# Multi-year Longitudinal Diabetes Analysis

Students: Yokey Li, Kayoko Watson, Kenny Zhang
Advisors: Prof. Ian Duncan, Prof. Xiyue Liao

## Abstract

Diabetes is the 7th leading cause of death in the United States. Prevalence is approximately 9.4% and diabetes accounts for 9.3% of national health expenditure. Although diabetes is a chronic disease, pre-diabetes is reversible. Our goal is to find important predictors that affect transition from pre-diabetes to healthy status. We use machine learning techniques to develop models and predictions based on a dataset provided by the Vitality Group. Random forest, Generalized Additive model, logistic regression and penalized logistic regression are the main methods we discuss here. We provide AUC, sensitivity, accuracy and other measures for different models.

## Introduction

We received our dataset from Vitality Group, a company that provides health and wellness solutions for employer groups, insured groups, and individuals all around the world. The dataset contains 347,928 observations, and 141,235 people with 2, 3 or 4 years of records. Each observation contains 38 explanatory variables. All the data are collected from 2012 to 2015. Our goal is to find the most relevant factors in transitioning from pre-diabetes to healthy status. We use Fasting Plasma Glucose (FPG) as a criteria for judging pre-diabetes. We remove all the people who once have diabetes or who always maintain a healthy status. We code our response variable into three categories, namely success transition from pre-diabetes to healthy, failure transition from pre-diabetes to pre-diabetes and worst transition from healthy to pre-diabetes (see Figure 2). Most of our research focuses on the first two scenarios.

## Methods and Analysis

We code our response variable by looking at the trend of pre-diabetes status (see Figure 2). Then we implement the following methods

- Balance the dataset using undersampling.
- Fit logistic regression and regularization methods including LASSO, ridge and elastic net using nested cross validation.
- Use LASSO to get a subset of important predictors.
- Use important predictors to fit Generalized Additive models.
- Use important predictors to fit Bagging, Boosting and Random Forest models.
- Use Gower distance and perform k-medoids clustering.

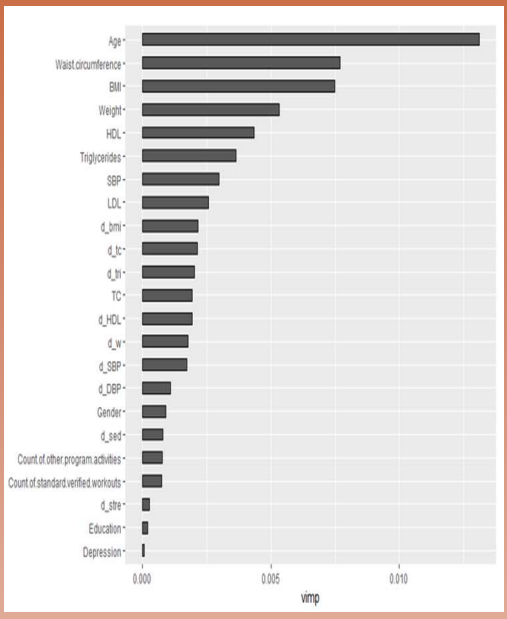


**Figure 1. Random forest.**

## Results

Table 1 gives us general results on how different models perform on the test set using different metrics. Random forest performs relatively well. Note that the generalized additive model is second best under most metrics and has a lower FPR (False Positive Rate) than Random Forest. We want a lower FPR in order to avoid a failure transition to be predicted as success. The Generalized Additive Model also gives us a useful trend estimation on each variable (Figure 3). Thus, we recommend the Generalized Additive Model in this sort of research in the future. Bagging, Boosting and Random Forest also produce a variable of importance plot (Figure 1). We can see that variables like age, BMI, HDL, and change in BMI play important roles in transitioning from pre-diabetes to healthy. The clustering visualization in Figure 4 shows that four clusters is optimal in our dataset. After clustering, we can see a pattern that male and female are very distinct and that within each gender, there are two clusters that agree on success transition or failure transition respectively. This suggests that a gender-specific approach might be helpful for further research.

