Frog Species Distribution Model*

STAT 432 Project

Kun Hu
kunh
STAT 432

Vishwadeepsinh Sarvaiya
vjs6
STAT 432

Zixuan Wang
zixuan14
STAT 432

ABSTRACT

Frogs, as representative species of amphibian, are indicators of ecological health and proxies for biological diversity. These keystone species help define the entire ecosystem. Without them, the ecosystem would be entirely different or cease to exist altogether. A Species Distribution Model was generated using species occurrence data and terrestrial climate data in Australia.

After joining the two datasets, exploratory data analysis was performed to understand the story behind data. Observations were made regarding correlation between features. Feature Engineering was carried out to leverage spatial data and sampling methods were implemented to address class imbalance. With the data ready for model fitting, five different classification algorithms were implemented to obtain the most suitable model for this case. Random Forest was found to be the model with best accuracy. Finally, the model was optimized with hyper-parameter tuning

Model evaluation is obtained using different scoring metrics. The study is concluded with the acknowledgement that this model fitting was done under the assumption that recorded data implies the true presence or absence of the particular species at a given location.

1. Introduction

The goal is to predict the occurrence of a single species of frog for a region using occurrence data and terrestrial climate data at a coarse spatial resolution. Frogs are an indicator species. This means they are a go-to for scientists wanting to find out more about the environmental health of a particular ecosystem. Frogs have permeable skin, which means they are very sensitive to pollutants, and because they can live on both land and in the water, they are a good indicator of the health of these two different environments. This particular species identify as a keystone species, which means that it is a critical species for the functioning of ecosystem. Without the keystone species, the ecosystem would be entirely different or cease to exist altogether. As indicators of ecological health and proxies for biological diversity, the disappearance of frogs is of great concern. Where frogs occur, we see healthy, thriving, resilient ecosystems. Where frogs have disappeared, we see ecosystems in poor health. All the 2030 Sustainable Development Goals (SDGs) are underpinned by healthy ecosystems. This means we will not reach our goals if we do not prevent and reverse the loss of healthy ecosystems.

Frogs are poorly served by existing species distribution models. They have very localized distributions, more restricted than suggested by a potentially suitable habitat, and therefore existing models struggle to represent their range accurately.

The output of this project is a species distribution model of one species of frog. The problem was converted into a classification problem by treating occurrence data of target species as presence and occurrence data of other species as the absence point. Throughout the project, various techniques were used to optimize the performance of this model.

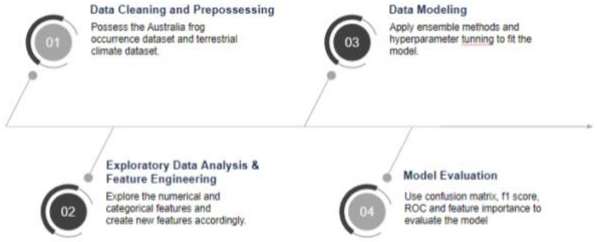

Figure 1: Project Overview

2. Literature Review

Species distribution modeling (SDM) is a methodology – a set of procedures, definitions, and techniques – built on a foundation of core ecological [1] and biogeographical [2] concepts about the relationship between species distributions (or other biotic response variables describing aspects of biodiversity) and the physical (abiotic) environment [1]. Species distribution models are one of the most widely used ecological tools, a cornerstone in many countries worldwide of environmental regulation and conservation [2].

In a study (Echelpoel et al., 2015), a concise overview of five different SDM techniques was provided along with advantages/drawbacks and future perspectives. Firstly, each of the techniques were introduced by providing a general understanding of the algorithm, conceptual framework and case studies where the

technique proved to be useful. The five techniques are as follows: Decision Trees, General Linear Models (GLM), Artificial Neural Networks (ANN), Fuzzy Logic, Bayesian Belief Networks (BBN).

