Machine Learning Project Presentation

Ice classification in the Greenland

Luc YAO – Richard CHEAM

université PARIS-SACLAY

Table of contents

Study context

- In recent years, the highest number of infrasonic detections were found in northwestern Greenland and it was concluded that infrasound measurements can help in gaining more insights into glacier dynamics.
- So, higher detection rates can indicate increased glacier movement or calving events.

-> **The aim of this study is to evaluate the ability of various predictive classifier to evaluate the (low or high) quantity of ice.**

Infrasound is inaudible sound (lower than 20hz) and created when large volumes of air are displaced in the atmosphere.

The data

Information about data:

- Provided by European Centre for Medium-Range Weather Forecasts (ECMWF)
- 11 features in total -> Continuous variables except variable time
- 4 target variables -> Work only on one target variable Y1
- Dataset is clean

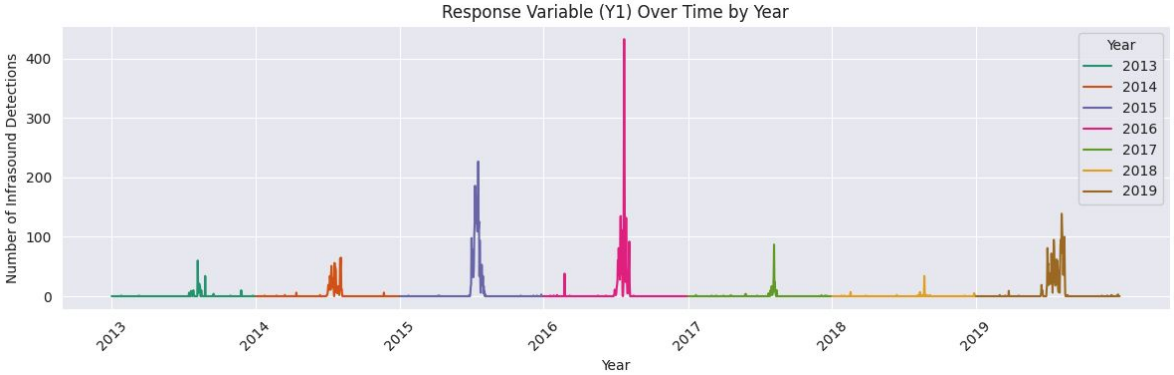


Figure 1: Values of Y1 over time

- Methodical jump during mid-year
- Highest between 2016-2017

- Mostly 0
- 5% of 2256 between interval [1,10)
- Equally shared for the others

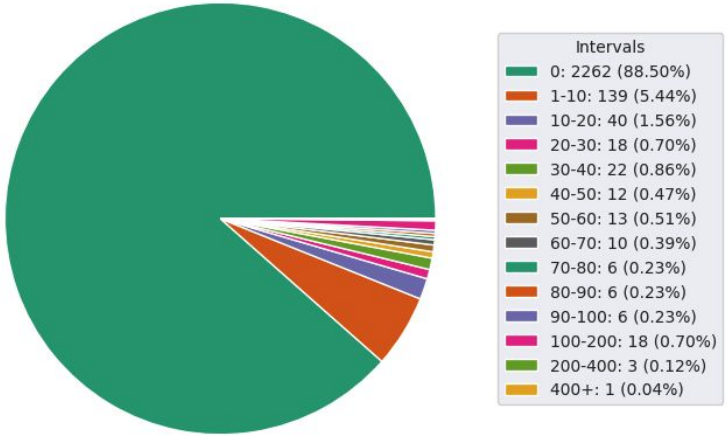


Figure 2: Occurrence of Y1 in % by intervals

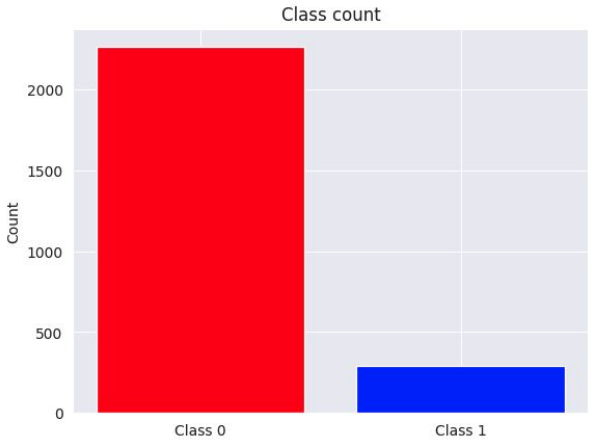
	t2m	u10	v10	SST	SIC	r1_MAR	r2_MAR	r3_MAR	r4_MAR	r5_MAR
count	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000	2556.000000
mean	-10.190040	0.139648	0.626351	-0.858922	73.267808	18.789275	11.522362	1.328740	4.382248	5.191286
std	10.340583	5.013640	3.955417	1.446165	29.250724	47.703600	27.942124	3.393452	12.971518	13.406972
min	-32.019122	-13.846656	-12.316128	-1.692462	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	-19.877500	-3.610168	-2.079422	-1.689860	70.000000	0.123798	0.121437	0.000000	0.000000	0.000000
50%	-9.596479	-0.187084	0.912780	-1.689860	84.601769	0.481870	0.481870	0.000000	0.000000	0.000000
75%	0.167824	3.807440	3.483397	-0.297573	90.362319	4.080174	3.961181	0.004413	0.010148	0.000150
max	7.840619	14.640913	12.811255	6.054536	99.500682	479.722174	281.673389	23.241791	115.876574	88.054318

