Capstone Three Project Proposal

PlantTraits Regression from Images+Ancillary Data | Kaggle Competition

Github Repository (to be populated)

Problem Statement (competition)

Predict a broad set of plant traits (6) from crowd-sourced plant images and ancillary data (based on images' geotags).

Problem Statement (project)

Predict plant traits from images and data. Measure performance with subset of competition training data. Achieve reasonable "baseline" accuracy with preliminary work and then improve accuracy and/or decrease model complexity/training time with subsequent iterations.

Provide analysis about model performance and limitations of dataset and target trait predictions. Identify limitations of certain images or geodata characteristics and nominate automatic identification and cleaning processes. Determine effective outlier detection, data pre-processing, and model evaluation methods.

Measure and report on utility of individual/combined data streams, model architectures, and prediction strategies:

- Geodata | feature selection, feature decomposition, clustering?
    - Plant photograph, Soil, Sun, Temperature, Precipitation, Satellite, Radar
- Images
    - Image augmentation, downsampling, segmentation
- Multi-output vs single-target regression

Context

- This project will expand on work based on a recent paper and is hosted by the Fine Grained Visual Categorization Workshop. Paper: "Deep learning and citizen science enable automated plant trait predictions from photographs"

- Plants can serve as biomarkers for our ecosystems' health, relative to climate change and other factors. Measurement of plant traits from citizen sourced data may enable new understandings of our planet's health.
- "These plant traits, although available for each image, may not yield exceptionally high accuracies due to the inherent heterogeneity of citizen science data. The various plant traits describe chemical tissue properties that are loosely related to the visible appearance of plants in images. Despite the anticipated moderate accuracies, the overarching goal is to explore the potential of this approach and gain insights into global changes affecting ecosystems. Your contribution to uncovering the wealth of data and the distribution of plant traits worldwide is invaluable."

Data sources **|** competition data

The competition provides ~55k training images (512x512x3), provided from iNaturalist. Associated with each image is ancillary "geodata" provided from a variety of sources. The geodata contains 175 features in the broad categories of Climate, Soil, Satellite, and Radar. The image and geodata should be used to predict six continuous plant traits (stem specific density, leaf area per leaf mass, plant height, seed dry mass, leaf nitrogen per area, leaf area). The target plant traits are provided from the TRY database, and are mean values for a species. Some entries contain standard deviation information, as well.

The test set for the competition will not be used for this project. It contains almost 14k additional images and associated geodata. Previous FGVC (competition host) competitions and associated literature may be referenced.

Success Criteria

The competition uses the mean R2 of the 6 prediction errors to score competitors. I will take care to identify a meaningful model metric that will be robust to error for extreme values. Additionally, model complexity and training time will be considered. After extensive EDA and outlier treatment and as the evaluation metric(s) are chosen, baseline performance will be established from the following:

- "Baseline 0"
  - Simple regression models for geodata data only
  - Simple CNN for image data only
- "Baseline 1"

    - Combined image and geodata model, loosely follow method used in paper

Next, I will try to change various parameters to reduce error and also computational cost. Any meaningful insights into model building for this specific application can be useful. The goal of this project is not to achieve the highest accuracy possible, but to “explore the potential of this approach and gain insights into global changes affecting our ecosystems.”

Solution space scope, Deliverables

Initially I will focus on data and modeling insights that lead to improved prediction accuracy. While developing a performant model is not the main goal, the steps taken to improve performance should be generally applicable to the citizen science / image processing effort.

As performance starts to plateau, I will look for ways to streamline training and analysis. For example, what data augmentation may help and what is not necessary? Can images be downsampled further? Are complex architectures required? How should the model handle separate data streams and how should it predict multiple outputs?

All of these findings will be summarized into two deliverables: a written report and a summary slide deck.

Solution space constraints

The competition hosts mention that the data is inherently heterogeneous. I believe they are referring to the training set of photographs provided. The paper referenced by the competition spent extensive efforts in curating their training and testing set of photographs, and the same efforts were probably not employed for this competition.

Additionally, the target traits are provided from the TRY database and are simply the means for a species (traits were matched to photographs with a specie’s name). So target prediction is not necessarily specific to the image.

All of these may limit the accuracy of regression models. Hopefully I will be able to see some improvements early in my process. The competition discourages specie’s identification, but as my work will not be used for submission, this could be a final aid to achieve solutions.

