The Killer Shrimp Invasion Challenge on Kaggle: An online competition tackling the spread of invasive marine species through machine learning

Adrian Bumann, Chalmers University of Technology (Sweden), adrian.bumann@chalmers.se
Robin Teigland, Chalmers University of Technology (Sweden), robin.teigland@chalmers.se
Jannes Germishuys, Combine AB (Sweden), jurie.germishuys@combine.se
Benedikt Ziegler, Combine AB (Sweden), benedikt.ziegler@combine.se
Martin Mattson, Medins Havs och Vattenkonsulter AB (Sweden), martin.mattsson@medinsab.se
Eddie Olsson, RISE - Research Institutes of Sweden (Sweden), eddie.olsson@ri.se
Robert Rylander, RISE - Research Institutes of Sweden (Sweden), robert.rylander@ri.se
Yixin Zhang, University of Gothenburg (Sweden), yixin.zhang@ait.gu.se
Torsten Linders, University of Gothenburg (Sweden), torsten.linders@gu.se

Background.

The world faces numerous complex marine challenges, such as overfishing, the spread of invasive species, and rising sea levels. These challenges are interconnected with the 17 UN Sustainable Development Goals, and in particular Goal #14 - Life below water. A paradox for any action related to the ocean is that while there is an enormous lack of ocean data, there is also an abundance of online marine and geo data that could be used to develop solutions through the application of artificial intelligence (AI). Open innovation and crowdsourcing could be a solution to such complex problems; however, applying data science to solve marine challenges through these open strategies is limited.

Responding to the above, Ocean Data Factory Sweden (ODF Sweden), a data-driven innovation consortium in Gothenburg, developed an online competition, The Killer Shrimp Invasion Challenge (Figure 1). The competition was launched on the data science competition platform, *Kaggle,* in the spring of 2020 to address the spread of the invasive Killer Shrimp through applying machine learning (ML) to online data. This paper will describe the Killer Shrimp use case, the launch of the Kaggle competition, competition results and reflections on this form of tackling marine challenges.

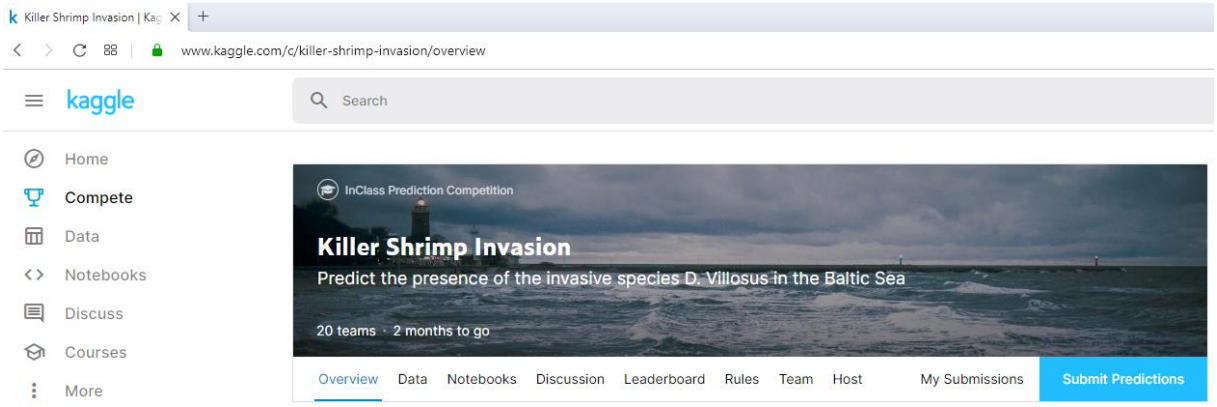

Figure 1: Landing page for the Killer Shrimp Invasion Challenge on Kaggle

Developing the Killer Shrimp Use Case.

In the fall of 2019, ODF Sweden started its first use case: predicting the spread of the invasive species, *Dikerogammarus Villosus* (aka the Killer Shrimp) in the Baltic Sea region. As this shrimp devastates any local ecosystem it invades, ODF Sweden and one of its partners, the Swedish Agency for Marine and Water Management (SwAM), decided to investigate whether ML methods could help predict the areas of the Baltic Sea suitable for the Killer Shrimp.

