Project Proposal

MSDS-696 • Data Science Practicum II

Tyler Wilson
Regis University

Contents

1. Problem or Situation
2. Research Question
3. Data
4. Methodology
5. Project Timeline

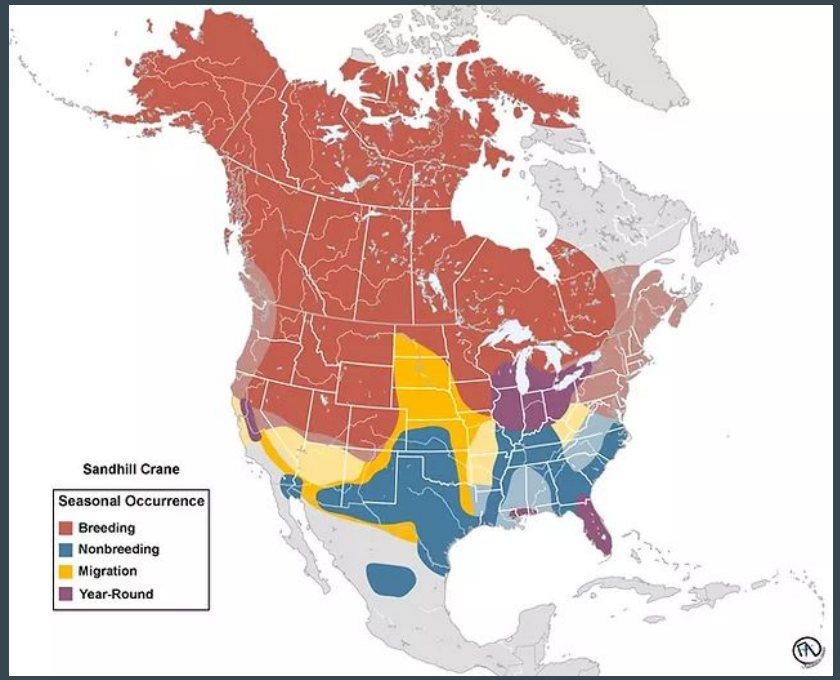
Problem or Situation

The Cornell Lab of Ornithology at Cornell University provides eBird as a way for bird watches to log their observations. There is a tremendous amount of data available and many potential uses for it. . The data is available for research and academic purposes through requests.

While there is a huge amount of data available, how it is used is open ended, but has historically been used in research papers and to help make conservation decisions.

Research Question

Were will a given bird species most likely be seen at any date and how many other bird watchers may be present in that area? Use machine learning and a Shiny app to answer and show results.



The Dataset

This project will mainly use data provided by eBird. It may be subsetted to only include North America and/or other areas that have sufficient data for model training and testing.

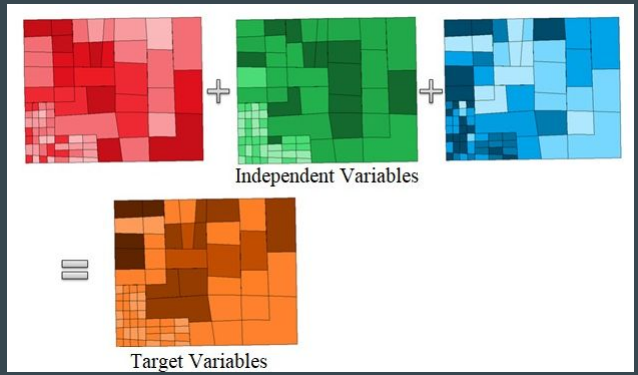
The worldwide dataset was provided as a tab delimited text file which is about 430 GB.

Part of my project will be learning how to handle large datasets that are far larger than system memory. I plan to use Apache Arrow to help with this.

Methodology

At this point, it is too early to determine which flavor of machine learning will be best to answer the research question. Possibilities include:

- Geographical Weighted Regression
- Random Forset classification
- Maximum Likehood
- Hot Spot analysis
- Space Time Pattern Mining
- Neural network



The method/methods that will be used will be chosen once exploratory data analysis (EDA) is complete.

Timeline

Task	Expected Date Complete
EDA/data splitting—-----	2022-10-30
Model exploration—-----	2022-11-6
Initial models made—---	2022-11-13
Parameter/hyper-parameter tuning—-------	2022-11-20
Shiny App creation—----	2022-11-27
Documentation—---------	2022-12-04/Project due date