Lecture 8

Model deployment

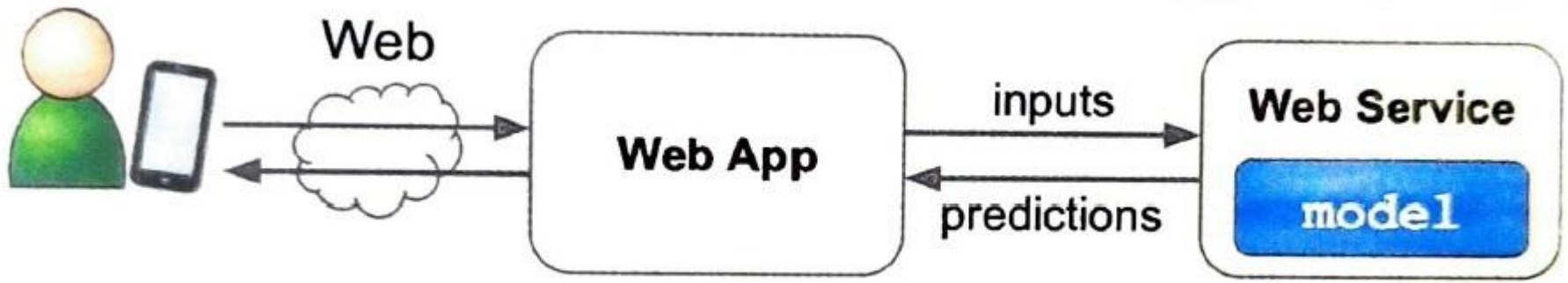
Suman Samui

Launch, Monitor, and Maintain your System

- Reached the final stage → Got approval to launch
- Let's polish the code, write documentation and test
- Deploy the final model to your production environment

Model deployment as web service

- Save scikit-learn model including the full preprocessing and prediction pipeline
- Load the trained model within your production environment
- Make prediction by calling predict( ) method → Inference



Model Deployment in cloud

- Save your model using joblib and upload it on Google Cloud Storage
- It takes JSON requests containing the input data and returns JSON responses containing predictions
- Deploying the tensorflow model is also quite similar

What next: Monitoring and maintenance

- Evaluate your model's live performance at regular intervals and trigger alerts when it drops.

- We can automate a few things:

1. Collect fresh data regularly and label it
2. Write a script to train the model (in regular interval)
3. Evaluate both new and previously saved model