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

Plant Traits Regression – Enabling Citizen Science

adapted from Kaggle

Project Repository

Presentation | Report

Contents

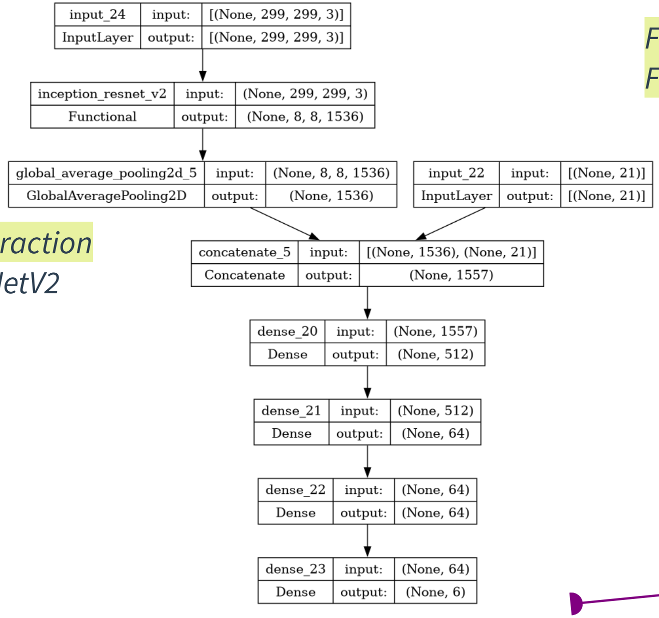
Project Summary

predict 6 plant traits from an image and geodata associated with its location

194801088.jpeg

Plant Image

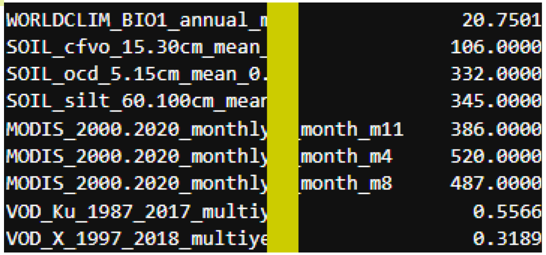
Feature Extraction
Ex: MobileNetV2



Multi-output Regression Model

Feature Engineering,
Feature Decomposition

Ancillary Geodata



WORLDCLIM_BIO1_annual_m		20.7501
SOIL_cfvo_15.30cm_mean_		106.0000
SOIL_ocd_5.15cm_mean_0.		332.0000
SOIL_silt_60.100cm_mear		345.0000
MODIS_2000.2020_monthly	month_m11	386.0000
MODIS_2000.2020_monthly	month_m4	520.0000
MODIS_2000.2020_monthly	month_m8	487.0000
VOD_Ku_1987_2017_multiy		0.5566
VOD_X_1997_2018_multiye		0.3189

Traits Prediction

Stem density
Leaf area
Height
Seed mass
Leaf area, nitrogen

Project Overview

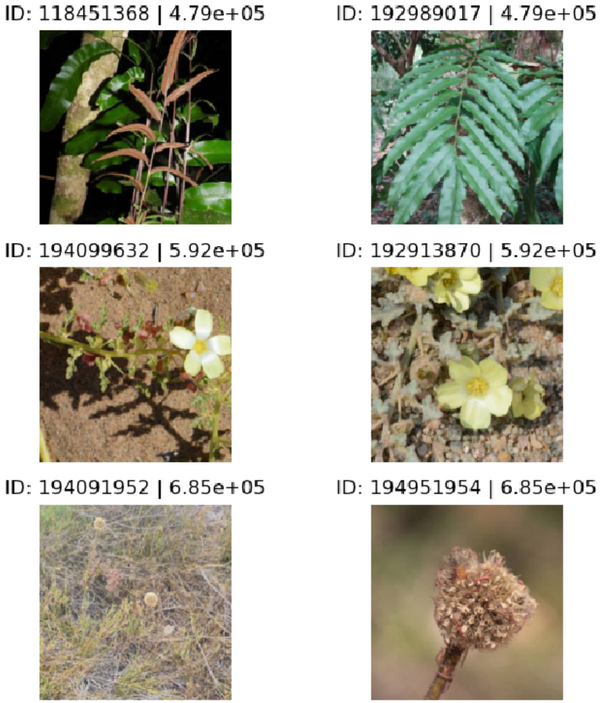
- **Problem Statement: Create a process and model to predict 6 plant traits from plant images and associated 'GeoData'.**
  - Images from iNaturalist, 512 x 512 RGB
  - GeoData based on image coordinates: Soil, Climate, Satellite, Radar measurements (*WORLDCLIM, SOIL, VOD, MODIS*)
  - Plant Traits targets: "chemical tissue properties loosely related to plant images" | *TRY database*
    - High accuracy not expected due to inherent heterogeneity of traits and plant images
- **Outcomes:**
  - Good start for outlier cleaning based on plant trait targets
  - Loosely optimized model architecture and multi-output regression strategy
  - Image+GeoData models performed better than GeoData only. Reasonable overall accuracy.

