Chlorophyll and Fishery Measures

Mercy Amankwah Gabe Khan Noah Rahman
Joseph Schmidt David Tweedle

Shrimp Modeling

This project aimed to model shrimp catch data using environmental factors, specifically from the SEAMAP program, which provides comprehensive fishery-independent data. Initial focus on finned fish data was abandoned due to modeling challenges like zero-inflation and the unpredictability of large catches. The project then shifted to shrimp, particularly pink, brown, and white shrimp, where data was more complete and manageable.

Data Preprocessing

Key steps in preprocessing included:

- Merging data using `STATIONID` and eliminating duplicates.
- Filtering out erroneous data (e.g., temperatures above 100°C).
- Handling missing or irrelevant features.
- Filling blank shrimp catch data with zeros (since zero catches were left blank according to the Trawling Operations Manual)
- Performing exploratory analysis to identify predictive variables.

Modeling Approach

The baseline model returned the mean of the training data (MSE = 2.4884). Several models were tested, and we have returned the MSE for the best performing model after feature selection and parameter tuning:

- **Without Imputation:**
  - XGBoost: MSE = 1.6099
  - Histogram Gradient Boosting: MSE = 1.3901
- **With Imputation:**
  - RandomForestRegressor: MSE = 1.5493
  - HistGradientBoostingRegressor: MSE = 1.4342
- **Neural Networks**:
  - ReLU Network: MSE = 1.8331
  - Tanh Network: MSE = 1.8194

Conclusions

The **HistGradientBoostingRegressor** emerged as the best-performing model with an MSE of 1.3901 for the validation set. Fine-tuning did not lead to further improvements, and the model was finalized with an MSE of 1.3514 on the test set. SHAP analysis revealed that ****location, time, chlorophyll levels, temperature**, and **bottom oxygen** were the most predictive factors. Low oxygen was particularly detrimental to shrimp populations. Surface chlorophyll was more predictive due to missing data at deeper layers.

Lastly, we trained **HistGradientBoostingRegressor** models with a quantile loss to generate prediction intervals. For this, it was necessary to transform the data to address the fact that it was zero-inflated and had a long tail. After feature selection and parameter tuning, the intervals for the final model were slightly narrow, with the 90% interval containing 84% of the data.

Biodiversity modeling

Typically, a more healthy ecosystem has a high variety of species and is “well-balanced” (i.e. no invasive species takes over the ecosystem). Therefore, another approach with the data was to model the species diversity by using species IDs (unique tag for each species) and their extrapolated counts during the survey expeditions. The main challenges with this approach are:

1. Creating a good metric, or metrics, for a diversity ecosystem
2. Gathering the data in such a way that preserves useful information

Data Preprocessing

Key steps in preprocessing include:

- Grouping 700,000 individual species data rows into averaged 7,000 daily catches
- Filtering averaged data to remove spatial and temporal outliers (i.e. removing averaged data from two very far cruises that occurred on the same day)
    - Note this was only done for cruises whose start & end date/location were significantly different when compared to the majority
- Collecting number of unique species IDs and extrapolated catch to calculate different metrics of biodiversity measures like spread, evenness, or balance

Modeling Approach and Outcomes:

Shannon entropy was used to capture biodiversity with values [0,4] with higher values being indicative of more diverse ecosystems. All models tended to plateau to an RMSE of ~0.5 and expanding the models quickly lead to overfitting (using train vs validation plots)

- Linear regression (baseline): RMSE = 0.637
- XGBoost: RMSE = 0.515
- Neural Network: RMSE = 0.538
- **CNN: RMSE = 0.507**
- Multiregressor XGBoost (Shannon entropy only): RMSE = 0.525

Conclusions

The CNN model has the best performing model and has a structure of 2 layers with 64 neurons/layer using 2D conv layers separating spatial and temporal features from the other data. As with the other models, expanding the models and tuning hyperparameters to improve performance quickly lead to overfitting. This indicates the data may need to be expanded to more samples or augmented with richer data (i.e. satellite data) to better capture and learn spatiotemporal aspects of the features.