Assessment of Key Risk Factors Driving Diabetes on NHANES Dataset

Zhiyi Sun

Introduction

Diabetes is a chronic medical condition in which the body's ability to produce or use insulin is impaired. The imbalance in insulin production or use can lead to high blood sugar levels, resulting in serious health consequences, including heart disease, stroke, kidney disease, blindness, and nerve damage. About 34.2 million people of all ages (about 1 in 10) have diabetes in the U.S., and the number of people who are diagnosed with diabetes increases with age[1].

Factors that increase risks of diabetes differ depending on the type of diabetes people ultimately develop[1]. Additionally, the development of diabetes is influenced by a combination of genetic, environmental, and lifestyle factors. The long latent period of diabetes makes it difficult to identify the risk factors that may have contributed to the development of the disease. There may be unidentified risk factors that are not yet known or fully understood, which requires extensive research in the future to guide the appropriate management of people with diabetes.

Questions of interest

Given those difficulties in dealing with diabetes, we planned to identify risk factors that may lead to diabetes by using feature selection methods and tried to make predictions of the risk of diabetes. To better deal with the possible computational challenge, we made a comparison between logistic regression and other machine learning algorithms to improve both efficiency and accuracy. Through our project, we can find models that would be used as references for future diabetes diseases prediction, and identify significant features of risks for clinicians to better assist biomedical researches in healthcare industry.

Data and Methods

Dataset

The **N**ational **H**ealth and **N**utrition **E**xamination **S**urvey (**NHANES**) dataset (2005-2014) from the National Center for Health Statistics (NCHS) is a collection of data ranging from *demographics, medical history, physical examinations, biochemistry* to *dietary* and *lifestyle questionnaires*, integrating to 5 csv files (Demographic, Diet, Examination, Questionnaire, and Lab Results).[2][3] As a comprehensive and reliable resource, this dataset has been widely used in researches on healthcare issues. It includes known features contributing to diabetes, such as blood pressure, serum cholesterol level, alcohol consumption, weight, etc., that are relevant to our research question. The 10-year data contribute to the long-term management of diabetes. The **sample size** of our original dataset is 50000+, with more than 200 variables. Our response variable **DIQ010** is a question asking "Has a doctor or other health professional ever told you that you had diabetes?", which is used to represent diabetes in this study.

Data preparation

Before feature selection, we **1)** combined dataset using inner_join; **2)** excluded null and missing values in response variable (DIQ010); **3)** excluded non-numeric values; **4)** removed columns that have over 50% missing values; **5)** replaced missing values with most frequent values. After that, we **6)** splitted the data into training set (80%), and test set (20%), both for response and predictors.

Statistical Analysis

Feature selection

XGBClassifier was used for feature selection. XGBClassifier is a popular machine learning algorithm used for solving classification problems in various industries, such as credit scoring, fraud detection, and customer churn prediction. As an implementation of the gradient boosting decision tree algorithm, XGBClassifier combines many weak models (i.e. shallow decision trees) into a single strong model by using boosting (a type of ensemble learning), where multiple models are trained and their predictions are combined to create a more accurate and robust model. The features selected were treated as inputs of the subsequent model.

The imbalanced dataset, in which the classes have highly uneven sample sizes (which is often seen in healthcare dataset), can be a problem when building machine learning models as they can be biased towards the majority class. Therefore, we introduce the method of **SMOTE** (Synthetic Minority Oversampling Technique). The basic idea is to take samples of the feature space for each target class and its nearest neighbors, then generates new examples that combine features of the target case with features of its neighbors to increase the number of cases in the minority class [4]. In this study, both before and after using SMOTE, the XGBClassifier method was used. We observed that the result in the confusion matrix is more reasonable after using SMOTE.

Model building

We used the following methods: **1) Logistic Regression**: a type of supervised learning algorithm that is used to predict a binary outcome (i.e. a outcome that has two possible values, such as “yes” or “no”) based on a given set of the independent variables, by using a logit function. **2) Support Vector Machine (SVM)**: SVM works by finding the best hyperplane (i.e. a decision boundary) that can linearly separate the data points into different classes,and is able to handle high-dimensional data efficiently and is robust to overfitting. **3) Random Forest**: is an ensemble learning method, in which each decision tree is trained on a random subset of the data, and then work together to make predictions. The final prediction is made by averaging the predictions of all the individual trees. Multiple decision trees in a random forest helps to improve the generalizability of the model and reduce overfitting. **4) Gradient Boosting Decision Tree (GBDT)**: is an extension of the gradient boosting algorithm that uses decision trees as the base learners (i.e. the individual models that are combined to make the final prediction).

