Whale Sighting Prediction Model

Beliz Pekkan (2066519)

Introduction

Climate change threatens marine life, especially cetaceans like whales, dolphins, and porpoises. Rising sea surface temperatures disrupt their distribution, habitats, and migration patterns (van Weelden et al., 2021). Understanding and predicting these impacts on cetacean migrations is crucial for conservation efforts and to further marine biology research.

This study aims to predict shifts in the migration patterns of blue whales (Balaenoptera musculus) caused by climate change. Advanced machine learning algorithms, such as linear regression, random forest, and XGBoost, were used to analyze historical data and sea surface temperature changes.

Data And Visualisation

The research utilizes data from the OBIS-SEAMAP database provided by the Marine Geospatial Ecology Lab at Duke University. A subset of the data from 1990 to the present was used, totaling 12,073 data points out of approximately 17,000. This selection focuses on recent trends for more accurate predictions.

Data exploration involved creating a global map of sightings (Figure 1) and a Kernel Density Estimation (KDE) plot (Figure 2) to analyze migration patterns.

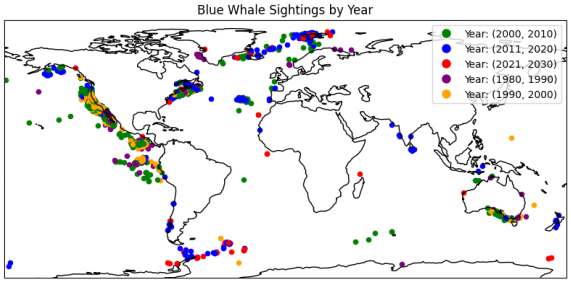

Figure 1: Blue Whale Sightings Map From 1903 Till 2023

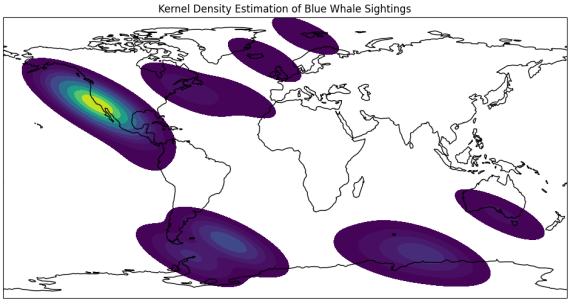

Figure 2: Blue Whale Sightings Kernel Density Estimation Map From 1903 Till 2023

These plots reveal important conclusions about persistent sighting areas and potential migration patterns.

Another KDE plot was created for a specific time range (Figure 3), showing different results and suggesting a slight pattern change.

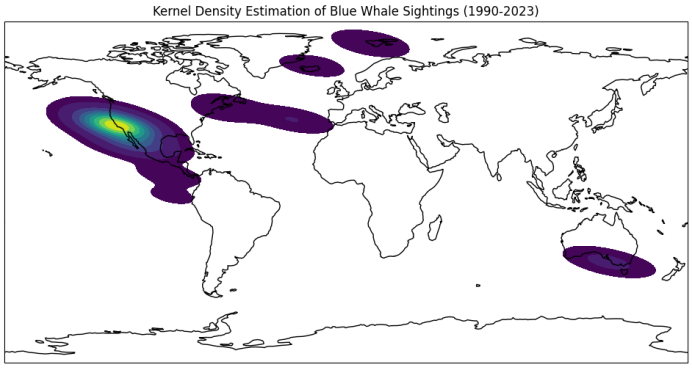

Figure 3: Blue Whale Sightings Kernel Density Estimation Map From 1990 Till 2023

Additionally, sea surface temperature data from NOAA was incorporated using an algorithm to align it with whale sighting records. This feature engineering step enhanced the ability to discover distinct patterns and see clearly the impacts of climate change.

Methodology

This study uses a step-by-step, multi-model approach to analyze blue whale migration data. The approach includes a linear regression model, a random forest model, and an XGBoost model with multi-output regressors. These models help improve the accuracy of predictions by gradually transitioning from more superficial to more complex models.

The selection of AI methodologies for this study focused on robust and adaptable machine learning models for ecological data analysis. Linear regression was chosen as a baseline for understanding basic trends. The random forest model handles complex ecological datasets and captures nonlinear relationships. XGBoost efficiently manages the multifaceted nature of this type of data. These methods provide a comprehensive approach to deciphering complex patterns in blue whale migration data and addressing climate change impacts on marine ecosystems.

Linear Regression Model

The linear regression model provides a baseline understanding of whale sighting data by modeling the relationship between time, sighting frequencies, and sea surface temperature.

Random Forest Model

The random forest model handles complex, non-linear ecological data relationships. It captures subtle patterns and interactions among variables, such as sea surface temperature. In the random forest model, the parameters n_estimators=100 and max_depth=None strike a balance between model accuracy and computational efficiency.

