

# Monitoring technical performance directly

MACHINE LEARNING MONITORING CONCEPTS



Hakim Elakhrass

Co-founder and CEO of NannyML

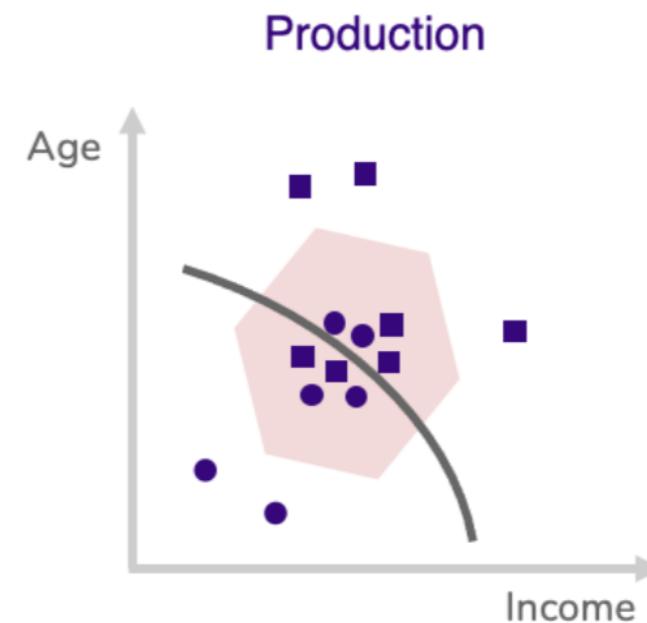
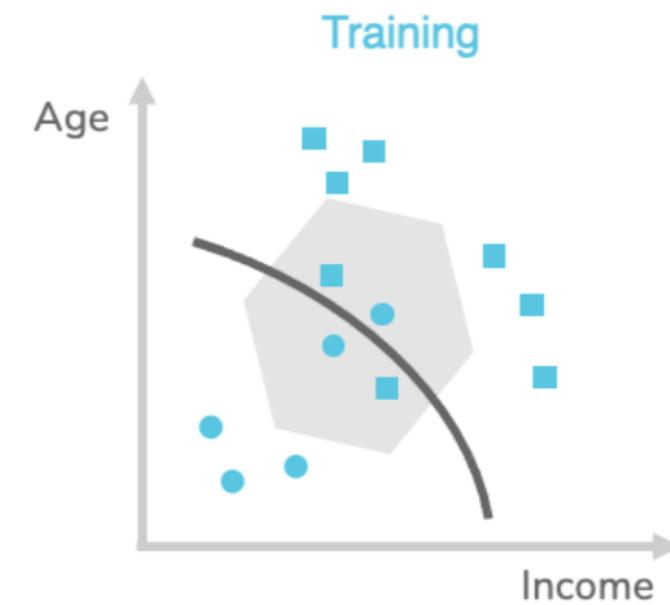
# Covariate shift - performance relationship

Three covariate shifts:

- Shifts to certain regions
    - more high-income people applying for loans
  - Shifts to underrepresented regions
    - having 10% of tech people applying for loans instead of 0.5% from training data
  - Shifts to less certain regions
    - shift from 20% to 40% of middle-income people which are close to decision boundary
- > No impact
- > Unknown impact
- > Negative impact

