IEOR 242: Applications in Data Analysis, Spring 2018

Homework Assignment #2

February 3, 2018

Problem 1: (15 points)

James Chapter 4, Exercise 6.

Problem 2: (15 points)

James Chapter 4, Exercise 7.

Problem 3: Framingham Heart Study (Adapted from Bertsimas Chapter 7) (70 points)

Heart disease is the leading cause of death worldwide. About 7.4 million people died from coronary heart disease (CHD) in 2012, which is 13% of all deaths that year across the globe.

In the late 1940s, the U.S. government took steps to study cardiovascular disease. In order develop high quality data for their study, they decided to track a large cohort of initially-healthy people over time. The town of Framingham, Massachusetts (a suburb of Boston) was selected as the site for the study, which commenced in 1948. The study enrolled 5,209 participants aged 30-62. Participants were given a questionnaire and a medical exam every two years. They also collected data on the participants' physical characteristics and behavioral characteristics, in addition to the medical test data. Over the years, the study has expanded to include multiple generations and has collected many more factors including genetic information. This data is now famously known and is simply called the Framingham Heart Study.

In this exercise, you are asked to build models using Framingham Heart Study data in order to predict CHD and to make recommendations to better prevent heart disease. The dataset is in the file `framingham.csv`. There are 3,658 observations, with each observation representing the data from a particular study participant. There are 16 variables in the dataset, which are described in Table 1. You will be asked to predict `TenYearCHD` (whether the patient experiences coronary heart disease within 10 years of their first examination). As a consequence of your modeling efforts, you should be able to identify *risk factors*, which are the variables that increase the risk of CHD.

a) (40 points) To lower the risk of CHD, physicians can prescribe preventive medication such as blood-pressure-lowering or cholesterol-lowering medications. Many policy makers, when recommending certain preventive medications to patients at risk of developing CHD, rely on

Table 1: Variables in the dataset `framingham.csv`.

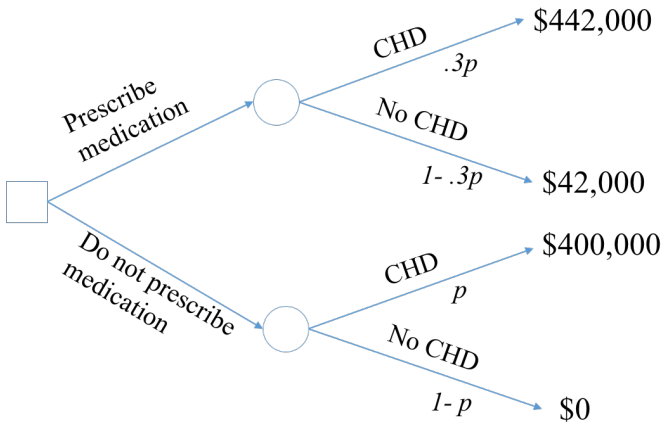
Variable	Description
`male`	Gender of patient
`age`	Age (in years) at first examination
`education`	Some high school, high school/GED, some college/vocational school, college
`currentSmoker`	Is a current smoker
`cigsPerDay`	Number of cigarettes per day
`BPMeds`	Is on blood pressure medication at time of first examination
`prevalentStroke`	Previously had a stroke
`prevalentHyp`	Currently hypertensive
`diabetes`	Currently has diabetes
`totChol`	Total cholesterol (mg/dL)
`sysBP`	Systolic blood pressure
`diaBP`	Diastolic blood pressure
`BMI`	Body Mass Index, weight (kg)/height $(m)^2$
`heartRate`	Heart rate (beats/minute)
`glucose`	Blood glucose level (mg/dL)
`TenYearCHD`	Experienced coronary heart disease within 10 years of first examination

evidence-based analysis that weighs the pros and cons of such interventions. Health economic evaluation is a commonly applied methodology for decision-making that takes both medical costs and health benefits (a monetized version of improved life longevity) into consideration. In fact, many countries establish clinical practice guidelines using such formalized health economic evaluation methodologies (the National Institute for Health and Clinical Excellence in England, for example).