Reference Paper: "*Deep learning and citizen science enable automated plant trait predictions from photographs*"

Data Wrangling, Preprocessing

- Images and GeoData provided by Kaggle competition organizers (FGVC11 workshop at CVPR 2024)
  - *Not saved in project repository*
- Outliers detected and removed based on plant trait targets. Distributions assessed alongside images for extreme values.
- GeoData grouped by measurement technique, normalized, then reduced via PCA
  - Seasonal features engineered for Satellite and Radar measurements

Example: Leaf Area Outliers

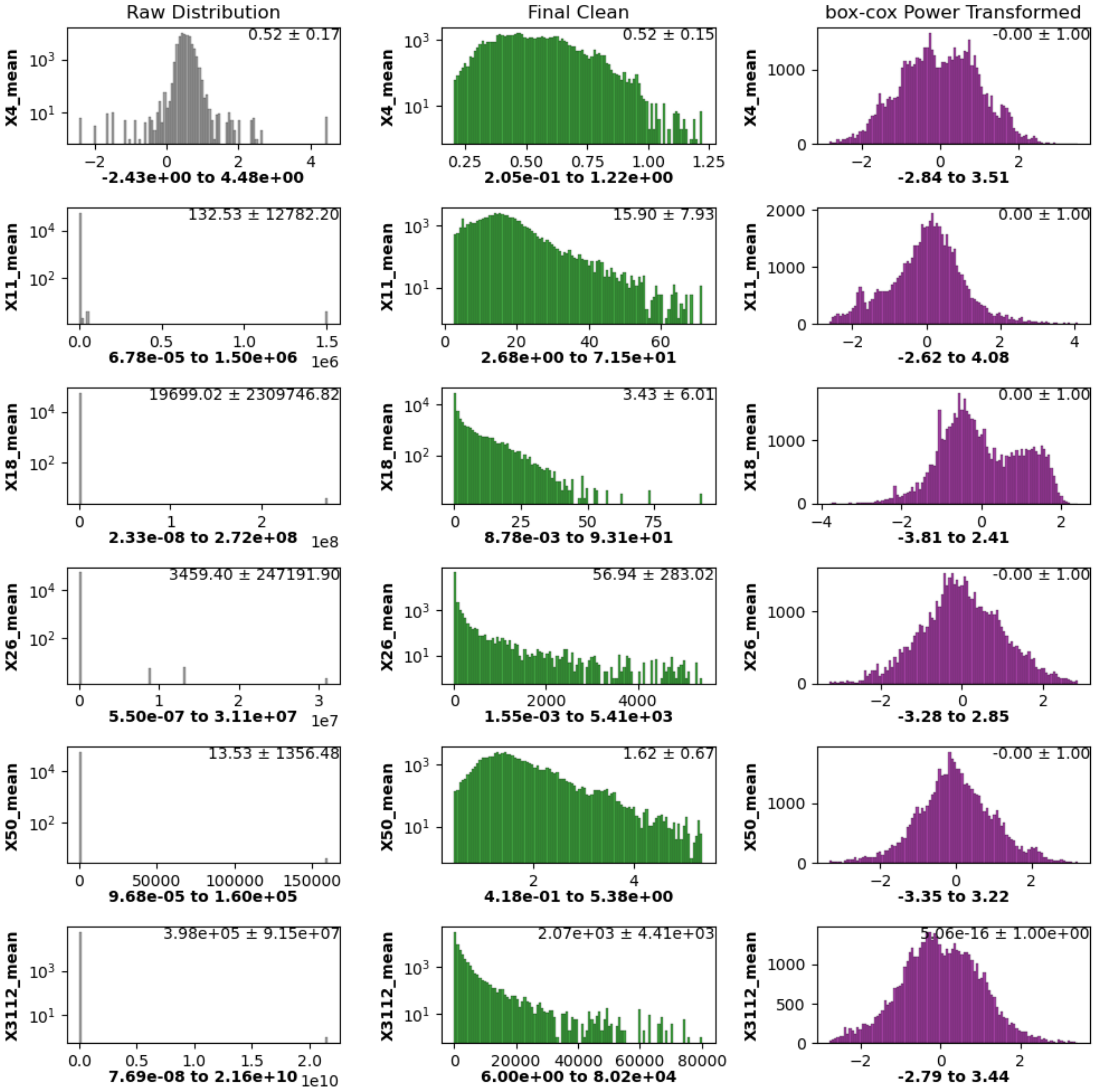


Area = 4.8e5 (cm^2?)