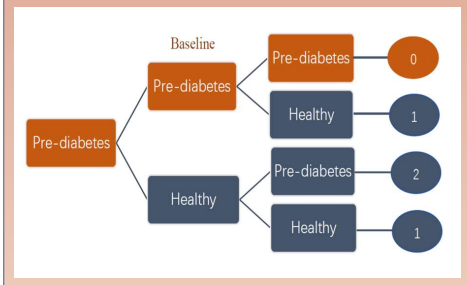


**Figure 2. Coding Response Variable.**

| | Precision | Recall | Accuracy | FPR | AUC | F-Measure |
|---|---|---|---|---|---|---|
| **GLM** | 0.768 | 0.599 | 0.611 | 0.364 | 0.655 | 0.673 |
| **Lasso** | 0.770 | 0.597 | 0.611 | 0.359 | 0.655 | 0.672 |
| **Ridge** | 0.770 | 0.591 | 0.608 | **0.356** | 0.655 | 0.668 |
| **Elastic Net** | **0.771** | 0.597 | 0.612 | 0.358 | 0.655 | 0.673 |
| **Bagging** | 0.760 | 0.606 | 0.609 | 0.385 | 0.656 | 0.674 |
| **Boosting** | 0.759 | 0.600 | 0.606 | 0.383 | 0.651 | 0.670 |
| **Random Forest** | 0.765 | **0.616** | **0.617** | 0.380 | **0.673** | **0.683** |
| **GAM** | 0.770 | 0.609 | 0.617 | 0.366 | 0.665 | 0.680 |

**Table 1. Test results from different models.**

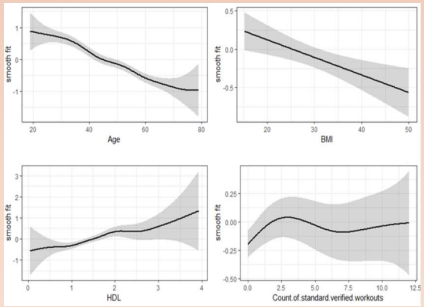


**Figure 3. GAM smooth plot.**

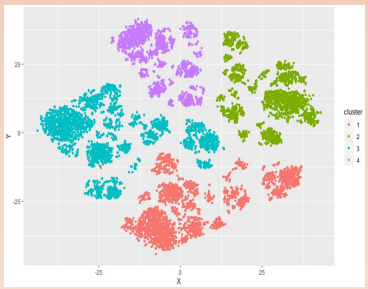


**Figure 4. Clustering Visualization.**

| Factors | p-value |
|---|---|
| **BMI** | 0.0016 |
| **Waist Circum.** | 0.0065 |
| **Change in BMI** | 3.64e-08 |
| **HDL** | 0.0004 |
| **Count of Other Program Activities** | 0.0003 |
| **Change in HDL** | 0.0017 |
| **Change in SBP** | 0.0022 |
| **Change in Sed** | 0.0099 |

**Table 2. Important Modifiable Factors**

## Conclusions

In almost all our models, age is an important predictor for transitioning from pre-diabetes to healthy. However, we cannot do much about aging. Table 2 provides a list of modifiable factors that are important for transitioning from pre-diabetes to healthy according to our research. We found important modifiable factors such as BMI, Waist Circumferences, HDL, count of activities, etc. Some have increasing trends, some have decreasing trends. In conclusion, a healthy lifestyle with proper exercise and nutrition intake is essential for getting away from pre-diabetes or in general, diabetes.

## Contact

Department of Statistics and Applied Probability
- University of California, Santa Barbara
Email: pstat296_2018-19@pstat.ucsb.edu



## References

1. Max, Kuhn. (2018, November 20). Classification and Regression Training.
2. Sebastian, Raschka. (2015, June 25). Machine Learning FAQ.
3. Centers for Disease Control and Prevention. (2017) National Diabetes Statistics Report, 2017. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services.
4. Alboukadel, Kassambara. (2018, November 3). Penalized Regression Essentials: Ridge, Lasso & Elastic Net.
5. National Center for Chronic Disease Prevention and Health Promotion. (2017). CDC National Diabetes Statistics Report.
6. Centers for Medicare and Medicaid Services. (2019). National Health Expenditures 2017 Highlights.
7. Daniel P. Martin. (2016, June 22).Clustering Mixed Data Types in R.