Capstone I Proposal

Background

High hospital readmission rates are associated with poor patient outcomes and high financial costs for both private and government-sponsored insurance programs. Cardiovascular disease is the leading cause of rehospitalization for patients in the United States. This capstone project will explore the relationship between variables related to cardiovascular disease and hospital readmissions. The goal is to create a machine learning model that predicts the likelihood of a patient with heart disease being readmitted within the next year, by isolating factors that are correlated with readmission. Thereafter, we can identify ways to address potentially causative factors for rehospitalization.

The Client

In effort to reduce financial burden and improve patient outcomes, our client, the US Department of Health and Human Services, has asked us to identify ways to reduce hospital readmission rates for patients with CHF and other heart-related diagnoses.

The Data

We will use data from the National Health and Nutrition Examination and Survey (NHANES), which provides a cross-sectional survey of health status of the US population each year. NHANES is made up of a series of health and nutrition surveys conducted by the Centers for Disease Control and Prevention since the early 1960's. Every year, approximately 5,000 individuals of all ages are interviewed in their homes and complete the health examination component of the survey. The NHANES target population is the noninstitutionalized civilian resident population of the United States. The NHANES design has changed periodically to sample larger numbers of certain subgroups of public health interest to increase the reliability and precision of estimates of health status indicators for these population subgroups. In this survey, non-Hispanic Asians, Hispanics, non-Hispanic blacks, older adults, and low income whites/others have been over-sampled.

For our training data, we will use the dataset from NHANES 2015-2016. In 2015-2016, 15,327 persons were selected for NHANES from 30 different survey locations. Of those selected, 9,971 completed the interview and 9,544 were examined. The data is publicly available on the CDC's website, at https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015.

There are several datasets available for the 2015-2016 Survey. All datasets can be linked by the SEQN variable which is tagged to each individual survey participant.

We will obtain data from the following portions of the survey:

- **Demographics Data** (DEMO_J) (variables we will use include: gender, age, education level, income)
- **Questionnaire Data**:
    - Acculturation (ACQ_J) (languages spoken at home)

- Medical Conditions (CDQ_J) and Age of Onset of condition (CHF, Heart Attack, Stroke, Angina) - Our sample dataset will include only patients who indicated that they have one of these conditions.
- Medicare or Medical Insurance (HIQ_J)
- Hospitalization information (HUQ_J - This will be used for the Y variable. HUQ071: Overnight patient in last year? HUQ080: Number of times overnight patient in last year?)

The Approach

The approach will be to identify adults with a history of cardiovascular disease who have other characteristics in common. We will then determine what combination of characteristics might be expected for a patient at-risk for rehospitalization (that is, with a likelihood of having greater than one hospitalization per year).

Python numpy and pandas packages will be used for data wrangling. The datasets will be reduced to the variables of interest, merged, then cleaned and checked for inconsistencies or missing/duplicate data.

Matplotlib and seaborn packages can provide exploratory data analysis through plots and visual representation of statistical values. Correlations and comparisons will be viewed here. Most questions will be answered here, and will provide insight to the predictive model.

Deliverables:

The final deliverables will include:

1. Python code on GitHub, including:

   - Data wrangling

   - Data exploration analysis

   - 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 PowerPoint 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.