Technique	Advantages	Drawbacks
Decision Trees	- Transparent modeling technique - Can deal with small datasets.	- Data-driven - Large datasets can create complex trees.
GLM	- Easy to use - Useful for specialized problems	- Data-driven - Assumes presence of specific distribution of response variable.
ANN	- High tolerance for noise and errors. - Ability to recognize relations between predictors and response variables in absence of system knowledge.	- Lack of guidelines for optimal design - Low ecological relevance
Fuzzy Logic	- Absence of strict boundary values. - Ability to address uncertainty by possibility approach.	- Increased complexity with increasing number of predictors. - Information loss.
BBN	- Addresses uncertainty - Complements empirical data with knowledge.	- No temporal dynamics - Information loss - Construction of knowledge-based rules is time intensive.

Table 1: Advantages/Drawbacks of SDM techniques

Insights on future perspectives were discussed by pointing out some of the limitations of current data-driven techniques that they rely on observational data without substantial integration of existing ecological knowledge and do not take in consideration climate change [3].

A review study [4] that consolidates data-driven modeling and expert/domain knowledge by establishing best-practice standards for models in biodiversity assessments. These standards were claimed to provide a hierarchy of reliability and ensure transparency. Best-practice standards were claimed to be generally applicable to a variety of available data and modeling approaches and reflect the evidence required for the particular type of question addressed or decision being taken. The framework proposed by the study consists of four levels:

Levels	Description
Gold	Ideal data, next generation modeling approach (seldom available)
Silver	Imperfect but best-available data, current cutting-edge approaches
Bronze	Minimum standard of data and procedure currently acceptable
Deficient	Use of data and procedure currently unacceptable in practice

Table 2: Levels of best-practice standards

The best-practice standards were obtained by studying 400 different SDMs based on 15 parameters. The parameters are displayed on the vertical axis in the following plot:

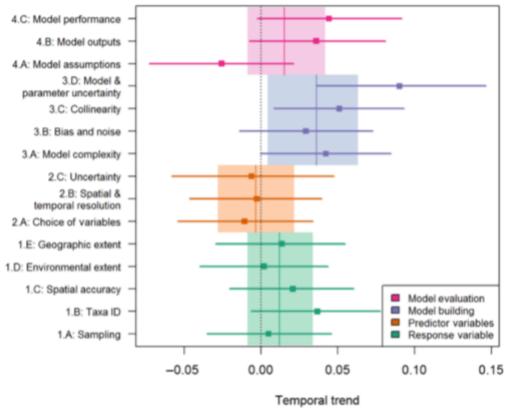

Figure 2: Boxplot on parameters (from Araújo et, al.)

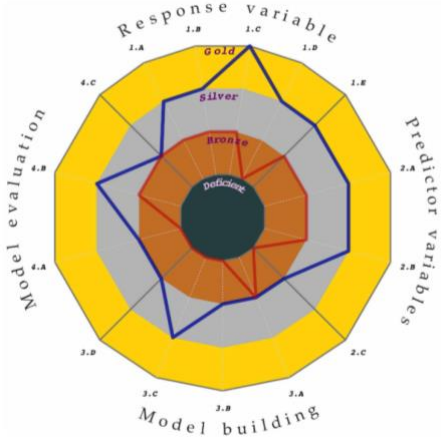

Figure 3: Best-practice standards achieved (from Araújo et, al.)

The researches then discussed strengths/limitations and feasibility of these best-practice standards. The study was concluded with an encouragement to improve upon these standards periodically [4].

3 Dataset Description

For the Australia frog occurrence dataset, we limited our search to frogs in the Greater Sydney area found between the start of 2015 to the end of 2019. For response variable transformation, we used the occurrence points of other species (crinia signifera) as absence points for the target species. The terrestrial climate dataset was merged with the occurrence dataset based on the geo-coordinates. The merged data set contains 9418 subjects with no missing values.

3.1 Data Dictionary and Data Cleaning

There are eight variables and one response. "occurrenceStatus," which is a binomial variable that returns '1' if the mentioned species is found and '0' otherwise. "eventDate" which is the date of the observation, an ordinal variable. "decimalLatitude" and

"decimalLongitude" which are the latitude and longitude of the location, a continuous variable. "species," the species of the frogs, which is nominal. "ppt_mean," which is accumulated precipitation per month in millimeters, a continuous variable. "soil_mean", which is the moisture in soil measured at the end of the month, represented in millimeters, a continuous variable. "tmax_mean", the maximum of the air temperature, measured two meters above the surface, which is a continuous variable. As well as "tmin_mean", the minimum of the air temperature, also measured two meters above the surface, which is a continuous variable.

