How to detect silent ML failure?

An introduction to ML Monitoring

by Wojtek Kuberski

Agenda

Data drift and concept drift: what are they?

Performance estimation (without access to targets)

Data and concept drift detection

Setting the Stage: Loan Default Prediction

Credit Scores and Customer Information → Loan Defaults

Target:

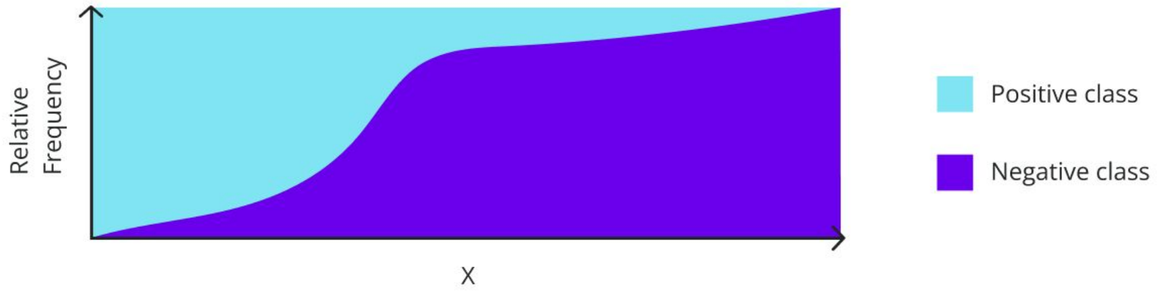
non-payment within 1 year

Technical Metric:

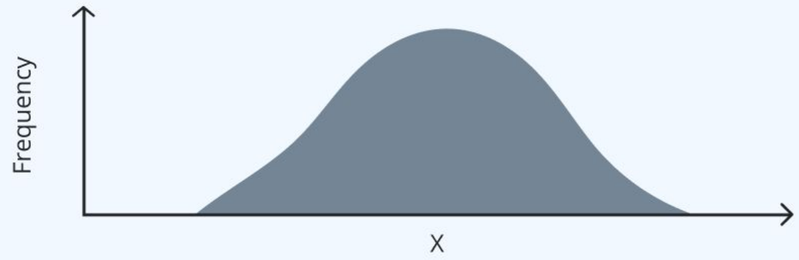
ROC AUC

The basics: true pattern, sampling and data

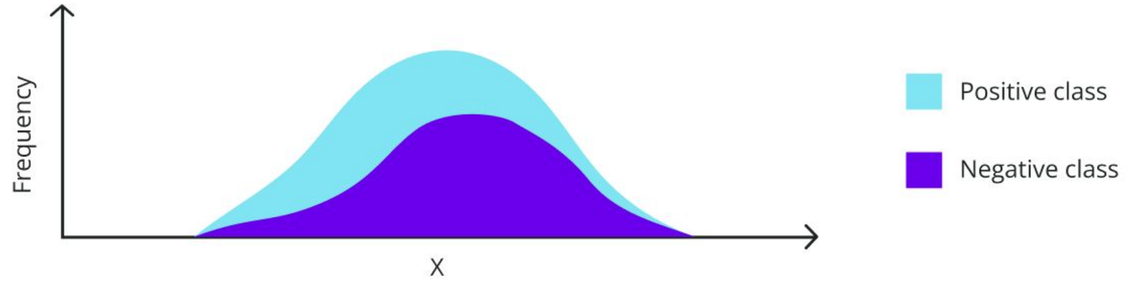
True pattern



Sampling

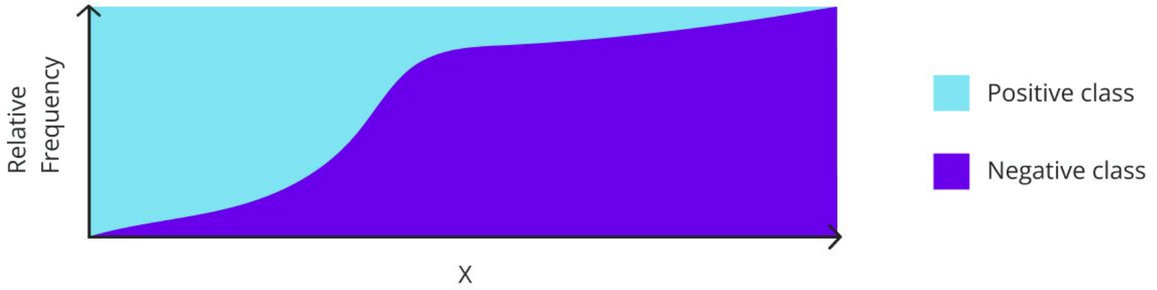


Data

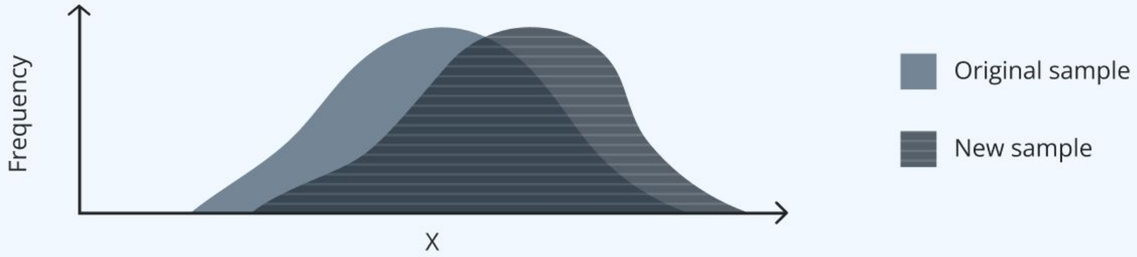


Data drift

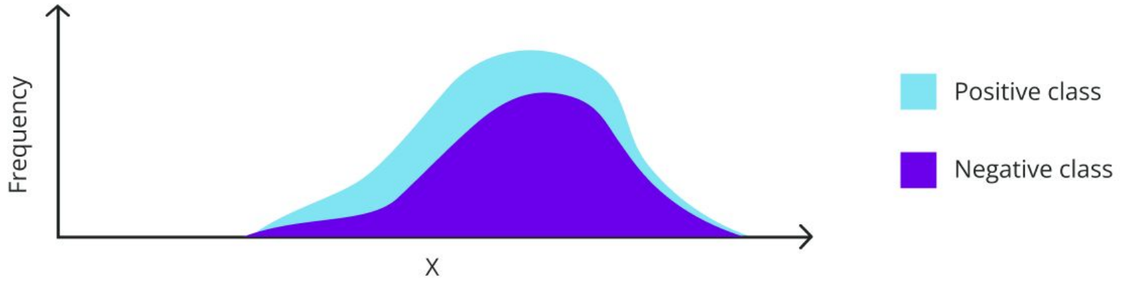
True pattern (unchanged)



Sampling

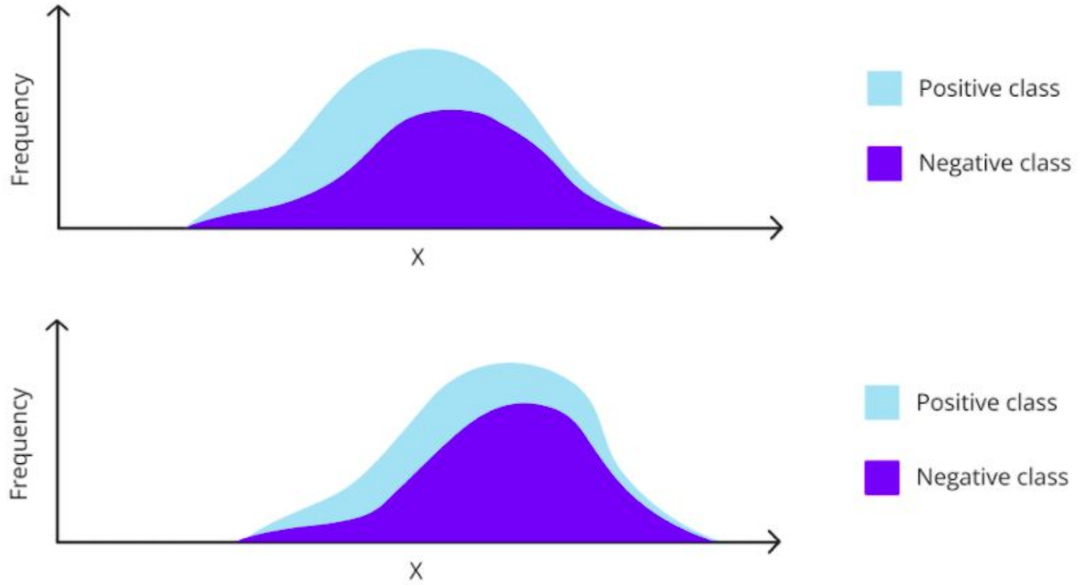


Data



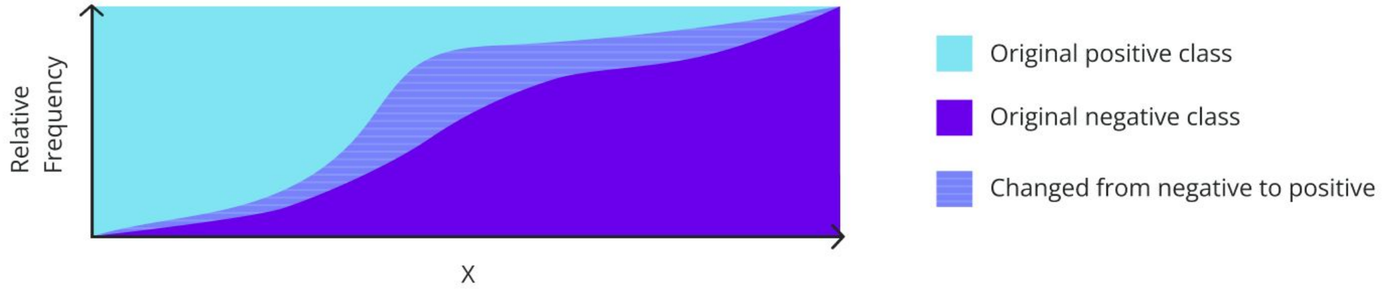
Data drift

Change joint model input distribution - P(x)

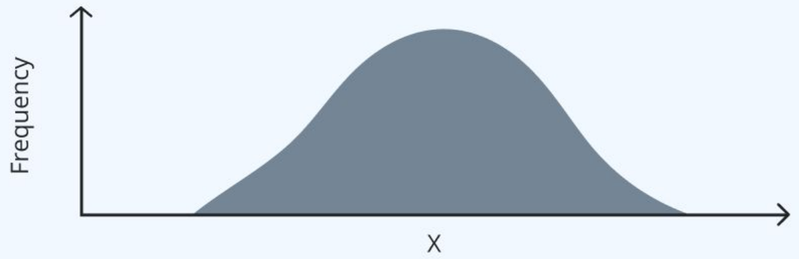


Concept drift

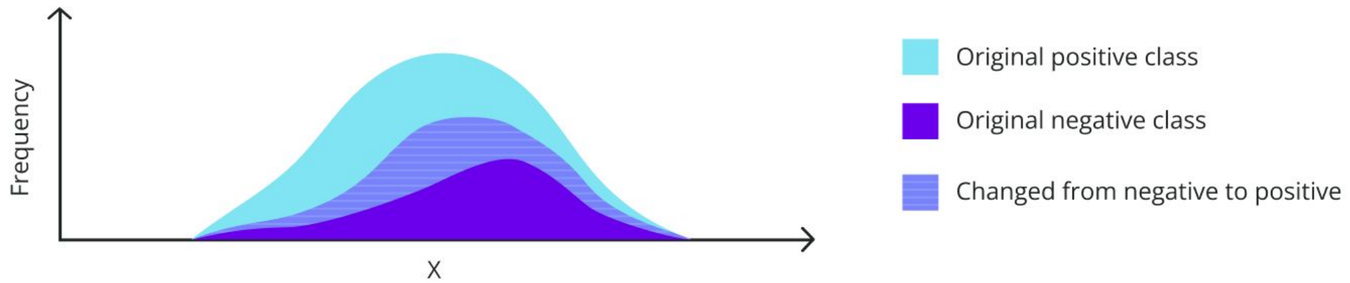
True pattern



Sampling (unchanged)

