Capstone 2: Ebird Data Visualization and Machine Learning

Part 1: Loading the data

This project began by downloading the california dataset from ebird’s publicly available data (https://ebird.org/data/download/ebd?showSuccessMsg=true). This file was 31 GB on its own, so I further trimmed it down to only include species commonly included in the ‘Birds of Prey’ category (and removing irrelevant columns), which cut the file down to 550 MB. This file was still fairly large, so I used chunking to read the csv file into a pandas dataframe.

Part 2: EDA

I did a brief analysis of sightings over time to see if there were any interesting conclusions to be drawn from that. Unfortunately, this data only seems to indicate the increased usage of ebird over time, rather than any real-world changes in bird population levels. As I do not have access to ebird’s internal data, I cannot normalize this to their number of users in a given year. It would be interesting to see in a decade or so, if usership has leveled out and broad conclusions can be drawn from this data.

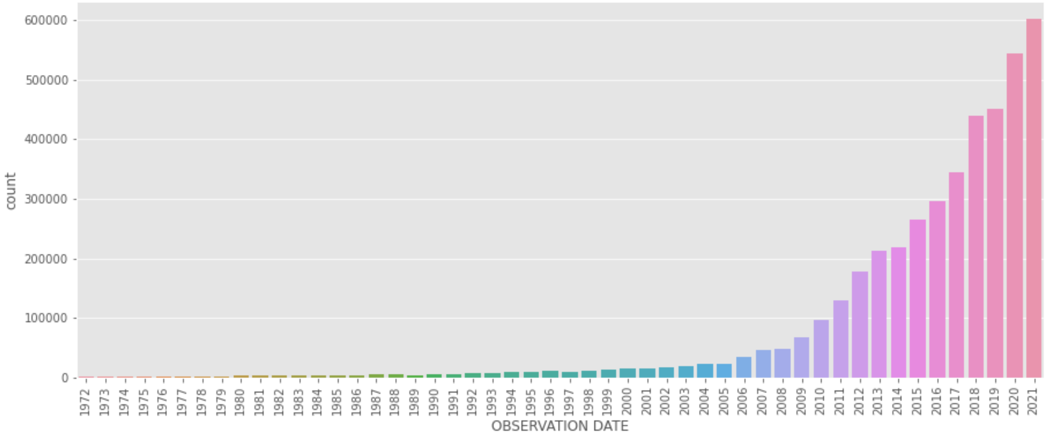
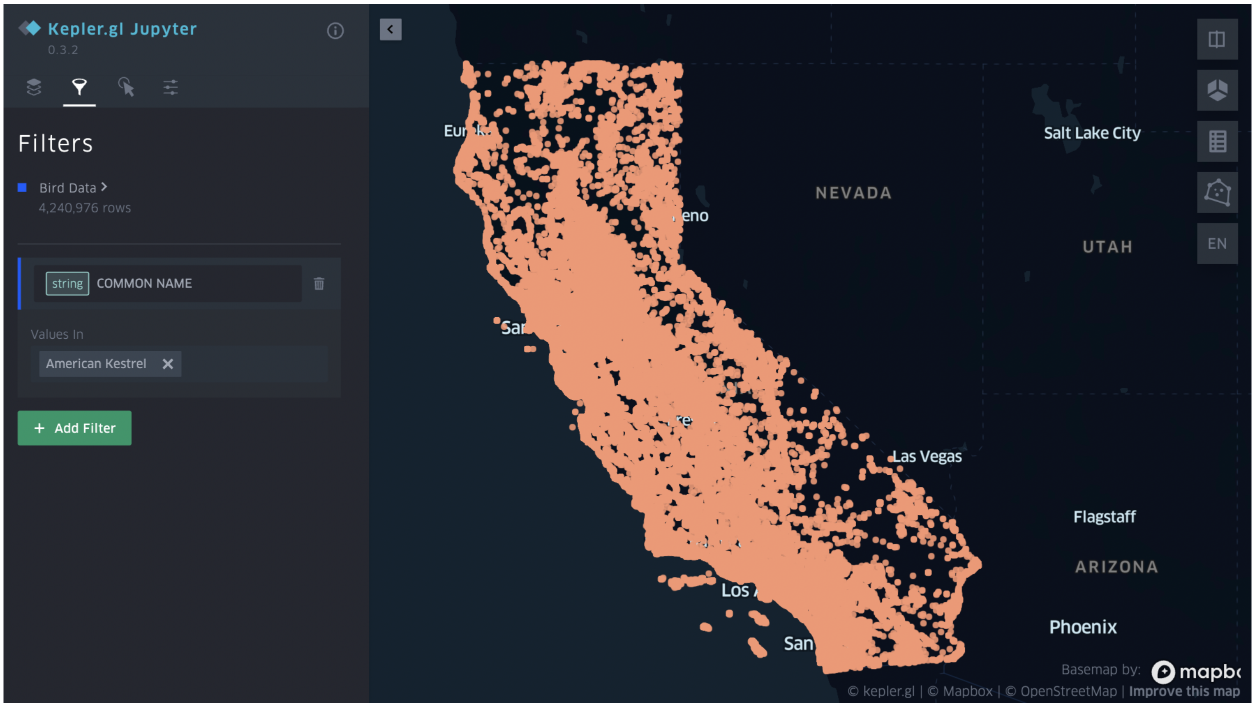


