ASRC FEDERAL

Continuous evaluation of machine learning models deployed in production

Borislav Karaivanov | February 2023

Agenda

1. Creation of an ML model.
2. Are we done?
3. Data drift
4. Example
5. Is maintenance costly?
6. Conclusion

Creation of an ML model

What does it take to create an ML model?

- **Creating an ML model**
  - Extract data (gather large amount of data from one or more databases, or data lakes, or scrape the web, etc.)
  - Transform data (clean, augment, normalize, etc.)
  - Load data (store transformed, ready-to-train data in a database)
  - Split data into train, validation, and test datasets
  - Design the architecture of ML model(s)
  - Train ML model for many iterations, validate after each step
  - Fine-tune hyperparameters until convergence with great validation accuracy
  - Verify the ML model is highly accurate on the test dataset as well
  - Deploy the best ML model to production
  - Job well done? Ready to start the next project?

Are we done?

Are we done? Not yet.

- **Data drift may cause ML model to "decay"**
  - Model decay
    - ML model stays exactly the same
    - its relevance diminish
  - Production data space may drift away from training data space
  - Data space may
    - Stay same – ideal, nothing to do
    - Grow – retrain, underfitting, architecture or hyperparameter changes
    - Shift – retrain
    - Shrink – no need to do anything
    - Grow and shift – retrain, underfitting, architecture or hyperparameter changes
    - Shrink and shift – retrain, overfitting

Data drift

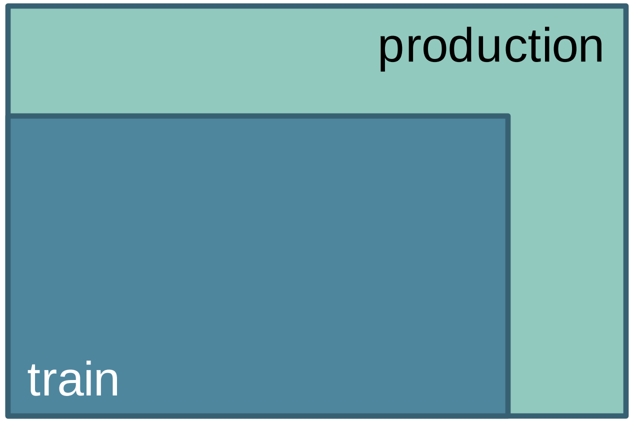
Stay same

- **Data space stayed the same**
  - All type of training data occurs in production
  - All type of production data has been used in training

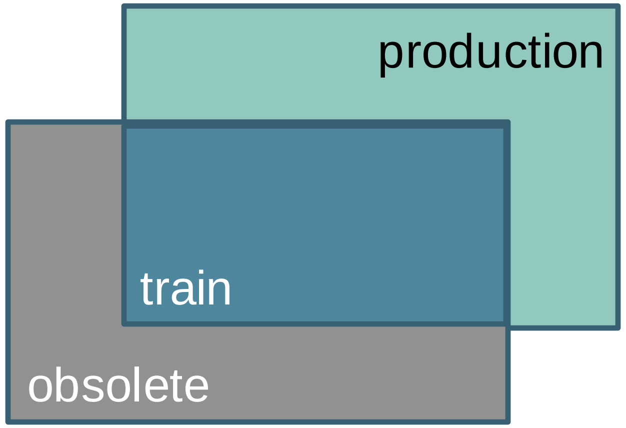
- **ML engineer needs to**
  - Sit back and relax
  - (Rarely the case)

Grow

- **Data space grew**
  - All type of training data occurs in production
  - Some type of production data has not been used in training
- **ML engineer needs to**
  - Expand training dataset with production data previously not seen
  - Modify architecture and/or hyperparameters of ML model
    - ML model was optimally selected for old training dataset. It will not have enough “neurons” to learn new, larger training dataset and may underfit
    - New ML model may require different hyperparameters to achieve target accuracy
  - Retrain a new ML model with new training data



Shift

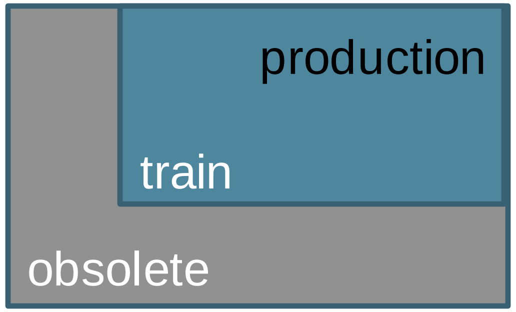


- **Data space shifted**
  - Some type of training data occurs in production
  - Some type of production data has not been used in training
  - Some type of training data no longer occurs in production
- **ML engineer needs to**
  - Expand training dataset with production data previously not seen
  - Reduce training dataset by removing obsolete training data
  - Retrain a new ML model with new training data

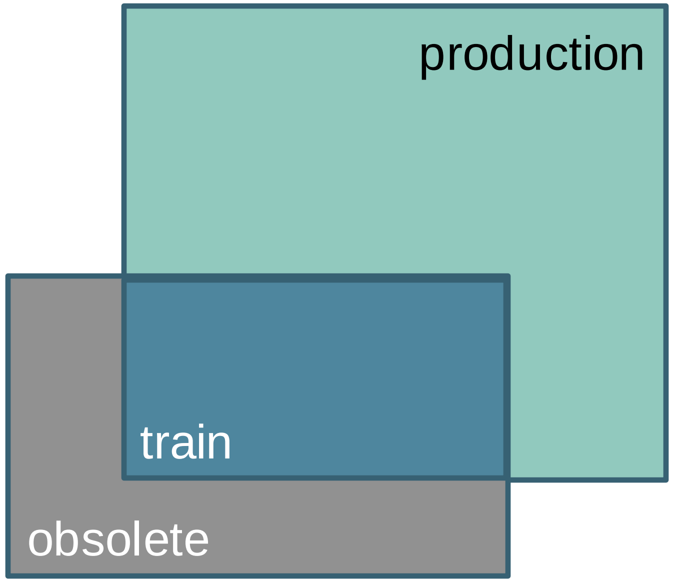
Shrink

- **Data space shrunk**
  - Some type of training data occurs in production
  - All type of production data has not been used in training
  - Some type of training data no longer occurs in production

- **ML engineer needs to**
  - No need to do anything
  - Anomaly detector still recognizes all production data as nominal
  - Smaller ML model can be designed and trained to achieve same accuracy with smaller computational cost

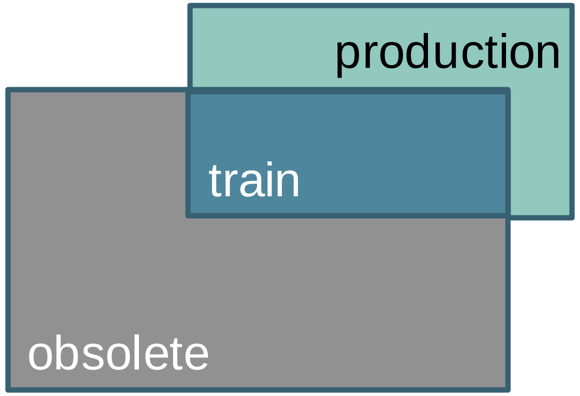


Grow and Shift



- **Data space grew and shifted**
  - Some type of training data occurs in production
  - A lot of production data has not been used in training
  - Some type of training data no longer occurs in production
- **ML engineer needs to**
  - Expand training dataset with production data previously not seen
  - Reduce training dataset by removing obsolete training data
  - Modify architecture and/or hyperparameters of ML model
    - ML model was optimally selected for old training dataset. It will not have enough “neurons” to learn a larger training dataset and may underfit
    - New ML model may require different hyperparameters to achieve target accuracy
  - Retrain a new ML model with new training data

Shrink and Shift



- **Data space shrunk and shifted**
  - Some type of training data occurs in production
  - Some type of production data has not been used in training
  - A lot of training data no longer occurs in production
- **ML engineer needs to**
  - Expand training dataset with production data previously not seen
  - Reduce training dataset by removing obsolete training data
  - Modify architecture and/or hyperparameters of ML model
    - ML model was optimally selected for old training dataset. It will have too many “neurons” to learn a smaller training dataset and may overfit
    - New ML model may require different hyperparameters to achieve target accuracy
  - Retrain a new ML model with new training data

