UNIVERSITY of HOUSTON
COLLEGE of NATURAL SCIENCES & MATHEMATICS

Multi-Class Prediction of Species' Extinction Risk Using Deep Learning, Support Vector Machine & Neural Networks

Dylan Berens, Kathiana Rodriguez, Shruti Yenamagandla, Carl Aguinaldo, Dominic McDonald
University of Houston College of Natural Sciences & Mathematics - Houston, TX

Team Data Acids: #0076

Contact Information:
Name: Dylan Berens
University of Houston
Phone: (713) 826-8885
Email:
dylan.berens700@gmail.com

ABSTRACT

Background: Protecting biodiversity is important for maintaining healthy ecosystems. Traditional methods of determining species' risk of endangerment requires a lot of time and can be inconsistent when measuring thousands of species because scientists have to track population trends, habitats, and biological characteristics. Our goal through this research was to develop an automated, efficient, and robust way of identifying species that may be at risk.

Methods: In our research, we tested Support Vector Machines (SVM) and Neural Networks (scikit-learn & TensorFlow) to predict species' risk of endangerment, preprocessing the dataset by dropping, imputation, mapping and scaling.

Results: We tested 3 different models: scikit-learn's SVM, scikit-learn's Neural Network using MLPClassifier, and a deep learning model using TensorFlow's Keras API. The highest accuracy we achieved was with the SVM model at 0.82, then the deep learning model at 0.79, then MLPClassifier at 0.77.

Conclusions: In conclusion, we were able to determine that SVM yielded the bet results with deep learning not too far behind. This means that both traditional machine learning and deep learning methods can be effective in flagging and automating species' risk.

BACKGROUND

Protecting species diversity is essential in ensuring healthy ecosystems. In order to determine species' risk, biologists traditionally need to track factors like population trends, habitats, and biological characteristics manually. This way of measuring risk is slow and may be inconsistent for large groups of species. To challenge this problem, we applied machine learning and deep learning models to expedite and automate the species' risk flagging process. Capacity to predict animal species' threatened status can be used to identify actions minimize the conditions leading to animal species' extinction and thereby preserve biodiversity.

OBJECTIVE

Our objective was to assess machine learning and deep learning models to achieve predictive modeling of whether a species is at risk of extinction based on taxonomic rank, population data and ecological information, optimizing model performance through hyperparameter tuning.

METHODS

Dataset

- We chose our dataset because it included many features (including biological classification, population information, and habitat data) that we sought to assess when applying our models. While our dataset was limited to species in New Zealand, it included over 8000 entries, providing enough data to train and test our models effectively.

Data Preprocessing

- Our data preprocessing involved dropping, imputation, mapping and scaling. We employed label encoding and one hot encoding to convert our categorical data into numerical data to ensure functionality in our models. After the train_test_split we applied MinMaxScaling.

Support Vector Machine (SVM)

- We created a model using sklearn's SVM. We split the training and testing set and scaled the features. We then tuned the parameters for SVM using GridSearchCV to improve the accuracy of our model. Using the best parameters, we were able to get an accuracy of 0.82 with SVM.

Deep Learning (TensorFlow Neural Network)

