95885 Data Science and Big Data

Project Proposal

Team 03: Daniel Lesser, Carol Xiang

Health care has become a topic of national importance in recent years, with ballooning costs and increased consumer awareness.  Specifically, obesity and diabetes are some of the most prevalent, preventable diseases that afflict individuals in the U.S. We believe there is an opportunity to provide better health care and reduce costs by deploying supervised machine learning on historic hospital visits to learn more about patient readmission as well as ancillary diseases that diabetic patients often suffer from.

According to the Centers for Disease Control and Prevention (CDC), as of 2015, 30.3 million people in US are diagnosed with diabetes and another 84.1 million people have prediabetes, a condition that if not treated often turns into type 2 diabetes within five years. In other words, more than a third of the US population is burdened with some level of diabetic condition. Furthermore, patients with diabetes usually have coexisting complications, such as cardiovascular disease and kidney disease. For example, 36.5% of adults aged 20 years or older with diabetes also has kidney disease. Diabetes is a detractor to the economy, whether by direct medical costs or costs due to reduced productivity, estimated at $327 billion in 2017. Among those, hospital readmission accounts for a significant portion of this economic burden.

Given those statistics, we find it important to help hospitals identify patients with a high risk of readmission. When a doctor is informed with lab tests, medical history, chronic conditions, and demographic information about the patient, how can the doctor make a decision so that the patient is less likely to be readmitted within 30 days? Since diabetes is usually concurrent with other symptoms, how can doctors better diagnose ancillary diseases so that patients can be informed about their conditions and get treatment for both diseases? In a traditional setting, doctors are capable of making these decisions based on their expertise and knowledge about patient’s conditions. However, our machine learning model will serve as an advisory component for them when they are under tight time constraints and high pressure.

The dataset we will leverage is from the UCI Machine Learning Repository. The data has approximately 100,000 diabetic inpatient encounters at 130 US hospitals between 1999 and 2008. Each observation records an individual diabetic patient’s first hospital admission within the time frame. Historical information as well as follow-up appointments are recorded as features for each patient. There are 55 features in total. The features can be segmented into the following categories:

* Identifiers: a unique identifier of an encounter and a unique identifier of a patient
* Demographics of the patient, e.g. race, gender, age, weight
* General information regarding the admission, e.g. admission type (emergency, urgent, newborn, etc.), discharge disposition (discharged to home, expired, etc.), time in hospital, payer code
* Patient’s medical history, e.g. hospital visits in the year preceding the encounter, number of medications
* Lab tests, e.g. A1c test result, Glucose serum test result
* Medications: 24 categorical features that indicate for each medication whether the patient took the medication and if yes, how the dosage was adjusted following the encounter.
* Diagnosis: number of diagnoses; categorical Diagnosis 1, 2 and 3 (coded primary, secondary and additional secondary diagnosis)
* Readmitted: days to inpatient readmission with three possible values - within 30 days, after 30 days, and no record of readmission.

    Some features have missing values. For example, 97% of weight values are missing. We will not be able to use those features with most of entries as empty. On the other hand, for the features like payer code which have 40% values missing, we may be able to impute the missing value from other observations using kNN or multinomial logistic regression.

    The dataset is inherently imbalanced such that about 90% of observations were not followed by a readmission within 30 days. Thus, before running machine learning algorithms, we will balance the dataset using an oversampling method such as SMOTE.

    Given the dataset and stated learning objectives, this task is most suited to a supervised learning approach.  Specifically, this will be a multi-class classification task for predicting readmission type and ancillary diseases.  Prior research on this dataset employed multivariate logistic regression, which we intend to use as a benchmark. Our group will also attempt to use Random Forest and SVM approaches to do classification.

    The experience portion of the project will be the 100k observations, which represent unique patient visits to a U.S. hospital.  This data will be split into train, validation, and test datasets in order to build the machine learning models.

    Performance will be measured through several evaluation metrics.  Accuracy will be used as a baseline metric. However, accuracy alone may be a misleading indicator of performance as there is an imbalance in values for our output labels.  For instance, only ~10% of observations see a readmission within 30 days. Because of this, we intend to use a precision-recall curve as well. It is important to review both precision and recall due to different desired outcomes for various parties.  Administrators may want a high precision score in order to only spend extra resources on patients that really need them. Conversely, physicians may prefer a high recall score in order to provide extra care to every patient that is at risk of readmittance.

    Health care in the United States is an industry rife with opportunity for optimization, better patient outcomes, and cost reduction.  By applying machine learning techniques to the diabetes epidemic, we hope to reduce patient recidivism and improve patient awareness to ancillary diseases that may be affecting them.