Biodiversity Across the Contiguous United States

Objective:

The United States of America is often renowned for its diverse landscapes. From steep, snowy mountains in the north, to swamps and sandy beaches in the south. Rolling mountains cover the east coast, while high, jagged mountains characterize the west coast. Each of these landscapes serve as a sanctuary for different types of plants and animals that make the United States so unique and broad. National parks, protected areas throughout the states, serve as a safe place for those species to exist and thrive. With such diverse landscapes, is there a difference in biodiversity across them? Looking at the United States, machine learning techniques will be used to identify similarities and differences of biodiversity and conservation status between states and categories.

Background:

The IUCN Red List, a comprehensive source of information about the population status of numerous species, has been active since 1964. As of the last update, more than 128,500 species were featured in this list. "The IUCN Red List is crucial not only for helping to identify those species needing targeted recovery efforts, but also for focusing the conservation agenda by identifying the key sites and habitats that need to be protected,"[2]. A global-level assessment is needed for each species included in the IUCN Red List, with updated assessments as years pass with any changes in population or condition. Meaning, an assessment of a species that includes the population numbers around the entire globe. Teams that gather information for these assessments are almost always apart of the Regional Red List Assessment group. In addition to total population size, literature written about the species must also be presented. Checks of each assessment are also performed by a specific group.

