UNIVERSITÀ DEGLI STUDI DI TRIESTE

DISTRIBUTION SHIFT

A Study on Their Effects on Statistical Models and Strategies for Mitigation

Andrea Spinelli, Giacomo Amerio,
Giovanni Lucarelli, Tommaso Piscitelli

University of Trieste

Table of contents

Introduction

Dataset shift

- **Dataset shift** is a common problem in machine learning.
- It occurs when the distribution of the training data differs from the distribution of the test data.
- This can lead to a decrease in the performance of the model.

The two most common and well-studied causes of dataset shift are:

- Sample selection bias
- non stationary environments

Aims of project

This project aims to evaluate the impact of simple **Covariate shift** in the input distribution on the performance of robust models within the context of a synthetic binary classification task.

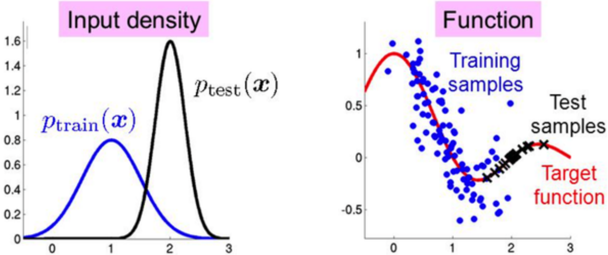
Key questions addressed in this study include:

- How do different types of covariate shifts affect the performance of robust models?
- Are certain models inherently more robust to simple covariate shifts?
- What strategies can be employed to improve model performance following such shifts?

Covariate shift

Can be formally defined as follows. Consider an input variable X and a response variable Y, where $X \to Y$ represents the relationship between the two. Let P_{tra} denote the probability distribution of the training data and P_{tst} denote the probability distribution of the test data. A covariate shift occurs when:

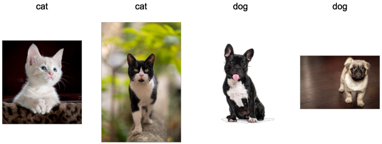
$$P_{\text{tra}}(Y \mid X) = P_{\text{tst}}(Y \mid X) \quad \text{but} \quad P_{\text{tra}}(X) \neq P_{\text{tst}}(X)\,.$$



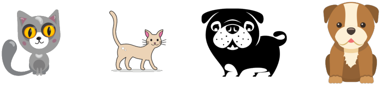
Example

Consider a model designed to distinguish between cats and dogs:

Training set:



Test set:



- Model will not accurately distinguish between cats and dogs because the feature distribution will differ.
- Changes in the input distribution can significantly impact the model's accuracy.

Inaccurate model

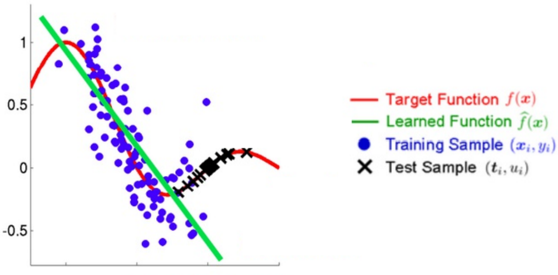


Figure 1: Example of inaccurate model.

In this study, we analyze the effects of distribution shift on different statistical models and propose strategies for its mitigation.

Data Generation

Training Dataset: Features

The dataset consists of $n = 10^4$ observations with 3 features and 1 binary target variable.

Training Dataset: Features

The dataset consists of $n = 10^4$ observations with 3 features and 1 binary target variable.

Features:

- $X_{\text{train}} = (X_{\text{train}\,1}, X_{\text{train}\,2}, X_{\text{train}\,3}) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{train}}, \boldsymbol{\Sigma}_{\text{train}})$
- $\mu_{\text{train}\,i} \sim \mathcal{U}_{[0,1]}$ for $i = 1, 2, 3$
- $[\boldsymbol{\Sigma}_{\text{train}}]_{i,j} \sim \mathcal{U}_{[-1,1]}$ for $i, j = 1, 2, 3$

<u>Note</u>: The $\boldsymbol{\Sigma}$ randomly generated has been transformed to a symmetric and positive semidefinite matrix by computing $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$.

