International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Enhancing Logistic Regression for Predicting Habitat Suitability of Scottish Crossbills (Loxia Scotica)

Mike Jayson F. Llovit[1], Mark Joshua L. Morales[2], Jonathan C. Morano[3], Khatalyn E. Mata[4]

[1,2]Student, College of Information Systems Technology Management Department, Pamantasan ng Lungsod ng Maynila
[3,4]Faculty, College of Information Systems Technology Management Department, Pamantasan ng Lungsod ng Maynila

Abstract
The prediction of habitat suitability for migratory avian species in response to climate change is a critical challenge in ecological conservation. This study focuses on the Scottish Crossbill (Loxia scotica), employing an enhanced Logistic Regression algorithm to address the limitations of traditional approaches. A significant issue with the existing algorithm was the use of default hyperparameters, including L2 regularization, a Limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) solver, and limited iterations. These settings constrained the algorithm's generalizability and ability to adapt to the complexity of habitat prediction data. To overcome these challenges, the researchers utilized the PyCaret package, which facilitates comprehensive hyperparameter tuning by systematically exploring combinations of key parameters such as regularization strength and solver types, alongside cross-validation for robust performance evaluation. The integration of PyCaret significantly improved the algorithm's performance. Compared to the existing Logistic Regression algorithm, the enhanced algorithm exhibited an 11.4% increase in accuracy, a 4.74% rise in the Area Under the Curve (AUC), and a 9% improvement in the F1-score during Test Set Evaluation. Specifically, the enhanced algorithm achieved an accuracy of 88%, an AUC of 88.99%, and an F1-score of 88%. These results highlighted the enhanced algorithm's predictive capabilities and its robustness in identifying suitable habitats. The enhanced algorithm's ability to predict habitat suitability more effectively underpins its potential for aiding conservation planning. By implementing systematic hyperparameter tuning, the enhanced algorithm not only achieves higher prediction accuracy but also minimizes bias and variance, paving the way for more reliable predictions of habitat suitability under changing climate conditions.

Keywords: Logistic Regression, hyperparameters, PyCaret, habitat suitability, Scottish Crossbill.

Chapter One
INTRODUCTION
1.1 Background of the Study
Birds are integral components of ecosystems, playing crucial roles in pollination, seed dispersal, and main

taining ecological balance. However, their survival heavily relies on the availability of suitable habitats that provide essential resources for foraging, nesting, and breeding. According to Plumer (2014), animals, including bird species, face an increased risk of extinction as their geographic range contracts due to the loss or degradation of suitable habitats. This phenomenon has been consistently observed throughout Earth's history, where species that thrive in specific habitats, such as grasslands or forests, experience decreased chances for survival when forced to relocate to areas lacking those vital resources.

The Scottish Crossbill (*Loxia scotica*) was selected as one of the two species for this study due to its unique ecological significance and current conservation challenges. These birds are endemic to the Caledonian Forests of Scotland, making them the only terrestrial vertebrate species endemic to the UK. Remarkably, they possess distinctive crossed mandibles that are specially adapted for prying open cones and extracting seeds, with strong bill muscles capable of cracking even the toughest cones (Forestry.com, n.d.). However, in an article by the BBC, the Scottish Crossbill is at risk of extinction because the climate is becoming increasingly unsuitable for its survival. The State of the UK's Birds in 2017 highlights that the average summer temperature is nearly 1°C higher than in the 1980s, indicating the potential impact of global warming on this species' habitat (BBC, 2017).

Habitat suitability is a critical aspect that determines the successful establishment and persistence of a species within a given environment. As mentioned by Baling et al. (2016), multiple factors contribute to habitat suitability, including resource availability, presence of invasive species, landscape connectivity, and climate. These factors collectively shape the quality and carrying capacity of a habitat, influencing the ability of a species to thrive and reproduce.

A study by Townsend and Aldstadt (2023), titled “Habitat suitability mapping using logistic regression analysis of long-term bioacoustic bat survey dataset in the Cassadaga Creek watershed (USA)”. Utilized both presence and pseudo-absence data points, allowing for a direct comparison of bat activity with given environmental variables using logistic regression analysis on a species-specific basis. Through this methodology, the researchers analyzed and spatially mapped the probability of bat presence based on specific environmental variables. The logistic regression model proved to be an effective machine learning tool for assessing habitat suitability, as it enabled the quantification of the likelihood of a species occurring in a particular area based on the environmental conditions present. This study demonstrated the potential of logistic regression in habitat suitability mapping, providing a framework for similar applications in other species and ecosystems. Given the success of this study in utilizing logistic regression to analyze the probability of bat presence based on environmental variables and map habitat suitability, this algorithm has been chosen for this study.

The researchers aims to solve one of the problems of the Logistic Regression and enhance the Logistic Regression which according to Thanda (2023), is the statistical technique used to predict the relationship between the dependent variable (Y) and the independent variable (X), where the dependent variable is binary in nature. Moreover, it is a statistical tool that can be used to analyze habitat selection and distribution of species, and to develop habitat suitability models.

In Logistic Regression, it is crucial to explore the various hyperparameters that influence the performance of Logistic Regression models and develop a systematic approach to tuning these parameters for enhanced accuracy and reliability (GeeksforGeeks, 2024). Thus, the researchers aim to find the solution throughout to be able to obtain the accurate results while utilizing the algorithm for the target of the study.

1.2 Statement of the Problem

1.2.1 General Statement of the Problem

Logistic Regression is a type of classification algorithm that is used to find the probability of success and failure event. Among its many advantages is that, it is easier to implement, interpret, and train. However, Logistic Regression performs best with the right set of hyperparameters which is difficult to identify. Thus, the researchers intend to solve the following problem:

1.2.2. Specific Statement of the Problem

In training the algorithm, the Logistic Regression requires careful tuning of hyperparameters. It is difficult to analyze the right hyperparameters which fit a specific algorithm. A good set of hyperparameters are necessary in order to achieve the good performance of the algorithm.

According to Melanee Group (2023), hyperparameter tuning is a crucial part of the machine learning process as it can significantly affect the performance of the model. By selecting the right hyperparameters such as learning rate and regularization, among others, a model can achieve optimal performance. However, the process of hyperparameter tuning can be challenging and time-consuming, as there are many hyperparameters to consider, and the search space can be vast. Therefore, it is essential to strike a balance between optimizing the model's performance and avoiding an endless cycle of trying to optimize. (Melanee Group, 2023).

1.3 Objective of the Study

1.3.1 General Objective

To implement a more effective comprehensive hyperparameter tuning for predicting the species' habitat suitability with the use of enhanced algorithms that are proposed by the researchers, and ultimately achieve greater precision and interpretations in predictions. Specifically, the researchers seeks:

1.3.2 Specific Objectives

To utilize a package that is used for fitting logistic regression algorithm and define a tuning grid with sets of hyperparameters to evaluate and figure out a good set of hyperparameters from the different combinations of hyperparameters for building a prediction algorithm.

1.4 Significance of the Study

The study helped enhance the accuracy and precision of Logistic Regression and provided crucial information and knowledge about the techniques that were used, and was able to create a working habitat suitability prediction algorithm for Scottish Crossbills (Loxia scotica) and Turtle Doves (Streptopelia Turtur). The study had benefited the following:

Environmentalists and Conservationists

The study provided environmentalists and conservationists with a powerful tool to predict the habitat suitability of birds more accurately. This capability can inform targeted conservation strategies, enabling more effective protection of birds and their habitats.

Policymakers

The study's findings will inform policymakers in developing regulations and guidelines that promote sustainable practices and protect bird species by understanding the potential impacts of human activities on habitat suitability and sustainability.

General Public

By presenting complex habitat suitability data in an easily understandable and engaging manner, the study

can foster greater public awareness and appreciation for the challenges faced by birds and potentially inspiring more individuals to support conservation efforts.

Future Researchers

The findings of this study can serve as a foundation for further advancements and enabling more sophisticated models and techniques to be developed, ultimately leading to a deeper understanding of habitat suitability and its implications.

1.5 Scope and Limitations

This study aims to enhance the capability of logistic regression in predicting habitat suitability for the Scottish Crossbill (Loxia scotica). The algorithm will leverage data obtained from two primary sources: the UK Met Office and the Global Biodiversity Information Facility (GBIF). The GBIF, an international network and research infrastructure, provides open access to data about life on Earth, making it a valuable resource for this project. The findings of this research will contribute valuable insights to conservation organizations working to preserve the Scottish Crossbill and mitigate its risk of extinction due to habitat loss and climate change.

It is important to acknowledge that the data obtained from the UK Met Office and GBIF may have inherent limitations. The availability and quality of data can vary across regions, potentially affecting the algorithm's performance in certain geographical areas. Additionally, the study's focus on the Scottish Crossbill may limit the generalizability of the findings to other bird species or ecological systems. Furthermore, the complexity of habitat requirements and the dynamic nature of environmental factors can introduce challenges in accurately predicting habitat suitability. Nevertheless, the proposed approach of enhancing logistic regression holds promise in improving our understanding of habitat suitability modeling and informing conservation efforts for this unique species.

1.6 Definition of Terms

This section of the study specifically defines the key terms that are used within the study. The following terms are:

Algorithm - a set of defined steps designed to perform a specific objective

Geospatial Analysis - is an analysis using geospatial data, which is data with some sort of geographical component. For example: an address, longitude and latitude coordinates, or a state/country name. Geospatial analysis can include a wide variety of methods, including mapping or calculations based upon area and distance.

