Predicting Diabetes Status in NHANES Data Set

Lakshmi Ganesan, Yingtong Liu, Brady Ryan

20/12/2020

Introduction

Diabetes is a top ten leading cause of death in the United States. It is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. In this project, we work on building prediction models to classify diabetes status in the National Health and Nutrition Examination Survey (NHANES) data set.

Data

The data set we use here is the National Health and Nutrition Examination Survey (NHANES) which examines a nationally representative sample of 10,000 people every two years from 1999-2018. The data set includes demographic, socio-economic, dietary and health-related data, along with medical, dental and physiological measurements, and laboratory tests. Findings from this survey are used to determine risk factors and prevalence of diseases, to develop public health policy, and expand health knowledge for the country.

For this project, we used NHANES data files from 1999 to 2018. We first linked together demographic data, questionnaire data, and part of examination data and dietary data, to create data sets for each two year collection period. We then created a final data set using variables that were commonly available across all data sets, and we removed columns that had more than 10,000 missing values. For the final dataset, we use a time frame from 2007 to 2018, because this maximizes the amount of available useful variables.

The variables we used to predict diabetes status are gender, age, race, ratio of family income to poverty, body mass index (BMI), high blood pressure, if the person had ever had asthma, whether the person had been taking medication for anemia in the past three months, regular/irregular pulse status during examination, self-reported value of how healthy the diet is, self-reported general health condition, minutes of sedentary activity per week, if the individual had seen a mental health professional in the last year, and if the individual had ever told a doctor they had trouble sleeping.

We then combined all the data tables from after 2007 and removed individuals with missing values in any of the 14 predictors or for diabetes status. We also removed individuals who reported "don't know", "refused", or had any other value besides the ranges described in the NHANES documentation in any of the variables. We used 70% of the final data for training and 30% for testing. This corresponded to 20,060 samples in the training set and 8,596 samples in the test set. The overall prevalence of diabetes in the final data was approximately 12.5%. To ensure valid model training and testing, we forced the prevalence of diabetes in the training set to be approximately equal to that of the test set.

Models and Results

We then chose to fit five different models. To do this we used naive Bayes classifier, gradient boosting machine (GBM), decision trees, random forests and gradient boosted decision trees (XGBoost). All five models are implemented in the caret package in R and we used the corresponding functions from that package to fit our models. To account for the imbalance in our dataset, we chose to use area under the ROC curve (AUC)

as a performance metric to both choose the best model in the training set, and to compare the subsequent models on the test set. This is a better metric by which to compare classifiers in the presence of imbalanced classes since it penalizes false positives and false negatives, both of which are very important to account for when classifying a serious disease such as diabetes. To further account for the imbalance, we chose to also fit weighted models as well as up-sampled, down-sampled, and synthetic minority over-sampling technique (SMOTE) models. Weights for cases were defined to be 1/(# of cases) * 0.5 = 2.01 x 10^(-4), and weights for controls were defined to be 1/(# of controls) * 0.5 = 2.85 x 10^(-5). The up-sampling method involves randomly duplicating observations from the "diabetes" class to better show its signal. The down-sampling method involves sampling less of the observations from the "no diabetes" class to again show the signal of the cases better. SMOTE works by generating new instances of the "diabetes" class by looking at the already input cases. Finally, we used repeated k-fold cross-validation with 5 folds and 5 repeats during model training to select the best model.

