Bayesian Inference Model
to Rank Response Technologies in Arctic Oil Spills

Tanmoy Das (Date: Apr 04, 2024)

Outline

Problem Statement & Data

- Oil spill & Arctic
- Rank spill response technologies
- Data Collection and Preparation

Modeling

- Build multiclass multilabel ML model (i.e. Bayesian Inference)

Result & Discussion

- Model validation and model evaluation
- Metric: ROC, LRAP, compare with OLR
- Model deployment: Web app, Docker

Problem Statement

Decision-makers are trying to understand Arctic conditions and aim to improve spill response

Background

- **Arctic oil spills** can have devastating impact on delicate species & Indigenous community
- Several **factors** for spill response effectiveness: remoteness, harsh arctic conditions e.g. presence of ice, cold temperature

- This ML project is a sub-element of a Decision Support Tool (DST) for oil spill response facility setup and asset allocation
- Acknowledgement: Funded by MEOPAR & NSERC, Canada. CCG aims to improve spill response in Canadian Arctic

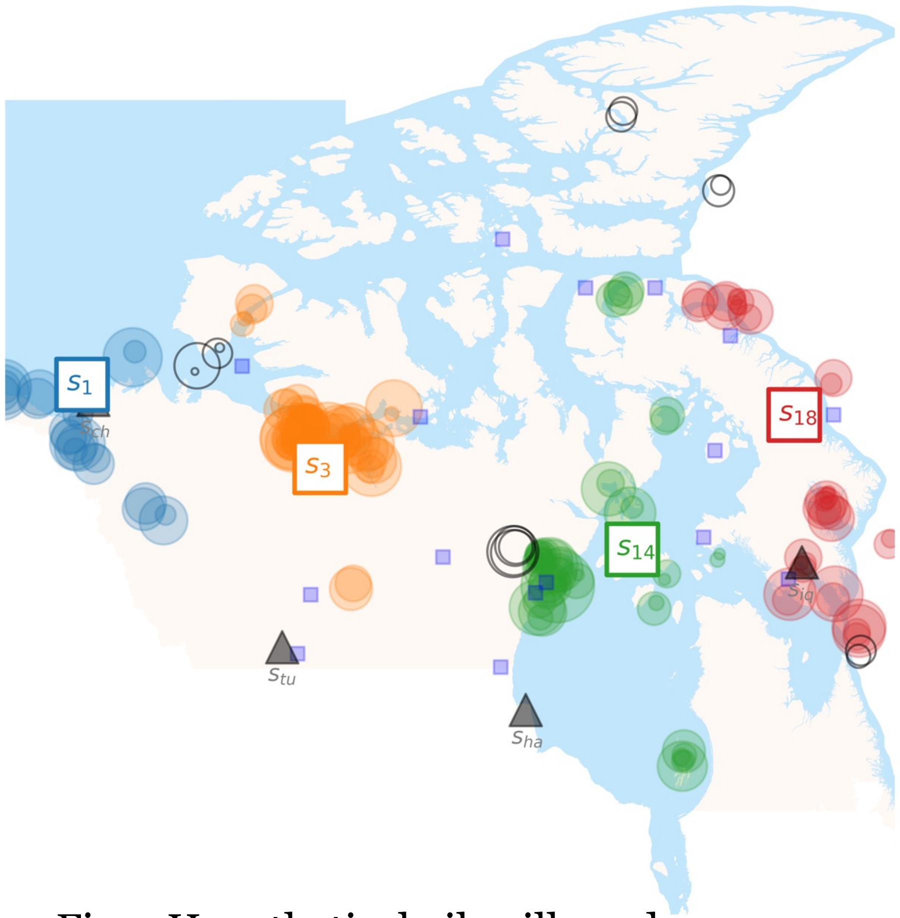


Fig 1. Hypothetical oil spills and response facilities in Canadian Arctic

Technologies & Challenges

Response Technologies

1. Mechanical Containment and Recovery (MCR): Removing oil from water surface by physical barriers and mechanical devices e.g. skimmers
2. Chemical Dispersant Use (CDU): Spraying chemical products into oil spill to disperse it into the water, and to accelerate dispersion
3. In-situ Burning (ISB): Controlled-burning of oil in its original place of spill

Limitation of existing systems

- Subjective judgment and labor-intensive models are available
- Limited data-driven models for Arctic
- No oil spill database of spill and environmental conditions vs response technologies in Arctic

