Capstone Three Report *[repository]*
Plant Traits Regression from Images and Ancillary Geodata

Background

This project is adapted from a Kaggle competition, hosted by the Fine-Grained Visual Categorization (11) workshop at the IEEE/CVF Computer Vision and Pattern Recognition Conference (2024). The Kaggle competition has recently concluded, in June 2024.

Plant life provides insight into our ecosystems, particularly into potential challenges brought by climate change. The competition organizers have provided thousands of plant images from "citizen scientists" using iNaturalist. Location information from the images was used to provide ancillary data describing the area's geography, "geodata". The hosts are expanding on previous studies; and one in particular was followed as a guide for this project, "*Deep learning and citizen science enable automated plant trait predictions from photographs.*" [see Bibliography]

Problem Statement

The problem statement is slightly adapted from the competition. Plant images, associated geodata, and plant traits will be used to train a model to predict the six plant traits (multi-output regression). For this project, only the training data from the competition will be used. As the plant traits are not directly apparent or related to the images, high accuracy will not be expected. However, the final model should have some success in predicting traits and will be considered for complexity and training time.

As will be discussed in the following section, model performance will rely heavily on data cleaning and outlier detection. A secondary goal of this project is to establish processes for data curation that can be improved by subject matter experts.

Data

Summary

Plant data was collected from a variety of sources. First, images were provided by iNaturalist, "a social network for sharing biodiversity information to help each other learn about nature" (app store description). All of the images were 512x512x3. Images were selected that contained GPS coordinates, and geodata was provided for the general area of each image ("derived from globally available raster data"). The geodata consisted of 163 features, describing soil, sun, temperature, precipitation, and water content. There were four measurement groups:

- **WORLDCLIM |** "Climate" – temperature and precipitation
- **SOIL |** "Soil" – various soil properties at various depths
- **MODIS |** "Satellite" – optical reflectance (sunlight) at various frequencies
- **VOD |** "Radar" – water content and plant biomass

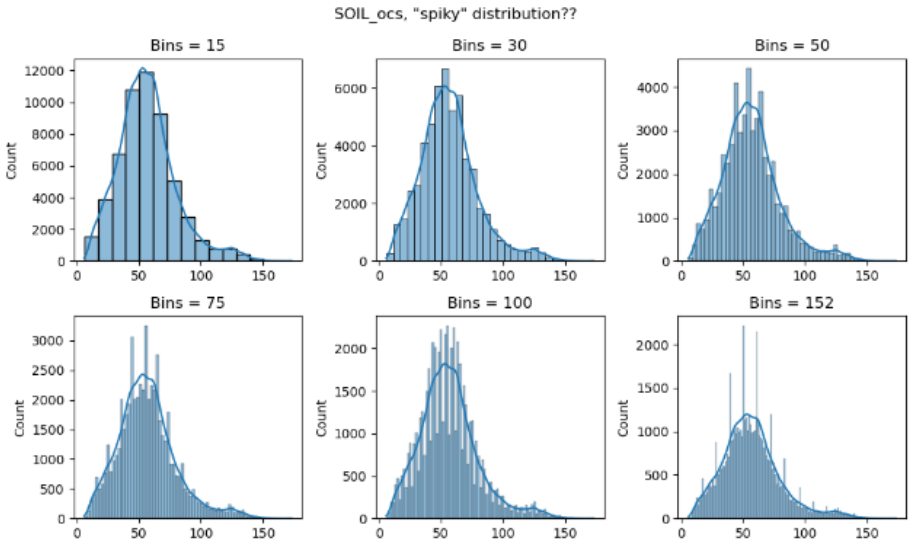
These groupings were maintained for future decomposition. Lastly, six plant traits were provided as targets for regression, from the TRY database. Briefly, they describe:

1. Stem specific density *X4_mean*
2. Leaf area per leaf mass *X11_mean*
3. Plant height *X18_mean*
4. Plant seed dry mass *X26_mean*
5. Leaf nitrogen per area *X50_mean*
6. Leaf area *X3112_mean*

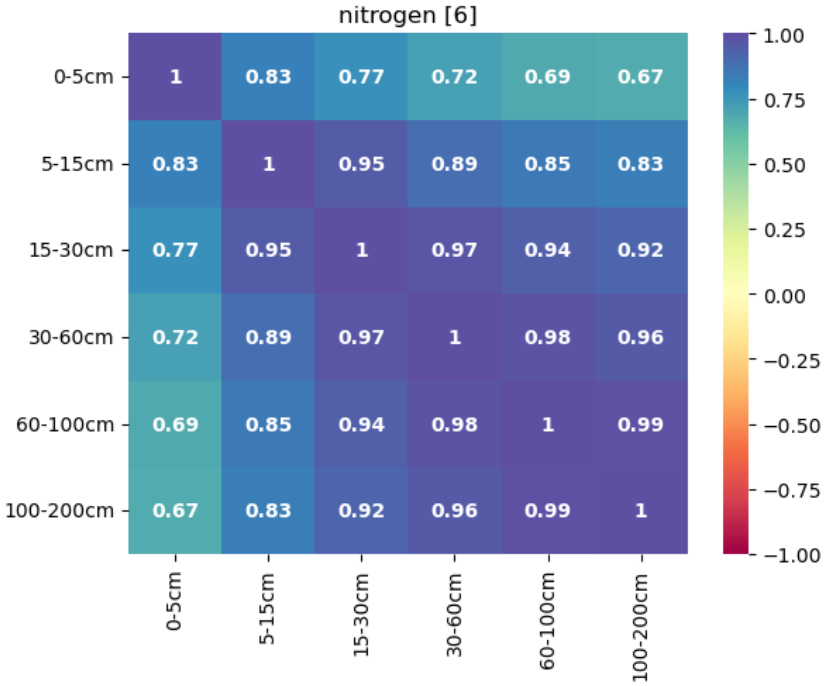
Based simply on the names, it should be expected for some of these values to be related to one another. The provided data was not checked for bogus target values, and therefore extensive cleaning of the training data is required.

