Name : Mun Chong Soo

**Level 1 : Local Frog Discovery Tool**

* This challenge is to predict the occurrence of a single species of frog for a single location using a single data source at a coarse spatial resolution.
* The output will be a species distribution model of one species of frog. Species distribution models are one of the most widely used ecological tools, a cornerstone in many countries worldwide of environmental regulation and conservation.
* Why frogs? 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.
* Because they have permeable skin, frogs 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.
* 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.
* 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 won’t reach our goals if we don’t prevent and reverse the loss of healthy ecosystems.

**Challenge Objective**

* Build a Species Distribution Model for the frog species - "Litoria Fallax" across Australia using TerraClimate variables.

**Dataset Used**

* For Target variable - Frog occurrence datasets (Australia)
* For Predictor variables - TerraClimate dataset (Climatic variables) available on Microsoft Planetary Computer portal (You will find additional information for the data to be used in the Data Description tab).

Table of Contents

1. Key elements of approach …………………………………………………………….. 3

2. Innovative approach …………………………………………………………………... 3

3. Target dataset & predictor dataset(s) used in the data preparation …………………… 3

4. Feature engineering and features selection approach ………………………………… 4

5. Model Validation ……………………………………………………………………… 4

6. Algorithms Used ……………………………………………………………………… 5

7. Feature Importance Plot ………………………………………………………………. 5

8. Model Selection ………………………………………………………………………. 6

9. F1-score on unseen data ……………………………………………………………… 6

10. Other elements of approach ………………………………………………………….. 6

11. Breakthroughs that helped improve the overall score ………………………………… 6

12. Hardest thing about this challenge ……………………………………………………. 6

13. Inference Runtime and ways to improve it …………………………………………… 6

1. Describe the key elements of your approach and assumptions of your approach if any

* **Binning the Terraclimate data in 2-year interval**
  + Binning the Terraclimate data in 2 years interval seems to work the best for 2014-2015, 2016-2017, 2018-2019. For example, joining Terraclimate data from 2014-2015 with frog data from 2014-2015 and calculate the feature metrics for Terraclimate data. Repeat this for every other 2-year interval data until 2019.
* **Calculating other metrics other than mean for Terraclimate data**
  + Other metrics like standard deviation, median, min, max have almost the same predictive power on the target label. Standard deviation has the highest importance among them.
* **Sample weighting instead of downsampling / upsampling for imbalanced data**
  + Sample weighting assigns higher weightage to the minority class label during training. This ensures that even though the training algorithms see less of the minority class label, their relative importance will not be ignored.
  + As compared to downsampling, we will not lose any information on the majority class.
  + As compared to upsampling, we will not add possibly any noise to the dataset which could downgrade the model performance.

1. What is innovative and unique about your approach?

* **Feature engineering other metrics other than mean**
* To enrich the feature dataset even more, generating more features is defnitely the way to go. Feature engineering mean, median, standard deviation, min, max was done to populate the dataset with even more features that could potentially improved the model performance.

1. Target dataset & predictor dataset(s) used in the data preparation stage along with the time frames. For example – Frog dataset, Terraclimate, JRC Water variables etc
   * **Binning the Terraclimate data in 2-year interval**

* Binning the Terraclimate data in 2-year interval seems to work the best for 2014-2015, 2016-2017, 2018-2019. For example, joining Terraclimate data from 2014-2015 with frog data from 2014-2015 and calculate the feature metrics for Terraclimate data. Repeat this for every other 2-year interval data until 2019.
* During inference time, taking only the Terraclimate data from 2014-2019 and make prediction based on that feature set

1. Describe your approach to data preparation steps like missing value treatment, outlier treatment, checking correlation etc. Feature engineering and features selection approach
   * **Without Interpolating the terraclimate feature maps**
     + Retrieving the terraclimate feature maps without interpolating / warping it afterwards seem to work best. It is probably because warping the feature maps may introduce noise as well as losing some important information in it.
   * **Feature Engineering & Feature Selection**
     + Feature engineering different metrics like mean, median, standard deviation, min, max using the Terraclimate feature maps.
     + Calculating all these different metrics for each of the 18 Terraclimate variables :
       1. **aet** (Actual Evapotranspiration, monthly total), units = mm
       2. **def** (Climate Water Deficit, monthly total), units = mm
       3. **pet** (Potential evapotranspiration, monthly total), units = mm
       4. **ppt** (Precipitation, monthly total), units = mm
       5. **ppt\_station\_influence**
       6. **vap\_station\_influence**
       7. **tmax\_station\_influence**
       8. **tmin\_station\_influence**
       9. **q** (Runoff, monthly total), units = mm
       10. **soil** (Soil Moisture, total column - at end of month), units = mm
       11. **srad** (Downward surface shortwave radiation), units = W/m2
       12. **swe** (Snow water equivalent - at end of month), units = mm
       13. **tmax** (Max Temperature, average for month), units = C
       14. **tmin** (Min Temperature, average for month), units = C
       15. **vap** (Vapor pressure, average for month), units = kPa
       16. **ws** (Wind speed, average for month), units = m/s
       17. **vpd** (Vapor Pressure Deficit, average for month), units = kpa
       18. **PDSI** (Palmer Drought Severity Index, at end of month), units = unitless

