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). In this capstone project, I have 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 to what extent these new variables effectively predict readmission.

Clients in this project are stakeholders associated with healthcare institutions interested in studying and reducing the readmission rates for diabetes patients. The results of my analysis can potentially help them determine whether monitoring HbA1c levels should be a significant part of treatment. Said stakeholders might also be interested in monitoring the new features that I am considering—which have been taken from the existing dataset—in order for them to determine if they prove to be significant in practice.

2. Data Attributes and Limitations

The dataset in use contains a few demographic variables: race, gender, age, and weight. Weight had to be dropped as a feature because 92% of data points were missing this value. There are also several features pertaining to individual patient visits: admission type, discharge circumstances, number of days spent in the hospital, insurer, diagnoses, and attending physician’s specialty. Among these features, insurer had to be removed due to missingness.

The features pertaining to diagnoses also warrant brief mention in the data-wrangling process. The three relevant features all had numerical values that could be grouped according to condition (e.g., circulatory). To create more meaningful values – and to expedite the one-hot encoding process to come – I grouped the numerical codes accordingly and then converted them to their related condition.

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

Finally, a few limitations that are typical of clinical data appear. The biggest one is missing values in demographic variables, which prevented me from answering questions pertaining to patient weight and insurer. Unfortunately, there are not reasonable steps to take or additional data to import to help compensate for this missing information. Another limitation is that some conventionally continuous features appear as categorical ones. This makes hypothesis tests difficult in exploratory analysis.

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 variable for ease of use in algorithms and exploratory analysis. Also, many patients had multiple hospital visits within the data set, which means that certain observations might not be independent. (A follow-up visit depends on an initial visit, for instance.) For these patients, I limited observations to only the earliest encounter to preserve independence (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

I aim 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 use Naïve Bayes and AdaBoost in comparison with logistic regression. If no convincing differences emerge, I may also look at Bayesian Networks, Random Forest, and Neural Networks.

There are two additional features that I will use in my final approach. First is the use of HbA1c testing not only with four different, categorical outcomes but as a binary variable, too. The original researchers found predictive value in HbA1c as a four-category variable; I want to know it remains useful in binary form as well. Second, I want to determine if number of patient visits, my newly created variable, has statistical significance in final models since it showed the potential to do so in my exploratory analysis.