AI-DRIVEN ANALYTICS & SOCIETY
MIDTERM ASSIGNMENT, SPRING 2024
DUE DATE: APRIL 26, 2024

Midterm Overview

The midterm assignment consists of a hospital observation unit case requiring the use of responsible machine learning. For this case, you are the Medical Director of an observation unit (OU) at a hospital, who is interested in improving the current operations of the unit. Hospital observation units (OUs) are meant to host patients for relatively short periods of time (usually following an emergency department (ED) encounter), during which healthcare providers can observe and treat a patient while assessing the need for an inpatient hospital admission. Not all observation-status patients are placed in the OU – depending on the patient's circumstances and the availability of beds, a patient requiring observation-level care can be placed either in the OU or in an inpatient ward. If an observation-status patient is initially placed in the OU but is not discharged from the OU within 48 hours, a determination is made as to whether the patient's condition suggests a significantly longer expected stay and whether the patient meets medical necessity for inpatient medical services. In such a case, the OU issues a request for a bed in an inpatient ward, and the patient is transferred to the ward. The full process is shown in the figure below.

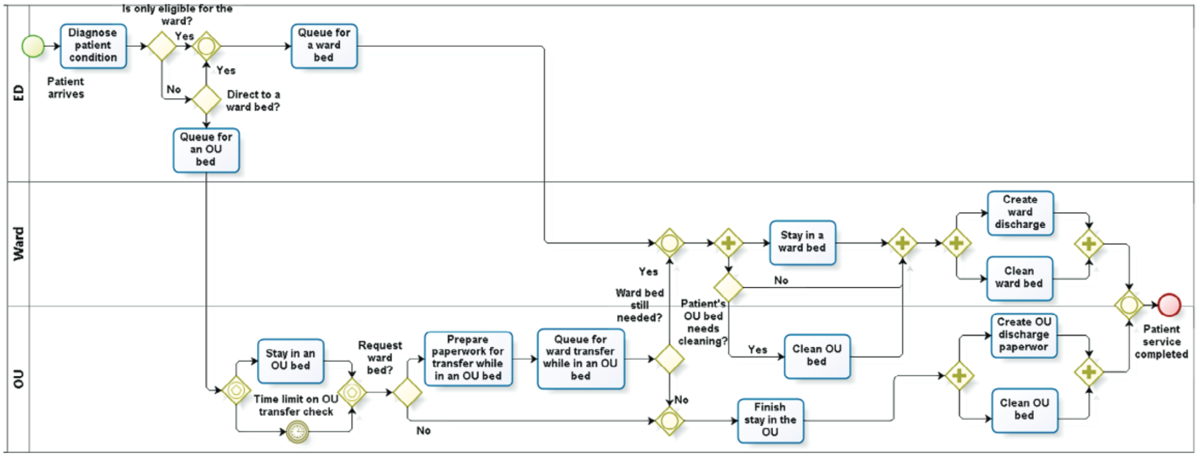
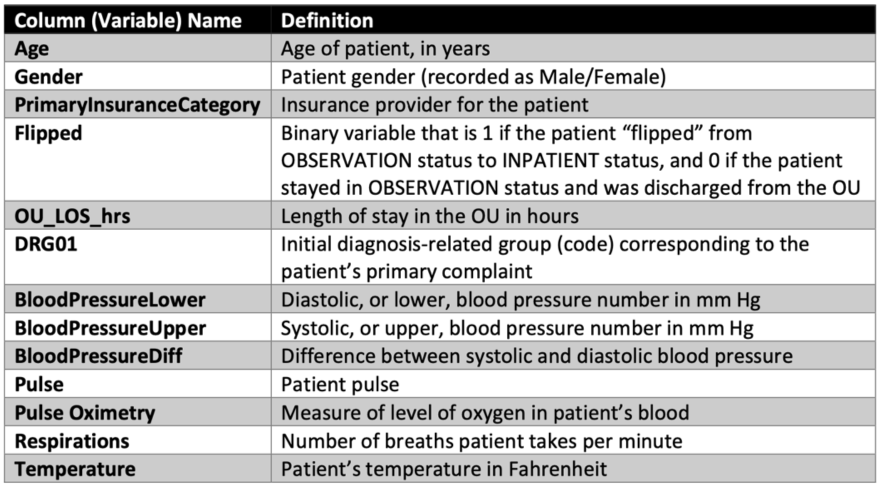