Probably real value
Looks quite leafy

Increasing leaf area

Area = 6.9e5
Unlikely real value
Doesn't look leafy

Possible TRY TraitID: 410 "Leaf area per plant"

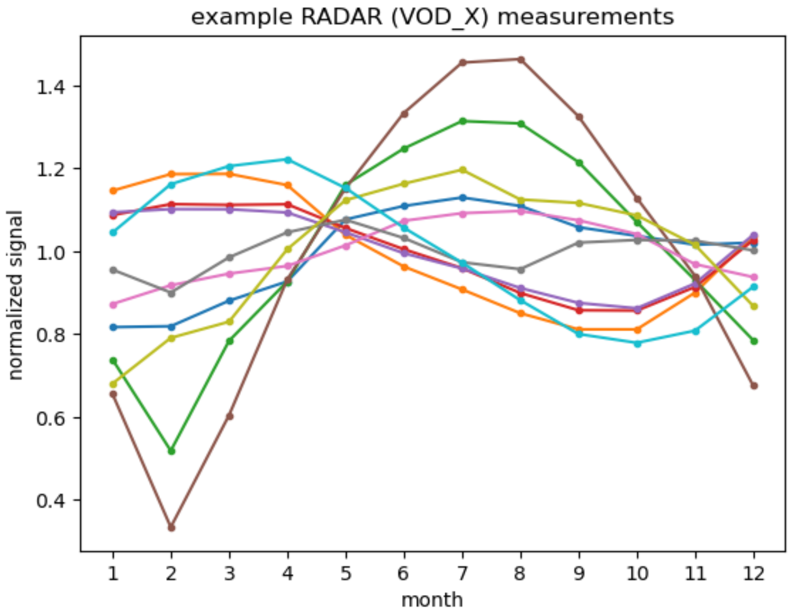


Plant Trait Target outlier removal

- **As shown in previous slide, plant images investigated to establish reasonable cutoffs for each trait.**
- **Then, additional outliers were removed based on log transformed standard deviations away from mean (≥ 3 removed).**
- **Finally, a power transformation was applied to provide a more normal distribution for each trait.**
  - Targets were inverse transformed prior to residuals analysis.

