

A Machine Learning Approach to Predicting Species Extinction Risk

Yogith Kishore & Robert Chun

Received July 14, 2025

Accepted October 18, 2025

Electronic access October 31, 2025

Despite the consequence of species going extinct not currently being visible, its impact in our future can be devastating to biodiversity, ecosystem stability, and our own society. Predicting extinction rates and identifying the primary causes of a species decline are crucial first steps to improving our conservation efforts. An accurate understanding and insight into factors driving species extinction can equip conservationists with the tools needed to preserve endangered species. As the number of extinct species continues to rise, a variety of issues become apparent. As species decline, so does biodiversity. With reduced biodiversity, many species – and by extension humans – could suffer from reduced resilience, causing us to be more easily influenced by diseases and climate change. To protect species at risk of extinction, conservationists need to identify which species are most vulnerable by understanding the factors that drive their decline. Utilizing AI models is a crucial first step in this process. This study utilizes the Random Forest Regressor model, trained on Animal Information Dataset published on Kaggle containing characteristics across a range of species, to help conservationists understand which species are at a greater risk of extinction¹. To ensure robustness, our model was tested using k-fold cross-validation to calculate RMSE values, such as those for the Amur Tiger and Alaotra Grebe being 2.94 and 8.85 years respectively. The model provides a relative timeline for species decline, allowing conservationists to more strategically allocate their resources. This paper explores the development of a prediction model designed to predict extinction rates, ultimately aiding conservationists in their efforts to protect endangered species.

Introduction

The decline and eventual extinction of species has become a major issue in modern times. A variety of species have gone extinct recently; some such species being the Alaotra Grebe which went extinct back in 2012, the Bramble Cay Melomys in 2016, and the Maui Akepa in 2018^{2,3}. The permanent loss of species can lead to ecosystem destruction which can cause irreversible damage to our planet and negatively impact our lives. Due to the structure of ecosystems, every species has a role to play, such as in food production and resource accessibility, making their extinction have a direct negative impact on us. When a species goes extinct, they cause an imbalance in the food chain⁴. The predators that used to consume them now have a reduced food source, which could lead to their starvation. Without natural predators, prey can grow uncontrollably, leading to imbalances in the ecosystem. As the number of species declines, so does biodiversity⁴. With reduced biodiversity, many species – and humans – could suffer from reduced resilience, causing us to be more easily influenced by diseases and climate change. Some human-induced extinction factors include climate change, pollution, and changes in habitat⁵. Understanding these factors is key to understanding extinction events. Not all extinction factors are man-made, yet people can still mitigate their effects. For example, an abundance of predators, a lack of resources, or the

inability to reproduce all play significant roles, and are a few examples of natural extinction causes.

Many modern prediction models related to extinction prediction focus mainly on a very specific species⁶. An underworked part of extinction prediction models is creating a general model that could give extinction dates for a variety of animals. Unlike standard extinction prediction models that focus on a single species, this model uses a multi-species dataset to predict extinctions across a wider taxonomic range.

Related Works

Comparative studies have shown species traits can be associated with extinction risk⁵. Across taxa, characteristics such as offspring production, taxonomic group, and social group size have correlations with extinction risk⁷. These results helped narrow down our search for datasets when deciding on characteristics to utilize for training the model.

Knowing this, past research has been done to approximate the International Union for Conservation and Nature (IUCN) Red Lists threat categories for species that are Data Deficient^{6,8}. Approaches typically tend to combine traits to identify species that are most at risk, using feature importance to drive their predictions^{6,9}.

This paper seeks to follow trait-based ML predictions but in a different way: developing a general species extinction prediction model's to predict lifespans rather than IUCN statuses¹⁰.

¹ Westview High School.

² San Jose State University

Methods

To predict the extinction date of species, there are several key steps. Firstly, identifying the best type of model for the most accurate prediction. Next, collecting proper data to make accurate predictions. Lastly, training the model on the data to predict their relative extinction date.

To decide on the best type of model, a variety of different models were researched to better understand their pros, cons, and varying use cases. After understanding multiple different models such as XGBoost, K-Nearest Neighbors (KNN), and Linear Regression, the model that was chosen for this research was the Random Forest Regressor model^{11,12}.