* Calculating the 5 different metrics : mean, median, standard deviation, min, max on these 18 variables would result in 5\*18 = 90 different features
* All these 90 features will then be used to train a binary classification model

1. Describe your approach to model validation
   * Randomized Search 5-Fold Cross Validation
     + The dataset is too huge for a brute-force grid search cross validation. Randomized search cross validation seems to be a more reasonable choice for doing model validation
     + The hyperparameters for randomized search cross valdiation is as follows:
       1. n\_iter = 100
       2. cv = 5
       3. scoring = roc\_auc

The number of iterations is 100, meaning 100 different combinations of the hyperparamers will be tried and for each of the combination, 5-fold cross validation is used. In total, 5\*100 = 500 models were trained. It took quite abit of time to do randomized search cross validation.

1. List the training methods you used (algorithms, model selection approach, model results)

* Algorithms Used
  + XGBoost
    - F1-score : 0.76
  + Random Forest
    - F1-score : 0.65

1. Describe your highest-performing features (please include an importance plot)

* **Feature Importance Plot of XGBoost**
  + There are 5 different importance types of XGBoost :
  1. ‘weight’: the number of times a feature is used to split the data across all trees.
  2. ‘gain’: the average gain across all splits the feature is used in.
  3. ‘cover’: the average coverage across all splits the feature is used in.
  4. ‘total\_gain’: the total gain across all splits the feature is used in.
  5. ‘total\_cover’: the total coverage across all splits the feature is used in.
* ‘gain’ importance type was used to generate the feature importance plots using XGBoost as the model
* Turns out that the highest-performing features are srad\_stad (standard deviation of downward surface shortwave radiation), vap\_mean (mean of vapor pressure), and tmin\_median (median of min temperature).

Table

Description automatically generated with medium confidence

1. Describe the evolution of your approach for selecting the best model and decisions you made at each point
   * **Model Selection**
     + Ensemble models were proven to work well on tabular data, as we have seen almost all winning models on Kaggle structured data competition have been ensemble models such as Random Forest and XGBoost. XGBoost has gain enormous amount of traction in recent years, due to its unprecedented performance on tabular data.
     + Random forest was used initially for model training but the f1-score was decent, having f1-score of 0.65. Whereas XGBoost performed better with an overall f1-score of 0.76.
2. Prediction accuracy score on the unseen data (submission template .csv file) on which the model was validated

* **Prediciton accuracy score**
  + The trained XGBoost model was saved as a pickle file, which then be used to predict the target label on the unseen data. The F1-score of the model on the unseen data is 0.76.

1. Any other element of your approach that you would like to highlight?

* **Filter Frog Response Variables**
* Filter frog data that has ‘uncertaintyInMeters’ less than 100. The assumption is that the latitude and longitude of the frog record may not be accurate since its ‘uncertaintyInMeters’ is huge. This may help reduce noise in the dataset

1. Describe the most important breakthrough that helped you to improve your score

* **Using different models**
* Trying out different ensemble models like Random Forest and XGBoost. XGBoost offered a much better performance boost as compared to Random Forest.

1. What was the hardest thing about solving this problem?

* **Data Understanding and Domain Knowledge**
  + Most of the time were spent trying to understand the data source and making sense out of it. Having a good understanding of the data source may help in data preparation, data manipulation as well as feature engineering. All of these would eventually boost the model performance.

1. How long does it take to apply the model to the test set to generate a results file (inference runtime)?Could this be made more efficient?

* **Inference Time on Test set**
* The inference time on test set was about 5 minutes. The reason that it takes this long is that the whole feature map(All longitudes & longitudes of the bounding box) was first extracted from the terraclimate data and this whole feature map was then fed into the trained XGBoost model for prediction. The output from this model is a predicition map of the entire region. However, not all of the points on the prediction map are of our interest since our test set only contains a subset of it. So the conclusion is that to make it more efficient, we could perhaps take only the latitude & longitude of the test set and map those latitudes & longitudes to the terraclimate data to retrieve the relevant features, which could then be fed to the trained XGBoost model for prediction.