Seasonal Feature Engineering: RADAR + Satellite

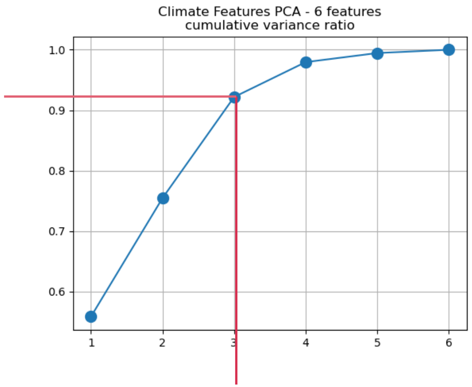
- Measurements are monthly means for various frequency bands
  - Satellite data provides sun reflectance information (5 bands, named by number)
  - Radar data provides water content, plant biomass information (3 bands: C, K_u, X)
- Inspired by provided climate data (BIOCLIM), engineered 3 features to replace each monthly measurement:
  - Max month – Min month (delta min,max)
  - Annual average
  - Seasonality (*standard deviation / mean*) [%]



Feature Decomposition – Principal Components Analysis

163 features → 21 PCA components

- **PCA decomposition gave better model performance than other feature selection/elimination methods (tested with ElasticNet)**
  - Sequential, Recursive, SelectFromModel
  - Makes sense as it should take more information into account
- **Each feature group considered separately**
  - Soil (10), Climate (3), Satellite (5), Radar (3)
- **PCA components chosen to capture at least 90% of feature group variation. Components analyzed to ensure input features had appropriate representation.**
- **Engineered Satellite and Radar features used instead of monthly means.**
- **Correlations with target traits compared for raw features and for PCA features**
  - Typically similar or higher for PCA



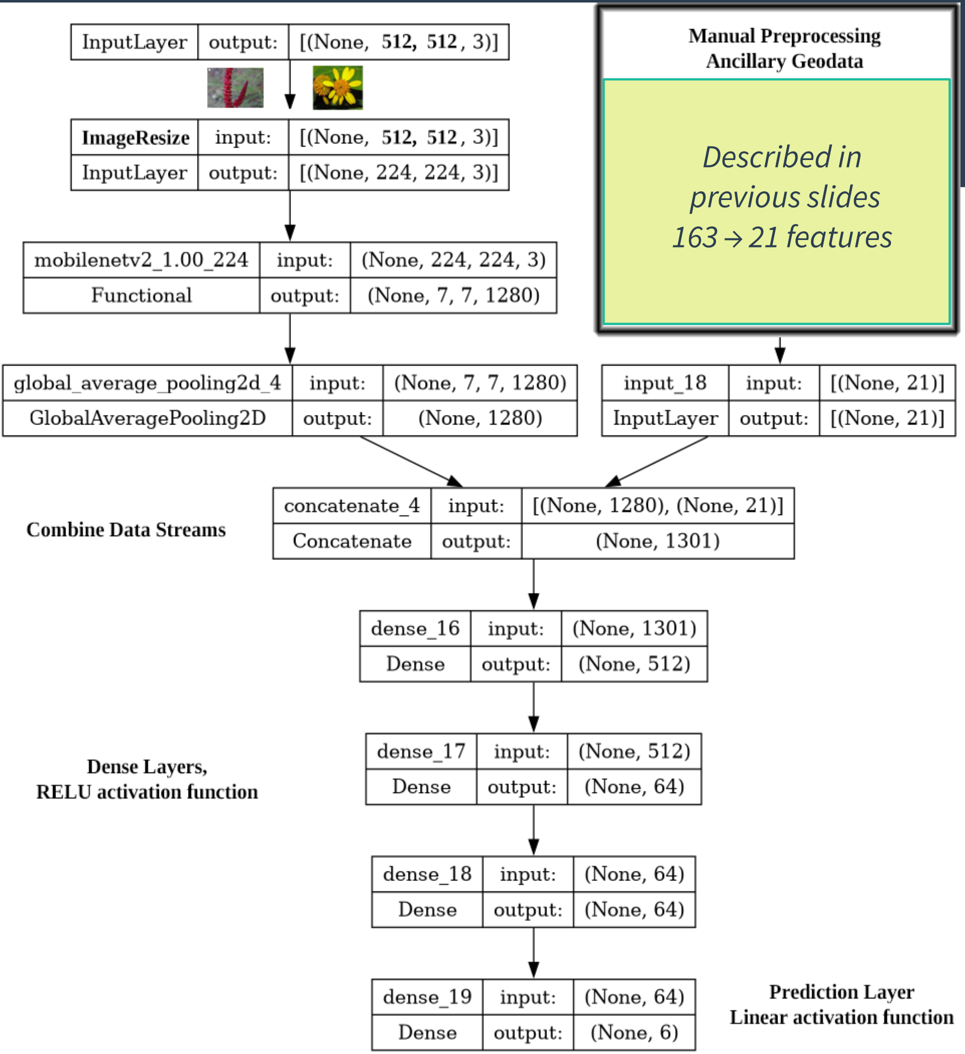
	PCA_1	PCA_2	PCA_3
BIO1	0.856	0.265	-0.609
BIO12	0.790	-0.476	0.856
BIO13.BIO14	0.846	0.166	1.000
BIO15	0.348	1.000	0.232
BIO4	-1.000	0.081	0.621
BIO7	-0.995	0.260	0.463

