Diabetes prediction using NHANES

Partha Pratim Das

Roll-MD2114

Introduction-

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

The NHANES program began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics. In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year.

The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

Findings from this survey will be used to determine the prevalence of major diseases and risk factors for diseases. Information will be used to assess nutritional status and its association with health promotion and disease prevention. NHANES findings are also the basis for national standards for such measurements as height, weight, and blood pressure. Data from this survey will be used in epidemiological studies and health sciences research, which help develop sound public health policy, direct and design health programs and services, and expand the health knowledge for the Nation.

Here we try to build classifiers for classifying whether people are diabetic or not. And to do that, we pick variablescorresponding to various factors that are linked with diabetes.

Use of the Data-

Information from NHANES is made available through an extensive series of publications and articles in scientific and technical journals. For data users and researchers throughout the world, survey data are available on the internet and on easy-to-use CD-ROMs.

Research organizations, universities, health care providers, and educators benefit from survey information. Primary data users are federal agencies that collaborated in the design and development of the survey. The National Institutes of Health, the Food and Drug Administration, and CDC are among the agencies that rely upon NHANES to provide data essential for the implementation and evaluation of program activities. The U.S. Department of Agriculture and NCHS cooperate in planning and reporting dietary and nutrition information from the survey.

NHANES' partnership with the U.S. Environmental Protection Agency allows continued study of the many important environmental influences on our health.

NHANES' record of important accomplishments is made possible by the thousands of Americans who have participated.

- Past surveys have provided data to create the growth charts used nationally by pediatricians to evaluate children's growth. The charts have been adapted and adopted worldwide as a reference standard – and have recently been updated using the latest NHANES figures.
- Blood lead data were instrumental in developing policy to eliminate lead from gasoline and in food and soft drink cans. Recent survey data indicate the policy has been even more effective than originally envisioned, with a decline in elevated blood lead levels of more than 70% since the 1970s.
- Overweight prevalence figures have led to the proliferation of programs emphasizing diet and exercise, stimulated additional research, and provided a means to track trends in obesity.
- Data have continued to indicate that undiagnosed diabetes is a significant problem in the United States. Efforts by government and private agencies to increase public awareness, especially among minority populations, have been intensified. These are just a few examples of what survey findings have meant. The current program promises continuing contributions and some new initiatives.
- Information collected in this survey will help the Food and Drug Administration decide if there is a need to change vitamin and mineral fortification regulations for the Nation's food supply.
- National programs to reduce hypertension and cholesterol levels continue to depend on NHANES data to steer education and prevention programs toward those at risk and to measure success in curtailing risk factors associated with heart disease, the Nation's number one cause of death.
- New measures of lung function will further our understanding of respiratory disease and better describe the burden of asthma in the United States.

Because NHANES is now an ongoing program, the information collected contributes to annual estimates in topic areas included in the survey. For small population groups and less prevalent conditions and diseases, data must be accumulated over several years to provide adequate estimates. The new continuous design also allows increased flexibility in survey content.

Results of NHANES benefit people in the United States in important ways. Facts about the distribution of health problems and risk factors in the population give researchers important clues to the causes of disease. Information collected from the current survey is compared with information collected in previous surveys. This allows health planners to detect the extent various health problems and risk factors have changed in the U.S. population over time. By identifying the health care needs of the population, government agencies and private sector organizations can establish policies and plan research, education, and health promotion programs that help improve present health status and will prevent future health problems.

Choosing variables

The response, the factors affecting the response and the variables corresponding to them were chosen as follows:-

- **Response :**
  - diabetic : LBXGH(glycohemoglobin) is used to classify whether a person is diabetic or not. If $LBXGH > 6.4$, then the person is diabetic, otherwise they are not. 1 represents diabetic, 0 non-diabetic.
- **Gender :**
  - gender : RIAGENDR from the demographic dataset gives the gender of the person. 1 represents

male and 2 represents female.

- **Age :**
  - age : RIDAGEYR from the demographic component gives the age in years.
- **Family Income :**
  - Family_income : INDFMPIR from the demographic component gives the income ratio and is used as a measurement of family income for the person.
- **Obesity :**
  - ArmCircum : BMXARMC from the examination component gives the arm circumference.
  - SaggitalAbdominal : BMDAVSAD gives the average saggital obdominal diameter.
- **Hypertension :**
  - Hypertension : BPXSY1,BPXSY2,BPXSY3,BPXDI1,BPXDI2,BPXDI3 are averaged out to give average blood pressure readings. All observations with systole BP > 120 or diastole BP >80 are classified as hypertensive, denoted by 1. Otherwise, they are considered as not hypertensive and denoted by 0. All these variables are in the laboratory component.
- **Grip Strength :**
  - GripStrength : MGDCGSZ from the examination component gives the grip strength of the person.
- **Triglycerides :**
  - triglycerides : LBXTR from the laboratory component gives the grip strength of the person.

```
library(tidyverse)
library(caret)
library(MASS)
library(class)
library(ROSE)
demographic <- read_csv("Data(13-14)/demographic.csv",col_select =
                          list(SEQN,RIAGENDR,RIDAGEYR,INDFMPIR))
names(demographic) <- c("ID","gender","age","Family_income")
examination <- read_csv("Data(13-14)/examination.csv",col_select =
                          list(SEQN,BMXARMC,BMDAVSAD,MGDCGSZ,BPXSY1,BPXSY2,
                               BPXSY3,BPXDI1,BPXDI2,BPXDI3))
names(examination) <- c("ID","ArmCircum","SaggitalAbdominal","GripStrength",
                        "SystoleBP1","SystoleBP2","SystoleBP3",
                        "DiastoleBP1","DiastoleBP2","DiastoleBP3")
labs <- read_csv("Data(13-14)/labs.csv",col_select = list(SEQN,LBXGH,LBXTR))
names(labs) <- c("ID","glycoHemoglobin","triglycerides")
```