As prior work, let us suppose that a colleague of yours has completed a health economics study analyzing the costs and benefits of a recently approved medication aimed at preventing CHD. The colleague determined that patients who experience CHD within the next 10 years are expected to incur a lifetime cost of $400,000 associated with the disease; this cost includes both the costs of treatment for CHD, $150,000, as well as a cost intended to capture the decreased quality and length of life experienced by patients with CHD, which is $250,000. Your colleague has determined that patients who take the preventative medicine being studied

will have their probability of developing CHD within the next 10 years reduced by 70%; that is, if their current 10-year risk of developing CHD is p without taking the medication, then their 10-year risk with the medicine would instead be $.3p$. Regardless of whether a patient eventually develops CHD, there is a $42,000 cost associated with taking this recently approved medication. A decision tree capturing your colleague's analysis is shown in Figure 1.

Figure 1: Decision tree for prescribing the approved medication to prevent CHD. The leaf nodes represent cost values.



Using all of the provided independent variables, build a logistic regression model to predict the probability that a patient will experience CHD within the next 10 years. Use a randomly selected subset of 70% of the data to train your model. In R, be sure to use the `sample.split` function, which ensures that training and test sets have approximately equal proportions of people with `TenYearCHD` to people without `TenYearCHD` (i.e., *stratified/proportional sampling*). Please answer the following questions concerning your model.

i) What is the fitted logistic regression model? Do not provide output from R, but instead state the equation used by the model to make predictions.

ii) What are the most important risk factors for 10-year CHD risk identified by the model? Pick one of these variables and use odds ratios to describe its impact on a patient's predicted odds of developing CHD in the next 10 years.

iii) Suppose that you wish to determine the optimal strategy for assigning which patients receive the medication. Given your colleague's analysis of the costs and benefits associated with the recently approved treatment, identify a threshold value of p such that it is optimal to prescribe the medication to a patient if and only if her 10-year CHD risk exceeds p.

iv) Describe the test set performance of the logistic regression model, using the threshold identified in part (*iii*) to separate patients into those who are at high risk for CHD (risk exceeding the threshold p) and those who are at low risk for CHD (risk below the threshold p). Describe the model's accuracy, True Positive Rate (TPR), and False Positive Rate (FPR). (Remember that your reported results should be accessible to a non-technical audience.)

v) If patients are prescribed the medication using the strategy implied by the model, use the test set data to provide an estimate(s) for the expected economic cost per patient. You should first report your estimate assuming that the CHD outcomes in the test set are not affected by the treatment decision. Is this assumption reasonable? You should then adjust your estimate in a way that takes into account the fact that the treatment decision impacts a patient's risk of developing CHD. (Hint: keep in mind that the probability estimates output by your logistic regression model do not necessarily reflect the *true* probability of a patient developing CHD.)

vi) Consider a simple baseline model that predicts none of the patients are at high risk for CHD and therefore does not recommend treatment for any of the patients. Describe the test set performance of the baseline model in terms of accuracy, TPR, and FPR, as well as expected economic cost per patient.

vii) Use an example to explain how to use the model in a real clinical setting. Suppose a new patient arrives, and the physician accesses the patient's electronic medical records and retrieves the following about the patient:

```
Female, age 51, college education, currently a smoker with an average of
20 cigarettes per day.  Not on blood pressure medication, has not had stroke,
but has hypertension.  Not diagnosed with diabetes; total Cholesterol at
220.  Systolic/diastolic blood pressure at 140/100, BMI at 31, heart rate
at 59, glucose level at 78.
```

What is the predicted probability that this patient will experience CHD in the next ten years? Based on the threshold from the decision tree, should the physician prescribe the preventive medication for this patient?

b) (10 points) Show the ROC curve for your logistic regression model on the test set and describe how this curve may be helpful to decision-makers looking to further study the medication you have considered so far in this homework as well as other possible medications for preventing CHD. Describe a few points from the ROC curve that you find interesting. What is the area under the curve (AUC) for your model in the test set?

c) (15 points) Again, using all of the provided independent variables, use linear discriminant analysis to train a model (on the same training set) to predict the probability that a patient will experience CHD within the next 10 years. Produce a plot with the test set ROC curves for both the logistic regression model and the LDA model. Report the test set AUC for the LDA model as well. Based on examining this plot and comparing the AUC values, which model do you recommend and why?

d) (5 points) Are there any aspects of the analysis performed thus far that raise ethical concerns? If so, suggest at least one way that this analysis could be changed to address such concerns.