Capstone 1 In Depth Analysis

Predicting Prediabetes and Diabetes from National Health and Nutrition Examination Survey (NHANES) Data 2007 – 2016

https://github.com/tcardwell/Capstone-1/blob/master/Capstone%201%20ML.ipynb

Introduction

There are two goals for this project. The first is developing a model to predict prediabetes from the health and nutrition data contained in the NHANES survey. The second is developing a model to predict diabetes from the same data.

The Centers for Disease Control and Prevention (2017) stated that 9.4% of the U.S. population has diabetes and nearly a quarter of those are undiagnosed. Just over a third of U.S. adults had prediabetes in 2015 and only 11.6% of them had been told by their doctor. New research is showing that damage to body systems begins with prediabetes, so there are huge benefits to the population and to the health care system to find ways to increase the number of people with prediabetes and diabetes who become aware of their condition.

Based on previous analysis, the models are expected to predict rates similar to the cited statistics.

Part I

As the goal for this project is classification, the algorithms chosen were Logistic Regression and tree-based algorithms, including Decision Tree, Random Forest, and various Boosting algorithms including Adaboost, Gradient Boosting, and XGBoost. Support Vector Machine and k-Nearest Neighbors were ruled out due to the very high dimensionality of the data. The chosen algorithms were first used with default parameters, then results were reviewed and selected models were tuned by optimizing hyperparameters with RandomizedSearchCV().

All features from the NHANES data that were tracked during the entire period from 2007-2016 were gathered. Those that had more than 40% missing values were eliminated, as were those that were strictly administrative in nature. Features used to create the target were eliminated as were other features highly correlated to them. Finally, several duplicate features were eliminated as well as some that were so closely correlated, they were nearly duplicates.

Several of the NHANES continuous variables are self-reported data and have code values well above the data range signaling the participant did not know the answer or refused to answer the question. These indicator codes were stripped from continuous variables so they did not skew the analysis.

The target for the first goal was derived from laboratory results using diagnostic criteria from the American Diabetes Association (n.d.). Three classes were defined: no diabetes, prediabetes, and diabetes. The initial goal was to predict the prediabetes category. As the data included sufficient targets for all three categories, the models were built as multiclass classifiers.

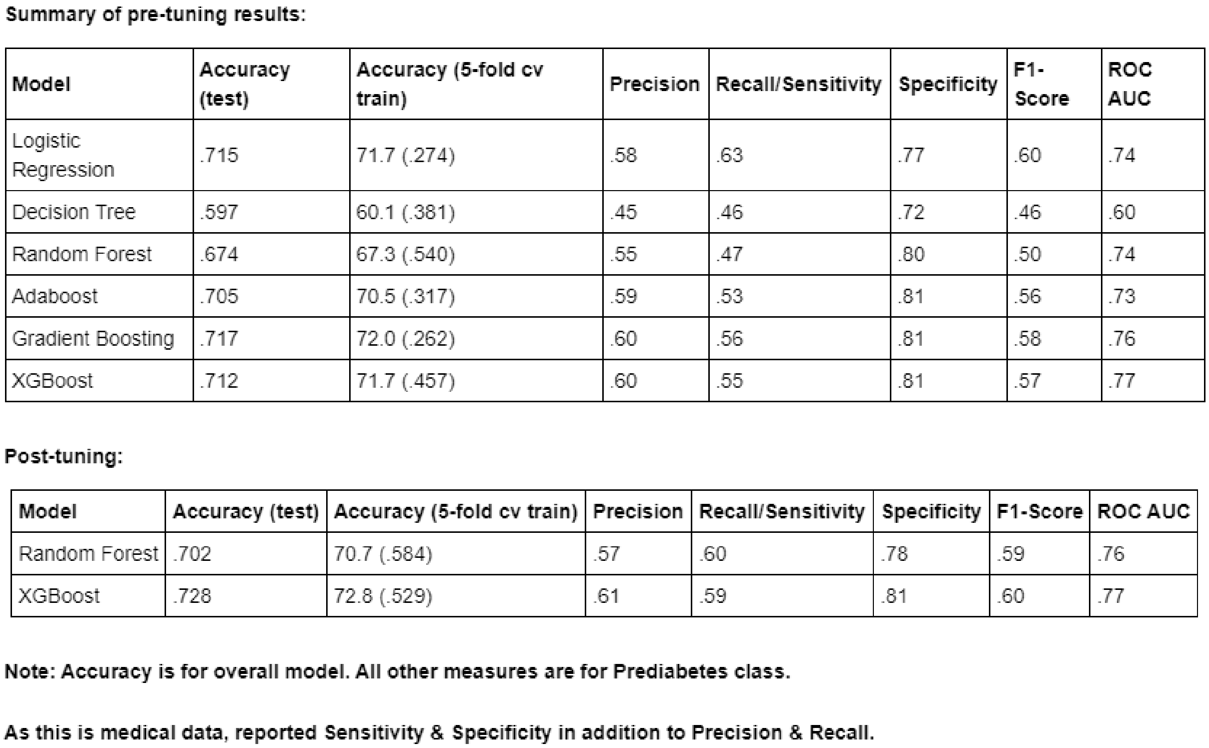
The data was split into training and test sets with 30% reserved for testing.

Missing values in the remaining features were imputed using IterativeImputer() with the default BayesianRidge() estimator using 5 other features. IterativeImputer() uses correlation coefficients to choose which features to use. The imputation algorithm was fit on the training data, then used to transform both the training and test datasets.

For Logistic Regression, the continuous features were normalized with StandardScaler() so large valued features would not dominate the regression. All categorical nominal features were encoded to dummy variables. Once again, the transforms were fit on the training data, then used to transform the training and test datasets. Categories in the test dataset that did not appear in the training data were ignored.

Evaluation criteria had to span both regression and tree-based models. Initial criteria chosen were ROC AUC, Precision and Recall. Accuracy, Specificity and F1-score were also reported. As the models were multiclass, a one vs all approach was taken to plotting ROC curves and calculating AUC values.

Part I Results



Summary of pre-tuning results:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Logistic Regression	.715	71.7 (.274)	.58	.63	.77	.60	.74
Decision Tree	.597	60.1 (.381)	.45	.46	.72	.46	.60
Random Forest	.674	67.3 (.540)	.55	.47	.80	.50	.74
Adaboost	.705	70.5 (.317)	.59	.53	.81	.56	.73
Gradient Boosting	.717	72.0 (.262)	.60	.56	.81	.58	.76
XGBoost	.712	71.7 (.457)	.60	.55	.81	.57	.77

Post-tuning:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Random Forest	.702	70.7 (.584)	.57	.60	.78	.59	.76
XGBoost	.728	72.8 (.529)	.61	.59	.81	.60	.77

Note: Accuracy is for overall model. All other measures are for Prediabetes class.

As this is medical data, reported Sensitivity & Specificity in addition to Precision & Recall.

Based on the evaluation criteria, the tuned XGBoost model is best, although none of these models is a strong predictor of prediabetes.

The tuned XGBoost model’s predictions on the test data are:

- No Diabetes = 58.8%
- Prediabetes = 31.1%
- Diabetes = 10.2%.

These results are very similar to the statistics reported by the Centers for Disease Control and Prevention (CDC).

Part II

This portion of the analysis was very similar to Part I with a slightly different target. The NHANES data contains self-reported prescription medication data that includes ICD-10 medical codes indicating the reason for the medication. This ICD-10 code data was used as the target for the second half of the analysis. These codes were tracked starting in 2013 so there are only four years data out of the total ten years gathered.

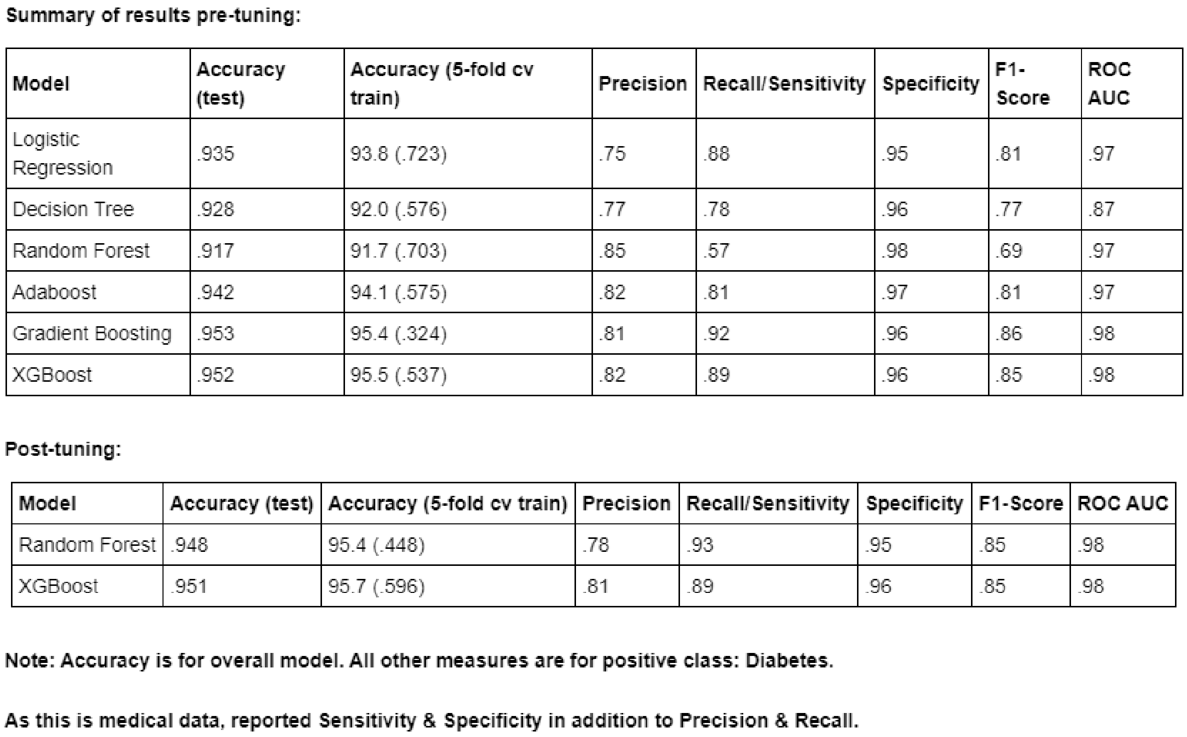
Data preparation steps were identical to those in Part I aside from target definition. As the 40% threshold for null values was determined based on ICD-10 code data, some different feature columns were eliminated due to excessive nulls. In addition, the features used to determine the target for Part I were included in Part II as they were not used to derive the new target.

