Sprint 2 Plan

Understanding Healthcare Data

Sprint Completion Date: 4/26/2020

Revision 1.1, Date: 4/12/2020

Goal: Start generalizing the models to determine whether they can be used on any chronic disease, as well as Dockerizing the pipeline in order to be run at Anthem.

User Stories

User Story 1: *As an insurance worker, I want to be able to assess the risk of patients getting a chronic disease by training and running a general machine learning model.*

Task 1: Generate a dataset of 100k patients (5 points)

- Sift through 100k patients to find only patients with MI and apply a data pipeline to create an image of patient records.

Task 2: Convert the extracted data into a csv format suitable for use in a neural network. (10 points)

- Use the previously created data pipeline and adapt to handle converting data for large amounts of patients.
- Refactor the pipeline to add additional conditions

Task 3: Create and train an RNN and CNN to train on the dataset. (10 points)

Task 4: Use SHAP to conduct significance analysis tests (10 points)

Total for user story: (25 story points)

User Story 2: *As an insurance worker, I want to be able to run these models without much hassle.*

Task 5: Create an initial Docker container (5 points)

- Place a simple script to extract some features from data

Task 6: Test to see if Hummingbird is an option (2 points)

- Check to see if Docker is installed, along with necessary libraries

Total for user story: (25 story points)

Initial Task Assignment:

Cassidy Norfleet - task 1, 2

Brendan Reilly-Langer task 1, 2

Aman Prasad - task 3, 4, 5, 6

Harshitha Arul Murugan - task 3, 4, 5, 6