Unlike some of the previously mentioned models, Random Forest Regressor excels at handling smaller amounts of data, whereas Linear Regression which might underfit due to its simplicity or XGBoost which can overfit due to recognizing quirks rather than general patterns^{11,13}. In addition, it excels at handling large amounts of categorical data where other models such as Linear Regression and KNN struggle^{11,14}. Finally, due to its use of multiple decision trees, it is able to more easily simulate complex ecological relationships by better accounting for various characteristic combinations compared to other model types.

In short, RandomForestRegressor splits testing data into several decision trees, unlike other models that just rely on one decision tree. By using multiple decision trees and comparing them to one another, the model enhances its ability to make accurate predictions¹¹. Then, when the model is supplied with data, it can compare the input data with the preprocessed data to find patterns to make predictions of the expected outcome.

To make predictions, adequate data on the topic is needed. After extensive searching, we selected the dataset, Animal Information Dataset, published by Sourav Banerjee on Kaggle, an online platform used for data science competitions that simultaneously hosts a large library of free use data. This specific dataset was chosen as it contained 16 characteristics spanning 206 species which would allow the model to be usable in a general setting. The dataset spans a wide range of taxa – such as mammals, reptiles, birds, and aquatic species – allowing the model to train on a diverse set of ecological traits. However, a majority of the data have lifespans under 30 years which could overrepresent short-lived species.

The data is often unusable in its base form, as it is not tailored towards the model. As such, preprocessing the data is required. In the end, the columns that remained, which represent the training characteristics, were Animal, Height (cm), Weight (kg), Color, Lifespan (years), Diet, Habitat, Predators, Average Speed (km/h), Countries Found, Conservation Status, Family, Gestation Period (days), Top Speed (km/h), Social Structure, and Offspring per Birth. These characteristics can be found in Figure 1. The model utilizes these combined characteristics to

	A	B	C	D	E	F	G	H
1	Animal	Height (cm)	Weight (kg)	Color	Lifespan (years)	Diet	Habitat	Predators
2	Aardvark	105-130	40-65	Grey	23	Insectivore	Savannas, Grasslands	Lions, Hyenas
3	Aardwolf	40-50	8-14	Yellow-brown	8	Insectivore	Grasslands, Savannas	Lions, Leopards
4	African Elephant	270-310	2700-6000	Grey	65	Herbivore	Savanna, Forest	Lions, Hyenas
5	African Lion	80-110	120-250	Tan	12	Carnivore	Grasslands, Savannas	Hyenas, Crocodiles
6	African Wild Dog	75-80	18-36	Multicolored	11	Carnivore	Savannas	Lions, Hyenas
7	Alpine Ibex	67-101	19-120	Brown	12	Herbivore	Mountains, Alpine	Wolves, Golden Eagles
8	Amazon Rainforest Frog	2-13	Up to 0.5	Various	4	Insectivore	Amazon Rainforest	Birds, Snakes
9	American Bison	152-186	318-1,000	Brown	17	Herbivore	Grasslands, Plains	Wolves, Grizzly Bears
10	Anteater	52-91	22-41	Brown, White	13	Insectivore	Grasslands, Forests	Jaguars, Pumas
11	Asian Horse	140-160	380-1000	Various	27	Herbivore	Middle East, North Africa	Humans, Predators
12	Arabian Oryx	70-90	65-90	White	17	Herbivore	Desert	Lions, Leopards
13	Arctic Fox	25-30	2.5-9	White	5	Omnivore	Tundra	Polar Bears
14	Arowana	Up to 120	Up to 6.7	Silver, Gold, Green	13	Carnivore, Piscivore	Freshwater Rivers, Lakes	Birds, Larger Fish
15	Asian Elephant	200-300	2000-5000	Grey	60	Herbivore	Grasslands, Forests	Tigers, Leopards
16	Atlantic Puffin	25-30	500-620	Black, White	23	Carnivore, Piscivore	North Atlantic, Arctic	Gulls, Birds of Prey
17	Atlantic Spotted Dolphin	200-250	120-140	Gray, White	25	Carnivore	Oceans, Coastal Areas	Sharks, Orcas
18	Australian Fur Seal	112-160	18-40	Brown, Gray	20	Carnivore	Coastal Waters	Sharks, Orcas
19	Axolotl	Up to 30	Up to 300	Various	13	Carnivore	Lakes, Canals	Fish, Birds
20	Aye-Aye	35-37	2.2-2.7	Black, Brown	15	Omnivore, Insectivore	Rainforests	Birds of Prey
21	Baird's Tapir	76-107	150-400	Brown, Black	27	Herbivore	Rainforests, Grasslands	Jaguars, Crocodiles
22	Bald Eagle	70-102	3-6	Brown, White	25	Carnivore	Forests, Lakes, Coasts	Wolves, Raccoons
23	Banded Palm Civet	41-71	2-5	Brown, Black, White	13	Omnivore	Forests	Birds of Prey
24	Barbary Macaque	Up to 75	5-13	Brown, Gray	17	Herbivore, Omnivore	Forests, Mountains	Leopards, Eagles
25	Basking Shark	Up to 1100	400-700	Gray, Brown	25	Carnivore	Oceans	Orcas, Great White Sharks
26	Bearded Dragon	Up to 60	Up to 600	Various	10	Omnivore	Deserts, Woodlands	Birds, Snakes
27	Bengal Fox	35-40	2-4	Yellowish-gray	9	Omnivore	Grasslands	Wolves, Birds of Prey
28	Bengal Tiger	90-110	220-260	Orange, Black	13	Carnivore	Grasslands, Mangroves	Humans, Crocodiles
29	Black Rhinoceros	132-180	800-1,400	Gray, Black	43	Herbivore	Grasslands, Forests	Lions, Hyenas
30	Blotfish	Up to 30	Up to 10	Pinkish-gray	35	Carnivore	Deep-sea	Not Applicable
31	Blotfish	Not Applicable	Not Applicable	Pinkish-gray	35	Carnivore	Deep-sea	Not Applicable
32	Blue Jay	22-30	0.07-0.12	Blue, White	8	Omnivore	Forests, Urban Areas	Birds of Prey, Snakes
33	Blue Morpho Butterfly	Up to 15	Up to 0.028	Blue	0.5	Herbivore	Rainforests	Birds, Bats

