15.071: The Analytics Edge **Fall 2022**
Homework 3: Logistic Regression

Out: October 3; Due: October 19.
Please post the assignment in pdf format with file name "Lastname_15071_HW3.pdf".
For each question, please include the main `R` *commands that you used in your submission.*

Problem 1: Predicting Survivors on the Titanic [20 pts]

You will predict build a model to predict whether or individual passengers on the Titanic survive the shipwreck. The data is contained in the file `titanic.csv`. There are 891 observations, each corresponding to a passenger. There are 12 variables in the dataset, which are described in Table 2. However, for this problem, you will only use the variables `Survived`, `Pclass`, `Sex`, and `SibSp`.

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

Variable	Description
`PassengerId`	A numerical ID assigned to each passenger
`Survived`	Whether or not the passenger survived the shipwreck (1 if survived, 0 if not)
`Pclass`	Ticket class (1 = first, 2 = second, 3 = third)
`Name`	Name of passenger
`Sex`	Sex of passenger (male, female)
`Age`	Age of passenger
`SibSp`	Number of siblings/spouses aboard the Titanic
`Parch`	Number of parents/children aboard the Titanic
`Ticket`	Ticket number
`Fare`	Ticket fare
`Cabin`	Cabin number
`Embarked`	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Import the data into `R` and ensure that categorical variables are properly represented using the following code.

```
data = read.csv("titanic.csv")
data$Pclass <- factor(data$Pclass)
data$Sex <- factor(data$Sex)
```

a) Using the independent variables `Pclass`, `Sex`, and `SibSp`, construct a logistic regression model to predict the probability that a passenger on the Titanic survives the shipwreck. **[5 pt]**

b) Explain in words how (according to your model) the passenger's ticket class, sex, and number of siblings/spouses aboard relates to their probability of surviving. **[5 pt]**

c) Now numerically, all else being equal, how does having two additional siblings/spouses aboard affect the odds of a passenger surviving? **[3 pt]**

d) Numerically, how does going from third class to second class affect the odds of a passenger surviving? **[4 pt]**

e) Numerically, how does going from third class to second class affect the probability of a passenger surving? (*Hint: This is a trick question. Explain why.*) **[3 pt]**

Problem 2: Framingham Heart Study [80 pts]

Heart disease is the leading cause of death worldwide. About 17.9 million people died from coronary heart disease (CHD) in 2016—over 25% of all deaths that year across the globe.

In the late 1940s, the U.S. government took steps to tackle heart disease. As part of this effort, it decided to track a large cohort of initially-healthy people over time. The town of Framingham, MA was selected as the site for the study. The study started in 1948, comprising 5,209 participants. Participants were given a questionnaire and a medical exam every two years. Data were also collected on the participants' physical and behavioral characteristics. Over the years, the study has expanded to include multiple generations and many factors—including genetic information. This dataset is now known as the Framingham Heart Study.

The data is contained in the file `framingham.csv`. There are 3,658 observations, each corresponding to a study participant. There are 16 variables in the dataset, which are described in Table 2. You will aim to predict `tenYearCHD`—that is, whether a patient experiences CHD within 10 years of his/her first examination. You will also aim to identify *risk factors*, in order to make recommendations to prevent CHD.

To lower the risk of CHD, physicians can prescribe preventive medication that lowers blood pressure or cholesterol. Recommending preventive medications requires evidence-based analysis that weighs the pros and cons of such interventions. A common methodology is known as *health economic evaluation*, which accounts for medical costs and health benefits (a monetized metric of improved life longevity). In fact, many countries establish clinical practice guidelines using such formalized health economic evaluation methodologies (e.g., the National Institute for Health and Clinical Excellence in England).

A colleague of yours has just completed a health economics study to assess a recently approved medication. The study has estimated that patients who experience CHD within the next 10 years are expected to incur a lifetime cost of $165,000 associated with the disease—including the costs of treatment ($80,000) as well as lower quality of life and life expectancy ($85,000). The medicine is expected to lower patients' risk of developing CHD within the next 10 years by a factor of 2.3. Regardless of whether a patient develops CHD, the preventive medication costs $7,500. A decision tree capturing the study's analysis is shown in Figure 1, where p denotes the probability that a patient will develop CHD within the next 10 years without medication.

a) What is the expected cost borne by a patient who does not take the preventive medication, as a function of p? And what is the expected cost borne by a patient who takes the preventive medication, as a function of p? For which values of p would you recommend the medication? **[10 pt]**