Feature Exploration

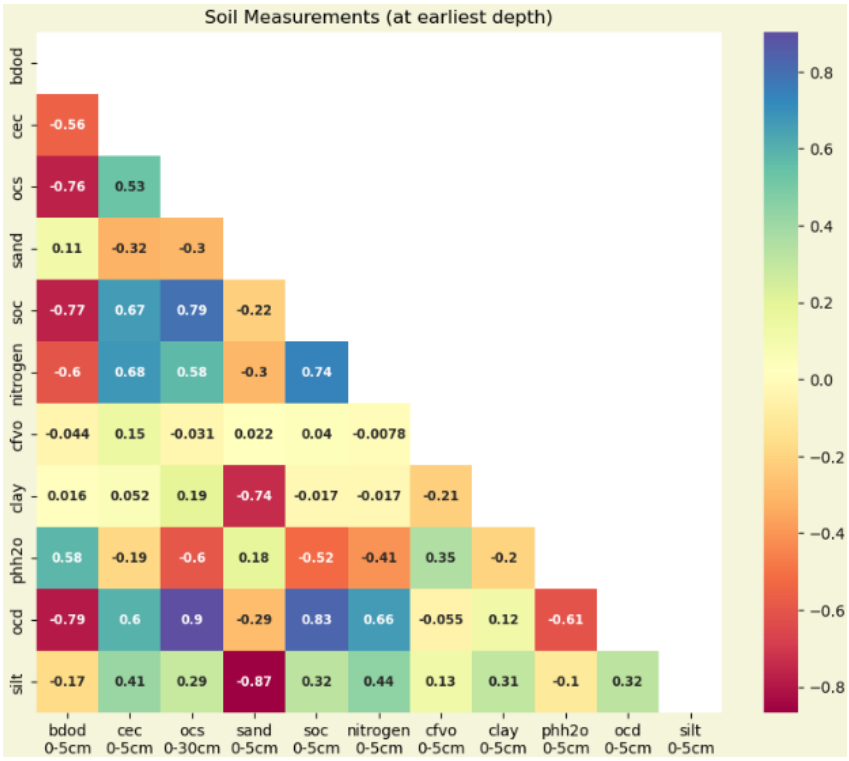
Each geodata group was explored independently, and finally the various measurements were analyzed together for intercorrelations. Distributions and outliers were checked, as well as correlation heat maps. Soil features will be discussed in this section as an example. The soil group contained 61 features, 11 different measurements at various depths: silt, soc (soil organic carbon), sand, cec (cation exchange capacity), clay, bdod (bulk density), ocs (organic carbon stock), phh20 (water acidity), nitrogen, cvfo (?), and ocd (organic carbon density). Some feature's histograms showed "spiky" shapes, indicating that rounded measurements may be more common than specific measurements.



Measurement values at various depths were closely and directly related to one another, and had stronger correlations at closer depths. The correlation heatmap below shows correlation coefficients for various depths of soil nitrogen.



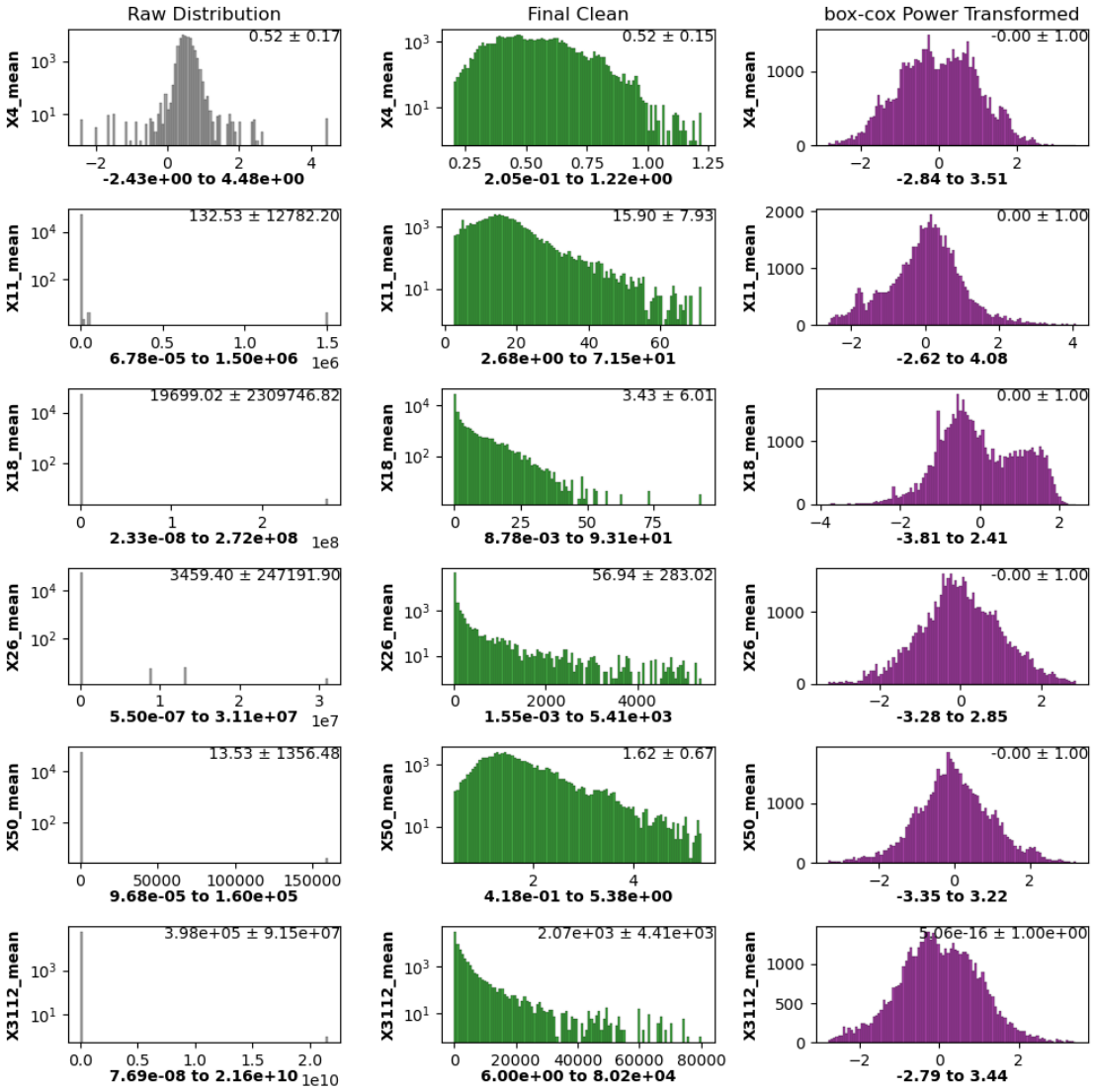
Some of the measurements had strong positive or negative correlations with each other. The next heatmap shows the correlation coefficients of different soil measurement at the shallowest depth (0-5cm).



Plant Trait Target Outlier Cleaning

The first step to data preparation was detecting and removing target outliers. The geodata features' relationships to the target traits should not be assessed with faulty data present. Outlier detection was performed without a great degree of domain knowledge. Distributions of each target trait were assessed, then associated plant images were examined for the most extreme values. In this way, upper-bound cutoffs were established for each trait. Lower-bounds were not explored, however negative values were dropped for Stem specific density. The images below provide an example for how a cutoff for Plant height was established:

Above each plant image is its ID and value for Plant height (meters?), and they are presented in order of increasing height (left to right, top to bottom). The first two rows of images look reasonable, until ***196144622***, with ***157.7*** for its height. Based on these images, a cutoff of **150** was established for Plant height. Take note how some images are not representative of the entire plant, and may present issues for models.



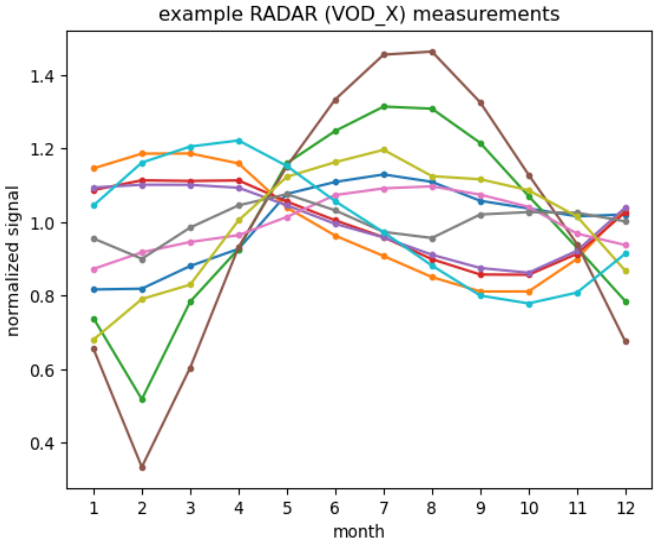
After establishing cutoffs and removing samples, additional cleaning was performed based on sample deviation. The referenced paper's method was used: log10-transform the target trait and remove samples greater than three standard deviations away from the mean. Lastly, the targets were power-transformed to achieve more normal distributions. The histograms above for each plant trait target, show distributions before and after cleaning and power-transformation (box-cox).

