EY Biodiversity Challenge

Final Report

Project By:
Samarth Verma
Haeun Kim
Jayasree Lakshmi Narayanan

1. Activities worked this week

Over the past week, our primary focus was to improve the performance by enhancing data preprocessing, refining features and evaluating more advanced machine learning models.

This week, we concentrated on optimizing our classification model to achieve better predictive performance and decided on the final model. We began by identifying variables that had conceptual relevance to frog presence and low collinearity. Based on correlation analysis and domain understanding we finalised the following features: ['tmax', 'def', 'ppt', 'ws', 'q', 'soil', 'vpd', 'pet','Occurrence Status']. Second, to address class imbalance (1: 3,792 vs. 0: 2,520), we applied RandomOverSampler to balance the target variable. This oversampling makes the model's ability improve significantly. We tried to change the ratio charging the penalty to '1', but it didn't help. Lastly, we addressed potential outliers in the data using the Z-score method to reduce their impact on the model.

Initially we trained and fine tuned two models - XGBoost and LightGBM. Building on the progress, we further experimented with additional models this week, including ExtraTreesClassifier. After tuning hyperparameters for all models and comparing performance:

Model	LightGBM	XGBoost	**ExtraTreesClassifier**
Test Accuracy	0.8248	0.8296	**0.8310**
F1 Score	0.82	0.83	**0.83**

The **ExtraTreesClassifier** emerged as the final model, with **highest accuracy 0.8310** with strong generalization and reduced overfitting.

2. What was learned

There are the three keys learning through this week activities:

- We learned that trying multiple advanced models and comparing their performance is essential to selecting the most robust one. Although XGBoost and LightGBM performed well, experimenting further with ExtraTreesClassifier revealed even better accuracy.
- The impact of hyperparameter tuning was clearly evident. Fine-tuning parameters for each model especially in depth, number of estimators, and regularization led to noticeable improvements in performance. This emphasized that default settings may not be optimal, and tuning can significantly boost accuracy.

- By combining effective feature engineering, class balancing, and outlier handling, we observed a cumulative effect on model improvement. The final preprocessing pipeline, including Z-score for outlier removal and RandomOverSampler for balancing, provided a strong foundation for model success.
- In this model, it seems overfitting exists, but it shows the highest f1 score for the validation result. So, we decided this model as our final model.

3. How you improved the model

- Feature Selection: Adding and subtracting the variables impact on the model performance a lot. After multiple trials with different combinations of variables, we found an optimal subset of features.
- Balanced Sampling: Using RandomOverSampler balanced the target variable's class, and it helped to improve the model performance.
- Removing Outliers: Using the Z-score, we removed the extreme outliers by minimizing the loss of the dataset.
- Model Enhancement: Trained and Fine-tuned three models - LightGBM, XGBoost and ExtraTreesClassifier. Through hyperparameter Tuning and performance comparison, we observed a steady increase in accuracy and F1 Score.
- Final Model Selection: After evaluating All models, **ExtraTreesClassifier** delivered the best performing model with **accuracy: 0.8310 and F1 Score: 0.83**, making it the final model choice.

The week	Output
Benchmark Output	Train: precision recall f1-score support 0 0.59 0.63 0.61 1715 1 0.65 0.61 0.63 1962 accuracy 0.62 3677 macro avg 0.62 0.62 0.62 3677 weighted avg 0.62 0.62 0.62 3677 Test: precision recall f1-score support 0 0.57 0.60 0.58 735 1 0.63 0.61 0.62 841 accuracy 0.60 1576 macro avg 0.60 0.60 0.60 1576 weighted avg 0.60 0.60 0.60 1576
	Train:

Week 7	Accuracy: 0.8951 Classification Report: precision recall f1-score support 0 0.91 0.86 0.89 1741 1 0.88 0.92 0.90 1958 accuracy 0.90 3699 macro avg 0.90 0.89 0.89 3699 weighted avg 0.90 0.90 0.89 3699 Test: Accuracy: 0.7396 Classification Report: precision recall f1-score support 0 0.73 0.68 0.71 729 1 0.74 0.79 0.77 857 accuracy 0.74 1586 macro avg 0.74 0.74 0.74 1586 weighted avg 0.74 0.74 0.74 1586
Week 8	Accuracy: 0.7956 Classification Report: precision recall f1-score support 0 0.80 0.78 0.79 2669 1 0.79 0.81 0.80 2669 accuracy 0.80 5338 macro avg 0.80 0.80 0.80 5338 weighted avg 0.80 0.80 0.80 5338 Accuracy: 0.7582 Classification Report: precision recall f1-score support 0 0.70 0.72 0.71 771 1 0.80 0.79 0.79 1123 accuracy 0.76 1894 macro avg 0.75 0.75 0.75 1894 weighted avg 0.76 0.76 0.76 1894

Week 9	LGBM: Training Accuracy: 0.9295 Training Classification Report: precision recall f1-score support 0 0.93 0.93 0.93 2884 1 0.93 0.93 0.93 2888 accuracy 0.93 5772 macro avg 0.93 0.93 0.93 5772 weighted avg 0.93 0.93 0.93 5772 Test Accuracy: 0.8220 Test Classification Report: precision recall f1-score support 0 0.82 0.82 0.82 724 1 0.82 0.82 0.82 720 accuracy 0.82 1444 macro avg 0.82 0.82 0.82 1444 weighted avg 0.82 0.82 0.82 1444 XGBoost: Training Accuracy: 0.9492 Training Classification Report: precision recall f1-score support 0 0.95 0.95 0.95 2884 1 0.95 0.95 0.95 2888 accuracy 0.95 5772 macro avg 0.95 0.95 0.95 5772 weighted avg 0.95 0.95 0.95 5772 Test Accuracy: 0.8179 Test Classification Report: precision recall f1-score support 0 0.82 0.81 0.82 724 1 0.81 0.82 0.82 720 accuracy 0.82 1444 macro avg 0.82 0.82 0.82 1444 weighted avg 0.82 0.82 0.82 1444

Week 10
Final model

LGBM:

Training Accuracy: 0.9286

Training Classification Report:

	precision	recall	f1-score	support
0	0.93	0.92	0.93	2884
1	0.92	0.93	0.93	2888
accuracy			0.93	5772
macro avg	0.93	0.93	0.93	5772
weighted avg	0.93	0.93	0.93	5772

Test Accuracy: 0.8248

Test Classification Report:

	precision	recall	f1-score	support
0	0.83	0.82	0.82	724
1	0.82	0.83	0.82	720
accuracy			0.82	1444
macro avg	0.82	0.82	0.82	1444
weighted avg	0.82	0.82	0.82	1444

XGBoost:

Training Accuracy: 0.9494

Training Classification Report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	2884
1	0.95	0.95	0.95	2888
accuracy			0.95	5772
macro avg	0.95	0.95	0.95	5772
weighted avg	0.95	0.95	0.95	5772

Test Accuracy: 0.8296

Test Classification Report:

	precision	recall	f1-score	support
0	0.83	0.82	0.83	724
1	0.82	0.84	0.83	720
accuracy			0.83	1444
macro avg	0.83	0.83	0.83	1444
weighted avg	0.83	0.83	0.83	1444

ExtraTreesClassifier:

Training Accuracy: 0.9653

Training Classification Report:

	precision	recall	f1-score	support
0	0.97	0.96	0.97	2884
1	0.96	0.97	0.97	2888
accuracy			0.97	5772
macro avg	0.97	0.97	0.97	5772
weighted avg	0.97	0.97	0.97	5772

Test Accuracy: 0.8310

Test Classification Report:

	precision	recall	f1-score	support
0	0.83	0.84	0.83	724
1	0.84	0.82	0.83	720
accuracy			0.83	1444
macro avg	0.83	0.83	0.83	1444
weighted avg	0.83	0.83	0.83	1444

5. What each team member contributed

- Haeun Kim: Trained the model, analyzed and enhanced the model accuracy, and wrote the weekly report.
- Jayasree Lakshmi Narayanan: Trained the model, analyzed and enhanced the model accuracy, and wrote the weekly report.
- Samarth Verma: Evaluated final model performance and created an improvement report.
- Vikramaditya Sriramachandra: