Improving rare events predictions by oversampling a tabular data with a mix of categorical and continuous variables

Dr. Alla Sapronova, Lead Data Scientist at StormGeo

Dr. Esa-Matti Tastula, Data Scientist at StormGeo

Dr. Gard Hauge, CTO at StormGeo

Dr. Nina Winther-Kaland, Research Director

StormGeo
Navigate tomorrow – today

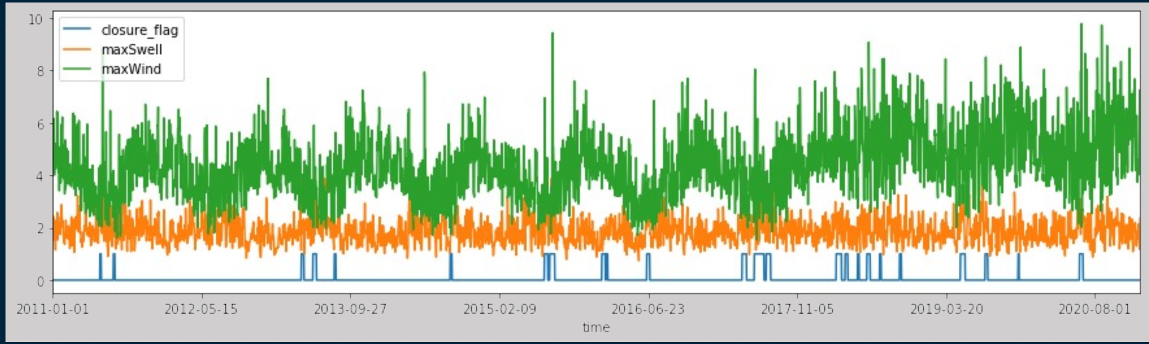
// Goal and Approach

- Use reported data about port closures and modeled weather and wave data to build <u>model that can accurately predict</u> if a given port is closed on a given data.

- We've used port closure information from together with historical weather and wave data (ERA5) from the ECMWF to train and validate several machine learning models.

- 1000 cases, of which 20% are port closures.
- Closures data from 20 ports, 10 years history

StormGeo
Navigate tomorrow – today



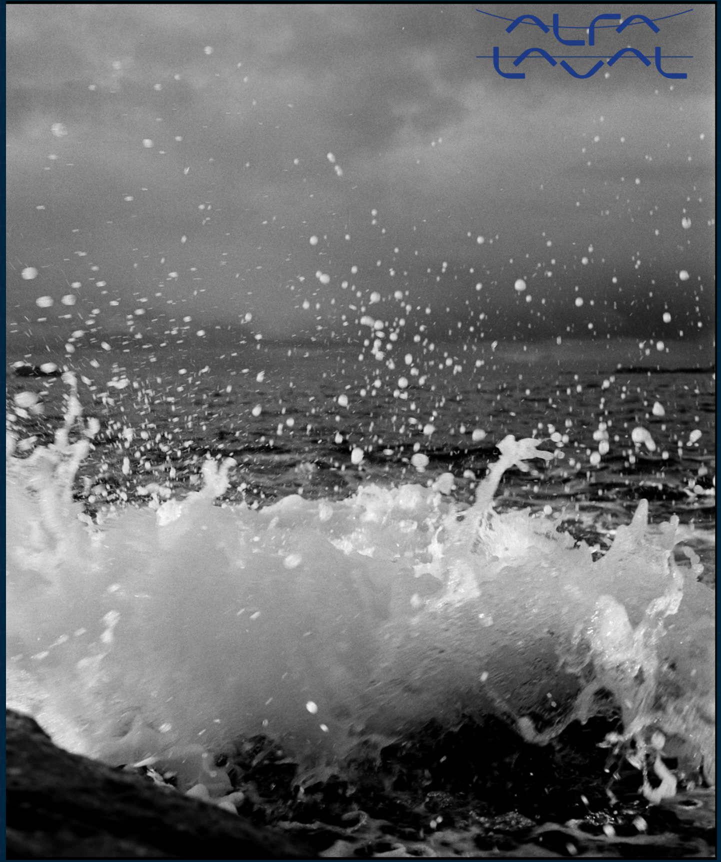
// Choosing the Right Features for the Model

StormGeo
Navigate tomorrow – today

We experimented with

- **maximum daily 10-m wind speed**
- **maximum daily significant wave height**
- mean daily swell period
- mean daily wave direction
- **port as explaining features in the model**

The RF shows the highest feature importance score for 10-m wind speed, maximum daily significant wave height and port name (category)



// Results on test dataset for RF (XGB)

Default 50% threshold (50% or greater likelihood that a port is closed will appear closed in the output)

Can adjust the threshold to minimize the number of False Open

		PREDICTED	
		OPEN	CLOSED
ACTUAL	OPEN	TRUE OPEN **187**	FALSE CLOSED (OVERPREDICTED) **6**
	CLOSED	FALSE OPEN (UNDERPEDICTED) 14	TRUE CLOSED 43

		PREDICTED	
		OPEN	CLOSED
ACTUAL	OPEN	TRUE OPEN **187 ⇨ 196**	FALSE CLOSED (OVERPREDICTED) **6 ➡ 21**
	CLOSED	FALSE OPEN (UNDERPEDICTED) 14 ➡ 5	TRUE CLOSED 43 ⇨ 28