Climate Example
6 → 3

	Total (R2)
PCA_1	0.506617
PCA_3	0.088865
PCA_2	0.023678

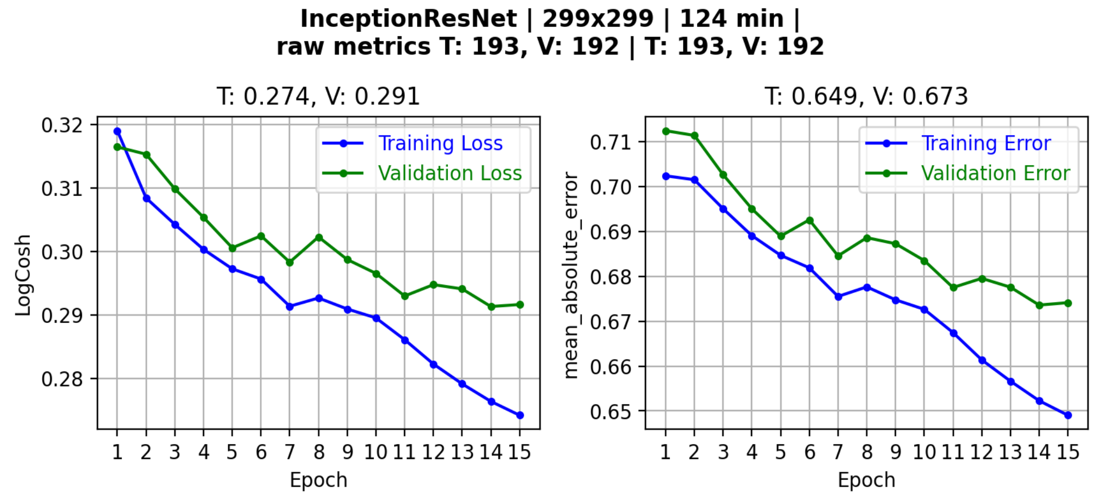
Sum of each target's R^2

Model (CNN) Architecture



- Example shown for adapting MobileNetV2
  - Same idea for InceptionResNetV2
- Concatenated image branch and geodata branch after individual feature extraction / decomposition
- Fully connected dense layers for final prediction
- Combining two data streams for one final prediction was found to be more accurate than separate predictions, combined after the fact (ex: VotingRegressor).
  - *Some tabular data (geodata) only models were more accurate when predicting traits sequentially*