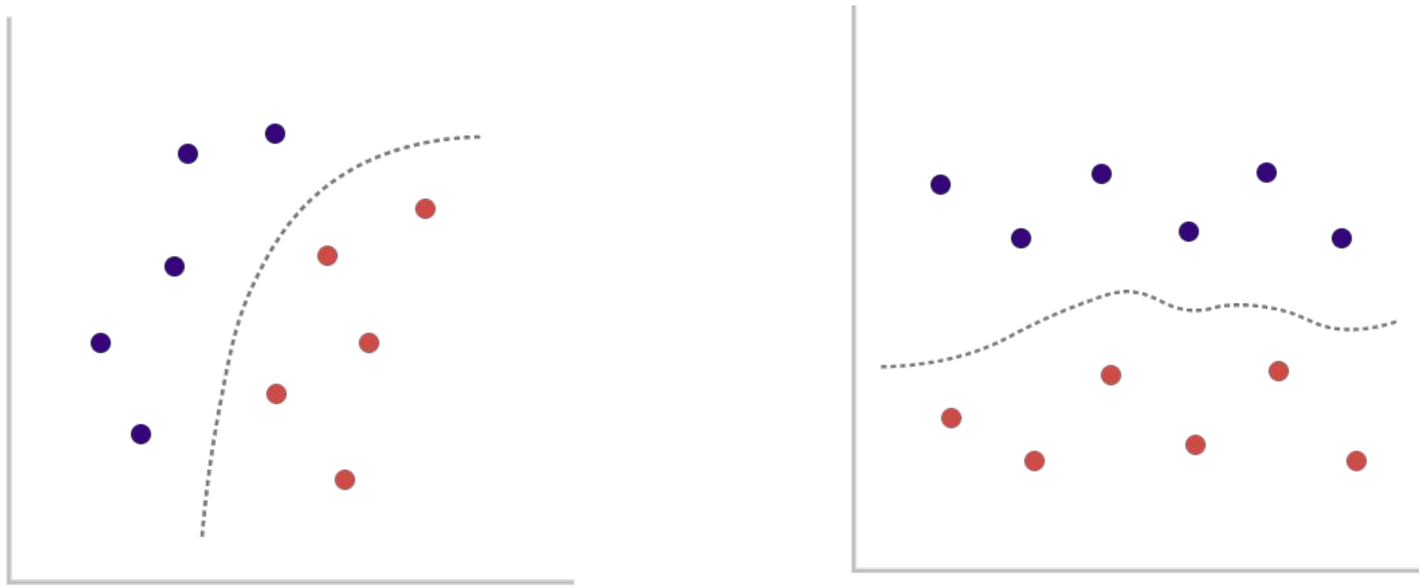


Data



Concept drift

Change in the underlying concept (pattern) between target and model inputs - P(y|x)



Training Data

Production Data

Performance matters.

