Local Frog Discovery Tool

Sneha Kumar

Image source: https://upload.wikimedia.org/wikipedia/commons/8/87/Litoria_fallax.jpg

General Approach

Data Preparation

Sub-sampling datasets & performing data manipulation and visualization techniques to find and address bias, imbalance, skewed distributions

Base Model Selection

Train 6 different base models. Narrow down 3 best models with highest accuracies from in-sample evaluation

Hypertuning I

The hypertuning of parameters are conducted for 3 selected models, which are evaluated using test-training split and cross validation

Ensemble Learning

3 selected models with optimized parameters are used as base estimators for voting classifier

Hypertuning II

The hypertuning of weights are conducted to perform soft voting within the best estimators.

Model Evaluation

Final model is evaluated once again with in-sample evaluation and cross-validation

Innovative Elements of Approach

01

Conducting a **preliminary selection** within 6 base models allows us to **sieve out higher-performing models,** which is more **efficient** when optimizing models and perform ensemble learning

02

Hypertuning of parameters is performed for **both the base estimators and the ensemble model** to ensure that the fitted model accurately represents the dataset and is more likely to pick up isolated frog occurrences

03

Soft voting was found to produce better results. Since **localized frog distributions are restricted**, voting with predicted probability of output class can provide a more accurate prediction of the class.

Datasets

1. Target Dataset

- **Frog occurrence** dataset
- Sub-sample: entire region of Australia, from 2016 to 2020

2. Predictor Dataset

- **TerraClimate** dataset
- Sub-sample: entire region of Australia, from 2016 to 2020
- 4 metrics: mean maximum monthly air temperature (tmin_mean), mean minimum monthly air temperature (tmax_mean), mean accumulated precipitation (ppt_mean), soil moisture (soil_mean)

Data Preparation

Sampling Bias

- Bias: Heavy bias in urban areas since frogs are more likely to be encountered by humans and frogs also cluster around towns, parks, and bushes
- Solution: Pseudo-absence points. Occurrence values are represented by binary values of 0 (absence of Litoria Fallax) or 1 (presence of Litoria Fallax)

Class Balancing

- Imbalance: There are significantly more data points with an occurrence label of 0 than that of 1
- Solution: Down-sampling. Absence points are sampled to match the number of presence points

Feature Engineering

- Skewness: Present in all predictor variables, except for tmin_mean
- Solution: standardization. Variables are scaled during the model selecting and building stage in order to ensure rare/extreme cases are covered, where isolated frog occurrences may occur (crucial data points for frog conservation)

Data Preparation

Removing NA values

After joining the datasets, observations (rows) containing NA values for any of the variables were omitted from the training data

Predictor & response variables

- Prediction and response variables are separated into a dataframe (X) and array (y).
- Longitude, latitude, and response varaibles are drop from X
- y contains only the occurrence class labels (0 or 1)

Train-test Splitting

X and y are split into training and testing datasets. This is helpful to validate the performance of our models while hypertuning parameters

Base Model Selection & Hypertuning I

6 base models

- 6 base models which generally work the best for SDMs are created

(Gradient boosting, K-nearest neighbors, Random forest, Naive Bayes, Support Vector Machine, Logistic Regression)

- All variables are scaled to be standardized with the help of machine learning pipelines

Evaluation

- Each of the models are fitted (with default parameters)
- In-sample evaluation accuracy calculated *(using val.score)*
- In-sample evaluation is sufficient to efficiently eliminate models that are not likely to work well with the provided nature of data

Top 3 Models

- Parameters hypertuned for top 3 models with highest accuracies

(Random forest, Gradient boosting, Support Vector Machine)

- Grid search of specified parameters and cross validation are used to select the best-performing parameters.
- New models with selected parameters are validated with the testing dataset.

Ensemble Learning & Hypertuning II

Ensemble Algorithm

- 3 models with hypertuned parameters are selected as our base estimators for ensemble learning
- Voting classifier was found to be the best choice. Voting of predicted output between multiple models would ensure that isolated frog occurences are better accounted for (if one base model doesn't pick it up, another base model may pick it up)

Soft Voting

- Soft voting uses predicted probability (instead of a binary output) of output class from each model to produce the final output
- This is helpful to take into account the restricted localized distributions of frogs
- The weights for soft voting are hypertuned. The best weights were found to be 2,2,1, which allows to prioritize the better-performing models in the voting.

Evaluation

- The final model is a voting classifier with base estimators of gradient boosting, random forest, SVC)
- In-sample evaluation: F1-score, accuracy, to ensure the model is well-trained
- Out-sample evaluation: 5-fold cross validation, to ensure model performs well on unseen data

Model Evaluation

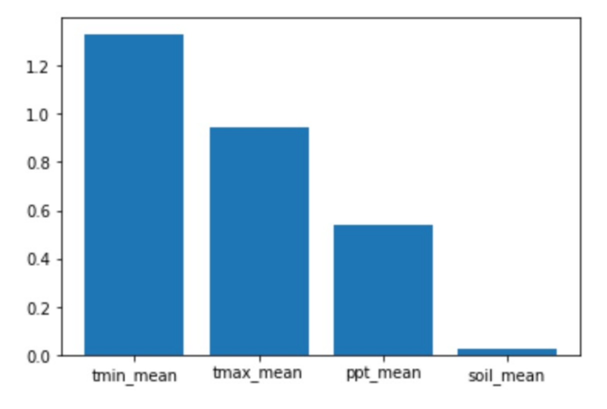
Evaluation & Validation

F1 score: 0.86
Accuracy score: 0.86
Cross-validation (mean) score: 0.84
Performance on platform: 0.74

Feature Importance

Most important features: tmin_mean, tmax_mean
Somewhat important: ppt_mean
Very little importance: soil_mean

Importance Plot



THANK YOU

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution