Testing ML Systems: Code, Data and Models

Machine Learning Architects Basel

Bassem Ben Hamed

December 2022

ML ARCHITECTS
BASEL

Agenda

- Types of Tests
- Unit Testing
- Data and Model Testing

Types of Tests

How to test ML artifacts (Code, Data and Models) to ensure a reliable ML system?

There are three majors types of tests which are utilized at different points in the development cycle:

1. Unit tests: tests on individual components that each have single responsability (ex. function that filters a list)
2. Integration tests: tests on the combined functionality of teh individual components (ex. data proprocessing)
3. Regression tests: tests based on errors we've seen before to ensure new changes don't reintroduce them

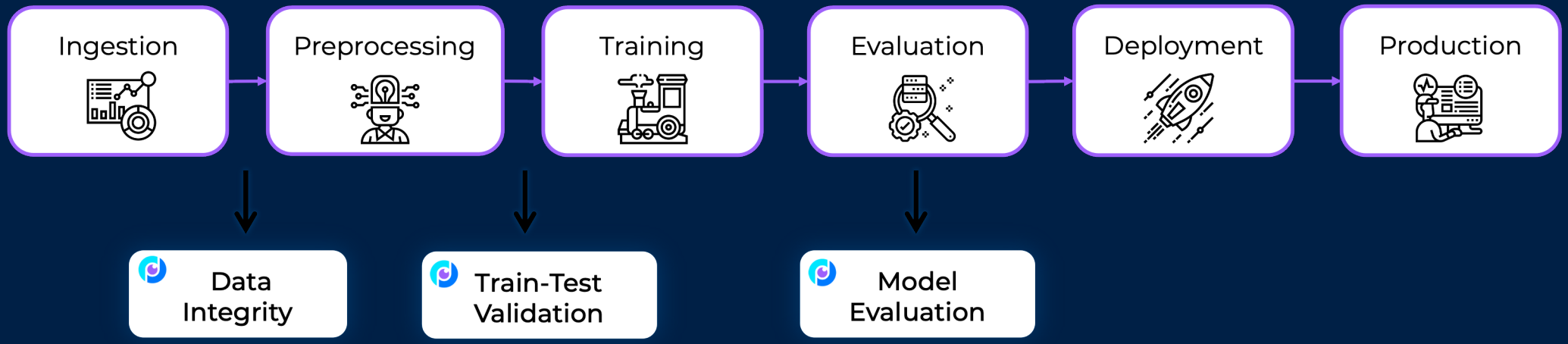
Unit Testing

It aims to test each component of the system in isolation (for example, test the functionality of a single class or function). In our case we used Pytest, a python library in order to test different components

Source : https://docs.pytest.org/en/7.2.x/

Data and Model Testing

One of the best existing for testing ML systems is Deepchecks. Deepcheks is a python tool which aims to build test suites for validating data and ML models. Deepchecks accompanies you through various validation and testing needs such as verifying the data's integrity, inspecting its distributions, validating data splits, evaluating your model and comparing different models.



Where we can use Deepchecks?

Source : https://docs.deepchecks.com/stable/getting-started/welcome.html

Deepchecks: Which Types of Checks Exists?

They are checks for different phases in the ML workflow:

- Data Integrity
- Train-Test Validation (Distribution, Drift and Methodology Checks)
- Model Performance Evaluation

