Capstone 2 Proposal
Prediction of Intensive Care Unit Mortality Using Natural Language Processing

Background

As the world copes with the Covid-19 pandemic, hospital mortality rates have naturally come to the forefront. In-hospital mortality has long been used as a measure of hospital quality. The intensive care unit (ICU) has the highest mortality rate of all the units in a hospital. Of the approximately 4 million ICU admissions per year in the United States, the average mortality rate ranges from 8-19%, or about 500,000 deaths annually.

Electronic Health Records (EHR) have become nearly ubiquitous in medicine over the past decade. They provide a wealth of information that traditionally was used mainly to document care for reporting, liability, and billing purposes. As they witnessed data science and machine learning making a great impact on business outcomes in other industries, healthcare systems and many providers have become interested in utilizing EHR data for prediction and to support clinical decision-making.

One way to take advantage of EHR data is to analyze the data present in caregiver notes. Natural language processing (NLP) can use words and phrases from text, like *sepsis* and *status worsening,* to be included in predictive models. The objective of this project is to use NLP to predict in-hospital mortality.

Client

The client is a healthcare system looking to reduce in-hospital mortality rates in order to improve patient outcomes and its quality of care scores.

Data

MIMIC-III is a large, freely-available database comprising de-identified health-related data associated with over 40,000 patients who stayed in ICUs of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality.

We will use data from the following MIMIC III datasets:
ADMISSIONS - contains admission discharge dates (has unique identifier HADM_ID)
NOTEEVENTS - contains caregiver notes on each hospitalization (has identifiers SUBJECT_ID and HADM_ID)
PATIENTS - contains information about patient's date of birth and death, if applicable (has unique identifier SUBJECT_ID)

The Approach

The approach will be to first explore and clean the data. Python pandas, numpy, matplotlib, and seaborn will be used for data wrangling and exploration. Next, an algorithm will be built to capture data from all caregiver notes written during the first 24 hours of the ICU admission. To allow the machine learning model to accept the data, text data from caregiver notes will then be processed into numbers using the Bag-of-Words approach.

The DOD_HOSP (date of death in hospital) variable will be used to classify whether a patient expires while in the hospital (1=death, 0=no death). Correlations and comparisons will be viewed, many questions will be answered, and some insight to the predictive model will be obtained at this point.

Next, the data will be split into training and test sets, ready to implement into machine learning models.

Deliverables:

The final deliverables will include:

1. Python code on GitHub, including:

   - Data wrangling (including text data pre-processing)

   - Data exploration

   - Machine learning model development

2. A final paper, located in GitHub, which explains the problem, approach, and findings. Ideas for further research will also be included, along with recommendations on how the Client can use the findings.
3. A Slide Deck for the project, also located on GitHub
4. An online video, office hour presentation, or blog post to share with other Springboard participants.