This course is part of the **Machine Learning Engineering for Production (MLOps) Specialization**

Deploying Machine Learning Models in Production

4.7 78 ratings • 12 reviews

Laurence Moroney +1 more instructor

Offered By

DeepLearning.AI

Go To Course Already enrolled

6,593 already enrolled

About this Course

89,669 recent views

In the fourth course of Machine Learning Engineering for Production Specialization, you will learn how to deploy ML models and make them available to end-users. You will build scalable and reliable hardware infrastructure to deliver inference requests both in real-time and batch depending on the use case. You will also implement workflow automation and progressive delivery that complies with current MLOps practices to keep your production system running. Additionally, you will continuously monitor your system to detect model decay, remediate performance drops, and avoid system failures so it can continuously operate at all times.

Understanding machine learning and deep learning concepts is essential, but if you're looking to build an effective AI career, you need production engineering capabilities as well. Machine learning engineering for production combines the foundational concepts of machine learning with the functional expertise of modern software development and engineering roles to help you develop production-ready skills.

Week 1: Model Serving Introduction
Week 2: Model Serving Patterns and Infrastructures
Week 3: Model Management and Delivery
Week 4: Model Monitoring and Logging

SKILLS YOU WILL GAIN

TensorFlow Serving | Model Monitoring | Model Registries | Machine Learning Operations (MLOps) | Generate Data Protection Regulation (GDPR)

Flexible deadlines
Reset deadlines in accordance to your schedule.

Shareable Certificate
Earn a Certificate upon completion

100% online
Start instantly and learn at your own schedule.

Course 4 of 4 in the
Machine Learning Engineering for Production (MLOps) Specialization

Advanced Level
• Some knowledge of AI / deep learning
• Intermediate Python skills
• Experience with any deep learning framework (PyTorch, Keras, or TensorFlow)

Approx. 33 hours to complete

English
Subtitles: English

Instructors

Instructor rating 4.65/5 (24 Ratings)

Laurence Moroney
Instructor
Lead AI Advocate, Google
338,299 Learners
15 Courses

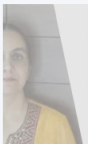
Robert Crowe
Instructor
TensorFlow Developer Engineer, Google
16,539 Learners
3 Courses

Offered by

DeepLearning.AI

DeepLearning.AI is an education technology company that develops a global community of AI talent.

DeepLearning.AIs expert-led educational experiences provide AI practitioners and non-technical professionals with the necessary tools to go all the way from foundational basics to advanced application, empowering them to build an AI-powered future.



I've been leading a happier life since I discovered Coursera. The courses and knowledge helped me become more comfortable and confident.

— *Sonal T.*

Coursera's rigorous assignments and broad range of subjects encourage me to keep up with my courses. The quality of the teachers keeps me coming back.

— *Sandra O.*

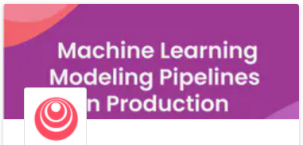
With Coursera, I learned the fundamenta to transition to my current career.

Related Courses | Degrees | Recommended

Other courses in this Specialization

Introduction to Machine Learning in Production
DeepLearning.AI
1 COURSE

Machine Learning Data Lifecycle in Production
DeepLearning.AI
1 COURSE



Machine Learning Modeling Pipelines in Production
DeepLearning.AI
1 COURSE

Syllabus - What you will learn from this course

WEEK 1

5 hours to complete

Week 1: Model Serving: Introduction

Learn how to make your ML model available to end-users and optimize the inference process

7 videos (Total 34 min), 5 readings, 5 quizzes SEE LESS

7 videos

Course Overview 4m

Introduction to Model Serving 6m

Introduction to Model Serving Infrastructure 5m

Deployment Options 3m

Improving Prediction Latency and Reducing Resource Costs 5m

Creating and deploying models to AI Prediction Platform 2m

Installing TensorFlow Serving 6m

5 readings

Ungraded Labs - Best Practices 5m

Ungraded Lab - Introduction to Docker 20m

Optional: Build, train, and deploy an XGBoost model on Cloud AI Platform 45m

Ungraded Lab - Tensorflow Serving with Docker 20m

Ungraded Lab - Serve a model with TensorFlow Serving 30m

3 practice exercises

Introduction to Model Serving 30m

Introduction to Model Serving Infrastructure 30m

TensorFlow Serving 30m

WEEK 2

10 hours to complete

2

Week 2: Model Serving: Patterns and Infrastructure

Learn how to serve models and deliver batch and real-time inference results by building scalable and reliable infrastructure

