Diabetes Prediction model: NHANES data 2013-2014

Hannah Rosenblum, James Ng, Purnima Sharma

Contents

Introduction

This project aims to study any association between diabetes and several covariates in participants ages 1 and older, using NHANES data, and selecting an optimal prediction model among linear, non-linear, parametric and non-parametric models. The main objective is building a binary classification model with supervised learning. Certain factors of special interest were any association with participant's race, age, cholesterol and lifestyle factors, among others. Data was extracted for the year 2013 - 2014 from the cdc.gov website, https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013.

Data description

Specifically, association was assessed between diabetes and the following covariates:

- Gender: Participant's gender (male or female)
- Age: Age at screening, with possible values of 0 to 79, or 80+ (years)
- Race: 6 categories for race include Mexican American, other Hispanic, White, Black, Asian and other.
- bmi: body mass index (kg/mˆ2)
- hdl: High-density lipoprotein (mg/dL)
- Blood pressure (mm Hg): Both systolic and diastolic, first-round measurements
- waist: Waist circumference measurement (cm)
- Sedentary activity (lifestyle, minutes): time spent sitting in a given day, not including sleeping.
- Education level: highest degree of adults 20+ years of age, with 7 categories.
- Marital status: Categories include married, widowed, divorced, separated, never married, living with partner, refused, and don't know
- Depression: severity on a scale of 0 to 3 treated as a continuous variable, with 0 as not at all depressed
- Sleep: amount of sleep in hours on a given night on weekdays or workdays

The outcome of "diabetes" dependent-variable was based on classification of the participants into two groups of those with diabetes and those who did not have diabetes. Individuals answered the question "other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?", and were classified as having diabetes if they answered yes.

Motivation

Motivation was provided by the fact that diabetes is one of the major leading causes of death in the United States. As stated by the CDC site's National Diabetes Statistics Report of 2020, 34.2 million Americans are diabetic, while 7.3 million were undiagnosed. Furthermore, increase in type 2 diabetes among children is a growing concern according to the CDC. With prevalence of diabetes and prediabetes on the rise, it was of interest to find factors that might affect the diabetes status. Later years post-2013 were tried for the data, however were unavailable for the variables of interest possibly due to continuing updates.

Data cleaning

After extracting and merging the necessary files by participant's Id number, variables of interest were retained in a dataframe. Gender, race, education level, marital status and the response variable Diabetes were converted to factors from numeric data type. Missing entries for the response of diabetes status were removed. 185 "borderline " reported cases, 5 with "don't know" responses and 1 with "refused" response were also removed given the small scale of these categories, which accounted for less than 2% of the data, and in order to focus on the majority of binary responses of presence or absence of diabetes. The cleaned dataset contained 9,578 observations of 18 variables, including the binary outcome variable diabetes.

EDA

Exploratory data analysis was performed for all 18 initial variables, including the outcome of response, using density plots and bar graphs. Summary statistics were analyzed for all variables, to get an overview of the data and check for extent of missing values. Density plots were used to check for relationships between diabetes and other numeric variables. Categorical variables were visualized separately, using bar graphs instead.