Training Dataset: Target Variable

Building the **target variable** $Y \in \{0, 1\}$:

1. $$z = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \beta_{ii} x_i^2 + \sum_{i=1}^{2} \sum_{j=i+1}^{3} \beta_{ij} x_i x_j \,, \quad \beta_{\cdot} \sim \mathcal{U}_{[-1,1]}$$

2. $$p = \frac{1}{1 + e^{-z}}$$

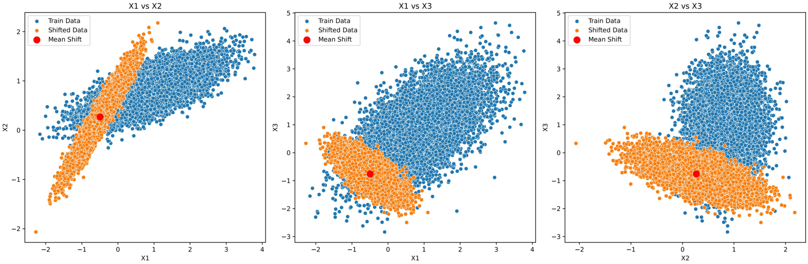
3. $$Y \sim \mathrm{Be}(p)$$

Testing Dataset

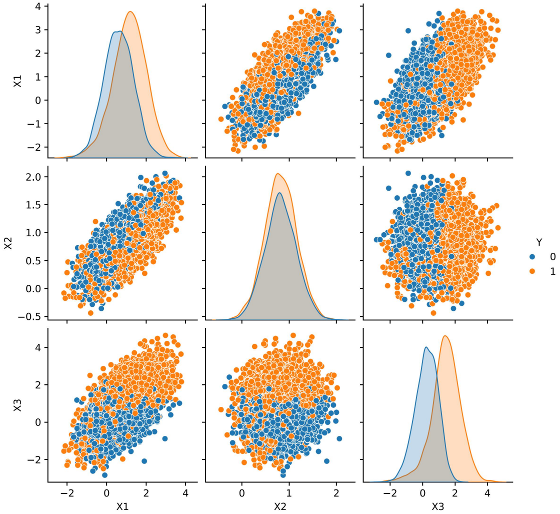
Same dataset structure as the train set, but:

- $X_{\text{shift}} = (X_{\text{shift}\,1}, X_{\text{shift}\,2}, X_{\text{shift}\,3}) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{shift}}, \boldsymbol{\Sigma}_{\text{shift}})$
- $\boldsymbol{\mu}_{\text{shift}} = \mathcal{Q}_{0.05}(X_{\text{train}})$
- $[\boldsymbol{\Sigma}_{\text{shift}}]_{i,j} \sim \mathcal{U}_{[-0.5,0.5]}$ for $i, j = 1, 2, 3$

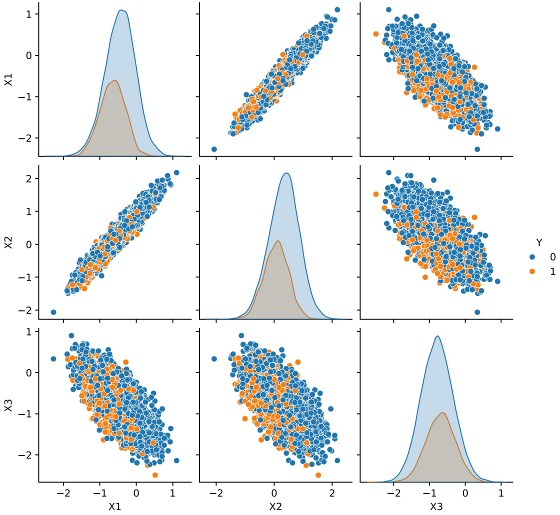
Original and Shiftd Features



Label Distribution in Train Set



Label Distribution in Shifted Test Set



Note: IR from 1.19 to 2.36

Testing Mixture

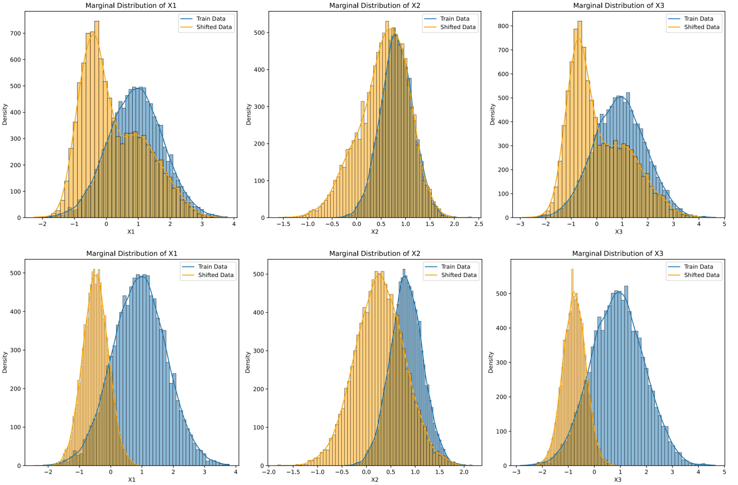
Series of datasets using **statistical mixtures** of the training features distribution and the fully shifted distribution.