Data cleaning and tidying

- All observations with any missing covariates aside from BP readings or with income 0 are removed. BP readings are present for all who had triglyceride readings taken, so removing observations with missing values in triglycerides variable is enough.
- Variables from the above dataset are modified to form the final set of variables.

```
data <- inner_join(demographic,examination,by="ID")
data <- inner_join(labs,data,by="ID")

diabetes_data<- as_tibble(data[complete.cases(data),])
rm(list = c("demographic","examination","labs","data"))
diabetes_data1 <- diabetes_data %>%
  mutate(systole=(SystoleBP1+SystoleBP2+SystoleBP3),
         Diastole=(DiastoleBP1+DiastoleBP2+DiastoleBP3),
         diabetic=ifelse(glycoHemoglobin>=6.4 ,1,0))%>%
  dplyr::select(-1,-2)%>%filter(Family_income!=0)

#Calculate division constants for averaging out diastole
i <- with(diabetes_data1,which(DiastoleBP1==0|DiastoleBP2==0|DiastoleBP3==0))
div <- rep(3,nrow(diabetes_data1))
for (j in 1:nrow(diabetes_data1)){
  if (j %in% i) div[j]=2
}
rm(i,j)

#The final dataset
diabetes_data2 <- diabetes_data1 %>%
  mutate(systole=systole/div,Diastole=Diastole/div,Hypertension=ifelse(systole>120 | Diastole>80,1,0))%
  dplyr::select(-contains("stole"))%>%relocate(diabetic,.after = Hypertension)
rm(div)
head(diabetes_data2)
```

```
## # A tibble: 6 x 9
##   triglycerides gender   age Family_income ArmCircum SaggitalAbdominal
##           <dbl>  <dbl> <dbl>         <dbl>     <dbl>              <dbl>
## 1            51      1    72          4.51      33.5               25.6
```

```
## 2            64      2    61          5         38                26.7
## 3            24      2    26          5         25.8              14.5
## 4            14      2    33          2.1       26.5              15.1
## 5            57      1    16          1.58      30.7              17
## 6           148      1    32          0.29      34.5              22.8
## # ... with 3 more variables: GripStrength <dbl>, Hypertension <dbl>,
## #   diabetic <dbl>
```

```
#Some useful variables for later use
response <- diabetes_data2$diabetic
hyperT <- diabetes_data2$Hypertension
gender <- diabetes_data2$gender

#Converting categorical responses into factors
diabetes_data2$diabetic <- factor(diabetes_data2$diabetic)
diabetes_data2$gender <- factor(diabetes_data2$gender)
diabetes_data2$Hypertension <- factor(diabetes_data2$Hypertension)
```

Splitting data into training and testing

- The dataset is split into 75 : 25 ratio of training data and testing data.

```
########Data splitting########
set.seed(123)
training.samples <- diabetes_data2$diabetic %>%
  createDataPartition(p = 0.75, list = FALSE)

train.data <- diabetes_data2[training.samples, ]
train.response <- response[training.samples]
train.hyperT <- hyperT[training.samples]
train.gender <- gender[training.samples]

test.data <- diabetes_data2[-training.samples, ]
test.response <- response[-training.samples]
test.hyperT <- hyperT[-training.samples]
test.gender <- gender[-training.samples]
```

- The data is then normalised by centering and scaling the quantitative variables and leaving the qualititative variables as they were.

```
########Transform data###############
preproc.param <- train.data %>%
  preProcess(method = c("center", "scale"))
# Transform the data using the estimated parameters
train.transformed <- preproc.param %>% predict(train.data)
test.transformed <- preproc.param %>% predict(test.data)
```

Simplifying assumptions

The NHANES dataset has a complex survey design, due to which the samples are not identically and independently distributed. This issue is ignored in the proceeding analysis.

Classifiers

Brief Description of LDA and QDA-

LDA (Linear Discriminant Analysis) is used when a linear boundary is required between classifiers and QDA (Quadratic Discriminant Analysis) is used to find a non-linear boundary between classifiers. LDA and QDA work better when the response classes are separable and distribution of X=x for all class is normal. In general LDA is less flexible than QDA.