summary statistics

Data Frame Summary

raw_data
Dimensions: 9578 x 18
Duplicates: 319

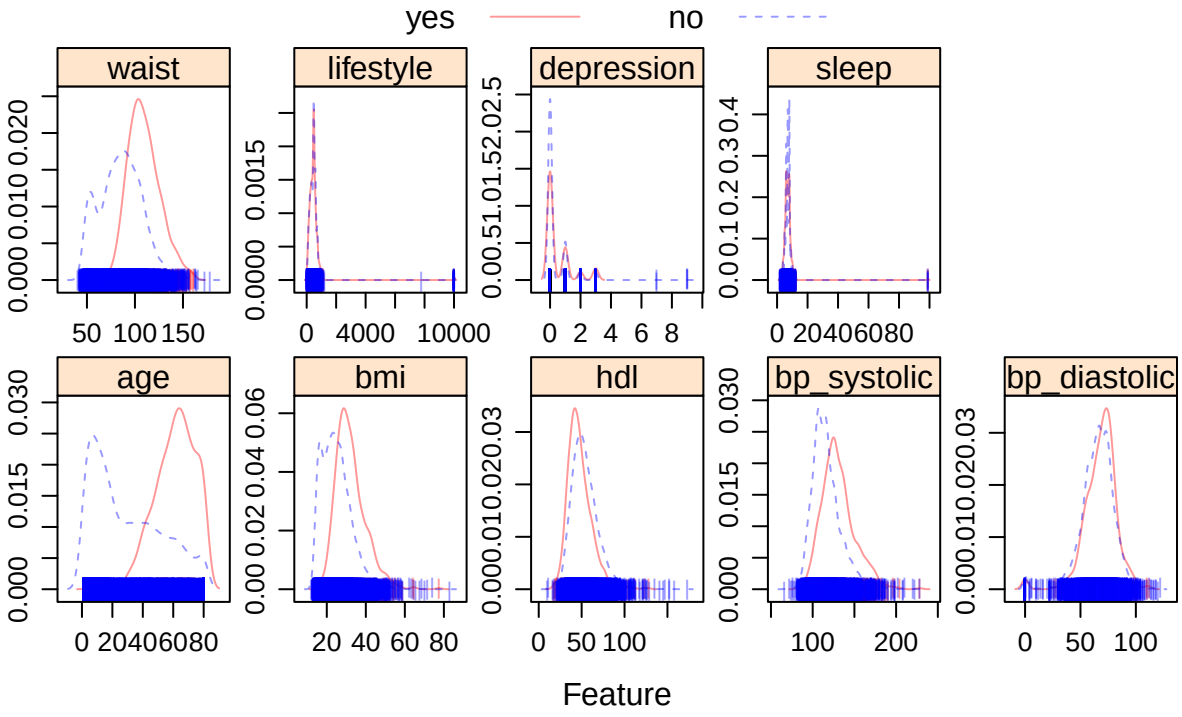
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	gender [factor]	1. 1 2. 2	4706 (49.1%) 4872 (50.9%)	IIIIIIIII IIIIIIIII	0 (0.0%)
2	age [numeric]	Mean (sd) : 32.4 (23.9) min < med < max: 1 < 28 < 80 IQR (CV) : 41 (0.7)	80 distinct values	: : : : : . : : : : : : : : . . : : : : : : : : : :	0 (0.0%)
3	race [factor]	1. 1 2. 2 3. 3 4. 4 5. 6 6. 7	1616 (16.9%) 893 (9.3%) 3449 (36.0%) 2148 (22.4%) 1033 (10.8%) 439 (4.6%)	III I IIIIIII IIII II	0 (0.0%)
4	bmi [numeric]	Mean (sd) : 25.6 (7.9) min < med < max: 12.1 < 24.6 < 82.9 IQR (CV) : 10.4 (0.3)	436 distinct values	: . : : : : : : : : . : : : : .	706 (7.4%)
5	hdl [numeric]	Mean (sd) : 53.2 (15.2) min < med < max: 10 < 51 < 173 IQR (CV) : 19 (0.3)	116 distinct values	: : . : . : : : : : : .	2128 (22.2%)
6	ldl [numeric]	Mean (sd) : 106 (34.9) min < med < max: 14 < 103 < 375 IQR (CV) : 46 (0.3)	194 distinct values	: . : : : . : : : . : : : .	6553 (68.4%)
7	triglyceride [numeric]	Mean (sd) : 111.7 (115.9) min < med < max: 13 < 88 < 4233 IQR (CV) : 73 (1)	344 distinct values	: : : : :	6515 (68.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
8	insulin [numeric]	Mean (sd) : 13.4 (18.7) min < med < max: 0.1 < 9.3 < 682.5 IQR (CV) : 9.1 (1.4)	1716 distinct values	:	6567 (68.6%)
9	glucose [numeric]	Mean (sd) : 114 (45.5) min < med < max: 40 < 104 < 604 IQR (CV) : 44 (0.4)	227 distinct values	:	7294 (76.2%)
