Academic writing (01)

Guangyao Zhao

2022-08-30

Contents

Ref: Interpretable tree-based ensemble model for predicting beach water quality

1. Using SHAP, we perform global and local feature importance analyses with the best models to predict the most important factors for beach closure, to evaluate the robustness of those important factor, and to dissect their interactions with other environment variables.

2. Machine learning models are trained on data collected by humans, meaning that these models reflect human biases and prejudices. When a black-box model fails, we are neither able to know why it fails nor to trace the source of failure from the bias in training data. Besides predictability, another challenge for the utility of a machine learning model is its transparency, i.e., the ability of humans to understand how environmental features influence the prediction of FIB concentrations. Both predictability and transparency are key constituting elements of model interpretability.

3. It' critical to monitor beach quality and make advisory and closing notifications.

4. Its SHAP value increases nearly monotonically with its value and is minimally affected by other environmental factors.

5. We aim to overcome the under-appreciated issue of model interpretability by developing machine learning models that are both predictive and transparent.

6. We tailored the dataset by choosing the days when both types of data were measured. Therefore, gap filling/data imputation was not needed.

7. GridSearchCV was used to tune the parameters by exhaustively searching over pre-specified parameter values for each classifier.

8. Overall, no significant improvement was seen to justify the necessity of variable selection.

9. SHAP values can reveal local, case-dependent contributions of the predictors.

10. Detection and quantification of features interactions has been a central topic in machine learning.