

ML Model Deployment (Cloud)

Week 5: Deployment on Cloud

Batch code: LISUM31, Submission date: Mar 9th, 2024 , Submitted to Data Glacier



by Xiyuan Wu


Table of contents

- [Continued Progress](#)
- [File Needed](#)
- [Deploy Model](#)
- [Example Output](#)
- [Thank You](#)

Continued Progress

Last week, we successfully deployed the model locally, and this week is a continuation of that progress as we will deploy the model to the cloud. You can choose any cloud service you want; I will choose Google Cloud Platform (GCP) for this project.

If you haven't seen last week's slides (make sure to read last week's slides first) and files from last week, you can find them here:

 github.com



Data Science Intern: Week 4

Approach

1

Get File Ready

We will use the files from last week, and we will need new files this week: Dockerfile, .dockerignore, and requirements.txt file.

2

Setup Google Cloud

We will start by creating an account, setting up a project, and properly configuring a billing account.

3

Cloud Build & Deploy

After we have everything set up, we can start building and deploying our model!

File Needed

We will use the files from last week (model.py, model.pkl, app.py, and index.html). In addition to those, we will need some new files:

Requirements.txt

requirements.txt is a text file in Python projects that lists all of the package dependencies the project needs to run successfully. It is used with pip, the Python package installer, to automatically install all the required libraries and their specific versions with a single command, ensuring consistency across different development and production environments. We can use following code:

```
pip install -r requirements.txt
```

Docker File

We need two files for Docker: one is Dockerfile and the other one is .dockerignore.

These should be the same for different models, unless you change the environment. In most cases, we can obtain both files' code from Google's official documentation, though you may need to make some small changes if you have different file names and function names. Note that the .dockerignore file is often blocked by GitHub, so I've included its code in the README file. You can find the code in Google's official documentation:



 Google Cloud

Quickstart: Deploy a Python service to Cloud Run | Cloud ...

Learn how to deploy a service to Cloud Run using Python and a container image.



Now, we have all the files we need. You can also find all the files under the week 5 folder:

 github.com

Data Science Intern: Week 5



Deploy Model

1 Setup Google Cloud


As I mentioned before, I will use GCP to deploy this model. We will start by creating a GCP account (if you don't have one). After creating an account, create a new project. After creating a new project, click on the dashboard to navigate to the dashboard.

Make sure to set up a billing account; this step is very important! Add your billing account to your project. You do have a 90-day free trial and \$300 credits for new accounts. If you already have an account and have passed the free trial period, you will need to add your credit card. But don't worry, it won't charge you before the model deployment. After deploying the model, you can close public access so it won't charge you. Remember, they won't charge you if you don't use it.

2 Install API & Download App

In the project's dashboard, we need to make sure we have installed both the Cloud Run API and the Cloud Build API. Install them if you haven't done so yet.

After installing the APIs, we need to download the Google Cloud SDK. Navigate to the provided link, choose your operating system, download the app, and follow the instructions on the screen.



Google Cloud

Install the gcloud CLI | Google Cloud CLI Documentati...

This page contains instructions for choosing and maintaining a Google Cloud CLI installation. The Google Cloud CLI includes t...

3 Build & Deploy

By now, we are ready to deploy our model! The last step is to build and deploy. Make sure you have downloaded all the files to your local PC, and then open the terminal by typing cmd. Type the following two commands:

```
gcloud builds submit --tag gcr.io/<project_id>/<function_name>
gcloud run deploy --image gcr.io/<project_id>/<function_name> --platform managed
```

The first command is for building the model, and the second command is for deploying the model. Replace project_id with your project ID, and function_name with your function name (the function name you defined in app.py).

- When building the model, if your dataset/model is large, it will take longer to process.


4 Success

If you see a link pop up, congratulations! You have successfully deployed your model on GCP. Before you open the link, you can go back to the dashboard and click on Cloud Run, where you will see your model has been deployed successfully.

<input type="checkbox"/>		Name	Req/sec	Region	Authentication	Ingress	Recommendation	Last deployed
<input type="checkbox"/>		getprediction	0.01	us-west1	Allow unauthenticated	All		2 hours ago

- After the model is deployed, by default, the link is open to the public. If you want to avoid charges, make sure to remove all user permissions.

