MGMTMSA 408 – Operations Analytics

Homework 4 – Healthcare Modeling (Question Sheet)

Due Friday June 7, 2024 at 1pm PST

Note: Please use R to solve this problem.

1 Diabetes risk prediction

Diabetes is a disease characterized by high blood sugar levels sustained over a long period of time. If untreated, diabetes can lead to numerous complications, such as stroke, chronic kidney disease and blindness. It is estimated that 425 million people worldwide have diabetes, and the total costs of diabetes to the United States healthcare system are over $200 billion a year.

In this problem, let us suppose that we are working for a healthcare provider to create a risk prediction model for diabetes. The idea is to use demographic information (e.g., age, education level, home ownership status, etc.) as well as basic medical information (e.g., direct cholesterol, blood pressure, etc.) to make a probabilistic prediction of whether the patient has diabetes.

We will use data from the National Health and Nutrition Examination Survey (NHANES) to understand which demographic and clinical factors are predictive of diabetes. This data is contained in the file `nhanes-diabetes-final.csv`. Each observation corresponds to a subject who completed a survey and who had certain lab measurements taken. The dependent variable we will try to predict is `Diabetes`, which is a binary variable that is 1 if the subject has been diagnosed with diabetes, and 0 otherwise. The remaining variables are summarized in the table below.

Part 1: Understanding the data

Load the `nhanes-diabetes-final.csv` data set into a dataframe named `diabetes`. Answer the following questions using the whole data set. Do not build any predictive models. Be sure to justify your answer by including the output of any appropriate commands in R.

a) How many individuals are there in the data set?

b) What fraction of individuals have diabetes?

c) Which level of education is associated with the highest risk of diabetes? Which is associated with the lowest risk?

d) Obesity is defined as an individual having a body mass index (BMI; weight in kilograms divided by height in meters squared) of over 30. Based on the data, is obesity a risk factor for diabetes?

Variable	Description
`Gender`	One of male or female
`Age`	Age of subject (note: all subjects are between 18 and 60)
`Race3`	Subject's race (coded as: 1 = Asian, 2 = Black, 3 = Hispanic, 4 = Mexican, 5 = Other, 6 = White)
`Education`	One of: 8th Grade, 9 - 11th Grade, College Grad, High School, Some College
`MaritalStatus`	One of: Divorced, LivePartner, Married, NeverMarried, Separated, Widowed
`HHIncome`	Household income of subject
`HomeOwn`	Category of subject's home ownership (one of Other, Own, Rent)
`Weight`	Subject's weight in kg
`Height`	Subject's height in cm
`Pulse`	Subject's pulse
`BPSysAve`	Average systolic blood pressure
`BPDiaAve`	Average diastolic blood pressure
`Testosterone`	Total testosterone level of subject
`DirectChol`	Direct HDL cholesterol measurement
`TotChol`	Total cholesterol measurement
`UrineVol1`	Urine volume in mL at first test
`SleepHrsNight`	Number of hours of sleep per night
`SleepTrouble`	Whether subject has trouble sleeping (yes/no)
`PhysActive`	Is the subject physically active (yes/no)
`PhysActiveDays`	Num. days per week of vigorous phys. activity
`LittleInterest`	Num. days per year in which subject had little interest in doing things (one of None, Several, Majority or AlmostAll)
`Depressed`	Num. days per year in which subject reported depression (one of None, Several, Majority or AlmostAll)
`TVHrsDay`	Num. hours of TV per day
`CompHrsDay`	Num. hours of computer use per day
`AlcoholDay`	Avg. num. alcoholic drinks consumed on days that subject drinks alcohol
`SmokeNow`	Subject currently smokes (yes/no)
`Smoke100`	Has subject smoked at least 100 cigarettes in entire life?

Part 2: A first logistic regression model

Next, we will develop an initial predictive model. Set your seed to 40 and split the data randomly into a training and a testing set. Use a 70-30 split and ensure that the relative proportion of the two levels of the dependent variable is preserved in the two sets. Estimate a logistic regression model from the training set using only the gender, age, household income and home ownership variables. Answer the following questions:

a) Which variables are statistically significant at the $\alpha = 0.05$ level? (For categorical variables, include the variable if at least one of its dummy variables is significant.)

b) Consider a 50 year old man who lives in a rented apartment, with a household income of \$66,000. What are the log-odds of him having diabetes? What are the odds of him having diabetes? What is the predicted probability of him having diabetes?

Part 3: A richer logistic regression model

Now, estimate a logistic regression model using all of the independent variables. Use the same training and testing sets from Part 2. Answer the following questions:

a) Use the model to make predictions on the test set. Use a threshold of 0.5. What is the test set accuracy of the model?

b) You show your model to a stakeholder at the healthcare provider who does not understand machine learning very well. When they see the result in (a), they become very excited. Explain why this excitement is unwarranted.

c) In class, we discussed another metric for quantifying predictive performance of classification models. What is that metric? What is the value of that metric for this model? Explain why this metric is more appropriate to use for this problem.

d) Besides the difference in the number of variables, and notwithstanding the difference in test set performance between this model and the model in Part 2, why might the healthcare provider prefer the model in Part 2?

Part 4: A smaller logistic regression model

Next, estimate a L1-regularized (LASSO) logistic regression model to obtain a logistic regression smaller than the one in Part 3. Use five-fold cross validation, and set your random number seed to 2000 beforehand. Use the same training and testing sets from Part 2. For all of the questions below, use `s = "lambda.min"` when accessing predictions or coefficients.

a) How many variables does your model use?

b) What is the test set accuracy of your model? (Use a threshold of 0.5.)

c) What is the test set AUC of your model?

Part 5: A random forest model

Next, let's develop a better model. Set the seed to 2000 beforehand, and estimate a random forest model. Use the `randomForest` package in R (**do not** give the additional input parameter `importance = TRUE`). Use all of the independent variables.

a) What is the test set accuracy of your model? Use a threshold of 0.5.

b) What is the test set AUC of your model?

c) Compare your answers in (a) and (b) to your answers in Part 3 and Part 4. What does the difference in performance imply about the underlying relationship between the independent variables and the risk of diabetes?

d) Calculate the sensitivity and the specificity of the model at a threshold of 0.20.

e) Suppose that the random forest model, at a threshold of 0.20, were to be used for a new population of patients for which it is known that 10% of the patients have diabetes. (Note that this is different from the current data set.) Based on your answer to (d), what accuracy would you expect the model to have in this new population of patients?

Part 6: Operationalizing the model

Suppose that the healthcare provider is interested in using your random forest model to identify individuals in the test set to screen for diabetes. The healthcare provider has collected the independent variables listed on page 1 for those individuals, but does not know whether these individuals have diabetes or not.

a) How could you use the predictions of your random forest model to make this decision?

b) Suppose the healthcare provider will enroll 150 individuals from the test set. Based on your answer to (a), determine this set of individuals. How many of these individuals actually have diabetes?

c) Suppose that instead of using your model, the healthcare provider selects the 150 individuals from the test set at random. Simulate this selection policy 100 times. Averaging over the 100 repetitions, how many of the 150 selected individuals have diabetes? *Hint*: You may find the following R code snippet useful:

```
> indices = sample(c(1:nrow(diabetes.test)), 150, replace = FALSE)
```

where `diabetes.test` is the test set, 150 is the number of individuals to sample and `replace = FALSE` indicates to sample without replacement.

d) Comparing your answers to (b) and (c), is the model useful? Explain your answer.

e) How does your answer to part (b) change if you use the model in Part 4 of this question? Is your answer lower or higher than part (b)? Explain why this makes sense.