AI Explainability WG

June 2024

June 2024 updates

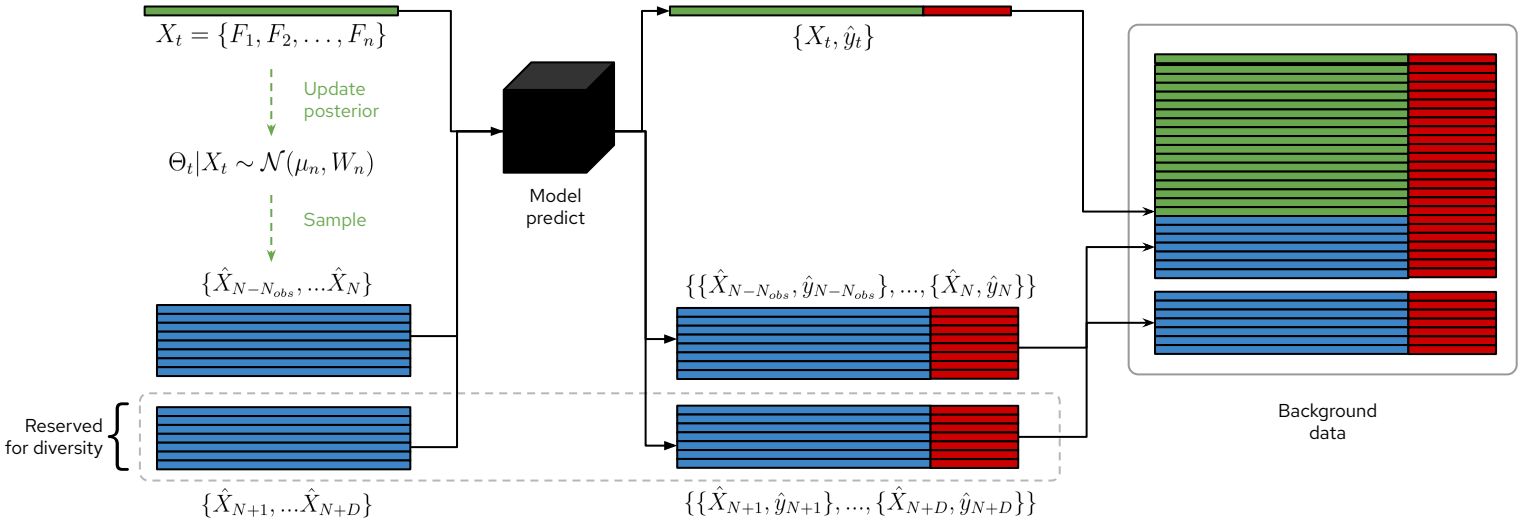
- TrustyAI core / service
  - 0.15.0 release
  - https://github.com/trustyai-explainability/trustyai-explainability/releases/tag/v0.15.0
  - quay.io/trustyai/trustyai-service:v0.15.0
- TrustyAI operator
  - 1.21.0 release
  - https://github.com/trustyai-explainability/trustyai-service-operator/releases/tag/v1.21.0
  - quay.io/trustyai/trustyai-service-operator:v1.21.0
- TrustyAI KServe explainer
  - 0.1.0 release
  - https://github.com/trustyai-explainability/trustyai-kserve-explainer/releases/tag/v0.1.0
  - quay.io/trustyai/trustyai-kserve-explainer:v0.1.0

What’s new?

TrustyAI - What's new?

- **TrustyAI core/service 0.15.0**
  - *Explainers*
    - Add online Gaussian estimation [#559]
    - Add SHAP streaming background generator [#561]
- **TrustyAI operator 1.21.0**
  - Update builder and base tags [#242]

TrustyAI - SHAP streaming support



Current work

TrustyAI - current work

- **TrustyAI core/service**
  - Initial support for image data in data drift
    - Jensen-Shannon divergence method
  - Support for database storage
    - Support for PVC- and DB-mode
  - HTTPS endpoint support
- **TrustyAI operator**
  - TLS-enabled Kubernetes `Service` for payload consumer
  - Operator support for PVC- and DB-mode
  - KServe readiness
- **Testing/CI**
  - Migrating CI to Python-based tests

TrustyAI - Database support

PVC-mode

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: TrustyAIService
metadata:
  name: trustyai-service
spec:
  storage:
       format: "PVC"
       folder: "/inputs"
       size: "1Gi"
  data:
       filename: "data.csv"
       format: "CSV"
  metrics:
       schedule: "5s"
```

DB-mode

```
apiVersion: trustyai.opendatahub.io/v1alpha1
kind: TrustyAIService
metadata:
  name: trustyai-service
spec:
  storage:
       format: "DATABASE"
       databaseConfigurations: db-credentials
  metrics:
       schedule: "5s"

—
apiVersion: v1
kind: Secret
metadata:
  name: db-credentials
type: Opaque
stringData:
  databaseKind: mysql
  databaseUsername: "foo"
  databasePassword: "bar"
  databaseService: "mariadb-service"
  databasePort: "3306"
  databaseGeneration: "update"
```

Roadmap

TrustyAI 2024 roadmap

- **December 2023 - March 2024 (proposal / discussion)**
  - Out-of-distribution (OOD) metrics completion
    - Finalise OOD metrics, namely data uploading aspect
      - Provide data connection for data upload
  - Explainability service endpoints completion
    - Formalise explainability payloads schema
    - Refining handling of synthetic payloads to avoid interference with metrics
    - Handle potential large computational times in a service setting
  - Detoxification at the library and service level
    - Integrate detoxification with Python TrustyAI (Jupyter main target)
    - Token scoring at the service level
  - Database backend
    - Address scalability
    - Replace PVC with DB?
  - Expand supported types (*eg* image data)
    - Metrics and explainability for non-tabular data
  - Model drift/data drift/anomaly detection
  - Improve handling of unsupported model serving runtimes

Legend

Not started

In progress

Completed

TrustyAI 2024 roadmap

- **June 2024 - August 2024 (proposal / discussion)**
  - KServe explainer integration
  - Detoxification fine-tuning
  - Saliency Explainers

Legend
Not started
In progress
Completed