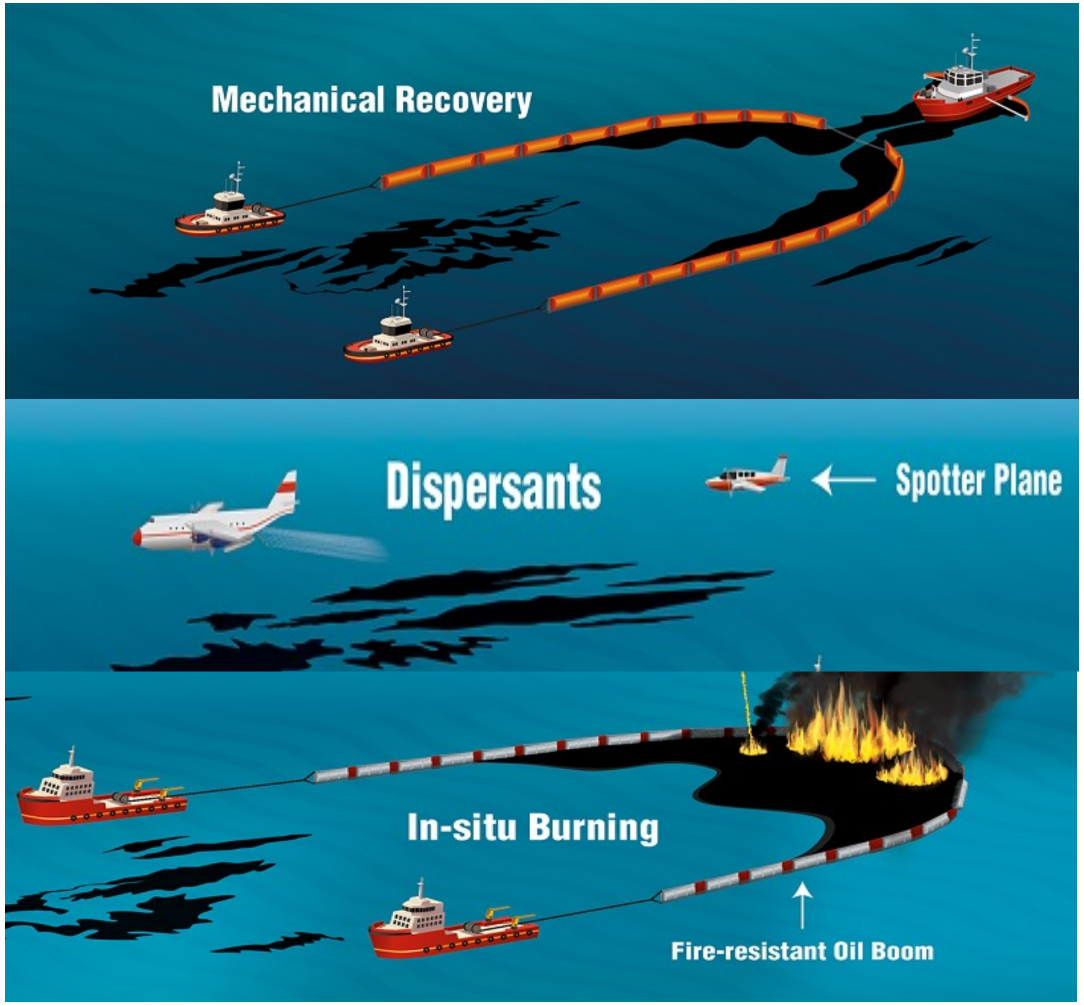


Fig 2. Oil Spill Response Technologies

Classification Task

Rank spill response technologies

Input

Key Factors for Cleanup Operations
Environmental: water temperature, ice presence, wind speed
Oil Properties: spill size, slick thickness
Other Arctic condition: Remoteness

Output

Response Systems	Classes
MCR	Ok, Consider, Go next season, Unknown
CDU	Ok, Consider, Not recommended, Unknown
ISB	Ok, Consider, Unknown

Data (Oil Spill Scenarios)

- **Challenges**
  - No relevant data for Canadian Arctic
  - EOS tool generate one record in each run (manual)
  - Class imbalance
- **Data Generation (**Monte Carlo Simulation)
  1. Distributions of variables are used from literature
  2. EOS Tool is applied in a Monte Carlo Simulation
  3. 3 target variables: MCR, CDU, ISB
     - Ranks in terms of *'OK' , 'Consider', 'Go next season', 'Unknown' and 'Not recommended'*

Multilabel (or multioutput

multiclass

Implementation: Excel VBA, Python to extract data from EOS Tool

- Shape of dataset
  - 30 features, 3 targets
  - 3100 oil spill scenarios

Sample Dataset

Scene no.	Oil spill size	Dispersion	Persistence	MCR ($E_{ss}, E_{sl}, E_{sw}, E_{sb}$)	Sufficient mixing energy	Sea surface	seawater	CDU ($E_{ss}, E_{sl}, E_{sw}, E_{sb}$)	Soot pollution	Residue recovery	displacement	ISB ($E_{ss}, E_{sl}, E_{sw}, E_{sb}$)
Scene 1	Medium	39	0	0,0,0,1	yes	Large	Large	1, 0, -1,-1	YES	yes	no	1,0,0,-1
Scene 2	Medium	53	1	0,-1,0,0	no	Large	Large	0,0,0,1	YES	no	yes	0,0,0,0
Scene 3	Large	78	1	1,0,0,-1	yes	Large	Large	0,0,0,1	YES	yes	yes	-1,0,0,0
Scene 4	Large	91	0	0,0,0,-1	yes	Large	Large	0,0,0,-1	YES	yes	no	0,0,0,-1
Scene 5	Small	45	1	1,0,0,0	no	Large	Large	0,0,0,0	NO	no	no	0,0,0,0

Water temp	wind	air temp	ice	Oil density	Oil viscosity	Retaining capacity	Shoreline length	Distance to community	MCR Options	CDU Options	ISB Options
-23	93	-39	53	36	8	0	4.315217	486	OK	Not recommended	Consider
-13	76	-32	9	45	15	1	3.095539	181	Unknown	Not recommended	OK
-6	63	-25	23	48	9	1	3.114366	26	OK	OK	OK
-4	17	-9	78	18	5	0	2.286095	313	OK	Not recommended	Consider
-16	59	-14	97	36	2	1	2.95920	311	Unknown	Not recommended	OK