Feature Name	Description	Type
occurrenceStatus	**1: if the mentioned species found, 0: otherwise**	**Binomial**
eventDate	Date for observation	Ordinal
decimalLatitude	Location	Continuous
decimalLongitude	Location	Continuous
species	Species of frog	Nominal
ppt_mean	Accumulated precipitation (ppt) – Cccumulated monthly in millimeters	Continuous
soil_mean	Soil moisture (soil) – Soil moisture in millimeters at end of month	Continuous
tmax_mean	Maximum air temperature (tmax) – 2 meters above surface	Continuous
tmin_mean	Minimum air temperature (tmin) – 2 meters above surface	Continuous

Table 3: Working Data Dictionary

4 Exploratory Data Analysis & Feature Engineering

4.1 Numerical Variables

4.1.1 Distribution of Terraclimate Variables. Correlations were discovered between the occurrences of a species and Terraclimate variables by plotting the distribution of 4 terra-climate variables according to the location on the maps and the occurrence of Litoria Fallax. We noticed that the higher temperatures and soil moisture, the more likely the species of Litoria Fallax appear. There is also a positive correlation between maximum air temperature and the soil's moisture. We will discuss these observations in detail in the following paragraphs.

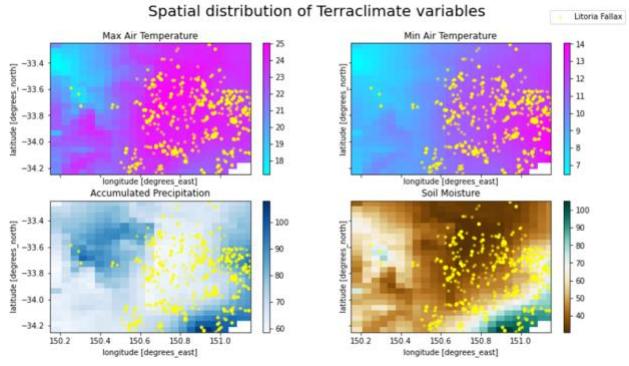

Figure 4: Distribution of Terraclimate Variables on Maps. Correlation between the occurrences of a species and Terraclimate variables.

4.1.2 Location Variables - Coordinate System Conversion. We decided to convert our current measurements to a coordinate system after reviewing related literature. We took the readings of the locations in latitude and longitude and transformed them into radius coordinates by taking the two readings' squares and adding them together.

	decimalLatitude	decimalLongitude	ppt_mean	soil_mean	tmax_mean	tmin_mean	loc_sumsquare	loc_atan2
count	9418.000000	9418.000000	9418.000000	9418.000000	9418.000000	9418.000000	9418.000000	9418.000000
mean	-33.774927	150.805720	75.340248	55.048435	22.864126	11.352835	154.541721	-0.220327
std	0.203942	0.295894	9.932862	15.187789	1.984061	1.955527	0.310490	0.001171
min	-34.249929	150.150402	59.283333	30.616667	17.160002	6.461668	153.805984	-0.223946
25%	-33.841313	150.623268	67.966667	42.200001	22.821669	10.618335	154.359527	-0.220628
50%	-33.756317	150.918000	73.933334	56.266666	23.786665	11.918336	154.645301	-0.220346
75%	-33.675679	151.062409	82.516670	62.683334	23.933336	12.873335	154.805678	-0.219519
max	-33.251121	151.150000	107.949997	105.150002	25.020002	14.045002	154.942023	-0.216864

Table 4: Continuous Variables after Location Data Transformation. Addition of the sums of squared latitude and longitude, arctan of latitude and longitude to the continuous variables.