Stakeholders *adapted from competition*

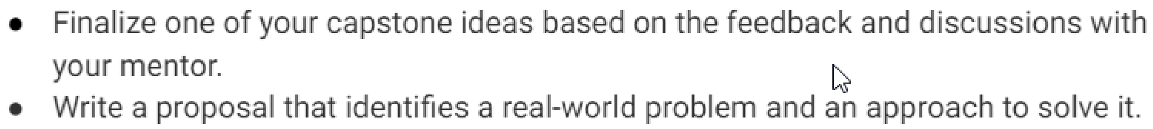
The primary stakeholders will be the scientists and their associated institutions. The concept exemplified in the paper is being expanded to more traits and a broader utility. Continuing to prove this concept will help expand the efforts of measuring our ecological health in relation to climate change. More specific implications may come from more accurate models' applications and insights. The scientists will be able to see how an area's fauna changes over time. Lastly, continued understanding, interest, and development of these models should help funding efforts from associated institutions.

Another stakeholder could be iNaturalist. From their website, they describe themselves:

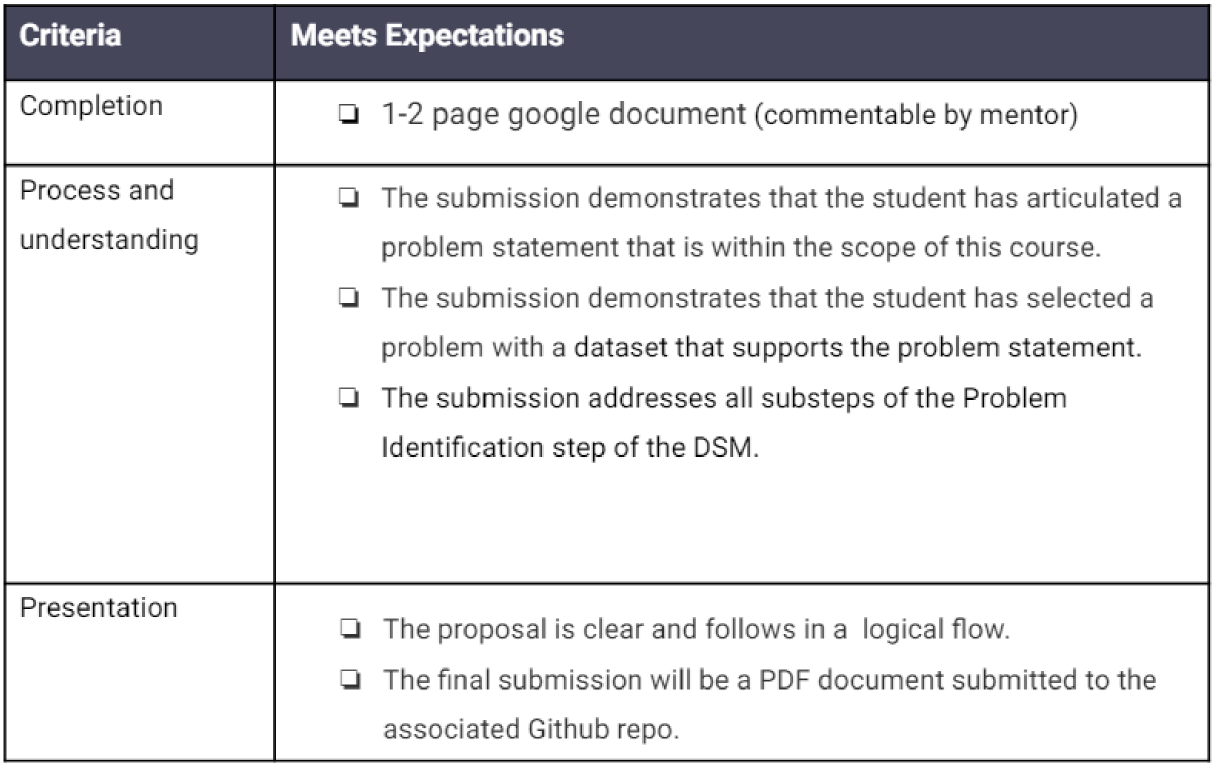
"iNaturalist is an online social network of people sharing biodiversity information to help each other learn about nature".

Their data is driven by its users, and maybe more projects like these will encourage financial support and investment from larger groups. Similarly the Plant Trait Database (TRY) may receive favorable attention and gain data insights from the project's efforts.

Rubric **S**pecific, **M**easurable, **A**ction-oriented, **R**elevant, **T**imebound



- Finalize one of your capstone ideas based on the feedback and discussions with your mentor.
- Write a proposal that identifies a real-world problem and an approach to solve it.



Criteria	Meets Expectations
Completion	□ 1-2 page google document (commentable by mentor)
Process and understanding	□ The submission demonstrates that the student has articulated a problem statement that is within the scope of this course. □ The submission demonstrates that the student has selected a problem with a dataset that supports the problem statement. □ The submission addresses all substeps of the Problem Identification step of the DSM.
Presentation	□ The proposal is clear and follows in a logical flow. □ The final submission will be a PDF document submitted to the associated Github repo.