Exploratory Data Analysis

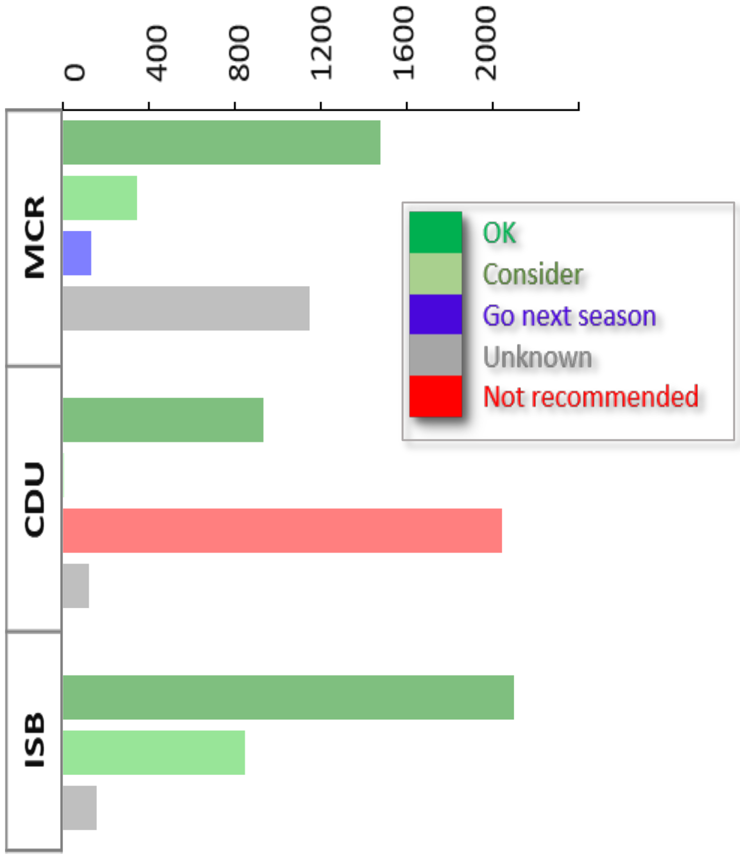


Fig 3. Distribution of classes of the target variables of the dataset

Data cleaning & Feature engineering

Feature Engineering

- Normalization for numeric features
- One Hot Encoding for categorical feature (dropping first entry to avoid curse of dimensionality)

Table 4: Encoding and Ranking of Target Variables

Subclasses of targets	Numerical Encoding	Rank
OK	10	1
Consider	8	2
Go next season	2	3
Unknown	-2	4
Not recommended	-10	5


Model

BIMReTA: Bayesian Inference Model to Rank Response Technology in Arctic

Naïve bayes

1. Calculate the prior probability for given class labels
2. Find Likelihood probability with each attribute for each class
3. Put these value in Bayes Formula and calculate posterior probability.
4. See which class has a higher probability, given the input belongs to the higher probability class

$$\mathrm{P(target \mid X)} = \frac{\mathrm{P(X \mid target).\,P(target)}}{\mathrm{P(X)}}$$

$$\mathcal{P}(f) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}f^T\Sigma^{-1}f)$$

$$\mathcal{P}_{ideal}(v_k \succ u_k | f(v_k), f(u_k)) = \begin{cases} 1 & \text{if } f(v_k) \succ f(u_k) \\ 0 & \text{otherwise} \end{cases}$$

- Bayesian (Naïve Bayes) works fine for multiclassification & smaller dataset (we have 3100 records)
- Expert opinion can be integrated in Bayesian model

Bayesian Inference Model

Pseudocode of the model

Input: Training data and Test data

1. Build a frequency table using training data ∀ feature Vector (x1, x2…, xn) against every class Ci, where i=1 to k
2. Create the likelihood table ∀ feature against each class Ci
3. Compute the conditional probabilities for Test Data ∀ Class:

$$P(Ci|x1, x2, \ldots, xn) = \prod_{j=1}^{n} P(xj|Ci).\frac{P(Ci)}{P(x1,x2,\ldots,xn)}$$

$$\text{for } 1 \leq i \leq k$$

4. Calculate $\max_i P(C_i|x1, x2\ldots, xn)$

Output: Class with maximum probability value

- **Assumption**
  - Feature variables are independent to each other
- No hyperparameter to tune in Naïve Bayes

How did I optimize the process?