Fig. 1 Re-formatted data from Animal Information Dataset published by Sourav Banerjee on Kaggle.

formulate accurate predictions. In addition to removing certain unnecessary columns, minor edits to the data were necessary to ensure usability such as removing hyphens for certain characteristics, changing capitalization, and other small data fixes. Feature engineering steps were taken to remove all null values from the data, re-format categorical inputs (e.g. USA and US), and apply one-hot encoding to avoid ranking categorical data in a similar style to numerical values^{15,16}. When using data to make predictions, it is also important to classify pieces of data as numerical or categorical. Numerical is data that uses numbers (integers & floats), while categorical uses words (strings). The model must know which characteristics use numerical and categorical, as the model treats them differently when making predictions. Numerical data can be compared much more easily, as the model can easily compare different numbers to each other. Since the model used in this research was Random Forest Regressor, which utilizes decision trees which are not sensitive to feature magnitude, no feature scaling was used¹¹.

Scikit-learn version: 1.6.1 was used with the default hyperparameters, some of which included n_estimators: 100 (number of trees), criterion: squared_error (how splits are evaluated), min_samples_split: 2 (minimum samples needed to split a node), and many others¹⁵. While this study did not focus on hyperparameter tuning due to the purpose of the research being to identify the feasibility of a general extinction prediction model, future work could explore optimizing these values to further improve accuracy in the model's predictions. Categorical data is treated differently as words cannot be compared with one another the same way numbers can. Assigning numerical values to countries could help improve accuracy. For example, assigning the US and USA the value of 1 can help to classify the United States of America as just one piece of data, rather than two pieces of data, as in categorical terms, the US and USA would act as different values despite meaning the same thing. Accurately classifying data as numerical and categorical before running the model can significantly boost the model's

accuracy. In this study, one-hot encoding was utilized to prevent the implicit ranking of categories.

Not all the data is used to train the model. The data is split into a training and testing category. By using all the data on training, the prediction model may be too rigid, and struggle making predictions when the input data is not what is directly from the training data. This is known as overfitting and can lead to very inaccurate predictions due to the model not being flexible enough to make accurate predictions. Underfitting, on the other hand, works in the opposite way. Using too little training data can cause the pattern recognition to be weak, causing the model's accuracy to suffer. The most common train-test splits are 80% train and 20% test, 67% train and 33% test, and 50% train and 50% test¹⁷. Finding a balance between splitting the training and testing data is key to accuracy and consistency of the model. When running the model, the model trains using the training data, and then tries making predictions using the testing data, checking its predictions with the actual values of the testing data.