Hyperparameters - are parameters whose values control the learning process and determine the values of algorithm parameters that a learning algorithm ends up learning.

Logistic Regression - Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The algorithm delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Migration - the process of animals traveling to a different place, usually when the season changes.

Prediction - to say that an event or action will happen in the future, especially as a result of knowledge or experience.

Chapter Two
REVIEW OF RELATED LITERATURE

This chapter presents the review of related literature and studies after the thorough research done by the researchers. This chapter will help in familiarizing information that are relevant and similar to the present study.

2.1 Habitat Suitability and Species Conservation

According to Crawford et al. - in their journal article titled "Expert-Informed Habitat Suitability Analysis for At-Risk Species Assessment and Conservation Planning", habitat suitability models (HSMs), also known as species distribution models, are now commonly used tools for estimating the distribution of species, their habitats, and potential threats. HSMs utilize measures of environmental and landscape attributes (e.g., soil characteristics, canopy cover, fragmentation, rainfall) in areas where a species was known to occur over a specific time scale to project where similar conditions exist throughout the species' range. Known species locations (presence data) can be collected from records in natural history collections, systematic surveys, or opportunistic observations. The projected habitat distributions can then be used to understand species-habitat relationships, predict where potentially suitable habitat is likely to occur, and prioritize areas for surveys, habitat management, parcel acquisition or designation, species translocation, and other applications related to conservation planning. These results have direct applications to management and conservation planning: partners can tailor site-level management based on attributes associated with high habitat suitability for species of concern; allocate survey effort in areas with suitable habitat but no known species records; and identify priority areas for management, land acquisitions, or other strategies based on the distribution of species records, suitable habitat, and land protection status.

The HSMs provide spatially-explicit information about habitat suitability that can directly inform and improve conservation planning and decision-making for these at-risk species.

HSMs can help with conservation planning in the following ways:

1. Inform site-level management: The model results identify habitat features associated with high suitability, allowing managers to tailor management practices accordingly.
2. Guide survey efforts: The models can identify areas with suitable habitat but no known species records, helping to prioritize survey locations.
3. Prioritize conservation actions: The models map the distribution of suitable habitat, species records, and land protection status, allowing partners to identify priority areas for management, acquisitions, or other conservation strategies.

According to Mancino et al. - in their article titled "Increase of nesting habitat suitability for green turtles in a warming Mediterranean Sea", understanding how habitat suitability for species may change in the future due to climate change can help conservation efforts in several ways, by identifying areas that may become more suitable, conservation efforts can focus on monitoring and protecting those areas before species begin using them. This allows conservation to be proactive rather than reactive. The article also finds that some areas in the western Mediterranean may become more suitable for green sea turtle nesting in the future and focusing conservation in those areas now can help the species if their range expands westward.

By identifying areas that may become suitable habitat in the future, conservation efforts can:

1. Monitor potential new habitat areas to document if/when species begin using them due to climate shifts.

2. Detect and address new threats in areas expected to become suitable habitats before they impact species.
3. Manage coastal areas proactively by focusing conservation actions on high-priority future habitats to protect it before it is needed.

2.2 Climate Change and Avian Habitats

According to Mallon and Wormworth - in their study of "Bird Species and Climate Change", this study was able to discuss the current scientific understanding of anthropogenic climate change impacts on global bird species now, and projected future effects. Climate change will have serious negative consequences for many bird populations and has already been linked to population declines and major reproductive declines. Looking to the future, the most serious of possible impacts - extinctions of entire bird species - are predicted.

Climate change puts many bird species at risk of extinction, even those currently considered safe and the stronger the climate change the stronger the risk. With a global mean surface temperature increase of 1-2°C above pre-industrial levels, many unique and threatened ecological systems will be at risk and numerous species will face extinction. Risk is dependent on the species. The golden bowerbird, like many other bird species in the Wet Tropics of Australia's northeast, is particularly vulnerable. Its suitable habitat would decrease 63 percent with less than 1°C of future warming, illustrating why this zone's climate scenario has been called "an impending environmental catastrophe".

Among particularly vulnerable groups -- migratory, Arctic, Antarctic, island, wetland, mountain and seabirds -- heightened impacts are expected. The threat of climate change to migratory birds is equal to the sum of all other human-caused threats combined with 84 percent of migratory bird species facing some type of climate change threat. For example, the Arctic-breeding red-breasted goose, already globally vulnerable, is expected to lose 99 percent of its tundra breeding habitat due to climate change (Zöckler and Lysenko, 2000). Birds that are habitat specialists are at higher risk than generalists. Birds breeding in arid environments and those with low population numbers, poor dispersal ability, already poor conservation status, and restricted or patchy habitats or limited climatic ranges are also at elevated risk from climate change.

The overall extinction risk of climate change to birds is still being quantified. However, first-cut estimates present the possibility of the extinction of more than a third of European bird species under a maximum (>2.0°C) climate change scenario, if birds cannot shift to new climatically suitable ranges. Indeed their capacity to shift is subject to considerable uncertainty given Europe's heavily modified landscape. One candidate for extinction is the **Scottish crossbill**, expected to lose 100 percent of its current habitat. In the Australian Wet Tropics bioregion, mid-range climate change is predicted to threaten almost three-quarters of rainforest birds there with extinction in the next 100 years. (Mallon & Wormworth, n.d.)

2.3 Logistic Regression in Habitat Suitability Modeling

According to Fieberg et al. - in their study of "A 'How to' guide for interpreting parameters in habitat-selection analyses", this study was able to discuss how Logistic Regression can be used for habitat suitability. Much of the confusion surrounding the interpretation of parameters in habitat-selection functions can be attributed to the use of logistic regression to model *use-availability* data (Keating & Cherry, 2004). Logistic regression is most easily understood as a model for binary random variables that can take on one of two values (0 or 1) with probability that depends on one or more explanatory variables

(Hosmer, Lemeshow, & Sturdivant, 2013).

Consider, for example, a study designed to infer how various environmental characteristics influence whether a habitat patch (e.g. a contiguous area of forest) will be used by one or more animals. In this case, we may randomly select n habitat patches and monitor them to determine if they are used ($y_i = 1$) or not ($y_i = 0$) for $i = 1, 2, \ldots, n$. Logistic regression allows us to model the probability that each patch will be used, $P(y_i = 1) = p_i$, as a logit-linear function of k patch-level predictors ($X_{i1}, \ldots, X_{ik}$) and regression parameters ($\beta_0, \beta_1, \ldots, \beta_k$):

$$\text{logit}(p_i) = \log\left[\frac{p_i}{(1-p_i)}\right] = \beta_0 + X_{i1}\beta_1 + \ldots X_{ik}\beta_k, \quad y_i \sim \text{Bernoulli}(p_i),$$

After having fit a model, we can exponentiate the regression coefficients, $\exp(\beta_j)$ for $(j = 1, \ldots, k)$, to quantify how the odds of patch i being used, $p_i/(1-p_i)$, change as we increase the jth predictor by 1 unit while holding all other predictors constant. We can also use the inverse-logit transformation (Equation 1) to estimate the probability that patch i will be used, given its set of spatial predictors:

$$p_i = \frac{\exp\left(\beta_0 + X_{i1}\beta_1 + \ldots X_{ik}\beta_k\right)}{1 + \exp\left(\beta_0 + X_{i1}\beta_1 + \ldots X_{ik}\beta_k\right)}.$$

The logit transformation ensures that p_i will be constrained between 0 and 1 for all values of the predictor variables.

Contrast this approach with how logistic regression is used to study habitat selection. In a typical habitat-selection study, logistic regression models are fit to separate samples of used and available sample units, usually points; these groups are not mutually exclusive (i.e. available habitat may also be used). In this case, y_i is no longer a Bernoulli random variable since p_i depends on the ratio of used to available points (which is under control of the analyst). That is, the probability that a location will be a 'used point' decreases with the number of user-generated 'available' locations. (Fienberg et al., 2021)

2.4 Limitations and Challenges of Logistic Regression

According to W.D. - in their article of "Scikit-learn's Defaults are Wrong", By default, logistic regression in scikit-learn runs w L2 regularization, I don't know if it's true that a plurality of people doing logistic regressions are using L2 regularization and lambda = 1, but the point is that it doesn't matter. Unregularized logistic regression is the most obvious interpretation of a bare bones logistic regression, so it should be the default, and RegularizedLogisticRegression could have its own class:

```
class RegularizedLogisticRegression(LogisticRegression):
    def __new__(cls, *args, **kwargs):
        if 'penalty' not in kwargs:
            kwargs['penalty'] = 'l2'
        return super().__new__(cls, *args, **kwargs)
```

If you're not normalizing your data, then you really can't penalize the parameters in a sensible way. This follows very straightforwardly from the math of regularization, explained in Appendix A of this post. Why is this a problem? Because one might expect that the most basic version of a function should broadly work for most cases. Except that's not actually what happens for LogisticRegression. Scikit-learn requires you to either preprocess your data or specify options that let you work with data that has not been preprocessed

in this specific way. You cannot simply put your data into sklearn's logistic regression for exploratory purposes and get sensible results. In other words, the ostensible simplicity and lack of fuss of these default parameters for machine learning creates an odd road bump in the case where you really want simplicity, i.e. exploratory analysis. If you type "logistic regression sklearn example" into Google, the first result does not mention that this preprocessing is necessary and does not mention that what is happening is not logistic regression but specifically penalized logistic regression. Furthermore, the lambda is never selected using a grid search. Someone learning from this tutorial who also learned about logistic regression in a stats or intro ML class would have no idea that the default options for sklearn's LogisticRegression class are wonky, not scale invariant, and utilizing untuned hyperparameters.