4.1.3 Outlier Detection. We plotted boxplots for all eight continuous variables and spotted outliers in all continuous variables, except in longitude, and sum squared location.

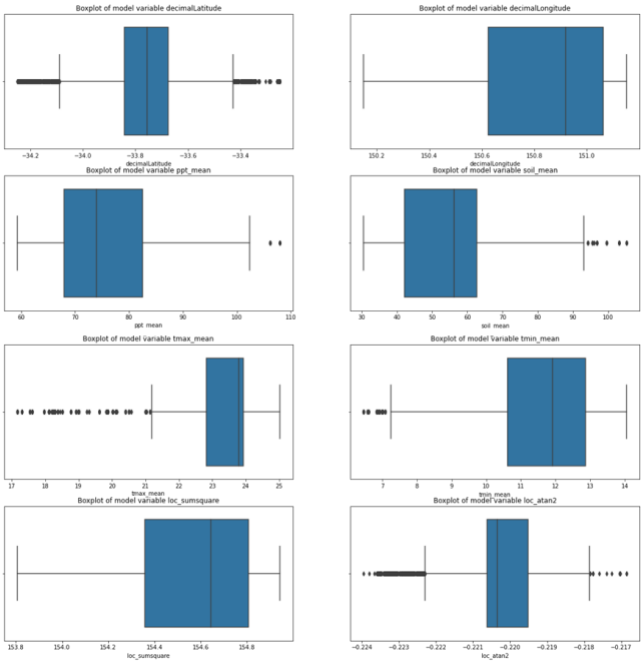

Figure 5: Outlier Detection Box Plots. There are outliers in all continuous variables except for longitude and sum squared location.

4.1.4 Correlation Analysis. We carried out a correlation analysis for all continuous variables. We detected strong and positive correlations in the continuous variables.

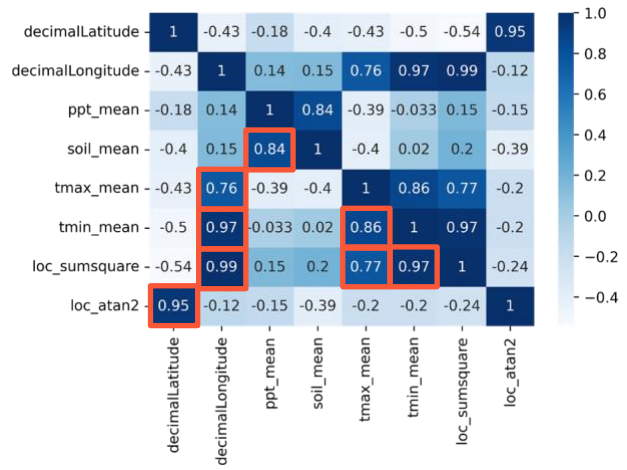

Figure 6: Correlation Analysis Matrix. Some strong and positive correlations existed in the continuous variables.

4.1.4 T-test. A t-test is any statistical hypothesis test in which the test statistic follows a student's t-distribution under the null hypothesis. Before applying the T-test, we checked if the variables followed a normal distribution. We concluded that none of the variables follows a normal distribution; therefore, we could not execute a t-test without variable transformation.

Continuous Variable	Follows normal distribution?
decimalLatitude	No
decimalLongitude	No
ppt_mean	No
soil_mean	No
tmax_mean	No
tmin_mean	No
loc_sumsquare	No
loc_atan2	No

Table 5: T-test. None of the variables follows a normal distribution, therefore we were not able to perform t-test without variable transformation.

4.2 Categorical Variables

4.2.1 Date Variable Transformation. We only had one categorical variable, the event date, in the string format. Therefore, we separated it into three categorical variables – weekday, month, and year, which is a process called data transformation.

	species	eventDate	weekday	month	year
0	Litoria Fallax	2017-11-12	7	11	2017
1	Crinia Signifera	2019-09-19	4	9	2019
2	Litoria Fallax	2019-11-03	7	11	2019
3	Litoria Fallax	2019-11-04	1	11	2019
4	Litoria Fallax	2015-06-16	2	6	2015

Table 6: Date Variable Transformation. Transformation from strings to 3 categorical variables – weekday, month and year.