10	bp_systolic [numeric]	Mean (sd) : 117.9 (18) min < med < max: 66 < 116 < 228 IQR (CV) : 20 (0.2)	71 distinct values	:	2571 (26.8%)
11	bp_diastolic [numeric]	Mean (sd) : 65.7 (15) min < med < max: 0 < 66 < 122 IQR (CV) : 16 (0.2)	59 distinct values	:	2571 (26.8%)
12	waist [numeric]	Mean (sd) : 86.9 (22.5) min < med < max: 40.2 < 87.4 < 177.9 IQR (CV) : 31.6 (0.3)	1030 distinct values	:	1091 (11.4%)
13	lifestyle [numeric]	Mean (sd) : 478.5 (642.1) min < med < max: 0 < 480 < 9999 IQR (CV) : 300 (1.3)	36 distinct values	:	2625 (27.4%)
14	education [factor]	1. 1 2. 2 3. 3 4. 4 5. 5 6. 7 7. 9	442 (7.9%) 761 (13.6%) 1261 (22.6%) 1715 (30.7%) 1406 (25.1%) 2 (0.0%) 5 (0.1%)	I II IIII IIIIII IIIII	3986 (41.6%)
15	married [factor]	1. 1 2. 2 3. 3 4. 4 5. 5 6. 6 7. 77 8. 99	2866 (51.3%) 419 (7.5%) 637 (11.4%) 170 (3.0%) 1096 (19.6%) 401 (7.2%) 2 (0.0%) 1 (0.0%)	IIIIIIIII I II III I	3986 (41.6%)
16	depression [numeric]	Mean (sd) : 0.4 (0.8) min < med < max: 0 < 0 < 9 IQR (CV) : 0 (2.1)	0 : 3955 (75.5%) 1 : 876 (16.7%) 2 : 205 (3.9%) 3 : 194 (3.7%) 7 : 2 (0.0%) 9 : 3 (0.1%)	IIIIIIIIIIIIII III	4343 (45.3%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
17	sleep [numeric]	Mean (sd) : 7 (3.2) min < med < max: 2 < 7 < 99 IQR (CV) : 2 (0.5)	12 distinct values	: : : : :	3300 (34.5%)
18	diabetes [factor]	1. yes 2. no	737 (7.7%) 8841 (92.3%)	I IIIIIIIIIIIIIIII	0 (0.0%)

As noted above in the summary table, several of the variables for laboratory data had high missing values. The variables ldl, triglyceride, insulin and 2hr glucose-test had close to 70% of the data missing. In an effort to retain a large enough sample size, the four variables were not retained for further analysis in this project.

Density plots



Density plots of several numeric covariates showed differences in distributions of the two classes. Plots of systolic blood pressure, waist circumference measurement, age, and body mass index seemed significantly different between those with diabetes and those without. Most significant difference seemed to be among different age groups, with a density curve of responses with no diabetes showing right-skewness, and those with diabetes skewed to the left along with a strong shift towards higher age.

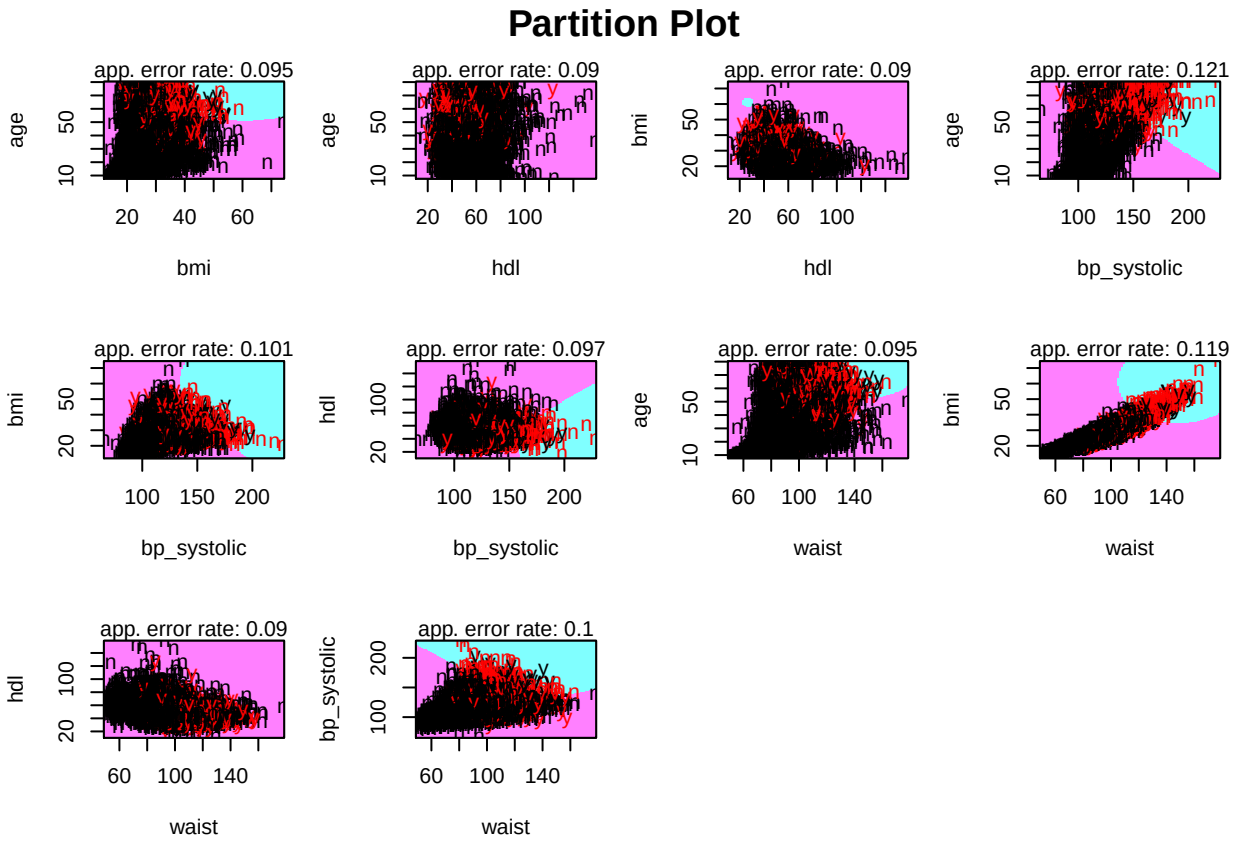
Bar plots

Bar plots, not shown, were analyzed for categorical variables gender, race ,education level, and marital status. Presence of diabetes did not seem to be gender-dependent, and with slight differences based on education level. Proportion of positive diabetes cases did seem to vary among different races, and based on marital

status. There seemed to be a significantly higher proportion of non-diabetics among individuals who were never-married or divorced.

Finally, paired partition plots were also examined for several variables using training data, to analyze misclassification rate. Methods using linear discriminant analysis, and non-linear boundaries such as quadratic discriminant analysis and Naive Bayes method were used, giving similar results in terms of error rates. Shown below is partition plots using Naive Bayes method for paired visuals on age, bmi, hdl, systolic blood pressure, and waist.

Partition plots



Missing data

The four Variables with high proportions of missing data, most of which were close to 70%, were removed. For the remaining variables with missing values, most were close to 25%, except marital status, education and depression level, which were approximately 40% missing. Assuming that the data was missing at random, and that single imputation might lead to bias and might not preserve relationships between variables; for those reasons imputation was not considered and the missing values were removed. The final sample consisted of 4,246 participants, still a fairly large dataset.

