

Machine Learning Systems Design

Deployment and Monitoring

Lecture 20: Model Monitoring



CE 40959 Spring 2023

Ali Zarezade

[SharifMLSD.github.io](https://github.com/SharifMLSD)

Agenda

1. Natural Labels & Feedback Loops
2. Causes of ML System Failures

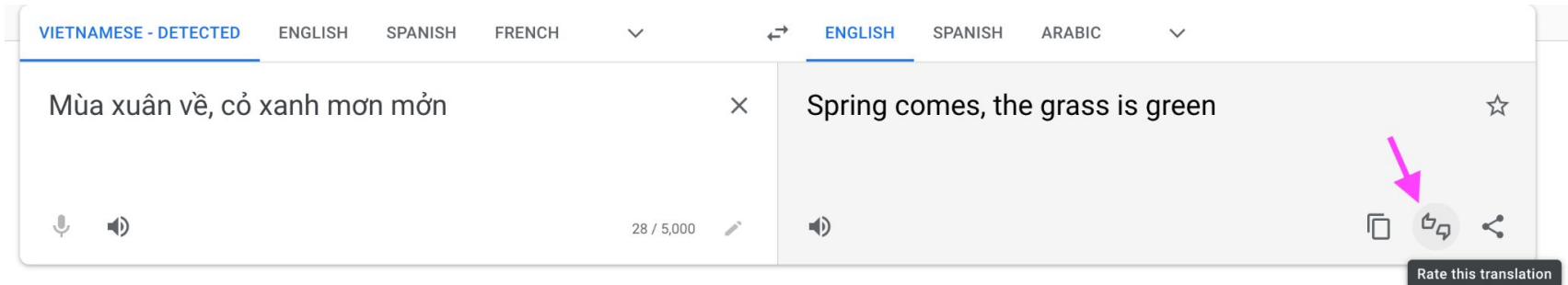
1. Natural Labels & Feedback Loops

Natural labels

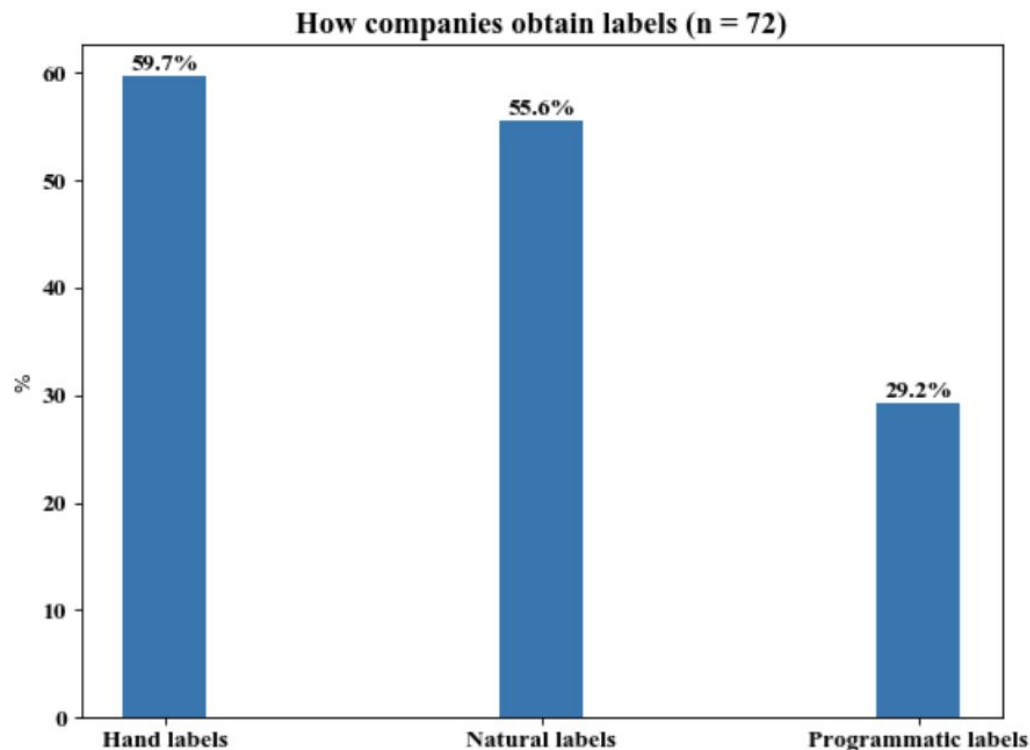
- The model's predictions can be automatically evaluated or partially evaluated by the system.
- Examples:
 - ETA
 - Ride demand prediction
 - Stock price prediction
 - Ads CTR
 - Recommender system