// Results on test dataset for RF (XGB)

StormGeo
Navigate tomorrow – today

Default 50% threshold (50% or greater likelihood that a port is closed will appear closed in the output)

Can adjust the threshold to minimize the number of False Open

		PREDICTED	
		OPEN	CLOSED
ACTUAL	OPEN	TRUE OPEN 187	FALSE CLOSED 6
	CLOSED	FALSE OPEN 14	TRUE CLOSED 43

```
Base Accuracy: 0.9335664335664335
Base classification report:
              precision    recall  f1-score

           0       0.94      0.97      0.96
           1       0.88      0.78      0.83

    accuracy                           0.93
   macro avg       0.91      0.88      0.89
weighted avg       0.93      0.93      0.93
```

		PREDICTED	
		OPEN	CLOSED
ACTUAL	OPEN	TRUE OPEN 196	FALSE CLOSED 21
	CLOSED	FALSE OPEN 5	TRUE CLOSED 28

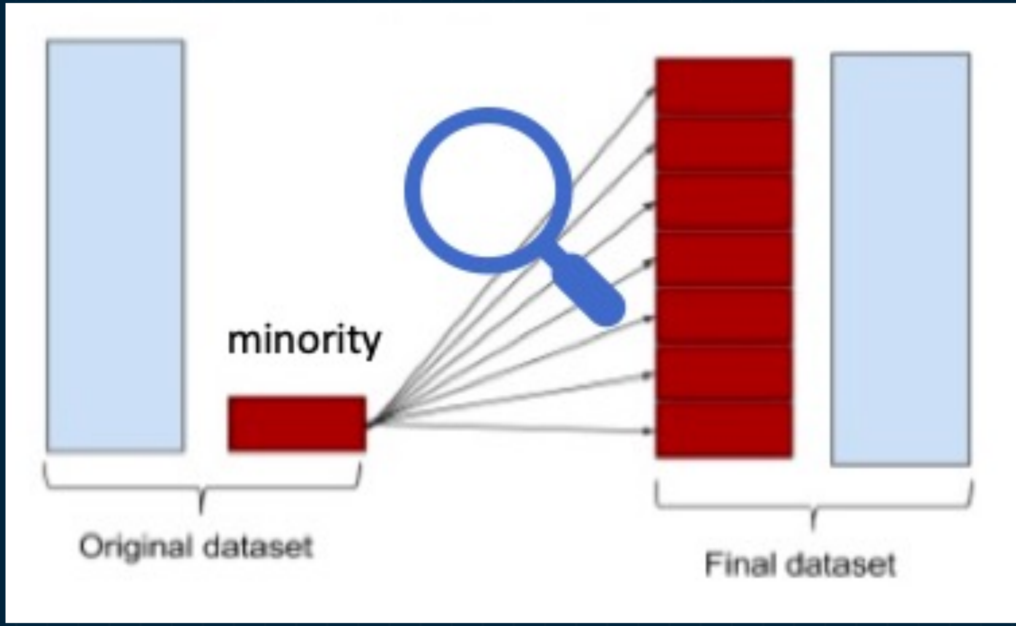
```
Accuracy of fake data model: 0.8526315789473684
Classification report of fake data model:
              precision    recall  f1-score

           0       0.88      0.87      0.87
           1       0.82      0.83      0.82

    accuracy                           0.85
   macro avg       0.85      0.85      0.85
weighted avg       0.85      0.85      0.85
```