Data drift != performance drop

Optimized in training

Business impact proxy

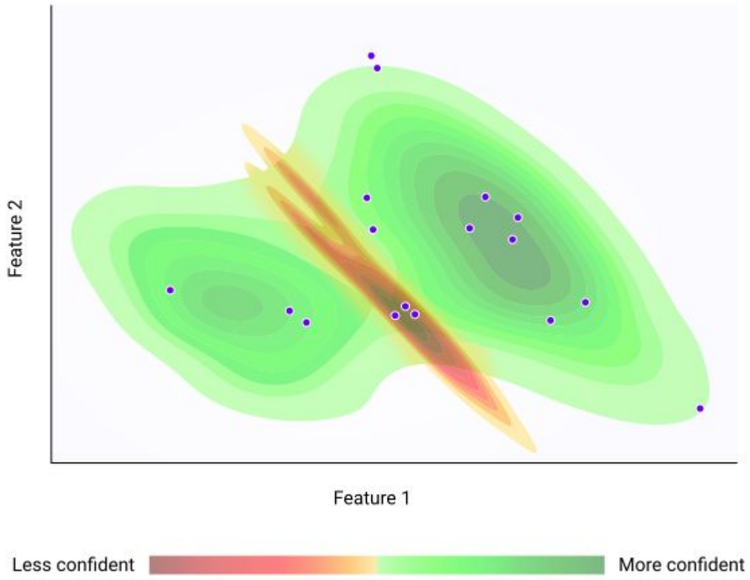
Do we Have Ground Truth? No

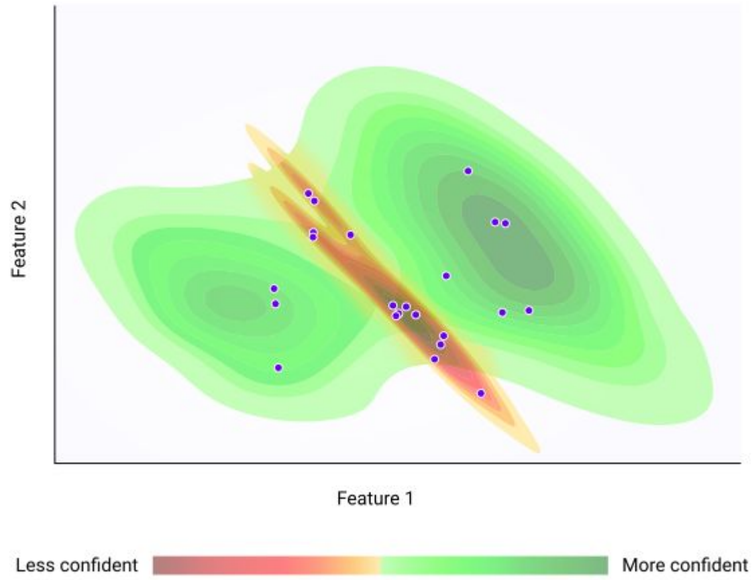
Delayed Data

Incomplete Labels

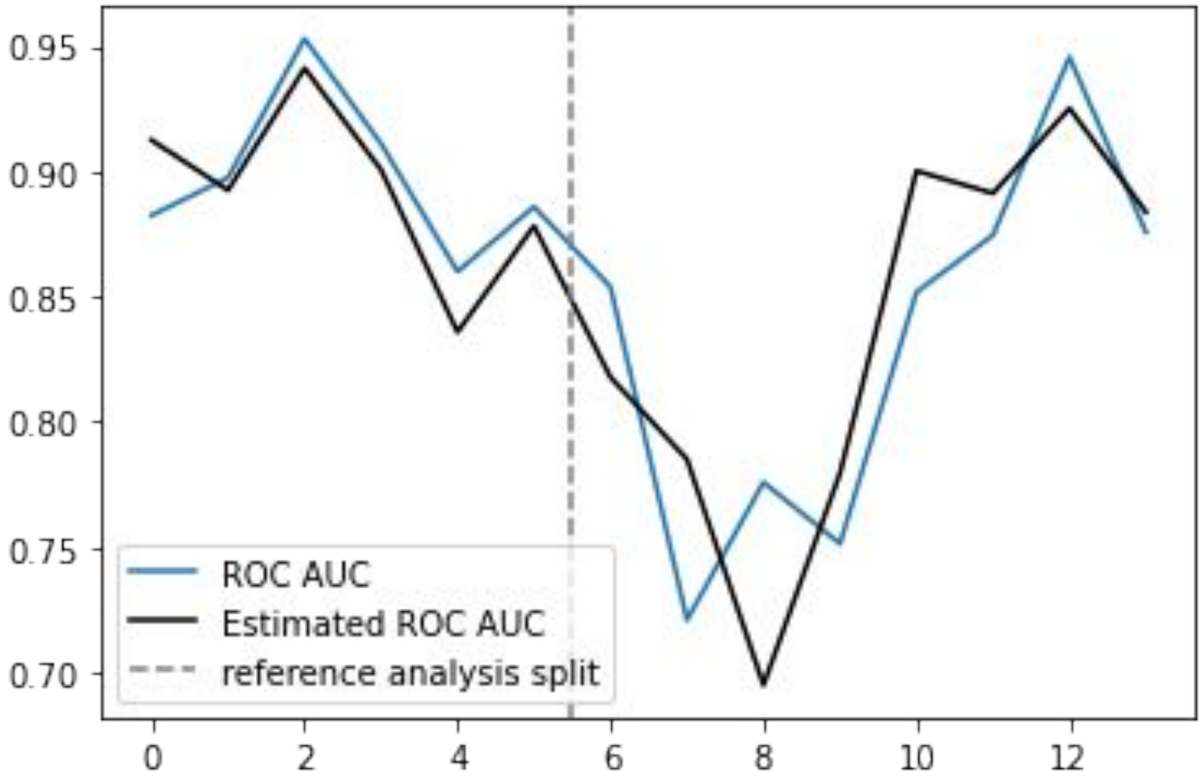
Automation Use Cases

Performance Estimation





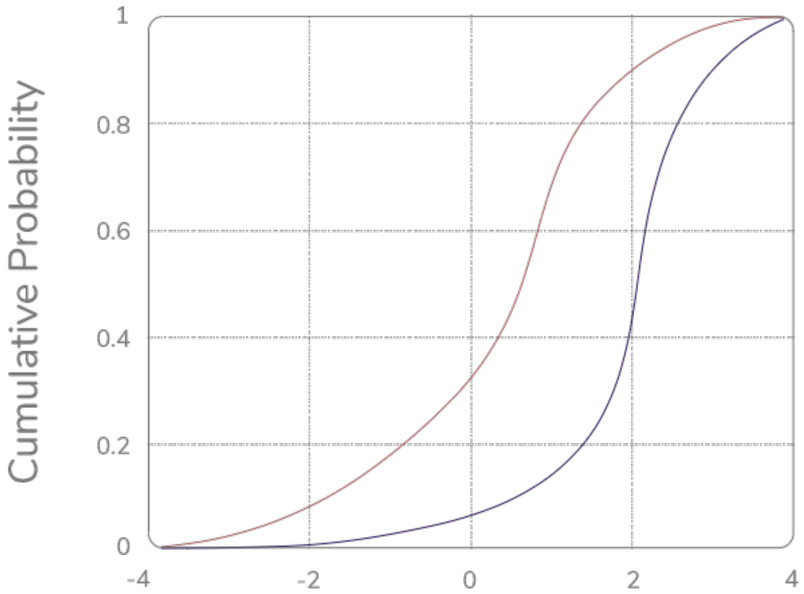
Performance Estimation - California Housing dataset



Data Drift Detection

Univariate

- K-S test
- Chi squared



Multivariate - data reconstruction

Dimensionality reduction + inverse transform

- PCA
- UMAP
- VAE

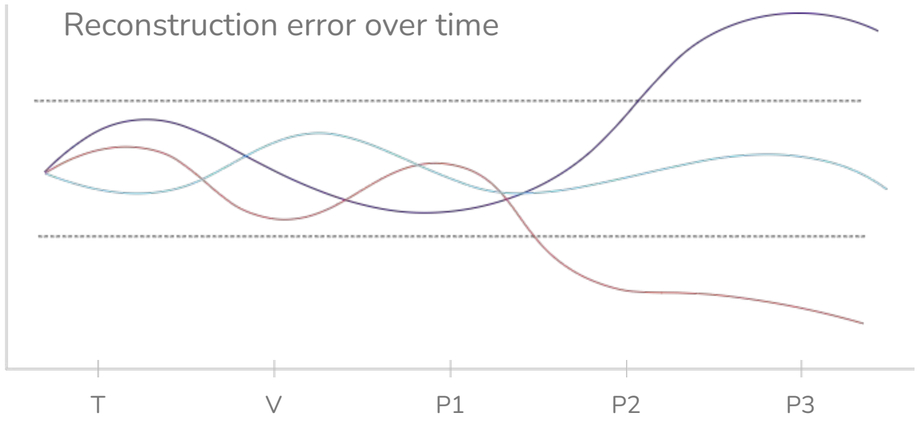
Data reconstruction

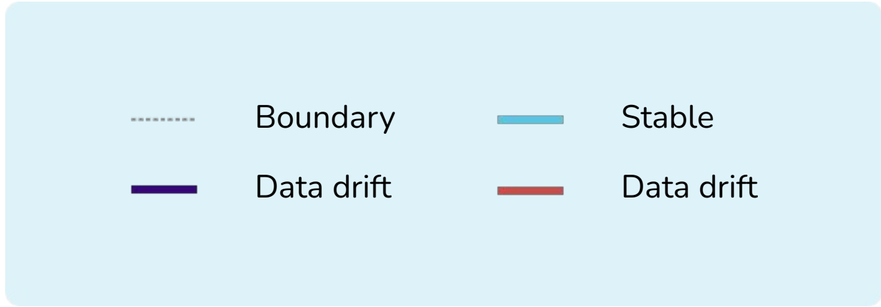
Requirements

- Encoding needs to learn the internal structure of the data
- Encoding needs to reduce the dimensionality of the data
- Inverse transformation needs to be possible
- The latent structure needs to map stably to original space

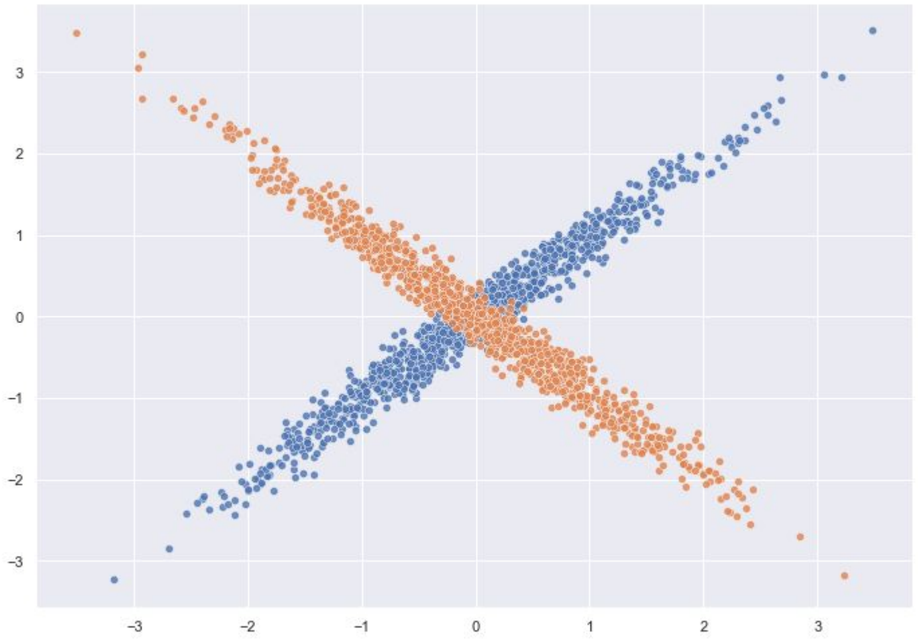
Reconstruction error

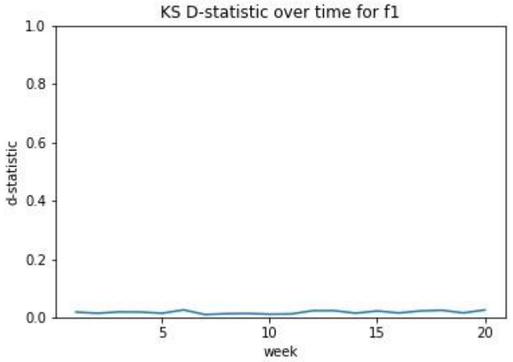
- Measure the dislocation of points before and after reconstruction
- Any distance metric could work
- RE = Mean($D_{Euclidean}$($P_{Original}$, $P_{Reconstructed}$))

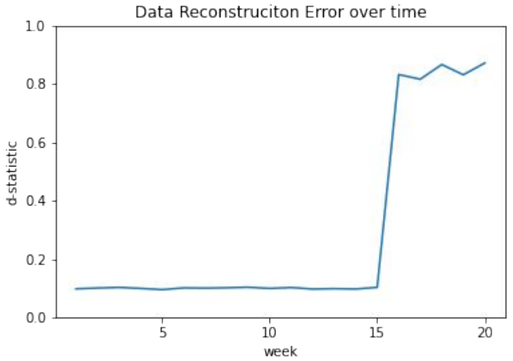




PCA Reconstruction error - examples







Monitoring Summary

Data drift does not always lead to drop in performance

Production targets are often not available to calculate performance

Performance estimation without target data is the key to ML monitoring

Thanks for Listening!

Would you Like to Learn More about Detecting Silent Model Failures?

https://www.nannyml.com

Let's talk! wojtek@nannyml.com

Or add me on LinkedIn:

https://www.linkedin.com/in/wojciech-kuberski

Check out our Github!

Github link:

- https://github.com/NannyML/nannyml

Documentation:

- https://docs.nannyml.com/

Blog:

- https://medium.com/nannyml/monitoring-as-a-first-step-to-observability-3776d9bd5829