Predicting Species Density using Human and Environmental Factors

Team 32: Erika Ordog, Rohan Sarkar, Kristen Scheckelhoff
Erdős Data Science Boot Camp, Spring 2023

Overview

Goals

We are interested in exploring the effects of environmental and human factors on animal species distribution. Some animal species are known to adapt and thrive in the face of expanding urbanization, but with this increased human activity also comes competition for resources. Identifying key factors which influence species density in urban areas can lead to policy recommendations for more sustainable urban growth.

Dataset

We used the mammal population densities dataset[1] compiled by Tucker et al. for a recent article[2] published in the journal Ecography. The authors concluded that mammal populations densities were higher in urban areas, but also noted that this may be due to related factors, e.g. humans choosing to build in areas with abundant natural resources. The dataset includes the Landsat Normalized Difference Vegetation Index (NDVI), a possible environmental predictor of species density.

Methods

Exploratory Data Analysis

After some initial exploration, we selected a subset of 6 species with sufficiently many data points and similar habitats: *Axis axis* (chital deer), *Loxodonta africana* (African elephant), *Panthera pardus* (leopard), *Panthera tigris* (tiger), *Sus scrofa* (wild boar), and *Syncerus caffer* (African buffalo), for a total of 1057 data points. Paired plots across the available features did not immediately reveal any strong trends, so we tested a variety of models.

Linear Regression

We fit linear regression models for species density using NDVI, Human Footprint, and Species Richness (all at the 1-degree level of granularity) as the predictive features, as well as log transforms of both the independent and dependent variables, and a quadratic regression. None of these models were significantly stronger than the baseline, so we moved on to nonlinear regression.

Nonlinear Models

We tested Random Forest, XGBoost, and K-Nearest Neighbors, using all available relevant features from the dataset to try to best predict species density, and used k-fold cross-validation to validate the performance. These models generally outperformed both the baseline (average density) and the linear regression models.

Results

We obtained the lowest error when predicting species density on the test set with the Random Forests regression model. This model also revealed that the five most important features impacting species density are (1) cropland, (2) night lights, (3) human population density, (4) NDVI, and (5) human footprint. In future iterations, it would be interesting to break down the analysis by individual species, to see if environmental features or human impact features play a greater role in determining density. We would also like to incorporate additional datasets containing information about natural resource availability.

Model	Cross-Val RMSE	Test Set RMSE
Baseline	8.50	7.23
Linear	7.95	6.18
XGBoost	5.49	6.22
Random Forests	4.83	5.30
KNN	4.68	6.00

References

[1] Marlee A. Tucker, Luca Santini, Chris Carbone, and Thomas Mueller. Mammal population densities at a global scale are higher in human-modified areas. Dryad, Dataset, 2020. https://doi.org/10.5061/dryad.m63xsj40d.

[2] Marlee A. Tucker, Luca Santini, Chris Carbone, and Thomas Mueller. Mammal population densities at a global scale are higher in human-modified areas. *Ecography*, 44(1):1–13, 2021.