

Lab-2: Fairness using AIF360

Introduction

Machine learning models are increasingly used to inform high stakes decisions about people. Although machine learning, by its very nature, is always a form of statistical discrimination, the discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage. Biases in training data, due to either prejudice in labels or under-/over-sampling, yields models with unwanted bias.

AI Fairness 360 is an open source toolkit, developed by IBM Research, that can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. The goal of this lab is to introduce the toolkit's bias detection capabilities and to show how to help mitigate the bias in a machine learning model using Adversarial Debiasing technique.

For more information see links below:

AIF360 Demo: <https://aif360.mybluemix.net>

AIF360 GitHub: <https://github.com/IBM/AIF360>

AIF360 API Docs: <https://aif360.readthedocs.io/en/latest/>

Objectives

In this notebook you will utilize AIF360 to detect and mitigate bias on the Compas dataset which is used to assess the likelihood that a criminal defendant will reoffend.

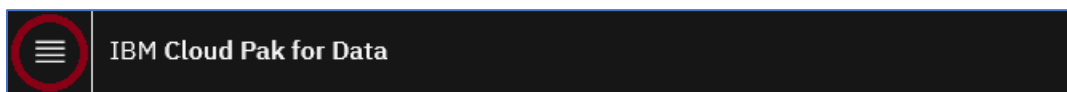
Upon completing this lab, you will learn:

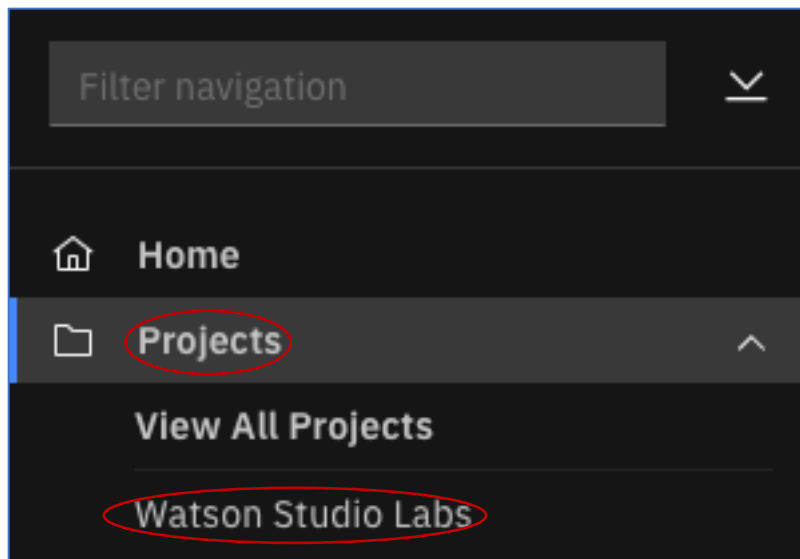
1. How to load datasets from the toolkit package
2. Check the dataset for bias
3. Mitigate existing bias by using Adversarial Debiasing technique
4. Train on both original and corrected dataset and compare results

Lab Steps

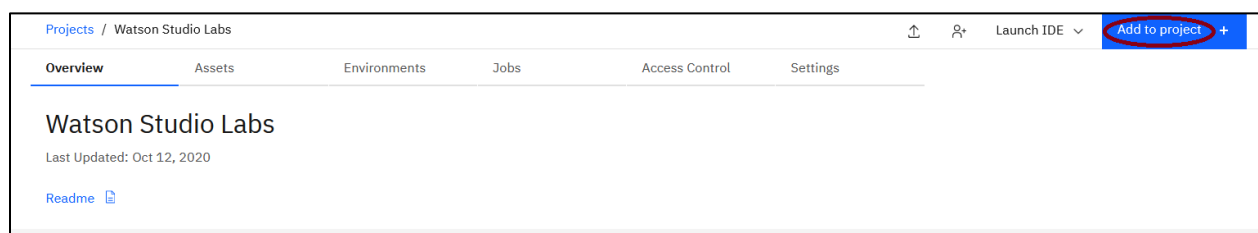
Step 1 - Create a Jupyter Notebook

1. Click on the hamburger icon , then click on **Projects**, and then **Watson Studio Labs** (or whatever you named the project)

