Predicting the spread of invasive marine species with open data and machine learning: Process and Challenges

IMDIS 2021 Submission 46

Bumann, Teigland, Germishuys, Ziegler, Mattson, Olsson, Rylander, Lindh, Zhang, Linders

Background

- Invasive species a major challenge
- Increasing amounts of open data & technological advancements in AI/ML

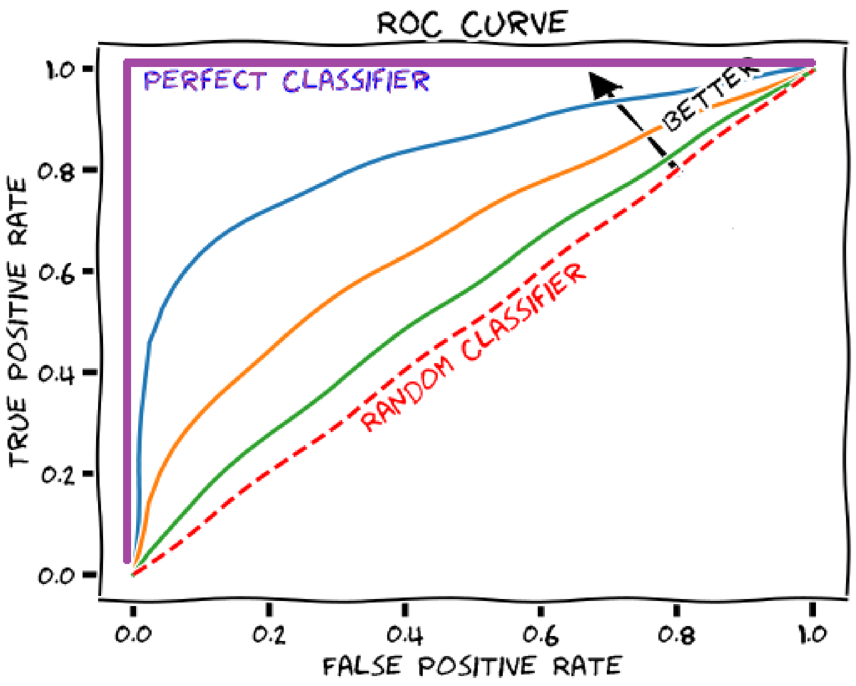
- -> overview of one use case applying ML
  - “Killer Shrimp” (Dikerogammarus Villosus) increasingly spreading in Baltic Sea
  - How does the Shrimp spread in the Baltic Sea under different conditions?

Data

- Open datasets
  - 1) port locations in Europe (EMODNET Human Activities)
  - 2) ocean surface temperatures and salinity
    - for Baltic Sea (SMHI - Swedish Meteorological and Hydrological Institute)
    - and North Sea regions (SeaDataNet)
  - 3) presence data of D. Villosus (GBIF, approx. 3000 data points)
    - Pseudo-absence data from Baltic Sea (approx. 2.8m data points)
  - 4) marine data layers (Bio-Oracle)
  - 5) ocean temperature and salinity (Marine Copernicus)

Process

- Cleaning the data
- Addressing presence-absence imbalance (3000 vs 2.8m)
  - Naïve classifier would have 99.9% accuracy
  - Thus aim for more False Positives (FP) than False Negatives (FN)
  - AUROC (Area under Receiver Operating Curve) to evaluate FN-FP balance



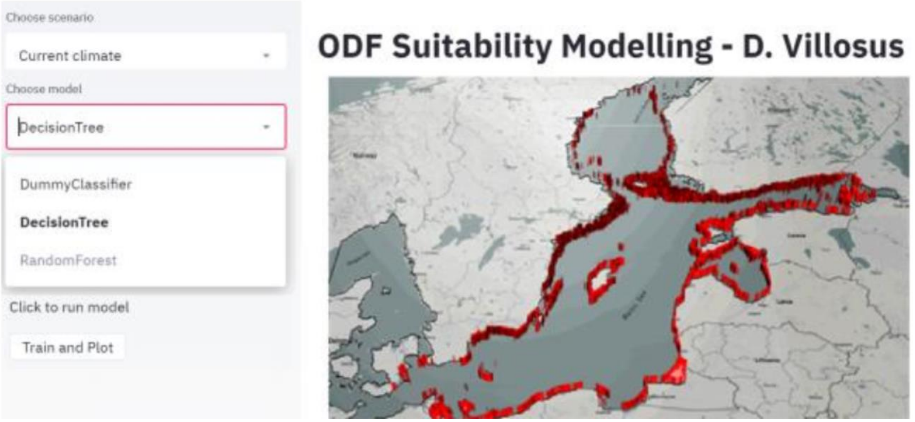
ML Models Used

- Tree-based models (single & ensemble)
  - Suitable given no feature selection or pre-processing
  - Easy to interpret
- Deep feed-forward Neural Network
  - Capture more complex features than by tree-based models alone
- Result
  - Neural Network outperforms
  - Tree-based: tend to predict majority class -> higher F1 score

Model	Accuracy	AUROC	F1	Recall
Majority Classifier	0.999	0.500	0.000	0.000
Decision Tree	0.999	0.917	0.833	0.833
Random Forest	0.999	0.917	0.810	0.833
Deep Neural Network	0.999	0.958	0.059	0.917

Evaluation & Visualization

- predictions for each cell in raster grid
  - areas of Åland & East Sweden:
- Using free platform Heroku for interactive visualization
- Published code on Kaggle & Github



Challenges

- Open data siloed on multiple platforms
- Easy to be overconfident in model predictions
- Differing Coordinate Reference Systems
  - Python packages, e.g. GDAL & Rasterio, help
- Complex context

Conclusion

- Interesting show case
  - ML in marine context has potential
  - Biggest challenge not ML but data
  - In this case: too many assumptions for practical use

- Derived recommendations
  - Users: ensure smooth collaboration of data & marine scientists
  - Data Providers: improve & align documentation standards