



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

1. Introduction
2. Data Generation
3. Performance Degradation
4. Performance Enhancement

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.
- It is related to another field of study "**transfer learning**"
- This can lead to a decrease in the performance of the model.

Most common cause of dataset shift

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

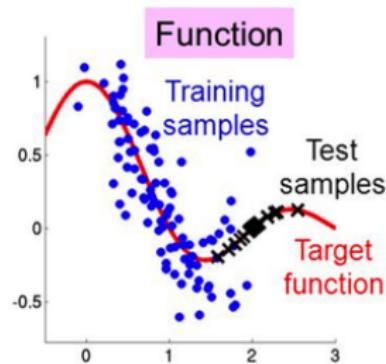
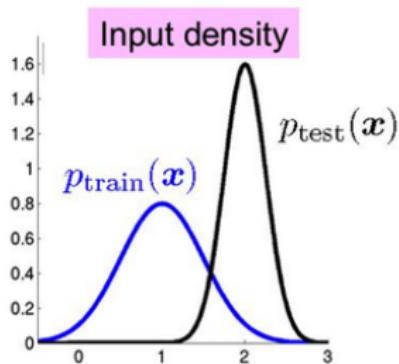
- Sample selection bias
- non stationary environments

In our project, we focused on one of the various forms of Dataset shift, **Covariate shift**, that is one of the most extensively researched.

Covariate shift

Can be formally defined as follows. Consider an input variable X and a response variable Y , where $X \rightarrow Y$ 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:

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



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

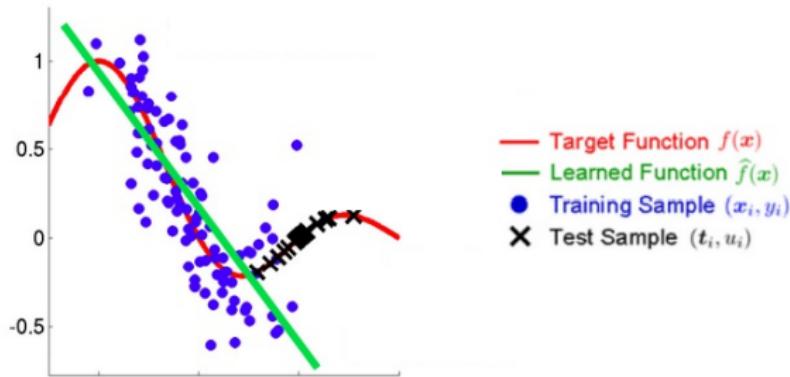


Figure 1: Example of inaccurate model.

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 target variable.

Training Dataset: Features

The dataset consists of $n = 10^4$ observations with 3 features and 1 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 $i, j = 1, 2, 3$

Note: 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.

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

2.

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

3.

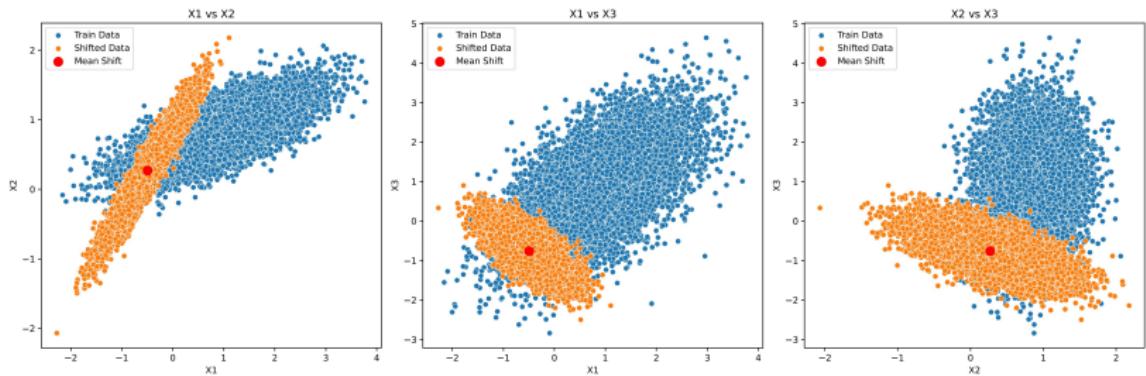
$$Y \sim \text{Be}(p)$$

Testing Dataset

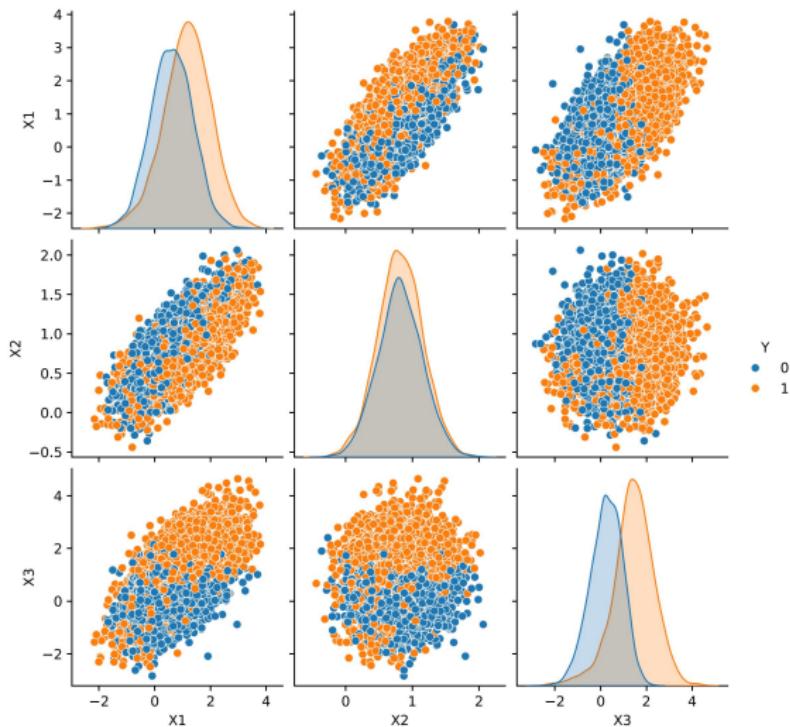
Same dataset structure as the train set, but:

- $X_{\text{shift}} = (X_{\text{shift}1}, X_{\text{shift}2}, X_{\text{shift}3}) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{shift}}, \boldsymbol{\Sigma}_{\text{shift}})$
- $\boldsymbol{\mu}_{\text{shift}} = \mathcal{Q}_{0.05}(X_{\text{train}})$
- $\boldsymbol{\Sigma}_{\text{shift}}$ generated in the same way as $\boldsymbol{\Sigma}_{\text{train}}$ (but $\boldsymbol{\Sigma}_{\text{shift}} \neq \boldsymbol{\Sigma}_{\text{train}}$)

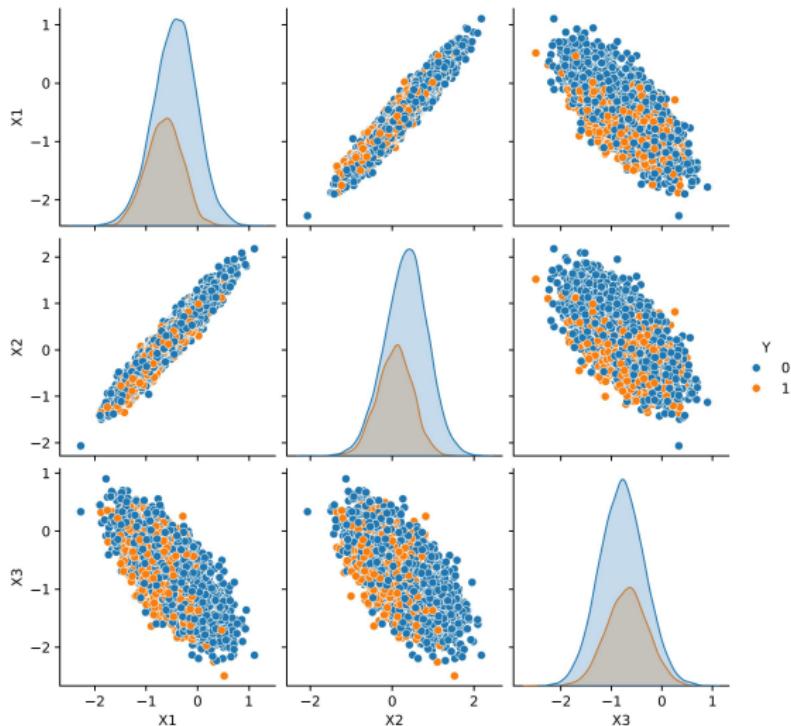
Original and Shifted Features



Label Distribution in Train Set



Label Distribution in Shifted Test Set



Testing Mixture

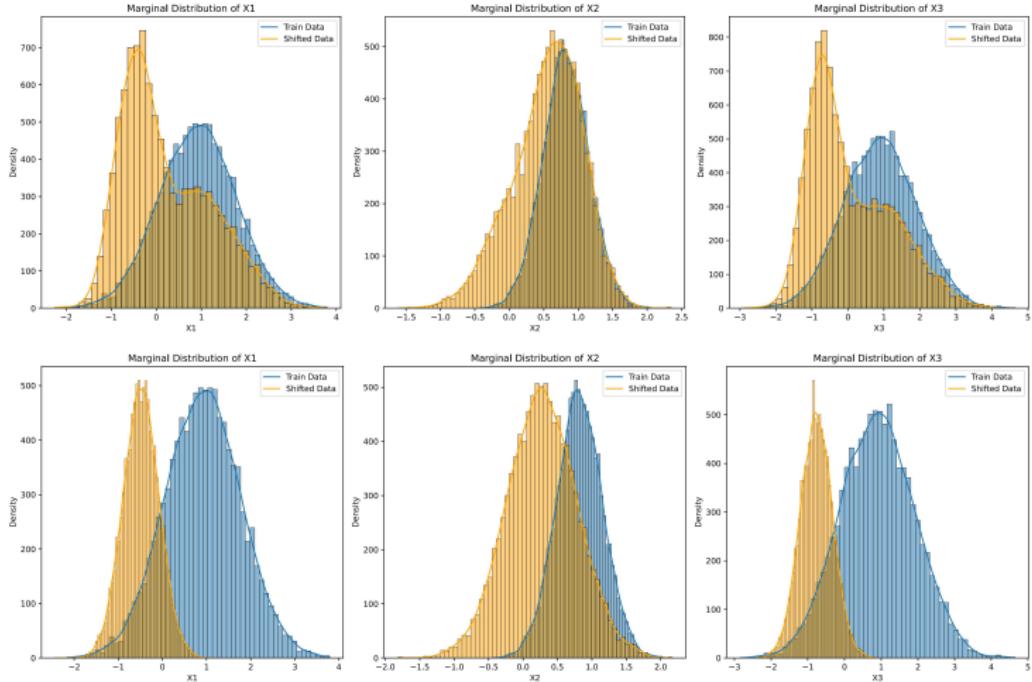
Series of datasets using **statistical mixtures** of the training features distribution and the fully shifted distribution.

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

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

Y_α 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

Questions?

References i