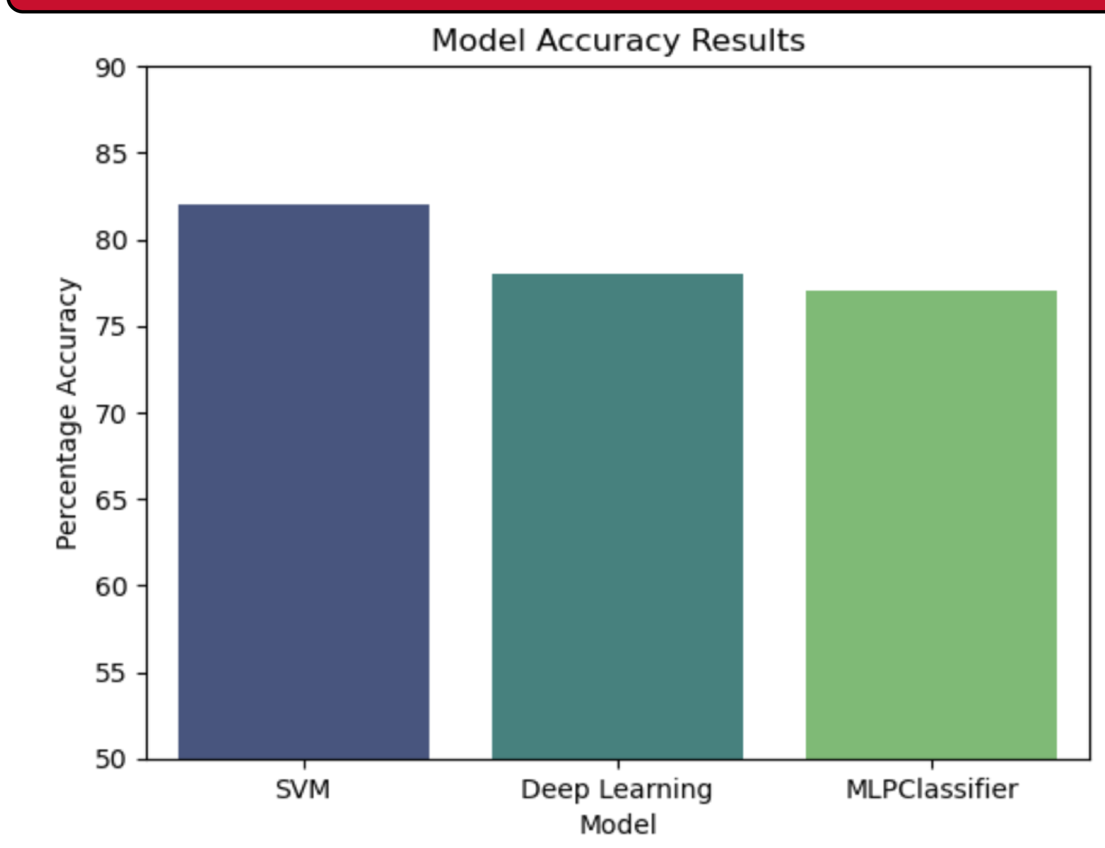
- Using TensorFlow's Keras API, we created a deep learning model. By leveraging a parameter grid to accomplish hyperparameter tuning, we arrived at the model configuration with an accuracy of 0.79 (4 layers, applying 'relu' and 'softmax' activations).

Neural Network

- We implemented a Neural Network model using sklearn's MLPClassifier. By applying GridSearch for hyperparameter tuning, we achieved an accuracy of 0.77.

