ML Model Deployment

[**Background**](#_danhcxty3zcd) **1**

[**High level details**](#_dulxsdukpsj2) **1**

[Available Deployment Options](#_bjv9gs2crq57) 1

# Background

This document provides details on the ML Model deployment details for Vertex AI and Document AI. Along with the deployment details it lists the other available options for the deployment and pros/cons of each.

# High level details

The ML Model deployment in ICICI would be performed based on the retraining activities for both Vertex AI and Document AI. Currently, the microservices API fetches the Model version ID for each request and maintains the provenance record in the CloudSql Database.

## Available Deployment Options

The availability of the Model deployment and serving varies for Vertex AI and Document AI, following are the comparisons for both:

* Vertex AI
  + Supports majority of the options, ie., Gcloud, Python SDK, Rest, etc
* Document AI
  + Limited Support, ie, only Rest calls

It is recommended that we build a single script which can accommodate both Vertex AI and Document AI deployments, and based on the above variation, **Python is recommended for below reasons as compared to a Shell script:**

| **Python** | **Shell Script** |
| --- | --- |
| Better support for REST calls, Document AI currently only supports Rest api invocations. | Difficult to embed Rest calls, it would require curl and manipulations to achieve the same goal |
| Supports ADC(app default credentials) out of the box for Rest API invocations | It will have to run print-access-token to fetch the token, and when run on a GCE, it may have additional ‘....’ which may need truncation |
| Better user experience and the flexibility around options on Vertex and Document AI deployments | Similar user experience however the code readability might be little difficult(as it may perform deployment for Vertex and Document AI both) based on Shell knowledge of the user/deployer |
| Extension of the python script would be easier | Bash logics are relatively difficult to extend with future requirements |
| Requires Python and packages to be installed | Requires Gcloud CLI which internally has a dependency on Python |