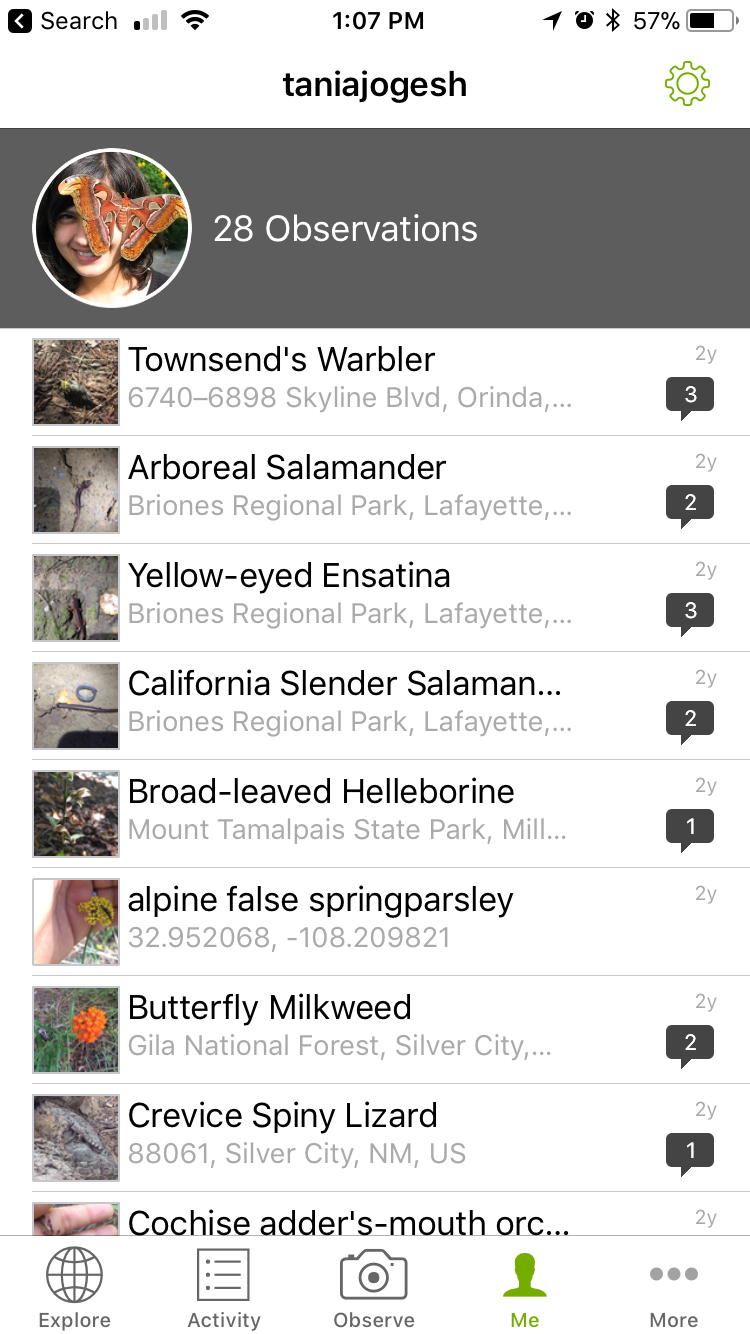
PREDICTING THE ACCURACY OF CITIZEN COLLECTED BIODIVERSITY DATA

**The Problem**

We share our plant with 20-30 billion species, and each individual species is a treasure trove of data. The food we eat, the plants in our gardens and even some of the most important medicines in human history are a product of other living species. Biodiversity is akin to a library that we are only beginning to understand and this library can hold the key to some of the most important problems in the future. It is then, imperative that we effectively document this data, to understand, which species share our planet and where there occur so we can catalog the valuable information around us.

iNaturalist is a mobile application run by the California Academy of Sciences that allows people to upload photos and information on species (plant and animals) they observe in their daily life. It is a fantastic application that relies of regular people, who look at birds in their garden or hike the National Parks of America to document and collect biodiversity data. However, not everyone is an expertly trained biologist, and some organisms are inherently difficult to identify.

Here, I develop a classifier that predicts if an identification posted via the Citizen Science Biodiversity app iNaturalist (https://www.inaturalist.org/) is likely to be correct. iNaturalist is a mobile application that allows users to upload photos and information on natural history observations (plant and animal sightings). The online community can subsequently verify observations uploaded by users. However, there are a lot of observations and not all plants/animals/fungi are easy to identify. Instead of relying solely on the community to verify an ID, my classifier will help iNaturalist decide which observations are more likely to be correct or misidentified.