- Full Model: Considering all features
- Reduced Model: A small model using only useful features

Result

Describing the outcome of my actions and the impact they had

Model validation and model evaluation

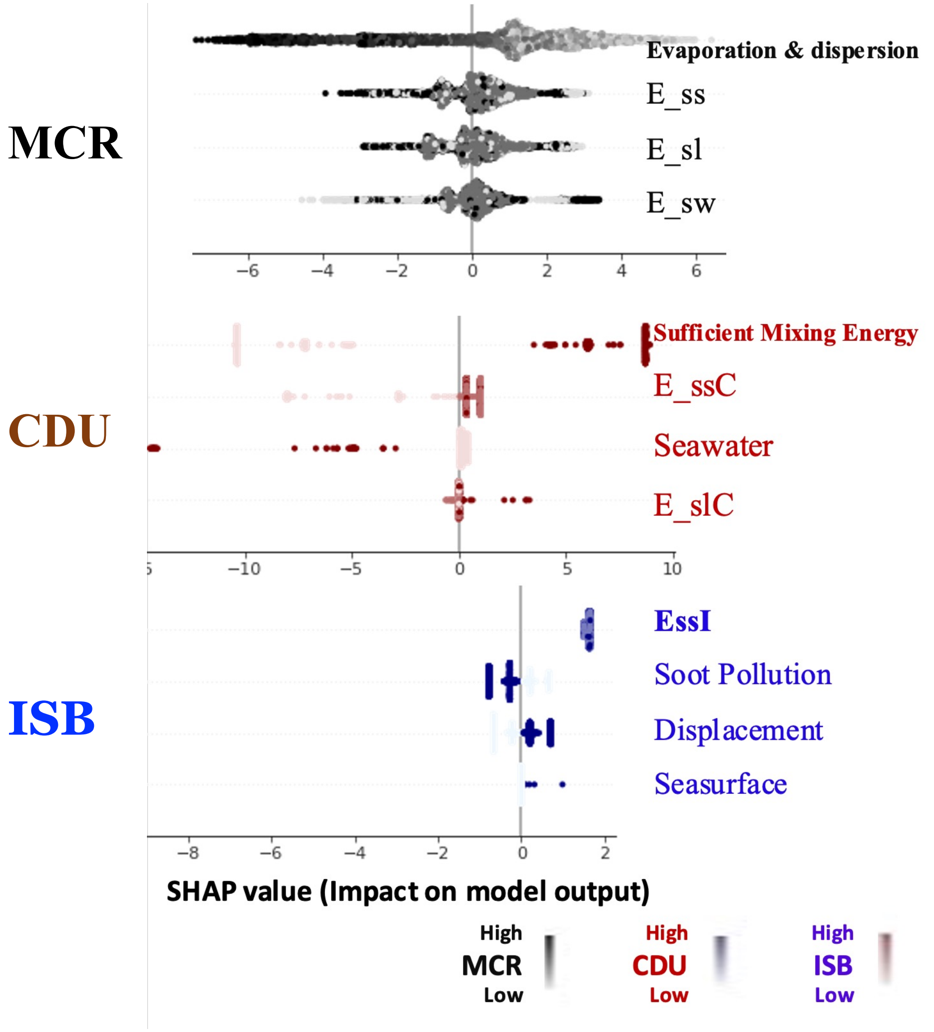


Fig 6: Summary plot of crucial features of **full model**

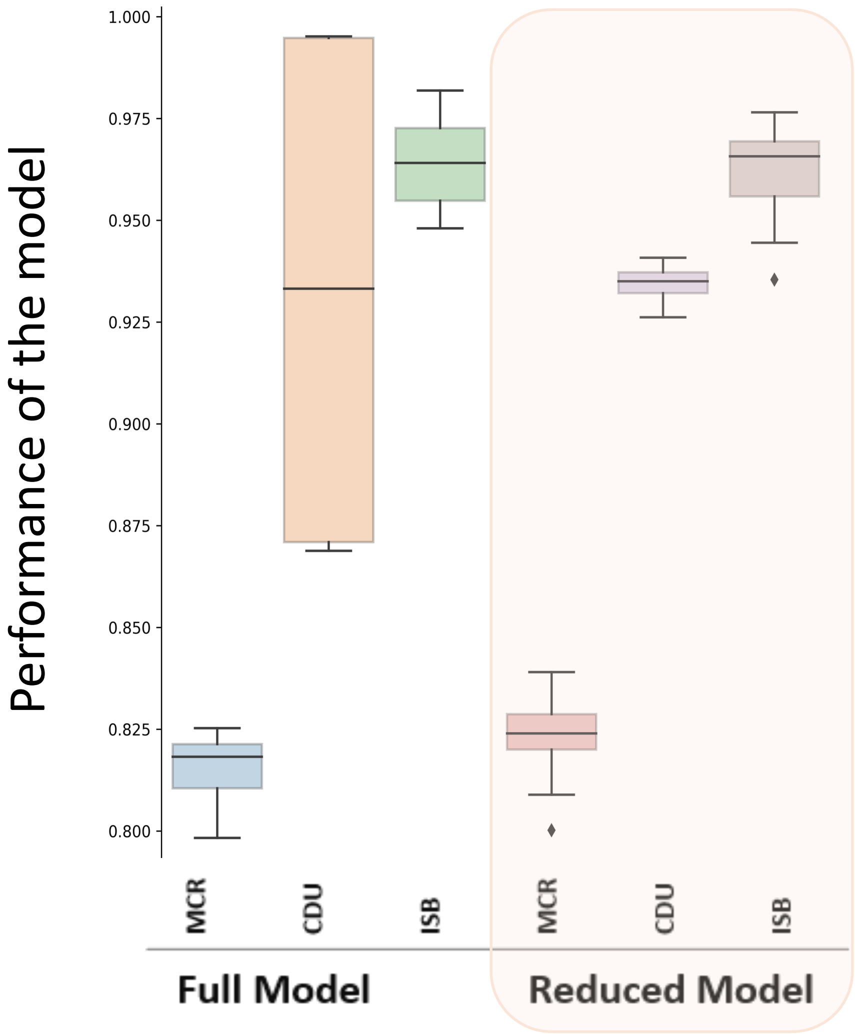


Fig 8: Comparing ROC-AUC of full and reduced models for k-fold cross validation set

