Homework #3

Preventing Hospital Readmissions

A key performance metric for hospitals is the *30-day unplanned readmission rate*, which is the proportion of patients who are discharged from the hospital but who had an unplanned readmission to that hospital or another one within 30 days. This metric is used as a performance indicator for hospitals, and programs like the Hospital Readmissions Reduction Program (HRRP) apply financial penalties (up to a 3% reduction in payments) to American hospitals that underperform on this metric in certain key patient populations. These penalties are significant motivators for hospitals, and are predicted to cost hospitals more than $500 million in withheld payments during 2018.

Hospitals have many approaches at their disposal to reduce unplanned readmissions. Some of these strategies, like confirming patient follow-up plans prior to discharge or asking patients to verbally repeat their treatment directions, are low cost and can be applied to all discharged patients. However, other approaches are more involved and costly, such as providing patients with a 30-day medication supply upon discharge or "telehealth" interventions in which hospital staff or contractors contact patients routinely after discharge via telephone or video call to receive updates on their status. Given the cost of these more intensive interventions, they are typically only appropriate for patients at elevated risk of readmission.

You are employed by a mid-sized hospital in the northeast United States and are tasked with analyzing the feasibility of a telehealth intervention involving telephone and video calls with diabetic patients who have been recently discharged from the hospital. The goal of this intervention is to reduce the 30-day readmission rate among diabetic patients discharged from your hospital. Your hospital is in the early planning stages for this intervention, and there has not been a pilot study of the telehealth intervention run at your institution. However, it is clear that the intervention must be limited to the patient subset at high risk of readmission due to the intervention's high cost, as you estimate the intervention will cost approximately $1,000/patient. As a result, a key component of your telehealth strategy will be targeting diabetic patient discharges with a high risk of 30-day unplanned readmission.

Unfortunately, your hospital's IT system does not accurately document 30-day readmissions, as this requires significant follow-up with discharged patients to determine if they were readmitted at another facility. As a result, you have decided to study risk of readmission among diabetic patients using a publicly available dataset of more than 100,000 hospital discharges of more than 70,000 diabetic patients from 130 hospitals across the United States from 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

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., "Impact of HbA1c Measurement on Hospital Readmission

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 of diabetic patient discharges is provided in the file **`readmission.csv`**. The file contains the following variables:

- **Outcome**: The variable `readmission` indicates if the patient had an unplanned readmission within 30 days of their discharge.
- **Patient characteristics**: The variables `race`, `gender`, and `age` capture basic demographic information about the patients.
- **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 specific 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.

NOTE: For the questions in this problem, provide your numerical answer as well as an explanation of how you arrived at your answer.

a) (5 points) Open the data file **`readmission.csv`** in R. Perform some exploratory data analysis and report two interesting insights you gained from your analysis.

b) To *construct the* CART *model, first split the readmission dataset into a training set and a test* set, putting 75% of the data in the training set. This can be done by running the following sequence of commands:

```
> set.seed(998)
> library(caret) # library to split training and testing
> split = createDataPartition(readmission$readmission, p = 0.75, list =
FALSE)
> readm.train = readmission[split,]
> readm.test = readmission[-split,]
```

Using the training set, fit a CART model using the `cp` parameter 0.002. When fitting the model, be sure to use the loss matrix you developed in part *c*).

i) (10 points) Construct an image of your fitted tree. What types of patients are selected for the telehealth intervention? Which variables are used to make the selections? How well does this result fit (or not fit) with your intuition regarding the risk of readmission? Include an image of your tree in your explanation.

ii) (4 points) Use the model to make predictions on the test set. How many patients in the test set would be selected for the telehealth intervention, and what would be the associated costs?

iii) (4 points) Assuming the telehealth intervention prevented readmission in 25% of the true positive cases (cases where the model selected the telehealth intervention and the patients would have been readmitted without the intervention), how many readmissions is the telehealth intervention estimated to prevent?