Predicting Hospital Readmission of Patients with Diabetes

Hongyun (Kevin) Wang

Problem

What are the risk factors for readmission of patients with diabetes and how to mitigate the risk? Hospital readmission rate is a good measure for health care quality and cost. Patients with diabetes currently represent about 9% of the US population, but they account for approximately 25% of hospitalizations (over eight million per year). The 30-day readmission rate of diabetic patients is 14.4-22.7%, which is higher than that of all hospitalized patients (8.5-13.5%). The burden of diabetes among hospitalized patients is substantial, growing and costly and readmissions contribute a significant portion of this burden. Therefore, reducing 30-day readmissions of patients with diabetes has the potential to greatly reduce healthcare costs while simultaneously improving care.

Clients

Medical insurance companies, hospitals and patients with diabetes are clients of this project. Medical insurance companies can use the results from this project to reduce cost and improve health care quality. Hospitals can evaluate risk factors and make right decisions to discharge patients at appropriate time to achieve lower readmission rate. This could result in better hospital resources allocation for other patients. For patients with diabetes, reduced times of hospitalization will help them reduce frustration due to multiple hospitalizations and help them recover sooner.

Data

The data used for this project will be acquired from UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008
The data are from two csv files. One csv file with ID mapping data includes 3 features' numeric IDs and corresponding text descriptions. Another csv file with diabetic data has 101,766 rows and 50 columns. Each row is a unique encounter (case) and different encounters could be from the same patient. The 50 columns include personal information, admission type, time in hospital, laboratory test results, medications and readmitted. The readmitted column has three outcomes/classes: no readmission, readmission within 30 days and readmission beyond 30 days. The readmitted column will be the target label.

Data features:

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%

Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%

Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.	0%

Approach (subject to changes and updates)

First, I will prepare the data for analysis by cleaning and wrangling the data. This includes mapping the numeric IDs to text descriptions for some features, encoding missing values (?) with NA, removing encounters with missing values, one-hot encoding for categorical features.

Second, to explore the data, I will use mainly using data visualization and description statistics methods. This includes plotting the counts for each feature and the target label to find highly associated features with the target label.

After identifying highly associated features, I will split the data with train and test set and start to build models using machine learning algorithms including logistic regression, SVM, random forest and boosting. Different models will be evaluated using accuracy, recall, precision and ROC curve.

Deliverables

My deliverables will be a final report and slide deck published to my GitHub account. The source codes on data acquisition, data cleaning and wrangling, exploratory data analysis and machine learning build will be published to this account too.