Conditions Summary

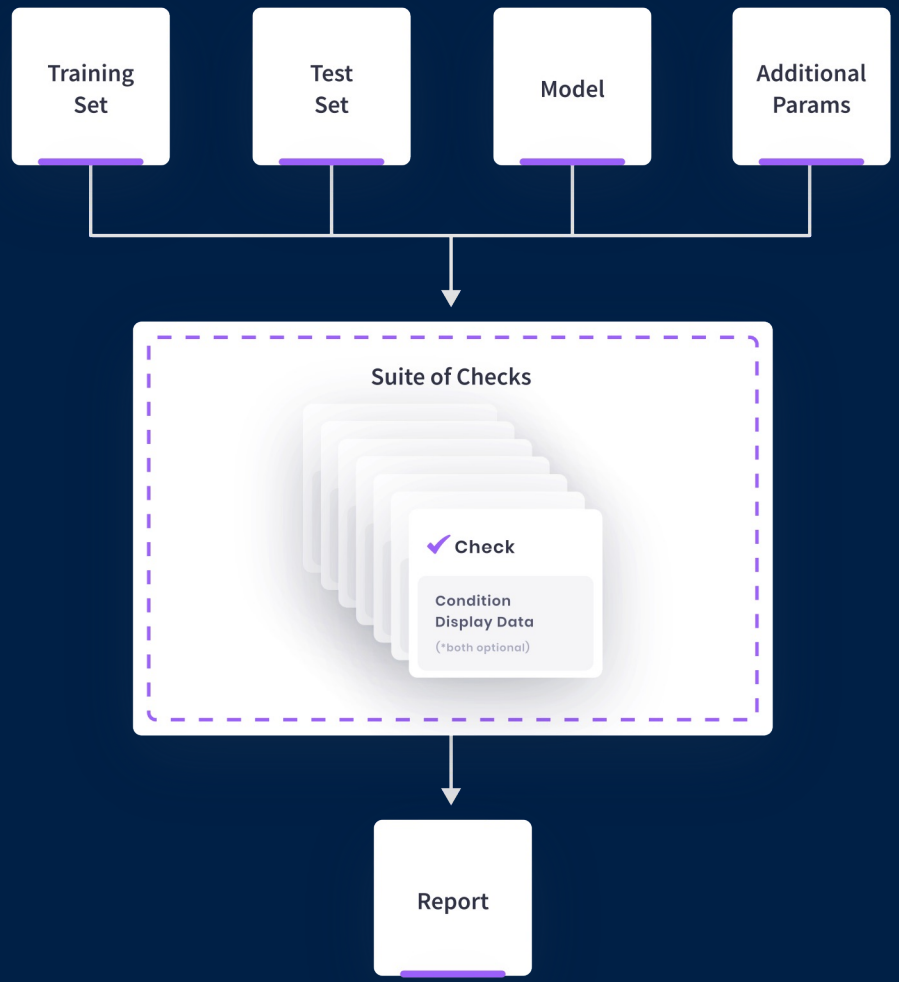
Status	Check	Condition	More Info
✖	Performance Report	Train-Test scores relative degradation is not greater than 0.1	Precision for class 1 (train=1 test=0.87) Recall for class 2 (train=1 test=0.83)
✖	Single Feature Contribution Train-Test	Train features' Predictive Power Score (PPS) is not greater than 0.7	Features in train dataset with PPS above threshold: petal width (cm), petal length (cm)
!	Model Error Analysis	The performance of the detected segments must not differ by more than 5.00%	Change in Accuracy in features: petal length (cm), petal width (cm) exceeds threshold.
✓	ROC Report - Test Dataset	Not less than 0.7 AUC score for all the classes	
✓	ROC Report - Train Dataset	Not less than 0.7 AUC score for all the classes	
✓	Single Feature Contribution Train-Test	Train-Test features' Predictive Power Score (PPS) difference is not greater than 0.2	
✓	Datasets Size Comparison	Test-Train size ratio is not smaller than 0.01	
✓	Whole Dataset Drift	Drift value is not greater than 0.25	
✓	Train Test Label Drift	PSI and Earth Mover's Distance for label drift cannot be greater than 0.2 or 0.1 respectively	

Source : https://www.reddit.com/r/MachineLearning/comments/rxczzl/p_deepchecks_an_opensource_tool_for_high/

Deepchecks: What Do You Need in Order to Start

Depending on your phase and what you wish to validate, you'll need a subset of the following:

- Raw data (before pre-processing such as OHE, string processing, etc.), with optional labels
- The model's training data with labels
- Test data (which the model isn't exposed to) with labels
- A supported model (e.g. scikit-learn models, XGBoost, any model implementing the predict method in the required format)



Source : https://github.com/deepchecks/deepchecks

References

https://deepchecks.com/how-to-test-machine-learning-models/

https://blog.testproject.io/2022/01/17/machine-learning-testing-for-beginners-the-all-in-one-guide/

https://www.jeremyjordan.me/testing-ml/

Implementing reliable
machine learning solutions

ML ARCHITECTS
BASEL

Operating Models – Technologies – Culture & Skills

Consulting – Engineering - Training