Feature Analysis Using TensorFlow Data Validation and Facets

2 hours · Free · Rate Lab

Overview

Bias can manifest in any part of a typical machine learning pipeline, from an unrepresentative dataset, to learned model representations, to the way in which the results are presented to the user. Errors that result from this bias can disproportionately impact some users more than others.

TensorFlow Data Validation (TFDV) is one tool you can use to analyze your data to find potential problems in your data, such as missing values and data imbalances - that can lead to Fairness disparities. The TFDV tool analyzes training and serving data to compute descriptive statistics, infer a schema, and detect data anomalies. Facets Overview provides a succinct visualization of these statistics for easy browsing. Both the TFDV and Facets are tools that are part of the Fairness Indicators.

In this lab, we use TFDV to compute descriptive statistics that provide a quick overview of the data in terms of the features that are present and the shapes of their value distributions. We use Facets Overview to visualize these statistics using the Civil Comments dataset.

Learning Objectives

- Use TFRecords to load record-oriented binary format data
- Use TFDV to generate statistics and Facets to visualize the data
- Use the TFDV widget to answer questions
- Analyze label distribution for subset groups

Setup

For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Make sure you signed into Qwiklabs using an **incognito window**.

2. Note the lab's access time (for example, 02:00:00) and make sure you can finish in that time block.

> There is no pause feature. You can restart if needed, but you have to start at the beginning.

3. When ready, click START LAB.

4. Note your lab credentials. You will use them to sign in to the Google Cloud Console.

5. Click **Open Google Console**.

6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.

> If you use other credentials, you'll get errors or **incur charges**.

7. Accept the terms and skip the recovery resource page.

> Do not click **End Lab** unless you are finished with the lab or want to restart it. This clears your work and removes the project.

Launch AI Platform Notebooks

To launch AI Platform Notebooks:

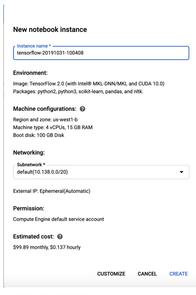
Step 1

Click on the Navigation Menu. Navigate to AI Platforms, then to Notebooks.

Step 2

On the Notebook instances page, click + NEW INSTANCE. Select TensorFlow 2.x without GPUs.

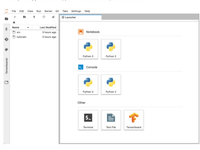
In the pop-up, confirm the name of the deep learning VM, move to the bottom of the window and click Create.



The new VM will take 2-3 minutes to start.

Step 3

Click **Open JupyterLab**. A JupyterLab window will open in a new tab.

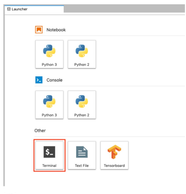


Clone course repo within your AI Platform Notebooks instance

To clone the `training-data-analyst` notebook in your JupyterLab instance:

Step 1

In JupyterLab, click the Terminal icon to open a new terminal.



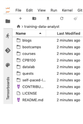
Step 2

At the command line prompt, type in the following command and press Enter.

```
git clone https://github.com/GoogleCloudPlatform/training-data-analyst
```

Step 3

Confirm that you have cloned the repository by double clicking on the `training-data-analyst` directory and ensuring that you can see its contents. The files for all the *Jupyter* notebook-based labs throughout this course are available in this directory.



Feature Analysis Using TensorFlow Data Validation and Facets

Step 1

In the notebook interface, navigate to **training-data-analyst > courses > machine_learning > deepdive2 > introduction_to_tensorflow > labs** and open **adv_tfdv_facets.ipynb**.

Step 2

In the notebook interface, click on **Edit > Clear All Outputs** (click on Edit, then in the drop-down menu, select Clear All Outputs).

Carefully read through the notebook instructions and fill in lines marked with #TODO where you need to complete the code as needed

Tip: To run the current cell you can click the cell and hit **shift+enter**. Other cell commands are found in the notebook UI under **Run**.

- Hints may also be provided for the tasks to guide you along. Highlight the text to read the hints (they are in white text).
- If you need more help, you may take a look at the complete solution by navigating to **training-data-analyst > courses > machine_learning > deepdive2 > introduction_to_tensorflow > solutions** and open **adv_tfdv_facets.ipynb**.

End your lab

When you have completed your lab, click **End Lab**. Qwiklabs removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the **Support** tab.