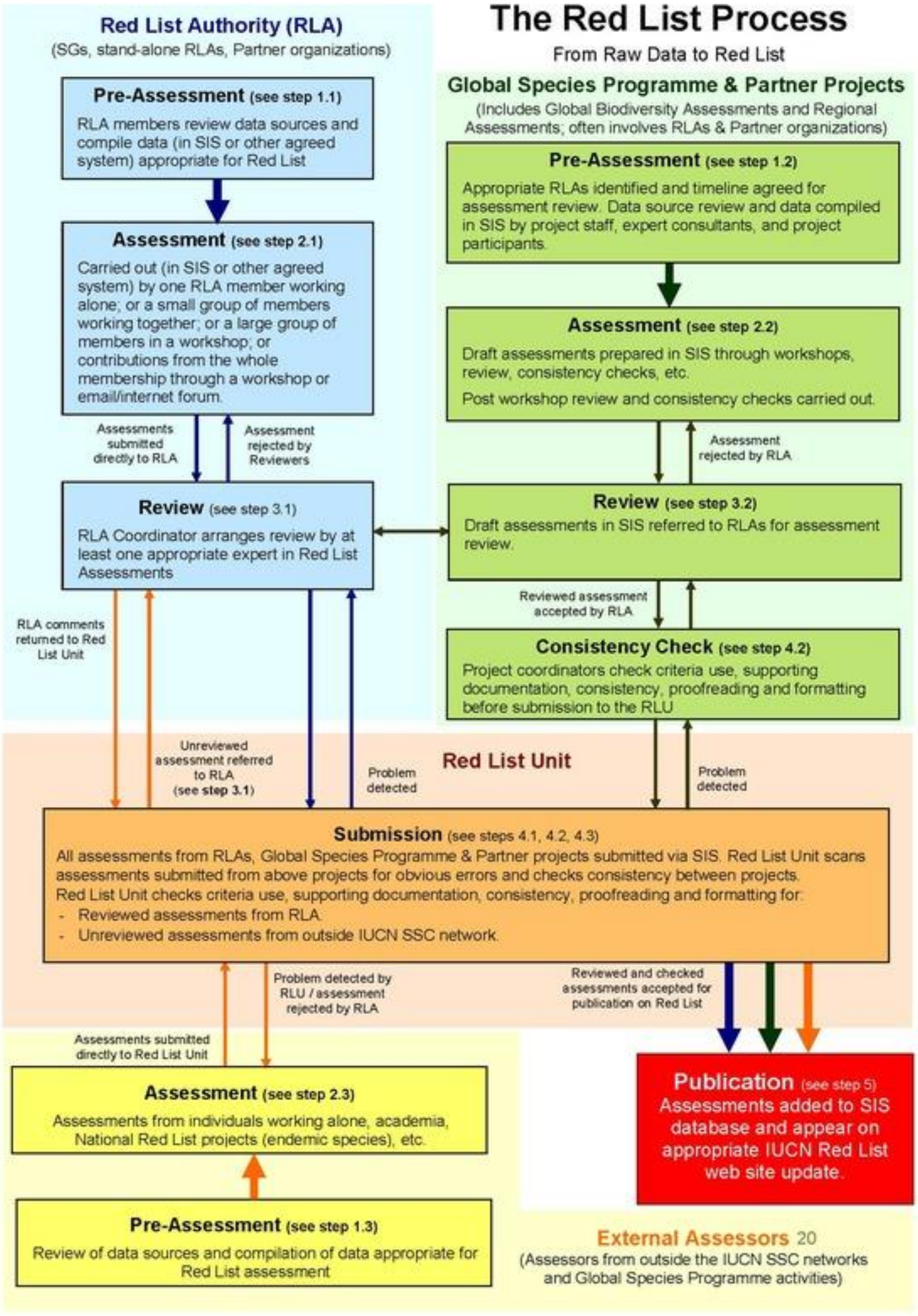

Red List Authority (RLA)
(SGs, stand-alone RLAs, Partner organizations)
The Red List Process
From Raw Data to Red List
Global Species Programme & Partner Projects
(Includes Global Biodiversity Assessments and Regional Assessments; often involves RLAs & Partner organizations)
Pre-Assessment (see step 1.1)
RLA members review data sources and compile data (in SIS or other agreed system) appropriate for Red List
Pre-Assessment (see step 1.2)
Appropriate RLAs identified and timeline agreed for assessment review. Data source review and data compiled in SIS by project staff, expert consultants, and project participants.
Assessment (see step 2.1)
Carried out (in SIS or other agreed system) by one RLA member working alone; or a small group of members working together; or a large group of members in a workshop; or contributions from the whole membership through a workshop or email/internet forum.
Assessment (see step 2.2)
Draft assessments prepared in SIS through workshops, review, consistency checks, etc.
Post workshop review and consistency checks carried out.
Assessments submitted directly to RLA
Assessment rejected by Reviewers
Assessment rejected by RLA
Review (see step 3.1)
RLA Coordinator arranges review by at least one appropriate expert in Red List Assessments
Review (see step 3.2)
Draft assessments in SIS referred to RLAs for assessment review.
Reviewed assessment accepted by RLA
RLA comments returned to Red List Unit
Consistency Check (see step 4.2)
Project coordinators check criteria use, supporting documentation, consistency, proofreading and formatting before submission to the RLU
Unreviewed assessment referred to RLA (see step 3.1)
Problem detected
Red List Unit
Problem detected
Submission (see steps 4.1, 4.2, 4.3)
All assessments from RLAs, Global Species Programme & Partner projects submitted via SIS. Red List Unit scans assessments submitted from above projects for obvious errors and checks consistency between projects.
Red List Unit checks criteria use, supporting documentation, consistency, proofreading and formatting for:
- Reviewed assessments from RLA.
- Unreviewed assessments from outside IUCN SSC network.
Problem detected by RLU / assessment rejected by RLA
Reviewed and checked assessments accepted for publication on Red List
Assessments submitted directly to Red List Unit
Assessment (see step 2.3)
Assessments from individuals working alone, academia, National Red List projects (endemic species), etc.
Publication (see step 5)
Assessments added to SIS database and appear on appropriate IUCN Red List web site update.
Pre-Assessment (see step 1.3)
Review of data sources and compilation of data appropriate for Red List assessment
External Assessors 20
(Assessors from outside the IUCN SSC networks and Global Species Programme activities)

Figure 1: Flowchart of IUCN Red List assessment process. This figure was gathered from https://www.iucnredlist.org/assessment/process

Dataset:

This dataset is accessed from Kaggle

Direct link: https://www.kaggle.com/nationalparkservice/park-biodiversity

Context:

This data is compounded of multiple national parks under the National Park Service that record data on animal and plant species. There is a National Park Species Portal that displays all the information for species used in this study.

Content:

The beginning shape of the datasets were (56,6) for the parks dataset and (119248,14) for the species dataset.

The features included in the parks dataset are:

Park Code: Abbreviation of park name

Park Name: Full name of the national park

State: State abbreviation of where the park is located

Acres: Total acreage included in property

Latitude: Numerical latitude of park location

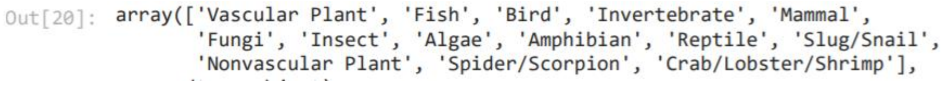
Longitude: Numerical longitude of park location

The features included in the species dataset are:

Species ID: The identification number for the individual species

Park Name: Full name of the national park

Category: The type of plant or animal



```
Out[20]: array(['Vascular Plant', 'Fish', 'Bird', 'Invertebrate', 'Mammal',
                'Fungi', 'Insect', 'Algae', 'Amphibian', 'Reptile', 'Slug/Snail',
                'Nonvascular Plant', 'Spider/Scorpion', 'Crab/Lobster/Shrimp'],
```

Order: The order of the scientific name

Family: The family of the scientific name

Scientific Name: The genus and species of the organism

Common Names: Any common name known

Record Status: Approved, In Review, or None

Occurrence: Present, Not Confirmed, Not Present (False Report), Not Present (Historical Report), Not Present