Feature Engineering

Two strategies were employed to reduce the amount of geodata features. The first was to replace the monthly means for satellite and radar measurements with "seasonal" features. The graph below shows a few random samples monthly means for a radar feature.

Following the definitions for the climate features (see detailed reference), three engineered features replaced the 12 monthly means for each measurement:

- *annual mean*: average of all monthly measurements
- *delta*: difference between maximum and minimum measurements
- *seasonality*: monthly standard deviation divided by the mean, expressed as a percentage



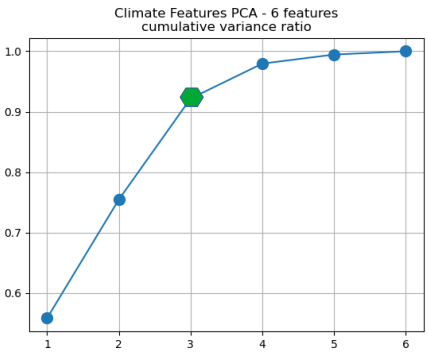
The three radar measurements were reduced from 36 to 9 features, and the five satellite measurements were reduced from 60 to 15 features. The engineered features typically had higher correlations with the plant trait targets than the monthly means.

Feature Normalization and Selection/Decompostion

Principal component analysis was employed for the second step of dimensionality reduction. All features were normalized by scaling to unit variance prior to PCA. Both CatBoost and ElasticNet multi-target regression models showed better performance using PCA components than with various feature selection methods (Recursive, Sequential, select from model weights). Additionally two methods of PCA decomposition were tested: either considering all the geodata features together or considering each feature group separately. Better performance was found when preserving feature groups.

Feature Group	**Original Features**	**PCA components**
Climate	6	3
Soil	61	10
Radar	36 → 9 *engineered features*	3
Satellite	60 → 15 *engineered features*	5

The number of PCA components for each feature group was determined to capture at least 90% of the original feature variation and for the original features to have relatively equal representation. For the soil, radar, and satellite groups, it was natural to let the number of PCA components be the same as the number of different measurements. Below is a summary for the Climate features.



	PCA_1	PCA_2	PCA_3
BIO1	0.856	0.265	-0.609
BIO12	0.790	-0.476	0.856
BIO13.BIO14	0.846	0.166	1.000
BIO15	0.348	1.000	0.232
BIO4	-1.000	0.081	0.621
BIO7	-0.995	0.260	0.463

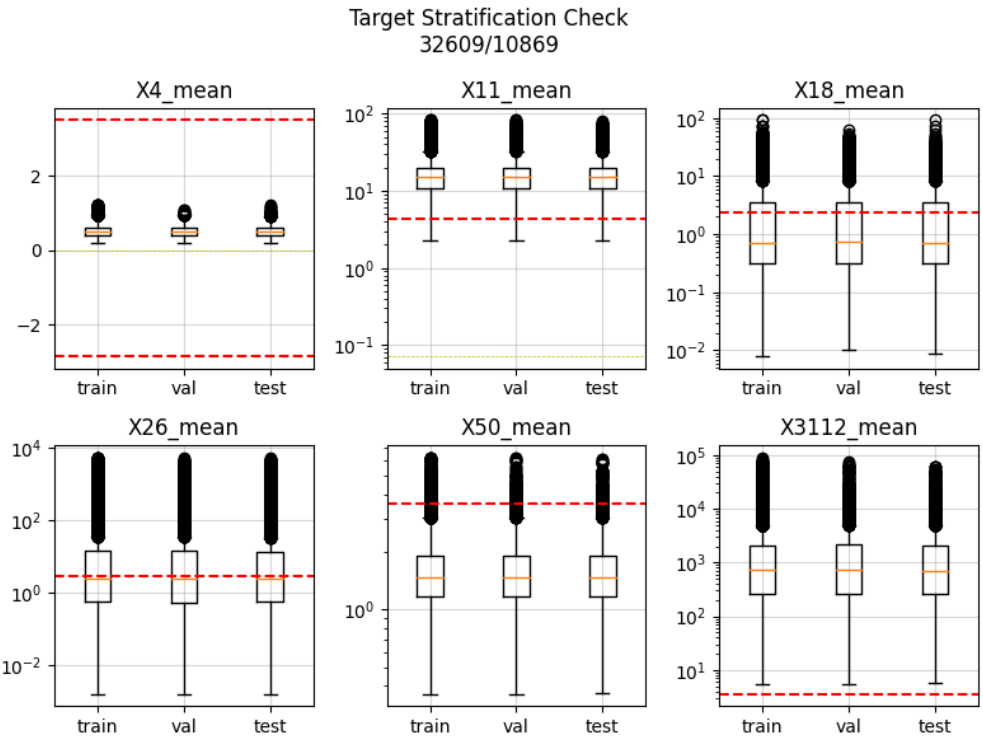
	Total (R2)
PCA_1	0.506617
PCA_3	0.088865
PCA_2	0.023678

In summary, the 163 geodata features were reduced to 21 PCA components. This helps reduce model training time and complexity.