ED
Ward
OU
Patient arrives
Diagnose patient condition
Is only eligible for the ward?
Yes
No
Yes
Direct to a ward bed?
Queue for a ward bed
Queue for an OU bed
Stay in an OU bed
Time limit on OU transfer check
Request ward bed?
No
Prepare paperwork for transfer while in an OU bed
Queue for ward transfer while in an OU bed
Ward bed still needed?
Yes
No
Finish stay in the OU
Stay in a ward bed
Patient's OU bed needs cleaning?
No
Yes
Clean OU bed
Create ward discharge
Clean ward bed
Create OU discharge paperwor
Clean OU bed
Patient service completed

Figure 1. Patient Process Flow through the Observation Unit and Hospital Wards

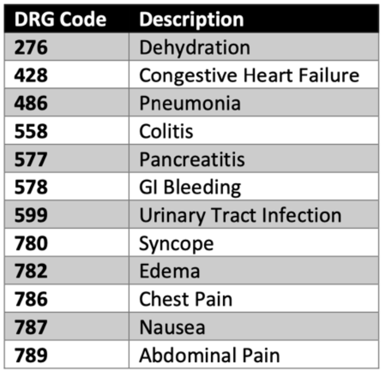
The major challenge in the case is how to identify the appropriate patients to place in the OU. Case misclassification can be a substantial problem for OU operations, leading to increased average length of stay and decreased hospital capacity. A data set (OUData.csv) sampled from randomly from OU arrival data at a hospital (extracted from the hospital's Electronic Health Record (HER) system) is provided with the case. Each row is a medicine service patient visit with a unique identifier. Each column contains information about the visit and the recorded vital signs. The following is a data dictionary for each of the columns:



Column (Variable) Name	Definition
Age	Age of patient, in years
Gender	Patient gender (recorded as Male/Female)
PrimaryInsuranceCategory	Insurance provider for the patient
Flipped	Binary variable that is 1 if the patient "flipped" from OBSERVATION status to INPATIENT status, and 0 if the patient stayed in OBSERVATION status and was discharged from the OU
OU_LOS_hrs	Length of stay in the OU in hours
DRG01	Initial diagnosis-related group (code) corresponding to the patient's primary complaint
BloodPressureLower	Diastolic, or lower, blood pressure number in mm Hg
BloodPressureUpper	Systolic, or upper, blood pressure number in mm Hg
BloodPressureDiff	Difference between systolic and diastolic blood pressure
Pulse	Patient pulse
Pulse Oximetry	Measure of level of oxygen in patient's blood
Respirations	Number of breaths patient takes per minute
Temperature	Patient's temperature in Fahrenheit

Table 2. Data Dictionary for the OUData.csv

Similarly, the following table contains a listing of initial primary diagnosis-related group (DRG) codes and descriptions.



DRG Code	Description
276	Dehydration
428	Congestive Heart Failure
486	Pneumonia
558	Colitis
577	Pancreatitis
578	GI Bleeding
599	Urinary Tract Infection
780	Syncope
782	Edema
786	Chest Pain
787	Nausea
789	Abdominal Pain

Figure 3. Initial Primary DRG Codes and Descriptions

This assignment consists of three parts:

1. <u>Cleaning the data and data visualization</u>: Like many real-world data sets, some of the data are missy and/or dirty; you will be first tasked with cleaning data. Then, to get comfortable with the data set, you will be tasked with performing descriptive analytics.
2. <u>Predictive model building and assessment</u>: You will build predictive models for classifying patients as either "appropriate" or "inappropriate" for placement in the OU and evaluate the relative performance of your various models (however you choose to define that).
3. <u>Recommendations, ethics, and responsible ML</u>: You will reflect on how to operationalize your predictive model as a concrete recommendation to the hospital, the ethical implications of your prescription, and how you might design "more responsible" machine learning models.

You are allowed to work alone or in teams of 2. You should submit a PDF report, the R code, and all outputs (cleaned CSVs, figures from the visualizations, etc.) to Canvas.