To make predictions of species expected extinction dates, there are a series of steps that need to be taken. Firstly, it is necessary to separate the characteristics into X and Y datasets. The Y represents what information that is being predicted, and in this case, the lifespan of the species, as that can be used to predict extinction dates using basic math. The rest of the characteristics mentioned earlier were put into the X list. As previously mentioned, the X and Y datasets were split into training and testing data. After testing the model, a split of 80% training data and 20% testing data worked best for this specific experiment when trying to make accurate predictions. Then the training data is used to fit the model, causing the model to learn about the data and create the patterns mentioned earlier. Now, when inputting new data (all characteristics mentioned earlier excluding lifespan) about animals not previously in the dataset, the model can recognize patterns in the data to give estimations about expected lifespans of certain species. With the predicted lifespan, the following information can be calculated:

- Death per year = Population / lifespan
- Birth Per Year = Offspring Per Birth * #Of Times Reproducing Per Year
- Rate of Change of the Population of Species = Births Per Year - Deaths Per Year

With this information, it is possible to graph an exponential equation which can represent the population of species by using the equation

$$\text{Future Population} = \text{Current Population} \times \text{Growth Rate}^{\text{Time}}$$

where the X-axis represents time and the Y-axis represents the population. There are two methods that can be employed to

check for accuracy of the model, checking the root mean squared error (RMSE) or the R-squared value. The mean squared error is a way of quantifying the overall error of the model. To do this, it averages the squared differences between the predicted and actual values of the model. By utilizing sklearn's `cross_val_score` function, we were able to calculate the RMSE value for both species. This was done by performing 10-fold cross-validation and computing the negative mean squared average for each fold, before averaging those results and taking the square root to obtain the RMSE^{15,17-20}. For the Amur Tigers and Alaotra Grebe graphs that are shown in the Results section of this paper, the RMSE values of 2.94 and 8.85 were obtained respectively. Considering the lifespan range in the dataset spanned from 0.25 to 125 years, an error of 2.94 and 8.85 years is a relatively small portion of the total range. This suggests that the model performs well when predicting species lifespans. The R-squared (R^2) value measures the amount of error that can be explained by the independent variables¹⁹. By using scikit-learn's `r2_score` function, the R^2 value was calculated as one minus the sum of squared differences between the predicted and actual values divided by the sum of the squared differences between actual values and their mean. An R^2 value closer to 1 shows that the model accounts for a greater variance in the data, while a value closer to 0 shows that the model performs the same as predicting the mean. For the Amur Tiger and Alaotra Grebe, they got an R^2 of 0.84 (83.76%) and 0.85 (84.67%) respectively. This signifies that in both cases, the model explains over 83% of the variation in lifespan, demonstrating a strong prediction.

Since the RMSE and R^2 values highly depend on the training data, prediction parameters, and range of values, comparing them across models can be misleading. As such, a direct comparison between these values was avoided in this study¹⁹.

Results

To test the Random Forest Regressor model, the extinction dates of multiple species were predicted to test for accuracy and validity. The Amur Tiger (otherwise known as the Siberian Tiger) is a species of tiger that is currently critically endangered. Based on prediction trends, the Amur Tigers were expected to go extinct around the year 2055 as shown in Figure 2. Due to the graph's exponential shape when relating time and population, the population numbers will never reach 0. However, estimations can be made on when species may go extinct based on how low their populations become. Based on all the tested characteristics being input to the model, the Amur Tigers are predicted to go extinct around that time frame, unless major action gets taken to support them. To validate these findings, two series of tests were run.