Table 1: Summary statistics of X

	Y1
count	2556.000000
mean	3.525430
std	18.977537
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	433.000000

encoding

(0 if not detected, 1 otherwise)



- Class 0 = 88.5%
- 294 of class 1

Exploratory data analysis (EDA)

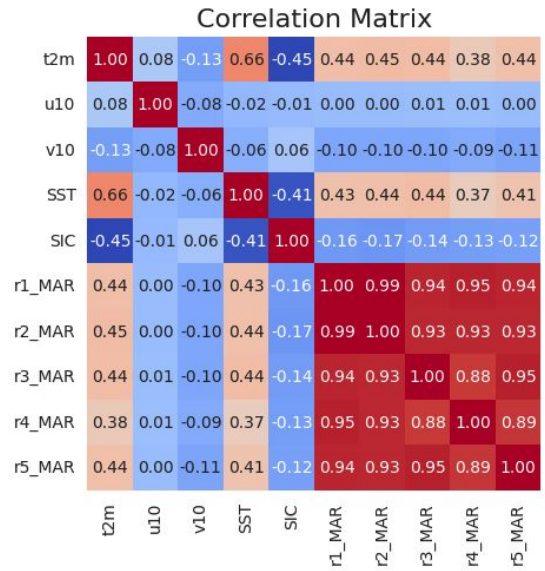


Figure 3: Correlation matrix of features

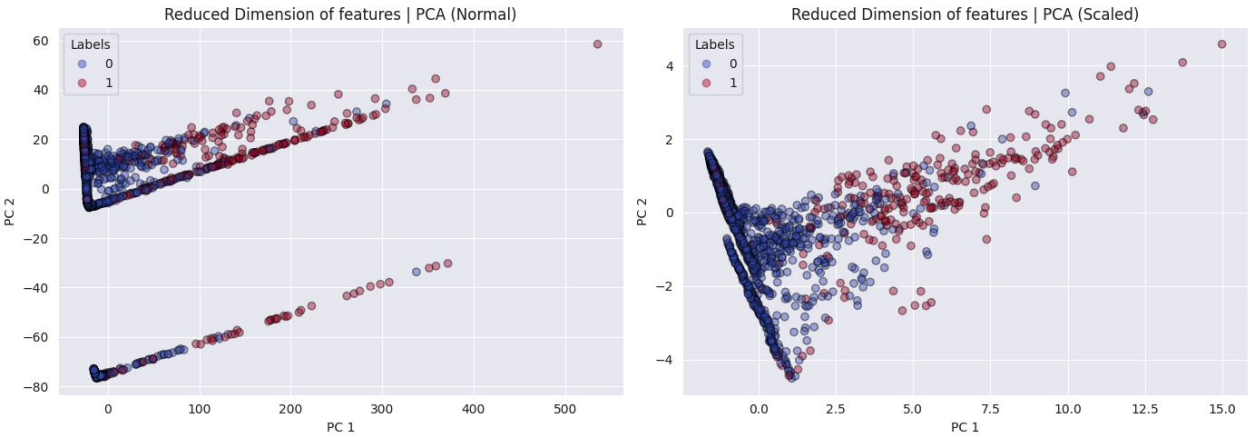


Figure 4: 2D space plot of initial dataset

- Greenland liquid water discharge simulated by Region Climate Models are highly correlated

