15.072: The Analytics Edge **Fall 2021**
Short Homework 3: CART

Out: October 04; Due: October 19, 11:59 pm.

Preventing Hospital Readmissions

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

One intervention to reduce readmission is to arrange "telehealth" interventions, in which health care providers contact patients routinely after discharge. You are working for a mid-sized hospital in the northeast United States, and are tasked to analyze the feasibility of such telehealth interventions with recently-discharged diabetic patients—with the ultimate goal of reducing the 30-day readmission 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's IT system does not document 30-day readmissions, as this requires significant follow-up with discharged patients. Therefore, you will use a publicly-available dataset to study the risk of readmission among diabetic patients. 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.[1] In the dataset, 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 of discharge, 0 otherwise.
- **Patient characteristics**: `race`, `gender`, and `age` capture demographic information about the patients.
- **Recent medical system use**: `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.

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. doi:10.1155/2014/781670

Run the following command to read in the data and to split the dataset into a training set (75%) and a test set (25%) using stratified sampling:

```
readmission = read.csv("readmission.csv")
set.seed(144)
split = createDataPartition(readmission$readmission, p = 0.75, list = FALSE)
readm.train <- readmission[split,]
readm.test <- readmission[-split,]
```

a. The hospital's management estimates the cost of a 30-day unplanned readmission at $35,000. Public reports indicate that telehealth interventions reduce the incidence of 30-day unplanned readmissions in the treated population by around 25%. Define the loss matrix for your CART model. **[15 pts]**

   (i) Estimate the costs of true negatives, false positives, false negatives, and true positives, and define the loss matrix for your CART model accordingly.

   Hint: Make sure that your loss matrix has zeros along the diagonal.

b. Construct a CART model on the training set with three values of the `cp` parameter (including the default value of 0.01). **[20 pts]**

   (i) Include an image of your three trees.

   (ii) What types of patients receive the telehealth intervention? Does that fit with your intuition?

c. Evaluate the predictive performance of your three trees from Question b. **[20 pts]**

   (i) For each tree, report the number of patients subject to telehealth interventions on the test set.

   (ii) For each tree, report the expected number of prevented readmissions on the test set.

   (iii) For each tree, report the accuracy on the test set.

   (iv) For each tree, report the true positive rate on the test set.

   (v) For each tree, report the false positive rate on the test set.

   (vi) For each tree, report the total cost on the test set.

d. Evaluate the practical performance of your three trees from Question b. **[15 pts]**

   (i) Report the cost savings (in absolute and relative terms) of each tree from Question b. on the test set, against current practice with no telehealth interventions.

e. Due to budget restrictions, the hospital can only apply telehealth interventions to at most 5% of its diabetic population. Modify your model accordingly, and assess the resulting decision rule. **[30 pts]**

   (i) How can you modify the parameters in your CART model to reduce the proportion of patients subject to the telehealth intervention?

   (ii) Fit a few different CART models (using a `cp` value of 0.001), and choose a tree that will meet the budget constraint—at least approximately. What parameters have you chosen?

   (iii) Evaluate your new tree on the test set. How many patients receive telehealth interventions? How many readmissions do you expect to prevent? What is the net value of these interventions?

   (iv) Estimate the cost savings (in absolute and relative terms) of the budget-constrained tree, against current practice and against the tree obtaind with no budget constraint.

   (v) What are the implications for the hospital's management?