4.2.2 K-means Clustering. We conducted a K-means Clustering on six continuous variables – decimalLatitude, decimalLongitude, ppt_mean, soil_mean, tmax_mean, and tmin_mean. Upon observation, K=2 is selected. We also learned that the corresponding within-group error (sum of squares of the distances to the centroids) is 32053.45 and that the corresponding between-group error (silhouette score) is 0.52.

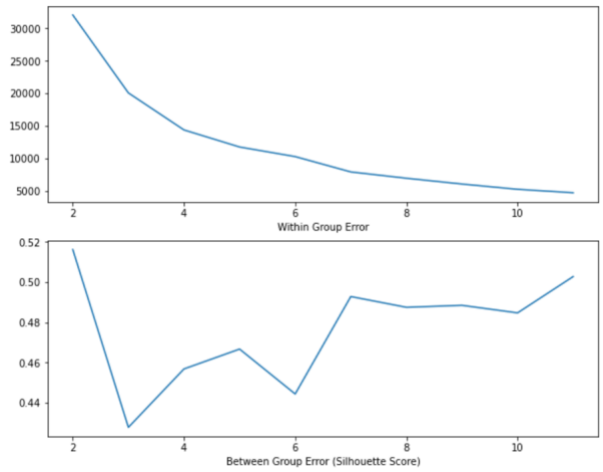

Figure 7: K-means Clustering. K-means Clustering on 6 continuous variables. Upon observation, K=2 is selected.

We plotted the occurrence of Litoria Fallax on a map (right) and our K-mean clustering labeled data (left). This label will be used for further analysis, which is mentioned later.

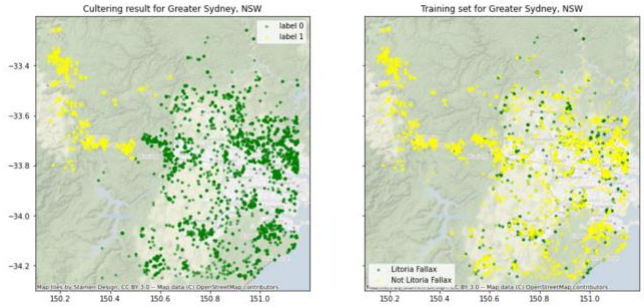

Figure 8: Distributions of different clusters and comparison with the training dataset.

4.1.3 Cross Tabulation Analysis. We executed a Cross-tabulation analysis and Chi-square tests on four categorical variables, which

means that only weekdays are independent. First, we performed a cross-tabulation analysis to compare the occurrence of Litoria Fallax (orange bars) across four categorical variables. We detected no difference in the occurrence of Litoria Fallax between different weekdays. The occurrence of Litoria Fallax is the lowest for June and July, goes up consecutively until the end of the year, then goes down. We also noticed that the occurrence of Litoria Fallax increased from 2015 to 2017, sharply decreasing afterward, which led us to investigate the causes of this during this time.

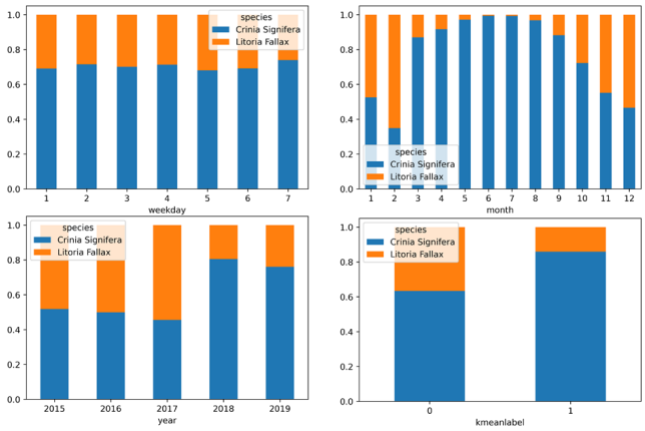

Figure 9: Cross Tabulation Analysis on weekday, month, year and K-mean label.

