Milestone Report of Capstone Project 1

Hongyun Wang

1/08/2020

Contents

1 Problem statement

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.

After a paper published in 2014 on HbA1c correlation with diabetic patients readmission, the result generated by the paper and the data and statistics used are of great influence and the test is then used widely. In this project, rather than verifying the result generated by the paper, I would like to examine diabetes patients readmission rate using a different approach, classificaion.

2 Description of the data set

2.1 Background

The dataset used for this project was acquired from UCI Machine Learning Repository at (https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008)

The dataset used in this study is from the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. Health Facts is a voluntary program offered to organizations which use the Cerner Electronic Health Record System. The database contains data systematically collected from participating institutions electronic medical records and includes encounter data (emergency, outpatient, and inpatient), provider specialty, demographics (age, sex, and race), diagnoses and in-hospital procedures documented by ICD-9-CM codes, laboratory data, pharmacy data, in-hospital mortality, and hospital characteristics. All data were identified

in compliance with the Health Insurance Portability and Accountability Act of 1996 before being provided to the investigators. Continuity of patient encounters within the same health system (EHR system) is preserved.

The Health Facts data the study used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States: Midwest (18 hospitals), Northeast (58), South (28), and West (16). Most of the hospitals (78) have bed size between 100 and 499, 38 hospitals have bed size less than 100, and bed size of 14 hospitals is greater than 500.

The database consists of 41 tables in a fact-dimension schema and a total of 117 features. The database includes 74,036,643 unique encounters (visits) that correspond to 17,880,231 unique patients and 2,889,571 providers. Because this data represents integrated delivery network health systems in addition to stand-alone hospitals, the data contains both inpatient and outpatient data, including emergency department, for the same group of patients. However, data from out-of-network providers is not captured.

2.2 Data wrangling and cleaning

2.2.1 remove encounters with duplicated patient number and encounters with discharge to a hospice or patient death

The preliminary dataset contained multiple inpatient visits for some patients and the observations could not be considered as statistically independent, an assumption of the logistic regression model. We thus used only one encounter per patient; in particular, we considered only the first encounter for each patient as the primary admission and determined whether or not they were readmitted within 30 days. Additionally, we removed all encounters that resulted in either discharge to a hospice or patient death, to avoid biasing our analysis. After filtering on the encounters, 71,050 rows were left comparing to the initial rows of 101,766.

2.2.2 Re-label the response variable

The current response variable 'readmitted' has 3 categories: <30, >30 and NO. This study is interested in those patients that are readmitted within 30 days after discharge. Then, the <30 category was coded as 1, other two categories were coded as 0.

2.2.3 Check missing values

Check the percentage of missing values for each variable and observation. Three variables have >39% missing values: `weight`, `medical_specialty` and `payer_code`. Other variables have <3% missing values. The variable weight with >50% missing values were dropped. Before dropping variables, the highest percentage of missing values in observations is 10%. This number decreases to 8% after dropping variables. Currently keep the observations with missing values. Three variables, `weight`, `payer_code` and `medical_specialty` were dropped.

2.2.4 Check outliers (numerical variables)

8 numerical varaibles were checked for outliers. There are no missing values but there are outliers in those 8 variables according to the boxplot of each numerical variable. Currently keep all those outliers in each variable.

2.2.5 Examine categorical variables

Categorical variables have following types:
- `race`, `gender`, `age`. Those are basic demographic information.
- `admission_type_id`, `discharge_disposition_id`, `admission_source_id` are numerical but they are IDs and should be treated as categorical.
- `diag_1`, `diag_2`, `diag_3` have several hundred distinct values. Those could be dropped for analysis. Those diagnose information are partly captured in numerical variable number_diagnoses.
- `max_glu_serum`, `A1Cresult`. Special lab test results.

- 23 generic medications and a special one `diabetesMed`. `change` is a binary variable indicating whether there was a change in diabetic medications (either dosage or generic name).

After checking the levels of each categorical variable, variables with only 1 level were droped. They are `citoglipton`, `glimepiride-pioglitazone`, `examide`. Other variables with >50 levels were dropped too. They are `diag_1`, `diag_2`, `diag_3`.