Figure 1: Sightings over time (Birds of Prey in California)

I also built a Kepler map that could be used to view the locations of birds in California at the times they were sighted and logged in ebird’s system. Filtering allows the user to see the distributions of various bird species. The American Kestrel (figure 2), for example, is both extremely common and widespread, able to be seen nearly everywhere in California. The Bald Eagle (figure 3), by comparison, is similarly widespread but much less common, with a much sparser map over the same length of time. Birds like the Barred Owl (figure 4) and the Northern Goshawk (figure 5) have clear borders on their typical habitats, plus some individual wanderers that have been seen outside those areas. One could also tell from this tool that the Aplomado Falcon (figure 6) is not a common or local species, as there is only one California sighting in the whole of the ebird dataset. Indeed, the typical habitat of the Aplomado Falcon is parts of South and Central America, and the bird seen in California would have been what is know as a 'vagrant bird' (ie. birds that have gotten very, *very* lost).



Kepler.gl Jupyter
0.3.2

Filters

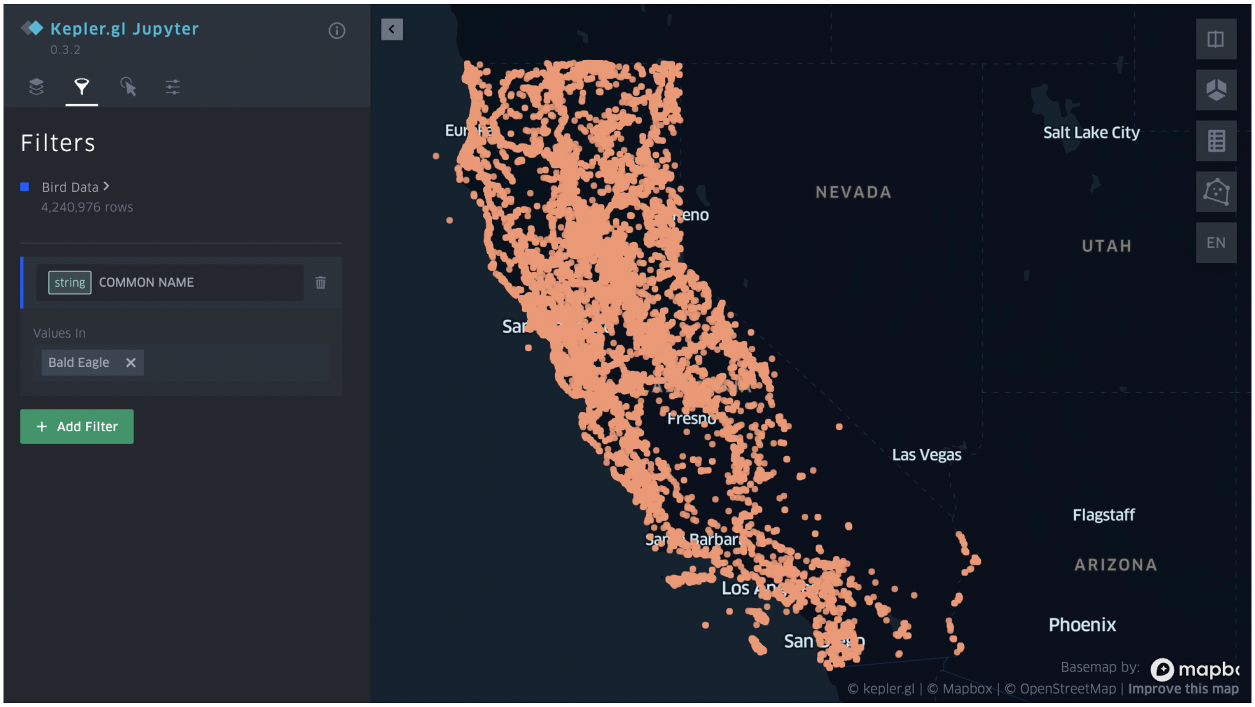
Bird Data >
4,240,976 rows

string COMMON NAME

Values In
American Kestrel ✕

+ Add Filter

Figure 2: Distribution of the American Kestrel



Kepler.gl Jupyter
0.3.2

Filters

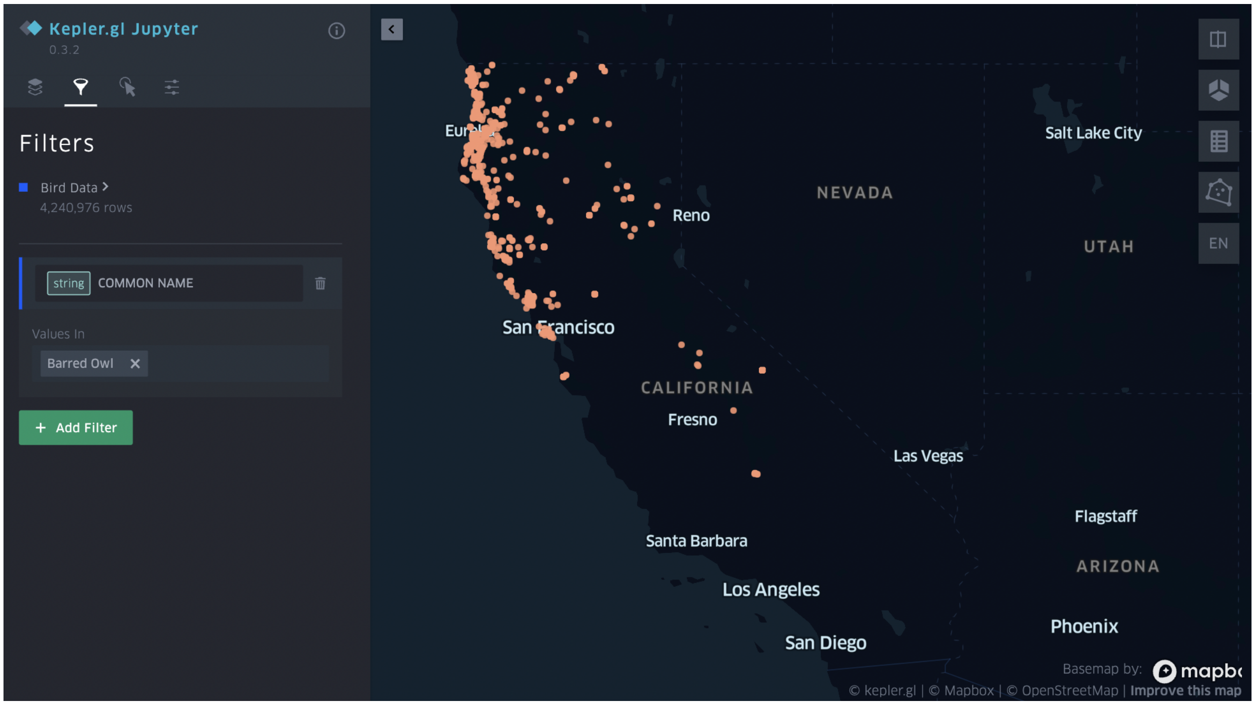
Bird Data >
4,240,976 rows

string COMMON NAME

Values In
Bald Eagle ✕

+ Add Filter

Figure 3: Distribution of the Bald Eagle



Kepler.gl Jupyter
0.3.2

Filters

Bird Data
4,240,976 rows

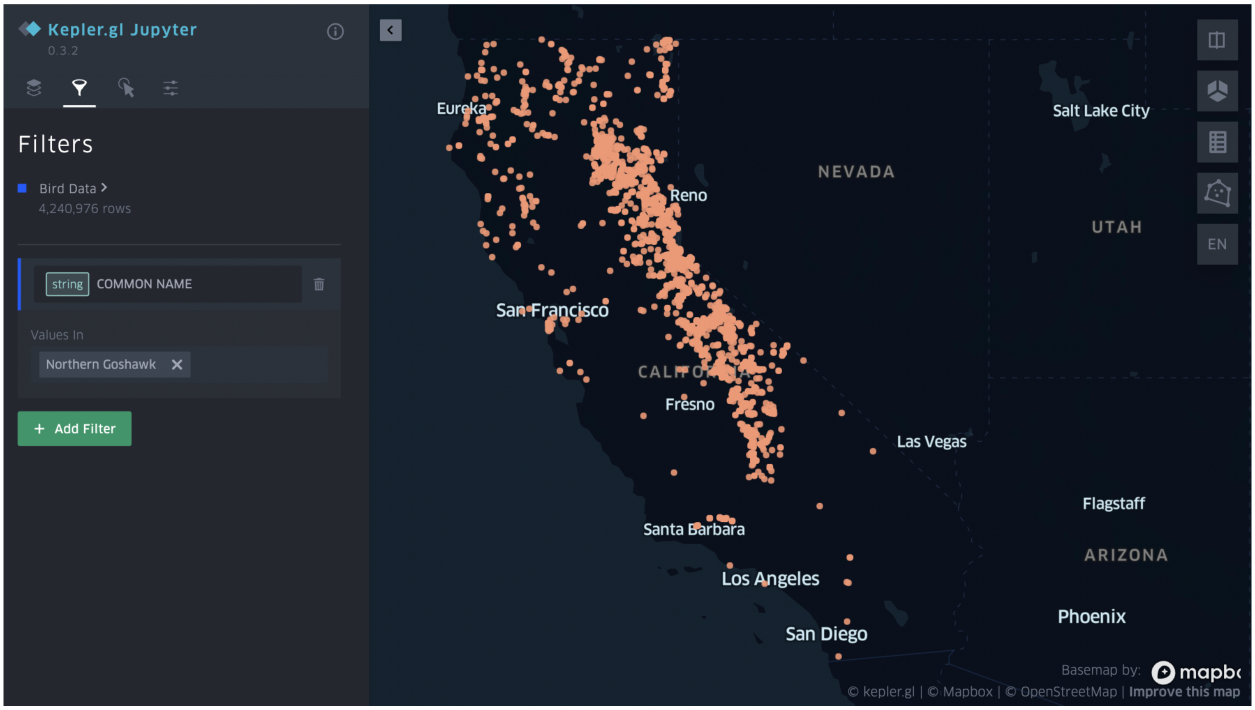
string COMMON NAME

Values In

Barred Owl ✕

+ Add Filter

Figure 4: Distribution of the Barred Owl



Kepler.gl Jupyter
0.3.2

Filters

Bird Data
4,240,976 rows

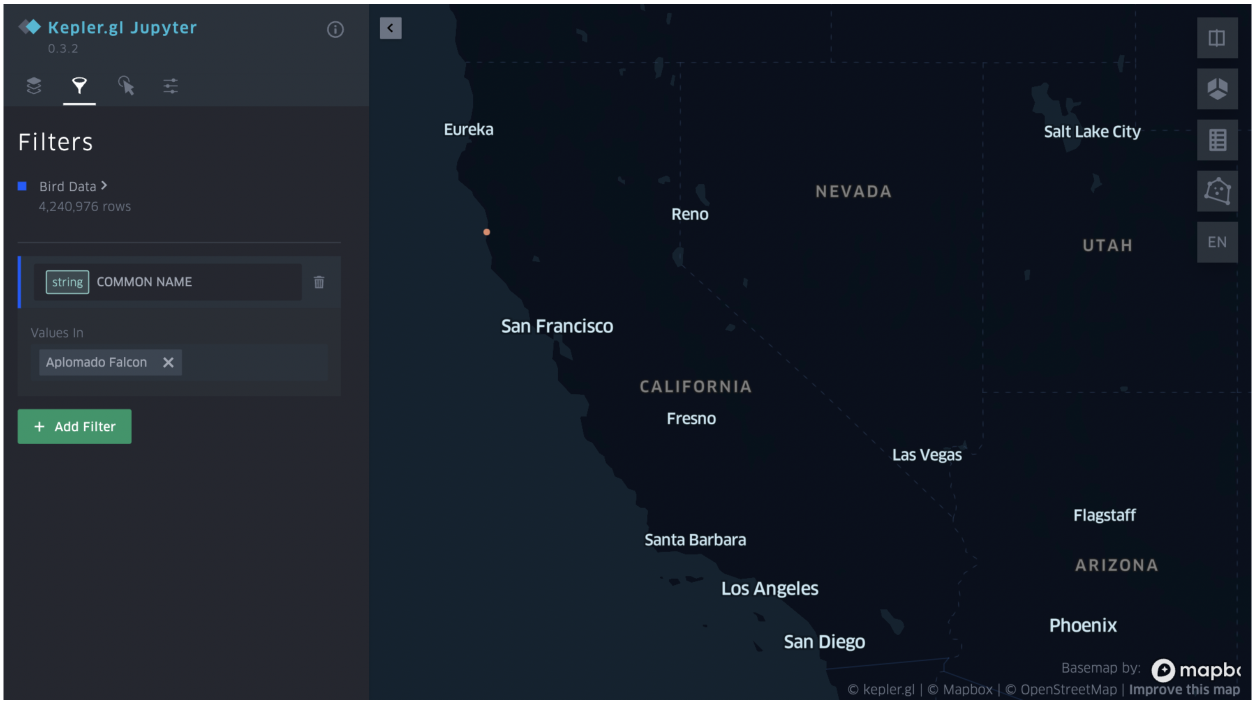
string COMMON NAME

Values In

Northern Goshawk ✕

+ Add Filter

Figure 5: Distribution of the Northern Goshawk



Bird Data
4,240,976 rows

string COMMON NAME

Values In
Aplomado Falcon

Figure 6: Distribution of the Aplomado Falcon

Part 3: Machine Learning