Preprocessing Summary

Training, testing, and validation splits were 6:2:2, resulting in 32,609 samples for training and 10,869 samples for validation and hold-out testing. Stratification was not imposed, but checked after the fact.

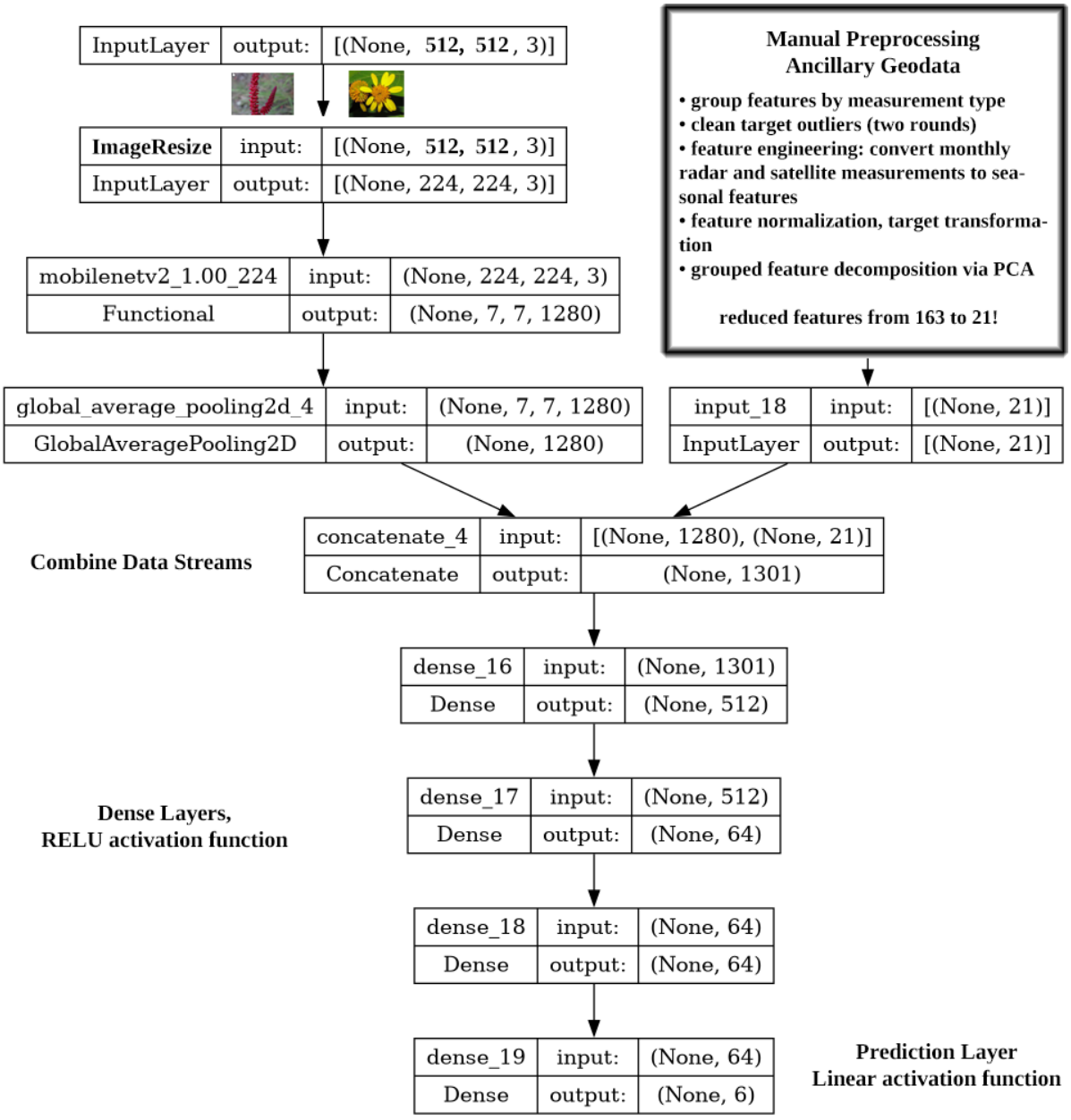
1. Remove samples with negative targets. Outlier detection and cleaning of plant trait targets. Remove based on visual assessment upper-bound cutoffs, then based on deviance. (*~2% samples removed*)
2. Engineer seasonal features for radar and satellite measurements, drop originals.
3. Split training and hold-out testing sets. Feature normalization, decomposition, and target transformation will only be fit to the training set, then applied to both sets.
4. Normalize features to unit variance (set mean to 0, standard deviation to 1).
5. Perform feature-group specific PCA decomposition.
6. Power-transform plant trait targets.
7. Split training and validation sets for model training.