**The Data**

iNaturalist’s user data and observations are freely available to the public and can be queried and obtained via an API. A subset of these data include whether or not an observation was accurately classified by experts in the community so the data are already labeled for classification. I downloaded 10,000 records from their database (I can download more but am limited by computing power) for observations from 2016-2017.

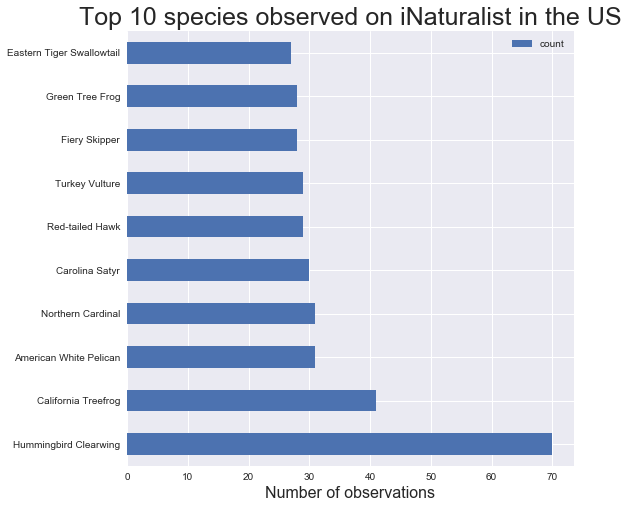
**The Approach**

I acquired the data via an API in json format and I downloaded the data with each observation in a single json file. I did this to make it easier for spark to infer the schema and so that the data were imported as a single observation per row in the Spark Data Frame. These data were really messy because a lot of the data was nested each value was stored as an array.

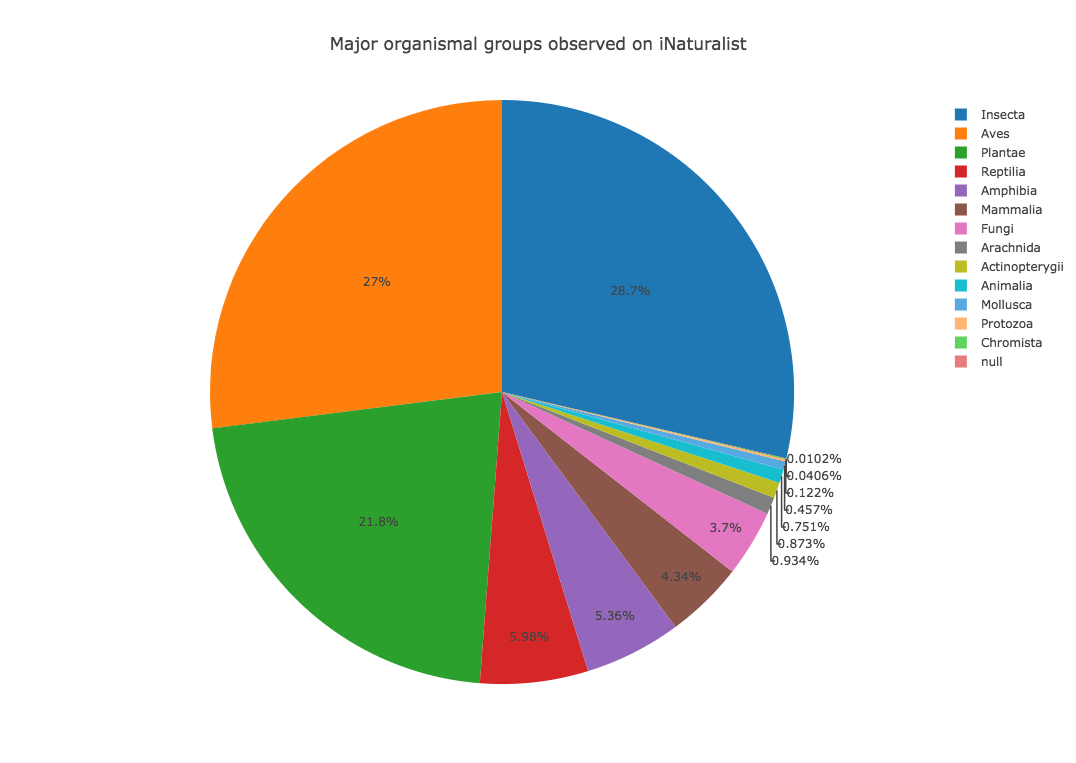
* To flatten the data structure and format the data I selected columns of interest into a Spark Data Frame table. However, all data points were still inputted as arrays. I tried mapping a lambda function to extract list values in both Spark and Pandas. Because of lazy evaluation (even with data caching), it was much faster to process the data in pandas and write it back to Spark.
* I conducted most of my exploratory data analysis using the Spark DataFrame API