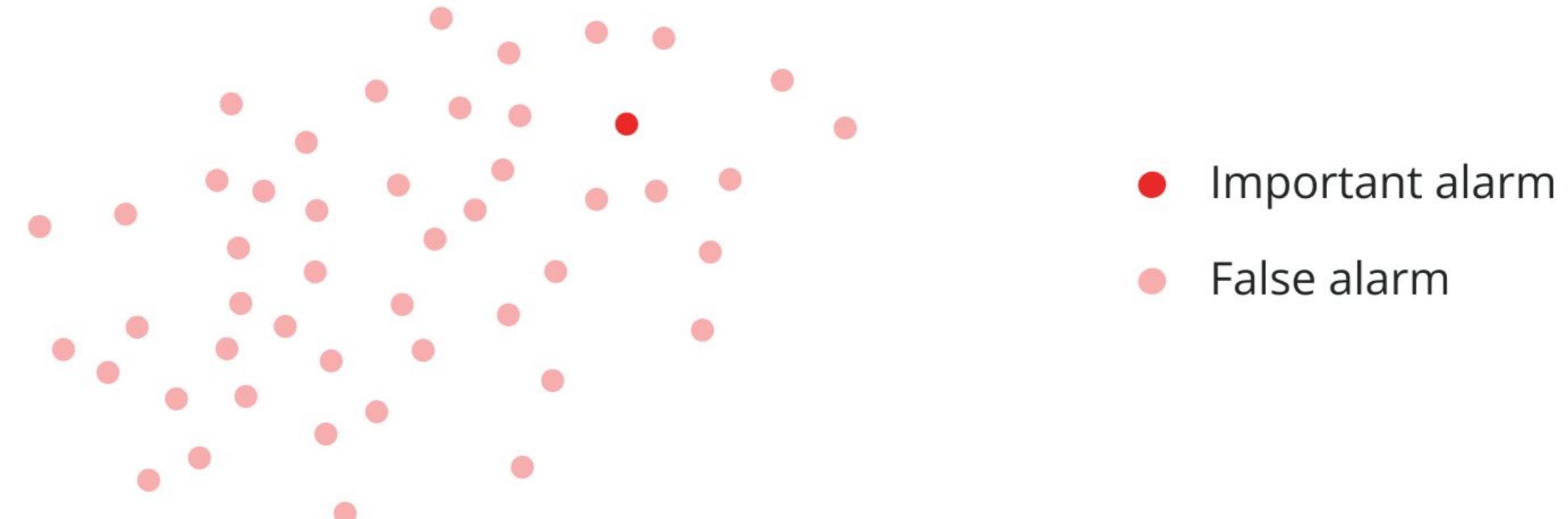
# Guaranteed negative impact

Covariate shift to uncertain regions always negatively impacts performance



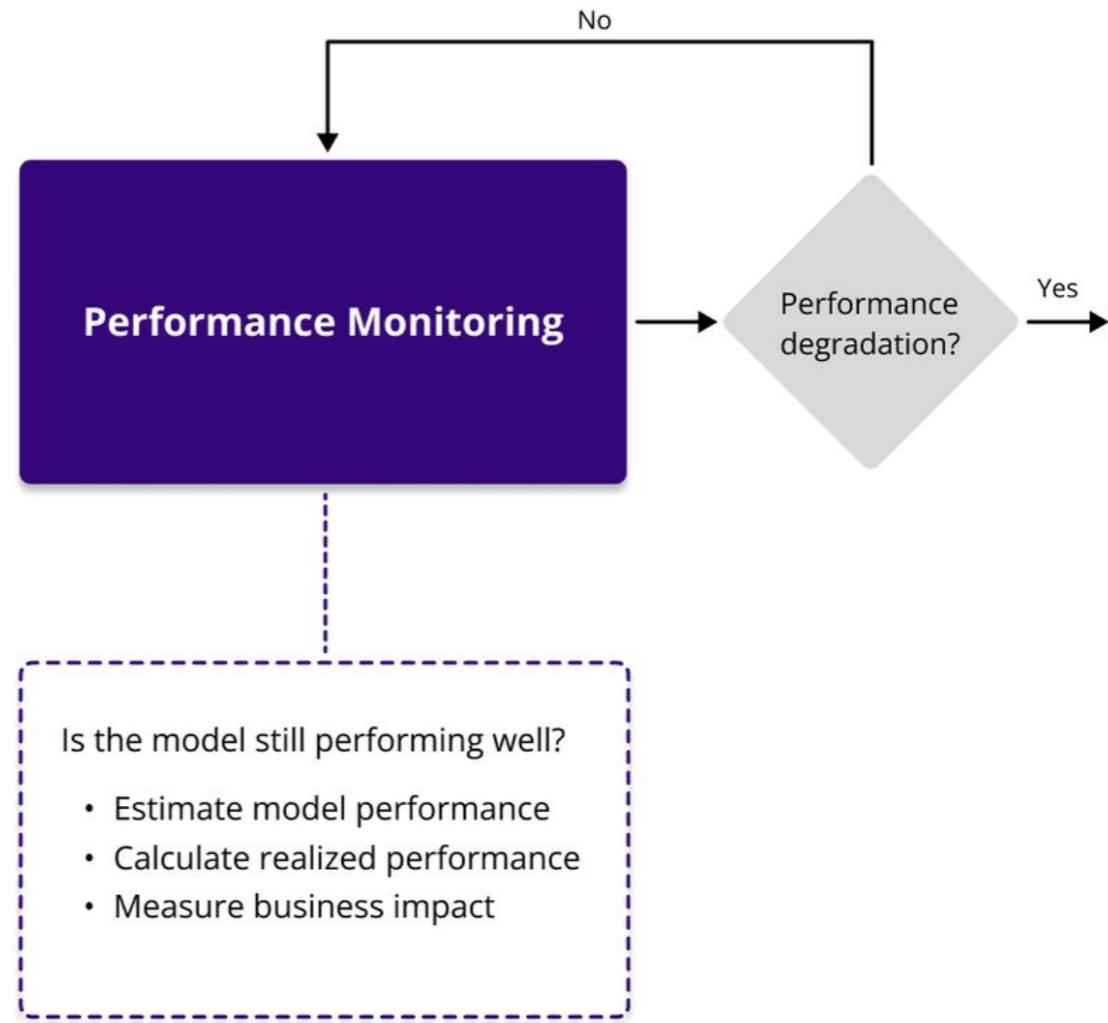
# False alerts problem

Alert fatigue



# The importance of technical performance

- Direct metric of how well the model performs the task at hand
- Reflects any silent model failure
- Removes the overload of false alerts