Modeling

As mentioned, models were trained using the geodata only to determine ideal preprocessing conditions. These models were much faster to train and test, but not nearly as accurate as the CNN models developed to also process the plant images. Popular CNN architectures developed in part for ImageNet were tested: VGG16, VGG19, InceptionV3, InceptionResNetV2, MobileNetV2, and various EfficientNet (B0-B5) models. All were available through Tensorflow / Keras.

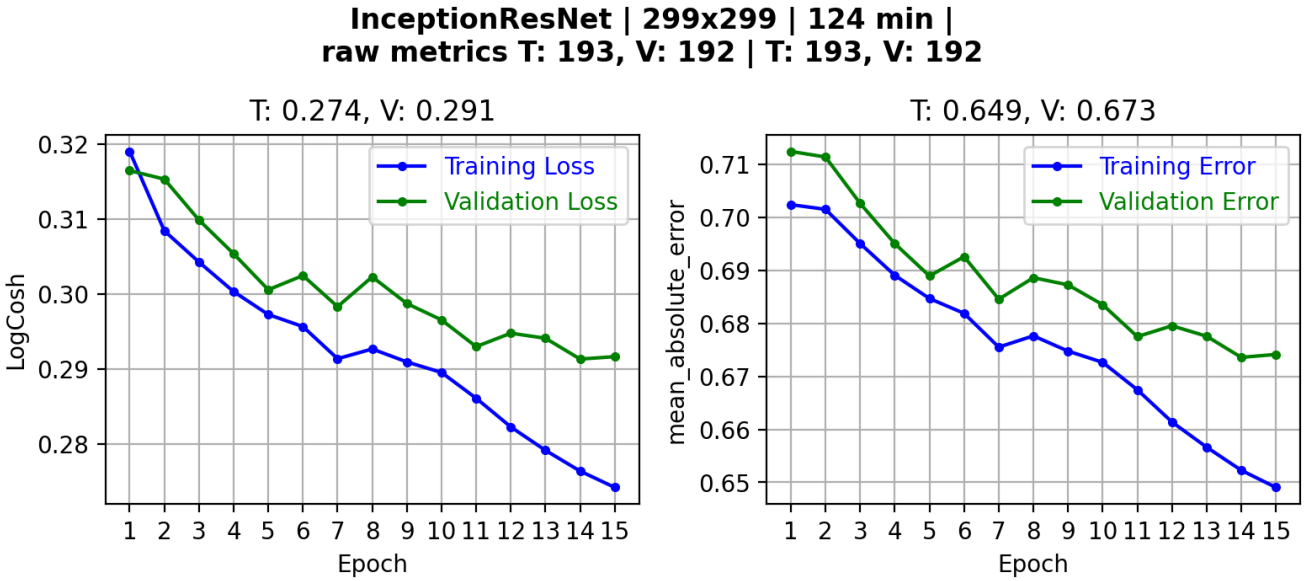
The following schematic shows how these models were employed to extract features from the plant images, and how those features were combined with the geodata to generate plant trait predictions.

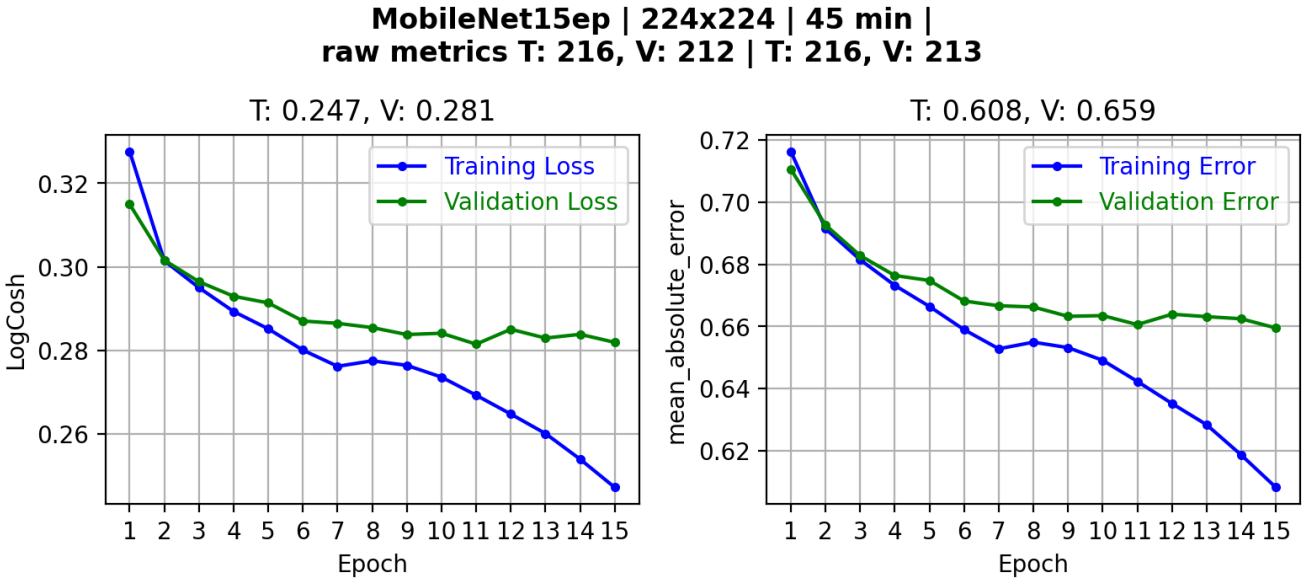


The first set of experiments compared the transferred model architectures. Plant trait regression performance did not correlate with ImageNet performance. The best two models used MobileNetV2 or

InceptionResNetV2 architectures. The latter model was used by the reference paper. Additional experiments were run with these two models. Not all hyperparameters were explored. Model refinements are discussed later.