The first test took the Amur Tigers population from 15 years ago, and tried to predict its current population, assuming none of the factors affecting their growth or decline changed. This

Predictions for new data (Lifespan in years): [13.89]
 Average RMSE: 15.50090171984873
 R² score (variance explained): 0.8783750766361967
 Percent variance explained: 87.83750766361968

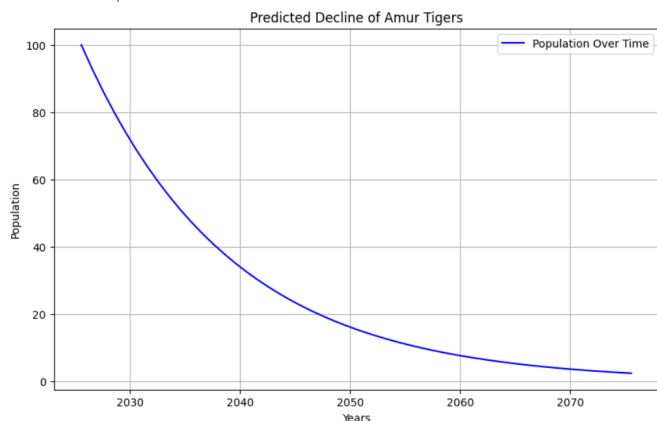


Fig. 2 Graph showing predicted population of Amur Tigers over the span of 50 years

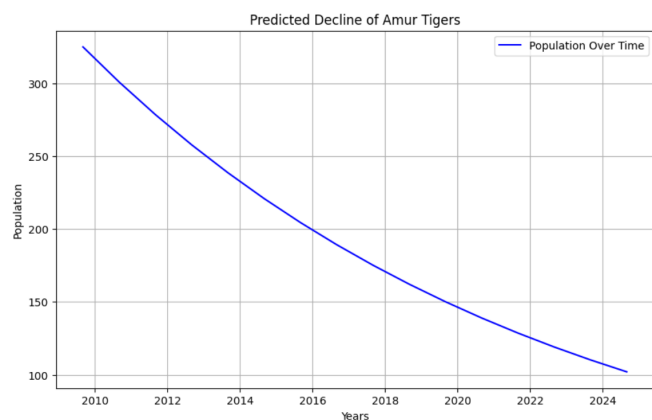


Fig. 3 The graph shows the predicted population of Amur Tigers in the modern day based on data from 15 years ago

graph shows that by following the trend of their population 15 years ago (325 Amur Tigers), a prediction can be made for the present day: there should be roughly 100 tigers left.

The second test tried to predict the rough extinction date of an already extinct species. To do this, a prediction was made for the Alaotra Grebe, a species of bird which unfortunately went extinct in 2012. Their population from 50 years ago was taken and used to predict their population to find when they may go extinct. Based on the exponential curve, a safe prediction is that their species may go extinct around 2015, as that is when their population numbers become dangerously low. With their actual date of extinction matching the predicted extinction, it is safe to assume that the model is relatively accurate when predicting extinction risk of species.

Luckily for Amur Tigers, Russia has been making conservation efforts to increase their population. Russia's efforts have been effective, as in recent times, the Amur Tiger population

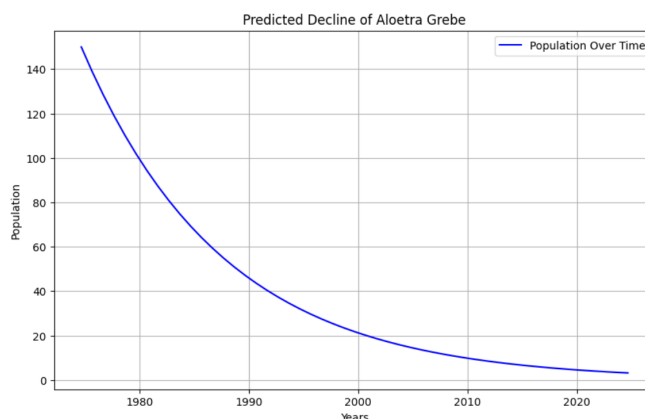


Fig. 4 Graph showing how accurate the model is by predicting the relative extinction date of the Alaotra Grebe, an already extinct species

Predictions for new data (Lifespan in years): [30.88]
 Average RMSE: 15.400797653545942
 R² score (variance explained): 0.8539628950421485
 Percent variance explained: 85.39628950421485