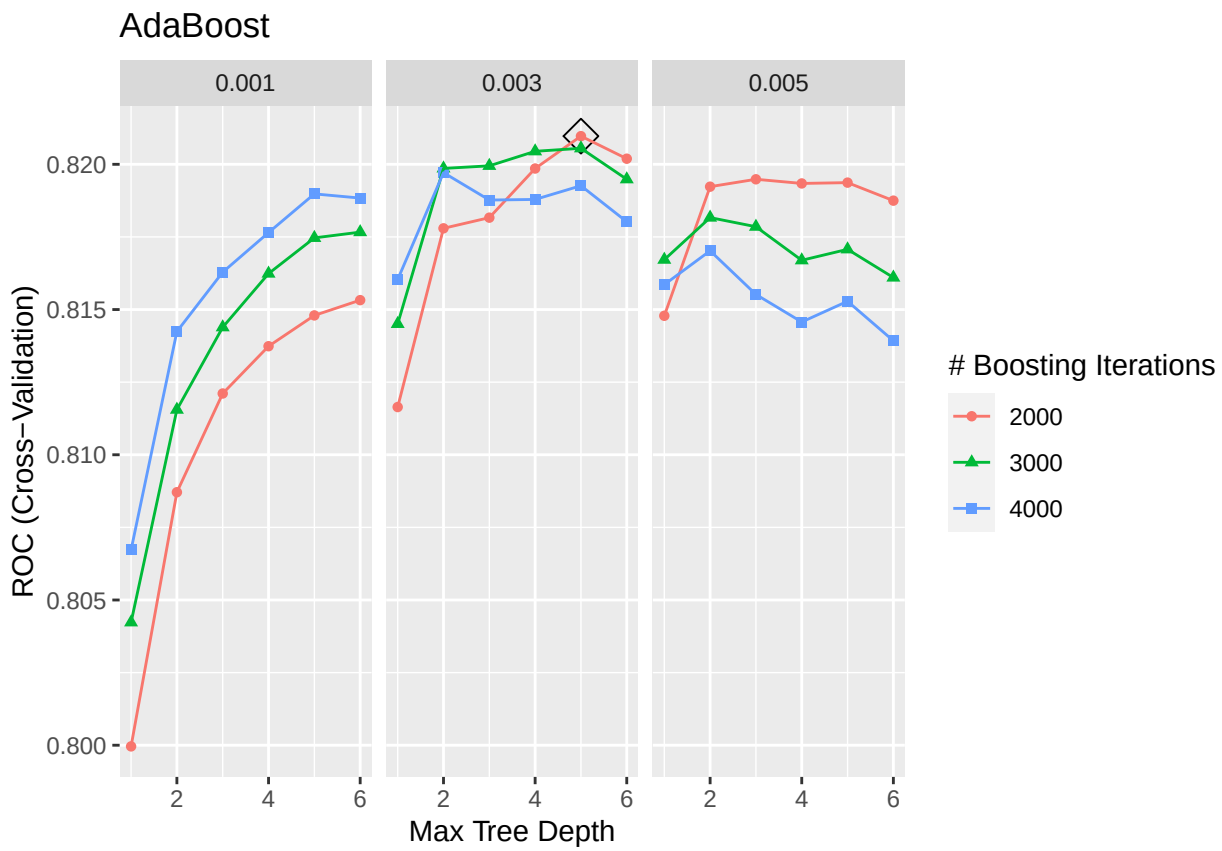
Models

Methods

Given a fairly large sample size the data was partitioned into train and test for prediction accuracy analysis, with 80% to 20% split. Several model-building techniques were used to predict the risk of diabetes using the 15 variables associated with it. Gender, age, race, bmi, hdl, systolic and diastolic blood pressure, waist circumference, lifestyle, education, marital status, depression and sleep were used as the predictor variables in the model building process. Given that the outcome was binary, generalized linear and non-linear methods such as logistic regression and discriminant analysis were tried, along with ensemble methods for trees, and the SVM (support vector machines) linear and non-linear boundary methods. For comparability of cross-validation performance, all models were fitted using the `caret` package. Built-in 10-fold cross validation method was used to get the area under the curves for comparisons, along with the summary statistics of performances on training data.

Due to ease of interpretability and the binary nature of the outcome, glm (logistic regression) and penalized logistic regression models were fitted to assess linear decision boundaries. Various combinations of the two tuning parameters were tried for the penalized logistic model to find the specific values that worked well for the given data. Generalized additive model (GAM) was considered but not used due to the length of time for its execution. Non-linear MARS model, with similar performance to GAM, was retained. Various combinations for tuning grid were tried to get its optimal performance.

