Lessons Learned
Deploying and Monitoring AI Models in Production at Major Tech Companies
Who are we?

Harish Doddi
CEO

Jerry Xu
CTO
Today’s Enterprise AI life cycle

Development: Discovery

"The Playground"

Production

"The Battleground"

Optimization

"The Battleground"
Lesson 1

You either NEVER deploy a model, or you have to do it over and over again
Are your models decaying?

Deploy and Done

- Model decays over time

![Graph showing performance decreasing over time]

Result: Model performance decreases

Deploy over and over...

- Model replenishes

![Graph showing performance fluctuating but consistent over time]

Result: Model performance consistent
ML model cycle is a continuously optimizing process

- Concept drift
- New concept comes up
- ...
Connecting Machine Learning to Software world

Before

Software deployment
once a 1 or 2 years

Now

Software deployment
every day

BUT

Machine Learning models
deploy
very slow

Future

Machine Learning models will
deploy
very frequent
and fast
Lesson 2

Models may go wrong, you need to monitor them
South Park and Alexa
Monitoring Learning: Post mortem is the only option

Without Model Monitoring

The problem occurs | The team detects the problem and decides what to do

With Model Monitoring

The problem occurs | The team decides what to do
Notify asap
Monitoring for Machine Learning Models

- **Model Performance monitoring**
  - Confusion Matrix
  - Gain and Lift charts
  - Kolomogorov Smirnov chart
  - Area Under the ROC curve
  - Gini Coefficient
  - Concordant – Discordant ratio
  - Root Mean Squared Error (RMSE)
  - etc
- **Model Timeout monitoring**
- **Infrastructure monitoring**
- **Organization KPI monitoring**
- **Deployment monitoring**
Lesson 3

Your real work starts AFTER you deploy the model to production
Enterprise AI Life Cycle

Exploration → Training → Deploy

X
Enterprise AI Life Cycle After Deployment

1. **Deploy**
   - Blue Green Deployment
   - Rollback
   - Canary

2. **A/B Testing**
   - Split traffic
   - Shadowing

3. **SLA**
   - Monitor performance
   - Fall back strategy
   - Alerting

4. **Model Selection**
   - Model routing
   - Challenger
   - KPI based selection

5. **Anomaly detection**
   - Feature distribution
   - Model result
Data science is a scarce resource, you need to make sure you organize well.
Deployment Learning: Rise of new engineering role

There is a hyper-competitive WAR FOR TALENT that is projected to get much worse.
Teams face cross-functional inefficiencies

BEST CASE SCENARIO:
with a world-class team, 1 model deployed per quarter

• Teams operate in silos, don’t speak the same language
• Errors due to lack of communication
• Engineering has to write stand-alone scripts
Machine learning offers a fantastically powerful toolkit for building complex systems quickly. … it is remarkably easy to incur massive ongoing maintenance costs at the system level when applying machine learning.

- Boundary erosion
- Entanglement
- Hidden feedback loops
- Undeclared consumers
- Data dependencies
- Changes in the external world
- System-level anti-patterns
Lesson 5

Be prepared, your number of models will increase
Deployment Learning: 1 model vs Multiple models

Growing Use of Deep Learning at Google

Number of directories containing model description files

Across many products/areas
- Apps
- Maps
- Photos
- Gmail
- Speech
- Android
- YouTube
- Translation
- Robotics Research
- Image Understanding
- Natural Language Understanding
- Drug Discovery
Cost per model increases significantly if no automation

As the number of models increases, the cost also increases.
Lesson 6

Senior people are needed AFTER deploying to production
Software Development vs ML Model Development

Senior People

Requirements → Design → Implementation → Testing → Evolution

Senior People

Requirements → Data Preparation → Training / Testing → Deploy to Production → Monitoring and Optimization
Lesson 7

Don’t be married to a single framework
Build/Bring Your Own Models, Frameworks, Languages
Innovators Pavilion
Booth P4

Thank you!

harish@datatron.com