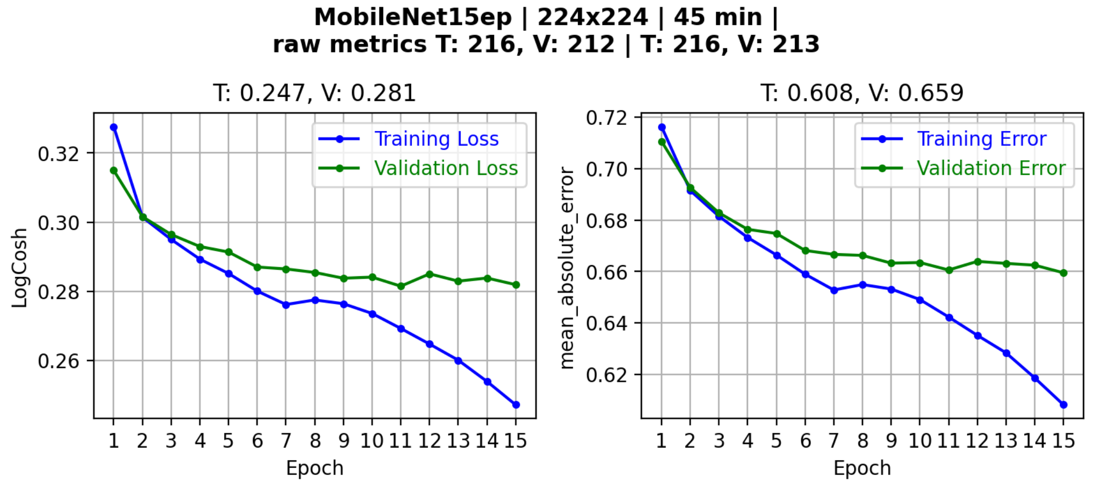
Modeling Training Results



History plot metrics are for transformed targets

Title text provides 'raw metrics' (LogCosh | MAE) for the training and validation sets

Hold out test scores presented on next slide.



Best two modeling attempts shown.

Conclusions from previous experiments:

- **Existing architectures provide good baseline performance. (see: Keras table)**
  - ImageNet ranking does not correlate to this task. **MobileNetV2** outperformed all tested models except for **InceptionResNetV2**.
- **Image augmentation and integrated resize layer help model performance.**
- **Fully retraining model weights boosts performance without taking more time.**
- **LogCosh good loss metric for raw target traits, limits influence of extreme values.**
- **Room for optimization and hyper-parameter tuning.**

Final Model

justification for MobileNetV2

Metric	MobileNet	IncResNet	CatBoost
MAE	212	192	263
LogCosh	212	192	263
Time	45 mins	124 mins	40s

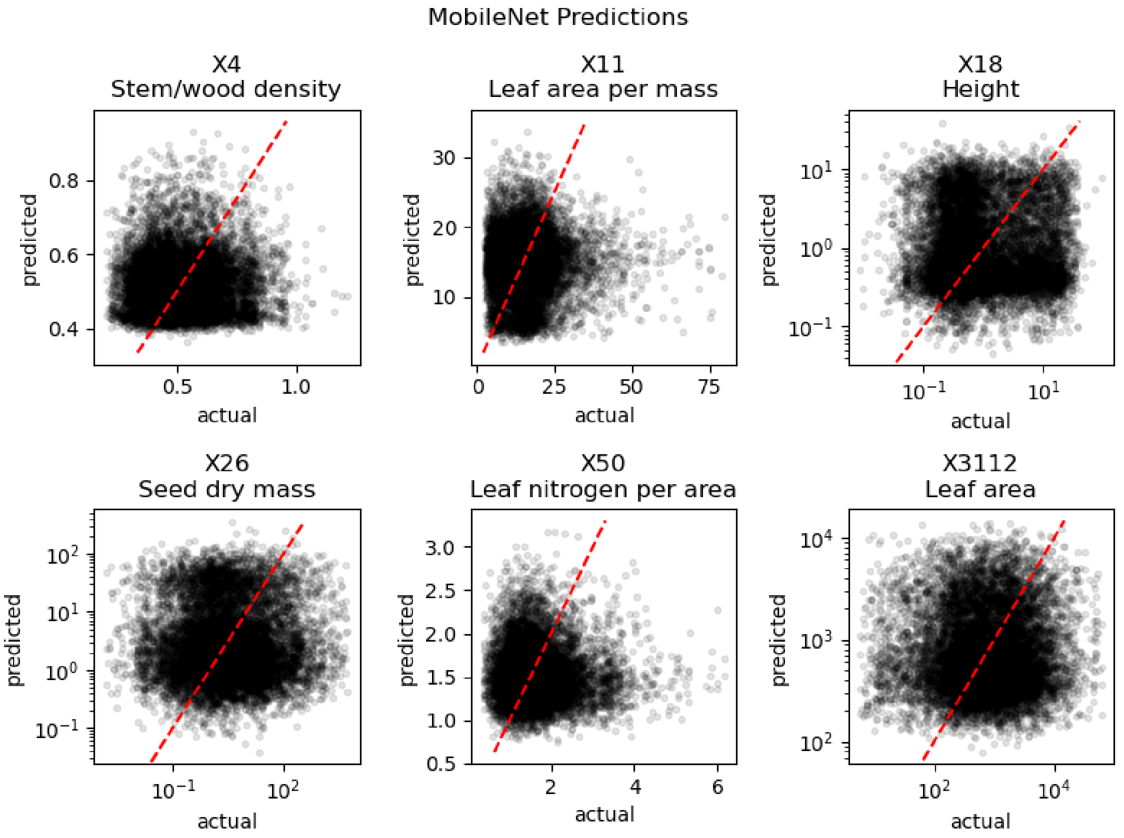
Hold Out test results (20% of data)