4.1.4 Chi-square test. We carried out chi-square tests on four categorical variables. The only independent variable is 'weekday,' which confirmed our observation mentioned in the previous paragraphs.

Variable	P-value	Decision	Result
weekday	0.1375736	Accept H0	Independent
month	2.89E-65	Reject H0	**Dependent**
year	6.86E-25	Reject H0	**Dependent**
kmeanlabel	8.45E-16	Reject H0	**Dependent**

Table 7: Chi-square tests on 4 categorical variables. Only weekday is independent.

5. Data Modeling

The dataset was randomly split into training and testing sets before fitting the data into the model. The training set contains 70% of the dataset and the testing set contains 30% of the dataset. 11 features were used to fit the model. Among the 11 features, *Month* and *weekday* features are categorical and *Loc_sumsquare*, *loc_atan2*, *decimalLongitude*, *decimalLatitude*, *tmax_mean*, *tmin_mean*, *ppt_mean*, *soil_mean* are continuous features. Class imbalance bias exists in the original dataset. To address imbalanced data problem, we applied different resampling methods so that the number of the target species matches that of the absent species. Different sampling methods were evaluated based on the accuracy score calculated from 10-fold cross validation. We used the same cross validation method to evaluate the different ensemble methods and conducted hyperparameter tuning for the best performed method. The confusion matrix, f1 score, ROC and AUC, and feature importance were used to evaluate the performance of the final model.

5.1 Imbalanced Data

Based on the frog species distribution of the training set shown in Figure 10, the number of the other species is 4609, which is 2.3 times as large as that of the target species. Therefore, class imbalance bias exists in the dataset. Five resampling were applied to address the class imbalance bias. The distributions of the resampled training set are shown in Table 8.

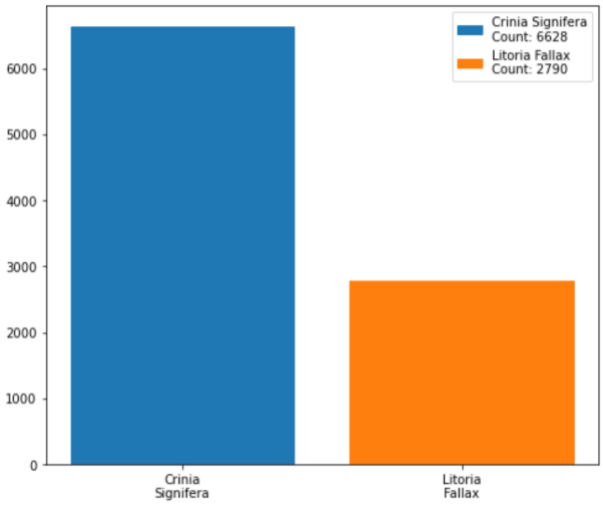

Figure 10: Frog species distribution of the training set

	Other Species	Target Species
Orignial Distribution	4609	1983
Random Under Sampling	1983	1983
Random Over Sampling	4609	4609
TomekLinks Under Sampling	4554	1983
SMOTETomek Sampling	4564	4564
ADASYN Sampling	4880	4609

Table 8: The distributions of the resampled training set

For each of the resampling method, 10-fold cross validation was performed the mean accuracy was calculated to select the best performed resampling method. The mean accuracy and the corresponding variance of different resampling methods is shown in Figure 11. Random Over Sampling, with 0.92 as its mean accuracy, reached the highest mean accuracy compared with other resampling methods. Therefore, the training set resampled by Random Over Sampling was used to fit into the further models.

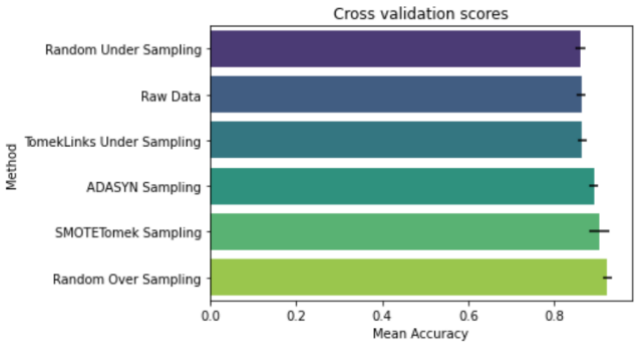

Figure 11: Mean accuracy of different resampling methods

5.2 Model Selection

The resampled training dataset was fitted into five classification trees ensemble methods. The five classification trees ensemble methods are Gradient Boosting Trees, Adaptive Boosting Trees, Extra Trees, XGBoosting Trees and Random Forest. The parameters of all these classifiers were set to be default and the mean accuracy of the 10-fold cross validation was used as the evaluation metric of the performance of different classifiers.

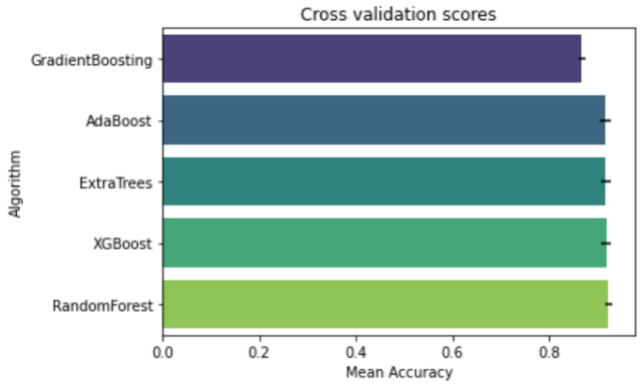

Figure 12: Mean accuracy of different classifiers

Random Forest Classifier, with 0.92 as its mean accuracy, reached the highest mean accuracy compared with other classification trees ensemble methods. Therefore, Random forest Classifier was used as the ultimate algorithm. The mean accuracy and the corresponding variance of different classifiers is shown in Figure 12.

5.3 Hyperparameter Tuning

Grid search was applied to find the optimal parameters. The parameters used for grid search were shown in Table #.

Parameters	Values
max_depth	10, **50**, 100, None
max_features	1, **3**, 10
min_samples_split	**2**, 3, 10
min_samples_leaf	**1**, 3, 10
n_estimators	**1000**

Table 9: Parameter grid for Random Forest Classifier (The selected values are bolded in the table)

According to scikit-learn, "the max_depth means the maximum depth of the tree. If None, then nodes are expanded until all leaves contain less than min_samples_split. The max_features means the number of features to consider when looking for the best split. The min_samples_split means the minimum number of samples required to split an internal node. The min_samples_leaf means the minimum number of samples required to be at a leaf node. The bootstrap means whether boostrap samples are used when building trees. The n_estimators is the number of trees in the forest".
Based on the grid search result, the max_depth was set to be 50; the mx_features was set to be 1, the min_samples_split was set to be 2, the min_samples_leaf was set to be 1. Boostrap was set to be True. All the selected values are bolded in Table 9.

5.4 Model Evaluation

The model was trained based on the best selected parameters using the training data and was tested using the testing set. Accuracy rate, precision rate, recall rate and f1 score were used to evaluate the performance of the Random Forest Classifier, as shown in Table 10.

Metrics	Values (%)
Accuracy Rate	86.31
Precision Rate	73.60
Recall Rate	80.16
F1 score	77.20

Table 10: Evaluation metrics of Random Forest Classifier

The accuracy rate measures the number of correct predictions over all predictions and is equal to 86.31%. The precision rate describes the number of correct positive predictions over all positive predictions and is equal to 73.60%. The recall rate describes the number of correct positive predictions over all positive instances and is equal to 80.16%. F1 score combines both precision and recall rate and is equal to 77.20%. The evaluation metrics proved that the Random Forest can predict the new data well.
The ROC (receiver operating characteristic) curve was used to demonstrate the performance of the Random Forest Classifier

throughout all the thresholds, as shown in Figure 13. The AUC (area under the ROC curve) was also calculated. The AUC of the Random Forest Classifier is 0.93.

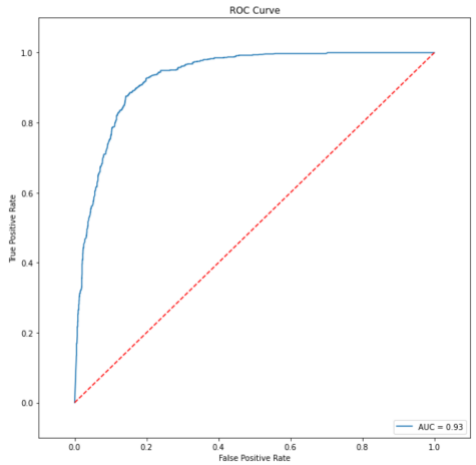

Figure 13: ROC and AUC of the Random Forest Classifier

The feature importance plot was shown in Figure 14. Among all the features, *month*, *loc_sumsquare*, *decimalLongitude*, *tmax_mean*, *decimalLatitude* are the top 5 important features used in the Random Forest Classifier. Two of the top 5 important features were not provided by the original dataset, proving that our feature engineering did help improve the model performance.

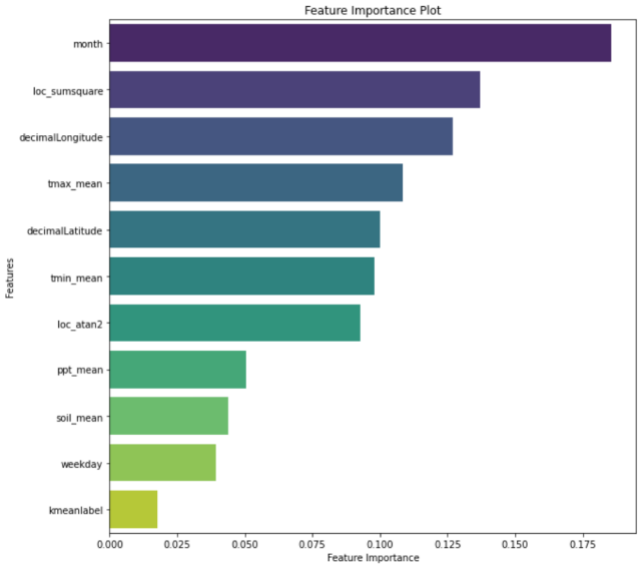

Figure 14: Feature importance plot

6 Conclusion

This Project aimed at developing a multi-dimensional frog density distribution model using statistical learning methods to predict the density of a frog species – "Litoria Fallax" with respect to other frog species. We applied the random over sampling method to address the class imbalance problem and trained the data using Random Forest Classifier. After performing hyperparameter tuning, the model reached 96.31% accuracy rate and 77.20% F1 score based on the testing set. The AUC reached 0.93. Overall, we believe that the proposed species density distribution model can provide great help on predicting the target species using the occurrence and climate data.

REFERENCES

[1] Whittaker RH, Levin SA, and Root RB (1973) Niche, habitat, and ecotope. American Naturalist 107: 321–338.

[2] Holdridge LR (1947) Determination of world formations from simple climatic data. Science 105: 367–368

[3] Van Echelpoel, W., Boets, P., Landuyt, D., Gobeyn, S., Everaert, G., Bennetsen, E.,Mouton, A., Goethals, P.L.M., 2015. Chapter 6 - species distribution models for sustainable ecosystem management. E.M.. In: Park, Y.-S., Lek, S., Baehr, C., Jørgensen, S.E.B.T.-D. (Eds.), Advanced Modelling Techniques Studying Global Changes in Environmental Sciences. Elsevier, pp. 115–134 http://dx.doi.org/10.1016/B978-0-444-63536-5.00008-9

[4] M. B. Araújo, R. P. Anderson, A. M. Barbosa, C. M. Beale, C. F. Dormann, R. Early, R. A. Garcia, A. Guisan, L. Maiorano, B. Naimi, R. B. O'Hara, N. E. Zimmermann, C. Rahbek, Standards for distribution models in biodiversity assessments. Sci. Adv. 5, eaat4858 (2019).