Performance was better when fully retraining model weights, instead of simply transferring pre-trained ImageNet weights or training over them. Surprisingly, the training time for fully retraining model weights was as fast or faster than simple transfer learning. Employing a learnable image resize layer also increased accuracy, instead of resizing images during dataset generation. Images were resized to each model's default parameters. Image augmentation also improved model performance, random brightness, contrast, orientation, and rotations were applied after images were resized. Batch sizes were chosen to maximize throughput with the available resources. Initial learning rates and learning rate schedules were determined by the batch size.





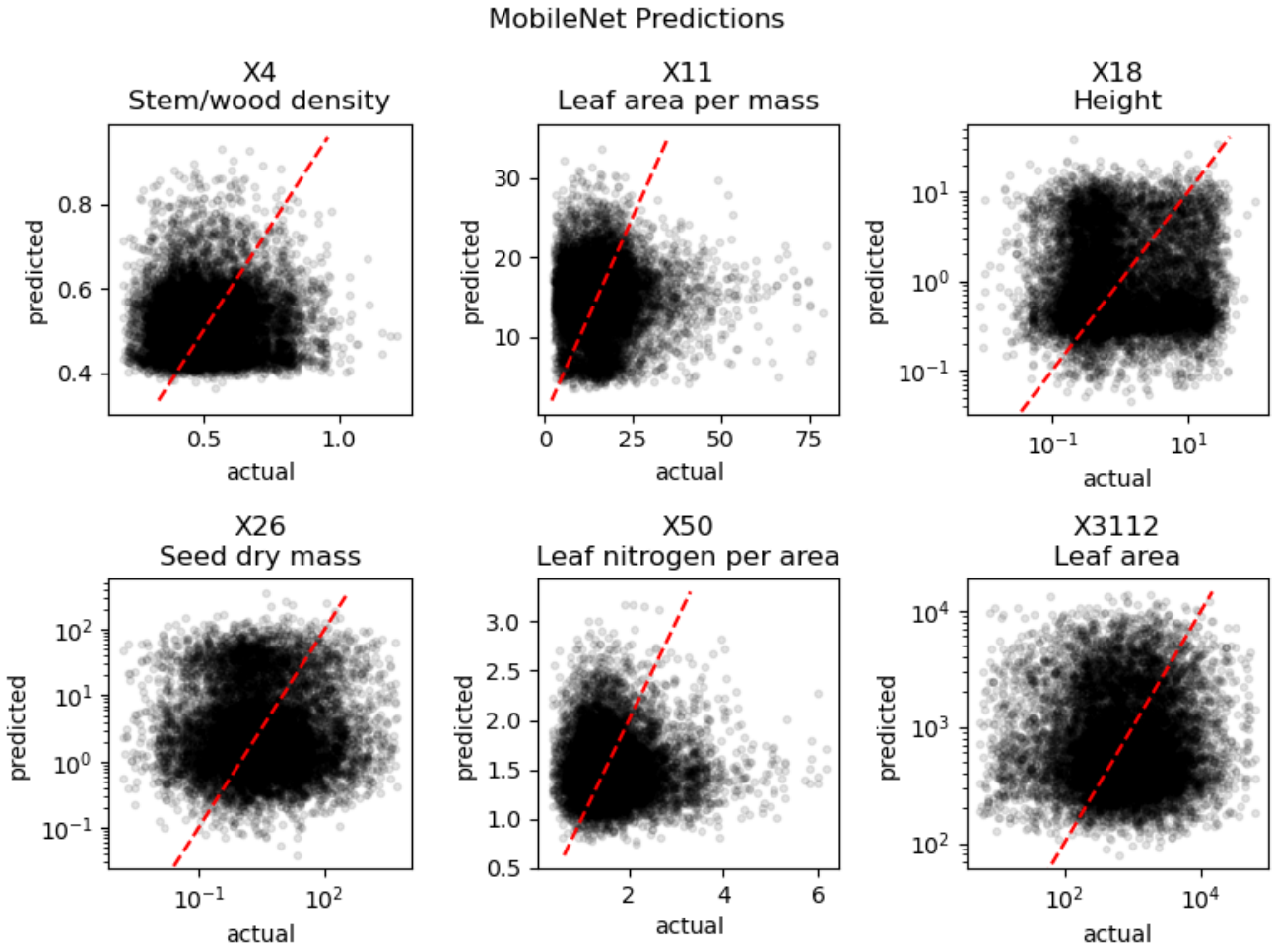
CNN training was done using Kaggle's virtual notebooks to take advantage of their GPU acceleration. Models were trained for at least 10 epochs during early experiments, and for at least 15 epochs for final testing. Mean absolute error was chosen for the model metric, and LogCosh error was chosen for the loss function. The latter was critical to use in early development, particularly if target transformations were not employed. Otherwise extreme target values over-influenced results.