- **The paper this competition was inspired by ultimately choose InceptionResNetV2 for their CNN architecture. While this also gave me the best performance, I will select MobileNetV2 for my final model.**
- **Accuracy was not much worse, and training time was *much faster*. Additionally, the model itself is lighter (45 MB vs 840 MB).**
- **As my final work has not been fully optimized, the model using MobileNetV2's architecture will be more suitable for tuning, assuming development on a similar system.**
  - More powerful systems could consider models I was not able to train.
- **The following slides will present residual analysis from the MobileNet model.**

Residual Analysis 1

errors (mostly) randomly distributed for the 6 plant traits

- Prediction range typically lower than actual range.
  - Does this indicate insufficient or poor outlier detection/cleaning?
- Plant Height seems to be more difficult to predict accurately, not quite random error.
  - **L** shape apparent. Accuracy worse for taller plants.
- *Note log scale for Height, Seed dry mass and Leaf Area*



Residual Analysis 2

image investigation – plant height (*X18_mean*) example | *worst predictions*

- Largest (4) over-predictions on top row, largest (4) under-predictions on bottom row.
- Makes sense for model to struggle with images showing only subset of plant.
- Should future preprocessing / cleaning consider image quality?
  - Ex: if ImageNet identifies an image as a “hand”, don’t use.

Residual Analysis 3

image investigation – leaf area (*X3112_mean*) example | *best predictions*

- Best (8) predictions based on absolute error.
- Encouraging to see accuracy at different magnitudes of leaf area (*trait range ~ 10^4*).
- Two of best predictions are from image containing a hand.
  - Maybe cleaning should not be as simple as suggested on the last slide.
  - Do hands affect some traits more than others?

Model Refinement

model and data cleaning still have room for improvement

- Continue developing cleaning methods for training data. Focus on images. More specific target outlier assessment.
  - Wide variety of images should be part of a training set, given “citizen-science” problem statement. But not all the images need be used for model training.
  - Given my lack of domain knowledge, identifying plant trait outliers or faulty data was “handwavy”. Subject matter expertise should improve this process.
- Target Trait plasticity
  - Plant traits are provided as a mean measurement. Previous work found benefit to “plasticizing” these values with the traits’ standard deviations, something which I did not try for this project.
- CNN model hyperparameter tuning. More training epochs.
  - Given long training time, I focused on “big” pieces for experimentation, like the choice of transferred CNN architecture.
  - More handles can be turned to tune model performance, like choice of optimizer, loss and metric functions, learning rate schedules, feature averaging techniques, dropout layers, activation functions, etc. . .
- Species identification: Should help model performance, as species and close relations will have similar traits.
- Transformer Models: I focused on CNN architectures for this project.

Project Implications

- My work, in addition to that of all the Kaggle competitors, provides an important start to utilizing citizen sourced data for measuring the health of Earth's ecosystems.
- High accuracy was not expected as part of the project design, but analysis of the images with the highest residuals could improve future data cleaning efforts.
- The data cleaning, feature engineering, feature decomposition, and feature combination streams employed prior to model training will benefit future, similar projects.

Acknowledgments

- **Aditya Bhattacharya**
  Thank you for your mentorship and guidance throughout the project!

- **Kaggle Competition, Paper Authors**
  The competition and paper provided a great template for this work, and the competitors helped me get started with the data and modeling. Thank you Rich Olson for your helpful discussions and shared work.

- **Springboard Curriculum**
  The guided capstone project served as a useful guide for this project.

Questions?