THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
EUROPE

Uncovering Project and Community Insights Using Data Driven Methods

Oindrilla Chatterjee
Senior Data Scientist
Red Hat

SCHEDULE

1. Metrics
2. Project AI4CI
3. Operate First Community Cloud
4. AI tools for open source projects
5. Time to Merge Model

Let’s talk metrics

Why?

- Better allocate resources
- Evaluate a project's success
- Advocate for the project
- Analyze the development community
- Growth and journey of community

How?

- Sources like repositories, communication channels, website, service endpoints
- Open end-to-end workflows, from collection to deployment

ML to aid project development?

- Predicting time-to-merge to get an idea of "estimated effort"
- Predicting time to respond on issues
- Optimal stopping point for long running tests

Not here to tell you

Which metrics to track

Why use metrics to support your open source community

Thursday, September 15 • 11:00 - 11:40

Community Data: What to Measure and Why? - Cali Dolfi, Red Hat

Tuesday, September 13 • 09:55 - 10:35

Dev Team Metrics that Matter - Avishag Sahar, LinearB

Here to tell you

How AI driven metrics can help

How to use open source ML tools to achieve this

How this can be done on an open community cloud

AI4CI: Open Source AIOps toolkit

https://github.com/aicoe-aiops/ocp-ci-analysis

Problem

- Need for **AIOps** - Automated monitoring, analysis, alerting with Ops (CI/CD, development processes)
- **Open Source data** originating from real world production systems is a rarity for public datasets.
- Lack of AI driven metrics for open source community health.

Opportunity

- **Open operations data** made available by running open source software and applications in production.
- Data includes CI/CD data, code, telemetry, logs, operational dashboards.
- Eg: Kubernetes testing infrastructure, Fedora make their testing data available open source.

Solution

- Collection of intelligent and open source **data science tools** to collect and analyze the CI/CD data.
- **AIOps models** like Github time-to-merge service, optimal stopping time prediction, build log classifier
- KPI and Metric dashboards
- Goal is to foster an open source AIOps community with open ops data, AI tools and services.

AI4CI supports CI/CD and software dev processes

What is AI4CI?

Collection of **Open Source AIOps tools** including scripts, notebooks, pipelines, dashboards and data sources.

Data collection

Collection of open operations data from Kubernetes testing platforms eg: Testgrid, Github, and Prow.

Metrics

Collects metrics and **KPIs** and visualization dashboards.

ML Services

ML services which can support CI/CD processes.

Open source AIOps template

Resource for open source AIOps communities (notebooks, scripts, automated ML pipelines, dashboards, services tools)

The Operate First Community Cloud

- Operate First makes **operations open source**.
- An initiative centered around learning and developing code and practices in an open **production community cloud**.
- Deploy and maintain apps in an open environment leading to **open operations data** which include logs, issues,metrics.

www.operate-first.cloud/

Open Source data science and engineering tools on Operate First

OPEN DATA HUB

OPERATE FIRST

Show All

JupyterHub

A multi-user version of the notebook designed for companies, classrooms and research labs

Argo

Kubernetes native workflows, events, CI and CD

Superset

A modern, enterprise-ready business intelligence web application

Prometheus

Systems monitoring and alerting toolkit

Grafana

Visualization and analytics software

Spark

Unified analytics engine for large-scale data processing

Seldon

Platform for rapidly deploying machine learning models on Kubernetes.