As previously explained, LogisticRegression's default options don't "work" with typical, unnormalized data. What's even crazier is that LogisticRegression's default options don't work on most data, even when normalized, unless lambda = 1 maximizes whatever score you're evaluating your model on. Even if it makes sense for all logistic regressions to be penalized and have lambda > 0, it does not follow that lambda = 1 is a good default. To be clear, this is totally arbitrary, and to get the lambda you want, you need to use something like grid search to tune your model to the lambda that maximizes whatever score you're using to evaluate it. Of course, you don't run into this issue if you just represent LogisticRegression as an unpenalized model. You run into the "issue" that your model is no longer penalized, but you know exactly what you're getting and it's totally intuitive. Yes, lambda = 0 is "wrong" if all models should be penalized, but lambda = 1 is also wrong for most models.

According to Scott Dallman - in their study of "Cheat ML Model Creation with PyCaret", in machine learning, optimizing the hyperparameters of a model is crucial for achieving the best performance. Logistic regression, a popular classification algorithm, has several hyperparameters like regularization strength and penalty type that can be tuned for better results. After preparing the data, let's use pycarets compare_models() function. It's used for model selection and performance evaluation. When we apply compare_models() to a dataset, PyCaret trains and evaluates the performance of several machine learning models on that dataset. The function automatically applies several data preprocessing techniques and hyperparameter tuning methods to each model in order to find the best-performing model for that particular dataset. The output of compare_models() is a table that shows the performance metrics of all the evaluated models, ranked in order of their performance. This function helps us to quickly compare and select the best model for our dataset, which can save us time and effort in the machine learning workflow.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
lr	Logistic Regression	0.9718	0.9971	0.9718	0.9780	0.9712	0.9573	0.9609
knn	K Neighbors Classifier	0.9718	0.9830	0.9718	0.9780	0.9712	0.9573	0.9609
qda	Quadratic Discriminant Analysis	0.9718	0.9974	0.9718	0.9780	0.9712	0.9573	0.9609
lda	Linear Discriminant Analysis	0.9718	1.0000	0.9718	0.9780	0.9712	0.9573	0.9609
lightgbm	Light Gradient Boosting Machine	0.9536	0.9935	0.9536	0.9634	0.9528	0.9298	0.9356
nb	Naive Bayes	0.9445	0.9868	0.9445	0.9525	0.9438	0.9161	0.9207
et	Extra Trees Classifier	0.9445	0.9935	0.9445	0.9586	0.9426	0.9161	0.9246
gbc	Gradient Boosting Classifier	0.9355	0.9792	0.9355	0.9416	0.9325	0.9023	0.9083
xgboost	Extreme Gradient Boosting	0.9355	0.9868	0.9355	0.9440	0.9343	0.9023	0.9077
dt	Decision Tree Classifier	0.9264	0.9429	0.9264	0.9502	0.9201	0.8886	0.9040
rf	Random Forest Classifier	0.9264	0.9909	0.9264	0.9343	0.9232	0.8886	0.8956
ada	Ada Boost Classifier	0.9155	0.9947	0.9155	0.9401	0.9097	0.8720	0.8873
ridge	Ridge Classifier	0.8227	0.0000	0.8227	0.8437	0.8186	0.7320	0.7454
svm	SVM - Linear Kernel	0.7618	0.0000	0.7618	0.6655	0.6888	0.6333	0.7048
dummy	Dummy Classifier	0.2864	0.5000	0.2864	0.0822	0.1277	0.0000	0.0000

Figure 2.1 Comparison of models utilizing PyCaret for a dataset

From the output above you can see that Logistic Regression seems to be the best model to use so we will focus on that to create a full model.

Modeling:

We can train machine learning models using PyCaret. PyCaret provides a wide range of classification algorithms such as logistic regression, decision tree, random forest, and gradient boosting. To train a model using PyCaret, we can use the create_model function. For example:

```
lr = create_model('lr')
```

We have now trained a logistic regression model using the create_model function. PyCaret automatically performs hyperparameter tuning using cross-validation to find the best hyperparameters for the model. We can also specify the hyperparameters manually using the tuned_parameters parameter.

The library also provides a wide range of ensemble models such as stacking, blending, and bagging. To train an ensemble model using PyCaret, we can use the ensemble_model function. For example:

```
stacker = ensemble_model(lr)
```

In the above example, we have trained a stacking ensemble model using the logistic regression model as the base estimator.

According to Jiang & Xu - in their study of "Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastis Using Clinical Data", in 2020, female breast cancer surpassed lung cancer as the most commonly diagnosed cancer worldwide, with an estimated 2.3 million new cases in 2020. Breast cancer remains one of main cancer-related causes of death in women globally and was responsible for 685,000 deaths worldwide in 2020. Breast cancer is the second leading cause of cancer death among US women after lung cancer, estimated to account for 43,600 deaths in 2021. It is the number one cause of cancer-related deaths for US women aged 20 to 59.

Women rarely die of breast cancer confined to the breast or draining lymph nodes; rather, they die mainly due to metastasis, a condition in which cancer spreads to other vital organs, such as the lung and brain. Metastatic breast cancer (MBC) is the cause of over 90% of breast cancer related deaths and remains a largely incurable disease.

Although most newly diagnosed breast cancer cases are not metastatic, all patients are at risk of developing metastatic cancer in the future, even if they are free of cancer for years after the initial treatment. The ability to effectively predict, for each individual patient, the likelihood of later metastatic occurrence is important, because the prediction can guide treatment plans tailored to a specific patient to prevent metastasis and to help avoid under- or over-treatment.

Various learning methods have been developed and applied in biomedical prediction. For instance, machine learning and language processing have been used to identify breast cancer local recurrence. A **logistic regression model** was developed for cancer classification and prediction.

In this research, we took the empirical approach to study the performance of DFNN models when predicting BCM using clinical data. We refer to these models as the

DFNN (deep feedforward neural network) models throughout the text. We applied the DFNN method to learn prediction models from LSM datasets. These models can be used to predict 5-, 10-, and 15-year BCM. The performance of a DFNN model is affected by the number of hidden layers and number of nodes per hidden layer, which are called hyperparameters. In addition, there are other hyperparameters that can be used to adjust the prediction performance of deep learning. For example, "epochs" is a hyperparameter we consider. One epoch means a deep learning model is trained by each of the training set samples exactly once. The learning might not converge when epochs is too low, and model overfitting tends to get severe

when it is too high. Tuning hyperparameters is the process of identifying the set of hyperparameter values that are expected to produce the best prediction model out of all sets of hyperparameter values examined. Grid search is designed to conduct hyperparameter tuning in a systematic way by going through a possible set of hyperparameter values automatically during learning. In this study, we optimized DFNN model performance by conducting hyperparameter tuning via grid search.

To evaluate the performance of the DFNN, we compared our DFNN models with the ones that we trained using nine other well-known machine-learning methods. We applied hyperparameter tuning and grid search to optimize model performance for each of the nine comparison methods. We conjectured that the performance of our DFNN models with grid search would be comparable to that of other machine learning methods when predicting the binary status of BCM. We posit this conjecture, because deep learning is a very powerful tool for prediction and has been successful in other applications, such as image recognition. In this study, we use the DFNN models to predict 5-, 10-, and 15-year BCM by learning from non-image clinical EHR data. Through literature searching, we found some deep learning-related studies that use image data to predict BCM, but we have not found a study that resembles ours. (Jang & Xu, 2022)

2.5 Conservation Efforts for the Scottish Crossbills

According to Forestry.com - in their study about "Scottish Crossbill", this study discusses the overall information about Scottish Crossbills and the conservation efforts for them. The Scottish Crossbill's *(Loxia scotica*) reflects both its genus (*Loxia*) and its specific epithet (*scotica*), indicating its association with Scotland. The genus name "*Loxia*" is derived from the Greek word "*loxos*," meaning "*crosswise*" or "*oblique*," which aptly describes the crossed tips of the bird's distinctive bill. The species name "*scotica*" highlights its connection to Scotland, emphasizing its status as an endemic bird found primarily in the Caledonian pine forests of the Scottish Highlands.

The long-term survival of the Scottish Crossbill depends on continued conservation efforts. By addressing the threats they face and protecting their habitat, we can ensure that these fascinating birds continue to thrive in the Caledonian Forest for generations to come.

Threats and Concerns:

- Habitat loss and fragmentation: Forestry practices, such as clear-cutting and the planting of non-native conifers, can reduce the availability of suitable mature Scots pine trees.
- Climate change: Changes in temperature and precipitation patterns can impact cone production and affect the long-term viability of their habitat.
- Predation: Predators like squirrels, crows, and raptors can threaten nests and chicks.
- Human activities: Disturbance from recreational activities and tourism can disrupt breeding and foraging behavior.

