Data Drift Detection

Machine Learning Operations (MLOps) Final Project

Course Instructor: Ishai Rosenberg

Reichman University

The Team

Guy Freund, Algorithm Developer at General Motors

Danielle Ben Bashat, Data Scientist at Elbit Systems

Elad Prager, Software Engineer at General Motors

Introduction

We were given two cases from the *financial services domain*:

Client A: Portuguese banking institution (Direct marketing campaigns)

- Classification: predict if a given bank's client will *subscribe to a term deposit*
- Business Goal: preserve and increase the *conversion rate over-time*

Client B: German banking institution (Credit risks)

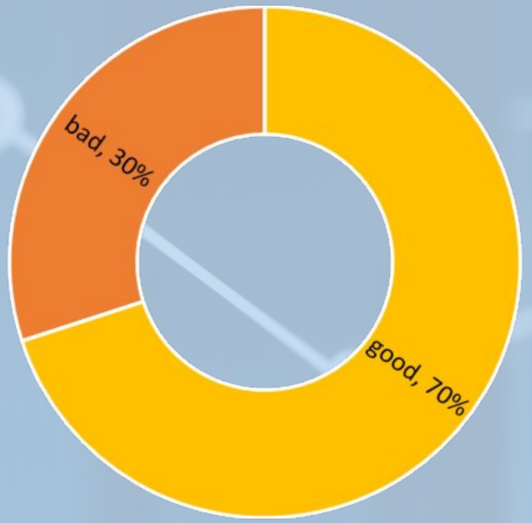
- Classification: predict if the risk of a given bank's client is *good or bad*
- Business Goal: *decrease the company's risk* when giving credits to clients

Datasets Overview

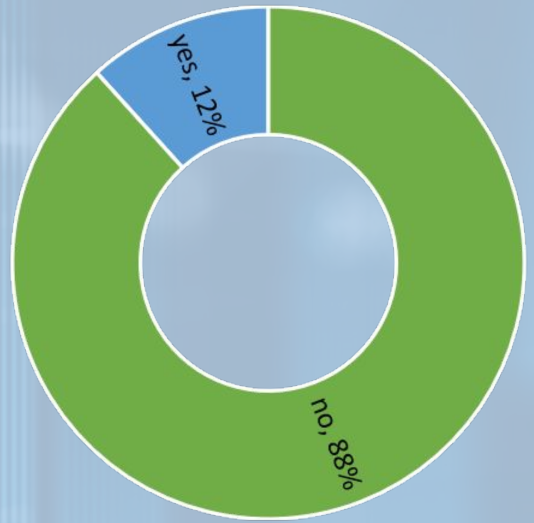
- For the portuguese banking institution we’re given:
  Tabular data with *45,211 observations*. Total of 16 features, where 9 are categorical and 7 are numerical
- For the German banking institution we’re given:
  Tabular data with *1,000 observations*. Total of 20 features, where 13 are categorical and 7 are numerical

Label Distribution

German Banking ‘y’ Label



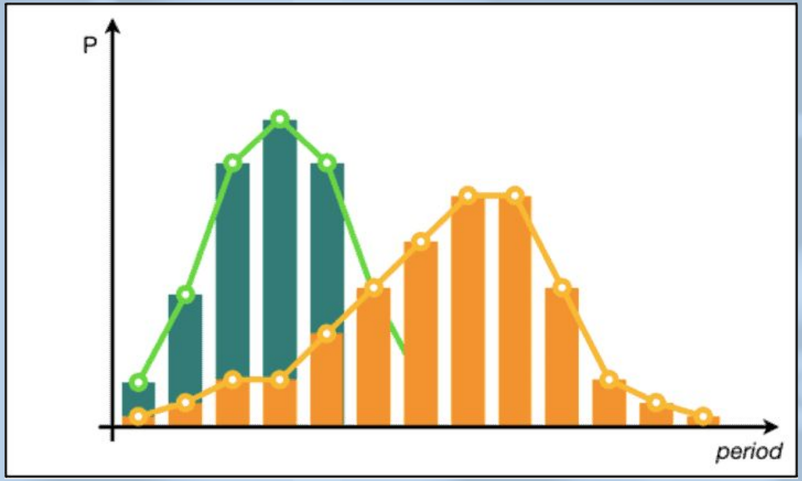
Portuguese Banking ‘y’ Label

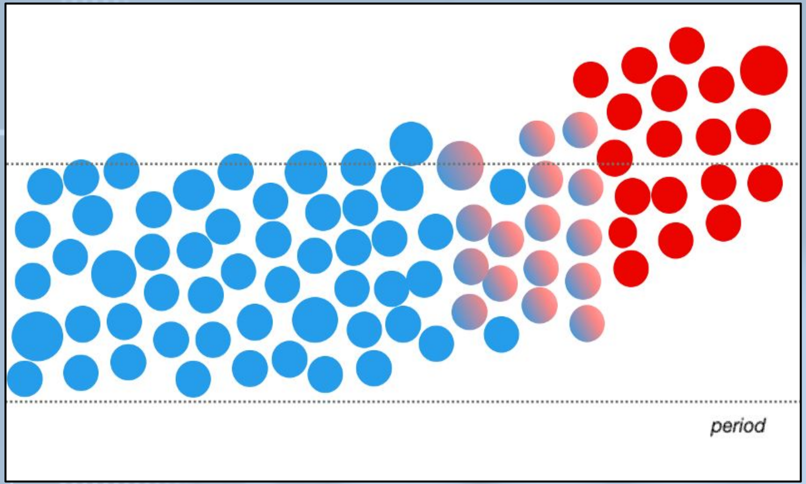


Data Drift Overview

- Over-time, *unexpected changes* to the real world data are likely to happen which might lead to a *performance degradation*.
- While designing a robust solution, we relate to the *time changing data*, which serves as our main goal in the project.

Data Drift Illustration



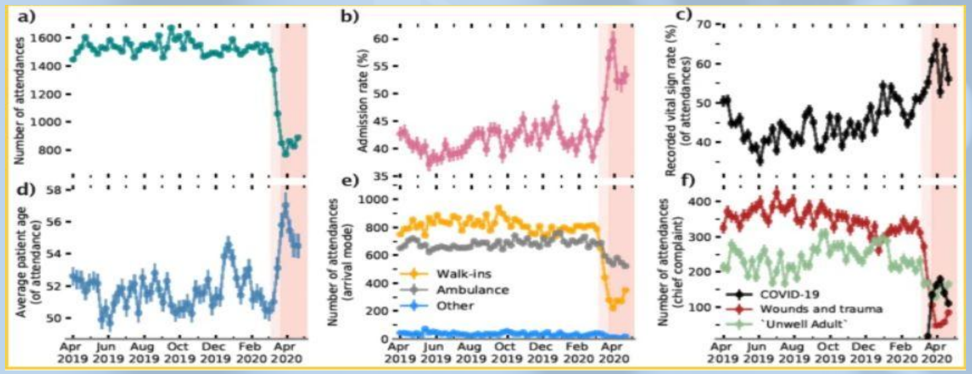


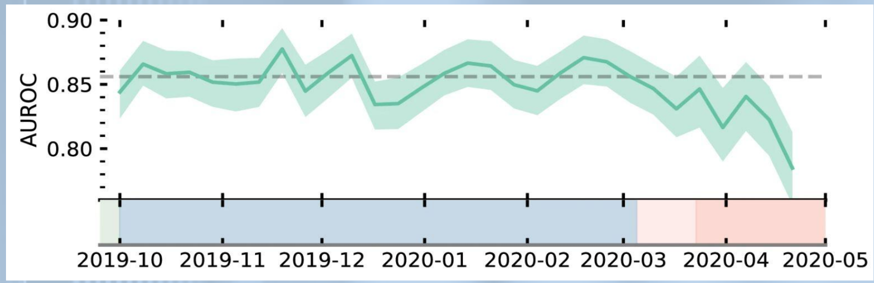
Data Drift Detection Motivation

- Provide useful information to stakeholders, such as: how to invest the *company's resources* effectively over-time, and as a result to boost the company *revenues and customer engagement* in the long term.
- Allow the business to run smoother during significant *worldwide events and crises*, when a formation of data drifts are more common.

Data Drift Real Case

- Using explainable ML to characterise data drift and identify *patients at high risk* of readmission to hospital at the point of attendance to an Emergency Department during COVID-19



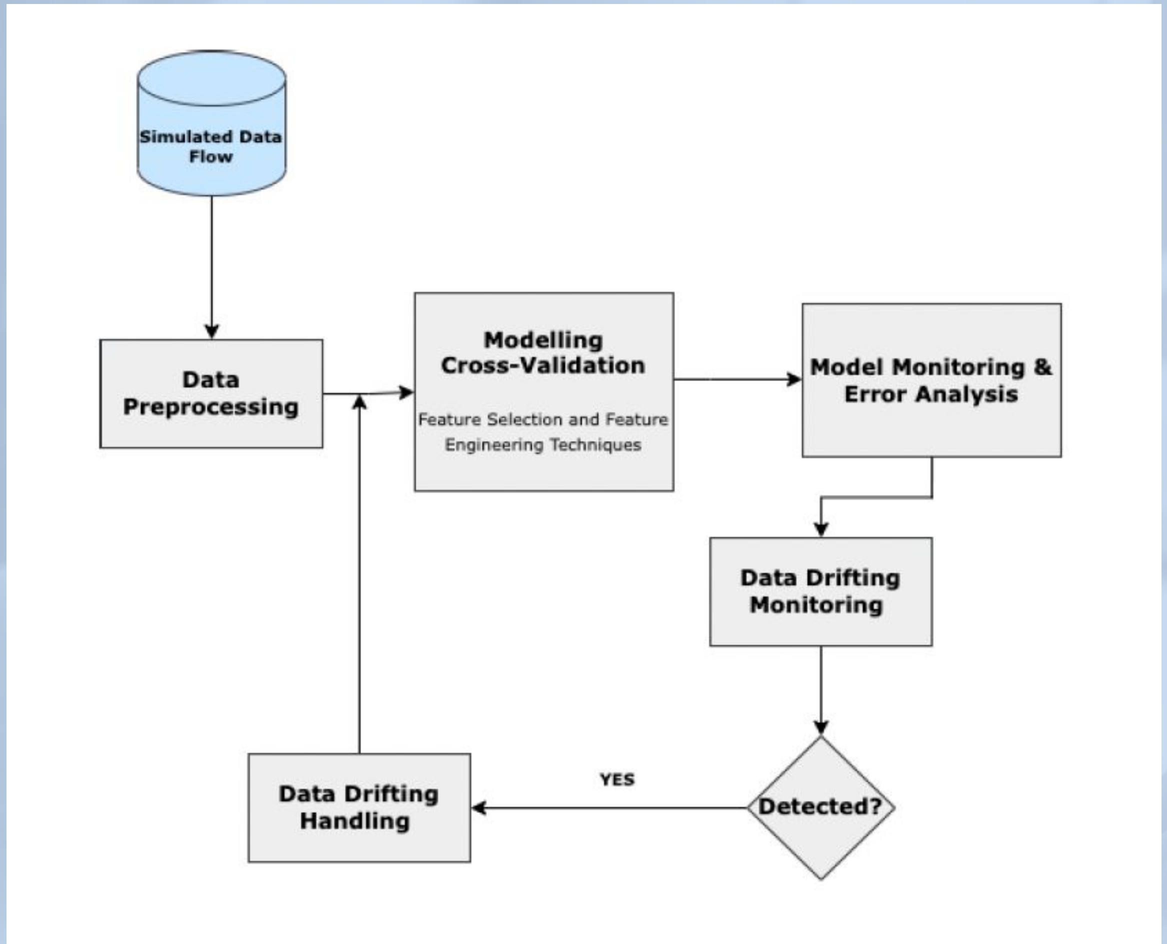


https://www.nature.com/articles/s41598-021-02481-y

Our Pipeline

- Preprocessing
- XGBoost Model
- Synthetic Data
- Data Drift Detection
- Re-train
- Evaluation

Diagram Flow



Preprocessing

- Handle Missing Data
- Data Ingestion
- Categorical Features Encoding
- Support Inverse Processing (decode the categorical features)
- Data Splitting
- Calculate and Save Feature Metrics: Mean, Variance, etc.

Modeling

- XGBoost
- Calculate Model Performance Metrics: Accuracy, F1, Recall, Precision, AUC
- Save Model as Pickle

Synthetic Data Generation

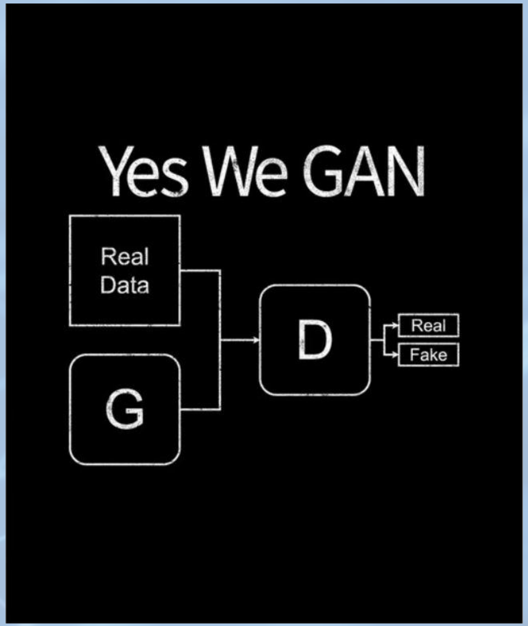
Generate Synthetic Data using:

- Conditional GAN (cGAN)
- SMOTE-NC

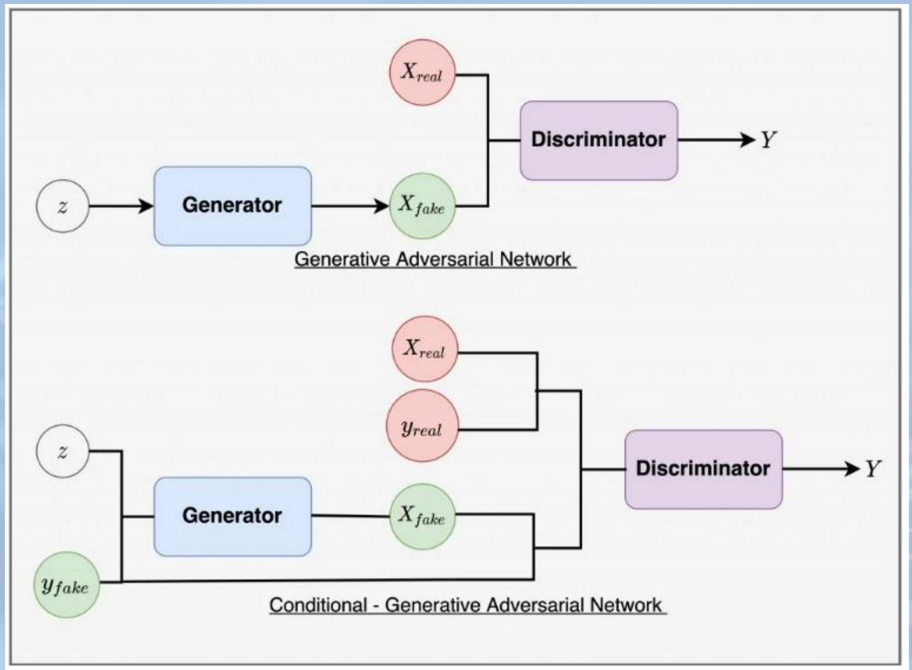
Synthetic Data Type:

- Create new unseen ‘normal’ data
- Create ‘drifted’ data

GAN Illustration



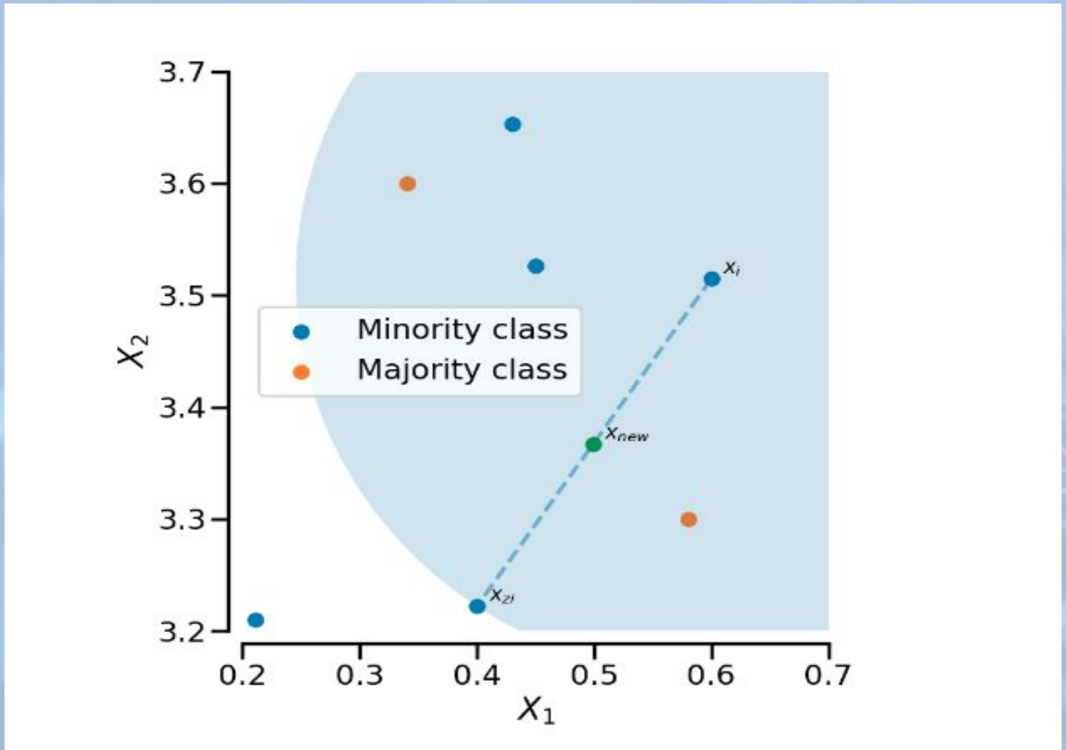
zoom in



Generative Adversarial Networks for synthetic data generation ©ydataai/ydata-synthetic

SMOTE-NC

Generating a new synthetic datapoint using SMOTE based on k-nearest neighbors.©imbalanced-learn



Synthetic Data Sanity Check

GAN Synthetic Data:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration
0	39	entrepreneur	single	secondary	no	1793	no	no	unknown	12	mar	76
1	39	retired	divorced	secondary	yes	1782	yes	yes	cellular	12	aug	75
2	39	housemaid	single	primary	no	1796	yes	yes	telephone	12	feb	76
3	39	housemaid	married	unknown	no	1785	yes	yes	telephone	12	jul	75
4	39	unknown	divorced	primary	yes	1798	yes	no	telephone	12	may	76
5	39	student	divorced	tertiary	yes	1800	no	yes	unknown	12	jun	76
6	39	student	married	secondary	yes	1799	no	yes	telephone	12	mar	76
7	39	admin.	divorced	primary	no	1779	no	no	cellular	12	jul	76
8	39	entrepreneur	married	primary	yes	1778	yes	no	telephone	12	apr	76

128 rows × 17 columns Open in new tab

```
dataset.raw_df
```

Original Data:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185
3	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199
4	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226
5	35	management	single	tertiary	no	747	no	no	cellular	23	feb	141
6	36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341
7	39	technician	married	secondary	no	147	yes	no	cellular	6	may	151
8	41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57

4521 rows × 17 columns Open in new tab

Data Generation

In each iteration of the pipeline:

- The generator samples with probability of 50% if to generate ‘drifted’ or ‘non-drifted’ data.
- Currently, we support two data drift types:
  - Statistical based drifting.
  - Number of nulls.

Data Drift Generation

Several parameters are sampled with range based configurations:

- The data drift **types.**
- **Number of features** to be drifted.
- The **percentage drift** of the statistical and null values.

For the numeric features: we transform the 'normal' synthesized data to a desired Mean and a Standard Deviation.

Suppose we start with feature x with mean μ_1 and non-zero std σ_1:
We do the following transformation:

- $x' = \mu_2 + (x - \mu_1) * \sigma_2 / \sigma_1$.
- Then, we get a new mean μ_2 with std σ_2

Data Drift Detection

- As part of the automatic step, we will manage data drift detection based on several known techniques
- Enable us to better understand the behavior pattern of our data over time

Data Drift Detection - Methodology

Decide on data drift based on:

- Statistical Based Approach
- Model Based Approach

Return a boolean result which is based on a (configured) weighted sum of the statistical and model-based detectors results

Data Drift Detection - Statistical Based Approach

- Calculate metrics such as: mean, variance, number of nulls for each feature in respect to its type (categorical or numeric).
- Compare the training feature metrics to the deployment feature metrics and decide heuristically whether a data drift has occurred.

Data Drift Detection - Model Based Approach

- Train a classifier on the concatenation of the training and the deployment datasets and test if the model is able to significantly differentiate between the sources. (Labels are: training and deployment)
- If the model accuracy is **not** as a coin-flip, then the data is **drifted**.

Data Drift Handling - Retraining The Model

- If a data drift was detected - we re-train the model on two sampled datasets concatenation.
- We sample the **training data** from the train set of the deployment model training phase and sample the **deployment data** and concatenate for a new dataset.

Evaluation

- Original Production Model Evaluation: on the **training dataset** vs the **deployment dataset**.
  - When data drift occured - expect to detect degradation in performance.
- Original Production Model vs Retrained Model Evaluation: on the new concatenated dataset (from sampled training dataset and deployment dataset)
  - When data drift occured - expect to detect increase in performance of the **retrained** production model.

Evaluation - Results

Original Deployment Model

		Accuracy	F1	Recall	Precision	AUC
GERMAN DATASET	Training Phase Dataset	76%	84%	90%	78%	0.65
	Deployment Phase Dataset*	70%	82%	90%	70%	0.5
PORTUGUESE DATASET	Training Phase Dataset	90%	47%	37%	65%	0.67
	Deployment Phase Dataset*	88%	12%	5%	60%	0.51

* is drifted = True

Evaluation - Results

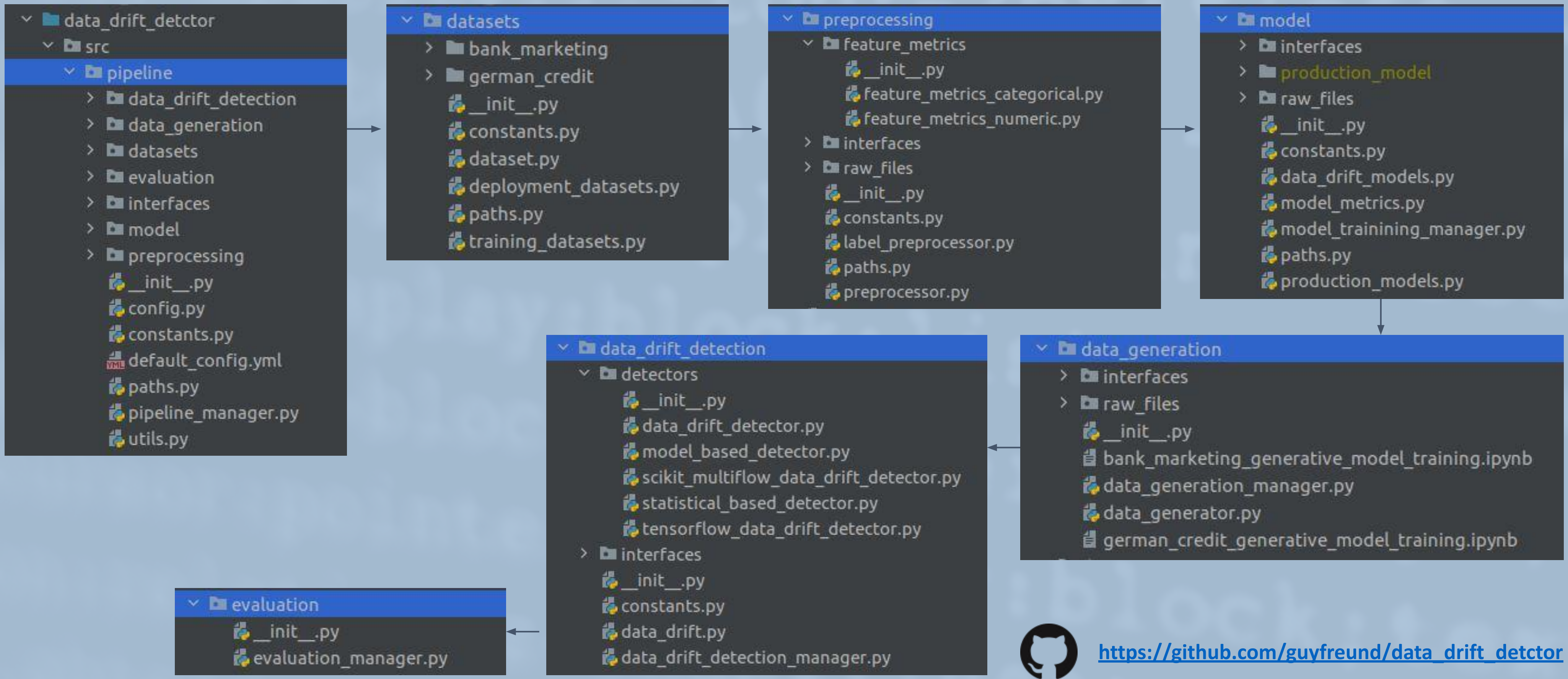
On The Concatenated Dataset

from sampled training dataset and deployment dataset phases

Dataset	Model	Accuracy	F1	Recall	Precision	AUC
GERMAN DATASET	Original Production Model	73%	84%	90%	72%	0.55
	Retrained Model*	76%	83%	88%	79%	0.67
PORTUGUESE DATASET	Original Production Model	89%	38%	27%	65%	0.62
	Retrained Model*	90%	45%	35%	62%	0.67

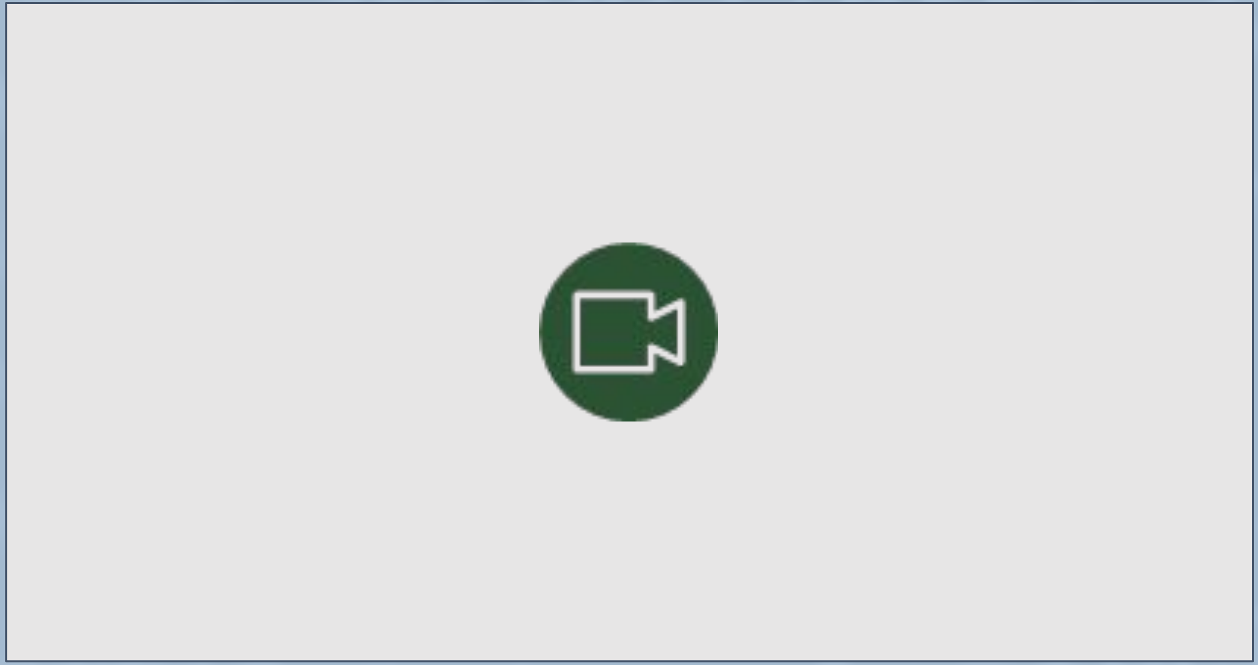
* is drifted = True

The Architecture and Code



https://github.com/guyfreund/data_drift_detctor

Live Demo



https://www.youtube.com/watch?v=YAIAqJdWYwU

Future Work

- Improve and fine tune the CGAN synthetic data generation model
- Improve data drift generation - detection, especially for categorical features: as new unseen categorical values, change in their distribution etc.
- Continuous development & integration
- Live metrics

Thank You