1. A key performance for hospitals is the *30-day unplanned readmission rate*—the proportion of patients discharged from the hospital who had an unplanned readmission within 30 days. Programs like the Hospital Readmissions Reduction Program (HRRP) apply penalties (up to a 3% reduction in payments) to underperforming U.S. hospitals—resulting in withheld payments in excess of $500 million in 2018.

Hospitals can employ some low-cost strategies to reduce unplanned readmissions, such as confirming patient follow-up plans prior to discharge and asking patients to verbally repeat their treatment directions. However, other approaches are more involved and costly. One example is to arrange “telehealth” interventions, in which health care providers contact patients routinely after discharge. Given the cost of these interventions, they are only appropriate for patients at elevated risk of readmission.

You are working for a mid-sized hospital in the northeast United States and are tasked to assess the impact of telehealth interventions on diabetic patients—with the ultimate goal of reducing the 30-day read- mission rate. The intervention will cost approximately $1,200 per patient. Clearly, it must be limited in scope, and a key component of your strategy will be targeting the “right” patients.

Unfortunately, your hospital does not document 30-day readmissions, as this requires significant follow-up with discharged patients. You will thus use a publicly available dataset to study readmission risk. The dataset includes over 100,000 hospital discharges of over 70,000 diabetic patients from 130 hospitals across the United States during the period 1999–2008. All patients were hospital inpatients for 1–14 days, and received both lab tests and medications while in the hospital. The 130 hospitals represented in the dataset vary in size and location: 58 are in the northeast United States and 78 are mid-sized (100–499 beds).

The dataset is provided in the “readmission.csv” file. It contains the following variables:

· **readmission**: 1 if the patient had an unplanned readmission within 30 days, 0 otherwise

· **Patient characteristics**: race, gender, and age capture demographic information.

· **Recent medical system use**: The variables numberOutpatient, numberEmergency, and numberInpatient capture the number of times the patient used the medical system in the last year.

· **Diabetic treatments**: A number of variables capture the patient’s diabetic treatments: acarbose, chlorpropamide, glimepiride, glipizide, glyburide, glyburide.metformin, insulin, metformin, nateglinide, pioglitazone, repaglinide, and rosiglitazone.

· **Admission information**: The variables admissionType and admissionSource contain information about how the patient was admitted to the hospital. The variable numberDiagnoses captures the number of diagnoses the patient had recorded for their admission. There are also a number of variables that indicate whether a patient was diagnosed with various conditions when admitted: diagAcuteKidneyFailure, diagAnemia, diagAsthma, diagAthlerosclerosis, diagBronchitis, diagCardiacDysrhythmia, diagCardiomyopathy, diagCellulitis, diagCKD, diagCOPD, diagDyspnea, diagHeartFailure, diagHypertension, diagHypertensiveCKD,diagIschemicHeartDisease, diagMyocardialInfarction, diagOsteoarthritis, diagPneumonia, and diagSkinUlcer.

· **Treatment information**: timeInHospital is the number of days the patient was in the hospital, and numLabProcedures, numNonLabProcedures, and numMedications capture the amount of care the patient received in the hospital.

*a)* Open the data file “readmission.csv”in R. Perform some exploratory data analysis on the full data set and report two interesting insights you gained from your analysis.

*YOUR SOLUTION: …*

*b)* Based on conversations with the hospital’s management, you estimate the cost of a 30-day unplanned readmission at $35,000. From published information at a similar institution, you estimate that tele- health interventions will reduce the incidence of 30-day unplanned readmissions in the treated popuation by 25%. Given the cost of $1,200 per intervention, what are:

– the “loss” of a false positive, as compared to a true negative; and

– the “loss” of a false negative, as compared to a true positive?

Define the loss matrix for your CART model.

*YOUR SOLUTION: …*

*c)* Fit a CART model using a cp parameter of 0.001 and the loss matrix defined in Question b. Include an image of your tree.

*YOUR SOLUTION: …*

*d)* Assess the model’s predictive performance using the test set. What is the accuracy, true positive rate and false positive rate? Contrast the decisions resulting from your model and those resulting from current practice (under which no patient is subject to a telehealth intervention). Provide summary statistics to explain how the decisions differ, and discuss the costs and benefits of each approach. Make sure to compare the total monetary costs of patient readmission. [7 pts]

*YOUR SOLUTION: …*

*e)* Can cross validation improve your model or is a cp of .001 optimal? [3 pts]

*YOUR SOLUTION: …*

*e)* So far, you have selected the subset of patients that maximizes total net value, given the cost of the telehealth intervention and the willingness to pay for a prevented readmission. However, in practice you might have to adjust your model to try to directly improve the true positive rate (TPR) and/or the false positive rate (FPR).

