Predicting readmission probability for diabetes inpatients

STAT 471/571/701 Modern Data Mining

Due midnight, April 7th, 2019

Contents

Instructions

- This project is due at **11:59pm on Sunday, April 7th, 2019.** No **late** submissions are allowed.
- It is an individual project, amounting to **20%** of your final grade. See the *Collaboration* section at the bottom of this document.
- There is no single correct answer. You will be graded on the general quality of your work.
- The entire write up should not be more than **8** pages. All the R-codes should be hidden. You may put any supporting documents, graphics, or other exhibits into an Appendix, which is not counted in the 8 page limit.

Introduction

Background

Diabetes is a chronic medical condition affecting millions of Americans, but if managed well, with good diet, exercise and medication, patients can lead relatively normal lives. However, if improperly managed, diabetes can lead to patients being continuously admitted and readmitted to hospitals. Readmissions are especially serious - they represent a failure of the health system to provide adequate support to the patient and are extremely costly to the system. As a result, the Centers for Medicare and Medicaid Services announced in 2012 that they would no longer reimburse hospitals for services rendered if a patient was readmitted with complications within 30 days of discharge.

Given these policy changes, being able to identify and predict those patients most at risk for costly readmissions has become a pressing priority for hospital administrators.

Goal of the study

In this project, we shall explore how to use the techniques we have learned in order to help better manage diabetes patients who have been admitted to a hospital. Our goal is to avoid patients being readmitted within 30 days of discharge, which reduces costs for the hospital and improves outcomes for patients. If we could identify important factors relating to the chance of a patient being readmitted within 30 days of discharge, effective intervention could be done to reduce the chance of being readmitted. Also if we could predict one's chance being readmitted well, actions can be taken.

The data

The original data is from the Center for Clinical and Translational Research at Virginia Commonwealth University. It covers data on diabetes patients across 130 U.S. hospitals from 1999 to 2008. There are over 100,000 unique hospital admissions in this dataset, from ~70,000 unique patients. The data includes demographic elements, such as age, gender, and race, as well as clinical attributes such as tests conducted, emergency/inpatient visits, etc. Refer to the original documentation for more details on the dataset. Three former students Spencer Luster, Matthew Lesser and Mridul Ganesh, brought this data set into the class and did a wonderful final project. We will use a subset processed by the group but with a somewhat different objective.

Characteristics of the Data Set

All observations have five things in common:

1. They are all hospital admissions
2. Each patient had some form of diabetes
3. The patient stayed for between 1 and 14 days.
4. The patient had laboratory tests performed on him/her.
5. The patient was given some form of medication during the visit.

The data was collected during a ten-year period from 1999 to 2008. There are over 100,000 unique hospital admissions in the data set, with ~70,000 unique patients.

Description of variables

The dataset used covers ~50 different variables to describe every hospital diabetes admission. In this section we give an overview and brief description of the variables in this dataset.

1) Patient identifiers:

a. `encounter_id`: unique identifier for each admission
b. `patient_nbr`: unique identifier for each patient

2) Patient Demographics:

`race`, `age`, `gender`, `weight` cover the basic demographic information associated with each patient. `Payer_code` is an additional variable that identifies which health insurance (Medicare /Medicaid / Commercial) the patient holds.

3) Admission and discharge details:

a. `admission_source_id` and `admission_type_id` identify who referred the patient to the hospital (e.g. physician vs. emergency dept.) and what type of admission this was (Emergency vs. Elective vs. Urgent).
b. `discharge_disposition_id` indicates where the patient was discharged to after treatment.

4) Patient Medical History:

a. `num_outpatient`: number of outpatient visits by the patient in the year prior to the current encounter
b. `num_inpatient`: number of inpatient visits by the patient in the year prior to the current encounter
c. `num_emergency`: number of emergency visits by the patient in the year prior to the current encounter

5) Patient admission details:

a. `medical_specialty`: the specialty of the physician admitting the patient
b. `diag_1`, `diag_2`, `diag_3`: ICD9 codes for the primary, secondary and tertiary diagnoses of the patient. ICD9 are the universal codes that all physicians use to record diagnoses. There are various easy to use tools to lookup what individual codes mean (Wikipedia is pretty decent on its own)
c. `time_in_hospital`: the patient's length of stay in the hospital (in days)
d. `number_diagnoses`: Total no. of diagnosis entered for the patient
e. `num_lab_procedures`: No. of lab procedures performed in the current encounter
f. `num_procedures`: No. of non-lab procedures performed in the current encounter
g. `num_medications`: No. of distinct medications prescribed in the current encounter