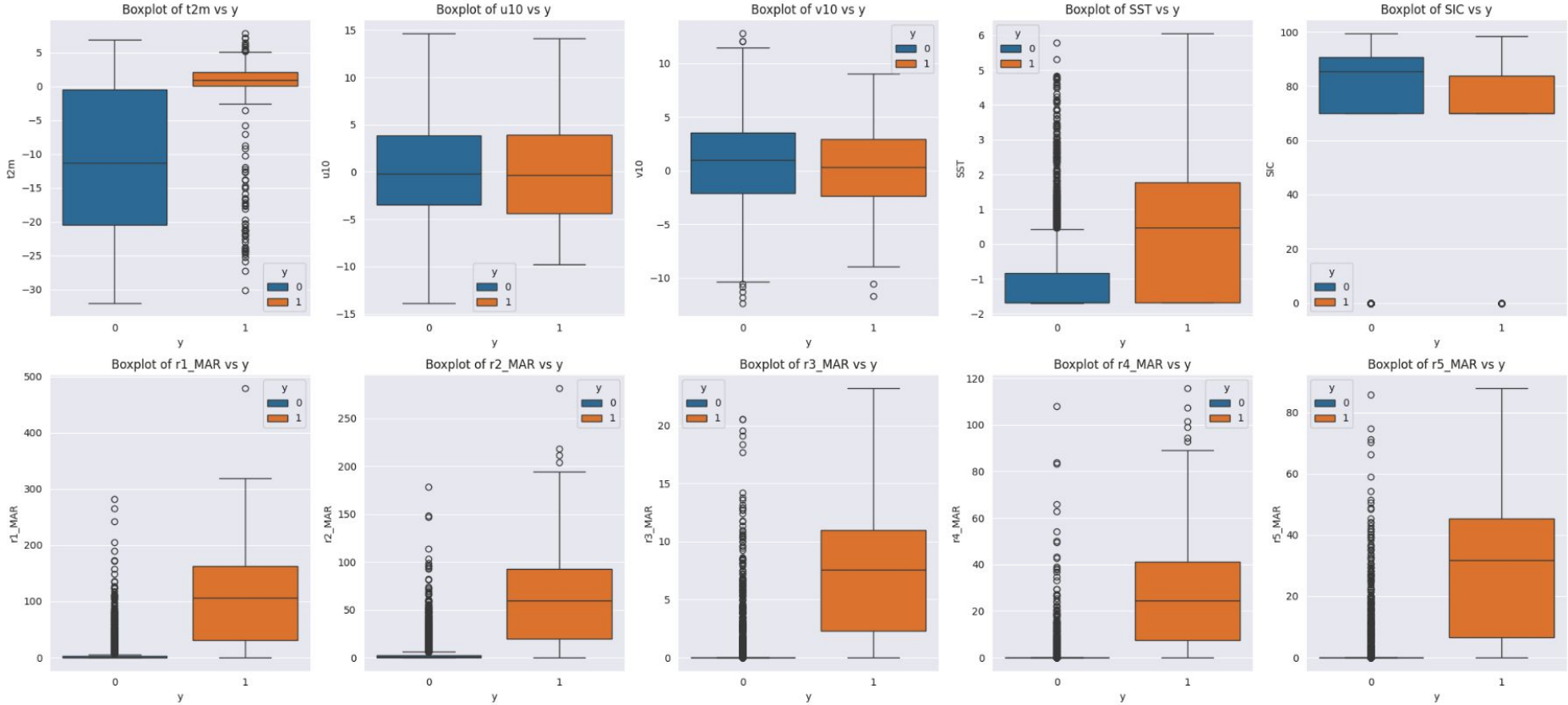


Figure 5: Boxplot of each feature by class

Weighted Error

		Predicted	
		0	1
Observed	0	TN	FP
	1	FN	TP

- ➔ Accuracy : same weight for FP and FN
- ➔ Weighted error : different weight for FP and FN

Approximation of the expectation :

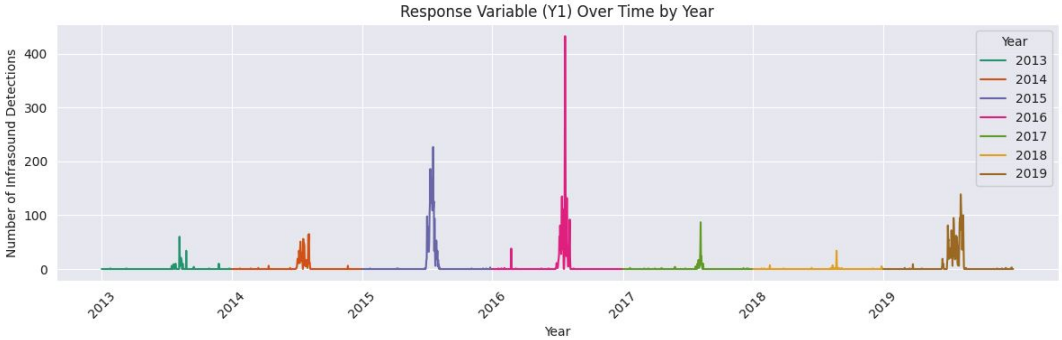
$$\mathbb{E}_{X,Y}[L(Y, f(X))] = \frac{1}{n_{test}} \sum_{i \in Test} L(y_i, f(x_i))$$

where $$L(Y, \hat{Y}) = \begin{cases} 1 & \text{if } Y = 0 \text{ and } \hat{Y} = 1 \text{ (False Positive)} \\ 5 & \text{if } Y = 1 \text{ and } \hat{Y} = 0 \text{ (False Negative)} \\ 0 & \text{otherwise} \end{cases}$$

