Improving Data Quality

45 minutesFree

Overview

Machine learning models can only consume numeric data, and that numeric data should be *1*s or *0*s. Data is said to be *messy* or *untidy* if it is missing attribute values, contains noise or outliers, has duplicates, wrong data, or upper/lower case column names, or is essentially not ready for ingestion by a machine learning algorithm.

In this lab, you will present and solve some of the most common issues of *untidy* data. Note that different problems will require different methods, and they are beyond the scope of this notebook.

What you learn

In this lab, you will:

- Resolve missing values.
- Convert the Date feature column to a datetime format.

- Rename a feature column, remove a value from a feature column.
- Create one-hot encoding features.
- Understand temporal feature conversions.

Setup

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

1. Sign in to Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example, 02:00:00), and make sure you can finish within that time. 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 (**Username** and **Password**). 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 receive errors or **incur charges**.
7. Accept the terms and skip the recovery resource page.

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

Set up your environment

Enable the Vertex AI API

1. In the Google Cloud Console, on the **Navigation menu**, click **Vertex AI**.

2. Click **Enable Vertex AI API**.

Launch Vertex AI Notebooks instance

1. In the Google Cloud Console, on the **Navigation Menu**, click **Vertex AI > Workbench**.

2. On the Notebook instances page, click **New Notebook > TensorFlow Enterprise > TensorFlow Enterprise 2.6 (with LTS) > Without GPUs**.

3. In the **New notebook** instance dialog, confirm the name of the deep learning VM, if you don't want to change the region and zone, leave all settings as they are and then click **Create**. The new VM will take 2-3 minutes to start.

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

5. You will see "Build recommended" pop up, click **Build**. If you see the build failed, ignore it.

Clone course repo within your Vertex AI Notebooks instance

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

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

2. At the command-line prompt, run the following command:

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

   Copied!

   content_copy

3. To confirm that you have cloned the repository, double-click on the training-data-analyst directory and ensure that you can see its contents. The files for all the Jupyter notebook-based labs throughout this course are available in this directory.

Improving Data Quality

Step 1

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

Step 2

In the notebook interface, click **Edit > 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.

Tips

- To run the current cell, click the cell and press **SHIFT+ENTER**. Other cell commands are listed 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, look at the complete solution by navigating to **training-data-analyst > courses > machine_learning > deepdive2 > launching_into_ml > solutions**, and open **improve_data_quality.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.