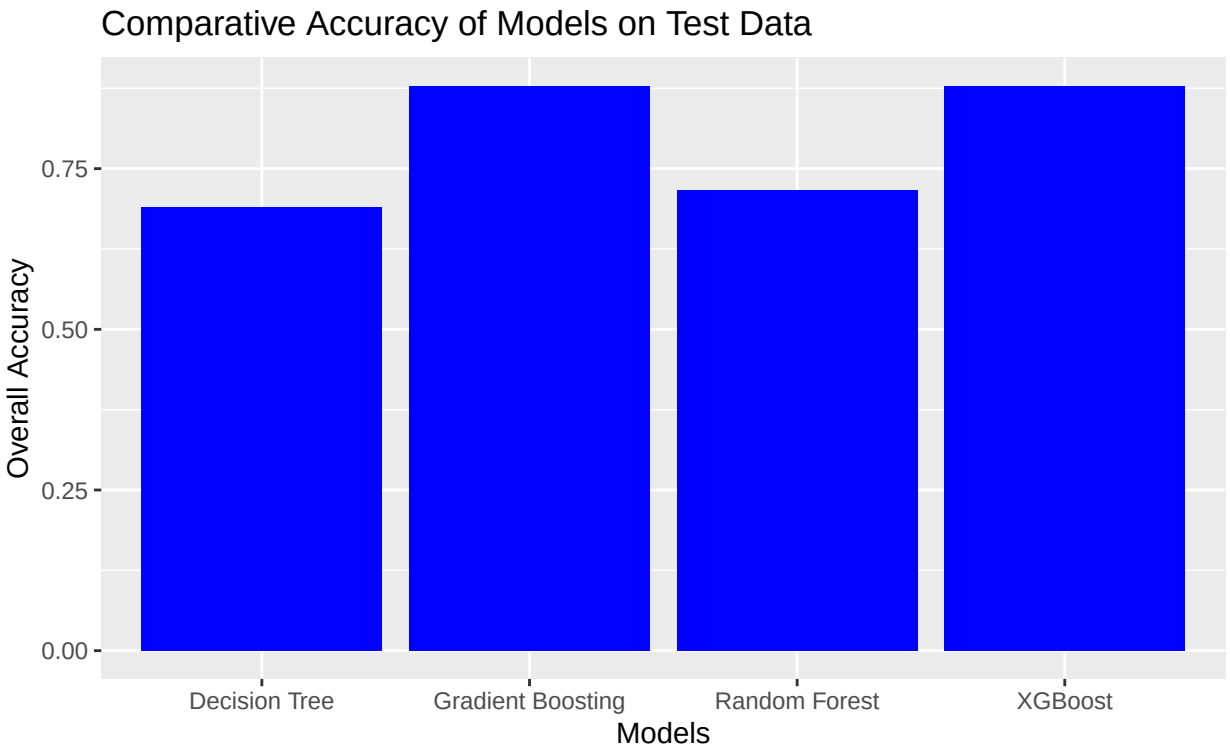
Primarily, we fit a naive Bayes classifier. It describes the relationship of conditional probabilities of statistical quantities. In Bayesian classification, we are interested in finding the probability of a label given some observed features. Applying on the original, the weighted model and the three resampling methods, we got AUC 0.831 for original. Even though the AUC scores suggest that it is a modest model for analyzing this dataset, after inspecting the predictions, we found the sensitivity is quite low and no information rate is higher than the accuracy, which implies the naïve Bayes classifier is improper for this dataset because of the inherited conditional dependence.

After fitting the naive Bayes classifier, we decided to fit a decision tree model. Decision trees are trees that contain nodes which split the data into subsets based on some condition. Decision tree machine learning algorithms decide the optimal way to split these nodes to best classify the data point. The underlying functions used by the caret function to fit decision trees are found in the rpart R package. As with all other models, we used weighting, up-sampling, down-sampling, and SMOTE to try and improve performance. Our original model had an AUC of 0.7429, with weighting giving 0.7645, down-sampling 0.7694, up-sampling 0.7663, and SMOTE 0.7693. Hence in this case the best method was to down-sample.