# **Let's practice!**

**MACHINE LEARNING MONITORING CONCEPTS**

# Availability of ground truth

MACHINE LEARNING MONITORING CONCEPTS



Hakim Elakhrass

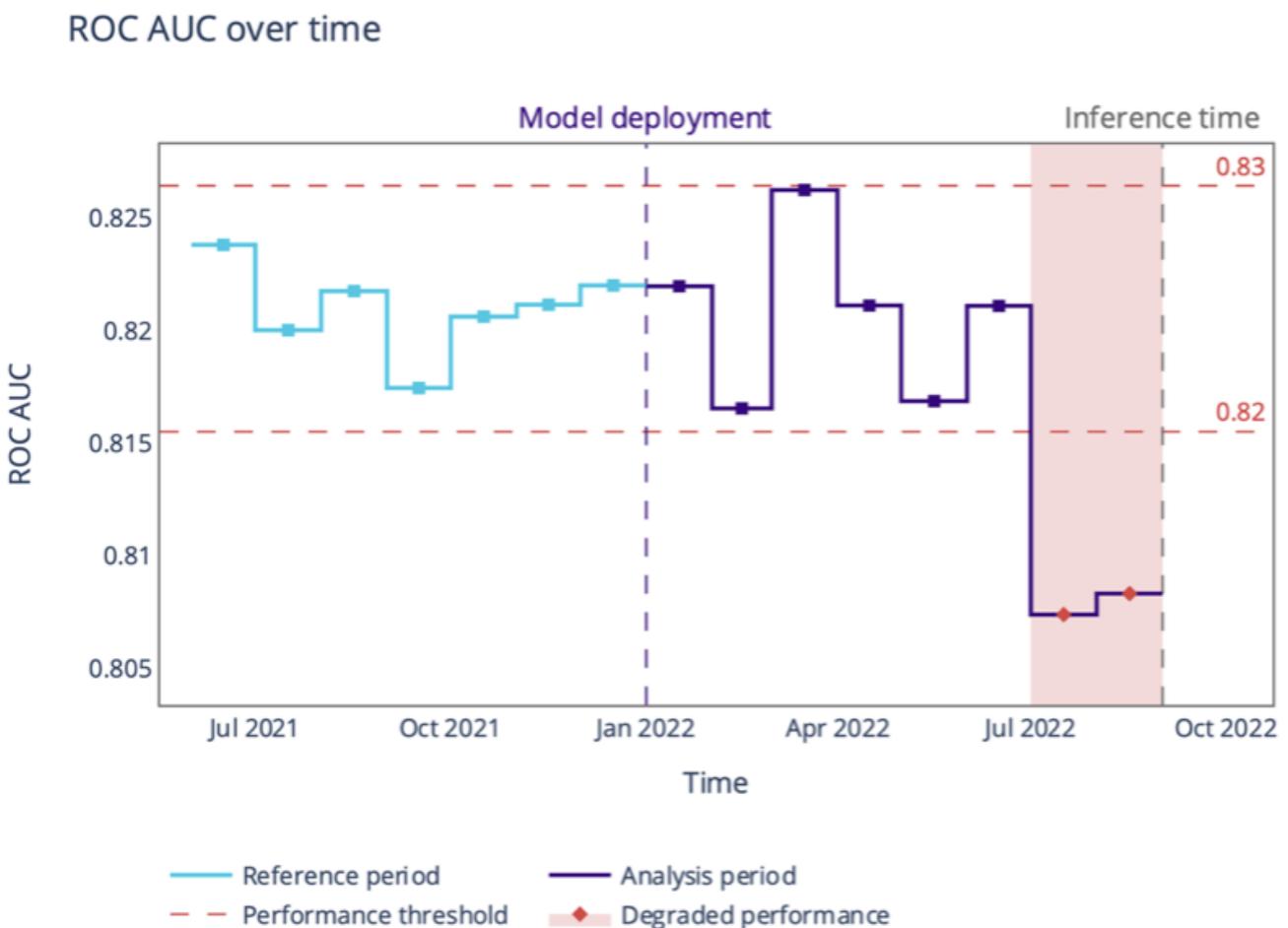
Co-founder and CEO of NannyML

# Instant ground truth



- Measures actual performance
- Easy evaluation
- Accurate