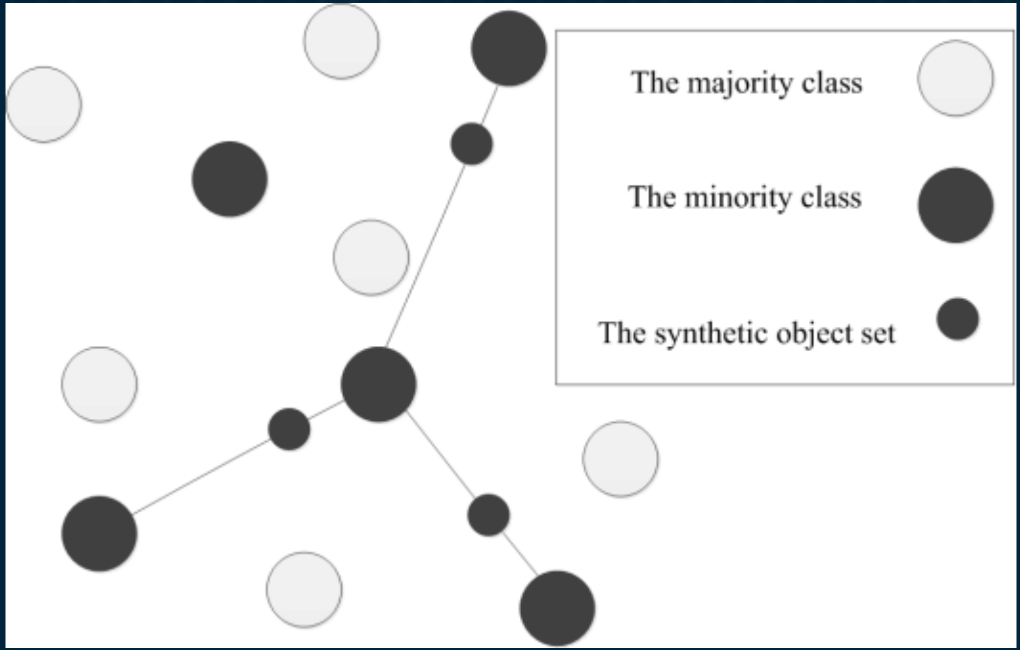
// Synthesizing More Data

- Too little data on closures leads to under-predicting:
- SMOTE, ADASYN

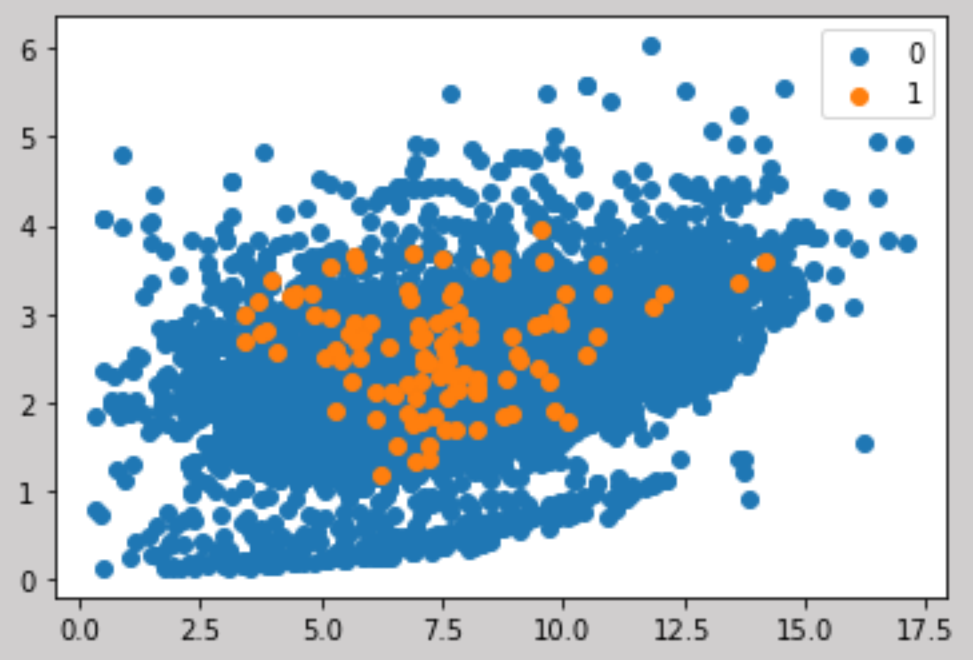
What: oversample - produce more underrepresented data (closure)

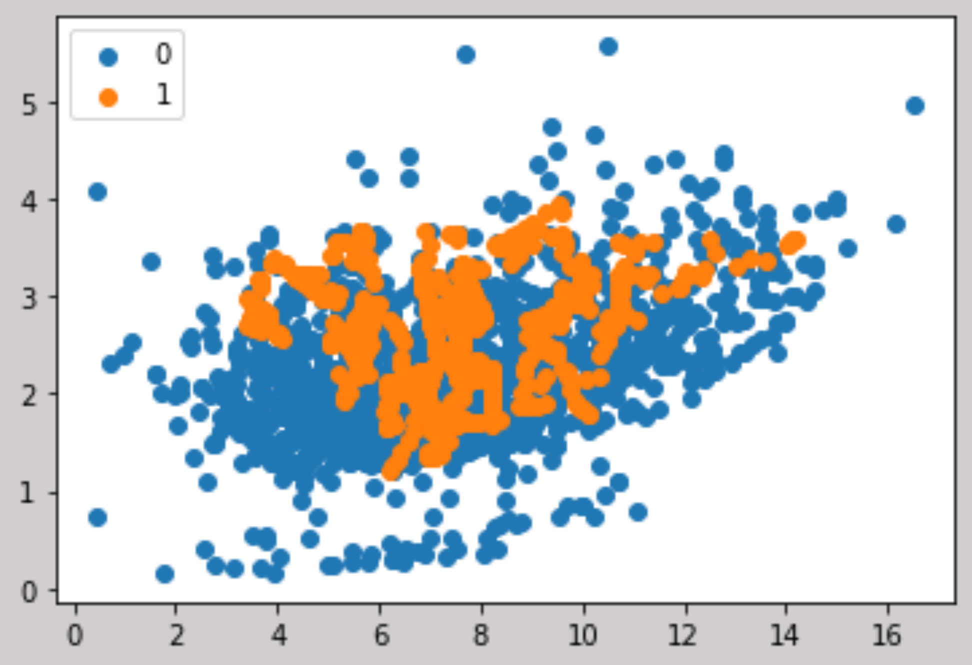
How: calculate points located close to existing points