2.2.6 Summary of cleaned dataset

The cleaned data set has 71,050 encounters and 41 variables:
- 8 numerical variables
- 30 categorical variables
- 1 encounter id variable
- 1 patient number variable
- 1 response variable

3 Initial findings from exploratory data analysis

3.1 summary of findings

In primary study (no categorical level combination), 7 categorical variables have significant association with Readmission (p-value < 0.01). They are `repaglinide`, `age`, `diabetesMed`, `change`, `insulin`, `discharge_disposition_id` and`glipizide`. Except `age` and `discharge_disposition_id`, other 5 variables are all medication related, including medication types and medication changes.

`A1cresult`is not signifcant at primary study. After combining >7 category to norm category for `A1Cresult`, the association test generate a p-value of 0.0499. When adding one more variable `change` and create 3-way table, the p-value of association bwteeen >8 `A1Cresult/change` of medications and readmission is 0.0999. But the p-value of association bwteeen normal `A1Cresult/change` of medications and readmission is 0.762.

Except two variables `num_procedures` and `number_outpatient` have p-values as 0.6626 and 0.0234, other 6 variables have p-value close to 0. The differences of mean of those 6 numerical variables between two Readmission categories are significant.

For 8 numerical variables, there are 28 paris of correlation coefficients. Among those 28 pairs of correlation coefficients, 3 pairs of numerical variables have corrrelation coeffieients greater than 0.3. They are `num_medications` and `time_in_hospital` (r=0.47), `num_procedres` and `num_medications` (r=0.4), and `num_lab_precedures` and `time_in_hospital` (r=0.33).

3.2 visuals and statistics to support findings

Scatterplot of variables `num_procedures`, `number_diagnoses`, `num_medications`,`num_lab_procedures`. Those four numerical variables are about how many procedures/diagnoses/medications have been given and they are in similar scales.

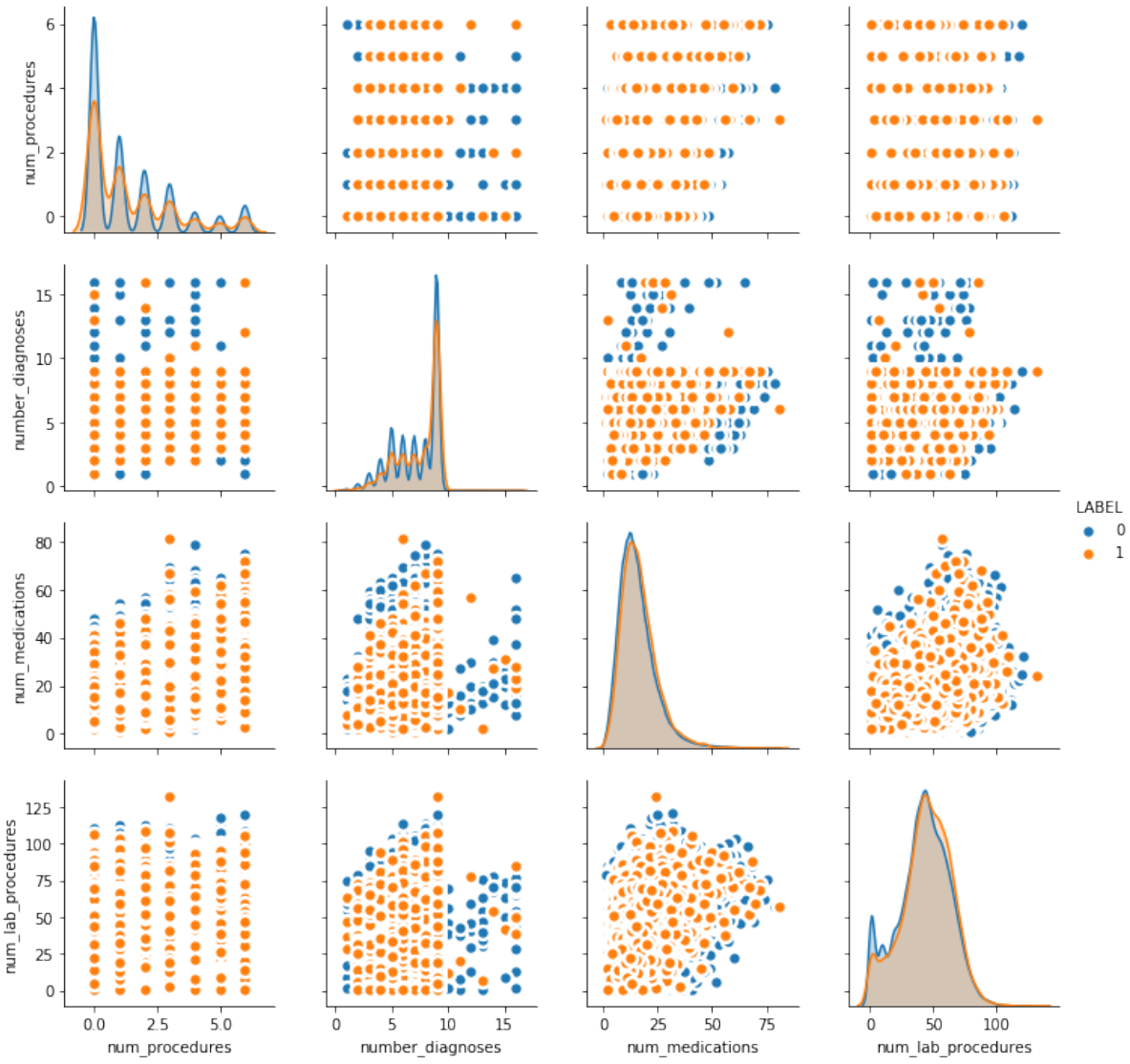
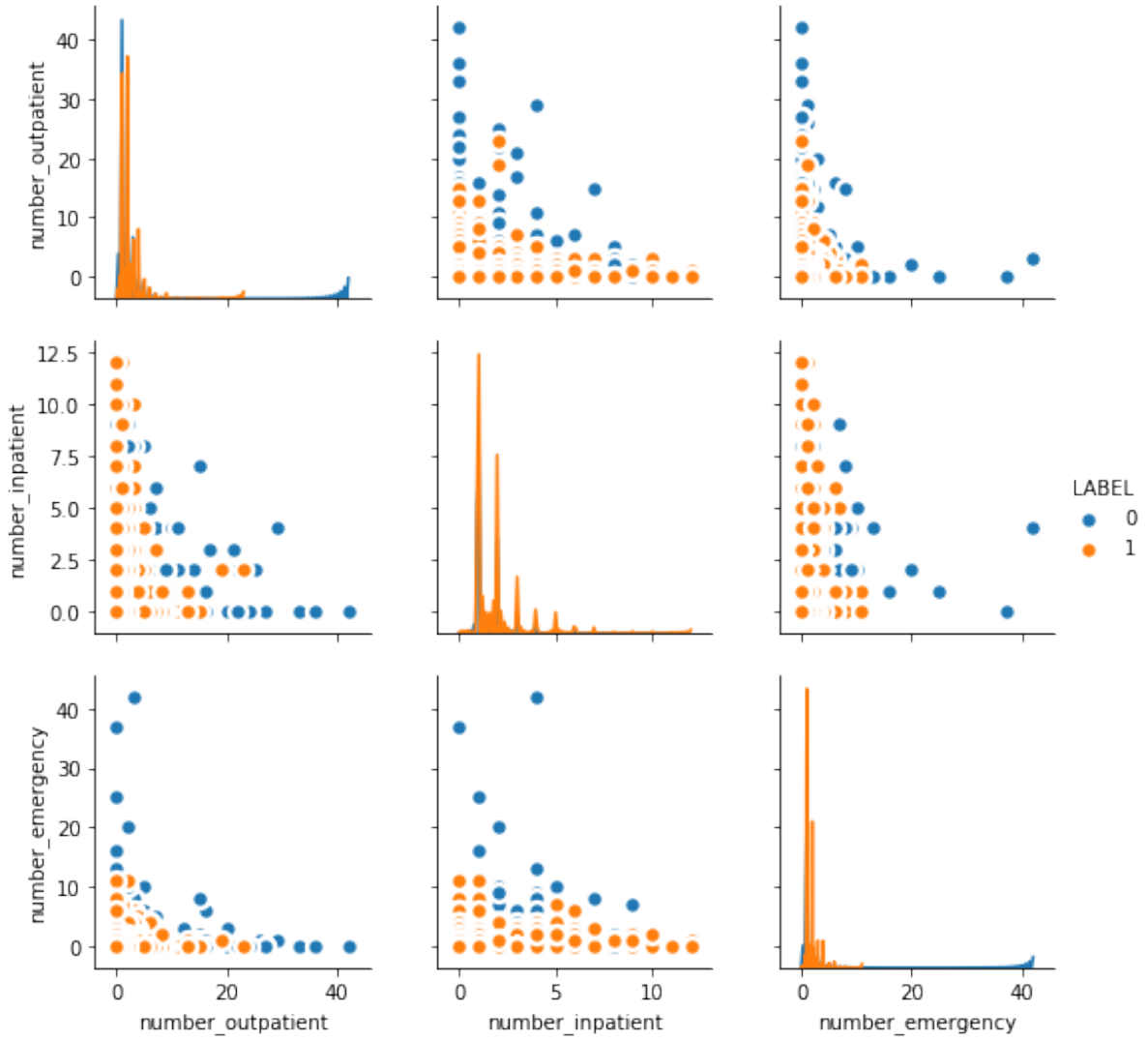
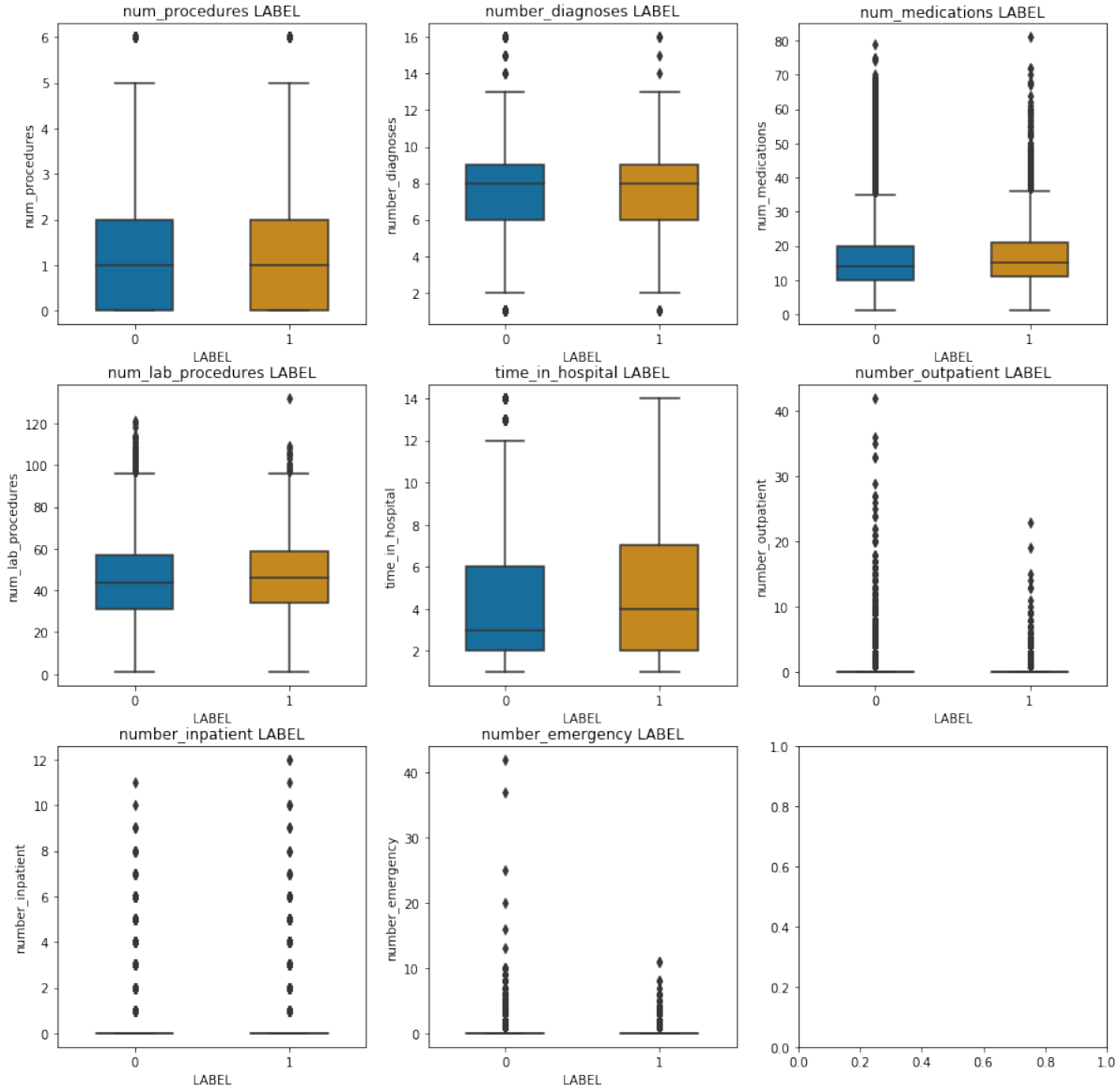


