Predicting Hospital Readmittance Via Machine Learning Classification Techniques Using Structured and Unstructured Data

Evan Canfield & Christine Personett

DSBA 6156: Applied Machine Learning

October 25, 2019

Insights Seeking

Hospital readmission rates are an important consideration for hospitals in the United States. Readmission rates are directly connected to Medicare reimbursement, with Medicare being the largest payer of hospital services in the country. Hospitals with high readmission rates incur financial penalties. Readmissions are defined as any return to the hospital within 30 days of being discharged. Currently, the LACE Index is to predict readmissions and provides a score based on four features: the patient's length of stay, acuity, co-morbidities, and emergency department visits within the last 6 months. This tool is used by physicians currently, and within some studies show poor discrimination in distinguishing between patients that will and will not be readmitted.

We will be exploring the prediction of hospital readmissions through machine learning methods applied to both structured and unstructured data. In analysis of structured data, we will be employing modeling methods such as logistic regression, naive bayes, support vector machines, and random forest. In analyzing unstructured data, we will process digitized doctor reports through natural language processing methods. Through the analysis of both structured and unstructured methods, we plan to improve on the prediction of patients likely to be readmitted.

Data Gathering

The MIMIC-III data is a relational database of patient information gathered from Beth Israel Deaconess Medical Center. It contains twenty-six tables with information on patient stays, diagnoses, procedures, prescriptions, lab results and vital signs among other attributes. There are some missing measurements in the data and some measurements that have different values assigned in across different tables. We will be joining tables on keys provided to access all of the required data.

Data Access

We completed the CITI Program course, "Data or Specimens Only Research" on October 16th and 17th. Then we submitted a data use agreement and credentialing application with Physionet. Access was granted on October 24, 2019.

Features Names and Number

There are 26 tables in the database, they vary in size from hundreds to millions of rows. The largest table has 330,712,483 rows. In analyzing the available structured data, we will primarily be using the Admissions table to access individual patients and their unique hospital visit. It has several data types including integers, date-time, strings and binary values.

For unstructured data analysis, we will be analyzing the Noteevent table, which collects digitized doctor reports for each patient. The Noteevent table contains 2,083,180 observations, with the digitized reports stored as free text.

Analysis

The analysis of hospital readmissions will consist of two parts: analysis of structured and unstructured data. After processing, the structured data provided in the MIMIC-III dataset will be analyzed with logistic regression, naive bayes, random forest, and support vector machine in order to predict hospital readmittance.

The unstructured data in the data set is in the form of digitized doctor's reports. These reports will be processed with natural language processing methods. Once processed, the notes will be analyzed with similar classification models to predict hospital readmittance.

The outputs of each method, structured and unstructured, will then be scored and compared to determine which model provides the most desirable and robust results.

Expected Accomplishment

The Lace Index is currently one method used to identify patients likely to be readmitted. Multiple studies have analyzed the LACE method. The predictive scoring capabilities of the models developed for this project will be compared to the available output of such studies.