**Capstone II - Milestone Report**

1. **Problem Statement: Why it is a useful question to answer and for whom.**

Diabetes is a metabolic disorder in which the body is incapable of producing any or enough of the hormone insulin to absorb blood glucose efficiently.

There are two types of diabetes, both of which can lead to a condition called hyperglycemia, or high blood sugar. Type II is more common and occurs when the body cannot use insulin properly, whereas Type I is an autoimmune disease in which the body cannot produce insulin. If left untreated, hyperglycemia can turn severe and cause complications affecting the heart, nerves, and kidneys. Self-management is necessary and possible by tracking blood sugar levels on a routine basis, and through medication.

Patient readmissions to hospitals are generally unplanned. Though not always preventable, readmission can be considered a marker of negligent care and is a waste of both hospital resources and spending. Patients with diabetes are known to have a high risk of 30-day readmission following initial hospitalization for hyperglycemia (American Diabetes Association 2019). To avoid readmission, hospitals must work to implement a structured protocol for the general admission and discharge of said patients, with focus on treating hyperglycemia and avoiding hypoglycemia (low blood sugar).

Initial evaluation of the patient should determine the type of diabetes or no previous history of diabetes, and an assessment of diabetes self-management knowledge. “Because inpatient insulin use and discharge orders can be more effective if based on an A1C level on admission, perform an A1C test on all patients with diabetes or hyperglycemia admitted to the hospital if the test has not been performed in the prior 3 months”. The A1C test (also known as the glycated hemoglobin test) gives an overall view of a patient’s average blood sugar level over several weeks or months. A1C levels can help assess the efficacy of current medication/treatment and allow the medical team to make changes (ADA 2019).

Reducing hospital readmissions—especially those resulting from poor inpatient or outpatient care—has long been a health policy goal because it represents an opportunity to lower health care costs, improve quality of care, and increase patient satisfaction. Stakeholders in the case of readmission are defined as “... those entities that are integrally involved in the healthcare system and would be substantially affected by reforms to the system” (Saint Joseph’s University 2011).
A stakeholder in this example would be a hospital that must decide between the following two options: discharge a patient with the risk of potential readmission in less than 30 days, or treat them as an inpatient for a longer period of time, increasing costs and using more resources (possibly unnecessarily), all in an effort to avoid readmission.

My focus in this project is to determine which factors cause a diabetes patient to be at risk for early readmission (readmission in less than 30 days), and to use these factors to build a successful predictive model.

## 2. Description of the dataset, how you obtained, cleaned, and wrangled it

The dataset can be found in the UC Irvine Machine Learning Repository, a collection of databases used for the empirical analysis of machine learning algorithms. It is available as an Excel file with 101,766 entries, representing 10 years (1999-2008) of clinical care at 130 hospitals and integrated delivery networks across the United States. The dataset contains both inpatient and outpatient data, in which hyperglycemia management took place.

The columns in the dataset are as follows:

**Table 1. Column Names and Descriptions**

| Column Name | Column Description |
|---|---|
| **encounter_id** | Unique identifier of an encounter |
| **patient_nbr** | Unique identifier of a patient |
| **race** | Race: Caucasian, Asian, African American, Hispanic, and Other |
| **gender** | Gender: Male, Female, and Unknown/Invalid |
| **age** | Age (grouped in 10-year intervals) |
| **weight** | Weight (in lbs) |
| **admission_type_id** | Integer identifier corresponding to 9 distinct types of admission |
| **discharge_disposition_id** | Integer identifier corresponding to 29 distinct types of discharge |
| **admission_source_id** | Integer identifier corresponding to 21 distinct types of admission source |
| **time_in_hospital** | Number of days between admission and discharge |
| **payer_code** | Integer identifier corresponding to 23 distinct types of insurance payers |
| **medical_specialty** | Integer identifier corresponding to 84 distinct specialties of admitting physicians |
| **num_lab_procedures** | Number of lab tests performed during the encounter |
| **num_procedures** | Number of procedures (other than lab tests) performed during the encounter |
| **num_medications** | Number of distinct generic names administered during the encounter |
| **number_outpatient** | Number of outpatient visits of patient in the year preceding the encounter |

