# Playbook: Deployment and Monitoring Guide

## Deployment Steps:

### 1. Clone Repository:

Clone the GitHub repository containing the source code for both the Training and Prediction systems.

https://github.com/Hackathon2024-March/voicedetectives/

### 2. Set Up Google Cloud Project:

Create a new project in Google Cloud Platform (GCP) or use an existing one.

Enable necessary APIs such as Cloud Run, Container Registry, and Cloud Storage.

### 3. Build Docker Images:

Build Docker images for both the Training and Prediction systems.

Tag the images appropriately.

```
gcloud builds submit --tag gcr.io/wells-fargo-genai24-8354/voice-inspectors
```

### 4. Push Images to Google Container Registry (GCR):

Push the built Docker images to Google Container Registry using gcloud CLI.

Ensure proper permissions and authentication for pushing images.

### 5. Deploy Training System:

Deploy the Training system as a Docker container on Google Cloud Run.

Specify required environment variables such as data storage location, credentials, etc.

### 6. Deploy Prediction System:

Deploy the Prediction system as a Docker container on Google Cloud Run.

Configure environment variables including model locations, API endpoints, etc.

## Monitoring Guide:

Monitor CPU and memory utilization of deployed containers in Google Cloud Run.

Set up alerts for abnormal resource consumption patterns.

### 1. Request Metrics Monitoring:

Monitor incoming requests to both Training and Prediction systems.

Track request latency, error rates, and throughput.

### 2. Model Performance Monitoring:

Monitor the performance of trained models in the Prediction system.

Track accuracy, precision, recall, and other relevant metrics.

### 3. Logging and Error Tracking:

Monitor logs generated by both Training and Prediction systems.

Set up log-based metrics and create alerts for critical errors.

### 4. Alerting and Notification:

Configure Stackdriver alerts to notify administrators of system failures or anomalies.

Define escalation policies for handling alerts and incidents.

### 5. Continuous Improvement:

Regularly review monitoring data and logs to identify areas for improvement.

Update models, configurations, or infrastructure based on monitoring insights.

## Post-Deployment Tasks:

### 1. Documentation and Knowledge Sharing:

Document deployment procedures, configurations, and monitoring setup.

Share knowledge with the team to ensure proper maintenance and troubleshooting.

### 2. Backup and Disaster Recovery:

Implement backup strategies for critical data and configurations.

Test disaster recovery procedures periodically to ensure resilience.

### 3. Security Auditing and Compliance:

Conduct regular security audits and vulnerability assessments.

Ensure compliance with relevant regulations and standards.

### 4. Scaling and Optimization:

Monitor system performance and scale resources as needed to handle increased load.

Optimize resource utilization and costs by rightsizing instances and configurations.

By following this playbook, you can effectively deploy and monitor the solution on Google Cloud Platform, ensuring reliability, performance, and scalability of both the Training and Prediction systems.