Nativeness: Native, Not Native, Unknown

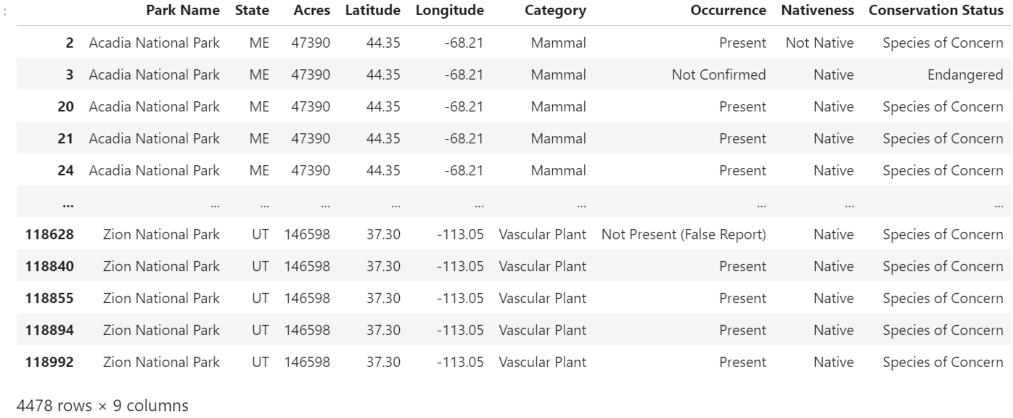
Abundance: Whether or not the species is seen often, to rare

Seasonality: When the species can be found in the park

Conservation Status: Where the species falls on the IUCN Red List

Initial Assessment of Data:

To answer the statistical questions being asked, the Conservation Status is the target feature. The park and species dataset was merged to overlay State, Acre, and Latitude and Longitude of each park for each organism.



	Park Name	State	Acres	Latitude	Longitude	Category	Occurrence	Nativeness	Conservation Status
2	Acadia National Park	ME	47390	44.35	-68.21	Mammal	Present	Not Native	Species of Concern
3	Acadia National Park	ME	47390	44.35	-68.21	Mammal	Not Confirmed	Native	Endangered
20	Acadia National Park	ME	47390	44.35	-68.21	Mammal	Present	Native	Species of Concern
21	Acadia National Park	ME	47390	44.35	-68.21	Mammal	Present	Native	Species of Concern
24	Acadia National Park	ME	47390	44.35	-68.21	Mammal	Present	Native	Species of Concern
...	...	...	...	...	...	...	...	...	...
118628	Zion National Park	UT	146598	37.30	-113.05	Vascular Plant	Not Present (False Report)	Native	Species of Concern
118840	Zion National Park	UT	146598	37.30	-113.05	Vascular Plant	Present	Native	Species of Concern
118855	Zion National Park	UT	146598	37.30	-113.05	Vascular Plant	Present	Native	Species of Concern
118894	Zion National Park	UT	146598	37.30	-113.05	Vascular Plant	Present	Native	Species of Concern
118992	Zion National Park	UT	146598	37.30	-113.05	Vascular Plant	Present	Native	Species of Concern

4478 rows × 9 columns

Figure 2: View of initial merged dataset before additional preprocessing.

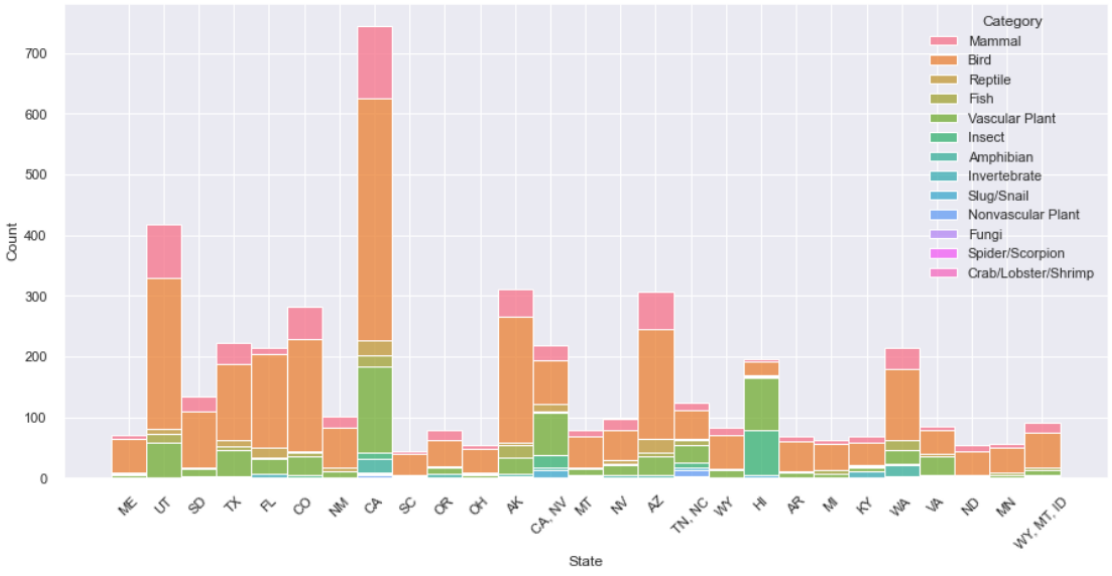


Figure 3: Barplot counting the total number of species per state with hue based on category of species.