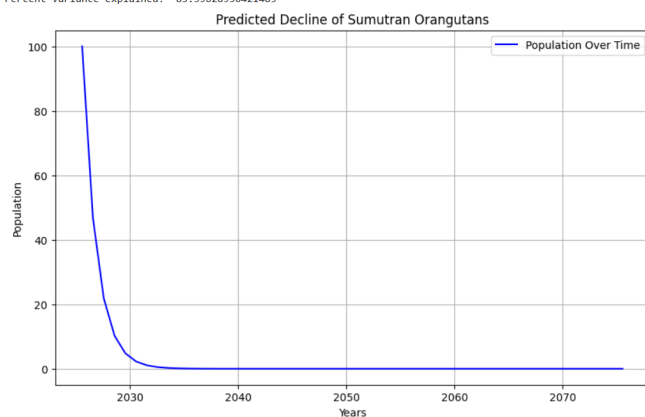


Fig. 5 Graph showing predicted decline of Sumatran Orangutans^{15,21}

has been steadily increasing. However, if Russia did not take the appropriate measures to maintain the tigers safety, these graphs reveal some potential scenarios that could have taken place. Figures 5 - 8 show example graphs of various species' predicted population over time graphs. In addition, understanding the factors behind species extinction is equally as important as understanding relative extinction dates. While analyzing the driving factors that led to reduced populations, feature importance revealed how the species family (e.g. Elephantidae, Felidae, Canidae, etc.), number of offspring per birth, and color were the strongest characteristics that drove extinction⁹. In contrast, traits such as social structure (solitary, herd-based, or group-based), top speed, and average speed had minimal effects relative to other characteristics when determining overall population extinction^{10,25}. These results suggest that ecological and reproductive factors play a more critical role than raw physical or behavioral traits in determining species survival.

While some features may be biologically important in survival

Predictions for new data (Lifespan in years): [23.96]
Average RMSE: 15.369476947792888
R² score (variance explained): 0.874906552697598
Percent variance explained: 87.4906552697598

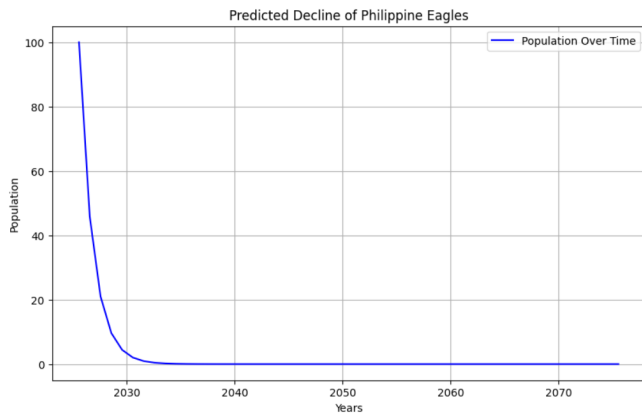


Fig. 6 Graph showing predicted decline of the Philippine Eagles^{15,22}

Predictions for new data (Lifespan in years): [56.02]
Average RMSE: 15.450944580227329
R² score (variance explained): 0.7937016406663014
Percent variance explained: 79.37016406663014

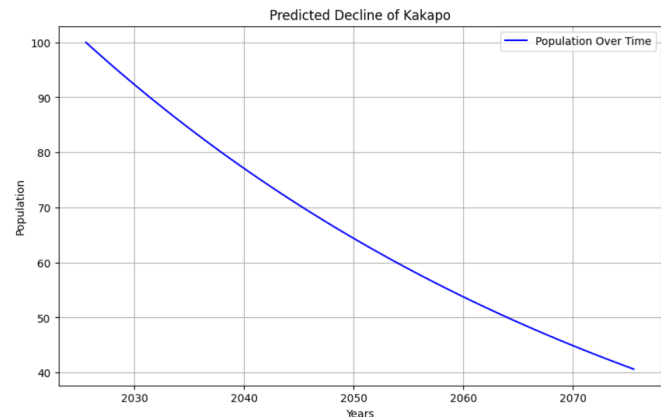


Fig. 8 Graph showing predicted decline of the Kakapo^{15,24}

Predictions for new data (Lifespan in years): [19.88]
Average RMSE: 15.59862665940624
R² score (variance explained): 0.8617747939338889
Percent variance explained: 86.1774793933889

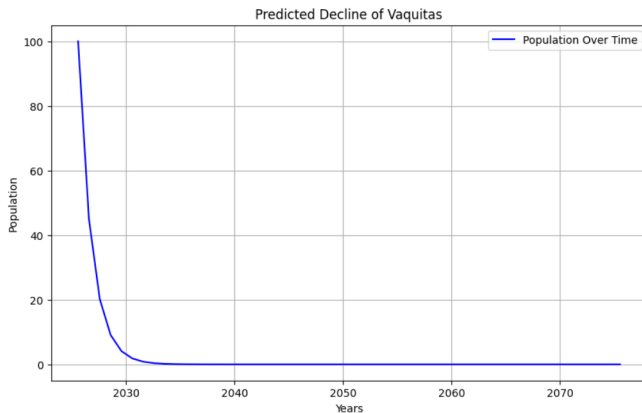


Fig. 7 Graph showing predicted decline of the Vaquita^{15,23}

such as social structure or speed, due to those traits showing reduced relative importance in this dataset, it may imply that there may be limitations in the data rather than those characteristics true ecological role.