**Exploratory Data Analysis**

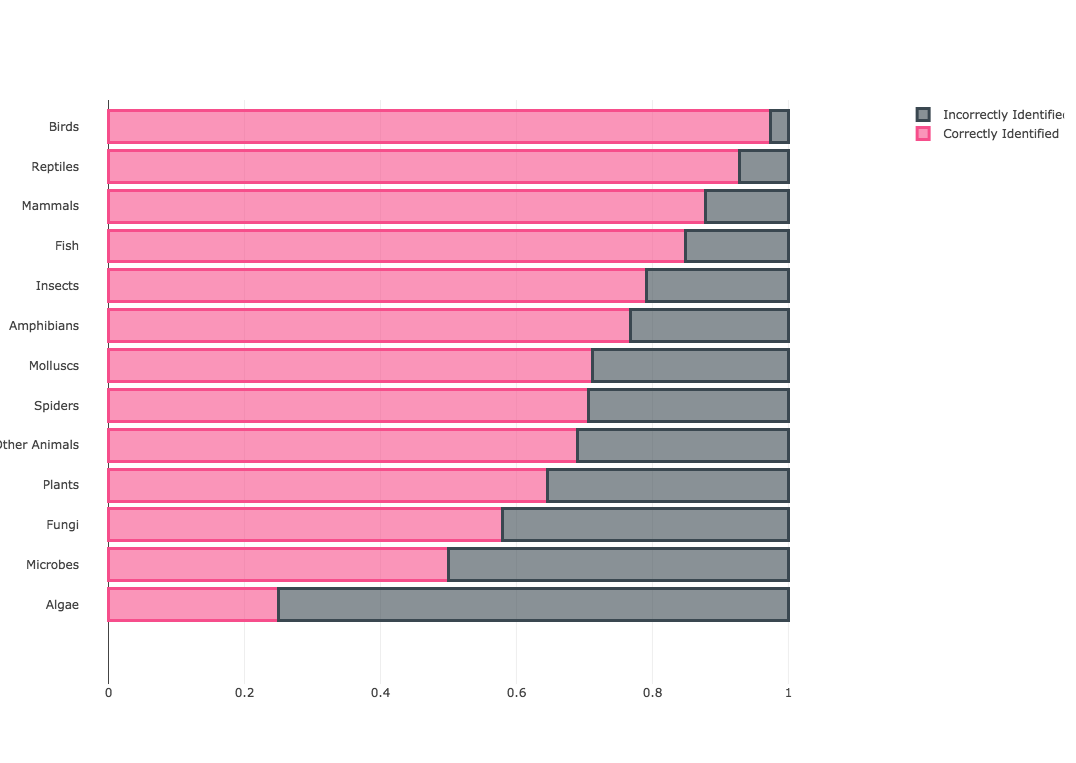
Observations on iNaturalist are likely to be biased in favor of species that are more charismatic. 80% of the “Top 10” list contains birds and butterflies!