# Production data - instant



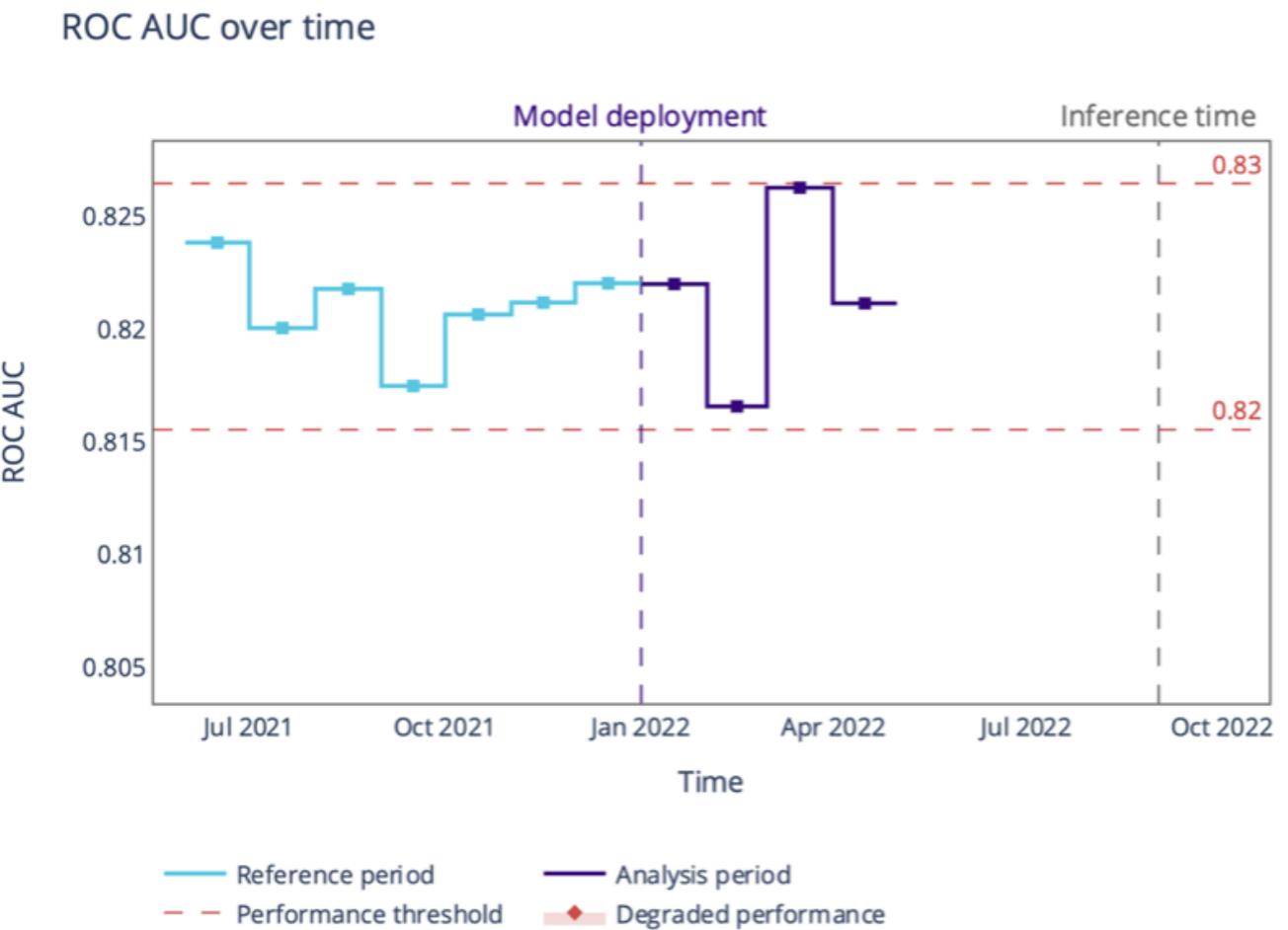
Period	Chunk	Predicted probability	Target	ROC AUC
Reference	Jun 2021	0.36	True	0.824
	...	...	False	...
Analysis	Dec 2021	0.32	False	0.821
	...	...	False	...
Analysis	Jan 2022	0.12	True	0.821
	...	...	False	...
	Nov 2022	0.16	True	0.807
	...	...	False	...