Discussion

This research showcases the feasibility of a general prediction model, specifically Random Forest Regressor, to predict extinction risk across a diverse range of species. The models alignment with both historical extinction (e.g. Alaotra Grebe) and future risk species (e.g. Amur Tiger) suggests accurate predictive findings.

The consistency between the model's predictions and real world conservation efforts (e.g. Russia's emphasis on Amur Tiger conservation) support the models accuracy. However, these results suggest correlations between characteristics and

extinction risk and are not meant to be used without proper ecological expertise. Additionally, the model's ability to predict across taxa is promising, but species-specific traits and unique ecological dynamics can still cause significant predictive error^{9,26}. While the model itself showcases reasonable accuracy in predicting the extinction date of species, it is difficult to compare it to a species-specific model as they serve distinct purposes. A species-specific model is specialized for a single organism and can produce higher accuracy predictions at the cost of time. A general model such as this one, on the other hand, can significantly reduce the time taken to make predictions on a wide range of taxa, allowing for better scalability, at the cost of reduced accuracy.

However, despite the model being accurate for many species, there are still several limitations that should be considered. One main factor is the variability in species ecosystem interactions. While the model does incorporate multiple biological factors such as reproductive rate, diet, and social structure, some species have unique biological responses that may potentially influence its extinction path in ways that may not be captured by the model. The model might oversimplify or completely omit crucial interactions between species, food accessibility, or disease prevalence which can vary widely even within a single habitat.

Additionally, the availability of data also posed serious limitations. Due to the limited data available to use, the model was most accurate when predicting the lifespan and extinction date of species with certain traits. When binned by lifespan, species with short lifespans (under 30 years) had a mean absolute error (MAE) of 1.21-1.53 but increased to 2.06 for species between ages 30 and 60 and increasing all the way to 7.01 for lifespans greater than 60 years. When creating bins based on conservation status, (e.g., least concern, critically endangered, etc.) the MAE was lower (0.96) when predicting species in more common categories (least concern) when compared to those in threatened

categories (e.g. Endangered: 2.25 or Vulnerable: 2.49). Many of these skews in prediction accuracy can most likely be attributed to the dataset having more available species in those specific categories, allowing for stronger predictions in those areas. Finally, due to the model having a set number of factors it is checking for and using to make predictions, it is difficult to account for unforeseen variables such as the impact of conservation areas, habitat loss/expansion, or natural disasters.

The potential impact of these findings can be used to highlight the role of predictive modeling in conservation biology. The ability to anticipate extinction timelines for species allows for preemptive action. By utilizing the insight of machine learning models, resources can be better allocated to more effectively and efficiently support species when needed. Models such as this one can add a data-driven perspective that complements conservationists expertise, helping to identify species that may experience rapid population declines without the proper support.

This study's model functions as an early warning system rather than a surefire predictor on species extinction risk. This model can act as a tool for biologists to use, complementing conservationists skills to better understand extinction trends. The predictive data generated can support the hypothesis that certain species are at a high risk of extinction. There are a few improvements that could be made to this project. Keeping track of more recent environmental impacts would greatly increase the accuracy of the prediction results. By incorporating factors relating to human impact such as global warming, pollution, and other modern factors, the results could be greatly influenced. With the increase in characteristics, the model would become more complex. In return, it would have increased accuracy as the model would be able to recognize patterns in the data more easily. By including additional factors, the model can better reflect the complex intricacies of modern environmental issues. Additionally, being able to link the predictions with real time systems such as satellite data could be used to make more dynamic predictions. By accounting for current weather conditions, changes in environments due to - deforestation, temperature rises, and increased sea levels - as well as monitoring the effect of protected areas (such as national parks, nature reserves, and wildlife sanctuaries) versus non-protected areas (wilderness), the real time systems would allow conservationists to immediately react to environmental changes.

