Peter Schmeiser Capstone Proposal II

Instandence rate of community acquired pneumonia in adults

For this capstone I would like to once again dive into the healthcare side of things. Fun fact, other than having babies pneumonia is the most common reason people check into a hospital. Sadly, in a lot of poorer areas of the world, such as South America, the mortality rate can be extremely high. Therefore finding some model that can even remotely predict infection and/or mortality rate would be extremely beneficial.

Question:

Can I create a model to successfully predict infection and or mortality rate for community acquired pneumonia in adults?

What are the data and do you have them, and have you looked at them?
The data is provided by Dryad Digital Repository and consists of 170+ metrics and entries for 2300 patients. The data will require a relatively large amount of cleaning to reduce nan's and to select from the 170+ metrics which are to be used. I plan on using my previous experience working in a pulmonary lab to select plausible metrics to use for modeling.

What is the MVP (minimum viable product)?

MVP: A linear or logistic regression model that successfully predicts infection or mortality

MVP+: A random forest of gradient boosted model that successfully predicts infection or mortality

Link to data:

https://datadryad.org/resource/doi:10.5061/dryad.r282vk6