| | |
|---|---|
| **number_emergency** | Number of emergency visits of the patient in the year preceding the encounter |
| **number_inpatient** | Number of inpatient visits of the patient in the year preceding the encounter |
| **diag_1, diag_2, diag_3** | Primary, secondary, and additional secondary diagnosis of patient (coded as the first three digits of the *International Classification of Diseases, 9th Revision*) |
| **number_diagnoses** | Number of diagnoses entered to the system |
| **max_glu_serum** | “>200,” “>300,” “normal,” and “none” if not measured |
| **A1Cresult** | ● “>8” if the result was greater than 8%<br>● “>7” if the result was greater than 7% but less than 8%<br>● “normal” if the result was less than 7%<br>● “none” if not measured |
| **24 Features for Medications** | Feature indicates whether the drug was prescribed or there was a change in the dosage:<br>● “up” if the dosage was increased during the encounter<br>● “down” if the dosage was decreased<br>● “steady” if the dosage did not change<br>● “no” if the drug was not prescribed |
| **change** | Indicates if there was a change in diabetic medications (either dosage or generic name) |
| **diabetesMed** | Indicates if any diabetic medication was prescribed |
| **readmitted** | Days to inpatient readmission:<br>● “<30” if the patient was readmitted in less than 30 days<br>● “>30” if the patient was readmitted in more than 30 days<br>● “No” for no record of readmission |

An initial exploration of the dataset included observing the value counts for all variables. Columns “encounter_id”, “weight”, “payer_code”, “medical_specialty”, and “max_glu_serum” were dropped for missing too many values. Next, all duplicate “patient_nbr” entries were dropped, and the first recorded encounter for each patient was retained.

The nominal target “readmitted” was converted to a binary variable by dropping all instances of readmissions “>30”, and attributing a value of 0 to no record of readmission, and 1 if a patient was readmitted in “<30” days. Major class imbalances were found for many features, as well as for the target.

There were three columns for diagnosis - “diag_1”, “diag_2”, and “diag_3”. A preliminary run of Random Forest Feature Importance indicated that the secondary and additional secondary diagnoses did not have a significant impact on classifying readmission. As such, the two columns were dropped. The values in “diag_1” were converted to integers, and then binned as follows:

| Group name | icd9 codes |
|---|---|
| Circulatory | 390–459, 785 |
| Respiratory | 460–519, 786 |
| Digestive | 520–579, 787 |
| Diabetes | 250.xx |
| Injury | 800–999 |
| Musculoskeletal | 710–739 |
| Genitourinary | 580–629, 788 |
| Neoplasms | 140–239 |
| Other (17.3%) | 780, 781, 784, 790–799 |
| | 240–279, without 250 |
| | 680–709, 782 |
| | 001–139 |
| | 290–319 |
| | E–V |
| | 280–289 |
| | 320–359 |
| | 630–679 |
| | 360–389 |

**Figure 1. Primary Diagnoses Values with Reference to Binned Categories. Adapted from “Impact of HbA1c on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records” by B.Strack, 2014, April 3, *BioMed Research International, https://www.hindawi.com/journals/bmri/2014/781670/tab2/***

After binning and removing any missing values, there were a total of 38,911 remaining entries. It was found that a majority of patients were initially admitted for circulatory system issues, followed by digestive and respiratory concerns.

### 3. Initial findings from exploratory analysis

Visual EDA followed the data wrangling process, and box plots were created for the following continuous variables in the dataset:

- time_in_hospital
- num_lab_procedures
- num_procedures
- num_medications
- number_outpatient
- number_emergency
- number_inpatient
- number_diagnoses

The variables were plotted against the target.

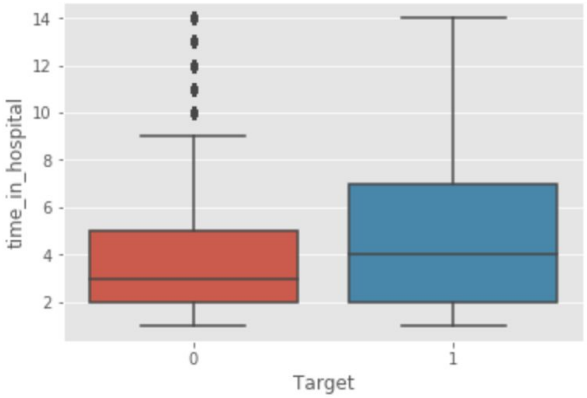



**Figure 2. Number of Days Spent in Hospital and Its Effect on Readmission**

There appears to be a significant difference in distribution between the positive and negative classes of patients regarding the number of days spent in the hospital between admission and discharge (Fig. 2). The median number of days spent for both groups is similar - 3 days for the negative class, and 4 days for the positive class. In general, patients who are not readmitted spend between 2 - 5 days in the hospital, with the maximum days being at 9, with outliers ranging all the way up to 14. For patients readmitted however, the number of days is between 2 - 7, with the maximum days topping off at 14. Thus, “time_in_hospital” would be a good feature for classifying the binary outcome.

