McCarthy ML Project Proposal

Kelly McCarthy

November 12 2022

1. Background: provide some brief background about the problem you are interested in tackling; why is it important, and to whom?

At the beginning of the semester, when you mentioned that we would conclude the course by being able to conduct a project on a topic of our choosing, I initially thought that I would love to do it on breast cancer data just because of family history. However, since we ended up getting the chance to work with breast cancer data in HW2, I felt that I would branch out and try to tackle a new topic, but I was still interested in investigating something in the medical realm. I came across a great resource, which I believe is where you retrieved the data for our breast cancer data set, the UCI Machine Learning Repository. They have a great collection of publicly available data sets. I went through the list to see what may be of interest to me. I came across the Thoracic Surgery Data Set, which contains information about patient symptoms and conditions prior to lung surgery as well as the 1 year survival period outcome. I would be interested in implementing logistic regression with L2 regularization. I believe this problem is a good candidate because the data set deals with predominantly binary outcomes (Y/N) for its 17 attributes. I could leverage these binary outcomes in order to make some prediction schema for the likelihood that someone survives 1 year post-surgery.

The reason I think that this problem is important is because of the relevance it presents for today's world, particularly the development of lung conditions in adolescents due to vaping. It appears that there was a huge push decades ago for an anti-smoking cigarettes campaign but no one foresaw its resurgence today. False narratives and manipulative advertising makes teens think vaping is a preferable option rather than cigarette smoking, or, they are aware of the risks, yet still suffer from the addictive quality of the product. Attribute 2: "Forced vital capacity," or the amount you are able to fully exhale after inhaling, would be an interesting statistic to compare for different age groups. Considering this data was collected between 2007-2011, I recognize that it does not reflect current or recent trends (within the past 5-8 years) of the rise of teen vaping, but I do hope these findings could be used to compare to results down the line. I would be curious to see if we see Attribute 16: "Age at surgery" decrease because more and more young people develop lung cancer.

2. Hypotheses and goals:

A lot of the articles I found online seemed to suggest that there are not yet robust or accurate models for assessing the prediction of patient survival for post-surgery or post-diagnosis. I would say that given the limited number of attributes in this data set, machine learning can help improve current practices by learning what symptoms are most predictive of a decreased likelihood of survival 1 year post-surgery. My hypothesis is that True values for Attribute 14: Smoking and values of OC13 and OC14 for Attribute 10: Size of original tumor, will result in the largest number of deaths 1-year post surgery. The reason for this is because tumor sizes OC13 and OC14 are on the larger end of the spectrum for tumor size. Once we can assess what attributes are most predictive, we can develop strategies to figure out how best to identify these symptoms and use them in conjunction with advancements in screening technologies in order to diagnose lung cancer and implement interventions earlier (https://www.nature.com/articles/s41598-020-67378-8).

I would say one of the limitations is the number of potential compounding factors that could affect patient outcome. Someone could have never smoked a day in their live but be suffering from Myocardial Infarction (MI - which is the lack of blood flow to the heart which can damage heart muscle), Peripheral arterial disease (PAD) (the narrowing/blockage of vessels that carry blood from heart to legs), genetic predisposition to small cell lung cancer, among many other factors.
See Radiogenomics analysis section of: https://www.nature.com/articles/s41598-020-67378-8

3. Sources of data:

I plan to use the Thoracic Surgery Data Set from the UCI website:
https://www.archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

I searched online for more publicly available datasets similar to this one but the closest thing I found were aggregated publications about "outcome data" but nothing that actually provided me the raw data in csv or alternative file formats.

4. Sources of compilation:

Just briefly looking through the data already, one problem I foresee is diversity of age representation. It appears most people in the data set are around ages 50-75 (plus or minus a few years). This is likely just due to the current state or average age that lung cancer develops in adults. However, it does restrict my ability to perform any analysis or draw conclusions across different age groups. I do not foresee missing data as being an issue because the description stated that there were no missing values from the set. There could be selection bias present. Since this data set derives from a university in Poland, I believe all patients in the data set are from the National Lung Cancer Registry in Poland. Although it may be including too many compounding factors for this project alone, it would be interesting to contrast how our physical environments, rates of smoking, genetic predisposition, or other factors differ among populations in Poland as opposed to other countries. We cannot know from this data set alone, but it is important to remember that the findings from this resource are not necessarily representative of global trends in post-surgical results for lung cancer patients.

5. Rough plan of analysis:

As mentioned earlier, I am interested in applying Logistic Regression with L2 Regularization since we are working to predict a binary outcome. I would be interested to further explore more recent topics such as decision trees or random forest algorithms. As stated in the previous question, I do not think that I can overcome the issue of selection bias in this project alone since that would require publicly available data sets from multiple other countries, derived around the same time period as this one from Poland. However, I do still think we can derive interesting results and explore the possibilities of applying machine learning methods to the Thoracic Surgery Data Set.