Some additional preprocessing work was necessary before I could build a predictive model. I removed any bird species that had less than 1000 sightings, and then added in color and size data for each species.

A note on the color and size data: The ebird dataset does not include any information about the specific appearance of the birds in the sightings reported. As a result, I had to spoof my own data, which I based on the adult appearance of each species (most raptors are not sexually dimorphic in appearance, but in the species that are, I based my color and size values on the adult male since they tend to be flashier and more recognizable). In reality, many species look very different at different stages of development, and will also have dark- or light-morph individuals which may be particularly common in specific areas. My algorithms, as a result, perform much better at categorizing individual sightings into species, than one would expect with actual real-world data. As a result, this is more a proof-of-concept project than anything that could actually be used in the real world (not to mention that better bird-species prediction software already exists.)

Once I had my color and size data, I converted them into dummy variables, normalized my lat/long data, split into test/train sets, and fed that into a Random Forest and XGBoost machine learning classifiers. Accuracy and F1 score for these models is shown below

Random Forest: Accuracy=0.824
Random Forest: f1-score=0.815

XGBoost: Accuracy=0.820
XGBoost: f1-score=0.799

Because these models (XGBoost in particular) were taking a significant amount of time to run, I cut down the number of rows to just 500,000 and reran the models, to see if I could improve the performance without impacting the accuracy too much.

XGBoost: Accuracy=0.817
XGBoost: f1-score=0.797

Random Forest: Accuracy=0.819
Random Forest: f1-score=0.810

Based on these results, it does seem like cutting down the size of the train dataset results in a much faster model, without much of a loss in accuracy or F1 score. Of these models, the best performing is the Random Forest.