Conservation efforts:

- Habitat protection: Organizations like the RSPB and NatureScot are working to protect and restore Caledonian Forest habitat.
- Sustainable forestry practices: Promoting forestry practices that maintain a diverse range of age classes and species of trees is crucial for the long-term survival of the crossbill.
- Monitoring and research: Ongoing monitoring of the crossbill population and research into their habitat needs are essential for effective conservation efforts.
- Public awareness: Raising awareness about the Scottish Crossbill and their importance to the Caledonian Forest can encourage support for conservation efforts. (Forestry, 2023)

2.6 Conservation Efforts for the Turtle Doves

According to RSPB.org - in their article titled "Lifeline thrown to the UK's Turtle Doves as another year of no hunting along their migration route is declared", one of the UK's fastest-declining wild bird species is the Turtle Dove. This globally threatened migratory bird has suffered steep declines in the UK, and in neighboring countries like Belgium and the Netherlands, since the 1970s, primarily due to changes to farming practices but with the situation made worse by unsustainable hunting in south–west Europe.

All UK-breeding Turtle Doves spend the winter in West Africa, migrating via south-west Europe in both autumn and spring. Here in the UK, Turtle Doves breed in key areas of southern and eastern England, with the first few returning birds spotted in the UK last week. Hunting of the birds has taken place for many years in France, Spain and Portugal, and prior to 2018, around one million Turtle Doves were being hunted each autumn across these three countries alone. Meanwhile, agricultural changes here at home have caused a loss of suitable habitat for the birds that make it to the UK to raise the next generation, leaving just 2,100 breeding territories remaining in the UK according to a 2021 study.

The RSPB and partners are working with growing numbers of fantastic farmers and landowners here in the UK as part of Operation Turtle Dove to reverse the fortunes of this beloved summer visitor. Driving forward the restoration and creation of more Turtle Dove breeding habitat - from thick thorny scrub and hedgerows to nest in, to plenty of flower and weed seeds that provide a source of energy- there is huge enthusiasm for helping this bird. Businesses, conservation groups, volunteer and community led initiatives are all helping, alongside farmers and landowners, to support these beloved birds, focusing effort in "Turtle Dove friendly zones" in eastern and south-eastern England. Creating habitat features – from hedgerows to ponds and wildflower lawns – even in gardens and local greenspaces can benefit Turtle Doves on their return to the UK, as well as a whole host of other wildlife.

According to RSPB.org - in their article titled "Giving the gift of hope: How an army of farmers and volunteers are helping to save rare Turtle Doves", hundreds of UK farmers, landowners and volunteers are helping to give the gift of hope for Turtle Doves, working with Operation Turtle Dove to provide better nesting and feeding habitat for the rare birds across southern and eastern England.

- A record year of collaborative effort across southern and eastern England is helping to turn around the fortunes of Turtle Doves, a globally threatened migratory dove.
- Operation Turtle Dove aims to boost numbers through the improvement of breeding habitat and food availability here in the UK, harnessing the power of hundreds of farmers, landowners and volunteers through science-led conservation.
- This year, the project is celebrating a record year of effort, with the number of farmers, landowners and volunteers involved rising to the highest ever since the project began in 2012.

Operation Turtle Dove is a partnership between the RSPB, Natural England, Pensthorpe Conservation Trust and Fair to Nature, which has led to the creation of 620 foraging and supplementary feeding sites for these special birds this year alone, a figure almost double the number provided in 2022. It has been working to create these feeding areas, maintain dense scrub and hedgerows as nesting sites, provide ponds for drinking and washing, and supply seed food – all of which have been shown to benefit Turtle Doves in focused trials. Now these conservation tools – as a tested formula for success - are being carefully rolled out by expert staff to improve the fortunes of these summer visitors right across the southern and eastern of England.

2.7 Ecological Modeling and Open Data Sources

According to Lajeunesse and Fourcade in their journal titled "Temporal analysis of GBIF data reveals the restructuring of communities following climate change", the journal uses occurrence data from the Global Biodiversity Information Facility (GBIF) to investigate how climate change has impacted species assemblages over time for different animal taxa. GBIF provides a large amount of geo-referenced species observation data that has allowed the researchers to analyze changes in community composition across broad spatial and temporal scales, even for taxa that typically lack long-term monitoring programs. Some key ways GBIF helped the research include:

1. Providing millions of occurrence records spanning 1990-2019 for 9 animal taxa from Europe and North America. This allowed analyzing changes over a 30-year period across a large geographic area.
2. The unstructured occurrence data from varied sources in GBIF allowed investigating community dynamics for most studied taxa without existing long-term data, opening up new research opportunities.
3. The abundance of geo-referenced records facilitated calculating community temperature indices (CTI) for "assemblages" in grid cells over time to detect changes in composition related to warming temperatures.

According to Fer et al. in their article "Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration", open data sources play a vital role in ecological modeling. They provide a vast amount of information that can be utilized to develop, validate, and improve models. The use of open data sources in ecological modeling helps to overcome model-data bottlenecks and accelerates the pace of discovery. Open data sources facilitate making data more accessible, scalable, and transparent, enabling the integration of expertise from the entire community, including both modelers and empiricists.

While data plays a critical role in modeling activities, its sheer volume and diversity can make it challenging to locate and obtain. To make data FAIR (Findable, Accessible, Interoperable, and Reusable), data producers are encouraged to use consistent naming structures and open file formats, such as comma-separated values or netCDF. Additionally, data should be stored in repositories that support standard, searchable metadata and machine-readable Application Programming Interfaces (APIs). When these repositories are part of jointly searchable networks, developers can leverage a single set of tools to access multiple sources.

Additionally, on the big data side, approaches for scientifically and computationally interacting with high-volume, high-velocity data become increasingly available. While it is important to generalize these cutting-edge tools and share with the community, modeling activities frequently involve a subset of data (e.g., a specific region or period) for which time to transfer data often exceeds the time to process it. Community cyberinfrastructure also provides a medium where a diverse array of data delivered by Internet of Things techniques can be integrated into models in a sensible manner. As developers combine cloud-based cyberinfrastructure tools with cutting-edge data platforms, this would free the users from their local constraints altogether. Empowering more groups to interact with large datasets brings its own push toward progress in terms of scientific proficiency and diversity.

Chapter Three

DESIGN AND METHODOLOGY

This chapter presents the methodologies used to test the further enhancement of the existing Logistic Reg-

ression. The resources used for this study will be discussed with the goal of providing a thorough comprehension of the researchers' methodologies. The emphasis is on how these results can enhance the current algorithm.

3.1 Research Design

The research utilizes a quantitative research design as it involves the use of logistic regression which is a mathematical and statistical method, to analyze numerical or measurable data related to habitat suitability. Logistic regression is particularly suited for binary classification problems, making it an ideal choice for assessing whether specific habitats are suitable for species like Turtle Doves (*Streptopelia turtur*) and Scottish Crossbills (*Loxia scotica*). The researchers aim to optimize the logistic regression algorithm by analyzing various hyperparameters, such as learning rate, regularization strength, and other algorithm-specific settings. The study involves testing multiple configurations to determine which combination of hyperparameters yields the best performance in identifying suitable habitats for these migratory avian species. This approach ensures that the algorithm is accurate and capable of providing meaningful insights into habitat suitability patterns, thus contributing to more effective conservation strategies.

3.2 Overview of the Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for binary classification problems, where the goal is to predict the probability of an outcome belonging to one of two classes. It models the relationship between one or more independent variables or predictors and a dependent variable or binary outcome using the logistic function, which maps predictions to a range between 0 and 1. The algorithm estimates the probability of an event occurring, and a threshold (commonly 0.5) is applied to classify observations.

3.2.1 Pseudocode of the Initial Logistic Regression Algorithm

Step 1: Load and Preprocess the Data

- Load climate and bird occurrence datasets.
- Generate pseudo-absences.
- Split data into training, validation, and test sets.

Step 2: Feature Engineering

- Drop irrelevant columns.
- Apply one-hot encoding for categorical variables.
- Align training, validation, and test sets.

Step 3: Train the Logistic Regression Algorithm

- Initialize a Logistic Regression algorithm with default hyperparameters and balanced class weights.
- Fit the algorithm to the training set.
- Evaluate the algorithm on the validation set.

PROBLEM: Limited hyperparameter tuning was performed, relying on default settings.

Step 4: Model Evaluation

- Predict on the validation and test sets.
- Plot the ROC curve.

The first step is the Loading and Preprocessing of Data, where the researchers begin by importing climate data from local NetCDF files. This dataset contains detailed information about various climate variables such as temperature, rainfall, wind speed, snow coverage, and ground frost. Data from the UK Met Office serves as the primary source for climate-related variables, ensuring geographic specificity to Great Britain.

