Predicting Lung Cancer Survival

Eli Parker, Pooja Patel, and Christopher Wilhite

Problem Statement and Background.

The data of our project originates from the National Institute of Cancer Imaging Archives, where real-life patients' CT scans have been taken to image their lungs. The data we are using is a conversion of the CT scans to a binary segmentation mask that is then used to create the numbers that represent the physical numbers. One of our informal success measures is to make sure the results from testing the models make logical sense. This project is beneficial for anyone who is or directly deals with a lung cancer patient. This has a large impact because it can be used to determine the best type of treatment through patterns allowing medical personnel to predict treatment reactions. Better solutions to the treatment of Lung Cancer could lead to a higher quality of life for the patients as well as a longer life expectancy. We are unaware of any related work and do not possess any external knowledge about lung cancer or the models used to predict survival times. However, we are actively researching the topic to further understand the variables used within our dataset.

Data and Exploratory Analysis.

We were given a set of training data and testing datasets. Each of the datasets consisted of clinical data that held patients' personal information related to cancer and radiomics data that held patients' valuable information about cancer that cannot be identified with the naked eye. We have our datasets in two main CSV's: clinical_data and radiomics. The clinical_data CSV had one header row, while the radiomics CSV came with 3, which posed a problem to us that took multiple attempts to fix. To resolve this hurdle, we concatenated the three header rows together to make one. After concatenating, we had to trim some jumble off the end. In order to clean the data, we started by dropping out any N/A's or data rows that were left empty using a python script. We also deleted any rows, using a python script, where Tstage was greater than 4, Mstage was greater than 1, or Nstage was greater than 3. After displaying the clinical data, we found that the Histology had two different types of not otherwise specified data. We also found some cases where the same data was represented in multiple datasets due to capitalization differences. We simply reinserted these values with the same capitalization. To understand the data, we took the now clean data and started by dropping the PatientID column to run summary statistics. Within the summary statistics, a lot of the characteristics go all over the place, so we were able to hand-pick some to correlate with each other. We made a pair plot with the seaborn library to make a sort of matrix of correlations between an x-axis of variables Mstage, Nstage, Tstage, age and a y-axis of variables ...Compactness1, ...Compactness2, ...SurfaceVolumeRatio, and ...ClusterTendency. With these plots, the data seemed to be all over the place, and the –stage variables did not seem to give much insight. The age column in the pair plot gave clusters of points which we thought could lead to a meaningful correlation, so we made scatterplots with a line of best fit on each. Three of these plots had a flat line of best fit unfortunately, but the age vs. cluster tendency plot had a slightly positive slope, which could mean that the older the patient, the larger the cluster tendency could be. After this, we did our value cleaning for the Histology column and ran a histogram of the types of histology we have. The adenocarcinoma and squamous cell carcinoma histology seem to have more representation in our cleaned data. To avoid this bias, we can run statistics based on each type of histology.

Next Steps

We are choosing to start trying out the Cox regression model and a linear regression model for the data. We have an output set to judge whether or not these models are effective, and we can use these two as a simple baseline to see what works.