Capstone Project 2 Proposal

Problem/Usefulness of Solution: The project is for a company called Neurolex Labs. Neurolex analyzes voice samples to help diagnose diseases such as schizophrenia, Parkinson's disease, Alzheimer's disease, and depression. In this case, we are analyzing human voice samples to determine the age of the speaker. We may want to know the age range of the individual so we can get specialized care for people in that age range. Moreover, knowing the age of the individual can give us a better context for the nature of their disease, and how likely they are to have a particular disease.

Data: The name of the dataset is called the Common Voice Dataset by Mozilla. It contains about 200,000 voice samples. It contains the feature we are interested in, which is the age of the speaker. In order to featurize this audio data, we will be using either pyAudioAnalysis or Librosa.

Problem Approach: This is a supervised learning problem. To solve it, we will be using a neural network. We will likely be using a recurrent neural network, but we may choose to use a convolutional neural network instead. It is unclear what variables we will be using as predictors just yet, because we need to featurize the data first. It is certain, however, that the variable we are trying to predict is the age range of the speaker (i.e. twenties, thirties, etc.). The dataset provides both a training and a test set, so we will be using these for their intended purposes.

Deliverables:

- Code for:
    - Data wrangling (acquisition and cleaning)
    - Initial exploratory data analysis
    - Statistical analysis

    - Deep Learning model
- Written report
- Presentation and Slide Deck