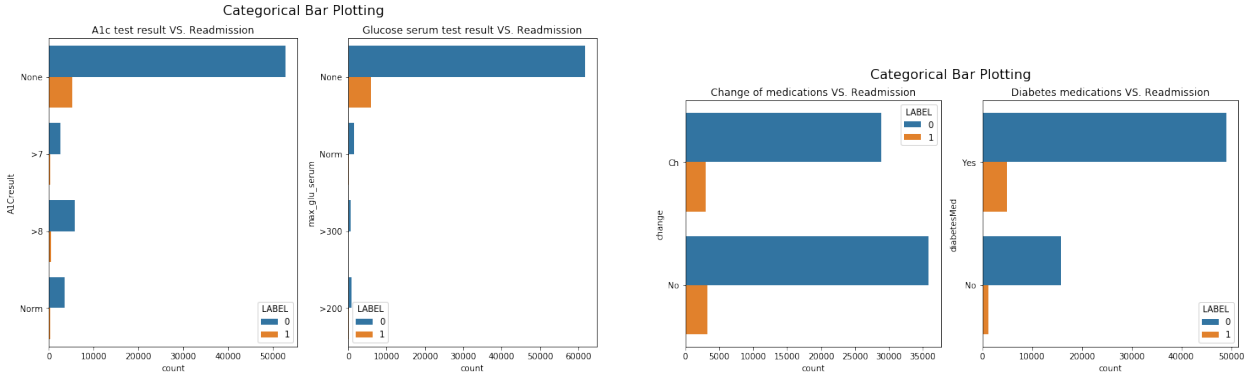
Figure 1: Scatterplot of numerical variables



Scatterplot of variables `number_outpatient`, `number_inpatient`, `number_emergency`, `num_lab_procedures`. Those three numerical variables are about different type of clinic/hospital visits in the year precdeing the encounter and they are in similar scales.



Readmissions tend to be associated with low number_emergency and number_outpatient. Those readmissions have longer time in hospital then those without readmissions. Then it suggests that the longer time stay (between 1 day and 14 days) in hospital, the more chance to have readmission.



From two-way contingency table, readmission is overrepresented in group (High A1C result and No change

of medication) when comparing encounter percentage of 7.29% to 3.11%. But this conclusion needs to be proved in a statistical test.

	LABEL	0	1	All
A1Cresult_1	change			
>7	Ch	1297	125	1422
	No	1341	121	1462
High	Ch	3739	348	4087
	No	2046	161	2207
None	Ch	22360	2359	24719
	No	30539	2840	33379
Norm	Ch	1482	139	1621
	No	1969	184	2153
All		64773	6277	71050

Figure 2: 2-way table