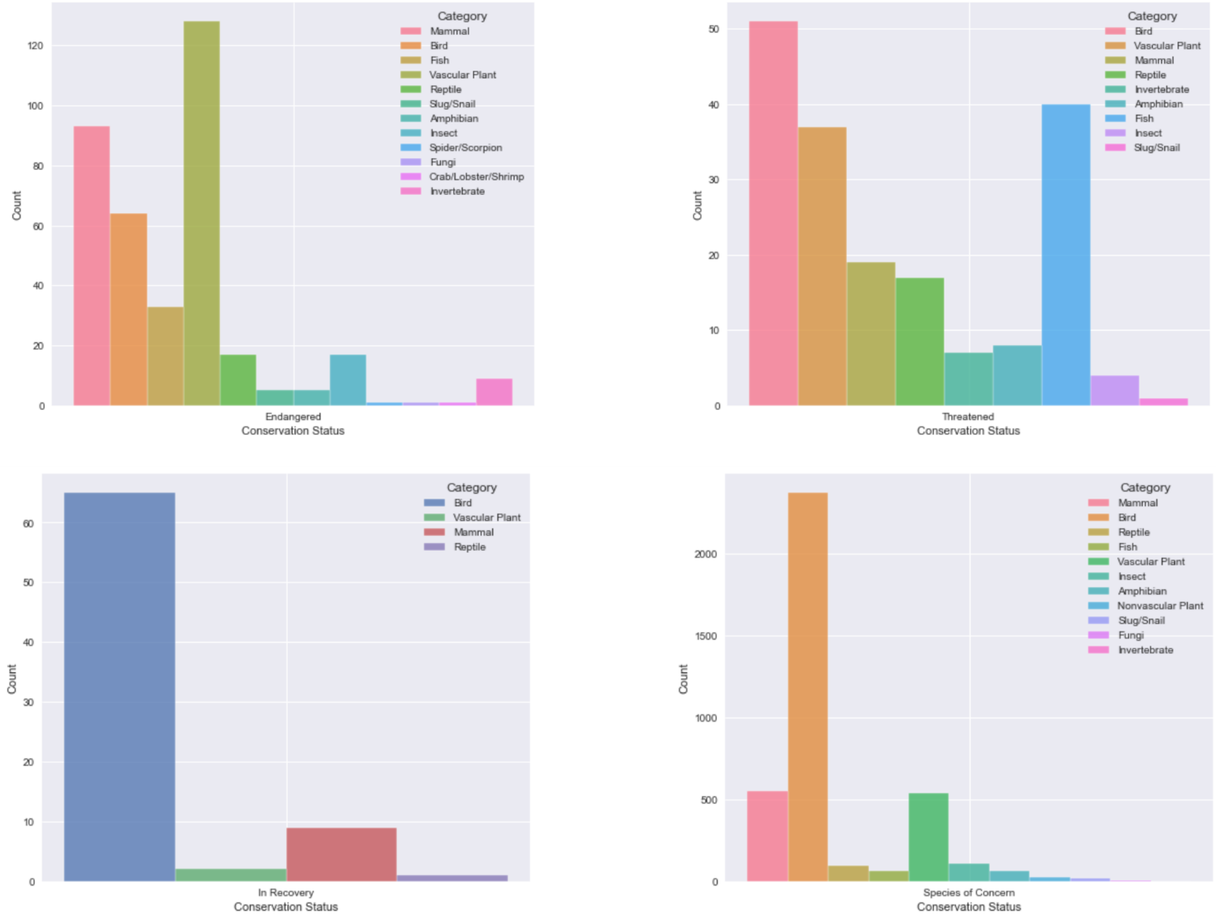


Figure 4: Subplotted barplots showing the number of species within each IUCN status. Hue is based on category of species.

It brings up some alarm that only birds, vascular plants, mammals, and reptiles are in recovery when the other conservation status' have many more species categories within them.

Feature Selection:

Two feature selection techniques used in this analysis are correlation using one-hot encoding and chi-square test. Since most of the columns are categorical, testing using one-hot encoding as well as chi-squared to confirm or deny similar results. Category and nativeness was encoded, and used the label encoder for park name, state, and conservation status.

