GPH-GU2338 Final Project

zz3449 Zitao Zheng, zw2929 Zitong Wang

4/21/2021

Introduction

Nowadays, diabetes had become one of the most serve chronic health condition that challenging the well-being of human. Darrell J. Gaskin and LaVeist (2014) had mentioned that socio-economic contributes to the disparities in diabetes. Thus, base on that motivation, in our project, we would like to discover whether there exists a relationship between diabetes and people's socio-economic status (i.e family income, education background and etc.) or not by comparing 6 existing classification methods (logistic regression, KNN and random forest) and one variable selection method (selecting base on lasso). The data set that will be used for this project is from the GSS library of University of Chicago.

Data and Experiment Setup

```
## Rows: 1,088
## Columns: 8
## $ age      <dbl> 42, 59, 43, 62, 55, 59, 34, 44, 75, 40, 34, 40, 37, 56, 68...
## $ educ     <dbl> 16, 13, 12, 8, 12, 19, 14, 16, 12, 20, 15, 14, 12, 16, 12,...
## $ maeduc   <dbl> 16, 6, 12, 12, 20, 12, 14, 12, 12, 19, 17, 16, 18, 16, 12,...
## $ race     <fct> 1, 2, 1, 1, 1, 1, 3, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ income   <fct> 12, 10, 12, 5, 12, 12, 11, 12, 12, 12, 12, 12, 12, 12, 12,...
## $ hyperten <chr> "Yes", "No", "No", "Yes", "No", "Yes", "No", "No", "Yes", ...
## $ diabetes <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0...
## $ gender1  <chr> "Male", "Female", "Male", "Female", "Male", "Female", "Fem...
```

Methods

Logsitic Regression

The confusion matrix of logistic regression is:

```
##    glm.pred
##       0   1
##   0 214  83
##   1  12  18
```

The test error that computed for logistic regression is: 0.2905199

Variable Selection of Logistic Regression

Hosmer, Jovanovic, and Lemeshow (1989) had mentioned that, for logistic regression, since R^2 is not an proper goodness-of-fit measure of logistic regression, the plausible model selection method that we should consider is based on maximum likelihood method, It should be plausible by using **bestglm library**. However, based scope of our class we will perform the selection using lasso and cross-validation.

```
## (Intercept)         age        educ     income6    income12 hypertenYes
##           1           2           3          11          17          18
## gender1Male
##          19
```

From the process we mentioned previously, the best lambda value that we have for selection is: 0.0088247. Based on the prediction, the variables that chose are: **age, educ, income6 (5000 - 5999), income12 (20000 - 24999), hypertenYes, gender1Male**.

Linear Discriminant Analysis and Quadratic Discriminant Analysis

We will use the variables that selected above for this section.

LDA

The confusion matrix of LDA is:

```
##
##       0   1
##   0 296   1
##   1  30   0
```

The test error rate for classification using LDA method is: 0.0948012

QDA

The confusion matrix of QDA is:

```
##
##       0   1
##   0 295   2
##   1  26   4
```

The test error rate for classification using QDA method is: 0.0856269

KNN with Cross Validation

After a 10-fold cross validation, the k that helped achieving the minimum test error rate is K = 5, the minimum test error rate that we have is: 0.088685

Random Forest and Bagging

Random Forest

After 10-fold cross-validation, the best mtry that we have is: 2 and the validation error is: 0.3080697

The confusion matrix of Random Forest is:

```
##    rf.pred
##       0   1
##   0 188 109
##   1   7  23
```

The test error rate for classification using random forest with best number of tree is: 0.3547401

Bagging

After 10-fold cross-validation, the validation error is: 0.3087378

The confusion matrix of Random Forest is:

```
##    bg.pred
##       0   1
##   0 219  78
##   1  14  16
```

The test error rate for classification using bagging with best number of tree is: 0.2813456

Results

Classifiers	Test error rate
Logsitic Regression	0.2905199
LDA	0.0948012
QDA	0.0856269
KNN	0.088685
Random Forest	0.3547401
Bagging	0.2813456

Comparing six different classifiers, we can found the best classifier is QDA which test error rate is 0.0856269.And we can also know educ(education), age, hyperten(Told have hypertension or high blood pressure) and gender are variables which help us to determine people have diabetes or not. Hyperten has a slightly greater impact on the predicted results.

Discussion

Limitations of the data are also under our considerations. The variables offered is not comprehensive enough.Take hormone drugs or not? Have offspring or not, BMI and family diabetes history can also be the factors that we can detect. Clinical data from the hospital might be more accurate and more comprehensive.

For this diabetes topic, we might need more diabetes related knowledge. Like gender, females are more likely to get diabetes. The potential reason may be pregnant. Pregnant women tend to have higher blood sugar, which is also a risk factor for diabetes.This a point we might ignore. For education, highly educated people are less likely to get diabetes. Education is actually a complex variable which might represent people's living condition or if have a good health care. There are more social factors that we can do the deep research to detect the relationship between diabetes and these various factors. And we also consider problems like determination of variables to use might be biased. etc.

Decomposition

Zitong and Zitao had contributed equally towards this project that Zitao was mainly in charge of **data cleaning, discussion of results as well as conclusion** while Zitong was in charge of the **method section (selection of models, tuning of models)**.

Reference

Darrell J. Gaskin, Emma E. McGinty, Roland J. Thorpe Jr, and Thomas A. LaVeist. 2014. "Disparities in Diabetes: The Nexus of Race, Poverty, and Place." *American Journal of Public Health* 104 (11): 2147–55.

Hosmer, David W., Borko Jovanovic, and Stanley Lemeshow. 1989. "Best Subsets Logistic Regression." *Biometrics* 45 (4): 1265–70. http://www.jstor.org/stable/2531779.