To complement this, bird observation records are retrieved from the Global Biodiversity Information Facility (GBIF), which provides spatial and temporal information about the presence of the species of interest. The researchers then clean and preprocess these datasets by handling missing values, filtering for relevant geographical regions, and removing duplicate or erroneous entries. In presence-only datasets, pseudo-absence points are generated through random sampling to balance the presence-absence distribution, enabling the algorithm to differentiate suitable and unsuitable habitats. Finally, the combined dataset is split into training, validation, and test sets using stratified sampling to maintain the class distribution. Second step, the researchers prepared the dataset for algorithm training by focusing on transforming and aligning features. First, irrelevant columns, such as identifiers or text-based entries that do not contribute to predictive analysis, are removed to reduce redundancy. Categorical variables, such as locality names, are encoded using one-hot encoding, ensuring they are transformed into a format suitable for numerical algorithms like logistic regression. The final step in feature engineering ensures that the columns across the training, validation, and test sets are consistent. This alignment is crucial to prevent mismatches that could disrupt the training or evaluation phases of the algorithm. This step does not include feature importance analysis, meaning all variables are retained, which could potentially introduce noise and overfitting. Third step, the researchers initialized and trained the logistic regression algorithm. To address class imbalance in the dataset—where the number of absence points significantly outweighs the presence points—the class_weight parameter is set to 'balanced'. This adjustment ensures that both classes contribute equally to the algorithm's loss function during training. However, this step reveals the limitations in the initial approach which is the Limited Hyperparameter Tuning. The default hyperparameters are used, restricting the algorithm's ability to generalize across datasets. For example, regularization parameters that control overfitting are not optimized, resulting in suboptimal performance. Despite the limitation, the algorithm is trained on the training dataset, and its predictive performance is evaluated on the validation set. Metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) are computed to assess the algorithm's predictive capability. A Receiver Operating Characteristic (ROC) curve is plotted to visualize the trade-off between sensitivity and specificity. Fourth step, the researchers assessed the algorithm's generalizability by applying it to both the validation and test datasets. Predictions are generated, and performance metrics are computed to evaluate the algorithm's effectiveness in identifying suitable habitats for the target species. These metrics include:

- **Accuracy**: The proportion of correct predictions over the total dataset.
- **Precision**: The ratio of true positive predictions to all predicted positives, measuring the algorithm's ability to avoid false alarms.
- **Recall**: The ratio of true positives to all actual positives, evaluating the algorithm's ability to detect all relevant instances.
- **F1-score**: A harmonic mean of precision and recall, balancing their trade-offs.
- **AUC-ROC**: The area under the ROC curve, which measures the algorithm's ability to distinguish between presence and absence classes.

The researchers also plot the ROC curve to provide a graphical representation of the algorithm's performance, showcasing the trade-off between the false positive rate (FPR) and true positive rate (TPR). These evaluations highlight the strengths and weaknesses of the initial logistic regression algorithm, setting the stage for enhancements to address identified challenges.

3.2.2 Simulation of the Problem in the Initial Logistic Regression Algorithm

Problem: It is difficult to find the right set of hyperparameters for Logistic Regression.

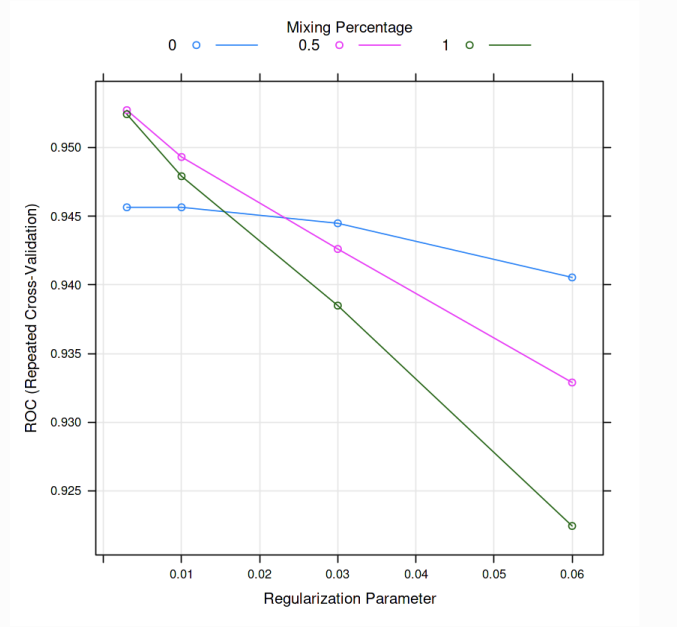


Figure 3.1 Simulated challenge of selecting optimal hyperparameters

Analysis:

The chart highlights the challenge of selecting optimal hyperparameters for logistic regression, as different configurations significantly affect algorithm performance. The x-axis represents the penalty parameter (controlling regularization strength), while the y-axis shows the ROC metric from repeated cross-validation, where higher values indicate better performance. The lines represent varying mixing percentages (0, 0.5, and 1), reflecting the combinations of regularization types like L1, L2, or ElasticNet. For a mixing percentage of 0, the ROC remains relatively stable, whereas at 0.5 and 1, performance drops sharply as the penalty parameter increases.

This demonstrates the sensitivity of logistic regression to hyperparameters, particularly in balancing overfitting and underfitting. The variability emphasizes the need for robust optimization techniques to systematically explore and identify the best hyperparameter settings for achieving optimal algorithm performance.

3.2.3 Pseudocode of the Proposed Enhancement of the Logistic Regression Algorithm

Step 1: Load and Preprocess the Data

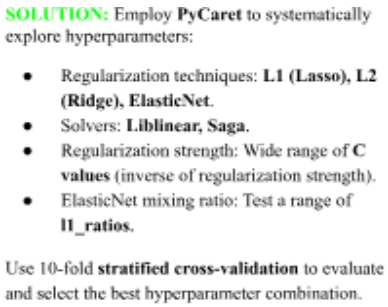
- Load climate and bird occurrence datasets.
- Generate pseudo-absences.
- Split data into training, validation, and test sets.

Step 2: Feature Engineering

- Drop irrelevant columns.
- Apply one-hot encoding for categorical variables.
- Align training, validation, and test sets.

Step 3: Train the Enhanced Logistic Regression Algorithm

- Address Imbalanced Data.
- Feature Selection with RFE.
- Normalization.
- Comprehensive Hyperparameter Tuning.
- Train the Final Algorithm.



Step 4: Model Evaluation

- Evaluate the trained algorithm on the training set using metrics such as: Accuracy, precision, recall, F1-score, and AUC-ROC.
- Test the algorithm on the test dataset to assess its generalizability.
- Visualize the algorithm's performance with ROC curves for all datasets (training, validation, test).

The researchers implemented multiple improvements over the initial algorithm to address its limitations, focusing on feature selection, data balancing, normalization, and hyperparameter tuning. Each sub-step is described below in detail:

1. Data Balancing with SMOTE

In the enhanced algorithm, the imbalance in the dataset is addressed using the **Synthetic Minority Oversampling Technique (SMOTE)**. SMOTE generates synthetic samples for the minority class (presence) based on its feature-space similarities. Unlike simple class weighting, SMOTE enables the algorithm to learn diverse patterns from the minority class, improving its **recall** and **F1-scores**. This ensures the algorithm generalizes better to unseen data while retaining the ability to predict the presence of the species.

2. Feature Selection with Recursive Feature Elimination (RFE)

The enhanced algorithm employs **Recursive Feature Elimination (RFE)** to systematically identify and retain the most important features. By using logistic regression as the estimator, RFE ranks features based on their predictive contribution and iteratively eliminates the least important ones. This process reduces noise and ensures the algorithm focuses on key predictors such as **locality**, **climate variables**, and **seasonal effects**, mitigating the risk of overfitting.

Mathematically, logistic regression uses coefficients βi to weight the features:

$$z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Where:

- β0: Intercept term.
- βi : Coefficient for feature xi .

Features with coefficients close to zero or with minimal impact on the prediction are systematically removed during RFE.

3. Normalization

To ensure that features with larger scales (e.g., rainfall) do not dominate smaller-scale features (e.g., ground frost), **Z-score normalization** is applied. The normalization formula is:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X: Original feature value.
- μ: Mean of the feature.
- σ: Standard deviation of the feature.

This scales all numerical features to a standard range, improving algorithm stability and interpretability.

4. Comprehensive Hyperparameter Tuning

The enhanced algorithm explores a wide range of hyperparameters to optimize its performance, focusing on the penalty type, regularization strength, and solver type. The tuning process employs **PyCaret** to automate this exploration and uses **10-fold stratified cross-validation** to evaluate configurations.

Regularization in Logistic Regression

Regularization introduces a penalty to the loss function to prevent overfitting. Three regularization techniques are explored:

- **L1 Regularization (Lasso)**: Adds the absolute values of the coefficients as a penalty:

$$\text{Penalty} = \lambda \cdot \sum_{i=1}^{n} |\beta_i|$$

- **L2 Regularization (Ridge** e coefficients as a penalty:

$$\text{Penalty} = \lambda \cdot \sum_{i=1}^{n} \beta_i^2$$

- **ElasticNet**: Combines L1 a

$$\text{Penalty} = \alpha \cdot \lambda \cdot \sum_{i=1}^{n} |\beta_i| + (1-\alpha) \cdot \lambda \cdot \sum_{i=1}^{n} \beta_i^2$$

Where:

- λ: Regularization strength.
- α : Balance between L1 and L2 regularization.

Hyperparameters Tuned

1. **Regularization Parameter (C)**: C is the inverse of λ, controlling the strength of regularization. A logarithmic search range of [10-5,10-3] is explored.
2. **Penalty Types**: L1, L2, and ElasticNet penalties are evaluated.
3. **ElasticNet Mixing Ratio (α\alphaα)**: Explored across a range of [0,1], from pure ridge to pure lasso regularization.
4. **Solver Types**: The algorithm uses solvers such as liblinear an d saga to handle L1, L2, and ElasticNet penalties efficiently.

5. Cross-Validation

To ensure robust evaluation, the enhanced algorithm uses **10-fold stratified cross-validation**. The dataset is split into 10 subsets, and the algorithm is trained on 9 while being validated on the remaining one. This process is repeated for all folds, and the results are averaged to assess performance.