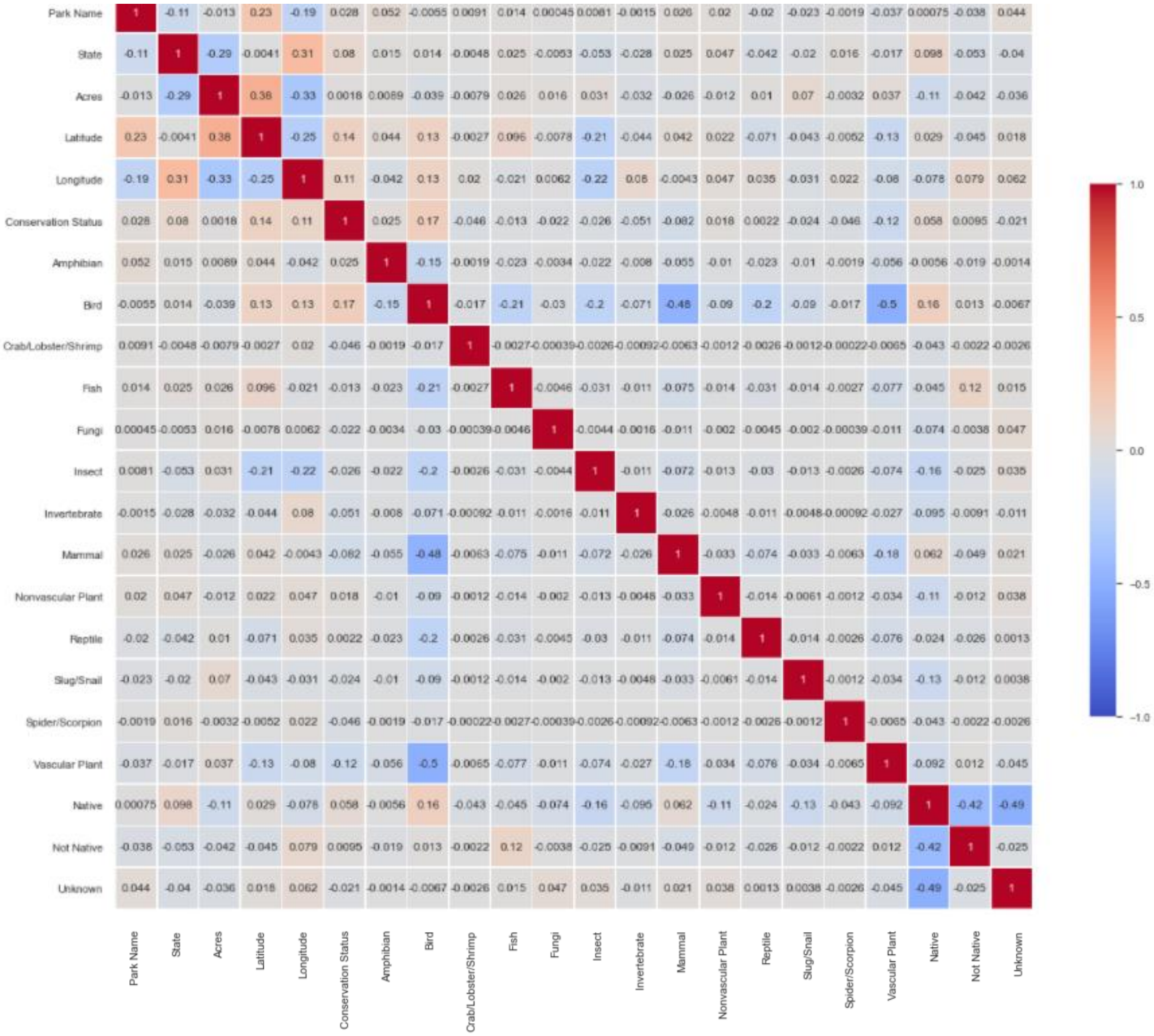


Figure 5: Heat matrix of all features. Significant correlations include conservation status, latitude, longitude, acres, state, park name, native, not native, unknown, slug/snail, nonvascular plant, mammal, insect, bird, reptile.

Correlation refers to a relationship between variables. The goal of feature selection is to disregard variables with no relationships or that are independent of all other variables. Variables mentioned within Figure 5 figure caption all have a relationship to another variable to some degree. Further assessments need to be done to confirm a relationship with the target variable, conservation status.

Since the correlation geared for numerical data a chi-squared test would help confirm correlation results since the chi-squared test is made more for categorical data. Latitude and longitude were removed as there was an error that there could not be negative values within the chi-squared test.

