Nate C. Carnes, Ph.D.
UCSD Extension Machine Learning Engineering Bootcamp

Capstone Project Proposal

Problem: The problem this project seeks to solve is to predict when and whether a patient hospitalized with a mental, behavioral or neurodevelopmental disorder will be readmitted to the hospital in the same year. The purpose of hospitalization for patients with a mental health diagnosis (as the primary diagnosis, as opposed to a comorbid diagnosis) is to identify and arrest the degradation associated with the disorder through a combination of information, medication, and outpatient treatment. A patient who is nonetheless readmitted in a similar or declining condition represents a failure of this treatment approach. Identifying who will likely be readmitted (as well as when and why) has the potential to reveal vulnerabilities, inform practitioners' treatment approaches, and prevent the rehospitalization of patients with various mental health disorders.

Data: This problem will be solved using the Healthcare Cost and Utilization Project (HCUP) 2018 Nationwide Readmissions Database (NRD). The NRD is a nationally representative sample of discharges for patients with and without repeat hospital visits in a year and those who have died in the hospital. This dataset contains *N* = 17,686,511 hospital encounter-level records, of which *N* = 1,044,120 include a primary diagnosis code for any mental health disorder, and there are over 100 variables containing demographic, hospital, medical, and financial information. This data was purchased from HCUP after completing the necessary training and application forms; it involves a data use agreement that restricts access because of the sensitivity of the data.

Approach: This is a supervised learning problem because all of the tabular data has class labels (such as whether or not a patient was readmitted within three months). After data processing and exploratory data analyses, classification algorithms will be employed to learn and classify the class labels, which could be treated as binary or multiclass depending on the construction of the class labels. Multiclass labels could reveal both whether a patient is readmitted and when that occurs (within a binned set of time periods); alternatively, multiclass labels could reveal the frequency of readmission. All of the available information in the dataset could be leveraged as predictors, but both feature selection (perhaps via regularization) and feature generation (perhaps via interaction term encoding) will be employed, and the most important set of predictors are likely to be the ICD-10-CM/PCS aggregate codes. Multiple classification algorithms (e.g., logistic regression, random forest with boosting, deep learning) will be employed and compared according to different evaluation metrics including accuracy, precision, recall, F1 score, and Matthews correlation coefficient.

1. Obtaining data
2. Combining data tables and restructuring for readmission analysis
3. Exploratory data analysis
    a. Cleaning data (including missingness)
    b. Recoding data (including ICD-10 codes)
    c. Transforming data (including normalization and standardization)
4. Feature engineering

5. Data analysis
    a. Baseline methods (logistic regression, kNN, SVM)
    b. Ensemble methods (random forest, xgboost/catboost, deep learning)
    c. Stacking methods
6. Application development
7. Report write up

Deliverable: A final deliverable will be a detailed report of the data processing, exploratory data analysis, and findings across different classification algorithms (as well as a description of the problem and evaluation of the solution); all code will be made available. It is important to note that the data will *not* be accessible to protect the confidentiality of these medical records. An additional final deliverable could be an application deployed as a web service with an API that allows users (such as practitioners) to retrieve a predicted class label (and the associated class label probabilities) for a single patient using the final trained algorithm to improve treatment.

Computation: The supervised learning models employed in this project will likely consume substantial computational power given the number of models, the amount of data, and the models themselves; however, stratified sampling will be employed in the earlier modeling stages to reduce the needed computational power for model development, and then a virtual machine with greater processing power and memory can be employed in the later modeling stages to train and evaluate the models with all of the available data (using cross-validation).