Final Model Structure

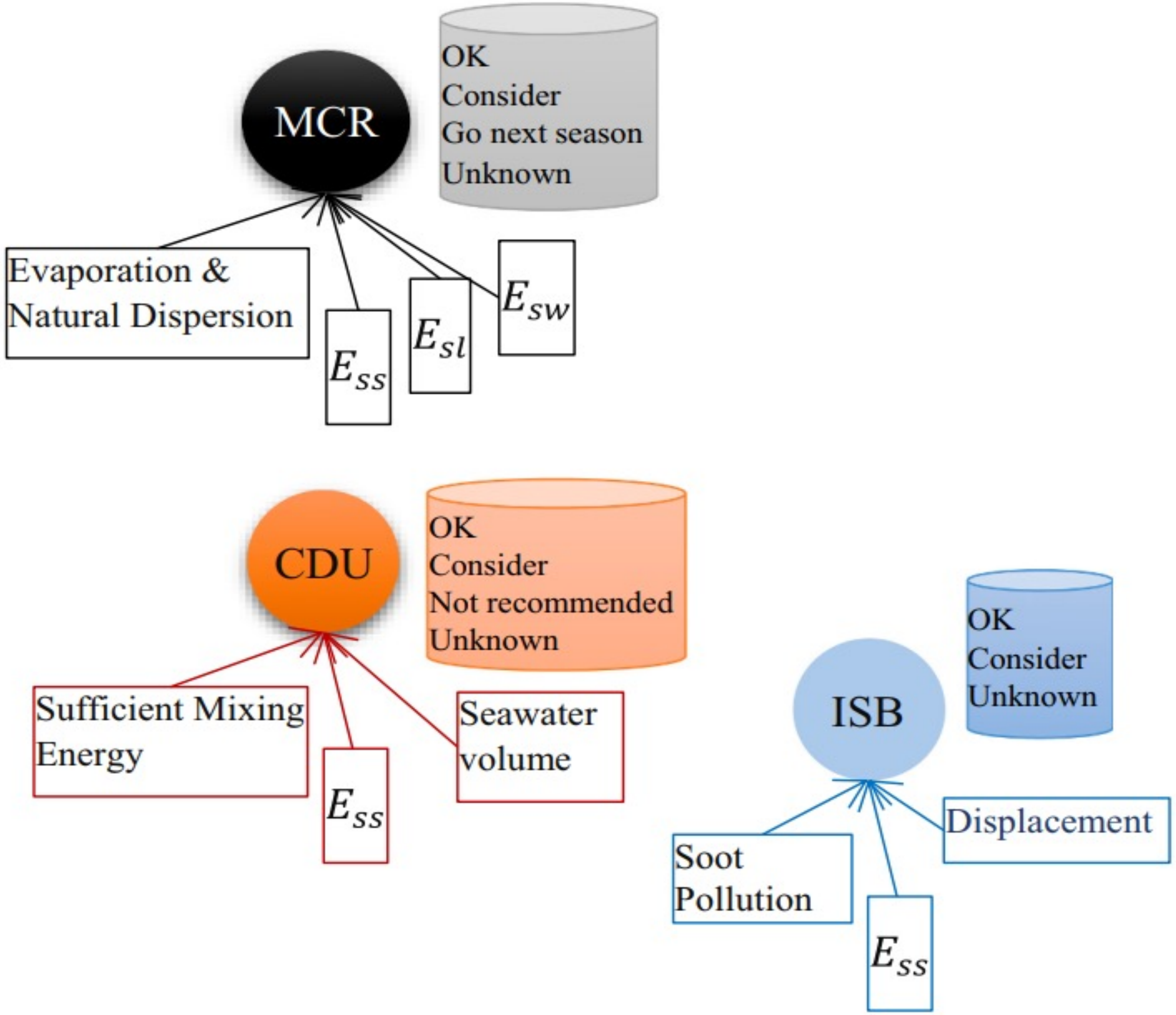
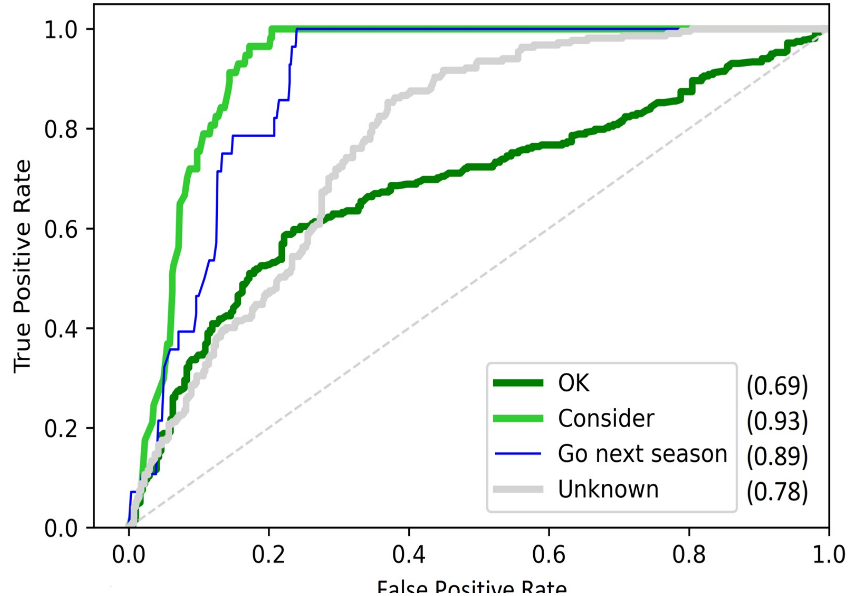
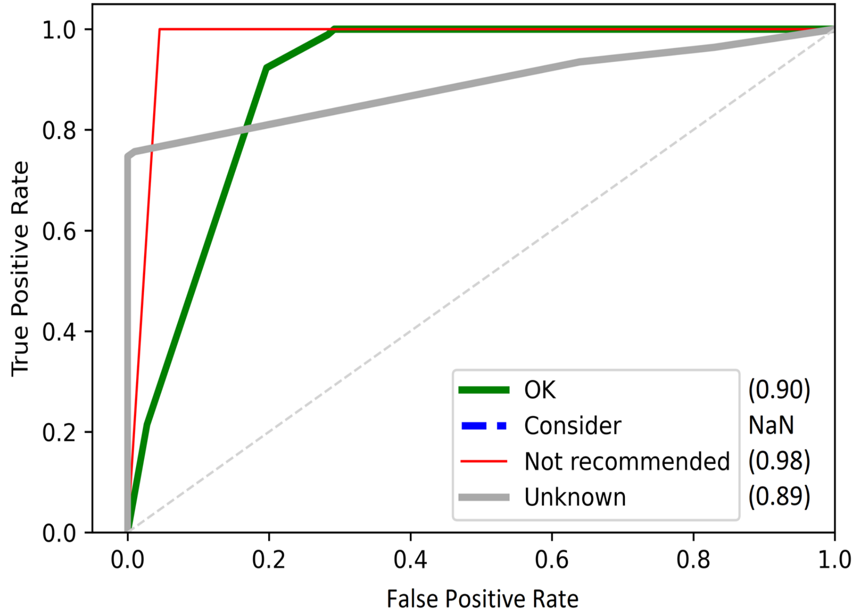