Models used

- Naive Bayes
- Logistic regression
- Decision Tree
- Random Forest
- Adaboost
- Stacking

Testing plan



- “time” → “year” / “month” / “day”
- dropping “year”, “u10” and “v10”

Cross-validation method used

model	accuracy	weighted error	AUC
Naive Bayes	0.923329	0.198748	0.841264
Logistic regression	0.942889	0.229264	0.804968
Decision Tree	0.909634	0.279343	0.772506
Decision Tree (optimal alpha)	0.937422	0.244131	0.792995
Random Forest	0.943673	0.215962	0.817247
AdaBoost	0.936634	0.240219	0.796992
Stacking	0.943655	0.211268	0.821686

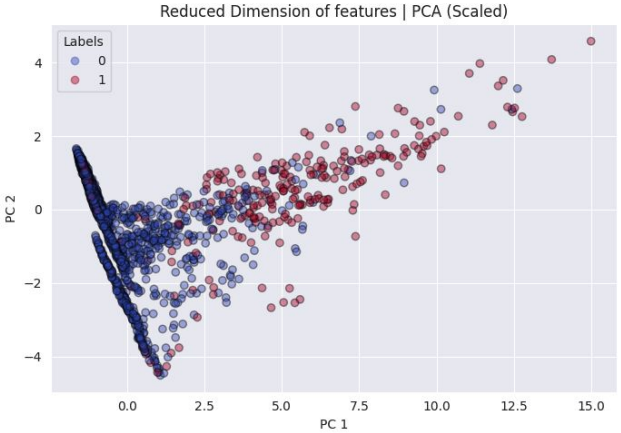
Performance metrics for various classification models

To go further: Data Augmentation

Three methods were used:

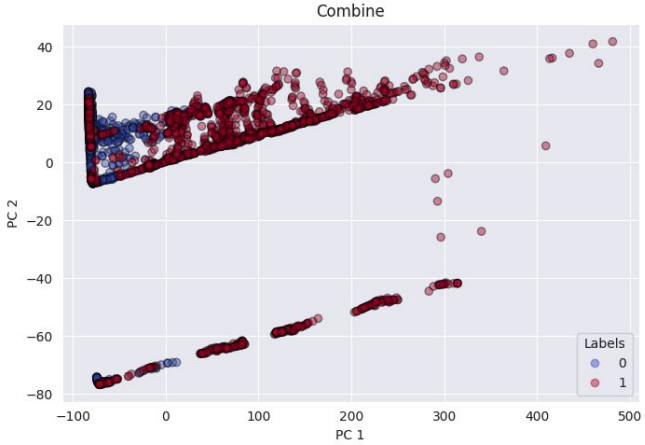
- Random Undersampling
- Oversampling with SMOTE (Synthetic Minority Oversampling Technique)
- Combination of SMOTE and Edited Nearest Neighbors Undersampling (SMOTEENN)

-> SMOTEENN is the best approach



Before: 2556 obs

After: 3745 obs



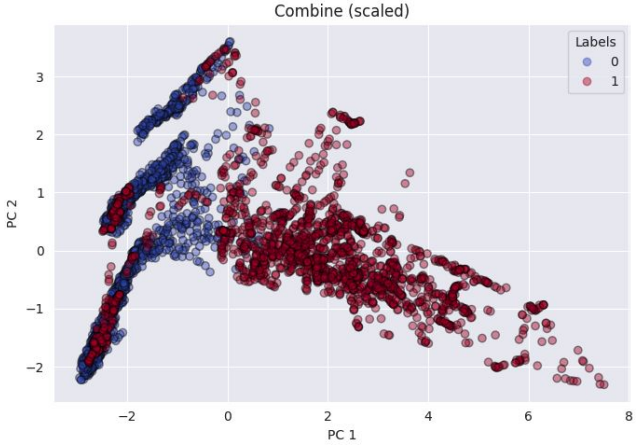


Figure 4: 2D space plot of augmented dataset

model	accuracy	weighted error	AUC
Naive Bayes	0.888652	0.452069	0.892264
Logistic regression	0.907610	0.437383	0.912806
Decision Tree	0.967162	0.090254	0.968184
Decision Tree (optimal alpha)	0.968765	0.086515	0.969832
Random Forest	0.981311	0.077437	0.981970
AdaBoost	0.928177	0.316422	0.931377

Performance metrics for various classification models with data augmentation

Thank you for your attention.