Kafka

Distributed event streaming platform

Airflow

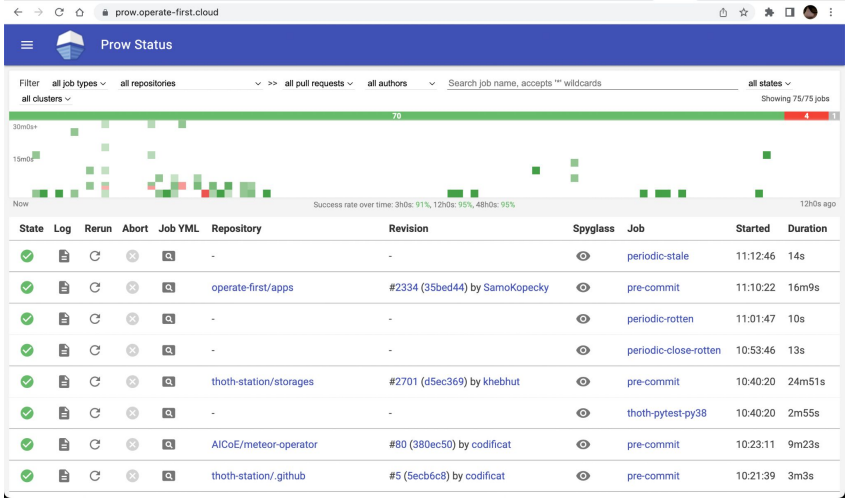
Platform to programmatically author, schedule, and monitor workflows

Hue

Data exploration platform for Hive and S3 storage

Open Source data science and engineering tools on Operate First

- Operations Data like: Logs, Metrics, Github issues, PRs, architectural decisions, blueprints.
- These open operations datasets can help **enable AI tools** to assist with cloud operations.



