PLTHIYAGU (HTTPS://ID.ANALYTICSVIDHYA.COM/ACCOUNTS/PROFILE/)

tics
a (https://www.analyticsvidhya.com/blog/)

GORY/BEGINNER/)

BLOG/CATEGORY/MACHINE-LEARNING/)

ystem Design

OG/AUTHOR/KPIGATE15/), JANUARY 29, 2021 (HTTPS://WWW.ANALYTICSV...

Article Videos

This article was published as a part of the [Data Science Blogathon (https://datahack.analyticsvidhya.com/contest/data-science-blogathon-4/)].

Introduction

As ML applications are maturing over time and becoming an indispensable component of industries for making faster and accurate decisions for critical and high-value transactions.

For example,

1. Recommendation system to increase click-through rate for e-commerce.

2. Increasing engagement time users for social media apps.

riving cars.

esponse is accurate, fast, and reliable. Scalability, maintainability, ove towards making ML one of the main components of enterprise- end system keeping requirements of ML becomes important.

m, data infrastructure, and hardware for ML Learning system to meet specific requirements of reliability, scalability, maintainability, and adaptability.

rrect function at the desired level of performance under a specified systems where learning is done from data is tricky as its failure ply give garbage outputs i.e. it may produce some outputs even ing ground truth during training.

ssage e.g. **There is some technical issue, the team is working on,**

es in language translation system (English to Hindi or vice versa), even though the model has not seen some words, it still might give some translation, which does not make sense.

Scalability

As the system grows (in data volume, traffic volume, or complexity), there should be reasonable ways of dealing with that growth. There should be automatic provision to scale compute/storage capacity because for some critical applications even 1 hour of downtime/failure can cause in loss of millions of dollars/credibility of the app.

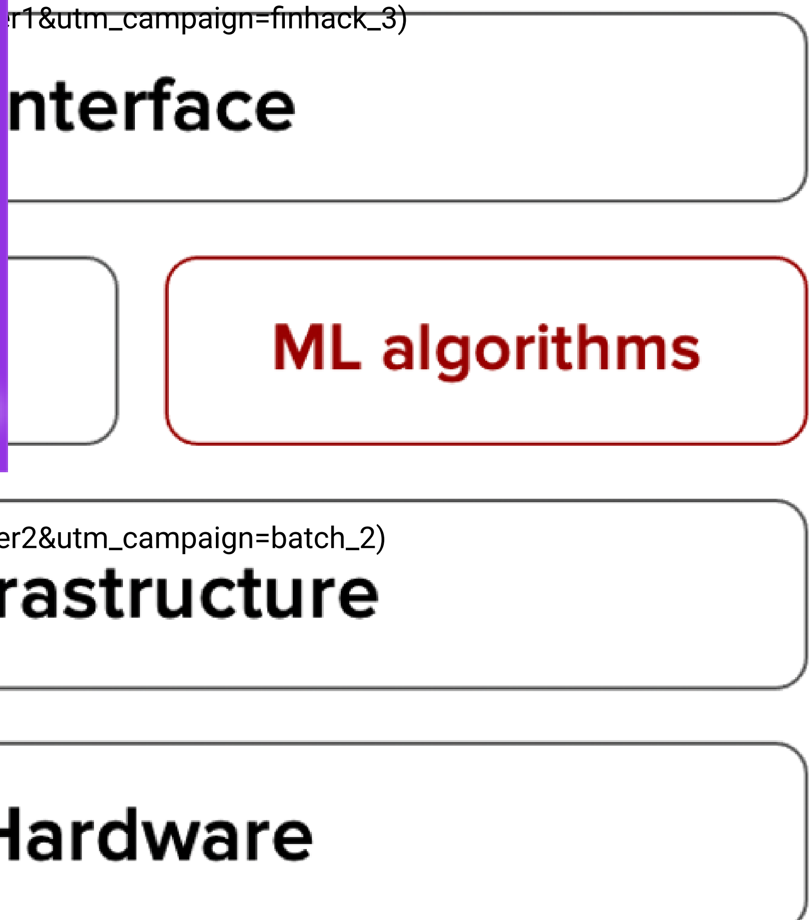
As an example, if for an e-commerce website on a prime day, if certain feature failure i.e. not working as expected, that may turn into revenue loss in order of millions.

Maintainability

...get change, and hence model performance. There should be a ...here is any model drift/data drift, and once the significant drift is ...able updated ML model without disturbing the current operation of

...the ML systems are the availability of new data with added features ...sion rate vs engagement time from customers for e-commerce. ...pdates without any service interruptions.

Schematic Diagram for ML System



For, designing the ML system we need to consider all the above-Data, ML algorithms, related infrastructure, hardware, and Interface.

Few of the high-level considerations for design,

ema – ranges of the feature values well captured to avoid any un-
response e.g. human age/height have expected value range, it
- 10 feet, etc.

in the system should have some usefulness in terms of predictive
e added has an associated handling cost.

3. **No features's cost is too much**– trade-off of cost vs benefits must be done for each feature added to remove features which has too little to add and too much to handle.

s – features used should match the project requirements e.g. there
the model e.g. sex, age, race, etc.

ontrols – e.g. personally identifiable information (PII) should be
data may lead to legal consequences.

s will help in adding new features to improve system performance if
em.

s e.g. one hot encoding/binning features or any other
transformations code must be tested to avoid any intermediate values going off the expected range e.g. handling of unseen levels in one-hot encoded features.

Model

1. **Model specs are reviewed and submitted**– proper versioning of the model learning code is needed for faster re-training.

2. **Offline and online metrics correlate**– model metrics (log loss, mape, mse) should well correlated with the objective of application e.g. revenue/cost/time.

3. **All hyperparameters have been tuned**– hyperparameters, such as learning rates, number of layers, layer sizes, max depth, and regularization coefficients must be tuned for the use case. Because the choice of

hyperparameter values can have a dramatic impact on prediction quality.

ow frequently re-train models based on changes in data distribution
nodel in production.

g models with baseline i.e. simple linear model with high-level
l testing and doing cost/benefit trade-off analysis against

a slices-model performance must be vetted against sufficiently

nclusion-model features should be well-vetted against importance
in forecasting as in some applications certain features may bias results towards certain categories mostly for fairness purposes.

e same data should produce two identical models. Generally, there
sion of the system/infra used. But, there should not be any major

to unit test model algorithm correctness and model API service
code/response.

3. **The ML pipeline is Integration tested**- complete ML pipeline – assembling of training data, feature generation, model training, model verification, and deployment to a serving system must be tested for the correct function of the ML system.

4. **Model quality is validated before serving**-After a model is trained but before it actually serves the real requests, an offline/online system needs to inspect it and verify that its quality is sufficient.

5. **The model is debuggable**-When someone finds a case where the model is behaving bizarrely, it should be well logged to make it easy to debug.

6. **Serving models can be rolled back-** considering the behavior of ML systems which performance very much depends on non-stationary quality/distribution of input data, there should be well designed fallback mechanism if something went wrong in ML response.

any changes in the downstream inputs of the ML system should be
ML performance deterioration.

uality and distribution should remain statistically constant, and if
lel should be refreshed/re-trained accordingly.

ure generation code for both training and inference should be the

4. Models are numerically stable-Invalid or implausible numeric values that can potentially crop up during model
training without triggering explicit errors, and knowing that they have occurred can speed diagnosis of the

l-model has not experienced regressions in training speed, serving

k for model performance metrics as soon as in production, after
ilable. If there is any significant model performance drift is

into ML system design is how to serve prediction, batch vs online
equirements.

Batch prediction

1. Periodical e.g. hourly, weekly, etc.

2. Processing accumulated data when you don't need immediate results (e.g. recommendation systems)

3. High throughput

4. Finite: need to know how many predictions to generate

Online prediction

ata sample is generated e.g., fraud detection

quirements, prediction can be served in a hybrid way- batch & online.
queries are precomputed and stored e.g., Restaurant
in each restaurant, item recommendations uses online predictions

Another consideration that comes, is cloud vs edge computing,

transfer

sistant

Edge Computing

1. Done on edge devices (browsers, phones, tablets, laptops, smartwatches, etc.)

2. Need high power hardware for system -memory, compute power, energy for doing computations

3. E.g., Unlocking with fingerprint, faces

Benefits of Edge Computing

1. Can work without (Internet) connections or with unreliable connections

2. Don’t have to worry about network latency

3. Fewer concerns about privacy

n is in production, the ML model can continually improve based on

e, here are some of the highlights of the offline/online learning

depending on the use case

2. Batch Size-larger sample size, the order of millions

)

2. Batch Size – Micro batches- hundreds of samples

3. Most relying on A/B testing kind of validation

4. Twitter #hashtag trending

Conclusion

As above described, we see creating production-level ML systems brings on a lot of challenges not found in small POC or even large offline research experiments e.g., system-level requirements- reliability, scalability, adaptability, and ML specific requirements – data, algorithms, model monitoring-model drift/covariate shift, model serving infrastructure – cloud/edge, model serving type-batch/online and model retraining -offline/online. Each step based on criticality and business requirements is important for the success of the ML-based systems.

l process of data science is – Data Scientist completes POC for the ased on business logic and specific model metrics, and then ship to is puts lots of system-level challenges e.g., the language of de in R while production system needs in python/java, python ue, latency requirements i.e., prediction request takes longer time

ist needs to have high-level visibility about the production system, pplication objective/function.

References

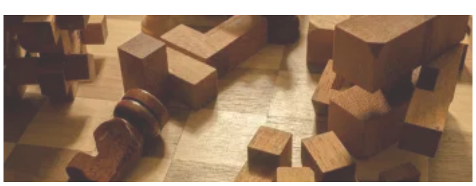
ps://research.google/pubs/pub46555/)

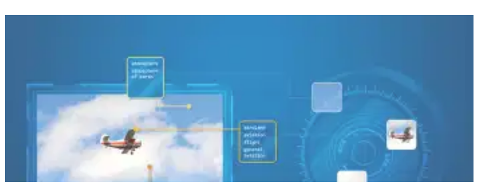
ml (https://stanford-cs329s.github.io/syllabus.html)

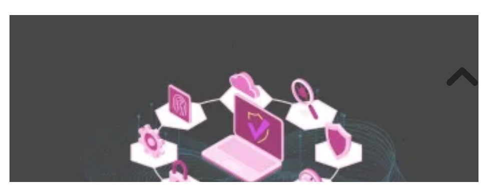
d by Analytics Vidhya and is used at the Author’s discretion.

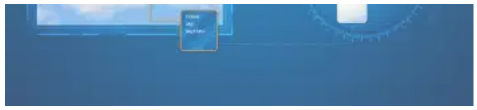
PP GET IT ON Google Play

Related Articles








/06/2 www.analyticsvidhya.com/blog/2020/11/automl- (https://www.analyticsvidhya.com/blog/2021/01/security-
g-ai-more-accessible-to-
sses/)
: Making AI more Accessible
nesses
/www.analyticsvidhya.com/bl
0/11/automl-making-ai-more-
ble-to-businesses/)

threats-to-machine-learning-systems/)
Security Threats to Machine Learning Systems
(https://www.analyticsvidhya.com/blog/2021/01/security-threats-to-machine-learning-systems/)

YA.COM/BLOG/TAG/BLOGATHON/), MACHINE LEARNING SYSTEM DESIGN
MACHINE-LEARNING-SYSTEM-DESIGN/)

NEXT ARTICLE

How to create your own Question and Answering API(Flask+Docker +BERT) using haystack framework

(https://www.analyticsvidhya.com/blog/2021/01/how-to-create-your-own-question-and-answering-apiflaskdocker-bert-using-haystack-framework/)

•••

PREVIOUS ARTICLE

Kaggle Grandmaster Series – Exclusive Interview with Kaggle Discussion Grandmaster Raju Kumar Mishra (#Rank 48)

(https://www.analyticsvidhya.com/blog/2021/01/kaggle-grandmaster-series-exclusive-interview-with-kaggle-discussion-

aster-raju-kumar-mishra-rank-48/)

(https://www.analyticsvidhya.com/blog/author/kpigate15/)

Kpigate15 (Https://Www.Analyticsvidhya.Com/Blog/Author/Kpigate15/)

Name (required)

Email (required)

Website

POPULAR POSTS

Commonly used Machine Learning Algorithms (with Python and R Codes) (https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)

Making Exploratory Data Analysis Sweeter with Sweetviz 2.0 (https://www.analyticsvidhya.com/blog/2021/01/making-exploratory-data-analysis-sweeter-with-sweetviz-2-0/)

Inferential Statistics – Sampling Distribution, Central Limit Theorem and Confidence Interval (https://www.analyticsvidhya.com/blog/2021/01/inferential-statistics-sampling-distribution-central-limit-theorem-and-confidence-interval/)

Introductory guide on Linear Programming for (aspiring) data scientists (https://www.analyticsvidhya.com/blog/2017/02/lintroductory-guide-on-linear-programming-explained-in-simple-english/)

ustering Techniques (Skill test Solution)
017/02/test-data-scientist-clustering/)

achine Learning [Solution: SkillPower – Machine Learning, DataFest
blog/2017/04/40-questions-test-data-scientist-machine-learning-
fest-2017/)

sics of Deep Learning (along with solution)
017/01/must-know-questions-deep-learning/)

hould add to their Portfolio
020/12/10-data-science-projects-for-beginners/)

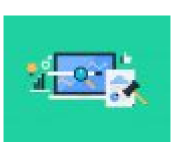
(https://datahack.analyticsvidhya.com/contest/ltfs-data-science-finhack-3/?
r1&utm_campaign=finhack_3)

(https://ascendpro.analyticsvidhya.com/?utm_source=Blog&utm_medium=stickybanner2&utm_campaign=batch_2)

CAREER RESOURCES

16 Key Questions You Should Answer Before Transitioning into Data Science (https://www.analyticsvidhya.com/16-key-questions-data-science-career-transition/?&utm_source=Blog&utm_medium=CareerResourceWidget)

NOVEMBER 23, 2020



Here's What You Need to Know to Become a Data Scientist! (https://www.analyticsvidhya.com/blog/2021/01/heres-what-you-need-

to-know-to-become-a-data-scientist/?
&utm_source=Blog&utm_medium=CareerResourceWidget)
...UARY 22, 2021

...se 7 Signs Show you have Data Scientist Potential!
...ps://www.analyticsvidhya.com/blog/2020/12/these-7-signs-show-...have-data-scientist-potential/?
...n_source=Blog&utm_medium=CareerResourceWidget)
...EMBER 3, 2020

... To Have a Career in Data Science (Business Analytics)?
...ps://www.analyticsvidhya.com/blog/2020/11/how-to-have-a-career-...ata-science-business-analytics/?
&utm_source=Blog&utm_medium=CareerResourceWidget)
NOVEMBER 26, 2020

...uld I become a data scientist (or a business analyst)?
...ps://www.analyticsvidhya.com/blog/2020/11/become-data-...ntist-business-analyst/?
...n_source=Blog&utm_medium=CareerResourceWidget)
...EMBER 24, 2020

How to create your own Question and Answering API(Flask+Docker +BERT) using haystack framework (https://www.analyticsvidhya.com/blog/2021/01/how-to-create-your-own-question-and-answering-apiflaskdocker-bert-using-haystack-framework/)

JANUARY 29, 2021

A Look at Machine Learning System Design (https://www.analyticsvidhya.com/blog/2021/01/a-look-at-machine-learning-system-design/)

JANUARY 29, 2021

Kaggle Grandmaster Series – Exclusive Interview with Kaggle Discussion Grandmaster Raju Kumar Mishra (#Rank 48) (https://www.analyticsvidhya.com/blog/2021/01/kaggle-grandmaster-series-exclusive-interview-with-kaggle-discussion-grandmaster-raju-kumar-mishra-rank-48/)

JANUARY 28, 2021

A Beginner's Guide Bayesian Inference (https://www.analyticsvidhya.com/blog/2021/01/a-beginners-guide-bayesian-inference/)

Download App (https://play.google.com/store/apps/details?id=com.analyticsvidhya.android) (https://apps.apple.com/us/app/analytics-vidhya/id1470025572)

Analytics Vidhya

About us (https://www.analyticsvidhya.com/about-me/)

Our Team (https://www.analyticsvidhya.com/about-me/team/)

Careers (https://www.analyticsvidhya.com/about-me/career-analytics-vidhya/)

Contact us (https://www.analyticsvidhya.com/contact/)

Data Science

Blog (https://www.analyticsvidhya.com/blog/)

Hackathon (https://datahack.analyticsvidhya.com/)

Discussions (https://discuss.analyticsvidhya.com/)

Apply Jobs (https://www.analyticsvidhya.com/jobs/)

Companies

Post Jobs (https://www.analyticsvidhya.com/corporate/)

Trainings (https://courses.analyticsvidhya.com/)

Hiring Hackathons (https://datahack.analyticsvidhya.com/)

Advertising (https://www.analyticsvidhya.com/contact/)

Visit us

(https://www.linkedin.com/company/analytics-vidhya/) (https://www.facebook.com/AnalyticsVidhya/) (https://www.youtube.com/channel/UCH6gDteHtH4hg3o2343iObA)

Privacy Policy Terms of Use Refund Policy