EY 2022 Better Working World Data Challenge

Level 1: Local Frog Discovery Tool

(Supporting Document)

BY: Team Data Ninjas

INTRODUCTION

Life in all its various forms is biodiversity. This variety of life is fundamental to the function of ecosystems, the health of forests – and even our prosperity. This challenge is about building computational models to locate biodiversity, specifically frogs, since frogs are a go-to for scientists wanting to study the health of a particular ecosystem.

PROBLEM STATEMENT

Globally, biodiversity loss is believed to be a threat to the ability to achieve the UN Sustainable Development Goals, including poverty alleviation and food, water, and energy security. As a result, there is an urgent need to prioritize the geographical areas and species that most need help. There is therefore a need to develop a species distribution model (SDM) for one Australian frog species using only variables from the TerraClimate dataset.

PROJECT AIM

To develop a species distribution model for litoria fallax across Australia using weather data from the TerraClimate dataset.

KEY ELEMENTS OF APPROACH

- Data Sampling

  In an attempt to search for frogs in the entire Australia, we generated a bounding box for Australia by picking coordinates from the lower left corner and upper right corner of Australia and choosing a ten-year window from the start of 2010 to the end of 2019. This gives a varied landscape of bushland, plains, rivers, and urban areas, and a larger dataset.

- Species

  To include diversity in our training dataset and make the machine learn differences between the target(Litoria Fallax) and other species, we have used all five species instead of only two. This we believe will help make better accurate classification.

- Sampling bias

  To address sampling bias, pseudo-absence technique was used. This technique uses the occurrence points of other species (Crinia Signifera, Crinia Glauerti, Ranoidea Australis, Austrochaperina) as absence points for the target species(litoria fallax), which was coded to 1 if the occurrence species is litoria fallax, and 0 if the species is not litoria fallax.

- Class Balancing

  To solve the class imbalance nature of the dataset, we used SMOTE(Synthetic Minority Over-sampling Technique) to up-sample the target species so that their numbers match that of the other species. This technique we believe is effective as synthetic samples are generated for the minority class.

- Null values

  Statistical imputation was used to fill the null values of predictors variables with median and mean based on the occurrence. We have chosen to fill in the null values instead of deleting them because the deletion will lead to loss of useful information for the training.

- Model Selection

  To develop an algorithm or species distribution model (SDM) that will reduce the cost and time required to locate biodiversity, while also improving accuracy, we had to choose a model that is very stable, very fast, robust, and one that can easily predict the distribution of frogs more accurately. We believe Extra trees algorithm was the best option for this problem and for satisfying the needs of an SDM.

UNIQUENESS OF OUR APPROACH

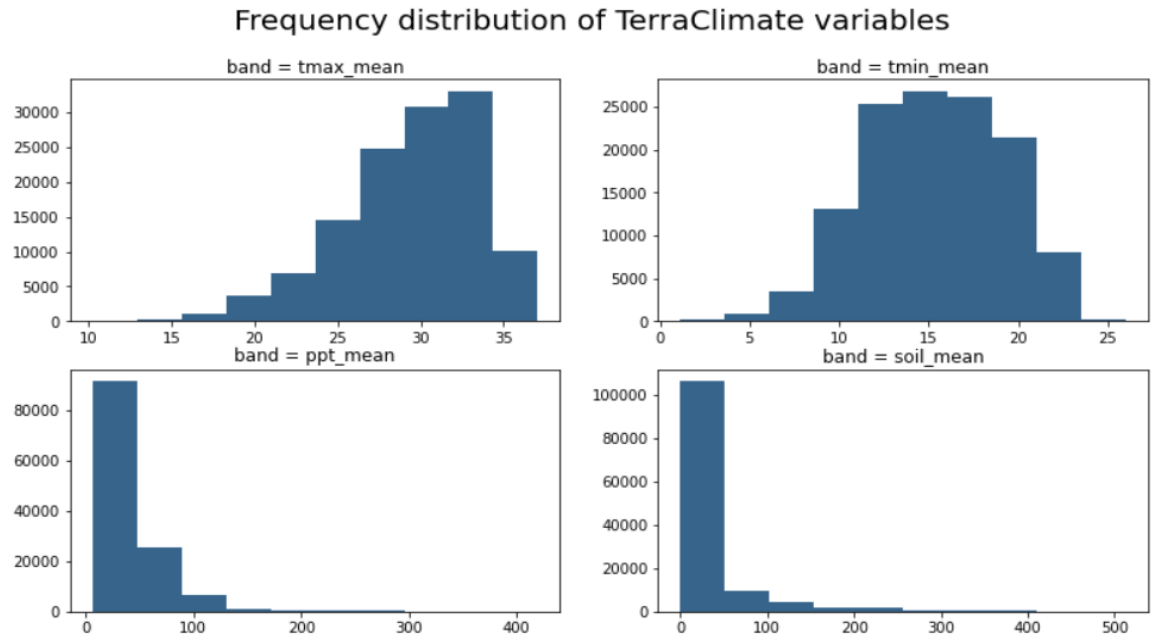
The model selected to achieve the end goal has the ability to reduce any biases and variety that the unseen datasets might have, this makes its ability to generalize on unseen data efficient. One other unique feature is the hyper parameters of the model which enhances its optimization ability or performance on unseen data. And finally, the techniques used to solve the various issues present in the dataset.

DATASETS

The target dataset is made up **78, 318 records** of the frog dataset and the predictor dataset is the TerraClimate dataset with the maximum monthly temperature, the minimum monthly temperature, the mean monthly precipitation, and the mean soil moisture as features making up the predictor dataset used for the training. The time frame used is a ten-year window from the start of 2010 to the end of 2019.

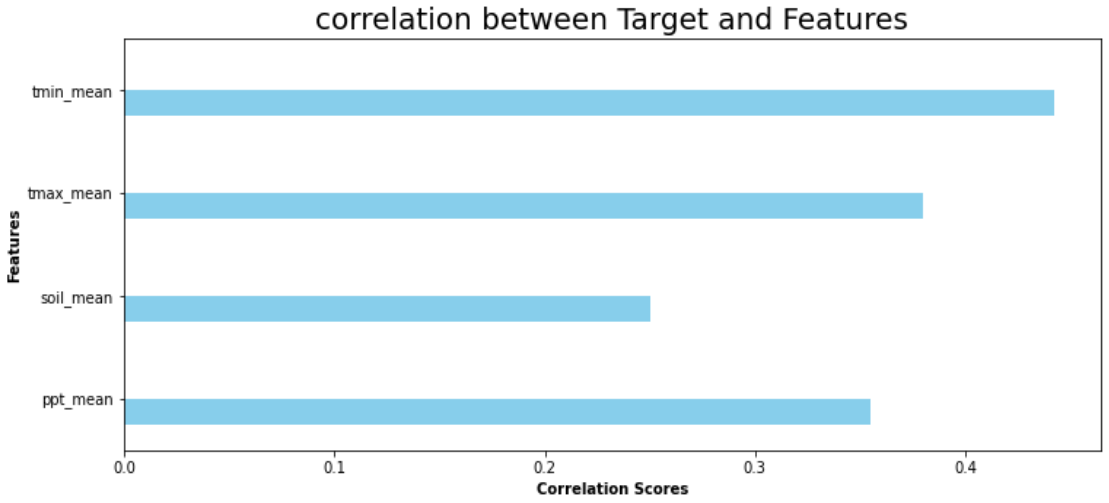
DATA PREPARATION

The distribution of the predictor data set was explored and three of the features were skewed with one almost normally distributed (tmin_mean feature) , hence missing values were filled based on the target species and other species and the distribution of the features. Hence, median values were used for skewed features and mean for the normally distributed.



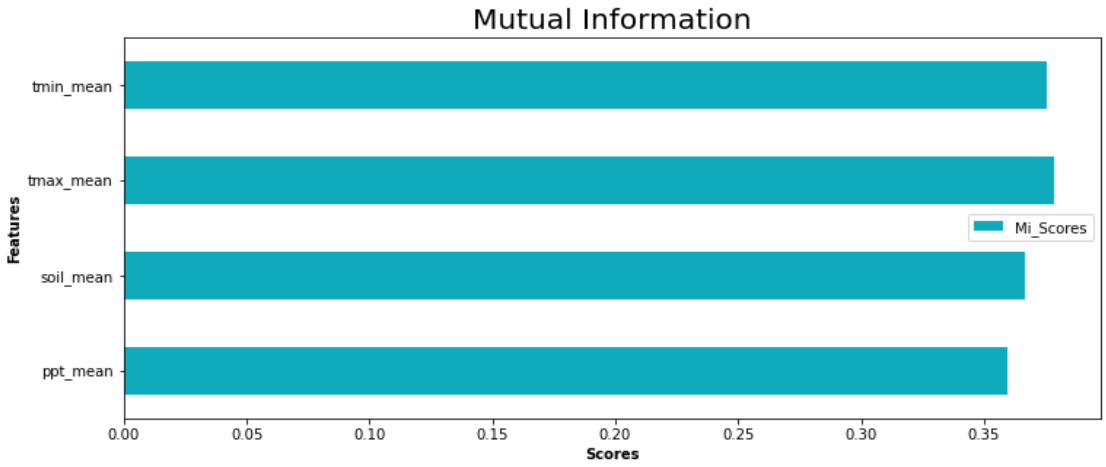
Point biserial correlation was used to find if there was any relationship between our target and the features and if the relationship was strong or not. And we found out that there is a positive relationship between the target and the predictors, even though the relationship is not strong.