Mathematically, for K-fold cross-validation:

$$\text{Mean Metric} = \frac{1}{K} \sum_{i=1}^{K} \text{Metric}_i$$

This approach reduces variance in evaluation metrics and provides a reliable estimate of the algorithm's generalization ability.

3.3 Methodology

The research aims to enhance the logistic regression algorithm by leveraging PyCaret to achieve the accuracy for predicting the habitat suitability of Great Britain migratory avian species, specifically Turtle Doves (*Streptopelia turtur*) and Scottish Crossbills (*Loxia scotica*). The methodology follows several steps to ensure accurate predictions, utilizing PyCaret's automated machine learning capabilities for data preprocessing, hyperparameter tuning, and algorithm evaluation to streamline the workflow and improve algorithm performance.

3.3.1 Data Collection

In terms of data collection, this study utilizes two comprehensive datasets: the eBird Observational Dataset (EOD), published by the Cornell Lab of Ornithology and distributed via the Global Biodiversity Informatics Facility (GBIF), and the HadUK-Grid dataset from the UK Met Office. To facilitate data acquisition, specific functions are employed to streamline the process. Preprocessing is conducted using Python libraries such as pandas and numpy, enabling the data to be reshaped, aggregated, and aligned with the spatial and temporal requirements of the study.

3.3.2 Evaluation Metrics

In assessing the performance of the Logistic Regression algorithm, the researchers utilized a set of evaluation metrics to comprehensively measure its predictive capabilities and reliability. Accuracy is calculated as the ratio of correctly predicted samples (True Positives + True Negatives) to the total number of samples, providing an overall measure of the algorithm's effectiveness. Precision focuses on the accuracy of positive predictions by dividing True Positives by the sum of True Positives and False Positives, indicating how well the algorithm avoids false alarms. Recall, also known as Sensitivity, measures the algorithm's ability to identify actual positives by dividing True Positives by the sum of True Positives and False Negatives. Lastly, the F1-Score harmonizes Precision and Recall into a single metric by computing their weighted average, ensuring a balance between these two crucial aspects. These evaluation metrics offered the researchers a robust framework for analyzing the Logistic Regression algorithm's performance and identifying areas for optimization. Thus, the following are the mathematical formulas for the evaluation metrics:

- **Accuracy:**

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

- **Precision:**

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall (Sensitivity):**

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **F1-Score:**

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 3.2 Mathematical formulas for the evaluation metrics

3.3.2 Data Visualization

The researchers utilized Matplotlib to visualize the performance of the enhanced Logistic Regression algorithm and its optimized hyperparameters. Through a series of plots, they analyzed metrics such as accuracy, precision, recall, and AUC-ROC, showcasing the algorithm's performance across training, validation, and test datasets. One of the primary functions employed was plot_roc_curves, which generated Receiver Operating Characteristic (ROC) curves for each dataset, visually highlighting the trade-off between true positive rates and false positive rates for the predictions. These plots also included AUC values to quantify the overall performance of the algorithm.

Additionally, the researchers implemented comprehensive preprocessing steps, including addressing imbalanced data through Synthetic Minority Oversampling Technique (SMOTE), feature selection via Recursive Feature Elimination (RFE), and normalization to improve algorithm performance. The effectiveness of the optimized Logistic Regression was evaluated using metrics like F1-score and AUC-ROC, which were visualized through PyCaret's in-built plotting tools.

The integration of these enhancements demonstrated a significant improvement in prediction accuracy and model generalization, as confirmed by higher silhouette-like scores derived from the model evaluation.

3.4 Requirement Analysis

The logistic regression algorithm enhanced with PyCaret for predicting habitat suitability of Great Britain migratory avian species was developed on a MacBook Pro 2020 with 8GB LPDDR4X of RAM, and a 256GB SSD. It is a model of Intel Core i5 (10th Generation).

In terms of programming language and libraries, the research utilized Python due to its versatility and extensive support for machine learning applications. Specific Python packages such as "xarray", "pandas", "numpy", "seaborn", and "matplotlib" were employed for data preprocessing, analysis, and visualization. The PyCaret library was leveraged for its automated machine-learning capabilities, enabling efficient model building, tuning, and evaluation. Development and testing were primarily conducted using Google Colab, which provided a cloud-based environment for seamless collaboration and access to computational resources.

The climate dataset used for this research was from and validated by the UK Met Office and the Cornell Lab of Ornithology, while the bird occurrence dataset used was from and validated by the Cornell Lab of Ornithology, the datasets are comprised with bird observation data and climate data spanning multiple decades, including variables such as location coordinates, temperature, and precipitation. These datasets were processed and aligned to evaluate habitat suitability accurately, leveraging PyCaret's built-in functionalities for streamlined experimentation and performance optimization.

3.5 Conceptual Framework

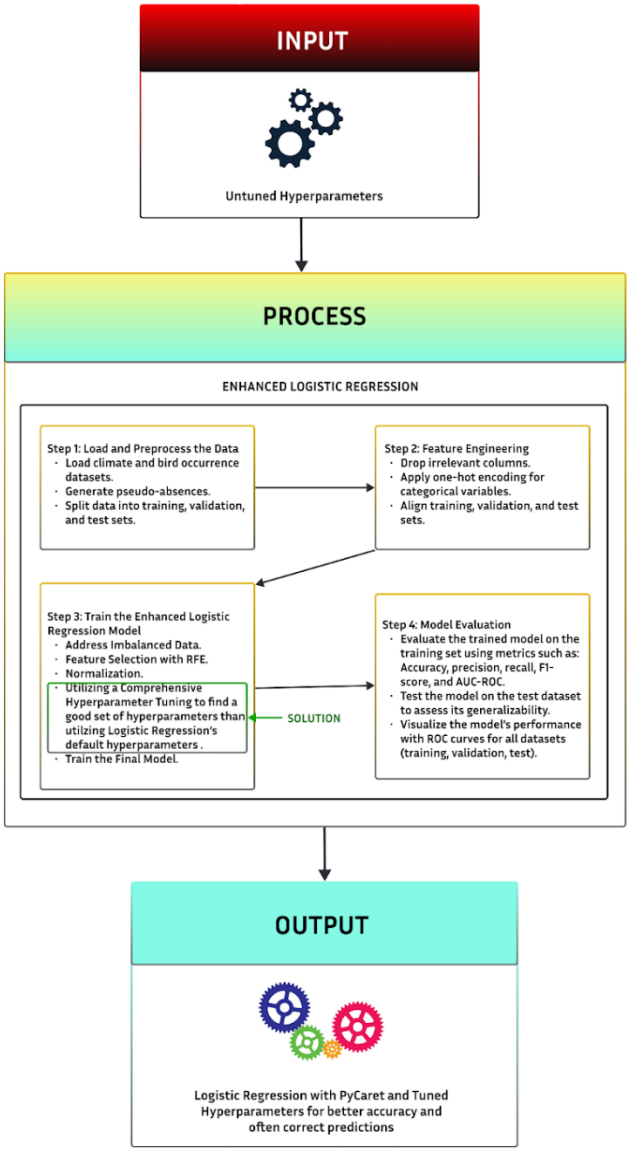


Figure 3.3 Conceptual Framework

This section presents the conceptual framework designed by the researchers, following the Input-Process-Output (IPO) model. The Input phase focuses on identifying critical components such as data, resources, and variables. In the Process phase, these inputs are transformed using specific methodologies, computational techniques, or procedural steps. Finally, the Output phase represents the results or outcomes generated through the process, such as an enhanced performance, or optimized solutions.

The figure illustrates the workflow of the enhanced Logistic Regression algorithm following the Input-Process-Output (IPO) framework. It begins with the Input stage, where untuned hyperparameters serve as the starting point. The hyperparameters significantly influence the algorithm's performance and require optimization. In the Process stage, the transformation begins with loading and preprocessing the data, including importing climate and bird occurrence datasets, generating pseudo-absences, and splitting the

data into training, validation, and test sets. Feature engineering follows, involving the removal of irrelevant columns, one-hot encoding for categorical variables, and aligning datasets for consistency. The algorithm is then trained by addressing imbalanced data, performing Recursive Feature Elimination (RFE) for feature selection, and normalizing the data. A key enhancement in this step is the integration of PyCaret for comprehensive hyperparameter tuning, replacing default Logistic Regression hyperparameters to optimize performance. Finally, the algorithm undergoes evaluation by calculating metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The evaluation extends to testing the algorithm on the test dataset and visualizing its performance through ROC curves for training, validation, and test datasets. In the Output stage, the process culminates in a Logistic Regression algorithm optimized with tuned hyperparameters, resulting in improved accuracy and reliability in predictions. This optimized output addresses the challenges posed by untuned parameters, enhancing the algorithm's ability to accurately predict habitat suitability for avian species.

Chapter Four
RESULTS AND DISCUSSION

In this chapter, results from the problems and the proposed objective is discussed by the proponents. The findings in this study are interpreted and described as new insights emerged from the results of the study's problems.