6) Clinical Results:

a. `max_glu_serum`: indicates results of the glucose serum test
b. `A1Cresult`: indicates results of the A1c test

7) Medication Details:

a. `diabetesMed`: indicates if any diabetes medication was prescribed
b. `change`: indicates if there was a change in diabetes medication
c. `24 medication variables`: indicate whether the dosage of the medicines was changed in any manner during the encounter

8) Readmission indicator:

Indicates whether a patient was readmitted after a particular admission. There are 3 levels for this variable: "NO" = no readmission, "< 30" = readmission within 30 days and "> 30" = readmission after more than 30 days. The 30 day distinction is of practical importance to hospitals because federal regulations penalize hospitals for an excessive proportion of such readmissions.

To save your time we are going to use some data sets cleaned by the group. Thus, we provide two datasets:

`diabetic.data.csv` is the original data. You may use it for the purpose of summary if you wish. You will see that the original data can't be used directly for your analysis, yet.

`readmission.csv` is a cleaned version and they are modified in the following ways:

1) `Payer code`, `weight` and `Medical Specialty` are not included since they have a large number of missing values.

2) Variables such as `acetohexamide`, `glimepiride.pioglitazone`, `metformin.rosiglitazone`, `metformin.pioglitazone` have little variability, and are as such excluded. This also includes the following variables: `chlorpropamide`, `acetohexamide`, `tolbutamide`, `acarbose`, `miglitor`, `troglitazone`, `tolazamide`, `examide`, `citoglipton`, `glyburide.metformin`, `glipizide.metformin`, and `glimepiride.pioglitazone`.

3) Some categorical variables have been regrouped. For example, `Diag1_mod` keeps some original levels with large number of patients and aggregates other patients as `others`. This process is known as 'binning.'

4) The event of interest is **readmitted within < 30 days**. Note that you need to create this response first by regrouping **Readmission indicator**!

Research approach

From the *Goals* section above, your study should respond to the following:

Analyses suggested

1) Identify important factors that capture the chance of a readmission within 30 days.

The set of available predictors is not limited to the raw variables in the data set. You may engineer any factors using the data, that you think will improve your model's quality.

2) For the purpose of classification, propose a model that can be used to predict whether a patient will be a readmit within 30 days. Justify your choice. Hint: use a decision criterion, such as AUC, to choose among a few candidate models.

3) Based on a quick and somewhat arbitrary guess, we estimate **it costs twice as much** to mislabel a readmission than it does to mislabel a non-readmission. Based on this risk ratio, propose a specific classification rule to minimize the cost. If you find any information that could provide a better cost estimate, please justify it in your write-up and use the better estimate in your answer.

Suggestion: You may use any of the methods covered so far in parts 1) and 2), and they need not be the same. Also keep in mind that a training/testing data split may be necessary.

The write up

As you all know, it is very important to present your findings well. To achieve the best possible results you need to understand your audience.

Your target audience is a manager within the hospital organization. They hold an MBA, are familiar with medical terminology (though you do not need any previous medical knowledge), and have gone through a similar course to our Modern Data Mining with someone like your professor. You can assume thus some level of technical familiarity, but should not let the paper be bogged down with code or other difficult to understand output.

Note then that the most important elements of your report are the clarity of your analysis and the quality of your proposals.

A suggested outline of the report would include the following components:

1) Executive Summary

- This section should be accessible by people with very little statistical background (avoid using technical words and no direct R output is allowed)
- Give a background of the study. You may check the original website or other sources to fill in some details, such as to why the questions we address here are important.
- A quick summary about the data.
- Methods used and the main findings.
- You may use clearly labelled and explained visualizations.
- Issues, concerns, limitations of the conclusions. This is an especially important section to be honest in - we might be Penn students, but we are statisticians today.

2) Detailed process of the analysis

i) Data Summary /EDA

- Nature of the data, origin
- Necessary quantitative and graphical summaries

- Are there any problems with the data?
- Which variables are considered as input

ii) Analyses

- Various appropriate statistical methods: e.g. glmnet and/or trees
- Comparisons various models
- Final model(s)

iii) Conclusion

- Summarize results and the final model
- Final recommendations

Maintain a good descriptive flow in the text of your report. Use Appendices to display lengthy output.

iii) Appendix

- Any thing necessary to keep but for which you don't want them to be in the main report.

Collaboration

This is an **individual** assignment. We will only allow private Piazza posts for questions. If there are questions that are generally useful, we will release that information.