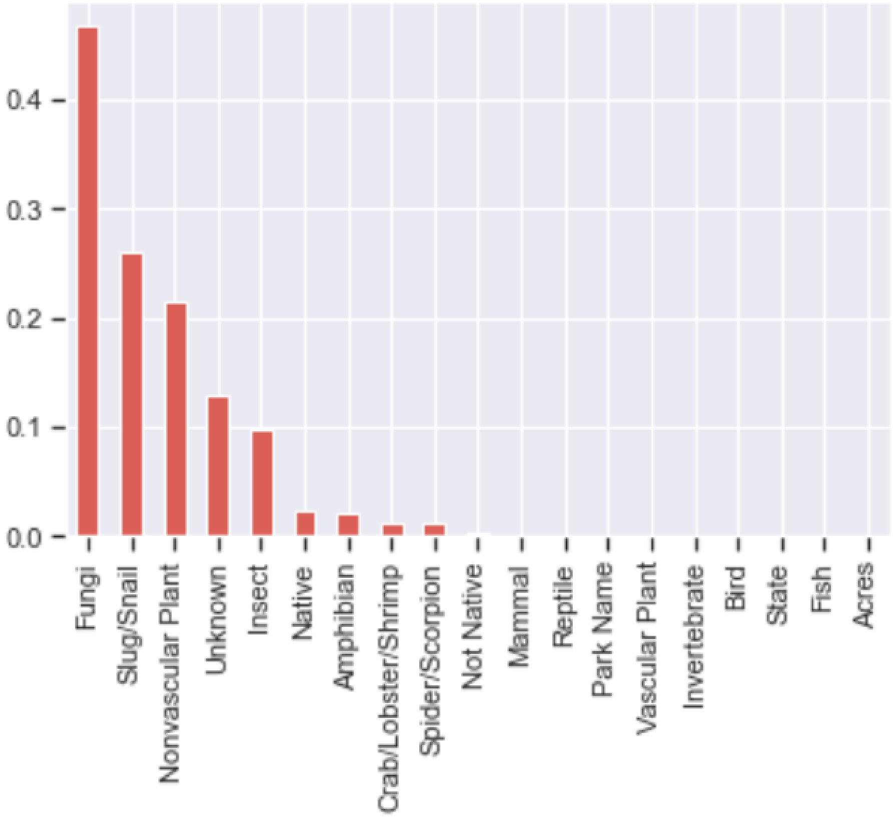


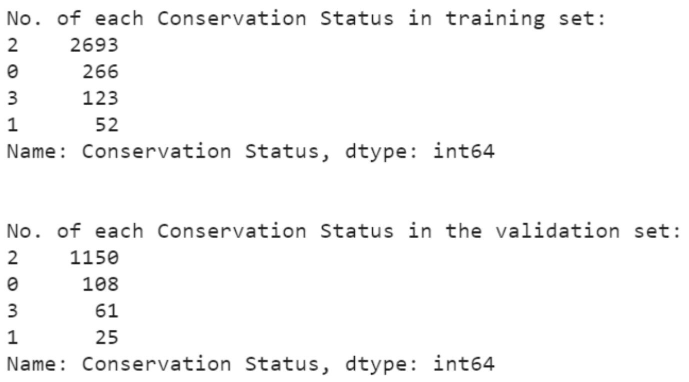
Figure 6: Barplot representation of Chi-Squared results. Features were tested against conservation status. High p-values were observed in fungi, slug/snail, nonvascular plant, unknown, and insect.

With the p-values of fungi, slug/snail, nonvascular plant, unknown, and insect being higher (above 0.05), it is suggested that these values are independent of conservation status. Since these features are considered independent of conservation status, they do not need to be included since the statistical questions being asked are based upon the conservation status.

In the heatmap above, the columns listed were combinations of all other columns, hence why there is some overlap. While the chi-squared test indicates Slug/Snail is independent of Conservation Status, it does have a slightly negative correlation with Native. Just an example of why the results for each test differ slightly.

Testing and Training Split:

When splitting the dataset into testing and training sets, the training set resulted in 3134 samples while there were 1344 samples in the testing/validation set. Below describes the amount of each conservation label per set:



```
No. of each Conservation Status in training set:
2    2693
0     266
3     123
1      52
Name: Conservation Status, dtype: int64


No. of each Conservation Status in the validation set:
2    1150
0     108
3      61
1      25
Name: Conservation Status, dtype: int64
```

Level 0 indicates the “Endangered” conservation status. Level 1 is the species “In Recovery”. Level 2 are “Species of Concern”. Lastly, level 3 is “Threatened” species.

Logistic Regression Model:

The logistic regression run on the data resulted in unimpressive results. Additional models will be tested to find a better fit.