Throughout the use case, ODF Sweden applied a principle of openness. Primary data were national operational monitoring data from the Pan-European Copernicus Marine Services

(marine.copernicus.eu). Datasets included presence data from the North Sea and Baltic Sea (roughly 3000 data points), pseudo-absence data from the Baltic Sea (2.8 million data points), and environmental rasters for key environmental drivers informed by subject experts. The project was documented on Kaggle Jupyter notebooks and Github was used as a code repository.

Launching the Kaggle Challenge.

Once the team was satisfied with the first use case results, we explored how to launch the Kaggle challenge. Our motivation for a Kaggle challenge was threefold: 1) to create awareness of our efforts to monitor and predict marine invasive species, 2) to encourage others to improve our ML solution, and 3) to build an engaged global community from diverse backgrounds to contribute to our future challenges. We chose Kaggle as it is one of the largest, most diverse data science communities. Kaggle is specifically designed as a platform for data scientists from anywhere "to solve predictive modeling problems through data competitions". Founded in 2010, Kaggle was acquired by Google LLC in 2017. A featured Kaggle competition costs between $80,000 and $200,000 including both rewards and Kaggle's hosting fees. An example challenge is the Deepfake Detection Challenge by AWS, Facebook, Microsoft, Partnership on AI's Media Integrity Steering Committee, and academics with 33,007 entries by 2281 teams competing for $1 million in prizes. In the marine area, NOAA Fisheries is offering $12,000 in a three-year contest to develop algorithms to count sea lions in aerial photos.

Kaggle also supports free in-class competitions for educational purposes. Due to our interest in exploring the Kaggle platform for innovation, we chose to run an in-class prediction competition. We conducted the following steps to launch our challenge: 1) defined our machine learning problem as "to predict the presence of the Killer Shrimp in areas of the Baltic Sea", 2) prepared and uploaded a training dataset with locations, physical parameters and shrimp presence to enable competitors to develop their ML models, 2) prepared and uploaded an example of a submission file to enable competitors to understand what they should submit (a simplified set of predictions assuming all locations with temperatures above 5 degrees contain the Killer Shrimp), 4) chose the evaluation metric of "Area Under Receiver Operating Curve" (AUROC) between the predicted probability and the observed target, 5) uploaded a test dataset to evaluate the competitors' models on unseen data, i.e. predicting the presence of the Killer Shrimp in various areas, with 30% of the test set used to calculate the public leaderboard during the competition and the remaining 70% used to calculate the private leaderboard, i.e., the winners, at the end of the competition, 6) decided on a first prize of €150 and the opportunity for the winner to present their solution at an ODF Sweden grand meeting, and 7) determined the timeline to be just under three months. We then launched the challenge through our networks and on social media channels, such as Twitter, LinkedIn, Facebook, and Instagram.

Kaggle Competition Results and Reflections.

After one month of the three-month competition, our Challenge leaderboard showed that we had received 114 entries from 20 competitors in 20 teams from 12 different countries. Shortly after the launch, two competitors achieved a perfect score of 1.0000, with one achieving this after only one entry. As a perfect score is very unusual, this suggested that competitors found ways to overfit their models. We therefore added the requirement for competitors to upload their code as well. This decision was met with positive feedback on the discussion board. One top scorer explained which method he used and that he had submitted the perfect score to point out issues with the data setup. The competitor had also submitted legitimate models before and encouraged other users to do the same. This revealed some of the difficulties in hosting such a competition.

In summary, we are positively surprised and encouraged by the results to date despite the relatively small prize money. Our next steps will be to increase our understanding of Kaggle and other open innovation platforms, such as Zooniverse, as well as how we can better incorporate open innovation in the ML design process to tackle marine challenges.

References

Dick, J. T., Platvoet, D., & Kelly, D. W. (2011). Predatory impact of the freshwater invader Dikerogammarus villosus (Crustacea: Amphipoda). *Canadian Journal of Fisheries and Aquatic Sciences*, *59*(6), 1078-1084.

Majchrzak, A., & Malhotra, A. (2019). *Unleashing the Crowd: Collaborative Solutions to Wicked Business and Societal Problems*. Springer Nature.