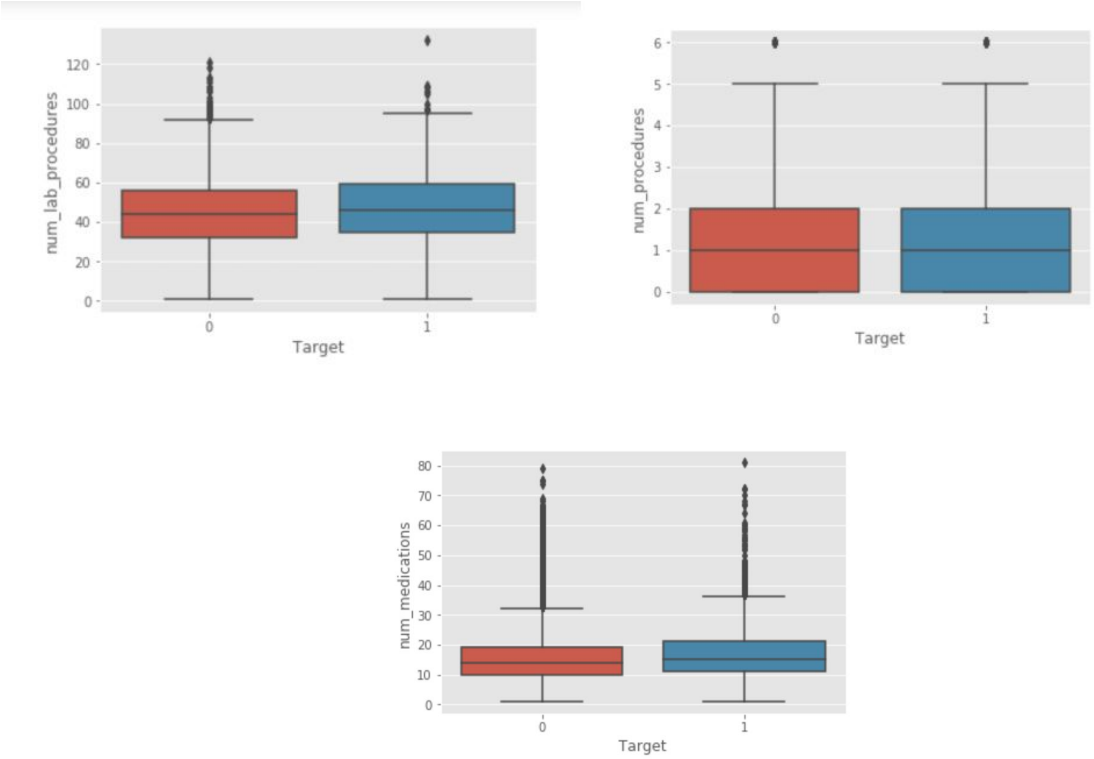
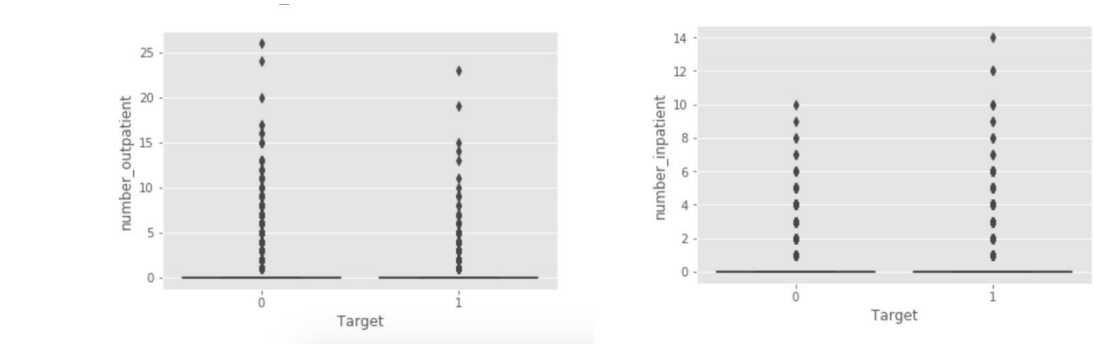



**Figure 3. The Numbers of Lab/Other Procedures and Medications Administered and their Effects on Readmission**

The distributions for the number of lab procedures and non-lab procedures performed, and the number of medications administered during an encounter are very similar for both classes (Fig. 3). These features would not classify the binary outcome well.



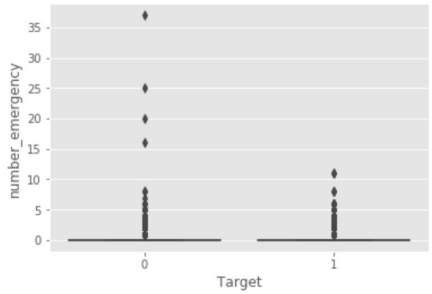



**Figure 4. Number of Visits in the Prior Year and their Effects on Readmission**

The distributions for the number of outpatient and inpatient visits in the year prior to a patient's current encounter are similar for both classes. The spread for the number of emergency visits is slightly different however. Whereas both classes show similar frequencies for 0 - 10 visits, there are multiple outliers present for the negative class of patients, where the range goes all the way up to 25 - 35 visits (Fig. 4).