4.1 Comparison of the ROC Curves (Existing vs. Enhanced Algorithm)

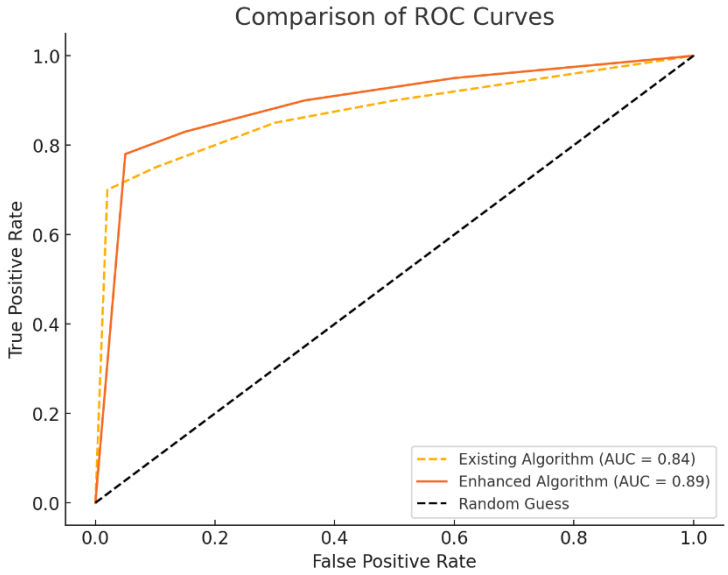


Figure 4.1 Comparing the ROC Curves (Existing vs. Enhanced Algorithm)

Figure 4.1 compares the ROC curves of the existing Logistic Regression algorithm and the enhanced algorithm during the test phase. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different probability thresholds, allowing for a visual assessment of the model's discriminative ability. The existing algorithm demonstrates a moderately good performance with an Area Under the Curve (AUC) of **84.25%**. This indicates that the algorithm has a reasonable ability to distinguish between positive and negative classes; however, its curve remains slightly below optimal, especially at higher False Positive Rates. In contrast, the enhanced algorithm shows a clear improvement with an AUC of **88.99%**, which represents a 4.74% increase in predictive capability. The curve for the

enhanced algorithm is consistently closer to the top-left corner of the plot, indicating superior performance in terms of True Positive Rate while maintaining a lower False Positive Rate. This improvement can be attributed to systematic hyperparameter tuning, the use of SMOTE for data balancing, and optimized regularization techniques implemented via PyCaret. The enhanced algorithm's higher AUC demonstrates its stronger ability to generalize to the test data and reduce bias. This improved performance suggests that the enhancements made—particularly in hyperparameter tuning and feature selection, allowing the algorithm to better capture the relationships within the test data while avoiding underfitting.

4.2 Comparison of Classification Metrics (Existing vs. Enhanced Algorithm)

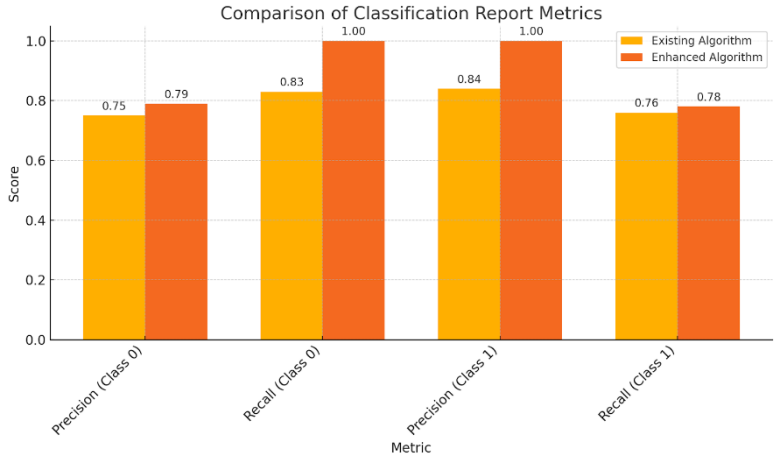


Figure 4.2 Comparing the Classification Metrics (Existing vs. Enhanced Algorithm)

Figure 4.2 presents a comparison of the precision and recall scores for both classes (Class 0 and Class 1) between the existing Logistic Regression algorithm and the enhanced algorithm. These metrics provide insights into the ability of each algorithm to correctly classify the two classes. For **Class 0** (negative class), the enhanced algorithm demonstrates a notable improvement. The **precision** increased from **0.75** to **0.79**, indicating fewer false positives. Similarly, the **recall** for Class 0 rose from **0.83** to **1.00**, meaning the enhanced algorithm was able to correctly identify all negative class samples during evaluation. For **Class 1** (positive class), the improvements are even more pronounced. The **precision** increased significantly from **0.84** to **1.00**, showing that all predicted positive cases were accurate with no false positives. The **recall** for Class 1 also improved from **0.76** to **0.78**, indicating a better balance between identifying true positives and minimizing false negatives. The enhancements observed in these metrics can be attributed to systematic hyperparameter tuning using PyCaret, data balancing with SMOTE to address class imbalance, and optimized feature selection. By refining the algorithm parameters and ensuring the algorithm focuses on the most relevant features, the enhanced algorithm achieved a better trade-off between precision and recall, particularly for Class 1, which is critical for identifying suitable habitats. These improvements highlight the enhanced algorithm's greater reliability and consistency in classifying both classes compared to the existing algorithm, thereby improving the overall performance of habitat suitability predictions.

4.3 Comparison of Acccuracy Results (Existing vs. Enhanced Algorithm)

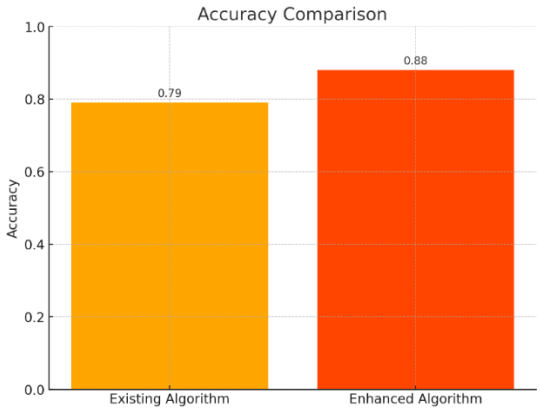


Figure 4.3 Comparing the Accuracy Results (Existing vs. Enhanced Algorithm)

Figure 4.3 illustrates the comparison of accuracy between the existing and enhanced Logistic Regression algorithms. The accuracy of the existing algorithm is **79%**, indicating a moderate performance in correctly predicting the habitat suitability for the target species. However, this level of accuracy suggests that the algorithm struggled to generalize effectively, likely due to the limitations of default hyperparameter settings and unbalanced data. In contrast, the enhanced algorithm achieved a significantly higher accuracy of **88%**, reflecting a substantial **9% improvement**. This increase can be attributed to several enhancements applied to the algorithm, including comprehensive hyperparameter tuning with the PyCaret package, which systematically optimized key parameters such as regularization strength and solver selection. Additionally, the use of SMOTE (Synthetic Minority Oversampling Technique) balanced the dataset, allowing the algorithm to capture patterns from both classes more effectively. Feature selection using Recursive Feature Elimination (RFE) further improved the algorithm's ability to focus on the most relevant predictors, reducing noise and enhancing its generalizability. The enhanced accuracy demonstrates that systematic improvements to the algorithm's configuration and data preprocessing pipeline significantly increase its ability to produce reliable predictions. This result underscores the importance of tuning machine learning algorithms to achieve optimal performance, particularly when addressing ecological challenges such as habitat suitability predictions.

4.4 Comparison of F1- Scores (Existing vs. Enhanced Algorithm)

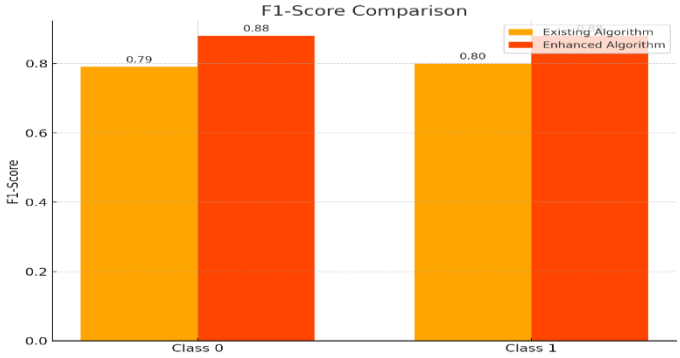


Figure 4.4 Comparing the F1- Scores (Existing vs. Enhanced Algorithm)

Figure 4.4 presents a comparison of the F1-scores for Class 0 and Class 1 between the existing algorithm and the enhanced algorithm. The F1-score serves as a balanced metric that combines precision and recall, making it particularly suitable for evaluating performance on imbalanced datasets. For Class 0, the existing algorithm achieved an F1-score of 0.79, while the enhanced algorithm demonstrated a notable improvement with an F1-score of 0.88. Similarly, for Class 1, the existing algorithm attained an F1-score

of 0.80, whereas the enhanced algorithm again outperformed it with an F1-score of 0.88. The improvement in the F1-scores for both classes reflects the enhanced algorithm's ability to better balance precision and recall. This improvement can be attributed to the comprehensive hyperparameter tuning, feature selection, and data balancing techniques (such as SMOTE) applied in the enhanced algorithm. By optimizing key parameters like regularization strength and solver types, the enhanced algorithm was able to minimize bias and variance while achieving more consistent predictions for both classes. These results further underscore the enhanced algorithm's reliability and its ability to generalize better across the dataset compared to the existing algorithm. The higher F1-scores for both classes indicate that the enhanced algorithm effectively reduces misclassifications, ensuring a more robust performance in predicting habitat suitability for the Scottish Crossbill.