****

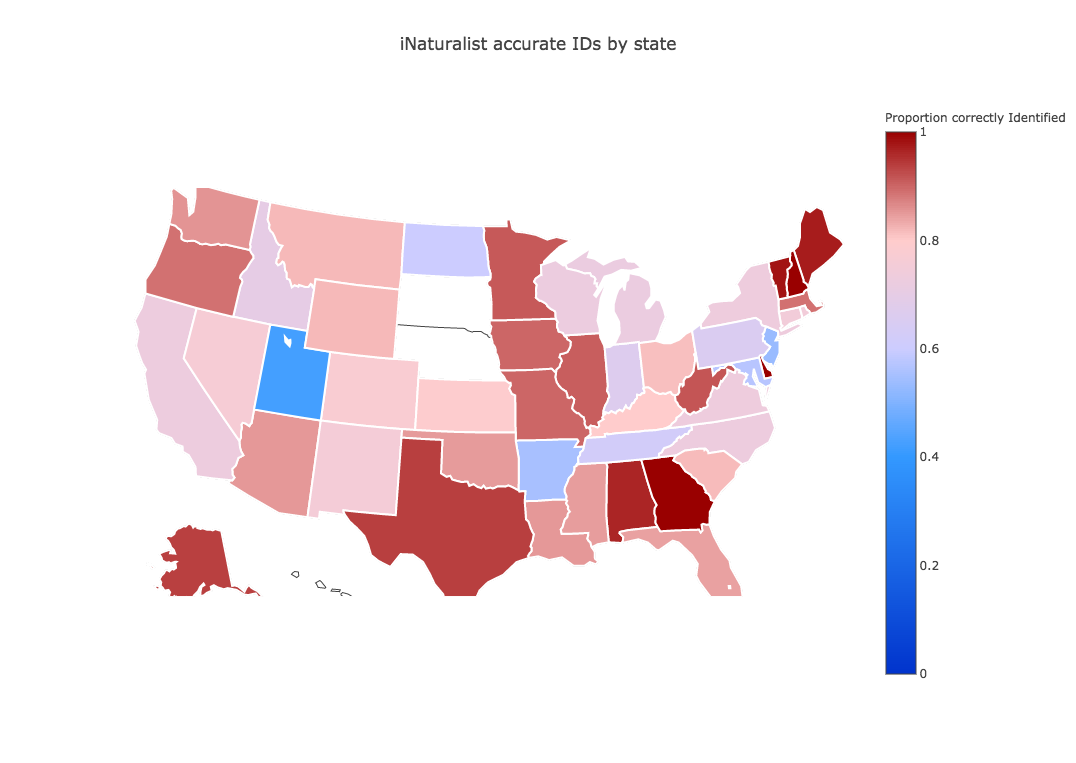
Similarly, charismatic organisms like Insects, Birds and Plants comprise the bulk of observations on iNaturalist

****

Indeed, the taxonomic group is important in determining the identification accuracy



It is also fascinating that some states have much higher proportions of accurately identified observations. States with fewer observations overall have low identification accuracy. It is also interesting that accuracy is higher is the west/mid-west compared to the east.



* Using groupby, count and crosstab, I explored the relationship between all the features of interest and the accuracy of identification
* I obtained information on the location state via reverse geocoding and used this to plot observation accuracy on a map. State will be used as a feature in the machine learning models

**Classification**

* The data contain a lot of categorical features. To deal with these I tested two approaches 1) labeled the data numerically 2) used a one hot encoder to get every category encoded.
* I tested three classifiers: 1.) Support Vector Machines 2.) Random Forest 3.) Gradient Boosting classifier. Why these three? Naïve Bayes and has high bias low variance so it is not as accurate and probably better for smaller data. Knn is dependent on scale and a lot of my features are categorical. Logistic regression is parametric and assumes a linear relationship between x and y, not necessarily true for this data so that leaves tree-based methods and SVM. I tried two tree-based methods, Random Forest and Gradient Boosting classifier along with a Support Vector Machine classifier.
* For each of these I optimized hyper-parameters individually or with a grid search. The Random Forest was the best model in terms of speed and accuracy

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **optimized**  **hyper-parameters** | **Categorical Features** | **Train accuracy** | **Test**  **accuracy** | **speed** |
| SVM | C=1, gamma =1 | labeled | 81.2% | 82.1% | 2.03s |
|  | one hot encoded | 87.4% | 83.7% | 21.86s |
| Random forest | default | labeled | 99.7% | 87.6%, | 1.13s |
|  | one hot encoded | 99.7% | 88.2% | 1.99s |
| Gradient Boosting | max depth = 9, learning rate = 0.01, subsample = 0.6, min\_leaf=1 | labeled | 99.7% | 87.1% | 33.62s |
|  | one hot encoded | 98.7% | 87.4% | 136.39s |

The random forest classifier is the winner with high accuracy (88.2%), high AUC scores (0.89) and shortest processing time.

**Conclusions**

* The final Random Forest classifier, which is 88% accurate, can by used by iNaturalist to automatically assign an accuracy rating to new observations.
* Observations with high accuracy can contribute to robust biodiversity data, whereas those predicted to be inaccurate could be tagged for expert identification.
* Accuracy classification can help streamline data processing and quickly identify good biodiversity data from bad.