LDA

- Even though there are categorical covariates and hence the data matrix is not normal, we still see how the LDA performs as a classifier :

```
model <- lda(diabetic~., data = train.transformed)
```

- Comparison of predictions vs observed for training data :

```
##           reference
## prediction0    1
##          0 1551  151
##          1   21   20
```

```
## Accuracy =  0.9013196
```

```
## Specificity =  0.1169591
```

```
## Sensitivity =  0.9866412
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 519  53
##          1   5   3
```

```
## Accuracy =  0.9
```

```
## Specificity =  0.05357143
```

```
## Sensitivity =  0.990458
```

QDA

- Similarly, we train a QDA classifier and evaluate its performance :

```
model <- qda(diabetic~., data = train.transformed)
```

- Comparison of predictions vs observed for training data :

```
##           reference
## prediction0    1
##          0 1533  138
##          1   39   33
```

```
## Accuracy =  0.8984509
```

```
## Specificity =  0.1929825
```

```
## Sensitivity =  0.9751908
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 502  52
##          1  22   4
```

```
## Accuracy =  0.8724138
```

```
## Specificity =  0.07142857
```

```
## Sensitivity =  0.9580153
```

Logistic

- We fit a logistic model for the response :

```
log.model0 <- glm(diabetic ~ ., data = train.transformed,
                  family = binomial(link="logit"))
```

- Comparison of predictions vs observed for training data :

```
##           reference
## prediction0   1
##          0 1555  153
##          1   17   18
```

```
## Accuracy =  0.902467
```

```
## Specificity =  0.1052632
```

```
## Sensitivity =  0.9891858
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 518  51
##          1   6   5
```

```
## Accuracy =  0.9017241
```

```
## Specificity =  0.08928571
```

```
## Sensitivity =  0.9885496
```

kNN

```
knn_model <- knn(train.transformed[,-9],test.transformed[,-9],cl=train.transformed$diabetic,k=5)
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 512  51
##          1  12   5
```

```
## Accuracy =  0.8913793
```

```
## Specificity =  0.08928571
```

```
## Sensitivity =  0.9770992
```

Problem of low sensitivity

For all of the above classifiers sensitivity is very low.Sensitivity is the probability of a person classified as diabetic, given that he truly is diabetic. This means that the classifiers aren't able to classify diabetic people as diabetic properly. This is happening because the diabetic population is only around 10% of the total population. And this imbalance causes all the classifiers to be biased towards classifying into the majority group. To combat this, we try oversampling to balance the data and then train our classifiers.

Balancing the dataset

```
over_data <- ovun.sample(diabetic~., data = train.transformed, method = "over", N = 2000)$data
```

LDA

- Even though there are categorical covariates and hence the data matrix is not normal, we still see how the LDA performs as a classifier :

```
model <- lda(diabetic~., data = over_data)
```

- Comparison of predictions vs observed for training data :

```
##           reference
## prediction0   1
##          0 1474  283
##          1   98  145
```

```
## Accuracy =  0.8095
```

```
## Specificity =  0.338785
```

```
## Sensitivity =  0.937659
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 497  36
##          1  27  20
```

```
## Accuracy =  0.8913793
```

```
## Specificity =  0.3571429
```

```
## Sensitivity =  0.9484733
```

QDA

- Similarly, we train a QDA classifier and evaluate its performance :

```
model <- qda(diabetic~., data = over_data)
```

- Comparison of predictions vs observed for training data :

```
##           reference
## prediction0   1
##          0 1508  310
##          1   64  118
```

```
## Accuracy =  0.813
```

```
## Specificity =  0.2757009
```

```
## Sensitivity =  0.9592875
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 492  48
##          1  32   8
```

```
## Accuracy =  0.862069
## Specificity =  0.1428571
## Sensitivity =  0.9389313
```

Logistic

- We fit a logistic model for the response :

```
log.model0 <- glm(diabetic ~ ., data = over_data,
                  family = binomial(link="logit"))
```

- Comparison of predictions vs observed for training data :

```
##           reference
## prediction0    1
##          0 1477  293
##          1   95  135
## Accuracy =  0.806
## Specificity =  0.3154206
## Sensitivity =  0.9395674
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 495  33
##          1  29  23
## Accuracy =  0.8931034
## Specificity =  0.4107143
## Sensitivity =  0.9446565
```

kNN

```
knn_model <- knn(over_data[,-9],test.transformed[,-9],cl=over_data$diabetic,k=5)
```

- Comparison of predictions vs observed for testing data :

```
##           reference
## prediction0  1
##          0 452  30
##          1  72  26
## Accuracy =  0.8241379
## Specificity =  0.4642857
## Sensitivity =  0.8625954
```

Related Links-

- Introduction to NHANES pdf icon[PDF – 1.2 MB]
- Institutional Review Board Approval
- Survey Contents 1999-2014 pdf icon[PDF – 2.3 MB]
- Survey Operations Schedule pdf icon[PDF – 45 KB]

- NHANES Data Release and Access Policy pdf icon[PDF – 79 KB]

Thanks Giving-

I would like to thank prof. Deepayan Sarkar for his continuous support to do assignment without which this project can not be successfully done.