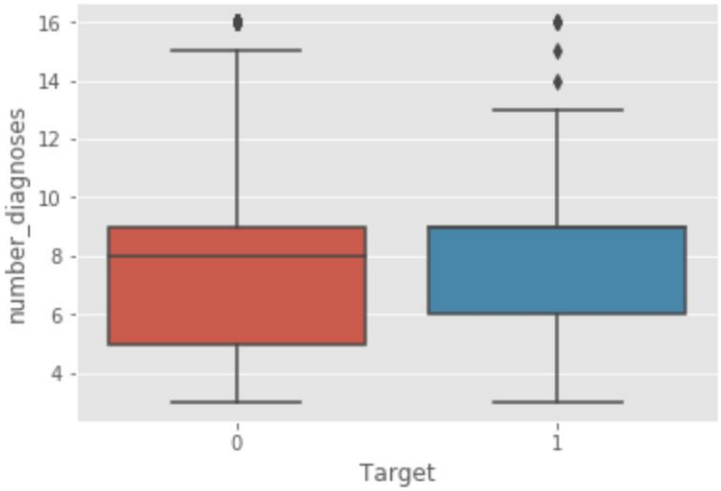



**Figure 5. Total Number of Diagnoses Entered into System and its Effect on Readmission**

The distributions for readmission in relation to the total number of diagnoses made is quite interesting. Whereas the maximum number of diagnoses for the negative class is higher than that of the positive class (15 diagnoses vs. 13), there is only a small overlap between the two boxplots, making the feature a good one for classifying the binary outcome (Fig. 5).

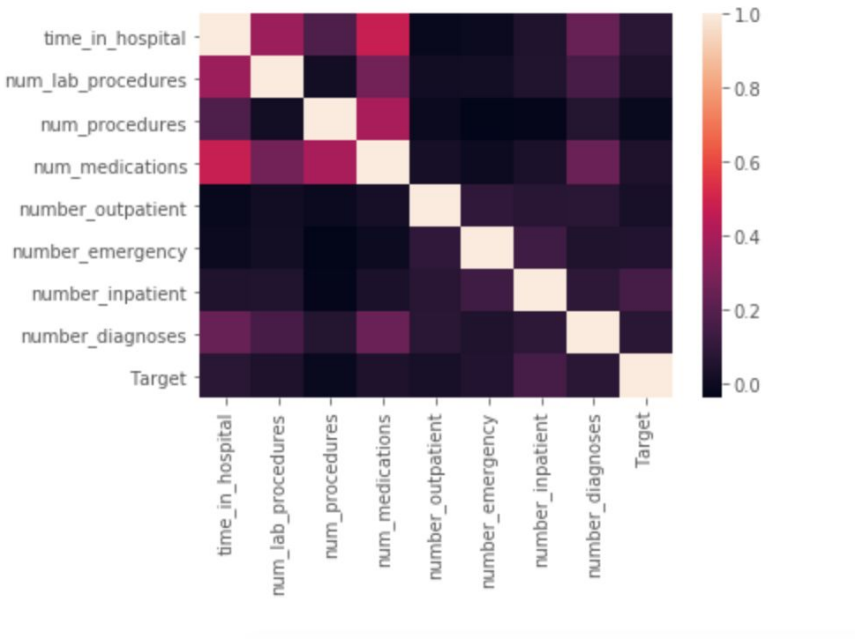



**Figure 6. Standard correlation matrix for all continuous variables**

A standard correlation matrix was graphed in an effort to find noticeable correlations and check for collinearity (Fig. 6). One immediate correlation that is seen is between the number of medications the patient is prescribed and the time they spend in the hospital.

This was followed by inferential statistical analysis to determine whether any perceived correlations were statistically significant.

Two-sample t-tests were conducted for three categorical variables - “gender”, “change”, and “diabetesMed”. A p-value of 0.053 was calculated for gender with regard to readmission, leading to the rejection of the alternate hypothesis at the 0.05 significance level, signifying that no difference in the rates of readmission can be proven for males and females.

Conversely, the null hypothesis was rejected for “change” (p-value = $1.7e^{-08}$), suggesting that there is a significant difference in the rate of readmission for patients with medication changes vs. no changes.

The null hypothesis was also rejected in favor of the alternate for “diabetesMed” (p-value = $4.15e^{-19}$), implying that there is a significant difference in the rate of readmission for patients with diabetes prescriptions vs. no prescriptions.