XGBoost Model

The XGBoost model efficiently handles large and complex datasets. It enhances the predictions of whale migration patterns and is robust against overfitting. In the XGBoost algorithm for multi-output

regression, the objective='reg:squarederror' minimizes the mean squared error between predicted and actual values.

Multi-Output Regressor

Multi-output regression models capture multiple dependent variables, providing a holistic view of migration patterns, including location, timing, and density.

Results

In this study, I used R^2 scores, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) as crucial metrics to evaluate regression model performances. R^2 measures how well the model replicates observed outcomes by explaining the proportion of total outcome variation. MSE and RMSE assess the average squared difference between estimated and actual values. MAE gives the average error magnitude regardless of direction. These metrics provide a comprehensive picture of accuracy and error magnitude, which is crucial for reliable ecological predictions.

	Average R^2 Scores	Test Mean Squared Error	Test Root Mean Squared Error	Test Mean Absolute Error
Linear Regression	-139.18	408566.63	639.19	361.15
Random Forest	-92.35	701792.79	837.73	353.28
XGBoost	-864.38	796346.13	597.45	402.04

Results from the linear regression model reveal significant limitations. Negative R^2 scores indicate that the model fails to capture the complexities of ecological data, performing worse than a simple mean line. High MSE, RMSE, and MAE further highlight deviations between model predictions and actual data.

Results from the Random Forest model suggest underperformance compared to linear regression—an average. R^2 score of -92.35 shows it is less accurate than a mean line, failing to capture underlying patterns. High Test MSE (701,792.79) and RMSE (837.73) reinforce this underperformance, indicating substantial deviations from actual data.

Results of the XGBoost model also show similar outcomes. Negative R^2 scores and high MSE (796346.13), RMSE (597.45), and MAE (402.04) indicate a significant deviation of model predictions from actual data.

Comparative Analysis of Models

Upon comparison, the XGBoost model demonstrated relatively better efficiency in capturing the complexities of whale migration patterns, as evidenced by its lower error metrics compared to the other models. The Random Forest model, typically robust in ecological data predictions, did not perform as expected, suggesting that the model's assumptions may not align well with the intricacies of this whale migration data.

This analysis highlights the complexities of ecological data modeling, emphasizing the need for continuous refinement of predictive models to enhance accuracy and reliability, especially in the context of climate change and its impact on marine life.

