10.1186/1472-6947-10-16.

6. Accuracy and precision. - Wikipedia, the free encyclopedia. Available at: https://en.m.wikipedia.org/wiki/accuracy_and_precision#iso_definition_.28iso_5725.29. Accessed April 10, 2016.

7. Cortes C, Vapnik V. Support-vector networks. Mach Learn Machine Learning 1995;20(3):273–297. doi:10.1007/bf00994018.

8. Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

9. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. The Lancet 2016;387(10027):1513–1530. doi:10.1016/s0140-6736(16)00618-8.

10. Collins GS, Mallett S, Omar O, Yu L-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine BMC Med 2011;9(1):103. doi:10.1186/1741-7015-9-103.

11. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak BMC Medical Informatics and Decision Making 2011;11(1):51. doi:10.1186/1472-6947-11-51.

12. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique, 2002; 16: 321-357