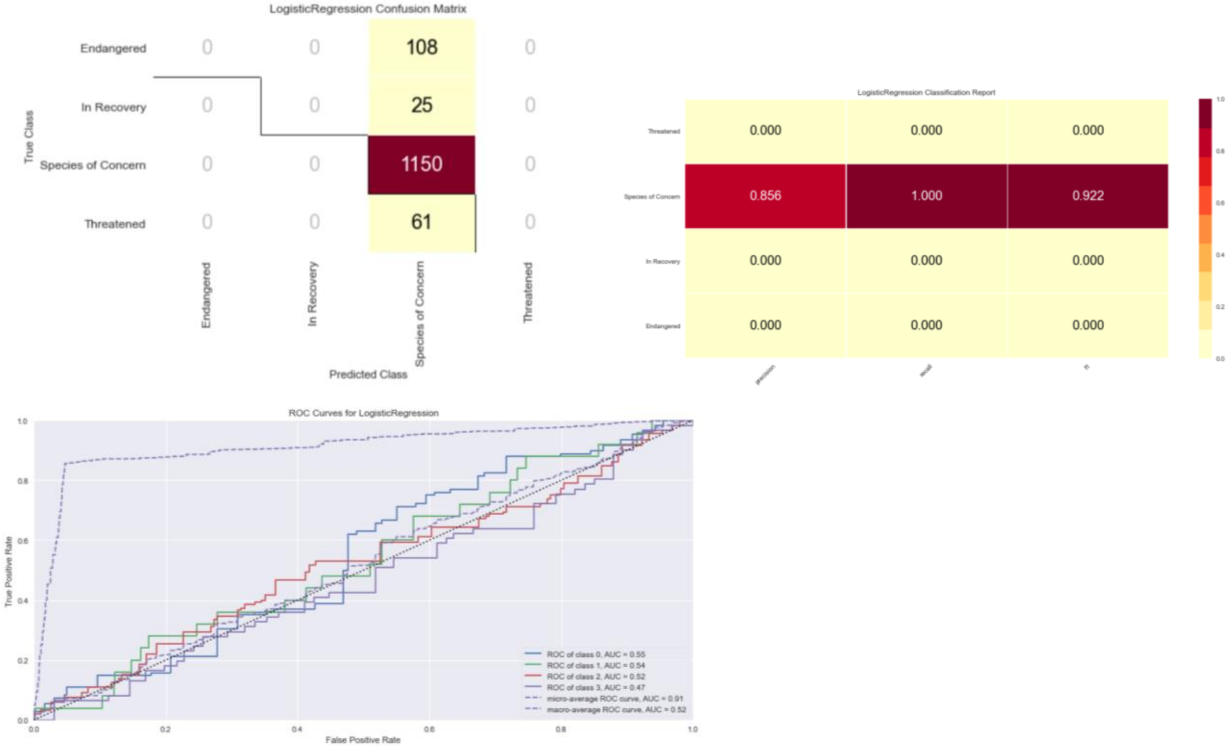
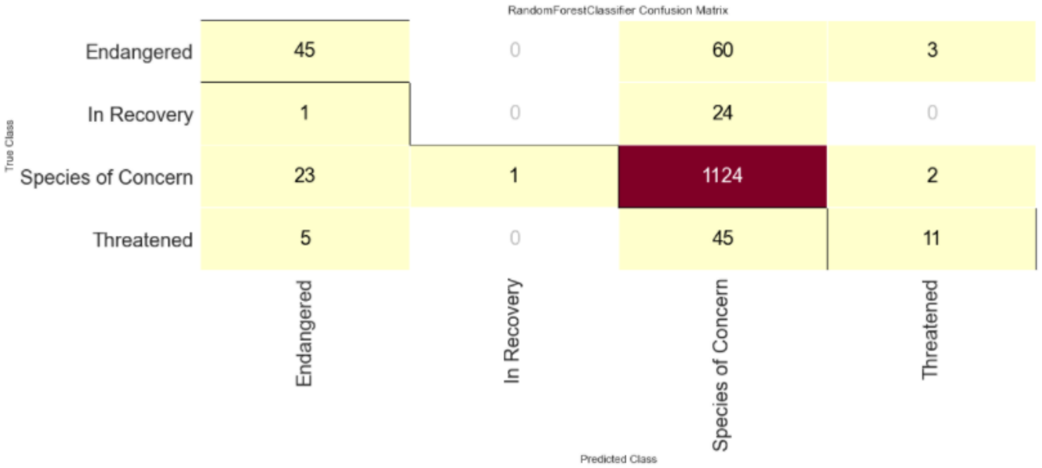


Figure 7: Comprehensive visualization of logistic regression. Species of concern was sole category to have results. ROC graph resulted in all other categories other than species of concern being along the average line.

Random Forest Model:



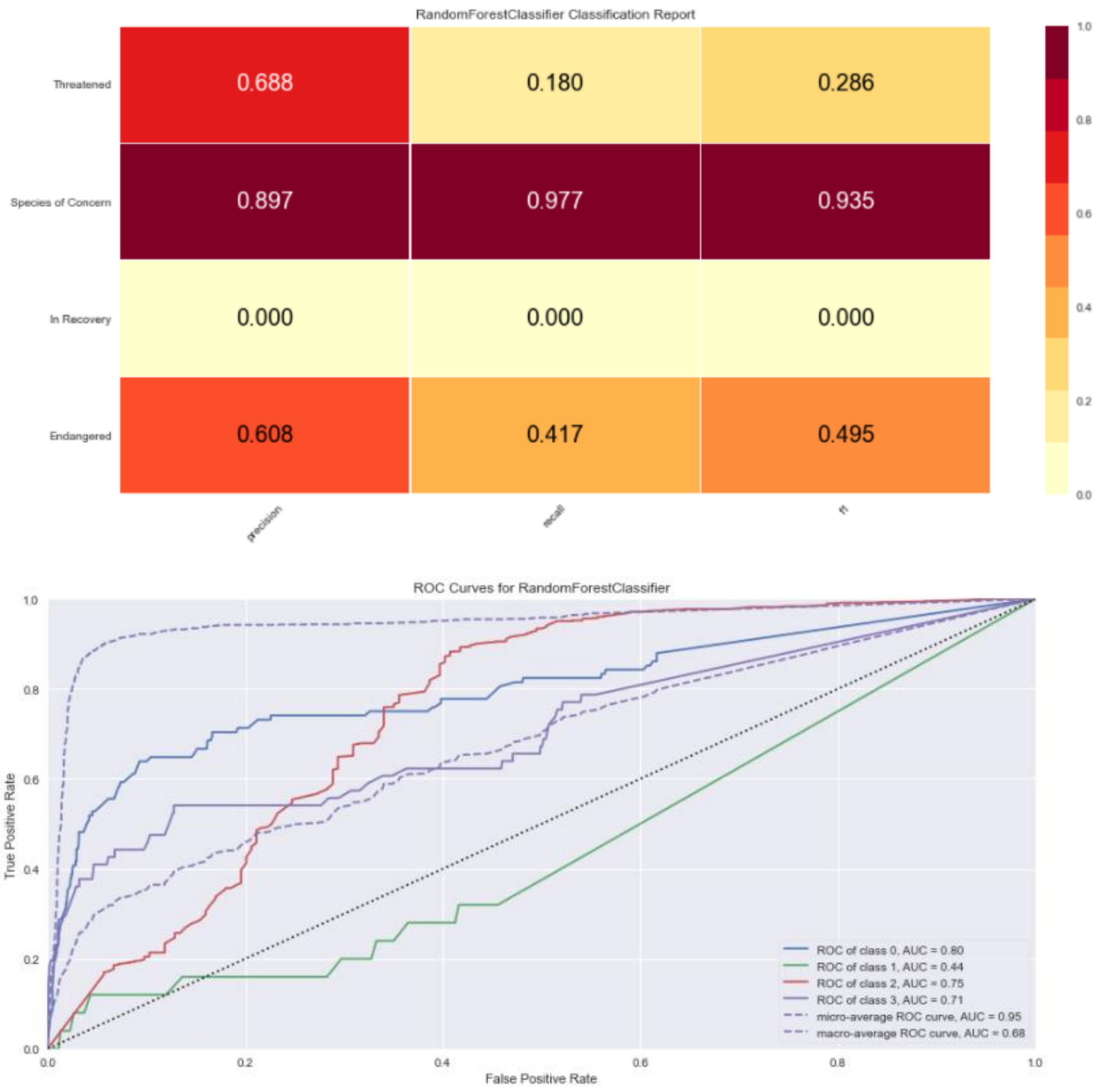


Figure 8: Random forest model results. Significant classification report with high p-values in every category.

The random forest model was found to have a better overall fit of the data and had better results in the classification report of precision, recall, and F1 score for all levels of conservation status but “In Recovery”. An inconclusive result for this level could be because the sample size for testing and training sets were below 100. The results of the logistic regression and the random forest suggest a nonlinear relationship between features and conservation status.

Hypertuning Random Forest:

The random forest classifier with 1000 estimators and a random state of 42 resulted in an accuracy of 87.79%. A random search was performed for hypertuning, which a parameter grid was created from. Then, a grid search was performed based on results of the random search. The ideal hyperparameters from the grid search are as follows: bootstrap set to False, max_depth of 20, max_features set to sqrt, min_samples_split of 8, and n_estimators 800. The accuracy with these hyperparameters decreased to 87.5%. The hypertuning resulted in a higher recall and F1 score for the threatened species, but a lower precision. In addition, all scores in the endangered species were lowered. Still, the species in recovery did not have enough data to result in any score from the random forest classifier.

Discussion:

Based on findings of this analysis, the largest amount of species found in the dataset resided in California. Of those, the majority of species were birds. Species that were not found to be related to conservation status were fungi, slug/snail, nonvascular plant, and insects. The largest category of conservation status was species of concern, also with birds as the highest count in this status. The status with the most variety of species categories is endangered species. The smallest status was in recovery with only four categories of species.

Because the entire dataset was narrowed down to 4,478 rows after data wrangling and cleaning, it was more difficult to fit models appropriately. A larger sample size with more even distribution of instances within classes would improve model fitting. The logistic regression model was the first model performed. The results of this model were limited, only showing an accurate prediction of conservation status in one status, species of concern. Since this was the largest status with 2,693 observations in the training set and 1,150 observations in the testing set, the model was able to fine tune this status more than the others which had smaller numbers of observations. In addition, the results of the logistic regression might indicate that there is a non-linear relationship within the data.

The second model performed on this data was a random forest. This model is more flexible with data than a logistic regression. It was able to make better predictions and a better overall result with an accuracy of 87.79%. The precision of the species of concern status increased from the logistic regression, as well as the threatened and endangered species both had a precision between 60 and 70%. While this model fit significantly better to the data than the logistic regression, the species in recovery still could not be predicted. This is most likely a result of the small sample of data. Hypertuning was performed on this model, however, the accuracy did not improve, so the original model was used.

Chi-squared tests were performed on both the category and state separate from the model analyses. In this, both category and state had a very significant p-value, 1.27^{-133} and 4.59^{-58} respectively. A p-value below 0.05 indicates a rejection of the null hypothesis that the feature is not associated with the target value. With this, both state and category reject the null hypothesis.

Conclusion:

The best fitting model for this dataset is the random forest model. Through association testing, both state and category rejected the null hypothesis that they are not associated with conservation status. Further studies can be done to prove association. Improvements would include more instances of species in recovery, threatened, and endangered. In addition, more accurate locations of latitude and longitude of species would also be beneficial versus the location of the national park. Inclusion of date would allow for a time series to evaluate the increase or decrease of species sightings.

Citations

Background & History. (n.d.). Retrieved from https://www.iucnredlist.org/about/background-history

Raw Data to Red List. (n.d.). Retrieved from https://www.iucnredlist.org/assessment/process