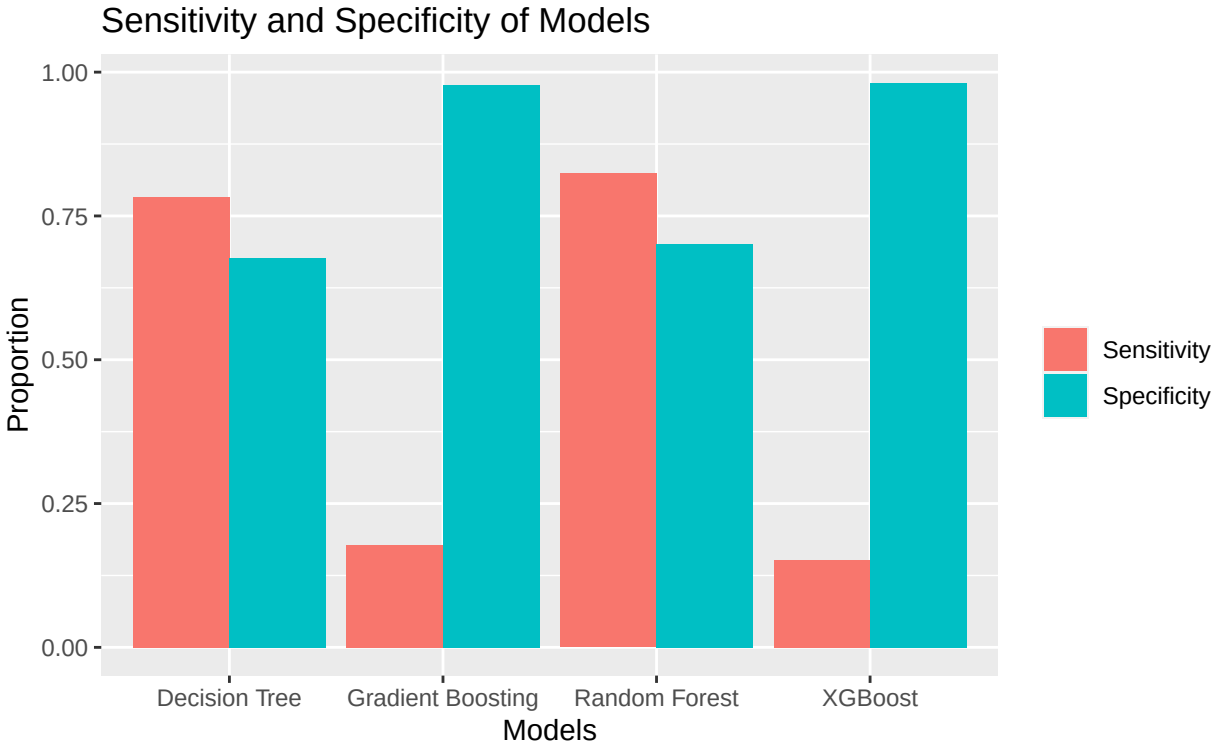
We then decided to fit a more complex model by using a gradient boosting machine. A GBM works by using an ensemble of decision trees to produce a more powerful model. It was hence expected that GBM would perform better than the decision tree model. Indeed, the original model had an AUC of 0.8363, with weighting giving 0.8258, down-sampling giving 0.8336, up-sampling giving 0.8363, and SMOTE giving 0.8266. Hence in this case the best method was simply the traditional GBM model.

Finally, we decided to also fit a random forest model. Random forests also use a collection of decision trees in order to classify data. The models used by caret can be found in the R package randomForest. We expected it to perform better than decision trees. Indeed, the original model gave an AUC of 0.82, weighting gave 0.824, down-sampling 0.8279, up-sampling 0.8269, and SMOTE giving 0.824. As with GBM, down-sampling gave the best results.

After fitting each of the models under the 5 different re-sampling/weighting techniques, we decided to take the model with the highest AUC from each of the models to do a further comparison. To do this, we looked at the confusion matrices of the models when predicting the test data set, their overall accuracy of prediction, and their sensitivity and specificity. The sensitivity of the prediction model summarizes how well the model predicts the "diabetes" class, whereas the specificity summarizes how well the model predicts the "no diabetes" class. Since the main goal of this project is to classify diabetes, we believe that more weight should be given towards the sensitivity of the model rather than the specificity. For comparison, we decided to leave out the naive Bayes classifier since the model simply predicted "no diabetes" for each observation except for one. Hence it is not a valid prediction model. Of the remaining four models, we plotted the prediction proportions as follows:



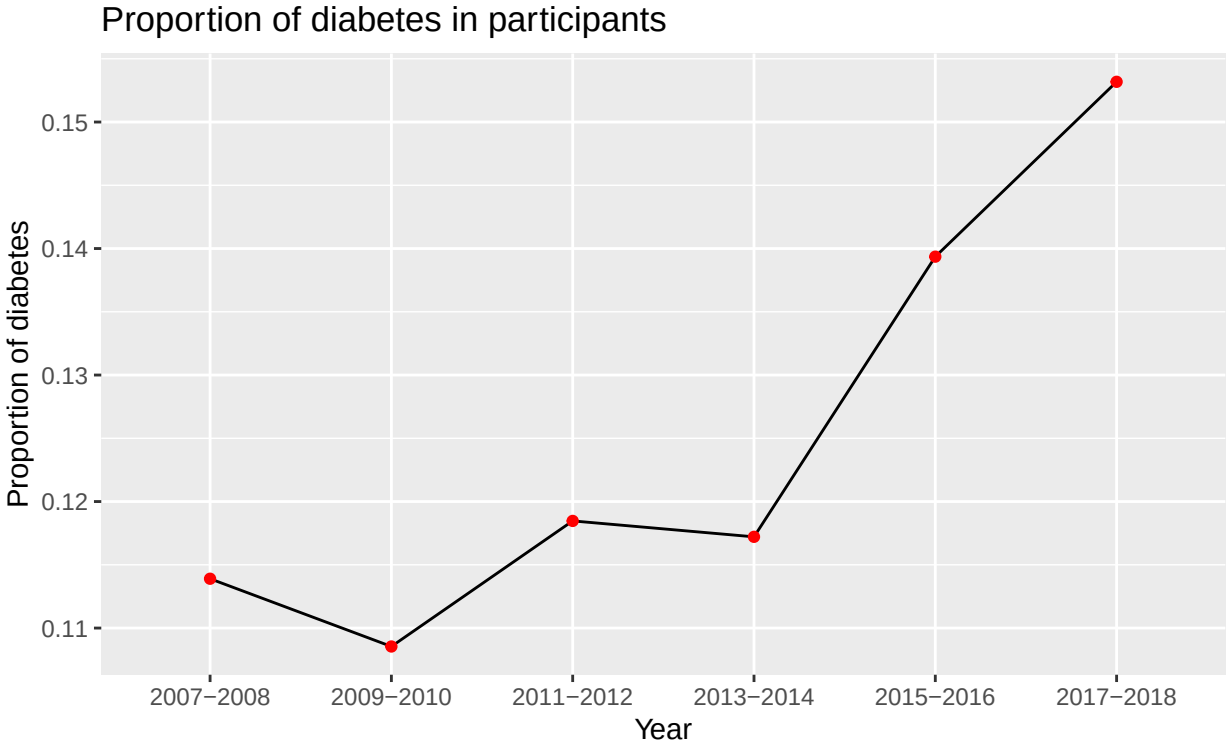
As we can see, XGBoost has the highest prediction at 0.8787, followed closely by GBM and then by random forest and decision trees. However, since we are doing imbalanced classification, the raw prediction probability is not the best measure for comparing models. To look more closely, we decided to look at the sensitivity/specificity of the models. Below is a side-by-side boxplot comparing these for each model.