Tooling for capturing metrics

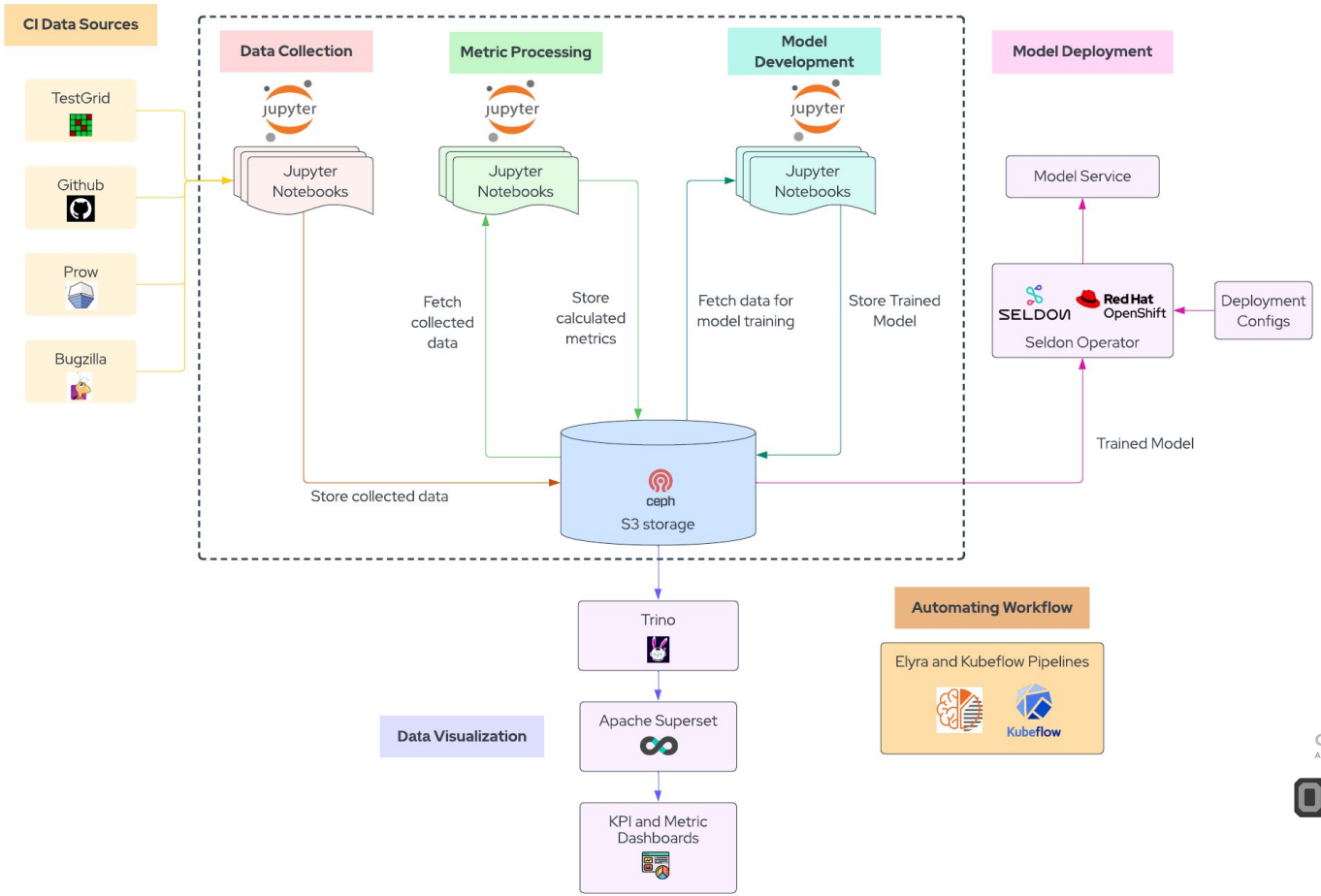
Code Repository

Github API,
Thoth MI-Scheduler,
CHAOSS-Augur

Feature exploration
& Engineering

Dashboard with key
metrics

AI4CI Architecture

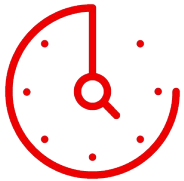


Running on

OPEN DATA HUB
AI Platform powered by Open Source

OPERATE FIRST

ML Service: Optimal Stopping Point Prediction

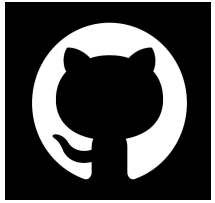


Sometimes tests/builds take longer than expected to run

Find an Optimal Stopping Point after which the test will fail.

We can better allocate and save resources.

ML Service: Time to Merge Prediction



A new PR is opened on a Git Repository

Predict the time that it will take to merge this open PR, and add a label on the PR

Learn from the actual time to merge and create a feedback loop

Time to merge prediction service for community health

- Identify **bottlenecks** in development process
- Leverage the rich **historical data** of consisting of Issues, Commits, PRs
- Give **new contributors** of an estimate of when their PR will be reacted upon
- Most importantly - develop an **AI driven mindset** for community health

Current workflow: Github time to merge prediction service

Collect data → **Features** → **Model Building** → **Creating Service** → **Bot integration**

Collect **data** from code repositories

Features like **size of the PR, types of files in the PR description of the PR**

Train repository or org specific model

Kubeflow based Model Pipelines

Creating **Model service** endpoint

Provide endpoint that can be accessed

Integrate service with repository using a **bot** to offer live feedback or a standalone site

Issues, PRs → **Discover project specific behavior** → **Model returns merge time estimate** → **Model API Endpoint** → **Provide estimates**