Final Results

Model results summary, error is for hold-out testing set

Model	Image Dimensions, Batch size	Time (min)	Mean Absolute Error
CatBoost (geodata only)	N/A	**0.65**	263
InceptionResNetV2	299x299, 64	124	**192**
MobileNetV2	224x224, 256	45	212

The model using InceptionResNetV2 architecture achieved the best results, but the model using MobileNetV2 architecture was chosen as the final model. Given the early development stage of this project, a model that can be trained faster and that is less complex is more appropriate. Building and iterating on this work will be much faster with the MobileNetV2 architecture and will be viable on a wider variety of workstations. The following predictions are from the MobileNetV2 model.



Model error appears to be randomly distributed, but looks like it under predictions may be more common. Plant height may show some patterns in its residuals and will be explored in more detail below. Images were investigated for the best and worst predictions for each model (see more).

The top row of images shows the largest over-predictions for Plant height, and the bottom row shows the biggest under-predictions. The bottom right image looks to be a single stem on the ground, but is representing one of the tallest plants in the cleaned dataset – *see cutoff discussion above*. The image is not necessarily representative of the plant at large, and perhaps a model should weigh the geodata factors more heavily for these images. The next two under-predictions contain hands with plant parts. Should images that are not primarily plants be cleaned in future work?

The next set of images shows the images associated with the model's best predictions for Leaf area. Recall that Leaf area had the greatest range of the plant traits, spanning almost 5 orders of magnitude even after outlier cleaning. The model seems to predict this trait well at a variety of magnitudes, hopefully highlighting the benefit of data cleaning and preprocessing. Interestingly, two of the best predictions were images containing hands. Are the hands less likely to affect Leaf area prediction?

Model Improvements, Implications

Data cleaning and feature processing methodologies were established in this project, and they ultimately contributed to a CNN based model with reasonable accuracy in predicting six plant traits. As discussed in the Kaggle competition, this is an important step to utilizing existing "data treasures" to study and understand our ecosystem in real time. Plants are important biomarkers for various ecosystems, meaning they be used as a proxy to measure broader environmental health and conditions.

The outlier detection and cleaning for this project was somewhat rudimentary. The process should be easily improved by subject matter experts with a basic understanding of the plant traits and their implications. For example, measurement units were not even identified for all plant traits. Additionally, the reference paper found benefit to "plasticizing" target traits. The TRY database provided mean values for the traits, and for some samples provided a standard deviation. Transforming target traits to live somewhere within this distribution, instead of simply being the mean, could improve model performance and should make predictions more realistic to individual samples.

Residual analysis indicated that certain types of images may have undue influence on trait predictions. Model predictions and their associated images can be studied in more detail, and perhaps an image cleaning step can be employed prior to model training. This step may be more critical as larger datasets are used. The competition provided ~55 thousand training images, but mentioned at least 20 million photographs were available from their sources. If an image is deemed insufficient, a model could lean more heavily on the ancillary geodata.

Some Kaggle competitors found benefit to utilizing the relationships between plant traits and certain plant trait / geodata feature relationships. This project relied on multi-output regression and hoped the model would learn and capture the appropriate relationships. But user guidance in model prediction should help, as the target traits are closely related to one another.

Given long training times, model hyperparameters were not tuned extensively. Choices of optimizer, metric and loss functions, learning rate schedules, feature averaging techniques, dropout layers, and activation functions for various layers could be tested in future experiments. Lastly, this project only focused on adapting CNN architectures for deep learning models. Other approaches could be explored, particularly transformer models.

Bibliography

"Citizen Science" papers

Schiller, C., Schmidtlein, S., Boonman, C., Moreno-Martínez, A., & Kattenborn, T. (2021). Deep learning and citizen science enable automated plant trait predictions from photographs. Scientific Reports, 11(1), 16395.
https://www.nature.com/articles/s41598-021-95616-0

Wolf, S., Mahecha, M. D., Sabatini, F. M., Wirth, C., Bruelheide, H., Kattge, J., ... & Kattenborn, T. (2022). Citizen science plant observations encode global trait patterns. Nature Ecology & Evolution, 1-10.
https://www.nature.com/articles/s41559-022-01904-x

Moles, A.T., Xirocostas, Z.A. Statistical power from the people. Nat Ecol Evol 6, 1802–1803 (2022).
https://www.nature.com/articles/s41559-022-01902-z

Kaggle Competition

Awsaf, AyushiSharma, HCL-Jevster, inversion, Martin Görner, Teja Kattenborn. (2024). PlantTraits2024 - FGVC11. Kaggle.
https://kaggle.com/competitions/planttraits2024

CNN Architectures

Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., & Chen, L. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510-4520.
https://arxiv.org/pdf/1801.04381