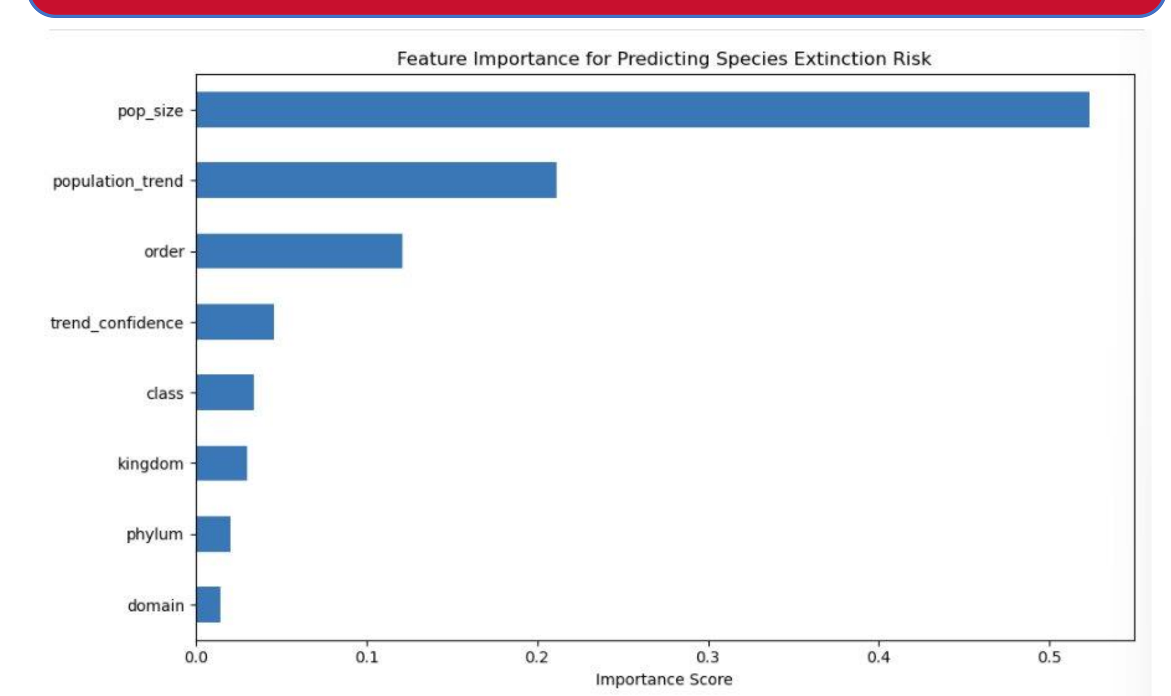
RESULTS

Figure 1. Model Accuracy Results



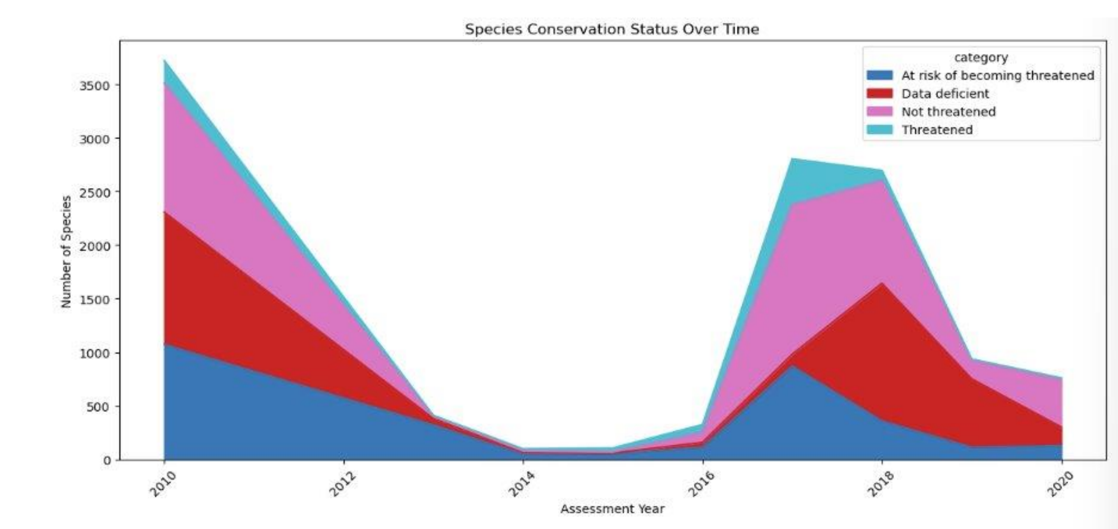
- SVM was our most accurate model with a score of 0.82 accuracy.
- Deep Learning and Neural Network-MLPClassifier trailed slightly at 0.79 and 0.77, highlighting that both machine learning and deep learning can be effective in predictive modeling for species' extinction risk.

Figure 2. Feature Importances for Predicting Species' Extinction Risk



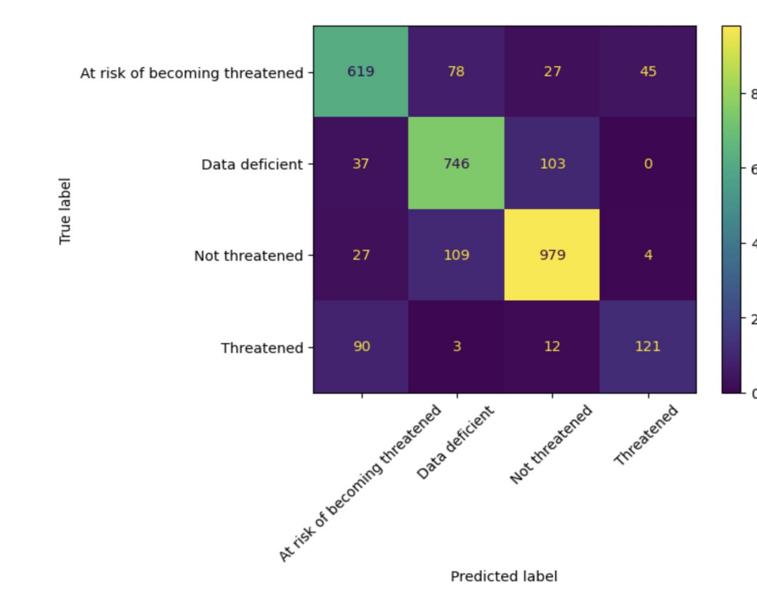
- Population size is the strongest predictor – species with small or declining populations are at highest risk.
- Population trend (increase/decrease) also plays a large role in determining target.
- Order is the strongest predictor amongst the higher Taxonomic Ranks

Figure 3. Species' Conservation Status Over Time



- Looking at this graph, ignoring the species that are data deficient, we can see that a large portion of the species are at risk of becoming threatened.
- There was a large dip in number of species being assessed from 2014 to 2016, however, all categories spiked again starting from 2016.

Figure 4. SVM Confusion Matrix



CONCLUSIONS

- In conclusion, our SVM model generated the best results, being able to correctly predict a species' risk 82% of the time. Our Deep Learning and MLPClassifier models had creditable accuracies of 79% and 77% respectively.
- Being able to predict species' risk of extinction is essential in assessing ecosystem health. Through our research, we demonstrated that it is possible to perform predictive modeling for species' extinction risk level by leveraging machine learning.