Fig 7. Structure of the BIMReTA model

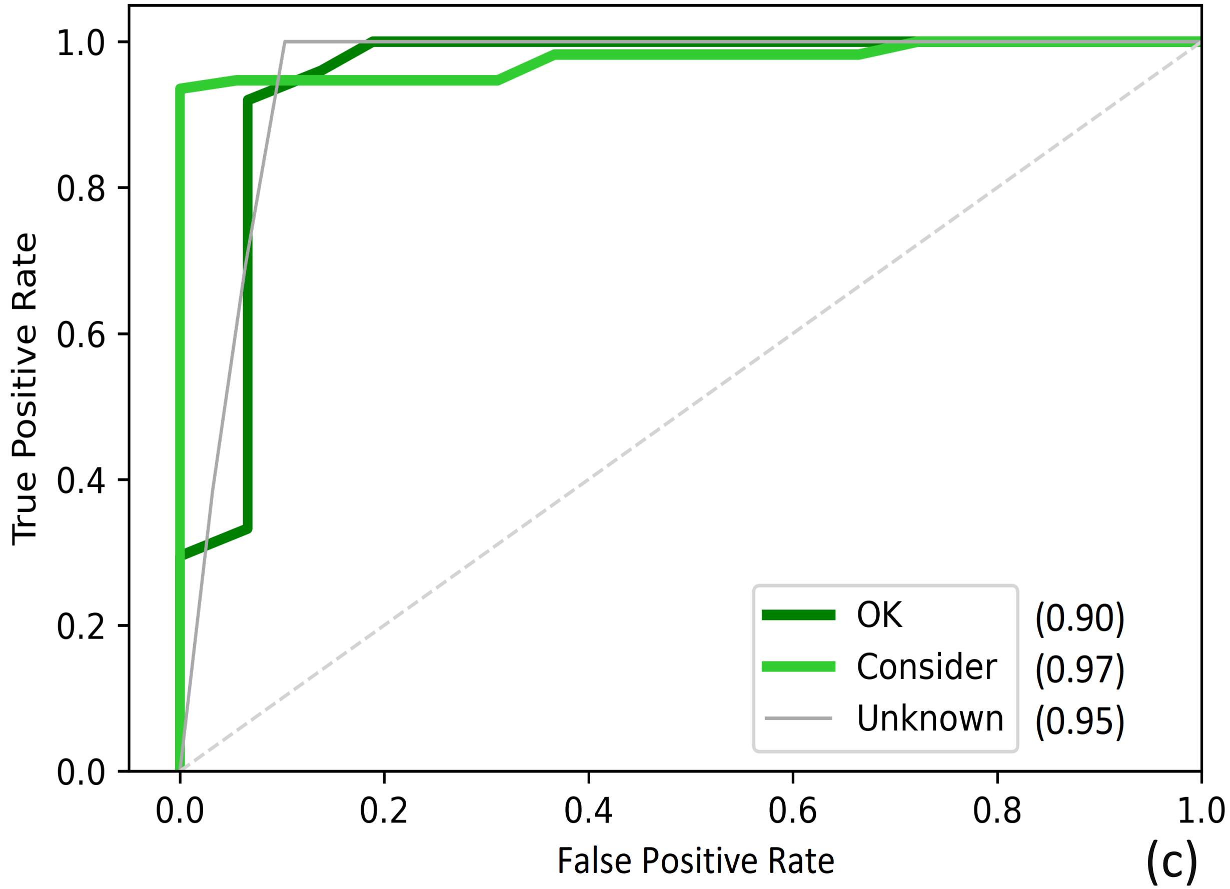
Model Performance (ROC curve)