Part I: Data Cleaning and Visualization (Descriptive Analytics: 5 points)

Open the R file CleanOUData.R available in Canvas. First, we will clean the missing data and fix any incorrect data types to produce OUDataClean.csv.

a. In words, explain what the code in subparts (i) and (ii) does.

b. Complete subparts (iii)-(xi) and output a cleaned CSV data file.

Open the R file OUDataVisualization.R available in Canvas. This contains realistic data for medicine service patients admitted to the observation unit (OU) over a period of time, including their age, gender, preliminary diagnosis-related group (DRG) at the time of admission to the OU, and whether they had to be sent to the inpatient wards eventually. Your task in this exercise will be to use various data summaries and visualizations to determine simple, intuitive rules for placing certain types of patients to the OU versus sending them directly to the wards. The main criteria we are going to use to determine whether a patient should be in the OU are:

1. The probability that the patient will "flip," i.e., that the patient's status will change from OBSERVATION to INPATIENT and the patient will need to be sent to the wards anyway.

2. The expected Length of Stay (LOS) of the patient in the OU.

Complete the following descriptive analytics tasks:

c. In words, explain what the code in subparts (xii)-(xiv) does.

d. In subparts (xv)-(xix), use appropriate visualizations to explore how a patient's Age, Gender, and primary complaint (DRG01) affect the likelihood that the patient will flip and his/her length of stay.

e. In words, explain what the code in subparts (xx)-(xxii) does and how it is different from the code in subparts (xii)-(xiv).

f. Consider combinations of variables in subparts (xxiii)-(xxv): Age-Gender, DRG-Gender, and Gender-PrimaryInsuranceCategory. What observations can you make about the profiles of patients that are more likely to flip and more likely to stay long in the OU?

g. In subpart (xv), let's focus only on two DRGs (780 and 782) and only two insurance types ("MEDICARE" and "MEDICARE OTHER"). Pooling all of this data, which types of primary complaints (780 or 782) are more likely to flip?

h. In subpart (xvi), cross-tabulate the number of patients from each DRG and insurance type that flipped. Based on the table, patients with which of the two DRGs (780 or 782) are a better fit for the OU, conditioning on each insurance type. Compare with your answer in part (g).

Part II: Building Predictive Models (Predictive Analytics: 5 points)

Open the file OUModelBuild.R available in Canvas. Your task will be to use machine learning algorithms to determine simple, intuitive rules for placing certain types of patients to OU versus sending them directly to the wards. The main criterion we are going to use to determine whether a patient should be placed in the OU is whether or not the patient will flip. Patients expected to flip should be sent directly to the wards, so as not to create congestion in the OU. The target variable in the data set is Flipped (a 0/1 binary variable), with 1 indicating the patient flipped.

Consider the following predictors to predict Flipped: Age (numerical), Gender (categorical), PrimaryInsuranceCategory (categorical), DRG01 (categorical), BloodPressureDiff (numerical), BloodPressureLower (numerical), BloodPressureUpper (numerical), Pulse (numerical), PulseOximetry (numerical), Respirations (numerical), and Temperature (numerical).

i. In words, explain what the code in subpart (xxvii) does and why this step is important in the evaluation of predictive models.

j. Run a logistic regression on the target variable using the predictors above with the R function `glm`.

k. Using `rpart`, create a full classification tree using the predictors indicated above. Prune the tree using `prune`. Plot the classification tree and interpret your solution using `rpart.plot`.

l. The confusion matrix is a 2x2 grid showing the true flip along one axis and the model's predicted flip along the other. Generate the confusion matrix on the test data to comment on the performance of your model, including a relative comparison to your model in part (j). You might find the following R functions useful in doing so:

   - `predict`
   - `table`
   - `addmargins`
   - `prop.table`

Part III: Ethics and Responsible Machine Learning (Prescriptive Analytics: 5 points)

Consider the predictive models you generated in Part II; you are now tasked with generating a prescriptive model for the hospital about whether to send a patient to OU or directly to the wards.

m. Suppose that you are the Medical Director of the Observation Unit and assume that the predictive model you created in part (k) was "good enough" in terms of predictive performance to use, i.e., that it predicts patients who will flip accurately for all practical purposes. Would you have any concerns about applying this model to decide whether to assign patients to the OU instead of sending them to the wards?

n. Open-ended: As a modeler, how would you change your process/model to alleviate possible concerns from part (m)?