Team D
Brenda Li

Methods:

My Kaggle competition required me to predict five soil properties given 3578 mid-infrared absorbance measurements and 16 spatial variables, and so for each outcome variable, I used LASSO incorporating all the provided predictor variables. For each variable, I experimented with a wide interval of lambda values at first and then narrowed it down to smaller intervals of lambda to search within those intervals for the optimal lambda value. I could have theoretically done this all in one run of CV but my computer did not cooperate when I tried too many lambda values. Thus, what I've included in my final code is my final rounds of CV for each variable, where I'm searching for the optimal lambda value within a smaller interval I've already narrowed it down to.

Other models I tried previously were simple linear regression (just to test the submission process) and ridge regression.

Issues:

The Kaggle data description states that the soil data was sampled by zones stratified by the climate zones of Africa and it acknowledged that there is certainly geographical clustering in the data. Even though the soil samples were geo-tagged upon collection, the organizers of the competition state that they do not wish to provide georeference information in this competition. While this potential issue is not something I can address with the material covered in this class, I did some quick research into the spatial variables provided to determine which variables could theoretically be used as proxies for location and I believe that the best variables to do so would be elevation and mean annual precipitation. If we could find data on the mean annual precipitation across different zones in Africa, we could incorporate that with data on the elevation of each region in Africa and approximate where each sample was taken. Using those location estimates, we could use machine learning methods that account for spatial autocorrelation to design our model.

Furthermore, the description of the data also states that the test data was not sampled from the same regions as the training data, which suggests that the effect of the issue of spatial autocorrelation would be even more prominent in the performance of my model. The latter two histograms in the EDA section of my final code confirm the difference in the source regions of where the data was sampled from for each dataset by comparing the distributions of the proxy location variables mentioned above.

Since the tools to address these issues are outside the scope of this course, my model operated under the (faulty) assumption that each data point is independent in the datasets and that the datasets are representative of the same population. That meant that my model would ultimately be overfit to the training data and that the CV process would overestimate the performance of my model

Extras:

I also conducted additional analysis of the final models that I ended up using, and this analysis can be found in the last section of the final project code.