Maintenance needed

- **In most cases data drift necessitates intervention for maintenance**
- **Modification of training dataset is based on domain knowledge of Operations engineers**
  - Domain and problem specific
  - User interface can help transfer domain knowledge from human to automated ETL routines
- **Modification of ML model architecture and hyperparameters cannot be easily automated**
  - Somewhat of an art
  - Results depend on experience and skills of ML engineers
  - Tools for automated model search perform inconsistently and generate ML models that are typically outperformed by those built by a good ML engineer

Example

Example

- **I trained a ML model for anomaly detection during North-South Station Keeping (NSSK) maneuvers**
  - monitored about 100 time series relevant to NSSK maneuvers
  - GOES-16 data from 2018, 2019, and first half of 2020
  - maneuvers were labeled as nominal or anomalous
  - training and validation datasets had only nominal maneuvers
  - test dataset had both nominal and anomalous maneuvers
  - ML model had 99% accuracy on test dataset

- **Later same ML model started reporting over 50% of maneuvers labeled nominal as anomalous**
  - GOES-16 data from second half of 2020 and 2021
  - maneuvers were labeled as nominal or anomalous by same Operations engineers

ASRC FEDERAL

Example

- **What has happened toward the end of 2020?**
  - Operations engineers helped me understand
  - one of the arcjet configurations traditionally used for NSSK maneuvers was dropped
  - a new arcjet configuration was introduced
  - samples of time series fed into anomaly detector have changed
  - all maneuvers with the new arcject configuration were seen as anomalous

- **Data drift was significant but well-understood**
  - data space has shifted; it neither grew nor shrunk
  - compiled a new training dataset
  - did not have to change architecture of ML model
  - needed to do hyperparameter tuning (weights in loss function)
  - trained a new ML model with similar test accuracy

Is maintenance costly?

Is maintenance costly?

- **Is maintenance costly?**
  - no, if properly planned
  - but it requires experienced ML engineer
  - not trivial to make it fully automated in most cases

- **What can be reused?**
  - entire ETL pipeline is coded during development; for maintenance it is used as is
  - class of ML model architectures that work well for problem at hand is discovered during development
  - training pipeline is coded during development and can be reused with minor adjustments
  - test pipeline can be reused likewise

ASRC FEDERAL

Is maintenance costly?

- **What needs attentions?**
  - training dataset needs to be modified
    - programmatically if data drift is well understood
    - manually if data needs to be labeled
  - data drift is addressed with minor changes to ML model architecture
  - hyperparameters tuning is needed and require experienced ML engineer to make changes based on training outcomes
  - trained a new ML model with similar test accuracy

Tools for cooperation

- **Operations engineers need tools to collaborate with ML engineers**
  - convenient to use
  - minimal additional effort required
  - information is programmatically accessible
- **Provided them with intuitive graphical user interface (GUI) to**
  - visually analyze time series identified as most anomalous by our detector
  - report each alarm as true positive or false positive
  - keep record of other observations they find relevant
  - manually approve or reject automatically generated suggestion that ML model needs maintenance

Conclusion

In conclusion

- **ML model for anomaly detection is not a once-and-for-all solution**
- **ML model requires continuous maintenance to avoid decay**
- **Data drift shows up as unusually high number of anomalies**
- **But only domain expert can tell if data drift is due**
  - intended change in operation routines, or
  - unexpected change in satellite hardware or surroundings
- **Maintenance includes continuous evaluation with help from domain experts**

In conclusion

- **Operations engineers can**
  - verify for data drift
  - understand nature of data drift
  - identify
    - obsolete training data
    - new training data

- **ML engineers can**
  - extract, transform, and load a new training data set
  - modify architecture and fine-tune hyperparameters
  - retrain ML model

- **Symbiotic cooperation between ML and Operations engineers is key to maintaining highly accurate ML models for anomaly detection** anaq!

Borislav Karaivanov

Sr Prin Software Engineer | ML & AI

e: **bkaraivanov@asrcfederal.com**