// Similarity score for SMOTE generated data

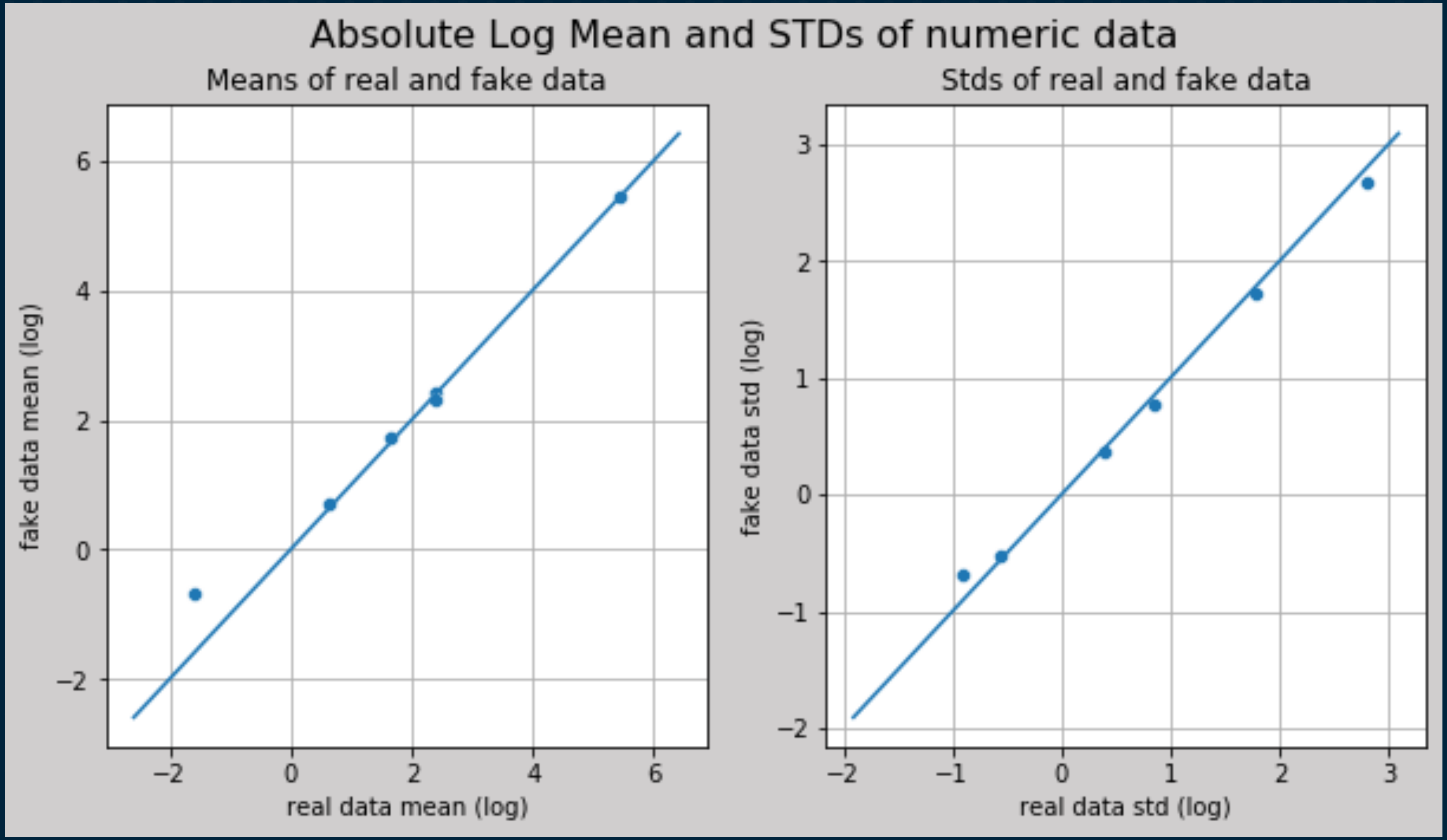




Results:

	result
Basic statistics	0.9757
Correlation column correlations	0.9753
Mean Correlation between fake and real columns	0.9100
1 - MAPE Estimator results	0.8860
Similarity Score	0.9367

