Dear Students and Mentors

I have the following ideas for my first capstone projects:

Please provide feedback regarding:

a) Is this project useful to demonstrate data science skills?

b) Is this project tackling an interesging question?

Project 1: Classification of three types of Pancreatic Cancer using a data-driven approach for MRI

Source: Cardenas Laboratory at University of Arizona, to be posted at OSF

Data Science Task:Supervised Classification

Description: The treatment of pancreatic cancer depends on the phenotype of the cancer, but it is very hard to get biological information without a biopsy. Furthermore, data-driven approaches are not commonly employed in imaging because of the high dimensionality of the data.

Goal: Create a model to predictive a tumor’s biological phenotype based on its imaging phenotypes.

Lots of data cleaning needed?: Yes

Project 2: Predict Low Grade Lenders who will NOT default at Lending Club,

Source: Lending Club and Kaggle, at

https://www.lendingclub.com/info/download-data.action

https://www.kaggle.com/wendykan/lending-club-loan-data

Data Science Task: Supervised Classification

Description: Lending Club (LC) publishes data for lenders that have been backed up by investors. Each lender is graded by LC and is also described by ~100 attributes. The interest rate is inversely correlated with the lender's grade. Thus, if you want to make more money, you need to invest in Lenders that are more likely to default.

Goal: Create a model to which lenders with grade <C will default.

Lots of data cleaning needed?: Yes

Project 3: Use multiple datasets from Medicare and Medicaid to predict hospital readmissions

Source: https://data.medicare.gov/data/hospital-compare

Data Science Task: Supervised Classification

The Centers for Medicare & Medicaid Services (CMS) created the Hospital Compare website to better inform health care consumers about a hospital’s quality of care. Hospital Compare provides data on over 4,000 Medicare-certified hospitals, including acute care hospitals, critical access hospitals (CAHs), children’s hospitals, and hospital outpatient departments. The data dictionary is about 80 pages long, but it contains pure gold.

Lots of data cleaning needed?: Yes (some)

Lots of data wrangling to combine and code categorical features?: Yes