8 videos (Total 44 min), 8 readings, 8 quizzes **SEE LESS**

8 videos

Model Serving Architecture 4m

Model Servers: TensorFlow Serving 3m

Model Servers: Other Providers 5m

Scaling Infrastructure 10m

Online Inference 6m

Data Preprocessing 4m

Batch Inference Scenarios 5m

Batch Processing with ETL 3m

8 readings

Documentation on model servers 10m

Ungraded Lab - Deploy a ML model with FastAPI and Docker 1h

Learn about scaling with boy bands 10m

Explore Kubernetes and KubeFlow 10m

Ungraded Lab: Intro to Kubernetes 1h 10m

Ungraded Lab - Latency testing with Docker Compose and Locust 45m

Data preprocessing 10m

Ungraded Lab (Optional): Machine Learning with Apache Beam and TensorFlow 45m

6 practice exercises

Model serving architecture 30m

Scaling Infrastructure 30m

Online Inference 30m

Data Preprocessing 30m

Batch inference scenarios 30m

Batch Processing with ETL 30m

WEEK

3

12 hours to complete

Week 3: Model Management and Delivery

Learn how to implement ML processes, pipelines, and workflow automation that adhere to modern MLOps practices, which will allow you to manage and audit your projects during their entire lifecycle

9 videos (Total 82 min), 10 readings, 6 quizzes **SEE LESS**

9 videos

Experiment Tracking 8m

Tools for Experiment Tracking 7m

Introduction to MLOps 11m

MLOps Level 0 5m

MLOps Levels 1&2 13m

Developing Components for an Orchestrated Workflow 13m

Managing Model Versions 7m

Continuous Delivery 7m

Progressive Delivery 7m

10 readings

Experiment Tracking 10m

MLOps Resources 10m

Ungraded Lab: Intro to Kubeflow Pipelines 2h

Architecture for MLOps using TFX, Kubeflow Pipelines, and Cloud Build 10m

Ungraded Lab: Developing TFX Custom Components 45m

Ungraded Lab - Model Versioning with TF Serving 40m

ML Model Management 10m

Ungraded Lab - CI/CD pipelines with GitHub Actions 1h

Continuous Delivery 10m

Progressive Delivery 10m

3 practice exercises

ML Experiments Management and Workflow Automation 30m

MLOps Methodology 30m

Model Management and Deployment Infrastructure 30m

WEEK

4

6 hours to complete

Week 4: Model Monitoring and Logging

Establish procedures to detect model decay and prevent reduced accuracy in a continuously operating production system

13 videos (Total 66 min), 7 readings, 5 quizzes SEE LESS

13 videos

Why Monitoring Matters 6m

Observability in ML 4m

Monitoring Targets in ML 4m

Logging for ML Monitoring 7m

Tracing for ML Systems 3m

What is Model Decay? 3m

Model Decay Detection 2m

Ways to Mitigate Model Decay 5m

Responsible AI 5m

Legal Requirements for Secure and Private AI 9m

Anonymization and Pseudonymisation 4m

Right to be Forgotten 8m

Specialization recap and farewell 1m

7 readings

Monitoring Machine Learning Models in Production 10m

Addressing Model Decay 10m

Responsible AI 10m

GDPR and CCPA 10m

Course 4 Optional References 10m

Acknowledgements 10m

Join our DeepLearning.AI Community! 10m

3 practice exercises

Model Monitoring and Logging 30m

Model Decay 30m

GDPR and Privacy 30m

Reviews

4.7 ★★★★★ 12 reviews

5 stars 76.19%
4 stars 20.23%
3 stars 1.19%
2 stars 2.38%

TOP REVIEWS FROM DEPLOYING MACHINE LEARNING MODELS IN PRODUCTION

★★★★★
by WH Sep 11, 2021

The most practical course for junior MLOPs engineers looking for the best productionization methodologie, and the tools that implement them.

★★★★☆
by PR Oct 6, 2021

Awesome course with very good instructors . However in instructions in graded google cloud labs could be improved.

View all reviews

About the Machine Learning Engineering for Production (MLOps) Specialization

Understanding machine learning and deep learning concepts is essential, but if you're looking to build an effective AI career, you need production engineering capabilities as well.

Effectively deploying machine learning models requires competencies more commonly found in technical fields such as software engineering and DevOps. Machine learning engineering for production combines

SHOW ALL

Start Learning Today

- This Course Plus the Full Specialization
- Shareable Certificates
- Self-Paced Learning Option
- Course Videos & Readings
- Practice Quizzes
- Graded Assignments with Peer Feedback
- Graded Quizzes with Feedback
- Graded Programming Assignments

Go To Course

6,593 already enrolled

Shareable on LinkedIn

You can share your Course Certificates in the Certifications section of your LinkedIn profile, on printed resumes, CVs, or other documents.

Frequently Asked Questions

- When will I have access to the lectures and assignments?
- What will I get if I subscribe to this Specialization?
- Is financial aid available?

More questions? Visit the **Learner Help Center**.