Bagging works especially well for algorithms that are considered weaker (ie. unstable) or more susceptible to variance (i.e. high variance), such as decision trees or KNN: **1) Bagging Decision Tree (Bagging on decision trees)**: implemented by creating bootstrap samples from the training dataset, and then building trees on the bootstrap samples to aggregate the outputs of all the trees and predicting the final output. **2) Bagging k-Nearest Neighbors**: despite bagging stable classifiers makes little difference, it worth a try when k is sufficiently small.

Deep learning methods are wildly used in public health today, such as predicting health conditions from electronic health records. We implemented **MLP**, which is a type of feedforward artificial neural network (ANN), where both classification and regression problems can be applied with. It can be thought of as generalized linear models that go through numerous phases of processing, non-linear mapping between input vectors and output vectors, before making a decision.

Model evaluation

Confusion matrix was applied to evaluate the performance of the classification algorithm. The rows of the matrix represent the actual class labels and the columns represent the predicted class labels. The elements in the matrix are TN (True Negative), FP (False Positive), FN (False Negative), TP (True Positive), respectively.

AUC-ROC curve is a performance measurement for the classification problems at various threshold settings. ROC (Receiver Operating Characteristics) is a probability curve and AUC (Area Under The Curve) represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes[5].

Computational techniques

Regarding computational challenges, when it comes to coding efficiency, we avoid using unnecessary loop for iteration, except for built-in functions. Also, local variables are used instead of global variables. The most up-to-date python version was adopted as it runs faster than previous versions. We also tried Biostat cluster to increase efficiency.

Results and Discussion

Our final dataframe has 48040 rows and 252 columns left after data preprocessing. The distribution of the response variable is shown in **Figure 1**. The dataset seems imbalanced as the negative responses are way more than positive ones. After balancing our dataset using SMOTE, we proceeded to feature selection using XGBClassifier in order to select the key features and fit them into our models. Top 10 features among 24 selected sorted by importance are listed (**Figure 2**).

