Alexander Olden

# Milestone Report: Assessing Diabetes Readmission with Machine Learning

## 1. Introduction to the Problem

Hyperglycemia is an excess of glucose in the bloodstream that is often associated with diabetes. Unsurprisingly, management of this condition matters greatly to hospitals and physicians. Prior to the data collection and analysis of Beata Strack et al., however, there was little nationwide research to serve as a baseline for tracking changes in hospitalized patients with hyperglycemia [1]. The Strack team's analysis focused on the regularity of measurement of HbA1c (glycated hemoglobin), which forms when red blood cells join with glucose in the body. Measuring HbA1c allows physicians to measure blood-sugar levels in patients with hyperglycemia.

Historically, measurement of HbA1c has been performed infrequently, which Strack et al. found to be the case in their data exploration. As such, the question of interest is how well one can predict hospital readmission within 30 days using the variables in this data (including HbA1c measurements). Additionally, I have since introduced other variables that the original researchers did not use, such as the number of patient visits to a medical facility. A secondary question of interest is whether or not these new variables effectively predict readmission.

Clients in this project are stakeholders associated with healthcare institutions interested in studying and reducing the readmission rates of diabetes patients. The results of my analysis will help them determine whether or not monitoring HbA1c levels should be a significant part of treatment. Said stakeholders will also be interested in monitoring the new features that I have created, should they prove to be significant.

## 2. Data Attributes and Limitations

The data set in use contains a few demographic variables: race, gender, age, and weight. Unfortunately, weight had to be dropped as a feature because 92% of its data were missing. There are also several categorical features pertaining to patient

admission and discharge: admission type, discharge circumstances, number of days spent in the hospital, insurer, diagnoses, and attending physician's specialty. Like weight, insurer had to be removed due to missingness. And as previously indicated, I will consider number of visits for a patient as a predictor, too.

Additionally, there are several features providing information about certain medications and clinical tests. The primary interest in this subset is whether or not the HbA1c test was performed, of course, but I also plan to consider HbA1c by its different result categories (3 different levels of HbA1c, and no test performed).

Finally, a few limitations that are typical of clinical data appear. The biggest one is missingness in demographic variables, which prevented me from answering questions pertaining to patient weight and insurer. The data is anonymous and was not collected from a certain demographic region with available, representative data, so there are not reasonable steps or additional data that could be imported to help answer relevant questions. Another limitation is the absence of numeric and continuous variables. Many features are categorical by convention, but even features that frequently appear as integers or continuously (such as age and certain test results) were rendered categorical. This makes hypothesis tests more challenging or untenable in some cases and will require mapping of string variables to numeric ones for final analyses.

Much of the data wrangling entailed handling missing data. For variables with large percentages (more than 50%) missing, I investigated the extent to which data were missing at random to ensure no underlying cause of missingness existed. Data appeared to be missing largely at random for all relevant variables, so I dropped them from the data set. A few other variables had smaller percentages (3% or less) missing, which is fairly insignificant in a data set with about 100,000 observations. In these cases, I simply dropped observations with data missing for the features.

The other data-wrangling steps performed were routine preparation for further analysis. I converted the outcome variable from a string to a binary for use in algorithms and exploratory analysis. And for patients with multiple hospital visits, I limited their observations to only the earliest encounter to keep observations independent (after capturing their total numbers of visits). Last, to further reduce bias, I removed encounters that resulted in death or discharge to a hospice facility.

## 3. Preliminary Exploration

To begin exploratory analysis, I considered the demographic variables that were not dropped, starting with patient age. A histogram showed that patient ages skew older in the data, which provided useful context. Additionally, I suspected that some features may correlate with age (such as number of medications and days spent in the hospital), so I checked for such relationships with different plots. No clear relationships emerged, so multicollinearity is not likely an issue with age and the variables pertaining to patient stays in a medical facility.

As a small sidebar, I also looked for a potential relationship between race and readmissions rates to see if historically underserved populations were also being underserved where diabetes readmission is concerned. No evidence emerged of that trend in the data, I am glad to report.

After looking at independent variables alone, I checked key variables for potential relationships with the outcome, readmission within 30 days. I initially found that the ratio of negative (not readmitted within 30 days) and positive (readmitted within 30 days) to be about 10:1. After seeing a possible relationship between having HbA1c levels tested and being readmitted, I performed a hypothesis test and found that patients with HbA1c levels tested were less likely to be readmitted within 30 days than were patients who were not tested, and that the relationship was not due to chance. (Note: a paired bar graph also showed a possible relationship when HbA1c was not in a binary format but rather split into its 4 original categories.)

Finally, I checked for a relationship between a patient's number of hospital visits with readmission. The resultant graph showed an exponential, negative trend between readmissions frequency and number of hospital visits, so the newly created variable could have some predictive power.

## 4. Approach

Overall, I still plan to solve this problem by comparing the results of a few machine-learning methods with the results that Strack, et al. saw from logistic regression. In particular, I will consider Naïve Bayes and AdaBoost in addition to logistic regression. Depending on these initial results, I may also look at Bayesian Networks, Random Forest, and Neural Networks.

There are two additional features that I now plan to use in my final approach, however. One is the use of HbA1c testing not only as a binary but as a categorical one with four different outcomes, too. I am interested to see if this feature maintains its significance in this form. Second, I want to determine if number of patient visits, my newly created variable, has significance in final models since it showed the potential to do so in my exploratory analysis.