The Killer Shrimp Invasion Challenge on Kaggle:

An online competition tackling the spread of invasive marine species through machine learning

IMDIS Submission 43

Background

- invasive species a major challenge on environmental, social and economic levels
- Crowdsourcing: utilize the public's workforce and creativity
- Machine learning: increasingly advanced capabilities and availability

- -> How to utilize crowdsourcing to engage the public in using machine learning (ML) in a marine context?

- This poster illustrates how Ocean Data Factory Sweden, a publicly funded innovation consortia, organized a challenge on Kaggle, a ML-focused crowdsourcing platform
- Goals:
  - To build awareness
  - To improve existing ML model (-> see submission 46)
  - To attract interested people for ODF Sweden

Figure 1 - the "Killer Shrimp"

Prologue

- ODF had already collected a large data set on the Killer Shrimp in the Baltic Sea
- Overarching question was
  - How does the Shrimp spread in the Baltic Sea under different conditions?

- 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)

Figure 2 - Data used

Preparing and launching the Kaggle Challenge

- Register challenge on Kaggle
  - "official" challenges are extremely expensive – good alternative are free "in-class" challenges
- define problem
  - "to predict the presence of the Killer Shrimp in areas of the Baltic Sea"
  - a catchy name "Killer Shrimp Invasion"
- upload
  - training dataset with locations, physical parameters and shrimp presence to enable competitors to develop their ML models
  - 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
  - example of a submission file to enable competitors to understand what they should submit
- choose
  - evaluation metric of "Area Under Receiver Operating Curve" (AUROC) between the predicted probability and the observed target
  - reward (150€)
  - timeline (~3 months)
- moderate

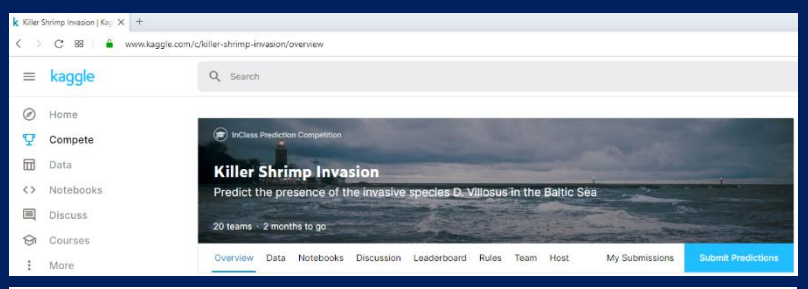
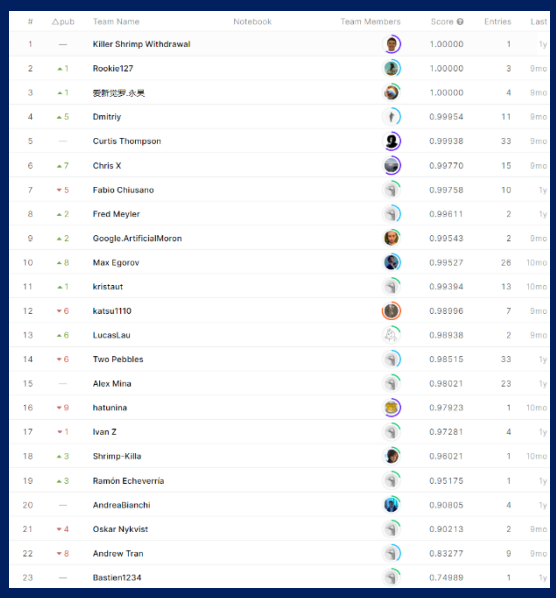


Figure 3 - Kaggle Challenge Header

Results

- 30 participants, mainly data scientists
  - Interest in "real-life" problems
  - A catchy name helps
  - Difficult with zero data science knowledge
  - needs good data
- Kaggle is a good platform…
  - For very streamlined, ML-focused competitions
  - Not to attract non-data scientists
  - But there are other interesting crowdsourcing alternatives (e.g. Zooniverse)
- Good moderation is key!
- Participants can be niftier than you think
  - Figured out how to "cheat" and received a perfect score with a poor ML model
  - This was done more out of curiosity than out of greed, and participants reacted positively to a subsequent rule change



#	Δpub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	Killer Shrimp Withdrawal			1.00000	1	1y
2	▲1	Rookie127			1.00000	3	9mo
3	▲1	爱新觉罗.永昊			1.00000	4	9mo
4	▲5	Dmitriy			0.99954	11	9mo
5	—	Curtis Thompson			0.99938	33	9mo
6	▲7	Chris X			0.99770	15	9mo
7	▼5	Fabio Chiusano			0.99758	10	1y
8	▲2	Fred Meyler			0.99611	2	1y
9	▲2	Google.ArtificialMoron			0.99543	2	9mo
10	▲8	Max Egorov			0.99527	26	10mo
11	▲1	kristaut			0.99394	13	10mo
12	▼6	katsu1110			0.98996	7	9mo
13	▲6	LucasLau			0.98938	2	9mo
14	▼6	Two Pebbles			0.98515	33	1y
15	—	Alex Mina			0.98021	23	1y
16	▼9	hatunina			0.97923	1	10mo
17	▼1	Ivan Z			0.97281	4	1y
18	▲3	Shrimp-Killa			0.96021	1	10mo
19	▲3	Ramón Echeverría			0.95175	1	1y
20	—	AndreaBianchi			0.90805	4	1y
21	▼4	Oskar Nykvist			0.90213	2	9mo
22	▼8	Andrew Tran			0.83277	9	9mo
23	—	Bastien1234			0.74989	1	1y

Figure 4- Overview of Participant Leaderboard

Reflections

- This was an interesting learning experience for ODF and participants
  - Yes, this type of competition works
  - Practical value in this case? Not so much, since it was designed more as proof of concept
- Advice for others
  - Think about whether and how much value you can get out of both ML and this type of open challenge
  - Prepare good and solid data sets
  - Think ahead how you want to spread the word about your challenge

Check out the challenge yourself!

Authors:
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