Predicting Coral Reef Regimes from Human and Natural Influences

Sallie S. Walecka[1], Austin K. Murphy[2], Tiffany Cheng[2]
{swalecka, amurphy5, tiffc} @stanford.edu
[1]Institute for Computational and Mathematical Engineering, [2]Statistics Department

Stanford Computer Science

I. Overview and Motivation

Climate change and other human-driven factors have put pressure on coral reefs, leading to transitions between regime types, including to those considered "degraded". A previous paper by *Jouffray J-B et al.,* used boosted regression trees to find the most influential predictors for each of the 4 distinct reef regimes. In comparison, we focus on accurately predicting regime types using structured human, biotic, and abiotic data from Hawaii as it will allow researchers to identify areas that need immediate restoration and to better understand the distribution of regimes globally.

As our baseline, we built a regularized logistic regression model and conducted an error analysis to understand potential model improvements. After that, we enhanced our logistic regression model and built more flexible models, such as SVMs and decision trees. Ultimately, we were able to improve our accuracy (i.e., F1 score), but found that similarities between some regimes put a ceiling on the accuracy we were able to achieve.

II. Data

- We used a structured data set (publicly available on Github from [1]) collected from 620 reef habitats across the Hawaii islands. It has 20 predictors from two categories: anthropogenic (e.g., commercial fishing, effluent, etc.) and biophysical (e.g., wave power, depth, etc.).
- Each observation (i.e., reef habitat location) is labelled with one of four possible regime types. The labels were derived by *Jouffray J-B et al.,* from a Gaussian Mixture Model using features, such as fish, coral, algae, and other coral reef organisms.

III. Features

- The original data set contains 20 features anthropogenic and biophysical features that are potential drivers of corral reef changes.
- **Complexity** and **depth** of the seafloor had many NA values which we imputed using K-means.
- After reading papers exploring coral reef drivers, we added several interaction terms between **wave power/complexity/depth** as well as **algae/irradiance** and **wave anomaly/chlorophyll anomaly**.
- We did additional feature engineering by using a radial basis and polynomial kernel for our SVM to see if projection into a high dimensional space improved prediction accuracy.

IV. Evaluation Metric: Micro-average of F1 Scores

- We use micro-averaging of F1 score as our performance metric for our multiclass classification problem as an alternative to accuracy. Micro-averaging accounts for the class imbalance of coral regimes.
- The micro-average F-score is calculated by taking the harmonic mean of the micro-average of precision: $\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FP_i}$ and of recall: $\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + FN_i}$

where *TP*, *FP*, and *FN* stand for True Positives, False Positives, and False Negatives, respectively, and *C* is the number of classes.

V. Models

Logistic Regression

The multinomial logistic regression loss function with L2 regularization is given as follows:

$$L(\theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} 1\{y^{(i)} = k\} \log P(y^{(i)} = k | x^{(i)}; \theta) + \lambda ||\theta||^2$$

Where θ corresponds to the predictor weights, *k* is the class indicator, and $y^{(i)}$ is the estimate of the class.

Support Vector Machine (SVM)

The optimization problem for a non-separable SVM problem is as follows:

$$\min_{\gamma,\omega,b} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i, i = 1, \ldots, n$$
$$\xi_i \geq 0, i = 1, \ldots, n.$$

where ξ_i corresponds to the size of the margin, $y^{(i)}(\omega^T x^{(i)}+b)$ is the functional margin, and *C* controls the relative weighting between the optimization goals of minimizing $||\omega||^2$ and ensuring that most examples have a functional margin of at least 1.

Decision Trees

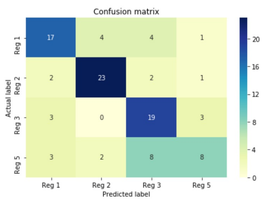
Decision trees look to find the optimal split across all variables that reduce the classification error in the model. The structural model for a classification decision tree is as follows:

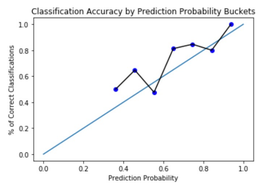
$$\hat{c}(x) = \sum_{m=1}^{M} \hat{c}_m 1\{x \in R_m\}$$

where $\hat{c}(x)$ is the predicted class, R_m are the mutually-exclusive regions in the predictor space, and $1\{x \in R_m\}$ is an indicator of whether the input feature *x* is in the respective region.

VI. Error Analyses

For our baseline model: Regularized Logistic Regression





VII. Results

Micro-average F1 Score results across the models

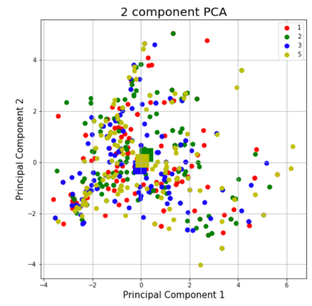
		Train	Validation
Logistic	drop NA	0.677	0.602
	mean NA	0.714	0.659
	imputed NA	0.687	0.637
	Regularized	0.695	0.621
	Reg + AF*	0.655	0.621
SVM	RBF	0.791	0.651
	POLY	0.714	0.613
Trees	RF + AF*	1	0.696
	GBT + AF*	0.826	0.692

* AF = additional features

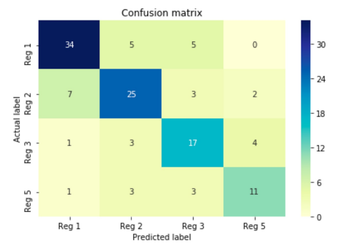
- Based on the results above, we chose a **Random Forest with additional features** from feature engineering as our final model. When we ran this model on our **test set** *(n = 124),* we got a micro-average F1 Score of 0.70.
- Train and Validation scores were taken from the means in 5-fold cross-validation with n=496.

VIII. Discussion

- We incrementally improved micro-averaged F1 score by a few percentage points with more flexible models, but were never able to achieve the quality of predictions that we expected based on the previous author's work.
- We decided to perform PCA on our data in order to determine how the class means relate to each other. As seen in the PCA plot projected onto the top 2 principal components, the classes are rather mixed and the class means are undistinguishable.
- In some of our models, we saw that observations from Regime 3 were often misclassified as being from Regime 5 and vice versa. In our best trained model's test-set confusion matrix, this was not observed.
- We were unable to find the original author's GMM output, which could provide insight into why our F1 score was not as great.



Best Model Results:
Random Forest Test Set Accuracy



IX. Future next steps

- Obtain output from GMM that classified the observations into their respective regimes to understand the confidence of the predictions.
- Conduct a more thorough error analyses of the flexible models that we built to understand their strengths.
- Enhance features by consulting a domain expert or finding additional data.

X. References

[1] D. et. al. Combining fish and benthic communities into multiple regimes reveals complex reef dynamics. *Scientific Reports,* (1):16943.
[2] G. J. et. al. Coral reef benthic regimes exhibit non-linear threshold responses to natural physical drivers. *Marine Ecology Progress Series*, 522:33–48, 2015.
[3] J. J.-B. et. al. Parsing human and biophysical drivers of coral reef regimes. *Proceedings of the Royal Society B: Biological Sciences*, 1896:286, 2019.
[4] W. L. M. et. al. Advancing the integration of spatial data to map human and natural drivers on coral reefs. *PLOS ONE*, 13(3):1–29, 03 2018.

XI. Acknowledgements

- We would like to thank Jean-Baptiste Jouffray (Ph.D Candidate at Stockholm University) for answering our clarification questions regarding his original paper, Anand Avati for his support as our project TA, and the CS 229 teaching staff for a informative quarter.