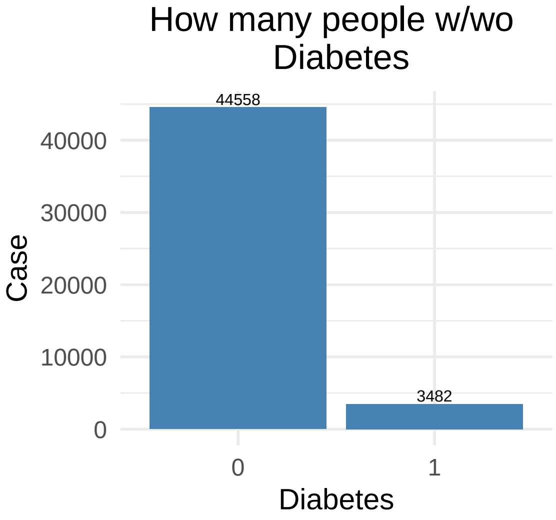

Figure 1

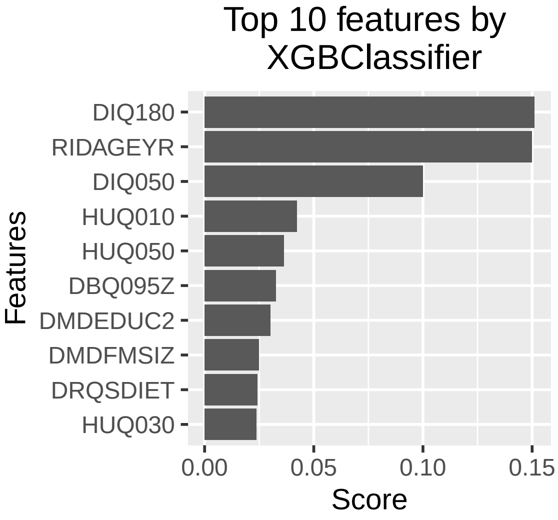

Figure 2

Logistic regression is the baseline model for prediction, which is implemented to compare with other machine learning models. We noticed that the accuracy score is about 0.845. In reality, this means we can predict whether a patient has diabetes based on the 24 selected features. We then used other machine learning models such as random forest, gradient boosting, etc, and made comparison between these models using their accuracy scores and operation times (**Figure 3**). The result indicated bagging decision tree has the highest accuracy score followed by random forest and bagging KNN. Besides, except for SVM has the lowest running efficiency, the running time of Gradient Boosting Decision Tree is faster than expected. Due to the sequential connection between individual trees, boosting algorithms are highly accurate. In gradient boosting decision trees we need to be careful about the number of trees we select, because having too many weak learners in the model may lead to the overfitting of data[6]. Therefore, tuning of hyperparameters in gradient boosting decision trees should be considered. Furthermore, Bagging on Decision Tree significantly improves the model performance, due to the weak classifier. It is worth noting that Bagging on KNN also shows a significant improvement in our study, which may due to the relatively low K[7][8]. However, the MLP model performance in our study is not as expected, which may be due to the setting of the hyperparameter tuning. The top five features selected by the random forest importance tree (**Figure 4**), which are age, general health condition, taking insulin and blood tested situation were consistent with the previous top 5 features selected by XGBClassifier.

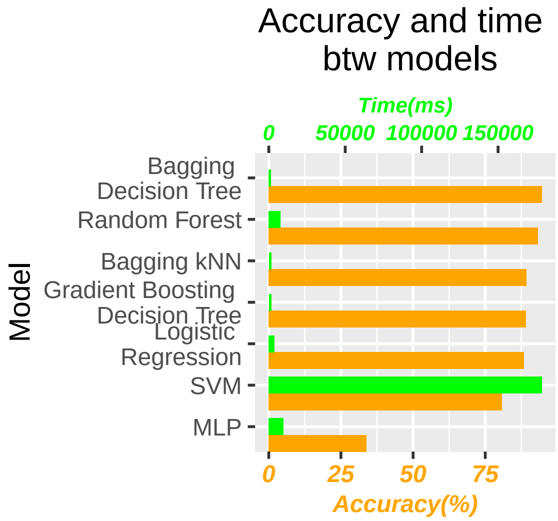

Figure 3

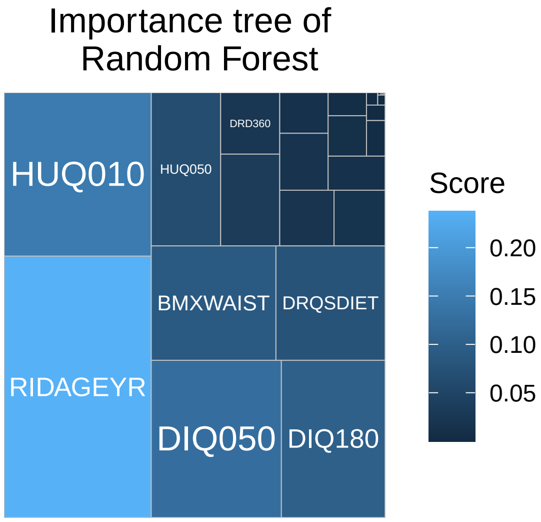

Figure 4

Lastly, five models which have accuracy scores higher than logistic regression model were selected, and their corresponding AUC-ROC curves were depicted as **Figure 5**. We found that random forest has the highest AUC value 0.97, which demonstrates its advantage at distinguishing whether the patient had diabetes.

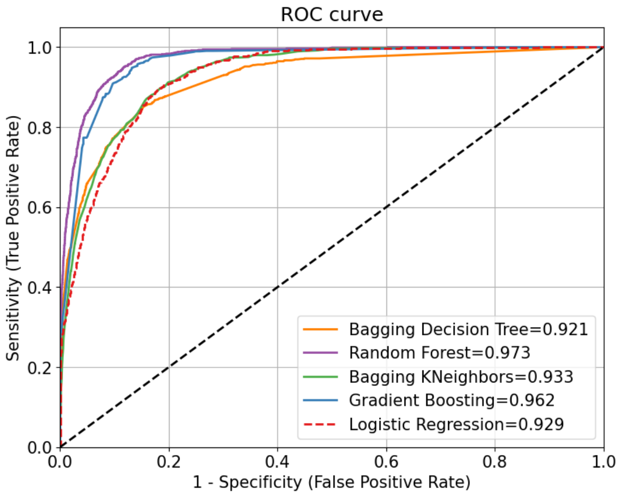

Figure 5: ROC-AUC curve

Conclusion

In conclusion, among all of the seven models, Bagging Decision Tree gives the highest accuracy (based on confusion matrix) on testing data set, while random forest would be preferred basing on AUC-ROC which gives the highest score 0.969. Age, general health condition, taking insulin and blood tested situation tend to be the commonly important features selected by XGBClassifier and random forest prediction model, which may indicate a relationship with diabetes based on NHANES (2005-2014) data. Besides, except for SVM has the lowest running efficiency, the running time of Gradient Boosting Decision Tree is faster than expected. Furthermore, Bagging Decision Tree illustrates a significant improvement of weaker classifiers implemented on our data (i.e. Bagging Decision Tree), and it worth mention that Bagging KNN also shows a surprisingly improvement of it in our study.

It is important to note that there are limitations to conclusions we can draw based on the data and the underlying model assumptions. The removal of missing values may lead to a biased representation of the data. Future studies should explore the use of interpolation for the missing values, such as replacing by the average or median, or predicting by classification or regression models[9]. Besides, some models (ie. gradient boosting and MLP) are implemented without appropriate tuning parameters which may affect variance-base trade-off or cause overfitting[10]. A more in-depth review of parameters tuning of certain cases is therefore necessary. Furthermore, the selected important features are based on model predictions, which may lose practical significance. In future studies, more literature reviews are needed to make inferences, especially basing on most recent NHANES data. Furthermore, future study considers C++ cross-compilation, multiprocessing, Polars (a lightning-fast dataFrame library) for efficiency improvement[11][12][13].

Broadly, this analysis was consistent with the trends in existing literature and maybe beneficial to the general public and future design of questionnaires. The study further underscores the need for future analyses to further examine the relationship between diabetes and widely influential features, as well as prediction.

Github link: https://github.com/scottsun417/NHANES.

References

[1] Virani SS, Alonso A, Benjamin EJ, et al; on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics- 2020 update: a report from the American Heart Association. Circulation. 2020;141:e1-e458. doi: 10.1161/CIR.0000000000000757.

[2] Nhanes - about the National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. Available at: https://wwwn.cdc.gov/Nchs/Nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2015.

[3] U.S. Healthcare Data. Kaggle. Available at: https://www.kaggle.com/datasets/maheshdadhich/us-healthcare-data?select=Nhanes_2013_2014.csv.

[4] Likebupt. Smote - Azure Machine Learning, Azure Machine Learning | Microsoft Learn. Available at:https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/smote.

[5] Narkhede, S. (2021) Understanding AUC - roc curve, Medium. Towards Data Science. Available at: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#.

[6] Lombardo, L., Cama, M., Conoscenti, C., Märker, M., & Rotigliano, E. J. N. H. (2015). Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). Natural Hazards, 79(3), 1621-1648.

[7] Zhou, Z. H., & Yu, Y. (2005). Adapt bagging to nearest neighbor classifiers. Journal of Computer Science and Technology, 20(1), 48-54.

[8] Tu, M. C., Shin, D., & Shin, D. (2009, December). A comparative study of medical data classification methods based on decision tree and bagging algorithms. In 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing (pp. 183-187). IEEE.

[9] Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy.

In Classification, clustering, and data mining applications (pp. 639-647). Springer, Berlin, Heidelberg.
[10] Windeatt, T. (2006). Accuracy/diversity and ensemble MLP classifier design. IEEE Transactions on Neural Networks, 17(5), 1194-1211.
[11] Zaytsev, Y. V., & Morrison, A. (2014). CyNEST: a maintainable Cython-based interface for the NEST simulator. Frontiers in neuroinformatics, 8, 23.
[12] Palach, J. (2014). Parallel programming with Python. Packt Publishing Ltd.
[13] Pola-Rs. Pola-Rs/Polars: Fast multi-threaded, hybrid-streaming DataFrame library in Rust: Python: Node.js, GitHub. Available at: https://github.com/pola-rs/polars