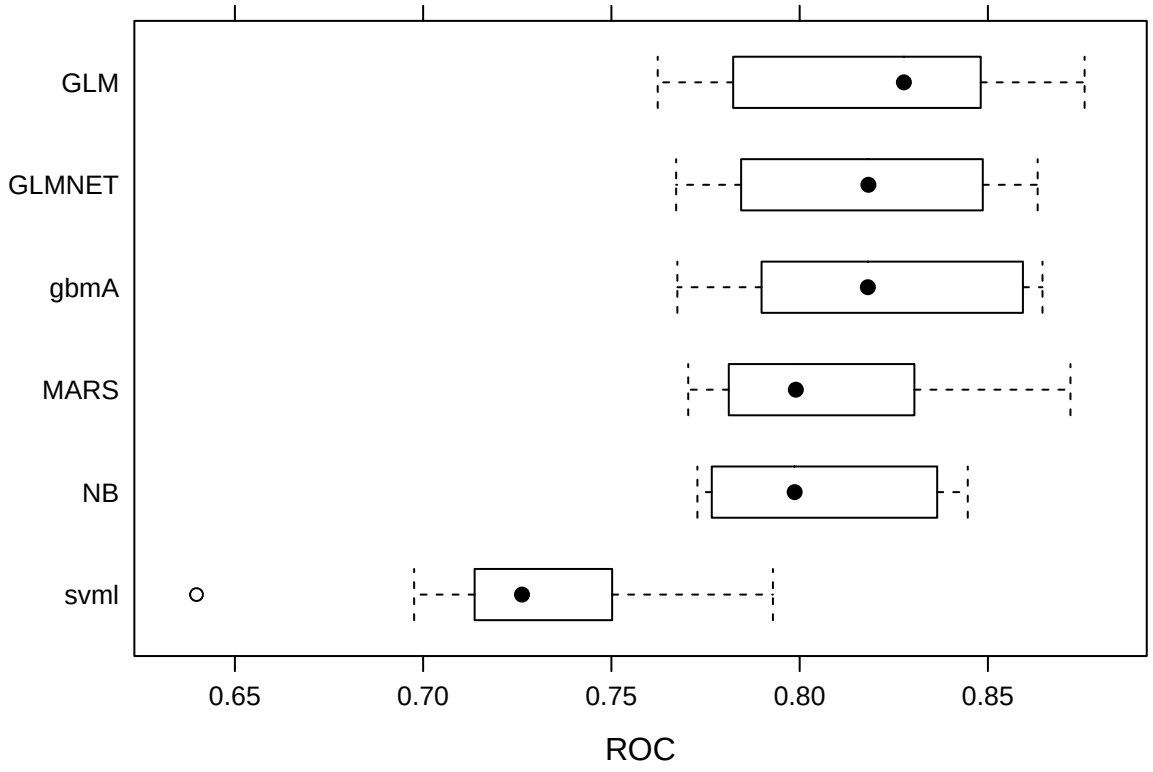
While no assumptions are needed for the predictors in linear logistic model, for discriminant analysis it is assumed that the predictors follow a normal distribution within each group of response, and that the variance-covariance matrix for response classes are the same for linear model (LDA), or could be different for non-linear QDA. Since four of the covariates in the dataset were categorical, these models were tried but excluded from analysis, also for the fact that these models work well for well-separated classes which was not apparent in the EDA. Under the assumption that features are independent within each class, Naives Bayes method (NB) was used due to its ability to handle mixed covariates, and due to its slightly better performance than LDA and QDA in terms of area under the ROC (receiver operating characteristic curve).



Models using non-parametric approach such as trees were also fitted using ensemble methods, which help improve prediction accuracy. Wisdom of crowds (bagging and random forest), and wisdom of weighted crowds of experts (Boosting) were the ensemble methods used. Single trees were not considered due to lack of predictive accuracy when compared to ensemble approach. For a single tree, a small change in data could cause a significant amount of change in the final estimated tree. For those reasons, it was left out. Random forest, bagging, boosting, and weighted boosting using `adaboost` were tried. Boosting model using `adaboost`, which minimizes the exponential loss function, was retained among the ensemble models due to its relatively better performance.

Lastly, support vector classifiers were also fitted, with linear and non-linear decision boundaries. Tuning grids were finalized after analyzing outputs of several combinations. The two models were compared using the ROC curve as a metric, with similar results. Nonlinear model was fitted using both the tuning parameters, giving a two-dimensional tuning grid. For that reason, a simpler linear model was selected for further comparison, considering the time of execution also as a deciding factor between the two models. Initially built using `kernlab` package, with predicted class probabilities requested the model iterations were showing as a part of the output. Using the `e1071` library instead to run the svm linear model corrected the problem. Neural networks were not considered due to their blackbox approach, which would result in non-transparency.

Model comparison



Comparing the models, their summary and boxplots showed that GLM model gave the highest median of 0.828 for ROC, a characteristic for evaluating performance based on sensitivity and specificity of a model at different thresholds for classification of response. Although the mean was marginally higher for `adaboost` model, simpler model was selected as the final model for ease of interpretability and as a better fit for a noisy data. Support vector machine linear model had the worst performance when compared to the other five final models. Based on these resampling comparisons and using ROC as our metric for model selection, GLM was chosen as the final model, and its prediction performance was then evaluated on test data.