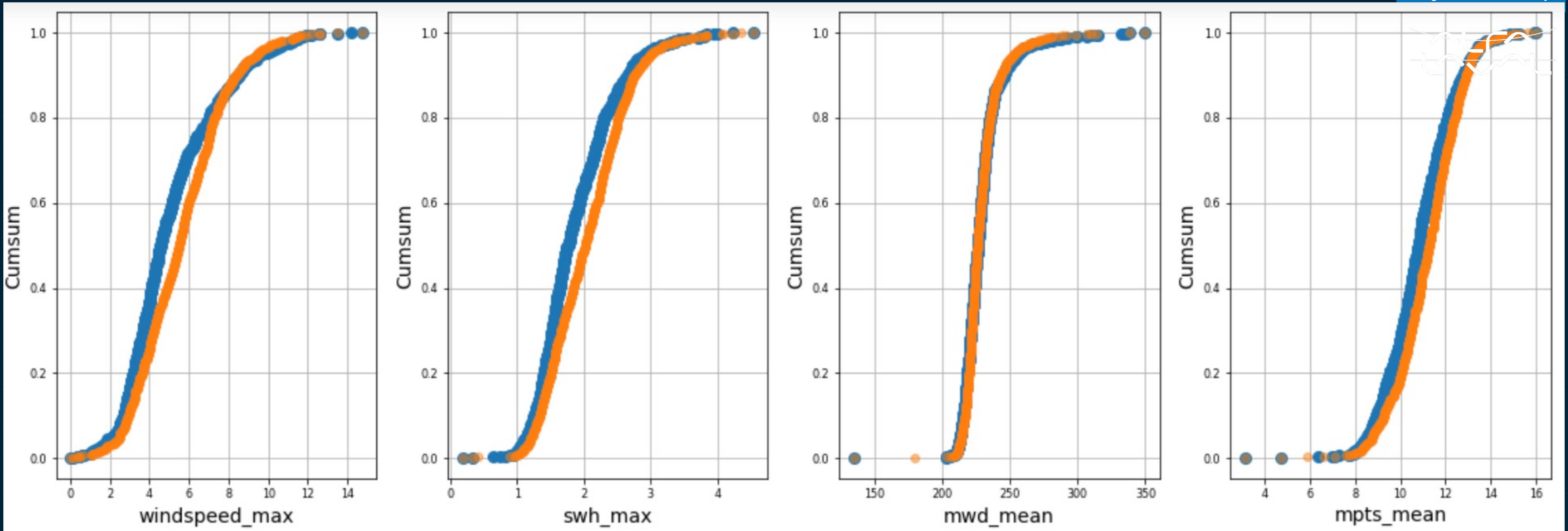
// Similarity for SMOTE- generated data



StormGeo
Navigate tomorrow – today

// Accuracy for RF trained on SMOTE

StormGeo
Navigate tomorrow – today

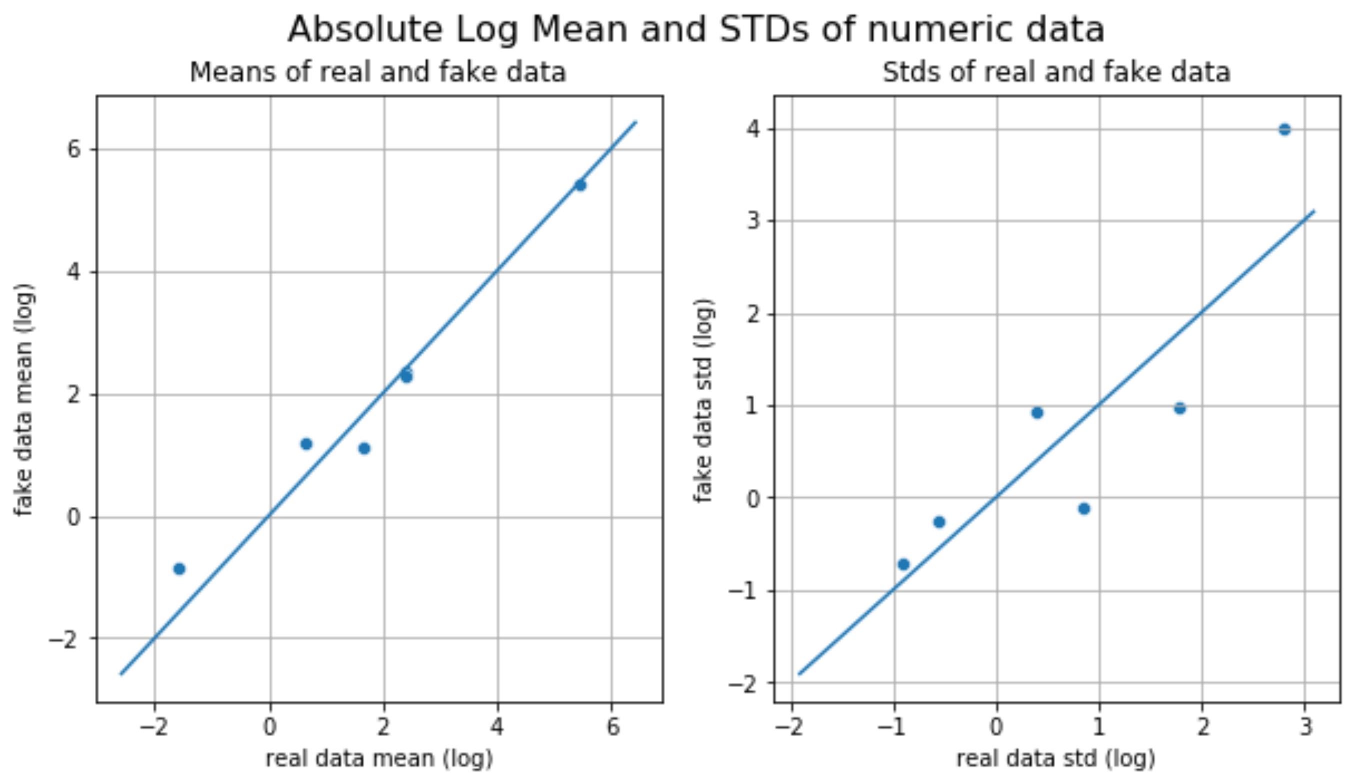



```
Accuracy for Random Forest on data: 89.48
Accuracy list: [90.56 88.81 86.71 90.56 91.26 89.16 88.81 86.01 90.56 92.31]
```

```
Accuracy for Random Forest on SMOTE: 92.97
Accuracy list: [91.96 94.41 93.71 91.96 91.61 93.71 92.31 94.41 92.31 93.36]
```

// Similarity for GAN- generated data

StormGeo
Navigate tomorrow – today



Results:

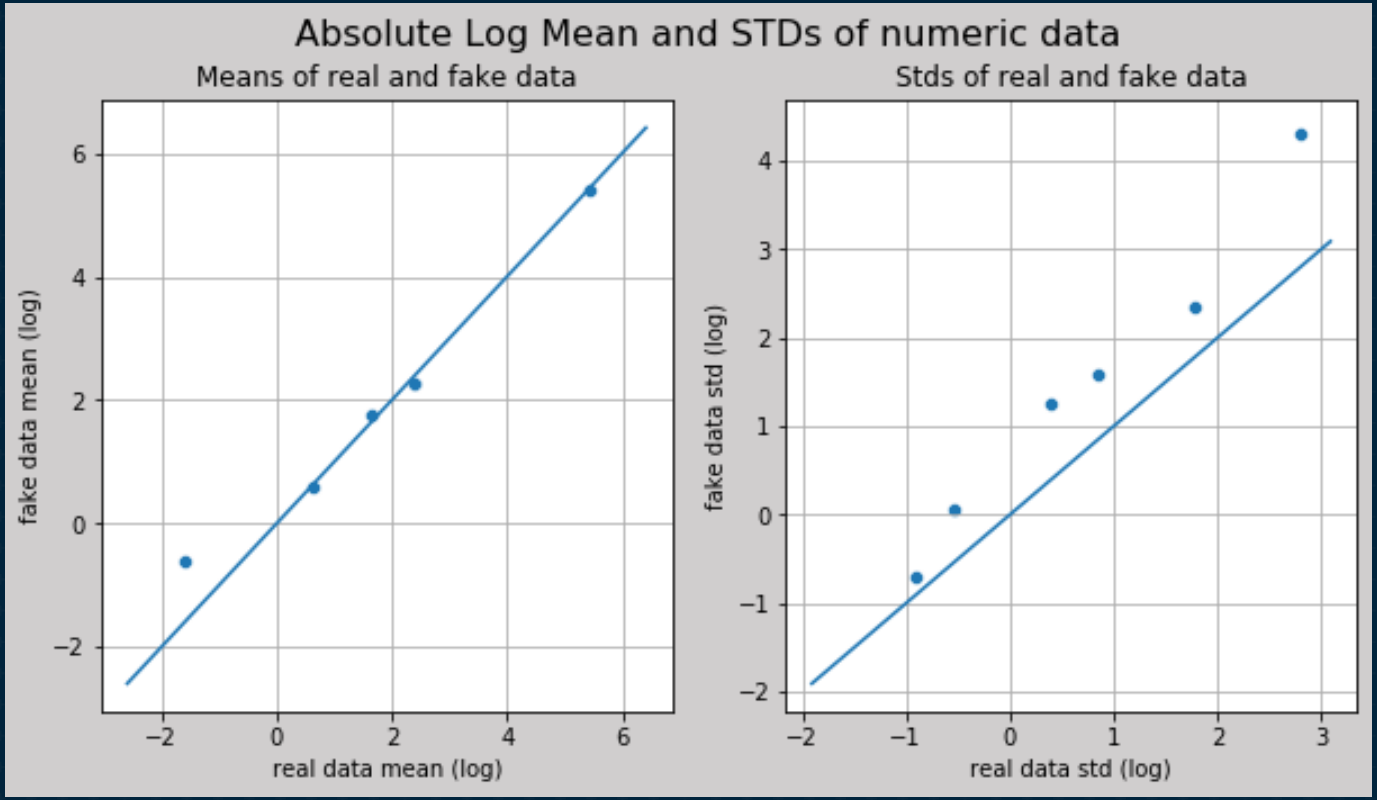
	result
Basic statistics	0.8496
Correlation column correlations	-0.0679
Mean Correlation between fake and real columns	0.8987
1 - MAPE Estimator results	0.6399
Similarity Score	0.5801

Accuracy of fake data model: 0.9052631578947369
Classification report of fake data model:

	precision	recall	f1-score	support
0	0.86	0.94	0.90	125
1	0.95	0.88	0.91	160
accuracy			0.91	285
macro avg	0.90	0.91	0.90	285
weighted avg	0.91	0.91	0.91	285

// Similarity for GAN- generated data, 2

StormGeo
Navigate tomorrow – today



Accuracy of fake data model: 0.9894736842105263
Classification report of fake data model:

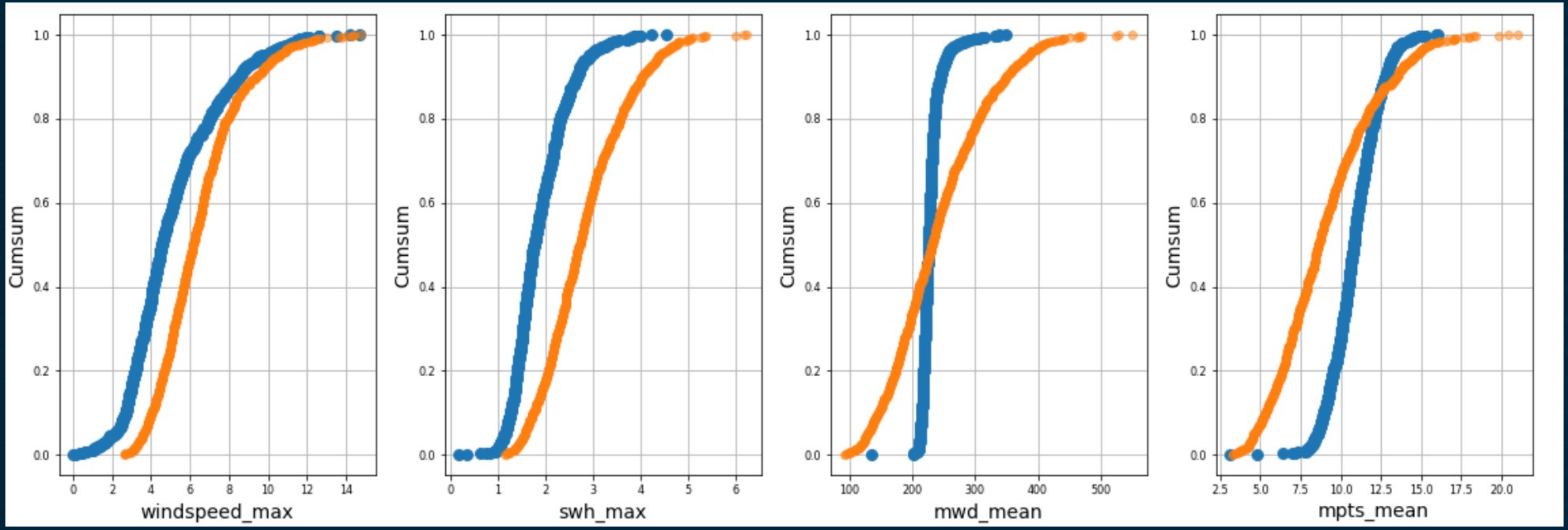
	precision	recall	f1-score	support
0	0.99	0.99	0.99	154
1	0.98	0.99	0.99	131
accuracy			0.99	285
macro avg	0.99	0.99	0.99	285
weighted avg	0.99	0.99	0.99	285

Results:

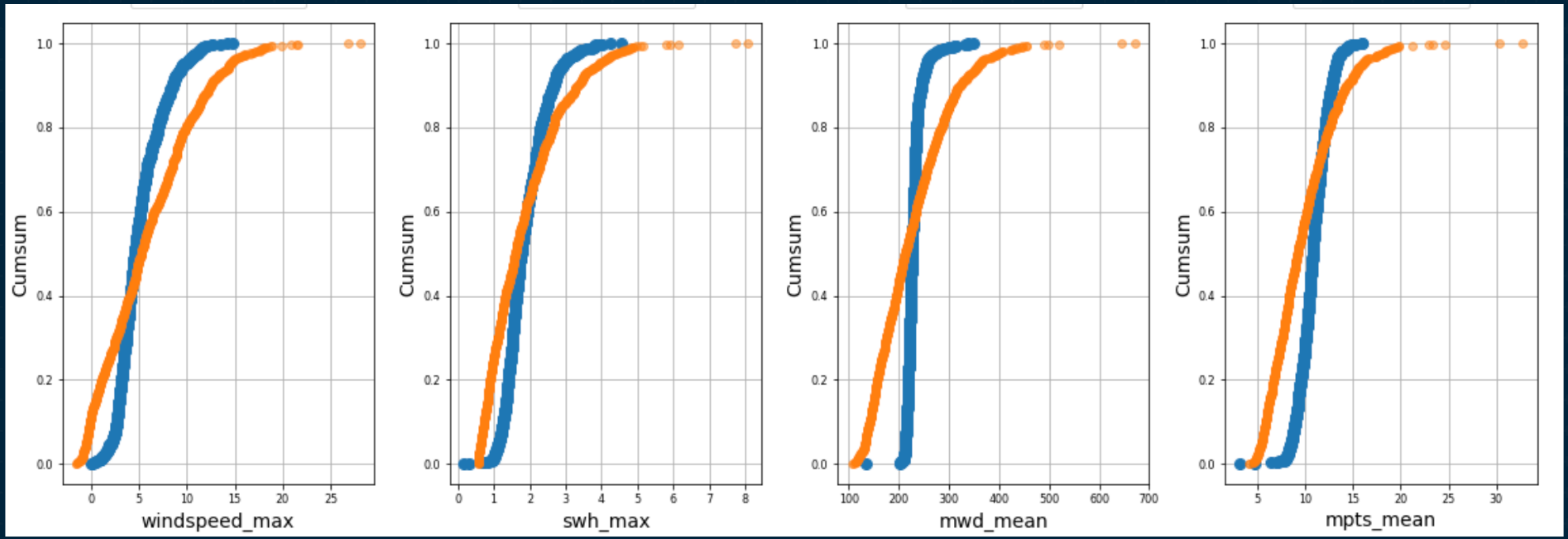
	result
Basic statistics	0.9235
Correlation column correlations	0.5228
Mean Correlation between fake and real columns	0.8848
1 - MAPE Estimator results	0.7226
Similarity Score	0.7634

// Goal and Approach

StormGeo
Navigate tomorrow – today




```
Accuracy of fake data model: 0.9052631578947369
Classification report of fake data model:
              precision    recall  f1-score

           0       0.86      0.94      0.90
           1       0.95      0.88      0.91

    accuracy                           0.91
   macro avg       0.90      0.91      0.90
weighted avg       0.91      0.91      0.91
```




```
Accuracy of fake data model: 0.9894736842105263
Classification report of fake data model:
              precision    recall  f1-score

           0       0.99      0.99      0.99
           1       0.98      0.99      0.99

    accuracy                           0.99
   macro avg       0.99      0.99      0.99
weighted avg       0.99      0.99      0.99
```