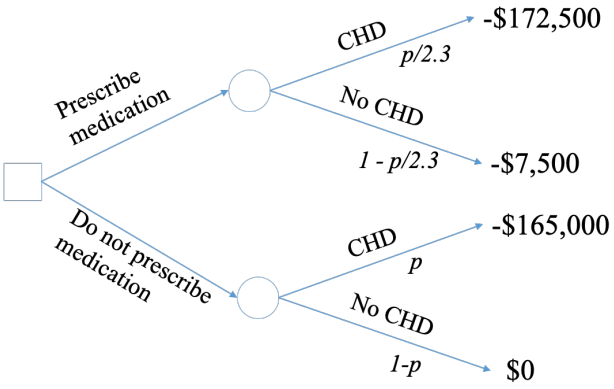
Import the data into `R` and split them randomly into a training set (containing 75% of the data) and a test set (containing the remaining 25% of the data). Use the following seed and commands:

```
library(caTools)
data = read.csv("framingham.csv")
data$TenYearCHD <- factor(data$TenYearCHD)
```

Table 2: Variables in the dataset `framingham.csv`.

Variable	Description
`male`	Gender of patient (1 if male, 0 if female)
`age`	Age (in years) at first examination
`education`	Some high school, high school/GED, some college/vocational school, college
`currentSmoker`	1 if patient is a current smoker, 0 otherwise
`cigsPerDay`	Number of cigarettes per day
`BPMeds`	1 if patient is on blood pressure medication at time of first examination, 0 otherwise
`prevalentStroke`	1 if patient previously had a stroke, 0 otherwise
`prevalentHyp`	1 if patient is currently hypertensive, 0 otherwise
`diabetes`	1 if patient currently has diabetes, 0 otherwise
`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`	1 if patient has experienced coronary heart disease within 10 years of first examination, 0 otherwise

Figure 1: Decision tree for prescribing the approved medication to prevent CHD.



```
data$male <- factor(data$male)
data$currentSmoker <- factor(data$currentSmoker)
data$BPMeds <- factor(data$BPMeds)
data$prevalentStroke <- factor(data$prevalentStroke)
data$prevalentHyp <- factor(data$prevalentHyp)
data$diabetes <- factor(data$diabetes)
set.seed(31)
N <- nrow(data)
idx = sample.split(data$TenYearCHD, 0.75)
train <- data[idx,]
test = data[!idx,]
```

b) Using all the independent variables in the dataset, construct a logistic regression model to predict the probability that a patient will experience CHD within the next 10 years. What are the most important risk factors for 10-year CHD identified by the model? Do these make intuitive clinical sense? **[10 pt]**

c) A physician retrieves the following information from a patient's electronic health records: the patient is a 55-year old college-educated male, smokes 10 cigarettes per day, is not on blood pressure medication, has not had a stroke, has hypertension, has not been diagnosed with diabetes, has a Cholesterol of 220, as a systolic blood pressure of 140 and a diastolic blood pressure of 100, has a BMI of 30, has a heart rate of 60, and has a glucose level 80. What is the predicted probability that this patient will experience CHD in the next ten years? Should the physician prescribe the medication? **[5 pt]**

d) In the clinical setting from the previous question, the physician would also like to discuss options with the patient—including medication but also other possible interventions. Using your model, estimate the probability that the patient will experience CHD within the next 10 years if he makes adjustments to his lifestyle and health. Provide some numbers that the physician could leverage in this discussion. What would be the physician's key talking points? **[5 pt]**

e) What are some possible issues with using the logistic regression model to determine the effect of lifestyle and health changes? **[5 pt]**

f) Using the threshold determined in Question a., build your prediction on the test set. Assess the out-of-sample prediction of the model by computing its confusion matrix, its accuracy, its True Positive Rate, and its False Positive Rate. Finally, if patients are prescribed the medication using the strategy implied by the model, what will be the expected economic cost for all patients in the test set? **[10 pt]**

[Recall that, among patients who receive the medication, the incidence of CHD is lowered by 2.3.]

g) Consider a "baseline" model (hat reflects current practice) and an "ideal" model (under which medication is only prescribed to patients that would otherwise develop CHD, assuming perfect *ex post* information on the test set). For each of these models, compute the expected economic cost for all patients in the test set. Use these numbers to assess the results obtained in Question e. Report and explain your results in a short paragraph addressed to a non-technical audience. **[10 pt]**

h) Construct the ROC curve for your logistic regression model on the test set. Discuss how this curve may be helpful to decision-makers looking to further study the medication under consideration 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? **[10 pt]**

i) In order for doctors to actually use a logistic regression model, they will need to be convinced of its sensibility on intuitive grounds. And they will be more easily convinced if the model uses a limited number of risk factors—otherwise, it becomes more difficult to interpret all of the model's coefficients. Choosing only three risk factors, fit a logistic regression model on the training data, and evaluate its performance on the test data. How well does your simplified model perform? **[10 pt]**

[You do not need to enumerate all possible combinations, and you do not need to find the "best" model. Just pick a sensible set of three risk factors that you believe might be important.]

j) Are there any aspects of the analysis that raise ethical concerns? If so, what are some ways that this analysis could be changed to address these concerns? **[5 pt]**