Final model: GLM

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1772   0.1424   0.2918   0.5104   2.1264
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   7.926e+00  7.603e-01  10.424  < 2e-16 ***
## gender2      -3.396e-01  1.398e-01  -2.429 0.015130 *
## race2        -1.555e-01  2.396e-01  -0.649 0.516436
```

```
## race3         4.706e-01  1.927e-01   2.442 0.014606 *
## race4        -2.126e-01  2.071e-01  -1.026 0.304758
## race6        -4.240e-01  2.655e-01  -1.597 0.110279
## race7         1.025e-01  4.537e-01   0.226 0.821237
## education2    1.569e-01  2.263e-01   0.693 0.488051
## education3    5.478e-01  2.189e-01   2.502 0.012340 *
## education4    4.838e-01  2.148e-01   2.252 0.024335 *
## education5    8.664e-01  2.343e-01   3.697 0.000218 ***
## married2      5.346e-02  1.989e-01   0.269 0.788132
## married3      1.703e-01  1.800e-01   0.947 0.343893
## married4     -1.123e-01  3.191e-01  -0.352 0.724982
## married5      3.894e-01  2.096e-01   1.858 0.063144 .
## married6      8.015e-01  3.735e-01   2.146 0.031886 *
## age          -4.899e-02  4.998e-03  -9.803  < 2e-16 ***
## bmi           6.495e-02  2.242e-02   2.896 0.003774 **
## hdl           2.708e-02  4.772e-03   5.675 1.39e-08 ***
## bp_systolic  -6.210e-03  3.465e-03  -1.792 0.073077 .
## bp_diastolic  1.140e-02  4.399e-03   2.592 0.009551 **
## waist        -6.666e-02  9.927e-03  -6.714 1.89e-11 ***
## lifestyle     6.979e-05  8.360e-05   0.835 0.403802
## depression   -1.151e-01  6.910e-02  -1.666 0.095755 .
## sleep        -3.465e-02  1.668e-02  -2.078 0.037736 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.2  on 3395  degrees of freedom
## Residual deviance: 1993.5  on 3371  degrees of freedom
## AIC: 2043.5
##
## Number of Fisher Scoring iterations: 6
```

Summary statistics of the model shows age, hdl, waist and education level of college versus below 9th grade as highly significant predictors of diabetes. Gender, race being American whites versus Mexicans, bmi, diastolic blood-pressure, and education level in general are also shown as significant predictors at 5% significance level. For instance, having attended college versus only up to 9th grade education level, changes the log odds of not having diabetes by 0.87, or by 2.4 times.

Model:

log(odds of no diabetes)= = 7.926 - 0.34 female - 0.049 age + 0.47 white + 0.065 bmi + 0.027 hdl + 0.011 diastolic - 0.0667 waist + 0.548 High_School+ 0.484 Some_college + 0.866 College_graduate + 0.8 living_with_partner - 0.0347 sleep

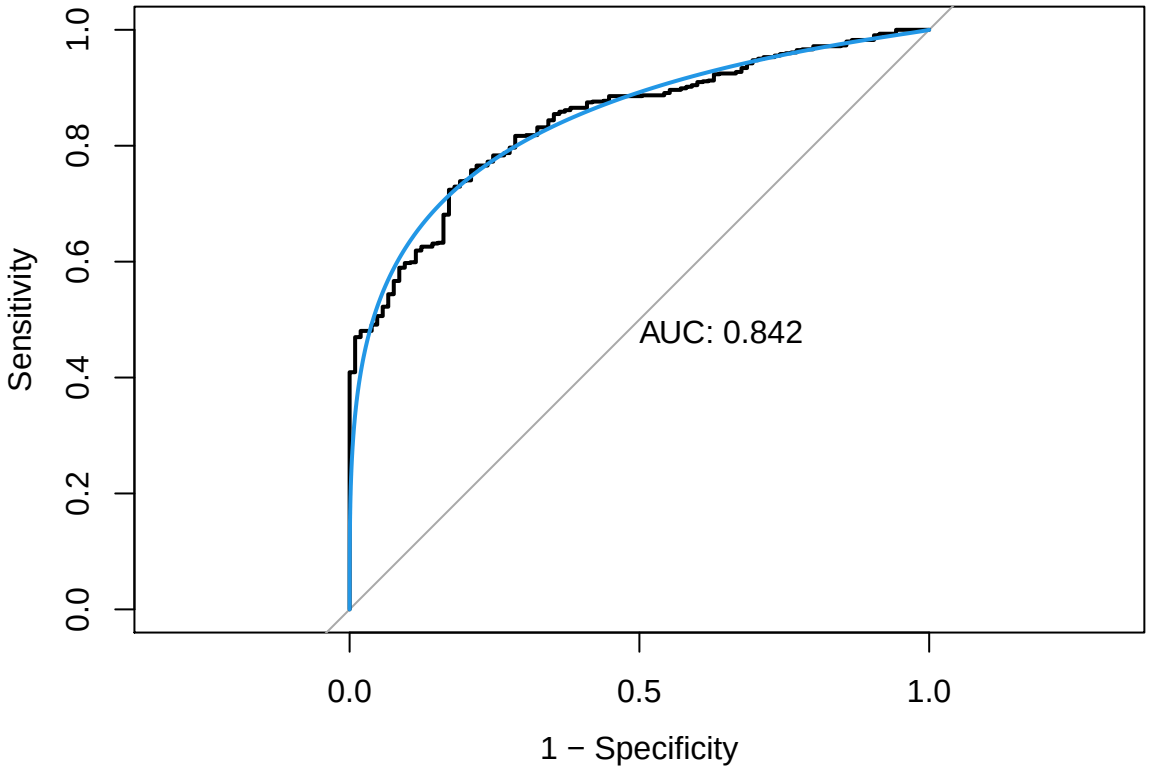
Model prediction performance

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction yes  no
##        yes  15  15
##        no   90 728
##
```

```
##                Accuracy : 0.8762
##                  95% CI : (0.8521, 0.8976)
##     No Information Rate : 0.8762
##     P-Value [Acc > NIR] : 0.526
##
##                   Kappa : 0.1769
##
##  Mcnemar's Test P-Value : 5.136e-13
##
##             Sensitivity : 0.9798
##             Specificity : 0.1429
##          Pos Pred Value : 0.8900
##          Neg Pred Value : 0.5000
##              Prevalence : 0.8762
##          Detection Rate : 0.8585
##    Detection Prevalence : 0.9646
##       Balanced Accuracy : 0.5613
##
##        'Positive' Class : no
##
```