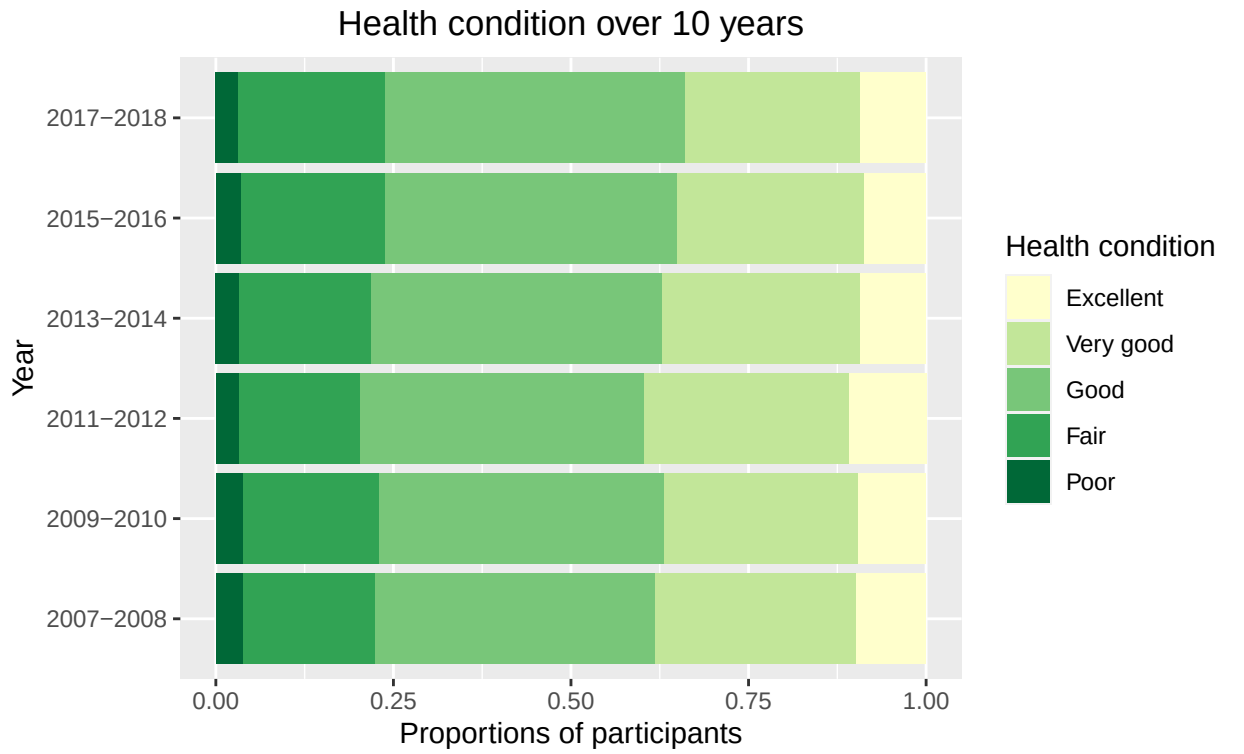


As can be seen in this plot, both the XGBoost and GBM models have very low sensitivity but high specificity. On the other hand, both decision tree and random forest have high sensitivity and reasonable specificity. In the models we have fit there is a tradeoff between predicting more true positives and the false positive rate. Hence although the random forest model was able to correctly classify 82.38% of the true diabetes cases, it incorrectly classified 29.82% of the controls as having diabetes. This is a very high false positive rate, and hurts the practical usage of this model. On the other hand, models such as XGBoost and GBM were unable to correctly classify a significant number of cases, but had very low false positive rates at only 1.833% and 2.233% respectively. However, the extremely low power of these models to correctly classify cases in our opinion makes them inferior to the random forest model.

Visualization

To give a general idea about the predicted value, diabetes, across ten years, we plotted the trend of the proportion of the participants who have diabetes at the time they were queried. We also examined health condition across ten years, which was categorized into five levels. There is no great change in the general health condition, but the proportion of diabetes participants has an overall increasing trend in these ten years. This was also done with diet health, and had a similar trend as overall health condition, but it is not shown for brevity.





Conclusion

To conclude, we believe that the random forest model can be a useful tool in a practical setting. All of the predictors are very easily obtained, and it correctly predicted true positives at a respectable rate. The down-sampled random forest model had AUC of 0.8279, overall prediction proportion of 71.7%, sensitivity of 82.38%, and specificity of 70.18%. It could potentially be used by individuals to attempt to classify their diabetes status. If classified as "yes", they could see a health professional for more formal testing. Due to relatively high false negative rates however, a negative result should not be taken to assume an individual does not have diabetes, or is not at high risk for diabetes.

There are some important limitations to what has been presented in this report. Primarily, although the NHANES data set is meant to be nationally representative, this requires oversampling from certain demographics groups. Hence the samples taken for the survey are not a simple random sample, and this may have introduced bias into some of our prediction models. Furthermore, our prediction method does not take into account the differences in type 1 and type 2 diabetes. The data in the NHANES data set reports only whether an individual states that they have diabetes, but does not distinguish between the two types. It is important to note this point since the two types may have different clinical manifestations and treatments. Finally, many of the data used in the prediction model were self-reported. Hence many of them may not be completely accurate as people may have difficulty judging their own health and diet qualities.

Finally, much future work could be conducted on this project. The data set could be further reduced to be able to include more common variables that may improve prediction quality. Similarly, the release for the 2019-2020 cycle is soon to happen, which would give more observations to train models on. Furthermore, more work on models could be done. Only five models were considered for this project and there are many more that could be considered. Weighting could be improved in the models to account for the fact that the data is not a simple random sample. Better feature selection criteria could also be used to remove noisy variables in models that do not use their own feature selection. Even with all this future work, we believe that in the scope of this project that the random forest model fitted is a good predictor of diabetes status.

Appendix

Throughout this project, Brady Ryan worked on reducing the final table to include only data after 2007 and finding predictors to use in the model, as well as fitting the GBM, decision tree, and random forest models, comparing the 4 full models (without naive Bayes) and completing the visualizations for the 4 models. Lakshmi Ganesan worked on downloading the approximately 500 different data files and merging them together, as well as fitting the XGBoost model. Yingtong Liu worked on the visualization of the data sets, and fitting the naive Bayes classifier. All three of us helped make decisions at every point.