	Occurence_Correlation	P_value
ppt_mean	0.355091	0.0
soil_mean	0.250313	0.0
tmax_mean	0.380096	0.0
tmin_mean	0.442001	0.0



For features selection, Mutual Information was used to see the amount of information we can obtain from the target variable given the features. And we came out that all the four features are informative features.

```
Feature 0: 0.359614
Feature 1: 0.366871
Feature 2: 0.378357
Feature 3: 0.375669
```



TRAINING METHODS

Algorithms tried includes:

Extra trees Classifier, Random Forest Classifier, Light GBM Classifier, and Logistic Regression.

These algorithms were used with their default parameters. A 10-fold cross-validation was done to determine the accuracy and standard deviation of accuracy for each of these algorithms:

	model_name	metric_name	std
3	LOG	0.7728	0.0285
0	LGBM	0.8533	0.0296
2	RAD	0.8566	0.0283
1	EXT	0.8572	0.0275

EVOLUTION AND DECISION FOR MODEL SELECTION

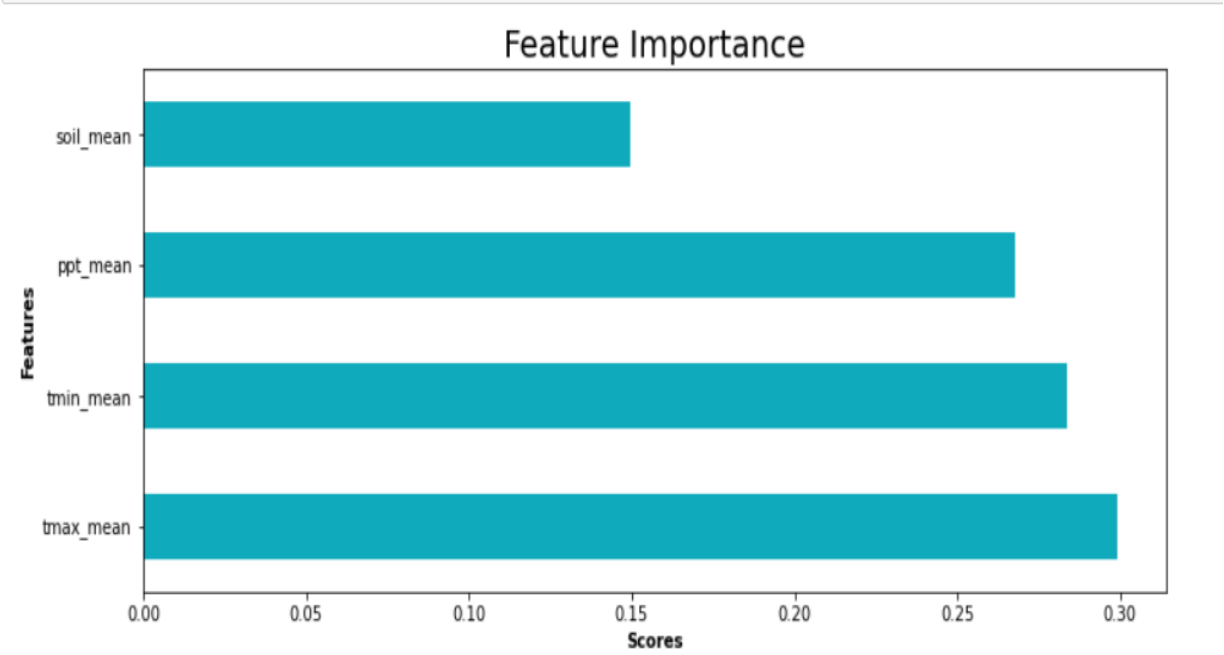
After evaluation of the different algorithms using the accuracy, the metric gives a quick idea of how robust a given model is. Based on this score, the best model – Extra trees Classifier was selected and parameter tuning using randomized search cross-validation to get the hyper parameters to enhance the performance and stability of this model.

MODEL VALIDATION

The model was trained on **80%** of the entire training data and tested or validated on the remaining **20%** of the data. This was to validate the performance of the trained model on unseen data and the accuracy and F1 score is determined.

Furthermore, a StratifiedKFold (10-folds) was used on the entire training data to split our data into folds, ensuring there is always a balanced number of frogs and non-frogs in each fold and to ascertain the performance of the model on different samples of the data. This allows the model to be trained and tested on these different samples during each fold, and the average of all accuracies are determined as the performance of the model.

HIGHEST PERFORMING FEATURES



The feature importance graph above represents the "importance" of each feature. The feature with the highest score is the tmax_mean (maximum temperature mean) followed by tmin_mean (minimum temperature mean). This means that these features will have a larger effect on the model.

ACCURACY-SCORE ON UNSEEN DATA

The accuracy score on the unseen data (submission template.csv file) is **75 %.** Which means that the model **correctly predicted 75 %** of both the presence and absence of frog species out of the total number of species introduced to it.

CHALLENGES

The first challenge we encountered was about the data sampling, how to get a good sample for the training. The second challenge was about the best technique to use to solve class imbalancing issues and finally, one minor challenge about this project has to do with the execution time when extracting theTerraClimate dataset (frog data and weather data) from Microsoft planetary system.