Based on the confusion matrix at 0.5 threshold level, the accuracy of the model was 0.8762, which was same in value to "No Information Rate". This indicates that the classification was not that meaningful at the selected level of 0.5 threshold. `Kappa`, which accounts for agreement by chance between observed and predicted classification, was also low at 0.1769 indicating that the probability of agreement could be by chance and not necessarily due to correct predictions.

To further analyze prediction performance at other threshold cutoffs, ROC curve with range of threshold values between 0 and 1 was plotted.



The high area under the curve at 0.842 indicates that the model seemed to perform well on classifying the test data for the given dataset.

Limitations

The models were limited in their accuracy due to the imbalance of the dataset.A larger dataset would have been needed to correct the skewness of class distribution, by including a greater range of data time-periods. Additionally, including greater set of covariates of potential influence, such as work conditions, exposure to environmental pollutants, etc. would also have helped to formulate a more precise model. Another limitation of models were that only the complete cases were used in the building process, under the assumption that data was missing at random. That might not have been the case for all missing data, for example the body weight data for participants who had limb amputations were set to missing. This factor was not a part of this data. Finally, extremely small sample sizes in several subcategories, which had to be excluded, were not handled well by the models.

Conclusion

For the given dataset, the final model selected performed fairly well in predicting the test-data responses, when evaluated based on area under the ROC curve. The findings were in support of the expectation that simple linear models do well with noisy data, which could be the issue with epidemiological data. A lot more laboratory data was missing than expected and the surveys were based on recollection, which could have lead to recall bias and under-reporting. These factors could have been the cause of noisy data that the other models were unable to account for properly.

The data included both type 1 and type 2 diabetes records without segregating the two groups. It was unfortunate to realize that the age distinctions between the two forms of the disease are disappearing; what was known as adult-onset diabetes can begin during childhood.

Significant association between diabetes and age, waist circumference as an indicator of weight, diastolic blood pressure and high density lipids was as expected, which has been a well-documented fact at this point. Relevance of college education and marital status were unexpected.

There is high rate of missingness in epidemiological studies and surveys than would be expected, making any inferences and predictions to be drawn from them tricky at best. This dataset showed that sometimes more involved and complicated models do not always mean a better model, a critical element to remember when model building and analyzing data from a study.