*i)* How would you modify the loss matrix from part b) to obtain a CART model with a higher TPR than the one in part c)?

*YOUR SOLUTION: …*

*ii)* How would you modify the loss matrix from part b) to obtain a CART model with a lower FPR than the one in part c)?

*YOUR SOLUTION: …*

2. For historical reasons the US has a system of taxing homeowners to fund a large fraction of local infrastructure such as local primary, middle and high schools, town and county administrations, town and county roads, …. The tax, called “property tax”, is based on an assessment (estimation, determination) of the value of each residence (home) and the lot (land) that belongs to it. Because the assessments become outdated after a few years, towns have to hire assessors and update the assessments every so often.

There are of course several factors that play a role in assessing the value of a property: square feet of livable area, size of the lot (land), quality and condition of the building, desirability of the area, … We will only consider square feet of livable area. The following exercise can be used to check for any property whether its assessed value is in line or out of line with other properties of similar size in terms of livable area. This dataset is called “Residential\_Property\_Assessments.csv.”

*a)* Show a scatterplot of *Assessment* against *Livable Area* (i.e., *Assessment* is the y variable, and *Livable Area* is the x variable). Add a main title along with axis labels.

*YOUR PLOT:*

*b)* Based on the scatterplot, is the association approximately linear?

*YOUR ANSWER: …*

*c)* Use R to find the equation for the regression line with *Assessment* as the dependent variable and *Living Area* as the independent variable.

*YOUR SOLUTION: …*

*d)* Interpret the slope. Initially, do so first formally according to the formulation from class. Then, give an informal interpretation in terms of the average value of a square foot of livable area.

*YOUR SOLUTION:*

*Formally: …*

*Informally: …*

*e)* Interpret the intercept, first formally [1pt], then explain why this is not meaningful

*YOUR SOLUTION: …*

*Formally: …*

*This is not meaningful because …*

*f)* What fraction of variation in *Assessment* is accounted for by *Livable Area*? Report the relevant quantity from the R output [2pt]. Finally, what fraction of variation is accounted for by other factors besides livable area, such as differences in lot size, condition and quality of the building, viability of the area…?

*YOUR SOLUTION: …*

*g)* Based on the fitted equation, what can you say about the predicted price for a residence with 2,500 sqft of livable area? Use R to calculate this and show your code

*YOUR SOLUTION: …*

*h)* In general, how much uncertainty is left over after making our predictions using the regression line? Quote the relevant standard deviation from R. Compare this quantity to the variability left over if we had instead simply predicted the average of Assessment for all individuals, rather than using information about Livable Area to improve our prediction.

*YOUR SOLUTION: …*

3. We have data on n=78 individuals who were diagnosed with breast cancer, had the tumor surgically removed, and had no identifiable trace of the disease in their bodies after the surgery. Five years later, for each of these individuals we have information on whether or not their cancer came back over the course of the five years versus whether or not they stayed in remission. The data set in “breastcancer.csv” contains 7 clinical covariates, such age, and tumor diameter, along with 4,348 genetic indicators of prevalence of certain genes. The goal is to see whether or not reoccurrence of breast cancer after surgery is predictable based on both clinical and genetic information. If possible, clinicians would like to identify those at high risk for recurrence and recommend adjuvant therapy post-surgery, such as chemotherapy, as a means of reducing the risk of recurrence. The data set contains 34 individuals whose cancer did come back (labeled as response=1), while the remaining 44 did not (response = 0). In terms of data structure, this data set differs from those we’ve previously studied in that the number of covariates is much larger than the actual number of patients in the study. We will now see that this introduces major complications which need to be handled with care.

Before proceeding, execute the provided code to partition the data set into training set and test set. The training set is of size 50, and the test set has 28 individuals.

*a)* Our first instinct would likely be to fit a multiple logistic regression model using glm()on the training set. Try this out in R. You’ll notice that the output of the summary command looks bizarre. How many slope coefficients do NOT have the value NA? How does this compare the size of the training set? The function is.na()will be helpful here.

*YOUR SOLUTION: …*

*b)* Nonetheless, proceed using the predict() function in R to predict the responses in the training set based on our model. Use the command round(..., digits = 10) to round the estimated probabilities to 10 decimal places. Looking at your training set, what does the distribution for the predicted probabilities look like?

*YOUR SOLUTION: …*

*c)* Construct the in-sample confusion matrix based on this algorithm. What are the False Positive Rate and True Positive Rate in-sample? And what would the AUC be?

*YOUR SOLUTION: …*

*d)* A system of n equations with n unknown parameters will always have at least one solution, while a system of n equations with more than n unknown parameters will have infinitely many solutions. In light of this, why is it that we were able to attain in-sample error rates observed in part c)?

*YOUR SOLUTION: …*

Situations where the number of covariates are larger than the number of observation provide further motivation for using regularized regression. We clearly would like to include all of the potentially relevant covariate information, but mathematical deficiencies of the conventional solution render this difficult. The regularization provides automatic variable selection, while making the observed fit less susceptible to overfitting. In the code you see we have given you code for running the “LASSO” version of logistic regression while cross-validating the tuning parameter, which is facilitated by the cv.glmnet function.

*e)* How many nonzero slope coefficients are there in the resulting penalized logistic regression?

*YOUR SOLUTION: …*

*f)* Compute the in-sample predictions from the regularized logistic regression. Present a histogram of the predictions, and compare this to what we saw in part b).

*YOUR SOLUTION: …*

*g)* Compute the out-of-sample AUC for the two we’ve developed here: multiple logistic regression and the regularized logistic regression. What do you find? Which one does best out of sample? And are there any algorithms that seem to perform worse than random coin flips?

*YOUR SOLUTION: …*

4. For this question, you may choose from two unstructured options. Choose only one option and do not submit answers for both. You need to complete steps 2-6 of the basic data analytics process outlined below for the option you choose. This question will be graded holistically and is worth 25 points

**Basic Data Analytics Process**

~~1. Generate a question and collect data.~~

2. Visualize, prepare data, and understand your data.

a. If needed, clean and normalize the data. Also, consider removing redundant or known correlated variables.

b. Provide insights you gain from data exploration.

3. Choose a model or models we covered in this class.

a. Explain why you chose the model or models to answer the primary question.

b. Explain whether you value accurate predictions above all else or more interpretable and actionable model(s).

4. Train and tune the model(s) on the training set.

5. Assess the model(s) on the test set.

6. Provide recommendations and analysis of your results.

Option 1

After a recent trip to a winery with some of my friends, who happen to be chemists, we developed a method that would enable us to alter physicochemical attributes of red wine to create the world’s highest quality wine. We hypothesized that we could even alter bagged wine to be delicious. Unfortunately, we did not take the time to figure out exactly what qualities we should alter to what levels. Fortunately, we believe you have the skills necessary to do so.

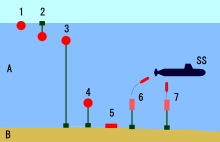
The file “redwine.csv” is related to red variants wine. The data only includes physicochemical (inputs) and sensory (the output) variables. There is no data about grape types, wine brand, wine selling price, etc.

Would you please help us understand what variables matter most and how we should adjust our wine to create the highest quality wine in all the land? [25pts if chosen]

*YOUR SOLUTION: …*

Option 2

As you are aware, I have spent some of my career beneath the waves on a submarine. Hazards come with the job, but what you may not know, is that that underwater mines are real and can be a threat to submarines. Rocks, on the other hand, are relatively harmless to pass over. (<https://en.wikipedia.org/wiki/Naval_mine>)



I would like to help my friends still serving on submarines by developing a model that could predict whether or not something is a mine or a rock from sonar data. They could then ignore the rocks and avoid the mines. Luckily, I found an amazing dataset on the internet to help me do just that.

The file "rockormine.csv" contains 111 patterns obtained by bouncing sonar signals off a “mine” at various angles and under various conditions and 97 patterns obtained from rocks under similar conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. The data set contains signals obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock.

Each pattern is a set of 60 numbers in the range 0.0 to 1.0. Each number represents the energy within a particular frequency band, integrated over a certain period of time. The integration aperture for higher frequencies occur later in time, since these frequencies are transmitted later during the chirp.

The label associated with each record contains the letter "R" if the object is a rock and "M" if it is a mine. The numbers in the labels are in increasing order of aspect angle, but they do not encode the angle directly.

Please help me keep my friends safe by developing a model that can best predict whether an object is a rock or a mine. While model accuracy is of utmost importance, do not penalize false positives or false negatives more than the other. Doing so could train your model to avoid every rock in the ocean, making it impossible for the submarine to even pull out of port.

*YOUR SOLUTION: …*