

Deployment

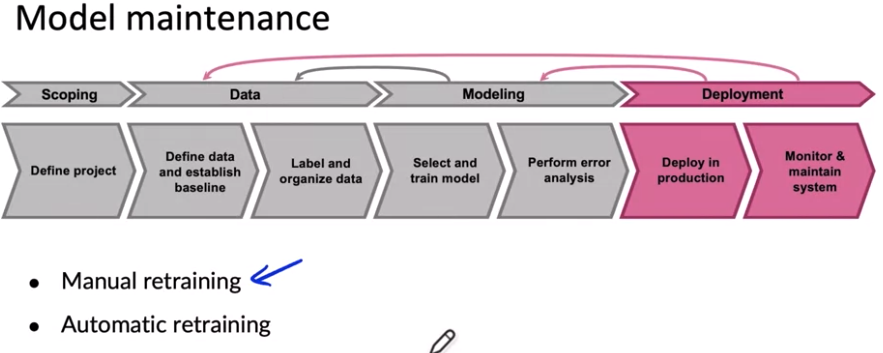
1. Concept Drift – Inflation changes house price
2. Data Drift – Covid 19 makes credit. Card purchases valid
3. Software Engineering issues – Batch or real time prediction
4. Cloud Vs Edge device or Browser
5. Compute resources – how much CPU or GPU needed?
6. Latency, throughput (QPS)
7. Logging and monitoring
8. Security and Privacy (HIPAA, GDPR)
9. First deployment Vs Maintenance. 1st deeply is 50% of work

Deployment Patterns

1. New product/capability
2. Automate/Assist with manual task
3. Replace previous ML system
4. Rollback ability
5. Gradual ramp up with monitoring
6. ML shadows human
7. Canary Deployment – Rollout to 10% of traffic. Rampup if you are confident in the model
8. Blue Green Deployment – Old version is Blue, Green version is new
   1. Have load balancer send blue to green. Rollback to blue if something goes green
   2. Gradual switch to green
9. Degrees of automation
   1. Human only
   2. Shadow mode
   3. AI assistance (human in the loop)
   4. Partial automation – confidence score, send to human if you are not confident
   5. Full automation

Monitoring

1. Dashboard, server load, non-null outputs, missing input values
2. Metrics that need to be monitored
3. Start with a large set of metrics then scale down
4. Software metrics - Memory, compute, latency throughput, server load (Software Metrics)
5. Input metrics - Statistical health, input metrics, Average input length, num missing values, average image brightness
6. Output metrics – Output is null, time user redoes search, # of times user switches to typing, CTR
7. Deployment is iterative like ML modeling
   1. ML model ->Experiment -> Error analysis
   2. Deploy -> Real traffic -> Performance analysis
8. Set threshold for alarms
9. Adapt metrics and thresholds over time
10. Manual retraining vs Automatic retraining



Pipeline Monitoring

1. Model 1 performance could impact model2 performance
2. User clicks -> user profile -> recommender system -> Recommender system, if click changes user profile model may change include recommender system

Monitor

\* software metrics

\* input metrics

\* output metrics

How quickly do data change

\* Depends on the problem

\* On average user data generally has slow drift

\* Covid-19 shocks are exception

\* B2B data shift fast. Enterprise data shift fast

<https://towardsdatascience.com/machine-learning-in-production-why-you-should-care-about-data-and-concept-drift-d96d0bc907fb>

<https://christophergs.com/machine%20learning/2020/03/14/how-to-monitor-machine-learning-models/>

<https://www.youtube.com/watch?v=06-AZXmwHjo>

Papers

Konstantinos, Katsiapis, Karmarkar, A., Altay, A., Zaks, A., Polyzotis, N., … Li, Z. (2020). Towards ML Engineering: A brief history of TensorFlow Extended (TFX). <http://arxiv.org/abs/2010.02013>

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2020). Challenges in deploying machine learning: A survey of case studies. <http://arxiv.org/abs/2011.09926>

Sculley, D., Holt, G., Golovin, D., Davydov, E., & Phillips, T. (n.d.). Hidden technical debt in machine learning systems. Retrieved April 28, 2021, from Nips.c<https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf>

<https://github.com/https-deeplearning-ai/MLEP-public>

<https://github.com/https-deeplearning-ai/MLEP-public/tree/main/course1/week1-ungraded-lab>