Chapter Five
CONCLUSIONS AND RECOMMENDATIONS
Conclusions

The study explored the development of a robust habitat suitability prediction algorithm for migratory avian species in response to climate change, focusing on Scottish Crossbills (*Loxia scotica*), in Great Britain. Using Logistic Regression as the foundational algorithm, the research evolved from a basic implementation with limited optimization to a comprehensive enhanced algorithm that leveraged advanced machine learning techniques.

The initial logistic regression algorithm provided a baseline for habitat suitability predictions, achieving moderate accuracy and AUC scores. However, its performance was hindered by challenges such as the inclusion of irrelevant features, imbalanced datasets, and limited hyperparameter tuning. The lack of systematic feature selection resulted in noise within the algorithm, reducing its predictive power and contributing to potential overfitting. Moreover, while class weighting partially mitigated data imbalance, it was insufficient to capture patterns in minority class distributions effectively. Hyperparameter tuning, which was conducted manually and relied on default parameters, constrained the algorithm’s ability to generalize across diverse datasets.

The enhanced logistic regression algorithm addressed these limitations through several methodological advancements. The integration of Recursive Feature Elimination (RFE) refined feature selection, enabling the algorithm to focus on the most significant predictors, including locality-specific and climate-based variables. The application of SMOTE (Synthetic Minority Oversampling Technique) resolved class imbalance by generating synthetic samples, improving the algorithm's recall and F1-scores. Comprehensive hyperparameter tuning, powered by PyCaret, allowed systematic exploration of regularization types, solver configurations, and parameter ranges, optimizing the algorithm for superior performance. Additional measures, such as Z-score normalization and stratified cross-validation, ensured feature scaling consistency and robust evaluation.

The results demonstrated a substantial improvement in the enhanced algorithm's predictive capability, with higher accuracy, AUC, and F1-score metrics across the training, validation, and test datasets. The enhanced algorithm exhibited a stronger ability to generalize to unseen data, making it a valuable tool for conservation efforts. It proved effective in identifying suitable habitats under current and future climate scenarios, providing critical insights for ecological planning and decision-making. Overall, this research highlights the importance of methodological refinements in improving habitat suitability predictions and

underscores the utility of logistic regression as a reliable modeling framework when coupled with advanced techniques.

Recommendations

Based on the findings of this study, several recommendations can be made to extend and apply the insights gained. First, future studies should consider incorporating additional data sources, such as higher-resolution satellite imagery, soil composition data, and species interaction networks, to further refine habitat suitability algorithms. These datasets could enhance the ecological validity of predictions by accounting for a broader range of environmental factors.

Second, while this study focused on logistic regression, exploring more complex machine learning algorithms, such as random forests, gradient boosting machines, or neural networks, could yield further improvements in predictive performance. These algorithms might better capture nonlinear relationships and interactions between features, particularly in larger datasets. However, their computational complexity and interpretability trade-offs should be carefully weighed against the gains in accuracy.

Third, future studies should emphasize the dynamic nature of climate change by integrating time-series algorithms or ensemble forecasting techniques. These approaches could provide a more comprehensive understanding of temporal trends in habitat suitability, aiding long-term conservation planning. Additionally, extending the predictions to include species dispersal patterns and migration pathways could provide a more holistic view of species survival under changing climatic conditions.

Lastly, for practical application, integrating the enhanced algorithm into decision-support systems for conservation planning is recommended. These systems could assist policymakers, conservationists, and land managers in identifying priority areas for habitat restoration and protection. Collaboration with stakeholders to co-develop user-friendly tools and visualizations, such as web-based GIS platforms, could bridge the gap between research and actionable outcomes.

LIST OF REFERENCES

1. Baling, M., Stuart-Fox, D., Brunton, D. H., & Dale, J. (2016). Habitat suitability for conservation translocation: The importance of considering camouflage in cryptic species. Biological Conservation, 203, 298–305. https://doi.org/10.1016/j.biocon.2016.10.002
2. BBC. (2017, December 5). Scottish crossbill faces climate change extinction. https://www.bbc.com/news/uk-scotland-42227188
3. Crawford, B. A., Maerz, J. C., & Moore, C. T. (2020). Expert-Informed Habitat Suitability Analysis for At-Risk Species Assessment and Conservation Planning. Journal of Fish and Wildlife Management, 11(1), 130–150. https://doi.org/10.3996/092019-JFWM-075
4. Dallman, S. (2023). Cheat ML Model Creation with PyCaret. https://medium.com/codex/accelerate-your-machine-learning-workflow-with-pycaret-304a4759562d
5. Dahu, B. M., Alaboud, K., Nowbuth, A. A., Puckett, H. M., Scott, G. J., & Sheets, L. R. (2023). The Role of Remote Sensing and Geospatial Analysis for Understanding COVID-19 Population Severity: A Systematic Review. International Journal of Environmental Research and Public Health, 20(5), Article 5. https://doi.org/10.3390/ijerph20054298
6. Fer, I., Gardella, A. K., Shiklomanov, A. N., Campbell, E. E., Cowdery, E. M., Kauwe, M. G. D., Desai, A., Duveneck, M. J., Fisher, J. B., Haynes, K. D., Parton, W. J., Poulter, B., Quaife, T., Raiho, A., Schaefer, K., Serbin, S. P., Simkins, J., Wilcox, K. R., Viskari, T., & Dietze, M. C. (2020,

November 6). Beyond ecosystem modeling: A roadmap to community cyberinfrastructure for ecological data-model integration—PMC. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7756391/

7. Fieberg, J., Signer, J., Smith, B., & Avgar, T. (2021). A 'How to' guide for interpreting parameters in habitat-selection analyses. Journal of Animal Ecology, 90(5), 1027–1043. https://doi.org/10.1111/1365-2656.13441
8. Ford, C. (2016, April 22). Visualizing the Effects of Logistic Regression | UVA Library. University of Virginia Library. https://library.virginia.edu/data/articles/visualizing-the-effects-of-logistic-regression
9. Forestry.com. (n.d.). Scottish Crossbill—Forestry.com. Forestry Editorial. Retrieved March 31, 2024, from https://forestry.com/animals/birds/scottish-crossbill/
10. GeeksforGeeks.org. (2024). How to Optimize Logistic Regression Performance. https://www.geeksforgeeks.org/how-to-optimize-logistic-regression-performance/
11. Jiang, X., & Xu, C. (2022). Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastasis Using Clinical Data. Journal of Clinical Medicine, 11(19), Article 19. https://doi.org/10.3390/jcm11195772
12. Kumar, A. (2023, December 6). Using GridSearchCV with Logistic Regression Models: Examples. Analytics Yogi. https://vitalflux.com/gridsearchcv-logistic-regression-machine-learning-examples/
13. Lajeunesse, A., & Fourcade, Y. (2023). Temporal analysis of GBIF data reveals the restructuring of communities following climate change. The Journal of Animal Ecology, 92(2), 391–402. https://doi.org/10.1111/1365-2656.13854
14. Mallon, Dr. K., & Wormworth, J. (n.d.). Bird Species and Climate Change. Retrieved April 21,2024, from https://library.sprep.org/sites/default/files/22_6.pdf
15. Mancino, C., Hochscheid, S., & Maiorano, L. (2023). Increase of nesting habitat suitability for green turtles in a warming Mediterranean Sea. Scientific Reports, 13(1), 19906. https://doi.org/10.1038/s41598-023-46958-4
16. Melanee Group. (2023, May 21). A Comprehensive Analysis of Hyperparameter Optimization in Logistic Regression Models. Medium. https://levelup.gitconnected.com/a-comprehensive-nalysis-of-hyperparameter-optimization-in-logistic-regression-models-521564c1bfc0
17. Plumer, B. (2014, September 9). Study: Climate change puts hundreds of bird species at risk. Vox. https://www.vox.com/2014/9/9/6128065/map-how-global-warming-could-put-hundreds-of-bird-species-at-risk
18. Thanda, A. (2023). What is Logistic Regression? A Beginner's Guide. https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#:~:text=Advantages%20of%20logistic%20regression&text=Training%20is%20the%20process%20of,as%20compared%20to%20other%20methods
19. Townsend, J. P., & Aldstadt, J. (2023). Habitat suitability mapping using logistic regression analysis of long-term bioacoustic bat survey dataset in the Cassadaga Creek watershed (USA). Science of The Total Environment, 895, 165077. https://doi.org/10.1016/j.scitotenv.2023.165077
20. Wang, X., Xu, Q., & Liu, J. (2023). Determining representative pseudo-absences for invasive plant distribution modeling based on geographic similarity. Frontiers in Ecology and Evolution, 11. https://doi.org/10.3389/fevo.2023.1193602

21. W.D. (2019). Scikit-learn's Defaults are Wrong. https://ryxcommar.com/2019/08/30/scikit-learns-defaults-are-wrong/
22. Yusuf, I. S., Tessera, K., Tumiel, T., Nevo, S., & Pretorius, A. (2021, November). On pseudo-absence generation and machine learning for locust breeding ground prediction in Africa. ResearchGate. https://www.researchgate.net/publication/356026991_On_pseudo-absence_generation_and_machine_learning_for_locust_breeding_ground_prediction_in_Africa