$$X_\alpha \sim \alpha \cdot \mathcal{N}(\boldsymbol{\mu}_{\text{shift}}, \boldsymbol{\Sigma}_{\text{shift}}) + (1-\alpha) \cdot \mathcal{N}(\boldsymbol{\mu}_{\text{train}}, \boldsymbol{\Sigma}_{\text{train}})$$

$$\alpha \in \{0.0, 0.1, \ldots, 1.0\}$$

$$Y_\alpha \text{ generated as before}$$

Note: $X_{0.0}$ and X_{train} come from the same distribution, but the former are used as fresh new data.



Top: $\alpha = 0.5$. Bottom: $\alpha = 1.0$.

Performance Degradation

Performance Enhancement

A Novel Pre-Processing Method Against Covariate Shift

1. **No Prior Shift Knowledge Needed**
   - Simplifies implementation by eliminating the need for shift estimation.
   - Adaptable to various datasets without additional shift information.
2. **Built-in Regularization**
   - Prevents overfitting by introducing controlled noise.
   - Enhances model generalization on unseen data.

- Random
- Augmentation
- Walk

Input: $Data_{\text{train}}$, $Size$, N, ε.

$Data_{\%} \leftarrow$ random subset of N% of Data_{train}

For x_i in $\text{Data}_{\%}$

$$x_i' \leftarrow \begin{cases} X_i + \varepsilon & \text{with probability } 0.5 \\ X_i - \varepsilon & \text{with probability } 0.5 \end{cases}$$

$$y_i' \leftarrow y_i$$

End For

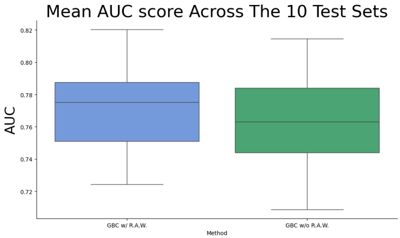
$Data_{\text{aug}} \leftarrow \text{Data}_{train} \cup Data_{\%}$

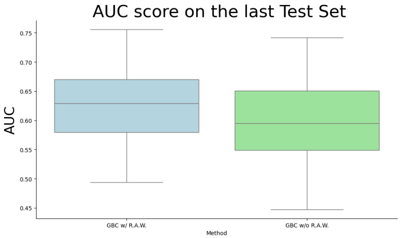
$Data_{\text{final}} \leftarrow \text{Downsample}(\text{Data}_{\text{aug}}, Size)$

Return $Data_{\text{final}}$

Classify With Gradient Boosting Using R.A.W.

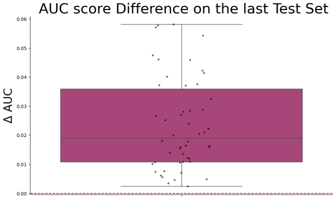
1. Apply the R.A.W. pre-processing method to the training data to address covariate shift.
2. Train a Gradient Boosting Classifier on the augmented dataset.
3. Evaluate the model's performance on shifted test sets.





A Statistical Analysis Of The Results

- H_0: $\Delta\bar{\mu} = \overline{\text{AUC}}_{\text{R.A.W.}} - \overline{\text{AUC}}_{\text{base}} = 0$
- H_1: $\Delta\bar{\mu} = \overline{\text{AUC}}_{\text{R.A.W.}} - \overline{\text{AUC}}_{\text{base}} \neq 0$
- **Test:** Paired t-test on 50 independent $\Delta\overline{\text{AUC}}$.



	$\Delta\bar{\mu}$	t-stat	p-value	95% CI
$\Delta\overline{\text{AUC}}$*	0.0083	8.75	1.39×10^{-11}	[0.006, 0.010]
$\Delta\overline{\text{AUC}}_{\text{last}}$**	0.0235	10.59	2.86×10^{-14}	[0.019, 0.028]

* Mean AUC score difference across all 10 shifted test sets.

** AUC score difference on the most shifted test set.

Questions?

References i