Shape Matters

Grant Cabrera, Shao Ting Chen, and Xianli Liu
Advisors: Ian Duncan and Xiyue Liao
UCSB Department of Statistic and Applied Probability
Sponsor: The Vitality Group

Abstract

It is well documented that diabetes and body mass index (BMI) are positively correlated. The purpose of this study is to investigate the relationship between diabetes and health factors. Using data from a wellness improvement company, The Vitality Group, we developed GLM and machine learning models to predict the prevalence of diabetes. We found that waist circumference is a significant predictor together with age, triglycerides, and BMI. Our results suggest that the distribution of body fat, namely in the waist, is strongly linked to diabetes which is even more significant than one's BMI.

Introduction

According to NIH (National Institute of Diabetes and Digestive and Kidney Diseases), diabetes is a disease where one's blood glucose levels, also called blood sugar levels, are too high.

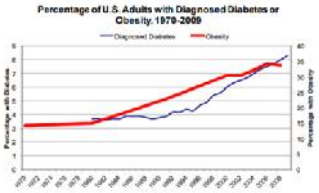


Figure 1

Recent observations suggest that BMI has peaked while prevalence of diabetes continues to rise, as per Figure 1. We observe the divergence between diabetes and obesity around 2006. This suggest a temporal relationship between the two and we wish to explore this relationship using other health measurements and factors. We also cluster our data set to discover any interesting sub populations.

Methods and Results

Data Pre-processing: Erroneous observations with abnormal age, weight, and height are removed. Set out-of-range values for each variable to NA.
MICE Algorithm: This was used for data imputation, since data set has a large number of missing values. Variables with missing entries, except for FPG and HbA1c, are all imputed.
Population Subset: We separated the data set into two subsets, one with one-year participants in this program, and another with two or more years participants. We analyzed the subset with one-year participants.
Models: Logistic Regression and Random Forests (Figure 3)
Clustering: Partition around medoids (PAM) algorithm (Figure 4)

From our random forest model, we get a variable of importance plot.

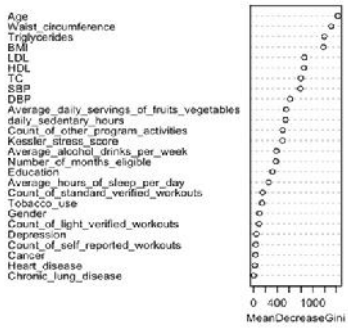
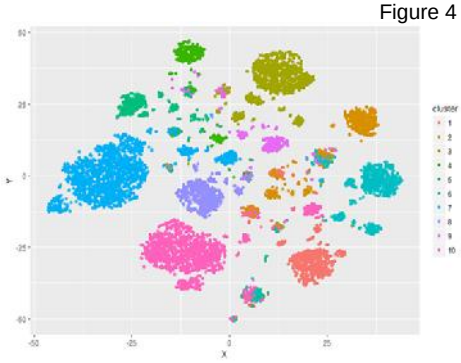


Figure 2

Model	Dataset Used	Predictors	Threshold	AUC	Sensitivity	Specificity	Accuracy	CV Accuracy
Logistic Regression	Whole dataset	All variables	0.08	0.7903	0.7127	**0.7380**	0.7359	
Logistic Regression	Balanced dataset	All variables	0.50	**0.8000**	**0.8486**	0.5753	**0.7465**	0.7279
Random Forests	Whole dataset	All variables	0.08	0.7882	0.7514	0.7041	0.7078	
Random Forests	Balanced dataset	All variables	0.50	0.7973	0.7675	0.6908	0.7445	
Logistic Regression	Whole dataset	Important variables	0.08	0.7897	0.7094	**0.7382**	0.7358	
Logistic Regression	Balanced dataset	Important variables	0.50	**0.7999**	**0.8486**	0.5757	**0.7466**	0.7262

Figure 3

Figure 4



Clusters

1. Male with postgraduate degree
2. Male college graduate; AA alcohol consumption
3. Female college dropout; high stress levels
4. Female college graduate; smoking history
5. Mostly female high school graduate; BA alcohol consumption
6. Male college dropout; AA BMI, AA waist circumference
7. Female college graduate; no smoking
8. Female with post graduate degree; BA BMI, BA waist circumference
9. Female college dropout; high depression rate with a smoking history
10. Male college graduate; no smoking, AA waist circumference

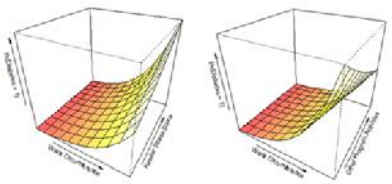


Figure 5

Based on our model, age and waist circumference resulted as the most significant predictors. Plotted in Figure 5 is thhe two most significant variables in our model to see how they affect the chances of having diabetes. If we fixed a low waist circumference here, in the red region, we could see that chances of having diabetes increased only slightly as one ages. However, if a person had a large waist circumference, around the yellow to almost-white region, the aging process greatly increased the chance of having diabetes. Waist circumference and BMI showed a similar.

Discussion

We used two methods, logistic regression and random forests, to build a model predicting the chances of having diabetes. We preferred the logistic regression model not only because it performed best, but also because it was easier to interpret, since each important variable had an estimated coefficient that could be utilized in other statistics or plots to better understand its relation to diabetes. We want to emphasize that there are other health factors such as age and waist circumference which are just as important as BMI in predicting diabetes.We also point out that our work is unfinished. Models can always be improved, and more sophisticated models can be tried on our data set. There is also work to be done on the two or more year longitudinal data.

.

References

[1] What Is Diabetes? National Institute of Diabetes and Digestive and Kidney Diseases, U.S. Department of Health and Human Services, 1 Nov. 2016, www.niddk.nih.gov/health-information/diabetes/overview/whatis-diabetes
[2] Statistics About Diabetes. American Diabetes Association, www.diabetes.org/diabetes-basics/statistics/.
[3] Logue, J., et al. Do Men Develop Type 2 Diabetes at Lower Body Mass Indices than Women? Diabetologia, vol. 54, no. 12, Dec. 2011, pp. 30033006., doi: https://doi.org/10.1007/s00125-011-2313-3.