Project Milestone 2: Data Acquisition & Understanding

Project Team: Karim Gowani, Ryan McGillicuddy, Yaseen Mohmand, Steven Worthington

Project Title: Predicting Hospital Readmission Rates for Diabetes.

Project Context

Our dataset is from the UC Irvine ML Repository and involves patient records of those diagnosed with diabetes from 1999 through 2008 at 130 US hospitals. We have downloaded this dataset and examined it. It has ~102K records, a binary target variable, and 47 features, the majority of which are categorical.

Our project will therefore focus on a classification problem, where the target is whether a given patient was readmitted to hospital early, i.e., within 30 days of discharge.

Our primary performance metric will be AUC, as not only is this metric invariant to class imbalance but it will also help rank patients and prioritize them for intervention in a clinical setting.

The rest of this paper will discuss the granularity of data, class imbalance, missingness, scaling, and additional preprocessing steps that we are planning to prepare data for downstream tasks including modeling.

Data Granularity

In a clinical setting, doctors and medical staff would like to answer the question, “given information from the current and previous hospitalizations, how likely is it for this patient to be readmitted to hospital early (within 30 days)?”. This question is inherently at the patient-level, but each record in the dataset is at the level of an ‘encounter’, which represents a patient hospitalization event (rather than an outpatient visit). A subset of 16.5% of patients have multiple encounters.

A patient-level perspective is more likely to be of benefit to clinicians, since answering the above question will help medical personnel prioritize follow-ups and interventions through the creation of patient risk profiles, which can identify patients at the highest risk level for early hospital readmittance. This information is actionable and can be used to mitigate negative health outcomes for these patients as well as increased costs for the hospital and insurance carrier. We will therefore aggregate data from the encounter-level to the patient-level.

For those patients with multiple encounters, however, features that vary at the encounter-level contain important information that we do not wish to discard. For example, if a patient was readmitted early relative to the immediately preceding encounter, it is perhaps more likely that the patient will be readmitted to hospital early again after the current encounter. Therefore, our strategy will be to select only the *final* encounter for these patients and create several new derived features that encapsulate the history of their *previous* encounters. Such features will include, but are not limited to, the number of previous inpatient encounters, whether the last encounter resulted in early readmission, and whether the patient ever had a high value of A1c.

In following this approach, we will have to make a (reasonable) assumption that encounters for each patient are in temporal order in the dataset because no explicit date information is provided.

Class Imbalance

About 11% of encounters belong to the positive class (readmitted within 30 days), so while there is imbalance, it is not severe. While our performance metric of interest - AUC - is robust to class imbalance, we will still try to address this issue in several ways. We will use stratified sampling in train/test splits, and we will attempt the standard techniques of undersampling and oversampling, as well as use of class weights built into the different ML models of interest, including Logistic Regression, CART, Random Forest, and XG Boost.

Missingness

There are only 7 (out of 47) relevant columns that contain missing values:

- **weight** is missing ~97% of its values, so this column can be safely dropped; no other numerical column has missing values.
- **medical specialty** is missing ~49% of its values, but may be relevant to the classification task, so we keep it and fill the missing values with 'unknown'..
- **payer code** (insurance carrier) is missing ~40% of the values, but because it does not seem to be relevant to the target - it is also a candidate for being eliminated altogether.

The remaining columns have less than 3% values missing so they can be managed. They are all categorical, including race and diagnosis codes. As is common for categorical variables, we will fill the missing values with 'unknown'.

Scaling

We are planning to use regularization with logistic regression, and will therefore scale the numeric variables (either through standardizing or normalizing); this way, the magnitude of the coefficients will not be affected by the different units (such as for variables 'days in hospital' vs. 'number of medications') in the penalty term of the regularization.

Additional Preprocessing

Several of the categorical variables have many categories each that should be easily collapsed to reduce dimensionality:

- At the most extreme, the first diagnosis column has 716 codes, only 23 of which represent more than 1% of the observations; similarly for the second and third diagnosis variables. In fact, these diagnosis codes should be grouped into types such as Circulatory (codes 390-459), respiratory (codes 460-519), digestive (520-579), etc.
- Admission type code has 8 categories, 3 of which make up less than 1% of observations and can be safely collapsed.
- Medical specialty has 72 categories, but only 9 represent more than 1% of the observations.
- Age buckets can be consolidated: Currently, each bucket includes only 10 years. Less than 1% of the observations fall into age < 20 and age > 90, for instance.

Furthermore, patients who were discharged with codes such as expired, hospice, transferred to another institution as inpatient, etc. should be filtered out as these types of discharge codes are of no practical relevance for predicting the target of early readmission. Trivially, encounter ID and patient ID are mere identifiers and should not be fed into any modeling.