// Summary

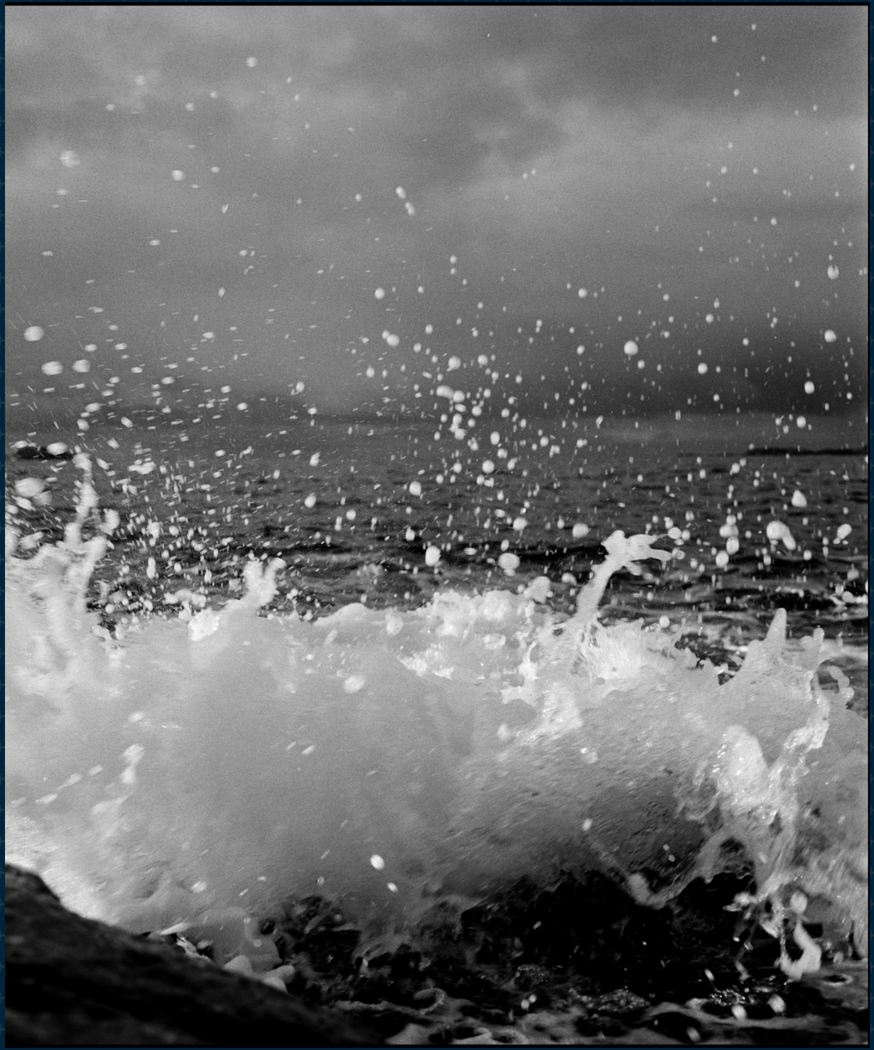
StormGeo
Navigate tomorrow – today

Overall success rate for closure status 90% (port open or closed), RF trained on on historical events

The data augmentation approach help to build an accurate ML predictive model for rare events forecasting

ML (RF classifier) trained on data generated with

- ADASYN
- SMOTE(C)
- WGAN
- DCGAN



// Summary

StormGeo
Navigate tomorrow – today

how the methods and amount of synthesized data affects predictive model's accuracy

RF, imbalanced data

Base Accuracy: 0.9335664335664335
Base classification report:

	precision	recall	f1-score
0	0.94	0.97	0.96
1	0.88	0.78	0.83
accuracy			0.93
macro avg	0.91	0.88	0.89
weighted avg	0.93	0.93	0.93

RF, SMOTE

Accuracy of fake data model: 0.9403508771929825
Classification report of fake data model:

	precision	recall	f1-score
0	0.94	0.96	0.95
1	0.94	0.92	0.93
accuracy			0.94
macro avg	0.94	0.94	0.94
weighted avg	0.94	0.94	0.94

RF, DCGAN

Accuracy of fake data model: 0.9894736842105263
Classification report of fake data model:

	precision	recall	f1-score
0	0.99	0.99	0.99
1	0.98	0.99	0.99
accuracy			0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99

RF, imbalanced, threshold

Accuracy of fake data model: 0.852631578947368
Classification report of fake data model:

	precision	recall	f1-score
0	0.88	0.87	0.87
1	0.82	0.83	0.82
accuracy			0.85
macro avg	0.85	0.85	0.85
weighted avg	0.85	0.85	0.85

RF, ADASYN

Accuracy of fake data model: 0.933333333333333
Classification report of fake data model:

	precision	recall	f1-score
0	0.94	0.94	0.94
1	0.93	0.92	0.92
accuracy			0.93
macro avg	0.93	0.93	0.93
weighted avg	0.93	0.93	0.93

RF, WGAN

Accuracy of fake data model: 0.9473684210526315
Classification report of fake data model:

	precision	recall	f1-score
0	0.95	0.96	0.95
1	0.94	0.94	0.94
accuracy			0.95
macro avg	0.95	0.95	0.95
weighted avg	0.95	0.95	0.95

// Summary

What evaluation metrics most suitable for data quality check and predictive models' assessment when using datasets containing synthetic data:

- Accuracy (F1-score)
- Confusion matrix
- Similarity score