ML Service: Github Time to Merge Model

COLLECT DATA

openshift / origin

srcopsmetrics 2.11.1

```
pip install srcopsmetrics
```

ML Service: Github Time to Merge Model

GITHUB PR DATA

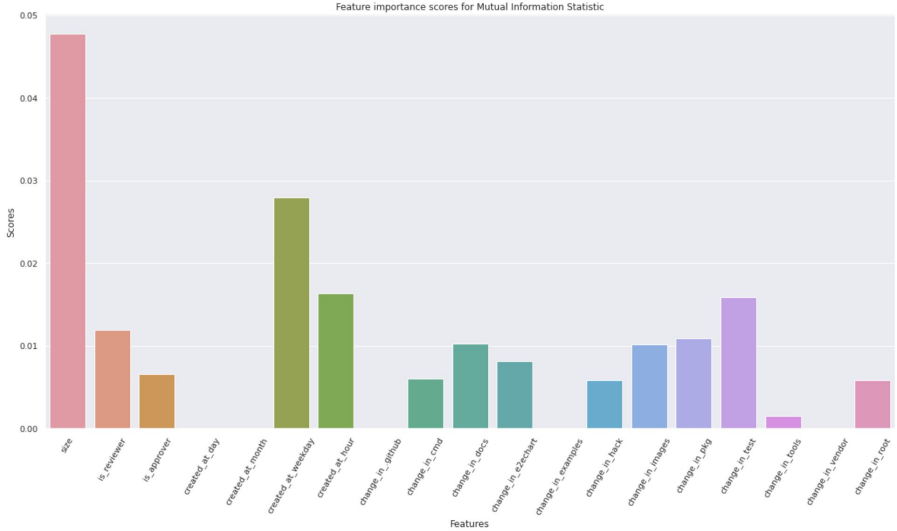
openshift / origin

FEATURE ENGINEERING

size of the PR,

types of files in the PR

description of the PR



ML Service: Github Time to Merge Model

GITHUB PR DATA → FEATURE ENGINEERING → MODEL TRAINING

openshift / origin

size of the PR,

types of files in the PR

description of the PR

Naive Bayes

XGBoost

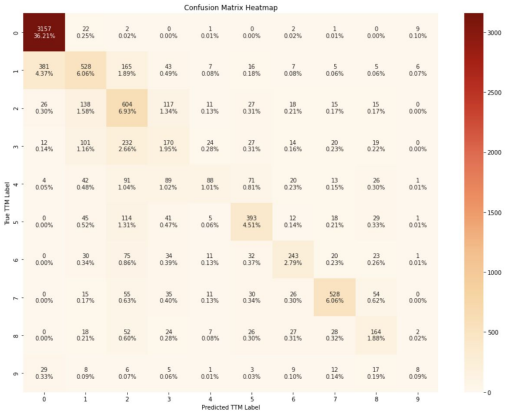
Random Forest

SVM

XG Boost

```
train_evaluate_classifier(xgbc, X_train_scaled, y_train, X_test_scaled, y_test)
```

	precision	recall	f1-score	support
0	0.87	0.99	0.93	3194
1	0.56	0.45	0.50	1163
2	0.43	0.62	0.51	971
3	0.30	0.27	0.29	619
4	0.53	0.20	0.29	445
5	0.63	0.60	0.61	658
6	0.64	0.52	0.57	469
7	0.88	0.70	0.75	754
8	0.47	0.47	0.47	348
9	0.29	0.08	0.13	98
accuracy			0.67	8719
macro avg	0.55	0.49	0.50	8719
weighted avg	0.66	0.67	0.66	8719



ML Service: Github Time to Merge Model

GITHUB PR DATA → FEATURE ENGINEERING → MODEL TRAINING → MODEL DEPLOYMENT

openshift / origin

size of the PR,

types of files in the PR

description of the PR

Naive Bayes

XGBoost

Random Forest

SVM

SELDON

RED HAT OPENSHIFT Container Platform

ML Service: Github Time to Merge Model

Services > Service details

S github-pr-ttm-seldon-github-pr-ttm-predictor

Managed by SD github-pr-ttm-seldon

Details YAML Pods

Service details

Name

github-pr-ttm-seldon-github-pr-ttm-predictor

Namespace

NS ds-ml-workflows-ws

Labels Edit

app.kubernetes.io/managed-by=seldon-core

seldon-app=github-pr-ttm-seldon-github-pr-ttm-predictor

seldon-deployment-id=github-pr-ttm-seldon

Pod selector

seldon-app=github-pr-ttm-seldon-github-pr-ttm-predictor

Annotations

1 annotation

MODEL DEPLOYMENT

SELDON

RED HAT® OPENSHIFT Container Platform

NEXT STEPS

- Service as a bot on Github PRs
- Live feedback from bot / service
- Iterate on time-to-merge models for better performance
- Toolified API for using the training service.

Resources

https://github.com/aicoe-aiops/ocp-ci-analysis/blob/master/docs/get-started.md

Open Data Sources	Notebooks	Dashboards	Model Endpoints	Automated Workflows	Video Playlist

https://tinyurl.com/aiforci

Thank you!

@oindrilla_chat

www.linkedin.com/in/oindrilla-chatterjee/