Alexander Olden

Proposal: Assessing Diabetes Readmission with Machine Learning

1. Background and Problem to be Solved

Hyperglycemia is an excess of glucose in the bloodstream that is often associated with diabetes. Unsurprisingly, management of this condition is of significant interest 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 had been performed infrequently, which Strack et al. found to be the case in their data exploration. As such, the question of interest here is how well one can predict hospital readmission within 30 days using the variables in this data (including HbA1c measurements). A potential secondary outcome is whether tracking HbA1c more closely would improve patient outcomes in the form of reduced hospital readmission.

2. Clients and Their Interest

Clients in this project are all stakeholders associated with healthcare institutions who are 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.

3. Data and Acquisition

Relevant data has been collected and aggregated by Strack et al. from hospitals across the country over the course of 10 years. I accessed it through The UCI Machine Learning Repository[1]. The set has about 100,000 observations and 55 variables. I have downloaded it as a CSV file and will read it into Python for exploration, machine learning, and visualization. I may also use R for this part.

While I will not have to do much data wrangling, it is worth noting that this data set has a few complicating features that are typical of clinical data. For one, some covariates (e.g., weight)

[1] http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

in the dataset have missing values. As such, I will need to consider imputing values or taking other steps to address missing data. Additionally, the data set uses both categorical and continuous variables.

4. Approach

I will 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 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.

For context, I have read three relevant papers discussing the use of clinical data to evaluate hospital readmission. One paper is the previously mentioned study by Strack, et al. A second one (Bhuvan, et al. [2]) also examined readmission with respect to diabetes. The third study (Hon, et al. [3]) looked at readmission concerning congestive heart failure by comparing different machine-learning approaches. Its aim was to stratify patients based on their risk for readmission within 30 days (the same time frame used by Strack, et al.). The table below compares salient parts of each paper.

Authors	Strack, et al.	Bhuvan, et al.	Hon, et al.
Medical Condition	Diabetes	Diabetes (interest in cost savings)	Congestive Heart Failure (model comparison)
Dataset Features	From hospitals across the country over 10 yrs., 100K observations, 55 variables	102K records from 130 US hospitals collected over 10 yrs., 50 potential risk factors to predict dichotomous outcome of readmission w/in 30 days	Non-public, obtained from CA state office; 500K unique patients; variables in 5 groups: demographics, cost/length of admission, clinical metrics of admission, comorbidities (from diagnosis codes), count data
Algorithms Used	Logistic regression	Naïve Bayes, Bayesian Networks, Random Forest, AdaBoost, Neural Networks	Logistic regression, decision trees, boosted trees (AdaBoost)
Main Results	Measuring HbA1c is associated with reduced readmission	Random forests best at identifying patients at high risk for both short-term and long-term readmissions	Logistic regression best w/ readmission as outcome; boosted tree best w/ cost and length of stay
Proposed Future Work	Look at finance – can findings help reduce readmission rates and care costs for patients?	A proposed cost-sensitive model showed that $252.76 million can be saved for 98,053 instances of diabetic patient encounters; look at model in real healthcare systems to evaluate health and financial outcomes	Investigate other machine-learning methods; improve models through feature engineering

5. What are your deliverables?

My deliverables will include (1) all relevant software code (2) a paper discussing processes and results, and (3) a slide deck. These items will be stored on Github.

References

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[2] Malladihalli S Bhuvan, Ankit Kumar, Adil Zafar, and Vinith Kishore. “Identifying Diabetic Patients with High Risk of Readmission.” Retrieved from https://arxiv.org. February 12 2016.

[3] Chun Pan Hon, Mayana Pereira, Shanu Sushmita, Ankur Teredesai, and Martine De Cock. “Risk Stratification for Hospital Readmission of Heart Failure Patients: A Machine Learning Approach.” Retrieved from http://faculty.washington.edu. October 2, 2016.