(a)



(b)



(c)

Fig 9. ROC curves of MCR, CDU and ISB are shown in (a), (b), (c)

Significance of the model

Use-cases

- CCG & other DM can utilize our model for strategic response
  - Deployed using Streamlit & Docker image
  - Source code: http://github.com/tanmoyie/Bayesian-Inference-Model
- Published in prestigious *Marine Pollution Bulletin* journal (link)
- Used as a critical sub-model for a DST for Canadian Arctic spill response

Benefit of such model

- BIMReTA requires significantly fewer input parameters to determine a suitable oil spill response technique in Arctic
- Faster than existing engineering model
  - Speed makes the model useful for integration in a strategic risk assessment
- Implementation advantage:
  - don't require much training data
  - fast & can make real-time prediction
  - simple, easy to implement, scalable

Recent updates

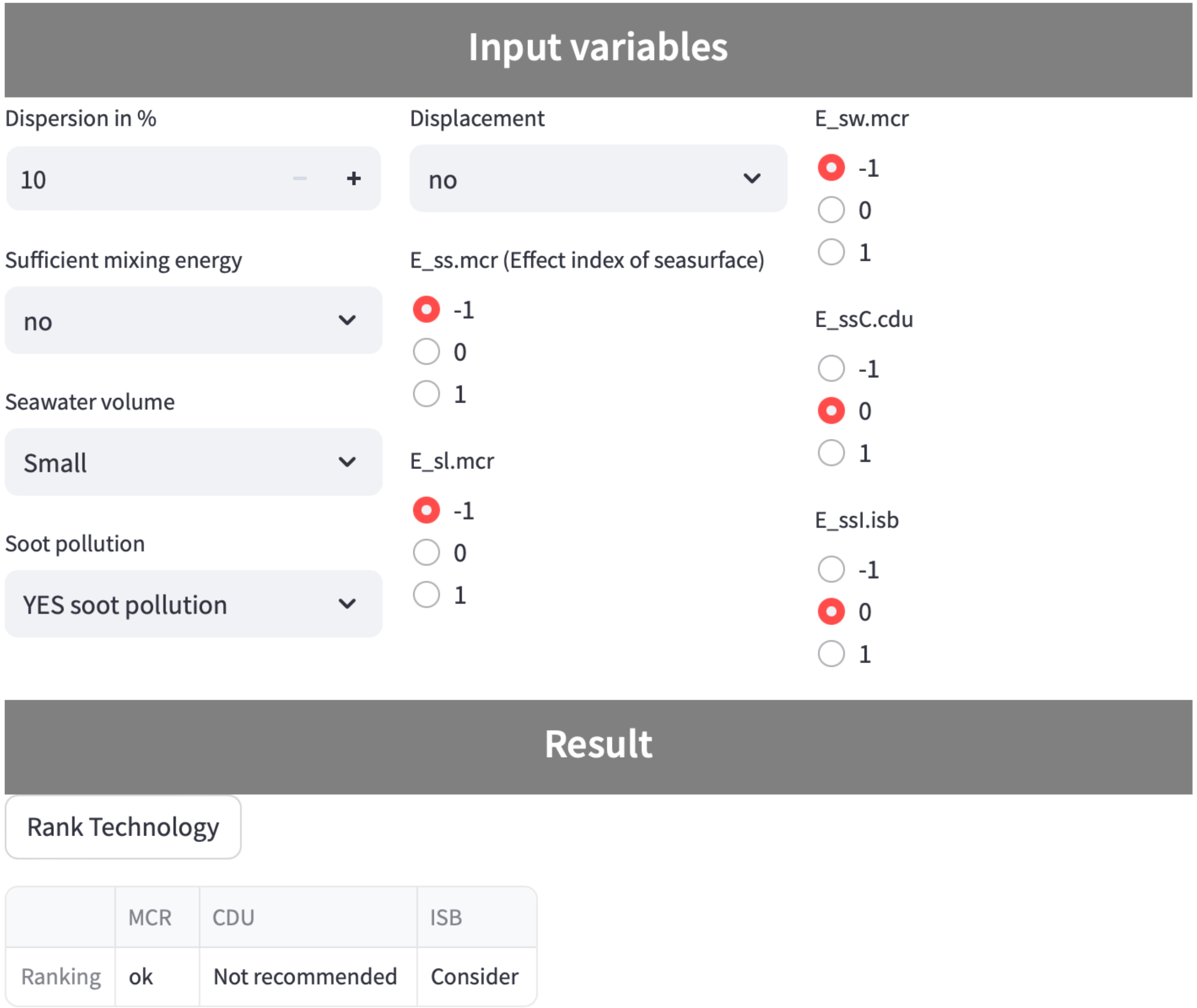


Fig11. The UI from Streamlit web app

- A simple Neural Network (NN) model is developed using Pytorch and compared with BIMReTA
- Although NN has better performance, stakeholders still prefer Bayesian since (i) easier to interpret, (ii) domain expert knowledge can be added (iii) no parameter to learn