Discussion And Conclusions

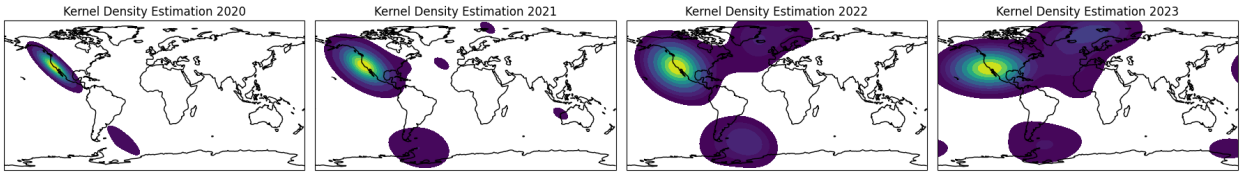

Figure 4: Kernel Density Estimations (2020-2023) for Comparison

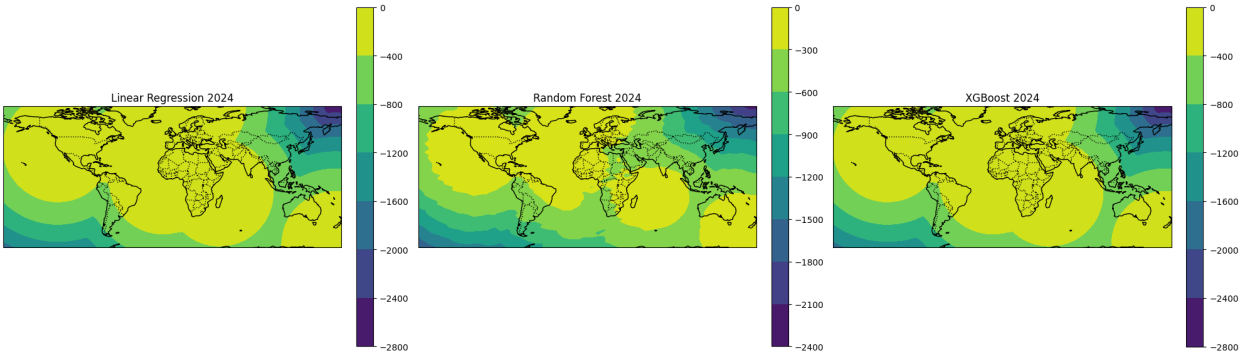

Figure 5: Model Estimations for 2024

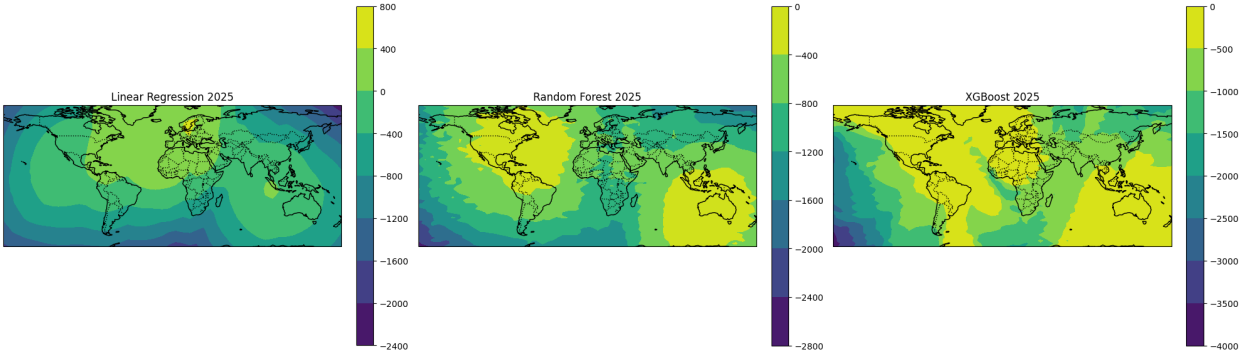

Figure 6: Model Estimations for 2025

In this project, the analysis of model outputs revealed vital insights into blue whale migration patterns. Linear Regression predicts higher whale densities in specific areas for 2024 and 2025, offering a foundational understanding of whale-environment interactions. However, its conservative prediction range suggests a shortfall in capturing the full complexity of these interactions. In contrast, the Random Forest model exhibits a broader spectrum of values and clear geographical patterns, highlighting its capability to grasp non-linear relationships and adapt to evolving ecological data. This adaptability is evident in the variability of predictions over the two years, though a negative R^2 score indicates potential inaccuracies in relation to observed data. Significantly, the

XGBoost model parallels Linear Regression in 2024 but diverges markedly in 2025, suggesting a notable shift in whale migration or densities, possibly due to its sensitivity to specific input features or its proficiency in detecting overlooked trends.

The application of this approach has both positive and negative consequences. It promises to enhance marine conservation by offering precise whale migration predictions, yet it must be coupled with broader ecological insights and conservation strategies to ensure a better understanding. The varied interpretations provided by the models demonstrate the inherent challenges in accurately predicting complex ecological phenomena, such as whale migration patterns.

Despite certain limitations, like the negative R^2 scores in some models, these findings can be instrumental in guiding conservation efforts. They enable scenario planning and underscore the necessity for robust, adaptable conservation strategies. Enhancing model performance through the incorporation of additional variables, increasing complexity, or applying different modeling techniques is essential. Continuous validation and collaboration with marine biologists remain vital for ensuring the accuracy and reliability of these models in marine conservation. Despite having setbacks, this idea might be improved upton to provide more solid predictions and be used as a foundational tool in the future to help marine biologists and conservationalists.

In conclusion, this project shows the complex and unpredictable nature of ecological data and the pressing need for advanced modeling techniques in ecological studies. While facing these challenges, the study significantly contributes to understanding marine mammal behavior amid climate change and the imperative of ongoing research and development in ecological modeling to support effective marine conservation strategies.

References

1. Van Weelden, M., Reijnders, P. J., & Van der Hiele, T. (2021). Climate change impact on cetaceans: A review. *Environmental Research Letters, 16*(9), 094005. https://doi.org/10.1088/1748-9326/ac1e62
2. OBIS-SEAMAP. (n.d.). *Ocean Biodiversity Information System Spatial Ecological Analysis of Mega Vertebrate Populations*. Retrieved from https://seamap.env.duke.edu/
3. NOAA National Centers for Environmental Information. (n.d.). *Sea Surface Temperature (SST)*. Retrieved from https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00516
4. Analytics Vidhya. (n.d.). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium*. Retrieved from https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e
5. Scikit-learn developers. (n.d.). *sklearn.linear_model.LinearRegression*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
6. Scikit-learn developers. (n.d.). *sklearn.ensemble.RandomForestRegressor*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
7. XGBoost developers. (n.d.). *XGBoost Documentation*. Retrieved from https://xgboost.readthedocs.io/en/stable/get_started.html
8. Scikit-learn developers. (n.d.). *sklearn.multioutput.MultiOutputRegressor*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputRegressor.html