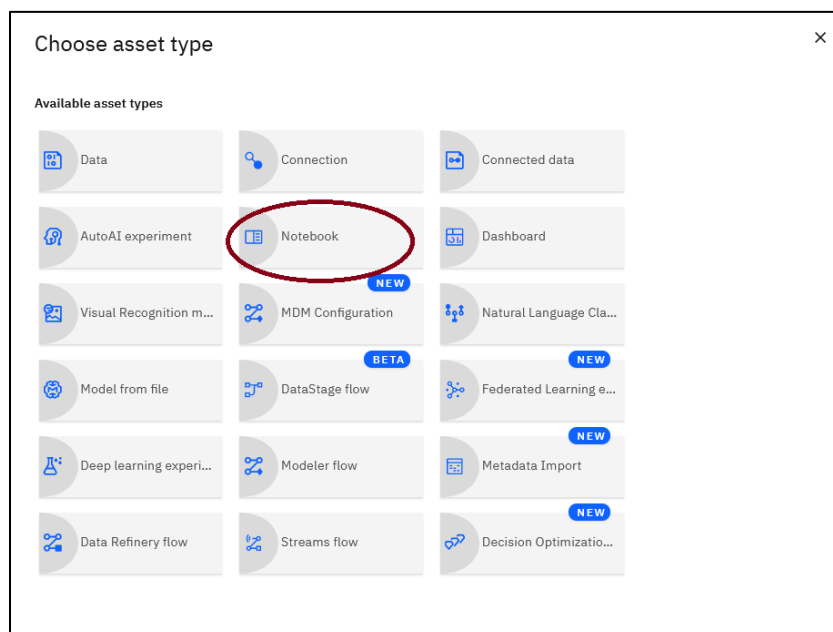




2. We are now going to create a notebook in our project. This notebook will be created from a url that points to the AIF notebook in the github repository. Click the **Add to project** link.



3. Click on **Notebook**



Click on **From URL** under **New Notebook**, enter **AIF360 Demo** for the **Name**, optionally enter a **Description**, leave the default for the **runtime**, and copy and paste the following url into the **Notebook URL** field.

https://github.com/bleonardb3/TR_POT_01-28-2021/blob/main/Lab-2/AIF360%20Demo.ipynb

Click **Create**.

Blank From file **From URL**

Name: AIF 360 Demo

Description (optional): Type your description here

Select runtime: Default Python 3.7 XS (2 vCPU 8 GB RAM)

The selected runtime has 2 vCPU and 8 GB RAM. It consumes 1 capacity unit per hour. [Learn more](#) about capacity unit hours and Watson Studio pricing plans.

Notebook URL: https://github.com/bleonardb3/TR_POT_01-28-2021/blob/main/Lab-1/AIF360%20Demo.ipynb

Cancel Create

4. Place the cursor in the first documentation cell.

My projects / Watson Studio Labs / AIF Demo

File Edit View Insert Cell Kernel Help

Not Trusted | Python 3.6

Detecting and Mitigating Bias Using AI Fairness 360

Using "Adversarial Debiasing"

Introduction

Machine learning models are increasingly used to inform high stakes decisions about people. Although machine learning, by its very nature, is always a form of statistical discrimination, the discrimination becomes objectionable when it places certain privileged groups at systematic advantage and certain unprivileged groups at systematic disadvantage. Biases in training data, due to either prejudice in labels or under-/over-sampling, yields models with unwanted bias.

The AI Fairness 360 Python package includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models. The AI Fairness 360 interactive demo provides a gentle introduction to the concepts and capabilities. The tutorials and other notebooks offer a deeper, data scientist-oriented introduction. The complete API is also available.

For more information see links below:

- AIF360 Demo: <https://aif360.mybluemix.net>
- AIF360 GitHub: <https://github.com/IBM/AIF360>
- AIF360 API Docs: <https://aif360.readthedocs.io/en/latest/>

5. Execute the code cells in the notebook. For those unfamiliar with Jupyter notebooks, read below.

A Jupyter notebook consists of a series of cells. These cells are of 2 types (1) documentation cells containing markdown, and (2) code cells (denoted by a bracket on

the left of the cell) where you write Python code, R, or Scala code depending on the type of notebook. Code cells can be run by putting the cursor in the code cell and pressing **<Shift><Enter>** on the keyboard. Alternatively, you can execute the cells by clicking on **Run icon** on the menu bar that will run the current cell (where the cursor is located) and then select the cell below. In this way, repeatedly clicking on **Run** executes all the cells in the notebook. When a code cell is executed the brackets on the left change to an asterisk ***** to indicate the code cell is executing. When completed, a sequence number appears. The output, if any, is displayed below the code cell.