	ROC-AUC [MCR, CDU, ISB]	**LRAP**
BIMReTA	[0.79, 0.93, 0.93]	[0.73, 0.9, 0.92]
NN	[0.91, 0.93, 0.94]	[0.91, 0.9, 0.92]

Project Communication & Team

Communication

- Assisting stakeholders by selecting, comparing, **and interpreting scenarios** of optimization runs
- Communicating the outcomes of ML model to broad audience
  - **Ocean Frontier Institute**
  - **IISE annual** conference, US.

Acknowledging my Team

- Formulating problem & model validation
  - My PhD supervisor
  - Interviews with domain experts
    - ECRC, CCG personnel & veterans
  - My teammates

My Contribution

- Building ML model (& analyse data)
- Writing technical report and validating model performance

Recap: Bayesian Inference Modelling

Problem Statement

- Our goal is to generate Arctic oil spill scenarios and determine appropriate oil spill response technology for Arctic harsh & remote conditions

Data (Spill Scenarios)

- Monte Carlo Simulation
- 3100 records, 30 features, 3 targets
- Preprocessing: MinMaxScaler, onehotencoding

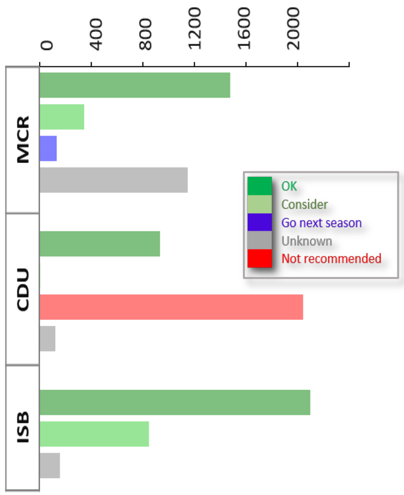


Fig3: Distribution of classes of target variables

Model

Bayesian Inference Modeling

1. Calculate prior probability for given class labels
2. Find Likelihood probability with each attribute for each class
3. Put these value in Bayes Formula and calculate posterior probability
4. Assign classes to higher probability criterion

$$\mathrm{P(target \mid X)} = \frac{\mathrm{P(X \mid target).\,P(target)}}{\mathrm{P(X)}}$$

$$\mathcal{P}(f) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}f^T\Sigma^{-1}f\right)$$

$$\mathcal{P}_{ideal}(v_k \succ u_k | f(v_k), f(u_k)) = \begin{cases} 1 & \text{if } f(v_k) \succ f(u_k) \\ 0 & \text{otherwise} \end{cases}$$

Result

Significance

- The model is intended to be used in strategic pollution preparedness and risk assessments
- The model requires fewer input parameters to determine a suitable spill response tech in Arctic
- Faster than existing engineering model

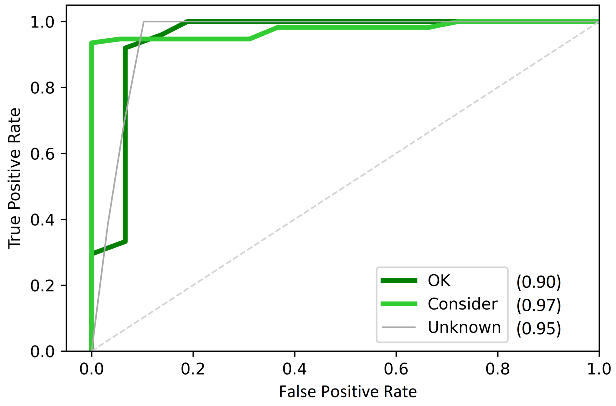


Fig9: ROC curve of multiclass of ISB

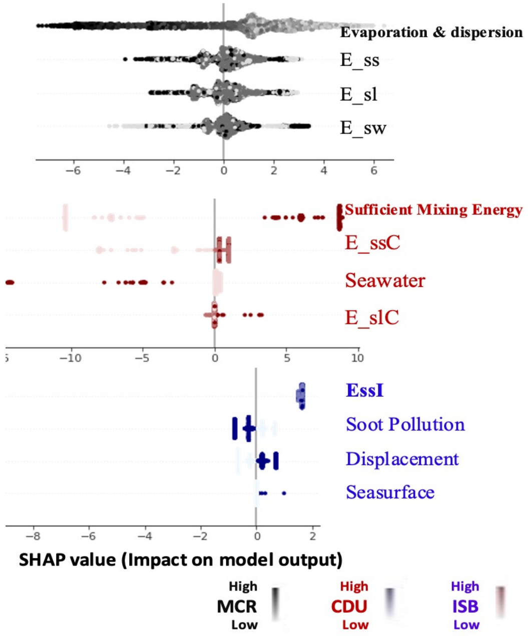


Fig10: ROC curve of multiclass of ISB

Questions?