The ICD-10 code prefix E11 was used to identify the Diabetes category per an ICD-10 code lookup tool (Centers for Disease Control and Prevention, 2019). The code prefix R73 in the absence of an E11 code was used to identify the prediabetes category. Some survey participants had both E11 and R73 codes. These were all assigned the diabetes category.

There were less than 2% prediabetes observations, most likely as prediabetes is usually controlled with diet and lifestyle changes rather than medication. As these 2% were a very small number of observations even before dividing into training and test sets, this portion of the analysis was made a binary classification problem with just the diabetes and no diabetes categories. Those in the prediabetes category were reassigned to the no diabetes category.

The same machine learning algorithms and evaluation criteria were used in Part II as in Part I.

Part II Results



Summary of results pre-tuning:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Logistic Regression	.935	93.8 (.723)	.75	.88	.95	.81	.97
Decision Tree	.928	92.0 (.576)	.77	.78	.96	.77	.87
Random Forest	.917	91.7 (.703)	.85	.57	.98	.69	.97
Adaboost	.942	94.1 (.575)	.82	.81	.97	.81	.97
Gradient Boosting	.953	95.4 (.324)	.81	.92	.96	.86	.98
XGBoost	.952	95.5 (.537)	.82	.89	.96	.85	.98

Post-tuning:

Model	Accuracy (test)	Accuracy (5-fold cv train)	Precision	Recall/Sensitivity	Specificity	F1-Score	ROC AUC
Random Forest	.948	95.4 (.448)	.78	.93	.95	.85	.98
XGBoost	.951	95.7 (.596)	.81	.89	.96	.85	.98

Note: Accuracy is for overall model. All other measures are for positive class: Diabetes.

As this is medical data, reported Sensitivity & Specificity in addition to Precision & Recall.

Gradient Boosting and XGBoost models performed the best. Tuning did not improve the XGBoost model in this case.

The tuned XGBoost model predicted 17.3% diabetes. This is above the statistic reported by the CDC (2017) but is close to the 15.8% of the project data participants taking medication for diabetes. The

NHANES data uses a complex survey design and oversamples selected communities. This could explain part of the difference between the 9.4% rate of diabetes in the U.S. and the 15.8% rate in this data.

Conclusion

The XGBoost model performed best in both parts of this analysis. The best prediabetes model has about a 60% rate of correctly predicting prediabetes. This leaves room for improvement but is still much better than random chance as the prediabetes class comprises only 33.7% of the data. The best diabetes model has about an 80% rate of correctly predicting diabetes. This class makes up only 15.8% of the original data. These precision rates may be partly due to the imbalance between the classes as there is a 54%/34%/13% ratio of no diabetes/prediabetes/diabetes observations in the dataset. Perhaps adjusting sampling to balance the classes would improve the models.

Predicting prediabetes proved more difficult than predicting diabetes. This is most likely because there is more overlap in features between the no diabetes and prediabetes classes than between no diabetes and diabetes. In addition, some in the prediabetes category may have only recently moved into that category and may not yet have developed elevated laboratory results that would assist the models in classification.

Finally, the NHANES dataset contains a large number of variables, many of which contributed noise to the models. Some of this noise was removed but certainly some remains. Perhaps the models would perform better with more of the noise removed through additional preprocessing of the data.

References:

Centers for Disease Control and Prevention. (2017) *National Diabetes Statistics Report.* American Diabetes Association. https://dev.diabetes.org/sites/default/files/2019-06/cdc-statistics-report-2017.pdf

American Diabetes Association. (n.d.) *Diagnosis.* https://www.diabetes.org/a1c/diagnosis

Centers for Disease Control and Prevention. (2019) *ICD-10-CM*. https://icd10cmtool.cdc.gov/?fy=FY2019