# Delayed ground truth



- Delay depends on the application
- A possible scenario is loan default prediction
- Unknown performance in the meantime
- Requires performance estimation

# Production data - delayed



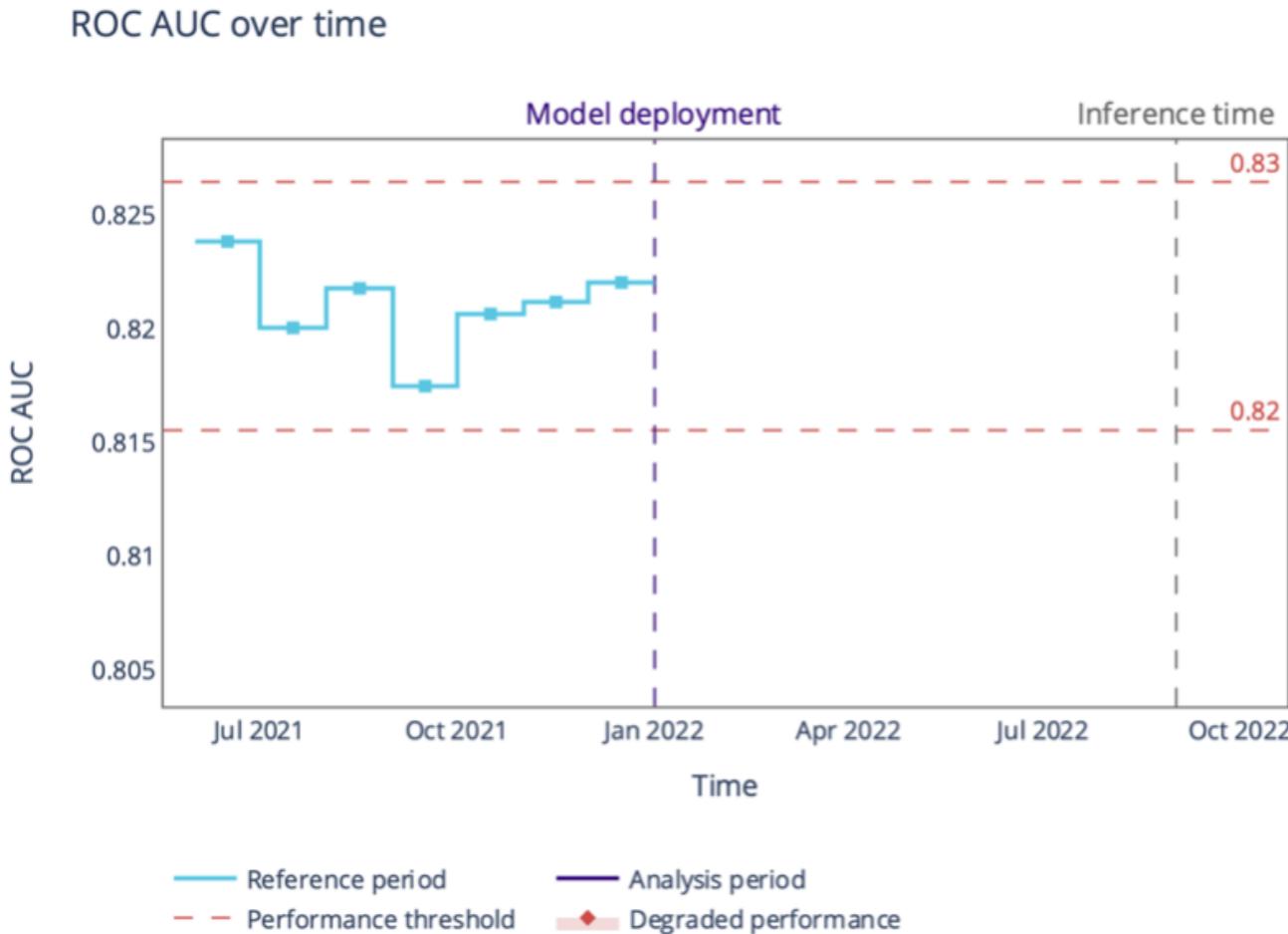
Period	Chunk	Predicted probability	Target	ROC AUC
Reference	Jun 2021	0.36	True	0.824
	... Dec 2021	0.02 0.32	False False	0.821
Analysis	Jan 2022	0.12 0.04	True False	0.821
	... Nov 2022	...	/	/

# Absent ground truth



- Present in fully-automated processes
- A possible scenario is insurance pricing
- Actual performance is unknown
- Requires performance estimation

# Production data - absent



Period	Chunk	Predicted probability	Target	ROC AUC
Reference	Jun 2021	0.36 ... 0.02	True ... False	0.824
	Dec 2021	0.32 ... 0.01	False ... False	0.821
Analysis	Jan 2022	0.12 ... 0.04	/	/
	Nov 2022	0.03 ... 0.01	/	/

