https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

ML System design - 1

Start at 9:05 pm

→

→ a

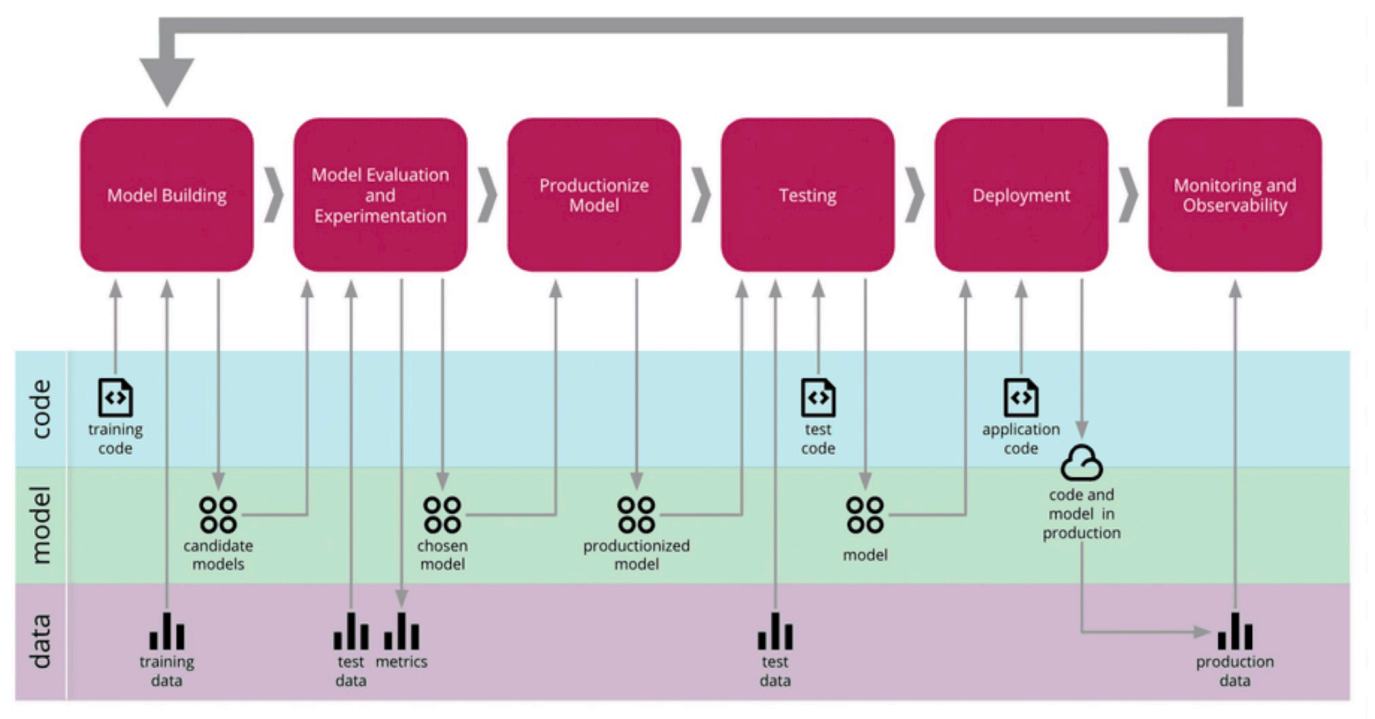
→ July 2020 ⟶ July 2023

→ 1) model change performance
2) more data available
3) New research could have been done
4) Buisness requirement ⟶ change

⇒ Developer → Software → System-design
↳ Solution architect

⇒ ML System Design →
→

Machine Learning Lifecycle



→ Classification problem → • DT • RF • Logistic • GBDT • SVM

→ on types of model + Hyper parameters

→ Production → you can use the entire dataset

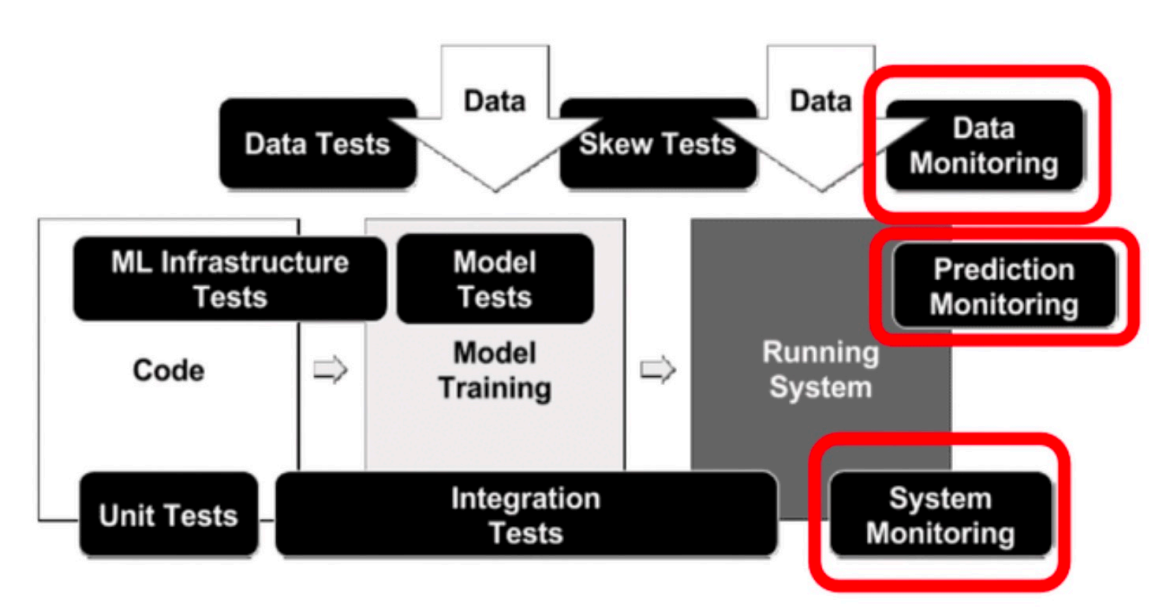
→ 80% → Train
20% → Test
→ Chosen model → Retrain on entire data → save

→ • Heavy Correlation task, preprocessing tasks

→ Monitor → • Performance • Drift • Aborting → Sending an alert

→ Model to check → Password Security → user Inputs (in go Logic)

⇒ Different testings in ML Pipeline →



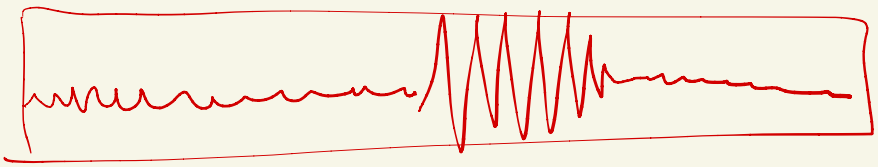
→ 0.05% extreme is outlier

→ 5 function → final function

→ unit testing → individually test all 5 function

→ Integration testing → test final function.

Monitoring → 1) Data monitoring



→ Prediction Monitoring → How good are the results

→ System Monitoring → . Request per second
· Latency →
· CPU / Memory / Network

⇒ "avg order per min" → RS for similar product

→ · Iphone → a black case with it

→ · Iphone + gaming pad or Iphone + black case

→ · Data for your model
+
Data to identify change events ⇒

Drift → away from normal

→ ALL ML model degrade over time when you deal with real world

→ Why it occurs →

→ Outdated Data →

→ Inefficient Model →

⇒ Two types of Drift →

- Concept drift → • For same data → expected output has changed.

→ 2020 → property price bangalo | → 1 BHK HSR (
2021 → metro started | → 1 BHK HSR

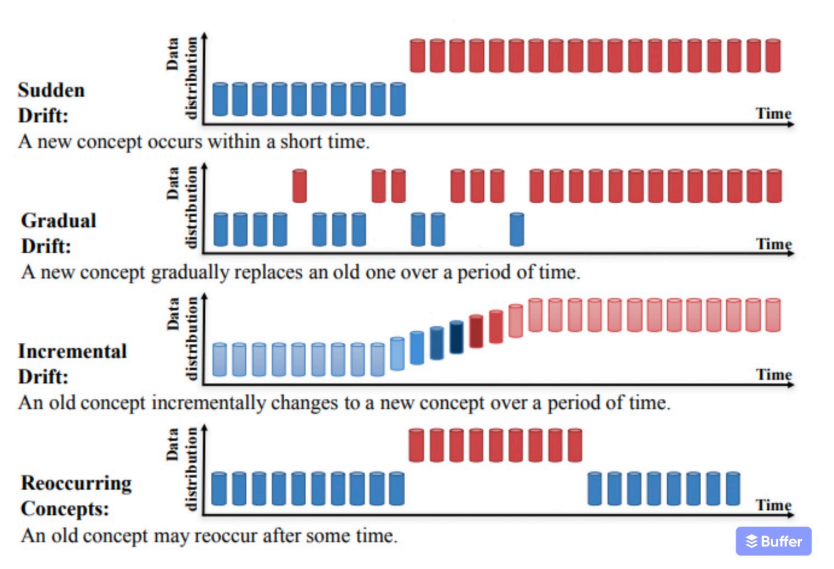
X → Y → this has changed

- Data drift → Inputs have changed
  - Demand of face mask, sanitizer ←
  -

grocery →

Reasons
→ • Data collection process has changed
  - Labeling can change →
  -
  -

High mod Low
High model Low



→ How to identify data drift

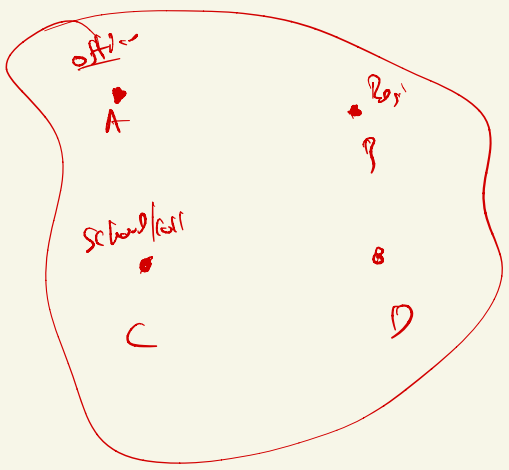
→ Visual Inspection

→ Statistical test → Chi-squared test
KS test

→ Model performance tracking →

→ Data quality checks →

⇒ → Cycle demand forecasting



Check the demand is
→ Drifting over time.

KS Test → Kolmogorov - Smirnov Test → Numerical Data

⇒ What is the probability that two give samples belong to the same probability distribution?

→ Dataset A

Dataset B

→ null hypothesis → Assume that A & B belong to same distribution

Accept →

Reject → ✓

⇒ Take data set A → Jan data

B → Feb data

⇒ Chi-Square Test → Categorical data

null hypothesis → The two groups have no significant difference.

→ • find difference in counts of each value

	Jan	Feb
→ A	10	4 ←
→ B	11	10 ←
→ C	12	13
→ D	10	9

⇒ Interview topics →

- ML Pipeline → All steps *
- Different testing] *
- Drift → • Concept vs Data *
  - Time base drift
  - KS Test & Chi-Square test

Doubts

1. identify factia
2. allow ml engineers to compare performance
3. → if same code, data, hyperparameters
4. Easily share your code