- This is a rough and general overview of the steps. For detailed instructions, you can watch this video:



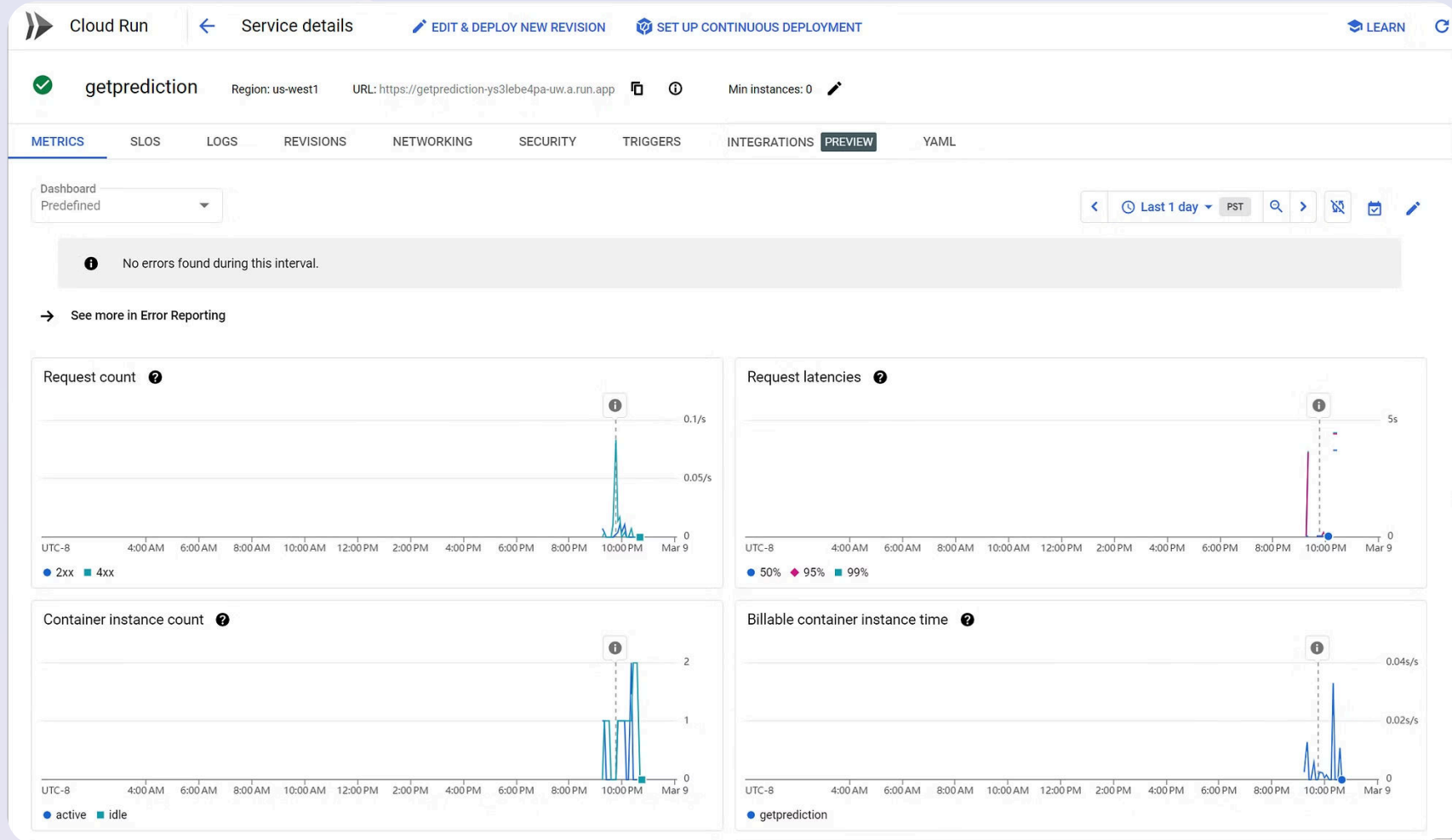
YouTube

How To Deploy ML Models With Google Cloud Run

Learn how to deploy Machine Learning / Deep Learning models with Google Cloud Run. We build a simple app with TensorFlo...

Example Output

Here's what my dashboard looks like after successfully deploying the model (I've already removed public access, so the link won't work on your side):



Thank You