# **Let's practice!**

**MACHINE LEARNING MONITORING CONCEPTS**

# Performance estimation

MACHINE LEARNING MONITORING CONCEPTS



**Hakim Elakhrass**

Co-founder and CEO of NannyML

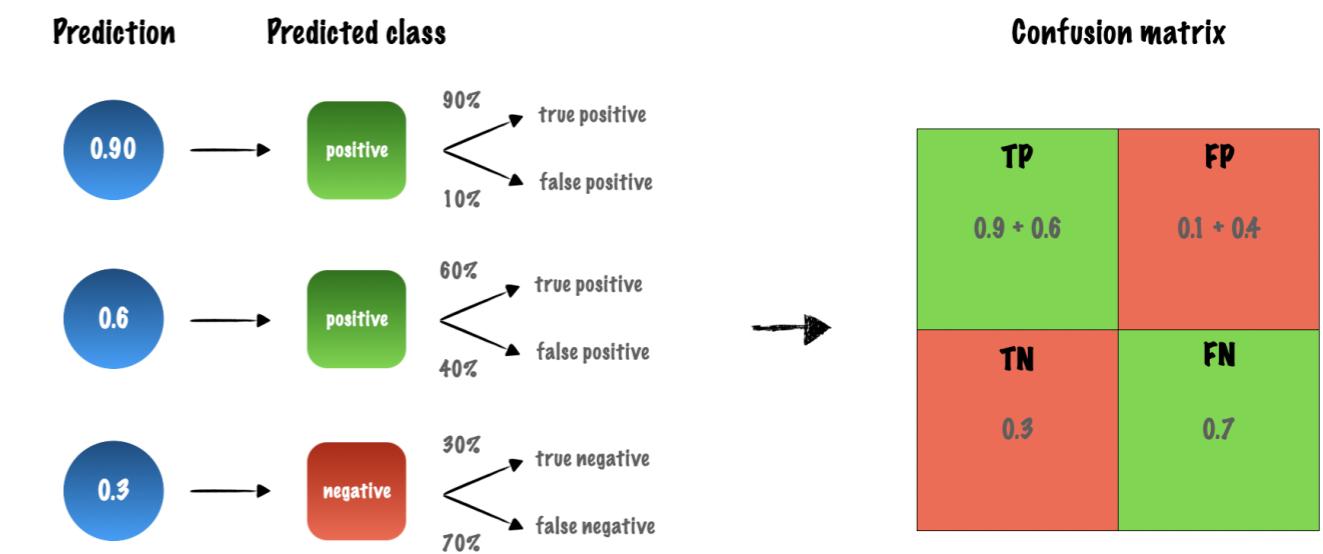
# Overview

Two performance estimation algorithms:

- CBPE - classification tasks
- DLE - regression tasks

# CBPE - How it works

- Estimates a confusion matrix
- Allows to calculate classification metrics like accuracy, precision, recall
- Captures the impact of covariate shift on the model



# CBPE - Considerations

- No covariate shift in the unseen regions
- No concept drift is present in the incoming data
- Requires a probability calibration

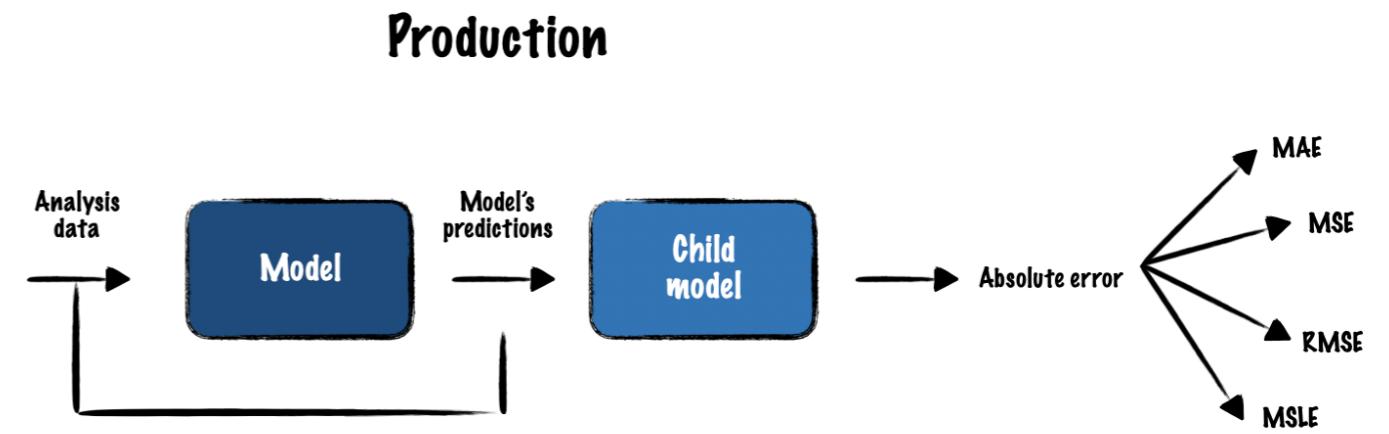
# DLE - How it works

- Predicts the absolute error of the model

- Uses an external child model

- Allows to calculate various regression metrics like MAE, MSE, MSLE

- Captures the presence of covariate shift in the input data



# DLE - Considerations

- No covariate shift in the unseen regions
- No concept drift is present in the incoming data
- Extra complexity

# **Let's practice!**

**MACHINE LEARNING MONITORING CONCEPTS**