Natural labels

- You can engineer a task to have natural labels



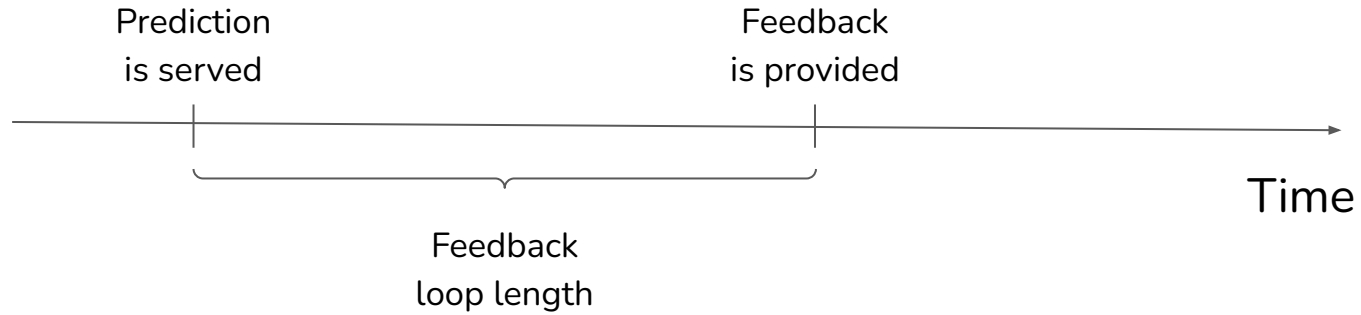
Natural labels: surprisingly common



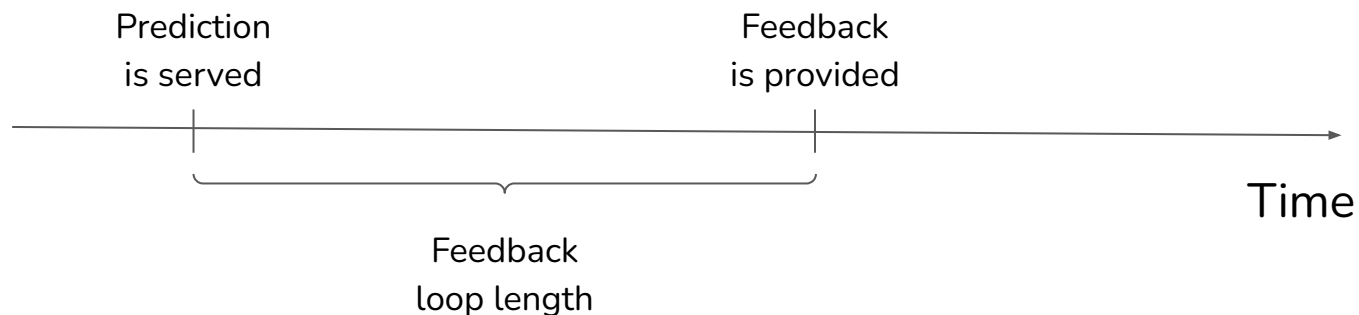
⚠ Biases ⚠

- Small sample size
- Companies might only use ML for tasks with natural labels

Delayed labels

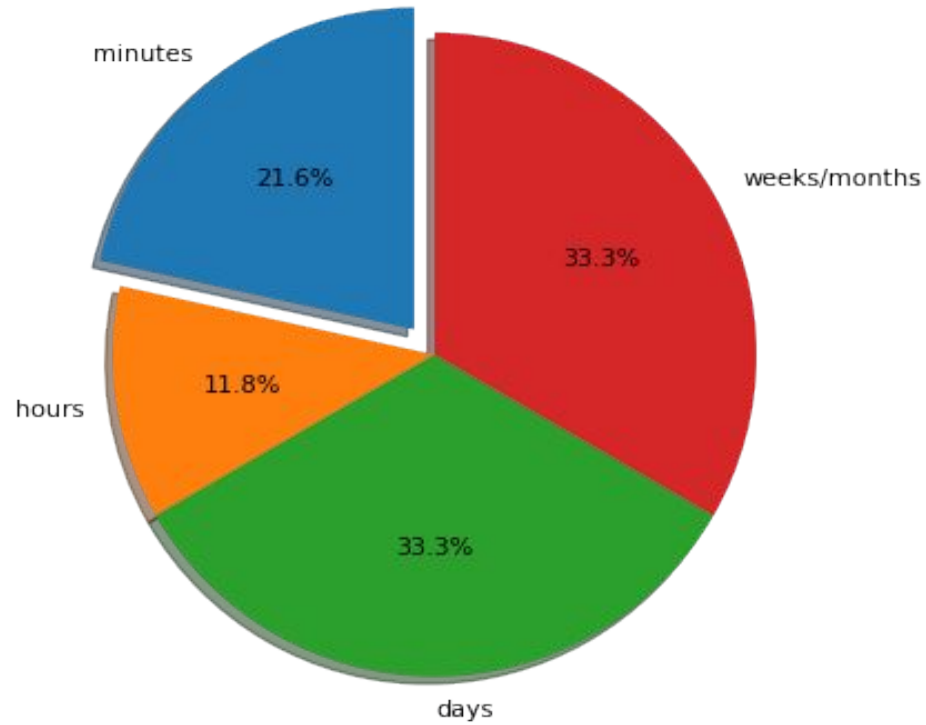


Delayed labels



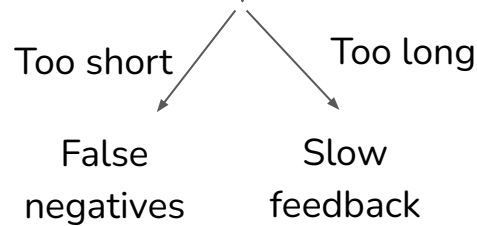
- Short feedback loop: minutes -> hours
 - Reddit / Twitter / TikTok's recommender systems
- Long feedback loop: weeks -> months
 - Stitch Fix's recommender systems
 - Fraud detection

Feedback loop length (n = 51)



! Labels are often assumed !

- Recommendation:
 - Click -> good rec
 - After X minutes, no click -> bad rec



Speed vs. accuracy
tradeoff



Labels are often assumed



- Recommendation:

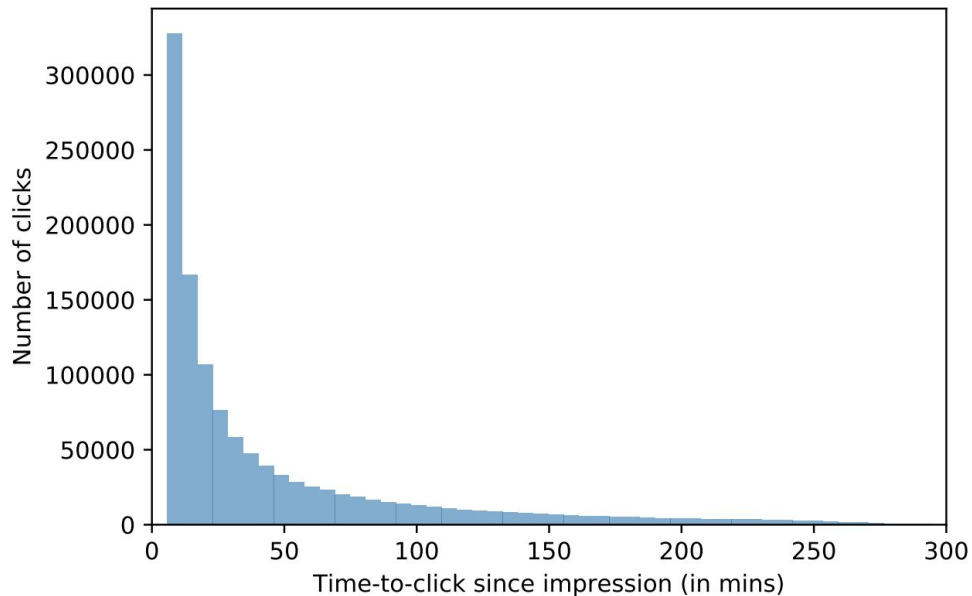
- Click -> good rec
- After X minutes, no click -> bad rec

Too short

Too long

False
negatives

Slow
feedback



2. Causes of ML System Failures

Amazon scraps secret AI recruiting tool that showed bias against women

That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

Japan's Henn na Hotel fires half its robot workforce

“Guests complained their robot room assistants thought snoring sounds were commands and would wake them up repeatedly during the night.”



What is an ML failure?

A failure happens when one or more expectations of the system is violated.

Two types of expectations:

- Operational metrics: e.g. average latency, throughput, uptime
- ML metrics: e.g. accuracy, F1, BLEU score

What is an ML failure?

A failure happens when one or more expectations of the system is violated

- Traditional software: mostly operational metrics
- ML systems: operational + ML metrics
 - Ops: returns an English translation within 100ms latency on average
 - ML: BLEU score of 55 (out of 100)

ML system failures

- If you enter a sentence and get no translation back -> ops failure
- If one translation is incorrect -> ML failure?

ML system failures

- If you enter a sentence and get no translation back -> ops failure
- If one translation is incorrect -> ML failure?
 - Not necessarily: expected BLEU score < 100
 - ML failure if translations are consistently incorrect

Ops failures

Visible

- 404, timeout, segfault, OOM, etc.



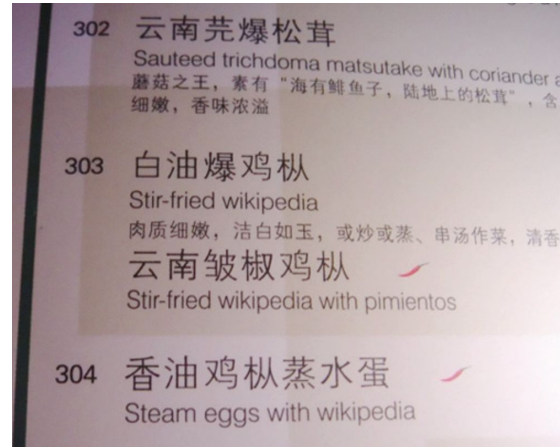
404 - PAGE NOT FOUND

A special message from Bernie



ML failures

Often invisible



Causes of ops failures (software system failures)

- Dependency failures
- Deployment failures
- Hardware failures
- Network failure: downtime / crash

Causes of ops failures (software system failures)

- Dependency failures
- Deployment failures
- Hardware failures
- Network failure: downtime / crash



60 / 96 ML systems failures are non-ML failures

(Papasian & Underwood, 2020)



As tooling & best practices around ML production mature,
there will be less surface for software failures

ML-specific failures (during/post deployment)

1. Production data differing from training data
2. Edge cases
3. Degenerate feedback loops

We've already covered problems
pre-deployment in previous lectures!

Production data differing from training data

- Train-serving skew:
 - Model performing well during development but poorly after production
- Data distribution shifts
 - Model performing well when first deployed, but poorly over time
 -  What looks like data shifts might be caused by human errors 

Production data differing from training data

- Train-serving skew:
 - Model performing well during development but poorly after production
- Data distribution shifts
 - Model performing well when first deployed, but poorly over time
 - ⚠️ What looks like data shifts might be caused by human errors ⚠️

} Common & crucial.
Will go into detail!

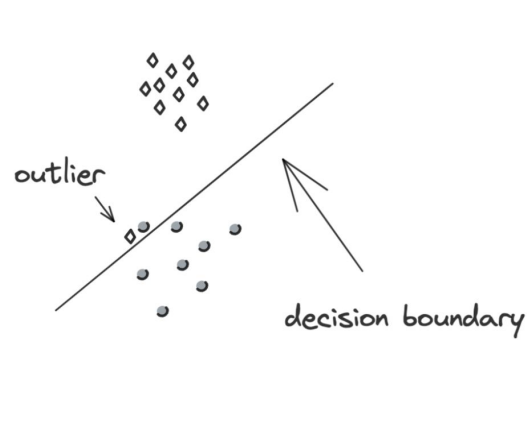
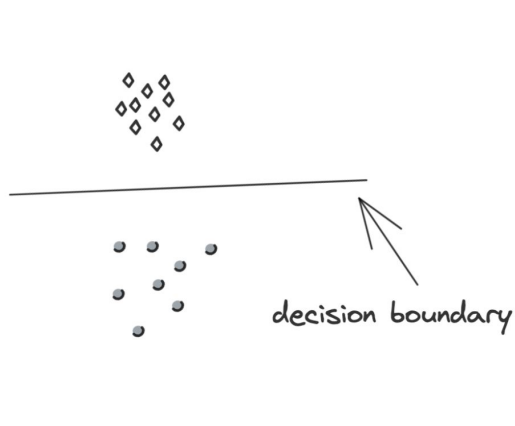
Edge cases

- Self-driving car (yearly)
 - Safely: 99.99%
 - Fatal accidents: 0.01%

Zoom poll: Would you
use this car?

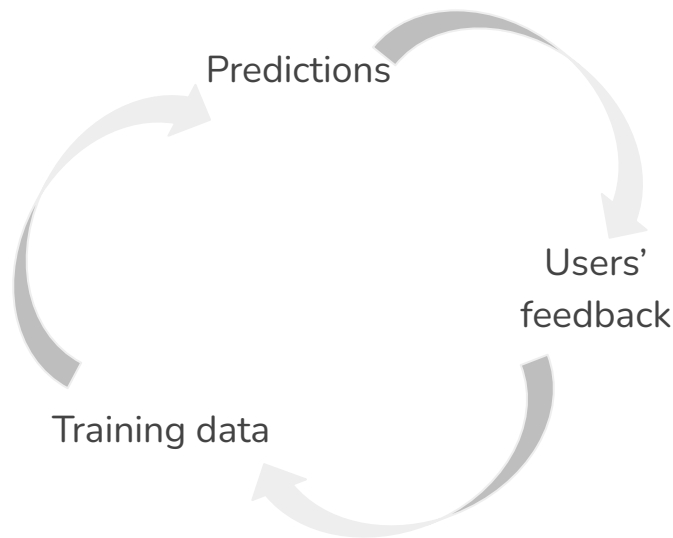
Edge case vs. outlier

- Outliers
 - Refer to inputs
 - Options to ignore/remove
- Edge cases
 - Refer to outputs
 - Can't ignore/remove



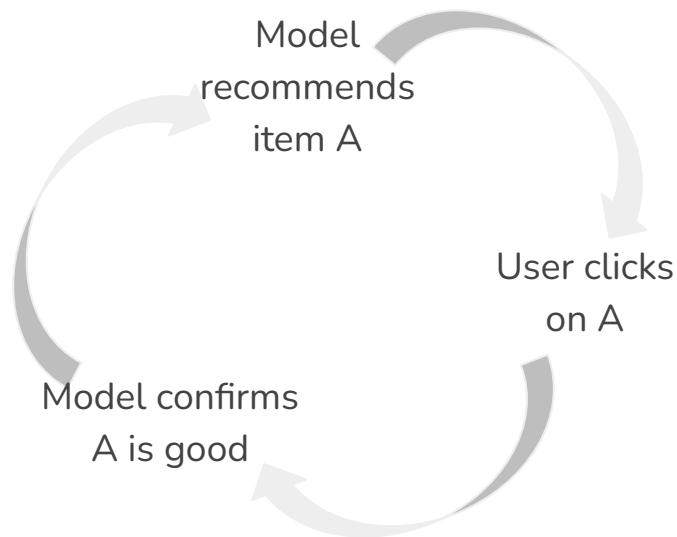
Degenerate feedback loops

- When predictions influence the feedback, which is then used to extract labels to train the next iteration of the model
- Common in tasks with natural labels



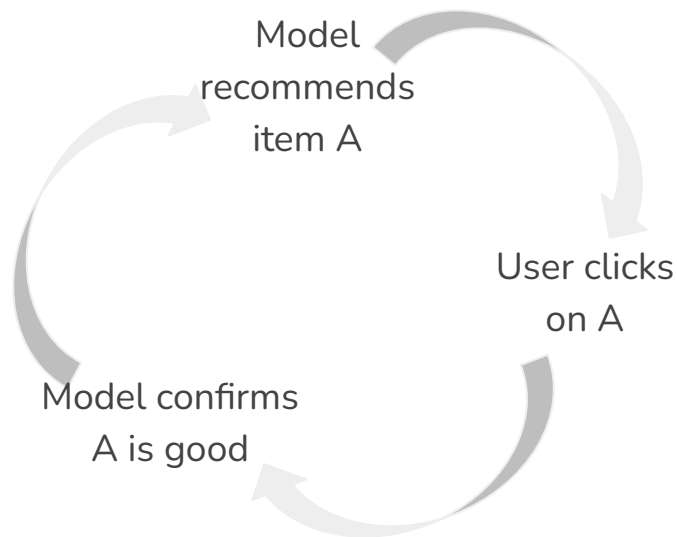
Degenerate feedback loops: recsys

- Originally, A is ranked marginally higher than B -> model recommends A
- After a while, A is ranked much higher than B



Degenerate feedback loops: recsys

- Originally, A is ranked marginally higher than B -> model recommends A
- After a while, A is ranked much higher than B



Over time,
recommendations
become more
homogenous

Degenerate feedback loops: resume screening

- Originally, model thinks X is a good feature
- Model only picks resumes with X
- Hiring managers only see resumes with X, so only people with X are hired
- Model confirms that X is good



Replace X with:

- Has a name that is typically used for males
- Went to Stanford
- Worked at Google

Degenerate feedback loops: resume screening

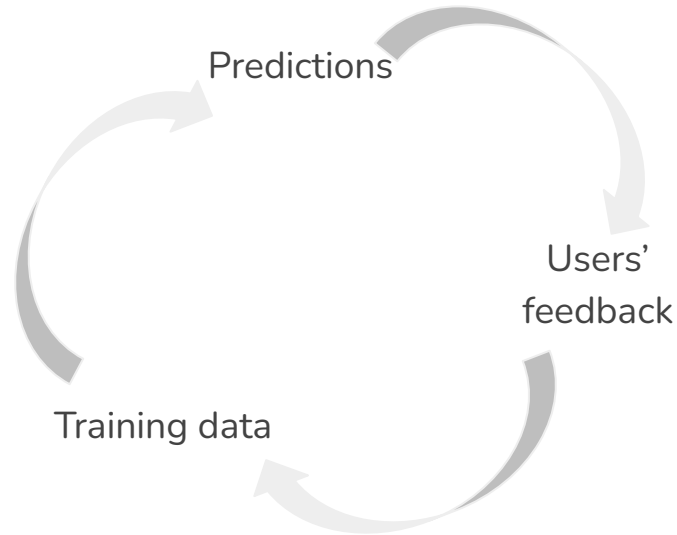
- Originally, model thinks X is a good feature
- Model only picks resumes with X
- Hiring managers only see resumes with X, so only people with X are hired
- Model confirms that X is good



Tracking feature importance might help!

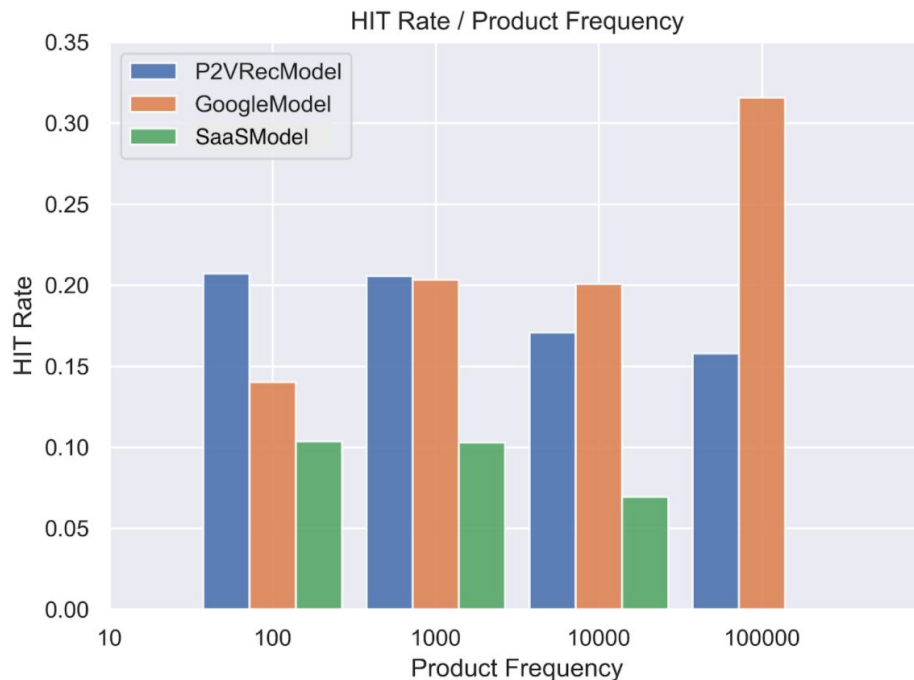
Detecting degenerate feedback loops

Only arise once models are in production -> hard to detect during training



Degenerate feedback loops: detect

- Average Rec Popularity (ARP)
 - Average popularity of the recommended items
- Average Percentage of Long Tail Items (APLT)
 - average % of long tail items being recommended
- Hit rate against popularity
 - Accuracy based on recommended items' popularity buckets



Degenerate feedback loops: mitigate

1. Randomization
2. Positional features

Randomization

- Degenerate feedback loops increase output homogeneity
- Combat homogeneity by introducing randomness in predictions

Randomization

- Degenerate feedback loops increase output homogeneity
- Combat homogeneity by introducing randomness in predictions
- Recsys: show users random items & use feedback to determine items' quality



Positional features

- If a prediction's position affects its feedback in any way, encode it.
 - Numerical: e.g. position 1, 2, 3, ...
 - Boolean: e.g. shows first position or not

Positional features: naive

ID	Song	Genre	Year	Artist	User	1st Position	Click
1	Shallow	Pop	2020	Lady Gaga	listenr32	False	No
2	Good Vibe	Funk	2019	Funk Overlord	listenr32	False	No
3	Beat It	Rock	1989	Michael Jackson	fancypants	False	No
4	In Bloom	Rock	1991	Nirvana	fancypants	True	Yes
5	Shallow	Pop	2020	Lady Gaga	listenr32	True	Yes

Positional features: naive

ID	Song	Genre	Year	Artist	User	1st Position	Click
1	Shallow	Pop	2020	Lady Gaga	listenr32	False	No
2	Good Vibe	Funk	2019	Funk Overlord	listenr32	False	No
3	Beat It	Rock	1989	Michael Jackson	fancypants	False	No
4	In Bloom	Rock	1991	Nirvana	fancypants	True	Yes
5	Shallow	Pop	2020	Lady Gaga	listenr32	True	Yes

Doesn't have this
feature during
inference?

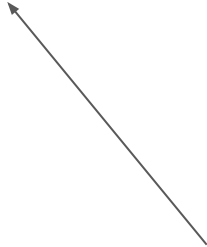
Positional features: naive

ID	Song	Genre	Year	Artist	User	1st Position	Click
1	Shallow	Pop	2020	Lady Gaga	listenr32	False	No
2	Good Vibe	Funk	2019	Funk Overlord	listenr32	False	No
3	Beat It	Rock	1989	Michael Jackson	fancypants	False	No
4	In Bloom	Rock	1991	Nirvana	fancypants	True	Yes
5	Shallow	Pop	2020	Lady Gaga	listenr32	True	Yes

Set to False during
inference

Positional features: 2 models

1. Predicts the probability that the user will **see and consider** a recommendation given its position.
2. Predicts the probability that the user will **click on the item given that they saw and considered it.**



Model 2 doesn't
use positional
features

How might degenerate feedback loops occur? (10 mins)

1. Build a system to predict stock prices and use the predictions to make buy/sell decisions.
2. Use text scraped from the Internet to train a language model, then use the same language model to generate posts.

Discuss how you might mitigate the consequences of these feedback loops.

Machine Learning Systems Design

Deployment and Monitoring

Next Lecture: Model Monitoring



CE 40959 Spring 2023

Ali Zarezade

[SharifMLSD.github.io](https://github.com/SharifMLSD)