As shown by the data, it is possible to create a general model that can predict the lifespans, population, and extinction risk of species. It was possible to decide which model was the most effective by running practice simulations and testing. By training the model with data that was preprocessed, predictions were made on the lifespan, population numbers, and extinction risk of declining species. Finally, it was possible to cross-check the work by running an additional series of tests to validate the findings.

This research may prove beneficial to conservationists, as

they can better allocate their conservation efforts to species that have a greater risk of extinction soon, allowing a more effective conservation effort regarding endangered species. As most modern prediction models are hyper-focused on a specific species in a small sub-region on the globe, this model strives to give an accurate, and more general, prediction for any species. Due to the current dataset having such limited information, the predictions for lifespans may not always be accurate, especially when species lifespans are too long or too short, as the current dataset does not have enough species with lifespans in those ranges to use as training data. As a result, when the model tries to run a simulation, the outcomes may be different than expected which could lead to inconsistencies while predicting extinction risk.

References

- 1 Kaggle, *Animal Information Dataset*, 2023, <https://www.kaggle.com/datasets/iamsouravbanerjee/animal-information-dataset>, Dataset.
- 2 BirdLife International, *IUCN Red List of Threatened Species: Tachybaptus rufolavatus*, 2022, <https://www.iucnredlist.org/species/22696558/208158571>.
- 3 BirdLife International, *IUCN Red List of Threatened Species: Loxops ochraceus*, 2024, <https://www.iucnredlist.org/species/103824084/250520617>.
- 4 O. R. Wearn *et al.*, *Science*, 2012, **337**, 228–232.
- 5 A. D. Davidson *et al.*, *Proceedings of the National Academy of Sciences*, 2012, **109**, 3395–3400.
- 6 S. D. Howard and D. P. Bickford, *Diversity and Distributions*, 2014, **20**, 837–846.
- 7 M. A. Zettlemoyer *et al.*, *American Journal of Botany*, 2019, **106**, 547–559.
- 8 G. H. De Oliveira Caetano *et al.*, *PLoS Biology*, 2022, **20**, e3001544.
- 9 A. Zizka *et al.*, *Diversity and Distributions*, 2021, **28**, 227–241.
- 10 S. E. Darrah *et al.*, *Diversity and Distributions*, 2017, **23**, 435–447.
- 11 L. Breiman, *Machine Learning*, 2001, **45**, 5–32.
- 12 D. R. Cutler *et al.*, *Ecology*, 2007, **88**, 2783–2792.
- 13 T. Chen and C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
- 14 T. Cover and P. Hart, *IEEE Transactions on Information Theory*, 1967, **13**, 21–27.
- 15 F. Pedregosa *et al.*, *Scikit-learn: Machine Learning in Python*, arXiv preprint, 2012, <https://arxiv.org/abs/1201.0490>.
- 16 W. McKinney, Proceedings of the Python in Science Conference, 2010, pp. 56–61.
- 17 R. Kohavi, Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 1995, pp. 1137–1143.

-
- 18 M. Stone, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 1974, **36**, 111–133.
 - 19 R. Anderson-Sprecher, *The American Statistician*, 1994, **48**, 113–117.
 - 20 R. Anderson-Sprecher, *The American Statistician*, 1994, **48**, 113–117.
 - 21 M. Nowak *et al.*, *IUCN Red List of Threatened Species: Pongo abelii*, 2017, <https://www.iucnredlist.org/species/121097935/259045437>.
 - 22 M. Nowak *et al.*, *IUCN Red List of Threatened Species: Pithecopaga jefferyi*, 2016, <https://www.iucnredlist.org/species/22696012/129595746>.
 - 23 L. Rojas-Bracho *et al.*, *IUCN Red List of Threatened Species: Phocoena sinus*, 2022, <https://www.iucnredlist.org/species/17028/214541137>.
 - 24 L. Rojas-Bracho *et al.*, *IUCN Red List of Threatened Species: Strigops habroptilus*, 2018, <https://www.iucnredlist.org/species/22685245/129751169>.
 - 25 M. Cardillo *et al.*, *Science*, 2005, **309**, 1239–1241.
 - 26 C. Strobl *et al.*, *BMC Bioinformatics*, 2007, **8**, 25.